Introvert: Human Trajectory Prediction via Conditional 3D Attention

Nasim Shafiee
Northeastern University
shafiee.n@northeastern.edu

Taskin Padir
Northeastern University
t.padir@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

Abstract

Predicting human trajectories is an important component of autonomous moving platforms, such as social robots and self-driving cars. Human trajectories are affected by both the physical features of the environment and social interactions with other humans. Despite recent surge of studies on human path prediction, most works focus on static scene information, therefore, cannot leverage the rich dynamic visual information of the scene. In this work, we propose Introvert, a model which predicts human path based on his/her observed trajectory and the dynamic scene context, captured via a conditional 3D visual attention mechanism working on the input video. Introvert infers both environment constraints and social interactions through observing the dynamic scene instead of communicating with other humans, hence, its computational cost is independent of how crowded the surrounding of a target human is. In addition, to focus on relevant interactions and constraints for each human, Introvert conditions its 3D attention model on the observed trajectory of the target human to extract and focus on relevant spatio-temporal primitives. Our experiments on five publicly available datasets show that the Introvert improves the prediction errors of the state of the art.

1. Introduction

Predicting future trajectories of humans in dynamic environments, such as streets, airports, shopping malls and sports fields, is an important task in computer vision with applications in autonomous driving, human-robot interaction, urban safety and advertising, among others [50, 11, 48, 21, 15]. Forecasting human motions, however, is an extremely difficult problem, due to physical, social and mental factors that collectively influence people's trajectories. In particular, as we move in an environment, we avoid physical constraints and obstacles, follow landmarks, yield right-of-way to nearby people, follow social norms and change our trajectory based on changes in the environment. This has motivated a large body of works in recent years that aim to model and incorporate various influencing factors for human trajectory prediction [1, 39, 23, 3, 13].

Prior Works and Challenges. Earlier works [14, 9, 10, 24, 25, 30, 36, 4, 49, 46, 51, 42, 53] have designed energy functions to model human-human interactions, also referred to as "social forces". Despite their relative success, such methods require careful feature and energy function design, which often could capture only simple interactions but not complex interactions in crowded environments. To mitigate these limitations, more recent methods have proposed data-driven approaches by leveraging advances in deep neural networks. In particular, sequence prediction methods based on recurrent neural networks (RNNs) model each person's trajectory by an RNN, whose latent states capture human motion, followed by social pooling that allows recurrent models of nearby trajectories to share their states [1, 13]. However, they cannot capture the influence of farther people to a target trajectory, while giving the nearby trajectories the same importance weights. To overcome these limitations, attention-based models have been integrated with RNNs [39, 3] and spatio-temporal graphs [41, 40, 33, 20] to weigh different trajectories by adjusting the importance of neighbors to each target human. However, most approaches discussed above rely on only kinematics data, which contains information about only moving agents in the scene.

Given that videos contains rich information about physical configuration of the scene and navigation constraints, several works have tried to use the visual context of the scene in conjunction with kinematics data for more effective predictions. This has been achieved by concatenating the states of all RNNs with visual features of a current frame extracted via CNNs [40, 26, 41], which could be followed by an attention model to select relevant features [39, 40]. However, existing works face multiple challenges. First, current methods extract visual information that is often shared and identical for all people moving in the environment. However, in practice, each person's trajectory depends on the region of the terrain where he/she is moving, physical constraints between the current position and the intended destination, as well as other humans relevant to the path. In other words, *different parts of the scene and visual features have different importance that depends on the target human*. Second, visual features obtained by *encoding one frame at*

a time cannot capture the complex interactions and social norms, which is why existing methods require to incorporate social interactions through pooling states of RNNs operating on kinematics data.

More importantly, from a computational stand point, *during inference time, one needs to first run a human detection and tracking algorithm for all people in the scene and then connect RNNs using nearest neighbors graphs or attention* to be able to predict the trajectory of a target human. This prohibits existing methods from being run in real-time at the inference time, especially in crowded environments with many humans, but one or a few targets of interests.

