IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTEGRATIVE COMPUTER VISION AND MULTIMEDIA ANALYTICS

Received December 28, 2019, accepted January 16, 2020, date of publication January 28, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969994

Pedestrian Motion Trajectory Prediction With Stereo-Based 3D Deep Pose Estimation and Trajectory Learning

JIANQI ZHONG[1], HAO SUN[2], WENMING CAO[1], AND ZHIHAI HE[2], (Fellow, IEEE)
[1]School of Information Engineering, Shenzhen University, Shenzhen 518060, China
[2]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

Corresponding author: Zhihai He (hezhi@missouri.edu)

This work was supported in part by the NSF CyberSEES under Grant 1539389.

ABSTRACT Existing methods for pedestrian motion trajectory prediction are learning and predicting the trajectories in the 2D image space. In this work, we observe that it is much more efficient to learn and predict pedestrian trajectories in the 3D space since the human motion occurs in the 3D physical world and and their behavior patterns are better represented in the 3D space. To this end, we use a stereo camera system to detect and track the human pose with deep neural networks. During pose estimation, these twin deep neural networks satisfy the stereo consistence constraint. We adapt the existing SocialGAN method to perform pedestrian motion trajectory prediction from the 2D to the 3D space. Our extensive experimental results demonstrate that our proposed method significantly improves the pedestrian trajectory prediction performance, outperforming existing state-of-the-art methods.

INDEX TERMS Trajectory prediction, deep learning, pose estimation, stereo vision.

I. INTRODUCTION

Human motion behaviors and trajectories are driven by human behavioral reasoning, common sense rules, social conventions, and interactions with others and the surrounding environment. Human can effectively predict short-term body motion of others and respond accordingly. The ability for a machine to learn these rules and use them to understand and predict human motion in complex environments is highly valuable with a wide range of applications in social robots, intelligent systems, and smart environments [1], [2]. In human trajectory prediction, the central task is: with motion trajectories of pedestrians observed during the past period of time, can we predict their future trajectories within the future short period of time, for example, 10 seconds?

Predicting human motion is a very challenging task [3]. An efficient algorithm for human trajectory prediction needs to model physical constraints of the environment on human motion and anticipate movements of other persons or vehicles and their social behaviors. Recently, a number of methods based on deep neural networks have been developed for human trajectory prediction [3], [4]. Earlier methods have been focused on learning dynamic patterns of moving agents (human and vehicles) [3] and modeling the semantics of the navigation environment [5]. Recent approaches focus more on interactions between all agents in the scene in order to predict the future trajectory for each agent. Methods have been developed to model human-human interactions [6], understand social acceptability using data-driven techniques based on Recurrent Neural Networks (RNNs) [4], [7], [8], and model the joint influence of all agents in the scene [9].

In this work, we observe that it is much more efficient to learn and predict pedestrian trajectories in the 3D space than the 2D image space, since the human motion occurs in the 3D physical world and and their behavior patterns are better represented in the 3D space. Therefore, its natural behavior and motion patterns are better represented by its 3D trajectory, instead of the 2D image coordinates. 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. Based on this observation, we propose to extend the existing deep learning-based pose estimation and trajectory prediction from the 2D image space to the 3D space. Specifically, we construct and calibrate a stereo camera system to capture pedestrian videos. We use a twin-network satisfying stereo consistency constraint to estimate the human pose in both video sequences simultaneously and reconstruct the trajectory in the 3D space.

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

We extend the SocialGAN method [9] for trajectory prediction from 2D to 3D. Our experimentally results demonstrate that, by estimating, learning, and predicting the human pose and trajectories in the 3D space instead of the 2D image space, our method is able to significantly reduce the trajectory prediction error by up to 71% with an average error reduction of 47%.

The rest of the paper is organized as follows. Section II reviews related work on human trajectory prediction. The proposed 3D stereo human trajectory learning and prediction method is presented in Section III. Section IV presents the experimental results, performance comparisons, and ablation studies. Section V summarizes our major contributions and concludes the paper.

