Reciprocal Twin Networks for Pedestrian Motion Learning and Future Path Prediction

Hao Sun, *Student Member, IEEE*, Zhiqun Zhao, *Student Member, IEEE*, Zhaozheng Yin, *Senior Member, IEEE*, and Zhihai He, *Fellow, IEEE*

***Abstract*—Modeling the moving behaviors and predicting the future paths of pedestrians, especially for those in complex scenes, remain a challenging problem in machine learning. We recognize that human motion trajectories, governed by social norms and constrained by physical structures of the surrounding environment, are both forward predictable and backward predictable. Motivated by this observation, we develop a new approach, called *reciprocal twin networks*, for human trajectory learning and prediction. We design two networks, a forward prediction network to predict future trajectory from past observations and a backward prediction that performs the trajectory prediction backward in time. The backward prediction network serves as the inverse operation of the forward prediction network, forming a reciprocal constraint. During the training stage, this reciprocal constraint allows them to be jointly learned for accurate and robust human trajectory prediction. During the inference stage, we borrow the concept of adversarial attack of deep neural networks, which iteratively modifies the input of the network to match the given or forced network output, and develop a new method, called *reciprocal attack for matched prediction*, to achieve accurate human trajectory prediction. Our experimental results on benchmark datasets demonstrate that our new method outperforms the state-of-the-art methods for human trajectory prediction.**

***Index Terms*—Human Trajectory Prediction, Deep Learning, Reciprocal Learning, Reciprocal Attack, Generative Adversarial Networks.**

I. Introduction

LEARNING and predicting human motion trajectories in complex environments plays an important role in autonomous driving systems [1]–[3], social robots [4], [5], human-machine interactions [6], [7], and smart environments [8], [9]. Human beings have the intelligence to understand the moving patterns and intentions of surrounding persons in the environment and act appropriately to avoid collision and follow social norms. Can a machine or robot do this? The problem of human trajectory prediction is different from person tracking [4]. It needs to learn the human decision and behaviors in complex environments to predict future motion trajectory during the next period of time (e.g., 5 seconds), instead of the next time instance. Researches recognize that human motion trajectories and motion patterns are governed by human perception, behavioral reasoning, common sense rules, social conventions, and interactions with others and the surrounding environment [1], [8].

Predicting human motion and modeling their common sense behaviors are a very challenging task [10]. An efficient algorithm for human trajectory prediction needs to accomplish the following tasks: (1) *Obeying physical constraints of the environment.* To walk on a feasible terrain and avoid obstacles or other physical constraints, we need to analyze the local and global spatial information surrounding the person and pay attention to important elements in the environment. (2) *Anticipating movements of other persons or vehicles and their social behaviors.* Some trajectories are physically possible but socially unacceptable. Human motions are governed by social norms, such as yielding right-of-way or respecting personal space. To capture and model them is a non-trivial task. (3) *Finding multiple feasible paths.* There are often a number of choices of moving trajectories for us to reach to the destination. This uncertainty poses significant challenges for accurate human trajectory prediction.

Recently, a number of methods based on deep neural networks have been developed for human trajectory prediction [10], [11]. Earlier methods have been focused on learning dynamic patterns of moving agents (human and vehicles) [10] and modeling the semantics of the navigation environment [12]. More recent approaches incorporate interactions between all agents in the scene into the analysis in order to predict the future trajectory for each agent. Methods have been developed to model human-human interactions [13], understand social acceptability using data-driven techniques based on Recurrent Neural Networks (RNNs) [11], [14], [15], and model the joint influence of all agents in the scene [16]. Methods have also been developed to predict multiple feasible paths of human [11], [15], [17].

In this work, we propose to explore the unique characteristics of human trajectories and develop a new approach, called *reciprocal learning* for human trajectory prediction. As illustrated in Fig. 1, we observe that the human trajectory is not only forward predictable, but also backward predictable. Imagine that the time is reversed and person is traveling backwards. As discussed in the above, the forward moving trajectories follow the social norm and obey the environmental constraints. So do the backward moving trajectories since the only difference between them is that their directions of time. From the training data, we can train two different prediction networks, the forward prediction network $\mathbf{F}_\theta$ which predicts

H. Sun, Z. Zhao, and Z. He are with the Department of Electrical Engineering and Computer Science, University of Missouri.

Z. Yin is with the Department of Computer Science, State University of New York, Stony Brook.

Corresponding author: Zhihai He, e-mail: hezhi@missouri.edu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

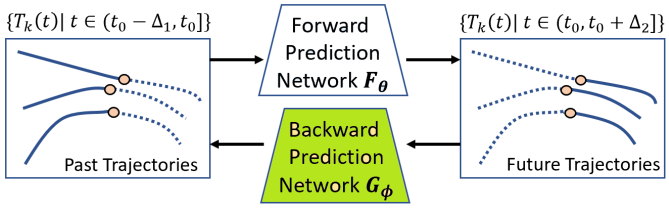


Fig. 1. The human trajectory is not only *forward* predictable, but also *backward* predictable. This leads to our new approach of reciprocal coupling and learning between the forward and backward prediction networks for accurate human trajectory prediction.

the future motion trajectories from past observations and the backward prediction network $\mathbf{G}_\phi$ which performs the prediction backwards. These two networks are inverse operations to each other. In other words, they satisfy a reciprocal constraint. Specifically, using the forward prediction network, we can predict the future trajectory $\mathbf{Y} = \mathbf{F}_\theta(\mathbf{X})$ from the observed or known trajectory $\mathbf{X}$. If the prediction $\mathbf{Y}$ is accurate, then $\mathbf{G}_\phi(\mathbf{Y})$ must be equal to $\mathbf{X}$.

Based on this observation and the unique reciprocal constraint, we develop a new approach called *reciprocal network learning* for accurate and robust prediction of human trajectories. We introduce the reciprocal prediction loss and establish an iterative procedure for training these two tightly coupled networks. We borrow the concept of the adversarial attacks of deep neural networks which iteratively modifies the input of the network to match a given target or forced network output. We integrate the reciprocal constraint with the adversarial attack method to develop a new matched prediction method for human trajectory prediction. Our experimental results on benchmark datasets demonstrate that our new method outperforms the state-of-the-art methods for human trajectory prediction.

The rest of the paper is organized as follows. Section II reviews related work on human trajectory prediction. The proposed reciprocal network learning and matched prediction are presented in Section III. Section V presents the experimental results, performance comparisons, and ablation studies. Section VI summarizes our major contributions and concludes the paper.

II. Related Work and Major Contributions

Existing methods for human trajectory prediction mainly focus on modeling human-human interactions and human-scene interactions. Human-human models focus on learning human movements and how human interacts with others [15], [16]. Human-scene models also try to learn the dynamic contents of the background scenes to extract some visual features to help better understand human motions [10], [12], [18]–[23]. In this section, we review existing work, including human-human models and human-scene models for human trajectory prediction. We also discuss related work in sequence prediction using Recurrent Neural Networks (RNNs) [24]. Our work is inspired by generative models [25], [26] and the idea of cycle consistence [27]–[30] in visual tracking, relevant papers in these two areas are also reviewed in this section.

A. Human-Human Models for Trajectory Prediction

A number of methods have been developed in the literature to model human social interactions and behaviors in crowded scenes, such as people attempting to avoid walking into each other. Helbing and Molnar [13] introduced the Social Force Model to characterize social interactions among people in crowded scenes using coupled Langevin equations. In recent methods based on LSTM (Long Short Term Memory) [15], social pooling was introduced to share features and hidden representations between different agents. The key idea is to merge hidden states of nearby pedestrians to make each trajectory aware of its neighbourhood. [31] found out that groups of people moving coherently in one direction should be excluded from the above pooling mechanism. [16] used a Generative Adversarial Network (GAN) to discriminate between multiple feasible paths. Their pooling mechanism relies on relative positions between all pedestrians with the target pedestrian. This model is able to capture different movement styles but does not differentiate between structured and unstructured environments. [32] predicted human trajectories using a spatio-temporal graph to model both position evolution and interactions between pedestrians.

B. Human-Scene Models for Trajectory Prediction

Another set of methods for human trajectory prediction have focused on learning the effects of physical environments. For example, human tend to walk along the sidewalk, around a tree or other physical obstacles. Sadeghian *et al.* [33] considered both traveled areas and semantic context to predict social and context-aware positions using a GAN (Generative Adversarial Network). Liang *et al.* [34] proposed to use abstract scene semantic segmentation features and multi-scale location encoding for better predicting multiple plausible trajectories. [35] designed a probabilistic model and introduced a dynamic attention-based state encoder to encode agent interactions. [36] extracted multiple visual features, including each person's body keypoints and the scene semantic map to predict human behavior and model interaction with the surrounding environment. [14] has studied attractions towards static objects, such as artworks, which deflect straight paths in several scenarios such as museums. [10] proposed a Bayesian framework to predict unobserved paths from previously observed motions and to transfer learned motion patterns to new scenes. In [37], the dynamics and semantics for long-term trajectory predictions have been studied. Scene-LSTM [38] divided the static scene into grids and predicted pedestrian's location using LSTM. The CAR-Net method [39] integrated past observations with bird's eye view images and analyzed them using a two-levels attention mechanism.

Another area of research is the prediction of human motion trajectory from a moving vehicle perspective, which has important applications in autonomous driving and robot navigation [40], [41]. [40] designed an RNN encoder-decoder architecture to encode observed human locations and the ego-vehicle's odometry data. [41] considered the human visual features to further improve the performance. In this work,

we focus on predicting the future human trajectory from a stationary surveillance camera.

