Trajectory Prediction in Heterogeneous Environment via Attended Ecology Embedding

Wei-Cheng Lai*
National Chiao Tung University
sean0514.eed04@nctu.edu.tw

Zi-Xiang Xia*
CyberLink
Vincent_Xia@cyberlink.com

Hao-Siang Lin
National Chiao Tung University
mrfish.eic08g@nctu.edu.tw

Lien-Feng Hsu
National Chiao Tung University
liyang.ee05@nctu.edu.tw

Hong-Han Shuai
National Chiao Tung University
hhshuai@nctu.edu.tw

I-Hong Jhuo
IBM
ihjhuo@gmail.com

Wen-Huang Cheng
National Chiao Tung University
whcheng@nctu.edu.tw

ABSTRACT

Trajectory prediction is a highly desirable feature for safe navigation or autonomous vehicle in complex traffic. In this paper, we consider the practical environment of predicting trajectory in the heterogeneous traffic ecology. The proposed method has various applications in trajectory prediction problems and also in applied fields beyond tracking. One challenge stands out of the trajectory prediction—heterogeneous environment. Particularly, many factors should be considered in the environments, i.e., multiple types of road-agents, social interactions and terrains. The information is complicated and large that may result in inaccurate trajectory prediction. We propose two social and visual enforced attention modules to circumvent the problem and a variant of an Info-GAN structure to predict the trajectory with multi-modal behaviors. Experimental results show that the proposed method significantly outperforms state-of-the-art methods in both heterogeneous and homogeneous real environments.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Computer systems organization** → **Robotic autonomy**.

KEYWORDS

trajectory prediction; trajectory forecasting; autonomous driving

ACM Reference Format:
Wei-Cheng Lai, Zi-Xiang Xia, Hao-Siang Lin, Lien-Feng Hsu, Hong-Han Shuai, I-Hong Jhuo, and Wen-Huang Cheng. 2020. Trajectory Prediction in Heterogeneous Environment via Attended Ecology Embedding. In *Proceedings of 28th ACM International Conference on Multimedia (MM '20).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413602

*Both authors contributed equally to the paper

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413602

1 INTRODUCTION

Predicting trajectories with multiple road-agents (*e.g.*, humans, vehicles, and bicycles) has been prominent in multimedia system, which requires an environment-understanding intelligent system. An accurate trajectory prediction facilitates a wide range of applications, e.g., detecting suspicious activity in public areas, managing the traffic congestion, providing safe navigation for autonomous driving or social robots [35, 40]. Conventional methods focus on using egocentric features, e.g., previous trajectory, for trajectory prediction, which lacks considering some surrounding information, i.e., multi-agents, in traffic environments. For example, some works in trajectory prediction mainly focus on road-agents in homogeneous environments [2, 4, 10, 11, 21, 30, 31, 34, 39], which only consists of a single type of road-agent in a scene. However, in real driving environment, it is necessary to differentiate the interactions among different types of road-agents such as the difference between pedestrians and bikes or bikes and trucks. Therefore, understanding the interactions among heterogeneous agents may play an important key in inferring trajectories.

Unfortunately, there is no straightforward way to fuse the interactions among heterogeneous agents. It is a challenging task due to the following facts: (1) *Complicated interactions between different road-agents.* When facing cars or pedestrians on the road, people would speculate their motion behaviors to avoid collisions. For example, a car may stay away from an aggressive bike rider, or a person may walk to someone who waves his hand. Therefore, this oftentimes introduces numerous variables and make the trajectory prediction more difficult [23]. (2) *Various interactions between road-agents and scene.* Different traffic scenes usually bring more diverse and complicated interactions to road-agents. For example, it is normal if the pedestrians walk on the sidewalk, while it is unusual for a truck driving on the sidewalk. Besides, trajectories often depend on spatial constraints. On the other hand, road-agents tend to follow the terrain in the traffic scene to avoid obstacles and collisions [30]. (3) *Different possible path generation.* When facing different traffic situations, people tend to consider numerous possible trajectory

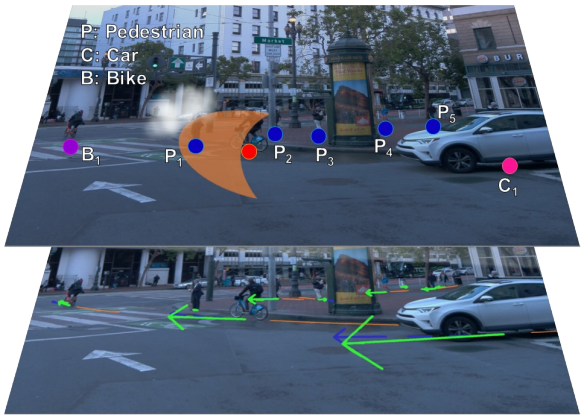

Figure 1: Illustration of the proposed model AEE-GAN . (top) the white portion indicates the attended region. The pedestrians in the horizon region (in orange region) gain more attention than other road-agents; (bottom) Three trajectories show in a traffic scene—the prediction result, the observed trajectory and the ground truth with the colors of green, orange and blue, respectively.

choices for determining the control action according to predefined criteria, e.g., risk minimization. Thus, the model should be able to predict a range of plausible trajectories, which is challenging for conventional prediction models [11].

Existing endeavors have demonstrated the strong correlations between inferring the trajectories and the heterogeneous environments [6, 8, 23]. For example, TrafficPredict [23] assumes that the same kind of road-agents shares similar behavior characteristics; thus they introduce an LSTM-based 4D Graph, which an instance layer and a category layer are constructed to capture the similarity between different individuals. TraPHic [6] models the different classes by taking into account the size of road-agents, and they introduce the convolutional LSTM model with the idea of horizon modeling, i.e., prioritizing the interactions in front of each road-agent. Nevertheless, the second and third challenges remain unsolved for the heterogeneous trajectory prediction task.

In addition, safety is an essential issue of autonomous navigation. It is difficult to reassure the passengers based on the number of performance from a trajectory prediction system. Therefore, besides pursuing high accuracy at the same time, system explainability [41] provides an essential component to convincing people during the process. Specifically, an attention mechanism is introduced in our model architecture design due to its enhancement toward the model and the well-explainability for prediction.