II. RELATED WORK

In this section, we review related work on trajectory prediction and pose estimation.

A. HUMAN TRAJECTORY PREDICTION

A number of methods have been developed in the literature to predict human trajectories in dynamic scenes. Helbing and Molnar [6] introduced the Social Force Model to characterize social interactions among people in crowded scenes using coupled Langevin equations. Social pooling [8] was introduced to share features and hidden representations between different moving agents. Reference [9] 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. [10] predicted human trajectories using a spatio-temporal graph to model both position evolution and interactions between pedestrians.

In human trajectory prediction, it is also very important to model the effects of physical environments. For example, people tend to walk along the sidewalk, around a tree or other physical obstacles. Sadeghian *et al.* [11] incorporated scene context to human trajectory prediction based on GAN (Generative Adversarial Network). Reference [12] 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. Reference [3] proposed a Bayesian framework to predict unobserved paths from previously observed motions and to transfer learned motion patterns to new scenes. Scene-LSTM [13] divided the static scene into grids and predicted pedestrian's location using LSTM. The CAR-Net method [14] integrated past observations with bird's eye view images and analyzed them using a two-levels attention mechanism.

B. HUMAN POSE ESTIMATION USING DEEP NEURAL NETWORKS

Our work is als related to human pose estimation. The task of human pose estimation is to determine the precise pixel locations of body keypoints from a single image. Since the work of *DeepPose* [15], human pose estimation has recently achieved significant progress with deep convolutional neural networks. Human pose estimation is often formulated as a regression problem, predicting locations of body joints from deep neural network features [15]. DeepPose uses a deep neural network (DNN) to directly regress the coordinates of body joints. Tompson *et al.* [16] argued that it is more efficient to use DNNs to regress heatmap images at multiple scales. While body models are not necessary for effective part localization, constraints between parts allow us to assemble independent detection results into an accurate body configuration. Detection-based methods are relying on powerful DNNS to detect body parts and then combine them into a human pose using a graphical model [17]–[19].

III. 3D STEREO HUMAN TRAJECTORY LEARNING AND PREDICTION

In this section, we present our stereo human pose tracking, trajectory learning, and prediction in the 3D space.

A. METHOD OVERVIEW

The objective of our work is to learn and predict the motion trajectories of pedestrians in the scene for the next period of time, say 10 seconds. We propose to perform the trajectory learning and prediction in the 3D space, instead of the conventional 2D image space used in all existing methods. As illustrated in Figure 1, we construct and calibrate a stereo camera system to observe the pedestrian scene. Each camera is capturing a live video, which will be analyzed by our deep neural network module for 2D pose estimation. We enforce stereo consistence between these two networks to improve the 3D pose estimation accuracy. With camera calibration, we reconstruct the 3D trajectories for all pedestrians in the scene. We then adapt the socialGAN network from 2D to 3D to predict the pedestrian trajectories for the next period of time (e.g. 10 seconds) based on the observed trajectories of the past time period (e.g., 8 seconds).

The standard formulation of trajectory prediction problem in the literature [10], [12] is in the 2D image space. 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. In this work, we extend the trajectory prediction from 2D to 3D. Specifically, let $\mathbf{X} = X_1, X_2, \cdots, X_N$ be the trajectories of all pedestrian 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 $X_n = (x_n^t, y_n^t, z_n^t)$ for time steps $t = 1, 2, \cdots, T_o$. The ground truth of future trajectory is given by $Y_n = (x_n^t, y_n^t, z_n^t)$ for time step $t = T_o + 1, \cdots, T_p$.

