Long-term Future Prediction under Uncertainty and Multi-modality

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Apratim Bhattacharyya, M.Sc.

Saarbrücken
2021

Day of Colloquium	24th of September, 2021
Dean of the Faculty	Prof. Dr. Thomas Schuster Saarland University, Germany

Examination Committee

Chair	Prof. Dr. Sebastian Hack
Reviewer, Advisor	Prof. Dr. Bernt Schiele
Reviewer, Advisor	Prof. Dr. Mario Fritz
Reviewer	Prof. Dr. Andreas Geiger
Reviewer	Prof. Dr. Sanja Fidler
Academic Assistant	Dr. Paul Swoboda

ABSTRACT

Humans have an innate ability to excel at activities that involve prediction of complex object dynamics such as predicting the possible trajectory of a billiard ball after it has been hit by the player or the prediction of motion of pedestrians while on the road. A key feature that enables humans to perform such tasks is *anticipation*. There has been continuous research in the area of Computer Vision and Artificial Intelligence to mimic this human ability for autonomous agents to succeed in the real world scenarios. Recent advances in the field of deep learning and the availability of large scale datasets has enabled the pursuit of fully autonomous agents with complex decision making abilities such as self-driving vehicles or robots. One of the main challenges encompassing the deployment of these agents in the real world is their ability to perform anticipation tasks with at least human level efficiency.

To advance the field of autonomous systems, particularly, self-driving agents, in this thesis, we focus on the task of future prediction in diverse real world settings, ranging from deterministic scenarios such as prediction of paths of balls on a billiard table to the predicting the future of non-deterministic street scenes. Specifically, we identify certain core challenges for long-term future prediction: long-term prediction, uncertainty, multi-modality, and exact inference.

To address these challenges, this thesis makes the following core contributions. Firstly, for accurate long-term predictions, we develop approaches that effectively utilize available observed information in the form of image boundaries in videos or interactions in street scenes. Secondly, as uncertainty increases into the future in case of non-deterministic scenarios, we leverage Bayesian inference frameworks to capture calibrated distributions of likely future events. Finally, to further improve performance in highly-multimodal non-deterministic scenarios such as street scenes, we develop deep generative models based on conditional variational autoencoders as well as normalizing flow based exact inference methods. Furthermore, we introduce a novel dataset with dense pedestrian-vehicle interactions to further aid the development of anticipation methods for autonomous driving applications in urban environments.

ZUSAMMENFASSUNG

Menschen haben die angeborene Fähigkeit, Vorgänge mit komplexer Objektdynamik vorauszusehen, wie z. B. die Vorhersage der möglichen Flugbahn einer Billardkugel, nachdem sie vom Spieler gestoßen wurde, oder die Vorhersage der Bewegung von Fußgängern auf der Straße. Eine Schlüsseleigenschaft, die es dem Menschen ermöglicht, solche Aufgaben zu erfüllen, ist die Antizipation. Im Bereich der Computer Vision und der Künstlichen Intelligenz wurde kontinuierlich daran geforscht, diese menschliche Fähigkeit nachzuahmen, damit autonome Agenten in der realen Welt erfolgreich sein können. Jüngste Fortschritte auf dem Gebiet des Deep Learning und die Verfügbarkeit großer Datensätze haben die Entwicklung vollständig autonomer Agenten mit komplexen Entscheidungsfähigkeiten wie selbstfahrende Fahrzeugen oder Roboter ermöglicht. Eine der größten Herausforderungen beim Einsatz dieser Agenten in der realen Welt ist ihre Fähigkeit, Antizipationsaufgaben mit einer Effizienz durchzuführen, die mindestens der menschlichen entspricht.

Um das Feld der autonomen Systeme, insbesondere der selbstfahrenden Agenten, voranzubringen, konzentrieren wir uns in dieser Arbeit auf die Aufgabe der Zukunftsvorhersage in verschiedenen realen Umgebungen, die von deterministischen Szenarien wie der Vorhersage der Bahnen von Kugeln auf einem Billardtisch bis zur Vorhersage der Zukunft von nicht-deterministischen Straßenszenen reichen. Insbesondere identifizieren wir bestimmte grundlegende Herausforderungen für langfristige Zukunftsvorhersagen: Langzeitvorhersage, Unsicherheit, Multimodalität und exakte Inferenz.

Um diese Herausforderungen anzugehen, leistet diese Arbeit die folgenden grundlegenden Beiträge. Erstens: Für genaue Langzeitvorhersagen entwickeln wir Ansätze, die verfügbare Beobachtungsinformationen in Form von Bildgrenzen in Videos oder Interaktionen in Straßenszenen effektiv nutzen. Zweitens: Da die Unsicherheit in der Zukunft bei nicht-deterministischen Szenarien zunimmt, nutzen wir Bayes'sche Inferenzverfahren, um kalibrierte Verteilungen wahrscheinlicher zukünftiger Ereignisse zu erfassen. Drittens: Um die Leistung in hochmultimodalen, nicht-deterministischen Szenarien wie Straßenszenen weiter zu verbessern, entwickeln wir tiefe generative Modelle, die sowohl auf konditionalen Variations-Autoencodern als auch auf normalisierenden fließenden exakten Inferenzmethoden basieren. Darüber hinaus stellen wir einen neuartigen Datensatz mit dichten Fußgänger-Fahrzeug-Interaktionen vor, um Antizipationsmethoden für autonome Fahranwendungen in urbanen Umgebungen weiter zu entwickeln.

ACKNOWLEDGEMENTS

First and foremost, I would like to begin by thanking my co-supervisors Prof. Dr. Bernt Schiele and Prof. Dr. Mario Fritz. They have been a constant source of inspiration during my time as a PhD student and have supported me through thick and thin. Most importantly, they have helped me become a better Computer Vision researcher. They have taught me how to conceptualize an idea and how to think critically about computer vision research. Likewise, I would like to thank Prof. Dr. Sanja Fidler and Prof. Dr. Andreas Geiger for serving as external reviewers as part of the thesis committee.

I would also like to thank all members of the Computer Vision and Machine Learning group at the Max Planck Institute for Informatics, Saarbrucken. The fruitful discussions especially during seminars and retreats were especially helpful in improving the quality of my research and by extension this thesis. I would like to thank Mateusz Malinowski for helping me get acquainted with research in the field of deep learning at the start of my PhD. Outside of work, I would also like to thank the members of the Computer Vision and Machine Learning group – Eldar Insafutdinov, Jan-Hendrik Lange, Evgeny Levinkov, Seong Joon Oh, Rakshith Shetty, Julian Steil, Yongqin Xian – for the fun filled activities especially during group retreats. I would also like thank my office mate Tribhuvanesh Orekondy for our many long discussions.

Moreover, I would like to thank my collaborators who helped tremendously in many aspects of this thesis – Shweta Mahajan, Dr. Christoph-Nikolas Straehle, Dr. Daniel Olmeda Reino and Dr. Stefan Roth. I am grateful to Christoph for supervising my internship at Bosch Center for AI and to Daniel for making my collaboration with Toyota Motor Europe possible.

Finally, I would like to thank my family and friends who always encouraged my journey as a computer vision researcher and stood by me through thick and thin. This thesis would not have been possible without the support of my parents who always encouraged me to pursue a career in science. I would like to especially thank Shweta for supporting me throughout my time as a PhD student. Lastly, I would also like to thank my friends – Marko, Kireeti, Ezekiel, Giorgi and Francesco for always supporting me and making my time as a PhD student a memorable experience.

CONTENTS

INTRODUCTION 1

Contents

"Prediction is very difficult, especially if it's about the future!"
– Niels Bohr

One of the key challenges facing artificial intelligence is the development of autonomous agents that can operate successfully in real world scenarios. In this context, an agent is a system that has the ability to function independently in order to achieve certain (pre-defined) goals, with at least human level performance. Particularly, an autonomous agent exhibits decision making abilities based on input signals from sensors, e.g. camera or lidar, so as to successfully accomplish a certain task (Maes, 1993). Potential applications of autonomous agents include home-assistant robots to perform day-to-day tasks such as cleaning in a domestic setting or autonomous vehicles to assist human drivers so as to eventually improve road safety. While considerable progress has been made to improve the decision making abilities of the aforementioned autonomous agents, yet these agents are limited to constrained environments and settings.

For example, current state of the art self-driving vehicles are still limited to very specific geographic locations or traffic conditions (Janai *et al.*, 2020; LeBeau, 2018). Specifically, self-driving vehicles are largely limited to highway environments where interactions with other traffic participants, such as pedestrians, are sparse. Therefore, the deployment of self-driving vehicles which maintain adequate safety distance in dense urban environments (Sauer *et al.*, 2018; Prakash *et al.*, 2020) with multi-agent interactions, e.g. pedestrians and bicyclists becomes challenging. Analogously, although there has been recent progress on robots designed for specific cases, e.g.

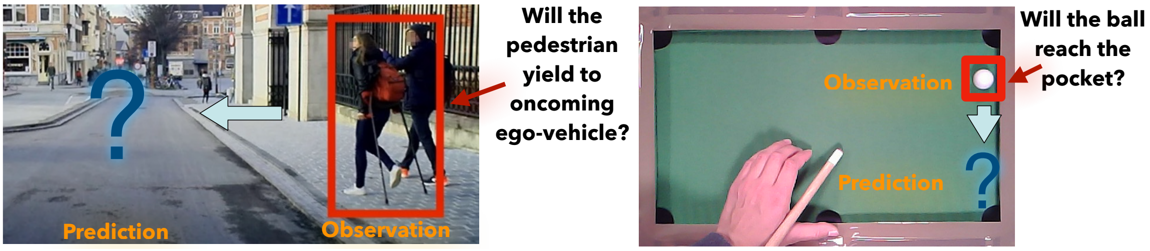

?
Prediction
Observation
Will the pedestrian yield to oncoming ego-vehicle?
Observation
Prediction
?
Will the ball reach the pocket?

Figure 1.1: Here we show routine tasks where anticipation is crucial. Left: A driver needs to anticipate whether the pedestrian would yield or step onto the path of the oncoming vehicle to avoid a collision. Right: A billiards player needs to anticipate the trajectory of the ball to score.

search and rescue (Raibert *et al.*, 2008) or drone racing (Madaan *et al.*, 2019), assistive robots widely deployable to home environments are still not available. Two important areas where the performance of home robots is limited are: navigation and object manipulation. For navigation in unknown environments robots must be able to avoid obstacles such as humans or other autonomous agents. For manipulation tasks, the robot must be able to understand the effect of its actions, e.g. how much force should be applied to move an object without damage.

Humans, in contrast, excel at the same tasks – driving in inner city environments, navigating in dense crowds or manipulating objects. A key component that enables the success of humans at these tasks is *anticipation*. While driving, anticipation allows us to maintain adequate safety distances and to avoid obstacles in order to prevent collisions. In Fig. 1.1 (left), we show a case where anticipation is crucial for making a decision while driving: to avoid collision it is essential to anticipate whether the pedestrian would yield or step onto the path of the oncoming vehicle so as to brake in time. Similarly, humans are good at manipulation tasks. In Fig. 1.1 (right), humans can anticipate the path of balls on the billiard table and are thus able to aim the strike such that the ball reaches the pocket. Further examples include sports such as soccer, where the goalkeeper can anticipate the motion of the players and of the ball to prevent goals from being scored.

In fact, recent work has shown that humans develop the ability to anticipate actions of other people and complex object dynamics from an early age. Green *et al.* (2014); Elsner *et al.* (2012) shows that even infants can anticipate the goal states and movements associated with human actions. Further, humans can anticipate the effect of social interactions. This explains why humans are good at anticipating the movements of pedestrians while driving (Fig. 1.1, left) even in urban environments with dense interactions. Similar results have been shown for complex object dynamics. Green *et al.* (2014) has shown that infants can already extrapolate the motion of objects such as moving balls. Further, Hamrick *et al.* (2011) shows that humans have comparable performance on prediction of complex object dynamics to that of physics simulation based oracles. This makes humans not only excel at routine tasks such as driving, object manipulation or playing sports but also helps humans generalize to new and unseen tasks and situations.

For autonomous agents to succeed at real world tasks such as autonomous driving (Mueller *et al.*, 2020) or to provide assistance in the household (Pirhonen *et al.*, 2020), abilities similar to or even exceeding humans in anticipating future events becomes particularly important. Therefore, in this thesis we focus on the task of anticipation – predicting future states – under diverse scenarios. These scenarios, which are important for the success of autonomous agents, can be divided into two (sometimes overlapping) categories. They are: 1. Deterministic scenarios which are governed by a set of deterministic rules, e.g. the laws of physics, and 2. Non-deterministic or agent-based scenarios whose future states are inherently ambiguous. Deterministic scenarios include, e.g. anticipation of future states of billiard tables (Fragkiadaki *et al.*, 2016), stability of object arrangements (Li *et al.*, 2017), dynamics of fluids (Bates *et al.*, 2019), among many others (Chuang *et al.*, 2018). These scenarios are important for robotic grasping and manipulation tasks. Non-deterministic scenarios include, e.g. traffic scenes, where the anticipation of paths of agents such as pedestrians or vehicles (Alahi *et al.*, 2016; Gupta *et al.*, 2018; Lee *et al.*, 2017b; Yu *et al.*, 2020b) are important for self-driving applications. In this thesis, owing to their importance and the potential for wide applicability, we focus on both of these scenarios.

Anticipating or predicting future states, even a few seconds into the future is very challenging (also attested by Niels Bohr in the opening quote of this chapter). The difficulty in prediction increases with increasing prediction horizons into the future. With the following key ingredients, we successfully address the challenges associated with long-term future prediction in this thesis: 1. Effectively utilizing observed information. 2. Reasoning over uncertain and multi-modal futures. 3. Inferring and optimizing the exact likelihood of future events under our models. Finally, in this thesis we show accurate predictions over long-time horizons in diverse scenarios including, 1. Predictions upto $\sim$1 second into the future on complex billiard ball scenarios with multiple balls, using only raw visual data. 2. Predictions of future pedestrian trajectories upto $\sim$4 seconds into the future in complex multi-modal street scenes with dense interactions between the traffic participants, including predictions in an "on-board" setting. 3. Predictions of full street scenes upto $\sim$1 second into the future, in similar complex multi-modal settings. 4. Reliable and calibrated uncertainty estimates to aid the decision making process of autonomous agents.

We now provide a comprehensive discussion of these challenges and include a brief overview of the contributions of the thesis which aims to improve long-term future prediction for autonomous agents (Table 1.1).

1.1 CHALLENGES OF FUTURE PREDICTION

Here, we discuss some of the challenges involved in anticipation or future prediction tasks, which can be broadly categorized into, 1. Long-term predictions: effective use of available information to maximize the accuracy of predictions in the long-term, 2. Uncertainty: capturing the distribution of likely futures in non-deterministic scenarios and ensuring that the uncertainty of the predictive distribution is calibrated, 3. Multi-modality: capturing the modes of the multi-modal distribution of likely futures, 4. Exact inference: inferring the exact likelihoods of future states under the model in order to maximize accuracy. Next, we discuss these challenges in more detail and also provide a brief overview of how this thesis addresses these challenges in Table 1.1.

1.1.1 Long-term Predictions

We divide the discussion on challenges of long-term predictions into two parts based on the type of scenario, deterministic or non-deterministic.

Long-term predictions in deterministic scenarios, such as billiard tables, is possible if the applicable deterministic rules (the laws of physics in case of billiard tables) and all associated physical quantities, e.g. the speed of the balls and coefficient of friction of the table are known apriori. However, accurately estimating the set of applicable deterministic rules and the corresponding physical quantities of interest is challenging. While it is definitely possible to hand-craft systems based on domain knowledge such systems would have to be tailor made for a particular scenario and would not generalize. Therefore, a recent group of work (Fragkiadaki *et al.*, 2016; Lerer *et al.*, 2016; Li *et al.*, 2017; Battaglia *et al.*, 2016; Watters *et al.*, 2017) have focused on implicitly learning these rules and associated physical quantities directly from the data without explicit human supervision. This enables such methods to be broadly applicable to diverse scenarios. In this thesis, we focus on such flexible methods to address the challenge of long-term predictions in Chapter 3 (Table 1.1).

In case of non-deterministic scenarios, it is still crucial to effectively integrate available information to maximize accuracy in the long-term (Yu *et al.*, 2020b). In case of street scenes, it is crucial to effectively integrate information from multiple sensors, e.g. RGB camera and Lidar, the effect of interactions among agents such as pedestrians and vehicles and in case of "on-board" prediction it is also crucial to integrate ego-motion information. We address these challenges in Chapters 4 to 6, 9 and 11 as illustrated in Table 1.1. Particularly for autonomous driving applications in dense urban environments, the interactions between the self-driving vehicle and the pedestrians (bicyclists) are of special interest (Gupta *et al.*, 2018; Mangalam *et al.*, 2020; Sadeghian *et al.*, 2019; Salzmann *et al.*, 2020). One of the challenges in the development of accurate models which can capture the effect of interactions in the distribution of likely future trajectories is the lack of large scale datasets which focus on pedestrian - vehicle interactions in dense urban environments. We deal with this challenge in Chapter 9, with a novel dataset which focuses on pedestrian - vehicle interactions.

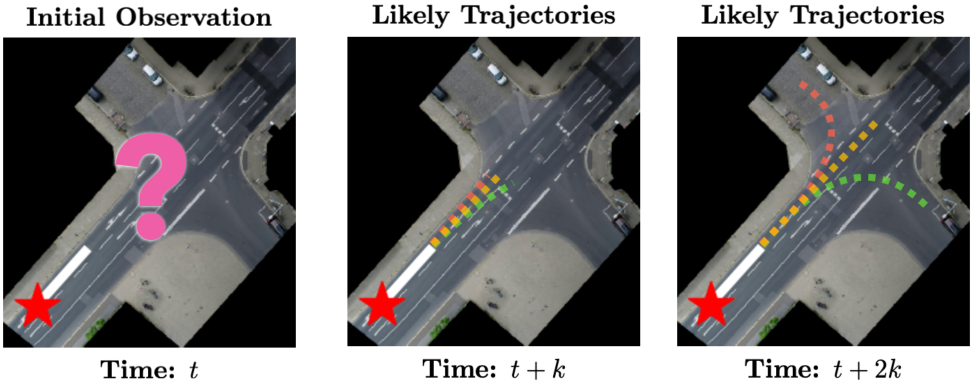


Figure 1.2: In case of non-deterministic agent based scenarios, such as street scenes, the future is highly uncertain and the distribution of likely future outcomes is multimodal. Here we show the distribution of likely future trajectories for a vehicle at an intersection, at k and $2k$ seconds after observation at time t (modes are shown in different colors). The uncertainty and multi-modality of the distribution of likely trajectories increases from into the future, highlighting the challenges of future prediction tasks.

1.1.2 Uncertainty

Many important real world scenarios are inherently non-deterministic, e.g. street scenes. Street scenes are inherently non-deterministic due to the involvement of external agents such as pedestrians or vehicles, whose future states are dependent upon decisions made by the agents themselves, which are hard to anticipate. This makes the future uncertain in such non-deterministic scenarios. The uncertainty increases into the future, with many possible distinct future outcomes even a few seconds into the future as shown in Fig. 1.2. Anticipation methods (Alahi *et al.*, 2016; Helbing and Molnar, 1995; Mathieu *et al.*, 2016) which aim to predict a single future outcome, e.g. the most likely outcome, do not perform well as the single predicted future can be far away from the true outcome. Therefore, it is crucial to accurately capture the distribution of likely future outcomes in such non-deterministic scenarios. Capturing the distribution of likely future outcomes comes with many challenges. An important challenge that we consider in this thesis is calibration (Gal and Ghahramani, 2016b; Kendall and Gal, 2017). Calibration here means that the uncertainty of the predicted distribution (variance) should correspond well to the observed (groundtruth) uncertainty. Calibration is important as it allows us to express confidence in the likelihood of occurrence of a predicted future state. Bayesian inference provides a theoretically grounded approach to obtain calibrated predictive uncertainties. However, standard approaches (MacKay, 1992; Neal, 2012) are computationally expensive. In this thesis, we focus on the challenge of development of scalable Bayesian inference methods for calibration in future prediction tasks in Chapters 4 and 5 as illustrated in Table 1.1.

	Contribution			
Paper	Long-term	Uncertainty	Multi-modality	Exact Inference
Long-Term Image Boundary Prediction Chapter 3, Bhattacharyya *et al.* (2018b)	✓			
Long-Term On-Board Prediction of People- -in Traffic Scenes Under Uncertainty Chapter 4, Bhattacharyya *et al.* (2018b)	✓	✓		
Bayesian Prediction of Future Street Scenes- -using Synthetic Likelihoods Chapter 5, Bhattacharyya *et al.* (2019a)	✓	✓✓	✓	
Accurate and Diverse Sampling of Sequences- -Based on a "Best of Many" Sample Objective Chapter 6, Bhattacharyya *et al.* (2018c)	✓	✓	✓	
Conditional Flow Variational Autoencoders for Structured Sequence Prediction Chapter 7, Bhattacharyya *et al.* (2019c)	✓	✓	✓✓	
Euro-PVI: Pedestrian Vehicle- -Interactions in Dense Urban Centers Chapter 9, Bhattacharyya *et al.* (2021)	✓	✓	✓✓	
Normalizing Flows With Multi-Scale- -Autoregressive Priors Chapter 10,Bhattacharyya *et al.* (2020a)			✓✓	✓
Haar Wavelet Based Block- -Autoregressive Flows for Trajectories Chapter 11,Bhattacharyya *et al.* (2020b)	✓	✓	✓✓	✓

Table 1.1: Overview of the main focus area of each methodological contribution of papers associated with this thesis.

1.1.3 Multi-modality

The next challenge that we consider in this thesis is multi-modality. In many real world scenarios, the distribution of likely futures is highly multi-modal. It is crucial to accurately capture all modes of the distribution of likely futures especially in safely critical applications such as autonomous driving, e.g. all likely paths that a pedestrian or vehicle can take need to be captured in order to avoid collisions (Fig. 1.2). Prior work on many important non-deterministic real-world scenarios with complex multi-modal distributions, e.g. trajectory prediction, have established latent variable models such as conditional generative adversarial networks and conditional variational autoencoders as state of the art (Gupta *et al.*, 2018; Lee *et al.*, 2017b). While such methods have shown promising performance, capturing multi-modal distributions still remains challenging, which we discuss next in more detail.

Conditional generative adversarial are trained using an adversarial min-max game formulation and they do not explicitly maximize the likelihood of the data under the model. This leads to the well know issue of mode collapse where one or more modes of the target distribution are missing (Arjovsky *et al.*, 2017; Srivastava *et al.*, 2017; Salimans *et al.*, 2016), which makes it challenging to fully capture multi-modal distributions. On the other hand, conditional variational autoencoders maximize a lower bound on the data log-likelihood under the model and the posterior distribution of latent variables is inferred using variational inference in a Bayesian framework. Thus, they have been shown to be better at capturing modes of the target data distribution (Bao *et al.*, 2017; Rosca *et al.*, 2017). Still, the standard objective for training conditional variational autoencoders does not sufficiently encourage diversity in predictions. Furthermore, the standard Gaussian prior used in conditional variational autoencoders also makes it challenging to fully capture distribution of likely future states. In this thesis, we focus on the challenge of improving the modelling flexibility of conditional variational autoencoders in Chapters 6 to 9 help better capture multi-modal data distributions in future prediction tasks.

1.1.4 Exact Inference

As noted above, conditional variational autoencoder based models only maximize a lower bound on the true data log-likelihood, the tightness of which is hard to control in practice (Cremer *et al.*, 2018; Huang *et al.*, 2020). The exact likelihood of a future state under the model cannot be optimized. This makes it challenging to accurately capture multi-modal data distributions. Recent work has therefore focused on exact inference models, e.g. autoregressive models (van den Oord *et al.*, 2016b; Salimans *et al.*, 2017) and normalizing flow (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) based models, however mostly for image data. Autoregressive models, while being highly flexible, are difficult to parallelize and thus suffer from slow sampling speeds. Normalizing flows on the other hand are easy to parallelize. However, normalizing flows suffer from comparatively low modelling flexibility due to the invertibility constraints on its internal layers. Long-term spatio-temporal correlations crucial for accurate long-term predictions are not fully captured. Therefore, it is challenging to design exact inference models for accurate long-term future prediction tasks. In this thesis, we deal with this challenge in Chapters 10 and 11 as illustrated in Table 1.1.

1.1.5 Summary

To summarize, in this thesis, we tackle the problem of anticipating the future upto several seconds in scenarios ranging from deterministic billiard tables to non-deterministic agent based street scenes. We target these scenarios because they are important for the success of autonomous agents in the real world, e.g. for autonomous driving. We address four important challenges: 1. Accurate long-term predictions; 2. Uncertainty of future states in the long-term; 3. Multi-modality of the distribution of likely future states in the long-term; 4. Exact inference of the likelihood of future states under the model.

We first deal with the challenge of long-term prediction through a method that exploits image boundaries and can be applied to both deterministic and non-deterministic scenarios. We show that prediction of image boundaries is easier than predictions directly in the RGB pixel space. We also propose the novel Euro-PVI dataset to foster the development of methods which can deal with the effect of interactions on the future states of pedestrians in dense urban environments for long-term predictions. To deal with the challenge of uncertainty, we develop scalable

Bayesian inference methods. We show that efficient Monte Carlo dropout based Bayesian inference methods can be successfully extended to the task of long-term on-board prediction of pedestrian trajectories, while yielding calibrated uncertainties. To deal with the challenge of multi-modality, we propose novel objectives and priors for Monte-Carlo Bayesian inference and conditional variational autoencoders. Finally in order to further improve accuracy and diversity of predictions, we turn our attention to the challenge of exact inference. We focus on normalizing flow based exact inference models, as they allow for efficient inference and sampling. To deal with the challenge of limited modelling flexibility of normalizing flow based model, we introduce efficient autoregressive structures which leads to competitive results in highly multi-modal scenarios. Next, we discuss these contributions in more detail.

1.2 MAIN THESIS CONTRIBUTIONS

This thesis contributes to the broad areas of Bayesian inference, generative modelling, trajectory prediction and scene prediction in general. Next, we group the contributions of the thesis and place them in context of prior work in the field. We additionally provide an overview in Table 1.1, where we highlight the main methodological contribution of each publication associated with this thesis.

1.2.1 Long-term Prediction from Raw Pixel Data

In this thesis, we propose the following approaches to improve long-term predictions in diverse real-world scenarios.

In Bhattacharyya *et al.* (2018b), discussed in Chapter 3, we propose an approach using image boundaries that enables long-term prediction in diverse deterministic and non-deterministic scenarios. Scene boundaries capture the important structure and extents of objects. Moreover, they can be accurately estimated (Khoreva *et al.*, 2016). This makes learning the dynamics of scene evolution and prediction considerably easier. Therefore, we propose to exploit image boundaries in addition to raw RGB pixel data. To predict future image boundaries, we propose a fully convolutional model, which includes a wide receptive field allowing the model to learn complex spatio-temporal dependencies. Accurate prediction at each time-step is additionally enabled by the lack of bottleneck layers. Global consistency of predictions is enabled by a shared context which allows for information sharing. We obtain accurate long-term predictions, in contrast to predictions directly on RGB data, which leads to very blurry results in the long-term. Accurate long-term predictions in deterministic scenarios and accurate short and medium-term predictions in non-deterministic agent based scenarios shows that our model developed a data-driven model of future motion and scene evolution over long time horizons.

In (Bhattacharyya *et al.*, 2021), discussed in Chapter 9, we propose a novel dataset with dense interactions to improve long-term prediction of pedestrian trajectories in dense urban environments. Anticipating the motion of agents in dense urban environments is made especially challenging due to the effect of interactions between different agents. Interactions add significant complexity to the distribution of the likely future trajectories which are already highly multi-modal. Most public datasets, e.g. nuScenes (Caesar *et al.*, 2020), Argoverse (Chang *et al.*, 2019), or Lyft L5 (Houston *et al.*, 2020), are primarily focused on trajectories of vehicles and vehicle-vehicle interactions – collected to a large extent on multi-lane roads in North America or Asia, with sparse interactions between the ego-vehicle and pedestrians or bicyclists. In Chapter 9, we propose the new European Pedestrian Vehicle Interaction (Euro-PVI) dataset which is collected in a dense urban environment in Brussels and Leuven, Belgium. The Euro-PVI dataset contains a rich and diverse set of interactions between the ego-vehicle and pedestrians (bicyclists). Sequences are recorded near busy urban landmarks, e.g. railway stations, narrow lanes or intersections where interactions are frequent and it is therefore challenging to predict pedestrian (bicyclist) paths.

1.2.2 Scalable Bayesian Inference for Uncertainty and Calibration

In this thesis, we propose the following Bayesian inference approaches to obtain calibrated uncertainty estimates in case of complex real-world non-deterministic scenarios.

In Bhattacharyya *et al.* (2018a), discussed in Chapter 4, we propose the first method for long-term prediction of pedestrian trajectories in an "on board setting". In such non-deterministic scenarios with multiple likely futures, a model which predicts a single future outcome would likely be inaccurate. In such settings, it is useful to consider the distribution of likely models that can explain (predict) the observed future, corresponding to the distribution of likely futures. This is also known as model or epistemic uncertainty (Kendall and Gal, 2017). In Bhattacharyya *et al.* (2018a), to capture the uncertainty in the distribution of likely futures, we aim to estimate this distribution of likely models in a Bayesian framework.

We propose a novel two stream Bayesian RNN encoder-decoder model in Bhattacharyya *et al.* (2018a), where the distribution of likely models (parameters) is inferred using dropout based Monte Carlo Bayesian inference. The two streams in our model jointly predict the pedestrian trajectory and the vehicle ego-motion for improved long-term prediction accuracy. The distribution of likely models is represented by a Bernoulli distribution (using dropout) on the weights matrices of the RNN encoder and decoder. The uncertainty introduced due to observation noise (aleatoric uncertainty) is captured by assuming a distribution of observation noise and estimating the sufficient statistics of the distribution. Our two stream Bayesian RNN encoder-decoder model leads to accurate predictions in this challenging "on-board" over a time horizon of 1 second on CityScapes (Cordts *et al.*, 2016). Moreover, we observe that the predicted uncertainty is calibrated. As the predicted uncertainty is calibrated, we also show that the predicted uncertainty upper bounds the error of the mean of the predictive distribution. Therefore, the predicted uncertainty helps us express trust in predictions and has the potential to serve as a basis for better decision making.

However, note that, in many real world scenarios, the distribution of likely future states is highly multi-modal. E.g. in case of a pedestrian crossing the street in front of a car the two most likely modes correspond to stopping and yielding to the oncoming car or pedestrian crossing the street and the oncoming car stopping and yielding to the pedestrian. However, we show in Bhattacharyya *et al.* (2019a) that the approach of Bhattacharyya *et al.* (2018a), does not perform well in case of multi-modal distributions. This is because the approach of Bhattacharyya *et al.* (2018a) uses the estimate of data log-likelihood as outlined in Gal and Ghahramani (2016b); Kendall and Gal (2017). This estimate of data log-likelihood does not lead to the recovery of the true model uncertainty (Osband, 2016), due to a conflation of risk and uncertainty (Osband, 2016). This limits the accuracy of the models over a plain deterministic (non-Bayesian) approach. In detail, this estimate of the data log-likelihood considers for every data point the average likelihood assigned by all models in the distribution of likely models. This forces every model to explain every data point well, pushing every model in the distribution of likely models to the mean. We address this problem in Bhattacharyya *et al.* (2019a), discussed in Chapter 5, through an objective leveraging synthetic likelihoods (Wood, 2010; Rosca *et al.*, 2017), obtained from a classifier. The synthetic likelihood estimate is based on whether the models explain (generate) data samples likely under the true data distribution. This removes the constraint on models to explain every data point – it only requires the explained (generated) data points to be likely under the data distribution. Thus, this allows models in the distribution of likely models to be diverse and deals with multi-modality. Our proposed method shows state of the art performance on the challenging task of predicting the future of street scenes in the Cityscapes dataset. More importantly, we show that the predictive distribution of our Bayesian model produces calibrated uncertainties.

1.2.3 Conditional Variational Autoencoders for Multi-modality

In this thesis, we propose the following to the improve accuracy and diversity of conditional variational autoencoders, to deal with complex multi-modal distributions for future prediction tasks: 1. A "Best of Many" samples objective (Bhattacharyya *et al.*, 2018c) (discussed in Chapter 6); 2. A flexible conditional normalizing flow based prior (Bhattacharyya *et al.*, 2019c) (discussed in Chapter 7); 3. A joint prediction framework to capture interactions across agents in a scene (Bhattacharyya *et al.*, 2021) (discussed in Chapter 9). 4. Building on the insights in Bhattacharyya *et al.* (2018c) we also propose a method to improve performance of VAE-GAN based methods for image generation in Chapter 8. Next, we discuss these contributions in more detail.

"Best of Many" samples objective. We identify two key limitations of the standard conditional variational autoencoder framework (Sohn *et al.*, 2015), popularly used for future prediction tasks (Lee *et al.*, 2017b). First, the standard objective hinders the learning of diverse predictions due to a marginalization over multi-modal futures. Second, a mismatch in latent variable distribution between training and testing leads to errors in model fitting. To overcome these limitations we propose a novel "Best of many" samples objective (Bhattacharyya *et al.*, 2018c). Compared to the standard conditional variational autoencoder objective, the recognition network of the conditional variational autoencoder now has multiple chances to draw samples with high posterior probability. This encourages diversity in the generated samples. Furthermore, the data log-likelihood estimate in our "Best of many" samples objective is tighter. Therefore, our "Best of many" sample objective loosens the constrains on the recognition network and allows it more closely match the latent prior distribution. We demonstrate improved accuracy as well as diversity of the generated samples on three diverse tasks: stroke completion on MNIST digits (D. De Jong, 2016), trajectory prediction on Stanford Drone Dataset (Cordts *et al.*, 2016) and precipitation nowcasting on HKO weather data (Shi *et al.*, 2017). On all three task we consistently outperform prior work.

Conditional normalizing flow based prior. Conditional variational autoencoders assume a standard Gaussian prior on the latent variables which induces a strong model bias (Hoffman and Johnson, 2016; Tomczak and Welling, 2018). This makes it challenging to capture multi-modal distributions. This also leads to missing modes due to posterior collapse (Bowman *et al.*, 2016; Razavi *et al.*, 2019a). Recent work (Tomczak and Welling, 2018; Wang *et al.*, 2017; Gu *et al.*, 2019) has focused on more complex Gaussian mixture based priors. Gaussian mixtures still have limited expressiveness and optimization is not straightforward, e.g. determining the number of mixture components. Normalizing flows are more expressive and enable the modelling of complex multi-modal priors. Here, we propose Conditional Flow Variational Autoencoders (CF-VAE) (Bhattacharyya *et al.*, 2019c) based on novel conditional normalizing flow based priors In order to model complex multi-modal conditional distributions over sequences. Our proposed CF-VAE outperforms prior work on diverse prediction tasks – stroke completion on MNIST, trajectory prediction on Stanford Drone (Cordts *et al.*, 2016) and highD (Krajewski *et al.*, 2018).

Joint prediction framework. State of the art conditional variational autoencoder based methods for trajectory prediction(Bhattacharyya *et al.*, 2018c, 2019c; Mangalam *et al.*, 2020; Salzmann *et al.*, 2020), encode interactions directly in the posterior. Thus, the latent space does not express interaction information from the input distribution which limits the accuracy of the generated future trajectories. To address this limitation, we develop a latent variable deep generative model in Bhattacharyya *et al.* (2021) which jointly models the distribution of future trajectories of the agents in the scene. This formulation leads to a "shared" latent space between agents, which can better capture the effect of interactions in the latent space and accurately represent the multi-modal distribution of trajectories. We demonstrate state of the art performance on pedestrian (bicyclist) trajectory prediction on nuScenes (Caesar *et al.*, 2020) and Euro-PVI (Bhattacharyya *et al.*, 2021).

"Best of Many" distribution matching. Variational autoencoders (VAE) in case of image data, maximize a data likelihood estimate based on the L_1/L_2 reconstruction cost which leads to lower overall sample quality (blurriness in case of image distributions). Therefore, there has been a spur of recent work (Donahue *et al.*, 2017; Larsen *et al.*, 2016; Rosca *et al.*, 2017) which aims integrate generative adversarial network

(GAN) in a variational autoencoder framework to improve generation quality of Variational autoencoders while covering all the modes. However, the reconstruction log-likelihood in the variational autoencoder objective is at odds with the divergence to the latent prior (Tabor *et al.*, 2018). This problem is further exacerbated with the addition of the synthetic likelihood term in the hybrid VAE-GAN objective – it is necessary for sample quality but it introduces additional constraints on the encoder/decoder. This leads to the degradation in the quality and diversity of samples. Here, we propose a novel objective (Bhattacharyya *et al.*, 2019b) for training hybrid VAE-GAN frameworks, which relaxes the constraints on the encoder by giving the encoder multiple chances to draw samples with high likelihood enabling it to generate realistic images while covering all modes of the data distribution. We demonstrate significant improvement over state of the art hybrid VAE-GANs and plain GANs on multi-modal synthetic data and on CIFAR-10 and CelebA datasets.

1.2.4 Exact Inference Models for Multi-modal Distributions

Recent work (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) considers normalizing flow based exact likelihood models to overcome limitations of generative adversarial networks and variational autoencoders. Still, normalizing flows suffer from limited modelling flexibility due to the invertibility constraints on the coupling layers. In this thesis, we propose the following contributions to improve the modelling flexibility of normalizing flows, 1. A multi-scale autoregressive prior (Bhattacharyya *et al.*, 2020a) (discussed in Chapter 9), 2. A block autoregressive scheme using Haar wavelets (Bhattacharyya *et al.*, 2020b) (discussed in Chapter 11). Next, we discuss these contributions in more detail.

Multi-scale autoregressive prior. We propose in Bhattacharyya *et al.* (2020a) multi-scale autoregressive priors for invertible flow models with split coupling flow layers, termed *mAR-SCF*. Our *mAR* prior is designed to capture long-term dependencies in multi-modal data distributions, directly in the latent space. Our multi-scale autoregressive prior is designed such that the computational cost of sampling grows linearly in the spatial dimensions in case of image data, compared to the quadratic cost of traditional autoregressive models. We show state-of-the-art density estimation results on MNIST, CIFAR-10, and ImageNet compared to prior invertible flow-based approaches and better sample quality as measured by the FID metric (Heusel *et al.*, 2017) and the Inception score (Salimans *et al.*, 2016), significantly lowering the gap to generative adversarial network based approaches.

Block autoregressive scheme using Haar wavelets. We introduce a block autoregressive exact inference model for future trajectory prediction using Haar wavelets. In more detail, we use a Haar-wavelet based invertible transform to recursively transform the trajectories into coarse and fine components in a multi-scale setup. Normalizing flows are applied at each scale of the multi-scale setup in a block auto-regressive structure. Normalizing flows at a certain scale model the fine components, and are conditioned on coarse trajectories from previous scale. The coarse trajectories provide global context, thereby modeling long term spatio-temporal correlations and leading to more accurate long-term predictions. We demonstrate the effectiveness of our approach for trajectory prediction on Stanford Drone (Cordts *et al.*, 2016) and Intersection Drone (Bock *et al.*, 2020), with improved accuracy over long time horizons.

1.3 CONTRIBUTIONS TO OTHER PROJECTS

Besides the main contributions discussed above, the author of this thesis actively collaborated in the following research projects mainly in an advisory role in the development of the core contributions of the respective projects. Next, we discuss these contributions and provide short descriptions of the projects.

Update Leaks. In this work (Salem *et al.*, 2020), we study the vulnerability of machine learning models to specific kinds of attacks that can reveal (private) information about the test-set. In many real-world settings, machine learning models are continuously updated with newly-collected data in an online learning scenario. This leads to the situation where a user would have access to the result of the model to the same query at two different points in time. In this work, we investigate whether the change in the result of the query (if any) would reveal information about the data used to update the model (updating set). This work investigates four different types of attacks to reveal different types of information about the updating set. These attacks can be divided into two cases where one or multiple data samples were used to update the model. Each case includes attacks to reconstruct the updating set. We propose to use a hybrid GAN based generative model for the reconstruction attacks.

The author of this dissertation helped in the development of the hybrid GAN based generative model for the reconstruction, which is partially based on Bhattacharyya *et al.* (2018c, 2019b) .

SampleFix. In this work (Hajipour *et al.*, 2019), we focus on the task of automatic program repair. Current approaches for automatic program repair aim to predict a single correct fix for an erroneous program. However, this is challenging due to the uncertainty associated with the true intention of the programmer. Therefore, in this work, we propose to instead deal with this uncertainty by learning the distribution over likely fixes. To this end, we propose a conditional variational autoencoder based model with a novel regularizer that encourages diversity in predictions. Our experiments show that our model is able to capture the distribution of likely fixes in case of common programming errors.

The author of this dissertation helped in the development of the conditional variational autoencoder and the novel regularizer which is based on Bhattacharyya *et al.* (2018c) for the generation of diverse fixes.

1.4 OUTLINE OF THE THESIS

In this section, we provide a short overview of each chapter in this thesis and relate the chapters to each other. With the exception of Chapters 1 and 12, all other chapters have been published in conferences or shown in a workshop. The author of this thesis, Apratim Bhattacharyya, is a lead author of the conference/workshop papers presented in Chapters 3 to 9 and 11 of this thesis. Chapter 10 is a joint work with Shweta Mahajan. The contributions of Shweta Mahajan are included in Chapter 10 for completeness. Prof. Dr. Bernt Schiele and Prof. Dr. Mario Fritz, the PhD co-supervisors of Apratim Bhattacharyya are co-authors of all the papers presented in this thesis. We discuss collaborations with other researchers per chapter in detail below.

Chapter 1: Introduction This chapter introduces the main research direction of the thesis, that is, anticipation of the future. This chapter highlights the main challenges associated with this research direction, 1. Long-term predictions, 2. Uncertainty, 3. Multi-modality, 4. Exact inference. This chapter also highlights the contributions of the thesis to the areas of Bayesian inference, generative models and datasets for anticipation in dense urban environments. This chapter also provides chapter wise outlines of each chapter included in the thesis.

Chapter 2: Related Work This chapter provides an extensive discussion of prior work pertaining to the main research directions of the thesis. We begin with a discussion of prior work on the area of long-term prediction in deterministic scenarios, in particular, the area of intuitive physics. We also include a discussion of methods with the aim to provide long-term predictions on highly non-deterministic scenarios, e.g. video frame prediction and street scene prediction. This is followed by a discussion of prior work on Bayesian inference methods with a focus on scalable methods which deal with the challenge of uncertainty. We then discuss generative models in more detail, in particular, conditional generative adversarial networks, conditional variational auto-encoders and exact inference methods, especially the successful application of these approaches to the challenge of multi-modality in diverse scenarios for anticipation including trajectory prediction. This chapter also provides an overview of existing datasets relevant to the problem of anticipation.

Chapter 3: Long-term Image Boundary Prediction. This chapter is based on the publication (Bhattacharyya *et al.*, 2018b) presented in AAAI Conference on Artificial Intelligence, 2018. This work is a collaboration with Mateusz Malinowski from MPI Informatics and Google Deepmind. In this chapter, we propose a method for long-term prediction that operates on raw image data for both deterministic (physics based) scenarios and non-deterministic agent based scenarios. Our method works directly on image boundaries for learning and prediction instead of the raw RGB images which allows for long-term prediction.

Chapter 4: Long-Term On-Board Prediction of People in Traffic Scenes under Uncertainty. This chapter is based on the publication Bhattacharyya *et al.* (2018a) presented at the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), 2018. In this chapter we present a Bayesian inference method to deal with uncertainty in the highly non-deterministic scenario of on-board pedestrian trajectory prediction. We show that our method produces calibrated uncertainty estimates, where the predicted uncertainty is empirically shown to upper bound the maximum observed error.

Chapter 5: Bayesian Prediction of Future Street Scenes using Synthetic Likelihoods. The chapter is based on the publication Bhattacharyya *et al.* (2019a) presented at the International Conference on Learning Representations, (ICLR), 2019. In this chapter we build upon the Bayesian inference method introduced in Chapter 4 to deal with multi-modal distributions. We introduce a novel objective function based upon synthetic classifier based likelihood estimates when encourages the coverage of all modes of a multi-modal distribution. We demonstrate state of the art results on future street scene prediction on the Cityscapes datasets.

Chapter 6: Accurate and Diverse Sampling of Sequences based on a "Best of Many" Sample Objective. This chapter is based on the publication Bhattacharyya *et al.* (2018c) presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. In this chapter we focus on improving conditional variational autoencoders in order to deal with highly multi-modal distributions encountered in future prediction problems. We introduce a novel "Best of Many" objective for training conditional variational autoencoders which explicitly encourages diversity in predictions. We demonstrate state of the art results on trajectory prediction results on Stanford Drone dataset and precipitation nowcasting on HKO.

Chapter 7: Conditional Flow Variational Autoencoders for Structured Sequence Prediction. This chapter is based on the workshop publication Bhattacharyya *et al.* (2019c) presented at the Machine Learning for Autonomous Driving and Bayesian Deep Learning, NeurIPS workshops 2019. This work is a collaboration with Christoph-Nikolas Straehle and Michael Hanselmann from Bosch Center for Artificial Intelligence, Renningen, Germany. In this chapter we further improve the conditional variational autoencoder framework presented in Chapter 6 for multi-modal distributions with a complex prior. The standard Gaussian prior in conditional variational autoencoders induces a strong model bias which makes it challenging to capture multi-modal distributions. Here we introduce a novel conditional normalizing flow based prior to capture complex multi-modal conditional distributions. We demonstrate state of the art results on trajectory prediction results on Stanford Drone and HighD datasets.

Chapter 8: "Best-of-Many-Samples" Distribution Matching. This chapter is based on the workshop publication Bhattacharyya *et al.* (2019b) presented at the Bayesian Deep Learning, NeurIPS workshop 2019. We build upon the insights gained in Chapter 6 and propose a "Best of Many" samples objective for image generation using a hybrid VAE-GAN based method. Similar to the findings in Chapter 6, we see that the "Best of Many" samples objective leads to higher diversity also in case of highly multi-modal image distributions. We demonstrate a better match to the groundtruth image distribution on CIFAR-10 and CelebA datasets compared to both hybrid VAE-GANs and plain GANs.

Chapter 9: Euro-PVI: Pedestrian Vehicle Interactions in Dense Urban Centers. This chapter is based on the publication Bhattacharyya *et al.* (2021) presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. This work is a collaboration with Daniel Olmeda Reino from Toyota Motor Europe. In this chapter we focus on the effect of agent-agent interactions, in particular pedestrian (bicyclist) - vehicle interactions, on the highly multi-modal future pedestrian trajectory distribution in dense urban environments. Firstly, we propose Euro-PVI, a novel dataset of pedestrian and bicyclist trajectories recorded in Europe with dense interactions with the ego-vehicle. Secondly, we propose a joint conditional variational autoencoder that models a shared latent space to better capture the effect of interactions on the multi-modal distribution of future trajectories. We demonstrate state of the art results both on nuScenes and Euro-PVI.

Chapter 10: Normalizing Flows with Multi-scale Autoregressive Priors. This chapter is based on the publication Bhattacharyya *et al.* (2020a) presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. This is a joint work with Shweta Mahajan and Stefan Roth from TU Darmstadt. In this chapter, we propose a normalizing flow based exact inference method to deal with highly multi-modal distributions. We propose an efficient multi-scale autoregressive prior to deal with the limited modelling flexibility of invertible normalizing flow based models. We demonstrate improved accuracy and image generation quality on CIFAR-10 and ImageNet datasets.

Chapter 11: Haar Wavelet based Block Autoregressive Flows for Trajectories. This chapter is based on the publication Bhattacharyya *et al.* (2020b) presented at the DAGM German Conference on Pattern Recognition (GCPR), 2020. This work is a collaboration with Christoph-Nikolas Straehle from Bosch Center for Artificial Intelligence, Renningen, Germany. Here we propose a multi-scale normalizing flow based model to deal with multi-modal trajectory distributions. We introduce a novel normalizing flow based block autoregressive approach that models trajectories at different spatio-temporal resolutions in a hierarchical manner using Haar wavelets. We demonstrate state of the art trajectory prediction results on Stanford Drone and Intersection Drone datasets.

Chapter 12: Conclusions and Future Perspectives. This chapter concludes the thesis by highlighting the key contributions and findings from each chapter. We also discuss their limitations and possible future directions of research to address these limitations. We also provide a broader outlook of future directions of research for the field.

RELATED WORK 2

Contents

The main focus of this thesis are the challenges of long-term prediction, uncertainty, multi-modality and exact inference associated with anticipation of the future. This chapter presents prior work in the area of long-term prediction in diverse scenarios, in particular, video frame prediction, street scene prediction and trajectory prediction. Followed by a discussion of methods that deal with uncertainty in predictive settings, with a focus on Bayesian inference methods. Finally we discuss prior work in generative modelling for multi-modal distributions. Here, we focus on conditional generative adversarial networks, conditional variational autoencoders and exact inference methods including both auto-regressive methods and normalizing flow based methods.

The following chapters (3-11) also discuss related work but restricted to the main area of focus of the respective chapter.

2.1 LONG-TERM PREDICTION

Here we discuss prior work on long-term prediction broadly categorized by the chosen representation (video, semantic segmentation, trajectory) and task (frame prediction, semantic segmentation prediction, trajectory prediction).

2.1.1 Video Frame Prediction

We begin with a discussion on the task of video frame prediction. More formally, given a sequence of t frames, $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_t\}$, the target is to predict the next k frames, $\mathbf{Y} = \{\mathbf{Y}_{t+1}, \cdots, \mathbf{X}_{t+k}\}$ based on $\mathbf{X}$ and any other conditioning information (Oprea *et al.*, 2020). Note that, here $\mathbf{X}_i$ and $\mathbf{Y}_i$ are both 3-dimensional tensors. Video frame prediction is challenging as it is quite general and videos can include both deterministic, e.g. physics based motion and non-deterministic agent based motion.

Most methods in literature consider the more general class of videos which includes non-deterministic motion. The first method using a deep neural architecture for video prediction was proposed by Ranzato *et al.* (2014). In this work, it was shown that existing methods for language modelling can be extended to the task of video prediction. Ranzato *et al.* (2014) also noted the difficulty of long-term prediction and the issue of blurriness associated with long-term prediction. Ranzato *et al.* (2014) sought to remedy the blurriness problem by discretizing the input through k-means atoms and predicting on this vocabulary instead. While it makes learning easier, it does not address the core issue of uncertainty. As there are already many likely future frames even a few time-steps into the future, the mean square error loss used in Ranzato *et al.* (2014) leads to the prediction of the average of the likely frames and thus leads to blurriness. The work of Mathieu *et al.* (2016) proposes using adversarial loss, which leads to improved results over Ranzato *et al.* (2014) as the adversarial loss prefers the prediction of a single sharp (thus likely) future frame over the average (thus very unlikely) future frame. Continuing on this line of work,Liang *et al.* (2017); Liu *et al.* (2017b); Patraucean *et al.* (2015) shows further improvement in sharpness through the use of optical flow information.Aigner and Körner (2018) improves over Ranzato *et al.* (2014) by using the progressively growing training scheme. Xu *et al.* (2018) proposes to decompose videos into high and low frequency content. The proposed model in Xu *et al.* (2018) consists of two streams which predict these two frequency bands separately for improved performance. Note that these methods do not fully address the core issue of uncertainty and thus truly long-term predictions are not possible. Another line of work considers more constrained scenarios where certain high level structures are present, which are easier to predict due to lower overall uncertainty. Villegas *et al.* (2017); Yang *et al.* (2018) proposes such methods. These works show promising results on videos where human pose is easily identifiable and is the appropriate high level structure to exploit – especially in videos where a single person performs an action such as hitting a baseball with a bat. In such cases, the sequence of possible future poses is highly constrained and thus easier to predict.

In contrast to the above mentioned approaches, Sutskever *et al.* (2008); Michalski *et al.* (2014) focus on deterministic scenarios, e.g. bouncing ball sequences, with the aim of long-term prediction with promising results. On the other hand, Kalchbrenner *et al.* (2017) focus on synthetic moving MNIST digits and Finn *et al.* (2016) focus on (partially deterministic) action conditioned video prediction of robotic arms.

More recently, Babaeizadeh *et al.* (2018); Denton and Fergus (2018) use generative models to explicitly deal with uncertainty and multi-modality. We provide a more in depth look at generative models in the following sections.

2.1.2 Intuitive Physics

Here we discuss methods which aim for long-term prediction in deterministic scenarios largely governed by the laws of physics. The knowledge of the relevant laws and the associated physical quantities would enable long-term prediction. While it is certainly possible to design hand-crafted methods particular to a certain environment based on domain knowledge to enable long-term prediction, such a system would not generalize and would require extensive manual intervention. On the other hand, humans develop an intuitive notion of physics from an early age, including notions of force, stability and object permanence (Smith and Casati, 1994; McCloskey, 1983; Liu *et al.*, 2017a) directly from observation. This intuitive notion of physics enables humans to excel at sports such as billiards which requires extensive reasoning and long-term prediction about the future state of the world. The challenges involved in developing such an "intuitive" understanding for autonomous agents has been recognised for a long time (Battaglia *et al.*, 2013). While pioneering work by Battaglia *et al.* (2013) took an simulation based approach, recent work (Fragkiadaki *et al.*, 2016; Lerer *et al.*, 2016; Li *et al.*, 2017) has taken a data driven approach to solve this problem, partly inspired by the ability of humans to develop this intuitive notion of physics largely from observation. Fragkiadaki *et al.* (2016) proposes an approach to predict future states of balls moving on a billiard table and Lerer *et al.* (2016); Li *et al.* (2017) aims to predict the stability of towers made out of blocks. These approaches present promising results – accurate long-term predictions several seconds into the future are shown. However, the specific tasks considered by these approaches are largely synthetic and far from the real world. Moreover, they do not operate on raw sensory input but input representations tailored to the specific task. Thus, generalization is limited. More recently, there is increasing interest in the development of "intuitive" physics approaches that operate directly on raw sensory input. In Ehrhardt *et al.* (2018), learning and prediction directly from raw visual observations is enabled by tracking dynamically-salient objects in videos using causality and equivariance. In Li *et al.* (2019) an end-to-end learning-based approach to predict stability directly from appearance is presented. Similarly, in Wang *et al.* (2018) real images are first converted to a synthetic domain representation that reduces complexity arising from lighting and texture which allows for intuitive learning of physics and accurate long-term prediction.

2.1.3 Trajectory Prediction

The task of trajectory prediction focuses on predicting the (future) trajectory $\mathbf{y} = \{\mathbf{y}_{t+1}, \cdots, \mathbf{y}_{t+k}\}$ of an agent, e.g. a pedestrian, k time-steps into the future, given the initial t time-steps, $\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$. Here, each $\mathbf{x}_i, \mathbf{y}_i$ is a vector of the position and/or velocity and acceleration of the agent. Most methods in literature focus on the prediction of trajectories of pedestrians, bicyclists and/or vehicles in highly non-deterministic traffic scenarios, due to the applications in the area of autonomous driving. Early approaches did not explicitly take into account uncertainty, similar to the area of video frame prediction. This therefore limited performance in the long-term, however to improve short to medium term performance recent approaches (Alahi *et al.*, 2016) aimed at capturing the effect of interactions.

One of the first approaches in this direction was the Social Forces model of Helbing and Molnar (1995). This line of work which considers the problem of traffic participant prediction in a social context, by taking into account interactions among traffic participants was followed up by recent work including but not limited to Alahi *et al.* (2016); Helbing and Molnar (1995); Yamaguchi *et al.* (2011); Robicquet *et al.* (2016). Social LSTM (Alahi *et al.*, 2016) was one of the first successful deep learning based approaches. To capture the effect of interactions over a simple LSTM based model, Social LSTM (Alahi *et al.*, 2016) introduces a social pooling layer to aggregate interaction information of nearby traffic participants. Although the proposed pooling layer improved accuracy, it introduced a significant computational overhead. An efficient extension of the social pooling operation was developed in Deo and Trivedi (2018). Other recent approaches have attempted to further improve the performance of the Social LSTM. Ma *et al.* (2019) proposed alternate instance and category layers to model interactions. Weighted interactions are proposed in Chandra *et al.* (2019). Another promising approach is the convolutional multi-agent tensor fusion scheme as proposed in Zhao *et al.* (2019) to capture interactions. The approach of Zhao *et al.* (2019) is also quite efficient as it is convolutional similar to Deo and Trivedi (2018). More recently, an attention based model to better capture interactions and to effectively integrate visual cues in path prediction tasks is proposed in Sadeghian *et al.* (2018).

Although the works discussed above have made great progress in effectively capturing the effect of interactions for accurate prediction, likely future trajectories are highly uncertain and the distribution of likely trajectories is highly multi-modal. The above discussed methods mostly assume a deterministic future and do not directly deal with the challenges of uncertainty and multimodality. To deal with the challenges of uncertainty and multimodality in anticipating future trajectories, recent works employ generative approaches to capture the distribution of future trajectories which we discuss in the following sections.

2.1.4 Street Scene Prediction

Anticipating the future of street scenes is important for applications such as autonomous driving. However, street scenes are highly non-deterministic. This makes the prediction of street scenes directly in RGB space challenging (due to issues pointed out in Section 2.1.1). Recent work therefore considers prediction in semantic segmentation space instead of RGB pixel space. That is, instead of predicting full RGB future frames, future semantic segmentations are predicted. Practically, future semantic segmentations are still very useful as the location and extents in the future are captured. Luc *et al.* (2017) proposes the first such method for the prediction of future semantic segmentations. The proposed model is fully convolutional with prediction at multiple scales and is trained auto-regressively. Accurate results upto 1 second into the future are presented, significantly more than RGB frame predictions. Jin *et al.* (2017) improves upon the model of Luc *et al.* (2017) using optical flow cues. Architecturally similar to Luc *et al.* (2017), the model of Jin *et al.* (2017) is fully convolutional model and is based on a Resnet-101 He *et al.* (2016) backbone with a single prediction scale. More recently, Luc *et al.* (2018) has extended the model of Luc *et al.* (2017) to the related task of future instance segmentation prediction. In Nabavi *et al.* (2018), a Convolutional LSTM based model is proposed, further improving short-term results over Jin *et al.* (2017). Fugosic *et al.* (2020); Saric *et al.* (2021) proposes models that predict future deep semantic features and predicts future motion based on the predicted features. Graber *et al.* (2021) proposes to use panoptic segmentation to help distinguish between background and foreground objects.

More accurate predictions in the medium to long-range over RGB predictions show the promise of future semantic segmentation prediction.

2.1.5 Relation to Our Work

Long-term prediction in both deterministic and non-deterministic scenarios is a primary goal of this thesis.

In Chapter 3, we propose a method for long-term video prediction for both of these scenarios. Our approach exploits image boundaries which captures crucial high frequency information in the scene. In case of deterministic scenarios, we show that our approach can develop an intuitive understanding of physics similar to approaches like Fragkiadaki *et al.* (2016); Lerer *et al.* (2016); Li *et al.* (2017), but directly from raw sensory input. Moreover, we show that our proposed approach leads to sharper video prediction in the long-term in comparison to approaches like Mathieu *et al.* (2016) . This is because image boundaries are easier to predict into the long-term and can be exploited by our approach to deal with blurriness issues.

In Chapter 4, we propose a method of trajectory prediction in an "on-board" setting in dense urban environments. A key ingredient of our success in long-term prediction in Chapter 4 is a two-stream model that jointly predicts pedestrian and vehicle ego-motion. However, the main focus of Chapter 4, is to develop an approach which can effectively capture uncertainty. Although interactions are important for accuracy in dense urban environments, we choose to address the issue of uncertainty as it has only been addressed in a limited manner by prior work. Rather than focusing solely on accuracy, an important observation from our work in Chapter 4, is that addressing uncertainty leads to the development of a more reliable method – we show empirically that our approach leads to a measure of uncertainty that upper bounds the maximum observed error.

In Chapter 5, we show long-term predictions on street scenes through future semantic segmentation prediction. In comparison to prior work such as Luc *et al.* (2017); Jin *et al.* (2017); Nabavi *et al.* (2018), we explicitly take into account the uncertainty associated with the prediction of future street scenes in a Bayesian framework. Building on our conclusions in Chapter 5, we see that our approach can deal with the uncertainty in the distribution of future street scenes even in case of multi-modal distributions.

In Chapter 9, we deal with the effect of interactions for long-term trajectory prediction of traffic participants such as pedestrians and bicyclists. Due to the particular importance in the area of autonomous driving we focus on pedestrian-vehicle interactions. We propose a novel dataset with dense pedestrian-vehicle interactions and a joint prediction framework to capture the effect of interactions across all agents in the scene.

2.2 MODELLING UNCERTAINTY

In this section, we discuss approaches that aim to capture uncertainty in predictive distributions. These methods are of particular interest for this thesis as the distribution of future states especially in case of non-deterministic, e.g. agent based motion, becomes highly uncertain. In particular, we focus on methods that aim to predict calibrated uncertainty estimates, that is, the uncertainty of the predictive distribution matches the uncertainty of the groundtruth distribution. Next, we discuss two main lines of research which aim to predict such calibrated uncertainty estimates.

2.2.1 Bayesian Methods

Bayesian methods aim to capture both observation (aleatoric) and model (epistemic) uncertainty (Gal and Ghahramani, 2016a). Model (epistemic) uncertainty deals with our ignorance about which model generated our data. Aleatoric uncertainty deals with noise inherent in the observation, e.g. sensor noise or motion noise. Note that heteroscedastic regression methods (Nix and Weigend, 1994; Le *et al.*, 2005) estimate aleatoric uncertainty by predicting the parameters of an assumed observation noise distribution. However, Bayesian methods provide an attractive framework to capture both epistemic and aleatoric uncertainty. Bayesian neural networks (MacKay, 1992; Neal, 2012) provide the opportunity to cast modern neural networks as Bayesian inference methods. Unlike standard neural networks, Bayesian neural networks place a distribution over the model parameters (weights). This offers a probabilistic view of deep learning and more importantly, provides an approach to obtain model (epistemic) uncertainty estimates – through uncertainty in the weight parameters. However, inference of the posterior distribution over weights in such networks is difficult. A popular approach is variational Inference (Blundell *et al.*, 2015). However, the approach of Blundell *et al.* (2015) incurs a considerable computational overhead. To deal with these computational issues, Gal and Ghahramani (2016b) showed that dropout training in deep neural networks approximates Bayesian inference in deep Gaussian processes. Extending these results it was shown in Gal and Ghahramani (2016a) that dropout training can be cast as approximate Bernoulli variational inference in Bayesian neural networks, with minimal computational overhead. These results were extended to recurrent neural networks in Gal and Ghahramani (2016c). The proposed Bayesian recurrent neural networks showed superior performance to standard recurrent neural networks with dropout in various tasks. More recently, Kendall and Gal (2017) presents a Bayesian deep learning framework jointly estimating aleatoric uncertainty together with epistemic uncertainty. The resulting framework provides state-of-the-art results on segmentation and depth regression benchmarks. This shows the promise of such Bayesian approaches to deal with complex real world problems.

2.2.2 Non-Bayesian Methods

In contrast to these Bayesian approaches, Lakshminarayanan *et al.* (2017) proposes an ensemble based approach for uncertainty estimation that is simple to implement, readily parallelizable, and requires very little hyperparameter tuning compared to Bayesian methods, with competitive uncertainty estimates. Another promising approach, as proposed by Malinin and Gales (2018), models uncertainty by parameterizing a prior distribution over predictive distributions. Other non-Bayesian approaches include Garriga-Alonso *et al.* (2019); Hendrycks *et al.* (2020); Thulasidasan *et al.* (2019) among others.

2.2.3 Relation to Our Work

We use Bayesian inference methods to deal with uncertainty in the distribution of likely future pedestrian trajectories and street scenes in Chapter 4 and Chapter 5. Specifically, we build upon the Bayesian inference method of Gal and Ghahramani (2016b). In Chapter 4, we propose a Bayesian inference method for "on-board" trajectory prediction. In detail, our proposed model is of a two stream Bayesian network that integrates ego-motion for improved pedestrian trajectory prediction. In Chapter 5, we specially consider the issue of multi-modality. Our model from Chapter 4, as it is based on the work of Gal and Ghahramani (2016b), cannot accurately capture multi-modal distributions. In Chapter 5, we propose a novel synthetic likelihood based objective for training Bayesian inference methods. This novel objective encourages diversity in the model distribution and allows us to better capture multi-modal data distributions.

