Which Way Are You Going? Imitative Decision Learning for Path Forecasting in Dynamic Scenes

Yuke Li
York University, Toronto, Canada
AutoNavi, Alibaba Group, Beijing, China
ykleewh@yorku.ca

Abstract

Path forecasting is a pivotal step toward understanding dynamic scenes, and it is an emerging topic in the computer vision field. This task is challenging due to the multimodal nature of the future, namely, there is more than one plausible prediction given histories. Yet, the state-of-the-art methods do not seem to be adequately responsive to this innate variability. Hence, how to better foresee the forthcoming trajectories in dynamic scenes has to be more thoroughly pursued. To this end, we propose a novel Imitative Decision Learning (IDL) approach. It delves deeper into the key that inherently characterizes the multimodality – the latent decision. The proposed IDL first infers the distribution of such latent decisions by learning from moving histories. A policy is then generated by taking the sampled latent decision into account to predict the future. Different plausible upcoming paths correspond to each sampled latent decision. This approach significantly differs from the mainstream literature that relies on a predefined latent variable to extrapolate diverse predictions. In order to augment the understanding of the latent decision and resultant multimodal future, we investigate their connection through mutual information optimization. Moreover, the proposed IDL integrates spatial and temporal dependencies into one single framework, in contrast to handling them with two-step settings. As a result, our approach enables simultaneous anticipating the paths of all pedestrians in the scene. We assess our proposal on the large-scale Stanford Aerial Pedestrian (SAP), ETH and UCY datasets. The experiments show that IDL introduces considerable marginal improvements with respect to recent leading studies.

1. Introduction

Path forecasting in dynamic scenes has surged as an intriguing topic because of the rising demands of emerging applications of artificial intelligence. For instance, robots and autonomous vehicles are required to react ``smartly´´as humans to their fast evolving environments. Thus, equipping them with the capability to forecast what will happen in the near future is imperative.

Figure 1. The multimodal nature of future paths in a dynamic scene: There are multiple plausible forthcoming paths (the dash red and cyan lines) based on identical historical moving records (the solid red and cyan lines). In this figure, we display three possibilities as an example.

Forecasting a future path refers to discerning an unknown upcoming trajectory by accessing records of prior movements. It entails effectively and efficiently processing the complex spatial dependencies (interactions among persons) and temporal dependencies (evolving motion patterns). Therefore, path forecasting is regarded as a multifaceted and complicated endeavor.

One issue that has been challenging for the task of path forecasting in dynamic scenes is the *multimodal nature* of the future: Given a set of historical observations, there will be more than one probable future (see Fig.1). Despite tremendous accomplishments that has been made to foresee a deterministic future[1, 44, 26, 45, 25], the majority of the existing studies fail to consider the multiple possibilities of future.

To date, state-of-the-art research has attempted to alleviate the issue of modeling this multimodality based on a predefined latent variable $z \sim N(0, 1)$. For instance, Social GAN [11] takes z as a part of the inputs. It encourages a diverse set of predictions by using a random sampled z for each forward pass. Lee *et al.* present the DESIRE [16] that approximates the distribution of future trajectories to the distribution of z. Nevertheless, to predefine z on the basis of $N(0, 1)$ is probably not able to fully assimilate the various factors affect the human trajectories within the dynamic scenes, such as the spatial and temporal dependen-

cies. The multimodal predictions of aforementioned studies might be compromised due to this point. Hence, the question of how to construct a framework that can better foresee the multimodal future remains an open challenge. Addressing this issue, rather than merely evoking a latent variable, we propose delving deeper into the substantial factor characterizing the multimodal future that has largely been neglected by previous works:

• *Latent decision*: The latent human decision internally determines the motion pattern of a person step by step to carve out the trajectory.

Our key insight is that the trajectory of a person is the outcome of his/her decisions, which are made upon all the related elements in dynamic scenes (e.g. the spatiotemporal dependencies). Exploring the latent decision leads to a richer understanding concerning the multimodality of future paths at a certain semantic level: The different plausible future moving course essentially accounts for each decision. Fig.1 presents an example to support our claim – a pedestrian may make different decisions and choose one route among all the possibilities. Thus, we propose investigating and mimicking the underlying human decision-making process to foresee the probable upcoming paths in dynamic scenes. Toward this end, we present a novel approach to the path forecasting problem, namely Imitative Decision Learning (IDL) based on the perspective of Generative Adversarial Imitation Learning (GAIL) [13]. Specifically, we first infer the distribution that corresponds to the latent decisions from historical observations, which consist of abundant information of how human decisions were made in dynamic scenes. The policy generating process then takes the sampled latent decision into consideration to perform forecasting. The connection of the latent decision to generated policy is augmented with the aid of optimizing their mutual information [5, 7]. Consequently, this optimization will yield informative feedback that highlights the intrinsic impact of the latent decision on the multimodal future paths. It is also noteworthy that the latent decision is learned in an unsupervised manner without any annotations.

Additionally, another common denominator of approaches with a predefined latent variable [11, 16] is that spatial and temporal dependencies are independently handled. These approaches assign a Long-short Term Memory (LSTM) to each person to obtain temporal information, and subsequently assign a social pooling term for spatial information. These methods overlook the fact that spatial and temporal information generally co-occur and affect each other. For example, in a dynamic scene, one person changing the direction of moving might cause another person following him/her to change direction or slow down to avoid collision. Therefore, these two parts are better jointly tied into a single model. In this study, our work enables processing the spatiotemporal dependencies at one shot. This setting also offers the capability of forecasting the paths of all people in dynamic scenes simultaneously.

