Shared Cross-Modal Trajectory Prediction for Autonomous Driving

Chiho Choi[1*] Joon Hee Choi[2*] Jiachen Li[1,3†] Srikanth Malla[1]
[1]Honda Research Institute USA
[2]Sungkyunkwan University
[3]University of California, Berkeley
{cchoi, smalla}@honda-ri.com jhchoi2019@skku.edu jiachen_li@berkeley.edu

Abstract

Predicting future trajectories of traffic agents in highly interactive environments is an essential and challenging problem for the safe operation of autonomous driving systems. On the basis of the fact that self-driving vehicles are equipped with various types of sensors (e.g., LiDAR scanner, RGB camera, radar, etc.), we propose a Cross-Modal Embedding framework that aims to benefit from the use of multiple input modalities. At training time, our model learns to embed a set of complementary features in a shared latent space by jointly optimizing the objective functions across different types of input data. At test time, a single input modality (e.g., LiDAR data) is required to generate predictions from the input perspective (i.e., in the LiDAR space), while taking advantages from the model trained with multiple sensor modalities. An extensive evaluation is conducted to show the efficacy of the proposed framework using two benchmark driving datasets.

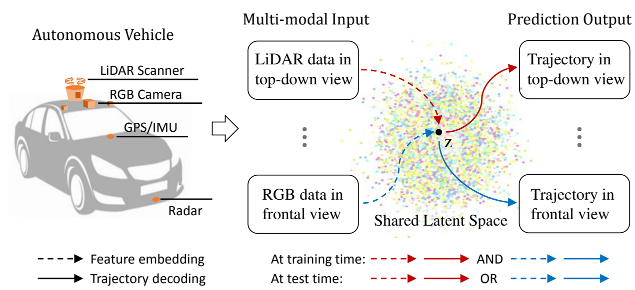


Figure 1: Given a set of multi-modal data (*e.g.*, LiDAR data, RGB images, etc.) obtained from an autonomous vehicle, the model is trained to embed complementary representations of different input modalities into a shared latent space. Output predictions are generated from different perspectives using a latent variable sampled from the learned embedding space. At test time, the proposed method takes a single input modality (*e.g.*, LiDAR data, red-dashed arrow) and predicts the future motion in the same space (*i.e.*, LiDAR-captured world space, red-solid arrow).

1. Introduction

Future trajectory prediction has become the central challenge to succeed in the safe operation of autonomous vehicles designed to cooperate with interactive agents (*i.e.*, pedestrians, cars, cyclists, etc.). It can benefit to the deployment of applications in autonomous navigation and driving assistance systems with advanced motion planning and decision making. Based on the fact that multi-modal sensors (*e.g.*, LiDAR scanner, RGB cameras, radar, etc.) are equipped in autonomous vehicles, we propose a cross-modal embedding framework that demonstrates the efficacy of the use of multiple sensor data for motion prediction.

Figure 1 illustrates an overview of the proposed approach. At training time, we embed multiple feature representations encoded from individual sensor data into a single shared latent space. Our model jointly optimizes the objective functions across different input modalities, so that the evidence lower bound of multiple input data over the likelihood can be jointly maximized. We provide a derivation of the objective of shared cross-modal embedding and its implementation using a CVAE-based generative model. At test time, the model takes a single input modality (*e.g.*, LiDAR data) and generates a future trajectory from the input perspective (*i.e.*, top-down view) using a latent variable sampled from the shared embedding space. In this way, we can benefit to the model training from the use of multiple input modalities[1], while keeping the same computational time for trajectory generation as if the single modality had been used. To the best of our knowledge, we are the first to employ multi-modal sensor data from a single framework

*Co-first author.
†Work done during Jiachen's internship at Honda Research Institute

[1]For example, top-down view LiDAR data and frontal view RGB images. However, the input modalities are not limited to these two types but also include stereo images, depth, radar, GPS, and many others equipped in autonomous vehicles, which can provide visual or locational information.

for trajectory prediction. Note that existing works solve the problem either in top-down view [23, 36, 9] with LiDAR data or in frontal view [46, 4, 30] with RGB images.

The proposed framework is clearly distinguishable to studies on a multi-modal pipeline for scene understanding such as detection [6, 22, 27], tracking [12, 50], and semantic segmentation [17, 41]. They have presented more accurate models by simply fusing different representations extracted from several sensor modalities. The generation of such joint representations, however, would not be desirable in driving automation systems due to the following issues: (i) during inference, it inherently increases the computation time proportional to the number of input modalities used; and (ii) with the anomalous LiDAR data, the model would fail in finding a solution, which is critical to operate self-driving vehicles. For the former issue, our proposed cross-modal embedding takes only a single input data during inference and thus does not influence the computational time, while it still benefits from the model trained with multiple input modalities. In the latter, our model provides alternative prediction solutions in frontal view using the RGB data, which will activate driving assistance functions (*i.e.*, ADAS) for safe vehicle operation, even with a sensor failure.

To this end, we generate multiple modes of future trajectories by sampling several latent variables from the learned latent space. However, such random sampling-based strategy [23, 9] is likely to predict similar trajectories, ignoring the random variables while generating predictions from the decoder. This posterior collapse[2] problem of VAEs is particularly critical to future prediction as it mitigates the diverse modes of system outputs. Therefore, we introduce a regularizer (i) that pushes the model to rely on the latent variables, predicting diverse modes of future motion; and (ii) that does not weaken the prediction capability of the decoder while preventing the performance degradation.

We address the following ideas in the proposed method:

- The objective of shared cross-modal embedding to jointly approximate a real distribution using multiple input sources is mathematically derived using the Kullback-Leibler divergence (Sec. 3.2).
- Shared cross-modal embedding is implemented based on our derivation to benefit from the use of multiple input modalities, while keeping the same computational time as if the single modality had been used (Sec. 3.2).
- The regularizer is designed for future prediction to mitigate posterior collapse of VAEs and to predict more diverse modes of motion behavior (Sec. 3.3).