2.3 GENERATIVE MODELS FOR MULTI-MODALITY

In this section, we discuss methods which aim to address the challenge of multi-modality in future prediction. Most such approaches are based on a type of generative model – most popularly conditional generative adversarial networks or conditional variational autoencoders. We begin with a discussion on both unconditional and conditional generative adversarial networks and variational autoencoders. We also discuss future prediction methods which are based on conditional generative adversarial networks or conditional variational autoencoders. Finally, we also discuss exact inference methods which unlike conditional generative adversarial Networks and conditional variational autoencoders allows for the computation of the exact likelihood of the data under the model.

2.3.1 Generative Adversarial Networks

Unconditional generative adversarial networks. The classic generative adversarial network formulation was proposed by Goodfellow *et al.* (2014). It consists of a generator which transforms a simple latent distribution to the target distribution. To learn this transformation, the generator is trained along with a discriminator in a adversarial setting. The discriminator is trained to distinguish between samples from the true data distribution and samples from the generator. In the ideal setting, the generator learns to match the true data distribution and the discriminator cannot distinguish between the true and generated data distributions. Experimental results in Goodfellow *et al.* (2014) already showed promising results on complex multi-modal image distributions. One of the first important improvements over the original formulation was proposed by Radford *et al.* (2016), through the introduction of a convolutional model architecture. However, the formulation of Goodfellow *et al.* (2014); Radford *et al.* (2016) has several shortcomings – importantly mode collapse. Mode collapse occurs when one or more modes of a multi-modal data distribution are not captured by the generative adversarial network. To address the challenge of multi-modality in future prediction, avoiding the issue of mode collapse is crucial.

Denoising Feature Matching (Warde-Farley and Bengio, 2017) deals with mode collapse by regularizing the discriminator using an autoencoder. MDGAN (Che *et al.*, 2017) uses two separate discriminators and regularizes using a auto-encoder. In EBGAN (Zhao *et al.*, 2017b), the discriminator is interpreted as an energy functional and is also cast in an auto-encoder framework, leading to improvements in semi-supervised learning tasks. BEGAN (Berthelot *et al.*, 2017) proposes a Wasserstein distance based objective to train such GANs with auto-encoder based discriminators. The proposed approach leads to smoother convergence. InfoGAN (Chen *et al.*, 2016) maximizes the mutual information between a small subset of latent variables and observations in an information theoretic framework. This leads to disentangled and more interpretable latent representations. PacGAN (Lin *et al.*, 2018) proposes to deal with the mode collapse problem by using the discriminator to distinguish between product distributions. D2GAN (Nguyen *et al.*, 2017) proposes to use two discriminators – one for the forward KL divergence between the true and generated distributions and one for the reverse. BourGAN (Xiao *et al.*, 2018) proposes to learn the distribution of the latent space (instead of assuming Gaussian) which reflects the distribution of the data. In Srivastava *et al.* (2017), a inverse mapping from from latent to data space is learned and the generator is penalized based on the inverted distribution to cover all modes. Ravuri *et al.* (2018) proposes a moment matching paradigm different from VAEs or GANs. Arjovsky *et al.* (2017); Gulrajani *et al.* (2017a) proposes GANs which minimize the Wasserstein distance between true and generated distributions. Miyato *et al.* (2018) demonstrates improved results by applying Spectral Normalization on the weights. In Tran *et al.* (2018), distance constraints are applied on top. In Adler and Lunz (2018), WGANs were extended to Banach Spaces to emphasize edges or large scale behavior. Orthogonally, Karras *et al.* (2018) focus on progressively learning to use more complex model architectures to improve performance.

Conditional generative adversarial networks and future prediction. In the majority of the future prediction tasks, we are interested in modelling conditional distributions, conditioned on observations and contextual information. Extension of generative adversarial network to the conditional case was proposed by Mirza and Osindero (2014). Conditional Generative Adversarial Networks suffer from similar issues as their unconditional counterparts, e.g. mode collapse. Fortunately, many of the methods described above for the unconditional case are applicable also to the conditional case to mitigate mode collapse. On the other hand, recent works such as Yang *et al.* (2019) specifically target the mode collapse problem in conditional generative adversarial networks. Yang *et al.* (2019) propose to explicitly regularize the generator to produce diverse outputs depending on latent codes. Liu *et al.* (2021) uses contrastive learning for diverse conditional image synthesis. Chrysos *et al.* (2019) leverages structure in the target space of the model to address the issue of robustness to noise. Brock *et al.* (2019) shows very high quality conditional image generation at high resolutions.

Conditional generative adversarial networks have been used in diverse scenarios for future prediction. Mathieu *et al.* (2016) was the first to propose the use of adversarial loss for future frame prediction. This basic formulation has been further improved by Lee *et al.* (2018); Kwon and Park (2019); Liang *et al.* (2017) among others. Conditional generative adversarial networks have also been widely used for trajectory prediction tasks to deal with multi-modality. Social-GAN (Gupta *et al.*, 2018) proposes to use social pooling in a conditional GAN setup to capture the effect of social interactions in the distribution of future pedestrian trajectories. Sophie (Sadeghian *et al.*, 2019) uses an attention mechanism in a conditional GAN setup to capture the effect of interactions. Graph attention networks are employed by Kosaraju *et al.* (2019) in a conditional GAN setup.

2.3.2 Variational Autoencoders

Unconditional variational autoencoders. Similar to generative adversarial networks, variational autoencoders are deep latent variable models. However, unlike generative adversarial networks which do not explicitly maximize the likelihood of the data under the model, variational autoencoders maximize a lower bound on the data log-likelihood. This property is potentially helpful for capturing all modes of complex multi-modal data distributions.

Variational learning has enabled learning of such deep directed graphical models with Gaussian latent variables on large datasets (Kingma and Welling, 2014; Kingma *et al.*, 2014; Rezende *et al.*, 2014). Model training is made possible through stochastic optimization by the use of a variational lower bound of the data log-likelihood and the re-parameterization trick. In Burda *et al.* (2016), a tighter lower bound on the data log-likelihood is introduced and multiple samples are used during training which are weighted according to importance weights. It is shown empirically that the IWAE framework can learn richer latent space representations. However, the standard variational autoencoder framework (Kingma and Welling, 2014) uses unimodal Gaussian prior and posterior distributions. This induces a strong model bias Hoffman and Johnson (2016); Tomczak and Welling (2018) which makes it challenging to capture multi-modal distributions. Thereafter, two lines of work have focused on developing either more expressive prior or posterior distributions.

Rezende and Mohamed (2015) propose normalizing flows to model complex posterior distributions. Kingma *et al.* (2016); Tomczak and Welling (2016); van den Berg *et al.* (2018) present more complex inverse autoregessive flows, householder and Sylvester normalizing flow based posteriors.

Nalisnick and Smyth (2017) which proposes a Dirichlet process prior and Goyal *et al.* (2017) which proposes a nested Chinese restaurant process prior. However, these methods require sophisticated learning methods. In contrast, Tomczak and Welling (2018) proposes a mixture of Gaussians based prior (with fixed number of components) which is easier to train and shows promising results on some image generation tasks. Chen *et al.* (2017) proposes an inverse autoregressive flow based prior which leads to improvements in complex image generation tasks like CIFAR-10. (Ziegler and Rush, 2019) proposes a prior for VAE based text generation using complex non-linear flows which allows for complex multi-modal priors.

While these approaches focus on the unconditional case, in this thesis we focus primarily on the conditional case, which we discuss in the following.

Conditional variational autoencoders and future prediction. Conditional variational autoencoders Sohn *et al.* (2015) extend the VAE variational autoencoder of Kingma and Welling (2014) to model conditional distributions. Conditional variational autoencoders, similar to variational autoencoders, maximize a lower bound on the conditional data log likelihood. Note that conditional variational autoencoders suffer from similar issues in modelling multi-modality as the unconditional variational autoencoders. In this thesis, we focus on these issues in the conditional case and propose improved objective functions and flexible priors to better model multi-modal distributions.

This conditional framework has been used for a variety of future prediction tasks. Xue *et al.* (2016); Li *et al.* (2018) propose to use conditional variational autoencoders for future frame prediction. Lee *et al.* (2017b); Felsen *et al.* (2018); Zhang *et al.* (2019); Mangalam *et al.* (2020); Salzmann *et al.* (2020) use conditional variational autoencoders for trajectory prediction. In more detail, DESIRE (Lee *et al.*, 2017b) uses a RNN refinement module over the plain conditional variational autoencoder setup. Predictions that "personalize" to agent behaviour are proposed in Felsen *et al.* (2018). A social graph network is used for conditioning in Zhang *et al.* (2019). It is shown in Mangalam *et al.* (2020) that conditioning additionally on the goal state can significantly improve accuracy. Finally, Trajectron++ (Salzmann *et al.*, 2020) uses a pooled scene graph in a conditional variational autoencoder framework to capture interactions. In this thesis, we propose methods to improve the accuracy and diversity of trajectory prediction methods in highly multi-modal scenarios. We also propose methods which can model the effect of interactions directly in the latent space for improved accuracy.

2.3.3 Hybrid and Alternative Approaches

Hybrid approaches. As both generative adversarial networks and variational autoencoders have their respective shortcomings in modelling complex multi-modal distributions, recent works have proposed hybrid methods which aim to combine their strengths to improve overall performance especially for image distributions. In Larsen *et al.* (2016), a VAE-GAN hybrid is proposed with discriminator feature matching – the variational autoencoders decoder is trained to match discriminator features instead of a L_1/L_2 reconstruction loss. ALI (Dumoulin *et al.*, 2017) proposes to instead match the encoder and decoder joint distributions – with limited success on diverse datasets. BiGAN (Donahue *et al.*, 2017), builds upon ALI to learn inverse mappings from the data to the latent space and demonstrate effectiveness on various discriminative tasks. Rosca *et al.* (2017) extends standard VAEs by replacing the log-likelihood term with a hybrid version based on synthetic likelihoods. The KL-divergence constraint to the prior is also recast to a synthetic likelihood form, which can be enforced by a discriminator (as in Makhzani *et al.* (2016); Tolstikhin *et al.* (2018)). The second improvement is crucial in generating realistic images at par with classic/Wasserstein generative adversarial networks. In this thesis, we propose improved objective functions for training such hybrid framework for improved diversity and mode coverage while maintaining competitive image quality.

Alternative approaches. Due to shortcoming of both generative adversarial networks and variational autoencoders, some recent works have taken alternative approaches to deal with the challenge of multi-modality. Rhinehart *et al.* (2018); Deo and Trivedi (2019) introduce push-forward policies and motion planning for generative modelling of trajectories. Determinantal point processes are used in Yuan and Kitani (2020) to better capture diversity of trajectory distributions.

2.3.4 Exaction Inference Methods

Exact likelihood models, especially autoregressive models, normalizing flows and very recently diffusion models, have been recently considered to overcome the limitations of generative adversarial networks and variational autoencoders mostly in the context of image synthesis.

Autoregressive models. These are a class of exact inference models that factorize the joint probability distribution over the input space as a product of conditional distributions, where each dimension is conditioned on the previous ones in a pre-defined order (Chen *et al.*, 2018; Domke *et al.*, 2008; Graves, 2013; Hochreiter and Schmidhuber, 1997; Parmar *et al.*, 2018; van den Oord *et al.*, 2016a,b). Recent autoregressive models, such as PixelCNN and PixelRNN (van den Oord *et al.*, 2016a,b), can generate high-quality images but are difficult to parallelize since synthesis is sequential. It is worth noting that autoregressive image models, such as that of Domke *et al.* Domke *et al.* (2008), significantly pre-date their recent popularity. Various extensions have been proposed to improve the performance of the PixelCNN model. For example, Multiscale-PixelCNN (Reed *et al.*, 2017) extends PixelCNN to improve the sampling runtime from linear to logarithmic in the number of pixels, exploiting conditional independence between the pixels. Chen *et al.* (2018) introduce self-attention in PixelCNN models to improve modelling power. Salimans *et al.* (2017) introduce skip connections and a discrete logistic likelihood model. WaveRNN (Kalchbrenner *et al.*, 2018) leverages customized GPU kernels to improve the sampling speed for audio synthesis. Menick and Kalchbrenner (2019) synthesize images by sequential conditioning on sub-images within an image. These methods, however, still suffer from slow sampling speed and are difficult to parallelize, therefore in this thesis we focus primarily on normalizing flow based models.

Normalizing flows. These models first introduced in Dinh *et al.* (2015), also allow for exact inference. They are composed of a series of invertible transformations, each with a tractable Jacobian and inverse, which maps the input distribution to a known base density, e.g. a Gaussian. Papamakarios *et al.* (2017) proposed autoregressive invertible transformations using masked decoders. However, these are difficult to parallelize just like PixelCNN-based approaches. Kingma *et al.* (2016) propose inverse autoregressive flow (IAF), where the means and variances of pixels depend on random variables and not on previous pixels, making it easier to parallelize. However, the approach offers limited generalization van den Oord *et al.* (2018). Recent work Behrmann *et al.* (2019); Dinh *et al.* (2017); Kingma and Dhariwal (2018) extends normalizing flows Dinh *et al.* (2015) to multi-scale architectures with split couplings, which allow for efficient inference and sampling. For example, Kingma and Dhariwal (2018) introduce additional invertible 1×1 convolutions to capture non-linearities in the data distribution. Hoogeboom *et al.* (2019) extend this to $d \times d$ convolutions, increasing the receptive field. Chen *et al.* (2019) improve the residual blocks of flow layers with memory efficient gradients based on the choice of activation functions. A key advantage of flow-based generative models is that they can be parallelized for inference and synthesis. Ho *et al.* (2019) propose Flow++ with various modifications in the architecture of the flows in Dinh *et al.* (2017), including attention and a variational quantization method to improve the data likelihood. The resulting model is computationally expensive as non-linearities are applied along all the dimensions of the data at every step of the flow, i.e. all the dimensions are instantiated with the prior distribution at the last layer of the flow. Yu *et al.* (2020a) proposes a Haar-wavelet based decomposition for scalability on high resolution image data. Lu and Huang (2020) extended normalizing flows to model conditional distributions. While comparatively efficient, such flow-based models have limited expressiveness compared to autoregressive models, which is reflected in their lower data log-likelihood. The goal of this thesis is to develop models that have the expressiveness of autoregressive models and the efficiency of flow-based models.

Diffusion models. These models are a recently proposed class of latent variable models based on nonequilibrium thermodynamics. Diffusion based models are parameterized Markov chains trained to convert a latent noise distribution to the target data distribution and have shown remarkable performance on image datasets such as CIFAR-10 (Ho *et al.*, 2020). While sampling from these models remains computationally expensive compared to flow based models, they offer a promising direction of future research.

Relation to our work. In Chapter 6 and Chapter 7, we primarily deal with the unimodal Gaussian constraint on both the prior and posterior of conditional variational autoencoders. In Chapter 6, we concentrate on the fact that this unimodal Gaussian constraint leads to a mismatch between the prior and posterior distributions in case of multi-modal data distributions. We argue that this gap cannot be fully closed by the standard conditional variational autoencoder objective function. We consider a new multi-sample objective which relaxes the constraints on the recognition network by encouraging diverse sample generation and thus leads to a better match between the prior and posterior latent variable distributions. In Chapter 7, we propose a normalizing flow based prior for conditional variational autoencoders. This removes the unimodal Gaussian constraints and leads to significantly improved performance on multi-modal pedestrian trajectory distributions. In Chapter 8, we propose a novel objective for hybrid VAE-GAN based models. Similar to our objective introduced in Chapter 6, our multi-sample objective in Chapter 8 relaxes the constraints on the recognition network and leads to a better match between the prior and posterior latent variable distributions. In Chapter 9, we propose a novel conditional variational autoencoder framework to model the effect of interactions on the distribution of pedestrian trajectories. Our formulation models the joint distribution of all agents, e.g. pedestrians, bicyclists in the scene, through a shared latent space across all agents in the scene.

In Chapter 10 and Chapter 11, we focus on exact inference models, in particular normalizing flows due to their advantage of efficiency. We focus on improving the modelling flexibility of normalizing flow based models. In Chapter 10, we propose an auto-regressive prior for normalizing flow based models. Our auto-regressive prior is applied channel-wise in a computationally efficient setup. In Chapter 11, we propose to use a Haar-wavelet based decomposition for normalizing flows on trajectory data. The Haar-wavelet based decomposition allows us to model the fine components of the decomposition conditioned on the coarse components which provides global context, leading to improved modelling flexibility.

2.4 DATASETS FOR TRAJECTORY PREDICTION

Datasets like ETH/UCY (Lerner *et al.*, 2007) and Stanford Drone (Pellegrini *et al.*, 2009) are among the first datasets in the field of pedestrian trajectory prediction. However, they are recorded using a bird's eye view camera or drone. In order to aid the development of autonomous driving capabilities, recent datasets have moved to a more realistic "on-board" setting – recorded from a (ego-)vehicle. The popular "on-board" datasets: nuScenes (Caesar *et al.*, 2020), Argoverse (Chang *et al.*, 2019) and Lyft L5 (Houston *et al.*, 2020) focus primarily on trajectories of nearby vehicles. nuScenes (Caesar *et al.*, 2020) and Argoverse (Chang *et al.*, 2019) do not include annotated pedestrian (bicyclist) trajectories in their test set. Although Lyft L5 (Houston *et al.*, 2020) includes pedestrian and bicyclist trajectories in the test set they are relatively rare (5.91% and 1.62% of all trajectories) as the chosen route does not include significant portions of dense urban environments. Moreover, in comparison to nuScenes (Caesar *et al.*, 2020) and Argoverse (Chang *et al.*, 2019), Lyft L5 (Houston *et al.*, 2020) has lower diversity in terms of locations as it recorded only along a fixed route ($\sim$6km) in Palo Alto, California. Additionally, Lyft L5 (Houston *et al.*, 2020) does not provide images from cameras and lidar point clouds, which are sources of rich contextual information. In contrast, PIE (Rasouli *et al.*, 2019), TITAN (Malla *et al.*, 2020) and TRAF (Chandra *et al.*, 2019) focus primarily on pedestrians (bicyclists). However, the trajectories are recorded as sequences of 2d bounding boxes in the image plane and are not localized in the 3d world. The ApolloScapes (Ma *et al.*, 2019) dataset does not include the trajectory of the ego-vehicle or contextual information, e.g. images from cameras or lidar point clouds. Finally, note that these datasets are recorded either in North America or Asia and no large scale trajectory datasets are available for Europe. Euro-PVI is the first large scale dataset recorded in Europe dedicated to the task of trajectory prediction and unlike the existing datasets focuses on interactions between the ego-vehicle and pedestrian (bicyclist).

Relation to our work. In Chapter 9, we propose a novel pedestrian trajectory prediction dataset, Euro-PVI. Most popular "on-board" datasets, e.g. (Caesar *et al.*, 2020; Chang *et al.*, 2019; Houston *et al.*, 2020), are recorded either in North America or Asia and no large scale trajectory datasets are available for Europe. Euro-PVI is the first large scale dataset recorded in Europe dedicated to the task of trajectory prediction and unlike the existing datasets focuses on interactions between the ego-vehicle and pedestrian (bicyclist).

LONG-TERM IMAGE BOUNDARY PREDICTION 3

Contents

Boundary estimation in images and videos has been a very active topic of research, and organizing visual information into boundaries and segments is believed to be a cornerstone of visual perception. While prior work has focused on estimating boundaries for observed frames, in this chapter our aim is to predict boundaries of future unobserved frames. This requires our model to learn about the fate of boundaries and corresponding motion patterns – including a notion of "intuitive physics". We experiment on natural video sequences along with synthetic sequences with deterministic physics-based and agent-based motions. While not being our primary goal, we also show that fusion of RGB and boundary prediction leads to improved RGB predictions.

3.1 INTRODUCTION

In this chapter, we propose the task of predicting future scene boundaries. Scene boundaries capture the important structure and extents of objects. Moreover, they can be accurately estimated Khoreva *et al.* (2016). Prediction of future scene boundaries requires understanding of object dynamics and motion patterns including an intuitive understanding of physical laws or "intuitive physics". In this work, we focus on two particular scenarios involving motion and local interactions. The first one, which we call physics-based motion (deterministic), can fully be described by the laws of physics, e.g. dynamics of billiard balls. The second one, which we call agent-based motion (non-deterministic), also involves understanding of intentions, e.g. dynamics of an ice-skater. Therefore, our methods have to deal with diverse situations, work on

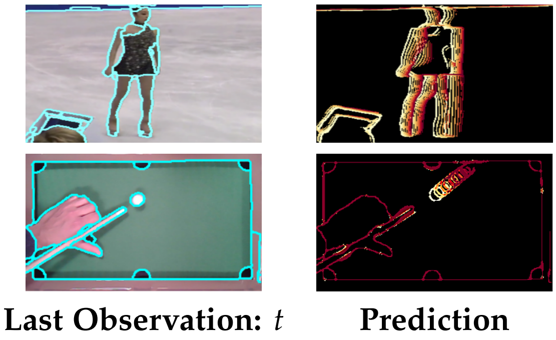


Figure 3.1: Predicted future boundary images, from t + 1 (Yellow) to t + 8 (Row 1), t + 18 (Row 2) (Red), superimposed.

raw pixels, and should be capable of long-term predictions. Fig. 3.1 shows example results of our method that accurately predicts future scene boundaries.

Recently, full future frame prediction of observed scenes has been studied (Mathieu *et al.*, 2016; Liu *et al.*, 2017b). But up to now, only very short range predictions of few frames have been shown, where blurriness/distortion artifacts occur in the predicted future frames – losing/incorrectly propagating high-frequency information. This high frequency information is crucial for meaningful predictions about the future, e.g. on a billiard table the location of a ball and table boundaries are necessary to infer the future state of the table. Boundaries capture this crucial high frequency information and are also known to reveal important structures of the visual scene (Wertheimer, 1923; Arbelaez *et al.*, 2011; Galasso *et al.*, 2013). Therefore, we argue that the task of future boundary prediction is a more suitable benchmark for understanding and predicting physics or agent-based motion.

The main contributions of this chapter are as follows, 1. We propose the novel task of future boundary prediction. 2. We propose the first method that predicts future boundaries based only on the raw pixels. 3. We evaluate our model on two scenarios involving deterministic physics-based (synthetic and real billiard sequences) and non-deterministic agent-based motion (VSB100, Galasso *et al.* (2013)). 4. Under the physics-based scenario, the method shows for the first time long-term predictions. 5. Under the agent-based scenario on VSB100 and UCF101, we show that the predicted boundaries can be used in a fusion scheme that improves RGB video prediction in the longer-term.

3.2 RELATED WORK

While we provide a broader discussion on related work in Chapter 2, here we discuss related work relevant to this chapter.

Video segmentation. Video segmentation as the task of finding consistent spatio-temporal boundaries in a video volume has received significant attention over the last years (Galasso *et al.*, 2014; Ochs *et al.*, 2014; Galasso *et al.*, 2013; Chang *et al.*, 2013), as it provides an initial analysis and abstraction for further processing. In contrast, our approach aims at predicting these boundaries into the future without any video observed for future frames.

3.3 MODEL

We present a model that observes a sequence of boundary images, where each pixel encodes the confidence of occurrence of an image boundary at that location and then predicts the boundary image(s) at the next time-step(s). An overview of our Convolutional Multi-Scale Context (CMSC) model is shown in Fig. 3.2.

We approach long term prediction by recursion, due to the advantage of efficiency. However, errors are potentially propagated and accumulated over time. In order to mitigate such effects, we need our model to be accurate and to consolidate information over time. Our model has been designed through analysis of prior work on the related task of frame prediction, to maximize accuracy. Furthermore, our model has many novel aspects which are key to long term prediction.

In order to generalize across diverse sequences while maintaining a tractable number of parameters, a patch based approach is adopted. Therefore, our model observes and predicts on patches rather than the complete input image. Alternatively, this can be seen as multiple replicas ("patch predictors") of our model predicting on patches of the input sequence. We now describe our model through an analysis of its various components.

3.3.1 Fully Convolutional

Our CMSC model consists of only convolutional layers. The input boundary image sequence is concatenated as channels and is read by the first convolutional layer. Convolutional layers can extract high quality location invariant features. In particular, they can extract information about the orientation and direction of motion of boundaries. Neurons at upper convolutional layers have larger receptive fields and can aggregate information. In fact, as shown by the work of Jain *et al.* (2007), the output layer should have a wide receptive field to preserve long range spatial and temporal dependencies and learn about interaction among boundaries in a spatio-temporal context. We therefore use several convolutional layers in our CMSC model. We also introduce pooling in between convolutional layers. Pooling further helps in the aggregation of information and increases receptive fields. However, excessive pooling (or tight bottlenecks with fully connected layers) have been shown to be successful in classification tasks, but also have shown by Ranzato *et al.* (2014) to induce image degradations for synthesis tasks. Therefore, it is crucial to use moderate pooling. Finally, we use up-sampling layers after pooling to maintain resolution.

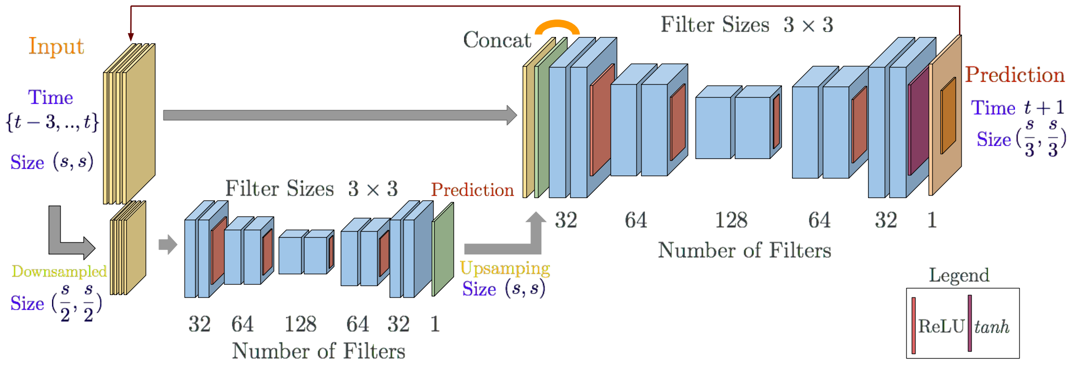


Figure 3.2: Convolutional Multi-scale with Context architecture (only 2 out of 4 scales illustrated).

3.3.2 Multiple Scale Prediction

Multi-Scale model architectures akin to a Laplacian pyramid have shown to be advantageous for generating natural images (Denton *et al.*, 2015) and predicting future rgb frames (Mathieu *et al.*, 2016). Such model architectures contain multiple levels which observes the input boundary image(s) at increasing (coarse to fine) scales. Down-sampling a boundary image would have the effect of smoothing and discarding details of a boundary image. Therefore, it would be easier to predict future boundary images at a coarser resolution. Our CMSC model also uses multiple scales (or levels). The input $I(L_{2k})$ to a certain level (L_{2k}) is the input boundary image sequence scaled to the current level X_{2k} and the boundary image O predicted by the previous coarser level (L_k). The boundary image predicted by the coarser level is upsampled $\hat{O}$ to the scale at the current level. We have,

$$I(L_{2k}) = \{X_{2k}, \hat{O}(L_k)\}$$

The coarse predicted boundary images $\hat{O}(L_k)$ act a guide for the next higher level of the model.

In detail, we use four levels, with scales increasing by a factor of two. Each level of the model consists of five sets of two convolutional layers. They are of a constant size 3x3. We introduce a pooling layer after the first three sets of convolutional layers. We double the number of convolutional kernels after pooling. There are 32, 64, 128, 64 and 32 filters respectively in each set. We upsample the convolutional maps after the third set to maintain resolution. We use *ReLU* non-linearities between every layer except the last. We use the *tanh* non-linearity at the end to ensures output in [0,1].

For accurate long term prediction, it is crucial to ensure global consistency through communication between the patch predictors. Consider a video of a moving ball. The trajectory of a ball might intersect with multiple patches. To correctly predict the motion far into the future, replicas of the model predicting on neighboring patches need to be consistent especially during transition of the ball between patches. Therefore, we describe next the final component of our CMSC model, the context, which ensures global consistency.

3.3.3 Context

Our CMSC model observes a central patch along with the directly neighbouring 8 patches. This neighbourhood is called the context. However, the model only predicts on the central patch. While predicting recursively, the model observes its previous output along with the the output of the neighboring patch predictors. This enables the learning of spatially consistent predictions while keeping the same number of parameters.

The addition of a context has the added advantage that the output layer neurons now have receptive fields that are uniform in size. Without context, the neurons at the boundary of the (2D) output layer have a smaller receptive field compared to the neurons at the center. This leads to a non-uniform (training and test) error distribution at the output layer neurons. In Fig. 3.3 we plot the average error at the output layer neurons of our CMSC model at increasing distance from the patch border, with and without context. Error increases consistently from patch center (right) to the patch border (left) without a context. Note that, the model of Mathieu *et al.* (2016) is also multi-scale and fully convolutional like CMSC, but it does not have pooling or context.

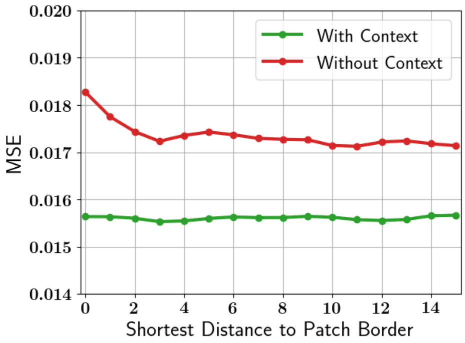


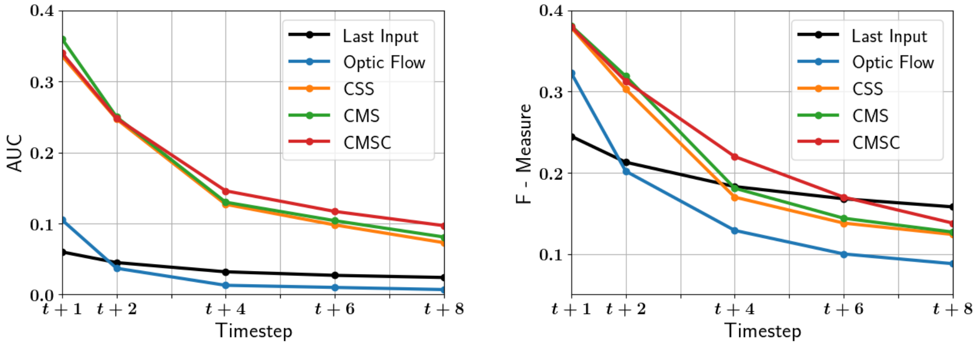
Figure 3.3: Our model without context has higher error near the patch boundary (red) vs. with context (green).

Next, we evaluate our CMSC model and the effectiveness of its various components.

3.4 EXPERIMENTS

We evaluate our CMSC model on natural video sequences involving agent-based motion and billiard sequences with only physics-based motion. We compare with various baselines and perform ablation studies to confirm design choices. We convert each video into 32x32 pixel patches. The CMSC model observes a central patch and eight neighbouring patches resulting in a context of size 96x96 pixels.

Training loss. We use L2 loss (mean square error) during training, which we optimize using the ADAM optimizer.



(a) Area under the curve.

(b) Best F-measure.

Figure 3.4: Evaluation of boundary prediction on VSB100.

Evaluation metric. As we want sharp and accurate boundaries, we use the established boundary precision recall (BPR) evaluation metric from the video segmentation literature (Galasso *et al.*, 2013). This metric is defined for a set P of predicted boundary images and G of corresponding ground truth boundary images as,

$$P = \frac{\sum_{B_p \in P, B_g \in G} \mid B_p \cap B_g \mid}{\sum_{B_p \in P} \mid B_p \mid}$$

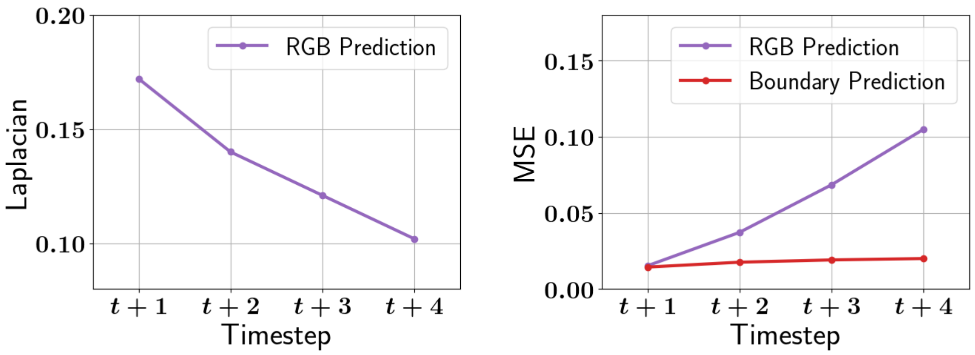
$$R = \frac{\sum_{B_p \in P, B_g \in G} \mid B_p \cap B_g \mid}{\sum_{B_g \in G} \mid B_g \mid}$$

$$F = \frac{2PR}{P+R},$$

where P is boundary precision, R is boundary recall and F is the combined F-measure. As we are interested in accurate predictions, predicted boundary pixels should be at most 1 pixel away from ground-truth boundary pixels to be correct.

3.4.1 Evaluation on Sequences with Agent-based Motion

Dataset and training. We use the VSB100 dataset which contains 101 videos with a maximum 121 frames each. The training set consists of 40 videos and the test set consists of 60 videos. The videos contain a wide range of objects of different sizes and shapes, including vehicles, humans and animals. The videos also have a wide variety of both object and camera motion. We use the hierarchical video segmentation algorithm in Khoreva *et al.* (2016) to segment these videos. The output is a ultra-metric contour map (ucm). Boundaries higher in the hierarchy typically correspond to semantically coherent entities like animals, vehicles etc and therefore their motion corresponds to object/camera motion. We discard boundaries belonging to the lowest level of the hierarchy (corresponding to an over-segmentation), as they



(a) Laplacian measure.

(b) Mean squared error.

Figure 3.5: RGB versus boundary prediction.

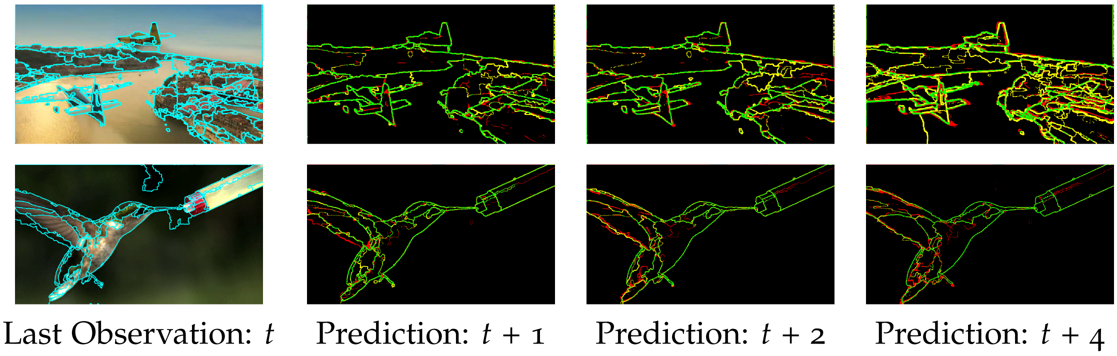


Figure 3.6: Rows top to bottom: Prediction on *airplane* and *hummingbird* sequences from VSB100. Correct boundaries predictions are encoded in green. Missed boundaries are encoded in yellow. Wrong boundaries are encoded in red.

are temporally very unstable. We use the ucm hierarchy as a confidence measure on boundary location at a pixel.

Experimental settings and baselines. The models are trained to predict boundaries of segmented VSB100 videos. Recall that the ground-truth boundaries (ucm) in VSB100 have different confidence values. Thus, we threshold the predictions before comparison to the groundtruth. We vary the threshold to obtain a precision-recall curve and report the area under the curve (AUC) along with the best F-measure across all thresholds. We include a "Last Input" baseline by using the last input frame as constant prediction and a "Optical flow" baseline. As many boundaries do not change between frames in the videos of VSB100, the last input is a strong baseline especially when we are predicting one step into the future. In case of the optic flow baseline, the optic flow is calculated between the last two input frames (at t - 1 and t) using the Epic flow method of Revaud *et al.* (2015). The boundary pixels at time t are propagated using the calculated flow to generate predictions at t + 1 to t + 8.

Results on VSB100. We perform an ablation study of our CMSC model and

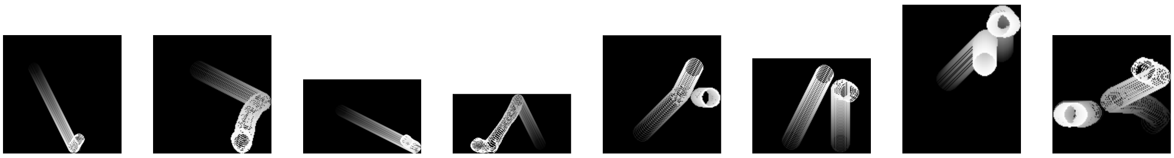
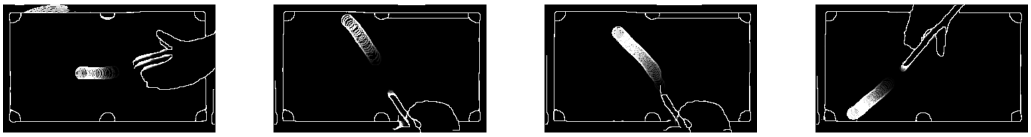


Figure 3.7: Trails produced by super-imposing predicted boundaries on synthetic sequences.



Trail up to t + 20 Trail up to t + 20 Trail up to t + 50 Trail up to t + 50

Figure 3.8: Trails produced by super-imposing predicted boundaries on real sequences.

we compare to, 1. A convolutional single scale model (CSS) 2. A convolutional multi-scale model (CMS), in addition to the baselines. Both models do not have a context. We report the quantitative results in Fig. 3.4(a) and Fig. 3.4(b) and the qualitative results in Fig. 3.6.

Quantitative evaluation: In the short term the CMS model (green lines) performs well. However, our CMSC (red lines) performs best in the longer term (both having the same number of parameters). This demonstrates the importance of the context for long-term prediction. The good performance of both of the mutli-scale models (CMS and CMSC) versus the single scale CSS model, shows that multiple scales lead to more accurate predictions. The performance advantage of our CMSC model over the last input baseline shows that the model learns to predict boundaries of moving objects while keeping static boundaries intact. The recall of the CMSC model declines with time as the future becomes increasingly uncertain. The poor performance of the "Optic flow" baseline is due to inaccurate flow information at object boundaries.

Qualitative evaluation: The boundaries produced by our CMSC model are sharp whenever the motion is smooth, e.g. the predictions in Fig. 3.6. However, the models are not able to deal with high uncertainty in the long-term often due to non-deterministic motion. The models in such situations react by blurring the boundaries, as a consequence of using the L2 training loss. While predicting recursively, this leads to loss of boundary confidence and eventual vanishing boundaries. The "Optic flow" baseline produces discontinuous (jagged) boundaries. Next we evaluate and compare RGB prediction to boundary prediction.

RGB verses boundary prediction. We report the sharpness of RGB frames (of VSB100) predicted by the adversarial model of Mathieu *et al.* (2016) (fine-tuned on VSB100) using the Laplacian measure Krotkov (2012) in Fig. 3.5(a). The Laplacian measure pools the gradient information of the image. We observe that the model

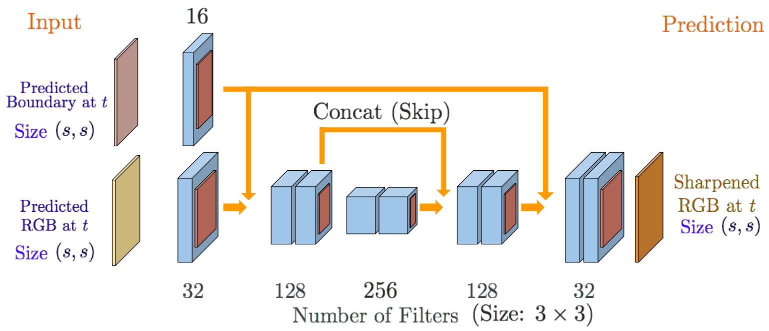


Figure 3.9: Our fusion model architecture.

of Mathieu *et al.* (2016) makes increasingly blurry predictions into the future. We also compare the mean squared error of RGB predictions of Mathieu *et al.* (2016) and predicted boundaries of our CMSC model in Fig. 3.5(b). We see a sharper increase in the error of RGB predictions compared to boundaries in the long term.

Step	Last Input	CMS	CMSC-BL	CMSC
$t+1$	0.141	0.282	0.957	**0.987**
$t+5$	0.038	0.101	0.841	**0.900**
$t+20$	0.002	0.066	0.347	**0.632**

Table 3.1: Evaluation on single ball billiard table worlds.

3.4.2 Evaluation on physics-based motion

Motion in the videos in the VSB100 dataset is frequently very complex as agent's actions quickly become non-deterministic and hence increasingly uncertain. Therefore, we also look at physics-based motion, which is still challenging yet it factors out the aforementioned issues. In this scenario, we evaluate the long-term prediction performance of the models on real and synthetic billiard ball sequences. We begin by describing our dataset.

Synthetic data generation. The synthetic billiard ball sequences are sampled from worlds which consist of balls moving on a frictionless surface with a border, akin to a billiard table. We used pygame to create such worlds and sample boundary images from them. The output images contain boundaries that can stem from ball(s) or the table and have binary confidence measures (indicating a boundary at that location). During evaluation, as the target is always a binary image, we report only the best F-measure obtained by thresholding the predicting boundary images and varying the threshold parameter. We sampled synthetic billiard sequences using the following parameters. 1. *Table size*: Side length randomly sampled from

	– Evaluation on three ball worlds –			– Evaluation on six ball worlds –		
Step	Last Input	CMSC-2B	CMSC	Last Input	CMSC-3B	CMSC
$t + 1$	0.246	0.967	**0.968**	0.250	0.962	**0.964**
$t + 5$	0.118	0.890	**0.892**	0.130	**0.875**	0.866
$t + 20$	0.090	0.664	**0.700**	0.115	0.511	**0.600**

Table 3.2: Evaluation on complex billiard table worlds.

{96,128,160,192,256} pixels. 2. *Ball velocity*: Randomly sampled from [{-3,..,3},{-3,..,3}] pixels. 3. *Ball size*: Constant, with a radius of 13 pixels. 4. *Initial Position*: Uniformly over the table surface.

Real data collection. We captured a novel data-set of real billiard sequences on a mini-billiard table. Frame rate was set to 120 per second to minimize motion blur. Each sequence consists of an actor (not visible) striking the ball with a cue stick once. The motion in the sequences of the dataset are that of the cue stick and the balls. We produce boundary images using the method of Maninis *et al.* (2018).

Evaluation on synthetic single ball worlds. We generate a training set using parameters in Section 3.4.2. However, to keep our training set as diverse as possible we prefer short sequences. We restrict each sequence to a maximum length of one or two collisions with walls and set a 50% bias of the initial position of the balls being 40 pixels from the walls. We sample 500 such sequences and train the models on these sequences. We then test the models on 30 independent test sequences. We again include the "last input" baseline as a constant predictor . We also include a "blind" Convolutional multi-scale Context model (CMSC-BL), which cannot see the table borders. This is a strong baseline as starting from 42% frames in the test set, there are no ball-wall collisions 20 steps into the future. To beat this baseline, our models need to learn the physics of ball-wall collisions. We report the results in Table 3.1.

Our CMSC model performs the best with accurate predictions 20 time-steps into the future – also exceeding the "blind" version (CMSC-BL) that cannot handle ball-wall collisions. The model without a context CMS, produces inaccurate results at patch borders and thus suffers heavily especially at larger time-steps.

Evaluation on synthetic two and three ball worlds. Worlds with more than one ball also involve harder to model physics of ball-ball collisions. To evaluate the models on such worlds we sample 100 training sequences each with two, three and six balls respectively with a maximum length of 200 frames. We use a curriculum learning approach (Bengio *et al.*, 2009), where we initialize the models with the weights learned on single, two and three ball worlds respectively. We test the models on 30 independent sequences containing two, three and six balls respectively. We report the results in Table 3.2. In each case, we also include CMSC models trained on single ball worlds (CMSC-1B), two ball worlds (CMSC-2B) and three ball worlds

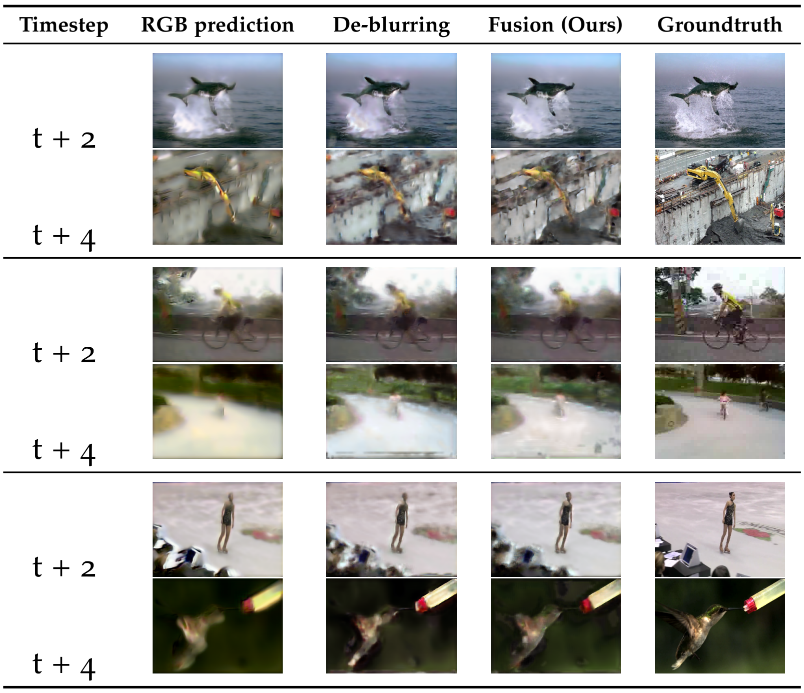


Figure 3.10: Sharpening RGB predictions using our Fusion scheme on VSB100 (top two rows) and on UCF101 (bottom two rows).

(CMSC-3B) respectively as baselines. To beat these strong baselines learning the physics of ball-ball collisions is necessary as in case of our two-ball and three-ball test sets, there are no ball-ball and 3-ball collisions 20 steps into the future for 92% and 98% of the starting frames (and no 6-ball collisions). Again, we see accurate prediction by the CMSC model even at 20 time-steps in the future.

Prediction over very long time scales. Although we evaluate only 20 timesteps into the future in Table 3.1 and Table 3.2, our models are stable over longer time-horizons. In Fig. 3.7, we predict 100 timesteps and visualize the boundary images by trails obtained by superposition. We notice a few failure cases where a ball reverse direction mid table and the ball(s) get deformed or disappear.

Step	Last Input	CMSC	Last Input(M)	CMSC(M)
$t + 1$	0.890	0.850	0.126	0.570
$t + 5$	0.855	0.804	0.085	0.541
$t + 20$	0.844	0.746	0.087	0.497

Table 3.3: Evaluation on real billiard sequences (M-masked).

Evaluation on real billiard sequences. Prediction on real billiard table sequences is a challenging test for our models. The table fabric causes rapid deceleration of the ball (compared to the constant velocity in the synthetic sequences). Spin is sometimes inadvertently introduced and a segmentation algorithm applied on the

	—— PSNR ——			—— Sharpness Loss ——			—— Laplacian Measure ——		
Step	**RGB prediction**	**De-blurring**	**Fusion (Ours)**	**RGB prediction**	**De-blurring**	**Fusion (Ours)**	**RGB prediction**	**De-blurring**	**Fusion (Ours)**
				VSB100					
$t + 2$	24.4	24.5	**25.1**	18.5	18.5	**18.6**	0.142	0.139	**0.155**
$t + 3$	22.2	22.9	**23.1**	18.2	18.2	**18.3**	0.121	0.109	**0.127**
$t + 4$	20.4	21.7	**22.3**	18.1	18.1	**18.2**	0.103	0.114	**0.118**
				UCF101					
$t + 2$	26.5	27.7	**28.2**	21.4	21.5	**21.7**	0.101	0.122	**0.136**
$t + 3$	23.4	25.1	**25.2**	20.5	20.8	**20.9**	0.095	0.093	**0.102**
$t + 4$	21.4	23.4	**23.8**	20.4	20.5	**20.6**	0.089	0.101	**0.112**

Table 3.4: Evaluation of our Fusion scheme. PSNR, Sharpness Loss and Laplacian measure: Higher is better.

observed frames introduce artifacts. The boundaries are not always consistent across frames of a sequence and they are jagged and change shape. We collect 350 real billiard sequences, with one ball, as our training set. To deal with deceleration, we experiment with increasing the number of input frames. We train our CMSC model with six input frames and pre-train on our synthetic one ball training set. We report the results of evaluation using F-measure on 30 independent sequences in Table 3.3. As many boundaries (e.g. table borders) remain static the last input baseline performs very well. For fair comparison we use a mask obtained with a ball tracker, Our method is able to propagate the motion of the ball and beats the last input baseline in the masked case. We show qualitative results in Fig. 3.8 as trails, where our model predicts 20 and 50 time-steps into the future.

3.5 SHARPENING RGB PREDICTIONS WITH FUSION

The sharp boundaries produced by our models raise the prospect of sharpening RGB predictions in a fusion scheme. We present our fusion architecture in Fig. 3.9, which fuses RGB predictions of Mathieu *et al.* (2016) with our boundaries. Note that, our approach can be used on top of any RGB frame prediction method and unlike Villegas *et al.* (2017) is video domain agnostic. It is inspired by prior work (Eigen *et al.*, 2013; Mao *et al.*, 2016) on deblurring/denoising. Like these models our fusion model is fully convolutional. Resolution is maintained by skip connections, as in Mao *et al.* (2016). Our fusion model takes as input the predicted RGB and boundaries at each timestep and is trained with L2 loss.

Datasets and metrics. We evaluate on both VSB100 and UCF101 datasets. We randomly select 30 and 20 videos from VSB100 to train our CMSC model and our fusion model. We test on the remaining 50 videos. Similarly we randomly select 1000, 500 (training) and 1000 (test) videos from UCF101. The UCF101 train/test set was segmented using the method of Maninis *et al.* (2018). We use PSNR, the sharpness loss measure from Mathieu *et al.* (2016) and the Laplacian measure as evaluation metrics.

Baselines. We include a baseline de-blurring model. It has the same architecture as our fusion model, except for the top block. This baseline aims to de-blur RGB predictions without observing our predicted boundaries.

Evaluation. We observe improved and sharper RGB predictions (see Table 3.4) *. Our fusion model learns to reintroduce lost high frequency information.

3.6 CONCLUSION

We propose the novel task of boundary prediction and demonstrate accurate results with our CMSC model. We argue for the key design choices, 1. A wide receptive field allowing the model to learn complex spatio-temporal dependencies. 2. Accurate prediction at each time-step with a fully convolutional setup without any bottleneck layers. 3. The context which allows for information sharing thus leading to global consistency. We obtain sharp predictions using L2 loss (in contrast to RGB prediction, which leads to very blurry results with L2 loss). Predictions by our CMSC model on diverse scenarios shows that it developed a data-driven model of future boundary motions over long time horizons. This includes dynamics of moving agents and billiard balls. Moreover, while not being our primary goal, our predicted boundaries lead to sharper RGB video predictions via a fusion-based approach.

* Corresponding results in Table 5 in Mathieu *et al.* (2016). We do not use motion masking as we would like our model to keep still boundaries intact.

LONG-TERM ON-BOARD PREDICTION OF PEOPLE IN TRAFFIC SCENES UNDER UNCERTAINTY 4

Contents

Progress towards advanced systems for assisted and autonomous driving is leveraging recent advances in recognition and segmentation methods. Yet, we are still facing challenges in bringing reliable driving to inner cities, as those are composed of highly dynamic scenes observed from a moving platform at considerable speeds. Anticipation becomes a key element in order to react timely and prevent accidents. In this chapter, we argue that it is necessary to predict at least 1 second and we thus propose a new model that jointly predicts ego motion and people trajectories over such large time horizons. We pay particular attention to modeling the uncertainty of our estimates arising from the non-deterministic nature of natural traffic scenes. Our experimental results show that it is indeed possible to predict people trajectories at the desired time horizons and that our uncertainty estimates are informative of the prediction error. We also show that both sequence modeling of trajectories as well as our novel method of long term odometry prediction are essential for best performance.

4.1 INTRODUCTION

While methods for automatic scene understanding have progressed rapidly over the past years, it is just one key ingredient for assisted and autonomous driving. Human capabilities go beyond inference of scene structure and encompass a broader type of scene understanding that also lends itself to anticipating the future.

As discussed in Chapter 1, anticipation is key in preventing collisions by predicting future movements of dynamic agents, e.g. people and cars in inner cities. It is

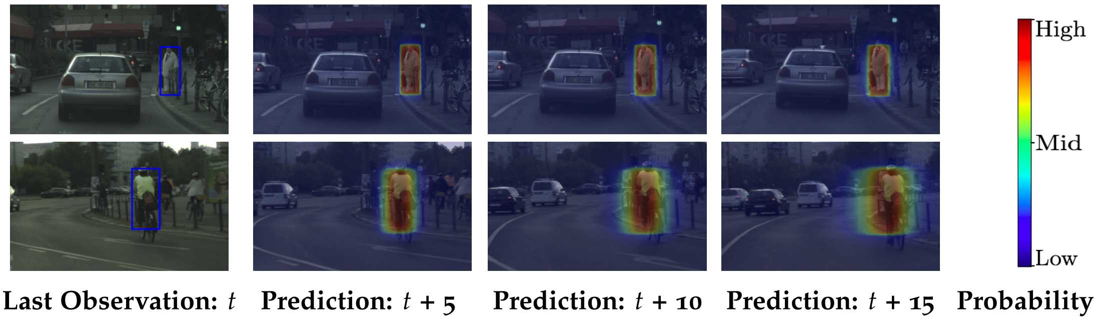


Table 4.1: Our predictive distribution upto $t + 15$ frames. The heat map encodes the probability of a certain pixel belonging to the person. The variance of the distribution encodes the uncertainty. *Row 1*: Low uncertainty. *Row 2*: High uncertainty.

also the key to operating at practical safety distances. Without anticipation, domain knowledge and experience, drivers would have to maintain an equally large safety distance to all objects, which is clearly impractical in dense and cluttered inner city traffic. Additionally, anticipation enables decision making, e.g. passing cars and pedestrians while respecting the safety of all participants. Even at conservative and careful driving speeds of 25miles/hour ($\sim$ 40km/hour) in residential areas, the distance traveled in 1 second corresponds roughly to the braking distance. Anticipation of traffic scenes on a time horizon of *at least* 1 second would therefore enable safe driving at such speeds.

We propose the first approach to predict people (pedestrians including cyclists) trajectories from on-board cameras over such long-time horizons with uncertainty estimates. Due to the particular importance for safety, we are focusing on the people class. While pedestrian trajectory prediction has been approached in prior work, we propose the first approach for on-board prediction. As predictions are made with respect to the moving vehicle, we formulate a novel two stream model for long-term person bounding box prediction and vehicle ego motion (odometry). In contrast to prior work, we model both *aleatoric* (observation) uncertainty and *epistemic* (model) uncertainty (Der Kiureghian and Ditlevsen, 2009) in order to arrive at an estimate of the overall uncertainty.

Our contributions in this chapter in detail are: 1. First approach to long-term prediction of pedestrian bounding box sequences from a mobile platform; 2. Novel sequence to sequence model which provides a theoretically grounded approach to quantify uncertainty associated with each prediction; 3. Detailed experimental evaluation of alternative architectures illustrating the importance and effectiveness of using a two-stream architecture; 4. Analysis of dependencies between uncertainty estimates and actual prediction error leading to an *empirical error bound*.

4.2 RELATED WORK

While we provide a broader discussion on related work in Chapter 2, here we discuss related work relevant to this chapter.

Human trajectory prediction. Recent works such as Keller *et al.* (2011); Rehder and Kloeden (2015) focus on the task of pedestrian trajectory prediction in 3D space. Initial trajectories and obstacle occupancy maps are obtained by dense stereo matching, assuming a linear road model of fixed width. However, 3D coordinates and obstacle maps obtained from stereo matching can be very noisy especially in unknown environments. Moreover, evaluation is on sequences with linear or no vehicle ego-motion. Our method does not depend upon unreliable 3D coordinates and needs no assumptions about scene geometry and vehicle ego-motion. Another class of models such as Helbing and Molnar (1995); Yamaguchi *et al.* (2011); Robicquet *et al.* (2016); Alahi *et al.* (2016); Trautman *et al.* (2013) consider the problem of pedestrian trajectory prediction in a social context by modelling human-human interactions. However, in the case of on-board prediction vehicle ego-motion dominates social aspects. Moreover, most methods are trained/tested on static camera datasets which are hand annotated with minimum observation noise. Apart from these, the class of models such as Hu *et al.* (2007); Kim *et al.* (2011); Morris and Trivedi (2011); Zhou *et al.* (2011); Zhang *et al.* (2013) aim at discovering motion patterns of humans and vehicles. Such methods cannot be used for trajectory prediction and do not consider vehicle ego-motion.

Assisted and autonomous driving. One of the earliest works on vehicle ego-motion (odometry) prediction or popularly, autonomous driving, was ALVINN by Pomerleau (1989). This work showed the possibility of directly predicting steering angles from visual input. This system used a simple fully-connected network. More recently, Bojarski *et al.* (2016) uses a convolutional neural network for this task and achieves a autonomy of 90% using a relatively small training set. However, the focus is on highway driving. Xu *et al.* (2017) proposes a FCN-LSTM that predicts the next vehicle odometry based on the visual input captured by an on-board camera and previous odometry of the vehicle. Here, a diverse crowd sourced dataset is used. However, these methods predict vehicle odometry (e.g. steering angle) only for the next time-step. In contrast, we focus on inner-city driving and predict multiple time-steps into the future. Santana and Hotz (2016) proposes a driving simulator that predicts the future in the form of frames, based on the current and past visuals observed from an on-board camera. It is well known that future frame prediction suffers from blurriness problems. In the long-term important details get lost (Mathieu *et al.*, 2016). We predict the future in terms of bounding box coordinates which remain well defined by design in the long-term.

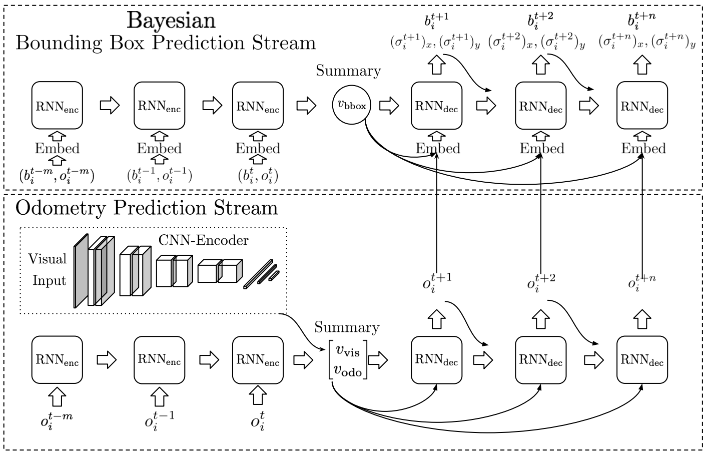


Figure 4.1: Two stream architecture for prediction of future pedestrian bounding boxes.

4.3 ON-BOARD PEDESTRIAN PREDICTION UNDER UNCERTAINTY

In order to anticipate motion of people in real-world traffic scenes from on-board cameras, we propose a novel approach that conditions the prediction of motion of people on predicted odometry (Section 4.3.4). Moreover, our approach models both *aleatoric* and *epistemic* uncertainty. Our model (see Fig. 4.1) consists of two specialized streams for prediction of pedestrian motion and odometry. The odometry specialist stream predicts the most likely future vehicle odometry sequence. The bounding box specialist stream consists of a novel Bayesian RNN encoder-decoder architecture to predict odometry conditioned distributions over pedestrian trajectories and to capture epistemic and aleatoric uncertainty. Bayesian probability theory provides us with a theoretically grounded approach to dealing with both types of uncertainties (Section 4.3.2).

We start by describing the bounding box prediction stream of our model and introduce our novel Bayesian RNN encoder-decoder which provides theoretically grounded uncertainty estimates.

4.3.1 Prediction of Pedestrian Trajectories

A bounding box corresponding to the i^{th} pedestrian observed on-board a vehicle at time step t can be described by the top-left and bottom-right pixel coordinates: $b_i^t = \{(x_{tl}, y_{tl}), (x_{br}, y_{br})\}$. We want to predict the distribution of future bounding box sequences $\mathrm{B_f}$ (where $|\mathrm{B_p}| = m$) of the pedestrian. We condition our predictions on the past bounding box sequence $\mathrm{B_p}$, the past odometry sequence $\mathrm{O_p}$ and the corresponding future odometry sequence $\mathrm{O_f}$ of the vehicle. The future odometry sequence $\mathrm{O_f}$ is predicted conditioned on the past odometry sequence $\mathrm{O_p}$ and on-board visual observation. Odometry sequences consists of the speed s^t and steering angle d^t of the vehicle, that is, $o^t = (s^t, d^t)$ at time-step t.

$$\begin{gathered} p(\mathrm{B_f} = [b_i^{t+1}, ..., b_i^{t+n}] \mid \mathrm{B_p}, \mathrm{O_p}, \mathrm{O_f}) \\ \mathrm{B_p} = [b_i^{t-m}, ..., b_i^t], \\ \mathrm{O_p} = [o^{t-m}, ..., o^t], \\ \mathrm{O_f} = [o^{t+1}, ..., o^{t+n}] \end{gathered}$$

The variance of the predictive distribution $p(\mathrm{B_f}|\mathrm{B_{p'}})$ provides a measure of the associated uncertainty.

We will describe a basic sequence to sequence RNN first and then extend it to predict distributions and provide uncertainty estimates. Our sequence to sequence RNN (Fig. 4.1) consists of two embedding layers, an encoder RNN and a decoder RNN. The input sequence consists of the concatenated past bounding box and odometry sequences $\mathrm{B_p}, \mathrm{O_p}$. The input embedding layer embeds the inputs sequence x_t into the representation $\hat{x}_t$. This embedded sequence is read by the encoder RNN (RNN$_{\text{enc}}$) which produces a summary vector v_{bbox}. This summary vector is concatenated with predicted odometry $\mathrm{O_f}$ and this summary sequence is embedded using the second embedding layer. This embedded summary sequence $\hat{v}$ (containing information about past pedestrian motion and future vehicle odometry) is used by the decoder RNN (RNN$_{\text{dec}}$) for prediction.

In the following, we extend this model to predict distributions and estimate uncertainty.

4.3.2 Bayesian of Modelling of Uncertainty

We phrase our novel RNN encoder-decoder model in a Bayesian framework (Kendall and Gal, 2017). We capture epistemic (model) uncertainty by inferring posterior distribution of models (here models are RNN encoder-decoders with varying parameters) $p(f|X,Y)$ likely to have generated our data $\{X,Y\}$, given the prior belief of the distribution of RNN encoder-decoders $p(f)$. The predictive probability over the future sequence $\mathrm{B_f}$ given the past sequence $\mathrm{B_p}$ is obtained by marginalizing over the posterior distribution of RNN encoder-decoders,

$$p(\mathrm{B_f}|\mathrm{B_p}, X, Y) = \int p(\mathrm{B_f}|\mathrm{B_f}, f)p(f|X,Y)df \tag{4.1}$$

However, the integral in Eq. (4.1) is intractable. But, we can approximate it in two steps (Gal and Ghahramani, 2016a,c; Kendall and Gal, 2017). First, we assume that our RNN encoder-decoder models can be described by a finite set of variables ω. Thus, we constrain the set of possible RNN encoder-decoders to ones that can be described with ω. Now, Eq. (4.1) can be equivalently written as,

$$p(\mathrm{B_f}|\mathrm{B_p}, X, Y) = \int p(\mathrm{B_f}|\mathrm{B_p}, \omega)p(\omega|X,Y)d\omega \tag{4.2}$$

Second, we assume an approximating variational distribution $q(\omega)$ which allows efficient sampling,

$$q(\mathrm{B_f}|\mathrm{B_p}) = \int p(\mathrm{B_f}|\mathrm{B_p}, \omega) q(\omega) d\omega \tag{4.3}$$

We choose the set of weight matrices $\{W_1, .., W_L\} \in \mathcal{W}$ of our RNN enocder-decoder as the set of variables ω. Then we define an approximating Bernoulli variational distribution $q(\omega)$ over the columns w_k^c of the weight matrices $W_k \in \mathcal{W}$,

$$\begin{aligned} q(W_k) &= M_k \cdot \mathrm{diag}([z_{i,j}]_{j=1}^{C_k}) \\ z_{i,j} &= \mathrm{Bernoulli}(p_i), i = 1, ..., L, j = 1, ..., K_{i-1}. \end{aligned} \tag{4.4}$$

where, M_k are the variational parameters. This distribution allows for efficient sampling during training and testing which we discuss in the following subsection.