In summary, the contributions of our work are:

1. We introduce a novel Imitative Decision Learning to the task of path forecasting in dynamic scenes. The proposed IDL delves deeper into the latent human decision to cover the space of diverse plausible future paths.
2. Our IDL can accommodate spatio-temporal dependencies in a single pass. Further, it allows to simultaneously predict the future trajectories for all persons in dynamic scenes.
3. We evaluate the proposed model on challenging large-scale video datasets, and show that significant gains can be attained with respect to trending works.

To the best of our knowledge, our work is the first study to imitate the underlying human decision-making process to uncover multimodality in the context of anticipating future paths in dynamic scenes.

The remainder of this paper is organized as follows: First, relevant works are discussed in Section 2. Section 3 details the framework of the proposed IDL. Section 4 conducts the experimental findings and discussions. Finally, we conclude in Section 5.

2. Related Work

The relevant literature has accumulated some efforts to overcome the challenges of path forecasting. The pioneering work [15] exploited the semantics of a single-person scenario in order to build a trajectory forecasting model. This work inspired some early studies that tried to solve the problem of path forecasting with scene-dependent motion patterns and handcrafted features, such as [2, 42, 47, 3, 35]. However, this raises a question about the applicability of these algorithms to different scenes.

Some studies focus on building a generalized predictive network for dynamic scenes motivated by data driven deep neural networks [8]. For instance, Behavior CNN [46] and FaF CNN [26] developed 3D Convolutional Neural Network (CNN) [14] based approaches for path predicting. Alahi *et al.* [1] proposed modeling the individual motion dynamic by assigning one LSTM [38] per pedestrian. Furthermore, a social pooling layer was adopted for processing spatial dependencies. The similar ideas are considered by the authors of [39, 45] with more sophisticated spatial information handling layers. Li. *et al.* [18, 19] attempted to processing the spatiotemporal information at the same time. The works of [22, 40, 48, 27] built networks upon ResNet [12, 50] and LSTM to predict highway traffic flow for preventing potential accidents. The methods presented in [28, 34] learned a reward correlated to the scene layout to find the best strategy for future trajectory constructions.

Nevertheless, a shared deficiency of these studies is that they are only able to predict a deterministic future path. In other words, the multimodal nature has not been taken into their considerations.

Recent generative models [9, 37] achieve cutting-edge performances on the task of synthesizing diverse images [29, 43, 6, 41, 20]. Inspired by these models, several recent approaches mitigated the issue of capturing multiple possibilities by harnessing a predefined latent variable $z \sim N(0,1)$, then incorporating it with generative networks. Gupta *et al.* [11] developed an approach to simulate multiple possible predictions based on perceiving the z along with past moving records. The authors of [16] considered deriving the distribution of the future path that is an approximation of $N(0,1)$. A major shortcoming of the research mentioned earlier is that z is predefined by $N(0,1)$ in the absence of proper reasoning and justification. Therefore, the prior works might inadequately fully digest the context of dynamic scenes and might fail to model the inherent multimodality of future.

Moreover, previous research involving z separately processed the spatial and temporal information. Their outcomes might be affected by not considering that these two components are dependent on each other. Thus, it is better to handle the spatiotemporal dependencies together.

In this study, we propose a novel IDL framework for path forecasting in dynamic scenes, which explores the latent decision to anticipate future paths. We would like to stress that our work significantly differs from the existing studies on path forecasting with following facts: (1) With respect to [11, 16], explicitly exploring the latent decision enables our approach better capturing mutimodality, rather than utilizing a predefined latent variable. (2) Unlike previous studies, our IDL accommodates the spatial and temporal factors in a single pass. We learn a single architecture for all persons in a scene instead of assigning one model per person.

3. Methodology

We carry out the path forecasting concern by mimicking the underlying human decision-making process. Subsection 3.1 introduces the problem formulation. We elaborate our framework outlining and formally deriving our objective in subsection 3.2. Subsection 3.3 details the implementation.

3.1. Problem Formulation

We map the labeled coordinates into a set of motion features $\mathcal{X}_t$ ($(t \in [t_1, t_k])$) and $\mathcal{GT}_{t'}$ ($(t' \in [t_{k+1}, t_{k+k'}])$). $\mathcal{X}_t$ and $\mathcal{GT}_{t'}$ refer to the moving histories and the ground truth future, respectively. These motion features encapsulate all individual motion patterns through displacement information [1]. Our proposed IDL proceeds by recovering a policy π from $\mathcal{X}_t$ ($(t \in [t_1, t_k])$) to generate $\mathcal{X}_{t'}$ ($(t' \in [t_{k+1}, t_{k+k'}])$). We produce future multimodal paths via incorporating the latent decision $\mathcal{S}$ in the process of recovering policy π.

Formally, we consider IDL by extending GAIL [13]. Borrowing the notations from GAIL, our states and actions correspond to $\mathcal{X}_t$ and $\mathcal{X}_{t'}$, respectively. The numerous labeled ground truth $\mathcal{GT}_{t'}$ are treated as the demonstration from experts. The latent decision $\mathcal{S} \sim p(\mathcal{S}|\mathcal{X}_t)$ is unknown and needs to be inferred. $p(\mathcal{S}|\mathcal{X}_t)$ denotes the distribution from which $\mathcal{S}$ is sampled. In the context of GAIL, the generator can be viewed as policy. Instead of solely obtaining a policy/generator π from the states $\mathcal{X}_t$, we propose to also learn from $\mathcal{S}$ for modeling the multimodality aspect. IDL quantifies the impact of $\mathcal{S}$ on predictions through optimizing the mutual information between π and $\mathcal{S}$ without supervision.