B. TWIN DEEP NEURAL NETWORKS WITH STEREO CONSTRAINT FOR 3D HUMAN POSE ESTIMATION

The first component of our proposed method is human body pose estimation and tracking. Existing regression-based pose

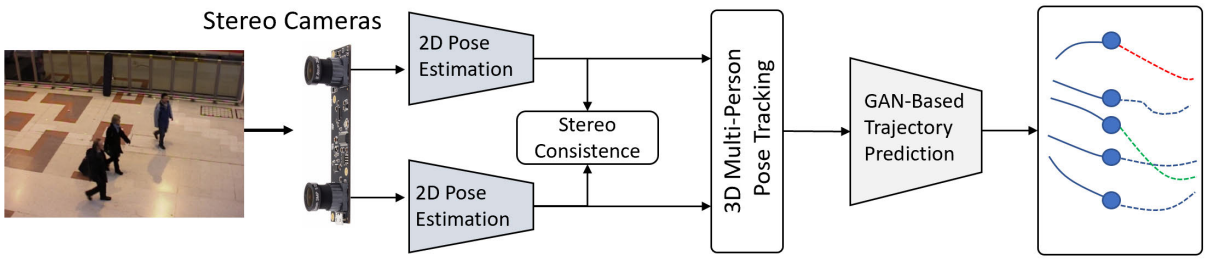


FIGURE 1. The overall framework of our proposed stereo 3D pedestrian pose tracking and trajectory prediction.

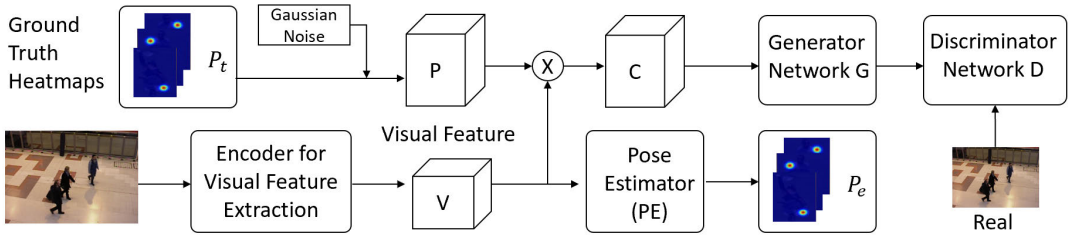


FIGURE 2. The overall framework of our proposed poseGAN for the joint human pose estimation and conditional image synthesis.

estimation methods work well with visible limbs. In this work, we are dealing with scenes with multiple pedestrians which often have significant body occlusion. To handle the partially occluded body joints and limbs, existing methods try to learn a body configuration model to infer their locations. In our experiments, we recognize that they cannot efficiently handle fully occluded limbs, which occur quite often in practical scenarios, especially when the person is moving around. Ideally, we wish that the training data contains images of all different poses of human body, including samples with fully occluded limbs. In this way, the deep neural network can be carefully designed and trained to predict the body joints of these fully occluded limbs. However, this is a nearly impossible task in practice since persons with free-style motion will have a wide variety of body poses being occluded by different objects, especially in highly cluttered environments. In the training data, some typical body poses are dominating while difficult cases are very rare. This pose a significant challenge for learning highly efficient human pose estimation.

1) GAN-BASED POSE ESTIMATION

To address this issue, we propose to incorporate generative adversary training into human pose estimation and train the following three networks jointly: the human pose estimator, the semantic data generator, and the semantic data discriminator. Specifically, the generative (G) network augments the training data, enforces the pose estimator to estimate more precisely of the joints. The discriminative (D) network evaluates the conditional pair, i.e., the training image and the estimated pose heatmaps, to enforce the G network to generate semantically-similar human pose images. An overview of the human pose estimation system is illustrated in Fig. 2. An RGB image is fed into the feature extraction module (FE) for visual feature extraction, and then fed into the pose estimation module to infer pose heatmaps. The visual features will be concatenated with the estimated heatmaps and the ground truth heatmaps to form an overall conditioned feature vector C, where the estimated heatmaps serve as the conditioning augmentation. The conditioned features C will be fed into the Discriminator and Generator networks to generate synthetic pose images. The discriminator is a matching-aware discriminator. The positive pair is the real image paired with the ground truth pose P_t concatenated with the extracted visual features V. The negative pairs have two groups. The first group is the real image paired with the **estimated pose** P_e concatenated with V. The second group is the synthetic image paired with P_t and V.