C. Recurrent Neural Networks for Sequence Prediction

This work is related to recurrent neural networks (RNNs) [24]. RNNs are widely used for sequence data analysis, *e.g.*, speech recognition [42]–[45], image captioning [46]–[51], machine translation [43] and video generation [52]. [53] recognized that the drawback of RNNs model is the lack of high-level and spatio-temporal structure. [15], [54], [55] have been proposed to learn complex interactions using multiple networks. [11] designed an RNN-based encoder-decoder framework and uses variational autoencoder (VAE) to predict the sequence. Alahi *et al.* proposed a so-called social pooling layer to capture the interactions of human within a certain range. [56] proposed a Hierarchical Concurrent Long Short-Term Concurrent Memory (H-LSTCM) to recognize and predict human interactions by utilizing a hierarchical LSTM to learn dynamic inter-related representations among all persons in a scene and designing a concurrent LSTM to aggregate these inter-related representations. [57] developed a recurrent architecture to predict the future sequences by jointly decomposing the memory states of an input sequence into a set of frequency components and then choosing a suitable set of state-frequency components.

D. Generative Networks and Cycle Consistency Learning

Generative Adversarial Networks (GANs) have been widely used and achieved impressive results in representation learning [58]–[60], image translation [61], [62] and image synthesis [63]–[66]. In this work, we adopt a GAN framework to force the generated future and past trajectories to be indistinguishable from the ground truth. Using transitivity as a way to regularize structured data has been explored. For example, in visual tracking, [28], [67] developed a forward-backward consistency constrain. In language processing, [68]–[70] studied human and machine translators to verify and improve translations based on back translation and reconciliation mechanisms. Cycle consistency has also been used for motion analysis [71], action prediction [72], 3D shape matching [73], dense semantic alignment [74], [75], depth estimation [76]–[78], and image-to-image translation [79], [80]. CycleGAN [80] introduces a cycle consistence constraint for learning a mapping to translate an image from the source domain into the target domain. Pang *et al.* [72] propose to use a bi-directional LSTM model for early actions prediction. It employs consistency learning by synthesizing future action and reconstructing observed action.

This work is related to reciprocal learning, which was recently developed for human re-identification [81], [82] and zero-shot image retrieval [83]. In this work, we introduce the reciprocal loss and design two tightly coupled prediction networks, the forward and backward prediction networks, which are jointly learned based on the reciprocal constraint. To the best of our knowledge, our work is the first to employ cycle-consistency in human trajectory prediction domain.

E. Major Contributions

The major contributions of this work can be summarized as follows. (1) We have established a forward and backward prediction network structure for human trajectory prediction, which satisfies the reciprocal prediction constraints. (2) Based on this constraint, we have developed a reciprocal learning approach to jointly train these two prediction networks in an collaborative and iterative manner. (3) Once the network is successfully trained, we have developed a new approach for network inference or testing by integrating the concept of adversarial attacks with the reciprocal constraint. It is able to iteratively refine the predicted trajectory by the forward network such that the reciprocal constraint is satisfied. (4) Our ablation studies have shown that the proposed new approach is very effective with significant contributions to the overall performance of our method, which outperforms other state-of-the-art methods in the literature.

III. Reciprocal Networks for Human Trajectory Prediction

In this section, we present our reciprocal network learning method for human trajectory prediction.

A. Problem Formulation

We follow the standard formulation of trajectory forecasting problem in the literature [32], [36]. With observed trajectories of all moving agents in the scene, including persons and vehicles, the task is to predict the moving trajectories of all agents for the next period of time, say 10 seconds, in the near future. Specifically, let $\mathbf{X} = X_1, X_2, \cdots, X_N$ be the trajectories of all human in the scene. Our task is to predict the future trajectories of all human $\hat{\mathbf{Y}} = \hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_N$ simultaneously. The input trajectory of human n is given by $X_n = (x_n^t, y_n^t)$ for time steps $t = 1, 2, \cdots, T_o$. The ground truth of future trajectory is given by $Y_n = (x_n^t, y_n^t)$ for time step $t = T_o + 1, \cdots, T_p$.

B. Method Overview

As illustrated in Fig. 1, in reciprocal learning, we are learning two coupling networks, the forward prediction network $\mathbf{F}_\theta$ which predicts the future trajectories $\mathbf{Y} = \mathbf{F}_\theta(\mathbf{X})$ from the past data $\mathbf{X}$, and the backward prediction network $\mathbf{G}_\phi$ which predicts the past trajectories $\mathbf{X} = \mathbf{G}_\phi(\mathbf{Y})$ from the future data $\mathbf{Y}$. It should be noted that, during training, both the past and future data are available. If both networks are well trained, then we should have following two reciprocal consistency constraints:

$$\mathbf{X} \approx \mathbf{G}_\phi(\mathbf{F}_\theta(\mathbf{X})), \tag{1}$$

$$\mathbf{Y} \approx \mathbf{F}_\theta(\mathbf{G}_\phi(\mathbf{Y})). \tag{2}$$

These two networks are able to help each other to improve the learning and prediction performance. Specifically, if the backward prediction network $\mathbf{G}_\phi$ is trained, we can use the reciprocal constraint (1) to double check the accuracy of the forward prediction network $\mathbf{F}_\theta$ and improve its performance during training. Likewise, if the forward prediction network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

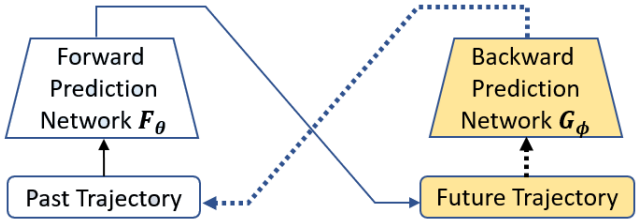


Fig. 2. Illustration of the proposed reciprocal learning approach.

$\mathbf{F}_\theta$ is trained, we can use (2) to improve the training performance of the backward prediction network $\mathbf{G}_\phi$. This results in a tightly coupled iterative learning and performance improvement process between these two prediction networks, as illustrated in Fig. 2. Once the forward and backward networks are successfully trained using the reciprocal learning approach, we develop a new network inference method called *reciprocal attack for matched prediction*. It borrows the concept of adversarial attacks of deep neural networks where the input is iteratively modified such that the network output matches a given target [84].

Our proposed idea is related to CycleGAN [80] which presents an approach for learning a mapping to translate an image from a source domain to a target domain. They also learn an inverse mapping and introduce the cycle consistence constraint. However, our approach is different from the CycleGAN method. We actually design two tightly coupled prediction networks, the forward and backward prediction networks, which are jointly learned based on the reciprocal constraint. During network inference, our approach introduces a new reciprocal attack method for matched prediction of human trajectory. The backward prediction network serves as a constraint to verify the prediction results generated by the forward prediction network so that it can iteratively optimize its prediction.

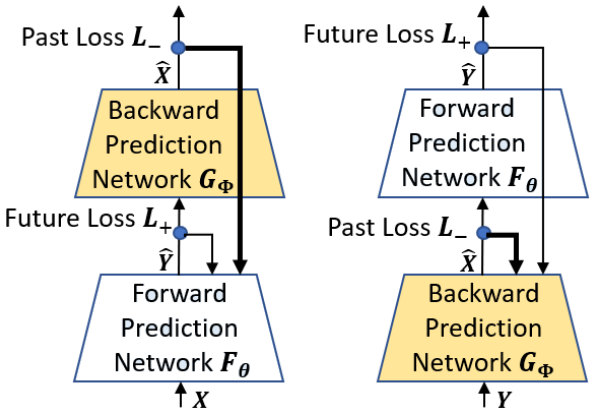


Fig. 3. Illustration of the training process of reciprocal learning.

C. Reciprocal Network Training

To successfully train the forward and backward prediction networks, we define two loss functions, J_- and J_+, to measure the prediction accuracy of the past and future trajectories. One reasonable choice will be the L_2 between the original trajectory and its prediction. These two loss functions will be updated alternatively and combined to guide the training of each of these two networks, as illustrated in Fig. 3. For example, when training the forward prediction network $\mathbf{F}_\theta$, the loss function used in existing literature is the prediction error of the future trajectory L_+. In reciprocal training, we first pre-train the backward prediction network $\mathbf{G}_\phi$ using the training data with all trajectories reversed in time. We then use this network to map the prediction result of $\mathbf{F}_\theta$, $\hat{\mathbf{Y}} = \mathbf{F}_\theta(\mathbf{X})$, back to the past trajectory, which is given by

$$\hat{\mathbf{X}} = \mathbf{G}_\phi(\hat{\mathbf{Y}}) = \mathbf{G}_\phi(\mathbf{F}_\theta(\mathbf{X})). \tag{3}$$

The past trajectory loss is then given by $L_- = ||\mathbf{X} - \hat{\mathbf{X}}||_2$. We refer to this loss as *reciprocal loss*. It will be combined with L_+ to form the loss function for the forward prediction network $\mathbf{F}_\theta$:

$$\begin{aligned} J_+[\theta] &= \lambda \cdot L_+ + (1-\lambda) \cdot L_- \\ &= \lambda \cdot ||\mathbf{Y} - \mathbf{F}_\theta(\mathbf{X})||_2 \\ &+ (1-\lambda) \cdot ||\mathbf{X} - \mathbf{G}_\phi(\mathbf{F}_\theta(\mathbf{X}))||_2. \end{aligned} \tag{4}$$

Similarly, we can derive the loss function for the backward prediction network $\mathbf{G}_\phi$:

$$\begin{aligned} J_-[\phi] &= \lambda \cdot L_- + (1-\lambda) \cdot L_+ \\ &= \lambda \cdot ||\mathbf{X} - \mathbf{G}_\phi(\mathbf{Y})||_2 \\ &+ (1-\lambda) \cdot ||\mathbf{Y} - \mathbf{F}_\theta(\mathbf{G}_\phi(\mathbf{Y}))||_2. \end{aligned} \tag{5}$$

In reciprocal training, we first pre-train the forward and backward prediction networks independently. Then, these two networks are jointly trained in an iterative manner based on the reciprocal constraint.

