MUSE-VAE: Multi-Scale VAE for Environment-Aware Long Term Trajectory Prediction

Mihee Lee
Rutgers University
ml1323@rutgers.edu

Samuel S. Sohn
Rutgers University
samuel.sohn@rutgers.edu

Seonghyeon Moon
Rutgers University
sm2062@cs.rutgers.edu

Sejong Yoon
The College of New Jersey
yoons@tcnj.edu

Mubbasir Kapadia
Rutgers University
mubbasir.kapadia@rutgers.edu

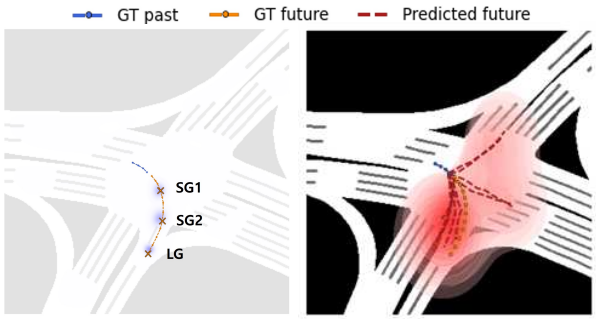
Vladimir Pavlovic
Rutgers University
vladimir@cs.rutgers.edu

Abstract

Accurate long-term trajectory prediction in complex scenes, where multiple agents (e.g., pedestrians or vehicles) interact with each other and the environment while attempting to accomplish diverse and often unknown goals, is a challenging stochastic forecasting problem. In this work, we propose MUSE-VAE, a new probabilistic modeling framework based on a cascade of Conditional VAEs, which tackles the long-term, uncertain trajectory prediction task using a coarse-to-fine multi-factor forecasting architecture. In its Macro stage, the model learns a joint pixel-space representation of two key factors, the underlying environment and the agent movements, to predict the long and short term motion goals. Conditioned on them, the Micro stage learns a fine-grained spatio-temporal representation for the prediction of individual agent trajectories. The VAE backbones across the two stages make it possible to naturally account for the joint uncertainty at both levels of granularity. As a result, MUSE-VAE offers diverse and simultaneously more accurate predictions compared to the current state-of-the-art. We demonstrate these assertions through a comprehensive set of experiments on nuScenes and SDD benchmarks as well as PFSD, a new synthetic dataset, which challenges the forecasting ability of models on complex agent-environment interaction scenarios.

1. Introduction

Human behavior forecasting is an essential problem studied in various research fields such as computer vision [14], computer graphics [15], robotics [10], and cognitive science [44]. The fundamental problem with predicting human motion is the inherent stochasticity stemming from the fact that human beings use numerous sources of information to make a wide variety of different decisions at any given moment, which all impact their future movement.



(a) Intermediate Short-term goals (b) Full trajectories

Figure 1. (a) The predicted trajectory heatmaps are overlaid in the semantic map. Ground Truth (GT) long-term goal (LG) and short-term goals (SG1 and SG2) are marked with 'x'. (b) Complete trajectory forecasting based on the predicted LG and SG. Each sequence of trajectories is obtained from a different pair of LG and SG predictions.

This movement uncertainty translates beyond the motion of the humans alone to the movement of objects controlled by humans, such as vehicles [6].

To embrace the uncertainty, in this paper, we focus on developing computational models, learned from data, that can predict a realistic multi-modal distribution of the future agent (humans, vehicles, etc.) trajectories. The models are designed in the context of two main factors that drive this uncertainty: the environment the agents occupy and the task they are attempting to accomplish.

However, direct forecasting of long-term trajectories is a challenging task. A person typically plans one's movement in a coarse-to-fine fashion: with a final destination in mind, through a sequence of intermediate goals or way-points, the movement is executed to reach those subgoals [8, 34]. State-of-the-Art (SOTA) methods [25, 43, 46] leverage this intuition to propose goal-conditioned prediction model. However, despite their effectiveness compared to traditional approaches [1, 14, 42], these models show limited ability to deal with complex environments [43], par-

ticularly as they affect the movement [46]. This often results in physically implausible trajectory predictions that violate agent-environment collision constraints. Moreover, the models frequently struggle to account for the diversity of the forecast goals and trajectories [25], which are driven by the uncertain, multi-modal nature of the problem.

To address this, we propose MUSE-VAE: a multi-scale, environment-aware model for long-term trajectory prediction which (1) takes a stage-wise, coarse-to-fine approach to trajectory prediction by predicting both the higher-level goals and the goal-conditioned trajectory, (2) avoids collision with obstacles without loss of spatial signal which can occur due to spatial reorganization when compressing 2D information into 1D features, and (3) learns a multi-modal predictive distribution across the stages, thus capturing the inherent uncertainty. MUSE-VAE embodies a three-step learning strategy across a Macro-stage and Micro-stage. The Macro-stage comprises of two steps for coarse predictions. We first predict the long-term goal, i.e., the last step of the given sequence based on heatmap trajectory representation. Given the long-term goal, sequential short-term goals are predicted as shown in Fig. 1a. After getting the goal positions in the Macro-stages, finally, our model produces the full trajectories in the Micro-stage as in Fig. 1b. Our main contributions are as follows: (**a**) We introduce a novel multi-scale learning strategy for CVAE-based probabilistic models in order to make environment-aware collision-free trajectory predictions. (**b**) Unlike the prior works, we show that one can learn trajectory distributions that can be well generalized in new scenes at test time, giving various reasonable predictions compliant to the environment without needing extra steps for diversity. (**c**) The proposed coarse-to-fine approach enables diverse and accurate trajectory predictions by forecasting the heading of the entire trajectories through goal prediction and then expanding it to granular and complete predictions.