3.2. Imitative Decision Learning

As mentioned previously, our work focuses on understanding and imitating the underlying human decision-making process to anticipate future paths in dynamic scenes. Fundamentally, our IDL can be viewed as jointly training (1) an inference sub-network $\mathcal{L}$ that extrapolates the latent decision, (2) a policy/generator π that recovers a policy to generate upcoming paths, (3) a statistics sub-network Q that discovers the impact of latent decision on predictions, and (4) a discriminator $\mathcal{D}$ that attempts to differentiate our generated outcomes from the expert demonstrations. We depict the structure and workflow of our proposed IDL framework in Figure 2. In what follows, we provide detailed descriptions of each part.

Latent Decision Inference: We invoke the point that the latent decision is the key behind the multimodal nature of the future. In order to grasp this point, we propose to first uncover the distribution of the latent decisions through learning from prior moving histories. It is owing to that the moving histories have a wealth of records on how human decisions were made under a highly complex dynamic scenario. In practice, we parameterize the distribution of latent decisions by means of the inference sub-network $\mathcal{L}$.

The existence of spatiotemporal dependencies suggests the fact that a person cannot decide his/her behavior without considering his/her neighbors in dynamic scenes. Hence, it is necessary to learn that a distribution can represent all individuals in the scene. To achieve this goal, we first input the motion features $\mathcal{X}_t$ into a pre-trained fully convolutional sub-module [24, 49] to extract a higher-level representation of $\mathcal{X}_t$ at time instance t ($t \in [t_1, t_k]$). A set of higher-level representations from t_1 to t_k are then fed into a temporal convolutional sub-module [33, 17] to produce a two-unit vector. We subsequently append a deconvolutional [30, 23] sub-module and a softmax layer on each unit of this two-unit vector. The final outcomes are treated as the

[1] Please refer to the supplementary material for a detailed description of the motion feature construction.

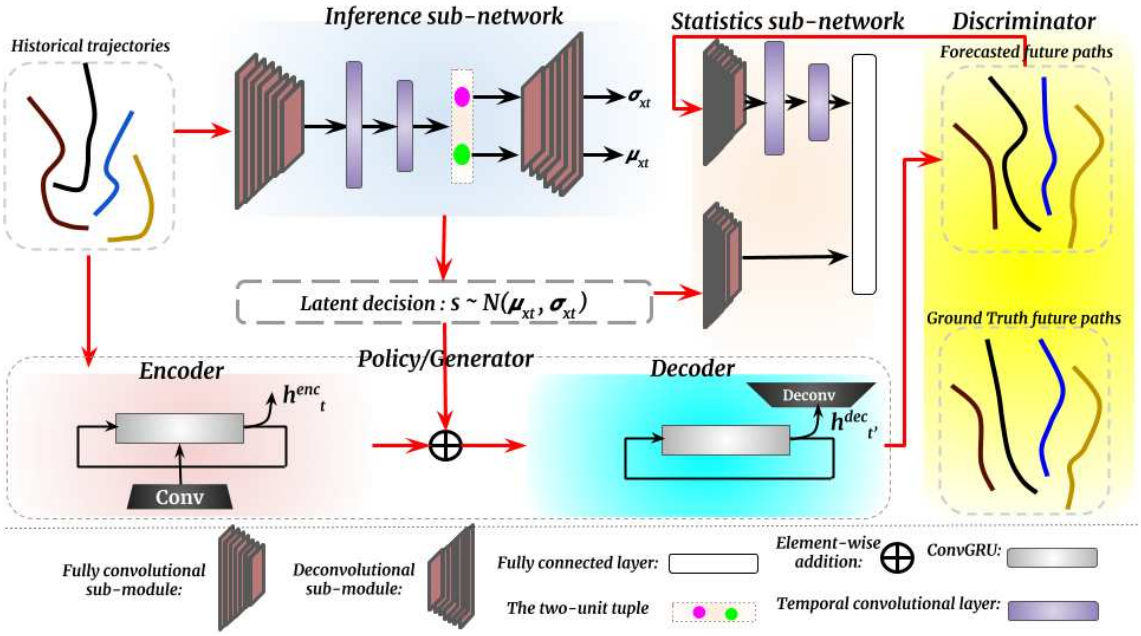

Figure 2. The detailed schematic diagram of our proposed IDL for forecasting future paths. The red arrows indicate the direction of information flow between each module. The black arrows suggest the direction of information flow inside a module. The historical trajectories are firstly input into the inference sub-network to infer the distribution of latent decisions. The temporal convolutional sub-module receives the output from the pre-trained convolutional sub-module and produces a two-unit vector. A pre-trained deconvolutional sub-module and a softmax layer read each unit to form the mean and derivation of a Gaussian distribution of latent decisions. Meanwhile, the encoder of our policy/generator π processes the historical trajectories by a ConvGRU layer. An element-wise addition product on the encoded hidden states $h_{t_k}^{enc}$ ($(t \in [t_1, t_k])$) and sampled latent decision $\mathcal{S}$ initializes the decoder. The final predictions are generated from the decoded hidden states $h_{t'}^{dec}$ ($t' \in [t_{k+1}, t_{k+k'}$) through a deconvolutional layer. The statistics sub-network reads prediction and latent decision to measure the significance of $\mathcal{S}$ on multimodal predictions. The discriminator distinguishes the predictions from the ground truth future paths (expert demonstrations).

mean $\mu_{\mathcal{X}_t}$ and variance $\sigma_{\mathcal{X}_t}$ of a conditional Gaussian distribution from which the latent decision $\mathcal{S}$ samples. It can be formulated as:

$$\mu_{\mathcal{X}_t}, \sigma_{\mathcal{X}_t} = Softmax(\mathcal{L}(\mathcal{X}_t), t \in [t_1, t_k]) \quad (1)$$

$$\mathcal{S} \sim p(\mathcal{S}|\mathcal{X}_t) = \mathcal{N}(\mu_{\mathcal{X}_t}, \sigma_{\mathcal{X}_t}) \quad (2)$$

During our experiments, we sample the latent decision $\mathcal{S}$ via the reparametrization trick. The process of inference offers the capability of attaining a profound comprehension of how human decisions handle the relevant factors, such as spatiotemporal information.