Generative Adversarial Networks (GAN) [20] consist of two models that are trained in an alternative manner. The generator G is optimized to reproduce the true data distribution p_{data} by generating data that are hard for the discriminator D to differentiate them from real data. Meanwhile, D is optimized to distinguish real data from synthetic ones generated by G. The overall training procedure is similar to a two-player min-max game with the following objective function,

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))], \quad (1)$$

where x is an instance from the true data distribution p_{data}, and z is a noise vector sampled from distribution p_z. Conditional GAN [21] is an extension of GAN where both the generator $G(z, c)$ and discriminator $D(x, c)$ receive additional conditioning variables c, thus allowing G to generate images conditioned on variables c.

The task of the Generator Network is to infer a RGB image from a stack of body part heatmaps with latent feature vectors, aiming to produce synthetic images with the same human pose but not necessarily with the same visual elements such as texture or color. It is complementary to geometric data augmentation that includes translation, rotation and scaling. The purpose is to train the semantic-aware generator that implicitly help the training of person-centric perception networks. Our hypothesis is that a generative network from adversarial training distills information from the distribution of training data, and is capable of reconstructing data with semantic meaning from the distilled visual elements. The task of the discriminator network is to distinguish the real and fake pairs of images and heatmaps. With adversary training, the generator, the pose estimator, and the discriminator will be learned jointly.

2) NETWORK TRAINING

Inspired by the work of [22], we use a modified version of the hourglass network [23] as our human pose estimator. The hourglass design is a state-of-the-art architecture for bottom-up and top-down inference with residual blocks. It processes input images at multiple scales with down-scaling and up-scaling. In this work, We find out that, by replacing the residual block in the hourglass network with an inception-residual [24] block, we can achieve improved accuracy in estimating occluded poses. The overall structure is shown in Fig. 3.

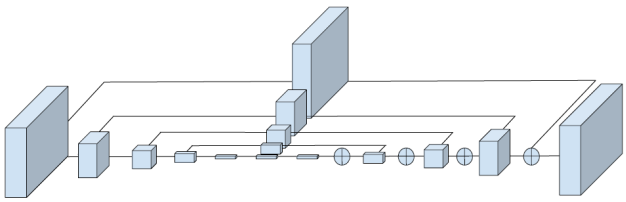


FIGURE 3. An illustration of hourglass design. Pixel-wise addition fuses the information from two branches while keeping the input and output resolution uniform. The illustration gives an example of a 4-level hourglass.

Let I be the input RGB image and P_t be the ground truth heatmap. We use P to denote the pose estimation network. $P(I)$ represents the predicted heatmap of body joints from the input image I and is denoted by P_e The *Generator* and *Discriminator* networks are denoted by G and D, respectively. $\otimes$ denotes the concatenation operation of tensors. The overall loss is the sum of GAN loss, the pose estimator loss, and L_2 loss:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{Pose}} + \lambda\mathcal{L}_2,$$

where the GAN loss is the sum of the generator loss and discriminator loss:

$$\mathcal{L}_{\text{GAN}} = \mathcal{L}_{\text{G}} + \mathcal{L}_{\text{D}}. \tag{2}$$

The pose estimator loss is the standard MSE (mean squared error) loss for heatmap regression:

$$\mathcal{L}_{\text{Pose}} = \|P(I) - P_t\|_2 + \|P(G(I, P_t)) - P_t\|_2. \tag{3}$$

In this work, we use a stereo camera system to observe the pedestrian trajectory. The human pose estimation results obtained from these two networks should satisfy the stereo consistency constraint. Specifically, let $P(I_A)$ and $P(I_B)$ be the pose estimation results from cameras A and B, respectively. let $\mathbf{M}_{A\to B}[\cdot]$ be the stereo mapping from A to B. The stereo consistency error is then given by

$$C_p = ||P(I_B) - \mathbf{M}_{A\to B}[P(I_A)]||_2 + ||P(I_A) - \mathbf{M}_{B\to A}[P(I_B)]||_2 \tag{4}$$

Then, the new pose estimator loss is given by $\mathcal{L}_{\text{Pose}}(P; I, P_t) + \alpha \cdot C_p$, where α is a weighting parameter whose default value is set to be 0.5 in our experiments.