Paper Contributions. In this paper, we develop an efficient framework for human trajectory prediction using a conditional 3D visual attention mechanism, which addresses the aforementioned challenges. We argue that the video itself (not an individual frame) contains all necessary information about the motions and interactions of humans as well as dynamic constraints, e.g., moving vehicles, and static constraints, e.g., buildings and sidewalks, of the environment. This can be seen from the fact that kinematic trajectories are extracted from videos, hence, cannot contain more information than the video itself. Thus, instead of modeling human-human interactions by connecting nearby or all recurrent models of human trajectories in the scene, we leverage the video to extract 3D visual interaction information (2 spatial and 1 temporal dimensions). This removes the need for running a detection and tracking algorithm for every human in the scene, hence, increases the efficiency at test time, where only the video and tracking of the target human would be needed.

We develop a sequence to sequence method that consists of two parallel encoding streams, which gather 3D visual and kinematics information relevant to a target human, and one decoding stream that predicts the future trajectory of the target human. To focus on relevant social interactions and physical constraints for each human, our visual encoder uses a conditional 3D attention mechanism that receives the input video and conditioning on the observed trajectory of the target human, extracts spatio-temporal primitives and learns to attend to most informative primitives. These extracted primitives could be e.g., parts of a sidewalk, few vehicles, distant landmarks as well as nearby or distant humans in the scene. By experiment on UCY [27] and ETH [35] datasets, we show that out method significantly improves the state of the art performance, reducing the average prediction error on 5 datasets from 0.41 to 0.34.

2. Related Works

Existing works on predicting human trajectories can be mainly divided into two categories: human-space interactions and human-human interactions. While the first group focuses on learning physical features of the environment, which influence trajectories of humans, the second group investigates the influence of humans on each other's paths.

Human-Space Interaction. Physical scene information, such as crosswalks and roads, has been exploited to address the task of human trajectory prediction. To infer feasible paths, [22] has proposed to leverage hidden Markov decision processes. On the other hand, [26] has employed static scene context to rank and refine the possible trajectories generated by an RNN-CVAE based framework. [40] extracts static scene information through a double attention mechanism to predict the future path of a target pedestrian. Also, [39] and [41] extract static scene information by considering the effect of neighboring pedestrians and multimodal output configuration through, respectively, attentive GAN and Info-VAE frameworks. Our work is similar to [40] in the sense that we also use a dual attention framework, however, our method extracts both static and dynamic scene features and it takes into account the interaction of a target human to other humans in the scene.

Human-Human Interaction. Research on predicting pedestrian behavior considers the interactions between pedestrians either as a crowd or as an individual. Social-force and its variants [14, 32, 2, 52, 38, 36] are the pioneer models that assist a pedestrian to go toward his goal while avoiding collisions. The main drawback of these methods is using handcrafted kinetic forces and energy potentials, which cannot capture complex interactions in crowded environments and cannot leverage data-driven approaches.

To predict the future trajectory of a human, recent works use data-driven models, particularly, deep neural networks, to encode the trajectory information and interactions between individuals. These interactions have been incorporated either through a pooling module [1, 13] or an attention module [39, 3]. Another trend to capture social interactions has been using graph representations, where the nodes and edges correspond to humans and their interactions [54, 23, 12, 55, 33, 20, 41]. On the other hand, our work focuses on capturing social interactions through dynamic 3D scene information.

Sequence Prediction using RNNs. Recurrent Neural Networks have been widely used for sequence generation in diverse range of natural language processing and computer vision applications. Recent studies on human trajectory prediction mostly employ RNNs to encode and decode kinematic trajectory information [1, 56, 16, 43, 45, 44, 6, 31, 34, 28, 29, 8, 37, 17, 47, 5]. However, as observed in [13, 39], RNNs cannot capture the spatio-temporal interactions among humans in the scene. One way to overcome this problem is augmenting RNNs by a pooling or an attention module to capture spatio-temporal interactions. Our work tackles this problem by leveraging dynamic scene features via a conditional 3D visual encoder based on attention [19, 18], which captures complex interactions.