D. Constructing the Forward and Backward Prediction Networks

Both the forward and backward networks share the same network structure. In the following, we use the forward prediction network $\mathbf{F}_\theta$ as an example to explain our network design. As illustrated in Fig. 4, we adopt the existing Social-GAN in [16] as our baseline prediction network. Our model consists of two key components: (1) a feature extraction module and (2) an LSTM (Long Short Term Memory)-based GAN (generative adversarial network) module.

1) Feature Extraction: In real world scenarios, human's selection of future path is affected by the surrounding environment, including other persons in the neighborhood and the physical scene. Our feature extraction module has three major components to extract human-specific, scene context and depth structure features.

(a) *Human-specific features*. The human scale feature captures the temporal pattern and dependency of each human trajectory. Given the observed trajectories $\mathbf{X} = X_1, X_2, \cdots, X_N$ of all human in the scene, the input trajectory of each human n is defined as $X_n = (x_n^t, y_n^t)$ from time steps $t = 1, 2, \cdots, T_o$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

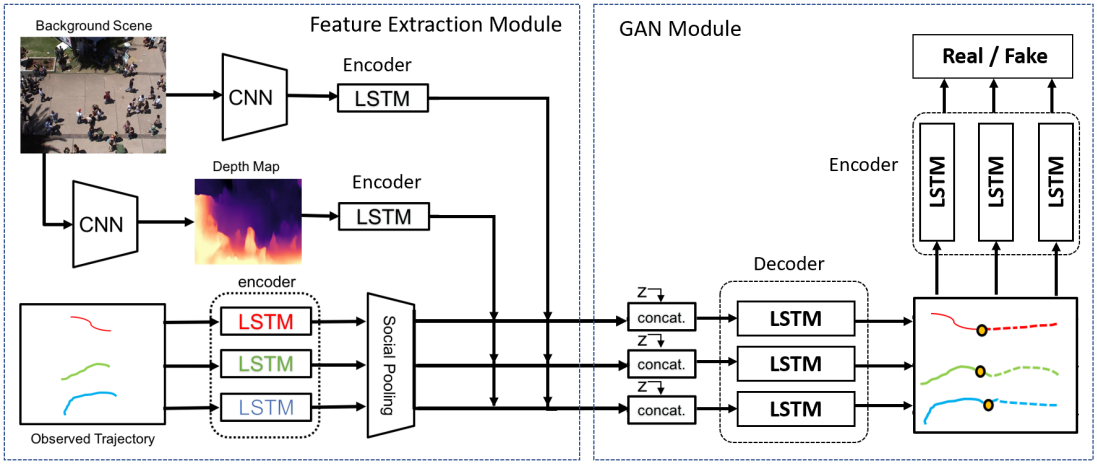


Fig. 4. Overview of our prediction model. Our model consists of two key components: (1) a feature extraction module, (2) an LSTM-based GAN module.

We first embed the coordinates of each human n into a fixed size vector e_n^t using a single layer MLP [85]:

$$e_n^t = \psi(x_n^t, y_n^t; W_{em}), \tag{6}$$

where $\psi(\cdot)$ is an embedding function with ReLU non-linearity [86] and W_{em} is the embedding weight. Then, we use an LSTM module to encode them into a high-dimensional feature $F_{h,n}^t$.

$$F_{h,n}^t = LSTM_{en}(F_{h,n}^{t-1}, e_n^t; W_{en1}), \tag{7}$$

where W_{en1} denotes the encoding weight which can be optimized during training process. Notice that W_{en1} is shared between all human in the scene. In order to capture the joint influence of all surrounding human's movements on the prediction of the target human n, we borrow the idea from [16] to build a social pooling module (SP) which encodes the human-human interactions. The relative distances between the target person and others are calculated. These distance vectors are concatenated with the hidden state in the LSTM network for each person and then embedded by an MLP and followed by a Max-Pooling function [87] to form the joint feature $F_{s,n}^t$.

$$F_{s,n}^t = SP(F_{h,1}^t, F_{h,2}^t, \cdots, F_{h,N}^t). \tag{8}$$

A maximum number of moving human in the scene is set and a default value of 0 is used if the corresponding agent does not exist in the current frame.

(b) *Scene context features*. As recognized in [17], [33], the environmental context affects the decision of the human in planning its next step of movement. Features of the current scene can be incorporated into the reasoning process. Similar to prior work [33], we use the VGGNet-19 network [88] pre-trained on the ImageNet [88] to extract the visual feature f^t of background scene I^t, which is then fed into an LSTM encoder to compute the hidden state tensor F_v^t.

$$f^t = VGG(I^t), \tag{9}$$

$$F_v^t = LSTM_{en}(F_v^{t-1}, f^t; W_{en2}), \tag{10}$$

where W_{en2} is the corresponding encoding weights.

(c) *Depth structure features*. As a unique feature of our proposed method, we propose to also incorporate the 3D scene depth structure into the reasoning process, which also improves the prediction accuracy of human trajectories. This is because the human motion occurs in the original 3D environment. Therefore, its natural behavior and motion patterns are better represented in the 3D instead of 2D coordinate system. For example, the trajectory of a person walking near the camera is much different from that of a person walking far away from the camera due to the camera perspective transform. To address this issue, we propose to estimate a depth map from a single image using existing depth estimation method [89]. We use the pre-trained model *Monodepth2*, denoted by $\mathbf{D}$ to perform monocular depth estimation and obtain the depth map $D^t = \mathbf{D}(I^t)$ of scene I^t, then use an LSTM to encode it into a depth feature F_d^t.

$$F_d^t = LSTM_{en}(F_d^{t-1}, D^t; W_{en3}), \tag{11}$$

where W_{en3} is the associated encoding weights. Fig. 5 presents qualitative depth estimation examples from Town Centre dataset [90] by *Monodepth2*.

Fig. 5. Examples of the input (left column) and the output (right column) of the monocular depth estimation [89]. The input image is from Town Centre dataset [90].

2) LSTM-based GAN for Trajectory Prediction: Inspired by previous work [16], [33], in this paper we use an LSTM based Generative Adversarial Network (GAN) module to generate human's future path as illustrated in Fig. 4. The generator is constructed by a decoder LSTM. Similar to the conditional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

GAN [9], a white noise vector Z is sampled from a multivariate normal distribution. Then, a merge layer is used in our proposed network which concatenates all encoded features mentioned above with the noise vector Z.

$$F_n^t = concat(F_s^t, F_v^t, F_d^t, Z), \tag{12}$$

We take F_n^t as the input to the LSTM decoder to generate the candidate future paths $\hat{Y}_n^t$ for each human.

$$\hat{Y}_n^t = LSTM_{de}(\hat{Y}_n^{t-1}, F_n^t; W_{de}), \tag{13}$$

where W_{de} is the decoding weights of LSTM.

The discriminator is built with an LSTM encoder which takes the input ${Y'}_n^t$ as randomly chosen trajectory from either ground truth Y_n^t or predicted trajectories $\hat{Y}_n^t$ and classifies them as "real" or "fake". Generally speaking, the discriminator classifies the trajectories which are not accurate as "fake" and forces the generator to generator more realistic and feasible trajectories.

$$L_n^t = LSTM_{en}({Y'}_n^t, h_{en}^t; W_{en4}), \tag{14}$$

where L_n^t is the predicted label from the discriminator for the chosen input trajectory to be "real"($L_n^t = 1$) or "fake"($L_n^t = 0$). h_{en}^t denotes the hidden state of the encoding LSTM and W_{en4} is the corresponding weights.

Within the framework of our reciprocal learning for human trajectory prediction, let $G^\theta : X \rightarrow Y$ and $G^\phi : Y \rightarrow X$ be the generators of the forward prediction network $\mathbf{F}_\theta$ and the backward prediction network $\mathbf{G}_\phi$, respectively. D^θ is the discriminator for $\mathbf{F}_\theta$. Its input Y' is randomly selected from either ground truth Y or the predicted future trajectory $\hat{Y}$. Similarly, D^ϕ is discriminator for $\mathbf{G}_\phi$. To train $\mathbf{F}_\theta$ and $\mathbf{G}_\phi$, we combine the adversarial loss with the forward prediction loss $J_+[\theta]$ and the backward prediction loss $J_-[\phi]$ in Eqs. (4) and (5) together to construct the overall loss function for $\mathbf{F}_\theta$ and $\mathbf{G}_\phi$, respectively:

$$\mathcal{L}_\theta = L_{GAN}^\theta + J_+[\theta], \quad \mathcal{L}_\phi = L_{GAN}^\phi + J_-[\phi], \tag{15}$$

where adversarial losses L_{GAN}^θ and L_{GAN}^ϕ are defined as:

$$\begin{aligned} L_{GAN}^\theta &= \min_{G^\theta} \max_{D^\theta} \mathbb{E}_{Y' \sim p(Y,\hat{Y})}[\log D^\theta(Y')] \\ &+ \mathbb{E}_{X \sim p(X), Z \sim p(Z)}[\log(1 - D^\theta(G^\theta(X, Z)))], \end{aligned} \tag{16}$$

$$\begin{aligned} L_{GAN}^\phi &= \min_{G^\phi} \max_{D^\phi} \mathbb{E}_{X' \sim p(X,\hat{X})}[\log D^\phi(X')] \\ &+ \mathbb{E}_{Y \sim p(Y), Z \sim p(Z)}[\log(1 - D^\phi(G^\phi(Y, Z)))]. \end{aligned} \tag{17}$$

IV. Reciprocal Attack for Matched Prediction of Human Trajectories

Once the forward and backward networks are successfully trained with the above loss functions based on the reciprocal learning approach, we are ready to perform prediction of the human trajectories. By taking advantage of the reciprocal property of the forward and backward networks, we develop a new network inference method called *reciprocal attack for matched prediction* to achieve improved performance in human trajectory prediction.