We demonstrate these contributions through experiments on both real and synthetic dataset. With various grounded evaluation metrics, we show that MUSE-VAE can produce predictions similar to GT trajectories while achieving less collision with the environment than the SOTA methods.

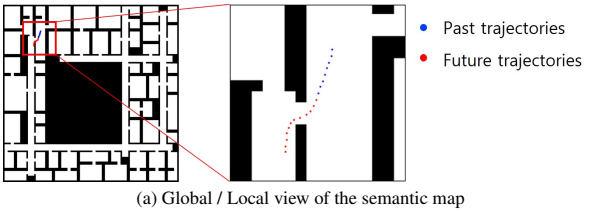
2. Related Work

The modeling of agent movement behavior, including individual humans, crowds, vehicles, etc., is a long-standing problem crossing the boundaries of multi-agent and computer vision communities. We focus on three relevant aspects: the forecasting of individual trajectories, the interplay between movement behavior and the environment, and the need for modeling thet uncertainty in motion prediction.

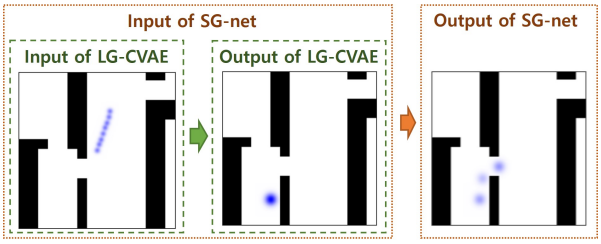
Sequence Learning The human trajectory has a sequence characteristic that changes in turn according to the passage of time. In order to capture the nature of the sequential information, many prior works [1, 14, 21, 32, 33, 42] utilize Recurrent Neural Networks (RNNs) [27] such as LSTMs and GRUs. However, RNN suffers from forgetting the past hidden states as the recursion goes. [12, 45] tackle the temporal aspects of human trajectory forecasting by adopting Transformer Networks [41]. Transformer solves the long-range dependency problem by processing the a sequence as a whole with self-attention and positional encoding. Y-net [25] solves the sequential trajectory learning problem with only convolution layers. They represent trajectories with multiple heatmaps, which are stacked with the semantic environment map image along the channel dimension and fed to their convolution networks as a whole. This way, they learn temporal movements with the environment without tradition sequence learning networks.

Environment Learning A decision about the trajectory taken towards a goal depends on the surrounding environment. Many prior approaches provide environmental information to their model for realistic trajectory predictions. [32,33,45,46] encode the environment layout and semantics as a representation of the scene image with a convolution network and use it to train their models along with trajectory features. While these approaches can learn the scene context surrounding the trajectory, they compress it into 1D feature vectors after CNNs and FCs layers, which can convey corrupted information in terms of spatial signals. Y-net [25] addresses this issue by aligning the semantic map with the trajectory heatmap spatially and processing them as a whole. Our model attempts more meaningful environmental learning without unnecessary information by focusing on a limited area around the trajectory rather than the entire scene while keeping the spatial signal by utilizing the heatmap trajectory representation.

Multimodal Learning The trajectory of an agent (human, vehicle, etc.) is affected by a number of factors such as the destination in mind, the surrounding environment, nearby agents and so on, which leads to an intrinsic uncertainty about the future behavior. Recent studies focus on learning the *distribution* of the human trajectory based on deep generative models, sidestepping the deterministic trajectory prediction. [17, 21, 33, 40, 45] adopt Conditional Variational AutoEncoders (CVAE) [35] and [14, 20, 32] introduce Generative Adversarial Network (GAN) [13] for learning of trajectory distribution where multiple predictions can be sampled. Trajectron++ [33] tackles the multimodal aspect of trajectory distributions by adopting a discrete latent distribution for the latent space, and Gaussian Mixture Model as the output distribution of the decoder in their CVAE framework. AgentFormer [45] promotes diversity of the predictions with a pairwise distance loss across predictions. However, this approach requires retraining whenever a different number of predictions are sought at test time. Y-net [25] utilizes K-means clustering of predictive discrete density maps at test time to achieve diverse



(a) Global / Local view of the semantic map



(b) Input / Output of the Macro-stage models

Figure 2. (a) The semantic map with 8 past / 12 future trajectories. Rather than the global map, we use the local map to focus on the nearby environment of the given trajectories. (b) Input and output format of Macro-stage models, LG-CVAE and SG-net. Trajectory heatmaps are overlaid with the local view semantic map. Here we assume 2 short-term goals at future time step 4 and 8 among 12 future steps. Thus, SG-net outputs 3 heatmaps; 2 for the short-term goals and 1 for the long-term goal.

prediction; however, the model does not explicitly learn the resolution-free multimodal trajectory density. Some prior works [25, 28, 43, 46] encourage the multimodality by proposing a goal-conditioned forecasting model under the assumption that one's movement depends primarily on the final goal position.

MUSE−VAE adopts a stage-wise training procedure to incorporate sequential information while maintaining a trajectory aligned with the environment. First, in the Macro-stage, future predictions are obtained by utilizing the heatmap representation of trajectories along with the semantic environment map, and then in the Micro-stage, RNN-based networks are used to facilitate sequence learning. The Micro-stage takes advantage of coarse predictions from the Macro-stage, reducing the long-range dependency problem and guiding the path to avoid obstacles. Adopting VAE in both Macro- and Micro-stages, our model learns the inherent uncertainty of forecasting, which can give a variety of plausible predictions.