Policy/Generator: Given the sequential characteristics of human decision-making process, we leverage an encoder-decoder structure based upon Convolutional Gated Recurrent Units (ConvGRU) [4] to implement our policy/generator π. ConvGRU is able to capture temporal information along with spatial context.

The definition of the encoding is as follows:

$$h_t^{enc} = Enc(Conv(\mathcal{X}_t), h_{t-1}^{enc}), t \in [t_1, t_k] \quad (3)$$

where Enc is the encoder sub-network and h_t^{enc} denotes the hidden states at time instance t ($t \in [t_1, t_k]$). $Conv$ pertains to a single convolutional layer that serves to remove the sparsity, which $\mathcal{X}_t$ might present.

Simply passing $h_{t_k}^{enc}$ to the decoder sub-network does not take into account the latent decision. This fails to expose the innate multimodality to predictions. To tackle this issue, we propose incorporating the $h_{t_k}^{enc}$ with the latent decision $\mathcal{S}$, which has a vital impact on policy/generator π for multimodal path forecasting. More specifically, the decoder reads the element-wise addition product $h_{t_k}^{enc} \oplus \mathcal{S}$ for initialization. Accordingly, each sample of the latent decision $\mathcal{S}$ eventually poses a different plausible prediction. We formulate the decoding process as:

$$h_{t'}^{dec} = Dec(h_{t'-1}^{dec}), t' \in [t_{k+1}, t_{k+k'}] \quad (4)$$

where Dec stands for decoder and t' ($t' \in [t_{k+1}, t_{k+k'}$) is the future time step. The hidden state $h_{t'}^{dec}$ at time instance t' is only determined upon the previous hidden state $h_{t'-1}^{dec}$. In order to obtain $\mathcal{X}_{t'}$, we append a single deconvolutional layer with a stride of 2. It serves to transform $h_{t'}^{dec}$ into the same size as the inputs:

$$\mathcal{X}_{t'} = Deconv(h_{t'}^{dec}), t' \in [t_{k+1}, t_{k+k'}] \quad (5)$$

The hidden states $h_{t'}^{dec}$ and pre-specified logarithmic standard deviations are set to form a Gaussian for the Proximal Policy Optimization (PPO) [36]. The objective of updating our policy/generator π will be described later.

Mutual Information Optimization: In this section, we explicitly delve into the essence that binds latent decision $\mathcal{S}$ to the policy/generator π to gain a clearer insight into the correlation between $\mathcal{S}$ and the multimodal future.

We proceed by optimizing the mutual information between $\mathcal{S}$ and π to establish their connection. As a result, we are able to quantitatively measure the significance of latent decision on predictions. The mutual information $I(\mathcal{S}, \pi)$ is termed as:

$$I(\mathcal{S}, \pi) = \mathcal{H}(\mathcal{S}) - \mathcal{H}(\mathcal{S}|\pi) \tag{6}$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy. A larger value of $I(\mathcal{S}, \pi)$ refers to a larger influence of the latent decision. Optimizing Eq.6 thereby incentivizes $\mathcal{S}$ and consolidates its impact on the predictions. Moreover, such optimization strengthens our semantic understanding of the impact of latent decision on multimodal future paths. With the assistance of the Mutual Information Neural Estimator [5], optimizing $I(\mathcal{S}, \pi)$ is equivalent to maximizing its lower bound L_I. Toward this end, we introduce a statistics sub-network $\mathcal{Q}$ to approximate L_I:

$$\begin{aligned} I(\mathcal{S}, \pi) \geq L_I = \mathbb{E}_{\mathcal{X}_{t'} \sim \pi, \mathcal{S} \sim p(\mathcal{S}|\mathcal{X}_t)} \mathcal{Q}(\mathcal{X}_{t'}, \mathcal{S}) \\ - log \mathbb{E}_{\mathcal{X}_{t'} \sim \pi, \hat{\mathcal{S}} \sim p(\mathcal{S}|\mathcal{X}_t)} (e^{\mathcal{Q}(\mathcal{X}_{t'}, \hat{\mathcal{S}})}) \end{aligned} \tag{7}$$

where $\mathcal{S}$ and $\hat{\mathcal{S}}$ are *i.i.d.* samples from $p(\mathcal{S}|\mathcal{X}_t)$. $\mathcal{X}_{t'}$ is the corresponding prediction of the sampled latent decision $\mathcal{S}$. We display the structure of statistics sub-network $\mathcal{Q}$ in Fig.2. The final result of L_I is obtained from a fully connected layer by reading the concatenation of the representations of predictions $\mathcal{X}_{t'}$ and the representations of $\mathcal{S}$. They are outputs from the temporal convolutional sub-module and pre-trained fully convolutonal sub-module, respectively.