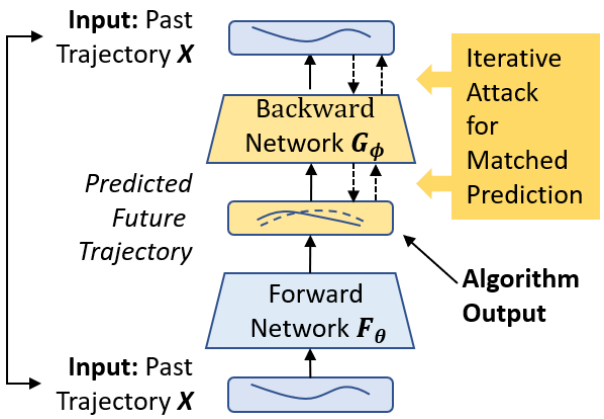


Fig. 6. Illustration of the proposed reciprocal attack method.

As illustrated in Fig. 6, $\mathbf{F}_\theta$ is our trained network for human trajectory prediction. With the past trajectories $\mathbf{X}$ as input, it predicts the future trajectories $\hat{\mathbf{Y}} = \mathbf{F}_\theta(\mathbf{X})$. During network testing or actual prediction, we do not know the ground truth of the future trajectory. How do we know if this prediction $\hat{\mathbf{Y}}$ is accurate or not? How can we further improve its accuracy? Fortunately, in our reciprocal learning framework, we have another network, the backward prediction network $\mathbf{G}_\phi$, which can be used to map the estimated $\hat{\mathbf{Y}}$ back to the known input $\mathbf{X}$. Our idea is that, if $\hat{\mathbf{Y}}$ is accurate, then its backward prediction $\hat{\mathbf{X}} = \mathbf{G}_\phi(\hat{\mathbf{Y}}) = \mathbf{G}_\phi(\mathbf{F}_\theta(\mathbf{X}))$ should match the original input $\mathbf{X}$. When the prediction $\hat{\mathbf{Y}}$ is not accurate, we can modify the prediction such that the above matching error is minimized. This leads to the following optimization problem:

$$\hat{\mathbf{Y}}^* = \underset{\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} + \Delta(t)}{\arg\min} ||\mathbf{X} - \mathbf{G}_\phi(\tilde{\mathbf{Y}})||_2. \tag{18}$$

Here, $\Delta(t)$ is the small perturbation or modification added to the existing prediction result $\hat{\mathbf{Y}}$. The above optimization procedure aims to find the best modification $\hat{\mathbf{Y}}^* = \hat{\mathbf{Y}} + \Delta(t)$ to minimize the matching error.

This optimization problem can be solved by adversarial attack methods recently studied in the literature of deep neural network attack and defense. In this work, we propose to borrow the idea from the famous Fast Gradient Sign method (FGSM) developed by Goodfellow *et al.* [84] to perform adversarial attacks. Essentially, it is the same error back propagation procedure as network training. The only difference is that network training modifies the network weights based on error gradients. However, the adversarial attack does not modify the network weights, it propagates the error all the way to the input layer to modify the original input image to minimize the loss.

This approach uses the sign of the gradient at each pixel to determine the direction of changing pixel value. In our case, we remove the sign function and directly use the gradient to update the input trajectory. With the matching error of human trajectories $E = ||\mathbf{X} - \mathbf{G}_\phi(\tilde{\mathbf{Y}})||_2$, we can perform multiple iterations of the modified FGSM attack on the prediction $\hat{\mathbf{Y}}$ such that the matching error is minimized. At iteration m, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

attacked trajectory (input) is given by

$$\hat{\mathbf{Y}}^m = \hat{\mathbf{Y}}^{m-1} - \epsilon \cdot \nabla_{\hat{\mathbf{Y}}} E(\mathbf{X}, \hat{\mathbf{Y}}^{m-1}), \tag{19}$$

with $\hat{\mathbf{Y}}^0 = \hat{\mathbf{Y}}$. ϵ is the magnitude of attacks [84]. $\nabla_{\hat{\mathbf{Y}}} E(\mathbf{X}, \hat{\mathbf{Y}}^{m-1})$ indicates the gradient of error function E with respect to the input $\hat{\mathbf{Y}}$. Intuitively, the updated trajectory $\hat{\mathbf{Y}}^m$ will minimize E. We then perform an exponential average of $\{\hat{\mathbf{Y}}^m\}$ to obtain the improved prediction

$$\hat{\mathbf{Y}}^* = \left[\sum_{m=1}^{M} e^{\alpha \cdot m} \cdot \hat{\mathbf{Y}}^m\right] / \sum_{m=1}^{M} e^{\alpha \cdot m}, \tag{20}$$

where M is the total iterations and α is a constant to control the relative weights between these different iterations of attacks.

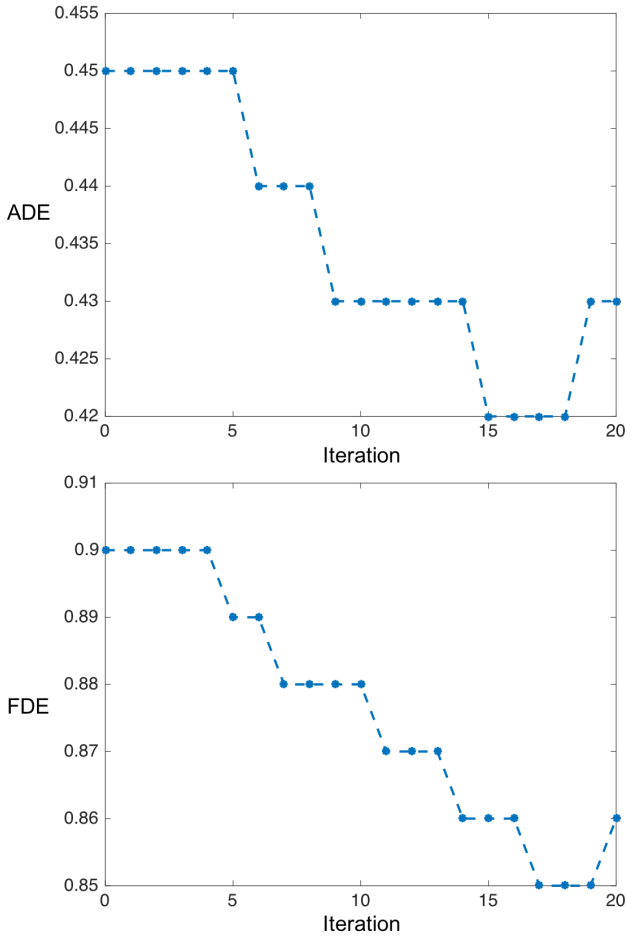


Fig. 7. Example of our proposed reciprocal attack performed on HOTEL dataset. X-coordinates in both figures indicate the iteration, while Y-coordinates indicate error metrics, ADE and FDE (see detailed explanation in Section V-C).

Fig. 7 shows an example of the trajectory prediction results using reciprocal attack in each iteration performed on the HOTEL dataset. Note that, the reciprocal attack is only performed during the testing phase. Once the forward and backward networks are well trained based on the reciprocal learning approach, we are ready to perform prediction of the human trajectories. We can see that, using reciprocal attack in an iterative manner, the error metrics, ADE and FDE (see definition in Section V-C) of trajectory prediction will decrease in a certain iteration, then they might increase after a few iterations. Since we do not know the ground truth, we choose to perform reciprocal attack for 20 iterations based on heuristic studies. Then an exponential average is performed on the result trajectories in each iteration to obtain the refined future trajectories prediction. The ablation studies in the following section will provide more results to demonstrate the effectiveness of this attacked-based matched prediction scheme.

V. Experimental Results

We provide extensive performance comparisons on benchmark datasets (ETH [91] and UCY [92]) between our work and state-of-the-art methods. We also conduct ablation to demonstrate the effectiveness of each algorithm component. To further evaluate the generalization capability of our method on predicting human future trajectories, we conduct experiments on two new datasets: Town Centre [90] and Grand Central Station [93].

A. Datasets

Performance comparisons and ablation studies are performed on the ETH [91] and UCY [92] datasets, which contain real world human trajectories and various natural human-human interaction scenarios. In total, 5 sub-datasets, ETH, HOTEL, UNIV, ZARA1 and ZARA2, are included in these two datasets. Each set contains bird's-eye view images and 2D locations of each human. In total there are 1536 humans in these 5 datasets. They contain challenging situations, including human collision avoidance, human crossing each other, and dynamic group behaviors. Each scene occurs in a unconstrained outdoor environment [33].