For an accurate approximation, we minimize the KL divergence between $q(\omega)$ and the true posterior $p(\omega|X, Y)$ as the training step. It can be shown that, (as in Gal and Ghahramani (2016b,a)),

$$\begin{aligned} \mathrm{KL}(q(\omega) \mid\mid p(\omega|X, Y)) &\propto \mathrm{KL}(q(\omega) \mid\mid p(\omega)) \\ &- \sum_t \int q(\omega) \log p(b_t^{t+n}|b_t^{t+n-1}, \mathrm{B_p}, \omega) d\omega. \end{aligned} \tag{4.5}$$

The first part corresponds to the distance to the prior model distribution and the second to the data fit. During training and prediction, we use Monte-Carlo integration to approximate the integrals Eq. (4.3) and Eq. (4.5) (see Section 4.3.5).

Aleatoric uncertainty can be captured along with epistemic uncertainty, by assuming a distribution of observation noise and estimating the sufficient statistics of the distribution. Here, we assume it to be a 4-d Gaussian at each time-step, $\mathcal{N}(b_i^{t+n}, \Sigma_i^t)$, where, $\Sigma = \mathrm{diag}\big((\sigma_x^{t+n})_i, (\sigma_y^{t+n})_i, (\sigma_x^{t+n})_i, (\sigma_y^{t+n})_i\big)$ in x and y directions in pixel space at time-step $t+n$. The predictive distribution of models parametrized by ω in our distribution $p(\omega|X, Y)$ is Gaussian at every time-step and complete the predictive distribution $p(\mathrm{B_f}|\mathrm{B_p}, X, Y)$ takes the form of a mixture of Gaussians at every time-step.

Uncertainty is the variance of our predictive distribution (Eq. (4.3)) and can be obtained through moment matching (Gal and Ghahramani, 2016b; Kendall and Gal, 2017). If we have T samples of future pedestrian bounding box sequences $\hat{\mathrm{B}}_\mathrm{f}$, with corresponding samples from the RNN encoder-decoder, the total uncertainty at time-step t is,

$$\begin{aligned} &\frac{1}{T}\Big(\sum_{i=1}^{T}(\hat{b}_i^t)^\intercal \hat{b}_i^t - \frac{1}{T}\big(\sum_{i=1}^{T}(\hat{b}_i^t)^\intercal\big)\big(\sum_{i=1}^{T}\hat{b}_i^t\big)\Big) \\ &\qquad + \frac{1}{T}\Big(\sum_{i=1}^{T}(\hat{\sigma}_i^t)_x + \sum_{i=1}^{T}(\hat{\sigma}_i^t)_y\Big). \end{aligned} \tag{4.6}$$

The first part of the sum corresponds to the epistemic uncertainty u_i^e and the second part corresponds to the aleatoric uncertainty u_i^a. We average the uncertainty across time-steps to arrive at the complete uncertainty estimate. Next, we describe how we sample from the Bernoulli distribution of RNN encoder-decoder weight matrices and the final sampling from the predictive distribution $p(\mathrm{B_f}|\mathrm{B_p}, \mathrm{O_p}, \mathrm{O_f})$.

4.3.3 Bayesian RNN Encoder-Decoder

The RNN encoder-decoder model of Section 4.3.1 contains four weight matrices. In detail, the two embedding layers contain two weight matrices W_{emi}, W_{ems}. The other two weight matrices belong to the encoder and decoder RNNs. We use an LSTM formulation as RNNs. Following Graves *et al.* (2013) the weight matrices of an LSTM can be concatenated into a matrix W and the LSTM can be formulated as in,

$$\begin{pmatrix} \underline{i} \\ \underline{f} \\ \underline{o} \\ \underline{\hat{c}} \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} \left(\begin{pmatrix} \hat{x}_t \\ h_{t-1} \end{pmatrix} \cdot W \right)$$
$$c_t = \underline{f} \odot c_{t-1} + \underline{i} \odot \underline{\hat{c}}\,, \quad h_t = \underline{o} \odot \tanh(c_t) \tag{4.7}$$

where $\underline{i}$ is the input gate, $\underline{f}$ is the forget gate, $\underline{o}$ is the output gate, c_t is the cell state, $\underline{\hat{c}}$ is the candidate cell state and h_t is the hidden state.

We define the Bernoulli variational distribution $q(\omega)$ (as in Eq. (4.4)) over the union of all the weight matrices of our model,

$$\omega = \{W_{emi}, W_{ems}, W_{enc}, W_{dec}\}\,. \tag{4.8}$$

where, W_{enc}, W_{dec} are the weight matrices of our RNN encoder and decoder.

Sampling from $q(W_{emi}), q(W_{ems})$ can be done efficiently by sampling random Bernoulli masks z_{emi}, z_{ems} and applying these masks after the linear transformations. In case of the input embedding,

$$\hat{x}_t = (x_t \cdot W_{emi}) \odot z_{emi} \tag{4.9}$$

Similarly, it was shown in Gal and Ghahramani (2016c) sampling weight matrices of a LSTM (here, $q(W_{enc}), q(W_{dec})$) can be efficiently performed by sampling random Bernoulli masks z_x, z_h and applying them at each time-step, while the LSTM encoder and decoder are unrolled,

$$\begin{pmatrix} \underline{i} \\ \underline{f} \\ \underline{o} \\ \underline{\hat{c}} \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} \left(\begin{pmatrix} x_t \odot z_x \\ h_{t-1} \odot z_h \end{pmatrix} \cdot W \right) \tag{4.10}$$

Sampling from our predictive distribution $p(\mathrm{B_{future}}|\mathrm{B_{past}}, \mathrm{O_{future}}, \mathrm{O_{past}})$ is done by first sampling weights matrices of our Bayesian RNN encoder-decoder. Then the parameters of the Gaussian observation noise distribution at each time-step is predicted. For this, we use the hidden state sequence h^t_{dec} of the RNN_{dec} and an additional linear transformation,

$$h^{t+n}_{\text{dec}} = \text{RNN}_{\text{dec}}(h^{t+n-1}_{\text{dec}}, v_{bbox}; z_x, z_h)$$
$$\hat{b}^{t+n}_i,\ (\hat{\sigma}_i{}^{t+n})_x,\ (\hat{\sigma}^{t+n}_i)_y = W_{bbox} * h^{t+n}_{\text{dec}} + bias_{bbox}.$$

We then sample from the predicted Gaussian distribution.

Next, we describe the second stream of our two-stream model – our model for long-term odometry prediction.

4.3.4 Prediction of Odometry

We use a similar RNN encoder-decoder architecture used for bounding box prediction, but without the embedding layers. We do not place a distribution over the weights but learn a single point estimate. We condition the predicted sequence O_{f} on the past odometry sequence O_{p} and last visual observation on-board the vehicle. The past odometry O_{past} is input to an encoder RNN which produces a summary vector v_{odo}. The past odometry of the vehicle O_{p} gives a strong cue about the future velocity especially in the short term (~100ms). We use the same LSTM formulation described previously as the RNN encoder; with the final hidden state h^t as the summary. The last visual observation can help in the longer term prediction of odometry, e.g. visual cues about bends in the road, obstacles etc. Similar to Xu *et al.* (2017); Bojarski *et al.* (2016), we employ a convolutional neural network (CNN-encoder) to embed the visual information provided by the currently observed frame; a visual summary vector v_{vis}. Next we describe our CNN-encoder architecture.

CNN-encoder. Our CNN-encoder should extract visual features to improve longer-term (multi-step versus single-step in Xu *et al.* (2017); Bojarski *et al.* (2016)) prediction. Therefore, we use a more complex CNN compared to Bojarski *et al.* (2016) and during training we learn the parameters from scratch, unlike Xu *et al.* (2017) which uses a pre-trained VGG network. Our CNN-encoder has 10 convolutional layers with *ReLU* non-linearities. We use a fixed, small filter size of 3x3 pixels. We use max-pooling after every two layers. After max-pooling we double the number of convolutional filters; we use {32,64,128,256,512} convolutional filters. The convolutional layers are followed by three fully connected layers with 1024, 256 and 128 neurons and *ReLU* non-linearities. The output of the last fully connected layer is the visual summary v_{vis}.

The odometry and visual summary vectors are concatenated $v = \{v_{odo}, v_{vis}\}$ and read by the RNN decoder (RNN_{dec}). We use the same LSTM formulation described previously as the RNN-decoder. As before, the hidden state of the LSTM decoder is used for predicting the future odometry sequence through a linear transformation.

$$h^{t+n}_{\text{dec}} = \text{RNN}_{\text{dec}}(h^{t+n-1}_{\text{dec}}, \{v_{odo}, v_{vis}\})$$
$$o^{t+n}_i = W_{odo} * h^{t+n}_{\text{dec}} + bias_{odo}.$$

We next describe training and inference in the complete two-stream model.

4.3.5 Training and Inference

Training. The two streams are trained separately. As the odometry prediction stream predicts point estimates, it is trained first by minimizing the MSE over the training set. The Bayesian bounding-box prediction stream is trained by estimating (Monte-Carlo) and minimizing the KL divergence of its approximate weight distribution $q(\omega)$ (Eq. (4.5)). More specifically, 1. We sample a mini-batch of size T of examples from the training set. 2. For each example, weights $\{W_{emi}, W_{ems}, W_{enc}, W_{dec}\}$ are sampled from $q(\omega)$ Eq. (4.8), by sampling Bernoulli masks as in Eq. (4.9) and Eq. (4.10). 3. For each example, the predicted means $\hat{\mathrm{B}}_{\mathrm{f}}$ and variances $\hat{\sigma}$ of the heteroscedastic models parameterized by ω are inferred. 4. The KL divergence **??** can be equivalently minimized by (similar to Gal and Ghahramani (2016b); Kendall and Gal (2017)) the following loss,

$$\Big(\frac{1}{4\,n\,N} \sum_{i=1}^{N} \sum_{j=1}^{n} \|\hat{b}_i^{t+j} - b_i^{t+j}\|_2^2 \, (\hat{\Sigma}_i^t)^{-2} \Big) + \lambda \|\mathcal{W}\|_2 + \log \hat{\sigma}_i^2$$

where, $\mid \mathrm{B}_{\mathrm{f}} \mid = n$ and N pedestrians. The left part is the equivalent of the negative log likelihood term in Eq. (4.5). The middle part is weight regularization parameterized by λ, equivalent to the KL term in Eq. (4.5). The right part is additional regularization as in Kendall and Gal (2017), to ensure finite predicted variance.

The ADAM optimizer (Kingma and Ba, 2015) is used during training. For training sequences longer than $|\mathrm{B}_{\mathrm{p}}| + |\mathrm{B}_{\mathrm{f}}|$ ($|\mathrm{O}_{\mathrm{p}} + \mathrm{O}_{\mathrm{f}}|$ respectively) we use a sliding window to convert to multiple sequences. Moreover, as the sequences in the training set are of varying lengths, we use a curriculum learning (CL) approach. We fix the length of the conditioning sequence $|\mathrm{B}_{\mathrm{p}}|, |\mathrm{O}_{\mathrm{p}}|$ and train for increasing longer time horizons $|\mathrm{B}_{\mathrm{f}}|, |\mathrm{O}_{\mathrm{f}}|$ (initializing the model parameters with those for shorter horizons). This allows us to train on a larger part of the Cityscapes training set (also on sequences shorter than $|\mathrm{B}_{\mathrm{p}}| + |\mathrm{B}_{\mathrm{f}}|$ of the final model) and leads to faster convergence.

Inference. Given a bounding box sequence B_{p} and corresponding odometry sequence (and visual observation), the odometry prediction stream is first used to predict O_{f}. We sample from the predictive distribution (Eq. (4.3)) by, 1. Sampling the weight matrices $\{W_{emi}, W_{ems}, W_{enc}, W_{dec}\}$ of the Bayesian bounding box prediction stream from the (learned) approximate distribution $q(\omega)$, by sampling Bernoulli masks as in Eq. (4.9) and Eq. (4.10), 2. The $\mathrm{RNN}_{\mathrm{dec}}$ is unrolled to obtain a sample $\{\mathrm{B}_{\mathrm{f}}, \hat{\sigma}_x, \hat{\sigma}_y\}$. The associated uncertainty is obtained using multiple samples as in Eq. (4.6).

4.4 EXPERIMENTS

We evaluate our model on real-world on-board street scene data and show predictions over a 1 second time horizon along with the associated uncertainty.

Dataset and evaluation metric. We evaluate on the Cityscapes dataset (Cordts *et al.*, 2016) which contains 2975 training, 500 validation and 1525 test video sequences

		MSE			$\mathcal{L}$		
			$\lvert B_p \rvert$			$\lvert B_p \rvert$	
Method	Odometry	4	6	8	4	6	8
Kalman Filter	None	1938	1289	1098	x	x	x
LSTM	None	692	663	650	8.11	7.99	7.77
LSTM-Aleatoric	None	772	758	750	5.92	5.81	5.54
LSTM-Bayesian	None	**647**	**624**	**618**	**4.31**	**4.26**	**4.13**
LSTM-Bayesian	Ground-truth	374	358	343	3.94	3.93	3.88

Table 4.2: Bounding box prediction error with varying $|B_p|$.

Method	MSE	$\mathcal{L}$
Social LSTM Alahi *et al.* (2016)	1514	5.63
LSTM-Bayesian	695	3.97
LSTM-Bayesian (centers)	648	x

Table 4.3: Bounding box center prediction error.

of length 1.8 seconds (30 frames). The video resolution is 2048×1024 pixels. The sequences were recorded on-board a vehicle in inner cities. Each sequence has associated odometry information. Pedestrian tracks were automatically extracted using the tracking by detection method of Tang *et al.* (2016). Detections were obtained using the Faster R-CNN based method of Zhang *et al.* (2017). This mimics real world autonomous/assisted driving systems where detections/tracks are obtained with a state-of-the-art detector/tracker and we have to deal with noise introduced by the detector and on rare occasions false positives from the pedestrian detector and tracker failures. We use as evaluation metric MSE in pixels (of the mean of the predictive distribution) and the negative log-likelihood $\mathcal{L}$. The $\mathcal{L}$ metric measures the probability assigned to the true sequence by our predictive distribution. We report these metrics averaged across all time-steps and plots per time-step. We use a dropout rate of 0.35, $\lambda = 10^{-4}$ (tuned on validation set) and use 50 Monte-Carlo samples across all Bayesian models.

Evaluation of bounding box prediction. We independently evaluate the first Bayesian LSTM stream of our two stream model, without conditioning it on predicted odometry. We predict 15 time-steps into the future and report the results in Table 4.2. We compare its performance with, 1. A linear Kalman filter baseline. 2. A homoscedastic LSTM encoder-decoder model (LSTM). 3. A heteroscedastic LSTM encoder-decoder (LSTM-Aleatoric). Finally, as an Oracle case, we compare against a Bayesian version in which the LSTM encoder can see the past odometry and the LSTM decoder can see the true future odometry at every time-step. We also vary the length of the conditioning sequence $|B_p|$ (training/test sets constant across varying $|B_p|$). In Table 4.2, we see that the homoscedastic LSTM model (2nd row)

Method	Visual	Speed (m/sec)	Angle (degrees)
Constant	None	1.62	26.85
Kalman Filter	None	0.053	2.44
LSTM	None	0.056	0.94
LSTM	RGB	0.048	0.88

Table 4.4: Odometry prediction error (MSE), $|O_p| = \{8\}$.

outperforms the linear Kalman filter (1st row). This shows that many bounding box sequences have a complex motion and therefore cannot be modelled by a Kalman filter. We see that the heteroscedastic LSTM (LSTM-Aleatoric, 3rd row) outperforms the homoscedastic LSTM (2nd row) with respect to the $\mathcal{L}$ metric. This means that the heteroscedastic LSTM learns to capture uncertainty and assigns higher probability to the true bounding box sequence. However, when epistemic uncertainty is not modelled, aleatoric uncertainty tried to compensate (as in Kendall and Gal (2017)) and this leads to poorer MSE. Finally, our Bayesian LSTM (4th row) outperforms all other methods. This can be attributed to two factors, 1. The richer Gaussian mixture model fitted by the Bayesian LSTM can capture aleatoric and epistemic uncertainty and fits the data distribution better (evidenced by $\mathcal{L}$ metric). 2. Additional introduced regularization (dropout and weight). Furthermore, we see that increasing the length of the conditioning sequence improves model performance. However, the performance gain saturates at $|B_{past}| = 8$. Henceforth, we will report results using $|B_p| = \{4, 8\}$ in the following. Finally, the odometry oracle case outperforms our Bayesian LSTM by a large margin. This shows that knowledge of vehicle odometry is crucial for good performance.

Comparison with Social LSTM (Alahi *et al.*, 2016). We compare our Bayesian LSTM model with the vanilla LSTM [†] model of Alahi *et al.* (2016) (with 128 neurons) that predicts trajectories independently in Table 4.3. Both models are trained to predict sequences of bounding box centers (length 15, given 8). Our Bayesian LSTM model performs better as it is more robust to mistakes during recursive prediction. The model of Alahi *et al.* (2016) observes true past pedestrian coordinates during training. However, during prediction it observes its own predictions causing errors to be propagated through multiple steps of prediction. Furthermore, we compare both methods to the centers obtained from the predictions of our Bayesian LSTM (second row of Table 4.2). The results show that we can improve upon bounding box center prediction by predicting bounding boxes.

Evaluation of odometry prediction. We train our odometry prediction LSTM encoder-decoder on the visual and odometry data of the Cityscapes training set. As many sequences have close to zero steering angle, we augment the training set to improve prediction performance. We reflect the steering angle and flip last observed image left to right of sequences with non-zero average steering angle.

[†] The version with social pooling did not converge on our dataset.

			MSE		$\mathcal{L}$	
			$\lvert B_p \rvert$		$\lvert B_p \rvert$	
Method	Streams	Visual	4	8	4	8
Kalman Filter	x	None	1938	1098	x	x
LSTM-Bayesian	One	None	572	546	4.03	3.97
LSTM-Bayesian	Two	RGB	**532**	**505**	**3.99**	**3.92**

Table 4.5: Evaluation of our Bayesian two stream model (Fig. 4.1).

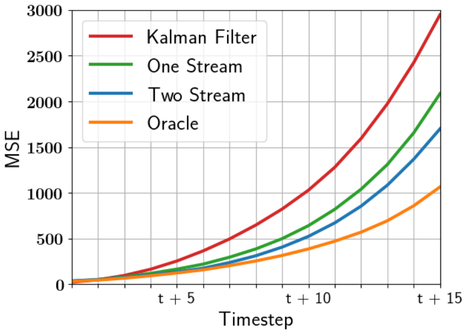


Table 4.6: MSE per time-step of models in Table 4.2 row 1, 4, 5 and Table 4.5 row 3.

This increases the training data with non-zero steering angles by a factor of two. We use MSE between the predicted future vehicle velocity and steering angles as evaluation metric. The velocity is in meters per second and angle in degrees. We include as baselines: 1. A constant steering predictor that predicts the last observed odometry. 2. A linear Kalman filter. 3. Our LSTM encoder-decoder without visual observation ($v = \{v_{odo}\}$). The third baseline is an ablation study. We observe no significant performance difference between $|O_p| = \{4\}$ and $|O_p| = \{8\}$. We evaluate 15 time-steps into the future and report the results in Table 4.4. We observe that the constant angle predictor performs significantly worse compared to the other baselines. This shows that the Cityscapes test set includes a significant number of non-trivial sequences with complex vehicle trajectories. We observe that the Kalman filter is able to quite accurately predict the vehicle speed. This is because most vehicles are travelling with constant speed or accelerating/decelerating smoothly. However, the performance of the linear Kalman filter is worse compared to the LSTM models with respect to steering angle. This means that many sequences have non-linear vehicle trajectories. The superior performance of our model compared to the RNN baseline without visual observations, especially in the long-term shows that our CNN encoder extracts information useful for long-term prediction.

Evaluation of our two-stream model. We perform an ablation study of our two-stream model (Fig. 4.1) and compare with a single-stream Bayesian LSTM encoder-decoder model where the encoder observes the concatenated past bounding box and velocity sequence $\{B_p, O_p\}$ and the decoder predicts the future bounding

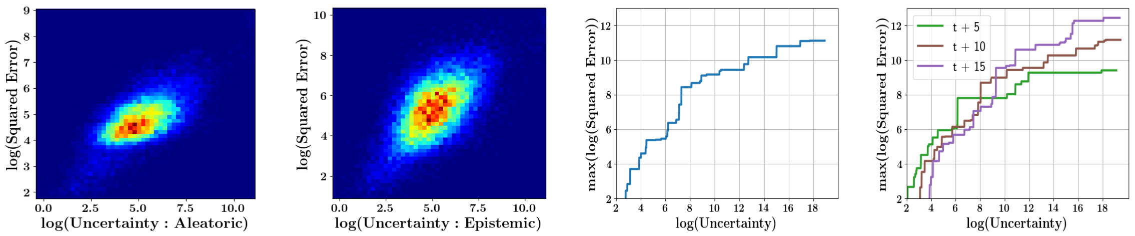


Figure 4.2: Quality of our uncertainty metric: plots 1 and 2 - uncertainty versus squared error, plots 3 and 4 - uncertainty versus *maximum* observed squared error.

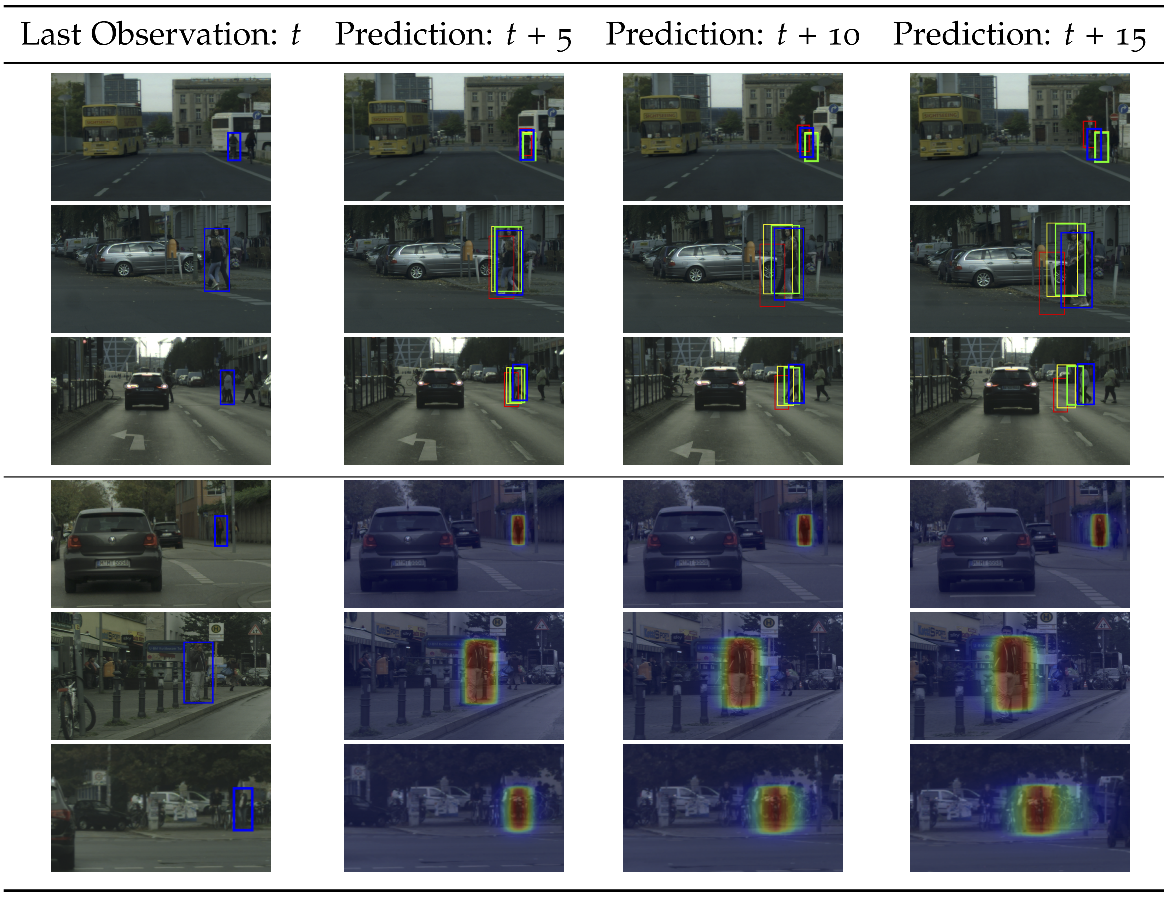


Figure 4.3: **Rows 1-3**: Point estimates. Blue: Ground-truth, Red: Kalman Filter (Table 4.2 row 1), Yellow: One-stream model (Table 4.2 row 4), Green: Two-stream model (mean of predictive distribution, Table 4.5 row 3). **Rows 4-6:** Predictive distributions of our two-stream model as heat maps.

box sequence B_f. This model does not see predicted future odometry. We evaluate the models and report the results in Table 4.5 and plot the MSE per time-step Table 4.6. The results show that jointly predicting odometry with pedestrian bounding boxes (3rd row) significantly improves performance (2nd row). The predicted odometry helps our two-stream model recover a significant fraction of the performance of the Oracle case in Table 4.2 row 5. The limiting factor here is that the odometry is difficult to predict in certain situations, e.g. at T-intersections. Apart from cases with inaccurate odometry prediction, the residual error of our two-stream (and the Oracle case) on a large part is due to the noise of the pedestrian detector and tracker failures. We show qualitative examples in Fig. 4.3. Row 1 shows point estimates under linear vehicle ego-motion and Rows 2, 3 non-linear vehicle ego-motion. Our two-stream model (mean of predictive distribution) outperforms other methods in the second case. Rows 4-5 shows the predictive distributions of the two-stream model under linear vehicle and pedestrian motion. The distribution is symmetric and has high aleatoric uncertainty which captures detection noise and possible pedestrian motion. Row 6 shows a case of a skewed distribution with high epistemic uncertainty which captures uncertainty in vehicle motion.

Quality of our uncertainty metric. We evaluate our uncertainty metric in Fig. 4.2. The first two plots show the aleatoric and epistemic uncertainty to the squared error of the mean of the predictive distribution of our two-stream model. We use log-log plots for better visualization as most sequences have low error (note, $\log(530) \approx 6.22$ the MSE of our two stream model, Table 4.5). We see that the epistemic and aleatoric uncertainties correlate well with the squared error. This means that for sequences where the mean of our predictive distribution is far from the true future sequence, our predictive distribution has a high variance (and vice versa). Therefore, for sequences with multiple likely futures, where the mean estimate would have high error, our model learns to predict diverse futures. In the third plot of Fig. 4.2, we plot the *maximum* log squared error (of the mean of the predictive distribution) observed at a certain predicted uncertainty level (sum of aleatoric and epistemic) in the test test. In the fourth plot, we plot the uncertainty with the maximum observed squared error at time-steps $t + \{5, 10, 15\}$. In both cases, uncertainty and observed maximum error is well correlated. This shows that *the predicted uncertainty upper bounds the error of the mean of the predictive distribution*. Therefore, the predicted uncertainty helps us express trust in predictions and has the potential to serve as a basis for better decision making.

4.5 CONCLUSION

We highlight the importance of anticipation for practical and safe driving in inner cities. We contribute to this important research direction the first model for long term prediction of pedestrians from on-board observations. We show predictions over a time horizon of 1 second. Predictions of our model are enriched by theoretically grounded uncertainty estimates. Key to our success is a Bayesian approach and long term prediction of odometry. We evaluate and compare several different architecture choices and arrive at a novel two-stream Bayesian LSTM encoder-decoder.

5 BAYESIAN PREDICTION OF FUTURE STREET SCENES USING SYNTHETIC LIKELIHOODS

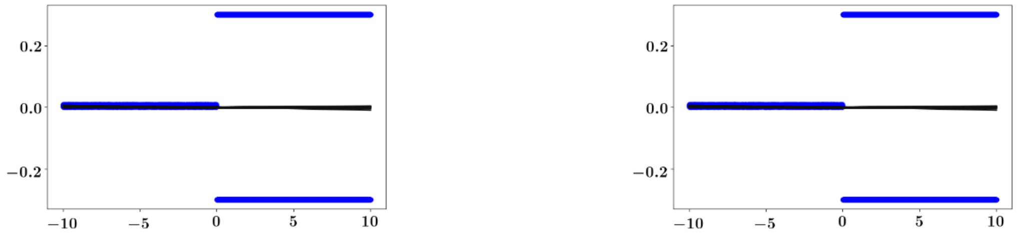
Contents

In real-world scenarios, future states become increasingly uncertain and multi-modal, particularly on long time horizons. As discussed in Chapter 4, dropout based Bayesian inference provides a computationally tractable, theoretically well grounded approach to learn likely hypotheses/models to deal with uncertain futures and make predictions that correspond well to observations – are well calibrated. However, it turns out that such approaches fall short to capture complex real-world scenes, even falling behind in accuracy when compared to the plain deterministic approaches. This is because the used log-likelihood estimate discourages diversity. In this work, we propose a novel Bayesian formulation for anticipating future scene states which leverages synthetic likelihoods. Unlike the formulation in Chapter 4, our novel Bayesian formulation encourages the learning of diverse models to accurately capture the multi-modal nature of future scene states. We show that our approach achieves accurate state-of-the-art predictions and calibrated probabilities through extensive experiments for scene anticipation on Cityscapes dataset.

5.1 INTRODUCTION

The future states of street scenes are inherently uncertain and the distribution of outcomes is often multi-modal. This is especially true for important classes like pedestrians. Recent works on anticipating street scenes (Luc *et al.*, 2017; Jin *et al.*, 2017; Nabavi *et al.*, 2018) do not systematically consider uncertainty.

Bayesian inference provides a theoretically well founded approach to capture both



Standard Dropout (Kendall & Gal (2017))

Standard Dropout (Kendall & Gal (2017))

Figure 5.1: Blue: Groundtruth distribution. Black: Models sampled at random from the model distribution.

model and observation uncertainty but with considerable computational overhead. A recently proposed approach (Gal and Ghahramani, 2016b; Kendall and Gal, 2017), as discussed in Chapter 4, uses dropout to represent the posterior distribution of models and capture model uncertainty. This approach has enabled Bayesian inference with deep neural networks without additional computational overhead. Moreover, it allows the use of any existing deep neural network architecture with minor changes.

However, when the underlying data distribution is multimodal and the model set under consideration do not have explicit latent state/variables (as most popular deep deep neural network architectures), the approach of Gal and Ghahramani (2016b); Kendall and Gal (2017) is unable to recover the true model uncertainty (see Fig. 5.1 and Osband (2016)). This is because this approach is known to conflate risk and uncertainty (Osband, 2016). This limits the accuracy of the models over a plain deterministic (non-Bayesian) approach. The main cause is the data log-likelihood maximization step during optimization – for every data point the average likelihood assigned by all models is maximized. This forces every model to explain every data point well, pushing every model in the distribution to the mean. We address this problem through an objective leveraging synthetic likelihoods (Wood, 2010; Rosca *et al.*, 2017) which relaxes the constraint on every model to explain every data point, thus encouraging diversity in the learned models to deal with multi-modality.

In this chapter, 1. We develop the first Bayesian approach to anticipate the multi-modal future of street scenes and demonstrate state-of-the-art accuracy on the diverse Cityscapes dataset without compromising on calibrated probabilities, 2. We propose a novel optimization scheme for dropout based Bayesian inference using synthetic likelihoods to encourage diversity and accurately capture model uncertainty, 3. Finally, we show that our approach is not limited to street scenes and generalizes across diverse tasks such as digit generation and precipitation forecasting.

Note that, as this chapter is based on the work Bhattacharyya *et al.* (2019a), we compare to prior work on street scene prediction; Luc *et al.* (2017); Nabavi *et al.* (2018). We provide an overview of more recent work in Chapter 2.

5.2 BAYESIAN MODELS FOR PREDICTION UNDER UNCERTAINTY

We phrase our models in a Bayesian framework, to jointly capture model (epistemic) and observation (aleatoric) uncertainty (Kendall and Gal, 2017). We begin with model uncertainty.

5.2.1 Model Uncertainty

Let $\mathrm{x} \in \mathrm{X}$ be the input (past) and $\mathrm{y} \in \mathrm{Y}$ be the corresponding outcomes. Consider $f : \mathrm{x} \mapsto \mathrm{y}$, we capture model uncertainty by learning the distribution $p(f|\mathrm{X}, \mathrm{Y})$ of generative models f, likely to have generated our data $\{\mathrm{X}, \mathrm{Y}\}$. The complete predictive distribution of outcomes y is obtained by marginalizing over the posterior distribution,

$$p(\mathrm{y}|\mathrm{x}, \mathrm{X}, \mathrm{Y}) = \int p(\mathrm{y}|\mathrm{x}, f) p(f|\mathrm{X}, \mathrm{Y}) df\,. \tag{5.1}$$

However, the integral in Eq. (5.1) is intractable. But, we can approximate it in two steps (Gal and Ghahramani, 2016b). First, we assume that our models can be described by a finite set of variables ω. Thus, we constrain the set of possible models to ones that can be described with ω. Now, Eq. (5.1) is equivalently,

$$p(\mathrm{y}|\mathrm{x}, \mathrm{X}, \mathrm{Y}) = \int p(\mathrm{y}|\mathrm{x}, \omega) p(\omega|\mathrm{X}, \mathrm{Y}) d\omega\,. \tag{5.2}$$

Second, we assume an approximating variational distribution $q(\omega)$ of models which allows for efficient sampling. This results in the approximate distribution,

$$p(\mathrm{y}|\mathrm{x}, \mathrm{X}, \mathrm{Y}) \approx p(\mathrm{y}|\mathrm{x}) = \int p(\mathrm{y}|\mathrm{x}, \omega) q(\omega) d\omega\,. \tag{5.3}$$

For convolutional models, Gal and Ghahramani (2016a) proposed a Bernoulli variational distribution defined over each convolutional patch. The number of possible models is exponential in the number of patches. This number could be very large, making it difficult to optimize over this very large set of models. In contrast, in our approach (Eq. (5.4)), the number of possible models is exponential in the number of weight parameters, a much smaller number. In detail, we choose the set of convolutional kernels and the biases $\{(W_1, b_1), \ldots, (W_L, b_L)\} \in \mathcal{W}$ of our model as the set of variables ω. Then, we define the following novel approximating Bernoulli variational distribution $q(\omega)$ independently over each element $w^{i,j}_{k',k}$ (correspondingly b_k) of the kernels and the biases at spatial locations $\{i, j\}$,

$$\begin{aligned} q(W_K) &= M_K \odot Z_K \\ z^{i,j}_{k',k} &= \text{Bernoulli}(p_K),\ \ k' = 1, \ldots, |K'|,\ \ k = 1, \ldots, |K|\,. \end{aligned} \tag{5.4}$$

Note, $\odot$ denotes the hadamard product, M_k are tuneable variational parameters, $z^{i,j}_{k',k} \in Z_K$ are the independent Bernoulli variables, p_K is a probability tensor equal to the size of the (bias) layer, $|K|$ ($|K'|$) is the number of kernels in the current (previous) layer. Here, p_K is chosen manually. Moreover, in contrast to Gal and Ghahramani (2016a), the same (sampled) kernel is applied at each spatial location leading to the detection of the same features at varying spatial locations. Next, we describe how we capture observation uncertainty.

5.2.2 Observation Uncertainty

Observation uncertainty can be captured by assuming an appropriate distribution of observation noise and predicting the sufficient statistics of the distribution (Kendall and Gal, 2017). Here, we assume a Gaussian distribution with diagonal covariance matrix at each pixel and predict the mean vector $\mu^{i,j}$ and co-variance matrix $\sigma^{i,j}$ of the distribution. In detail, the predictive distribution of a generative model draw from $\hat{\omega} \sim q(\omega)$ at a pixel position $\{i, j\}$ is,

$$p^{i,j}(\mathrm{y}|\mathrm{x}, \hat{\omega}) = \mathcal{N}\big((\mu^{i,j}|\mathrm{x}, \hat{\omega}), (\sigma^{i,j}|\mathrm{x}, \hat{\omega})\big)\,. \tag{5.5}$$

We can sample from the predictive distribution $p(\mathrm{y}|\mathrm{x})$ (Eq. (5.3)) by first sampling the weight matrices ω from Eq. (5.4) and then sampling from the Gaussian distribution in Eq. (5.5). We perform the last step by the linear transformation of a zero mean unit diagonal variance Gaussian, ensuring differentiability,

$$\hat{\mathrm{y}}^{i,j} \sim \mu^{i,j}(\mathrm{x}|\hat{\omega}) + z \times \sigma^{i,j}(\mathrm{x}|\hat{\omega}), \quad \text{where } p(z) \text{ is } \mathcal{N}(0, I) \text{ and } \hat{\omega} \sim q(\omega)\,. \tag{5.6}$$

where, $\hat{\mathrm{y}}^{i,j}$ is the sample drawn at a pixel position $\{i, j\}$ through the liner transformation of z (a vector) with the predicted mean $\mu^{i,j}$ and variance $\sigma^{i,j}$. In case of street scenes, $\mathrm{y}^{i,j}$ is a class-confidence vector and sample of final class probabilities is obtained by pushing $\hat{\mathrm{y}}^{i,j}$ through a softmax.

5.2.3 Training

For a good variational approximation (Eq. (5.3)), our approximating variational distribution of generative models $q(\omega)$ should be close to the true posterior $p(\omega|\mathrm{X}, \mathrm{Y})$. Therefore, we minimize the KL divergence between these two distributions. As shown in Gal and Ghahramani (2016b,a); Kendall and Gal (2017), the KL divergence is given by (over i.i.d data points),

$$\begin{aligned}
\mathrm{KL}(q(\omega) \,||\, p(\omega|\mathrm{X}, \mathrm{Y})) &\propto \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int q(\omega) \log p(\mathrm{Y}|\mathrm{X}, \omega) d\omega \qquad (5.7)\\
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int q(\omega) \Big(\int \log p(\mathrm{y}|\mathrm{x}, \omega) d(\mathrm{x}, \mathrm{y}) \Big) d\omega.\\
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \Big(\int q(\omega) \log p(\mathrm{y}|\mathrm{x}, \omega) d\omega \Big) d(\mathrm{x}, \mathrm{y}).
\end{aligned}$$

The log-likelihood term at the right of (Eq. (5.7)) considers every model for every data point. This imposes the constraint that every data point must be explained well by every model. However, if the data distribution (x, y) is multi-modal, this would push every model to the mean of the multi-modal distribution (as in Fig. 5.1 where only way for models to explain both modes is to converge to the mean). This discourages diversity in the learned modes. In case of multi-modal data, we would not be able to recover all likely models, thus hindering our ability to fully capture model uncertainty. The models would be forced to explain the data variation as observation noise (Osband, 2016), thus conflating model and observation uncertainty. We propose to mitigate this problem through the use of an approximate objective using synthetic likelihoods (Wood, 2010; Rosca *et al.*, 2017) – obtained from a classifier. The classifier estimates the likelihood based on whether the models $\hat{\omega} \sim q(\omega)$ explain (generate) data samples likely under the true data distribution $p(\mathrm{y}|\mathrm{x})$. This removes the constraint on models to explain every data point – it only requires the explained (generated) data points to be likely under the data distribution. Thus, this allows models $\hat{\omega} \sim q(\omega)$ to be diverse and deal with multi-modality. Next, we reformulate the KL divergence estimate of Eq. (5.7) to a likelihood ratio form which allows us to use a classifier to estimate (synthetic) likelihoods,

$$\begin{aligned}
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \Big(\int q(\omega) \log p(\mathrm{y}|\mathrm{x}, \omega) d\omega \Big) d(\mathrm{x}, \mathrm{y}). \\
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \Big(\int q(\omega) \big(\log \frac{p(\mathrm{y}|\mathrm{x}, \omega)}{p(\mathrm{y}|\mathrm{x})} + \log p(\mathrm{y}|\mathrm{x}) \big) d\omega \Big) d(\mathrm{x}, \mathrm{y}). \qquad (5.8) \\
&\propto \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \int q(\omega) \log \frac{p(\mathrm{y}|\mathrm{x}, \omega)}{p(\mathrm{y}|\mathrm{x})} d\omega \; d(\mathrm{x}, \mathrm{y}).
\end{aligned}$$

In the second step of Eq. (5.8), we divide and multiply the probability assigned to a data sample by a model $p(\mathrm{y}|\mathrm{x}, \omega)$ by the true conditional probability $p(\mathrm{y}|\mathrm{x})$ to obtain a likelihood ratio. We can estimate the KL divergence by equivalently estimating this ratio rather than the true likelihood. In order to (synthetically) estimate this likelihood ratio, let us introduce the variable θ to denote, $p(\mathrm{y}|\mathrm{x}, \theta = 1)$ the probability assigned by our model ω to a data sample (x, y) and $p(\mathrm{y}|\mathrm{x}, \theta = 0)$ the true probability of the sample. Therefore, the ratio in the last term of Eq. (5.8) is,

$$\begin{aligned}
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \int q(\omega) \log \frac{p(\mathrm{y}|\mathrm{x}, \theta = 1)}{p(\mathrm{y}|\mathrm{x}, \theta = 0)} d\omega \; d(\mathrm{x}, \mathrm{y}). \\
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \int q(\omega) \log \frac{p(\theta = 1|\mathrm{x}, \mathrm{y})}{p(\theta = 0|\mathrm{x}, \mathrm{y})} d\omega \; d(\mathrm{x}, \mathrm{y}). \quad \text{(Using Bayes theorem)} \qquad (5.9) \\
&= \mathrm{KL}(q(\omega) \,||\, p(\omega)) - \int \int q(\omega) \log \frac{p(\theta = 1|\mathrm{x}, \mathrm{y})}{1 - p(\theta = 1|\mathrm{x}, \mathrm{y})} d\omega \; d(\mathrm{x}, \mathrm{y}).
\end{aligned}$$

In the last step of Eq. (5.9) we use the fact that the events $\theta = 1$ and $\theta = 0$ are mutually exclusive. We can approximate the ratio $\frac{p(\theta=1|\mathrm{x},\mathrm{y})}{1-p(\theta=1|\mathrm{x},\mathrm{y})}$ by jointly learning a discriminator $D(\mathrm{x}, \hat{\mathrm{y}})$ that can distinguish between samples of the true data distribution and samples $(\mathrm{x}, \hat{\mathrm{y}})$ generated by the model ω, which provides a synthetic estimate of the likelihood, and equivalently integrating directly over $(\mathrm{x}, \hat{\mathrm{y}})$,

$$\approx \mathrm{KL}(q(\omega) \mid\mid p(\omega)) - \int\int q(\omega) \log\Big(\frac{D(\mathrm{x}, \hat{\mathrm{y}})}{1 - D(\mathrm{x}, \hat{\mathrm{y}})}\Big) d\omega \; d(\mathrm{x}, \hat{\mathrm{y}}). \tag{5.10}$$

Note that the synthetic likelihood $\big(\frac{D(\mathrm{x}, \hat{\mathrm{y}})}{1 - D(\mathrm{x}, \hat{\mathrm{y}})}\big)$ is independent of any specific pair (x, y) of the true data distribution (unlike the log-likelihood term in Eq. (5.7)), its value depends only upon whether the generated data point $(\mathrm{x}, \hat{\mathrm{y}})$ by the model ω is likely under the true data distribution $p(\mathrm{y}|\mathrm{x})$. Therefore, the models ω have to only generate samples $(\mathrm{x}, \hat{\mathrm{y}})$ likely under the true data distribution. The models need not explain every data point equally well. Therefore, we do not push the models ω to the mean, thus allowing them to be diverse and allowing us to better capture uncertainty.

Empirically, we observe that a hybrid log-likelihood term using both the log-likelihood terms of Eq. (5.10) and Eq. (5.7) with regularization parameters α and β (with $\alpha \geq \beta$) stabilizes the training process,

$$\alpha \int\int q(\omega) \log\Big(\frac{D(\mathrm{x}, \hat{\mathrm{y}})}{1 - D(\mathrm{x}, \hat{\mathrm{y}})}\Big) d\omega \; d(\mathrm{x}, \hat{\mathrm{y}}) + \beta \int\int q(\omega) \log p(\mathrm{y}|\mathrm{x}, \omega) d\omega \; d(\mathrm{x}, \mathrm{y}). \tag{5.11}$$

Note that, although we do not explicitly require the posterior model distribution to explain all data points, due to the exponential number of models afforded by dropout and the joint optimization (min-max game) of the discriminator, empirically we see very diverse models explaining most data points. Moreover, empirically we also see that predicted probabilities remain calibrated. Next, we describe the architecture details of our generative models ω and the discriminator $D(\mathrm{x}, \hat{\mathrm{y}})$.

5.2.4 Model Architecture for Street Scene Prediction

The architecture of our ResNet based generative models in our model distribution $q(\omega)$ is shown in Fig. 5.2. The generative model takes as input a sequence of past segmentation class-confidences $\mathrm{s_p}$, the past and future vehicle odometry $\mathrm{o_p}, \mathrm{o_f}$ ($\mathrm{x} = \{\mathrm{s_p}, \mathrm{o_p}, \mathrm{o_f}\}$) and produces the class-confidences at the next time-step as output. The additional conditioning on vehicle odometry is because the sequences are recorded in frame of reference of a moving vehicle and therefore the future observed sequence is dependent upon the vehicle trajectory. We use recursion to efficiently predict a sequence of future scene segmentations $\mathrm{y} = \{\mathrm{s_f}\}$. The discriminator takes as input $\mathrm{s_f}$ and classifies whether it was produced by our model or is from the true data distribution.

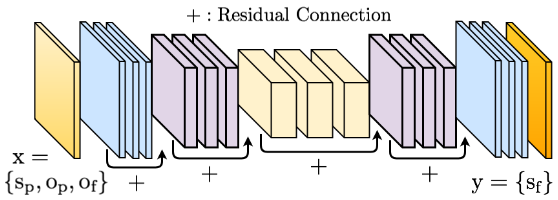


Figure 5.2: The architecture of our ResNet based generative models for street scene prediction in our model distribution $q(\omega)$.

In detail, generative model architecture consists of a fully convolutional encoder-decoder pair. This architecture builds upon prior work of Luc *et al.* (2017); Jin *et al.* (2017), however with key differences. In Luc *et al.* (2017), each of the two levels of the model architecture consists of only five convolutional layers. In contrast, our model consists of one level with five convolutional blocks. The encoder contains three residual blocks with max-pooling in between and the decoder consists of a residual and a convolutional block with up-sampling in between. We double the size of the blocks following max-pooling in order to preserve resolution. This leads to a much deeper model with fifteen convolutional layers, with constant spatial convolutional kernel sizes. This deep model with pooling creates a wide receptive field and helps better capture spatio-temporal dependencies. The residual connections help in the optimization of such a deep model. Computational resources allowing, it is possible to add more levels to our model. In Jin *et al.* (2017), a model is considered which uses a Res101-FCN as an encoder. Although this model has significantly more layers, it also introduces a large amount of pooling. This leads to loss of resolution and spatial information, hence degrading performance. Our discriminator model consists of six convolutional layers with max-pooling layers in-between, followed by two fully connected layers.

5.3 EXPERIMENTS

Next, we evaluate our approach on synthetic 2D data, MNIST digit generation and street scene anticipation on Cityscapes.

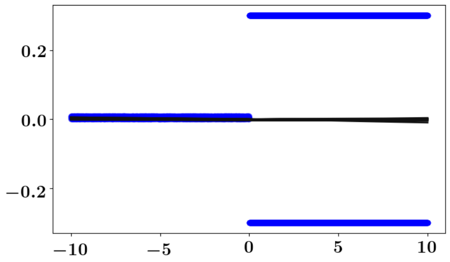


Figure 5.3: Blue: Data points. Black: Sampled models $\hat{\omega} \in q(\omega)$ learned by the Bayes-S approach. All models fit to the mean.

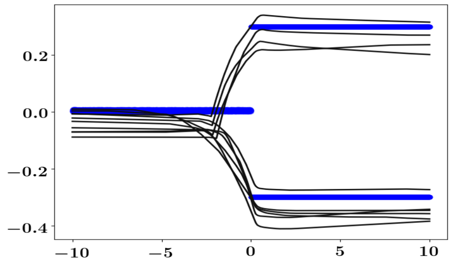
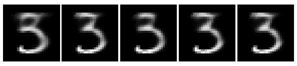
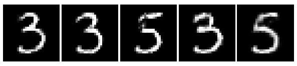
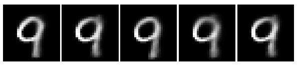
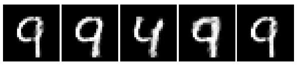


Figure 5.4: Blue: Data points. Black: Sampled models $\hat{\omega} \in q(\omega)$ learned by the Bayes-SL approach. We recover models covering both modes.

5.3.1 Multi-modal 2D Data.

We show results on simple multi-modal 2d data as in the motivating example in the introduction. The data consists of two parts: $x \in [-10, 0]$ we have $y = 0$ and $x \in [0, 10]$ we have $y = (-0.3, 0.3)$. The set of models under consideration is a two

Groundtruth	Bayes-S Samples	Bayes-SL Samples
		
		

Method	Top-10%
Mean	80.1±0.4
Bayes-S	79.3±0.4
CVAE	82.5±0.5
Bayes-SL	**83.1±0.6**

Table 5.1: **Left**: MNIST generations: The models see the non grayed-out region of the digit. The samples are generated from models drawn at random from $\hat{\omega} \sim q(\omega)$. **Right**: Top-10% accuracy on MNIST generation.

	Timestep		
Method	+0.06sec	+0.18sec	+0.54sec
Last Input (Luc *et al.* (2017))	x	49.4	36.9
Luc *et al.* (2017) (ft)	x	59.4	47.8
Last Input (Nabavi *et al.* (2018))	62.6	51.0	x
Nabavi *et al.* (2018)	71.3	60.0	x
Last Input (Ours)	67.1	52.1	38.3
Bayes-S (mean)	71.2	64.8	45.7
Bayes-WD (mean)	73.7	63.5	44.0
Bayes-WD-SL (mean)	**74.1**	64.8	45.9
Bayes-WD-SL (ft, mean)	x	**65.1**	**51.2**
Bayes-WD-SL (top 5%)	**75.3**	65.2	49.5
Bayes-WD-SL (ft, top 5%)	x	**66.7**	**52.5**

Table 5.2: Comparing mean predictions to the state-of-the-art.

Method	mIoU
Dilation10 (Luc *et al.*, 2017)	68.8
PSPNet (Nabavi *et al.*, 2018)	75.7
PSPNet (Ours)	76.9

Table 5.3: Comparison of segmentation estimation methods on Cityscapes validation set.

hidden layer neural network with 256-128 neurons with 50% dropout. We show 10 randomly sampled models from $\hat{\omega} \sim q(\omega)$ learned by the Bayes-S approach in Figure 5.3 and our Bayes-SL approach in Figure 5.4 (with $\alpha = 1, \beta = 0$). We assume constant observation uncertainty (=1). We clearly see that our Bayes-SL learns models which cover both modes, while all the models learned by Bayes-S fit to the mean. Clearly showing that our approach can better capture model uncertainty.

5.3.2 MNIST Digit Generation

Here, we aim to generate the full MNIST digit given only the lower left quarter of the digit. This task serves as an ideal starting point as in many cases there are multiple likely completions given the lower left quarter digit, e.g. 5 and 3. Therefore, the learned model distribution $q(\omega)$ should contain likely models corresponding to these completions. We use a fully connected generator with 6000-4000-2000 hidden units with 50% dropout probability. The discriminator has 1000-1000 hidden units with leaky ReLU non-linearities. We set $\beta = 10^{-4}$ for the first 4 epochs and then reduce it to 0, to provide stability during the initial epochs. We compare our synthetic likelihood based approach (Bayes-SL) with, 1. A non-Bayesian mean model, 2. A standard Bayesian approach (Bayes-S), 3. A Conditional Variational Autoencoder (CVAE) (architecture as in Sohn *et al.* (2015)). As evaluation metric we consider (oracle) Top-k% accuracy (Lee *et al.*, 2017b). We use a standard Alex-Net based classifier to measure if the best prediction corresponds to the ground-truth class – identifies the correct mode – in Table 5.1 (right) over 10 splits of the MNIST test-set. We sample 10 models from our learned distribution and consider the best model. We see that our Bayes-SL performs best, even outperforming the CVAE model. In the qualitative examples in Table 5.1 (left), we see that generations from models $\hat{\omega} \sim q(\omega)$ sampled from our learned model distribution corresponds to clearly defined digits (also in comparison to Figure 3 in Sohn *et al.* (2015)). In contrast, we see that the Bayes-S model produces blurry digits. All sampled models have been pushed to the mean and show little advantage over a mean model.

5.3.3 Cityscapes Street Scene Anticipation

Next, we evaluate our apporach on the Cityscapes dataset – anticipating scenes more than 0.5 seconds into the future. The street scenes already display considerable multi-modality at this time-horizon.

Evaluation metrics and baselines. We use PSPNet (Zhao *et al.*, 2017a) to segment the full training sequences as only the 20$^{\text{th}}$ frame has groundtruth annotations. We always use the annotated 20$^{\text{th}}$ frame of the validation sequences for evaluation using the standard mean Intersection-over-Union (mIoU) and the per-pixel (negative) conditional log-likelihood (CLL) metrics. We consider the following baselines for comparison to our Resnet based (architecture in Fig. 5.2) Bayesian (Bayes-WD-SL) model with weight dropout and trained using synthetic likelihoods: 1. Copying the last seen input; 2. A non-Bayesian (ResG-Mean) version; 3. A Bayesian version with standard patch dropout (Bayes-S); 4. A Bayesian version with our weight dropout (Bayes-WD). Note that, combination of ResG-Mean with an adversarial loss did not lead to improved results (similar observations made in Luc *et al.* (2017)). We use grid search to set the dropout rate (in Eq. (5.4)) to 0.15 for the Bayes-S and 0.20 for Bayes-WD(-SL) models. We set $\alpha, \beta = 1$ for our Bayes-WD-SL model. We train all models using Adam (Kingma and Ba, 2015) for 50 epochs with batch size 8. We use one sample to train the Bayesian methods as in Gal and Ghahramani (2016a) and use 100 samples during evaluation.

Comparison to state of the art. We begin by comparing our Bayesian models to state-of-the-art methods (Luc *et al.*, 2017; Nabavi *et al.*, 2018) in Table 5.2. We use the mIoU metric and for a fair comparison consider the mean (of all samples) prediction of our Bayesian models. We always compare to the groundtruth segmentations of the validation set. However, as all three methods use a slightly different semantic segmentation algorithm (Table 5.3) to generate training and input test data, we include the mIoU achieved by the Last Input of all three methods. Similar to Luc *et al.* (2017), we fine-tune (ft) to predict at 3 frame intervals for better performance at +0.54sec. Our Bayes-WD-SL model outperforms baselines and improves on prior work by 2.8

	Timestep			
	$t + 5$		$t + 10$	
Method	mIoU	CLL	mIoU	CLL
Last Input	45.7	0.86	37.1	1.35
ResG-Mean	59.1	0.49	46.6	0.89
Bayes-S	58.8	0.48	46.1	0.80
Bayes-WD	59.2	0.48	46.6	**0.79**
Bayes-WD-SL	**60.2**	**0.47**	**47.1**	**0.79**

Table 5.4: Evaluation on capturing uncertainty (using mIoU top 5%).

	Timestep	
	$t + 5$	$t + 10$
Method	mIoU	mIoU
CVAE (First)	58.7	45.5
CVAE (Mid)	58.9	46.6
CVAE (Last)	59.2	46.8
Bayes-WD-SL	**60.2**	**47.1**

Table 5.5: Ablation study and comparison to a CVAE baseline.

mIoU at +0.06sec and 4.8 mIoU/3.4 mIoU at +0.18sec/+0.54sec respectively. Our Bayes-WD-SL model also obtains higher relative gains in comparison to Luc *et al.* (2017) with respect to the Last Input Baseline. These results validate our choice of model architecture and show that our novel approach clearly outperforms the state-of-the-art. The performance advantage of Bayes-WD-SL over Bayes-S shows that the ability to better model uncertainty does not come at the cost of lower mean performance. However, at larger time-steps as the future becomes increasingly uncertain, mean predictions (mean of all likely futures) drift further from the ground-truth. Therefore, next we evaluate the models on their (more important) ability to capture the uncertainty of the future.

Evaluation of predicted uncertainty. Next, we evaluate whether our Bayesian models are able to accurately capture uncertainty and deal with multi-modal futures, upto $t + 10$ frames (0.6 seconds) in Table 5.4. We consider the mean of (oracle) best 5% of predictions (Lee *et al.*, 2017b) of our Bayesian models to evaluate whether the learned model distribution $q(\omega)$ contains likely models corresponding to the groundtruth. We see that the best predictions considerably improve over the mean predictions – showing that our Bayesian models learns to capture uncertainty and deal with multi-modal futures. Quantitatively, we see that the Bayes-S model performs worst, demonstrating again that standard dropout (Kendall and Gal, 2017) struggles to recover the true model uncertainty. The use of weight dropout improves the performance to the level of the ResG-Mean model. Finally, we see that our Bayes-WD-SL model performs best. In fact, it is the only Bayesian model whose (best) performance exceeds that of the ResG-Mean model (also outperforming state-of-the-art), demonstrating the effectiveness of synthetic likelihoods during training. In Fig. 5.5 we show examples comparing the best prediction of our Bayes-WD-SL model and ResG-Mean at $t + 9$. The last row highlights the differences between the predictions – cyan shows areas where our Bayes-WD-SL is correct and ResG-Mean is wrong, red shows the opposite. We see that our Bayes-WD-SL performs better at classes like cars and pedestrians which are harder to predict (also in comparison to Table 5 in Luc *et al.* (2017)). In Fig. 5.6, we show samples from randomly sampled models $\hat{\omega} \sim q(\omega)$, which shows correspondence to the range of possible movements of bicyclists/pedestrians. Next, we further evaluate the models with the CLL metric

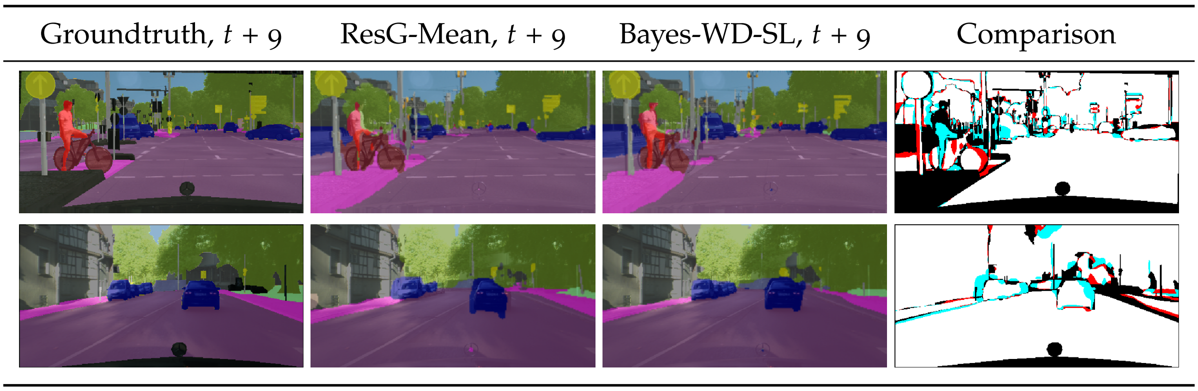


Figure 5.5: Bayes-WD-SL (top 1) vs ResG-Mean. Cyan: Bayes-WD-SL is correct and ResG-Mean is wrong. Red: Bayes-WD-SL is wrong and ResG-Mean is correct, white: both right, black: both wrong/unlabeled.

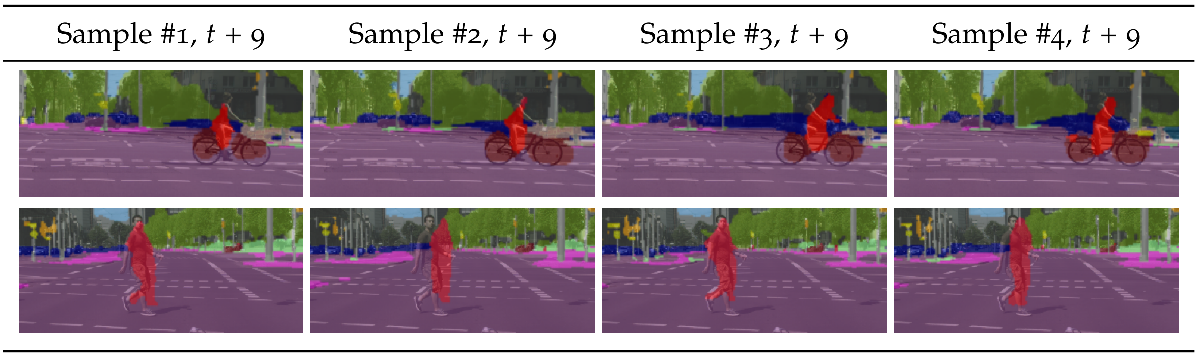


Figure 5.6: Random samples from our Bayes-WD-SL model corresponds to the range of likely movements of bicyclists/pedestrians.

in Table 5.4. We consider the mean predictive distributions (Eq. (5.3)) up to $t + 10$ frames. We see that the Bayesian models outperform the ResG-Mean model significantly. In particular, we see that our Bayes-WD-SL model performs the best, demonstrating that the learned model and observation uncertainty corresponds to the variation in the data.

Comparison to a CVAE baseline. As there exists no CVAE (Sohn *et al.*, 2015) based model for future segmentation prediction, we construct a baseline as close as possible to our Bayesian models based on existing CVAE based models for related tasks (Babaeizadeh *et al.*, 2018; Xue *et al.*, 2016). Existing CVAE based models (?Xue *et al.*, 2016) contain a few layers with Gaussian input noise. Therefore, for a fair comparison we first conduct a study in Table 5.5 to find the layers which are most effective at capturing data variation. We consider Gaussian input noise applied in the first, middle or last convolutional blocks. The noise is input dependent during training, sampled from a recognition network. We observe that noise in the last layers can better capture data variation. This is because the last layers capture semantically higher level scene features. Overall, our Bayesian approach (Bayes-WD-SL) performs the best. This shows that the CVAE model is not able to effectively leverage Gaussian noise to match the data variation.

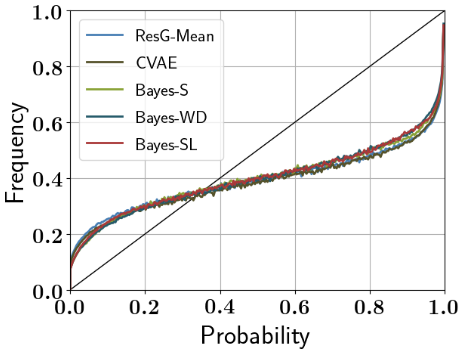


Figure 5.7: Uncertainty calibration at $t + 10$.

Uncertainty calibration. We further evaluate predicted uncertainties by measuring their calibration – the correspondence between the predicted probability of a class and the frequency of its occurrence in the data. As in Kendall and Gal (2017), we discretize the output probabilities of the mean predicted distribution into bins and measure the frequency of correct predictions for each bin. We report the results at $t + 10$ frames in Fig. 5.7. We observe that all Bayesian approaches outperform the ResG-Mean and CVAE versions. This again demonstrates the effectiveness of the Bayesian approaches in capturing uncertainty.

5.4 CONCLUSION

We propose a novel approach for predicting real-world semantic segmentations into the future that casts a convolutional deep learning approach into a Bayesian formulation. One of the key contributions is a novel optimization scheme that uses synthetic likelihoods to encourage diversity and deal with multi-modal futures. Our proposed method shows state of the art performance in challenging street scenes. More importantly, we show that the probabilistic output of our deep learning architecture captures uncertainty and multi-modality inherent to this task. Furthermore, we show that the developed methodology goes beyond just street scene anticipation and creates new opportunities to enhance high performance deep learning architectures with principled formulations of Bayesian inference.