3. Proposed Method

The trajectory prediction problem is formulated as follows. Assume that we are given $t_p > 0$ timestamps, the past trajectory positions $x = \{x_i^t\}_{t=1}^{t_p}$ of agent i in scene S, where $x_i^t \in \mathbb{R}^2$ denotes the 2D world coordinates of the agent i at time t. Our goal is to predict the future trajectory of the same agent during $t_f > 0$ future timestamps, $y = \{y_i^t\}_{t=t_{p+1}}^{t_{p+f}}$ in the sense of their distribution. $y_i^t \in \mathbb{R}^2$ is the future 2D position in the same coordinate system as x_i^t. This prediction should take into account the environmental context S, i.e., $p(y|x, S)$. We propose our MUlti-Scale Environment-aware model, MUSE−VAE for coarse-to-fine trajectory forecasting. The Macro-stage is defined as a coarse prediction of the future trajectories, and the Micro-stage is defined as a fine prediction based on the coarse prediction. In the Macro-stage, only a subset of the future steps are predicted as the long-term and short-term goals. We denote the long-term goal as the final step at $t_{LG} = t_{p+f}$ and the short-term goals as some intermediate steps $t_{SG} \in \{t_{p+1}, \ldots, t_{p+f-1}\}$. The Macro-stage aims to obtain rough predictions that are well aligned with the scene for collision avoidance against environmental obstacles. Based on the coarse prediction, the Micro-stage generates a fine-grained prediction of all t_f future steps. In this stage, we adopt the RNN [27] to efficiently learn the sequential features of trajectories.

In Sec. 3.1, we introduce the coarse prediction stage, Macro-stage, and elaborate on how the primary Macro-stage model, Long-term Goal Conditional VAE (LG-CVAE), and the subsequent Macro-stage model, Short-term Goal network (SG-net), are formulated. Sec. 3.2 introduces the Micro-stage, the fine prediction stage, used to refine predictions of complete forecast trajectories.

3.1. Macro-stage: Coarse Prediction Stage

One of the most important factors in the uncertainty of the future behavior is the future heading of an individual. One way to narrow the possibilities is to be aware of the surroundings and learn patterns from the past. [33, 45, 46] learn a representation of the environment, defined in image space, by encoding the semantic map of the scene into a 1D flattened feature, which can introduce distortion of spatial information of the scene. For alignment between trajectories and the semantic map, we represent trajectories x in the pixel space as suggested in Y-net [25], using a Gaussian heatmap, denoted by I_x. The Gaussian filter has a variance of 4, and we create the homography matrices to map the world coordinates in meters to the image-based coordinates in pixel. Trajectories in t_p past timestamp are all represented in a single heatmap, while each future step is represented as one heatmap per step. The trajectory heatmap size matches the size of the semantic map.

Typically, the full environment information of a given scene is not necessary for long-term trajectory prediction. Often, the scene proximal to an agent's current location is sufficient. Thus, we focus only on the local semantic map, with trajectory heatmaps created as illustrated in Fig. 2a. The local map is centered at the last observed agent location. The inputs and outputs of the Macro-stage are illustrated in Fig. 2b. The input of the long-term goal prediction model, LG-CVAE, consists of concatenated (local semantic map, past trajectory heatmap) and outputs one long-term goal heatmap. The short-term goal prediction model, SG-

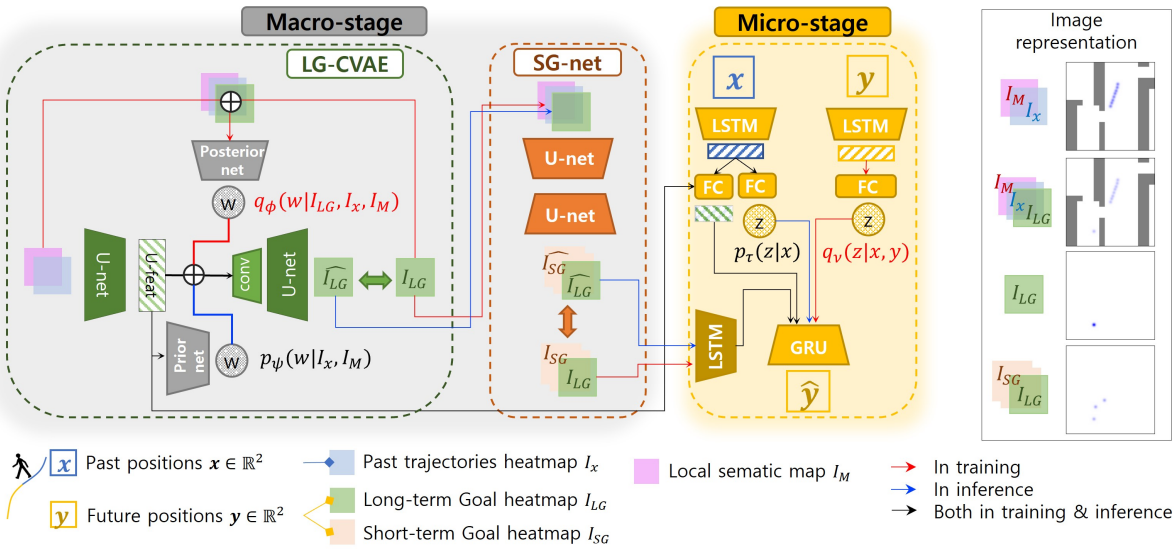


Figure 3. MUSE-VAE architecture. LG-CVAE is the first stage which predicts the long-term goal based on CVAE framework. Conditioned on the long-term goal, SG-net predicts the waypoints from the past trajectories to the long-term goal. We group these two stages as Macro-stage where the predictions are made in heatmap representation to keep the spatial signal along with the semantic map. Finally in Micro-stage, full trajectories are obtained with RNN-based CVAE. More implementation details are in the Supplementary Materials.

net, has the input of concatenated (local semantic map, past trajectory heatmap, long-term goal heatmap) and outputs $N_{SG}+1$ heatmaps, where N_{SG} is the number of short-term goals[1]. The local semantic map I_M can be determined as $f(S, x_i^{t_p}, \mathcal{H}, n)$ where f is the function that converts the global scene information S and homography $\mathcal{H}$ into a local image-based representation of size (n, n) pixels centered at the last observed location $x_i^{t_p}$ of agent i.