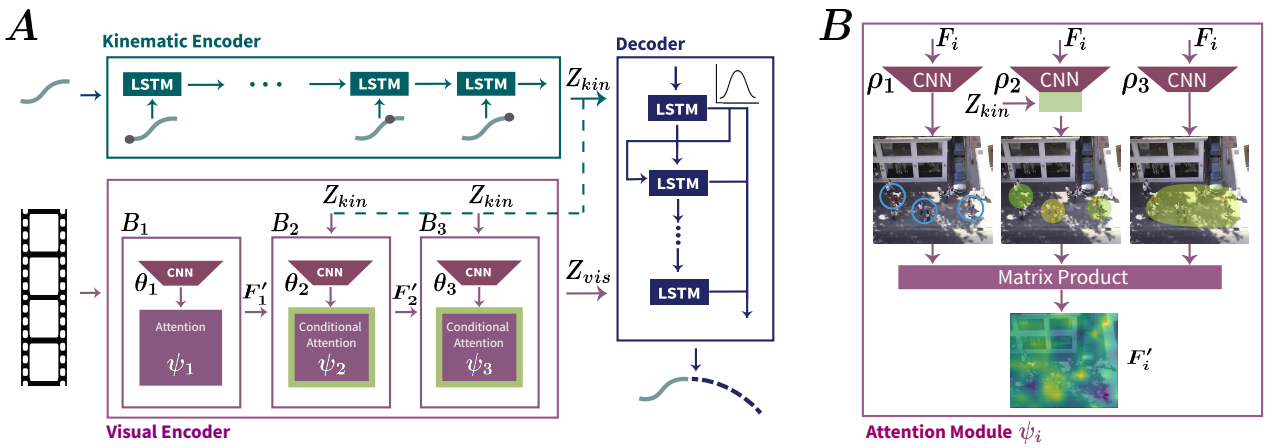

Figure 1. (A) Introvert is a sequence to sequence model, which consists of i) kinematic encoder, ii) 3D visual encoder using conditional 3D attention, iii) trajectory decoder. (B) Structure of the conditional 3D attention module.

3. Trajectory Prediction via Conditional 3D Attention

In this section, we develop a sequence to sequence framework for human trajectory prediction that leverages video data directly to infer human-dependent interactions using a conditional 3D attention mechanism.

3.1. Problem Settings

Trajectory prediction is the problem of estimating the positions of humans in the future, given their previous positions and the visual information of the scene. Assume we have multiple training videos, each containing several human trajectories in t_f frames. Similar to prior works, we assume that each training video is preprocessed by a human detection and tracking algorithm to obtain the spatial coordinates of each person across the t_f video frames (during testing, our method only requires the trajectory of the target human). We denote the 2D position of human p at frame t by $\boldsymbol{u}_t^{(p)} = (x_t^{(p)}, y_t^{(p)}) \in \mathbb{R}^2$. Assume we observe trajectories and the scene from frame 1 to t_o and the goal is to predict the trajectories in frames $t_o + 1$ to t_f.

For a person p, we denote the sequence of observed and future positions, respectively, by

$$\mathcal{T}_o^{(p)} = (\boldsymbol{u}_1^{(p)}, \dots, \boldsymbol{u}_{t_o}^{(p)}), \quad \mathcal{T}_f^{(p)} = (\boldsymbol{u}_{t_o+1}^{(p)}, \dots, \boldsymbol{u}_{t_f}^{(p)}). \quad (1)$$

We also denote the sequence of observed frames by $\mathcal{V}_o = (I_1, \dots, I_{t_o})$, which correspond to top-view or angle-view video frames of the scene.