6 ACCURATE AND DIVERSE SAMPLING OF SEQUENCES BASED ON A "BEST OF MANY" SAMPLE OBJECTIVE

Contents

As discussed in Chapters 1 and 5, real-world scenarios demand a model of uncertainty for future prediction problems. This is because predictions become increasingly uncertain – in particular on long time horizons. While impressive results have been shown on point estimates, scenarios that induce multi-modal distributions over future sequences remain challenging. Our work addresses these challenges in a Gaussian latent variable model for sequence prediction. Our core contribution in this chapter is a "Best of Many" sample objective that leads to more accurate and more diverse predictions that better capture the true variations in real-world sequence data. Beyond our analysis of improved model fit, our models also empirically outperform prior work on three diverse tasks ranging from traffic scenes to weather data.

Moreover, following the contributions in this chapter, in Chapter 7 we introduce a novel conditional normalizing flow based prior to further improve performance of Gaussian latent variable models for multi-modal distributions. Further, in Chapter 8, we build a hybrid VAE-GAN model using the insights gained in this chapter for multi-modal distributions. Finally, in Chapter 9, we propose a joint inference framework for Gaussian latent variable models to capture the effect of interactions, among agents such as pedestrians or vehicles in traffic scenes, on the multi-modal distribution of future trajectories.

6.1 INTRODUCTION

Many future prediction tasks ranging from autonomous driving to precipitation forecasting can be formulated as sequence prediction problems. Given a past sequence of events, probable future outcomes are to be predicted.

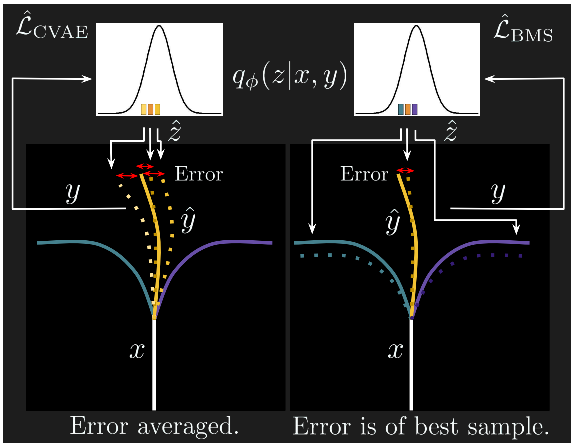


Figure 6.1: Comparison between our "Best of Many" sample objective and the standard CVAE objective.

Recurrent Neural Networks (RNN) especially LSTM formulations are state-of-the-art models for sequence prediction tasks (Alahi *et al.*, 2016; Xu *et al.*, 2017; Finn *et al.*, 2016; Shi *et al.*, 2015). These approaches predict only point estimates. However, many sequence prediction problems are only partially observed or stochastic in nature and hence the distribution of future sequences can be highly multi-modal. Consider the task of predicting future pedestrian trajectories. In many cases, we do not have any information about the intentions of the pedestrians in the scene. A pedestrian after walking over a zerba crossing might decide to turn either left or right. A point estimate in such a situation would be highly unrealistic. Therefore, in order to incorporate uncertainty of future outcomes, we are interested in *structured predictions*. Structured prediction implies learning a one to many mapping of a given fixed sequence to plausible future sequences (Sohn *et al.*, 2015). This leads to more realistic predictions and enables probabilistic inference.

Recent work Lee *et al.* (2017b) has proposed deep conditional generative models with Gaussian latent variables for structured sequence prediction. The Conditional Variational Auto-Encoder (CVAE) framework (Sohn *et al.*, 2015) is used in Lee *et al.* (2017b) for learning of the Gaussian Latent Variables. We identify two key limitations of this CVAE framework. First, the currently used objectives hinder learning of diverse samples due to a marginalization over multi-modal futures. Second, a mismatch in latent variable distribution between training and testing leads to errors in model fitting. We overcome both challenges which results in more accurate and diverse samples – better capturing the true variations in data. Our main contributions in this chapter are, 1. We propose a novel "Best of many" sample objective for which we provide an analytic derivation. 2. We analyze the benefits of our model analytically as well as show an improved fit for the latent variables compared to prior approaches. 3. We also show for the first time that this modeling paradigm extends to full-frame images sequences with diverse multi-modal futures. 4. We demonstrate improved accuracy as well as diversity of the generated samples on three diverse tasks: stroke completion, Stanford Drone Dataset and HKO weather data. On all three datasets we consistently outperform the state of the art and baselines.

Note that, as this chapter is based on the work Bhattacharyya *et al.* (2018c), we compare to prior work with Gaussian latent variables on the Stanford Drone dataset; Lee *et al.* (2017b). We provide an overview of more recent work, e.g. Pajouheshgar and Lampert (2018); Gupta *et al.* (2018); Zhao *et al.* (2019); Sadeghian *et al.* (2019); Deo and Trivedi (2019); Mangalam *et al.* (2020) in Chapter 2.

6.2 RELATED WORK

While we provide a broader discussion on related work in Chapter 2, here we discuss related work relevant to this chapter.

Structured output prediction. Stochastic feed-forward neural networks (SFNN) (Tang and Salakhutdinov, 2013) model multi-modal conditional distributions through binary stochastic hidden variables. During training multiple samples are drawn and weighted according to importance-weights. However, due to the latent variables being binary SFNNs are hard to train on large datasets. There have been several efforts to make training more efficient for binary latent variables (Raiko *et al.*, 2015; Gu *et al.*, 2016; Mnih and Rezende, 2016; Lee *et al.*, 2017a). However, not all tasks can be efficiently modelled with binary hidden variables. In Sohn *et al.* (2015), Gaussian hidden variables are considered where the re-parameterization trick can be used for learning on large datasets using stochastic optimization. Inspired by this technique we model Gaussian hidden variables for structured sequence prediction tasks.

Recurrent neural networks. Recurrent Neural Networks (RNNs) are state of the art methods for a variety of sequence learning tasks (Graves, 2013; Sutskever *et al.*, 2014). In this work, we focus on sequence to sequence regression tasks, in particular, trajectory prediction and image sequence prediction. RNNs have been used for pedestrian trajectory prediction. In Alahi *et al.* (2016), trajectories of multiple people in a scene are jointly modelled in a social context. However, even though the distribution of pedestrian trajectories are highly multimodal (with diverse futures), only one mean estimate is modelled. Lee *et al.* (2017b) jointly models multiple future pedestrian trajectories using a recurrent CVAE sampling module. Samples generated are refined and ranked using image and social context features. While our trajectory prediction model is similar to the sampling module of Lee *et al.* (2017b), we focus on improving the sampling module by our novel multi-sample objective function. Convolutional RNNs (Shi *et al.*, 2015) have been used for image sequence prediction. Examples include, robotic arm movement prediction (Finn *et al.*,

2016) and precipitation now-casting (Shi *et al.*, 2015, 2017). In this work, we extend the model of Shi *et al.* (2015) for structured sequence prediction by conditioning predictions on Gaussian latent variables. Furthermore, we show that optimization using our novel multi-sample objective leads to improved results over the standard CVAE objective.

6.3 structured sequence prediction with gaussian latent variables

We begin with an overview of deep conditional generative models with gaussian latent variables and the CVAE framework with the corresponding objective (Sohn *et al.*, 2015) used for training. Then, we introduce our novel "Best of many" samples objective function. Thereafter, we introduce the conditional generative models which serve as the test bed for our novel objective. We first describe our model for structured trajectory prediction which is similar to the sampling module of Lee *et al.* (2017b) and consider extensions which additionally conditions on visual input and generates full image sequences.

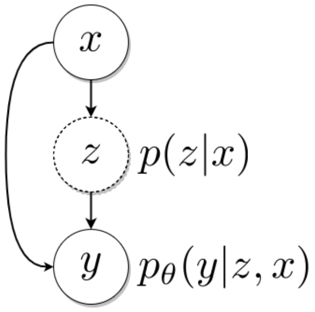


Figure 6.2: Conditional generative models.

We consider deep conditional generative models of the form shown in Fig. 6.2. Given an input sequence x, a latent variable $\hat{z}$ is drawn from the conditional distribution $p(z|x)$ (assumed Gaussian). The output sequence $\hat{y}$ is then sampled from the distribution $p_\theta(y|x,z)$ of our conditional generative model with parameterized by θ. The latent variables z enables one-to-many mapping and the learning of multiple modes of the true posterior distribution $p(y|x)$. In practice, the simplifying assumption is made that z is independent of x and $p(z|x)$ is $\mathcal{N}(0,I)$. Next, we discuss the training of such models.

6.3.1 Conditional Variational Auto-encoder Training Objective

We would like to maximize the data log-likelihood $p_\theta(y \mid x)$. To estimate the data log-likelihood of our model p_θ, one possibility is to perform Monte-Carlo sampling of the latent variable z. For T samples, this leads to the following estimate,

$$\hat{\mathcal{L}}_{\text{MC}} = \log\Big(\frac{1}{T}\sum_{i=1}^{T} p_\theta(y|\hat{z}_i, x)\Big),\ \ \hat{z}_i \sim \mathcal{N}(0, I). \tag{6.1}$$

This estimate is unbiased but has high variance (Mnih and Rezende, 2016). We would underestimate the log-likelihood for some samples and overestimate for others, especially if T is small. This would in turn lead to high variance weight updates.

We can reduce the variance of updates by estimating the log-likelihood through importance sampling during training. As described in Sohn *et al.* (2015), we can sample the latent variables z from a recognition network q_ϕ using the re-parameterization trick (Kingma and Welling, 2014). The data log-likelihood is,

$$\begin{aligned} &\log(p(y \mid x)) = \\ &\log\Big(\int p_\theta(y|z, x)\,\frac{p(z|x)}{q_\phi(z|x, y)}\,q_\phi(z|x, y)\,dz\Big). \end{aligned} \tag{6.2}$$

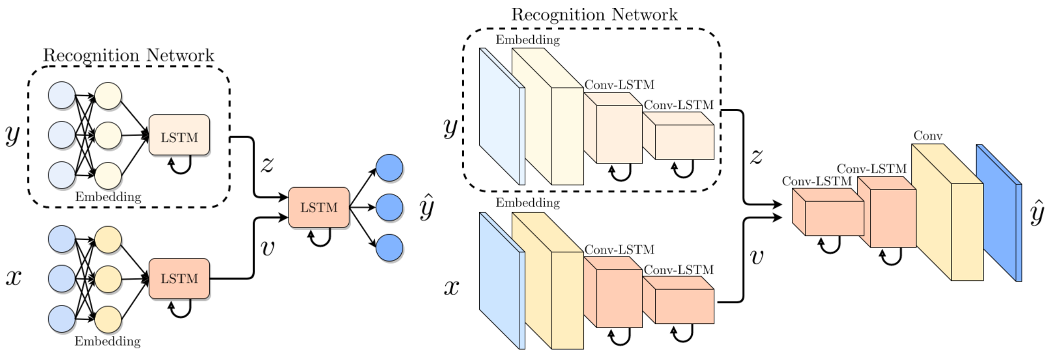
The integral in Eq. (6.2) is computationally intractable. In Sohn *et al.* (2015), a variational lower bound of the data log-likelihood Eq. (6.2) is derived, which can be estimated empirically using Monte-Carlo integration (also used in Lee *et al.* (2017b)),

$$\begin{aligned} \hat{\mathcal{L}}_{\text{CVAE}} = \frac{1}{T}\sum_{i=1}^{T} &\log\ p_\theta(y|\hat{z}_i, x) \\ - D_{\text{KL}}(q_\phi(z|x, y) \parallel p(z|x)),\ &\hat{z}_i \sim q_\phi(z|x, y). \end{aligned} \tag{6.3}$$

The lower bound in Eq. (6.3) weights all samples ($\hat{z}_i$) equally and so they must all ascribe high probability to the data point (x, y). This introduces a strong constraint on the recognition network q_ϕ. Therefore, the model is forced to trade-off between a good estimate of the data log-likelihood and the KL divergence between the training and test latent variable distributions. One possibility to close the gap introduced between the training and test pipelines, as described in Sohn *et al.* (2015), is to use an hybrid objective of the form $(1-\alpha)\hat{\mathcal{L}}_{\text{MC}} + \alpha\,\hat{\mathcal{L}}_{\text{CVAE}}$. Although such a hybrid objective has shown modest improvement in performance in certain cases, we could not observe any significant improvement over the standard CVAE objective in our structured sequence prediction tasks. In the following, we derive our novel "Best of many" samples objective which on the one hand encourages sample diversity and on the other hand aims to close the gap between the training and testing pipelines.

6.3.2 "Best of Many" Samples Objective

Here, we propose our objective which unlike Eq. (6.3) does not weigh each sample equally. Consider the functions $f_1(z) = {}^{p(z|x)}/_{q_\phi(z|x,y)}$ and $f_2(z) = p_\theta(y|z, x) \times q_\phi(z|x, y)$ in Eq. (6.2). We cannot evaluate $f_2(z)$ directly for Monte-Carlo samples.



(a) Our model for structured trajectory prediction.

(b) Our model for structured image sequence prediction.

Figure 6.3: Our model architectures. The recognition networks are only available during training.

Notice, however, that both $f_1(z)$ and $f_2(z)$ are continuous and positive. As $q_\theta(z|x,y)$ is normally distributed, the integral above can be very well approximated on a large enough bounded interval $[a,b]$. Therefore, we can use the First Mean Value Theorem of Integration Comenetz (2002) ‡, to separate the functions $f_1(z)$ and $f_2(z)$ in Eq. (6.2),

$$\log(p_\theta(y|x)) \geq \log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) + \log\Big(\frac{p(z'|x)}{q_\phi(z'|x,y)}\Big),\ z' \in [a,b]. \tag{6.4}$$

To do this, we set $f_1(z) = {}^{p(z|x)}/_{q_\phi(z|x,y)}$ and $f_2(z) = p_\theta(y|z,x) \times q_\phi(z|x,y)$ in Eq. (6.2). The integral in Eq. (6.2) can be very well approximated on a large enough bounded interval $[a,b]$. Thus, the integral in Eq. (6.2) can be expressed using the First Mean Value Theorem as,

$$\Big(\int_a^b p_\theta(y|z,x)\, \frac{p(z|x)}{q_\phi(z|x,y)}\, q_\phi(z|x,y)\, dz\Big) = \frac{p(z'|x)}{q_\phi(z'|x,y)}\ \Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big). \tag{6.5}$$

Taking log on both sizes of Eq. (6.5) leads to Eq. (6.4). We can further lower bound Eq. (6.4) leading to,

$$\log(p_\theta(y|x)) \geq \log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) + \min_{z' \in [a,b]} \Big(\log\Big(\frac{p(z'|x)}{q_\phi(z'|x,y)}\Big)\Big) \tag{6.6}$$

‡The First Mean Value Theorem of Integration states that, if $f_1 : [a,b] \to \mathbb{R}$ is continuous and f_2 is an integrable function that does not change sign on $[a,b]$, then $\exists z' \in (a,b)$ such that,

$$\int_a^b f_1(z)\, f_2(z)\, dz = f_1(z')\ \int_a^b f_2(z)\, dz \tag{S1}$$

However, the minimum in Eq. (6.6) is difficult to estimate. Therefore, we use the following approximation. From Eq. (6.4), we know that $\exists z' \ \in (a, b)$ which lower bounds the data log-likelihood. To maximize this data log-likelihood, we would like to maximize $\log(f_1(z'))$. However, as we do not know z', we instead choose to maximize it for a set of N points in (a, b),

$$\log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) + \log\Big(\frac{p(z_1'|x)}{q_\phi(z_1'|x,y)}\Big) + \ldots + \log\Big(\frac{p(z_N'|x)}{q_\phi(z_N'|x,y)}\Big). \tag{6.7}$$

As values of both p and q_ϕ are bounded above by 1, the value of the function $f_2(z_i') = {}^{p(z_i'|x)}/_{q_\phi(z_i'|x,y)}$ is likely to be low when is p low and q_ϕ is high. Therefore, to give more importance to such points z_i', we weight each point by $q_\phi(z_i'|x,y)$,

$$\begin{aligned}\log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) + q_\phi(z_1'|x,y) \times \log\Big(\frac{p(z_1'|x)}{q_\phi(z_1'|x,y)}\Big)\\ + \ldots + q_\phi(z_N'|x,y) \times \log\Big(\frac{p(z_N'|x)}{q_\phi(z_N'|x,y)}\Big).\end{aligned} \tag{6.8}$$

Flipping the sign before the terms in the second part of Eq. (6.8),

$$\begin{aligned}\log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) - q_\phi(z_1'|x,y) \times \log\Big(\frac{q_\phi(z_1'|x,y)}{p(z_1'|x)}\Big)\\ - \ldots - q_\phi(z_N'|x,y) \times \log\Big(\frac{q_\phi(z_N'|x,y)}{p(z_N'|x)}\Big).\end{aligned} \tag{6.9}$$

If we choose a sufficiently large set of points $z_i' \in (a, b)$, we can collect the terms in the second part of Eq. (6.9) and replace them with a single integral,

$$\log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) - \int_a^b q_\phi(z|x,y) \times \log\Big(\frac{q_\phi(z|x,y)}{p(z_|x)}\Big)\, dz. \tag{6.10}$$

The second integral in Eq. (6.10) is the KL divergence between the two distributions $q_\phi(z|x,y)$ and $p(z|x)$,

$$\log\Big(\int_a^b p_\theta(y|z,x)\, q_\phi(z|x,y)\, dz\Big) - D_{\text{KL}}(q_\phi(z|x,y) \parallel p(z|x)). \tag{6.11}$$

We can estimate the data log-likelihood term in Eq. (6.11) using Monte-Carlo integration. This leads to the following "many-sample" objective,

$$\hat{\mathcal{L}}_{\text{MS}} = \log\Big(\frac{1}{T}\sum_{i=1}^{T} p_\theta(y|\hat{z}_i, x)\Big) - D_{\text{KL}}(q_\phi(z|x,y) \parallel p(z|x)),\ \hat{z}_i \sim q_\phi(z|x,y). \tag{6.12}$$

Compared to the CVAE objective Eq. (6.2), the recognition network q_ϕ now has multiple chances to draw samples with high posterior probability ($p_\theta(y \mid z, x)$). This encourages diversity in the generated samples. Furthermore, the data log-likelihood Eq. (6.2) estimate in this objective is tighter as $\hat{\mathcal{L}}_{\text{MS}} \geq \hat{\mathcal{L}}_{\text{CVAE}}$ follows from the Jensen's inequality. Therefore, this bound loosens the constraints on the recognition network q_ϕ and allows it more closely match the latent variable distribution $p(z \mid x)$. However, as we focus on regression tasks, probabilities are of the form $e^{-\text{MSE}(\hat{y},y)}$. Therefore, in practice the Log-Average term can cause numerical instabilities due to limited machine precision in representing the probability $e^{-\text{MSE}(\hat{y},y)}$. Therefore, we use a "Best of Many" samples approximation of Eq. (6.12). We can pull the constant $1/T$ term outside the average in Eq. (6.12) and approximate the sum with the maximum,

$$\begin{aligned}\hat{\mathcal{L}}_{\text{MS}} = &\log\Big(\sum_{i=1}^{T} p_\theta(y|\hat{z}_i, x)\Big) - \log(T) \\ &- D_{\text{KL}}(q_\phi(z|x,y) \parallel p(z|x)),\ \hat{z}_i \sim q_\phi(z|x,y)\end{aligned} \tag{6.13}$$

$$\begin{aligned}\hat{\mathcal{L}}_{\text{MS}} \geq \hat{\mathcal{L}}_{\text{BMS}} = &\max_i \big(\log(p_\theta(y|\hat{z}_i, x))\big) - \log(T) \\ &- D_{\text{KL}}(q_\phi(z|x,y) \parallel p(z|x)),\ \hat{z}_i \sim q_\phi(z|x,y).\end{aligned} \tag{6.14}$$

Similar to Eq. (6.12), this objective encourages diversity and loosens the constraints on the recognition network q_ϕ as only the best sample is considered. During training, initially p_θ assigns low probability to the data for all samples $\hat{z}_i$. The $\log(T)$ difference between Eq. (6.12) and Eq. (6.14) would be dominated by the low data log-likelihood. Later on, as both objectives promote diversity, the Log-Average term in Eq. (6.12) would be dominated by one term in the average. Therefore, Eq. (6.12) would be well approximated by the maximum of the terms in the average. Furthermore, Eq. (6.14) avoids numerical stability issues.

6.3.3 Model Architectures for Structured Sequence Prediction

We base our model architectures on RNN Encoder-Decoders. We use LSTM formulations as RNNs for structured trajectory prediction tasks (Fig. 6.3(a)) and Convolutional LSTM formulations (Fig. 6.3(b)) for structured image sequence prediction tasks. During training, we consider LSTM recognition networks in case of trajectory prediction (Fig. 6.3(a)) and for image sequence prediction, we consider Conv-LSTM recognition networks (Fig. 6.3(b)). Note that, as we make the simplifying assumption that z is independent of x, the recognition networks are conditioned only on y.

Model for structured trajectory prediction. Our model for structured trajectory prediction (see Fig. 6.3(a)) is similar to the sampling module of Lee *et al.* (2017b). The input sequence x is processed using an embedding layer to extract features and the embedded sequence is read by the encoder LSTM. The encoder LSTM produces a summary vector v, which is its internal state after reading the input sequence x. The decoder LSTM is conditioned on the summary vector v and additionally a sample of the latent variable z. The decoder LSTM is unrolled in time and a prediction is generated by a linear transformation of it's output. Therefore, the predicted sequence at a certain time-step $\hat{y}^t$ is conditioned on the output at the previous time-step, the summary vector v and the latent variable z. As the summary v is deterministic given x, we have,

$$\begin{aligned} p_\theta(y|x) &= \sum_t p_\theta(y^{t+1}|y^t, v)\, p(v|x) \\ &= \sum_t p_\theta(y^{t+1}|y^t, x) \\ &= \int \sum_t p_\theta(y^{t+1}|y^t, z, x)\, p_\theta(z|x)\, dz. \end{aligned} \tag{6.15}$$

Conditioning the predicted sequence at all time-steps upon a single sample of z enables z to capture global characteristics (e.g. speed and direction of motion) of the future sequence and generation of temporally consistent sample sequences $\hat{y}$.

Extension with visual input. In case of dynamic agents e.g. pedestrians in traffic scenes, the future trajectory is highly dependent upon the environment e.g. layout of the streets. Therefore, additionally conditioning samples on sensory input (e.g. visuals of the environment) would enable more accurate sample generation. We use a CNN to extract a summary of a visual observation of a scene. This visual summary is given as input to the decoder LSTM, ensuring that the generated samples are additionally conditioned on the visual input.

Model for structured image sequence prediction. If the sequence (x, y) in question consists of images e.g. frames of a video, the trajectory prediction model Fig. 6.3(a) cannot exploit the spatial structure of the image sequence. More specifically, consider a pixel $y_{i,j}^{t+1}$ at time-step $t+1$ of the image sequence y. The pixel value at time-step $t+1$ depends upon only the pixel $y_{i,j}^t$ and a certain neighbourhood around it. Furthermore, spatially neighbouring pixels are correlated. This spatial structure can be exploited by using Convolutional LSTMs (Shi *et al.*, 2015) as RNN encoder-decoders. Conv-LSTMs retain spatial information by considering the hidden states h and cell states c as 3D tensors – the cell and hidden states are composed of vectors $c_{i,j}^t$, $h_{i,j}^t$ corresponding to each spatial position. New cell states, hidden states and outputs are computed using convolutional operations. Therefore, new cell states $c_{i,j}^{t+1}$, hidden states $h_{i,j}^{t+1}$ depend upon only a local spatial neighbourhood of $c_{i,j}^t$, $h_{i,j}^t$, thus preserving spatial information.

We propose conditional generative models networks with Conv-LSTMs for structured image sequence prediction (Fig. 6.3(b)). The encoder and decoder consists of two stacked Conv-LSTMs for feature aggregation. As before, the output is conditioned on a latent variable z to model multiple modes of the conditional distribution $p(y \mid x)$. The future states of neighboring pixels are highly correlated. However, spatially distant parts of the image sequences can evolve independently. To take into account the spatial structure of images, we consider latent variables z which are 3D tensors. As detailed in Fig. 6.3(b), the input image sequence x is processed using a convolutional embedding layer. The Conv-LSTM reads the embedded input sequence and produces a 3D tensor v as the summary. The 3D summary v and latent variable z is given as input to the Conv-LSTM decoder at every time-step. The cell state, hidden state or output at a certain spatial position, $c^t_{i,j}$, $h^t_{i,j}$, $y^t_{i,j}$, is conditioned on a sub-tensor $\underline{z}_{i,j}$ of the latent tensor z. Spatially neighbouring cell states, hidden states (and thus outputs) are therefore conditioned on spatially neighbouring sub-tensors $\underline{z}_{i,j}$. This coupled with the spatial information preserving property of Conv-LSTMs detailed above, enables z to capture spatial location specific characteristics of the future image sequence and allows for modeling the correlation of future states of spatially neighboring pixels. This ensures spatial consistency of sampled output sequences $\hat{y}$. Furthermore, as in the fully connected case, conditioning the full output sequence sample $\hat{y}$ is on a single sample of z ensures temporal consistency.

6.4 EXPERIMENTS

We evaluate our models both on synthetic and real data. We choose sequence datasets which display multimodality. In particular, we evaluate on key strokes from MNIST sequence data (D. De Jong, 2016) (which can be seen as trajectories in a constrainted space), human trajectories from Stanford Drone data (Robicquet *et al.*, 2016) and radar echo image sequences from HKO (Shi *et al.*, 2015). All models were trained using the ADAM optimizer, with a batch size of 32 for trajectory data and 4 for the radar echo data. All experiments were conducted on a single Nvidia M40 GPU with 12GB memory. For models trained using the $\hat{\mathcal{L}}_{\text{CVAE}}$ and $\hat{\mathcal{L}}_{\text{BMS}}$ objectives, we use $T = \{10, 10, 5\}$ samples during training on the MNIST Sequence, Stanford Drone, and HKO datasets respectively.

Method	CLL
LSTM	136.12
LSTM-MC	102.34
LSTM-CVAE	96.42
LSTM-BMS	**95.63**

Table 6.1: Evaluation on the MNIST Sequence dataset.

6.4.1 MNIST Sequence

The MNIST sequence dataset consists of pen strokes which closely approximate the skeleton of the digits in the MNIST dataset. We focus on the stroke completion task. Given an initial stroke the distribution of possible completions is highly multimodal. The digits 0, 3, 2 and 8, have the same initial stroke with multiple writing styles for each digit. Similarly for the digits 0 and 6, with multiple writing styles for each digit.

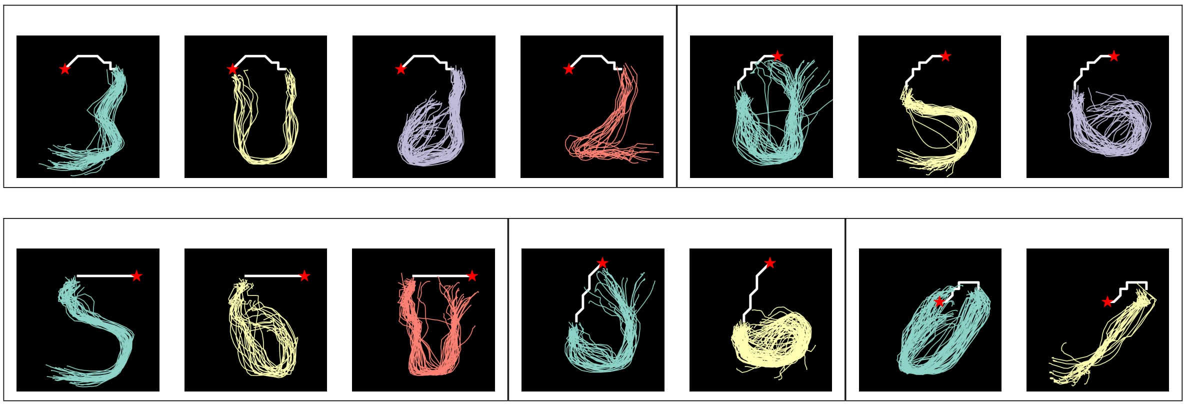


Figure 6.4: Diverse samples drawn from our LSTM-BMS model trained using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective, clustered using k-means. The number of clusters is set manually to the number of expected digits based on the initial stroke on the MNIST Sequence dataset.

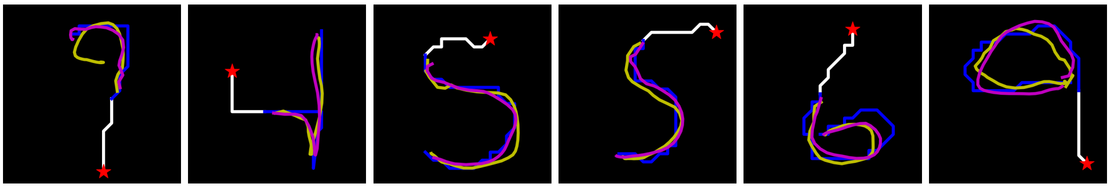


Figure 6.5: Top 10% of samples drawn from the LSTM-BMS model (magenta) and the LSTM-CVAE model (yellow), with the groundtruth in (blue) on the MNIST Sequence dataset.

Method	Visual	Err at 1.0(sec)	Err at 2.0(sec)	Err at 3.0(sec)	Err at 4.0(sec)	CLL
LSTM	x	1.08	2.57	4.70	7.20	134.29
LSTM	RGB	0.84	1.95	3.86	6.24	133.12
DESIRE-SI-IT4	RGB	1.29	2.35	3.47	5.33	x
LSTM-CVAE	RGB	**0.71**	1.86	3.39	5.06	127.51
LSTM-BMS	RGB	0.80	**1.77**	**3.10**	**4.62**	**126.65**

Table 6.2: Evaluation on the Stanford Drone dataset. Euclidean error measured at (1/5) resolution (DESIRE-SI-IT4 is from Lee *et al.* (2017b)).

We fix the length of the initial stroke sequence at 10. We use the trajectory prediction model from Fig. 6.3(a) and train it using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective (LSTM-BMS). We compare it against the following baselines: 1. A vanilla LSTM encoder-decoder regression model (LSTM) without latent variables; 2. The trajectory prediction model from Fig. 6.3(a) trained using the $\hat{\mathcal{L}}_{\text{MC}}$ objective (LSTM-MC); 3. The trajectory prediction model from Fig. 6.3(a) trained using the $\hat{\mathcal{L}}_{\text{CVAE}}$ objective (LSTM-CVAE). We use the negative conditional log-likelihood metric (CLL) and report the results in Table 6.1. We use $T = 100$ samples to estimate the CLL.

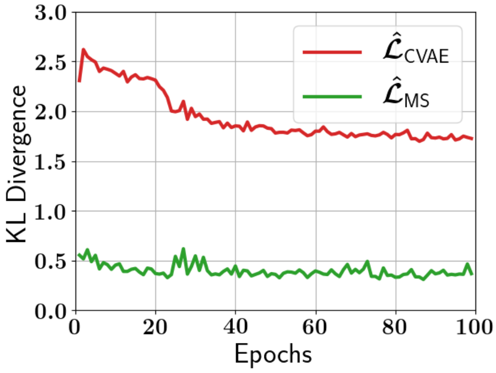


Figure 6.6: KL Divergence during training on the MNIST Sequence dataset.

We observe that our LSTM-BMS model achieves the best CLL. This means that our LSTM-BMS model fits the data distribution best. Furthermore, we see that the latent variables sampled from our recognition network $q_\phi(z \mid x, y)$ during training better matches the true distribution $p(z \mid x)$ used during testing. This can be seen through the KL divergence in Fig. 6.6 $D_{\text{KL}}(q_\phi(z \mid x, y) \parallel p(z \mid x))$ during training of the recognition network trained with the $\hat{\mathcal{L}}_{\text{BMS}}$ objective versus that of the $\hat{\mathcal{L}}_{\text{CVAE}}$ objective. We observe that the KL divergence of the recognition network trained with the $\hat{\mathcal{L}}_{\text{BMS}}$ to be substantially lower, thus, reducing the mismatch in the latent variable z between the training and testing pipelines.

We show qualitative examples of generated samples in Fig. 6.4 from the LSTM-BMS model. We show $T = 100$ samples per test example. The initial conditioning stroke is shown in white. The samples drawn are diverse and clearly multimodal. We cluster the generated samples using k-means for better visualization. The number of clusters is set manually to the number of expected digits based on the initial stroke. In particular, our model generates corresponding to 2, 3, 0 (1st example), 0, 6 (2nd example) and so on.

We compare the accuracy of samples generated by our LSTM-BMS model versus the LSTM-CVAE model in Fig. 6.5. We display the mean of the oracle top 10% of samples (closest in euclidean distance w.r.t. groudtruth) generated by both models. Comparing the results we see that, using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective leads to the generation of more accurate samples.

Method	Rainfall-MSE	CSI	FAR	POD	Correlation	NLL
Shi *et al.* (2015)	1.420	0.577	0.195	0.660	0.908	x
Conv-LSTM-CVAE	1.259	0.651	**0.155**	0.701	0.910	132.78
Conv-LSTM-BMS	**1.163**	**0.670**	0.163	**0.734**	**0.918**	**132.52**

Table 6.3: Evaluation on HKO radar image sequences.

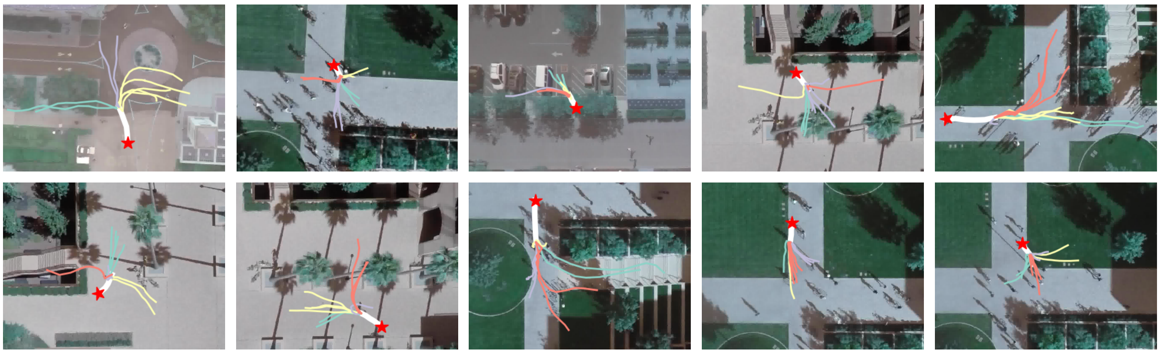


Figure 6.7: Diverse samples dawn from our LSTM-BMS model trained using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective, color-coded after clustering using k-means with four clusters on the Stanford Drone dataset

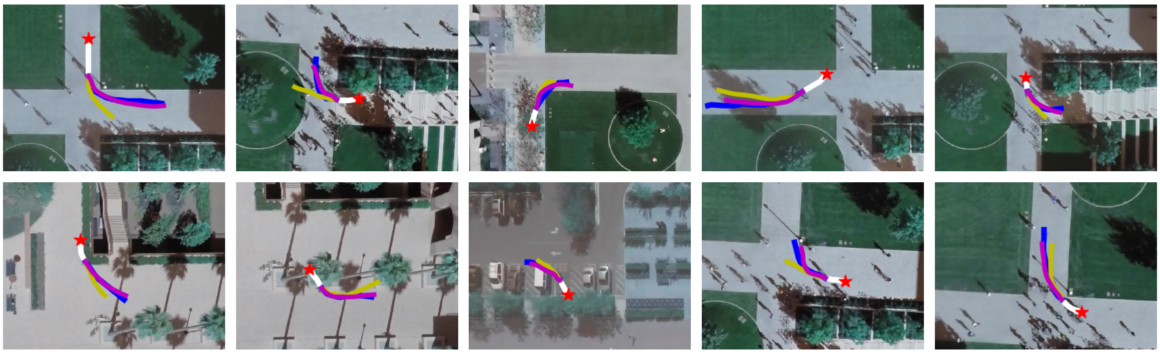


Figure 6.8: Top 10% of samples drawn from the LSTM-BMS model (magenta) and the LSTM-CVAE model (yellow), with the groundtruth in blue on the Stanford Drone dataset

6.4.2 Stanford Drone

The Stanford Drone dataset consists of overhead videos of traffic scenes. Trajectories of various dynamic agents including Pedestrians and Bikers are annotated. The paths of such agents are determined by various factors including the intention of the agent, paths of other agents and the layout of the scene. Thus, the trajectories are highly multi-modal. As in Robicquet *et al.* (2016); Lee *et al.* (2017b), we predict the trajectories of these agents 4.8 seconds into the future conditioned on the past 2.4 seconds. We use the same train-test split as in Lee *et al.* (2017b). We encode trajectories as relative displacement from the initial position. The trajectory at each time-step can be seen as the velocity of the agent.

We consider the extension of our trajectory prediction model (in Fig. 6.3(a)) discussed in Section 6.3.3 conditioned on the last visual observation from the overhead camera. We use a 6 layer CNN to extract visual features. We train this model with the $\hat{\mathcal{L}}_{\text{BMS}}$ objective and compare it to: 1. A vanilla LSTM encoder-decoder regression model with and without visual observation (LSTM); 2. The state of the art

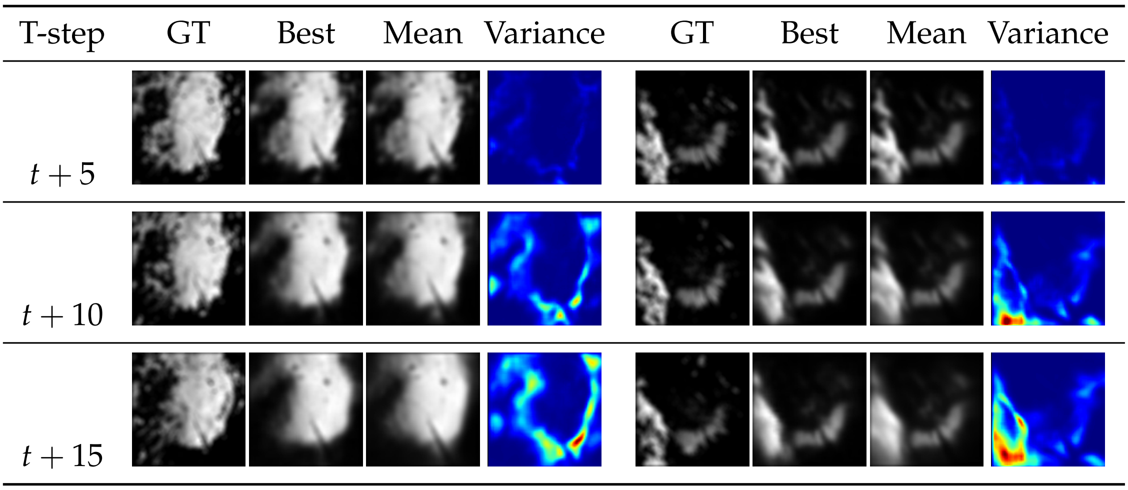


Figure 6.9: Statistics of samples generated by our LSTM-BMS model on the HKO dataset (T-step is Timestep and GT is groundtruth).

DESIRE-SI-IT4 model from Lee *et al.* (2017b); 3. Our extended trajectory prediction model Fig. 6.3(a) trained using the $\hat{\mathcal{L}}_{\text{CVAE}}$ objective (LSTM-CVAE).

We report the results in Table 6.2. We report the CLL metric and the euclidean distance in pixels between the true trajectory and the oracle top 10% of generated samples at 1, 2, 3 and 4 seconds into the future at (1/5) resolution (as in Lee *et al.* (2017b)). Our LSTM-BMS model again performs best both with respect to the euclidean distance and the CLL metric. This again demonstrates that using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective enables us to better fit the groundtruth data distribution and enables the generation of more accurate samples. The performance advantage with respect to DESIRE-SI-IT4 (Lee *et al.*, 2017b) is due to 1. Conditioning the decoder LSTM in Fig. 6.3(a) directly on the visual input at higher (1/2 versus 1/5) resolution (as our LSTM-CVAE outperforms DESIRE-SI-IT4), 2. Our $\hat{\mathcal{L}}_{\text{BMS}}$ objective (as our LSTM-BMS outperforms both DESIRE-SI-IT4 and LSTM-CVAE).

We show qualitative examples of generated samples ($T = 10$) in Fig. 6.7. We color code the generated samples using k-means with four clusters. The qualitative examples display high plausibility and diversity. They follow the layout of the scene, the location of roads, vegetation, vehicles etc. We qualitatively compare the accuracy of samples generated by our LSTM-BMS model versus the LSTM-CVAE model in Fig. 6.8. We see that the oracle top 10% of samples generated using the $\hat{\mathcal{L}}_{\text{BMS}}$ objective are more accurate and thus more representative of the groundtruth.

6.4.3 Radar Echo

The Radar Echo dataset (Shi *et al.*, 2015) consists of weather radar intensity images from 97 rainy days over Hong Kong from 2011 to 2013. The weather evolves due to varity of factors, which are difficult to identify using only the radar images, with varied and multimodal futures. Each sequences consists of 20 frames each of resolution 100×100, recorded at intervals of 6 minutes. We use the same dataset split as Shi *et al.* (2015) and predict the next 15 images given the previous 5 images.

We compare our image sequence prediction model in Fig. 6.3(b) trained with the $\hat{\mathcal{L}}_{\text{BMS}}$ (Conv-LSTM-BMS) objective to one trained with the $\hat{\mathcal{L}}_{\text{CVAE}}$ (Conv-LSTM-CVAE) objective. We additionally compare it to the Conv-LSTM model of Shi *et al.* (2015). In addition to the CLL metric (calculated per image sequence), we use the following precipitation nowcasting metrics from Shi *et al.* (2015), 1. Rainfall means squared error (Rainfall-MSE), 2. Critical success index (CSI), 3. False alarm rate (FAR), 4. Probability of detection (POD), and 5. Correlation. For fair comparison we estimate these metrics using $T = 1$ random samples from the Conv-LSTM-CVAE and Conv-LSTM-BMS models.

We report the results in Table 6.3. Both the Conv-LSTM-CVAE and Conv-LSTM-CMS models perform better compared to Shi *et al.* (2015). This is due to the use of embedding layers for feature extraction and the use of 2×2 max pooling in between two Conv-LSTM layers for feature aggregation (compared no embedding layers or pooling in Shi *et al.* (2015)). Furthermore, the superior CLL of the Conv-LSTM-BMS model demonstrates it's ability to fit the data distribution better. We show qualitative examples in Fig. 6.9 at $t + 1$, $t + 5$ and $t + 10$. We generate $T = 50$ samples and show the sample closest to the groundtruth, the mean of all the samples and the per-pixel variance (uncertainty) in the samples. The qualitative examples demonstrate that our model produces highly accurate and diverse samples.

6.5 CONCLUSION

We have presented a novel "Best of many" sample objective for Gaussian latent variable models and show its advantages for learning conditional models on multi-modal distributions. Our analysis shows indeed the learnt latent representation is better matched between training and test time – which in turn leads to more accurate samples. We show the benefits of our model on trajectory as well as image sequence prediction using three diverse datasets: MNIST strokes, Stanford drone and HKO weather. Our proposed appraoch consistently outperforms baselines and state of the art in all these scenarios.

7 CONDITIONAL FLOW VARIATIONAL AUTOENCODERS FOR STRUCTURED SEQUENCE PREDICTION

Contents

Prior work for structured sequence prediction based on latent variable models, introduced in Chapter 6, imposes priors with limited expressiveness or prior which are difficult to optimize, e.g. determining the number of Gaussian mixture components. This makes it challenging to fully capture the multi-modality of the distribution of the future states. In this chapter, we introduce *Conditional Flow Variational Autoencoders (CF-VAE)* using our novel conditional normalizing flow based prior to capture complex multi-modal conditional distributions for effective structured sequence prediction. Moreover, we propose two novel regularization schemes which stabilizes training and deals with posterior collapse for better fit to the target data distribution. Our experiments on three multi-modal structured sequence prediction datasets – MNIST Sequences, Stanford Drone and HighD – show that the proposed method obtains state of art results across different evaluation metrics.

7.1 INTRODUCTION

Conditional Variational Autoencoders (CVAE) (Sohn *et al.*, 2015; Bayer and Osendorfer, 2014; Chung *et al.*, 2015) have been very successful in future prediction problems – from prediction of pedestrians trajectories (Lee *et al.*, 2017b; Bhattacharyya *et al.*, 2018c; Pajouheshgar and Lampert, 2018) to outcomes of robotic actions (Babaeizadeh *et al.*, 2018). In complex real world environments, the distribution of future sequences is diverse and highly multi-modal. As discussed in Chapter 6, CVAEs model diverse

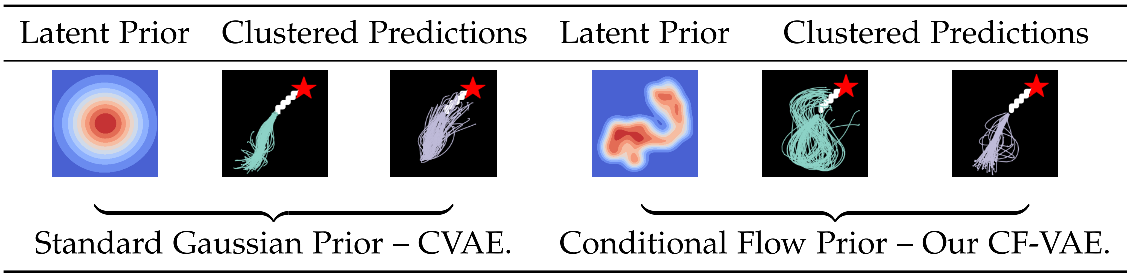


Figure 7.1: Clustered stroke predictions on MNIST sequences. Our multi-modal Conditional Normalizing Flow based prior (right) enables our regularized CF-VAE to capture the two modes of the conditional distribution, while predictions with unimodal Gaussian prior (left) have limited diversity. Note, our 64D CF-VAE latent distribution is (approximately) projected to 2D using tSNE and KDE.

futures by factorizing the distribution of future states using a set of latent variables which are mapped to likely future states. However, CVAEs assume a standard Gaussian prior on the latent variables which induces a strong model bias (Hoffman and Johnson, 2016; Tomczak and Welling, 2018) which makes it challenging to capture multi-modal distributions. This also leads to missing modes due to posterior collapse (Bowman *et al.*, 2016; Razavi *et al.*, 2019a).

Recent work (Tomczak and Welling, 2018; Wang *et al.*, 2017; Gu *et al.*, 2019) has therefore focused on more complex Gaussian mixture based priors. Gaussian mixtures still have limited expressiveness and optimization suffers from complications, e.g. determining the number of mixture components. Normalizing flows are more expressive and enable the modelling of complex multi-modal priors. Recent work on flow based priors (Chen *et al.*, 2017; Ziegler and Rush, 2019), have focused only on the unconditional (plain VAE) case. However, this is not sufficient for CVAEs because in the conditional case the complexity of the distributions are highly dependent on the condition.

In this chapter, 1. We propose *Conditional Flow Variational Autoencoders (CF-VAE)* based on novel conditional normalizing flow based priors in order to model complex multi-modal conditional distributions over sequences. In Fig. 7.1, we show example predictions of MNIST handwriting stroke of our CF-VAE. We observe that, given a starting stroke, our CF-VAE model with data dependent normalizing flow based latent prior captures the two main modes of the conditional distribution – i.e. 1 and 8 – while CVAEs with fixed unimodal Gaussian prior predictions have limited diversity. 2. We propose a regularization scheme that stabilizes the optimization of the evidence lower bound and leads to better fit to the target data distribution. 3. We leverage our conditional flow prior to deal with posterior collapse which causes standard CVAEs to ignore modes in sequence prediction tasks. 4. Finally, our method outperforms the state of the art on three structured sequence prediction tasks – handwriting stroke prediction on MNIST, trajectory prediction on Stanford Drone and HighD.

Note that, as this chapter is based on the work Bhattacharyya *et al.* (2019c), we compare to prior work on the Stanford Drone dataset; Lee *et al.* (2017b); Pajouheshgar and Lampert (2018); Gupta *et al.* (2018); Zhao *et al.* (2019); Sadeghian *et al.* (2019); Deo and Trivedi (2019). We provide an overview of more recent work, e.g. Mangalam *et al.* (2020) in Chapter 2.

7.2 RELATED WORK

While we provide a broader discussion on related work in Chapter 2, here we discuss related work relevant to this chapter.

Posterior collapse. Posterior collapse arises when the latent posterior does not encode useful information. Most prior work (Yang *et al.*, 2017; Dieng *et al.*, 2019; Higgins *et al.*, 2017) concentrate on unconditional VAEs and modify the training objective – the KL divergence term is annealed to prevent collapse to the prior. Fu *et al.* (2019) extends KL annealing to CVAEs. However, KL annealing does not optimize a true lower bound of the ELBO for most of training. Zhao *et al.* (2017c) also modifies the objective to choose the model with the maximal rate. Razavi *et al.* (2019a) propose anti-causal sequential priors for text modelling tasks. Bowman *et al.* (2016); Gulrajani *et al.* (2017b) proposes to weaken the decoder so that the latent variables cannot be ignored, however only unconditional VAEs are considered. Wang and Wang (2019) shows the advantage of normalizing flow based posteriors for preventing posterior collapse. In contrast, we study for the first time posterior collapse in conditional models on datasets with minor modes.

7.3 CONDITIONAL FLOW VARIATIONAL AUTOENCODER (CF-VAE)

Our Conditional Flow Variational Autoencoder is based on the conditional variational autoencoder (Sohn *et al.*, 2015) which is a deep directed graphical model for modeling conditional data distributions $p_\theta(\mathrm{y}|\mathrm{x})$. Here, x is the sequence up to time t, $x = [x^1, \cdots, x^t]$ and y is the sequence to be predicted up to time T, $y = [y^{t+1}, \cdots, y^T]$. CVAEs factorize the conditional distribution using latent variables z. In detail, $p_\theta(\mathrm{y}|\mathrm{x}) = \int p_\theta(\mathrm{y}|\mathrm{z},\mathrm{x})p(\mathrm{z}|\mathrm{x})dz$, where $p(\mathrm{z}|\mathrm{x})$ is the prior on the latent variables. During training, amortized variational inference is used and the posterior distribution $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$ is learnt using a recognition network. The ELBO is maximized, given by,

$$\log(p_\theta(\mathrm{y}|\mathrm{x})) \geq \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(p_\theta(\mathrm{y}|\mathrm{z},\mathrm{x})) - D_{\mathrm{KL}}(q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})||p(\mathrm{z}|\mathrm{x})). \tag{7.1}$$

In practice, to simplify learning, simple unconditional standard Gaussian priors are used (Sohn *et al.*, 2015). However, the complexity, e.g. the number of modes of the target distributions $p_\theta(\mathrm{y}|\mathrm{x})$, is highly dependent upon the condition x. An unconditional prior demands identical latent distributions irrespective of the complexity of the target conditional distribution – a very strong constraint on the recognition network. Moreover, the latent variables cannot encode any conditioning information and this leaves the burden of learning the dependence on the condition completely on the decoder.

Furthermore, on complex conditional multi-modal data, Gaussian priors have been shown to induce a strong model bias (Tomczak and Welling, 2016; Ziegler and Rush, 2019). It becomes increasingly difficult to map complex multi-modal distributions to unimodal Gaussian distributions, further complicated by the sensitivity of the RNNs encoder/decoders to subtle variations in the hidden states (Bowman *et al.*, 2016). Moreover, the standard closed form estimate of the KL-divergence pushes the encoded latent distributions to the mean of the Gaussian leading to latent variable collapse (Wang *et al.*, 2017; Gu *et al.*, 2019) while discriminator based approaches (Tolstikhin *et al.*, 2018) lead to underestimates of the KL-divergence (Rosca *et al.*, 2017).

Therefore, we propose conditional priors based on conditional normalizing flows to enable the latent variables to encode conditional information and allow for complex multi-modal latent representations. Next, we introduce our new conditional non-linear normalizing flows followed by our regularized Conditional Flow Variational Autoencoder (CF-VAE) formulation.

7.3.1 Conditional Normalizing Flows

Recently, normalizing flow (Tabak *et al.*, 2010; Dinh *et al.*, 2015) based priors for VAEs have been proposed (Chen *et al.*, 2017; Ziegler and Rush, 2019). Normalizing flows allows for complex priors by transforming a simple base density, e.g. standard Gaussian, to a complex multi-modal density through a series of n layers of invertible transformations f_i,

$$\epsilon \overset{f_1}{\longleftrightarrow} \mathrm{h}_1 \overset{f_2}{\longleftrightarrow} \mathrm{h}_2 \cdots \overset{f_n}{\longleftrightarrow} \mathrm{z}. \tag{7.2}$$

However, such flows cannot model conditional priors. In contrast to prior work, we utilize conditional normalizing flows to model complex conditional priors. Conditional normalizing flows also consists of a series of n layers of invertible transformations f_i (with parameters ψ), however we modify the transformations f_i such that they are dependent on the condition x,

$$\epsilon|\mathrm{x} \overset{f_1|\mathrm{x}}{\longleftrightarrow} \mathrm{h}_1|\mathrm{x} \overset{f_2|\mathrm{x}}{\longleftrightarrow} \mathrm{h}_2|\mathrm{x} \cdots \overset{f_n|\mathrm{x}}{\longleftrightarrow} \mathrm{z}|\mathrm{x}. \tag{7.3}$$

Further, in contrast to prior work (Lu and Huang, 2020; Atanov *et al.*, 2019; Ardizzone *et al.*, 2019a) which use affine flows (f_i), we build upon Ziegler and Rush (2019) and introduce conditional non-linear normalizing flows with split coupling. Split couplings ensure invertibility by applying a flow layer f_i on only half of the dimensions at a time. To compute Eq. (7.5), we split the dimensions z^D of the latent variable into halfs, $\mathrm{z}^L = \{1, \cdots, D/2\}$ and $\mathrm{z}^R = \{D/2, \cdots, d\}$ at each invertible layer f_i. Our transformation takes the following form for each dimension z^j alternatively from z^L or z^R,

$$f_i^{-1}(\mathrm{z}^j|\mathrm{z}^R, \mathrm{x}) = \epsilon^j = a(\mathrm{z}^R, \mathrm{x}) + b(\mathrm{z}^R, \mathrm{x}) \times \mathrm{z}^j + \frac{c(\mathrm{z}^R, \mathrm{x})}{1 + \left(d(\mathrm{z}^R, \mathrm{x}) \times \mathrm{z}^j + g(\mathrm{z}^R, \mathrm{x})\right)^2}. \tag{7.4}$$

where, $\mathrm{z}^j \in \mathrm{z}^L$. To ensure that the generated prior distribution is conditioned on x, in Eq. (7.4) and in the corresponding forward operation f_i, the coefficients $\{a, b, c, d, g\} \in \mathbb{R}$ are functions of both the other half of the dimensions of z *and* the condition x (unlike Ziegler and Rush (2019)). Finally, due to the expressive power of our conditional non-linear normalizing flows, simple spherical Gaussians base distributions were sufficient.

7.3.2 Variational Inference using Flow based Priors

Here, we derive the ELBO in Eq. (7.1) for our regularized CF-VAE with our conditional flow based prior. In the case of the standard CVAE with the Gaussian prior, the KL divergence term in the ELBO has a simple closed form expression. In case of our conditional flow based prior, we can use the change of variables formula to compute the KL divergence. In detail, given the base density $p(\epsilon|\mathrm{x})$ and the Jacobian J_i of each layer i of the transformation, the log-likelihood of the latent variable z under the prior can be expressed using the change of variables formula,

$$\log(p_\psi(\mathrm{z}|\mathrm{x})) = \log(p(\epsilon|\mathrm{x})) + \sum_{i=1}^{n} \log(|\det J_i|). \tag{7.5}$$

This change of variables allows us to evaluate the likelihood of latent variable z over the base distribution instead of the complex conditional prior and to express the KL divergence as,

$$\begin{aligned} -D_{\mathrm{KL}}(q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})||p_\psi(\mathrm{z}|\mathrm{x})) &= -\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})) + \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(p_\psi(\mathrm{z}|\mathrm{x})) \\ &= \mathcal{H}(q_\phi) + \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(p(\epsilon|\mathrm{x})) + \sum_{i=1}^{n} \log(|\det J_i|). \end{aligned} \tag{7.6}$$

where, $\mathcal{H}(q_\phi)$ is the entropy of the variational distribution. Therefore, the ELBO can be expressed as,

$$\begin{aligned} \log(p_\theta(\mathrm{y}|\mathrm{x})) \geq \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(p_\theta(\mathrm{y}|\mathrm{z},\mathrm{x})) + \mathcal{H}(q_\phi) + \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})} \log(p(\epsilon|\mathrm{x})) \\ + \sum_{i=1}^{n} \log(|\det J_i|) \end{aligned} \tag{7.7}$$

To learn complex conditional priors, we alternately optimize both the variational posterior distribution $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$ and the conditional prior $p_\psi(\mathrm{z}|\mathrm{x})$ in Eq. (7.7). This would allow the variational posterior q_θ to match the conditional prior and vice-versa so that the ELBO Eq. (7.7) is maximized. However, in practice we observe instabilities during training and posterior collapse. Next, we introduce our novel regularization schemes to deal with both these problems.

Posterior regularization for stability (pR). The entropy and the log-Jacobian of the CF-VAE objective Eq. (7.7)) are at odds with each other. The log-Jacobian

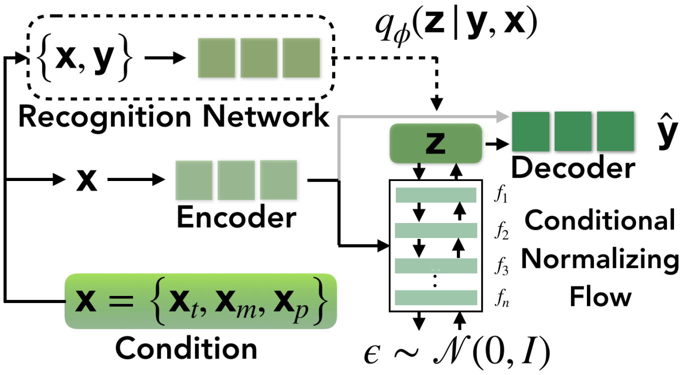

$q_\phi(\mathbf{z}|\mathbf{y},\mathbf{x})$
$\{\mathbf{x},\mathbf{y}\}$
Recognition Network
$\mathbf{x}$
Encoder
$\mathbf{x} = \{\mathbf{x}_t, \mathbf{x}_m, \mathbf{x}_p\}$
Condition
$\mathbf{z}$
$\hat{\mathbf{y}}$
Decoder
f_1
f_2
f_3
f_n
Conditional Normalizing Flow
$\epsilon \sim \mathcal{N}(0, I)$

Figure 7.2: CF-VAE. The decoder is regularized by removing conditioning (grey arrow) to prevent posterior collapse.

favours the contraction of the base density. Therefore, log-Jacobian at the right of Eq. (7.7) is maximized when the conditional flow maps the base distribution ($\epsilon \leftrightarrow$ z in Fig. 7.2) to a low entropy conditional prior and thus a low entropy variational distribution $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$. Therefore, in practice we observe instabilities during training. We observe that either the entropy or the log-Jacobian term dominates and the data log-likelihood is fully or partially ignored. Therefore, we regularize the posterior $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$ by fixing the variance to C. This leads to a constant entropy term which in turn bounds the maximum possible amount of contraction, thus upper bounding the log-Jacobian. This encourages our model to concentrate on explaining the data and leads to a better fit to the target data distribution. Note that, although $q_\phi(\mathrm{z}|\mathrm{x},\mathrm{y})$ has fixed variance, this does not significantly effect sample quality as the marginal $q_\phi(\mathrm{z}|\mathrm{x})$ can be arbitrarily complex due to our conditional flow prior. Moreover, we observe that the LSTM based decoders employed demonstrate robust performance across a wide range of values $\mathrm{C} = [0.05, 0.25]$.

Condition regularization for posterior collapse (cR). We observe missing modes when the target conditional data distribution has a major mode(s) and one or more minor modes (corresponding to rare events). This is because the condition x on the decoder is already enough to model the main mode(s). If the cost of ignoring the minor modes is out-weighed by the cost of encoding a more complex latent distribution reflecting all modes, the minor modes and the latent variables are ignored. We propose a regularization scheme by removing the additional conditioning x on the decoder, when the dataset in question has a dominating mode(s). This enabled by our conditional flow prior, which ensures that conditioning information is encoded in the latent space and $p_\theta(\mathrm{y}|\mathrm{z})$ can match $p_\theta(\mathrm{y}|\mathrm{x},\mathrm{z})$. Leading to a simpler factorization, $p_\theta(\mathrm{y}|\mathrm{x}) = \int p_\theta(\mathrm{y}|\mathrm{z}) p_\psi(\mathrm{z}|\mathrm{x}) dz$. Equivalently, this ensures that the latent variable z cannot be ignored by the CF-VAE and thus must encode useful information. Note that this regularization scheme is only possible due to our conditional prior, the unconditional Gaussian prior of CVAE would always need to condition the decoder.

The parallel work of Klushyn *et al.* (2019) also proposes a similar regularization scheme. However, we employ this regularization to deal with posterior collapse only in case of distributions with dominant modes.

Finally, we discuss the integration of diverse sources of contextual information into the conditional prior $p_{\psi}(\mathrm{z}|\mathrm{x})$ for even richer conditional latent distributions of our regularized CF-VAE.

7.3.3 Conditioning Priors on Contextual Information

For prediction tasks, it is often crucial to integrate sources of contextual information, e.g. past trajectories or environmental information, for accurate predictions. As these sources are heterogeneous, we employ source specific networks to extract fixed length vectors from each source.

Past trajectory. We encode the past trajectories using a LSTM to an fixed length vector x_t. For efficiency we share the condition encoder between the conditional flow and the CF-VAE decoder.

Environmental map. We use a CNN to encode environmental information to a set of region specific feature vectors. We apply attention conditioned on the past trajectory to extract a fixed length conditioning vector x_m, such that x_m contains information relevant to the future trajectory.

Interacting agents. To encode information of interacting traffic participants/agents, we build on Deo and Trivedi (2018) and propose a fully convolutional social pooling layer. We aggregate information of interacting agents using a grid overlaid on the environment. This grid is represented using a tensor, where the past trajectory information of traffic participants are aggregated into the tensor indexed corresponding to the grid in the environment. In Deo and Trivedi (2018) past trajectory information is aggregated using a LSTM. We aggregate the past trajectory information into the tensor using 1×1 convolutions as it allows for stable learning and is computationally efficient. Finally, we apply several layers of $k \times k$ convolutions to capture interaction aware contextual features x_p of traffic participants in the scene.

Due to the expressive power of our conditional non-linear normalizing flows, simple concatenation into a single vector $\mathrm{x} = \{\mathrm{x}_t, \mathrm{x}_m, \mathrm{x}_t\}$ was sufficient to learn powerful conditional priors.

7.4 EXPERIMENTS

We evaluate our CF-VAE on three popular and highly multi-modal sequence prediction datasets. We begin with a description of our evaluation metrics and model architecture.

Evaluation metrics. In line with prior work (Lee *et al.*, 2017b; Bhattacharyya *et al.*, 2018c; Pajouheshgar and Lampert, 2018; Deo and Trivedi, 2019; Bhattacharyya *et al.*, 2019a) (see also Chapter 6), we use the negative conditional log-likelihood (-CLL) and mean Euclidean distances of the oracle Top $n\%$ of N predictions. The oracle

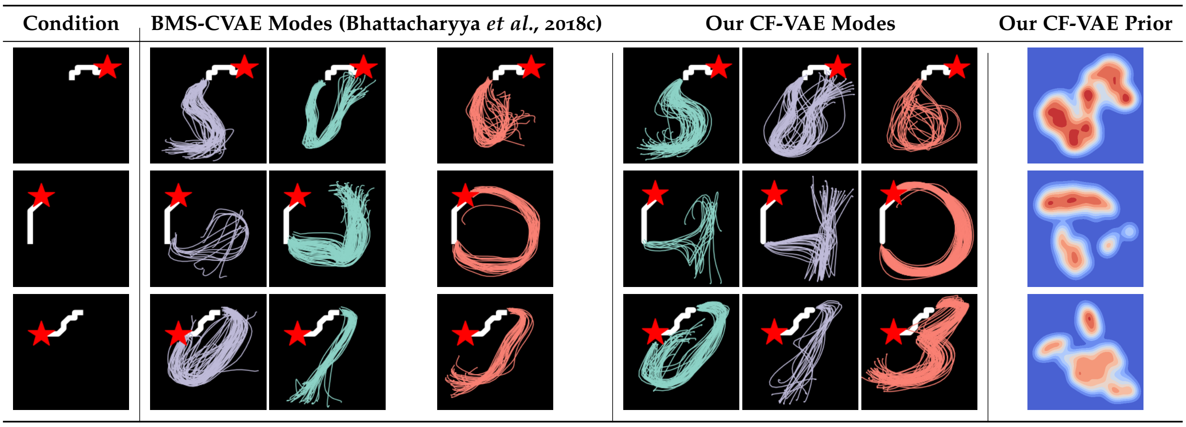


Figure 7.3: Random samples clustered using k-means. The number of clusters is set manually to the number of expected digits. The corresponding priors of our CF-VAE + pR on the right. Note, our 64D CF-VAE latent distribution is (approximately) projected to 2D using tSNE and KDE.