3.1.1 LG-CVAE: Long-term Goal Prediction Model

Where a person will go in the future depends primarily on the long-term goal position. Therefore, for different potential future trajectories, it is of paramount importance to predict different long-term goal positions in good quality. To model the inherent uncertainty with semantic map and heatmap trajectory representations, we combine U-net [30] and Conditional Variational AutoEncoder (CVAE) [35] as studied in [19]. Given the heatmap I_x of the past trajectories, the heatmap I_{LG} of the long-term goal, and the local semantic map I_M, the objective of the CVAE is to maximize the conditional distribution,

$$p(I_{LG}|I_x, I_M) = \int p_\theta(I_{LG}|w, I_x, I_M)p(w|I_x, I_M)dw. \quad (1)$$

The stochasticity of the conditional latent distribution $p(w|I_x, I_M)$ is propagated and contributes to the multi-modality of $p(I_{LG}|I_x, I_M)$. The LG-CVAE loss is defined as the negative evidence lower bound as follows.

$$\begin{aligned}\mathcal{L}_{I_{LG}} = &-\mathbb{E}_{q_\phi(w|I_{LG},I_x,I_M)}\left[\log p_\theta(I_{LG}|w, I_x, I_M)\right] \\ &+KL(q_\phi(w|I_{LG}, I_x, I_M)||p_\psi(w|I_x, I_M)),\end{aligned} \quad (2)$$

[1]The extra count corresponds to the long-term goal.

where $q_\phi(w|I_{LG}, I_x, I_M)$ and $p_\psi(w|I_x, I_M)$ are the posterior and the conditional prior distributions, respectively, assumed to be Gaussian for tractability. The output trajectory heatmap distribution $p_\theta(I_{LG}|w, I_x, I_M)$ has a Bernoulli distribution. Parameters of those densities are modeled using deep neural networks with the learning parameters ϕ, ψ, and θ, respectively, see Fig. 3. We use focal loss between the predicted heatmap $\widehat{I_{LG}}$ and the Ground Truth (GT) heatmap I_{LG} for the reconstruction loss to mitigate the imbalanced class issue in the trajectory heatmap representation.

Joint pixel-based environment-trajectory input (I_M, I_x) is encoded using a U-net architecture backbone [30], which shows excellent performance on semantic segmentation learning. The encoded U-net features of dimension (C, H, W), where the feature map has C channels, a height of H, and a width of W, are average-pooled in the spatial dimension, and outputs $(C, 1, 1)$ feature maps, which are eventually converted to a C-dimensional vector. It is concatenated with the latent factor w sampled from the latent distribution. The posterior and the prior latent distributions are obtained from the separated posterior and prior network respectively consisting of convolutional layers.

To avoid the posterior collapse [4,39] stemming from the strong U-net decoder, we pretrain the encoders and apply Free Bits [18] and KL annealing [5] strategies as studied in [22]. Additional implementation details are discussed in the Supplementary Materials.

3.1.2 SG-net: Short-term Goal Prediction Model

In the second stage of the Macro-stage, we predict the short-term goals based on the long-term goal prediction from LG-

CVAE. The purpose of SG-net is to give waypoints from the last observed step to the long-term goal that are well-aligned with the environment. The final stage in Sec. 3.2 Micro-stage processes the trajectory and the semantic map as 1D feature vectors separately. Therefore, predicting all fine-grained future steps using only long-term goal information increases the risk of making predictions that are not well aligned with the environment based on destroyed spatial signals. SG-net utilizes U-net to generate $N_{SG}+1$ heatmaps where N_{SG} is the number of short-term goals and 1 accounts for the long-term goal as illustrated in Fig. 2b. Unlike the LG-CVAE, this stage outputs the deterministic prediction based on the predicted long-term goal since we deal with the uncertainty of the fine trajectories other than long-term goals in the next stage. Thus, SG-net loss is simply reconstruction loss with focal loss as follows.

$$\begin{aligned}\mathcal{L}_{SG} = -\sum_{i=1}^{N_{SG}+1} \big(&\alpha(1-\widehat{I_{SGi}})^{\gamma} I_{SGi} log(\widehat{I_{SGi}}) \\ &+(1-\alpha)\widehat{I_{SGi}}^{\gamma}(1-I_{SGi})log(1-\widehat{I_{SGi}})\big),\end{aligned} \tag{3}$$

where I_{SG} is the GT trajectory heatmap and $\widehat{I_{SG}}$ is the predicted heatmap and $\alpha = 0.25, \gamma = 2$ as studied in [23].