3.2. Overview of Proposed Framework

To address the problem of human trajectory prediction, we develop a new sequence to sequence model using an encoder-decoder architecture. Our model consists of two parallel encoders: a kinematic and a visual encoder, see Figure 1. The kinematic encoder receives the observed trajectory information, $\mathcal{T}_o^{(p)}$, and produces a latent kinematic trajectory $\mathcal{Z}_{kin}^{(p)}$, which encodes the information of the observed positions. The visual encoder, on the other hand, receives the observed frames $\mathcal{V}_o$ and extracts conditional spatio-temporal context, $\mathcal{Z}_{vis}^{(p)}$, for each person, which captures the necessary physical constraints and social interactions required for predicting the future trajectories. To extract the spatio-temporal context $\mathcal{Z}_{vis}^{(p)}$, we use a 3D dual attention mechanism, consisting of i) multiple spatial attention modules that learn to extract and focus on global descriptors of the video, such as humans, crosswalks, cars and alleys; ii) descriptor attentions that finds the importance of each descriptor for each pixel in a frame. Given that the salient visual information used for moving in the environment for each human is different than others, we condition the dual attention mechanism on the latent kinematic trajectory of the person, $\mathcal{Z}_{kin}^{(p)}$, to capture human-specific visual encodings. The decoder receives the encoded information from kinematic and visual encoders and decodes them to the distribution of future trajectories of the target, $\mathcal{T}_f^{(p)}$.

Notice that, unlike prior works, the kinematics encoder of different humans in our framework do not interact. Instead, the interaction is captured through the visual stream by operating on the observed video as a whole, instead of processing each frame individually, and by conditioning the visual encoder on each person's observed trajectory. This allows our method to inherently capture the kinematic information of relevant scene elements and to have the flexibility of paying attention to physical constraints and humans that could be far.

Next, we discuss each component of our framework in details and then present our learning and inference strategy. For simplicity of notation, we drop the superscript p from variables, as it would be clear from the context.

3.3. Kinematic Encoder

To obtain suitable representations of trajectories, the kinematic encoder, $\mathrm{E}_k(\cdot)$, receives the observed trajectory of a target human as input of the form $\mathcal{T}_o^\delta = (\boldsymbol{u}_1, \boldsymbol{u}_2 - \boldsymbol{u}_1, \boldsymbol{u}_3 - \boldsymbol{u}_2, \ldots)$,

which consists of the coordinates of the start position and the relative displacements of the human between consecutive frames. We choose this format as it enables the model to better capture similarities between almost identical trajectories that may have different starting points. We transform each input vector using a fully-connected network, Φ, and pass it to a recurrent network (an LSTM) to capture dependencies between different coordinates of an observed trajectory. We denote sequence of outputs from the LSTM units by $\mathcal{Z}_{kin}$, which captures the latent kinematic trajectory.

3.4. Conditional 3D Visual Encoder

As discussed before, the observed video $\mathcal{V}_o$ contain information about physical and social constraints of all humans in the scene. Thus, we use a visual encoder, $\mathrm{E}_v(\cdot)$, to extract tailored visual information for each human in the scene, which we denote by $\mathcal{Z}_{vis}$. Our encoder consists of three consecutive conditional visual feature extraction and attention blocks $\{B_i\}_{i=1}^3$ that learn to extract increasingly complex and high-level features. Every block B_i is composed of a 3D CNN layer (denoted by Θ_i) followed by a conditional dual attention network (denoted by Ψ_i). While each 3D CNN extracts spatio-temporal information from the video, the conditional dual attention network focuses on relevant spatio-temporal regions in the video to each human by using his/her latent kinematic trajectory information, $\mathcal{Z}_{kin}$. In other words, the input to the visual encoder has 3 dimensions (2 spatial dimensions + 1 temporal dimension), hence, it processes the video through 3D CNNs and generate 3D attentions (2 spatial + 1 temporal dimension) for each video input.