Top $n\%$ metric measures not only the coverage of all modes but also discourages random guessing for a reasonably large value of n (e.g. $n = 10\%$). This is because, a model can only improve this metric by moving randomly guessed samples from an overestimated mode to the correct modes.

Conditional flow model architecture. Our conditional flow prior consists of 16 layers of conditional non-linear flows with split coupling. Increasing the number of conditional non-linear flows generally led to "over-fitting" on the training latent distribution.

7.4.1 MNIST Sequences

The MNIST Sequence dataset (D. De Jong, 2016) consists of sequences of handwriting strokes of the MNIST digits. The state-of-the-art approach is the "Best-of-Many"-CVAE (Bhattacharyya *et al.*, 2018c) in Chapter 6 with a Gaussian prior. We follow the evaluation protocol of Bhattacharyya *et al.* (2018c) (in Chapter 6) and predict the complete stroke given the first ten steps. We also compare with, 1. A standard CVAE with uni-modal Gaussian prior; 2. A CVAE with a data dependent conditional mixture of Gaussians (MoG) prior; 3. A CF-VAE without any regularization ; 4. A CF-VAE without the conditional non-linear flow layers (CF-VAE-*Affine*, replaced with affine flows (Lu and Huang, 2020; Atanov *et al.*, 2019)). We also experiment with a conditional MoG prior. We use the same model architecture (Bhattacharyya *et al.*, 2018c) across all baselines.

We report the results in Table 7.1. We see that our CF-VAE with posterior regularization (pR) performs best. It has a performance advantage of over 20% against the state of the art BMS-CVAE. We see that without regularization (pR) (C = 0.2) there is a 40% drop in performance, highlighting the effectiveness of

Table 7.1: Evaluation on MNIST Sequences.

Method	-CLL $\downarrow$
CVAE (Sohn *et al.*, 2015)	96.4
BMS-CVAE (Bhattacharyya *et al.*, 2018c)	95.6
CVAE + *increased capacity* (Ours)	94.5
CVAE + *conditional prior* (Ours)	88.9
MoG-CVAE, $M = 3$	84.6
CF-VAE - *no regularization* (Ours)	104.3
CF-VAE - *Affine* + pR, C = 0.2 (Ours)	77.2
CF-VAE + pR, C = 0.2 (Ours)	**74.9**

our proposed regularization scheme. We further illustrate the modes captured and the learnt multi-modal conditional flow priors in Fig. 7.3. We do not use condition regularization here (cR) as we do not observe posterior collapse. In contrast, the BMS-CVAE is unable to fully capture all modes – its predictions are pushed to the mean due to the strong model bias induced by the Gaussian prior. The results improve considerably with the multi-modal MoG prior ($M = 3$ components work best). We also experiment with optimizing the standard CVAE architecture. This improves performance only slightly (after increasing LSTM encoder/decoder units to 256 from 48, increasing the number of layers did not help). Moreover, our experiments with a conditional (MoG) AAE/WAE (Gu *et al.*, 2019) based baseline did not improve performance beyond the standard (MoG) CVAE, because the discriminator based KL estimate tends to be an underestimate (Rosca *et al.*, 2017). This illustrates that in practice it is difficult to map highly multi-modal sequences to a Gaussian prior and highlights the need of a data-dependent multi-modal priors. Our CF-VAE still significantly outperforms the MoG-CVAE as normalizing flows are better at learning complex multi-modal distributions (Kingma and Dhariwal, 2018). We also see that affine conditional flow based priors leads to a drop in performance (77.2 vs 74.9 CLL) illustrating the advantage of our non-linear conditional flows.

7.4.2 Stanford Drone

The Stanford Drone dataset (Robicquet *et al.*, 2016) consists of multi-model trajectories of traffic participants, e.g. pedestrians, bicyclists, cars captured from a drone. Prior works follow two different evaluation protocols, 1. (Lee *et al.*, 2017b; Bhattacharyya *et al.*, 2018c; Pajouheshgar and Lampert, 2018) (see also Chapter 6) use 5 fold cross validation, 2. (Robicquet *et al.*, 2016; Sadeghian *et al.*, 2018, 2019; Deo and Trivedi, 2019) use a single split. We evaluate using the first protocol in Table 7.2 and the second in Table 7.3.

Additionally, Pajouheshgar and Lampert (2018) suggest a "Shotgun" baseline. This baseline extrapolates the trajectory from the last known position and orientation

Method	Visual	Error @ 1sec	Error @ 2sec	Error @ 3sec	Error @ 4sec	-CLL ↓
"Shotgun" (Top 10%) (Pajouhesghar and Lampert, 2018)	None	0.7	1.7	3.0	4.5	91.6
DESIRE-SI-IT4 (Top 10%) (Lee *et al.*, 2017b)	RGB	1.2	2.3	3.4	5.3	x
STCNN (Top 10%) (Pajouhesghar and Lampert, 2018)	RGB	1.2	2.1	3.3	4.6	x
BMS-CVAE (Top 10%) (Bhattacharyya *et al.*, 2018c)	RGB	0.8	1.7	3.1	4.6	126.6
MoG-CVAE, $M = 3$ (Top 10%)	None	0.8	1.7	2.7	3.9	86.1
CF-VAE - *no regularization* (Ours, Top 10%)	None	0.9	1.9	3.3	4.7	96.2
CF-VAE + pR, C = 0.2 (Ours, Top 10%)	None	**0.7**	**1.5**	2.5	3.6	84.6
CF-VAE + pR, C = 0.2 (Ours, Top 10%)	RGB	**0.7**	**1.5**	**2.4**	**3.5**	**84.1**

Table 7.2: Five fold cross validation on the Stanford Drone dataset. Euclidean error at (1/5) resolution.

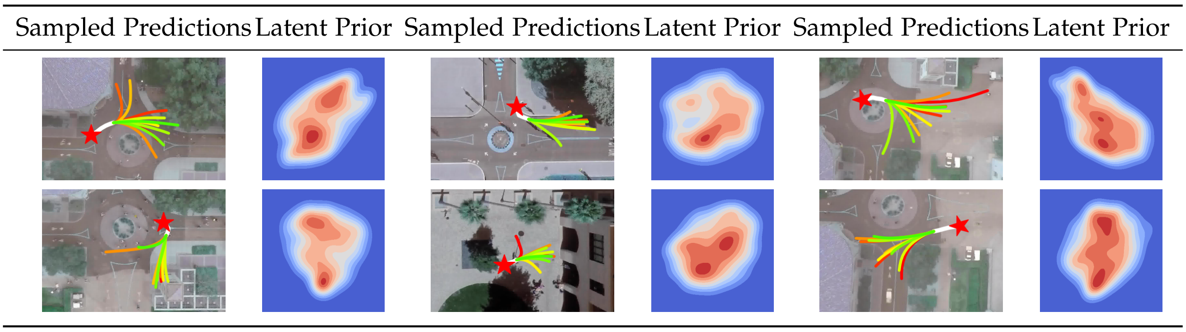


Figure 7.4: Randomly sampled predictions of our CF-VAE + pR model on the Stanford Drone. We observe that our predictions are highly multi-modal and are reflected by the Conditional Flow Priors. Note, our 64D CF-VAE latent distribution is (approximatly) projected to 2D using tSNE and KDE.

in 10 different ways – 5 orientations: $(0^\circ, \pm 8^\circ, \pm 15^\circ)$ and 5 velocities: None or exponentially weighted over the past with coefficients $(0, 0.3, 0.7, 1.0)$. This baseline obtains results at par with the state-of-the-art because it is a good template which covers the most likely possible futures (modes) for traffic participant motion in this dataset. We report the results using 5 fold cross validation in Table 7.2. We additionally compare to a mixture of Gaussians prior. We use the same model architecture as in Bhattacharyya *et al.* (2018c) (Chapter 6) and a CNN encoder with attention to extract features from the last observed RGB image. These visual features serve as additional conditioning (x_m) to our Conditional Flow model. We see that our CF-VAE model with RGB input and posterior regularization (pR) performs best – outperforming the state-of-the-art "Shotgun" and BMS-CVAE by over 20% (Error @ 4sec). We see that our conditional flows are able to utilize visual scene (RGB) information to improve performance (3.5 vs 3.6 Error @ 4sec). We also see that the MoG-CVAE and our CF-VAE + pR outperforms the BMS-CVAE, even without visual scene information. This again reinforces our claim that the standard Gaussian prior induces a strong model bias and data dependent multi-modal priors are needed for best performance. The performance advantage of CF-VAE over the MoG-CVAE again illustrates the advantage of normalizing flows at learning complex conditional multi-modal distributions. The performance advantage over the "Shotgun" baseline

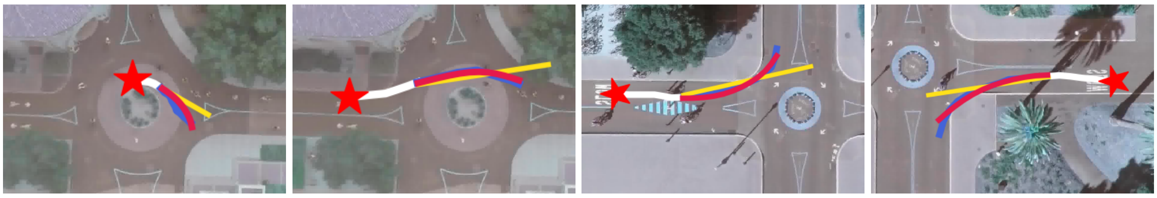


Figure 7.5: Comparison of our CF-VAE + pR (Red) and the "Shoutgun" baseline (Yellow) of (Pajouheshgar and Lampert, 2018), Groundtruth (Blue). Initial conditioning trajectory in white. Our CF-VAE not only learns to capture the correct modes but also generates more fine-grained predictions.

Method	mADE ↓	mFDE ↓
SocialGAN (Gupta *et al.*, 2018)	27.2	41.4
MATF GAN (Zhao *et al.*, 2019)	22.5	33.5
SoPhie (Sadeghian *et al.*, 2019)	16.2	29.3
Goal Prediction (Deo and Trivedi, 2019)	15.7	28.1
CF-VAE + pR, $C = 0.2$ (Ours)	**12.6**	**22.3**

Table 7.3: Evaluation on the Stanford Drone dataset on a single split (see also Table 7.2).

shows that our CF-VAE + pR not only learns to capture the correct modes but also generates more fine-grained predictions. The qualitative examples in Fig. 7.5 shows that our CF-VAE is better able to capture complex trajectories with sharp turns.

We report results using the single train/test split of (Robicquet *et al.*, 2016; Sadeghian *et al.*, 2018, 2019; Deo and Trivedi, 2019) in Table 7.3. We use the minimum Average Displacement Error (mADE) and minimum Final Displacement Error (mFDE) metrics as in (Deo and Trivedi, 2019). The minimum is over as set of predictions of size N. Although this metric is less robust to random guessing compared to the Top $n\%$ metric, it avoids rewarding random guessing for a small enough value of N. We choose $N = 20$ as in (Deo and Trivedi, 2019). Similar to the results with 5 fold cross validation, we observe 20% improvement over the state-of-the-art.

7.4.3 HighD

The HighD dataset (Krajewski *et al.*, 2018) consists of vehicle trajectories recorded using a drone over highways. In contrast to other vehicle trajectory datasets e.g. NGSIM it contains minimal false positive trajectory collisions or physically improvable velocities.

The HighD dataset is challenging because lane changes or interactions are rare $\sim$ 10% of all trajectories. The distribution of future trajectories contain a single main mode (linear continuations) along with several minor modes. Thus, approaches

Method	Context	ADE ↓	FDE ↓	-CLL ↓
Constant Velocity	None	1.09	2.66	x
FF (Diehl *et al.*, 2019)	None	0.45	1.09	x
GAT (Diehl *et al.*, 2019)	Yes	0.47	1.04	x
CVAE (Top 10%)	None	0.45	0.96	5.32
CVAE + *Cyclic KL* (Top 10%)	None	0.38	0.80	4.80
CF-VAE + pR, (Ours, Top 10%)	None	0.44	0.94	4.71
CF-VAE + {pR,cR}, (Ours, Top 10%)	None	0.30	0.57	3.64
CF-VAE + {pR,cR}, (Ours, Top 10%)	Yes	**0.29**	**0.55**	**3.42**

Table 7.4: Evaluation on the HighD dataset.

which predict a single mean trajectory (targeting the main mode) are challenging to outperform. In Table 7.4, we see that the simple Feed Forward (FF) model performs well and the Graph Convolutional GAT model of Diehl *et al.* (2019), which captures interactions, only narrowly outperforms the FF model. This dataset is challenging for CVAE based models as they frequently suffer from posterior collapse when a single mode dominates. This is clearly observed with our CVAE baseline in Table 7.4. To prevent posterior collapse, we use the cyclic KL annealing scheme proposed in Fu *et al.* (2019) (using a MoG prior did not help). This already leads to significant improvement over the deterministic FF and GAT baselines. We also observe posterior collapse with our CF-VAE model. Therefore, we regularize by removing additional conditioning (cR). Our CF-VAE + {pR,cR} with condition regularization significantly outperforms the CF-VAE + pR and CVAE baselines (with cyclic KL annealing), demonstrating the effectiveness of our condition regularization scheme (cR) in preventing posterior collapse. The addition of contextual information of interacting traffic participants using our convolutional social pooling network with 1×1 convolutions significantly improves performance, demonstrating the effectiveness of our conditional normalizing flow based priors.

7.5 CONCLUSION

In this chapter, we presented the first variational model for learning multi-modal conditional data distributions with Conditional Flow based priors – the Conditional Flow Variational Autoencoder (CF-VAE). Furthermore, we propose two novel regularization techniques – posterior regularization (pR) and condition regularization (cR) – which stabilizes training solutions and prevents posterior collapse leading to better fit to the target distribution. These techniques lead to a better match to the target distribution. Our experiments on diverse sequence prediction datasets show that our CF-VAE achieves state-of-the-art results across different performance metrics.

"BEST-OF-MANY-SAMPLES" DISTRIBUTION MATCHING 8

Contents

Generative Adversarial Networks (GANs) can achieve state-of-the-art sample quality in generative modelling tasks but suffer from the mode collapse problem. Variational Autoencoders (VAE) on the other hand explicitly maximize a reconstruction-based data log-likelihood forcing it to cover all modes, but suffer from poorer sample quality. Recent works have proposed hybrid VAE-GAN frameworks which integrate a GAN-based synthetic likelihood to the VAE objective to address both the mode collapse and sample quality issues, with limited success. This is because the VAE objective forces a trade-off between the data log-likelihood and divergence to the latent prior. The synthetic likelihood ratio term also shows instability during training. Based on our insights from Chapter 6, we propose a novel objective with a "Best-of-Many-Samples" reconstruction cost and a stable direct estimate of the synthetic likelihood. This enables our hybrid VAE-GAN framework to achieve high data log-likelihood and low divergence to the latent prior at the same time and shows significant improvement over both hybrid VAE-GANs and plain GANs in mode coverage and quality.

8.1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) have achieved state-of-the-art sample quality in generative modeling tasks. However, GANs do not explicitly estimate the data likelihood. Instead, it aims to "fool" an adversary, so that the adversary is unable to distinguish between samples from the true distribution and the generated samples. This leads to the generation of high quality samples

(Adler and Lunz, 2018; Brock *et al.*, 2019). However, there is no incentive to cover the whole data distribution. Entire modes of the true data distribution can be missed – commonly referred to as the mode collapse problem.

In contrast, the Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014) explicitly maximize data likelihood and can be forced to cover all modes (Bozkurt *et al.*, 2018; Shu *et al.*, 2018). VAEs enable sampling by constraining the latent space to a unit Gaussian and sampling through the latent space. However, VAEs maximize a data likelihood estimate based on the L_1/L_2 reconstruction cost which leads to lower overall sample quality – blurriness in case of image distributions. Therefore, there has been a spur of recent work (Donahue *et al.*, 2017; Larsen *et al.*, 2016; Rosca *et al.*, 2017) which aims to integrate GANs in a VAE framework to improve VAE generation quality while covering all the modes. Notably in Rosca *et al.* (2017), GANs are integrated in a VAE framework by augmenting the L_1/L_2 data likelihood term in the VAE objective with a GAN discriminator based synthetic likelihood ratio term.

However, Rosca *et al.* (2017) reports that in case of hybrid VAE-GANs, the latent space does not usually match the Gaussian prior. This is because, the reconstruction log-likelihood in the VAE objective is at odds with the divergence to the latent prior (Tabor *et al.*, 2018) (also in case of alternatives proposed by Makhzani *et al.* (2016); Arjovsky *et al.* (2017)). This problem is further exacerbated with the addition of the synthetic likelihood term in the hybrid VAE-GAN objective – it is necessary for sample quality but it introduces additional constraints on the encoder/decoder. This leads to the degradation in the quality and diversity of samples. Moreover, the synthetic likelihood ratio term is unstable during training – as it is the ratio of outputs of a classifier, any instability in the output of the classifier is magnified. We directly estimate the ratio using a network with a controlled Lipschitz constant, which leads to significantly improved stability. Our contributions in this chapter in detail are, 1. We propose a novel objective for training hybrid VAE-GAN frameworks, which relaxes the constraints on the encoder by giving the encoder multiple chances to draw samples with high likelihood enabling it to generate realistic images while covering all modes of the data distribution, 2. Our novel objective directly estimates the synthetic likelihood term with a controlled Lipschitz constant for stability, 3. Finally, we demonstrate significant improvement over prior hybrid VAE-GANs and plain GANs on highly muti-modal synthetic data, CIFAR-10 and CelebA.

8.2 NOVEL OBJECTIVE FOR HYBRID VAE-GANS

We begin with a brief overview of hybrid VAE-GANs followed by details of our novel objective.

Overview. Hybrid VAE-GANs (Fig. 8.1) are generative models for data distributions $\mathrm{x} \sim p(\mathrm{x})$ that transform a latent distribution $\mathrm{z} \sim p(\mathrm{z})$ to a learned distribution $\hat{\mathrm{x}} \sim p_\theta(\mathrm{x})$ approximating $p(\mathrm{x})$. The GAN (G_θ,D_I alone can generate realistic samples, but has trouble covering all modes. The VAE (R_ϕ,G_θ,D_L) can cover all modes of the distribution, but generates lower quality samples overall. VAE-GANs leverage the strengths of both VAEs and GANs to generate high quality samples while capturing all modes. We begin with a discussion of the prior hybrid VAE-GAN objectives (Rosca *et al.*, 2017) and its shortcomings, followed by our novel "Best-of-Many-Samples" objective with a novel reconstruction term and regularized stable direct estimate of the synthetic likelihood.

8.2.1 Shortcomings of Hybrid VAE-GAN Objectives

Hybrid VAE-GANs (Dumoulin *et al.*, 2017; Makhzani *et al.*, 2016; Rosca *et al.*, 2017; Zhao *et al.*, 2017c) maximizes the log-likelihood of the data ($\mathrm{x} \sim p(\mathrm{x})$) akin to VAEs. The log-likelihood, assuming the latent space to be distributed according to $p(\mathrm{z})$,

$$\log(p_\theta(\mathrm{x})) = \log\Big(\int p_\theta(\mathrm{x}|\mathrm{z})p(\mathrm{z})dz\Big). \tag{8.1}$$

Here, $p(\mathrm{z})$ is usually Gaussian. This requires the generator G_θ to generate samples that assign high likelihood to every example x in the data distribution for a likely $\mathrm{z} \sim p(\mathrm{z})$. Thus, the decoder θ can be forced to cover all modes of the data distribution $\mathrm{x} \sim p(\mathrm{x})$. In contrast, GANs never directly maximize the data likelihood and there is no direct incentive to cover all modes.

However, the integral in Eq. (8.1) is intractable. VAEs and Hybrid VAE-GANs use amortized variational inference using a recognition network $q_\phi(\mathrm{z}|\mathrm{x})$ (R_ϕ). The final hybrid VAE-GAN objective of the state-of-the-art α-GAN (Rosca *et al.*, 2017) which integrates a synthetic likelihood ratio term is,

$$\begin{aligned}\mathcal{L}_{\alpha\text{-GAN}} = \lambda\ \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \log(p_\theta(\mathrm{x}|\mathrm{z})) +\ \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \log\Big(\frac{D_\mathrm{I}(\mathrm{x}|\mathrm{z})}{1 - D_\mathrm{I}(\mathrm{x}|\mathrm{z})}\Big) \\ - D_\mathrm{KL}(p(\mathrm{z}) \parallel q_\phi(\mathrm{z}|\mathrm{x})).\end{aligned} \tag{8.2}$$

This objective has two important shortcomings. Firstly, as pointed in Bhattacharyya *et al.* (2018c); Tolstikhin *et al.* (2018) (see also Chapter 6), this objective severely constrains the recognition network as the average likelihood of the samples generated from the posterior $q_\phi(\mathrm{z}|\mathrm{x})$ is maximized. This forces all samples from $q_\phi(\mathrm{z}|\mathrm{x})$ to explain x equally well, penalizing any variance in $q_\phi(\mathrm{z}|\mathrm{x})$ and thus forcing it away from the Gaussian prior $p(\mathrm{z})$. Therefore, this makes it difficult to match the prior in the latent space and the encoder is forced to trade-off between a good estimate of the data log-likelihood and the divergence to the latent prior.

Secondly, the synthetic likelihood ratio term is the ratio of the output of D_I, any instability (non-smoothness) in the output of the classifier is magnified. Moreover, there is no incentive for D_I to be smooth (stable). For two similar images, $\{\mathrm{x}_1, \mathrm{x}_2\}$ with $|\mathrm{x}_1 - \mathrm{x}_2| \leq \epsilon$, the change of output $|D_\mathrm{I}(\mathrm{x}_1|\mathrm{z}_1) - D_\mathrm{I}(\mathrm{x}_1|\mathrm{z}_2)|$ can be arbitrarily large. This means that a small change in the generator output (e.g. after a gradient descent step) can have a large change in the discriminator output.

Next, we describe how we can effectively leverage multiple samples from $q_\phi(\mathrm{z}|\mathrm{x})$ to deal with the first issue. Finally, we derive a stable synthetic likelihood term (Rosca *et al.*, 2017; Wood, 2010) to deal with the second issue.

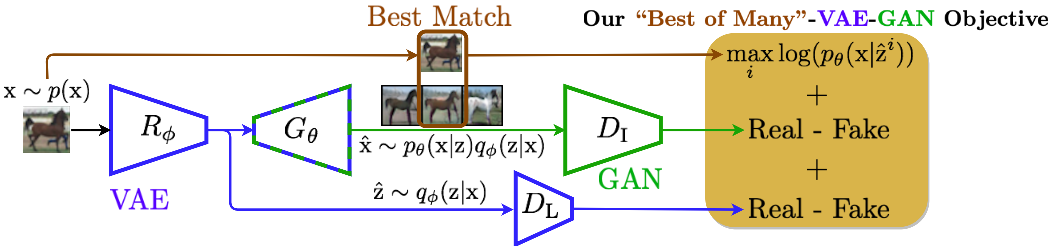

Best Match
Our "Best of Many"-VAE-GAN Objective
$\max_i \log(p_\theta(\mathrm{x}|\hat{\mathrm{z}}^i))$
+
Real - Fake
+
Real - Fake
$\mathrm{x} \sim p(\mathrm{x})$
R_ϕ
VAE
G_θ
$\hat{\mathrm{x}} \sim p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})$
$\hat{\mathrm{z}} \sim q_\phi(\mathrm{z}|\mathrm{x})$
D_I
GAN
D_L

Figure 8.1: Overview of our BMS-VAE-GAN framework. The terms of our novel objective Eq. (8.7) are highlighted at the right. We consider only the best sample from the generator G_θ while computing the reconstruction loss.

8.2.2 Leveraging Multiple Samples

Building upon Bhattacharyya *et al.* (2018c) (discussed in Chapter 6), we derive an alternative variational approximation of Eq. (8.1), which uses multiple samples to relax the constraints on the recognition network,

$$\mathcal{L}_{\mathrm{MS}} = \log\Big(\int p_\theta(\mathrm{x}|\mathrm{z})q_\phi(\mathrm{z}|\mathrm{x})\,dz\Big) - D_{\mathrm{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \tag{8.3}$$

In comparison to the α-GAN objective in Eq. (8.2) where the expected likelihood assigned by each sample to the data point x was considered, we see that in Eq. (8.3) the likelihood is computed considering all generated samples. The recognition network gets multiple chances to draw samples which assign high likelihood to x. This allows $q_\phi(\mathrm{z}|\mathrm{x})$ to have higher variance, helping it better match the prior and significantly reducing the trade-off with the data log-likelihood. Next, we describe how we can integrate a synthetic likelihood term in Eq. (8.3) to help us generate sharper images.

8.2.3 Integrating Stable Synthetic Likelihoods

Considering only L_1/L_2 reconstruction based likelihoods $p_\theta(\mathrm{x}|\mathrm{z})$ (as in Bhattacharyya *et al.* (2018c); Kingma and Welling (2014); Tolstikhin *et al.* (2018) and Chapter 6) might not be sufficient in case of complex high dimensional distributions, e.g. in case of image data this leads to blurry samples. Synthetic estimates of the likelihood Wood (2010) leverages a neural network (usually a classifier) which is jointly trained to distinguish between real and generated samples. The network is trained to assign low likelihood to generated samples and higher likelihood to real data samples. Starting from Eq. (8.3), we integrate a synthetic likelihood term with weight β to encourage our generator to generate realistic samples. The L_1/L_2 reconstruction likelihood (with weight α) forces the coverage of all modes. However, unlike prior work (Bhattacharyya *et al.*, 2019a; Rosca *et al.*, 2017) (as in Chapter 5), our synthetic likelihood estimator D_I is not a classifier. We first convert the likelihood term to a likelihood ratio form which allows for synthetic estimates,

$$\begin{aligned}\mathcal{L}_{\text{MS}} =&\alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - D_{\text{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})) \\ &\propto \alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{x}|\mathrm{z})}{p(\mathrm{x})}\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - D_{\text{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})).\end{aligned} \tag{8.4}$$

To enable the estimation of the likelihood ratio $p_\theta(\mathrm{x}|\mathrm{z})/p(\mathrm{x})$ using a neural network, we introduce the auxiliary variable y where, $\mathrm{y} = 1$ denotes that the sample was generated and $\mathrm{y} = 0$ denotes that the sample is from the true distribution. We can now express Eq. (8.4) (using Bayes theorem),

$$\begin{aligned}&=\alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{x}|\mathrm{z},\mathrm{y}=1)}{p(\mathrm{x}|\mathrm{y}=0)}\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - D_{\text{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \\ &=\alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{1-p(\mathrm{y}=1|\mathrm{x})}\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - D_{\text{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})).\end{aligned} \tag{8.5}$$

The ratio $p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})/1-p(\mathrm{y}=1|\mathrm{x})$ should be high for generated samples which are indistinguishable from real samples and low otherwise. In case of image distributions, we find that direct estimation of the numerator/denominator (as in Rosca *et al.* (2017)) exacerbates instabilities (non-smoothness) of the estimate. Therefore, we estimate this ratio directly using the neural network $D_{\text{I}}(\mathrm{x})$ – trained to produce high values for images indistinguishable from real images and low otherwise,

$$\begin{aligned}\mathcal{L}_{\text{MS-S}} \propto \alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} D_{\text{I}}(\mathrm{x}|\mathrm{z})\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) \\ -D_{\text{KL}}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})).\end{aligned} \tag{8.6}$$

To further ensure smoothness, we directly control the Lipschitz constant K of D_{I}. This ensures, $\forall \mathrm{x}_1, \mathrm{x}_2$, $|D_{\text{I}}(\mathrm{x}_1|\mathrm{z}_1) - D_{\text{I}}(\mathrm{x}_2|\mathrm{z}_2)| \leq K|\mathrm{x}_1 - \mathrm{x}_2|$ – the function is strictly smooth everywhere. Small changes in generator output cannot arbitrarily change the synthetic likelihood estimate, hence allowing the generator to smoothly improve sample quality. We constrain the Lipschitz constant K to 1 using Spectral Normalization (Miyato *et al.*, 2018). Note that the likelihood $p_\theta(\mathrm{x}|\mathrm{z})$ takes the form $e^{-\lambda\|\mathrm{x}-\hat{\mathrm{x}}\|_n}$ in Eq. (8.6) – a log-sum-exp which is numerically unstable. As we perform stochastic gradient descent, we can deal with this after stochastic (MC) sampling of the data points. We can estimate the log-sum-exp well using the max – the "Best-of-Many-Samples" (Nielsen and Sun, 2016),

$$\log\left(\frac{1}{T}\sum_{i=1}^{i=T} p_\theta(\mathrm{x}|\hat{\mathrm{z}}^i)\right) \geq \max_i \log(p_\theta(\mathrm{x}|\hat{\mathrm{z}}^i)) - \log(T)$$

Furthermore, in practice, we observe that we can improve sharpness of generated images by penalizing generator G_θ, using the least realistic of the T samples,

$$\log\left(\sum_{i=1}^{i=T} D_{\text{I}}(\mathrm{x}|\hat{\mathrm{z}}^i)\right) \geq \min_i \log\left(D_{\text{I}}(\mathrm{x}|\hat{\mathrm{z}}^i)\right)$$

Our final "Best-of-Many"-VAE-GAN objective takes the form (ignoring the constant $\log(T)$ term),

$$\mathcal{L}_{\text{BMS-S}} = \alpha \min_i \log\left(D_{\text{I}}(\text{x}|\hat{\text{z}}^i)\right) + \beta \max_i \log(p_\theta(\text{x}|\hat{\text{z}}^i)) - D_{\text{KL}}(q_\phi(\text{z}|\text{x}) \parallel p(\text{z})). \quad (8.7)$$

We use the same optimization scheme as in Rosca *et al.* (2017).

Approximation Errors. The "Best-of-Many-Samples" scheme introduces the $\log(T)$ error term. However, this error term is dominated by the low data likelihood term in the beginning of optimization (see Bhattacharyya *et al.* (2018c) and Chapter 6). Later, as generated samples become more diverse, the log likelihood term is dominated by the Best of T samples – "Best of Many-Samples" is equivalent.

Classifier based estimate of the prior term. Recent work (Makhzani *et al.*, 2016; Arjovsky *et al.*, 2017; Rosca *et al.*, 2017) has shown that point-wise minimization of the KL-divergence using its analytical form leads to degradation in image quality. Instead, KL-divergence term is recast in a synthetic likelihood ratio form minimized "globally" using a classifier instead of point-wise. Therefore, unlike Bhattacharyya *et al.* (2018c), here we employ a classifier based estimate of the KL-divergence to the prior. However, as pointed out by prior work on hybrid VAE-GANs (Rosca *et al.*, 2017), a classifier based estimate with still leads to mismatch to the prior as the trade-off with the data log-likelihood still persists without the use of the "Best-of-Many-Samples". Therefore, as we shall demonstrate next, the benefits of using the "Best-of-Many-Samples" extends to cases when a classifier based estimate of the KL-divergence is employed.

8.3 EXPERIMENTS

Next, we evaluate on multi-modal synthetic data as well as CIFAR-10 and CelebA. We perform all experiments on a single Nvidia V100 GPU with 16GB memory. We use as many samples during training as would fit in GPU memory so that we make the same number of forward/backward passes as other approaches and minimize the computational overhead of sampling multiple samples.

8.3.1 Evaluation on Multi-modal Synthetic data.

We evaluate in Tables 8.1 and 8.2 on the standard 2D Grid and Ring datasets, which are highly challenging due to their multi-modality. The metrics considered are the number of modes captured and % of high quality samples (within 3 standard deviations of a mode). The generator/discriminator architecture is the same as in Srivastava *et al.* (2017). We see that our BMS-VAE-GAN (using the best of $T = 10$ samples) outperforms state of the art GANs, e.g. Eghbal-zadeh *et al.* (2019) and the WAE / α-GAN baselines. The explicit maximization of the data log-likelihood enables our BMS-VAE-GAN and the WAE and α-GAN baselines to capture all modes in both the grid and ring datasets. The significantly increased proportion of high

	2D Grid (25 modes)		2D Ring (8 modes)	
Method	Modes	HQ%	Modes	HQ%
VEEGAN (Srivastava *et al.*, 2017)	24.6	40	8	52.9
GDPP-GAN (Elfeki *et al.*, 2019)	24.8	68.5	8	71.7
SN-GAN (Miyato *et al.*, 2018)	23.8±1.5	90.9±4.0	6.8±1.1	86.6±9.7
MD-GAN (Eghbal-zadeh *et al.*, 2019)	25	99.3±2.2	8	89.0±3.6
WAE (Arjovsky *et al.*, 2017)	25	65.4±3.8	8	35.8±4.7
α-GAN (Rosca *et al.*, 2017)	25	70.5±4.2	8	83.6±5.3
BMS-VAE-GAN (Ours) $T = 10$	25	**99.7±0.2**	8	**99.6±0.3**

Table 8.1: Evaluation on multi-modal synthetic data.

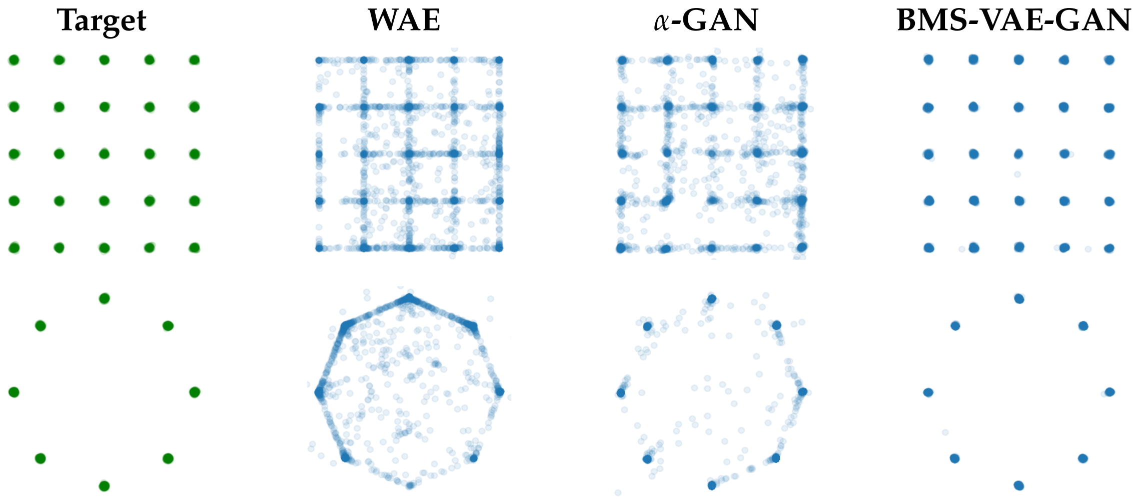


Table 8.2: Visualization of samples.

quality samples with respect to WAE and α-GAN baselines is due to our novel "Best-of-Many-Samples" objective. We illustrate this in Table 8.3. Following Rosca *et al.* (2017), we analyze the learnt latent spaces in detail, in particular we check for points (in red) which are likely under the Gaussian prior $p(z)$ (blue) but have low probability under the marginal posterior $q_\phi(z) = \int q_\phi(z|x)dx$. We use tSNE to project points from our 32-dimensional latent space to 2D. In Table 8.3 (Top Row) we clearly see that there are many such points in case of the WAE and α-GAN baselines (note that this low probability threshold is common across all methods). In Table 8.3 (Bottom Row) we see that these points lead to the generation of low quality samples (in red) in the data space. Therefore, we see that our "Best-of-Many-Samples" samples objective helps us match the prior in the latent space and thus this leads to the generation of high quality samples and outperforming both state of the art GANs and hybrid VAE-GAN baselines.

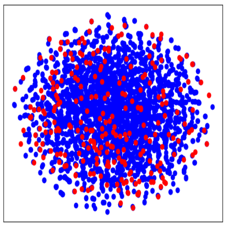
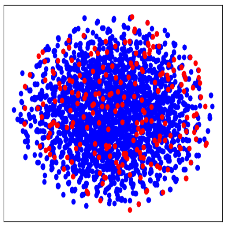
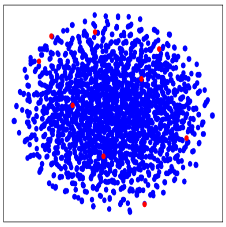
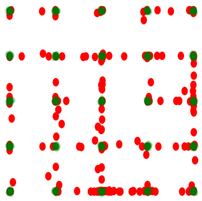
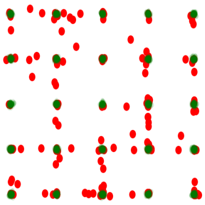
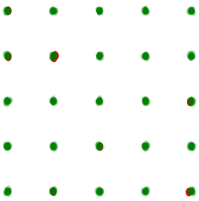
WAE	**α-GAN**	**BMS-VAE-GAN**
(Gaussian) Latent space samples z, $q_\phi(z) \ll p(z)$		
		
Corresponding data space samples, $p_\theta(x\|z)$		
		

Table 8.3: Effect of our novel objective in the latent space. **Top Row:** The standard WAE and α-GAN objectives lead to mismatch to the prior in the latent space. We show samples z (in red) which are highly likely under the standard Gaussian prior (blue) but have low probability under the learnt marginal posterior $q_\phi(z)$. **Bottom Row:** We show that such points z lead to low quality data samples (in red), which do correspond to any of the modes.

8.3.2 Evaluation on CIFAR-10

Next, we evaluate on the CIFAR-10 dataset. Auto-encoding based approaches (Kingma and Welling, 2014; Makhzani *et al.*, 2016) do not perform well on this dataset, as a simple Gaussian reconstruction based likelihood is insufficient for such highly multi-modal image data. This makes CIFAR-10 a challenging dataset for hybrid VAE-GANs.

Architecture details. We use two different types of architectures for the generator/discriminator pair G_θ, D_I : DCGAN based (Radford *et al.*, 2016) as used in Rosca *et al.* (2017) and the Standard CNN used in Miyato *et al.* (2018); Tran *et al.* (2018).

Experimental details and baselines. We use the ADAM optimizer (Kingma and Ba, 2015) and use a learning rate of 2×10^{-4}, $\beta_1 = 0.0$ and $\beta_2 = 0.9$ for all components. We use the same architecture of the latent space discriminator D_L as in α-GAN (Rosca *et al.*, 2017) (3-layer MLP with 750 neurons each). Values of $\log(D_\text{I}) \in [0, 2]$ work well.

We consider the following baselines for comparison against our BMS-VAE-GAN with a DCGAN generator/discriminator, 1. A standard DCGAN (Radford *et al.*, 2016), 2. The α-GAN model of (Rosca *et al.*, 2017). Furthermore, we compare our

Method	IvOM ↓
DCGAN (Radford *et al.*, 2016)	0.0084±0.0020
VEEGAN (Srivastava *et al.*, 2017)	0.0068±0.0001
SN-GAN (Miyato *et al.*, 2018)	0.0055±0.0006
α-GAN + SN (Ours) $T = 1$	0.0048±0.0005
BMS-VAE-GAN (Ours) $T = 30$	**0.0037±0.0005**

Table 8.4: IvOM on CIFAR-10.

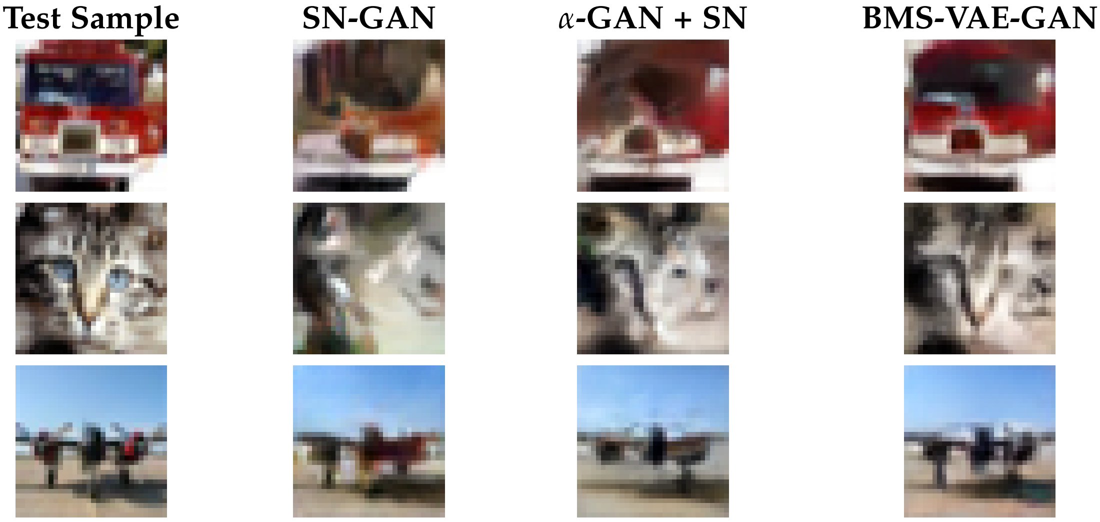


Figure 8.2: Closest generated images found using IvOM.

BMS-GAN with the Standard CNN generator/discriminator to, 1. SN-GAN (Miyato *et al.*, 2018), 2. BW-GAN (Adler and Lunz, 2018), 3. Dist-GAN (Tran *et al.*, 2018), 4. Our α-GAN + SN is an improved version of the α-GAN which includes Spectral Normalization for stable estimation of synthetic likelihoods. Again, the α-GAN and α-GAN + SN baselines are identical to the corresponding BMS-VAE-GAN except for the "Best-of-Many-Samples" reconstruction likelihood.

Discussion of results. We report results in Table 8.5. Please note that the higher latent space dimensionality (100) makes the latent spaces much harder to reliably analyze. Therefore, we rely on the FID and IoVM metrics. We follow the evaluation protocol of Miyato *et al.* (2018); Tran *et al.* (2018) and use 10k/5k real/generated samples to compute the FID score. The α-GAN (Rosca *et al.*, 2017) model with (DCGAN architecture) demonstrates better fit to the true data distribution (29.3 vs 30.7 FID) compared to a plain DCGAN. This again shows the ability of hybrid VAE-GANs in improving the performance of plain GANs. We observe that our novel "Best-of-Many-Samples" optimization scheme outperforms both the plain DCGAN and hybrid α-GAN(28.8 vs 29.4 FID), confirming the advantage of using "Best-of-Many-Samples". Furthermore, we see that our BMS-VAE outperforms the

Method	FID ↓
DCGAN Architecture	
WAE (Tolstikhin *et al.*, 2018)	89.3±0.3
BMS-VAE (Ours) $T = 10$	87.9±0.4
DCGAN (Radford *et al.*, 2016)	30.7±0.2
α-GAN (Rosca *et al.*, 2017)	29.4±0.3
BMS-GAN (ours) $T = 10$	**28.8**±**0.4**
Standard CNN Architecture	
SN-GAN (Miyato *et al.*, 2018)	25.5
BW-GAN (Adler and Lunz, 2018)	25.1
α-GAN + SN (Ours) $T = 1$	24.6±0.3
BMS-VAE-GAN (Ours) $T = 10$	23.8±0.2
BMS-VAE-GAN (Ours) $T = 30$	**23.4**±**0.2**
Dist-GAN (Tran *et al.*, 2018)	22.9
BMS-VAE-GAN (Ours) $T = 10$	**21.8**±**0.2**

Table 8.5: FID on CIFAR-10.

state-of-the-art plain auto-encoding WAE (Tolstikhin *et al.*, 2018).

We further compare our BMS-VAE-GAN to state-of-the-art GANs using the Standard CNN architecture in Table 8.5 with 100k generator iterations. Our α-GAN + SN ablation significantly outperforms the state-of-the-art plain GANs (Adler and Lunz, 2018; Miyato *et al.*, 2018), showing the effectiveness of hybrid VAE-GANs with a stable direct estimate of the synthetic likelihood on this highly diverse dataset. Furthermore, our BMS-VAE-GAN model trained using the best of $T = 30$ samples significantly improves over the α-GAN + SN baseline (23.4 vs 24.6 FID), showing the effectiveness of our "Best-of-Many-Samples". We also compare to Tran *et al.* (2018) using 300k generator iterations, again outperforming by a significant margin (21.8 vs 22.9 FID). The IoVM metric of Srivastava *et al.* (2017) (Table 8.4 and Fig. 8.2), illustrates that we are also able to better reconstruct the image distribution. The improvement in both sample quality as measured by the FID metric and data reconstruction as measured by the IoVM metric shows that our novel "Best-of-Many-Samples" objective helps us both match the prior in the latent space and achieve high data log-likelihood at the same time.

8.3.3 Evaluation on CelebA

Next, we evaluate on CelebA at resolutions 64×64 and 128×128.

Training and architecture details. As the focus is to evaluate training objectives for hyrid VAE-GANs, we use simple DCGAN based generators and discriminators for generation at both 64×64 and 128×128. Approaches like progressive growing

(Karras *et al.*, 2018) are orthogonal and can be applied on top.

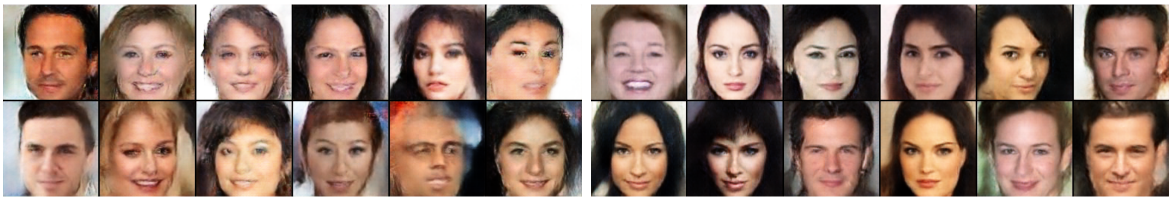
Baselines and experimental details. We consider the following baselines for comparison with our BMS-GAN with $T = \{10, 30\}$ samples, 1. WAE (Tolstikhin *et al.*, 2018) the state-of-the-art plain auto-encoding generative model, 2. α-GAN (Rosca *et al.*, 2017) the state-of-the-art hybrid VAE-GAN, 3. Our α-GAN + SN is an improved version of the α-GAN which includes Spectral Normalization for stable estimation of synthetic likelihoods. Note, the α-GAN baseline is identical to our BMS-GAN except for the "Best-of-Many" reconstruction likelihood. Moreover, we include several plain GAN baselines, 1. Wasserstein GAN with gradient penalty (WGAN-GP) Gulrajani *et al.* (2017a), 2. Spectral Normalization GAN (SN-GAN) Miyato *et al.* (2018), 3. Dist-GAN (Tran *et al.*, 2018).

To train our BMS-VAE-GAN and αR-GAN models we use the two time-scale update rule (Heusel *et al.*, 2017) with learning rate of 1×10^{-4} for the generator and 4×10^{-4} for the discriminator. We use the Adam optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.9$. We use a three layer MLP with 750 neurons as the latent space discriminator D_{L} (as in Rosca *et al.* (2017)) and a DCGAN based recognition network R_ϕ. We use the hinge loss to train D_{I} to produce high values for real images and low values for generated images, constraining $\log(D_{\text{I}}) \in [-0.5, 0.5]$ works well.

Method	FID ↓
Resolution: 64×64	
WAE (Tolstikhin *et al.*, 2018)	41.2±0.3
BMS-VAE (Ours) $T = 10$	39.8±0.3
DCGAN	31.1±0.9
WGAN-GP (Gulrajani *et al.*, 2017a)	26.8±1.2
BEGAN (Berthelot *et al.*, 2017)	26.3±0.9
Dist-GAN (Tran *et al.*, 2018)	23.7±0.3
SN-GAN (Miyato *et al.*, 2018)	21.9±0.8
α-GAN (Rosca *et al.*, 2017)	19.2±0.8
α-GAN + SN (Ours) $T = 1$	15.1±0.2
BMS-VAE-GAN (Ours) $T = 10$	14.3±0.4
BMS-VAE-GAN (Ours) $T = 30$	**13.6**±**0.4**
Resolution: 128×128	
SN-GAN (Miyato *et al.*, 2018)	60.5±1.5
αR-GAN (Ours) $T = 1$	45.8±1.4
BMS-GAN (Ours) $T = 10$	**42.7**±**1.2**

Table 8.6: FID on CelebA.

Discussion of results. We train all models for 200k iterations and report the FID scores (Heusel *et al.*, 2017) for all models using 10k/10k real/generated samples in Table 8.6. The pure auto-encoding based WAE (Tolstikhin *et al.*, 2018)



(a) Our α-GAN + SN ($T = 1$, 128×128)

(b) Our BMS-VAE-GAN ($T = 10$, 128×128)

Figure 8.3: CelebA Random Samples. Our "Best of Many" reconstruction cost leads to sharper results.

has the weakest performance due to blurriness. Our pure auto-encoding BMS-VAE (without synthetic likelihoods) improves upon the WAE (39.8 vs 41.2 FID), already demonstrating the effectiveness of using "Best-of-Many-Samples". We see that the base DCGAN has the weakest performance among the GANs. BEGAN suffers from partial mode collapse. The SN-GAN improves upon WGAN-GP, showing the effectiveness of Spectral Normalization. However, there exists considerable artifacts in its generations. The α-GAN of Rosca *et al.* (2017), which integrates the base DCGAN in its framework performs significantly better (31.1 vs 19.2 FID). This shows the effectiveness of VAE-GAN frameworks in increasing the quality and diversity of generations. Our enhanced α-GAN + SN regularized with Spectral Normalization performs significantly better (15.1 vs 19.2 FID). This shows the effectiveness of a regularized direct estimate of the synthetic likelihood. Using the gradient penalty regularizer of Gulrajani *et al.* (2017a) lead to a drop of 0.4 FID. Our BMS-VAE-GAN improves significantly over the α-GAN + SN baseline using the "Best-of-Many-Samples" (13.6 vs 15.1 FID). The results at 128×128 resolution mirror the results at 64×64. We see that by using the "Best-of-Many-Samples" we obtain sharper (Fig. 8.3(b)) results that cover more of the data distribution as shown by both the FID and IoVM.

8.4 CONCLUSION

We propose a new objective for training hybrid VAE-GAN frameworks which overcomes key limitations of current hybrid VAE-GANs. We integrate, 1. A "Best-of–Many-Samples" reconstruction likelihood which helps in covering all the modes of the data distribution while maintaining a latent space as close to Gaussian as possible, 2. A stable estimate of the synthetic likelihood ratio.. Our hybrid VAE-GAN framework outperforms state-of-the-art hybrid VAE-GANs and plain GANs in generative modelling on CelebA and CIFAR-10, demonstrating the effectiveness of our approach.

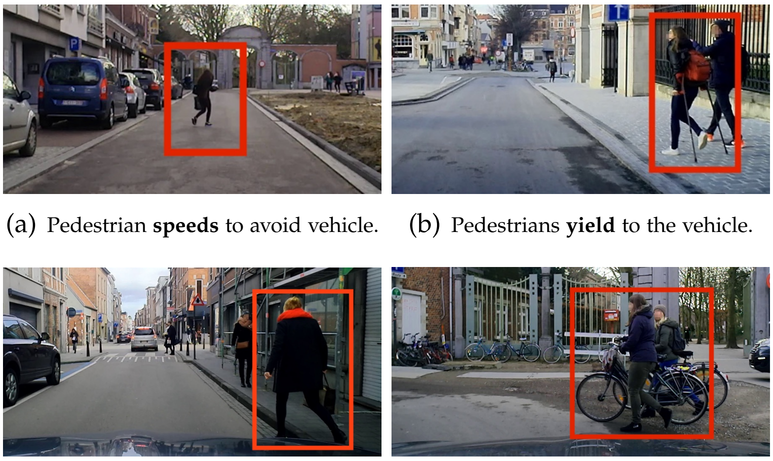
9 EURO-PVI: PEDESTRIAN VEHICLE INTERACTIONS IN DENSE URBAN CENTERS

Contents

Interactions between vehicle and pedestrian or bicyclist have a significant impact on the trajectories of traffic participants, e.g. stopping or turning to avoid collisions. Although recent datasets and trajectory prediction approaches have fostered the development of autonomous vehicles yet the amount of vehicle-pedestrian (bicyclist) interactions modeled are sparse. In this chapter, we propose Euro-PVI, a dataset of pedestrian and bicyclist trajectories. In particular, our dataset caters more diverse and complex interactions in dense urban scenarios compared to the existing datasets. To address the challenges in predicting future trajectories with dense interactions, we develop a joint inference model that learns an expressive multi-modal shared latent space across agents in the urban scene. This enables our Joint-β-cVAE approach to better model the distribution of future trajectories. We achieve state of the art results on the nuScenes and Euro-PVI datasets demonstrating the importance of capturing interactions between ego-vehicle and pedestrians (bicyclists) for accurate predictions.

9.1 INTRODUCTION

Notwithstanding recent progress in the development of reliable self-driving vehicles, dense inner city environments remain challenging. One of the key components for the success of self-driving vehicles in dense urban environments is anticipation (Bhattacharyya *et al.*, 2018c; Lee *et al.*, 2017b). Anticipating the motion of traffic participants in dense urban environments is made especially challenging due to the effect of interactions between different agents. For example, a pedestrian might turn onto the road to avoid an obstacle on the sidewalk which requires the vehicle to stop (Fig. 9.1(c)). Alternately, a pedestrian attempting to cross the road ahead of the ego-vehicle might continue or stop depending upon the distance and velocity of the



(a) Pedestrian **speeds** to avoid vehicle. (b) Pedestrians **yield** to the vehicle.

(c) Vehicle **slows** to avoid pedestrian. (d) Vehicle **yields** to the bicyclists.

Figure 9.1: Examples of interactions between the ego-vehicle and pedestrians (bicyclists) in Euro-PVI.

vehicle (cf. Fig. 9.2). Thus, interactions add significant complexity to the distribution of the likely future trajectories which is highly multi-modal.

Recently, datasets like nuScenes (Caesar *et al.*, 2020), Argoverse (Chang *et al.*, 2019), or Lyft L5 (Houston *et al.*, 2020) have greatly aided the development of trajectory prediction methods. However, these datasets are primarily focused on trajectories of vehicles and vehicle-vehicle interactions – collected to a large extent on multi-lane roads in North America or Asia, with sparse interactions between the ego-vehicle and pedestrians or bicyclists (e.g. Figure 4 in Houston *et al.* (2020)). Therefore, they do not represent trajectories in dense urban environments well where interactions between the trajectories of agents are prominent. Such scenarios are particularly common in inner-city environments in Europe.

In this chapter, we propose the new European Pedestrian Vehicle Interaction (*Euro-PVI*) dataset [§] which is collected in a dense urban environment in Brussels and Leuven, Belgium. The Euro-PVI dataset contains a rich and diverse set of interactions between the ego-vehicle and pedestrians (bicyclists). Sequences are recorded near busy urban landmarks, e.g. railway stations, narrow lanes or intersections (cf. Figs. 9.1 to 9.3) where interactions are frequent and it is therefore challenging to predict pedestrian (bicyclist) paths.

Further, in spite of the recent progress in trajectory prediction methods, accurately capturing the multi-modal distribution of future trajectories, e.g. in dense urban environments, remains challenging. Current state of the art (Bhattacharyya *et al.*, 2019c; Mangalam *et al.*, 2020; Salzmann *et al.*, 2020) generative models for trajectory prediction and the approaches presented in Chapters 6 and 7 encode interactions directly in the posterior. Thus, the latent space does not express interaction information from the input distribution which limits the accuracy of the generated future

[§] The dataset is available at www.europvi.mpi-inf.mpg.de trajectories. To address this limitation, we develop a latent variable deep generative model which jointly models the distribution of future trajectories of the agents in the scene. Our *Joint-β-Conditional Variational Autoencoder (Joint-β-cVAE)* models a "shared" latent space between agents, to better capture the effect of interactions in the latent space and accurately represent the multi-modal distribution of trajectories.

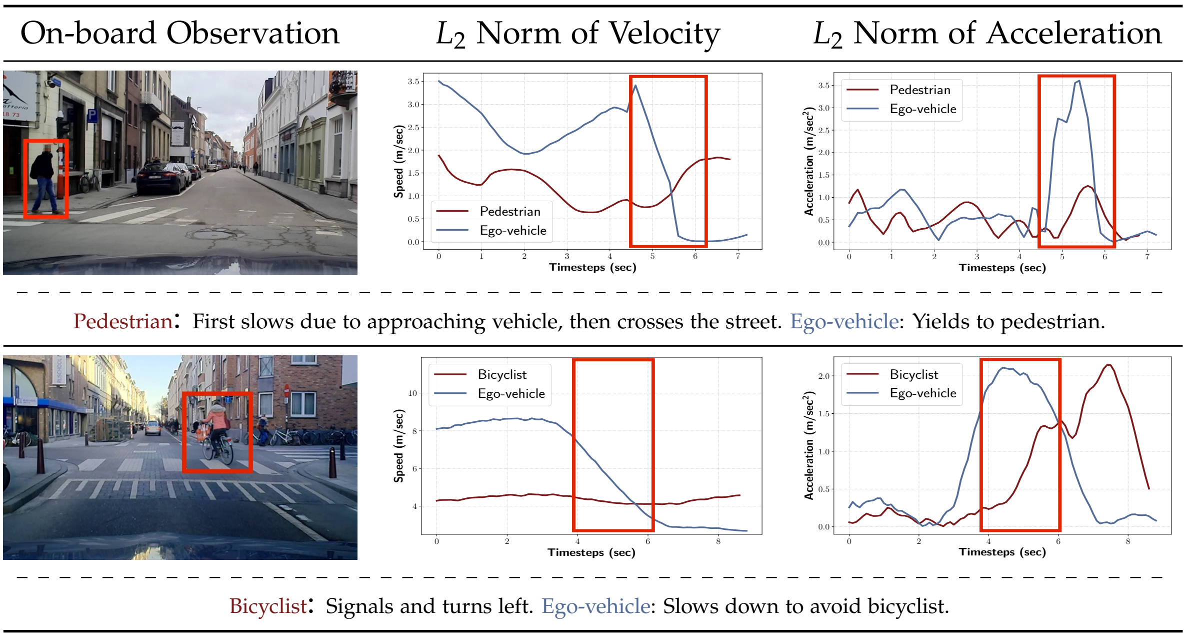
Our contributions in this chapter are, 1. We propose Euro-PVI, a novel dataset of pedestrian and bicyclist trajectories recorded in Europe with dense interactions with the ego-vehicle. 2. Our dataset facilitates research on dense interactions as we show that – in contrast to prior datasets – there is a pronounced performance gap between methods that model vehicle-pedestrian-interaction vs not. 3. To this end, we develop a latent generative model – Joint-β-cVAE – that models a shared latent space to better capture the effect of interactions on the multi-modal distribution of future trajectories. 4. Finally, we demonstrate state of the art performance on pedestrian (bicyclist) trajectory prediction on nuScenes and Euro-PVI.

9.2 THE EURO-PVI DATASET

In this section, we introduce our Euro-PVI dataset to advance the task of "on-board" trajectory prediction especially in dense urban environments. The dataset focuses on the role of the ego-vehicle - pedestrian (bicyclist) interactions present in a scene to predict the future trajectories in dense urban environments. We first concretely define an "interaction", which guides our data collection process and helps us select relevant sequences. Next, we provide details of the sensor setup and the data collection process. We then compare Euro-PVI to the existing datasets with respect to the density of interactions and provide detailed dataset statistics.

Interactions. We define an interaction between the ego-vehicle and a traffic participant (e.g. pedestrian or bicyclist) as an event where the presence of either (or both) the ego-vehicle or the traffic participant causes a change in velocity (change of speed/direction of motion i.e. acceleration) of the other. Examples of interactions include, 1. The ego-vehicle yielding to a pedestrian at a crosswalk (Fig. 9.2 top). 2. Non-verbal communication causes the ego-vehicle to slow down, as to avoid a bicycle which wants to turn (Fig. 9.2 bottom). We aim to record sequences which contain dense interactions between the ego-vehicle and vulnerable traffic participants, in particular, pedestrians and bicyclists.

Drive planning and scene selection. Euro-PVI is recorded in the dense urban scene of Brussels and Leuven, Belgium. The ego-vehicle is equipped with a 360° lidar, a positioning system and a set of front-facing synchronized cameras. The ego-vehicle is crewed by a pilot and a co-pilot. The pilot is instructed to drive freely over a predetermined area and to re-visit locations at times when dense concentration of pedestrians (bicyclists) is to be expected, such as transport hubs during peak hour. The duration of the driving sessions were up to 8 hours per day, over the course of two weeks. The co-pilot is tasked with identifying interactions and tagging the event. In the case of changes in trajectory or velocity of the ego-vehicle, the co-pilot



Pedestrian: First slows due to approaching vehicle, then crosses the street. Ego-vehicle: Yields to pedestrian.

Bicyclist: Signals and turns left. Ego-vehicle: Slows down to avoid bicyclist.

Figure 9.2: Examples of interactions in Euro-PVI. Spikes in the magnitude (L_2 norm) of acceleration resulting from interactions are marked.

asks the pilot for confirmation. The pilot is also instructed to spontaneously indicate that an interaction has happened. In Fig. 9.3 we show the geographical distributions of the trajectories in two example locations. We see that there is a high density of trajectories located around busy urban landmarks e.g. the railway station. We also show that interactions are not confined to road intersections with crosswalks where pedestrian (bicyclist) trajectories are simpler to predict, but also occur at locations without crosswalks (i.e. without designated crossing areas for pedestrians/bicyclists) where trajectories are more challenging to predict.

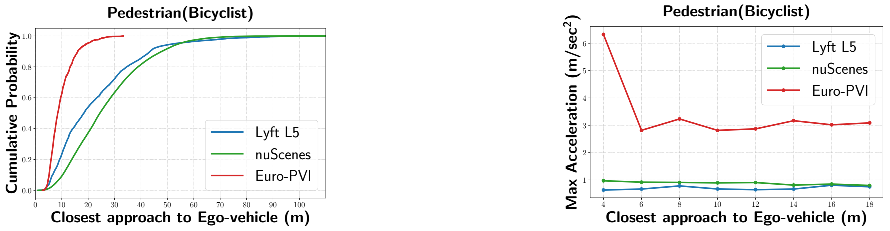
Interactions in urban environments. We now compare Euro-PVI to existing datasets for trajectory prediction with respect to the density of interactions, in particular to the two largest datasets – nuScenes (Caesar *et al.*, 2020) and Lyft L5 (Houston *et al.*, 2020). First, we compare the distances between the ego-vehicle and pedestrians (bicyclists) in the scene. Short distances are indicative of closely packed urban environments where interactions frequently occur. In Fig. 9.4 (left), we show the closest approach (proximity) of a pedestrian or bicyclist to the ego-vehicle. We see that in the nuScenes and Lyft L5 datasets, the majority of the pedestrians (bicyclists) do not approach the ego-vehicle closer than ~ 20 meters. Such large distances between the pedestrians (bicyclists) and the ego-vehicle, more than 4 typical car lengths, decreases the likelihood of interactions. In contrast, in Euro-PVI the majority of the pedestrians (bicyclists) approach the ego-vehicle as close as ~ 8 meters. Such short distances between the ego-vehicle and pedestrians (bicyclists) are indicative of the densely packed urban environment in which Euro-PVI is recorded which increases the likelihood of interactions.

Railway
Station

(a) Leuven railway station.

(b) Leuven city center.

Figure 9.3: Examples of aggregated spatial distribution of trajectories of pedestrians and cyclists around intersections and urban landmarks.



(a) Distance to the ego-vehicle.

(b) Maximum acceleration.