3.2. Micro-stage: Fine Prediction Stage

In the final stage of our model, we predict complete future trajectories at the micro level. Here we change the coordinate from the discrete pixel coordinate to continuous world coordinate for fine predictions. Even if guided by predicted long-term and short-term goals from SG-net, individual steps may also have the variability stemming from the surrounding environment. To deal with this uncertainty, we leverage CVAE in this step as well. As illustrated in Fig. 3, we set $p(z|x)$ as the prior conditioned on past trajectories x, which is learned to approximate the posterior latent distribution $p(z|x,y)$ where y denotes the future trajectories In test time, we sample the latent factor z from $p(z|x)$ to predict $p(y|z,x)$. While decoding future steps, our model use the long-term and short-term goal information from SG-net in the form of LSTM-encoded features. We apply the Teacher Forcing technique to correct the prediction by feeding the GT/predicted long-term and short-term goals during training/test time respectively. To reduce the gap between training and test time reconstructions, we provide an additional reconstruction loss from the prior distribution following [7, 36]. Thus, Micro-stage training loss with β-weighted ELBO [16] is formulated as follows.

$$\begin{aligned}&\mathcal{L}_{Micro} = -\mathbb{E}_{q_v(z|x,y)}\left[\log p_\eta(y|z,x)\right] \\ &-\mathbb{E}_{p_\tau(z|x)}\left[\log p_\eta(y|z,x)\right] + \beta KL(q_v(z|x,y)||p_\tau(z|x)),\end{aligned} \tag{4}$$

where both the latent distributions and the output trajectory distribution are assumed as Gaussian distributions. We feed the U-net features from LG-CVAE to the prior network of Micro-stage so that the Micro-stage also recognizes the environment.

4. Experiments

Sec. 4.1 introduces the datasets, evaluation metrics, and statistical analysis used in the experiments. Sec. 4.2 quantitatively evaluates SOTA models as well as MUSE-VAE. Sec. 4.3 compares the qualitative aspects of the predictions for intuitive assessment. In Sec. 4.4, each component of MUSE-VAE is analyzed by ablation studies.

4.1. Preliminaries

Datasets We used three datasets for the evaluation. The Stanford Drone Dataset (**SDD**) [29] is used in the TrajNet challenge [31] and prior works [25, 32]. The **nuScenes** Dataset [6] is a public autonomous driving dataset used by many prior arts [24, 26, 45]. In addition, we created a new Path Finding Simulation Dataset (**PFSD**) using environments borrowed from [38]. Unlike SDD and nuScenes, the spaces in PFSD are more complex to navigate. For more details, please refer to the Supplementary Materials.

Evaluation Metrics For the evaluation, we adopted the standard metrics of minimum Average Displacement Error (**ADE**) and Final Displacement Error (**FDE**). We also report the Kernel Density Estimate-based Negative Log Likelihood (**KDE NLL**) used in [17, 33] as a comprehensive indicator of the predictive performance. Finally, we assess the Environment Collision-Free Likelihood (**ECFL**) [37], the probability that an agent has a path free of collision with the environment. We use it to address a drawback of existing works, which often neglect the importance of forecasting that adheres to environment structures. We report ECFL in percent points, where 100% means no collisions. More details can be found in the Supplementary Materials.

Statistical Analysis / Model Ranking It is challenging to compare different models across multiple metrics. Therefore, we test the statistical significance of the results, using both traditional approach [9] and modern Bayesian analysis [2]. The Supplementary Materials provides the details.

4.2. Quantitative Results

We conduct experiments on the three datasets introduced in Sec. 4.1 and compare the performance of MUSE-VAE with Trajectron++ (T++) [33], Y-net [25], and AgentFormer (AF) [45] baselines, using their public code. Scene maps provided by PFSD and nuScenes show a much wider range of environments compared to SSD. Therefore, we provide a local view of the semantic map to all models including ours for a fair comparison. For all experiments in MUSE-VAE, we sample the latent factor z only once in Micro-stage, and we gain all diversity from the latent factor w in LG-CVAE by sampling it K times since we assume the uncertainty primarily depends on the long-term goal position.

Tab. 1 summarizes the experimental results on PFSD. Following the commonly used temporal horizon setting, we observe 3.2 sec (8 frames) and predict 4.8 sec (12 frames)

future trajectories. Considering the increased complexity of the local environment layouts of PFSD, we choose sampling number $K = 20, 50$ to investigate the learned trajectory distribution. Our model can achieve the best performance for all metrics in $K = 20, 50$ except for FDE in $K = 20$ where our model stands at second best. The KDE NLL scores of Y-net and AF indicate that their K predictions fail to reflect the true trajectory distribution. This is because the K predictions are not sampled from the learned distribution from their first training stage but sampled in the next stage by manipulating them to focus on the diversity. Y-net conducts a test time sampling trick based on K-means clustering to obtain diverse predictions. AF has the second stage training to apply the pairwise distance loss between K predictions for the diversity, which is inefficient since it requires re-training whenever K changes. On the other hand, MUSE-VAE can produce predictions within a low error range with GT trajectories, while reflecting the GT trajectory distribution (lower KDE NLL) and making realistic predictions reducing environment collisions (higher ECFL).

Tab. 2 shows the evaluation on SDD. It follows the same temporal horizon setup as PFSD. As in the prior works, we choose $K = 5, 20$ and errors are reported in pixel distance. MUSE-VAE can significantly outperform the state-of-the-art methods in ADE. Though our model shows the second best performance in FDE, MUSE-VAE largely ties up with the best method. For the same reason analyzed in PFSD, our model gives the best performance in KDE NLL. We can see that MUSE-VAE has slightly worse ECFL, which is still the second best, than Y-net. This is because the labeling of the scene provided from Y-net is incomplete[2], which adversely affects MUSE-VAE that relies heavily on the semantic map in Macro-stage predictions.