In addition, we design an interaction graph with a graph-level target (Sec. 3.1), introduce a new evaluation metric to measure prediction success (Sec. 4.2), and propose to use absolute motions in frontal view (Sec. 4.1).

[2]We do not carry out any study on mode collapse of GANs or related problems other than posterior collapse of VAEs where our work is built on.

Throughout the paper, we use the word 'multi-modality' to denote two different sources. First, *multi-modal input* represents input data obtained from different types of sensors. Second, *multi-modal prediction* depicts predicted trajectory outputs with multiple variations.

2. Related Work

Pedestrian Trajectory Prediction A majority of research on trajectory prediction [1, 15, 43, 49] has been conducted toward modeling the interactive behavior between humans. These works first encode the temporal information of individual humans and then find their correlation through a social module. Recently, social interactions have been modeled from the graph structure in [42, 19, 31]. Although these methods may be successful in interaction modeling, they overlook the environmental influences that may cause prediction failures in structured environments with stationary obstacles. Therefore, the subsequent work [7, 21] takes images as input to constrain their model using scene context.

Vehicle Trajectory Prediction in Top-down View Similar interaction modules are applied for vehicle trajectory prediction. Some approaches only consider the past motion of road agents [11, 33, 28, 25], and thus result in large errors with a complex road environment in traffic scenes. To alleviate such problems, [23, 36, 24, 9, 37, 39] input additional visual cues to condition their model on the road topology. However, they overlook the vehicle interactions against pedestrians, which is most critical to model the natural behavior of vehicles on the road for safe driving. We thus do not limit our scope to 'vehicle' trajectories and its interactions. Instead, we explicitly discover interactions of heterogeneous entities using the proposed interaction graph.

Vehicle Trajectory Prediction in Frontal View [4, 46, 30, 29] aim to predict the future trajectory of vehicles in a frontal view image space. They predict a target agent's *relative* trajectory with respect to the potential motion of ego-vehicle. Therefore, the predictions are valid only if the accurate ego-future is available. In practice, however, prediction of ego-motion is an another research topic [18] in the transportation domain, which makes hard to simply apply such systems to the real world driving applications. Therefore, we predict the *absolute* coordinates of trajectories with no effect of unknown ego-future in frontal view.

Multi-Modal Learning Learning representations of multiple input modalities have been explored in recent years. As described in [32], multi-modal learning can be categorized into three types. *Multi-modal fusion* takes multiple modalities as input and learns their joint representations. Basically, the same set of input types should be provided at test time as in [20, 45]. *Cross-modal learning* tries to learn more descriptive representations from one modality when auxiliary modalities are given at training time. During inference, the auxiliary modalities are not necessary as in [16, 8]. *Shared representation learning* learns the representation from one modality and performs the test on the other modality as shown in [47, 35]. The proposed cross-modal embedding aligns in between cross-modal learning and shared representation learning, similar in spirit to [2]. We aim to benefit from different modalities that are correlated to each other. However, rather than learning common representations, we train the model to embed different representations into the shared cross-modal latent space.

3. Proposed Method

Given a scenario with the trajectory data $\mathcal{T} = \{T^i | \forall i \in \{1, ..., K\}\}$ of K traffic agents, we split T^i into a past trajectory $\mathbf{x}^i = \{x_t^i | \forall t \in \{1, ..., \tau\}\}$ for the first τ observation time steps and a future trajectory $\mathbf{y}^i = \{x_t^i | \forall t \in \{\tau+1, ..., \tau+\delta\}\}$ for the next δ time steps, where x_t^i represents a 2D position of an arbitrary agent i at time t. Assuming that a visual sequence $\mathcal{I}$ is available during τ observation time steps, we compute the optical flow $\mathcal{O}$ by running TV-L1 [48] and segmentation map S from DeepLab-V2 [5] trained on Cityscapes [10]. Given $\{\mathcal{O}, S\}$ and $\{\mathbf{x}^i | \forall i \in \{1, ..., K\}\}$, our goal is to generate a trajectory prediction $\hat{\mathbf{y}}^k$ of the target agent k. To achieve this, we build a feature extraction module in Sec. 3.1 upon graph neural networks (GNNs) in order to learn social behaviors $\mathbf{c}^k$ of the target k toward all other traffic agents (*e.g.*, pedestrians, vehicles, etc.) as well as surrounding road structures. Then, we derive the objective of the proposed shared cross-modal embedding and show its implementation within CVAE in Sec. 3.2. The encoder $q(\mathbf{z}|\mathbf{y}^k, \mathbf{c}^k)$ is learned to embed $\mathbf{y}^k$ into the latent space, conditioning on the observed social behavior $\mathbf{c}^k$. The following decoder $p(\mathbf{y}^k|\mathbf{z}, \mathbf{c}^k)$ reconstructs the future locations $\mathbf{y}^k$ using $\mathbf{c}^k$ with a latent sample $\mathbf{z}$. Finally, in Sec. 3.3 we provide a solution for mode diversification addressing the posterior collapse issue.

3.1. Social Behavior Encoding

Input Layer for External Features The importance of external constraints on trajectory prediction is particularly pronounced for traffic agents in driving scenes. To model such environmental influences, the system should be able to recognize each object's static/dynamic states as well as the semantic context of the scene.

The image sequence $\mathcal{I}$ captured during the past time steps is used to generate two types of representations: a set of optical flow images $\mathcal{O}$ and a segmentation map S. The temporal changes of the objects from $\mathcal{O}$ are processed using the 3D convolutional neural network CNN_{3D} by extracting temporal representations $\mathbf{f}_T$ along the time axis:

$$\mathbf{f}_T = CNN_{3D}(\mathcal{O}; W_T), \tag{1}$$

where W_T is the learnable weight parameters.

In addition, a pixel-level segmentation map is obtained at the first time step of the given scenario. Among the estimated labels, we only leave the background structures such as road, sidewalk, vegetation, etc. to extract visual features from the stationary environment. The 2D convolutional neural network CNN_{2D} is used in this stream to take advantage of its spatial feature encoding:

$$\mathbf{f}_S = CNN_{2D}(S; W_S), \tag{2}$$

where W_S is the learnable parameters.