Figure 9.4: Left (a): Cumulative distribution sorted by distance to the ego-vehicle. Close proximity of the ego-vehicle and pedestrians(bicyclists) in Euro-PVI indicate dense traffic scenarios where interactions are likely. Right (b): Maximum acceleration sorted by distance to the ego-vehicle. High acceleration in close proximity of the ego-vehicle and pedestrians(bicyclists) indicate high likelihood of interactions.

However, close proximity only increases the likelihood of interactions, but does not necessarily lead to interactions. Note that by definition, interactions lead to a change in velocity i.e. acceleration. Therefore, in Fig. 9.4 (right) we plot the maximum acceleration experienced by the pedestrian (bicyclist) and ego-vehicle with increasing distance to the ego-vehicle and a pedestrian (bicyclist) respectively. In the case of nuScenes and Lyft L5 we do not see a strong dependence on distance. In contrast, in Euro-PVI both pedestrians (bicyclists) and the ego-vehicle experience the maximum acceleration close to the point of closest approach. This is again strongly indicative of dense interactions in Euro-PVI.

Qualitative examples of interactions. We provide example interactions in Euro-PVI along with the resultant acceleration of the involved agents in Fig. 9.2, e.g. top row: the pedestrian first slows down due to the approaching ego-vehicle and at the same time, the ego-vehicle sees the pedestrian and yields. This is visible as a spike in the velocity and acceleration plots. Similar spikes in acceleration can be observed due to interactions in the other examples in Fig. 9.2.

Additional statistics. We report dataset statistics and available labels of Euro-PVI

				Trajectory Instances		Labels			
Dataset	**Location**	**Scenes**	**Length (hrs)**	**Pedestrians**	**Bicyclists**	**Front Camera**	**Lidar**	**Seg. Maps**	**IMU**
nuScenes	North Am. & Asia	1000	5.5	9142	550	✓	✓	✓	✓
ApolloScapes	Asia	54	1.7	3065	1827 *	x	x	x	x
Lyft L5	North Am.	*170×10^3*	*1118*	*605×10^3*	*77×10^3*	x	x	x	✓
CityScapes	Europe	5000	2.5	0	0	✓	x	✓	✓
KITTI	Europe	71	1.5	380	150	✓	✓	✓	✓
KITTI-360	Europe	9	2.2	262	82	✓	✓	✓	✓
A2D2	Europe	23	0.9	260	248	✓	✓	✓	✓
Euro-PVI (Ours)	Europe	1077	2.2	6177	1581	✓	✓	✓	✓

Table 9.1: Comparison of dataset statistics. (Seg. Maps – semantic and/or instance segmentation, * includes motorbikes. nuScenes (Caesar *et al.*, 2020), ApolloScapes (Ma *et al.*, 2019), Lyft L5 (Houston *et al.*, 2020), CityScapes (Cordts *et al.*, 2016), KITTI (Geiger *et al.*, 2012), KITTI-360 (Xie *et al.*, 2016), KITTI-360 (Xie *et al.*, 2016), A2D2 (Geyer *et al.*, 2020))

in Table 9.1. In addition to the annotated pedestrian (bicyclist) trajectories, Euro-PVI contains 83k camera images and the corresponding lidar point clouds along with synchronized IMU data. In terms of size Euro-PVI is competitive with nuScenes (Caesar *et al.*, 2020) and ApolloSpaces (Ma *et al.*, 2019) and while Lyft L5 Houston *et al.* (2020) is significantly larger, it does not provide labels e.g. camera images or lidar point clouds. Furthermore, Euro-PVI surpasses the the largest autonomous driving datasets collected in Europe i.e. CityScapes (Cordts *et al.*, 2016), KITTI Geiger *et al.* (2012), KITTI-360 (Xie *et al.*, 2016) and A2D2 (Geyer *et al.*, 2020) in terms of number of instances of pedestrian (bicyclist) trajectories. CityScapes and A2D2 do not provide annotated 3D pedestrian or bicycle trajectories. KITTI (Geiger *et al.*, 2012) and KITTI-360 (Xie *et al.*, 2016) contains mostly linear motion with sparse interactions and thus commonly not used for trajectory prediction Ma *et al.* (2019); Rhinehart *et al.* (2018). In fact, Euro-PVI is the first large scale dataset with dense interactions (Fig. 9.4) dedicated to trajectory prediction in Europe, to the best of our knowledge.

9.3 JOINT-β-CVAE: JOINT MODEL FOR DENSE URBAN ENVIRONMENTS

Following the observations in Section 9.2, we find that vehicle-pedestrian (bicyclist) interactions are crucial to the task of future trajectory prediction in dense urban scenarios. In particular, inherent multi-modality of the distribution of future trajectories and the effect of interactions on this complex distribution, make accurately predicting the future trajectories in dense urban environments challenging.

Specifically, given a scene with n agents e.g. vehicles, pedestrians or bicyclists in a dense urban environment and the past observations $\mathbf{x}_i \in \mathbf{X}$ for each agent i, we model the future trajectories $\mathbf{y}_i \in \mathbf{Y}$ for each agent $i \in \{1, \ldots, n\}$ in the scene. Here the past observations $\mathbf{x}_i$ include the past trajectories and the past context corresponding to the past trajectory sequence. Prior work (Bhattacharyya *et al.*, 2019c, 2018c; Mangalam *et al.*, 2020; Salzmann *et al.*, 2020) and Chapters 6 and 7 models the conditional distribution $p_\theta(\mathbf{y}_i|\mathbf{X})$ parameterized by θ of the future trajectories $\mathbf{y}_i$ using the latent variables $\mathbf{z}_i$ in the standard conditional VAE formulation (Higgins *et al.*, 2017; Kingma and Welling, 2014; Sohn *et al.*, 2015),

$$p_\theta(\mathbf{y}_i|\mathbf{X}) = \int p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{X})\, p_\theta(\mathbf{z}_i|\mathbf{x}_i)\, d\mathbf{z}_i. \tag{9.1}$$

Here, the distribution $p_\theta(\mathbf{z}_i|\mathbf{x}_i)$ assumes conditional independence of the latent variables $\{\mathbf{z}_1, \dots, \mathbf{z}_n\} \in \mathbf{Z}$ given the past observation $\mathbf{x}_i$ of each agent. This assumption essentially ignores the motion patterns of interacting agents i.e. the ego-vehicle and other pedestrians (bicyclists) in the scene, which in real world dense urban scenarios is critical for the accurate prediction of future trajectories. The formulation in Eq. (9.1) therefore limits the amount of interactions between the agents that can be encoded in the latent space. Since the latent variables $\mathbf{z}_i$ are crucial for capturing diverse futures in such conditional models, it is important to express the effect of interactions in the latent space.

We now introduce our Joint-β-cVAE approach which aims to accurately capture the effect of interactions in the latent space for trajectory prediction in dense urban environments. Our proposed Joint-β-cVAE model, in contrast to prior conditional VAE based models (Bhattacharyya *et al.*, 2019c, 2018c; Mangalam *et al.*, 2020; Salzmann *et al.*, 2020) encodes the joint distribution of the latent variables across all agents in the scene. This allows our Joint-β-cVAE model to encode the dependence of the future trajectory distribution on interacting agents in the latent space, leading to more accurate modeling of the multi-modal future trajectory distribution.

Formulation. Our Joint-β-cVAE in Fig. 9.5 models the joint distribution of future trajectories $\mathbf{Y}$ across all n agents, using the latent variables $\mathbf{z}_i \in \mathbf{Z}$,

$$\begin{aligned} p_\theta(\mathbf{Y}|\mathbf{X}) &= \int p_\theta(\mathbf{Y}, \mathbf{Z}|\mathbf{X})\, d\mathbf{Z} \\ &= \int \prod_i^n p_\theta(\mathbf{y}_i, \mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})\, d\mathbf{Z} \\ &= \int \prod_i^n p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}) p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})\, d\mathbf{Z} \end{aligned} \tag{9.2}$$

where $p_\theta(\mathbf{Y}, \mathbf{Z}|\mathbf{X})$ is the joint distribution of the future trajectories and the latent variables for all agents. In the second step, without loss of generality, we autoregressively factorize the joint distribution over the n agents, where $\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}$ denotes the latent variables and trajectories for agents $\{1, \dots, i-1\}$. Note that, the factorization is agnostic to the choice of ordering of the agents. In contrast to Eq. (9.1), the prior distribution of the latent variables in Eq. (9.2) exhibits joint modeling of the latent variables $\mathbf{z}_i$, i.e. $p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})$.

We maximize the log-likelihood of the data under the model in Eq. (9.2) with variational inference using a joint variational posterior distribution $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$.

The joint posterior. To encode rich latent spaces shared between the n agents which capture the effect of interactions in dense urban environments, we propose a

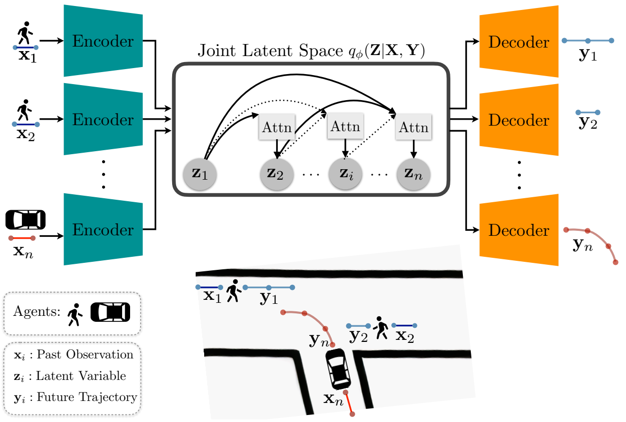


Figure 9.5: Our Joint-β-cVAE, which models a Joint latent space across all agents $\{1, \cdots, n\}$ in the scene. The posterior latent distribution $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$ factorizes auto-regressively and models the dependence of $\mathbf{z}_i$ on $\{\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y}\}$ using an attention mechanism.

joint posterior over all n agents in the scene, which auto-regressively factorizes,

$$q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{Y}) = q_\phi(\mathbf{z}_1|\mathbf{X}, \mathbf{Y}) \cdots q_\phi(\mathbf{z}_n|\mathbf{Z}_{<n}, \mathbf{X}, \mathbf{Y}). \tag{9.3}$$

The conditional distributions $q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})$ corresponding to agent i are normal distributions whose mean and variances are functions of $\mathbf{Z}_{<i}$, $\mathbf{X}$ and $\mathbf{Y}$. Intuitively, given the past trajectories of the agents $\{1, \ldots, n\}$ and the joint latent posterior distribution over the interacting agents $\{1, \ldots, i\}$, the latent posterior distribution corresponding to agent i encoding its interactions can be inferred conditioned on the latent information of the interacting agents $\{1, \ldots, i-1\}$. We use an attention mechanism inspired by Anderson *et al.* (2018), to model the dependence of the distribution of $\mathbf{z}_i$ on $\mathbf{Z}_{<i}$ and $\{\mathbf{X}, \mathbf{Y}\}$. The attention weights on $\mathbf{z}_j$ and $\{\mathbf{x}_j, \mathbf{y}_j\}$, for $j \neq i$, is additionally conditioned on the past observation and location of the agents (Fig. 9.5) – which allows to attend to agents j interacting with agent i.

In contrast, prior work (Bhattacharyya *et al.*, 2019c, 2018c; Mangalam *et al.*, 2020; Salzmann *et al.*, 2020) and Chapters 6 and 7 employ conditionally independent posteriors $q_\phi(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)$ across agents in a scene – encoding limited interactions in the latent space.

The joint prior. The prior term in Eq. (9.2), $p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})$, encodes the effect of interactions on the latent space of agent i through the dependence on $\mathbf{Z}_{<i}, \mathbf{X}$. In practice, we find that a simpler joint prior,

$$p_\theta(\mathbf{Z}|\mathbf{X}) = p_\theta(\mathbf{z}_1|\mathbf{X}) \cdots p_\theta(\mathbf{z}_n|\mathbf{Z}_{<n}, \mathbf{X}) \tag{9.4}$$

is sufficient for rich latent spaces that capture interactions. We parameterize the prior as a conditional normal distribution, where the mean and variance depends on $\{\mathbf{Z}_{<i}, \mathbf{X}\}$.

The ELBO. As the standard log evidence lower bound (ELBO) for cVAEs, proposed in (Kingma and Welling, 2014; Sohn *et al.*, 2015), experiences several issues, e.g.

	Interactions		**Best of** $N = 20 \downarrow$			**KDE NLL** $\downarrow$		
Method	**P-P**	**P-V**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Social-GAN	✓	–	0.04	0.11	0.21	-2.78	-1.40	-0.46
Social-GAN	✓	✓	0.04	0.11	0.21	-2.80	-1.41	-0.48
Sophie	✓	–	0.04	0.11	0.21	-2.59	-1.26	-0.41
Sophie	✓	✓	0.04	0.11	0.21	-2.63	-1.27	-0.42
Trajectron++	✓	–	**0.01**	0.08	0.15	-5.55	-3.87	-2.69
Trajectron++	✓	✓	**0.01**	0.08	0.15	-5.58	-3.96	-2.77
cVAE	–	–	0.05	0.12	0.23	-2.51	-1.20	-0.21
β-cVAE	–	–	**0.01**	0.08	0.17	-6.90	-4.10	-2.41
Joint-β-cVAE	✓	–	**0.01**	**0.06**	0.13	-7.50	-4.53	-2.95
Joint-β-cVAE	✓	✓	**0.01**	**0.06**	**0.13**	**-7.55**	**-4.59**	**-2.98**

Table 9.2: Evaluation on nuScenes. P-P and P-V: whether pedestrian - pedestrian or pedestrian - ego-vehicle interactions are modeled (Social-GAN Gupta *et al.* (2018), Sophie Sadeghian *et al.* (2019), Trajectron++ Salzmann *et al.* (2020), β-cVAE Higgins *et al.* (2017), Joint-β-cVAE (Ours)).

posterior collapse, we employ the ELBO formulation of β-VAE (Higgins *et al.*, 2017) to improve the representational capacity of the latent space and more accurately capture the effect of interactions in dense urban environments. With the formulation of the factorized variational distribution $q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})$ in Eq. (9.3) and the joint prior distribution in Eq. (9.4), the ELBO is given by,

$$\begin{aligned} \log(p_\theta(\mathbf{Y}|\mathbf{X})) \geq & \sum_i \mathbb{E}_{q_\phi} \log(p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i},\mathbf{Y}_{<i},\mathbf{X})) \\ & - \beta \sum_i D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i},\mathbf{X},\mathbf{Y})||p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i},\mathbf{X})\big). \end{aligned} \tag{9.5}$$

Additionally, we find it beneficial to model the observation noise σ^2 in the posterior distribution over the trajectories $p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i},\mathbf{Y}_{<i},\mathbf{X})$ as recommended in Lucas *et al.* (2019). During training, we alternately optimize both the posterior and the prior distributions such that the ELBO is maximized (Bhattacharyya *et al.*, 2019c; Tomczak and Welling, 2018) (see also Chapter 7).

We find in practice that is it sufficient to condition the decoder $p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i},\mathbf{Y}_{<i},\mathbf{X})$ only on $\{\mathbf{Z}_{\leq i},\mathbf{X}\}$. The rich latent space of our Joint-β-cVAE already models the effect of interactions, making it unnecessary to additionally condition on $\mathbf{Y}_{<i}$ for good performance.

9.4 EXPERIMENTS

In this section, we 1. Demonstrate the effectiveness of our Joint-β-cVAE method. 2. Provide additional experimental evidence to better highlight the differences in the density of interactions between the ego-vehicle and pedestrians(bicyclists) Euro-PVI

	Interactions		**Best of** $N=20\downarrow$			**KDE NLL** $\downarrow$		
Method	**P-P**	**P-V**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Social-GAN	✓	–	0.14	0.36	0.65	-0.38	0.74	1.55
Social-GAN	✓	✓	0.14	0.35	0.64	-0.41	0.73	1.52
Sophie	✓	–	0.11	0.30	0.58	-1.53	-0.22	0.53
Sophie	✓	✓	0.11	0.29	0.56	-1.71	-0.31	0.40
Trajectron++	✓	–	**0.09**	0.29	0.56	-2.75	-0.91	0.23
Trajectron++	✓	✓	**0.09**	0.28	0.54	-2.81	-1.00	0.15
cVAE	–	–	0.12	0.32	0.60	-1.40	-0.08	0.78
β-cVAE	–	–	0.10	0.30	0.56	-2.61	-0.78	0.31
Joint-β-cVAE	✓	–	**0.09**	0.29	0.53	-3.69	-1.29	0.02
Joint-β-cVAE	✓	✓	**0.09**	**0.27**	**0.51**	**-3.75**	**-1.38**	**-0.05**
Joint-β-cVAE + {Camera, Lidar}	✓	✓	*0.09*	*0.27*	*0.50*	*-3.78*	*-1.41*	*-0.13*

Table 9.3: Evaluation on Euro-PVI. P-P and P-V: whether pedestrian - pedestrian or pedestrian - ego-vehicle interactions are modeled (Social-GAN Gupta *et al.* (2018), Sophie Sadeghian *et al.* (2019), Trajectron++ Salzmann *et al.* (2020), β-cVAE Higgins *et al.* (2017), Joint-β-cVAE (Ours))).

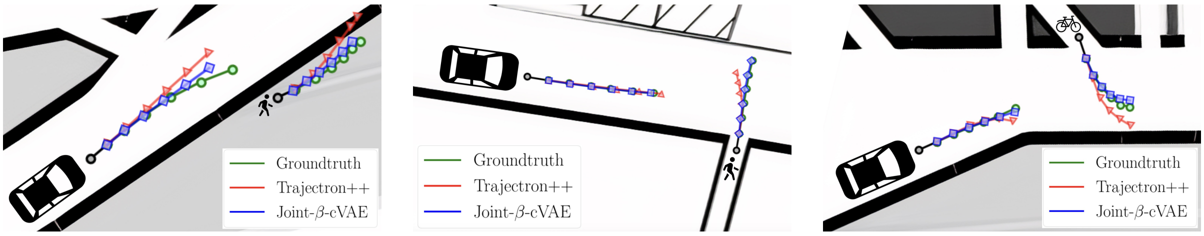


Figure 9.6: Qualitative examples on Euro-PVI. We compare the Best of $N = 20$ samples for Trajectron++ (red) and our Joint-β-cVAE (blue).

and current datasets. In order to address the above points, in addition to Euro-PVI, we evaluate on nuScenes (Caesar *et al.*, 2020). We choose nuScenes (Caesar *et al.*, 2020) as it is significantly larger compared to datasets like ApolloScapes (Ma *et al.*, 2019) and more diverse in comparison to Lyft L5 (Houston *et al.*, 2020), while possessing similar proximity/acceleration statistics (Fig. 9.4). We first evaluate our approach on the nuScenes dataset followed by the evaluation on our proposed Euro-PVI dataset.

Evaluation metrics. Following Salzmann *et al.* (2020), we report, 1. Best of N (FDE): The final (euclidean) displacement error in meters using the best of $N=20$ samples (Gupta *et al.*, 2018; Sadeghian *et al.*, 2019; Salzmann *et al.*, 2020). 2. KDE NLL: The (mean) negative log-likelihood of the groundtruth trajectory under the predicted distribution estimated using a Gaussian kernel (Ivanovic and Pavone, 2019; Thiede and Brahma, 2019), computed using the code provided by Salzmann *et al.* (2020). Both these metrics aim to measure the match of the predicted trajectory distribution to the groundtruth distribution (Bhattacharyya *et al.*, 2019c; Gupta *et al.*, 2018; Lee *et al.*, 2017b).

9.4.1 nuScenes

Following Salzmann *et al.* (2020), we split the training set into the training and validation splits. The original validation split is used as test set. We provide 1 - (*upto*) 5 secs of observation (historical context) and predict 3 seconds ahead Salzmann *et al.* (2020). We compare our Joint-β-cVAE approach to the following state of the art models: Social-GAN Gupta *et al.* (2018), Sophie Sadeghian *et al.* (2019) and Trajectron++ Salzmann *et al.* (2020). Additionally, in order to measure the density and influence of interactions between the ego-vehicle and pedestrians (bicyclists) on the trajectories in nuScenes (in comparison to Euro-PVI), we also evaluate the above methods without modeling ego-vehicle - pedestrians (bicyclists) interactions. Any significant difference in performance of these models would indicate the presence of dense ego-vehicle - pedestrian (bicyclist) interactions.

To illustrate the effectiveness of our Joint-β-cVAE, we also include two ablations of our Joint-β-cVAE model, 1. A simple cVAE model and, 2. A β-cVAE model, (neither of which can model interactions). These ablations are designed to show the effectiveness of our Joint-β-cVAE model in capturing interactions in the latent space.

We report results using both the Best of N and KDE NLL metrics in Table 9.2. The P-P and P-V columns in Table 9.2 indicate whether the pedestrian (bicyclist) - pedestrian (bicyclist) and pedestrian(bicyclist) - ego-vehicle interactions are modeled – using social pooling in case of Social-GAN (Gupta *et al.*, 2018), the attention mechanism in case of Sophie (Sadeghian *et al.*, 2019), the scene graph in case of Trajectron++ (Salzmann *et al.*, 2020) and with a shared latent space in case of our Joint-β-cVAE model. Trajectron++ outperforms both the conditional GAN based Social-GAN and Sophie models – partly due to the better modeling of interaction with the scene graph compared to Social-GAN and Sophie. Also, note that Trajectron++ is built using a cVAE backbone. Thus, the performance advantage of Trajectron++ also illustrates the effectiveness of cVAE based models in capturing the distribution of future trajectories. We see that our Joint-β-cVAE outperforms Trajectron++. Additionally, our Joint-β-cVAE outperforms both the simple cVAE and β-cVAE ablations, illustrating that our Joint-β-cVAE model can effectively model interactions in the latent space. The performance advantage of our Joint-β-cVAE model over Trajectron++ shows the advantage of a joint latent space that can model the effect of interactions, in comparison to independent latent spaces which model the effect of interactions only as an additional condition to the decoder. Finally, across all models, we see that models which additionally encode pedestrian (bicyclist) - ego-vehicle interactions in nuScenes do not show a significant improvement in performance. This further lends support to the fact that pedestrian (bicyclist) - ego-vehicle interactions are sparse in the nuScenes dataset.

9.4.2 Euro-PVI

We now evaluate the different models for trajectory prediction on our novel Euro-PVI dataset. We use 792 sequences for training, 100 sequences for validation and 185 sequences for testing. The train/val/test splits do not share pedestrian (bicyclist) instances. As on nuScenes, we compare our Joint-β-cVAE model with the Social-GAN (Gupta *et al.*, 2018), Sophie (Sadeghian *et al.*, 2019), Trajectron++ (Salzmann *et al.*, 2020) models. We also include the cVAE and β-cVAE ablations (which cannot model interactions), to establish whether our Joint-β-cVAE approach can model interactions in the latent space. We follow a similar evaluation protocol as in nuScenes, where we predict trajectories up to 3 seconds into the future. However, we provide a shorter observation of 0.5 seconds as quick reactions are essential in dense traffic scenarios.

We report the results in Table 9.3. As on nuScenes, we observe that our Joint-β-cVAE approach outperforms the competing methods. The performance advantage over Trajectron++ again illustrates the advantage of the joint latent space over all agents in the scene versus an independent latent space which cannot model the effect of interactions. Our Joint-β-cVAE outperforms the cVAE and β-cVAE baselines, which illustrates that our Joint-β-cVAE model can model the effect of interactions successfully in the latent space. Additionally, the performance gain on Euro-PVI (0.03m, Best of $N = 20$) of our Joint-β-cVAE model over Trajectron++ is larger than in nuScenes. This shows that our Joint-β-cVAE model can better capture the complex distribution of pedestrians (bicyclists) trajectories in dense urban environments under the effect of interactions. We further show that performance can be improved by conditioning our Joint-β-cVAE model on visual features from the camera and lidar. We provide qualitative examples comparing our Joint-β-cVAE model to Trajectron++ in Fig. 9.6. We see in Fig. 9.6 (left) that Trajectron++ incorrectly predicts that the pedestrian will step onto the road, while our Joint-β-cVAE correctly predicts that due to the oncoming ego-vehicle the pedestrian avoids stepping onto the road. Similarly, in Fig. 9.6 (middle) our Joint-β-cVAE correctly predicts that the pedestrian quickly crosses the street due to the oncoming ego-vehicle and in Fig. 9.6 (right) the bicyclist merges in front of the ego-vehicle which slows down.

Finally, across all methods, we see the gain in performance (using both the Best of N and KDE NLL metrics) across methods when pedestrian (bicyclist) - ego-vehicle (P-V) interactions are modeled in Table 9.3 is larger than in nuScenes. This provides further evidence of dense pedestrian(bicyclist) - ego-vehicle interactions in Euro-PVI compared to the sparse interactions in nuScenes. Additionally, in Fig. 9.7 we plot the Best of N error of our Joint-β-cVAE model along with the β-cVAE ablation versus the distance of the trajectory from the ego-vehicle for both nuScenes and Euro-PVI. We see that in case of nuScenes, the error does not depend strongly on distance. This mirrors the results in Fig. 9.4 (right) which again suggests that the interactions between pedestrians and the ego-vehicle are sparse. In contrast, in case of our Euro-PVI dataset, the error is largest when the distance between the pedestrian (bicyclist) and the ego-vehicle is smallest i.e. at close encounters. This again suggests the presence of dense pedestrian (bicyclist) - ego-vehicle interactions in Euro-PVI.

Method	Best of $N=20\downarrow$		
	$t+1$ sec	$t+2$ sec	$t+3$ sec
Trajectron++ Salzmann *et al.* (2020)	0.10	0.35	0.63
Joint-β-cVAE (Ours)	0.10	0.33	0.61

Table 9.4: Transferring models trained on nuScenes to Euro-PVI.

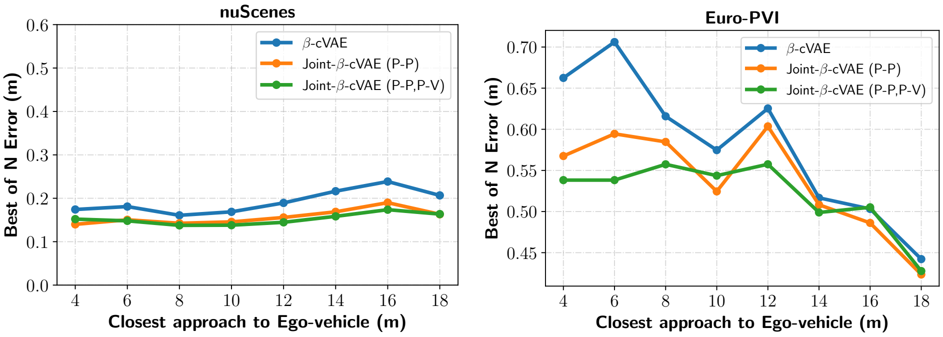


Figure 9.7: Error sorted by closest approach (proximity) to ego-vehicle. Higher error in close proximity to the ego-vehicle suggests dense interactions.

Transferring models from nuScenes. Finally, we experiment with transferring the best performing models on nuScenes i.e. Trajectron++ and our Joint-β-cVAE from nuScenes (with both P-P,P-V interactions) to Euro-PVI in Table 9.4. We observe a considerable drop in performance in the Best of N error in comparison to the performance of the models when they are *both* trained and evaluated on Euro-PVI (Table 9.3). This provides additional evidence that the distribution of trajectories and interaction patterns in Euro-PVI is significantly different compared to nuScenes.

9.5 CONCLUSION

We presented Euro-PVI, a new dataset with dense scenarios of vehicle-pedestrian (bicyclist) interaction and their trajectories to advance the task of future trajectory prediction which is crucial to the development of self-driving vehicles. We investigated the effect of interactions in urban environments on current state-of-the-art methods for existing nuScenes dataset which show a notable performance gap on our Euro-PVI dataset. To address this challenge of modeling complex interactions, we propose a Joint-β-cVAE approach. We demonstrate state of the art results both on nuScenes and on Euro-PVI. The performance advantage of our Joint-β-cVAE on Euro-PVI highlights the effectiveness of our approach in dense urban scenarios. The key to our success is a shared latent space between the interacting agents – which encodes the effect of intersections – in comparison to prior work which employ conditionally independent latent spaces. We believe that the Euro-PVI dataset along with

Joint-β-CVAE approach provides a new important dimension to the task of future trajectory prediction with dense ego-vehicle - pedestrian (bicyclist) interactions.

NORMALIZING FLOWS WITH MULTI-SCALE AUTOREGRESSIVE PRIORS 10

Contents

Flow-based generative models are an important class of exact inference models that admit efficient inference and sampling for image synthesis. Owing to the efficiency constraints on the design of the flow layers, e.g. split coupling flow layers in which approximately half the pixels do not undergo further transformations, they have limited expressiveness for modeling long-range data dependencies compared to autoregressive models that rely on conditional pixel-wise generation. In this chapter, we improve the representational power of flow-based models by introducing channel-wise dependencies in their latent space through multi-scale autoregressive priors (mAR). Our mAR prior for models with split coupling flow layers (mAR-SCF) can better capture dependencies in complex multimodal data. The resulting model achieves state-of-the-art density estimation results on MNIST, CIFAR-10, and ImageNet. Furthermore, we show that mAR-SCF allows for improved image generation quality, with gains in FID and Inception scores compared to state-of-the-art flow-based models.

10.1 INTRODUCTION

Deep generative models aim to learn complex dependencies within very high-dimensional input data, e.g. natural images (Brock *et al.*, 2019; Razavi *et al.*, 2019b) or audio data (Dieleman *et al.*, 2018), and enable generating new samples that are representative of the true data distribution. These generative models find application in various downstream tasks like image synthesis (Goodfellow *et al.*, 2014; Kingma and Welling, 2014; van den Oord *et al.*, 2016b) or speech synthesis

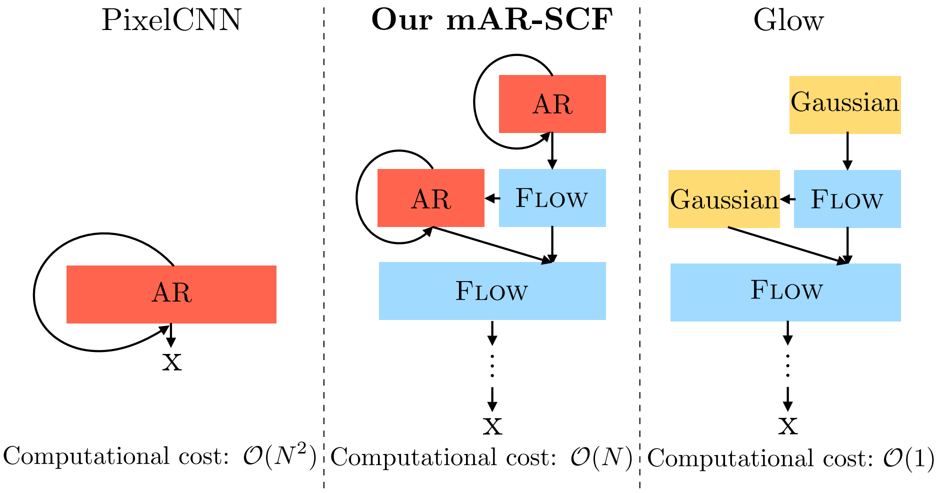

PixelCNN
AR
x
Computational cost: $\mathcal{O}(N^2)$
Our mAR-SCF
AR
AR
Flow
Flow
x
Computational cost: $\mathcal{O}(N)$
Glow
Gaussian
Gaussian
Flow
Flow
x
Computational cost: $\mathcal{O}(1)$

Figure 10.1: Our *mAR-SCF* model combines normalizing flows with autoregressive (AR) priors to improve modeling power while ensuring that the computational cost grows linearly with the spatial image resolution $N \times N$.

(Dieleman *et al.*, 2018; van den Oord *et al.*, 2018). Since it is not feasible to learn the exact distribution, generative models generally approximate the underlying true distribution. Popular generative models for capturing complex data distributions are Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014), which model the distribution implicitly and generate (high-dimensional) samples by transforming a noise distribution into the desired space with complex dependencies; however, they may not cover all modes of the underlying data distribution. Variational Autoencoders (VAEs) (Kingma and Welling, 2014) optimize a lower bound on the log-likelihood of the data. This implies that VAEs can only approximately optimize the log-likelihood (Rezende *et al.*, 2014).

Autoregressive models (Domke *et al.*, 2008; van den Oord *et al.*, 2016a,b) and normalizing flow-based generative models (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) are exact inference models that optimize the exact log-likelihood of the data. Autoregressive models can capture complex and long-range dependencies between the dimensions of a distribution, e.g. in case of images, as the value of a pixel is conditioned on a large context of neighboring pixels. The main limitation of this approach is that image synthesis is sequential and thus difficult to parallelize. Recently proposed normalizing flow-based models, such as NICE (Dinh *et al.*, 2015), RealNVP (Dinh *et al.*, 2017), and Glow (Kingma and Dhariwal, 2018), allow exact inference by mapping the input data to a known base distribution, e.g. a Gaussian, through a series of invertible transformations. These models leverage invertible split coupling flow (SCF) layers in which certain dimensions are left unchanged by the invertible transformation as well as Split operations following which certain dimensions do not undergo subsequent transformations. This allows for considerably easier parallelization of both inference and generation processes. However, these models lag behind autoregressive models for density estimation.

In this chapter, we *(i)* propose multi-scale autoregressive priors for invertible flow models with split coupling flow layers, termed *mAR-SCF*, to address the limited modeling power of non-autoregressive invertible flow models (Dinh *et al.*, 2017; Ho *et al.*, 2019; Kingma and Dhariwal, 2018; Razavi *et al.*, 2019b) (Fig. 10.1); *(ii)* we apply our multi-scale autoregressive prior after every Split operation such that the computational cost of sampling grows linearly in the spatial dimensions of the image compared to the quadratic cost of traditional autoregressive models (given sufficient parallel resources); *(iii)* our experiments show that we achieve state-of-the-art density estimation results on MNIST (LeCun *et al.*, 1998), CIFAR-10 (Krizhevsky *et al.*, 2009), and ImageNet (Russakovsky *et al.*, 2015) compared to prior invertible flow-based approaches; and finally *(iv)* we show that our multi-scale autoregressive prior leads to better sample quality as measured by the FID metric (Heusel *et al.*, 2017) and the Inception score (Salimans *et al.*, 2016), significantly lowering the gap to GAN approaches (Radford *et al.*, 2016; Wei *et al.*, 2018).

10.2 RELATED WORK

While we provide a broader discussion on related work in Chapter 2, here we discuss related work relevant to this chapter.

Methods with complex priors. Recent work (Chen *et al.*, 2017) develops complex priors to improve the data likelihoods. VQ-VAE2 integrates autoregressive models as priors (Razavi *et al.*, 2019b) with discrete latent variables (Chen *et al.*, 2017) for high-quality image synthesis and proposes latent graph-based models in a VAE framework. Tomczak and Welling (2018) propose mixtures of Gaussians with predefined clusters, and Chen *et al.* (2017) use neural autoregressive model priors in the latent space, which improves results for image synthesis. Ziegler and Rush (2019) learn a prior based on normalizing flows to capture multimodal discrete distributions of character-level texts in the latent spaces with nonlinear flow layers. However, this invertible layer is difficult to be optimized in both directions. Moreover, these models do not allow for exact inference. In this chapter, we propose complex autoregressive priors to improve the power of invertible split coupling-based normalizing flows (Dinh *et al.*, 2017; Ho *et al.*, 2019; Kingma and Dhariwal, 2018).

10.3 OVERVIEW AND BACKGROUND

In this chapter, we propose multi-scale autoregressive priors for split coupling-based flow models, termed *mAR-SCF*, where we leverage autoregressive models to improve the modeling flexibility of invertible normalizing flow models without sacrificing sampling efficiency. As we build upon normalizing flows and autoregressive models, we first provide an overview of both (see also Chapter 7).

Normalizing flows. Normalizing flows (Dinh *et al.*, 2015) are a class of exact inference generative models. Here, we consider invertible flows, which allow for both efficient exact inference and sampling. Specifically, invertible flows consist of a sequence of n invertible functions f_{θ_i}, which transform a density on the data $\mathbf{x}$ to a density on latent variables $\mathbf{z}$,

$$\mathbf{x} \overset{f_{\theta_1}}{\longleftrightarrow} \mathbf{h}_1 \overset{f_{\theta_2}}{\longleftrightarrow} \mathbf{h}_2 \cdots \overset{f_{\theta_n}}{\longleftrightarrow} \mathbf{z}. \tag{10.1}$$

Given that we can compute the likelihood of $p(\mathbf{z})$, the likelihood of the data $\mathbf{x}$ under the transformation f can be computed using the change of variables formula,

$$\log p_\theta(\mathbf{x}) = \log p(\mathbf{z}) + \sum_{i=1}^{n} \log |\det J_{\theta_i}|, \tag{10.2}$$

where $J_{\theta_i} = {}^{\partial \mathbf{h}_i}/_{\partial \mathbf{h}_{i-1}}$ is the Jacobian of the invertible transformation f_{θ_i} going from $\mathbf{h}_{i-1}$ to $\mathbf{h}_i$ with $\mathbf{h}_0 \equiv \mathbf{x}$. Note that most prior work (Chen *et al.*, 2019; Dinh *et al.*, 2015, 2017; Ho *et al.*, 2019; Kingma and Dhariwal, 2018) considers i.i.d. Gaussian likelihood models of $\mathbf{z}$, e.g. $p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mu, \sigma)$.

These models, however, have limitations. First, the requirement of invertibility constrains the class of functions f_{θ_i} to be monotonically increasing (or decreasing), thus limiting expressiveness. Second, of the three possible variants of f_{θ_i} to date (Ziegler and Rush, 2019), MAF (masked autoregressive flows), IAF (inverse autoregressive flows), and SCF (split coupling flows), MAFs are difficult to parallelize due to sequential dependencies between dimensions and IAFs do not perform well in practice. SCFs strike the right balance with respect to parallelization and modeling power. In detail, SCFs partition the dimensions into two equal halves *and* transform one of the halves $\mathbf{r}_i$ conditioned on $\mathbf{l}_i$, leaving $\mathbf{l}_i$ unchanged and thus not introducing any sequential dependencies (making parallelization easier). Examples of SCFs include the affine couplings of RealNVP (Dinh *et al.*, 2017) and MixLogCDF couplings of Flow++ (Ho *et al.*, 2019).

In practice, SCFs are organized into blocks (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) to maximize efficiency such that each f_{θ_i} typically consists of Squeeze, StepOfFlow, and Split operations. Squeeze trades off spatial resolution for channel depth. Suppose an intermediate layer $\mathbf{h}_i$ is of size $[\mathrm{C}_i, \mathrm{N}_i, \mathrm{N}_i]$, then the Squeeze operation transforms it into size $[4\,\mathrm{C}_i, \mathrm{N}_i/2, \mathrm{N}_i/2]$ by reshaping 2×2 neighborhoods into 4 channels. StepOfFlow is a series of SCF (possibly several) coupling layers and invertible 1×1 convolutions (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018).

The Split operation (distinct from the split couplings) splits an intermediate layer $\mathbf{h}_i$ into two halves $\{\mathbf{l}_i, \mathbf{r}_i\}$ of size $[2\,\mathrm{C}_i, \mathrm{N}_i/2, \mathrm{N}_i/2]$ each. Subsequent invertible layers $f_{\theta j > i}$ operate only on $\mathbf{r}_i$, leaving $\mathbf{l}_i$ unchanged. In other words, the Split operation fixes some dimensions of the latent representation $\mathbf{z}$ to $\mathbf{l}_i$ as they are not transformed any further. This leads to a significant reduction in the amount of computation and memory needed. In the following, we denote the spatial resolutions at the n different levels as $\mathrm{N} = \{\mathrm{N}_0, \cdots, \mathrm{N}_n\}$, with $N = \mathrm{N}_0$ being the input resolution. Similarly, $\mathrm{C} = \{\mathrm{C}_0, \cdots, \mathrm{C}_n\}$ denotes the number of feature channels, with $C = \mathrm{C}_0$ being the number of input channels.

In practice, due to limited modeling flexibility, prior SCF-based models (Dinh *et al.*, 2017; Ho *et al.*, 2019; Kingma and Dhariwal, 2018) require many SCF coupling layers in f_{θ_i} to model complex distributions, e.g. images. This in turn leads to high memory requirements and also leads to less efficient sampling procedures.

Autoregressive models. Autoregressive generative models are another class of powerful and highly flexible exact inference models. They factorize complex target distributions by decomposing them into a product of conditional distributions, e.g. images with $N \times N$ spatial resolution as $p(\mathbf{x}) = \prod_{i=1}^{N} \prod_{j=1}^{N} p(\mathbf{x}_{i,j} | \mathbf{x}_{\text{pred}(i,j)})$ (Domke *et al.*, 2008; Graves, 2013; Papamakarios *et al.*, 2017; van den Oord *et al.*, 2016a,b). Here, $\text{pred}(i, j)$ denotes the set of predecessors of pixel (i, j). The functional form of these conditionals can be highly flexible, and allows such models to capture complex multimodal distributions. However, such a dependency structure only allows for image synthesis via ancestral sampling by generating each pixel sequentially, conditioned on the previous pixels (van den Oord *et al.*, 2016a,b), making parallelization difficult. This is also inefficient since autoregressive models, including PixelCNN and PixelRNN, require $\mathcal{O}(N^2)$ time steps for sampling.

10.4 MULTI-SCALE AUTOREGRESSIVE FLOW PRIORS

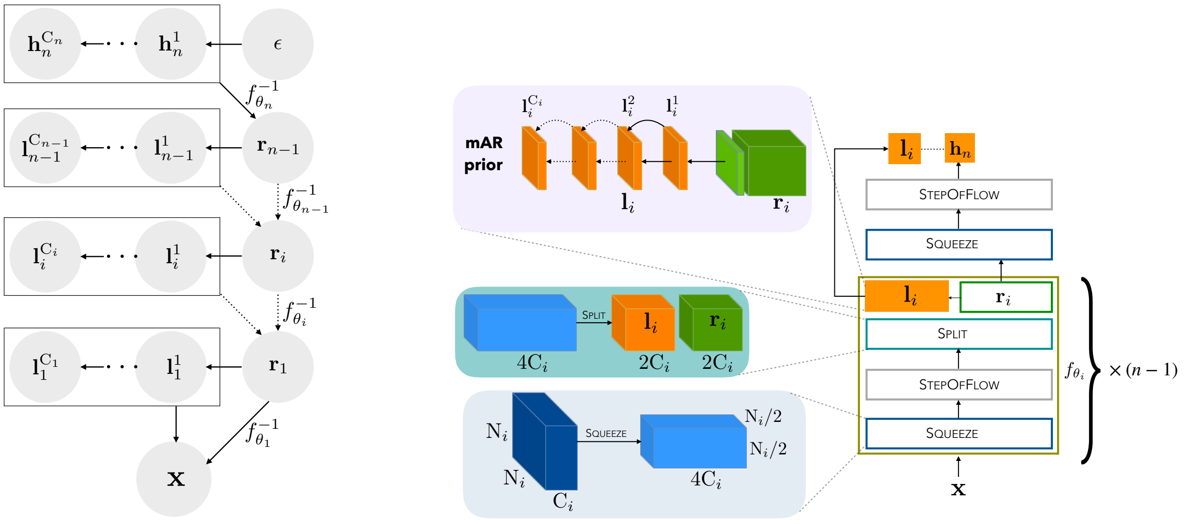
We propose to leverage the strengths of autoregressive models to improve invertible normalizing flow models such as (Dinh *et al.*, 2017; Kingma and Dhariwal, 2018). Specifically, we propose novel *multi-scale autoregressive priors for split coupling flows (mAR-SCF)*. Using them allows us to learn complex multimodal latent priors $p(\mathbf{z})$ in multi-scale SCF models, cf. Eq. (10.2). This is unlike Dinh *et al.* (2017); Ho *et al.* (2019); Kingma and Dhariwal (2018); Razavi *et al.* (2019b), which rely on Gaussian priors in the latent space. Additionally, we also propose a scheme for interpolation in the latent space of our *mAR-SCF* models.

The use of our novel autoregressive *mAR* priors for invertible flow models has two distinct advantages over both vanilla SCF and autoregressive models. First, the powerful autoregressive prior helps mitigate the limited modeling capacity of the vanilla SCF flow models. Second, as only the prior is autoregressive, this makes flow models with our *mAR* prior an order of magnitude faster with respect to sampling time than fully autoregressive models. Next, we describe our multi-scale autoregressive prior in detail.

Our *mAR-SCF* model uses an efficient invertible split coupling flow $f_{\theta_i}(\mathbf{x})$ to map the distribution over the data $\mathbf{x}$ to a latent variable $\mathbf{z}$ and then models an autoregressive *mAR* prior over $\mathbf{z}$, parameterized by ϕ. The likelihood of a data point $\mathbf{x}$ of dimensionality $[C, N, N]$ can be expressed as

$$\log p_{\theta,\phi}(\mathbf{x}) = \log p_{\phi}(\mathbf{z}) + \sum_{i=1}^{n} \log |\det J_{\theta_i}|. \qquad (10.3)$$

Here, J_{θ_i} is the Jacobian of the invertible transformations f_{θ_i}. Note that, as $f_{\theta_i}(\mathbf{x})$ is an invertible function, $\mathbf{z}$ has the same total dimensionality as the input data point



(a) Generative model for mAR-SCF.

(b) SCF flow with mAR prior.

Figure 10.2: Flow-based generative models with multi-scale autoregressive priors (*mAR-SCF*). The generative model (*left*) shows the multi-scale autoregressive sampling of the channel dimensions of $\mathbf{l}_i$ at each level. The spatial dimensions of each channel are sampled in parallel. $\mathbf{r}_i$ are computed with invertable transformations. The *mAR-SCF* model (*right*) shows the complete multi-scale architecture with the mAR prior applied along the channels of $\mathbf{l}_i$, i.e. at each level i after the SPLIT operation.

x.

10.4.1 Formulation of the *mAR* prior and *mAR-SCF* model. [¶]

***mAR* prior.** We now introduce our *mAR* prior $p_\phi(\mathbf{z})$ along with our *mAR-SCF* model, which combines the split coupling flows f_{θ_i} with an *mAR* prior. As shown in Fig. 10.2, our *mAR* prior is applied after every SPLIT operation of the invertible flow layers as well as at the smallest spatial resolution. Let $\mathbf{l}_i = \{\mathbf{l}_i^1, \cdots, \mathbf{l}_i^{C_i}\}$ be the C_i channels of size $[C_i, N_i, N_i]$, which do not undergo further transformation f_{θ_i} after the SPLIT at level i. Following the SPLIT at level i, our *mAR* prior is modeled as a conditional distribution, $p_\phi(\mathbf{l}_i|\mathbf{r}_i)$; at the coarsest spatial resolution it is an unconditional distribution, $p_\phi(\mathbf{h}_n)$. Thereby, we assume that our *mAR* prior at each level i autoregressively factorizes along the channel dimension as

$$p_\phi(\mathbf{l}_i|\mathbf{r}_i) = \prod_{j=1}^{C_i} p_\phi\left(\mathbf{l}_i^j \middle| \mathbf{l}_i^1, \cdots, \mathbf{l}_i^{j-1}, \mathbf{r}_i\right). \tag{10.4}$$

[¶] This section is based on the contributions of Shweta Mahajan from the joint work Bhattacharyya *et al.* (2020a). It is included in this thesis for completeness.

Furthermore, the distribution at each spatial location (m,n) within a channel $\mathbf{l}_i^j$ is modeled as a conditional Gaussian,

$$p_\phi(l^j_{i(m,n)}|\mathbf{l}_i^1,\cdots,\mathbf{l}_i^{j-1},\mathbf{r}_i) = \mathcal{N}\left(\mu^j_{i(m,n)},\sigma^j_{i(m,n)}\right). \tag{10.5}$$

Thus, the mean, $\mu^j_{i(m,n)}$ and variance, $\sigma^j_{i(m,n)}$ at each spatial location are autoregressively modeled along the channels. This allows the distribution at each spatial location to be highly flexible and capture multimodality in the latent space. Moreover from Eq. (10.4), our *mAR* prior can model long-range correlations in the latent space as the distribution of each channel is dependent on all previous channels.

This autoregressive factorization allows us to employ Conv-LSTMs (Shi *et al.*, 2017) to model the distributions $p_\phi(\mathbf{l}_i^j|\mathbf{l}_i^1,\cdots,\mathbf{l}_i^{j-1},\mathbf{r}_i)$ and $p_\phi(\mathbf{h}_n)$. Conv-LSTMs can model long-range dependencies across channels in their internal state. Additionally, long-range spatial dependencies within channels can be modeled by stacking multiple Conv-LSTM layers with a wide receptive field. This formulation allows all pixels within a channel to be sampled in parallel, while the channels are sampled in a sequential manner,

$$\hat{\mathbf{l}}_i^j \sim p_\phi\left(\mathbf{l}_i^j\middle|\mathbf{l}_i^1,\cdots,\mathbf{l}_i^{j-1},\mathbf{r}_i\right). \tag{10.6}$$

This is in contrast to PixelCNN/RNN-based models, which sample one pixel at a time.

***mAR-SCF* model.** We illustrate our *mAR-SCF* model architecture in Fig. 10.2(b). Our *mAR-SCF* model leverages the Squeeze and Split operations for invertible flows introduced in Dinh *et al.* (2015, 2017) for efficient parallelization. Following Dinh *et al.* (2015, 2017); Kingma and Dhariwal (2018), we use several Squeeze and Split operations in a multi-scale setup at n scales (Fig. 10.2(b)) until the spatial resolution at $\mathbf{h}_n$ is reasonably small, typically 4×4. Note that there is no Split operation at the smallest spatial resolution. Therefore, the latent space is the concatenation of $\mathbf{z} = \{\mathbf{l}_1,\ldots,\mathbf{l}_{n-1},\mathbf{h}_n\}$. The split coupling flows (SCF) f_{θ_i} in the *mAR-SCF* model remain invertible by construction. We consider different SCF couplings for f_{θ_i}, including the affine couplings of Dinh *et al.* (2017); Kingma and Dhariwal (2018) and MixLogCDF couplings (Ho *et al.*, 2019).

Given the parameters ϕ of our multimodal *mAR* prior modeled by the Conv-LSTMs, we can compute $p_\phi(\mathbf{z})$ using the formulation in Eqs. (10.4) and (10.5). We can thus express Eq. (10.3) in *closed form* and directly maximize the likelihood of the data under the multimodal *mAR* prior distribution learned by the Conv-LSTMs.

Next, we show that the computational cost of our *mAR-SCF* model is $\mathcal{O}(N)$ for sampling an image of size $[C,N,N]$; this is in contrast to the standard $\mathcal{O}(N^2)$ computational cost required by purely autoregressive models.

10.4.2 Analysis of sampling time.

We now formally analyze the computational cost in the number of steps T required for sampling with our *mAR-SCF* model. First, we describe the sampling process in

Algorithm 1 MARPS: Multi-scale Autoregressive Prior Sampling

```
Sample $\hat{\mathbf{h}}_n \sim p_\phi(\mathbf{h}_n)$
for $i \leftarrow n-1, \cdots, 1$ do
    {/*                SplitInverse                */}
    $\hat{\mathbf{r}}_i \leftarrow \hat{\mathbf{h}}_{i+1}$                     {Assign previous}
    $\hat{\mathbf{l}}_i \sim p_\phi(\mathbf{l}_i|\mathbf{r}_i)$                  {Sample mAR prior}
    $\hat{\mathbf{h}}_i \leftarrow \left\{\hat{\mathbf{l}}_i, \hat{\mathbf{r}}_i\right\}$                  {Concatenate}
    {/*              StepOfFlowInverse              */}
    Apply $f_i^{-1}(\hat{\mathbf{h}}_i)$                    {SCF coupling}
    {/*               SqueezeInverse               */}
    Reshape $\hat{\mathbf{h}}_i$                       {Depth to Space}
end for
$\mathbf{x} \leftarrow \hat{\mathbf{h}}_1$
```

detail in Algorithm 1 (the forward training process follows the sampling process in reverse order). Next, we derive the worst-case number of steps T required by MARPS, given sufficient parallel resources to sample a channel in parallel. Here, the number of steps T can be seen as the length of the critical path while sampling.

Lemma 10.4.1. *Let the sampled image $\mathbf{x}$ be of resolution $[C, N, N]$, then the worst-case number of steps T (length of the critical path) required by MARPS is $\mathcal{O}(N)$.*

Proof. At the first sampling step (Fig. 10.2(a)) at layer f_{θ_n}, our *mAR* prior is applied to generate $\mathbf{h}_n$, which is of shape $[2^{n+1}C, N/2^n, N/2^n]$. Therefore, the number of sequential steps required at *the last flow layer* $\mathbf{h}_n$ is

$$T_n = C \cdot 2^{n+1}. \tag{10.7}$$

Here, we are assuming that each channel can be sampled in parallel in one time-step.

From $f_{\theta_{n-1}}$ to f_{θ_1}, f_{θ_i} always contains a SPLIT operation. Therefore, at each f_{θ_i} we use our *mAR* prior to sample $\mathbf{l}_i$, which has shape $[2^i C, N/2^i, N/2^i]$. Therefore, the number of sequential steps required for sampling at layers $\mathbf{h}_i, 1 \leq i < n$ of our *mAR-SCF* model is

$$T_i = C \cdot 2^i. \tag{10.8}$$

Therefore, the total number of sequential steps (length of the critical path) required for sampling is

$$\begin{aligned} T &= T_n + T_{n-1} + \cdots + T_i + \cdots + T_1 \\ &= C \cdot \left(2^{n+1} + 2^{n-1} + \cdots + 2^i + \cdots + 2^1\right) \\ &= C \cdot \left(3 \cdot 2^n - 2\right). \end{aligned} \tag{10.9}$$

Now, the total number of layers in our *mAR-SCF* model is $n \leq \log(N)$. This is because each layer reduces the spatial resolution by a factor of two. Therefore, the total number of time-steps required is

$$T \leq 3 \cdot C \cdot N. \tag{10.10}$$

In practice, $C \ll N$, with $C = C_0 = 3$ for RGB images. Therefore, the total number of sequential steps required for sampling in our *mAR-SCF* model is $T = \mathcal{O}(N)$. □

It follows that with our multi-scale autoregressive *mAR* priors in our *mAR-SCF* model, sampling can be performed in a linear number of time-steps in contrast to fully autoregressive models like PixelCNN, which require a quadratic number of time-steps (van den Oord *et al.*, 2016b).

10.4.3 Interpolation.

A major advantage of invertible flow-based models is that they allow for latent spaces, which are useful for downstream tasks like interpolation – smoothly transforming one data point into another. Interpolation is simple in case of typical invertible flow-based models, because the latent space is modeled as a unimodal i.i.d. Gaussian. To allow interpolation in the space of our multimodal *mAR* priors, we develop a simple method based on Bregler and Omohundro (1994).

Let $\mathbf{x}_\text{A}$ and $\mathbf{x}_\text{B}$ be the two images (points) to be interpolated and $\mathbf{z}_\text{A}$ and $\mathbf{z}_\text{B}$ be the corresponding points in the latent space. We begin with an initial linear interpolation between the two latent points, $\left\{\mathbf{z}_\text{A}, \mathbf{z}^1_\text{A,B}, \cdots, \mathbf{z}^k_\text{A,B}, \mathbf{z}_\text{B}\right\}$, such that, $\mathbf{z}^i_\text{A,B} = (1-\alpha^i)\,\mathbf{z}_\text{A} + \alpha^i\,\mathbf{z}_\text{B}$. The initial linearly interpolated points $\mathbf{z}^i_\text{A,B}$ may not lie in a high-density region under our multimodal prior, leading to non-smooth transformations. Therefore, we next project the interpolated points $\mathbf{z}^i_\text{A,B}$ to a high-density region, without deviating too much from their initial position. This is possible because our *mAR* prior allows for exact inference. However, the image corresponding to the projected $\bar{\mathbf{z}}^i_\text{A,B}$ must also not deviate too far from either $\mathbf{x}_\text{A}$ and $\mathbf{x}_\text{B}$ either to allow for smooth transitions. To that end, we define the projection operation as

$$\begin{aligned}\bar{\mathbf{z}}^i_\text{A,B} = \arg\min \Big(& \left\|\bar{\mathbf{z}}^i_\text{A,B} - \mathbf{z}^i_\text{A,B}\right\| - \lambda_1 \log p_\phi(\bar{\mathbf{z}}^i_\text{A,B}) \\ & + \lambda_2 \min\left(\left\|f^{-1}(\bar{\mathbf{z}}^i_\text{A,B}) - \mathbf{x}_\text{A}\right\|, \left\|f^{-1}(\bar{\mathbf{z}}^i_\text{A,B}) - \mathbf{x}_\text{B}\right\|\right)\Big),\end{aligned} \tag{10.11}$$

where λ_1, λ_2 are the regularization parameters. The term controlled by λ_1 pulls the interpolated $\mathbf{z}^i_\text{A,B}$ back to high-density regions, while the term controlled by λ_2 keeps the result close to the two images $\mathbf{x}_\text{A}$ and $\mathbf{x}_\text{B}$. Note that this reduces to linear interpolation when $\lambda_1 = \lambda_2 = 0$.

10.5 EXPERIMENTS

We evaluate our approach on the MNIST (LeCun *et al.*, 1998), CIFAR-10 (Krizhevsky *et al.*, 2009), and ImageNet (van den Oord *et al.*, 2016b) datasets. In comparison

Method	Coupling	Levels	|SCF|	Channels	**bits/dim** (↓)
Glow (Kingma and Dhariwal, 2018)	Affine	3	32	512	1.05
Residual Flow (Chen *et al.*, 2019)	Residual	3	16	–	0.97
mAR-SCF (Ours)	Affine	3	32	256	1.04
mAR-SCF (Ours)	Affine	3	32	512	1.03
mAR-SCF (Ours)	MixLogCDF	3	4	96	**0.88**

Table 10.1: Evaluation of our *mAR-SCF* model on MNIST (using uniform dequantization for fair comparsion with Chen *et al.* (2019); Kingma and Dhariwal (2018)).

Method	Coupling	Levels	|SCF|	Channels	**bits/dim** (↓)
PixelCNN (van den Oord *et al.*, 2016b)	Autoregressive	–	–	–	3.00
PixelCNN++ (van den Oord *et al.*, 2016a)	Autoregressive	–	–	–	2.92
Glow (Kingma and Dhariwal, 2018)	Affine	3	32	512	3.35
Flow++ (Ho *et al.*, 2019)	MixLogCDF	3	–	96	3.29
Residual Flow (Chen *et al.*, 2019)	Residual	3	16	–	3.28
mAR-SCF (Ours)	Affine	3	32	256	3.33
mAR-SCF (Ours)	Affine	3	32	512	3.31
mAR-SCF (Ours)	MixLogCDF	3	4	96	3.27
mAR-SCF (Ours)	MixLogCDF	3	4	256	**3.24**

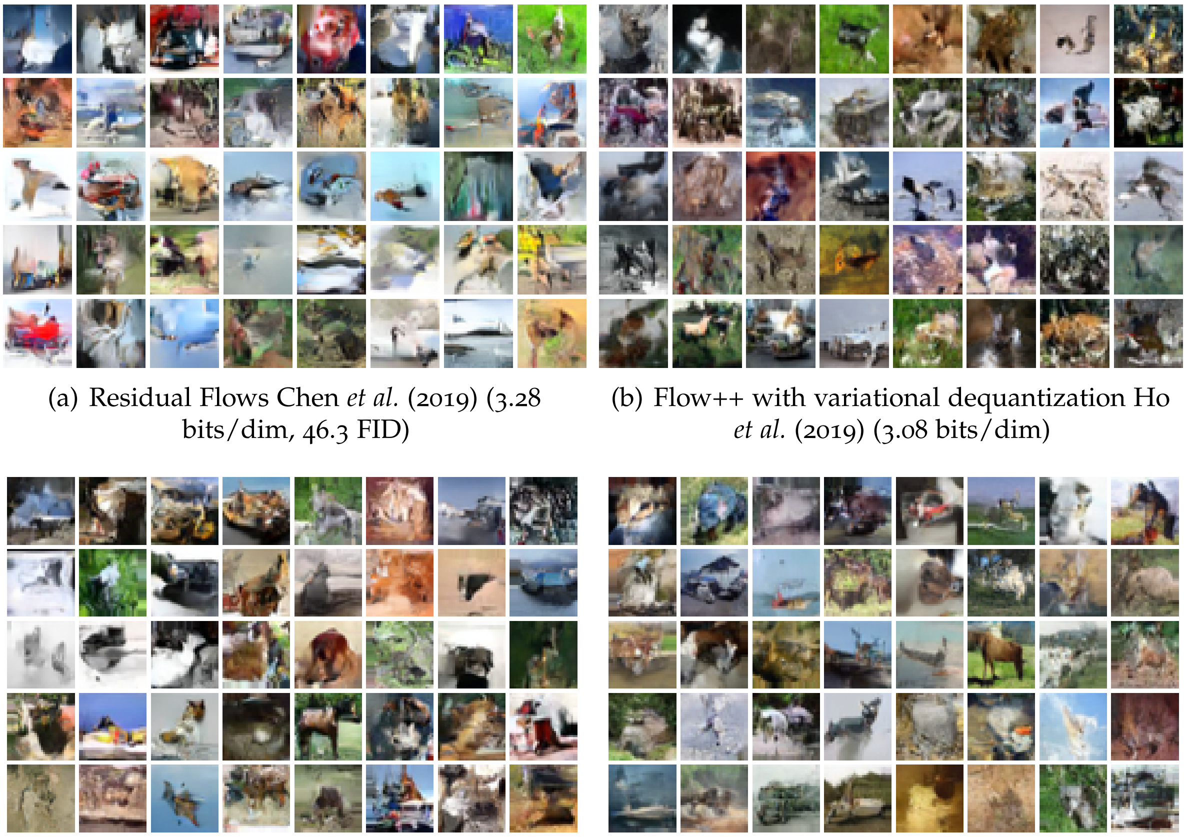
Table 10.2: Evaluation of our *mAR-SCF* model on CIFAR-10 (using uniform dequantization for fair comparsion with Chen *et al.* (2019); Kingma and Dhariwal (2018)).

to datasets like CelebA, CIFAR-10 and ImageNet are highly multimodal and the performance of invertible SCF models has lagged behind autoregressive models in density estimation and behind GAN-based generative models regarding image quality.

10.5.1 MNIST and CIFAR-10

Architecture details. Our *mAR* prior at each level f_{θ_i} consists of three convolutional LSTM layers, each of which uses 32 convolutional filters to compute the input-to-state and state-to-state components. Keeping the *mAR* prior architecture constant, we experiment with different SCF couplings in f_{θ_i} to highlight the effectiveness of our *mAR* prior. We experiment with affine couplings of Dinh *et al.* (2017); Kingma and Dhariwal (2018) and MixLogCDF couplings Ho *et al.* (2019). Affine couplings have limited modeling flexibility. The more expressive MixLogCDF applies the cumulative distribution function of a mixture of logistics. In the following, we include experiments varying the number couplings and the number of channels in the convolutional blocks of the neural networks used to predict the affine/MixLogCDF transformation parameters.

Hyperparameters. We use Adamax (as in Kingma and Dhariwal (2018)) with



(a) Residual Flows Chen *et al.* (2019) (3.28 bits/dim, 46.3 FID)

(b) Flow++ with variational dequantization Ho *et al.* (2019) (3.08 bits/dim)

(c) Our *mMAR-SCF* Affine (3.31 bits/dim, 41.0 FID)

(d) Our *mMAR-SCF* MixLogCDF (3.24 bits/dim, 41.9 FID)

Figure 10.3: Comparison of random samples from our *mAR-SCF* model with state-of-the-art models.

a learning rate of 8×10^{-4}. We use a batch size of 128 with affine and 64 with MixLogCDF couplings (following Ho *et al.* (2019)).

Density estimation. We report density estimation results on MNIST and CIFAR-10 in Tables 10.1 and 10.2 using the per-pixel log-likelihood metric in bits/dim. We also include the architecture details (# of levels, coupling type, # of channels). We compare to the state-of-the-art Flow++ (Ho *et al.*, 2019) method with SCF couplings and Residual Flows (Chen *et al.*, 2019). Note that in terms of architecture, our *mAR-SCF* model with affine couplings is closest to that of Glow (Kingma and Dhariwal, 2018). Therefore, the comparison with Glow serves as an ideal ablation to assess the effectiveness of our *mAR* prior. Flow++ (Ho *et al.*, 2019), on the other hand, uses the more powerful MixLogCDF transformations and their model architecture does not include Split operations. Because of this, Flow++ has higher computational and memory requirements for a given batch size compared to Glow. Furthermore, for fair comparison with Glow (Kingma and Dhariwal, 2018) and Residual flows (Chen *et al.*, 2019), we use uniform dequantization unlike Flow++, which proposes to use variational dequantization.