Generalization studies are performed on the Town Centre [90] and Grand Central Station [93] datasets. The Town Center dataset contains short videos with frequent human-human and human-scene interactions. It is originally used for human tracking tasks with bounding boxes for the head and body for each human. In the experiment, we use the center of the human body bounding box as the location, as in existing methods [17], [94]. The Grand Center Station dataset contains a long-duration video (more than 32 minutes) and consists of about 12,600 pedestrians with frequent human interactions. It is originally used for human behavior analysis.

B. Implementation Details

Our GAN model is constructed using the LSTM for the encoder and decoder. The generator and discriminator are trained iteratively with the Adam optimizer. We choose the batch size of 64 and the initial learning rate of 0.001. The whole model is trained for 200 epochs. The trajectories are embedded using a single layer MLP with dimension of 16. The encoder and decoder for the generator use an LSTM with the hidden state's dimension of 32. In the LSTM encoder for the discriminator, the hidden state's dimension is 48. In the pooling module, we follow the procedure and setting in [16].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
COMPARISONS OF DIFFERENT METHODS ON ETH (COLUMN 3 AND 4) AND UCY (COLUMN 5-7) DATASETS.

Metric	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
ADE	Linear [16]	1.33	0.39	0.82	0.62	0.77	0.79
	LSTM [15]	1.09	0.86	0.61	0.41	0.52	0.70
	S-LSTM [15]	1.09	0.79	0.67	0.47	0.56	0.72
	S-GAN [16]	0.81	0.72	0.60	0.34	0.42	0.58
	S-GAN-P [16]	0.87	0.67	0.76	0.35	0.42	0.61
	SoPhie [33]	0.70	0.76	0.54	0.30	0.38	0.54
	Scene-LSTM [94]	**0.36**	0.95	0.63	0.45	0.40	0.56
	Next [36]	0.73	**0.30**	0.60	0.38	0.31	0.46
	Ours	0.69	0.43	**0.53**	**0.28**	**0.28**	**0.44**
FDE	Linear [16]	2.94	0.72	1.59	1.21	1.48	1.59
	LSTM [15]	2.14	1.91	1.31	0.88	1.11	1.52
	S-LSTM [15]	2.35	1.76	1.40	1.00	1.17	1.54
	S-GAN [16]	1.52	1.61	1.26	0.69	0.84	1.18
	S-GAN-P [16]	1.62	1.37	1.52	0.68	0.84	1.21
	SoPhie [33]	1.43	1.67	1.24	0.63	0.78	1.15
	Scene-LSTM [94]	**0.67**	1.77	1.41	1.00	0.90	1.15
	Next [36]	1.65	**0.59**	1.27	0.81	0.68	1.00
	Ours	1.24	0.87	**1.17**	**0.61**	**0.59**	**0.90**

TABLE II
ABLATION EXPERIMENTS OF OUR FULL ALGORITHM WITHOUT DIFFERENT COMPONENTS. ERROR METRICS REPORTED ARE ADE AND FDE IN METER SCALE.

Metric	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
ADE	Our Method (Full Algorithm)	0.69	0.43	0.53	0.28	0.28	0.44
	- Without Reciprocal Learning	0.73	0.49	0.60	0.38	0.36	0.51
	- Without Depth Features	0.71	0.43	0.56	0.31	0.31	0.46
	- Without Reciprocal Attacks	0.70	0.45	0.55	0.32	0.30	0.46
FDE	Our Method (Full Algorithm)	1.24	0.87	1.17	0.61	0.59	0.90
	- Without Reciprocal Learning	1.31	0.97	1.22	0.73	0.70	0.99
	- Without Depth Features	1.30	0.88	1.19	0.63	0.62	0.92
	- Without Reciprocal Attacks	1.26	0.90	1.18	0.65	0.61	0.92

The maximum number of human surrounding the target human is set to 32. This value is chosen since in all datasets, none of them has more than 32 human in any frame. For the feature extraction part, following the prior work [33], we use the VGG feature with a size of 512 for the background scene, which is then embedded using a single MLP to a dimension of 16. For the depth map estimation, we use the pre-trained model *Monodepth2* from [89] and the depth feature is also embedded using a single layer MLP with a dimension of 16. The weight for our loss function is $\lambda = 0.5$. We perform the reciprocal attack for 20 iterations, the perturbation ϵ is set as 0.05.

C. Evaluation Metrics and Methods

Following the standard evaluation procedure [15], [95], we use the following two error metrics for performance evaluations. (1) Average Displacement Error (ADE) is the average L_2 distance between the ground truth trajectory and our prediction over all predicted time steps from $T_o + 1$ to T_p. (2) Final Displacement Error (FDE) is the Euclidean distance between the predicted final destination and the true final destination at end of the prediction period T_p. They are defined as:

$$\text{ADE} = \frac{\sum_{n \in \Psi} \sum_{t=T_o+1}^{T_p} \sqrt{(\hat{x}_n^t - x_n^t)^2 + (\hat{y}_n^t - y_n^t)^2}}{|\Psi| \cdot T_p}, \quad (21)$$

$$\text{FDE} = \frac{\sum_{n \in \Psi} \sqrt{(\hat{x}_n^{T_p} - x_n^{T_p})^2 + (\hat{y}_n^{T_p} - y_n^{T_p})^2}}{|\Psi|}, \quad (22)$$

where $(\hat{x}_n^t, \hat{y}_n^t)$ and (x_n^t, y_n^t) are the predicted and ground truth coordinates for human n at time t, Ψ is the set of human and $|\Psi|$ is the total number of human in the test set.

Following existing methods [15], [16], [33], we use the leave-one-out evaluation protocol on the ETH and UCY datasets. Specifically, four datasets are used for training and the remaining one is used for testing. Given the human trajectory for the past 8 time steps (3.2 seconds), our model predicts the future trajectory for next 12 time steps (4.8 seconds). In our generalization studies, following the previous work [94], we split the data of Town Centre and Grand Central Station into one half for training and the other half for testing. All location coordinates are normalized to [0, 1] for training and testing.

D. Comparison with Existing Methods

We compare our method against the following state-of-the-art methods: (1) *Linear* [16]: This method applies a linear regression to estimate linear parameters by minimizing the least square error [16]. (2) *LSTM* [15]: This is the baseline model for the LSTM-based method, which does not consider human-human interactions or background scene information. (3) *S-LSTM* [15]: This method models each human by an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III
AVERAGE PERCENTAGE OF COLLIDING HUMAN FOR EACH SCENE IN ETH AND UCY DATASETS. THE FIRST COLUMN REPRESENTS THE GROUND TRUTH.

	GT	Linear [16]	S-GAN [16]	SoPhie [33]	Ours
ETH	0.000	3.137	2.509	1.757	**1.512**
HOTEL	0.092	1.568	1.752	1.936	**1.547**
UNIV	0.124	1.242	**0.559**	0.621	0.563
ZARA1	0.000	3.776	1.749	**1.027**	1.094
ZARA2	0.732	3.631	2.020	1.464	**1.252**
Avg	0.189	2.670	1.717	1.361	**1.194**

LSTM and proposes a social pooling mechanism. Both *S-LSTM* and *LSTM* generate one trajectory for each observation. (4) *S-GAN* [16]: This is one of the first GAN-based methods. During the pooling stage, all human in the scene are considered. *S-GAN* and *S-GAN-P* are different only in whether the pooling mechanism is applied or not. The method chooses the best trajectory from 20 network predictions as the final test result. (5) *SoPhie* [33]: This work implements a so-called physical constrain described by background scene features. Also the attention mechanism is used in this GAN-based method. (6) *Scene-LSTM* [94]: This method imposes a two-level grid structure on the scene to incorporate the scene information with human movements. (7) *Next* [36]: This method introduces a LSTM-based predictor with pooling of multiple features. In the test part, besides using a single model, it follows [16] to train 20 different models with random initialization. In our comparison, we follow [36] to report the minimum ADE and FDE over 20 outputs.

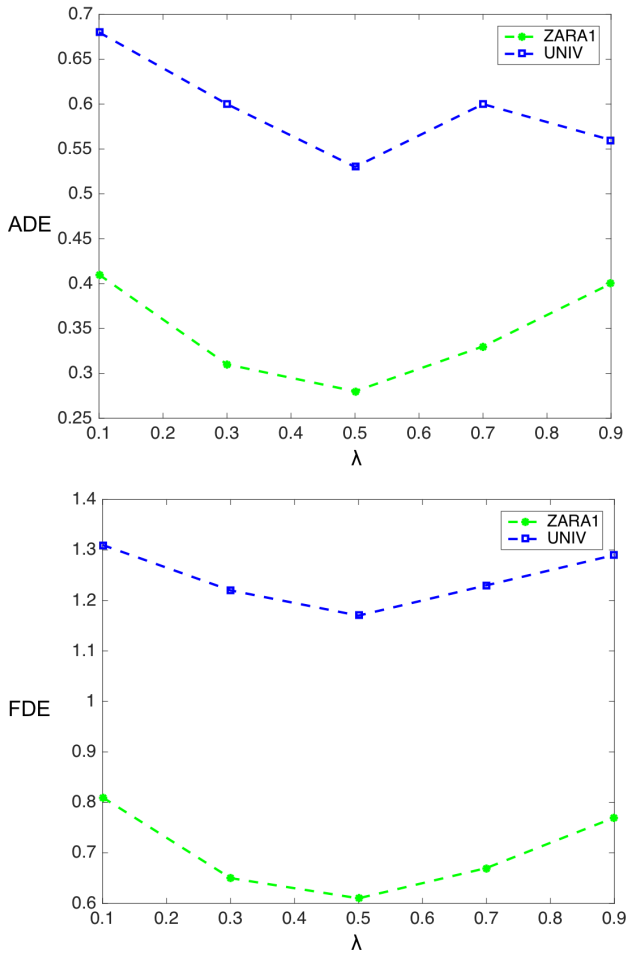


Fig. 8. Illustration of ADE and FDE changes with respect to different λ values on ZARA1 and UNIV dataset