We merge the temporal states $\mathbf{f}_T$ of static/dynamic objects with the spatial features $\mathbf{f}_S$ of the stationary context to generate spatio-temporal features

$$\mathbf{f}_E = \mathbf{f}_T + \mathbf{f}_S. \tag{3}$$

We further convert $\mathbf{f}_E \in \mathbb{R}^{d_C \times d_C \times d_E}$ to the external feature matrix $F \in \mathbb{R}^{K \times d_E}$ for the graph. K entities (of size d_E) are taken from one of cells in a $d_C \times d_C$ grid of $\mathbf{f}_E$, where the cell location corresponds to each agent i's original pixel location at time τ. For example, an agent shown in the first 32×32 sub-region of an original 256×256 image takes the feature vector from the $(1, 1)$-th cell in a 8×8 grid of $\mathbf{f}_E$.

Input Layer for Node Features Using the past motion history of traffic agents, we encode the node features. Assuming the task is to predict the future motion of the target agent k, we first discover its own intent by preforming the following procedure. The past states $\mathbf{x}^k$ is encoded into high dimensional feature representations $\mathcal{U}_k$ through the multi-layer perceptron (MLP). The encoded features are then combined with the local perception that contains mid-level semantic context $\Omega_{x_\tau^k}$ (nearby areas of x_τ^k) from former CNN_{2D}. By adding spatial locality, interactions of the target toward the local environment further constrain its motion intent. The subsequent LSTM captures the temporal dependency of motion states on the local environment by

$$\begin{aligned} \mathcal{U}^k &= \text{MLP}\left(\mathbf{x}^k; W_U\right), \\ \mathbf{h}_{t+1}^k &= \text{LSTM}\left(\mathcal{U}_t^k + \Omega_{x_\tau^k}, \mathbf{h}_t^k; W_K\right), \end{aligned} \tag{4}$$

where W_U and W_K is the learnable parameters of MLP and LSTM layer, and $\mathbf{h}_t^k$ denotes the hidden state of LSTM at time t. We define the last hidden state as $\mathbf{h}_{(0)}^k$ and use it to initialize the node features of the target in the graph.

We run a different feature encoding procedure for the rest of the agents $j \in \{1, ..., K\} \backslash \{k\}$ to model their relative motion toward the target agent k as follows:

$$\begin{aligned} \mathcal{V}^j &= \text{MLP}\left(\mathbf{x}^k - \mathbf{x}^j; W_V\right), \\ \mathbf{h}_{t+1}^j &= \text{LSTM}\left(\mathcal{V}_t^j, \mathbf{h}_t^j; W_J\right), \end{aligned} \tag{5}$$

where W_V and W_J is the learnable parameters of MLP and LSTM, and $\mathbf{h}_t^j$ denotes the hidden state of LSTM at time t. This process is simple yet effective to infer temporal changes of interactive behavior of individual agents. We use

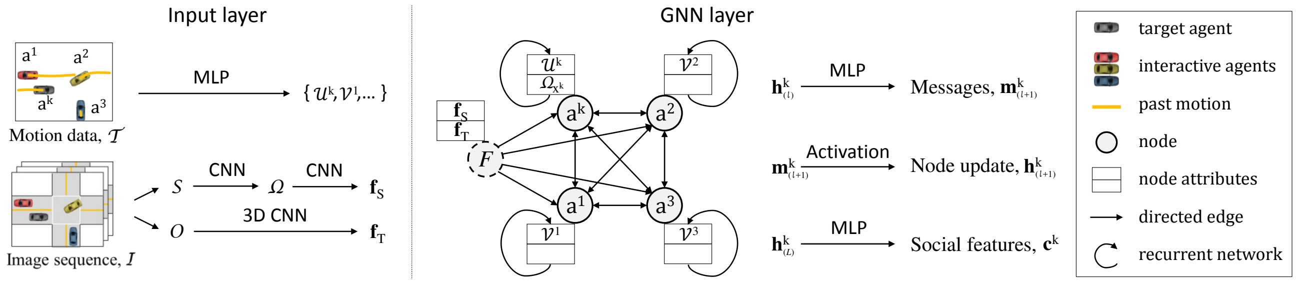


Figure 2: The illustration of the proposed feature encoder. Using the past image sequence, we model spatio-temporal factors given by external environments. The internal and social factors of the target agent is encoded from its past motion as well as surrounding local perceptual context. In the following GNN layer, we model agent-specific social behavior.

the last hidden state of each agent j as $\mathbf{h}^j_{(0)}$ for the graph.

GNN Layer The social behavior of the target agent is modeled from each agent's features and external environment features. We define a graph $G = (H, F)$, where $H \in \mathbb{R}^{K \times d_V}$, $H = \{\mathbf{h}^i | \forall i \in \{1, ..., K\}\}$ is a node feature matrix representing K node embeddings of size d_V. $F \in \mathbb{R}^{K \times d_E}$ is an external feature matrix $F = \{\mathbf{f}^i | \forall i \in \{1, ..., K\}\}$, where each entity represents outside influence on each node in the graph. Following the general message passing phases [14], we construct a GNN architecture:

$$H_{(l+1)} = M(H_{(l)}, F), \tag{6}$$

where M is the message propagation function that takes the node feature matrix $H_{(l)}$ updated by l times of the message passing phase. We initialize $H_{(0)} = \{\mathbf{h}^j_{(0)} | \forall j \in \{1, ..., K\} \backslash \{k\}\} \cup \{\mathbf{h}^k_{(0)}\}$ using the hidden states obtained from the input layer.