Based on these observations, in this paper, we propose a novel trajectory prediction model, namely, Attended Ecology Embedding-based Generative Adversarial Networks (*AEE-GAN*), to predict the paths for each road-agent in heterogeneous environment. Specifically, we first embed the types of road-agents into the input state space and utilize the visual information to implicitly learn the behaviors and interactions. Recurrent Visual Attention Enforcement (RVAE) is further proposed to extract and update related visual embedding from the ecology. Inspired by the recent success of the self-attention mechanism that encodes the relationship across image regions [38], RVAE first leverages the self-attention mechanism to encode the long-range dependencies for capturing contextual visual embedding. The long-range dependencies improve the ability to model the interaction between road-agents and the scene. Moreover, the predicted trajectory is recurrently fed into RVAE to update the visual embedding for the next prediction. In addition to the visual embedding, we further propose Social Enforcement (SE) to extract embedding of the interaction between road-agents. SE applies both soft attention and horizon attention to handle complicated interactions. Finally, after deriving the attended embedding, inspired by [4], we utilize the Info-GAN structure to predict the output trajectories. Different from [4], the proposed AEE-GAN provides the feedback to update the visual embedding. The main contributions of this paper are summarized in the following:

- We propose an attention-based model, Attended Ecology Embedding-based Generative Adversarial Networks (AEE-GAN) by considering both of the social interactions and visual information from the from traffic scene and agents in the practical environment.
- The attention mechanism is widely used in our model for the performance improvement. First, the recurrent visual attention extracts and updates visual information. Second, the proposed social enforcement utilizes the horizon attention to concentrate on important interactions and the attention weights shows the impact of the neighboring agents.
- Experimental results on 7 real benchmarks demonstrate that the proposed model outperforms state-of-the-art models by at least 23.5% and 45% in heterogeneous and homogeneous environments, respectively. These results with future updates can be reproduced by our released code.[1]

2 PRELIMINARIES

2.1 Problem Definition

Given the previous states of road-agents and scene information, the goal of trajectory prediction is to estimate the 2D positions of all road-agents in the future. Specifically, let I_t denote the input of scene information at time t, e.g., top-view image or front-view image. Furthermore, the state of road-agent i at time t is denoted as $X_t^i = (p_t^i, v_t^i, c^i)$, where $p_t^i = (x_t^i, y_t^i)$ and $v_t^i = (v_{x,t}^i, v_{y,t}^i)$ represent the 2D position and 2D velocity, respectively, and c^i denotes the category of the road-agent i, e.g., pedestrians, bicycles. Moreover, let τ denote the observation period. The input states of the N road-agents from the time steps $t = 1, \cdots, \tau$ can be represented as:

$$X_{1:\tau}^i = \{(p_t^i, v_t^i, c^i) | t = 1, ..., \tau\}, \quad 1 \le i \le N.$$

Given a user-specified prediction period T, the output is the coordinates of 2D trajectory prediction for each road-agent i between frames $\tau + 1$ and $\tau + T$, denoted as $\hat{Y}_{\tau+1:\tau+T}^i$, i.e.,

$$\hat{Y}_{\tau+1:\tau+T}^i = \{(x_t^i, y_t^i) | t = \tau + 1, ..., \tau + T\}, \quad 1 \le i \le N.$$

Therefore, the goal is to design a function f with the parameters Θ that can correctly predict the trajectories of each road-agent i

[1] https://github.com/ego2eco/AEE-GAN

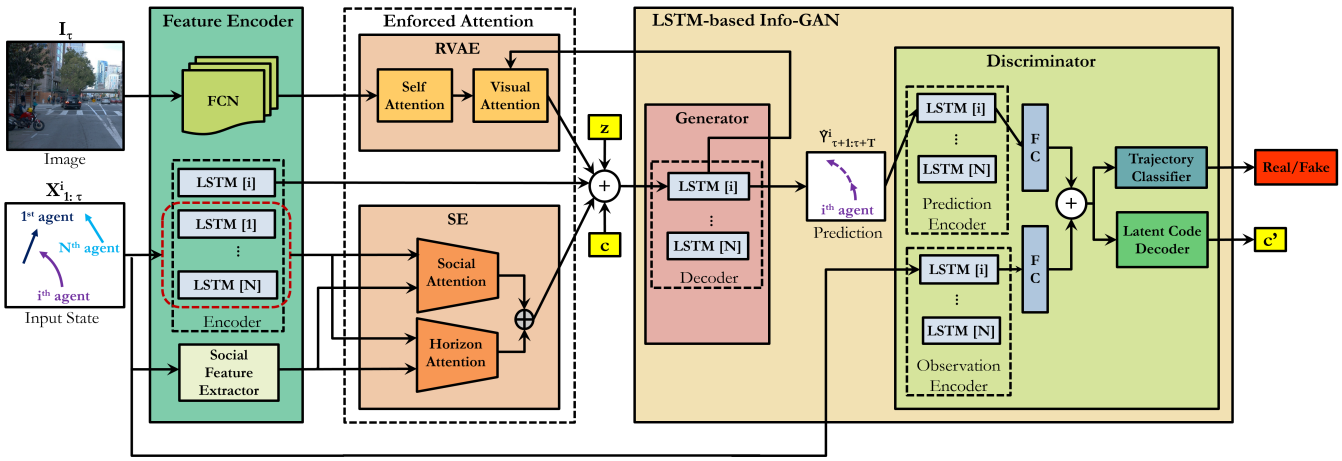

Figure 2: An overview of our proposed AEE-GAN architecture. AEE-GAN consists of three key modules: Feature Encoder, Enforced Attention and LSTM-based Info-GAN. Enhanced Attention module consists of Recurrent Visual Attention Enforcement (RVAE) and Social Enforcement (SE). In addition to the adversarial loss calculation in GAN structure, c and c' is leveraged to calculate the information loss.

between $\tau+1$ and $\tau+T$, i.e.,

$$\hat{Y}^{i}_{\tau+1:\tau+T} = f(X^{i}_{1:\tau}, \cup_{j\neq i} X^{j}_{1:\tau}, I_{\tau}; \Theta).$$