In comparison to Glow, we achieve improved density estimation results on both MNIST and CIFAR-10. In detail, we outperform Glow (e.g. 1.05 vs. 1.04 bits/dim on MNIST and 3.35 vs. 3.33 bits/dim on CIFAR-10) with $|\text{SCF}| = 32$ affine couplings and 3 levels, while using parameter prediction networks with only half (256 vs. 512) the number of channels. We observe that increasing the capacity of our parameter prediction networks to 512 channels boosts the log-likelihood further to 1.03 bits/dim on MNIST and 3.31 bits/dim on CIFAR-10. As this setting with 512 channels is identical to the setting reported in Kingma and Dhariwal (2018), this shows that our *mAR* prior boosts the accuracy by ~ 0.04 bits/dim in case of CIFAR-10. To place this performance gain in context, it is competitive with the ~ 0.03 bits/dim boost reported in Kingma and Dhariwal (2018) (cf. Fig. 3 in Kingma and Dhariwal (2018)) with the introduction of the 1×1 convolution. We train our model for ~ 3000 epochs, similar to Kingma and Dhariwal (2018). Also note that we only require a batch size of 128 to achieve state-of-the-art likelihoods, whereas Glow uses batches of size 512. Thus our *mAR-SCF* model improves density estimates and requires significantly lower computational resources (~ 48 vs. ~ 128 GB memory). Overall, we also observe competitive sampling speed (see also Table 10.3). This firmly establishes the utility of our *mAR-SCF* model.

For fair comparison with Flow++ (Ho *et al.*, 2019) and Residual Flows (Chen *et al.*, 2019), we employ the more powerful MixLogCDF couplings. Our *mAR-SCF* model uses 4 MixLogCDF couplings at each level with 96 channels but includes SPLIT operations unlike Flow++. Here, we outperform Flow++ and Residual Flows (3.27 vs. 3.29 and 3.28 bits/dim on CIFAR-10) while being equally fast to sample as Flow++ (Table 10.3). A baseline model without our *mAR* prior has performance comparable to Flow++ (3.29 bits/dim). Similarly on MNIST, our *mAR-SCF* model again outperforms Residual Flows (0.88 vs. 0.97 bits/dim). Finally, we train a more powerful *mAR-SCF* model with 256 channels with sampling speed competitive with Chen *et al.* (2019), which achieves state-of-the-art 3.24 bits/dim on CIFAR-10 [‡]. This is attained after ~ 400 training epochs (comparable to ~ 350 epochs required by Chen *et al.* (2019) to achieve 3.28 bits/dim). Next, we compare the sampling speed of our *mAR-SCF* model with that of Flow++ and Residual Flow.

Method	Coupling	Levels	|SCF|	Ch.	Speed (`ms` ↓)
Glow (Kingma and Dhariwal, 2018)	Affine	3	32	512	13
Flow++ (Ho *et al.*, 2019)	MixLogCDF	3	–	96	19
Residual Flow (Chen *et al.*, 2019)	Residual	3	16	–	34
PixelCNN++ (Salimans *et al.*, 2017)	Autoregressive	–	–	–	5e3
mAR-SCF (Ours)	Affine	3	32	256	**6**
mAR-SCF (Ours)	Affine	3	32	512	17
mAR-SCF (Ours)	MixLogCDF	3	4	96	19
mAR-SCF (Ours)	MixLogCDF	3	4	256	32

Table 10.3: Evaluation of sampling speed with batches of size 32.

[‡] *mAR-SCF* with 256 channels trained to convergence on CIFAR-10 achieves 3.22 bits/dim.

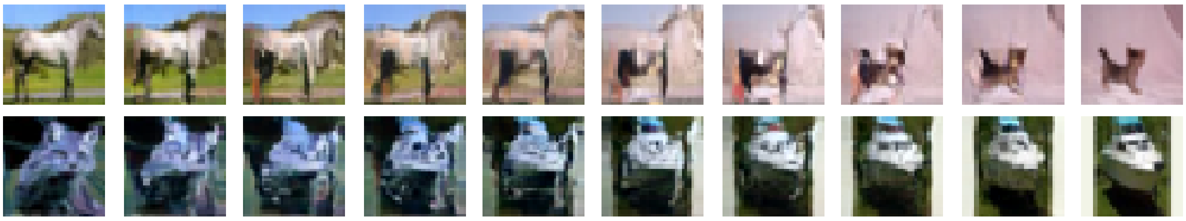
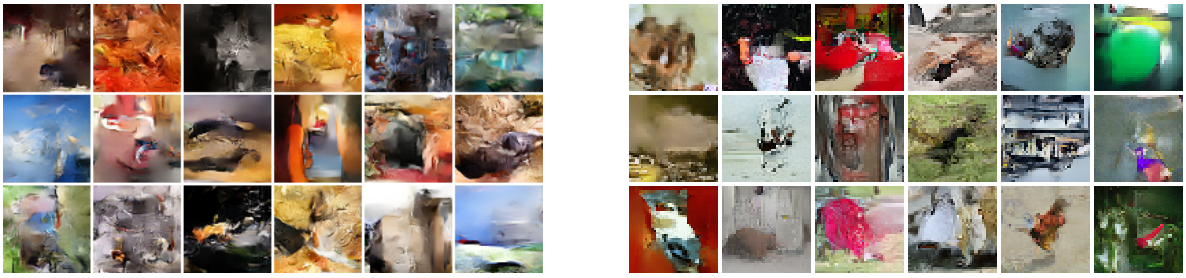


Figure 10.4: Interpolations of our *mAR-SCF* model on CIFAR-10.



(a) Residual Flows (Chen *et al.*, 2019) (3.75 bits/dim)

(b) Our *mMAR-SCF* (Affine, 3.80 bits/dim)

Figure 10.5: Random samples on ImageNet (64 × 64).

Sampling speed. We report the sampling speed of our *mAR-SCF* model in Table 10.3 in terms of sampling one image on CIFAR-10. We report the average over 1000 runs using a batch size of 32. We performed all tests on a single Nvidia V100 GPU with 32GB of memory. First, note that our *mAR-SCF* model with affine coupling layers in 3 levels with 512 channels needs 17 ms on average to sample an image. This is comparable with Glow, which requires 13 ms. This shows that our *mAR* prior causes only a slight increase in sampling time – particularly because our *mAR-SCF* requires only $\mathcal{O}(N)$ steps to sample and the prior has far fewer parameters compared to the invertible flow network. Moreover, our *mAR-SCF* model with affine coupling layers with 256 channels is considerably faster (6 vs. 13 ms) with an accuracy advantage. Similarly, our *mAR-SCF* with MixLogCDF and 96 channels is competitive in speed with Ho *et al.* (2019) with an accuracy advantage and considerably faster than Chen *et al.* (2019) (19 vs. 34 ms). This is because Residual Flows are slower to invert (sample) as there is no closed-form expression of the inverse. Furthermore, our *mAR-SCF* with MixLogCDF and 256 channels is competitive with respect to Chen *et al.* (2019) in terms of sampling speed while having a large accuracy advantage. Finally, note that these sampling speeds are two orders of magnitude faster than state-of-the-art fully autoregressive approaches, e.g. PixelCNN++ (Salimans *et al.*, 2017).

Sample quality. Next, we analyze the sample quality of our *mAR-SCF* model in Table 10.4 using the FID metric (Heusel *et al.*, 2017) and Inception scores (Salimans *et al.*, 2016). The analysis of sample quality is important as it is well-known that visual fidelity and test log-likelihoods are not necessarily indicative of each other

Method	Coupling	FID (↓)	Inception Score (↑)
PixelCNN (van den Oord *et al.*, 2016b)	Autoregressive	65.9	4.6
PixelIQN (Ostrovski *et al.*, 2018)	Autoregressive	49.4	–
Glow (Kingma and Dhariwal, 2018)	Affine	46.9	–
Residual Flow (Chen *et al.*, 2019)	Residual	46.3	5.2
mAR-SCF (Ours)	MixLogCDF	41.9	**5.7**
mAR-SCF (Ours)	Affine	**41.0**	**5.7**
DCGAN (Radford *et al.*, 2016)	Adversarial	37.1	6.4
WGAN-GP (Wei *et al.*, 2018)	Adversarial	*36.4*	*6.5*

Table 10.4: Evaluation of sample quality on CIFAR-10. Other results are quoted from Chen *et al.* (2019); Ostrovski *et al.* (2018).

Method	Coupling	|SCF|	Ch.	bits/dim (↓)	Mem (GB, ↓)
Glow (Kingma and Dhariwal, 2018)	Affine	32	512	4.09	∼ 128
Residual Flow (Chen *et al.*, 2019)	Residual	32	–	4.01	–
mAR-SCF (Ours)	Affine	32	256	4.07	∼ **48**
mAR-SCF (Ours)	MixLogCDF	4	460	**3.99**	∼ 80

Table 10.5: Evaluation on ImageNet (32 × 32).

(Theis *et al.*, 2016). We achieve an FID of 41.0 and an Inception score of 5.7 with our *mAR-SCF* model with affine couplings, significantly better than Glow with the same specifications and Residual Flows. While our *mAR-SCF* model with MixLogCDF couplings also performs comparably, empirically we find affine couplings to lead to better image quality as in Chen *et al.* (2019). We show random samples from our *mAR-SCF* model with both affine and MixLogCDF couplings in Fig. 10.3. Here, we compare to the version of Flow++ with MixLogCDF couplings and variational dequantization (which gives even better log-likelihoods) and Residual Flows. Our *mAR-SCF* model achieves better sample quality with more clearly defined objects. Furthermore, we also obtain improved sample quality over both PixelCNN and PixelIQN and close the gap in comparison to adversarial approaches like DCGAN (Radford *et al.*, 2016) and WGAN-GP (Wei *et al.*, 2018). This highlights that our *mAR-SCF* model is able to better capture long-range correlations.

Interpolation. We show interpolations on CIFAR-10 in Fig. 10.4, obtained using Eq. (10.11). We observe smooth interpolation between images belonging to distinct classes. This shows that the latent space of our *mAR* prior can be potentially used for downstream tasks similarly to Glow (Kingma and Dhariwal, 2018).

10.5.2 ImageNet

Finally, we evaluate our *mAR-SCF* model on ImageNet (32 × 32 and 64 × 64) against the best performing models on MNIST and CIFAR-10 in Table 10.5, i.e. Glow (Kingma and Dhariwal, 2018) and Residual Flows (Chen *et al.*, 2019). Our model with affine couplings outperforms Glow while using fewer channels (4.07 vs. 4.09 bits/dim). For comparison with the more powerful Residual Flow models, we use four MixLogCDF couplings at each layer f_{θ_i} with 460 channels. We again outperform Residual Flows (Chen *et al.*, 2019) (3.99 vs. 4.01 bits/dim). These results are consistent with the findings in Tables 10.1 and 10.2, highlighting the advantage of our *mAR* prior. Finally, we also evaluate on the ImageNet (64 × 64) dataset. Our *mAR-SCF* model with affine flows achieves 3.80 vs. 3.81 bits/dim in comparison to Glow Kingma and Dhariwal (2018). We show qualitative examples in Fig. 10.5 and compare to Residual Flows. We see that although the powerful Residual Flows obtain better log-likelihoods (3.75 bits/dim), our *mAR-SCF* model achieves better visual fidelity. This again highlights that our *mAR* is able to better capture long-range correlations.

10.6 CONCLUSION

We presented *mAR-SCF*, a flow-based generative model with novel multi-scale autoregressive priors for modeling long-range dependencies in the latent space of flow models. Our *mAR* prior considerably improves the accuracy of flow-based models with split coupling layers. Our experiments show that not only does our *mAR-SCF* model improve density estimation (in terms of bits/dim), but also considerably improves the sample quality of the generated images compared to previous state-of-the-art exact inference models. We believe the combination of complex priors with flow-based models, as demonstrated by our *mAR-SCF* model, provides a path toward efficient models for exact inference that approach the fidelity of GAN-based approaches.

HAAR WAVELET BASED BLOCK AUTOREGRESSIVE FLOWS FOR TRAJECTORIES 11

Contents

For pediction of trajectories such as that of pedestrians, conditional generative models like GANs and VAEs (as in Chapters 6, 7 and 9) have been leveraged for learning the distribution of likely future trajectories. Accurately modeling the dependency structure of these multimodal distributions, particularly over long time horizons remains challenging. Normalizing flow based generative models can model complex distributions admitting exact inference. These include variants with split coupling invertible transformations that are easier to parallelize compared to their autoregressive counterparts. To this end, we introduce a novel Haar wavelet based block autoregressive model leveraging split couplings, conditioned on coarse trajectories obtained from Haar wavelet based transformations at different levels of granularity. This yields an exact inference method that models trajectories at different spatio-temporal resolutions in a hierarchical manner. We illustrate the advantages of our approach for generating diverse and accurate trajectories on two real-world datasets – Stanford Drone and Intersection Drone.

11.1 INTRODUCTION

To capture the uncertainty of the real world for anticipation tasks such as trajectory prediction, it is crucial to model the distribution of likely future trajectories. Therefore recent work (Bhattacharyya *et al.*, 2018c, 2019c; Lee *et al.*, 2017b; Sadeghian *et al.*, 2019) (see also Chapters 6, 7 and 9) have focused on modeling the distribution of likely future trajectories using either generative adversarial networks (GANs, Goodfellow *et al.* (2014)) or variational autoencoders (VAEs, Kingma and Dhariwal (2018)). However, GANs are prone to mode collapse and the performance of VAEs depends on the tightness of the variational lower bound on the data log-likelihood which is hard to control in practice (Cremer *et al.*, 2018; Huang *et al.*, 2020). This makes it difficult to accurately model the distribution of likely future trajectories.

Normalizing flow based exact likelihood models (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) have been considered to overcome these limitations of GANs and VAEs in the context of image synthesis. Building on the success of these methods, recent approaches have extended the flow models for density estimation of sequential data, e.g. video (Kumar *et al.*, 2019) and audio (Kim *et al.*, 2019). Yet, VideoFlow (Kumar *et al.*, 2019) is autoregressive in the temporal dimension which results in the prediction errors accumulating over time Lee *et al.* (2018) and reduced efficiency in sampling. Furthermore, FloWaveNet (Kim *et al.*, 2019) extends flows to audio sequences with odd-even splits along the temporal dimension, encoding only *local* dependencies (Bhattacharyya *et al.*, 2020a; Huang *et al.*, 2020; Kirichenko *et al.*, 2020), (also discussed in Chapter 10). We address these challenges of flow based models for trajectory generation and develop an exact inference framework to accurately model future trajectory sequences by harnessing long-term spatio temporal structure in the underlying trajectory distribution.

In this chapter, we propose *HBA-Flow*, an exact inference model with coarse-to-fine block autoregressive structure to encode long term spatio-temporal correlations for multimodal trajectory prediction. The advantage of the proposed framework is that multimodality can be captured over long time horizons by sampling trajectories at coarse-to-fine spatial and temporal scales (Fig. 10.1). Our contributions are: 1. we introduce a block autoregressive exact inference model using Haar wavelets where flows applied at a certain scale are conditioned on coarse trajectories from previous scale. The trajectories at each level are obtained after the application of Haar wavelet based transformations, thereby modeling long term spatio-temporal correlations. 2. Our HBA-Flow model, by virtue of block autoregressive structure, integrates a multi-scale block autoregressive prior which further improves modeling flexibility by encoding dependencies in the latent space. 3. Furthermore, we show that compared to fully autoregressive approaches (Kumar *et al.*, 2019), our HBA-Flow model is computationally more efficient as the number of sampling steps grows logarithmically in trajectory length. 4. We demonstrate the effectiveness of our approach for trajectory prediction on Stanford Drone and Intersection Drone, with improved accuracy over long time horizons.

Note that, as this chapter is based on the work Bhattacharyya *et al.* (2020b), we

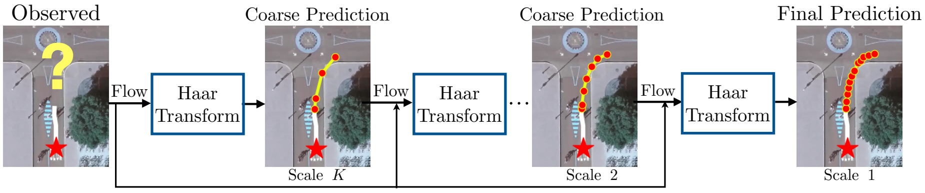


Figure 11.1: Our normalizing flow based model uses a Haar wavelet based decomposition to block autoregressively model trajectories at K coarse-to-fine scales.

compare to prior work on the Stanford Drone dataset; Lee *et al.* (2017b); Pajouheshgar and Lampert (2018); Gupta *et al.* (2018); Zhao *et al.* (2019); Sadeghian *et al.* (2019); Deo and Trivedi (2019). We provide an overview of more recent work, e.g. Mangalam *et al.* (2020) in Chapter 2.

11.2 BLOCK AUTOREGRESSIVE MODELING OF TRAJECTORIES

In this work, we propose a coarse-to-fine block autoregressive exact inference model, *HBA-Flow*, for trajectory sequences. We first provide an overview of conditional normalizing flows which form the backbone of our HBA-Flow model. To extend normalizing flows for trajectory prediction, we introduce an invertible transformation based on Haar wavelets which decomposes trajectories into K coarse-to-fine scales (Fig. 11.1). This is beneficial for expressing long-range spatio-temporal correlations as coarse trajectories provide global context for the subsequent finer scales. Our proposed HBA-Flow framework integrates the coarse-to-fine transformations with invertible split coupling flows where it block autoregressively models the transformed trajectories at K scales.

11.2.1 Conditional Normalizing Flows for Sequential Data

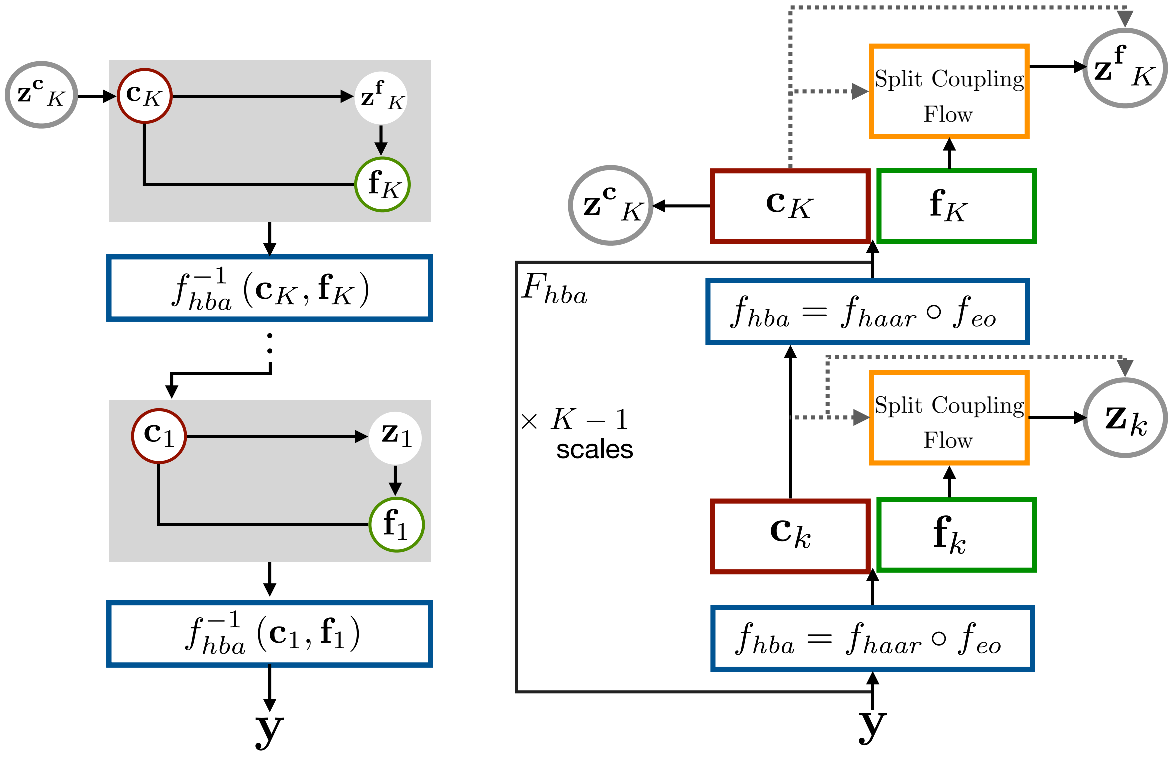
We base our HBA-Flow model on normalizing flows (Dinh *et al.*, 2015) which are a type of exact inference model (see also Chapters 7 and 10). In particular, we consider the transformation of the conditional distribution $p(\mathbf{y}|\mathbf{x})$ of trajectories $\mathbf{y}$ to a distribution $p(\mathbf{z}|\mathbf{x})$ over $\mathbf{z}$ with conditional normalizing flows (Ardizzone *et al.*, 2019b; Bhattacharyya *et al.*, 2019c) using a sequence of n transformations $g_i : \mathbf{h}_{i-1} \mapsto \mathbf{h}_i$, with $\mathbf{h}_0 = \mathbf{y}$ and parameters θ_i,

$$\mathbf{y} \overset{g_1}{\longleftrightarrow} \mathbf{h}_1 \overset{g_2}{\longleftrightarrow} \mathbf{h}_2 \cdots \overset{g_n}{\longleftrightarrow} \mathbf{z}. \tag{11.1}$$

Given the Jacobians $\mathbf{J}_{\theta_i} = {}^{\partial \mathbf{h}_i}/_{\partial \mathbf{h}_{i-1}}$ of the transformations g_i, the exact likelihoods can be computed with the change of variables formula,

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \log p(\mathbf{z}|\mathbf{x}) + \sum_{i=1}^{n} \log |\det \mathbf{J}_{\theta_i}|, \tag{11.2}$$

Given that the density $p(\mathbf{z}|\mathbf{x})$ is known, the likelihood over $\mathbf{y}$ can be computed exactly. Recent works (Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018) consider invertible split coupling transformations g_i as they provide a good balance between efficiency and modeling flexibility. In (conditional) split coupling transformations, the input $\mathbf{h}_i$ is split into two halves $\mathbf{l}_i$, $\mathbf{r}_i$, and g_i applies an invertible transformation only on $\mathbf{l}_i$ leaving $\mathbf{r}_i$ unchanged. The transformation parameters of $\mathbf{l}_i$ are dependent on $\mathbf{r}_i$ and $\mathbf{x}$, thus $\mathbf{h}_{i+1} = [g_{i+1}(\mathbf{l}_i|\mathbf{r}_i, \mathbf{x}), \mathbf{r}_i]$. The main advantage of (conditional) split coupling flows is that both inference and sampling are parallelizable when the transformations g_{i+1} have an efficient closed form expression of the inverse g_{i+1}^{-1}, e.g.



(a) Generative model.

(b) Multi-scale architecture.

Figure 11.2: Left: *HBA-Flow* generative model with the Haar wavelet Haar (1910) based representation F_{hba}. Right: Our multi-scale *HBA-Flow* model with K scales of Haar based transformation.

affine (Kingma and Dhariwal, 2018) or non-linear squared (Ziegler and Rush, 2019) and unlike residual flows (Chen *et al.*, 2019).

As most of the prior work, e.g. (Ardizzone *et al.*, 2019b; Dinh *et al.*, 2015, 2017; Kingma and Dhariwal, 2018), considers split coupling flows g_i that are designed to deal with fixed length data, these models are not directly applicable to data of variable length such as trajectories. Moreover, recall that for variable length sequences, while VideoFlow (Kumar *et al.*, 2019) utilizes split coupling based flows to model the distribution at each time-step, it is still fully autoregressive in the temporal dimension, thus offering limited computational efficiency. FloWaveNets (Kim *et al.*, 2019) split $\mathbf{l}_i$ and $\mathbf{r}_i$ along even-odd time-steps for audio synthesis. This even-odd formulation of the split operation along with the inductive bias (Kirichenko *et al.*, 2020; Huang *et al.*, 2020; Bhattacharyya *et al.*, 2020a) of split coupling based flow models is limited when expressing local and global dependencies which are crucial for capturing multimodality of the trajectories over long time horizons. Next, we introduce our invertible transformation based on Haar wavelets to model trajectories at various coarse-to-fine levels to address the shortcomings of prior flow based methods (Kumar *et al.*, 2019; Kim *et al.*, 2019) for sequential data.

11.2.2 Haar Wavelet based Invertible Transform

Haar wavelet transform allows for a simple and easy to compute coarse-to-fine frequency decomposed representation with a finite number of components unlike alternatives, e.g. Fourier transformations (Porwik and Lisowska, 2004). In our HBA-Flow framework, we construct a transformation F_{hba} consisting of mappings f_{hba} recursively applied across K scales. With this transformation, trajectories can be encoded at different levels of granularity along the temporal dimension. We now formalize invertible function f_{hba} and its multi-scale Haar wavelet based composition F_{hba}.

Single scale invertible transformation. Consider the trajectory at scale k as $\mathbf{y}_k = [\mathbf{y}_k^1, \cdots, \mathbf{y}_k^{T_k}]$, where T_k is the number of timesteps of trajectory $\mathbf{y}_k$. Here, at scale $k = 1$, $\mathbf{y_1} = \mathbf{y}$ is the input trajectory. Each element of the trajectory is a vector, $\mathbf{y}_k^j \in \mathbb{R}^d$ encoding spatial information of the traffic participant. Our proposed invertible transformation f_{hba} at any scale k is a composition, $f_{hba} = f_{haar} \circ f_{eo}$. First, f_{eo} transforms the trajectory into even $(\mathbf{e}_k)$ and odd $(\mathbf{o}_k)$ downsampled trajectories,

$$f_{eo}(\mathbf{y}_k) = \mathbf{e}_k, \mathbf{o}_k \text{ where, } \mathbf{e}_k = [\mathbf{y}_k^2, \cdots, \mathbf{y}_k^{T_k}] \text{ and } \mathbf{o}_k = [\mathbf{y}_k^1, \cdots, \mathbf{y}_k^{T_k - 1}]. \tag{11.3}$$

Next, f_{haar} takes as input the even $(\mathbf{e}_k)$ and odd $(\mathbf{o}_k)$ downsampled trajectories and transforms them into coarse $(\mathbf{c}_k)$ and fine $(\mathbf{f}_k)$ downsampled trajectories using a scalar "mixing" parameter α. In detail,

$$\begin{aligned} f_{haar}(\mathbf{e}_k, \mathbf{o}_k) = \mathbf{f}_k, \mathbf{c}_k \text{ where, } \mathbf{c}_k &= (1-\alpha)\mathbf{e}_k + \alpha \mathbf{o}_k \quad \text{and} \\ \mathbf{f}_k &= \mathbf{o}_k - \mathbf{c}_k = (1-\alpha)\mathbf{o}_k + (\alpha - 1)\mathbf{e}_k \end{aligned} \tag{11.4}$$

where, the coarse $(\mathbf{c}_k)$ trajectory is the element-wise weighted average of the even $(\mathbf{e}_k)$ and odd $(\mathbf{o}_k)$ downsampled trajectories and the fine $(\mathbf{f}_k)$ trajectory is the element-wise difference to the coarse downsampled trajectory. The coarse trajectories $(\mathbf{c}_k)$ provide global context for finer scales in our block autoregressive approach, while the fine trajectories $(\mathbf{f}_k)$ encode details at multiple scales. We now discuss the invertibility of this transformation f_{hba} and compute the Jacobian.

Lemma 11.2.1. *The generalized Haar transformation $f_{hba} = f_{haar} \circ f_{eo}$ is invertible for $\alpha \in [0, 1)$ and the determinant of the Jacobian of the transformation $f_{hba} = f_{haar} \circ f_{eo}$ for sequence of length T_k with $\mathbf{y}_k^j \in \mathbb{R}^d$ is* $\det \mathbf{J}_{hba} = (1-\alpha)^{(d \cdot T_k)/2}$.

Proof. First, note that f_{haar} in Eq. (11.4) is a linear system. To compute the Jacobian of f_{hba}, note that each element of the output fine $(\mathbf{f}_k)$ and coarse $(\mathbf{c}_k)$ trajectories can be equivalently written (using Eq. (11.3)) and Eq. (11.4)) in terms of the elements of the input trajectory $\mathbf{y}_k$. We can now rearrange the output by placing elements from $\mathbf{f}_k$ and $\mathbf{c}_k$ in an alternating fashion. This results in a Jacobian $J_{hba} \in \mathbb{R}^{d \cdot T_k \times d \cdot T_k}$ which is block diagonal, with a repeating block $\bar{J}_{hba}$ of the form,

$$\bar{J}_{hba} = \begin{pmatrix} (1-\alpha) & (\alpha - 1) \\ \alpha & (1-\alpha) \end{pmatrix}$$

This block repeats $(d\cdot T_k)/2$ times in J_H as the trajectory is of length T_k and each element of the trajectory has d dimensions. Therefore, the determinant of the Jacobian J_{hba} is $(1-\alpha)^{(d\cdot T_k)/2}$.

For $\alpha \in [0,1)$ we see that $\det J_{hba} > 0$. Thus, the linear system f_{haar} in Eq. (11.4) is non-singular and invertible. □

This property allows our HBA-Flow model to exploit f_{hba} for spatio-temporal decomposition of the trajectories $\mathbf{y}$ while remaining invertible with a tractable Jacobian for exact inference. Next, we use this transformation f_{hba} to build the coarse-to-fine multi-scale Haar wavelet based transformation F_{hba} and discuss its properties.

Multi-scale Haar wavelet based transformation. To construct our generalized Haar wavelet based transformation F_{hba}, the mapping f_{hba} is applied recursively at K scales (Fig. 11.2, left). The transformation f_{hba} at a scale k applies a low and a high pass filter pair on the input trajectory $\mathbf{y}_k$ resulting in the coarse trajectory $\mathbf{c}_k$ and the fine trajectory $\mathbf{f}_k$ with high frequency details. The coarse (spatially and temporally sub-sampled) trajectory ($\mathbf{c}_k$) at scale k is then further decomposed by using it as the input trajectory $\mathbf{y}_{k+1} = \mathbf{c}_k$ to f_{hba} at scale $k+1$. This is repeated at K scales, resulting in the complete Haar wavelet transformation $F_{hba}(\mathbf{y}) = [\mathbf{f}_1, \cdots, \mathbf{f}_K, \mathbf{c}_K]$ which captures details at multiple (K) spatio-temporal scales. The finest scale $\mathbf{f}_1$ models high-frequency spatio-temporal information of the trajectory $\mathbf{y}$. The subsequent scales $\mathbf{f}_k$ represent details at coarser levels, with $\mathbf{c}_K$ being the coarsest transformation which expresses the "high-level" spatio-temporal structure of the trajectory (Fig. 11.1).

Next, we show that the number of scales K in F_{hba} is upper bounded by the logarithm of the length of the sequence. This implies that F_{hba}, when integrated in the multi-scale block auto-regressive model, provides a computationally efficient setup for generating trajectories.

Lemma 11.2.2. *The number of scales K of the Haar wavelet based representation F_{hba} is $K \leq \log(T_1)$, for an initial input sequence $\mathbf{y}_1$ of length T_1.*

Proof. The Haar wavelet based transformation f_{hba} halves the length of trajectory $\mathbf{y}_k$ at each level k. Thus, for an initial input sequence $\mathbf{y}_1$ of length T_1, the length of the coarsest level K in $F_{hba}(\mathbf{y})$ is $|\mathbf{c}_K| = T_1/2^K \geq 1$. Thus, $K \leq \log(T_1)$. □

11.2.3 Haar Block Autoregressive Framework

HBA-Flow model. We illustrate our HBA-Flow model in Fig. 11.2. Our HBA-Flow model first transforms the trajectories $\mathbf{y}$ using F_{hba}, where the invertible transform f_{hba} is recursively applied on the input trajectory $\mathbf{y}$ to obtain $\mathbf{f}_k$ and $\mathbf{c}_k$ at scales $k \in \{1, \cdots, K\}$. Therefore, the log-likelihood of a trajectory $\mathbf{y}$ under our HBA-Flow model can be expressed using the change of variables formula as,

$$\begin{aligned}\log(p_\theta(\mathbf{y}|\mathbf{x})) &= \log(p_\theta(\mathbf{f}_1, \mathbf{c}_1|\mathbf{x})) + \log\, |\det{(\mathbf{J}_{hba})_1}\,| \\ &= \log(p_\theta(\mathbf{f}_1, \cdots, \mathbf{f}_K, \mathbf{c}_K|\mathbf{x})) + \sum_{i=1}^{K} \log\, |\det{(\mathbf{J}_{hba})_i}\,|.\end{aligned} \tag{11.5}$$

Next, our HBA-Flow model factorizes the distribution of fine trajectories w.l.o.g. such that $\mathbf{f}_k$ at level k is conditionally dependent on the representations at scales $k+1$ to K,

$$\begin{aligned}\log(p_\theta(\mathbf{f}_1, \cdots, \mathbf{f}_K, \mathbf{c}_K|\mathbf{x})) = \log(p_\theta(\mathbf{f}_1|\mathbf{f}_2, \cdots, \mathbf{f}_K, \mathbf{c}_K, \mathbf{x})) + \cdots \\ + \log(p_\theta(\mathbf{f}_K|\mathbf{c}_K, \mathbf{x})) + \log(p_\theta(\mathbf{c}_K|\mathbf{x})).\end{aligned} \tag{11.6}$$

Finally, note that $[\mathbf{f}_{k+1}, \cdots, \mathbf{f}_K, \mathbf{c}_K]$ is the output of the (bijective) transformation $F_{hba}(\mathbf{c}_k)$ where f_{hba} is recursively applied to $\mathbf{c}_k = \mathbf{y}_{k+1}$ at scales $\{k+1, \cdots, K\}$. Thus HBA-Flow equivalently models $p_\theta(\mathbf{f}_k|\mathbf{f}_{k+1}, \cdots, \mathbf{c}_K, \mathbf{x})$ as $p_\theta(\mathbf{f}_k|\mathbf{c}_k, \mathbf{x})$,

$$\begin{aligned}\log(p_\theta(\mathbf{y}|\mathbf{x})) =& \log(p_\theta(\mathbf{f}_1|\mathbf{c}_1, \mathbf{x})) + \cdots + \log(p_\theta(\mathbf{f}_K|\mathbf{c}_K, \mathbf{x})) \\ &+ \log(p_\theta(\mathbf{c}_K|\mathbf{x})) + \sum_{i=1}^{K} \log\, |\det{(\mathbf{J}_{hba})_i}\,|.\end{aligned} \tag{11.7}$$

Therefore, as illustrated in Fig. 11.2 (right), our HBA-Flow models the distribution of each of the fine components $\mathbf{f}_k$ block autoregressively conditioned on the coarse representation $\mathbf{c}_k$ at that level. The distribution $p_\theta(\mathbf{f}_k|\mathbf{c}_k, \mathbf{x})$ at each scale k is modeled using invertible conditional split coupling flows (Fig. 11.2, right) Kim *et al.* (2019), which transform the input distribution to the distribution over latent "priors" $\mathbf{z}_k$. This enables our framework to model variable length trajectories. The log-likelihood with our HBA-Flow approach can be expressed using the change of variables formula as,

$$\log(p_\theta(\mathbf{f}_k|\mathbf{c}_k, \mathbf{x})) = \log(p_\phi(\mathbf{z}_k|\mathbf{c}_k, \mathbf{x})) + \log\, |\det(\mathbf{J}_{sc})_k| \tag{11.8}$$

where, $\log\, |\det(\mathbf{J}_{sc})_k|$ is the log determinant of Jacobian $(\mathbf{J}_{sc})_k$ of the split coupling flow at level k. Thus, the likelihood of a trajectory $\mathbf{y}$ under our HBA-Flow model can be expressed exactly using Eqs. (11.7) and (11.8).

The key advantage of our approach is that after spatial and temporal downsampling of coarse scales, it is easier to model long-term spatio-temporal dependencies. Moreover, conditioning the flows at each scale on the coarse trajectory provides global context as the downsampled coarse trajectory effectively increases the spatio-temporal receptive field. This enables our HBA-Flows better capture multimodality in the distribution of likely future trajectories.

HBA-Prior. Complex multi-model priors can considerably increase the modeling flexibility of generative models (Bhattacharyya *et al.*, 2019c; Kim *et al.*, 2019; Kumar *et al.*, 2019). The block autoregressive structure of our HBA-Flow model allows us introduce a Haar block autoregressive prior (HBA-Prior) over $\mathbf{z} = [\mathbf{z}_1, \cdots, \mathbf{z^f}_K, \mathbf{z^c}_K]$ in Eq. (11.8), where $\mathbf{z}_k$ is the latent representation for scales $k \in \{1, \cdots, K-1\}$

and ${\mathbf{z}^{\mathbf{f}}}_K, {\mathbf{z}^{\mathbf{c}}}_K$ are the latents for the coarse and fine representations scales K. The log-likelihood of the prior factorizes as,

$$\begin{aligned}\log(p_\phi(\mathbf{z}|\mathbf{x})) = \log(p_\phi(\mathbf{z}_1|\mathbf{z}_2, \cdots, {\mathbf{z}^{\mathbf{f}}}_K, {\mathbf{z}^{\mathbf{c}}}_K, \mathbf{x})) + \cdots \\ + \log(p_\phi({\mathbf{z}^{\mathbf{f}}}_K|{\mathbf{z}^{\mathbf{c}}}_K, \mathbf{x})) + \log(p_\phi({\mathbf{z}^{\mathbf{c}}}_K|\mathbf{x})).\end{aligned} \tag{11.9}$$

Each coarse level representation $\mathbf{c}_k$ is the output of a bijective transformation of the latent variables $[\mathbf{z}_{k+1}, \cdots, {\mathbf{z}^{\mathbf{f}}}_K\, {\mathbf{z}^{\mathbf{c}}}_K]$ through the invertible split coupling flows and the transformations f_{hba} at scales $\{k+1, \cdots, K\}$. Thus, HBA-Prior models $p_\phi(\mathbf{z}_k|\mathbf{z}_{k+1}, \cdots, {\mathbf{z}^{\mathbf{f}}}_K, {\mathbf{z}^{\mathbf{c}}}_K, \mathbf{x})$ as $p_\phi(\mathbf{z}_k|\mathbf{c}_k, \mathbf{x})$ at every scale (Fig. 11.2, left). The log-likelihood of the prior can also be expressed as,

$$\begin{aligned}\log(p_\phi(\mathbf{z}|\mathbf{x})) = \log(p_\phi(\mathbf{z}_1|\mathbf{c}_1, \mathbf{x})) + \cdots + \log(p_\phi(\mathbf{z}_{K-1}|\mathbf{c}_{K-1}, \mathbf{x})) \\ + \log(p_\phi({\mathbf{z}^{\mathbf{f}}}_K|\mathbf{c}_K, \mathbf{x})) + \log(p_\phi({\mathbf{z}^{\mathbf{c}}}_K|\mathbf{x})).\end{aligned} \tag{11.10}$$

We model $p_\phi(\mathbf{z}_k|\mathbf{c}_k, \mathbf{x})$ as conditional normal distributions which are multimodal as a result of the block autoregressive structure. In comparison to the fully autoregressive prior in Kumar *et al.* (2019), our HBA-Prior is efficient as it requires only $\mathcal{O}(\log(T_1))$ sampling steps.

Analysis of sampling speed. From Eq. (11.6) and Fig. 11.2 (left), our HBA-Flow model autoregressively factorizes across the fine components $\mathbf{f}_k$ at K scales. From Lemma 11.2.2, $K \leq \log(T_1)$. At each scale our HBA-Flow samples the fine components $\mathbf{f}_k$ using split coupling flows, which are easy to parallelize. Thus, given enough parallel resources, our HBA-Flow model requires maximum $K \leq \log(T_1)$ i.e. $\mathcal{O}(\log(T_1))$ sampling steps and is significantly more efficient compared to fully autoregressive approaches e.g.VideoFlow (Kumar *et al.*, 2019), which require $\mathcal{O}(T_1)$ steps.

11.3 EXPERIMENTS

We evaluate our approach for trajectory prediction on two challenging real world datasets – Stanford Drone (Robicquet *et al.*, 2016) and Intersection Drone (Bock *et al.*, 2020). These datasets contain trajectories of traffic participants including pedestrians, bicycles, cars recorded from an aerial platform. The distribution of likely future trajectories is highly multimodal due to the complexity of the traffic scenarios e.g. at intersections.

Evaluation metrics. We are primarily interested in measuring the match of the learned distribution to the true distribution. Therefore, we follow (Bhattacharyya *et al.*, 2018c, 2019c; Lee *et al.*, 2017b; Pajouheshgar and Lampert, 2018), Chapters 6 and 7 and use Euclidean error of the top 10% of samples (predictions) and the (negative) conditional log-likelihood (-CLL) metrics. The Euclidean error of the top 10% of samples measures the coverage of all modes of the target distribution and is relatively robust to random guessing as shown in Bhattacharyya *et al.* (2019c).

Method	Visual	Er @ 1sec	Er @ 2sec	Er @ 3sec	Er @ 4sec	-CLL	Speed
"Shotgun" (Pajouheshgar and Lampert, 2018)	–	0.7	1.7	3.0	4.5	91.6	–
DESIRE-SI-IT4 (Lee *et al.*, 2017b)	✓	1.2	2.3	3.4	5.3	–	–
STCNN (Pajouheshgar and Lampert, 2018)	✓	1.2	2.1	3.3	4.6	–	–
BMS-CVAE (Bhattacharyya *et al.*, 2018c)	✓	0.8	1.7	3.1	4.6	126.6	58
CF-VAE (Bhattacharyya *et al.*, 2019c)	–	**0.7**	1.5	2.5	3.6	84.6	47
CF-VAE (Bhattacharyya *et al.*, 2019c)	✓	**0.7**	1.5	2.4	3.5	84.1	88
Auto-regressive (Kumar *et al.*, 2019)	–	**0.7**	1.5	2.6	3.7	86.8	134
FloWaveNet (Kim *et al.*, 2019)	–	**0.7**	1.5	2.5	3.6	84.5	**38**
FloWaveNet (Kim *et al.*, 2019) + HWD	–	**0.7**	1.5	2.5	3.6	84.4	**38**
FloWaveNet (Kim *et al.*, 2019)	✓	**0.7**	1.5	2.4	3.5	84.1	77
HBA-Flow (Ours)	–	**0.7**	1.5	2.4	3.4	84.1	41
HBA-Flow + Prior (Ours)	–	**0.7**	**1.4**	**2.3**	3.3	83.4	43
HBA-Flow + Prior (Ours)	✓	**0.7**	**1.4**	**2.3**	**3.2**	**83.1**	81

Table 11.1: Five fold cross validation on the Stanford Drone dataset. Lower is better for all metrics. Visual refers to additional conditioning on the last observed frame. Top: state of the art, Middle: Baselines and ablations, Bottom: Our HBA-Flow. (HWD is from Ardizzone *et al.* (2019b))

11.3.1 Stanford Drone

We use the standard five-fold cross validation evaluation protocol (Bhattacharyya *et al.*, 2018c, 2019c; Lee *et al.*, 2017b; Pajouheshgar and Lampert, 2018) and predict the trajectory up to 4 seconds into the future. We use the Euclidean error of the top 10% of predicted trajectories at the standard ($^1/_5$) resolution using 50 samples and the CLL metric in Table 11.1. We additionally report sampling time for a batch of 128 samples in milliseconds.

We compare our HBA-Flow model to the following state-of-the-art models: The handcrafted "Shotgun" model (Pajouheshgar and Lampert, 2018), the conditional VAE based models of Bhattacharyya *et al.* (2018c, 2019c); Lee *et al.* (2017b) (and Chapter 6) and the autoregressive STCNN model (Pajouheshgar and Lampert, 2018). We additionally include the various exact inference baselines for modeling trajectory sequences: the autoregressive flow model of VideoFlow (Kumar *et al.*, 2019), FloWaveNet (Kim *et al.*, 2019) (without our Haar wavelet based block autoregressive structure), FloWaveNet (Kim *et al.*, 2019) with the Haar wavelet downsampling of Ardizzone *et al.* (2019b) (FloWaveNet + HWD), our HBA-Flow model with a Gaussian prior (without our HBA-Prior). The FloWaveNet (Kim *et al.*, 2019) baseline serves as ideal ablations to measure the effectiveness of our block autoregressive HBA-Flow model. For fair comparison, we use two scales (levels) $K = 2$ with eight non-linear squared split coupling flows (Ziegler and Rush, 2019) each, for both our HBA-Flow and FloWaveNet (Kim *et al.*, 2019) models. Following (Bhattacharyya *et al.*, 2019c; Pajouheshgar and Lampert, 2018) and Chapter 7, we additionally experiment with conditioning on the last observed frame using a attention based CNN (indicated by "Visual" in Table 11.1).

We observe from Table 11.1 that our HBA-Flow model outperforms both state-

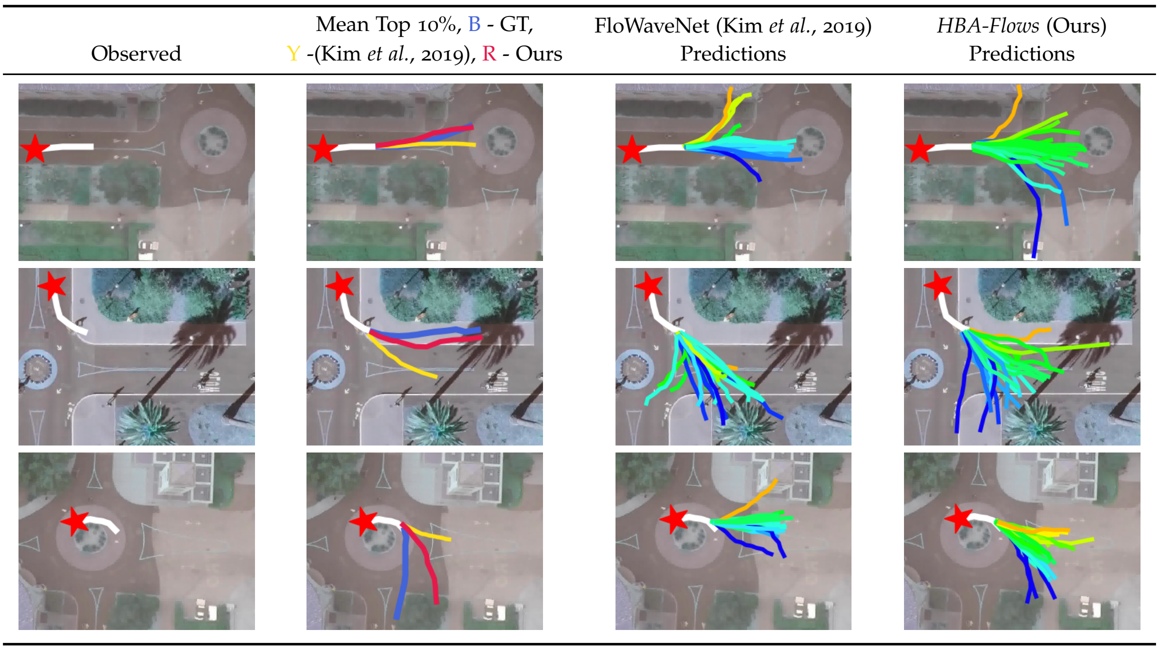

Observed
Mean Top 10%, B - GT, Y -(Kim *et al.*, 2019), R - Ours
FloWaveNet (Kim *et al.*, 2019) Predictions
HBA-Flows (Ours) Predictions

Figure 11.3: Mean top 10% predictions (Blue - Groudtruth, Yellow - FloWaveNet Kim *et al.* (2019), Red - Our *HBA-Flow* model) and predictive distributions on Intersection Drone dataset. The predictions of our HBA-Flow model are more diverse and better capture the multimodality the future trajectory distribution.

Method	mADE ↓	mFDE ↓
SocialGAN (Gupta *et al.*, 2018)	27.2	41.4
MATF GAN (Zhao *et al.*, 2019)	22.5	33.5
SoPhie (Sadeghian *et al.*, 2019)	16.2	29.3
Goal Prediction (Deo and Trivedi, 2019)	15.7	28.1
CF-VAE (Bhattacharyya *et al.*, 2019c)	12.6	22.3
HBA-Flow + Prior (Ours)	**10.8**	**19.8**

Table 11.2: Evaluation on the Stanford Drone using the split of Deo and Trivedi (2019); Sadeghian *et al.* (2019); Zhao *et al.* (2019).

of-the-art models and baselines. In particular, our HBA-Flow model outperforms the conditional VAE based models of Bhattacharyya *et al.* (2018c, 2019c); Lee *et al.* (2017b) in terms of Euclidean distance and -CLL. Further, our HBA-Flow exhibits competitive sampling speeds. This shows the advantage of exact inference in the context of generative modeling of trajectories – leading to better match to the groundtruth distribution. Our HBA-Flow model generates accurate trajectories compared to the VideoFlow (Kumar *et al.*, 2019) baseline. This is because unlike VideoFlow, errors do not accumulate in the temporal dimension of HBA-Flow. Our HBA-Flow model outperforms the FloWaveNet model of Kim *et al.* (2019) with comparable sampling speeds demonstrating the effectiveness of the coarse-to-fine block autoregressive structure of our HBA-Flow model in capturing long-range spatio-temporal dependencies. This is reflected in the predictive distributions and the top 10% of predictions of our HBA-Flow model in comparison with FloWaveNet (Kim *et al.*, 2019) in Fig. 11.3. The predictions of our HBA-Flow model are more diverse and can more effectively capture the multimodality of the trajectory distributions especially at complex traffic situations, e.g. intersections and crossings. We also observe in Table 11.1 that the addition of Haar wavelet downsampling (Ardizzone *et al.*, 2019b) to FloWaveNets (Kim *et al.*, 2019) (FloWaveNet + HWD) does not significantly improve performance. This illustrates that Haar wavelet downsampling as used in Ardizzone *et al.* (2019b) is not effective in case of sequential trajectory data as it is primarily a spatial pooling operation for image data. Finally, our ablations with Gaussian priors (HBA-Flow) additionally demonstrate the effectiveness of our HBA-Prior (HBA-Flow + Prior) with improvements with respect to accuracy. We further include a comparison using the evaluation protocol of Robicquet *et al.* (2016); Sadeghian *et al.* (2018, 2019); Deo and Trivedi (2019) in Table 11.2. Here, only a single train/test split is used. We follow Bhattacharyya *et al.* (2019c); Deo and Trivedi (2019) and use the minimum average displacement error (mADE) and minimum final displacement error (mFDE) as evaluation metrics. Similar to Bhattacharyya *et al.* (2019c); Deo and Trivedi (2019) the minimum is calculated over 20 samples. Our HBA-Flow model outperforms the state-of-the-art demonstrating the effectiveness of our approach.

11.3.2 Intersection Drone

We further include experiments on the Intersection Drone dataset (Bock *et al.*, 2020). The dataset consists of trajectories of traffic participants recorded at German intersections. In comparison to the Stanford Drone dataset, the trajectories in this dataset are typically longer. Moreover, unlike the Stanford Drone dataset which is recorded at a University Campus, this dataset covers more "typical" traffic situations. Here, we follow the same evaluation protocol as in Stanford Drone dataset and perform a five-fold cross validation and evaluate up to 5 seconds into the future.

We report the results in Table 11.3. We use the strongest baselines from Table 11.1 for comparison to our HBA-Flow + Prior model (with our HBA-Prior), with three scales, each having eight non-linear squared split coupling flows (Ziegler and Rush,

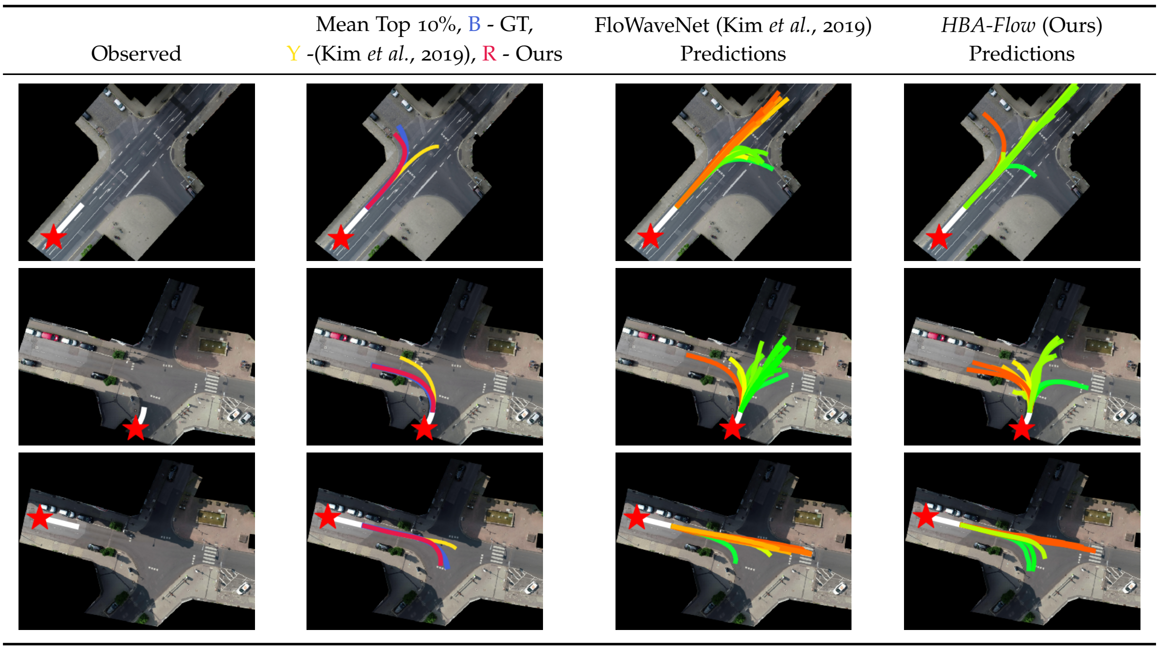

Observed
Mean Top 10%, B - GT, Y -(Kim *et al.*, 2019), R - Ours
FloWaveNet (Kim *et al.*, 2019) Predictions
HBA-Flow (Ours) Predictions

Figure 11.4: Mean top 10% predictions (Blue - Groudtruth, Yellow - FloWaveNet (Kim *et al.*, 2019), Red - Our *HBA-Flow* model) and predictive distributions on Intersection Drone dataset. The predictions of our HBA-Flow model are more diverse and better capture the modes of the future trajectory distribution.

Method	Er @ 1sec	Er @ 2sec	Er @ 3sec	Er @ 4sec	Er @ 5sec	-CLL
BMS-CVAE (Bhattacharyya *et al.*, 2018c)	0.25	0.67	1.14	1.78	2.63	26.7
CF-VAE (Bhattacharyya *et al.*, 2019c)	0.24	0.55	0.93	1.45	2.21	21.2
FloWaveNet (Kim *et al.*, 2019)	0.23	0.50	0.85	1.31	1.99	19.8
FloWaveNet (Kim *et al.*, 2019) + HWD	0.23	0.50	0.84	1.29	1.96	19.5
HBA-Flow + Prior (Ours)	**0.19**	**0.44**	**0.82**	**1.21**	**1.74**	**17.3**

Table 11.3: Five fold cross validation on the Intersection Drone dataset (HWD is from Ardizzone *et al.* (2019b)).

2019). For a fair comparison, we compare with a FloWaveNet (Kim *et al.*, 2019) model with three levels and eight non-linear squared split coupling flows per level. We again observe that our HBA-Flow leads to much better improvement with respect to accuracy over the FloWaveNet (Kim *et al.*, 2019) model. Furthermore, the performance gap between HBA-Flow and FloWaveNet increases with longer time horizons. This shows that our approach can better encode spatio-temporal correlations. The qualitative examples in Fig. 11.4 from both models show that our HBA-Flow model generates diverse trajectories and can better capture the modes of the future trajectory distribution, thus demonstrating the advantage of the block autoregressive structure of our HBA-Flow model. We also see that our HBA-Flow model outperforms the CF-VAE model (Bhattacharyya *et al.*, 2019c), again illustrating the advantage of exact inference.

11.4 CONCLUSION

In this chapter, we presented a novel block autoregressive *HBA-Flow* framework taking advantage of the representational power of autoregressive models and the efficiency of invertible split coupling flow models. Our approach can better represent the multimodal trajectory distributions capturing the long-range spatio-temporal correlations. Moreover, the block autoregressive structure of our approach provides for efficient $\mathcal{O}(\log(T))$ inference and sampling. We believe that accurate and computationally efficient invertible models that allow exact likelihood computations and efficient sampling present a promising direction of research of anticipation problems in autonomous systems.

12 CONCLUSIONS AND FUTURE PERSPECTIVES

Contents

RECENT work has lead to significant progress in the area of anticipation and future prediction to aid the development of autonomous agents which can deal with complex real-world scenarios and safety-critical environments, e.g. self-driving vehicles in dense urban landscapes (Janai *et al.*, 2020). The progress has largely been accelerated through the development of data-driven deep learning-based approaches (Mathieu *et al.*, 2016; Lee *et al.*, 2017b; Cordts *et al.*, 2016), for both deterministic and non-deterministic scenarios.

In case of deterministic scenarios, recent work has made significant progress on "intuitive" physics-based methods, facilitating anticipation capabilities for robotic manipulation and grasping (Battaglia *et al.*, 2016; Watters *et al.*, 2017). However, one of the key limitations of these approaches is the reliance on explicit object level information, e.g. velocity or acceleration, raw visual data cannot be directly leveraged. For non-deterministic scenarios, the main challenge for long-term predictions is that of uncertainty, multi-modality and exact inference – there are many likely futures and the distribution of likely futures is highly multi-modal. Recent work (Mathieu *et al.*, 2016; Finn *et al.*, 2016), which model the most likely future state while not addressing aforementioned issues of uncertainty, multi-modality and exact inference, do not perform well. Bayesian approaches (Gal and Ghahramani, 2016b; Kendall and Gal, 2017) have been applied to deal with uncertainty, however, performance on multi-modal distributions is lacking. To deal with multi-modality, generative adversarial networks (Goodfellow *et al.*, 2014) or variational autoencoders (Kingma and Welling, 2014) have been proposed, but capturing the fully diversity of the groundtruth data distribution still remains challenging (Lee *et al.*, 2017b; Salzmann *et al.*, 2020). To deal with the limitations of generative adversarial networks (Goodfellow *et al.*, 2014) or variational autoencoders (Kingma and Welling, 2014), exact inference models, e.g. normalizing flows (Kingma and Dhariwal, 2018; Ho *et al.*, 2019) are promising. Yet performance of such models are limited due to specific architectural requirements. Next, we discuss the contributions of this thesis which address these challenges in more detail.

12.1 PROGRESS TOWARDS PREDICTING THE FUTURE

The main goal of this thesis is to develop methods for long-term future prediction in diverse scenarios – both deterministic and non-deterministic scenarios. Prediction in both scenarios is important for the success of autonomous agents such as self-driving vehicles in the real world. We focus especially on dealing with the challenges of long-term prediction, uncertainty, multi-modality, and exact inference for in real-world scenarios such as street scenes. Next, we discuss our contributions with respect to these goals in each chapter.

12.1.1 Long-term Predictions

The first main contribution of this thesis is that of long-term prediction in diverse scenarios, ranging from deterministic billiard ball scenarios to street scenes and pedestrian trajectories. We find that the first key ingredient which enables long-term predictions is the effective utilization of observed (available) information. This can be in the form of image boundaries in the observed frames of a video sequence or interaction information in case of street scenes with multiple traffic participants.

In detail, we observe that the use of image boundaries from the observed frames in case of videos, leads to improved video prediction performance (Chapter 3). This is because image boundaries preserve high frequency information which is crucial for meaningful predictions about the future, while discarding details such as appearance and texture. This preserves important structures of the visual scene, while making training and long-term predictions much simpler. We also find that it is important to choose the correct model architecture for prediction: especially with a wide receptive field allowing the model to learn complex spatio-temporal dependencies, lack of bottleneck layers and a context for information sharing and global consistency. Using image boundaries we obtain sharp long-term predictions (upto $\sim$1 second into the future) in deterministic billiard ball scenarios with complex dynamics, showing that our model develops an intuitive understanding of physics. Furthermore, we show that with the fusion of image boundaries we can improve long term video frame prediction in non-deterministic scenarios even in the RGB prediction space on complex diverse datasets such as UCF-101 (Soomro *et al.*, 2012). We obtain improved results over methods such as (Mathieu *et al.*, 2016), that rely only on raw RGB pixel data. Similar to the video sequences, we observe that long-term prediction on street scenes can be improved by exploiting semantic segmentations of the observed sequence instead of relying only on raw RGB pixel data (Chapter 5). Additionally, we find that the choice of model architecture is also important – crucial to our success is a fully convolutional network with residual connections that preserves high frequency information. We show accurate long-term predictions upto 1 second into the future on complex multi-modal street scenes from Cityscapes (Cordts *et al.*, 2016). Thus, these approaches using image boundaries or semantic segmentations uses observed information more effectively and also allows for the model to be trained directly on raw visual data for widely applicability across diverse scenarios.

Additionally, we observe that long-term prediction of pedestrian trajectories can be improved by taking into account interaction information between traffic participants. To aid the development of interaction-aware trajectory prediction methods for dense urban environments, we propose a novel dataset (Chapter 9). The proposed Euro-PVI dataset contains dense scenarios of vehicle-pedestrian (bicyclist) interactions, unlike current datasets such as nuScenes (Caesar *et al.*, 2020) and Lyft L5 (Houston *et al.*, 2020) which focus mainly on trajectories of vehicles and vehicle-vehicle interactions. We see that approaches which do not model vehicle-pedestrian (bicyclist) interactions do not perform well on Euro-PVI. Thus, the Euro-PVI dataset highlights the importance of modelling interaction information – effectively utilizing observed information – for long-term trajectory prediction in dense urban environments.

12.1.2 Uncertainity and Calibration

The second main contribution of this thesis is that of calibrated estimates of uncertainty for long-term predictions in case of non-deterministic scenarios. We find that scalable Bayesian inference schemes based on dropout based Monte Carlo variational inference (Gal and Ghahramani, 2016b; Kendall and Gal, 2017) can be successfully applied to future prediction tasks for calibrated uncertainties and when trained using synthetic likelihood based objectives can also deal with complex multi-modal distributions.

In detail, we show that the Bayesian inference scheme of Gal and Ghahramani (2016b); Kendall and Gal (2017) can be extended to the complex real-world task of long-term "on-board" trajectory prediction. Crucial to our success is the choice of model architecture: a novel Bayesian two-stream recurrent model. Our two-stream model architecture jointly predicts odometry for improved long term prediction. This enables long-term predictions of at least 1 second on pedestrian trajectories from Cityscapes (Cordts *et al.*, 2016) dataset, with calibrated uncertainty estimates (in Chapter 4). However, we observe that the Bayesian inference scheme of Gal and Ghahramani (2016b); Kendall and Gal (2017) is not sufficient in case of multi-modal data distributions, e.g. predicting the future of street scenes, crucial for anticipation tasks in real-world scenarios. We show that the main bottleneck is the training objective which does not encourage diversity. Therefore, we propose a novel objective for training Bayesian inference models on multi-modal data distributions. Our novel optimization scheme uses synthetic likelihoods to deal with multi-modal distributions for future prediction tasks. We show that this encourages the Bayesian posterior distribution of models to be diverse and thus capture different modes of the multi-modal future distribution without collapsing to the mean. We apply this approach to the challenging task of long-term prediction of real-world semantic segmentations of at least 1 second into the future on street scenes from the Cityscapes (Cordts *et al.*, 2016) dataset (in Chapter 5).

To summarize, we observe that our Bayesian inference schemes provide calibrated uncertainty estimates in the long-term – even with our synthetic likelihood based optimization scheme uncertainties remain calibrated and competitive with the scheme of Gal and Ghahramani (2016b); Kendall and Gal (2017). Furthermore, in practice, we observe that the predicted uncertainties upper bound the maximum error during evaluation. This makes the predictions of our models more trustworthy and helps autonomous agents in decision making in complex real-world scenarios such as street scenes. Finally, we believe that the proposed Bayesian inference schemes are not limited to the task of pedestrian trajectory or street scene prediction and creates new opportunities to enhance high-performance deep learning architectures with principled formulations of Bayesian inference.