The proposed GNN structure for social behavior modeling can be considered as a family of pair message passing neural networks [3], where the function M takes a concatenation of two nodes as a pair. We design our model on top of this process with an additional graph-level target:

$$\begin{aligned} \mathbf{m}^k_{(l+1)} &= \sum_{i,j} \text{MLP}\Big(\text{Concat}(\mathbf{h}^i_{(l)} + \mathbf{f}^i, \\ &\qquad\qquad \mathbf{h}^j_{(l)} + \mathbf{f}^j, \;\; \mathbf{h}^k_{(l)}); W_M\Big), \\ \mathbf{h}^k_{(l+1)} &= \sigma\left(\mathbf{m}^k_{(l+1)}\right), \end{aligned} \tag{7}$$

where W_M is the learnable parameters of MLP, Concat(,,) denotes concatenation, k is a target agent, and i and j are the rest of agents. During the message passing phase, the relation between two nodes i and j is encoded with respect to the target node k by considering their external influences $\mathbf{f}^i$, $\mathbf{f}^j$. A summation operation generates messages invariant to the permutation of the nodes. Then, the features of the target node $\mathbf{h}^k_{(l+1)}$ in the graph are updated by a non-linearity function σ such as ReLU using the messages $\mathbf{m}^k_{(l+1)}$. After L updates, the output social behavior features $\mathbf{c}^k$ are generated by another MLP during the readout phase:

$$\mathbf{c}^k = \text{MLP}(\mathbf{h}^k_{(L)}; W_R), \tag{8}$$

where W_R is the learnable parameters. We use $\mathbf{c}^k_i$ for a certain input i. For notational brevity, we drop the target indicator k in the following sections. The input layer and GNN layer is illustrated in Figure 2, and details of the network architecture are shown in the supplementary material.

3.2. Shared Cross-Modal Framework

The main contribution of this work is that we propose a cross-modal embedding framework for future prediction. It aims to benefit from the use of multiple input modalities, while keeping the same computational complexity as if the single data type had been used for trajectory prediction. To implement such functionality, we derive our model within the CVAE framework to embed various types of representations into a single shared latent space. Instead of learning the latent space manifold from a single input, several complementary representations extracted from multiple data sources simultaneously characterize the cross-modal space at training time. By jointly learning the same scenario from different input perspectives, the generative process becomes more descriptive, which results in increasing the performance. At test time, a single modal input is used to sample the latent variables from the learned cross-modal space, taking advantages with other sensor modalities.

In the followings, we mathematically derive the objective function of shared cross-modal embedding and extend its derivation toward a generative model conditioned on the input observation.

Joint Optimization The objective of cross-modal embedding is to jointly approximate a real distribution $p(\mathbf{z})$ using a posterior $q_i(\mathbf{z}|\mathbf{y}_i)$ of multiple input sources $i \in$ {LiDAR, RGB, ...}, where $\mathbf{y}_i$ is the sample data point of input modality i, and $\mathbf{z}$ is the latent variable. Exploiting the fact that

$$KL(q(\mathbf{y})||p(\mathbf{y})) = -\int q(\mathbf{y}) \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) d\mathbf{y} \geq 0, \tag{9}$$

the Kullback-Leibler (KL) divergence associated with multiple approximates q_i is given by:

$$\sum_i KL(q_i(\mathbf{z}|\mathbf{y}_i) \mid\mid p(\mathbf{z}|\mathbf{y}_i)) = \sum_i -\int q_i(\mathbf{z}|\mathbf{y}_i)\log\left(\frac{p(\mathbf{z}|\mathbf{y}_i)}{q_i(\mathbf{z}|\mathbf{y}_i)}\right)d\mathbf{z} \geq 0. \quad (10)$$

By applying Baye's theorem and employing $\int q_i(\mathbf{z}|\mathbf{y}_i)d\mathbf{z} = 1$, Eqn. (10) can be revised as

$$\sum_i \left(-\int q_i(\mathbf{z}|\mathbf{y}_i)\log\left(\frac{p_i(\mathbf{y}_i|\mathbf{z})p(\mathbf{z})}{q_i(\mathbf{z}|\mathbf{y}_i)}\right)d\mathbf{z} + \log p(\mathbf{y}_i)\right) \geq 0. \quad (11)$$

Using the definition of the KL divergence and expected value and simple math, Eqn. (11) is converted to

$$\log\left(\prod_i p(\mathbf{y}_i)\right) \geq \sum_i \Big(-KL\left(q_i(\mathbf{z}|\mathbf{y}_i)||p(\mathbf{z})\right) + \mathbb{E}_{\sim q_i(\mathbf{z}|\mathbf{y}_i)}[\log p_i(\mathbf{y}_i|\mathbf{z})]\Big). \quad (12)$$

Therefore, maximizing the evidence lower bound (ELBO) of multiple input data over the likelihood jointly maximizes their evidence probability.

Cross-Modal Embedding The proposed cross-modal embedding framework is trained to jointly learn the shared latent space conditioned on multiple input observations such as $\mathbf{c}_{i\in\{\text{LiDAR,RGB,etc.}\}}$. The variational lower bound of the log-likelihood can be extended as a conditional form by

$$\log\left(\prod_i p(\mathbf{y}_i|\mathbf{c}_i)\right) \geq \sum_i \Big(-KL(q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)||p(\mathbf{z}|\mathbf{c}_i)) + \mathbb{E}_{\sim q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)}[\log p_i(\mathbf{y}_i|\mathbf{z},\mathbf{c}_i)]\Big), \quad (13)$$

where $q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)$ and $p_i(\mathbf{y}_i|\mathbf{z},\mathbf{c}_i)$ is implemented as a pair of an encoder and decoder for i-th input modality following the reparameterization trick of CVAE. $\mathbf{c}_i$ is the conditional observation. **The full derivation is provided in the supplementary material.** We draw the loss to minimize the negative ELBO while training the model as follows:

$$\mathcal{L}_E = \sum_i \Big(KL(q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)\|p(\mathbf{z}|\mathbf{c}_i)) - \mathbb{E}_{\sim q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)}[\log p_i(\mathbf{y}_i|\mathbf{z},\mathbf{c}_i)]\Big). \quad (14)$$

The network parameters of the encoder are learned to minimize the KL divergence between the prior distribution $p(\mathbf{z}|\mathbf{c}_i)$ and the approximates $q_i(\mathbf{z}|\mathbf{y}_i,\mathbf{c}_i)$. The second term is the log-likelihood of samples, which is considered as the reconstruction loss of the decoder. The decoder generates trajectories using the latent variables $\mathbf{z}$ sampled from the prior that is modeled as Gaussian distribution $\mathbf{z} \sim \mathcal{N}(0, \mathrm{I})$.