2.2 Related Work

Forecasting the trajectory of traffic objects in the dynamic scene has been widely-studied in the last decade. Several approaches for trajectory forecasting mainly aim at exploiting previous trajectories of individuals to compute the future motion, e.g., Kalman Filters [14] and Particle Filters [25]. However, these approaches only consider the individual trajectories and do not deal with complex environments such as traffic intersections and heterogeneous traffic objects. To deal with the complex environment, some of the previous studies take the scene information into account for predicting trajectories that comply the physical environment. Previous works [5, 15, 16, 28, 37, 42] made use of scene information such as map, image and semantic feature, to represent the complete and global information of the environment, allowing the model to predict more plausible and realistic trajectories. However, all above methods ignore the influence of the other agents' information in the scene on predicting the path for the road-agent. [12] introduces "social forces" to capture human motion with attractive and repulsive forces. This model has been further extended by many researchers [3, 24, 29, 36]. Nevertheless, most of these methods use handcrafted rules/features to explicitly formulate the motion of road-agents in the context, which may cause the limitation of their ability to generalize in a complex scene.

With the advance of deep learning technology, many works based on deep neural networks are proposed to directly learn the interaction between road-agents from the data. One of the popular deep neural networks for forecasting sequential behaviors is Recurrent Neural Nets [2, 11]. For example, Social-LSTM [2] constructs LSTMs for each person in the scene and a social pooling layer to aggregate the interactions of the neighboring humans. However, it limits the number of considered pedestrians within a predefined spatial grid. Therefore, Social-GAN [11] takes the influences from all road-agents in the traffic scene into account and uses the GAN model to generate different feasible trajectories. Nevertheless, instead of properly preserving the multi-modal behavior, the results are usually with high variances since it generates trajectories from the input noise. Thus, Social Ways [4] applies Info-GAN, which introduces latent code to enhance the multi-modality of prediction and computes the attentive social features to generate a more convincing result. Social-BiGAT [17] relies on BiGAN architecture to help reduce the variance of the predicted trajectory distributions and allow for better generalization. However, these methods focus on the trajectory prediction in homogeneous environment without considering the types of road-agents.

For heterogeneous environments, TrafficPredict [23] proposes to construct a 4D Graph for representing instances, their interactions, temporal information, and categorization of road-agents, and they apply LSTMs to predict the trajectory based on the 4D Graph. However, the LSTMs fall short of capturing the spatial dependencies between the encoded features from each individual. Therefore, TraPHic [6] uses Convolutional Social Pooling LSTM [10] to learn the locally useful interactions. Nevertheless, when the number of interactions becomes large, e.g., prediction on homogeneous datasets (many pedestrians), the performance significantly degrades due to the unnecessary interaction modeling. In addition to capturing the related features between the past trajectories, SoPhie [30] computes the attentive interaction from the trajectory and visual embedding derived from convolution layers. To the best of our knowledge, this is the first work that exploits the enforced visual attention to help model the long-range dependencies across visual regions for trajectories prediction. In addition, by combining the enforced social attention, our model learns to attend both socially and visually important information in the scene, leading our model to achieve state-of-the-art performance on both homogeneous and heterogeneous datasets.

3 AEE-GAN

To successfully predict trajectories of road-agents in a complicated environment, it is necessary to consider different kinds of information, e.g., the state and intention of other road-agents, surrounding obstacles. Current approaches predict the trajectory of a road-agent by attending to the interactions among road-agents in the scene. Nevertheless, the road-agents in most scenarios do not react to all the other road-agents around them, instead, they selectively concentrate on key interactions in front of them. Therefore, similar to humans, attention is a limited resource, and the model requires only paying attention to where is important for making the decision. To better model these, we propose a novel method for focusing on the salient features of the road-agents. Specifically, Fig. 2 illustrates the proposed model, which is comprised of three main modules: Feature Encoder, Enforced Attention and LSTM-based Info-GAN. Feature Encoder extract 1) the semantic information from the given scene at the last frame I_τ with a fully convolutional neural network, 2) the geometric social features between each road-agent for evaluating the influence of each road-agent [4], and 3) the state history of each road-agent, $X^i_{1:t_\tau}$ for modeling the temporal features along trajectories. Afterward, Enforced Attention is composed of two attention-based modules, namely, Recurrent Visual Attention Enforcement (RVAE) and Social Enforcement(SE). RVAE leverages the self-attention mechanism to provide refined scene features and further concentrates on salient spatial information from the refined visual feature and the predicted trajectory. On the other hand, SE makes use of the soft attention module to model the interactions between road-agents and enforce the attention on the key interaction in the front region of the road-agent. Finally, the LSTM-based Info-GAN module integrates all the feature vectors to predict multiple plausible trajectories for each road-agent and feedback to RVAE for better estimating the attention.

3.1 Feature Encoder

To obtain the information from the traffic scene, Feature Encoder first extracts the visual features at time t, denoted as v_t, through a Fully Convolutional Network (FCN). Here, we leverage the image from the last frame I_τ for extracting visual features as follows:

$$v_\tau = FCN(I_\tau; W_{FCN}), \tag{1}$$

where $FCN(\cdot)$ and W_{FCN} represent the Fully Convolutional Network and corresponding weights, respectively. Moreover, we use VGG-16 [33] as backbone and modify it into a fully convolutional network [22] since it has been proved to be effective for extracting semantic features and geometry contour [20]. W_{FCN} is initialized with the pre-trained weights on ImageNet and fine-tuned on the task of semantic segmentation on Cityscapes dataset [9].

Moreover, to capture the temporal information along the past trajectories, we use one LSTM layer to integrate the motion sequence of each road-agent. The motion feature of road-agent i at time t is denoted as $M_t(i)$, i.e.,

$$M_t(i) = LSTM_{enc}(h^{enc}_t(i), X^i_t; W_{enc}), \tag{2}$$

where $LSTM_{enc}$ denotes as the LSTM Encoder, $h^{enc}_t(i)$ is the hidden state in the LSTM Encoder for road-agent i at time t, $t \in [1, \tau]$.