E. Quantitative Results

Table I shows the comparison results of our method against existing methods on the above two performance metrics ADE and FDE. As illustrated in Table I, our method outperforms all other methods except on the ETH dataset against *Scene-LSTM* and on the Hotel dataset against *Next*. We can see that the *Linear* method has the lowest accuracy, it can only predict the straight trajectory and have very poor performance in videos with complicated human-human and human-environment interactions. *LSTM* performs better than *Linear* since it can handle more complicated trajectories. *S-LSTM* also outperforms the Linear model since it uses the social pooling mechanism, but it performs worse than *LSTM*. According to [16], the *S-LSTM* is trained on a synthetic dataset and fine-tuned on the real dataset to improve the accuracy. *Scene-LSTM* achieves better results than *S-LSTM* since it incorporates the scene information as well as human movements. Both *SoPhie* and *Next* outperform the *S-GAN* due to the use of background visual features and the attention module. Overall, our method achieves the best average error metrics in both ADE and FDE among all comparison methods.

To evaluate the performance of our method in predicting feasible paths in crowded scenes, we follow the procedure in previous papers [33] to report a new evaluation metric which is the percentage of *near-collisions* among humans. A collision is defined when the Euclidean distance between two human is smaller than 0.1m. We compute the average percentage of human near-collision in each frame of ETH and UCY datasets. The comparison results against the *Linear*, *S-GAN* and *SoPhie* are shown in Table III. We can see that our method outperforms these three methods on the ETH, HOTEL, and ZARA2 datasets, producing less human collision in the future time. On the other two datasets, UNIV and ZARA1, *S-GAN* and *SoPhie* perform slightly better than ours. However, they suffer from significant performance degradation on other datasets.

F. Ablation Studies

To systematically evaluate our method and study the contribution of each algorithm component, we perform a number of ablation experiments in Table II. Our algorithm has three major new components, the reciprocal learning, the incorporation of 3D depth map features, and the reciprocal attacks for matched prediction. In the first row of Table II, we list the ADE and FDE results for our method (full algorithm). The second row shows the results for our method without reciprocal training. The third row shows results without depth map features. The last row shows results without reciprocal attacks for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

TABLE IV
THE QUANTITATIVE RESULTS (ADE AND FDE) ON TOWN CENTRE AND GRAND CENTRAL STATION DATASETS WITH DIFFERENT PREDICTION LENGTHS OF FUTURE TRAJECTORIES.

Metrics	Datasets	Prediction Length	S-GAN [16]	S-GAN-P [16]	Scene-LSTM [94]	Ours
ADE	Town Center	12	0.22	0.21	0.09	**0.07**
		16	0.37	0.38	0.14	**0.09**
	Grand Central Station	12	0.21	0.40	0.11	**0.06**
		16	0.32	0.79	0.14	**0.07**
FDE	Town Center	12	0.46	0.42	0.18	**0.13**
		16	0.80	0.81	0.27	**0.18**
	Grand Central Station	12	0.45	0.74	0.17	**0.11**
		16	0.62	1.50	0.25	**0.15**

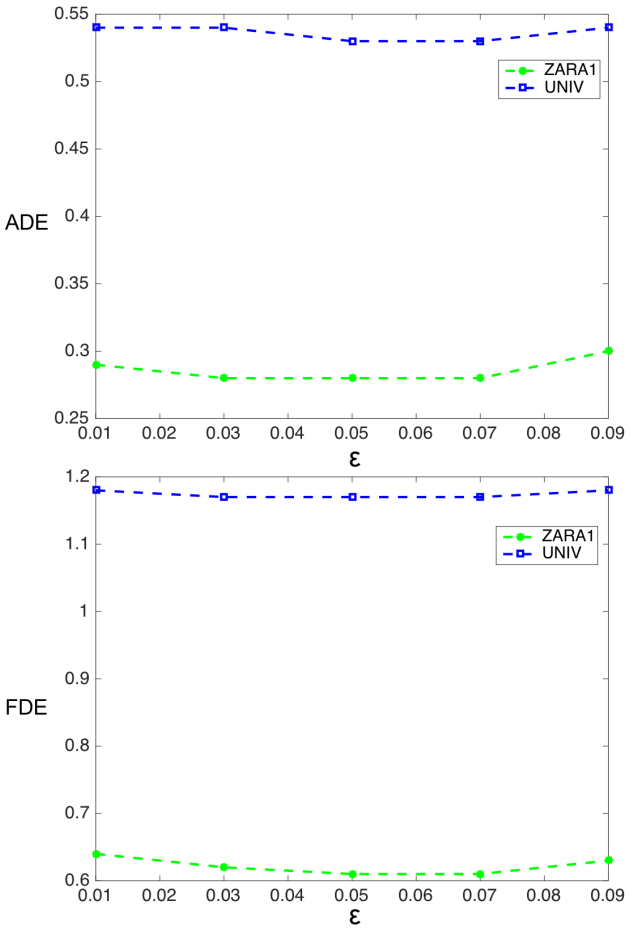


Fig. 9. Illustration of ADE and FDE changes with respect to different ϵ values on the ZARA1 dataset

prediction. We can clearly see that each algorithm component is contributing to the overall performance.

With the reciprocal consistence constraints, during training, our model forces the backward predicted trajectory to be consistent with the observed past trajectory, thus the predicted future trajectory which is the input of the backward network will be forced to be closer to the ground truth. As shown in the 2nd and 6th rows of Table II, the ADE increases to 0.51 from 0.44 and FDE increases to 0.99 from 0.90 on average when reciprocal consistence is excluded. By adding the depth features and reciprocal attacks, the prediction can be slightly refined to further improve the performance. Results in the 3rd and 7th rows shown in Table II shows the benefit of introducing the depth features since it can help the model to better understand human behavior and the background scene context. The reciprocal attack mechanism modifies the predicted trajectory in an iterative manner to match the original trajectory with the backward prediction network. The minor improvement of this proposed mechanism is clearly shown in the 4th and 8th rows of Table II. With all these ablation experimental results, we can conclude that all three algorithm components are critical in our proposed method.

To evaluate the influence of the parameter λ in Eqn. 4 and 5, we perform ablation experiments on ZARA1 and UNIV datasets with λ value from 0.1 to 0.9. Fig. 8 presents how ADE and FDE changes with respect to different λ values. As we can see in Fig. 8, both the forward trajectory loss and the past trajectory loss have contributions to overall performance. However, the forward trajectory loss plays a relatively more important role than the backward trajectory loss does. For example, when $\lambda = 0.1$, it indicates the weight for the forward trajectory loss is 0.1, while the weight for the backward trajectory loss is 0.9, the ADE for UNIV dataset with $\lambda = 0.1$ is greater than it with $\lambda = 0.9$.

We also perform ablation experiments to evaluate the influence of the parameter ϵ in Eqn. 19 which is the magnitude of reciprocal attack. As we discussed above, reciprocal attack minor refines the predicted forward trajectory to match the ground truth better. As we can see in Fig. 9, ϵ within a certain range has slight influence on the overall performance.

G. Qualitative Results

Fig. 10 shows successful and failure examples of our predicted trajectories. Following prior work *S-GAN* [16], we show the best predicted trajectory among 20 model outputs in the figure. We can see that our proposed method is able to correctly predict the future path. According to the background scene, we can see that our method can ensure that each human path follows the physical constrains of the scene, such as walking around obstacles, *e.g.* trees, and staying on sidewalks. Our method also shows the decent prediction results with human-human interactions. When persons walk in a crowded road, they can avoid each other when they merge from various directions and then walk towards a common direction.

The last row in Fig. 10 shows some failure cases which have relatively large error rates. For example, the person in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3076078, IEEE Transactions on Circuits and Systems for Video Technology