3.3. Multi-modal Prediction

In practice, the optimization of VAE and its variants is challenging itself because of the posterior collapse problem. The strong autoregressive power of the decoder often ignores the random variable $\mathbf{z}$ sampled from the learned latent space. Thus, the output is dominantly generated using the conditional input $\mathbf{c}$, still satisfying the minimization of the KL divergence and maximization of the log-likelihood in Eqn. (14). Such a problem alleviates the multi-modal nature of future prediction where multiple plausible trajectories are generated given the same past motion. To address posterior collapse, we consider the following challenges: (i) our technique helps to generate diverse responses from the decoder, which enables multi-modal prediction and (ii) it does not physically weaken the decoder to alleviate its prediction capability. In this sense, we design an auxiliary regularizer that makes the decoder to rely on the latent variable.

At training time, we assume that there exist N modes of trajectories for each query. Then, the latent variables $\mathbf{z}_n \sim q(\mathbf{z}_n|\mathbf{y},\mathbf{c}) = \mathcal{N}(\mu, \sigma^2)$ are sampled from the normal distribution with the mean μ and variance σ^2, where $n \in \{1, ..., N\}$. We consider the trajectories generated using these latent variables as N modes of prediction outputs. To maximize the diversity among predictions, the pair-wise similarity is evaluated using Gaussian kernel by

$$K = \exp\left(-\frac{D(\hat{\mathbf{y}}_i,\hat{\mathbf{y}}_j)}{2\sigma_G^2}\right), \quad (15)$$

where D measures a distance between predictions $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ with $i,j \in \{1, ..., N\}$ and σ_G^2 is the hyper-parameter of this kernel function. We find a maximum similarity K_{max} and minimize it during training. The regularizer then enforces the model to maximize the diversity among N predicted trajectories through the optimization without losing the prediction capability of the decoder.

As a result, the total objective function of the proposed approach is drawn as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_E + \lambda\sum_i K_{max,i} \quad (16)$$

where $i \in$ {LiDAR,RGB,...} is an indicator for input data modalities and λ balances multi-modality and accuracy ($\lambda = 10$ is used). To optimize the first term in Eqn. (16), we find $\hat{\mathbf{y}}_n$ of the mode n that best reconstructs the ground truth $\mathbf{y}$. In this way, the log-likelihood in Eqn. (14) encourages the decoder to generate accurate results, while preserving the mode diversity with the regularizer.

4. Experiments

4.1. Input Modalities

Any set of sensory data can be used as input to the proposed framework. For demonstration, however, we use two exemplary data types that are easily accessible from the existing benchmark datasets: (i) *LiDAR data* provide 3D scanning of the surrounding environment. Using 3D point clouds, we project every single point onto the ground plane in top-down view and predict trajectories of traffic agents in the LiDAR-captured world coordinates. (ii) *RGB images* captured from a frontal-facing camera provide rich and dense representations. We predict the trajectories from the egocentric perspective in the image space. Unlike relative trajectories in [44, 46], we propose to predict trajectories using the absolute locations, eliminating the effect of uncertain ego-future [30]. We provide its details with our data preparation in the supplementary material.

4.2. Datasets and Evaluation Metrics

Datasets Two benchmark driving datasets (KITTI [13] and H3D [34]) are used to evaluate the proposed approach comparing to self-generated baselines and state-of-the-art methods. The KITTI dataset was introduced for trajectory forecast in [23] to predict future motions of road agents in top-down view, and then [46] found their future locations in frontal view using this dataset. As in [23], we generate a set of trajectory segments with 6 sec long (2 sec for observation and 4 sec for prediction) using Road and City scenes in the Raw subset. We divide all videos into five sets and conduct 5-fold cross validation, following the split of [7]. In addition, the H3D [34] dataset is used to further validate the proposed approach on heterogeneous agents in highly congested urban environments. For evaluation, we divide 160 scenarios of H3D into the training (75%) and test set (25%) and use the same observation / prediction time as KITTI.

Metrics For the performance comparison, we mainly follow the standard evaluation metrics:

- *Average Distance Error (ADE)* is computed using L2 distance between the predicted trajectory and the ground truth for a certain time duration.
- *Final Distance Error (FDE)* shows L2 distance between the predicted location and the ground truth at a certain time step.

Both ADE and FDE are reported with 1 sec interval at future time steps. For multi-modal prediction, we sample 20 trajectories and find the best one with a minimum ADE at 4 sec in future. Note that the single- and multi-modal models are respectively denoted by a different suffix _S and _M.

In addition, we introduce a new metric that measures the rate of prediction success:

- *Success Rate (SR)* finds the fraction of scenarios where L2 distance between the predicted endpoint and ground truth is within a certain threshold value ϵ.

Under the assumption that the prediction would be successful if the error at the endpoint is within a certain threshold, this metric plots how many scenarios can be considered as 'success prediction'. SR thus is a more practical evaluation metric that tests the overall robustness of the algorithm.

Component			1.0 sec	2.0 sec	3.0 sec	4.0 sec
Env	Soc	Mul				
-	-	-	0.37 / 0.64	0.69 / 1.47	1.20 / 3.01	1.94 / 5.32
+	-	-	0.38 / 0.65	0.68 / 1.39	1.16 / 2.96	1.87 / 4.97
-	+	-	0.33 / 0.55	0.61 / 1.31	1.09 / 2.80	1.79 / 4.92
+	+	-	0.31 / 0.51	0.53 / 1.07	0.92 / 2.36	1.53 / 4.35
+	+	+ Fus	0.20 / 0.35	0.42 / 1.00	0.82 / 2.31	1.45 / 4.38
+	+	+ Emb	0.20 / 0.36	0.42 / 1.00	0.82 / 2.29	1.44 / 4.33

Table 1: Ablation study on the KITTI [13] dataset. ADE / FDE is reported in *meters*. Refer to Sec. 4.3 for description.