In addition to visual features and individual trajectory features, we further extract the social features, i.e., the interaction between different road-agents, to better model the behaviors. Specifically, similar to [4], we extract three features as the social features between road-agent i and j, denoted as $s^{i,j}$ by concatenating: (1) the Euclidean distance between road-agents i and j, (2) the bearing angle between road-agent i and j, and (3) the speculation distance between road-agent i and j.

3.2 Enforced Attention

After encoding the features, Recurrent Visual Attention Enforcement (RVAE) and Social Enforcement (SE) are proposed to derive the attended ecology embedding by attending the visual and social features, respectively.

Recurrent Visual Attention Enforcement. A problem of CNN is about its receptive field, a stick object that crosses the long-range of the image is hard to recognize. But the helpful information in the traffic scene almost are come from long-range objects. e.g., The road that constrains vehicle driving, the sidewalk that people usually walking on. Therefore, we exploit the self-attention mechanism [38] to model long-range visual features and enrich the multi-level dependencies toward the image region. Specifically, there are R locations in the visual features v_τ. Let v^p_τ and v^q_τ denote the C-dimension visual features of p^{th} and q^{th} locations, respectively, i.e., $v^p_\tau, v^q_\tau \in \mathbb{R}^C$. We first compute the self-attention weight of v^p_τ with regard to v^q_τ, denoted as $\alpha^{q,p}_{self}$, by projecting v^p_τ and v^q_τ into two feature spaces with two functions ϕ_1 and ϕ_2 [2], i.e.,

$$\alpha^{q,p}_{self} = \frac{exp(\phi_1(v^p_\tau) \cdot \phi_2(v^q_\tau))}{\sum_{p=1}^{R} exp(\phi_1(v^p_\tau) \cdot \phi_2(v^q_\tau))}. \tag{3}$$

We derive the attended feature o^q_τ by using the self-attention weight to summarize transformed visual features, and re-projecting it to the same dimension of the original visual features by ϕ_3, i.e.,

$$o^q_\tau = \phi_o\left(\sum_{p=1}^{R} \alpha^{q,p}_{self} \phi_3(v^p_\tau)\right), \tag{4}$$

where $\phi_3 : \mathbb{R}^{C'} \to \mathbb{R}^C$ and $\phi_o : \mathbb{R}^C \to \mathbb{R}^{C'}$ also are the learnable projection and implemented through 1×1 convolutions. Finally, the attended visual features of q^{th} are derived by multiplying a scaling factor ω to o^q_τ and add back to the input feature map, i.e., $\omega o^q_\tau + v^q_\tau$ where ω is a learnable parameter which is initialized as 0. We further encode the attended visual features through a convolution layer and a fully connected layer, and denoted as $\tilde{v}_\tau$.

After enriching the global information of the visual features, the enhanced visual features can be used to facilitate the trajectory prediction from time $\tau + 1$ to $\tau + T$. However, the enhanced visual features are not updated during time $\tau + 2$ to $\tau + T$, which might cause improper attention if the road-agent is far from the original attended region at time $\tau + T$. To update the related visual region with the trajectory prediction, we further recurrently compute the attention between the predicted results and the visual features. Specifically, the input attention takes the hidden state of the LSTM Decoder from generator ($h^{dec}_{t-1}(i)$) and the features after the self-attention module ($\tilde{v}^l_\tau$) to compute the attention coefficient. The

[2]Here, $\phi_1(v^p_\tau) = W_{\phi_1} v^p_\tau$ and $\phi_2(v^p_\tau) = W_{\phi_2} v^q_\tau$, where $W_{\phi_1} \in \mathbb{R}^{C' \times C}$, $W_{\phi_2} \in \mathbb{R}^{C' \times C}$ are two learnable matrices, and C' is the transformed dimension.

attention coefficient of visual features $\tilde{v}_\tau^l$ for predicted trajectory of road-agent i at time t, denoted as $\alpha_{v2p,t}^{i,l}$, is derived as follows:

$$\alpha_{v2p,t}^{i,l} = \frac{exp\left(h_{t-1}^{dec}(i) \odot W_{v2p}\tilde{v}_\tau^l\right)}{\sum_{l=1}^{L} exp\left(h_{t-1}^{dec}(i) \odot W_{v2p}\tilde{v}_\tau^l\right)}, \tag{5}$$

where $\odot$ represents Hadamard product operation, and W_{v2p} denotes the linear mapping of the visual features to predicted trajectory features. Finally, we multiply the weighted vector with the embedded visual features to obtain the attended visual ecology embedding for road-agent i at time t, i.e., $E_{vis,t}(i) = (E_{vis,t}^1(i), E_{vis,t}^2(i), ..., E_{vis,t}^l(i), ...E_{vis,t}^L(i))$, where $E_{vis,t}^l(i) = \alpha_{v2p,t}^{i,l}\tilde{v}_\tau^l$.

It is worth noting that the RVAE not only focuses on the relevant scene context in regard to road-agent i, but also enriches the heterogeneous modeling. The shape and type of the related road-agents in visual features are captured to specify their different interactions, which is suitable for both homogeneous and heterogeneous environments.

Social Enforcement. The interactions between road-agents also take part in the trajectory prediction. Similar to [30], we model the interaction between road-agent i and road-agent j in the traffic scene by applying soft attention in Social Attention module. Specifically, the attention coefficient of social features for road-agent i to the other road-agent j at time t, denoted as $\alpha_{soc,t}^{i,j}$, is derived as follows:

$$\alpha_{soc,t}^{i,j} = \frac{exp(s^{i,j} \cdot W_{p2s}h_t^{enc}(j))}{\sum_{k\neq i} exp(s^{i,k} \cdot W_{p2s}h_t^{enc}(k))}, \tag{6}$$

where W_{p2s} denotes the linear mapping of the predicted trajectory features to the social features.[3] This attention computes all the other road-agents in the traffic scene, even for the road-agents far away from road-agent i, which is too complicated to be well-learned. Therefore, to place more emphasis on the interaction in front of road-agent i, we further calculate the Horizon Attention, i.e., only considering the road-agent set $H_d^t(i)$, which contains the road-agents in a semicircle region in front of road-agent i with a radius of d meters at time t. The horizon attention coefficient of social features for road-agent i to the other road-agent j at time t, denoted as $\alpha_{hor,t}^{i,j}$, is derived as follows:

$$\alpha_{hor,t}^{i,j} = \begin{cases} \dfrac{exp(s^{i,j} \cdot W_{p2s'}h_t^{enc}(j))}{\sum_{k\neq i} exp(s^{i,k} \cdot W_{p2s'}h_t^{enc}(k))}, & \text{if } j \in H_d^t(i), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where $W_{p2s'}$ denotes the new linear mapping of the predicted trajectory features to the social features within the horizon. Finally, to enrich the information from the road-agents in horizon region, we combine the social and horizon attention information. Instead of concatenating the outputs, we use entrywise sum adding the horizon weighted features on the original social features to refine the attentive information of the horizon agents. The attended social ecology embedding for road-agent i to other road-agent j at time t is obtained by

$$E_{soc,t}^j(i) = ((\alpha_{soc,t}^{i,j} + \alpha_{hor,t}^{i,j})h_t^{enc}(j)). \tag{8}$$

We concatenate the N attended social ecology embedding of road-agent i to derive $E_{soc,t}(i)$.

3.3 Trajectory Prediction

Trajectory prediction is known as a multi-modal problem, i.e., given a trajectory history, the prediction should be able to generate multiple plausible trajectories since people tend to have many possible trajectory choices. A recent line of studies leverages the natural ability of GAN and proposes to use the input noise z together with the latent features of trajectory history to generate several possible trajectories [11, 30]. Nevertheless, it does not impose any restrictions on the generator, which may generate undesirable predictions, e.g., mode-collapsing trajectories when the generator tends to ignore z. Inspired by [4], we leverage the Info-GAN architecture by additionally taking a latent code c as input and making c meaningful via maximizing a lower bound of the mutual information between the distribution of c and the distribution of the generated trajectory. As such, the random values of c are inclined to change the predicted trajectories and alleviate the mode collapse. It is worth noting that the Info-GAN architecture here is different from previous work [4] since we feedback the predicted trajectory to RVAE for updating of visual features. As such, the attended visual ecology embedding are up-to-date for a better prediction.

Generator (G). Our generator uses one LSTM layer, $LSTM_{dec}$, to predict the future location of the road-agent. Specifically, the input vector to our generator is concatenated as follows:

$$K_t(i) = concat(h_t^{enc}(i), E_{vis,t}(i), E_{soc,t}(i), z, c). \tag{9}$$

That is, we take the noise vector z sampled from a multivariate normal distribution and a coding variable c as input while conditioning the Generator on the hidden states of $LSTM_{enc}$, the attended ecology visual embedding $E_{vis,t}(i)$ and attended ecology social embedding $E_{soc,t}(i)$. Accordingly, the generated future state for each road-agent is obtained by

$$\hat{Y}_t(i) = LSTM_{dec}(h_t^{dec}(i), K_t(i); W_{dec}). \tag{10}$$

Discriminator (D). The discriminator consists of one LSTM layer along with one FC layer for encoding the past trajectory and another LSTM layer along with one FC for encoding the future trajectory, which is either sampled from the ground truth or predicted by the generator. We concatenate the output of these two encoded vectors and use it to 1) classify whether the future trajectory is fake or real, and 2) reconstruct the latent code c' that maximizes the mutual information (please refer to Fig. 2).

Losses. Besides the adversarial loss function, we train another sub-network $Q(c|X_{1:\tau})$ with parameters θ_Q to evaluate the mutual information between the latent code c and the generated samples. As a result, the overall objective function V of our model is:

$$\begin{aligned} &min_{\theta_G,\theta_Q} max_{\theta_D} V(\theta_G, \theta_Q, \theta_D) = \\ &\quad \mathbb{E}[logD(Y_{\tau+1:\tau+T}^i | X_{1:\tau}^i; \theta_D)] + \\ &\quad \mathbb{E}[log(1 - D(G(z|h_t^{enc}(i), E_{vis,t}(i), E_{soc,t}(i); \theta_G); \theta_D))] - \\ &\quad \lambda\mathbb{E}[logQ(c|G(z|h_t^{enc}(i), E_{vis,t}(i), E_{soc,t}(i); \theta_G), \theta_Q)]. \end{aligned} \tag{11}$$

[3] $s^{i,j}$ can be updated through time but the computation is expensive when the number of road-agents is large.

4 EXPERIMENTS

In this section, we compare the performance of AEE-GAN against the various baselines on two heterogeneous datasets: Waymo [1] and Stanford drone dataset [29], as well as on two commonly-used homogeneous datasets, *i.e.,* ETH [27] and UCY [19]. Moreover, a qualitative analysis of the trajectories generated by our AEE-GAN is provided to demonstrate the effectiveness of the proposed RVAE and SE. For the implementation details and more results, please refer to the following link: https://ego2eco.github.io/.

Datasets. Two real datasets are used to evaluate the performance of AEE-GAN in heterogeneous environments. The first one is Waymo dataset [1] which is recently released and consists of front-view images of various real-world road-agents (i.e., vehicles, pedestrians and bikes) navigating in a wide range of environments, from dense urban centers to suburban landscapes, as well as data collected in different weather. The coordinates of Waymo dataset are provided in the world coordinate. To construct the dataset for trajectory prediction, similar to [29], we observe trajectories for 2.4 seconds and predict for 3.6 seconds, where each trajectory is sub-sampled every 0.3 seconds. Finally, 9750 instances are created and 5-fold cross validation is used for evaluation. Similar to Waymo, the other heterogeneous dataset is Stanford Drone Dataset (SDD) [29], which is a standard benchmark for trajectory prediction containing the categories of the road-agents. Different from Waymo dataset, the images are provided from the top-view angle and the coordinates are provided in pixel. Moreover, the scene context in SDD covers a variety of outdoor places and contains more spatial constraints such as intersection and roundabouts on a university campus. There are other datasets of heterogeneous traffic such as Apolloscape [13] and TRAF [7], unfortunately, due to the lack of raw image and camera matrices, respectively, we could not make use of them in our paper.