C. CONSTRUCTING THE FORWARD AND BACKWARD PREDICTION NETWORKS

Both networks for the left and right video sequences the same network structure and weights. As illustrated in Figure 4, we adopt the existing Social-GAN in [9] 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 EXTRACTOR

Specifically, we first use the LSTM encoder to capture the temporal pattern and dependency within each trajectory of human n and encode them into a high-dimensional feature $\mathbf{F}_h^t(n)$. 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 [9] to build a social pooling module which extracts the joint social feature $\mathbf{F}_s^t(n)$ of all human in the scene to encode the human-human interactions. The relative distance values 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 to form the joint feature. 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 at the current time.

As recognized in [11], [25], 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 [11], we use VGGNet-19 [26] pre-trained on the ImageNet [26] to extract the visual feature of background scene I^t, which is then fed into an LSTM encoder to compute the hidden state tensor $\mathbf{F}_v^t$.

2) LSTM-BASED GAN FOR TRAJECTORY PREDICTION

Inspired by previous work [9], [11], in this paper we use an LSTM based Generative Adversarial Network (GAN) module to generate human's future path as illustrated in Figure 4. The generator is constructed by a decoder LSTM. Similar to the conditional GAN [2], a white noise vector $\mathbf{Z}$ is sampled

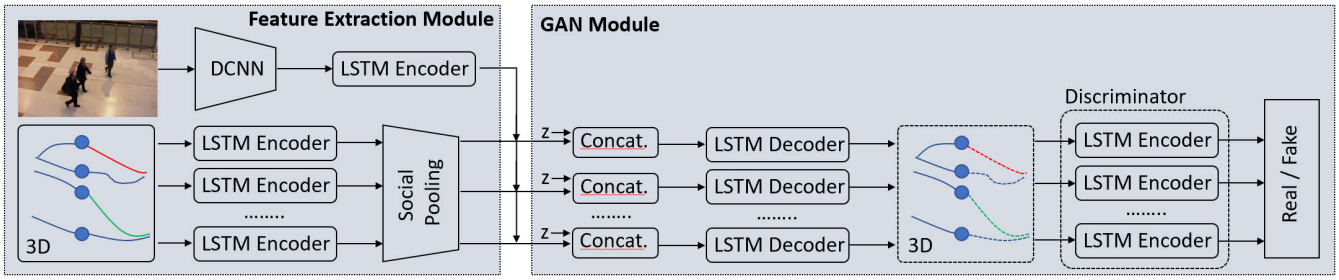


FIGURE 4. Overview of our prediction model. Our model consists of two key components: (1) Feature Extraction Module, (2) LSTM-based GAN module.

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 $\mathbf{Z}$. We take this as the input to the LSTM decoder to generate the candidate future paths for each human. The discriminator is built with an LSTM encoder which takes the input as randomly chosen trajectory from either ground truth or predicted trajectories 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.

Within our 3D method for human trajectory prediction, let $G^\theta : \mathbf{X} \rightarrow \mathbf{Y}$ be the generator of the prediction network $\mathbf{F}_\theta$. D^θ is the discriminator for $\mathbf{F}_\theta$. Its input $\mathbf{Y}'$ is randomly selected from either ground truth $\mathbf{Y}$ or the predicted future trajectory $\hat{\mathbf{Y}}$. To train the prediction network $\mathbf{F}_\theta$, we combine the adversarial loss with the trajectory prediction loss $J[\theta]$

$$\mathcal{L}_\theta = L^\theta_{GAN} + J[\theta], \tag{5}$$

where the trajectory prediction loss is defined as

$$J[\theta] = ||\mathbf{Y} - \mathbf{F}_\theta(\mathbf{X})||_2, \tag{6}$$

and the adversarial loss L^θ_{GAN} is defined as:

$$L^\theta_{GAN} = \min_G \max_D \quad \mathbb{E}_{\mathbf{Y}'\sim p(\mathbf{Y},\mathbf{\Phi})}[\log D(\mathbf{Y}')] + \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X}),\mathbf{Z}\sim p(\mathbf{Z})}[\log(1-D(G(\mathbf{X},\mathbf{Z})))]. \tag{7}$$