Discriminator and Objective: We advocate applying GAIL [13, 21] to train our framework. Thus our IDL retains the efficiency of gradient-based learning while formulating path forecasting as an occupancy measure matching problem. We propose using a discriminator $\mathcal{D}$ to distinguish $[\mathcal{X}_t, \mathcal{X}_{t'}]$ from $[\mathcal{X}_t, \mathcal{GT}_{t'}]$ to guide π. $[\mathcal{X}_t, \mathcal{X}_{t'}]$ and $[\mathcal{X}_t, \mathcal{GT}_{t'}]$ pertain to the combination of the past records and the predictions/ground truth, respectively. As a result, the discriminator can only be fooled if $\mathcal{X}_{t'}$ is consistent with $\mathcal{X}_t$. The objective of discriminator is as follows:

$$\begin{aligned} \mathcal{L}^{\mathbf{D}} =& \mathcal{D}([\mathcal{X}_t, \mathcal{X}_{t'}]) - \mathcal{D}([\mathcal{X}_t, \mathcal{GT}_{t'}]) \\ &+ \lambda(\| \nabla \mathcal{D}(\epsilon \mathcal{GT}_{t'} + (1-\epsilon)\mathcal{X}_{t'}) \|_2 - 1)^2 \end{aligned} \tag{8}$$

$(\| \nabla \mathcal{D}(\epsilon \mathcal{GT}_{t'} + (1-\epsilon)\mathcal{X}_{t'}) \|_2 - 1)^2$ is the gradient penalty term following Wasserstein GAN with Gradient Penalty (WGAN-GP) [10]. $\lambda > 0$ is a coefficient, and $\epsilon \sim U[0, 1]$ is a random parameter. The discriminator $\mathcal{D}$ consists of a single ConvGRU layer. We top four stacked convolutional layers upon the ConvGRU layer to obtain a score that reflects either $[\mathcal{X}_t, \mathcal{X}_{t'}]$ or $[\mathcal{X}_t, \mathcal{GT}_{t'}]$.

The policy/generator π receives the gradient from $\mathcal{D}$ through PPO [36] by maximizing the following objective:

$$\mathcal{L}^{\pi} = \mathcal{D}([\mathcal{X}_t, \mathcal{X}_{t'}]) + \eta L_I \tag{9}$$

where η is the parameter of L_I.

Algorithm 1 Imitative Decision Learning

Input:

1. Historical records for i-th sequence $\mathcal{X}_t^i$ ($t \in [t_1, t_k]$);
2. Ground truth future for i-th sequence $\mathcal{GT}_{t'}^i$ ($t' \in [t_1, t_{k+k'}]$);
3. The initial parameters of inference sub-network, policy/generator, statistics sub-network and discriminator.

Output: Learned policy/generator π

for i = 0,1,2... **do**

1. Sample and fix $\mathcal{S}^i \sim \mathcal{N}(\mu_{\mathcal{X}_t^i}, \sigma_{\mathcal{X}_t^i})$ (Eq.2) for each rollout.
2. Generate future paths $\mathcal{X}_{t'}^i$ ($t' \in [t_1, t_{k+k'}]$) (Eq.5).
3. Gradient descent on $\mathcal{D}$ to minimize $\mathcal{D}([\mathcal{X}_t^i, \mathcal{X}_{t'}^i]) - \mathcal{D}([\mathcal{X}_t^i, \mathcal{GT}_{t'}^i]) + \lambda(\| \nabla \mathcal{D}(\epsilon \mathcal{GT}_{t'}^i + (1-\epsilon)\mathcal{X}_{t'}^i) \|_2 - 1)^2$
4. Sample and fix $\hat{\mathcal{S}}^i \sim \mathcal{N}(\mu_{\mathcal{X}_t^i}, \sigma_{\mathcal{X}_t^i})$ independent of $\mathcal{S}^i$ for each rollout. Updating $\mathcal{Q}$ and $\mathcal{L}$ by maximizing $\mathbb{E}_{\mathcal{X}_{t'}, s^i}(\mathcal{Q}(\mathcal{X}_{t'}^i, s^i)) - log \mathbb{E}_{\mathcal{X}_{t'}^i, \hat{\mathcal{S}}^i}(e^{\mathcal{Q}(\mathcal{X}_{t'}^i, \hat{\mathcal{S}}^i)})$
5. Maximize $\mathcal{D}([\mathcal{X}_t^i, \mathcal{X}_{t'}^i]) + \eta L_I$ with PPO [36] to update policy/generator π.

end for

3.3. Implementations

Training Strategy: Among the variants of GAN [9], WGAN-GP [10] stands out due to its ability to overcome the weaknesses of mode collapse and unstable convergence. Hence, we form the objective of our proposed IDL by following WGAN-GP. η in Eq.9 λ in Eq.8 are empirically set as 0.1 and 10, respectively. We summarize the training strategy of our proposed IDL in Algorithm 1. The backpropagation through time and RMSProp [8] is adopted to optimize D with the learning rate initialized at 5×10^{-5}, and at 3×10^{-5} for updating L and $\mathcal{Q}$, respectively. To accelerate training, we initialize our policy/generator π from behavior cloning as [13, 21] suggest.

Network Configurations: An inference sub-network $\mathcal{L}$, a statistics sub-network $\mathcal{Q}$, a policy/generator π and a discriminator $\mathcal{D}$ form our proposed IDL approach. Please refer to the supplementary material for the details.

Our implementation is based on the PyTorch library [31]. The experiments were carried out on two Nvidia GeForce GTX 1080 Ti, supplied with 22 GB of memory in total.

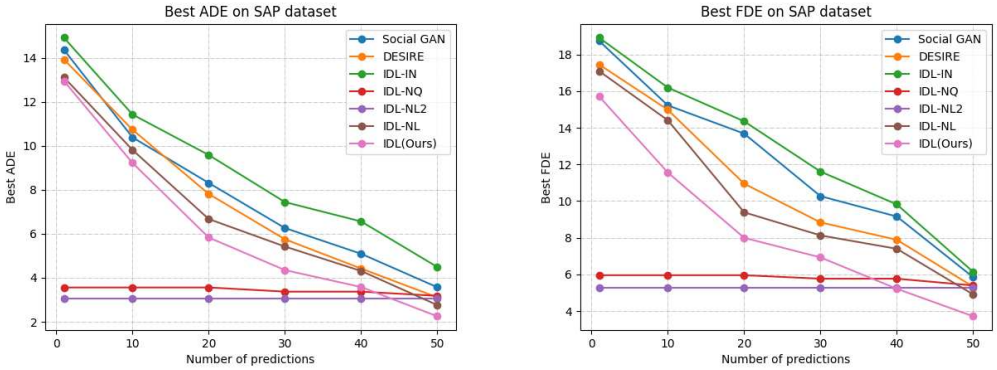

Figure 3. The stochasticity of results from different methods on SAP dataset. Each line presents the result with best ADE (top) and best FDE (bottom). The numbers on X-axis indicate the number of predictions, and the numbers on Y-axis denote the scores.