In addition to the heterogeneous dataset, we also evaluate the performance of AEE-GAN in the homogeneous environment on two commonly-used homogeneous datasets, i.e., ETH [27] and UCY [19] datasets. These two datasets contain annotated trajectories of the real-world pedestrians with a variety of social interaction scenarios, including collision avoidance behavior, group movement, group forming, and group dispersing. ETH dataset consists of two unique scenes, ETH and Hotel, while UCY dataset contains 3 unique scenes, Zara01, Zara02, and Univ. Each frame in both datasets includes top-view images and 2D locations of each person in the world coordinate. Both of the datasets take place in a relatively less constrained outdoor environment.

Evaluation Metrics. Similar to prior works [4, 11, 30], we use the following metrics to evaluate the proposed system:

(1) *Average Displacement Error (ADE):* Average L2 distance between ground truth and our prediction overall predicted time steps.
(2) *Final Displacement Error (FDE):* The distance between the predicted final destination and the ground truth final destination at the end of the prediction period T.

Baselines. We compare AEE-GAN against 4 state-of-the-art models on Waymo dataset, including 1) S-GAN-P [11], an LSTM-GAN hybrid network that applies generative adversarial architecture to Social LSTM [2], 2) S-Ways [4], a GAN predictor that applies the Info-GAN architecture to S-GAN [11], and 3) TraPHic [6], an LSTM-CNN hybrid network that models the interactions between heterogeneous road-agents 4) S-STGCNN [26], a GCN based model that through a Spatio-Temporal Graph CNN and following by the Time-Extrapolator CNN layers. For Stanford Drone Dataset, we compare AEE-GAN against 5 state-of-the-art models including 1) S-GAN-P, 2) SoPhie [30], a predictive model that applies attention mechanism to S-GAN [11], 3) Lin, a linear regressor, 4) S-LSTM [2], an LSTM-based network that models interaction with LSTM unit, and 5) DESIRE [18] which is an RNN-based conditional variational auto-encoder (CVAE) applying inverse optimal control (IOC). For homogeneous datasets, ETH and UCY, several homogeneous baselines are compared with AEE-GAN, including S-LSTM, S-GAN-P, SoPhie and S-Ways. Furthermore, to understand the effectiveness of our proposed Self Attention module and Horizon Attention module, we also compare three versions of our model in an ablation setting on all the datasets, i.e., our complete model (AEE-GAN), AEE-GAN without using Horizon Attention module (H_o) and AEE-GAN without using the Self Attention module (S_o).

Table 1: Quantitative results of AEE-GAN vs. baseline models on the front-view and heterogeneous dataset, Waymo. ADE and FDE are used as the errors metrics and reported in meters, which are separated by a slash.

	Baselines				Ours
Waymo	S-GAN-P	S-Ways	TraPHic	S-STGCNN	**AEE-GAN**
	6.05 / 11.26	5.87 / 10.17	5.79 / 10.87	3.71 / 6.05	**3.24 / 5.84**

4.1 Quantitative Results

Waymo Dataset. Table 1 shows the performance of the baselines and our algorithm on the heterogeneous front-view dataset in terms of ADE and FDE in meter space. S-STGCNN models the spatio-temporal dependencies at one shot and outperforms other baselines. However, AEE-GAN still outperforms S-STGCNN since AEE-GAN jointly consider both visual and social features. The results manifest that AEE-GAN is capable of modeling interaction between different road-agents by incorporating the type and the Enforced Attention. Furthermore, the long-range relation in the image generated by the self-attention mechanism allows for more accurate trajectory predictions on the heterogeneous dataset. Moreover, TraPHic is better than S-GAN-P and S-Ways because it incorporates the heterogeneous-based weighted interactions to model the interaction between heterogeneous road-agents.

Stanford Drone Dataset (SDD). Table 2 compares AEE-GAN to several baselines in terms of ADE and FDE in pixel space. AEE-GAN outperforms all other baselines, suggesting that combining the proposed RVAE and SE help AEE-GAN generate more accurate predictions. Lin performs the worst, while S-LSTM and S-GAN provide a slight improvement in accuracy. DESIRE provides a significant improvement due to the use of scene contextual information. Meanwhile, SoPhie performs better than other baselines since it incorporates both social and physical attention with GAN.

ETH and UCY. Here, we compare AEE-GAN against different baselines to demonstrate that the proposed AEE-GAN outperforms other homogeneous state-of-the-art methods on the homogeneous

Table 2: Evaluation on Stanford Drone Dataset (SDD). ADE and FDE of various models are reported in pixels. AEE-GAN outperforms other baselines since the combination of Social Enforcement module and Visual Enforcement module allows AEE-GAN model the interaction between road-agents and road-agents, as well as the interaction between road-agent and scene.

	Baselines					Ours
SDD	LIN	S-LSTM	S-GAN-P	DESIRE	SOPHIE	**AEE-GAN**
	37.11 / 63.51	31.19 / 56.98	28.31 / 42.63	19.25 / 34.05	16.27 / 29.38	**12.45 / 14.00**

Table 3: ADE and FDE of AEE-GAN and baselines are reported in meters on several homogeneous dataset, which are separated by a slash.

	Baselines					Ours
Dataset	S-LSTM	S-GAN-P	SOPHIE	S-WAYS	S-STGCNN	**AEE-GAN**
ETH	1.09 / 2.35	0.77 / 1.38	0.70 / 1.43	0.39 / 0.64	0.62 / 1.07	**0.32 / 0.44**
Hotel	0.79 / 1.76	0.44 / 0.89	0.76 / 1.67	0.39 / 0.66	0.41 / 0.51	**0.19 / 0.23**
Univ	0.67 / 1.40	0.75 / 1.50	0.54 / 1.24	0.55 / 1.31	0.62 / 1.07	**0.37 / 0.56**
ZARA1	0.47 / 1.00	0.35 / 0.69	0.30 / 0.63	0.44 / 0.64	0.40 / 0.61	**0.24 / 0.33**
ZARA2	0.56 / 1.17	0.36 / 0.72	0.38 / 0.78	0.51 / 0.92	0.31 / 0.49	**0.26 / 0.25**
AVG	0.71 / 1.53	0.53 / 1.02	0.53 / 1.15	0.45 / 0.83	0.48/ 0.73	**0.27 / 0.36**

Table 4: Ablation studies on Stanford Drone Dataset and Waymo Dataset in terms of ADE and FDE.