IV. EXPERIMENTAL RESULTS

A. 3D PEDESTRIAN TRAJECTORY DATASET

Our work is one of the first efforts to study the learning and prediction of pedestrian trajectory in the 3D space. To evaluate the performance of our algorithm and support future research on this topic, we establish a dataset of stereo videos with 3D pedestrian trajectories and will make it publicly available. This dataset contains 5 long stereo videos, each having a time duration of about 30 minutes and a resolution of 1920×1080 at 30 frames per second. Each stereo video has two synchronized video sequences captured by a stereo GoPro camera system constructed and calibrated by our self. Assisted by human pose detection and tracking tools, we manually examine the correctness of human pose detection and tracking results and correct errors. Based on the stereo camera calibration results, we combine the results from two stereo cameras and compute the human trajectory in the 3D coordinate system.

TABLE 1. Comparing the ADE between our method and the SocialGAN method on our test dataset.

Method	T1	T2	T3	T4	T5
SocialGAN [9]	118.96	147.29	144.34	158.46	142.40
Our Method	101.56	47.25	92.40	86.04	83.93
Error Reduction	-14.6%	-67.9%	-35.9%	-45.7%	-41.1%

B. EVALUATION METRICS AND PROTOCOL

We use the same error metrics in [8], [27] for performance evaluation. (1) Average Displacement Error (ADE) is the average L_2 distance between the ground truth 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:

$$\text{ADE} = \frac{\sum_{n\in\Psi}\sum_{t=T_o+1}^{T_p}\sqrt{((\hat{x}_n^t,\hat{y}_n^t)-(x_n^t,y_n^t))^2}}{|\Psi|\cdot T_p}, \tag{8}$$

$$\text{FDE} = \frac{\sum_{n\in\Psi}\sqrt{((\hat{x}_n^{T_p},\hat{y}_n^{T_p})-(x_n^{T_p},y_n^{T_p}))^2}}{|\Psi|}, \tag{9}$$

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 previous papers [8], [9], [11], we use the similar leave-one-out evaluation methodology. 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). All location coordinates are normalized to 0 to 1 for training and testing.

C. PERFORMANCE EVALUATIONS AND COMPARISON

Fig. 5 shows examples of body poses estimated by the proposed poseGAN method. With these accurately detected body joints in each frame of the two video sequences captured by the stereo camera system, we are able to construct the

FIGURE 5. Examples of human pose detection results.

TABLE 2. Comparing the FDE between our method and the SocialGAN method on our test dataset.

Method	T1	T2	T3	T4	T5
SocialGAN [9]	171.53	231.25	204.34	228.71	209.62
Our Method	125.26	65.56	113.13	125.72	127.26
Error Reduction	-26.9%	-71.6%	-44.6%	-45.0%	-39.3%

FIGURE 6. Illustration of our method predicting future 12 time steps trajectories, given previous 8 time steps ones.

3D trajectory of each body joint and track the person across frames. Following the evaluation protocol outlined in the above section, we train the 3D stereo trajectory prediction network, measure its performance using the ADE and FDE error metrics, and compare its performance with the stateo-of-the-art method SocialGAN [9]. The SocialGAN method is only able to learn and predict the trajectory from the image in the 2D coordinate system. We apply the method to both sequences, compute the prediction error for each sequence, and report the smaller one for performance comparison. Table 1 summarizes the ADE prediction error comparison with our method and SocialGAN. Table 1 summarizes the FDE prediction error comparison. The third rows show the percentage of error reduction. We can see that, by estimating, learning, and predicting the human trajectory in the 3D domain, our new method is able to significantly reduce the prediction error by up to 71.6% with an average of 47%. Figure 6 shows examples of our trajectory prediction with the observed trajectory shown in green, the predicted trajectory shown in blue, and the ground truth shown in red. We can see that our algorithm is able to accurately predict the motion trajectory of the pedestrians.