12.1.3 Multi-modality

The third main contribution of this thesis deals with the challenge of multi-modality in long-term future prediction tasks in non-deterministic scenarios, through novel formulations of conditional variational autoencoder based models. Specifically, we focused on non-deterministic scenarios with are especially important for the success of autonomous agents, e.g. pedestrian trajectories in dense urban environments. We find that the objective used for training and the latent prior distribution are crucial for good performance. Moreover, in case of trajectory prediction in dense urban environments the factorization of the latent prior distribution plays an important role in capturing interactions between agents.

In detail, we observe that the standard objective (Kingma and Welling, 2014; Lee *et al.*, 2017b) for training conditional variational autoencoders, does not encourage diversity in predictions making it challenging to fully capture multi-modal distributions and leading to issues such as missing modes. To deal with this issue, we proposed a novel "Best of many" sample objective for training conditional variational autoencoder based models (Chapter 6). We observe that our "Best of many" sample objective leads to a better match between the training time (aggregate) posterior and the test (sampling) time prior distributions. We show that conditional variational autoencoder based models when trained using our objective obtain improved results on diverse tasks including trajectory prediction upto 4 seconds into the future on Stanford drone dataset (Robicquet *et al.*, 2016) and precipitation nowcasting on the HKO weather dataset (Shi *et al.*, 2015). However, independent of the training objective, the latent prior distribution has an important role in determining modelling capacity of the conditional variational autoencoder model. We observe that the standard unimodal Gaussian prior imposes a strong model bias which makes it challenging to fully capture complex multi-modal distributions. Therefore, we proposed a novel – multi-modal – conditional normalizing flow based prior which allows improved modelling of complex multi-modal data distributions (Chapter 7). To further improve performance, we found regularization to be helpful, through two novel regularization techniques – posterior regularization and condition regularization. We show that posterior regularization helps to improve traning stability. On the other hand, condition regularization prevents posterior collapse leading to better fit to the target distribution. This allows us to further significantly improve performance on trajectory prediction tasks on Stanford Drone and highD datasets (Krajewski *et al.*, 2018) up to 4 seconds into the future. This confirms that both the training objective (with regularization) and latent prior distributions are crucial for best performance on complex multi-modal future prediction tasks.

Although the above mentioned approaches improved performance on pedestrian trajectory prediction tasks (e.g. on Stanford Drone and highD), they do not explicitly take into account the effect of interactions which are especially important in dense urban environments. The latent distribution of each agent is modelled independently, making it challenging to fully capture the complex multi-modal distribution of future trajectories. We observe that a shared latent space between all interacting agents which encodes the effect of intersections is especially helpful for good performance. To this end, we proposed a Joint-β-cVAE approach (Chapter 9) which infers the joint latent distribution across all agents in the scene. We observe improved performance with our Joint-β-cVAE on Euro-PVI for long-term predictions up to 3 seconds into the future, highlighting the effectiveness in capturing interactions in dense urban scenarios. Thus, these experiments show that the choice of factorization of the latent space also plays a crucial role in capturing interactions for accurate long-term predictions.

Finally, we also show that the insights gained by improving the modelling flexibility of conditional variational autoencoder based models for multi-modal data distributions (in Chapters 6 and 7), can also be applied to hybrid VAE-GAN frameworks for image datasets. In detail, we show that the integration of a "Best of Many" samples based objective (Chapter 8) improves performance of hybrid VAE-GAN frameworks. The "Best of Many" samples based objective helps in covering all the modes of the data distribution while maintaining a latent space (the aggregate posterior distribution) as close to Gaussian as possible – similar to our observations on the "Best of many" sample objective for trajectory prediction tasks (in Chapter 6). Moreover, we show that the performance is further improved by a stable estimate of the synthetic likelihood ratio term in the hybrid VAE-GAN objective. Our hybrid VAE-GAN framework trained with our novel objective, outperforms state-of-the-art hybrid VAE-GANs and plain GANs in generative modelling on CelebA and CIFAR-10, demonstrating the effectiveness of our approach.

12.1.4 Exact Inference

The fourth main contribution is the development of exact inference models for the task of future prediction in highly multi-modal non-deterministic scenarios. We focus on normalizing flows, as they allow for efficient inference and sampling – especially important for quick reactions of autonomous agents to changing environmental conditions. However, we observe that compared to alternative generative models such as conditional generative adversarial networks and conditional variational autoencoders, accurately modeling the dependency structure of of multimodal distributions, e.g. trajectories, particularly over long time horizons remains challenging with normalizing flows due to their limited modelling flexibility. Therefore, we focus on improving the modelling flexibility of normalizing flow based exact inference models.

In detail, we observe that one key factor which limits modelling flexibility normalizing flow models is the standard unimodal Gaussian prior (base) distribution. To deal with this limitation, we propose a novel autoregressive prior for normalizing flow based models to improve modelling flexibility (Chapter 10). Our multi-scale autoregressive prior for models with split coupling flow layers (mAR-SCF model) can better capture dependencies in complex multimodal data. This is because our computationally efficient mAR prior mitigates the limited expressive power of the split coupling flow layers. The model achieves state-of-the-art density estimation results on complex multi-modal datasets including MNIST, CIFAR-10, and ImageNet, while having completive sampling speeds in comparison to state of the art flow-based models. Furthermore, we show that our mAR-SCF model allows for improved image generation quality, with gains in standard image generation quality scores, e.g. FID (Heusel *et al.*, 2017) and Inception scores (Salimans *et al.*, 2016) compared to state-of-the-art flow-based models. Secondly, we observe that the modelling flexibility of normalizing flow based models can be further improved by integrating a (computationally efficient) partial auto-regressive structure. Specifically, we propose a multi-scale Haar-wavelet based decomposition for normalizing flow based trajectory prediction models (Chapter 11). Our block autoregressive model can leverage computationally efficient split couplings, while conditioning on coarse trajectories obtained from Haar wavelet based transformations at different levels of granularity for improve modelling flexibility and captures long-term correlations. This is because the coarse levels in the multi-scale structure provide global context to finer levels. Furthermore, sampling speeds remain competitive with respect to state of art trajectory prediction models. Finally, we show state of the art results upto 4 seconds into the future on two real-world datasets, Stanford Drone (Robicquet *et al.*, 2016) (improving over Chapters 6 and 7) and Intersection Drone (Bock *et al.*, 2020). To summarize, these experiments show that normalizing flows can outperform competing generative models including conditional generative adversarial networks or variational autoencoders, while providing exact likelihoods. Thus, looking ahead they are potentially the generative model of choice for future prediction tasks.

12.2 FUTURE PERSPECTIVES

The main goal of this thesis is that of development of accurate models for future prediction tasks. Despite the recent progress in this area, accurate long-term predictions for safely critical scenarios such as autonomous driving remain challenging. Next, we discuss future perspectives for the focus areas in this thesis, in particular for the challenges of long-term prediction, uncertainty, multi-modality and exact inference. Finally, in the last section we provide a broader outlook of the field.

12.2.1 Long-term Predictions

While the main challenges for long-term prediction in non-deterministic scenarios are that of uncertainty, multi-modality and exact inference, we differ the discussion on these topics to the next section. In this section, we begin with a discussion of perspectives on how to more effectively exploit available (observed) information for improving long-term predictions and thus can be applied to both deterministic and non-deterministic scenarios.

Alternative representations for video prediction. As discussed in Chapter 3, video frame prediction directly in RGB pixel space is challenging due to the difficulty in modelling the changes in appearance, e.g. texture and lighting. The key to our success on long-term prediction in Chapter 3 using image boundaries is due to the fact image boundaries discard details in appearance which makes it easier to learn (and predict) the dynamics of motion, while keeping the shape and extents of objects intact. However, a drawback of utilizing only image boundaries is that object class and instance information is lost. An interesting direction of future research is integrating image boundaries and semantic (instance) segmentations for video frame prediction. Semantic (instance) segmentations have been already explored for prediction in street scenes and shown to be easier to predict than RGB pixels. Semantic (instance) segmentations would provide object class and instance information which are missing from image boundaries. While prediction of image boundaries along with semantic (instance) segmentations would provide rich information about the future and is sufficient for many applications, for more complex tasks such as full video frame prediction appearance cues (e.g. color and texture) must also be integrated. One possibility is to formulate conditional generative models, conditioned on image boundary and segmentation information to fill in appearance details similar to Villegas *et al.* (2017). Another promising extension of this approach is to use a hierarchy of representations, where representations lower in the hierarchy capture high level features which are easy to predict in the long-term and prediction of each subsequent (finer) level with more appearance details are aided by the coarser representations.

Integrating 3d scene structure for street scene prediction. Current models for street scene prediction, e.g. the model described in Chapter 5, do not have an explicit understanding of the 3d scene structure. This makes it challenging to predict the motion of objects in the long-term because it limits the understanding of both motion patterns and interaction patterns with other objects in the scene. Therefore, long-term prediction accuracy can potentially be improved by additionally including 3d scene structure information along with RGB or semantic scene information. However, it is unclear how this can be done effectively. One potential option would be to include lidar segmentation maps as they are now widely available (Caesar *et al.*, 2020) and because they provide rich 3d scene information along with class information. However, lidar point clouds lack resolution at larger distances and full 3d scene reconstructions using automotive lidar still remain challenging. One potential solution could be to additionally include high definition map information for high level scene information, which is available in recent datasets (Caesar *et al.*, 2020; Chang *et al.*, 2019; Houston *et al.*, 2020). Secondly, lidar point clouds do not have an explicit notion of objects. Therefore, another direction for future research is that of including at least partial 3d reconstructions of street scenes using recent techniques such as occupancy networks (Mescheder *et al.*, 2019), which includes important objects such as vehicles, pedestrians, drivable surfaces and obstacles.

12.2.2 Uncertainty and Calibration

Obtaining calibrated uncertainty estimates remains a challenging task as described in Chapters 4 and 5. While the Bayesian inference approach in Chapter 5 already shows significant improvement over the naive non-Bayesian approaches, the results are still far from being perfectly calibrated. Here, we discuss two directions of future research which are promising for improving calibration of uncertainty estimates.

Ensemble based approaches. In contrast to the Bayesian inference methods presented in Chapters 4 and 5, ensemble based approaches (Lakshminarayanan *et al.*, 2017) for uncertainty estimation are readily parallelizable and require very little hyperparameter tuning with competitive uncertainty estimates. In more detail, performance of Bayesian inference based neural networks depends on the necessary approximations introduced due computational constraints. Performance further depends upon the chosen prior. Ensemble based approaches (Lakshminarayanan *et al.*, 2017) do not suffer from these disadvantages. Recent ensemble based approaches have shown promise in semantic segmentation tasks with fully convolutional architectures (K *et al.*, 2020). Extending such approaches to street scene prediction tasks using fully convolutional architectures as in Chapter 5 is promising. On the other hand, for many future prediction tasks such as trajectory prediction, recurrent neural networks have shown state-of-the-art performance (see also Chapter 4). Therefore, another direction of future research is the development of ensemble based approaches using recurrent neural networks for trajectory prediction tasks.

Alternatives to synthetic likelihoods. As discussed in Chapter 5, the Bayesian inference method in Chapter 4 does not perform well on multi-modal data distributions. The issue identified in Chapter 5 is that of the data log-likelihood term in the objective used for training the Bayesian inference method, which does not encourage diversity. The proposed synthetic likelihood term in Chapter 5 is designed to address this issue. However, the synthetic likelihood term is obtained from a classifier that is trained jointly. This makes training computationally more expensive and unstable. Therefore, an interesting direction of future research is the development of novel objectives which encourage diversity similar to our synthetic likelihood based objective. In this regard, combining normalizing flows with Bayesian inferences is an interesting direction of research Radev *et al.* (2020).

12.2.3 Multi-modality

Capturing multi-modality in complex real-world data distributions remains a challenging task as described in Chapters 6, 7 and 9 to 11. In Chapters 6 and 7 , we focused on improving the flexibility of conditional variational autoencoders to help capture complex multi-modal distributions. While exact inference models have advantages over conditional variational autoencoders (as discussed in Chapters 10 and 11), the performance of conditional variational autoencoders can be further improved. We discuss potential avenues for improvement in the following.

Autoregressive priors. We observed the most substantial gain in performance of conditional variational autoencoders on multi-modal distributions in Chapter 7 through the use of complex priors over the standard uni-modal Gaussian prior. While the conditional normalizing flow based prior in Chapter 7 offers the advantage of efficient inference, autoregressive priors can potentially offer more modelling flexibility as demonstrated on image data by Razavi *et al.* (2019b). The challenge especially for long-term future prediction tasks would be to balance the computational cost of an autoregressive prior, which grows with the prediction horizon, with the modelling flexibility.

Priors for interactions. As discussed in Chapter 9, in case of street scenes, interactions between the traffic participants are crucial in future prediction tasks. The joint prediction framework introduced in Chapter 9 accounts for the effect of interactions using a shared latent space. The latent distribution of each agent in the scene is, however, still uni-modal Gaussian distributed. The modelling flexibility of this joint prediction framework can be further improved through the use of more expressive conditional normalizing flow based priors.

12.2.4 Exact Inference

Exact inference models allow for the inference and thus optimization of the exact likelihood of a future outcome under the model, leading to improved accuracy over alternatives like conditional variational autoencoders as discussed in Chapters 10 and 11. Normalizing flow based exact inference models provide the advantage of efficiency, but suffer from limited modelling flexibility. Next, we discuss potential directions for research to improve the modelling flexibility of normalizing flow based exact inference models and integrating interactions to improve accuracy in case of street scenes.

Improving modelling flexibility of normalizing flow based models. In Chapter 11, the proposed normalizing flow model for trajectories uses affine coupling based flow layers. While such affine couplings are efficient, they are not particularly flexible. Thus, recent work (Ho *et al.*, 2019; Behrmann *et al.*, 2019; Chen *et al.*, 2019) has proposed more flexible flow layers which have shown promising performance on image data. A promising direction of future research is the development of similar more flexible flow layers for future prediction tasks. Particularly promising is the development of residual flow based methods (Chen *et al.*, 2019) for pedestrian trajectory prediction tasks. The affine couplings used in Chapter 11 employ 1d convolutions to model the transformation parameters of the flow. Therefore, it would be feasible to transform the model from Chapter 11 into a fully convolutional model that can be cast in the residual flow framework, for improved modelling flexibility.

Normalizing flow based models for interactions. In Chapter 11, the proposed normalizing flow model for trajectories does not take into account the effect of interactions on the distribution of future pedestrian trajectories. On the other hand, the model proposed in Chapter 9 proposes a joint prediction framework for conditional variational autoencoders to take into account the effect of interactions. A promising direction of future research would be to investigate similar joint prediction frameworks for normalizing flow based models. The joint prediction framework in Chapter 9 uses an autoregressive factorization of the latent variables corresponding to each agent and an attention mechanism to model the effect of interactions. An analogous approach can be used for flow based models, where 1×1 convolutions (Kingma and Dhariwal, 2018), coupling with an attention mechanism in the flow layers can be used to model the effect of interactions in a joint exact inference framework.

12.2.5 Broader Outlook of the Field

The broader goal of this thesis is the development of autonomous agents that are successful in real world situations for tasks such as autonomous driving or household assistance. Anticipation methods are crucial to this goal and such methods developed in this thesis illustrate the great progress in this field in recent years. However, for wide-spread deployment of autonomous agents in the real world there are still multiple hurdles to overcome.

Generalization to unseen scenarios. A person who has never played a game of billiards still can easily anticipate the motion of balls on the billiard table. On the other hand, current anticipation methods on the same task requires task specific training. To obtain flexible methods that mimic human ability to generalize across tasks, multi-task learning (Zamir *et al.*, 2018) approaches across anticipation tasks is an interesting direction of research. While generalizing across tasks is an important challenge, given a certain task, e.g. trajectory prediction, the dataset employed for training can still bias the learned method to the locations and specifics of that particular dataset (Wang *et al.*, 2020). Development of methods that can generalize across datasets and geographical locations remains an important avenue for future research.

Integration with planning and control modules. Most recent work on various anticipation tasks such as trajectory prediction are currently developed in isolation. However, to maximize the utility of anticipation approaches, integration with the planning and control modules of the autonomous agents would be helpful. Recent work has already shown the promise of such joint prediction, planning and control approaches (Casas *et al.*, 2021) which leads to improved results across all three tasks. Such joint leaning approaches are an important direction of future research.

Assisted versus fully autonomous driving. Autonomous driving is one of the main focus areas where anticipation plays a key role. However, fully autonomous driving still has many challenges to overcome before it can be deployed at a large scale. In the meantime, anticipation approaches can help in the area of assisted driving to improve safety. One particular case of interest is that of trajectory prediction approaches which work in parallel with the human driver to anticipate potentially dangerous situations and warn the driver in advance.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

J. Adler and S. Lunz (2018). Banach Wasserstein GAN, in *NeurIPS 2018*. Cited on pages 25, 100, 107, and 108.

S. Aigner and M. Körner (2018). FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing Autoencoder GANs, *CoRR*, vol. abs/1810.01325. Cited on page 18.

A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F. Li, and S. Savarese (2016). Social LSTM: Human Trajectory Prediction in Crowded Spaces, in *CVPR 2016*. Cited on pages 3, 5, 20, 47, 54, 55, 72, and 73.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, in *CVPR 2018*. Cited on page 118.

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik (2011). Contour Detection and Hierarchical Image Segmentation, *TPAMI*. Cited on page 32.

L. Ardizzone, J. Kruse, C. Rother, and U. Köthe (2019a). Analyzing Inverse Problems with Invertible Neural Networks, in *ICLR (Poster) 2019*. Cited on page 90.

L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe (2019b). Guided Image Generation with Conditional Invertible Neural Networks, *CoRR*, vol. abs/1907.02392. Cited on pages 143, 144, 149, 151, 152, 174, and 175.

M. Arjovsky, S. Chintala, and L. Bottou (2017). Wasserstein GAN. Cited on pages 6, 25, 100, 104, and 105.

A. Atanov, A. Volokhova, A. Ashukha, I. Sosnovik, and D. Vetrov (2019). Semi-Conditional Normalizing Flows for Semi-Supervised Learning, in *ICML Workshop 2019*. Cited on pages 90 and 94.

M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine (2018). Stochastic Variational Video Prediction, in *ICLR (Poster) 2018*. Cited on pages 18, 69, and 87.

J. Bao, D. Chen, F. Wen, H. Li, and G. Hua (2017). CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training, in *ICCV 2017*. Cited on page 7.

C. Bates, I. Yildirim, J. B. Tenenbaum, and P. W. Battaglia (2019). Modeling human intuitions about liquid flow with particle-based simulation, *PLoS Comput. Biol.*, vol. 15(7). Cited on page 3.

P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum (2013). Simulation as an engine of physical scene understanding, *Proceedings of the National Academy of Sciences*, vol. 110(45), pp. 18327–18332. Cited on page 19.

P. W. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu (2016). Interaction Networks for Learning about Objects, Relations and Physics, in *NIPS 2016*. Cited on pages 4 and 155.

J. Bayer and C. Osendorfer (2014). Learning Stochastic Recurrent Networks, *CoRR*, vol. abs/1411.7610. Cited on page 87.

J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J. Jacobsen (2019). Invertible Residual Networks, in *ICML 2019*. Cited on pages 28 and 164.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston (2009). Curriculum learning, in *ICML 2009*. Cited on page 40.

D. Berthelot, T. Schumm, and L. Metz (2017). BEGAN: Boundary Equilibrium Generative Adversarial Networks, *CoRR*, vol. abs/1703.10717. Cited on pages 24 and 109.

A. Bhattacharyya, M. Fritz, and B. Schiele (2018a). Long-Term On-Board Prediction of People in Traffic Scenes Under Uncertainty, in *CVPR 2018*. Cited on pages 9, 10, and 14.

A. Bhattacharyya, M. Fritz, and B. Schiele (2019a). Bayesian Prediction of Future Street Scenes using Synthetic Likelihoods, in *ICLR (Poster) 2019*. Cited on pages 6, 10, 15, 60, 93, and 102.

A. Bhattacharyya, M. Fritz, and B. Schiele (2019b). "Best-of-Many-Samples" Distribution Matching, *NeurIPS workshops*. Cited on pages 12, 13, and 15.

A. Bhattacharyya, M. Hanselmann, M. Fritz, B. Schiele, and C. Straehle (2019c). Conditional Flow Variational Autoencoders for Structured Sequence Prediction, *NeurIPS workshops*. Cited on pages 6, 10, 11, 15, 88, 112, 116, 117, 118, 119, 121, 141, 143, 147, 148, 149, 150, 151, and 152.

A. Bhattacharyya, S. Mahajan, M. Fritz, B. Schiele, and S. Roth (2020a). Normalizing Flows With Multi-Scale Autoregressive Priors, in *CVPR 2020*. Cited on pages 6, 12, 16, 130, 142, and 144.

A. Bhattacharyya, M. Malinowski, B. Schiele, and M. Fritz (2018b). Long-Term Image Boundary Prediction, in *AAAI 2018*. Cited on pages 6, 8, and 14.

A. Bhattacharyya, D. O. Reino, M. Fritz, and B. Schiele (2021). Euro-PVI: Pedestrian Vehicle Interactions in Dense Urban Centers, in *CVPR 2021*. Cited on pages 6, 8, 10, 11, and 16.

A. Bhattacharyya, B. Schiele, and M. Fritz (2018c). Accurate and Diverse Sampling of Sequences Based on a "Best of Many" Sample Objective, in *CVPR 2018*. Cited on pages 6, 10, 11, 13, 15, 73, 87, 93, 94, 95, 96, 101, 102, 104, 111, 116, 117, 118, 141, 148, 149, 151, and 152.

A. Bhattacharyya, C. Straehle, M. Fritz, and B. Schiele (2020b). Haar Wavelet Based Block Autoregressive Flows for Trajectories, in *GCPR 2020*. Cited on pages 6, 12, 16, and 142.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra (2015). Weight Uncertainty in Neural Network, vol. 37, pp. 1613–1622. Cited on page 22.

J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein (2020). The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections, pp. 1929–1934. Cited on pages 13, 148, 151, and 160.

M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba (2016). End to End Learning for Self-Driving Cars, *CoRR*, vol. abs/1604.07316. Cited on pages 47 and 52.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio (2016). Generating Sentences from a Continuous Space, pp. 10–21. Cited on pages 11, 88, 89, and 90.

A. Bozkurt, B. Esmaeili, D. H. Brooks, J. G. Dy, and J.-W. van de Meent (2018). Can VAEs Generate Novel Examples?, *NeurIPS Workshop*. Cited on page 100.

C. Bregler and S. M. Omohundro (1994). Nonlinear Image Interpolation using Manifold Learning, in *NIPS 1994*. Cited on page 133.

A. Brock, J. Donahue, and K. Simonyan (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis, in *ICLR 2019*. Cited on pages 25, 100, and 125.

Y. Burda, R. B. Grosse, and R. Salakhutdinov (2016). Importance Weighted Autoencoders, in *ICLR (Poster) 2016*. Cited on page 26.

H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom (2020). nuScenes: A Multimodal Dataset for Autonomous Driving, in *CVPR 2020*. Cited on pages 9, 11, 30, 112, 114, 116, 120, 157, 162, and 174.

S. Casas, A. Sadat, and R. Urtasun (2021). MP3: A Unified Model to Map, Perceive, Predict and Plan, *CoRR*, vol. abs/2101.06806. Cited on page 165.

R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha (2019). TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions, in *CVPR 2019*. Cited on pages 20 and 30.

J. Chang, D. Wei, and J. W. F. III (2013). A Video Representation Using Temporal Superpixels, in *CVPR 2013*. Cited on page 32.

M. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays (2019). Argoverse: 3D Tracking and Forecasting With Rich Maps, in *CVPR 2019*. Cited on pages 9, 30, 112, and 162.

T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li (2017). Mode Regularized Generative Adversarial Networks, in *ICLR (Poster) 2017*. Cited on page 24.

R. T. Q. Chen, J. Behrmann, D. Duvenaud, and J. Jacobsen (2019). Residual Flows for Invertible Generative Modeling, in *NeurIPS 2019*. Cited on pages 28, 128, 134, 135, 136, 137, 138, 139, 144, 164, 170, and 174.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel (2016). InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, in *NIPS 2016*. Cited on page 24.

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel (2017). Variational Lossy Autoencoder, in *ICLR 2017*. Cited on pages 26, 88, 90, and 127.

X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel (2018). PixelSNAIL: An Improved Autoregressive Generative Model, in *ICML 2018*. Cited on page 28.

G. G. Chrysos, J. Kossaifi, and S. Zafeiriou (2019). Robust Conditional Generative Adversarial Networks, in *ICLR (Poster) 2019*. Cited on page 25.

C. Chuang, J. Li, A. Torralba, and S. Fidler (2018). Learning to Act Properly: Predicting and Explaining Affordances From Images, in *CVPR 2018*. Cited on page 3.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio (2015). A Recurrent Latent Variable Model for Sequential Data, in *NIPS 2015*. Cited on page 87.

M. Comenetz (2002). *Calculus: the elements*, World Scientific Publishing Co Inc. Cited on page 76.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding, in *CVPR 2016*. Cited on pages 9, 11, 13, 53, 116, 155, 157, 158, and 174.

C. Cremer, X. Li, and D. Duvenaud (2018). Inference Suboptimality in Variational Autoencoders, vol. 80, pp. 1086–1094. Cited on pages 7 and 142.

E. D. De Jong (2016). *The MNIST Sequence Dataset., https://edwin-de-jong.github.io/blog/mnist-sequence-data/*. Cited on pages 11, 80, and 94.

E. Denton and R. Fergus (2018). Stochastic Video Generation with a Learned Prior, vol. 80, pp. 1182–1191. Cited on page 18.

E. L. Denton, S. Chintala, A. Szlam, and R. Fergus (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks, in *NIPS 2015*. Cited on page 34.

N. Deo and M. M. Trivedi (2018). Convolutional Social Pooling for Vehicle Trajectory Prediction, in *CVPR Workshops 2018*. Cited on pages 20 and 93.

N. Deo and M. M. Trivedi (2019). Scene Induced Multi-Modal Trajectory Forecasting via Planning, in *ICRA Workshop 2019*. Cited on pages 27, 73, 89, 93, 95, 97, 143, 150, 151, and 174.

A. Der Kiureghian and O. Ditlevsen (2009). Aleatory or epistemic? Does it matter?, *Structural Safety*, vol. 31(2), pp. 105–112. Cited on page 46.

F. Diehl, T. Brunner, M. Truong-Le, and A. C. Knoll (2019). Graph Neural Networks for Modelling Traffic Participant Interaction, in *IV 2019*. Cited on page 98.

S. Dieleman, A. van den Oord, and K. Simonyan (2018). The challenge of realistic music generation: Modelling raw audio at scale, in *NeurIPS 2018*. Cited on pages 125 and 126.

A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei (2019). Avoiding Latent Variable Collapse with Generative Skip Models, vol. 89, pp. 2397–2405. Cited on page 89.

L. Dinh, D. Krueger, and Y. Bengio (2015). NICE: Non-linear Independent Components Estimation, in *ICLR Workshop 2015*. Cited on pages 7, 12, 28, 90, 126, 127, 128, 131, 142, 143, and 144.

L. Dinh, J. Sohl-Dickstein, and S. Bengio (2017). Density estimation using Real NVP, in *ICLR 2017*. Cited on pages 7, 12, 28, 126, 127, 128, 129, 131, 134, 142, 143, and 144.

J. Domke, A. Karapurkar, and Y. Aloimonos (2008). Who killed the directed model?, in *CVPR 2008*. Cited on pages 28, 126, and 129.

J. Donahue, P. Krähenbühl, and T. Darrell (2017). Adversarial Feature Learning, in *ICLR (Poster) 2017*. Cited on pages 11, 27, and 100.

V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. C. Courville (2017). Adversarially Learned Inference, in *ICLR (Poster) 2017*. Cited on pages 27 and 101.

H. Eghbal-zadeh, W. Zellinger, and G. Widmer (2019). Mixture Density Generative Adversarial Networks, pp. 5820–5829. Cited on pages 104 and 105.

S. Ehrhardt, A. Monszpart, N. J. Mitra, and A. Vedaldi (2018). Unsupervised Intuitive Physics from Visual Observations, in *ACCV (3) 2018*. Cited on page 19.

D. Eigen, D. Krishnan, and R. Fergus (2013). Restoring an Image Taken through a Window Covered with Dirt or Rain, in *ICCV 2013*. Cited on page 42.

M. Elfeki, C. Couprie, M. Rivière, and M. Elhoseiny (2019). GDPP: Learning Diverse Generations using Determinantal Point Processes, vol. 97, pp. 1774–1783. Cited on page 105.

C. Elsner, T. Falck-Ytter, and G. Gredebäck (2012). Humans anticipate the goal of other people's point-light actions, *Frontiers in psychology*, vol. 3, p. 120. Cited on page 2.

P. Felsen, P. Lucey, and S. Ganguly (2018). Where Will They Go? Predicting Fine-Grained Adversarial Multi-agent Motion Using Conditional Variational Autoencoders, in *ECCV 2018*. Cited on pages 26 and 27.

C. Finn, I. J. Goodfellow, and S. Levine (2016). Unsupervised Learning for Physical Interaction through Video Prediction, in *NIPS 2016*. Cited on pages 18, 72, 73, and 155.

K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik (2016). Learning Visual Predictive Models of Physics for Playing Billiards, in *ICLR (Poster) 2016*. Cited on pages 3, 4, 19, and 21.

H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin (2019). Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing, in *NAACL-HLT (1) 2019*. Cited on pages 89 and 98.

K. Fugosic, J. Saric, and S. Segvic (2020). Multimodal Semantic Forecasting Based on Conditional Generation of Future Features, in *GCPR 2020*. Cited on page 21.

Y. Gal and Z. Ghahramani (2016a). Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference, in *ICLR (Workshop) 2016*. Cited on pages 22, 23, 49, 50, 61, 62, and 67.

Y. Gal and Z. Ghahramani (2016b). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in *ICML 2016*. Cited on pages 5, 10, 22, 23, 50, 53, 60, 61, 62, 155, 157, and 158.

Y. Gal and Z. Ghahramani (2016c). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks, in *NIPS 2016*. Cited on pages 23, 49, and 51.

F. Galasso, M. Keuper, T. Brox, and B. Schiele (2014). Spectral Graph Reduction for Efficient Image and Streaming Video Segmentation, in *CVPR 2014*. Cited on page 32.

F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele (2013). A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis, in *ICCV 2013*. Cited on pages 32 and 36.

A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison (2019). Deep Convolutional Networks as shallow Gaussian Processes, in *ICLR (Poster) 2019*. Cited on page 23.

A. Geiger, P. Lenz, and R. Urtasun (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite, in *CVPR 2012*. Cited on pages 116 and 174.

J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis, and P. Schuberth (2020). A2D2: Audi Autonomous Driving Dataset, *CoRR*. Cited on pages 116 and 174.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio (2014). Generative Adversarial Nets, in *NIPS 2014*. Cited on pages 24, 99, 125, 126, 141, and 155.

P. Goyal, Z. Hu, X. Liang, C. Wang, E. P. Xing, and C. Mellon (2017). Nonparametric Variational Auto-Encoders for Hierarchical Representation Learning, in *ICCV 2017*. Cited on page 26.

C. Graber, G. Tsai, M. Firman, G. J. Brostow, and A. G. Schwing (2021). Panoptic Segmentation Forecasting, *CoRR*, vol. abs/2104.03962. Cited on page 21.

A. Graves (2013). Generating Sequences With Recurrent Neural Networks, *arXiv:1308.0850*. Cited on pages 28, 73, and 129.

A. Graves, A. Mohamed, and G. E. Hinton (2013). Speech recognition with deep recurrent neural networks, in *ICASSP 2013*. Cited on page 51.

D. Green, O. Kochukhova, and G. Gredebäck (2014). Extrapolation and direct matching mediate anticipation in infancy, *Infant Behavior and Development*, vol. 37(1), pp. 111–118. Cited on page 2.

S. Gu, S. Levine, I. Sutskever, and A. Mnih (2016). MuProp: Unbiased Backpropagation for Stochastic Neural Networks, in *ICLR (Poster) 2016*. Cited on page 73.

X. Gu, K. Cho, J. Ha, and S. Kim (2019). DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder. Cited on pages 11, 88, 90, and 95.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville (2017a). Improved Training of Wasserstein GANs, in *NIPS 2017*. Cited on pages 25, 109, and 110.

I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taïga, F. Visin, D. Vázquez, and A. C. Courville (2017b). PixelVAE: A Latent Variable Model for Natural Images, in *ICLR (Poster) 2017*. Cited on page 89.

A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi (2018). Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks, in *CVPR 2018*. Cited on pages 3, 4, 6, 25, 73, 89, 97, 119, 120, 121, 122, 143, 150, and 174.

A. Haar (1910). Zur Theorie der orthogonalen Funktionensysteme, *Mathematische Annalen*, vol. 69(3), pp. 331–371. Cited on pages 144 and 171.

H. Hajipour, A. Bhattacharyya, and M. Fritz (2019). SampleFix: Learning to Correct Programs by Sampling Diverse Fixes, *CoRR*, vol. abs/1906.10502. Cited on page 13.

J. B. Hamrick, P. W. Battaglia, and J. B. Tenenbaum (2011). Probabilistic internal physics models guide judgments about object dynamics, in *CogSci 2011*. Cited on page 2.

K. He, X. Zhang, S. Ren, and J. Sun (2016). Deep Residual Learning for Image Recognition, in *CVPR 2016*. Cited on page 21.

D. Helbing and P. Molnar (1995). Social force model for pedestrian dynamics, *Physical review E*, vol. 51(5), p. 4282. Cited on pages 5, 20, and 47.

D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan (2020). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, in *ICLR 2020*. Cited on page 23.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in *NeurIPS 2017*. Cited on pages 12, 109, 127, 137, and 160.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, in *ICLR 2017*. Cited on pages 89, 117, 119, 120, and 174.

J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel (2019). Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design, in *ICML 2019*. Cited on pages 28, 127, 128, 129, 131, 134, 135, 136, 137, 156, 164, and 170.

J. Ho, A. Jain, and P. Abbeel (2020). Denoising Diffusion Probabilistic Models, in *NeurIPS 2020*. Cited on page 29.

S. Hochreiter and J. Schmidhuber (1997). Long Short-Term Memory, *Neural Computation*, vol. 9(8), pp. 1735–1780. Cited on page 28.

M. D. Hoffman and M. J. Johnson (2016). ELBO surgery: yet another way to carve up the variational evidence lower bound, in *NIPS Workshop 2016*. Cited on pages 11, 26, and 88.

E. Hoogeboom, R. van den Berg, and M. Welling (2019). Emerging Convolutions for Generative Normalizing Flows, in *ICML 2019*. Cited on page 28.

J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska (2020). One Thousand and One Hours: Self-driving Motion Prediction Dataset, *CoRR*. Cited on pages 9, 30, 112, 114, 116, 120, 157, 162, and 174.

W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank (2007). Semantic-based surveillance video retrieval, *IEEE Transactions on image processing*, vol. 16(4), pp. 1168–1181. Cited on page 47.

C. Huang, L. Dinh, and A. C. Courville (2020). Augmented Normalizing Flows: Bridging the Gap Between Generative Flows and Latent Variable Models, *CoRR*, vol. abs/2002.07101. Cited on pages 7, 142, and 144.

B. Ivanovic and M. Pavone (2019). The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs, in *ICCV 2019*. Cited on page 120.

V. Jain, J. F. Murray, F. Roth, S. C. Turaga, V. P. Zhigulin, K. L. Briggman, M. Helmstaedter, W. Denk, and H. S. Seung (2007). Supervised Learning of Image Restoration with Convolutional Networks, in *ICCV 2007*. Cited on page 33.

J. Janai, F. Güney, A. Behl, and A. Geiger (2020). Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art, *Found. Trends Comput. Graph. Vis.*, vol. 12(1-3), pp. 1–308. Cited on pages 1 and 155.

X. Jin, H. Xiao, X. Shen, J. Yang, Z. Lin, Y. Chen, Z. Jie, J. Feng, and S. Yan (2017). Predicting Scene Parsing and Motion Dynamics in the Future, in *NIPS 2017*. Cited on pages 21, 22, 59, and 65.

S. B. K, N. Hochgeschwender, P. Plöger, F. Kirchner, and M. Valdenegro-Toro (2020). Evaluating Uncertainty Estimation Methods on 3D Semantic Segmentation of Point Clouds, *CoRR*, vol. abs/2007.01787. Cited on page 162.

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu (2018). Efficient Neural Audio Synthesis, in *ICML 2018*. Cited on page 28.

N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu (2017). Video Pixel Networks, in *ICML 2017*. Cited on page 18.

T. Karras, T. Aila, S. Laine, and J. Lehtinen (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation, in *ICLR 2018*. Cited on pages 25 and 109.

C. G. Keller, C. Hermes, and D. M. Gavrila (2011). Will the Pedestrian Cross? Probabilistic Path Prediction Based on Learned Motion Features, in *DAGM-Symposium 2011*. Cited on page 47.

A. Kendall and Y. Gal (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, pp. 5574–5584. Cited on pages 5, 9, 10, 23, 49, 50, 53, 55, 60, 61, 62, 68, 70, 155, 157, and 158.

A. Khoreva, R. Benenson, F. Galasso, M. Hein, and B. Schiele (2016). Improved Image Boundaries for Better Video Segmentation, in *ECCV Workshops (3) 2016*. Cited on pages 8, 31, and 36.

K. Kim, D. Lee, and I. A. Essa (2011). Gaussian process regression flow for analysis of motion trajectories, in *ICCV 2011*. Cited on page 47.

S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon (2019). FloWaveNet : A Generative Flow for Raw Audio, in *ICML 2019*. Cited on pages 142, 144, 147, 149, 150, 151, 152, and 171.

D. P. Kingma and J. Ba (2015). Adam: A Method for Stochastic Optimization. Cited on pages 53, 67, and 106.

D. P. Kingma and P. Dhariwal (2018). Glow: Generative Flow with Invertible 1x1 Convolutions, in *NeurIPS 2018*. Cited on pages 7, 12, 28, 95, 126, 127, 128, 129, 131, 134, 135, 136, 138, 139, 141, 142, 143, 144, 156, 164, and 174.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling (2014). Semi-supervised Learning with Deep Generative Models, in *NIPS 2014*. Cited on page 26.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling (2016). Improved variational inference with inverse autoregressive flow, in *NIPS 2016*. Cited on pages 26 and 28.

D. P. Kingma and M. Welling (2014). Auto-Encoding Variational Bayes, in *ICLR 2014*. Cited on pages 26, 75, 100, 102, 106, 117, 118, 125, 126, 155, 156, and 158.

P. Kirichenko, P. Izmailov, and A. G. Wilson (2020). Why Normalizing Flows Fail to Detect Out-of-Distribution Data. Cited on pages 142 and 144.

A. Klushyn, N. Chen, B. Cseke, J. Bayer, and P. van der Smagt (2019). Increasing the Generalisaton Capacity of Conditional VAEs, in *ICANN (2) 2019*. Cited on page 92.

V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. D. Reid, H. Rezatofighi, and S. Savarese (2019). Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks, in *NeurIPS 2019*. Cited on page 25.

R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein (2018). The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems, in *ITSC 2018*. Cited on pages 11, 97, and 159.

A. Krizhevsky, G. Hinton, *et al.* (2009). Learning multiple layers of features from tiny images, Technical report, U. of Toronto. Cited on pages 127 and 133.

E. P. Krotkov (2012). *Active computer vision by cooperative focus and stereo*, Springer Science & Business Media. Cited on page 38.

M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma (2019). VideoFlow: A Flow-Based Generative Model for Video, *CoRR*, vol. abs/1903.01434. Cited on pages 142, 144, 147, 148, 149, and 151.

Y. Kwon and M. Park (2019). Predicting Future Frames Using Retrospective Cycle GAN, in *CVPR 2019*. Cited on page 25.

B. Lakshminarayanan, A. Pritzel, and C. Blundell (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, in *NeurIPS 2017*. Cited on pages 23 and 162.

A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther (2016). Autoencoding beyond pixels using a learned similarity metric, in *ICML 2016*. Cited on pages 11, 27, and 100.

Q. V. Le, A. J. Smola, and S. Canu (2005). Heteroscedastic Gaussian process regression, in *ICML 2005*. Cited on page 22.

P. LeBeau (2018). Waymo starts commercial ride-share service, *URL: https://www. cnbc. com/2018/12/05/waymo-starts-commercial-ride-share-service. html*. Cited on page 1.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.* (1998). Gradient-based learning applied to document recognition, *Proc. IEEE*, vol. 86(11), pp. 2278–2324. Cited on pages 127 and 133.

A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine (2018). Stochastic Adversarial Video Prediction, *CoRR*, vol. abs/1804.01523. Cited on pages 25 and 142.

K. Lee, J. Kim, S. Chong, and J. Shin (2017a). Simplified Stochastic Feedforward Neural Networks, *CoRR*, vol. abs/1704.03188. Cited on page 73.

N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker (2017b). DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents, in *CVPR 2017*. Cited on pages 3, 6, 10, 26, 67, 68, 72, 73, 74, 75, 78, 81, 83, 84, 87, 89, 93, 95, 96, 111, 121, 141, 143, 148, 149, 151, 155, 158, and 173.

A. Lerer, S. Gross, and R. Fergus (2016). Learning Physical Intuition of Block Towers by Example, in *ICML 2016*. Cited on pages 4, 19, and 21.

A. Lerner, Y. Chrysanthou, and D. Lischinski (2007). Crowds by Example, in *Computer Graphics Forum 2007*. Cited on page 30.

W. Li, A. Leonardis, J. Bohg, and M. Fritz (2019). Learning Manipulation under Physics Constraints with Visual Perception, *CoRR*, vol. abs/1904.09860. Cited on page 19.

W. Li, A. Leonardis, and M. Fritz (2017). Visual Stability Prediction for Robotic Manipulation, in *IEEE International Conference on Robotics and Automation (ICRA) 2017*. Cited on pages 3, 4, 19, and 21.

Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang (2018). Flow-Grounded Spatial-Temporal Video Prediction from Still Images, in *ECCV (9) 2018*. Cited on page 26.

X. Liang, L. Lee, W. Dai, and E. P. Xing (2017). Dual Motion GAN for Future-Flow Embedded Video Prediction, in *ICCV 2017*. Cited on pages 18 and 25.

Z. Lin, A. Khetan, G. C. Fanti, and S. Oh (2018). PacGAN: The power of two samples in generative adversarial networks, in *NeurIPS 2018*. Cited on page 24.

R. Liu, Y. Ge, C. L. Choi, X. Wang, and H. Li (2021). DivCo: Diverse Conditional Image Synthesis via Contrastive Generative Adversarial Network, in *CVPR 2021*. Cited on page 25.

S. Liu, T. Ullman, J. Tenenbaum, and E. S. Spelke (2017a). What's worth the effort: Ten-month-old infants infer the value of goals from the costs of actions, in *CogSci 2017*. Cited on page 19.

Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala (2017b). Video Frame Synthesis Using Deep Voxel Flow, in *ICCV 2017*. Cited on pages 18 and 32.

Y. Lu and B. Huang (2020). Structured Output Learning with Conditional Generative Flows, in *AAAI 2020*. Cited on pages 29, 90, and 94.

P. Luc, C. Couprie, Y. LeCun, and J. Verbeek (2018). Predicting Future Instance Segmentations by Forecasting Convolutional Features, *CoRR*, vol. abs/1803.11496. Cited on page 21.

P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun (2017). Predicting Deeper into the Future of Semantic Segmentation, in *ICCV 2017*. Cited on pages 21, 22, 59, 60, 65, 66, 67, and 68.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi (2019). Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse, in *NeurIPS 2019*. Cited on page 119.

Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha (2019). TrafficPredict: Trajectory Prediction for Heterogeneous Traffic-Agents, in *AAAI 2019*. Cited on pages 20, 30, 116, 120, and 174.

D. J. MacKay (1992). A practical Bayesian framework for backpropagation networks, *Neural computation*, vol. 4(3), pp. 448–472. Cited on pages 5 and 22.

R. Madaan, N. Gyde, S. Vemprala, M. Brown, K. Nagami, T. Taubner, E. Cristofalo, D. Scaramuzza, M. Schwager, and A. Kapoor (2019). AirSim Drone Racing Lab, in *Proceedings of Machine Learning Research 2019*. Cited on page 2.

P. Maes (1993). Modeling adaptive autonomous agents, *Artificial life*, vol. 1(1_2), pp. 135–162. Cited on page 1.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey (2016). Adversarial autoencoders, in *ICLR 2016*. Cited on pages 27, 100, 101, 104, and 106.

A. Malinin and M. J. F. Gales (2018). Predictive Uncertainty Estimation via Prior Networks, in *NeurIPS 2018*. Cited on page 23.

S. Malla, B. Dariush, and C. Choi (2020). TITAN: Future Forecast Using Action Priors, in *CVPR 2020*. Cited on page 30.

K. Mangalam, H. Girase, S. Agarwal, K. Lee, E. Adeli, J. Malik, and A. Gaidon (2020). It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction, in *ECCV 2020*. Cited on pages 4, 11, 26, 27, 73, 89, 112, 116, 117, 118, and 143.

K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. V. Gool (2018). Convolutional Oriented Boundaries: From Image Segmentation to High-Level Tasks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40(4), pp. 819–833. Cited on pages 40 and 42.

X. Mao, C. Shen, and Y. Yang (2016). Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections, *CoRR*, vol. abs/1606.08921. Cited on page 42.

M. Mathieu, C. Couprie, and Y. LeCun (2016). Deep multi-scale video prediction beyond mean square error. Cited on pages 5, 18, 21, 25, 32, 34, 35, 38, 39, 42, 43, 47, 155, and 156.

M. McCloskey (1983). Intuitive physics, *Scientific american*, vol. 248(4), pp. 122–131. Cited on page 19.

J. Menick and N. Kalchbrenner (2019). Generating High fidelity Images with subscale pixel Networks and Multidimensional Upscaling, in *ICLR 2019*. Cited on page 28.

L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger (2019). Occupancy Networks: Learning 3D Reconstruction in Function Space, in *CVPR 2019*. Cited on page 162.

V. Michalski, R. Memisevic, and K. R. Konda (2014). Modeling Deep Temporal Dependencies with Recurrent "Grammar Cells", in *NIPS 2014*. Cited on page 18.

M. Mirza and S. Osindero (2014). Conditional Generative Adversarial Nets, *CoRR*. Cited on page 25.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida (2018). Spectral Normalization for Generative Adversarial Networks, in *ICLR 2018*. Cited on pages 25, 103, 105, 106, 107, 108, and 109.

A. Mnih and D. J. Rezende (2016). Variational Inference for Monte Carlo Objectives, in *ICML 2016*. Cited on pages 73 and 75.

B. T. Morris and M. M. Trivedi (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach, *IEEE transactions on pattern analysis and machine intelligence*, vol. 33(11), pp. 2287–2301. Cited on page 47.

A. S. Mueller, J. B. Cicchino, and D. S. Zuby (2020). What humanlike errors do autonomous vehicles need to avoid to maximize safety?, *Journal of safety research*, vol. 75, pp. 310–318. Cited on page 3.

S. S. Nabavi, M. Rochan, and Y. Wang (2018). Future Semantic Segmentation with Convolutional LSTM, in *BMVC 2018*. Cited on pages 21, 22, 59, 60, 66, and 67.

E. T. Nalisnick and P. Smyth (2017). Stick-Breaking Variational Autoencoders, in *ICLR (Poster) 2017*. Cited on page 26.

R. M. Neal (2012). *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media. Cited on pages 5 and 22.

T. D. Nguyen, T. Le, H. Vu, and D. Q. Phung (2017). Dual Discriminator Generative Adversarial Nets, in *NIPS 2017*. Cited on page 24.

F. Nielsen and K. Sun (2016). Guaranteed Bounds on the Kullback-Leibler Divergence of Univariate Mixtures, *IEEE Signal Process. Lett.*, vol. 23(11), pp. 1543–1546. Cited on page 103.

D. A. Nix and A. S. Weigend (1994). Estimating the mean and variance of the target probability distribution, in *Neural Networks 1994*. Cited on page 22.

P. Ochs, J. Malik, and T. Brox (2014). Segmentation of Moving Objects by Long Term Video Analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36(6), pp. 1187–1200. Cited on page 32.

S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. G. Rodríguez, and A. A. Argyros (2020). A Review on Deep Learning Techniques for Video Prediction, *CoRR*, vol. abs/2004.05214. Cited on page 18.

I. Osband (2016). Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout, *NIPS Workshop on Bayesian Deep Learning*. Cited on pages 10, 60, and 63.

G. Ostrovski, W. Dabney, and R. Munos (2018). Autoregressive Quantile Networks for Generative Modeling, in *ICML 2018*. Cited on pages 138 and 174.

E. Pajouheshgar and C. H. Lampert (2018). Back to square one: probabilistic trajectory forecasting without bells and whistles, in *NeurIPS Workshop 2018*. Cited on pages 73, 87, 89, 93, 95, 96, 97, 143, 148, 149, and 169.

G. Papamakarios, I. Murray, and T. Pavlakou (2017). Masked Autoregressive Flow for Density Estimation, in *NeurIPS 2017*. Cited on pages 28 and 129.

N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran (2018). Image Transformer, in *ICML 2018*. Cited on page 28.

V. Patraucean, A. Handa, and R. Cipolla (2015). Spatio-temporal video autoencoder with differentiable memory, *CoRR*, vol. abs/1511.06309. Cited on page 18.

S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool (2009). You'll never walk alone: Modeling social behavior for multi-target tracking, in *ICCV 2009*. Cited on page 30.

J. Pirhonen, H. Melkas, A. Laitinen, and S. Pekkarinen (2020). Could robots strengthen the sense of autonomy of older people residing in assisted living facilities? - A future-oriented study, *Ethics Inf. Technol.*, vol. 22(2), pp. 151–162. Cited on page 3.

D. A. Pomerleau (1989). ALVINN, an autonomous land vehicle in a neural network, Technical report, Carnegie Mellon University, Computer Science Department. Cited on page 47.

P. Porwik and A. Lisowska (2004). The Haar-wavelet transform in digital image processing: its status and achievements, *Machine graphics and vision*, vol. 13(1/2), pp. 79–98. Cited on page 145.

A. Prakash, A. Behl, E. Ohn-Bar, K. Chitta, and A. Geiger (2020). Exploring Data Aggregation in Policy Learning for Vision-Based Urban Autonomous Driving, in *CVPR 2020*. Cited on page 1.

S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe (2020). BayesFlow: Learning complex stochastic models with invertible neural networks, *CoRR*, vol. abs/2003.06281. Cited on page 163.

A. Radford, L. Metz, and S. Chintala (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, in *ICLR 2016*. Cited on pages 24, 106, 107, 108, 127, and 138.

M. Raibert, K. Blankespoor, G. Nelson, and R. Playter (2008). Bigdog, the rough-terrain quadruped robot, *IFAC Proceedings Volumes*, vol. 41(2), pp. 10822–10825. Cited on page 2.

T. Raiko, M. Berglund, G. Alain, and L. Dinh (2015). Techniques for Learning Binary Stochastic Feedforward Neural Networks, in *ICLR (Poster) 2015*. Cited on page 73.

M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra (2014). Video (language) modeling: a baseline for generative models of natural videos, *CoRR*, vol. abs/1412.6604. Cited on pages 18 and 33.

A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos (2019). PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction, in *ICCV 2019*. Cited on page 30.

S. V. Ravuri, S. Mohamed, M. Rosca, and O. Vinyals (2018). Learning Implicit Generative Models with the Method of Learned Moments, vol. 80, pp. 4311–4320. Cited on page 25.

A. Razavi, A. van den Oord, B. Poole, and O. Vinyals (2019a). Preventing Posterior Collapse with delta-VAEs, in *ICLR (Poster) 2019*. Cited on pages 11, 88, and 89.

A. Razavi, A. van den Oord, and O. Vinyals (2019b). Generating Diverse High-Fidelity Images with VQ-VAE-2, in *NeurIPS 2019*. Cited on pages 125, 127, 129, and 163.

S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas (2017). Parallel Multiscale Autoregressive Density Estimation, in *ICML 2017*. Cited on page 28.

E. Rehder and H. Kloeden (2015). Goal-Directed Pedestrian Prediction, in *ICCV Workshops 2015*. Cited on page 47.

J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid (2015). EpicFlow: Edge-preserving interpolation of correspondences for optical flow, in *CVPR 2015*. Cited on page 37.

D. J. Rezende and S. Mohamed (2015). Variational Inference with Normalizing Flows, in *ICML 2015*. Cited on page 26.

D. J. Rezende, S. Mohamed, and D. Wierstra (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models, in *ICML 2014*. Cited on pages 26 and 126.

N. Rhinehart, K. M. Kitani, and P. Vernaza (2018). R2P2: A ReparameteRized Pushforward Policy for Diverse, Precise Generative Path Forecasting, in *ECCV 2018*. Cited on pages 27 and 116.

A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese (2016). Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes, in *ECCV (8) 2016*. Cited on pages 20, 47, 80, 83, 95, 97, 148, 151, 158, and 160.

M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed (2017). Variational Approaches for Auto-Encoding Generative Adversarial Networks, *CoRR*, vol. abs/1706.04987. Cited on pages 7, 10, 11, 27, 60, 63, 90, 95, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, and 110.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, vol. 115(13), pp. 211–252. Cited on page 127.

A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese (2019). SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints, in *CVPR 2019*. Cited on pages 4, 25, 73, 89, 95, 97, 119, 120, 121, 122, 141, 143, 150, 151, and 174.

A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese (2018). CAR-Net: Clairvoyant Attentive Recurrent Network, in *ECCV (11) 2018*. Cited on pages 20, 95, 97, and 151.

A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang (2020). Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning, in *USENIX Security Symposium 2020*. Cited on page 13.

T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen (2016). Improved Techniques for Training GANs, in *NIPS 2016*. Cited on pages 6, 12, 127, 137, and 160.

T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma (2017). PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications, in *ICLR (Poster) 2017*. Cited on pages 7, 28, 136, and 137.

T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone (2020). Trajectron++: Multi-Agent Generative Trajectory Forecasting With Heterogeneous Data for Control, in *ECCV 2020*. Cited on pages 4, 11, 26, 27, 112, 116, 117, 118, 119, 120, 121, 122, 123, 155, and 174.

E. Santana and G. Hotz (2016). Learning a Driving Simulator, *CoRR*, vol. abs/1608.01230. Cited on page 47.

J. Saric, S. Vrazic, and S. Segvic (2021). Joint Forecasting of Features and Feature Motion for Dense Semantic Future Prediction, *CoRR*, vol. abs/2101.10777. Cited on page 21.

A. Sauer, N. Savinov, and A. Geiger (2018). Conditional Affordance Learning for Driving in Urban Environments, in *CoRL 2018*. Cited on page 1.

X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, in *NIPS 2015*. Cited on pages 72, 73, 74, 79, 80, 82, 84, 85, and 158.

X. Shi, Z. Gao, L. Lausen, H. Wang, D. Yeung, W. Wong, and W. Woo (2017). Deep Learning for Precipitation Nowcasting: A Benchmark and a New Model, in *NeurIPS 2017*. Cited on pages 11, 74, and 131.

R. Shu, H. H. Bui, S. Zhao, M. J. Kochenderfer, and S. Ermon (2018). Amortized Inference Regularization, pp. 4398–4407. Cited on page 100.

B. Smith and R. Casati (1994). Naive physics, *Philosophical psychology*, vol. 7(2), pp. 227–247. Cited on page 19.

K. Sohn, H. Lee, and X. Yan (2015). Learning Structured Output Representation using Deep Conditional Generative Models, in *NeurIPS 2015*. Cited on pages 10, 26, 67, 69, 72, 73, 74, 75, 87, 89, 95, 117, and 118.

K. Soomro, A. R. Zamir, and M. Shah (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, vol. abs/1212.0402. Cited on page 156.

A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton (2017). VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning, in *NIPS 2017*. Cited on pages 6, 24, 104, 105, 107, and 108.

I. Sutskever, G. E. Hinton, and G. W. Taylor (2008). The Recurrent Temporal Restricted Boltzmann Machine, in *NIPS 2008*. Cited on page 18.

I. Sutskever, O. Vinyals, and Q. V. Le (2014). Sequence to Sequence Learning with Neural Networks, in *NIPS 2014*. Cited on page 73.

E. G. Tabak, E. Vanden-Eijnden, *et al.* (2010). Density estimation by dual ascent of the log-likelihood, in *Communications in Mathematical Sciences 2010*. Cited on page 90.

J. Tabor, S. Knop, P. Spurek, I. T. Podolak, M. Mazur, and S. Jastrzebski (2018). Cramer-Wold AutoEncoder, *CoRR*, vol. abs/1805.09235. Cited on pages 12 and 100.

S. Tang, B. Andres, M. Andriluka, and B. Schiele (2016). Multi-person Tracking by Multicut and Deep Matching, in *ECCV Workshops (2) 2016*. Cited on page 54.

Y. Tang and R. Salakhutdinov (2013). Learning Stochastic Feedforward Neural Networks, in *NIPS 2013*. Cited on page 73.

L. Theis, A. van den Oord, and M. Bethge (2016). A note on the evaluation of generative models, in *ICLR 2016*. Cited on page 138.

L. A. Thiede and P. P. Brahma (2019). Analyzing the Variety Loss in the Context of Probabilistic Trajectory Prediction, in *ICCV 2019*. Cited on page 120.

S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak (2019). On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks, in *NeurIPS 2019*. Cited on page 23.

I. O. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf (2018). Wasserstein Auto-Encoders, in *ICLR 2018*. Cited on pages 27, 90, 101, 102, 108, and 109.

J. M. Tomczak and M. Welling (2016). Improving variational auto-encoders using householder flow, in *NIPS Workshop 2016*. Cited on pages 26 and 90.

J. M. Tomczak and M. Welling (2018). VAE with a VampPrior, in *AISTATS 2018*. Cited on pages 11, 26, 88, 119, and 127.

N. Tran, T. Bui, and N. Cheung (2018). Dist-GAN: An Improved GAN Using Distance Constraints, in *ECCV (14) 2018*. Cited on pages 25, 106, 107, 108, and 109.

P. Trautman, J. Ma, R. M. Murray, and A. Krause (2013). Robot navigation in dense human crowds: the case for cooperation, in *ICRA 2013*. Cited on page 47.

R. van den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling (2018). Sylvester Normalizing Flows for Variational Inference, in *UAI 2018*. Cited on page 26.

A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves (2016a). Conditional Image Generation with PixelCNN Decoders, in *NIPS 2016*. Cited on pages 28, 126, 129, and 134.

A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu (2016b). Pixel Recurrent Neural Networks, in *ICML 2016*. Cited on pages 7, 28, 125, 126, 129, 133, 134, and 138.

A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis (2018). Parallel WaveNet: Fast High-Fidelity Speech Synthesis, in *ICML 2018*. Cited on pages 28 and 126.

R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee (2017). Learning to Generate Long-term Future via Hierarchical Prediction, in *ICML 2017*. Cited on pages 18, 42, and 161.

L. Wang, A. G. Schwing, and S. Lazebnik (2017). Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space, in *NIPS 2017*. Cited on pages 11, 88, and 90.

P. Z. Wang and W. Y. Wang (2019). Riemannian Normalizing Flow on Variational Wasserstein Autoencoder for Text Modeling, in *NAACL-HLT (1) 2019*. Cited on page 89.

Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. E. Campbell, K. Q. Weinberger, and W. Chao (2020). Train in Germany, Test in the USA: Making 3D Object Detectors Generalize, in *CVPR 2020*. Cited on page 165.

Z. Wang, S. Rosa, Y. Miao, Z. Lai, L. Xie, A. Markham, and N. Trigoni (2018). Neural Allocentric Intuitive Physics Prediction from Real Videos, *CoRR*, vol. abs/1809.03330. Cited on page 19.

D. Warde-Farley and Y. Bengio (2017). Improving Generative Adversarial Networks with Denoising Feature Matching, in *ICLR (Poster) 2017*. Cited on page 24.

N. Watters, D. Zoran, T. Weber, P. W. Battaglia, R. Pascanu, and A. Tacchetti (2017). Visual Interaction Networks: Learning a Physics Simulator from Video, in *NIPS 2017*. Cited on pages 4 and 155.

X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang (2018). Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect, in *ICLR 2018*. Cited on pages 127 and 138.

M. Wertheimer (1923). Laws of organization in perceptual forms, *A source book of Gestalt Psychology*. Cited on page 32.

S. N. Wood (2010). Statistical inference for noisy nonlinear ecological dynamic systems, *Nature*, vol. 466(7310), p. 1102. Cited on pages 10, 60, 63, 101, and 102.

C. Xiao, P. Zhong, and C. Zheng (2018). BourGAN: Generative Networks with Metric Embeddings, in *NeurIPS 2018*. Cited on page 24.

J. Xie, M. Kiefel, M. Sun, and A. Geiger (2016). Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer, in *CVPR 2016*. Cited on pages 116 and 174.

H. Xu, Y. Gao, F. Yu, and T. Darrell (2017). End-to-End Learning of Driving Models from Large-Scale Video Datasets, pp. 3530–3538. Cited on pages 47, 52, and 72.

J. Xu, B. Ni, Z. Li, S. Cheng, and X. Yang (2018). Structure Preserving Video Prediction, in *CVPR 2018*. Cited on page 18.

T. Xue, J. Wu, K. L. Bouman, and B. Freeman (2016). Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks, in *NIPS 2016*. Cited on pages 26 and 69.

K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg (2011). Who are you with and where are you going?, in *CVPR 2011*. Cited on pages 20 and 47.

C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin (2018). Pose Guided Human Video Generation, in *ECCV (10) 2018*. Cited on page 18.

D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee (2019). Diversity-Sensitive Conditional Generative Adversarial Networks, in *ICLR (Poster) 2019*. Cited on page 25.

Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick (2017). Improved Variational Autoencoders for Text Modeling using Dilated Convolutions, in *ICML 2017*. Cited on page 89.

J. J. Yu, K. G. Derpanis, and M. A. Brubaker (2020a). Wavelet Flow: Fast Training of High Resolution Normalizing Flows, in *NeurIPS 2020*. Cited on page 28.

W. Yu, Y. Lu, S. Easterbrook, and S. Fidler (2020b). Efficient and Information-Preserving Future Frame Prediction and Beyond, in *ICLR 2020*. Cited on pages 3 and 4.

Y. Yuan and K. M. Kitani (2020). Diverse Trajectory Forecasting with Determinantal Point Processes. Cited on page 27.

A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese (2018). Taskonomy: Disentangling Task Transfer Learning, in *CVPR 2018*. Cited on page 164.

H. Zhang, A. Geiger, and R. Urtasun (2013). Understanding High-Level Semantics by Modeling Traffic Patterns, in *ICCV 2013*. Cited on page 47.

L. Zhang, Q. She, and P. Guo (2019). Stochastic trajectory prediction with social graph network, *CoRR*. Cited on pages 26 and 27.

S. Zhang, R. Benenson, and B. Schiele (2017). CityPersons: A Diverse Dataset for Pedestrian Detection, pp. 4457–4465. Cited on page 54.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia (2017a). Pyramid Scene Parsing Network, in *CVPR 2017*. Cited on page 67.

J. J. Zhao, M. Mathieu, and Y. LeCun (2017b). Energy-based Generative Adversarial Networks, in *ICLR (Poster) 2017*. Cited on page 24.

S. Zhao, J. Song, and S. Ermon (2017c). InfoVAE: Information Maximizing Variational Autoencoders. Cited on pages 89 and 101.

T. Zhao, Y. Xu, M. Monfort, W. Choi, C. L. Baker, Y. Zhao, Y. Wang, and Y. N. Wu (2019). Multi-Agent Tensor Fusion for Contextual Trajectory Prediction, in *CVPR 2019*. Cited on pages 20, 73, 89, 97, 143, 150, and 174.

B. Zhou, X. Wang, and X. Tang (2011). Random field topic model for semantic region analysis in crowded scenes from tracklets, in *CVPR 2011*. Cited on page 47.

Z. M. Ziegler and A. M. Rush (2019). Latent Normalizing Flows for Discrete Sequences, in *ICML 2019*. Cited on pages 26, 88, 90, 91, 127, 128, 144, 149, and 151.