	Ours		
Dataset	**H_o**	**S_o**	**AEE-GAN**
Waymo	5.31 / 9.49	3.74 / 7.19	**3.24 / 5.84**
SDD	15.88 / 27.67	13.83 / 21.03	**12.46 / 14.01**
ETH	0.35 / 0.53	0.34 / 0.43	**0.32 / 0.44**
Hotel	0.30 / 0.36	0.22 / 0.23	**0.19 / 0.23**
Univ	0.46 / 0.74	0.51 / 0.96	**0.37 / 0.56**
ZARA1	0.28 / 0.47	0.25 / 0.34	**0.24 / 0.33**
ZARA2	0.27 / 0.46	0.28 / 0.47	**0.26 / 0.25**

dataset. The prediction errors for both ADE and FDE are reported in meter space. Table 3 shows that AEE-GAN averagely outperforms all the other models by at least 40% and 50.6% with regards to ADE and FDE in homogeneous environments. S-LSTM performs the worst since it only concerns the interaction between the neighboring people without taking all the people in the scene into account. S-GAN-P improves S-LSTM by applying the pooling module to model the interactions of all people involved in the scene and leveraging the generative ability of GAN architecture, while SOPHIE outperforms S-GAN-P because it applies both social attention and physical attention. On the other hand, S-WAYS outperforms other baselines since it applies the Info-GAN architecture to preserve the multi-modal property.

Ablation Studies. To analyze the effectiveness of our proposed Horizon Attention module and Self Attention module, we perform ablation studies on multiple homogeneous and heterogeneous datasets, as reported in Table 4. The results manifest that although both variants of our full method outperform all the other baselines (please refer to Tables 1 , 2 and 3), the variant that does not use Horizon Attention module (H_o) has a higher error than the variant with Horizon Attention module but without using Self Attention module (S_o). The result suggests that the Horizon Attention module contributes more to the accuracy even on the Stanford Drone Dataset. Meanwhile, by jointly strengthen the ability to model social interaction and the interaction between road-agent and scene by incorporating both Enforcement modules, AEE-GAN achieves the best performance. There is another more subtle ablation study performed on Waymo dataset, please refer to the supplementally material for the details.

4.2 Qualitative Results

The impact of Recurrent Visual Attention Enforcement. To demonstrate the effectiveness of the Recurrent Visual Attention Enforcement module for helping to predict the trajectory compliant to spatial constraints, we sample 30 random observation trajectories of the road-agents from the test set. AEE-GAN generates prediction trajectories by using the 30 random observations in the scene as the starting points. The distribution of these predicted trajectories is capable of identifying a moveable area according to the spatial constraints. Fig. 3 presents two unique scenes from SDD, showing an interpretable traversability map by using the distribution heatmap of the prediction trajectories. In the first scene, Nexus 6, we can see that the generated distribution is compliant to the central road and the path to the side, while in Nexus 9, our generated prediction trajectories are able to identify the central road as well as the path to other complicated sidewalks and intersection. Therefore, the Recurrent Visual Attention Enforcement module indeed facilitates the recognition of the traversable areas.

Horizon Attention. To show the effectiveness of our Horizon Attention module, we compare S-WAYS [4] which uses the same Social Attention module as our model and the variation of our model, VT_o, which solely uses the Horizon Attention module to strengthen the original Social Attention module. Here, UCY dataset is used to show the distribution of our attention weight and the effectiveness of our attention module for different social interaction scenarios since UCY dataset which contains many social behaviors. Fig. 4 visualizes the attention weights of the Social Enforcement module (a) and

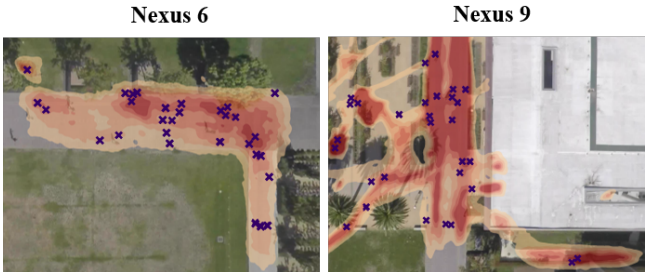

Figure 3: Visualization of the distribution of the prediction trajectories. The distribution of the trajectories is presented in red, and the 30 random starting samples are illustrated as blue crosses. The figure is best viewed in color.

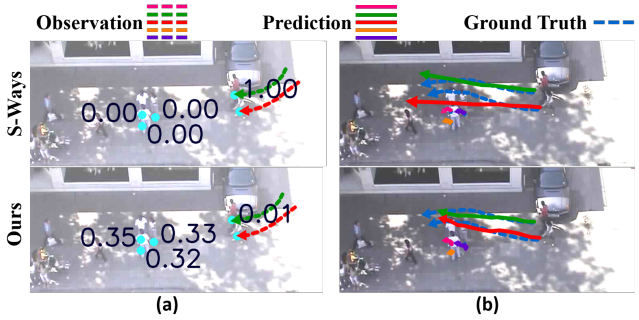

Figure 4: Visualization of the social attention coefficient $\alpha^{i,j}_{soc,t}$. Each number denotes the coefficient of other agents in regard to agent i (red line). Rows shows two prediction results predicted by S-Ways and AEE-GAN respectively.

the prediction trajectories (b). S-Ways fails to emphasize on the road-agents in front of the predicted agent but only focuses on the neighboring agent. As such, the red predicted trajectory occur collision. In contrast, our model provides a more appropriate attention weights distribution that allows the predicted agent to notice important regions and predict a movement to avoid the collision.