For the nuScenes dataset, following prior works, 2 sec (4 frames) observations and 6 sec (12 frames) predictions are made only for the vehicles and $K = 5, 10$ generations are investigated. Tab. 3 shows that our model consistently outperforms the others in every metric and sampling number. Compared to the previous two datasets, nuScenes has much narrower and strict navigable space, where our Macro stage can take the benefit of accurate LG and SG predictions aligned well with environment. On the other hand, since nuScenes is a real world dataset, many static past trajectories are also observed. Due to the fact that our model focuses on learning the trajectory distribution rather than simply having min ADE/FDE based on diverse samplings and generations, these real world data characteristics in nuScenes are well reflected in the trained model, which can lead to better performance across all metrics.

Statistical Analysis We computed average rankings of the methods, and T++, Y-Net, AF, and **Ours** obtain 3.42, 2.92, 2.33, **1.33**, respectively. We conducted the Friedman test [11] and confirmed that our method outperformed AF with statistical significance. We also conducted the Bayesian signed rank test [3] and confirmed that our method is either superior or at least on par versus the competitors. The Supplementary Materials explain this in further detail.

[2]The incomplete labels are discussed in the Supplementary Materials.

Table 1. Results on the PFSD with K = 20 and 50. With $t_p = 3.2$s (8 frames) and $t_f = 4.8$s (12 frames), errors are in meters.

K	Model	ADE ↓	FDE ↓	KDE NLL ↓	ECFL ↑
20	T++	0.17	0.37	-0.88	83.32
	Y-net	0.13	0.20	0.20	91.52
	AF	0.08	**0.11**	0.47	94.54
	Ours	**0.07**	0.12	**-1.46**	**96.95**
50	T++	0.14	0.25	-1.11	83.39
	Y-net	0.09	0.12	0.04	91.74
	AF	0.08	**0.09**	1.17	95.37
	Ours	**0.06**	**0.09**	**-1.68**	**97.02**

Table 2. Results on the SDD with K = 5 and 20. With $t_p = 3.2$s (8 frames) and $t_f = 4.8$s (12 frames), errors are in pixels.

K	Model	ADE ↓	FDE ↓	KDE NLL ↓	ECFL ↑
5	T++	11.11	24.42	8.74	86.94
	Y-net	11.49	20.19	8.98	**89.99**
	AF	11.47	**18.88**	8.57	89.02
	Ours	**9.60**	19.70	**8.43**	89.30
20	T++	8.16	16.40	7.37	86.88
	Y-net	7.84	11.94	8.05	**89.32**
	AF	8.35	**11.03**	7.48	87.30
	Ours	**6.36**	11.10	**7.21**	89.30

Table 3. Results on the nuScenes with K = 5 and 10. With $t_p = 2$s (4 frames) and $t_f = 6$s (12 frames), errors are in meters.

K	Model	ADE ↓	FDE ↓	KDE NLL ↓	ECFL ↑
5	T++	3.14	7.45	7.20	68.99
	Y-net	2.46	5.15	11.03	85.46
	AF	1.59	3.14	9.39	86.74
	Ours	**1.38**	**2.90**	**5.12**	**89.24**
10	T++	2.46	5.65	5.61	69.02
	Y-net	1.88	3.47	7.52	82.90
	AF	1.30	2.47	7.76	85.76
	Ours	**1.09**	**2.10**	**3.82**	**89.33**

4.3. Qualitative Results

We provide additional qualitative context to the quantitative metrics, in order to reveal the underlying factors that support each model's benefits and tradeoffs. In Fig. 4, we visualize several instances of predicted long and short-term goals as well as the trajectories in the context of different environments and movement behaviors, driven by the three datasets we used for evaluation. Specifically, Figs. 4a and 4b are instances from PFSD with $K = 20$, Figs. 4c and 4d are drawn for SDD with $K = 20$, and Figs. 4e and 4f come from nuScenes with $K = 10$. We take a look at in-

stances of a 'fork-in-the-road' scenario from each dataset to test ability of models to understand the multimodality of long-term goals conditioned on the environment. In Figs. 4a, 4c and 4e, we overlay predicted trajectories and goal heatmaps from Macro-stage over local semantic maps to demonstrate the ability of the models to make reasonable coarse predictions in the context of different environment features. The first column with the green border is the long-term goal prediction from LG-CVAE. The following three columns with the orange border are two short-term goals and one long-term goal from SG-Net. The two rows show two different predictions generated by sampling two different latent factors w in LG-CVAE, based on the same observation x. We can see that (1) the short-term goals align well with the given predicted long-term goal, and (2) long-term goal projections naturally vary because of the structure of the 'fork-in-the-road' scenario, which gives a generally bimodal uncertainty in the possible goal directions.

Figs. 4b, 4d and 4f illustrate complete trajectory predictions, where the images in the clock-wise order, from the top-left, correspond to the Micro-stage of MUSE-VAE, followed by T++, AF, and Y-net, respectively. Across all three datasets, we can observe that predictions of T++ and AF tend to lead to collisions with the environment. On the other hand, predictions of Y-net and our MUSE-VAE are well-aligned and collision-free. We attribute this to T++ and AF encoding the semantic map into a 1D-representation, which entangles the spatial signal, while our model and Y-net process the semantic map along with the trajectory heatmap in 2D. Although Y-net produces predictions that avoid collision with obstacles, in contrast to MUSE-VAE it yields trajectories with diverse duration, which often overshoot or undershoot the true trajectory horizon. This is because the goal predictions of Y-net are not made directly by the learned model; rather, they stem from the test time sampling trick, which is weakly conditioned on the past trajectory signal, particularly its velocity. On the other hand, our MUSE-VAE's goal predictions are not only well aligned with the environment structure in the Macro-stage, but also reflect learned dependency on the past trajectory sequence modeled by an RNN in the Micro stage.