4.3. Ablative Study

We first demonstrate our design choices through ablative studies conducted in top-down view using KITTI. We evaluate the baseline models on the following components:

- **Env**: External features ($\mathbf{f}_S$ and $\mathbf{f}_T$);
- **Soc**: Social influences of other agents;
- **Mul**: Multi-modal learning. **Fus**: multi-modal fusion with feature aggregation, **Emb**: proposed shared cross-modal embedding.

Table 1 compares ADE and FDE of six baseline models that are designed by adding (+) or dropping (-) these components. When one or more of information is missing, a significant performance drop is observed. The error of the model without any components is particularly larger than others by a huge margin. By considering environmental influences (**Env**), the performance improves toward long-term prediction ($4sec$). It clearly demonstrates the effectiveness of the environmental constraints on more distant areas. We observe that adding **Soc** outperforms previous baselines, which implies the role of social behavior encoding for trajectory prediction with the significant improvement at the short-term time steps. The impressive error drop is found by taking both **Env** and **Soc** into account. It demonstrates the validity of the proposed feature extractor. We highlight the efficacy of the use of multiple sensor modalities from **Mul**, where both **Fus** and **Emb** further improve the performance. Interestingly, the proposed cross-modal embedding (**Emb**) achieves even lower error than fusion-based counterpart (**Fus**). It indicates that our model benefits from complementary input modalities with cross-modal embedding, even though a single input data is used during inference.

Additionally, we show the efficacy of the proposed regularizer for multi-modal prediction in the bottom of Table 2. Without the proposed regularizer (**S-CM_10 w/o reg**), the performance improvement of the model with 10 samples is minimal against single-modal prediction (**S-CM_1**), which is interpreted as a posterior collapse problem. However, the model with the regularizer (**S-CM_10**) highly improves the

Method	N	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art					
Const-Vel [40]	1	0.34 / 0.56	0.85 / 1.79	1.60 / 3.72	2.55 / 6.24
S-LSTM [1]	1	0.53 / 1.07	1.05 / 2.10	1.93 / 3.26	2.91 / 5.47
Gated-RN [7]	1	0.34 / 0.62	0.70 / 1.72	1.30 / 3.34	2.09 / 5.55
DESIRE [23]	1	- / 0.51	- / 1.44	- / 2.76	- / 4.45
DESIRE [23]	20	- / **0.28**	- / 0.67	- / 1.22	- / 2.06
S-GAN [15]	20	0.29 / 0.43	0.67 / 1.34	1.26 / 2.94	2.07 / 5.22
S-STGCNN [31]	20	0.21 / 0.36	0.38 / 0.70	0.59 / 1.31	0.82 / 2.14
Trajectron++ [39]	20	0.19 / 0.33	0.34 / 0.65	0.53 / 1.18	0.78 / 1.96
Ours					
S-CM_1	1	0.20 / 0.36	0.42 / 1.00	0.82 / 2.29	1.44 / 4.33
S-CM_10 w/o reg	10	0.20 / 0.35	0.40 / 0.96	0.77/ 2.06	1.33 / 4.04
S-CM_10	10	0.18 / 0.31	0.32 / 0.61	0.49/ 1.09	0.75 / 1.99
S-CM_20	20	**0.17** / 0.29	**0.29 / 0.53**	**0.42/ 0.83**	**0.61 / 1.57**

Table 2: Quantitative comparison (ADE / FDE in *meters*) of our approach with the state-of-the-art methods. The KITTI dataset [38] is used to predict trajectories in top-down view. N denotes the number of samples used.

Method	N	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art					
Conv1D* [44]	1	24.38 / 44.13	- / -	- / -	- / -
FVL* [46]	1	17.88 / 37.11	- / -	- / -	- / -
Const-vel [40]	1	5.88 / 9.42	13.23 / 26.03	22.13 / 45.99	31.90 / 68.03
S-GAN [15]	1	7.54 / 10.51	13.39 / 23.74	21.37 / 43.82	31.76 / 71.55
S-GAN [15]	20	6.96 / 9.58	12.25 / 21.42	19.48 / 39.66	28.89 / 65.02
Ours					
S-CM_1 w/o Emb	1	4.17 / 7.85	8.22 / 17.68	13.63 / 30.48	19.63 / 44.97
S-CM_1	1	4.17 / 7.35	8.21 / 17.64	13.59 / 30.41	19.59 / 44.92
S-CM_10	10	3.25 / 5.57	5.71 / 10.74	8.25 / 15.74	11.22 / 24.85
S-CM_20	20	**3.19 / 5.42**	**5.52 / 10.02**	**7.51 / 12.26**	**9.59 / 19.76**

Table 3: ADE / FDE is evaluated in $pixels$. The KITTI [13] dataset is used to predict trajectories in frontal view. $*$ denotes the evaluation on relative motion from ego-vehicle. N denotes the number of samples used.

accuracy generating diverse output responses. We conclude that the proposed regularizer can ease posterior collapse for future prediction.

4.4. Quantitative Results

We first compare the performance of the proposed approach with the state-of-the-art methods using KITTI. In Table 2, we observe from single-modal prediction (N=1) that our **S-CM_1** outperforms all compared single-modal approaches including social interaction oriented methods [1] as well as scene context oriented methods [23, 7]. For multi-modal prediction, the proposed approach (**S-CM_10**) with N=10 already achieves overall lower ADE and FDE than other competitors in top-down view trajectory forecast. By sampling N=20 modes, we improve FDE at 4.0 sec over 19% against [39].