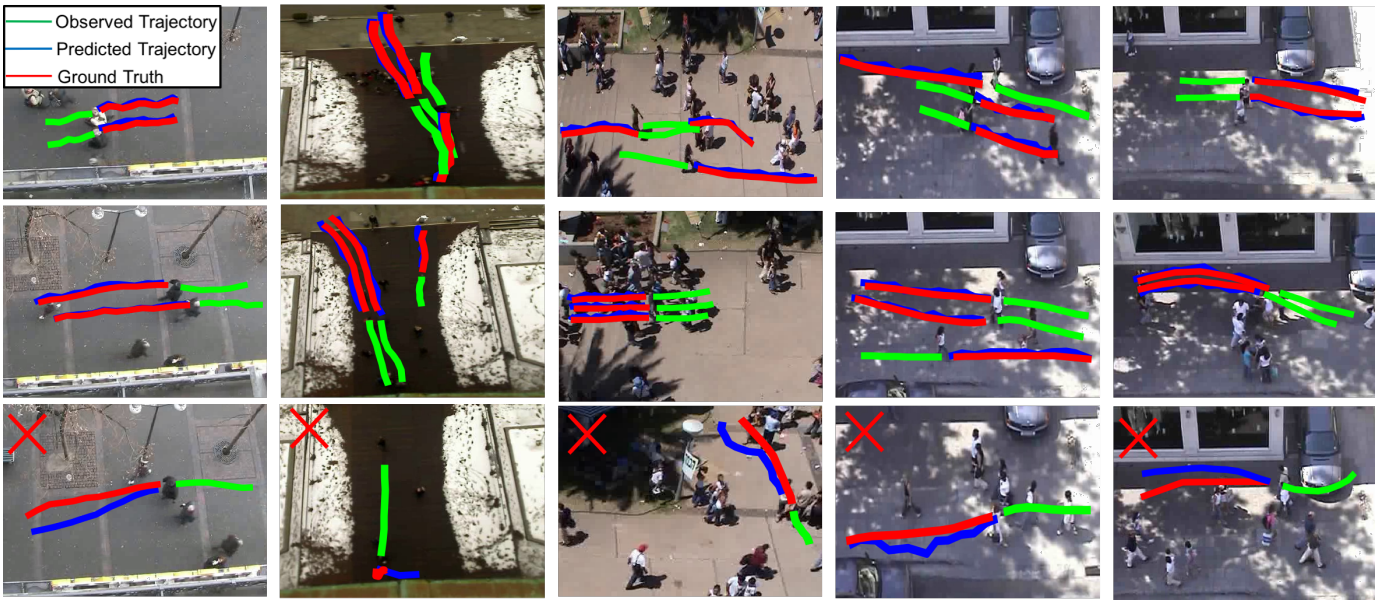


Fig. 10. Illustration of our method predicting future 12 time steps trajectories, given previous 8 time steps. The results are drawn under HOTEL, ETH, UNIV and ZARA1 and ZARA2 datasets from 1st column to 5th column, respectively.

TABLE V
THE QUANTITATIVE RESULTS ON ETH (COLUMN 3 AND 4) AND UCY (COLUMN 5-7) DATASETS ON THE TASK OF BACKWARD PREDICTION (PREDICTING THE TRAJECTORIES OF PREVIOUS 8 TIME STEPS, GIVEN THE TRAJECTORIES OF 12 FUTURE TIME STEPS).

Metric	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Avg
ADE	S-GAN [16]	0.57	0.27	0.39	0.22	0.24	0.34
	S-GAN-P [16]	0.56	0.31	0.37	0.24	0.27	0.35
	Ours	**0.50**	**0.22**	**0.31**	**0.20**	**0.18**	**0.28**
FDE	S-GAN [16]	1.05	0.68	0.74	0.42	0.43	0.67
	S-GAN-P [16]	1.07	0.72	0.71	0.43	0.49	0.68
	Ours	**0.95**	**0.44**	**0.65**	**0.40**	**0.37**	**0.56**

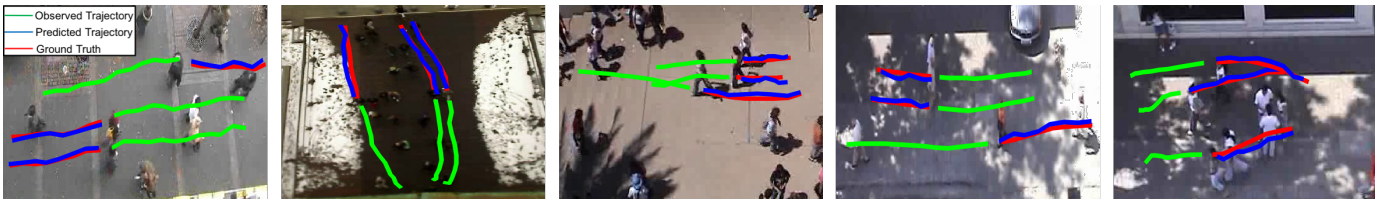


Fig. 11. Illustration of backward prediction (predicting previous 8 time steps trajectories, given future 12 time steps ones). The results are drawn under HOTEL, ETH, UNIV and ZARA1 and ZARA2 datasets from 1st column to 5th column respectively. Note that, we crop and resize the original image for better visualization.

the scene slows down or even stops for a while, or walks directly over the obstacles instead of walking around them. In most cases, our method still can predict the plausible path, even though the predicted path is not exactly the same as the ground truth. For example, for the first, third, and fifth cases in the last row, in our prediction, the person walks around another person or the tree in the road, which is quite reasonable in practice.

We do notice that our method could not beat all other state-of-the-art methods on the ETH and HOTEL datasets. We visualize more cases on these two datasets that our predicted future trajectories are not close enough to the ground truth in Fig. 12. In the first row, our method may recognize the snow and the small square around the tress as obstacles. Hence, in the predicted path, it looks like that the person walks around these obstacles, while the pedestrians actually walk over these areas. In the second row, we show some cases that the observed pedestrians take a straight path to the destination, while our predicted paths make a few direction changes due to the enforcement of human-human and human-scene constraints. But in these two cases, our method obtains a better prediction of the final destination. In the third row, we show some strange cases where the pedestrians walk directly to the nearby pedestrians and then take a sharp detour or stopping

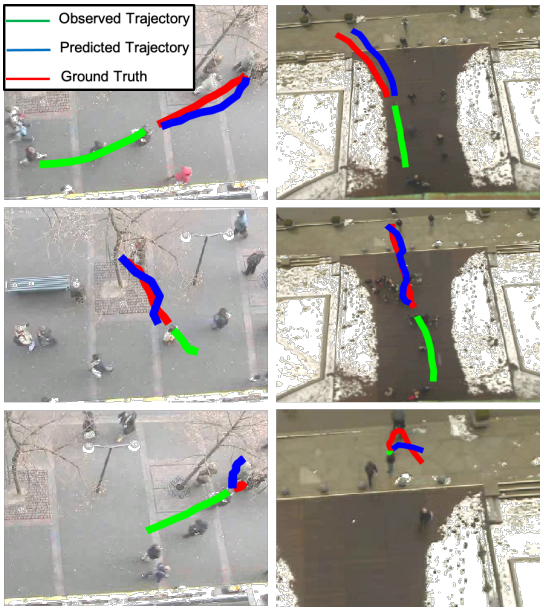


Fig. 12. Qualitative examples of some failure cases on the ETH and HOTEL datasets which have relatively large error rates. Note that, we crop and resize the original image for better visualization.

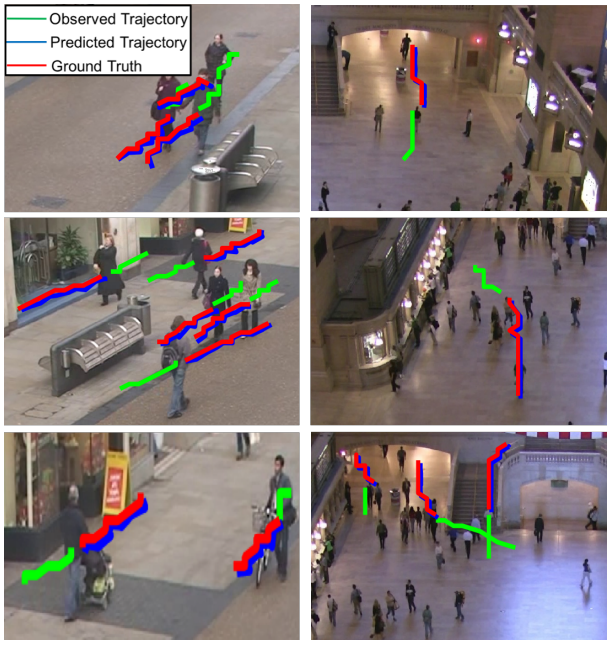


Fig. 13. Qualitative examples of our method predicting future 12 time steps trajectories, given previous 8 time steps ones on Town Centre (1st column) and Grand Central Station (2nd column) dataset. Note that, we crop and resize the original image for better visualization.

for a while. Our method could not work well in these cases and made some inaccurate but reasonable predictions. To address this issue, in our future work, we can develop a multi-level context-driven interaction model which considers both global scene and local context around the pedestrian.

H. Generalization: Evaluations on Town Centre and Grand Central Station Datasets

To further evaluate the generalizability of our method, we perform experiments on new datasets: Town Centre [90] and Grand Central Station [93]. Following the previous work [94], for each of these two datasets, we combine the training data from ETH and UCY datasets and 50% data from this dataset for training, the remaining data is used for testing. The objective is to predict trajectories in the next 12 and 16 time steps based on the trajectories of 8 previous time steps. The comparison results of our method with *S-GAN* [16] and *Scene-LSTM* [94] are shown in Table IV. We can clearly see that our method outperforms the existing methods in both datasets. Some qualitative examples on Town Centre and Grand Central Station datasets are presented in Fig. 13.

I. Backward Prediction Evaluation

We also conduct experiments of backward trajectory prediction (predict past trajectories by giving future trajectories) on the ETH and UCY datasets. We compare the ADE and FDE results with the *S-GAN* and *S-GAN-P* methods. The objective is to predict trajectories of the previous 8 time steps based on the trajectories of 12 future time steps. The prediction error results are shown in Table V. We can see that, our reciprocal learning method outperforms *S-GAN* and *S-GAN-P* on both ETH and UCY datasets. The results show that our reciprocal learning is able to accurately perform both forward and backward prediction of human trajectories. Several visual examples of our backward prediction on the ETH and UCY datasets are shown in Fig. 11.