V. CONCLUSION

In this work, we have extended the pedestrian trajectory learning and prediction from the 2D image space into the 3D physical space. To this end, we constructed and calibrated a stereo camera system. We developed a twin poseGAN network with stereo consistence constraint to detect human pose and construct their trajectories in the 3D space. We extended the SocialGAN from 2D into the 3D and demonstrated that our new method is able to significantly improve the trajectory prediction accuracy, reducing the prediction error by an average of 47%. The proposed system and method have important applications in advanced video surveillance and intelligent human-computer interaction applications.

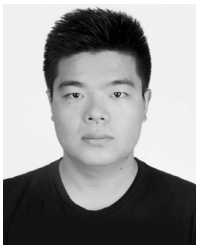
REFERENCES

[1] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 464–469.

[2] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.

[3] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *Proc. Eur. Conf. Comput. Vis.*, pp. 697–713, Springer, 2016.
[4] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 336–345.
[5] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. Eur. Conf. Comput. Vis.*, vol. 59. Berlin, Germany: Springer, 2012, p. 88.
[6] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 5, pp. 4282–4286, Jul. 2002.
[7] F. Bartoli, G. Lisanti, L. Ballan, and A. D. Bimbo, "Context-aware trajectory prediction," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1941–1946.
[8] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.
[9] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.
[10] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–7.
[11] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.
[12] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5725–5734.
[13] H. Manh and G. Alaghband, "Scene-LSTM: A model for human trajectory prediction," 2018, *arXiv:1808.04018*. [Online]. Available: https://arxiv.org/abs/1808.04018
[14] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "CAR-Net: Clairvoyant attentive recurrent network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 151–167.
[15] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.
[16] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. NIPS*, 2014, pp. 1799–1807.
[17] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. NIPS*, 2014, pp. 1736–1744.
[18] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.
[19] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784
[22] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3332–3341.
[23] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.
[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: https://arxiv.org/abs/1602.07261
[25] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1186–1194.
[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[27] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

JIANQI ZHONG received the B.S. degree in electronic science and technology from the Hunan Institute of Engineering, China, in 2017. He is currently pursuing the M.S. degree in integrated circuit engineering with Shenzhen University, Shenzhen, China. His research interests include pattern recognition and physiological function assessment.



HAO SUN received the B.E. degree in electronic engineering from the Nanjing University of Science and Technology, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA. His research interests include computer vision and machine learning.

WENMING CAO received the M.S. degree from the System Science Institute, China Academy Sciences, Beijing, China, in 1991, and the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 2003. From 2005 to 2007, he was a Postdoctoral Researcher with the Institute of Semiconductors, Chinese Academy of Sciences. He is currently a Professor with Shenzhen University, Shenzhen, China. He has authored or coauthored over 80 publications in top-tier conferences and journals. His research interests include pattern recognition, image processing, and visual tracking.

ZHIHAI HE (Fellow, IEEE) received the B.S. degree in mathematics from Beijing Normal University, Beijing, China, in 1994, and 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, 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, where he is currently a tenured Full Professor. His current research interests include image/video processing and compression, wireless communication, computer vision, and sensor networks. He is a member of the Visual Signal Processing and Communication Technical Committee, IEEE Circuits and Systems Society. He serves as a member of the Technical Program Committee or the Session Chair of a number of international conferences. He has received the Best Paper Award of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, in 2002, 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 was also a Guest Editor of the IEEE TCSVT Special Issue on Video Surveillance. He has served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and the *Journal of Visual Communication and Image Representation*.