Using the same cross-modal model, we examine the frontal view prediction capability in Table 3. Note that Conv1D [44] and FVL [46] predicts relative motion with respect to the future ego-motion. Their poor performance might be caused by the prediction difficulties with unknown ego-future. Although the proposed method (S_CM_1) further improved the accuracy without affecting the inference time, the effect seems less significant compared to that

Method	N	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art					
Const-Vel [40]	1	0.18 / 0.26	0.34 / 0.60	0.52 / 1.03	0.74 / 1.54
S-LSTM [1]	1	0.26 / 0.41	0.49 / 0.92	0.72 / 1.53	1.01 / 2.32
S-GAN [15]	1	0.27 / 0.37	0.45 / 0.77	0.68 / 1.29	0.94 / 1.91
S-GAN [15]	20	0.26 / 0.35	0.44 / 0.72	0.65 / 1.24	0.90 / 1.84
Gated-RN [7]	20	0.18 / 0.32	0.32 / 0.64	0.49 / 1.03	0.69 / 1.56
STGAT [19]	20	0.24 / 0.33	0.34 / 0.48	0.46 / 0.77	0.60 / 1.18
S-STGCNN [31]	20	0.23 / 0.32	0.36 / 0.52	0.49 / 0.89	0.73 / 1.49
Trajectron++ [39]	20	0.21 / 0.34	0.33 / 0.62	0.46 / 0.93	0.71 / 1.63
EvolveGraph [26]	20	0.19 / 0.25	0.31 / 0.44	0.39 / 0.58	0.48 / 0.86
Ours					
S-CM_1	1	0.14 / 0.25	0.27 / 0.54	0.43 / 0.95	0.62 / 1.45
S-CM_10	10	0.12 / 0.21	0.21 / 0.37	0.30 / 0.61	0.42 / 0.96
S-CM_20	20	**0.11 / 0.19**	**0.18 / 0.30**	**0.25 / 0.46**	**0.34 / 0.77**

Table 4: Quantitative results (ADE / FDE) are reported in *meters*. We use H3D [34] to evaluate the proposed method in top-down view. N denotes the number of samples used.

Method	N	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art					
Const-vel [40]	1	13.15 / 19.22	24.64 / 44.13	38.18 / 74.75	53.38 / 110.07
S-GAN [15]	1	12.91 / 17.05	20.57 / 33.53	29.70 / 54.41	40.71 / 84.51
S-GAN [15]	20	12.38 / 16.26	19.67 / 31.86	28.30 / 51.55	38.88 / 80.55
Ours					
S-CM_1	1	8.69 / 16.06	16.52 / 33.25	25.68 / 54.91	36.29 / 82.05
S-CM_10	10	6.62 / 11.36	10.69 / 18.12	14.25 / 24.51	18.22 / 36.92
S-CM_20	20	**6.25 / 10.51**	**9.61 / 15.26**	**12.07 / 18.66**	**15.05 / 30.56**

Table 5: Our approach is evaluated on ADE / FDE (in *pixels*) using the H3D [34] dataset. The proposed absolute motions (with ego-future elimination) are used to compute errors in frontal view. N denotes the number of samples used.

shown in top-down view (Table 1). Our insight is as follows: (i) the use of complementary features obtained from different input modality is not as impactful as it was for top-down prediction; and (ii) the performance improvement achieved by other aspects (*e.g.*, social behavior, semantic context, etc.) is already exceptional in frontal view, which makes the improvement with embedding less compelling. Nevertheless, the proposed method with cross-modal embedding generally shows higher accuracy against others.

We further evaluate our work using the H3D dataset in its highly congested environments. In top-down view as in Table 4, we found that the proposed model with a single sample (**S-CM_1**) already achieves the lower error than most of methods including very recent graph-based models ([31] w/o scene and [39] w/ scene). Our explicit modeling of relational interactions together with cross-modal embedding enables us to explore more discriminative behavior representations over these graph-based methods. The performance is further improved by sampling multiple predictions with the regularizer (**S-CM_20**). Compared to the best state-of-the-art method [26] that finds the dynamic evolution of interactions, our work improves the performance over 10% on FDE at 4.0 sec. Such lower errors demonstrates the generation of highly diverse yet acceptable future motions using our model, considering the road topology.

Subsequently, we evaluate our trajectory prediction framework for the task of frontal view forecast. In Table 5,

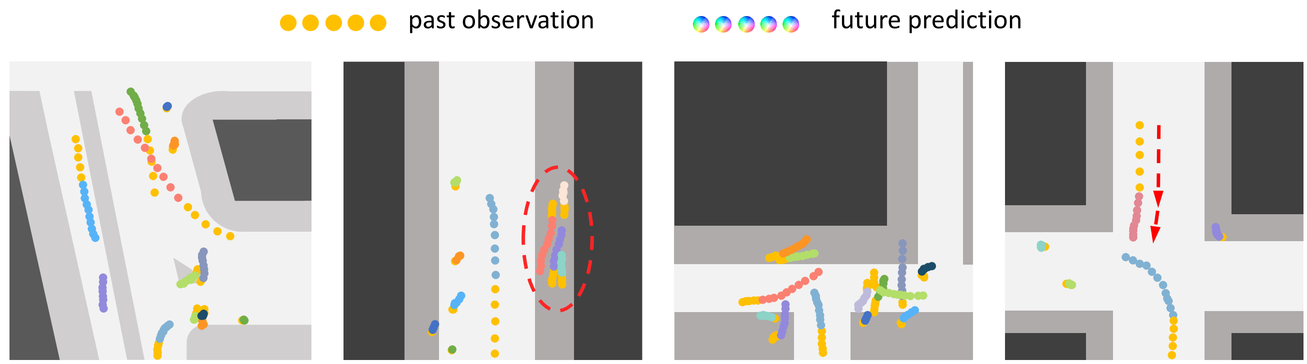
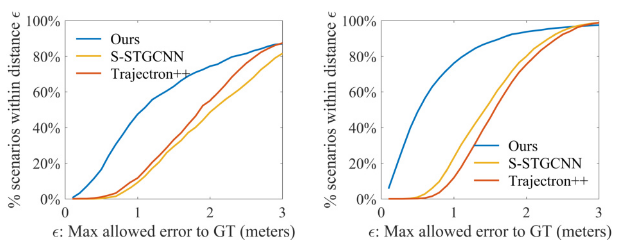


Figure 3: We visualize the top-1 prediction from highly interactive scenarios among heterogeneous traffic agents, such as human-human, human-vehicle, and vehicle-vehicle interactions.