4.3 Visualizations of RVAE

To investigate the performance of RVAE, we use Grad-Cam [32] to visualize the result of visual attention by the heatmap, which represents the attention weight toward the image pixels. Fig. 5 shows the visualization results on SDD and Waymo Datasets. The major difference between the two datasets is the camera angle, where SDD dataset is mostly presented in a top-down view, while Waymo Dataset is mostly presented in front view. This difference leads to an interesting comparison between the visual attention of two datasets. In Waymo, RVAE often focuses on the region which may pass in the future. In contrast, in the SDD dataset, the attended region of RVAE sometimes not only be the future path area but also be the region which might "potentially" affect the prediction result. i.e., the flowerbed in the road circle is a physical constraint that people can not pass (Fig. 5(b)). This is probably due to the image scene will not be changed in the top-down view dataset (SDD), and

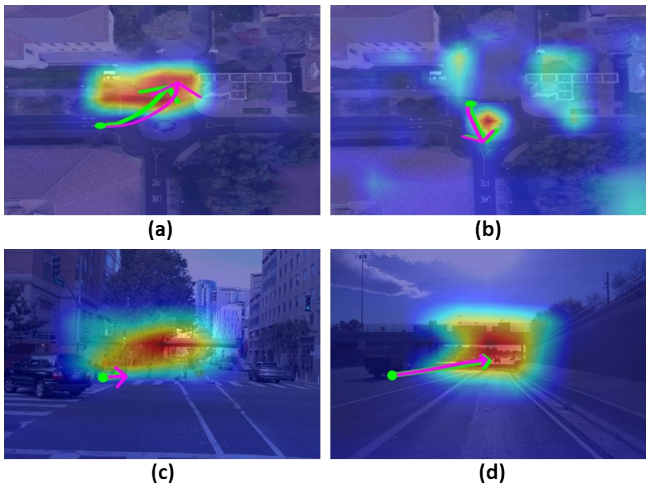

Figure 5: Visualizations of RVAE. (a)(b) show the results from SDD, and (c)(d) show the results from Waymo Dataset. The heatmap indicates that which region will be focused on when we predict the current agent's trajectory. The predicted results of trajectory and ground truth are colored in pink and green, respectively.

our model is, therefore, able learn to focus on the specific region which might affect prediction results. When the visual information is constantly changing, i.e., the front view (Waymo), the attention module only focuses on the region which is informative for the predicted trajectories based on the current image.

5 CONCLUSION

We conduct the problem of trajectory prediction in heterogeneous environment and propose AEE-GAN for attending to important features via Recurrent Visual Attention Enforcement and Social Enforcement. We design an Info-GAN architecture to generate multi-modal trajectories with recurrent feedback. Experimental results demonstrate that AEE-GAN achieves state-of-the-art performance on several trajectory prediction benchmarks. The proposed method, AEE-GAN provides better results in the top-view data than in front-view data due to the complete *surrounding information*. In the future, we plan to improve the front-view by incorporating more information from the ecology, e.g., 3D features, the traffic signals and audio signals as auxiliary information.

ACKNOWLEDGEMENTS

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2221-E-009-114-MY3, MOST-109-2634-F-009-018, MOST-109-2223-E-009-002-MY3, MOST-109-2218-E-009-025, MOST-109-2634-F-007-013, MOST-109-2218-E-002-015 and MOST-108-2218-E-002-055.

REFERENCES

[1] 2019. Waymo Open Dataset: An autonomous driving dataset.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.

[3] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. 2014. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2203–2210.

[4] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. 2019. Social ways: Learning multi-modal distributions of pedestrian trajectories with GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[5] Graeme Best and Robert Fitch. 2015. Bayesian intention inference for trajectory prediction with an unknown goal destination. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5817–5823.

[6] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. 2019. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8483–8492.

[7] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. 2019. TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Rohan Chandra, Uttaran Bhattacharya, Christian Roncal, Aniket Bera, and Dinesh Manocha. 2019. RobustTP: End-to-End Trajectory Prediction for Heterogeneous Road-Agents in Dense Traffic with Noisy Sensor Inputs. In *ACM Computer Science in Cars Symposium*. 1–9.

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

[10] Nachiket Deo and Mohan M Trivedi. 2018. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1468–1476.

[11] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2255–2264.

[12] Dirk Helbing and Peter Molnar. 1995. Social force model for pedestrian dynamics. *Physical review E* 51, 5 (1995), 4282.

[13] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. 2018. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 954–960.

[14] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82, 1 (1960), 35–45.

[15] Vasiliy Karasev, Alper Ayvaci, Bernd Heisele, and Stefano Soatto. 2016. Intent-aware long-term prediction of pedestrian motion. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2543–2549.

[16] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. 2012. Activity forecasting. In *European Conference on Computer Vision*. Springer, 201–214.

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

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

[19] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. 2007. Crowds by example. In *Computer graphics forum*, Vol. 26. Wiley Online Library, 655–664.

[20] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: Dual Shot Face Detector. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5725–5734.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[23] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. 2019. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6120–6127.

[24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1975–1981.

[25] Andreas Møgelmose, Mohan M Trivedi, and Thomas B Moeslund. 2015. Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations. In *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 330–335.

[26] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. *CVPR* (2020).

[27] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. 2010. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*. Springer, 452–465.

[28] Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. 2016. Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2096–2101.

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

[30] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1349–1358.

[31] Amir Sadeghian, Ferdinand Legros, Maxime Voisin, Ricky Vesel, Alexandre Alahi, and Silvio Savarese. 2018. Car-net: Clairvoyant attentive recurrent network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 151–167.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[33] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, Article arXiv:1409.1556 (Sep 2014), arXiv:1409.1556 pages. arXiv:1409.1556 [cs.CV]

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

[35] Jacob Walker, Abhinav Gupta, and Martial Hebert. 2014. Patch to the future: Unsupervised visual prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3302–3309.

[36] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. 2011. Who are you with and where are you going?. In *CVPR 2011*. IEEE, 1345–1352.

[37] Shuai Yi, Hongsheng Li, and Xiaogang Wang. 2016. Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance. *IEEE transactions on image processing* 25, 9 (2016), 4354–4368.

[38] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* (2018).

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

[40] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang. 2015. Learning collective crowd behaviors with dynamic pedestrian-agents. *International Journal of Computer Vision* 111, 1 (2015), 50–68.

[41] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. 2019. Explainable Video Action Reasoning via Prior Knowledge and State Transitions. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 521–529. https://doi.org/10.1145/3343031.3351040

[42] Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. 2009. Planning-based prediction for pedestrians. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3931–3936.