Figure 4. Qualitative comparisons on SAP dataset. The top left shows the observed records and the matching ground truth (G.T.). In order to have a clear visualization for better understanding the multimodality, we separately illustrate several trajectories and the diverse predicted paths apart from others from example 1 to example 5.

4. Experimental Analysis

4.1. Datasets and Experimental Settings

Three large-scale benchmark datasets are exploited to validate the performance of the proposed IDL on the task of path forecasting in dynamic scenes. Namely Stanford Aerial Pedestrian (SAP) dataset [35], ETH dataset and UCY dataset [32] are selected. The SAP dataset comprises long video sequences for eight scenes in total. It labels complete trajectories of different categorized moving objects from the time they enter the scene to the exit time, for instance the pedestrians, bicyclists and vehicles. The SAP dataset makes a reasonable foundation for realistically evaluating our experimental results as it manifests a highly dynamic scenario. The ETH dataset contains two scenes with 750 different pedestrians and is split into two sub-datasets (ETH and Hotel). The UCY dataset contains three sub-scenes with 786 people: UCY, ZARA-01 and ZARA-02. Both the ETH and the UCY datasets are uniformly annotated at a 0.4 second rate. The spatial coordinates of the annotations provided by all datasets are embedded to a dimension of 256×256 to feed into our network.

To ensure a fair comparison on the SAP dataset, we follow the experimental settings opted in DESIRE [16]. The entire SAP dataset is divided into 16,000 short video clips across scenes for our experiments. We train and test by observing past $k = 60$ frames (2 seconds), and forecasting subsequent $k' = 120$ frames (4 seconds). The evaluation criterion follows a randomized 5-fold cross-validation strategy on nonoverlapping videos clips. In the second part of our experiments on the ETH and UCY dataset, following [11], we extend the value of k to 8 time steps and k' to 12 time steps. This is equivalent to observing 3.2 seconds and predicting next 4.8 seconds. We utilize the leave-one-out cross-validation to evaluate our performance, which is training on 4 sub-datasets and testing on the remaining one as [1, 11].

As per comparing approaches, we assess the perfor-

mance of our IDL versus the two most recent state-of-the-art studies of the path forecasting task. Specifically, we select DESIRE [16], which performs the best on the SAP dataset, and Social GAN [11], which achieves cutting-edge results on the ETH and UCY datasets. The classic Social LSTM [1] is employed for comparisons as well. We also study the effectiveness of each part of our IDL through analyses against the following baselines:

1. IDL-NL: In order to highlight the merit of our latent decision inference, we use a predefined $\mathcal{S} \sim \mathcal{N}(0, 1)$ rather than inducing the distribution from histories. The rest of the framework remains unchanged.
2. IDL-NQ: We evaluate the necessity of mutual information optimization by comparing with IDL-NQ. This baseline drops the statistics sub-network from the IDL framework.
3. IDL-NL2: To test the impact of the latent decision on the forecasting of future paths, we construct the IDL-NL2 baseline by discarding both the inference sub-network and statistics sub-network.
4. IDL-IN: We term our last baseline as IDL-IN to validate the effectiveness of joint processing spatiotemporal dependencies. This baseline replaces the ConvGRU with vanilla GRU, and replaces the convolutional layers with fully connected layers in our IDL.

4.2. Quantitative Evaluation

We carry out the experiments by drawing samples of the latent decision $\mathcal{S}$ 50 times for each sequence of the SAP dataset, and 20 times for each sequence of the ETH and UCY datasets. In other words, we generate 50 predictions/rollouts on each sequence of the SAP dataset, and 20 predictions/rollouts for each sequence of the ETH and UCY datasets. In our quantitative evaluation, we aim to determine whether the extensive range of possible predictions produced by our proposed IDL includes the true future. We judge our experiments with the best Average Displacement Error (best ADE) and the best Final Displacement Error (best FDE) of the various approaches. Lower values suggest better results for both measurements. These two metrics are reasonable since they address to measure if the ground truth is approximated within a diverse set of multiple predictions. Fig.3 suggests that the probability of forecasting the true future ascends by creating more plausible upcoming futures, as it is likely to obtain a prediction that is closer to ground truth.

Table 1 summarizes the quantitative results of the best ADE and best FDE following [11, 16]. Our proposed IDL manifests the best performance against other approaches often by a considerable margin for both criteria. For instance, the reported best ADE and best FDE rate of IDL amount to 2.25 and 3.82 on SAP dataset. These scores outperform Social GAN [11] (by 1.32 and 2.03, respectively),

Best ADE / FDE on SAP dataset

	ADE	FDE
Social GAN[11]	3.57	5.85
DESIRE[16]	3.11	5.33
IDL-IN	4.49	6.14
IDL-NQ	3.08	5.36
IDL-NL2	3.04	5.27
IDL-NL	2.76	4.90
IDL (Ours)	**2.25**	**3.82**

Best ADE / FDE on ETH and UCY datasets

	ETH	HOTEL	UNIV	ZARA1	ZARA2
	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE
Social LSTM[1]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17
Social GAN[11]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84
IDL-IN	1.24/2.61	1.06/2.04	0.92/1.87	0.64/1.16	0.77/1.39
IDL-NQ	0.83/1.57	0.66/1.25	0.74/1.50	0.33/0.67	0.41/0.82
IDL-NL2	0.81/1.59	0.65/1.22	0.74/1.48	0.31/0.64	0.39/0.80
IDL-NL	0.75/1.51	0.60/1.06	0.69/1.42	0.28/0.61	0.35/0.73
IDL (Ours)	**0.59/1.30**	**0.46/0.83**	**0.51/1.27**	**0.22/0.49**	**0.23/0.55**

Table 1. The quantitative comparisons.