VI. Conclusion and Future Work

In this paper, we have explored the unique characteristics of human trajectories and developed a new approach, reciprocal network learning, for human trajectory prediction. Two networks, the forward and backward prediction networks, are tightly coupled together, satisfying the reciprocal constraint, which allows them to be jointly learned for accurate and robust human trajectory prediction. Based on this constraint, we borrowed the concept of adversarial attacks of deep neural networks, which iteratively modifies the input of the network to match the given or forced network output, and developed a new method for network testing, called *reciprocal attack for matched prediction*. It has further improved the prediction accuracy slightly. Extensive experimental results have demonstrated our approach achieves the state-of-art performance on public benchmark datasets. In our future work, we plan to explore multi-level context-driven interaction model which considers both global scene and local context around the pedestrian. We will also extend the proposed approach to a closely related research problem: predicting the human future trajectory from an egocentric view of a moving vehicle with the on-board camera.

Acknowledgements

This work was supported in part by National Science Foundation under grants 1647213, 1646065 and 1646162.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. An preliminary version of this work was presented at the IEEE Conference on Computer Vision and Pattern Recognition 2020.

REFERENCES

[1] S. Srikanth, J. A. Ansari, S. Sharma *et al.*, "Infer: Intermediate representations for future prediction," *arXiv preprint arXiv:1903.10641*, 2019.

[2] A. D. Berenguer, M. Alioscha-Perez, M. C. Oveneke, and H. Sahli, "Context-aware human trajectories prediction via latent variational model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[3] K. Chen, X. Song, and X. Ren, "Pedestrian trajectory prediction in heterogeneous traffic using pose keypointsbased convolutional encoder-decoder network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[4] X.-T. Truong and T. D. Ngo, "Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 4, pp. 1743–1760, 2017.

[5] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and X. Li, "A survey of human action analysis in hri applications," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[6] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, "A spatio-temporal crf for human interaction understanding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1647–1660, 2016.

[7] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 744–760, 2015.

[8] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 464–469.

[9] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

[10] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *European Conference on Computer Vision*. Springer, 2016, pp. 697–713.

[11] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.

[12] K. M. Kitani, B. D. Ziebart, and J. Andrew, "Bagnell, and martial hebert. activity forecasting," in *European Conference on Computer Vision. Springer*, vol. 59, 2012, p. 88.

[13] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[14] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1941–1946.

[15] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[16] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.

[17] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1186–1194.

[18] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *CVPR 2011*. IEEE, 2011, pp. 3441–3448.

[19] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, "Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6067–6076.

[20] P. Coscia, F. Castaldo, F. A. Palmieri, L. Ballan, A. Alahi, and S. Savarese, "Point-based path prediction from polar histograms," in *2016 19th International Conference on Information Fusion (FUSION)*. IEEE, 2016, pp. 1961–1967.

[21] K. Kim, D. Lee, and I. Essa, "Gaussian process regression flow for analysis of motion trajectories," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1164–1171.

[22] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on image processing*, vol. 16, no. 4, pp. 1168–1181, 2007.

[23] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011.

[24] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, 2001.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] P. Pan, F. Porikli, and D. Schonfeld, "Recurrent tracking using multifold consistency," in *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

[28] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2756–2759.

[29] I. K. Sethi and R. Jain, "Finding trajectories of feature points in a monocular image sequence," *IEEE Transactions on pattern analysis and machine intelligence*, no. 1, pp. 56–73, 1987.

[30] H. Wu, A. C. Sankaranarayanan, and R. Chellappa, "In situ evaluation of tracking algorithms using time reversed chains," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[31] N. Bisagno, B. Zhang, and N. Conci, "Group lstm: Group trajectory prediction in crowded scenarios," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[32] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[33] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "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*, 2019, pp. 1349–1358.

[34] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann, "The garden of forking paths: Towards multi-future trajectory prediction," *arXiv preprint arXiv:1912.06445*, 2019.

[35] C. Tang and R. R. Salakhutdinov, "Multiple futures prediction," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 398–15 408.

[36] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5725–5734.

[37] P. Coscia, F. Castaldo, F. A. Palmieri, A. Alahi, S. Savarese, and L. Ballan, "Long-term path prediction in urban scenarios using circular distributions," *Image and Vision Computing*, vol. 69, pp. 81–91, 2018.

[38] H. Manh and G. Alaghband, "Scene-lstm: A model for human trajectory prediction," *arXiv preprint arXiv:1808.04018*, 2018.

[39] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "Car-net: Clairvoyant attentive recurrent network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 151–167.

[40] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4194–4202.

[41] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6262–6271.

[42] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[43] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[44] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[45] L. Yu, J. Yu, and Q. Ling, "Bltrcnn based 3d articulatory movement prediction: Learning articulatory synchronicity from both text and audio inputs," *IEEE Transactions on Multimedia*, 2018.

[46] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[47] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2018.

[48] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5523–5531.

[49] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[51] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[52] Y. Yan, B. Ni, W. Zhang, J. Xu, and X. Yang, "Structure-constrained motion sequence generation," *IEEE Transactions on Multimedia*, 2018.

[53] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[54] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[55] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[56] X. Shu, J. Tang, G. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[57] H. Hu and G.-J. Qi, "State-frequency memory recurrent neural networks," in *International Conference on Machine Learning*, 2017, pp. 1568–1577.

[58] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[59] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[60] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Advances in Neural Information Processing Systems*, 2016, pp. 5040–5048.

[61] L. Chen, L. Wu, Z. Hu, and M. Wang, "Quality-aware unpaired image-to-image translation," *IEEE Transactions on Multimedia*, 2019.

[62] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[63] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Transactions on Multimedia*, 2019.

[64] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[65] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.

[66] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915.

[67] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *European conference on computer vision*. Springer, 2010, pp. 438–451.

[68] R. W. Brislin, "Back-translation for cross-cultural research," *Journal of cross-cultural psychology*, vol. 1, no. 3, pp. 185–216, 1970.

[69] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Advances in Neural Information Processing Systems*, 2016, pp. 820–828.

[70] M. Twain, *The jumping frog: in English, then in French, then clawed back into a civilized language once more by patient, unremunerated toil*. Courier Corporation, 1971.

[71] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1426–1433.

[72] G. Pang, X. Wang, J.-F. Hu, Q. Zhang, and W.-S. Zheng, "Dbdnet: learning bi-directional dynamics for early action prediction," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 897–903.

[73] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," in *Computer Graphics Forum*, vol. 32, no. 5. Wiley Online Library, 2013, pp. 177–186.

[74] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3d-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 117–126.

[75] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros, "Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1191–1200.

[76] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[77] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[78] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[79] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[80] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[81] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2376–2386, 2019.

[82] M. Shim, H.-I. Ho, J. Kim, and D. Wee, "Read: Reciprocal attention discriminator for image-to-video re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 335–350.

[83] F. Yang, Z. Wang, J. Xiao, and S. Satoh, "Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval." in *AAAI*, 2020, pp. 12 589–12 596.

[84] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[85] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[86] S. S. Talathi and A. Vartak, "Improving performance of recurrent neural network with relu nonlinearity," *arXiv preprint arXiv:1511.03771*, 2015.

[87] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on artificial neural networks*. Springer, 2010, pp. 92–101.

[88] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[89] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3828–3838.

[90] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*. IEEE, 2011, pp. 3457–3464.

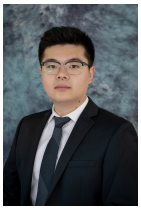
[91] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *European conference on computer vision*. Springer, 2010, pp. 452–465.

[92] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3542–3549.

[93] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2871–2878.

[94] M. Huynh and G. Alaghband, "Trajectory prediction by coupling scene-lstm with human movement lstm," in *International Symposium on Visual Computing*. Springer, 2019, pp. 244–259.

[95] 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*. IEEE, 2009, pp. 261–268.



Hao Sun received the B.S. degree in electronic engineering from Nanjing University of Science and Technology, Nanjing, China, in 2013. He received the M.S. degree in computer engineering in 2016, and the Ph.D. degree in electrical engineering in 2020 from University of Missouri, Columbia, MO, USA. His current research interests include computer vision and machine learning.

Zhiqun Zhao received the B.S. degree in communication engineering from University of Electronic Science and Technology of China, UESTC, Chengdu, China, in 2013. He is Currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering University of Missouri, Columbia, MO, USA. His current research interests include deep convolutional neural networks, low-rank approximation, adversarial examples and zero-shot learning.å

Zhaozheng Yin is a SUNY Empire Innovation Associate Professor at Stony Brook University. He is affiliated with the AI Institute, Department of Biomedical Informatics, and Department of Computer Science. His group has been working on Biomedical Image Analysis, Computer Vision, and Machine Learning. He is an IEEE senior member.

Zhihai He (Fellow, IEEE) received the B.S. degree in mathematics from Beijing Normal University, Beijing, China, in 1994, the M.S. degree in mathematics from the Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, in 1997, and the Ph.D. degree in electrical engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2001. In 2001, he joined Sarnoff Corporation, Princeton, NJ, USA, as a Member of Technical Staff. In 2003, he joined the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA, where he is currently a Robert Lee Tatum Distinguished Professor. His current research interests include multimedia networking, wireless sensor networks, computer vision, and machine learning. He is also a member of the Visual Signal Processing and Communication Technical Committee of the IEEE Circuits and Systems Society. He serves as a technical program committee member or a session chair of a number of international conferences. He was a recipient of the 2002 IEEE Transactions on Circuits and Systems for Video Technology Best Paper Award and the SPIE VCIP Young Investigator Award in 2004. He was the Co-Chair of the 2007 International Symposium on Multimedia over Wireless in Hawaii. He has served as an Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), the IEEE Transactions on Multimedia (TMM), and the Journal of Visual Communication and Image Representation. He was also the Guest Editor for the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) Special Issue on Video Surveillance.