(a) (b)

Figure 4: The success rate (SR) is plotted as a fraction of success prediction scenarios with respect to maximum allowed distance error, which is indicative of overall robustness of algorithms. SR is evaluated in top-down view on FDE at 4.0 sec. (a) KITTI and (b) H3D.

we observe that the performance of our single-modal prediction model (**S-CM_1**) is on par with multi-modal prediction model of Social-GAN (**S-GAN** with N=20). It implies that the prediction capability of the proposed framework is being at the level of the state-of-the-art. The significant improvement of error from our multi-modal prediction model **S-CM_20** further demonstrates the effectiveness of our objective function for optimization.

4.5. Evaluation with Success Rate

The standard evaluation metrics such as ADE and FDE do not capture the success or failure of predictions. We thus introduce SR that plots the proportion of scenarios that can be considered as 'successful prediction' with respect to the definition of *success*. We use the error threshold ϵ on the x-axis and measure the rate of success scenarios by FDE at 4.0 sec. Figure 4 compares our approach with two state-of-the-art methods [31, 39]. We observe from 4a that our approach performs better than others in terms of the correctness of predictions. Assuming that the real driving application is designed with a small prediction tolerance ($\epsilon = 1.5m$), our model is more reliable and credible with considerably higher success rate (63% compared to [39] of 33% or [31] of 29%). We also plot SR using the H3D dataset in 4b, which indicates that our prediction model can achieve much smaller errors in the majority of scenarios. Our method shows consistently higher success rate, validating the robustness of our prediction capability.

4.6. Qualitative Results in Top-down View

Figure 3 visualizes the top-1 prediction result of the proposed approach. Each scenario contains the heterogeneous agents (*i.e.*, cars, bus, pedestrians, cyclist, etc.) interactive one to another. We robustly forecast their future motions by taking advantages of the proposed social behavior modeling and cross-modal embedding. In between pedestrians, our approach models their motion behaviors and generates socially acceptable trajectories (dotted oval in the second column). In the last column, our model predicts that the car would turn left, which influences the behavior of on-coming vehicle that slows its speed (*i.e.*, yielding; dotted arrow). We conclude that the proposed graph accordingly considers relational interactions while predicting future motions. We provide the results of 20 prediction samples as well as qualitative results in frontal view in the supplementary material.

5. Conclusion

We proposed a solution to future trajectory forecast in driving scenarios. Assuming that the multiple sensory data is available for autonomous driving, our approach can benefit from the model trained using multiple input modalities. First, the GNN-based feature encoder extracts social behaviors of the target agent, considering its interactions toward all other traffic agents as well as surrounding road structures. Then, the relational behaviors obtained from multiple perspectives are embedded into a shared cross-modal latent space. We provided its derivation that jointly optimizes objective functions using the generative variational models. Finally, we designed an auxiliary regularizer to ease the posterior collapse problem for future prediction. We analyzed the significance of the proposed approach through the extensive evaluation, showing the improvement of the performance against the state-of-the-art methods.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2303–2314, 2017.

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[4] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4202, 2018.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[7] Chiho Choi and Behzad Dariush. Looking to relations to future trajectory forecast. In *Computer Vision (ICCV), 2019 IEEE International Conference on*. IEEE, 2019.

[8] Chiho Choi, Sangpil Kim, and Karthik Ramani. Learning hand articulations by hallucinating heat distribution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3104–3113, 2017.

[9] Chiho Choi, Srikanth Malla, Abhishek Patil, and Joon Hee Choi. Drogon: A trajectory prediction model based on intention-conditioned behavior reasoning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2020.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] Nachiket Deo and Mohan M Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1179–1184. IEEE, 2018.

[12] Davi Frossard and Raquel Urtasun. End-to-end learning of multi-sensor 3d tracking by detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 635–642. IEEE, 2018.

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[14] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

[15] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[17] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[18] Xin Huang, Stephen McGill, Brian C Williams, Luke Fletcher, and Guy Rosman. Uncertainty-aware driver trajectory prediction at urban intersections. *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[19] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6272–6281, 2019.

[20] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125. IEEE, 2016.

[21] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, pages 137–146, 2019.

[22] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[23] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[24] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. In *2019 IEEE Conference on Robotics and Systems (IROS)*, 2019.

[25] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Interaction-aware multi-agent tracking and probabilistic behavior prediction via adversarial learning. In *2019 IEEE In-*

ternational Conference on Robotics and Automation (ICRA). IEEE, 2019.

[26] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. In *Advances in neural information processing systems*, 2020.

[27] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[28] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. 2019.

[29] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020.

[30] Srikanth Malla, Isht Dwivedi, Behzad Dariush, and Chiho Choi. Nemo: Future object localization using noisy ego priors. *arXiv preprint arXiv:1909.08150*, 2019.

[31] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[32] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[33] Seong Hyeon Park, ByeongDo Kim, Chang Mook Kang, Chung Choo Chung, and Jun Won Choi. Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1672–1678. IEEE, 2018.

[34] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE, 2019.

[35] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *IJCAI*, pages 3846–3853, 2016.

[36] Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018.

[37] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2830, 2019.

[38] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[39] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[40] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

[41] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, pages 1–47, 2019.

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

[43] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5275–5284, 2018.

[44] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on multimedia*, pages 978–987. ACM, 2016.

[46] Yu Yao, Mingze Xu, Chiho Choi, David J Crandall, Ella M Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[47] Dong Yi, Zhen Lei, and Stan Z Li. Shared representation learning for heterogenous face recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.

[48] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.

[49] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.

[50] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2365–2374, 2019.