DESIRE [16] (by 0.86 and 1.51, respectively), IDL-IN (by 2.24 and 2.32, respectively), IDL-NQ (by 0.83 and 1.54, respectively), IDL-NL2 (by 0.79 and 1.45) and IDL-NL (by 0.51 and 1.08, respectively). The significant superiorities of our proposed IDL compared with other methods on the ETH and UCY datasets speak to its advantages.

In order to analyze the benefits of IDL in detail, we further conduct ablation studies from the following two aspects:

Latent Decision Exploration: The proposed inference sub-network investigates the latent decision from observed records. We report that the results reflect the benefit of inferring the distribution of the latent decision in Table 1. Our proposed IDL incurs remarkable advantages, by far, versus IDL-NL, which achieves the second best performance across the datasets. The IDL drastically advances the state-of-the-art methods Social GAN [11] and DESIRE [16] as well. These outcomes tip the balance steeply toward delving deeper into the latent decision in terms of forecasting future paths and away from the use of a predefined latent variable.

During our experiments, we find that IDL-NQ tends to generate predictions that are insensitive to the latent decision. Furthermore, these predictions are close to deterministic IDL-NL2 (refer to Fig.3). This result overwhelmingly demonstrates the value of considering the mutual information optimization for capturing multimodality.

The scores attained by our IDL outdo those of IDL-NL2 on all datasets. This finding provides the evidence that explicitly modeling the latent decision enables a better understanding of the multimodal nature. In fact, even the IDL-NL baseline yields better results than the IDL-NL2. Regarding

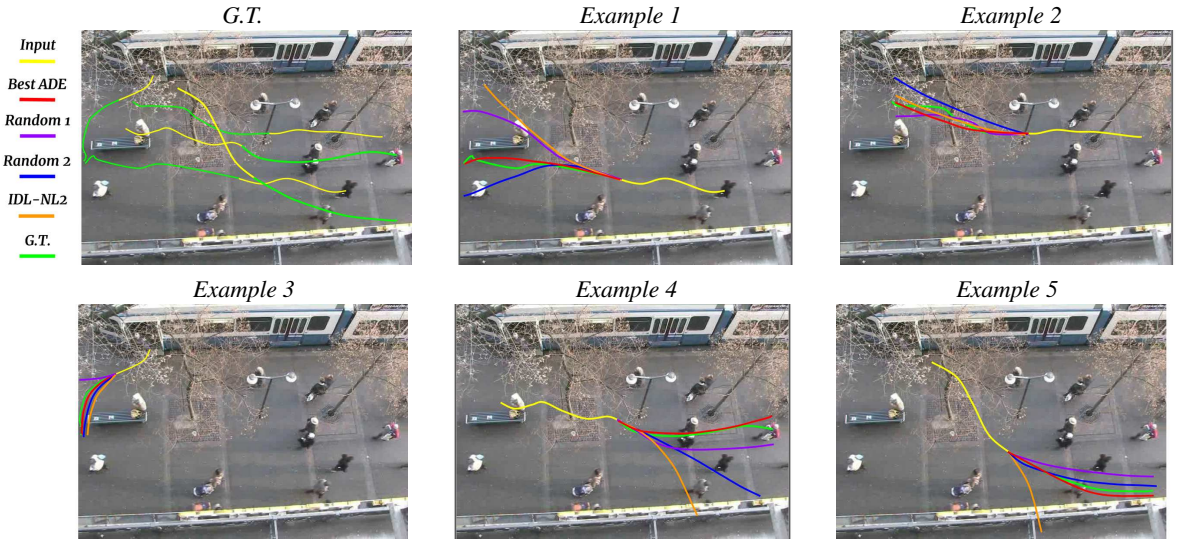

Figure 5. The visual results on ETH dataset. The top left depicts the entire trajectories combining the historical observations and ground truth future (G.T.) in a dynamic scene. For clearer visualizations, we isolate each input and corresponding multimodal predictions produced by our IDL and deterministic IDL-NL2 for comparison in each example.

Social LSTM [1], the second worst results are obtained on the ETH and UCY datasets due to lack of considering the multimodality.

Spatiotemporal Dependencies Processing: In this section, we verify the effectiveness of jointly processing spatiotemporal dependencies. A significant overall poor performance of the methods with two-step settings, i.e., Social GAN [11], DESIRE [16] and Social LSTM [1], compared to our IDL can be observed. This finding meets our expectations that combining the spatial and temporal dependencies into one single framework is a better strategy. Additionally, we assign one IDL-IN framework per person as [11, 16, 1]. However, failing to consider the spatial dependencies hinders IDL-IN achieving satisfactory results.

4.3. Qualitative Evaluation

Since neither of the best ADE/FDE perfectly captures the perceptual fidelity of the multimodal nature of future paths, we make additional qualitative evaluations.