4.4. Ablation Study

We analyze the effectiveness of each component in MUSE-VAE through an ablation study. Tab. 4 shows three ablated experiments with the complete model MUSE-VAE. **w/o SG-net** model has no SG-net in Macro-stage, and thus, the long-term goal prediction is directly fed to the Micro-stage. **w/o Micro-stage** model does not include the Micro-stage, implying all future trajectories are predicted in the SG-net by letting $N_{SG} = t_f - 1$. In **w/o LL-prior** model, we eliminate the log-likelihood from the prior distribution $p_\tau(z|x)$ to assess the utility of this term in reducing the gap between the training and the inference-time reconstruction.

Table 4. Ablation study on the PFSD with $K = 20$. With $t_p = 3.2$s (8 frames) and $t_f = 4.8$s (12 frames), errors are in meters.

Model	ADE ↓	FDE ↓	KDE NLL ↓	ECFL ↑
MUSE-VAE	0.07	0.12	-1.46	96.95
w/o SG-net	0.10	0.13	-0.48	91.88
w/o Micro-stage	0.13	0.12	-	99.24
w/o LL-prior	0.07	0.13	-0.96	95.34

Our model requires the LG prediction from LG-CVAE, necessitating its presence in all experiments. Thus, there is little observed variability in min FDE. The most notable difference in performance stems from **w/o Micro-stage**, the absence of which precludes evaluation of the KDE NLL score. In this case, complete trajectory predictions happen in the SG-net, defined in discrete pixel coordinates, thus limiting the accuracy of the forecasted trajectory[3]. On the other hand, an advantage of this model is the few collisions, indicated by ECFL, because all predictions are obtained from pixel coordinates well aligned with the environment. In **w/o SG-net**, Micro-stage has no information of waypoints other than the LG prediction from LG-CVAE. Thus, the KDE NLL value shows that distribution learning of **w/o SG-net** is not as good as a complete model. **w/o LL-prior** also degrades KDE NLL performance. This indicates that the reconstruction loss from the prior distribution during training allows the model to learn how to generate predictions that better reflect one's movement patterns given past trajectories. This thorough ablation study shows that it is crucial to consider both the Macro-stage for coarse predictions aligned well with the environment and the Micro-stage for fine predictions reflecting the past sequential states.

5. Conclusion

In this paper, we introduce MUSE-VAE a probabilistic model capable of recognizing the environment and generating multimodal predictions based on the coarse-to-fine approach. Our experimental results using various datasets and metrics show MUSE-VAE achieves both versatile and accurate forecasts that are well matched to environmental conditions. MUSE-VAE processes each agent independently, which cannot reflect agent-interaction. In the future work, we will take into consideration of multi agent-aware model that can avoid collisions with neighboring agents.

Acknowledgement

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119.

[3]Complete trajectory predictions in this case are made from the heatmap maxima.