Fig.4 and Fig.5 illustrate examples of path forecasting of our proposed IDL on the SAP and ETH datasets. We highlight the prediction that obtains the best ADE scores and two randomly selected results. It is worth noting that IDL simultaneously forecasts the future paths of all moving objects.In order to better understand the various possible future paths, we also visualize the deterministic output from IDL-NL2 baseline and ground truth. It is evident that our IDL generates diverse forthcoming paths. Such diversity can be traced back to different latent decisions. For instance, in example 1 of Fig.4, we observe that the predictions of ``random 1´´and ``random 2´´exhibit two different types of future possibilities (going straight and turning right, respectively). Meanwhile, IDL also successfully foresees the true future path of turning left, as the result of ``best ADE´´indicates. Conversely, the IDL-NL2 produces a deterministic path with a large discrepancy to ground truth due to disregard the inherent multimodality of future paths.

5. Conclusion

In this paper, we propose a novel Imitative Decision Learning approach for multimodal path forecasting in dynamic scenes. Our IDL delves deeper into the latent decision that shapes the multimodality to anticipate multiple plausible outcomes. Moreover, Our approach enables the processing of the spatiotemporal information in one unified framework. We extensively assess the performance of the proposed IDL on two large-scale datasets in a path forecasting challenge. We demonstrate that IDL is capable of producing diverse future paths as shown in our visual examples. Additionally, our IDL outperforms the recent prominent studies by quantitative justifications

We believe that our IDL can benefit future studies of real-world applications by imitating human decision-making process. For instance, one interesting direction would be to extend our framework to enable a self-navigating robot or an autonomous vehicle choosing an optimal path in dynamic scenes after foreseeing multiple possibilities.

6. Acknowledgement

The author was with York University, Canada. This work was funded by VISTA and Canada First Research Excellence Fund (CFREF). The author sincerely appreciates the consistent help and thought-provoking discussions from Prof. James Elder. The author is grateful to the insightful suggestions from Dr. Ling Wang and Dr. Henrique Morimitsu as well.

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 (CVPR)*, pages 961–971, 2016. 1, 2, 6, 7, 8

[2] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 2203–2210, 2014. 2

[3] Lamberto Ballan, Francesco Castaldo, Alexandre Alahi, Francesco Palmieri, and Silvio Savarese. Knowledge Transfer for Scene-specific Motion Prediction. In *European Conference on Computer Vision (ECCV)*, pages 697–713. Springer, 2016. 2

[4] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations (ICLR)*, 2016. 4

[5] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning(ICML)*, volume 80, pages 531–540, 10–15 Jul 2018. 2, 5

[6] Liqun Chen, Shuyang Dai, Yunchen Pu, Erjin Zhou, Chunyuan Li, Qinliang Su, Changyou Chen, and Lawrence Carin. Symmetric variational autoencoder and connections to adversarial learning. In *International Conference on Artificial Intelligence and Statistics*, pages 661–669, 2018. 3

[7] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 2, 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 3, 5

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 5

[11] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 6, 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pages 770–778, 2016. 2

[13] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016. 2, 3, 5

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, (1):221–231, 2013. 2

[15] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision (ECCV)*, pages 201–214. Springer, 2012. 2

[16] 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. 1, 2, 3, 6, 7, 8

[17] Peng Lei and Sinisa Todorovic. Temporal Deformable Residual Networks for Action Segmentation in Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6742–6751, 2018. 3

[18] Yuke Li. Pedestrian Path Forecasting in Crowd: A Deep Spatio-Temporal Perspective. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 235–243, New York, NY, USA, 2017. ACM. 2

[19] Yuke. Li. A Deep Spatiotemporal Perspective for Understanding Crowd Behavior. *IEEE Transactions on Multimedia*, 20(12):3289–3297, Dec 2018. 2

[20] Yuke Li. Video Forecasting with Forward-Backward-Net: Delving Deeper into Spatiotemporal Consistency. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 211–219. ACM, 2018. 3

[21] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pages 3812–3822, 2017. 5

[22] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 3

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[25] Pauline Luc, Camille Couprie, Yann LeCun, and Jakob Verbeek. Predicting Future Instance Segmentation by Forecasting Convolutional Features. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[26] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting With a Single Convolutional Net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3569–3577, 2018. 1, 2

[27] Zhongjian Lv, Jiajie Xu, Kai Zheng, Hongzhi Yin, Pengpeng Zhao, and Xiaofang Zhou. LC-RNN: A Deep Learning Model for Traffic Speed Prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3470–3476, 7 2018. 2

[28] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4636–4644. IEEE, 2017. 2

[29] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *International Conference on Machine Learning*, pages 3478–3487, 2018. 3

[30] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 3

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[32] S Pellegrini, A Ess, K Schindler, and L van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 2009. 6

[33] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[34] Nicholas Rhinehart and Kris M Kitani. First-Person Activity Forecasting With Online Inverse Reinforcement Learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 2

[35] 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. 2, 6

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4, 5

[37] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015. 3

[38] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised Learning of Video Representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 843–852, 2015. 2

[39] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1501–1510, 2017. 2

[40] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating Traffic Accidents With Adaptive Loss and Large-Scale Incident DB. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[41] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2018. 3

[42] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3302–3309. IEEE, 2014. 2

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 5, 2018. 3

[44] Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, and Xiaokang Yang. Structure Preserving Video Prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[45] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding Crowd Interaction With Deep Neural Network for Pedestrian Trajectory Prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[46] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian Behavior Understanding and Prediction with Deep Neural Networks. In *European Conference on Computer Vision (ECCV)*, pages 263–279. Springer, 2016. 2

[47] YoungJoon Yoo, Kimin Yun, Sangdoo Yun, JongHee Hong, Hawook Jeong, and Jin Young Choi. Visual Path Prediction in Complex Scenes With Crowded Moving Objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[48] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3634–3640, 2018. 2

[49] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 2