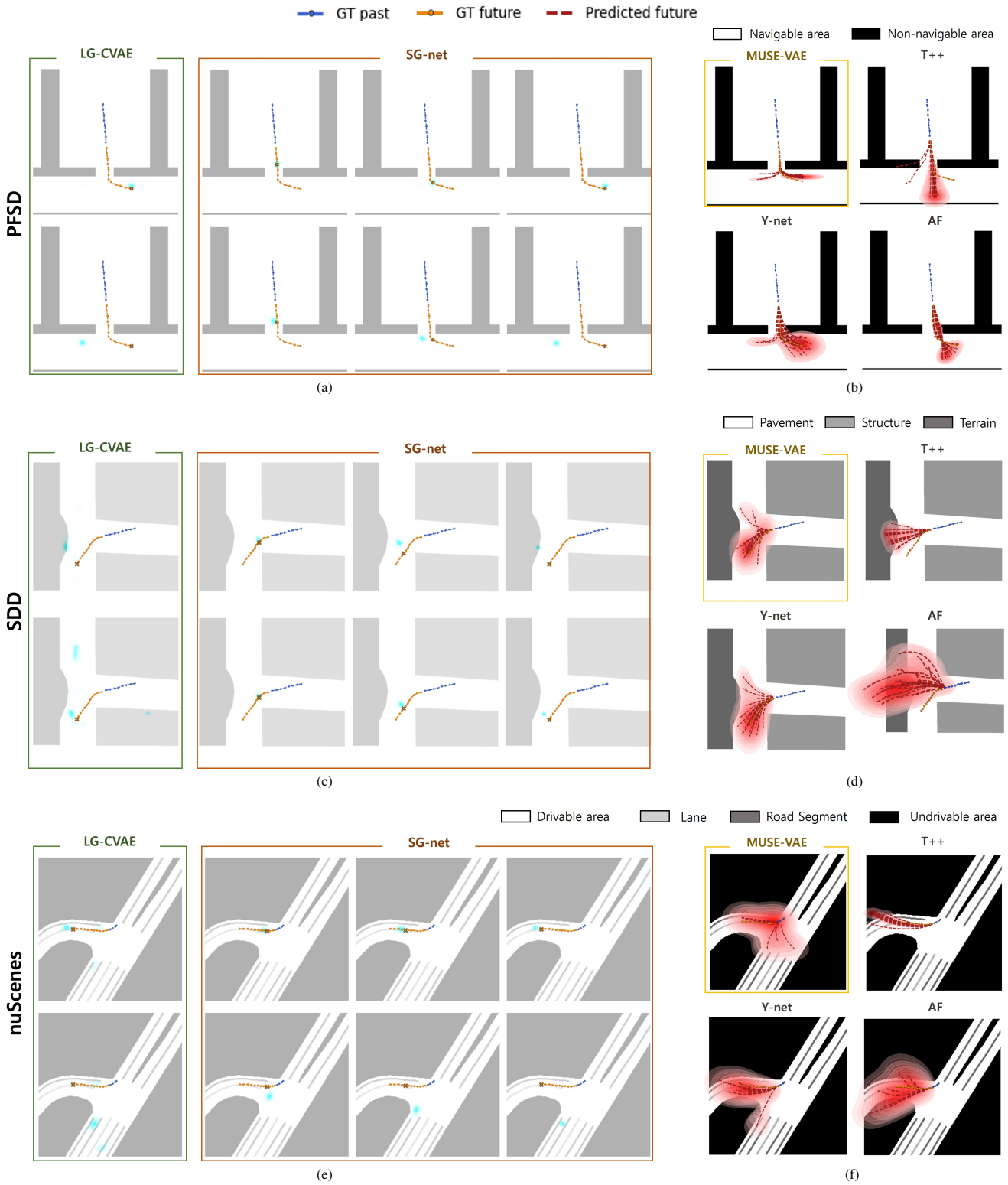


Figure 4. Left: Macro-stage results of (a) PFSD, (c) SDD, and (e) nuScenes respectively. In the first column, the Long-term Goal (LG) heat map prediction from LG-CVAE is overlaid on the local semantic map. The following three columns are two Short-term Goals (SG) and one LG from SG-Net. Here we show only two different sampling generations in each dataset. The blue and orange lines indicate GT past and GT future trajectories, respectively. GT LG and SGs are marked with 'x'. Right: Complete trajectory predictions of (b) PFSD, (d) SDD, and (f) nuScenes respectively. In each dataset, the 1st/2nd/3rd/4th image from top-left to bottom-right is from Micro-stage of ours/Trajectron++/Y-net/AgentFormer, respectively. The blue, orange, and red lines indicate GT past, GT future, predicted future trajectories, respectively.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1, 2

[2] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017. 5

[3] A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1026–II–1034. JMLR.org, 2014. 6

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. 4

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016. 4

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 1, 5

[7] Wang Yiwei Zhu Yiheng Cham Tat-Jen Cai Jianfei Yuan Junsong Liu Jun Cai, Yujun et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 5

[8] Eric Chown, Stephen Kaplan, and David Kortenkamp. Prototypes, location, and associative networks (plan): Towards a unified theory of cognitive mapping. *Cognitive Science*, 19(1):1–51, 1995. 1

[9] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. 5

[10] Gonzalo Ferrer, Anais Garrell, and Alberto Sanfeliu. Social-aware robot navigation in urban environments. In *2013 European Conference on Mobile Robots*, pages 331–336. IEEE, 2013. 1

[11] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. 6

[12] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342, 2021. 2

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[14] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. pages 2255–2264, 06 2018. 1, 2

[15] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physical review E*, 51(5):4282, 1995. 1

[16] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 5

[17] B. Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2375–2384, 2019. 2, 5

[18] Diederik P. Kingma, Tim Salimans, and Max Welling. Improved variational inference with inverse autoregressive flow. *ArXiv*, abs/1606.04934, 2017. 4

[19] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, 2018. 4

[20] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian D. Reid, Seyed Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 2019. 2

[21] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher Bongsoo Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2165–2174, 2017. 2

[22] Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. A surprisingly effective fix for deep latent variable modeling of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November 2019. 4

[23] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 5

[24] Yecheng Jason Ma, Jeevana Priya Inala, Dinesh Jayaraman, and Osbert Bastani. Diverse sampling for normalizing flow based trajectory forecasting. *ArXiv*, abs/2011.15084, 2020. 5

[25] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021. 1, 2, 3, 5

[26] Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. Covernet: Multimodal behavior prediction using trajectory sets. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14062–14071, 2020. 5

[27] Bastiaan Quast. rnn: a recurrent neural network in r. *Working Papers*, 2016. 2, 3

[28] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2821–2830, 2019. 3

[29] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016. 5

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015. 4

[31] Amir Sadeghian, Vineet Kosaraju, Agrim Gupta, Silvio Savarese, and A Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018. 5

[32] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. pages 1349–1358, 06 2019. 2, 5

[33] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. *Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data*, pages 683–700. 12 2020. 2, 3, 5

[34] Farnaz Sharif, Behnam Tayebi, György Buzsáki, Sebastien Royer, and Antonio Fernandez-Ruiz. Subcircuits of deep and superficial ca1 place cells support efficient spatial coding across heterogeneous environments. *Neuron*, 109(2):363–376, 2021. 1

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *Neural Information Processing Systems (NIPS)*, 2015. 2, 4

[36] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 5

[37] Samuel S. Sohn, Mihee Lee, Seonghyeon Moon, Gang Qiao, Muhammad Usman, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. A2x: An agent and environment interaction benchmark for multimodal human trajectory prediction. In *Motion, Interaction and Games*, MIG '21, New York, NY, USA, 2021. Association for Computing Machinery. 5

[38] Samuel S Sohn, Honglu Zhou, Seonghyeon Moon, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. Laying the foundations of deep long-term crowd flow prediction. In *European Conference on Computer Vision*, pages 711–728. Springer, 2020. 5

[39] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NIPS*, 2016. 4

[40] Yichuan Tang and Ruslan Salakhutdinov. Multiple futures prediction. In *NeurIPS*, 2019. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[42] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, 2018. 1, 2

[43] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J. Crandall. Stepwise goal-driven networks for trajectory prediction. *ArXiv*, abs/2103.14107, 2021. 1, 3

[44] Jan M Wiener, Simon J Büchner, and Christoph Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9(2):152–165, 2009. 1

[45] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5

[46] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, Congcong Li, and Dragomir Anguelov. TNT: target-driven trajectory prediction. *CoRR*, abs/2008.08294, 2020. 1, 2, 3