Conditional Dual Attention Network. Let $\mathcal{F}_i$ denote the output features of the 3D CNN in the i-th visual feature extraction block, B_i. We employ the double attention architecture proposed in [7] and modify it with kinematic conditioning for three layers of 3D CNN. The conditional dual attention network in each block i performs a two-step operation on $\mathcal{F}_i$ to produce its output, $\mathcal{F}_i'$. The first step extracts global video descriptors conditioned on the kinematic information of the person, which we denote by $g(\mathcal{F}_i|\mathcal{Z}_{kin}, \boldsymbol{u}_1)$. These global descriptors would correspond to the scene elements, such as subsets of pedestrians, landmarks, obstacles that are relevant to the trajectory of the target human. The second step, on the other hand, finds the relevance of each of these global descriptors to each pixel in each frame.

More specifically, the conditional dual attention network in each block B_i consists of three three convolution layers, $\{\rho_j\}_{j=1}^3$, with filter size equal to one. The first layer, ρ_1, refines the input $\mathcal{F}_i$ and expands the number of its channels to m. The second layer, ρ_2, learns n spatial attention modules conditioned on $\mathcal{Z}_{kin}$ to build n global visual primitives, each of size m, from the scene. Finally, ρ_3 corresponds to an attention vector on the usage of the global descriptors for each pixel at each frame. We can write this as,

$$\mathcal{F}_i' = \rho_1(\mathcal{F}_i)\, g(\mathcal{F}_i|\mathcal{Z}_{kin}, \boldsymbol{u}_1)^\top \sigma\big(\rho_3(\mathcal{F}_i)\big), \qquad (2)$$

σ denotes the softmax operation and $g(\mathcal{F}_i|\mathcal{Z}_{kin}, \boldsymbol{u}_1)$ denotes the global video descriptors conditioned on the kinematic information. We build g using the second later ρ_2 as

$$g(\mathcal{F}_i|\mathcal{Z}_{kin}, \boldsymbol{u}_1) = \sigma\big(\rho_2(\mathcal{F}_i)\big) \odot \sigma(\mu([\mathcal{Z}_{kin}, \boldsymbol{u}_1])), \qquad (3)$$

where μ is a fully-connected layer and $\odot$ denotes the Hadamard (entry-wise) product between the output of μ and each of n global attention map generated by ρ_2. We build the conditional visual feature vector, $\mathcal{Z}_{vis}$ by passing $\mathcal{F}_3'$, which is the output of the last block, to a fully-connected layer. We will use $\mathcal{Z}_{vis}$ in the decoder module to predict the future trajectory of the target human.

3.5. Trajectory Decoder

After encoding the the kinematics and visual information, we feed the fusion tensor $\begin{pmatrix} \mathcal{Z}_{kin} & \mathcal{Z}_{kin} \times \mathcal{Z}_{vis} \\ 1 & \mathcal{Z}_{vis} \end{pmatrix}$ to a maxpool layer followed by a linear layer to capture information from visual and kinematic streams for the decoder.

Next, the result is fed as a hidden vector to an LSTM in the decoder module. The output of each LSTM unit representing a future time instant, $t > t_o$, is then connected to a MLP, consisting of two fully-connected layers, that outputs a multivariate Gaussian distribution for the displacement

$$\delta\boldsymbol{u}_t \triangleq \boldsymbol{u}_t - \boldsymbol{u}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t),\ \boldsymbol{\Sigma}_t = \begin{pmatrix} \sigma_t^x & 0 \\ 0 & \sigma_t^y \end{pmatrix}, \qquad (4)$$

where the two coordinates are. assumed to be independent. Notice that predicting displacements instead of absolute positions, allows our model to better decode identical or similar trajectories with different start points.

Our method outputs trajectories in a stochastic mode. More specifically, we sample C sequences $\big(\delta\boldsymbol{u}_{t_o+1}, \ldots, \delta\boldsymbol{u}_{t_f}\big)$ from the learned Gaussian distributions to obtain K plausible trajectories that the target human may take in future.

The uncertainty in the predicted coordinates of each sampled trajectory comes from the accumulation of the uncertainty of the prediction in a specific time-step and its previous time-steps. These uncertainties allow the method to handle multi-modal nature of human trajectories, where often there exist multiple plausible paths.

3.6. Training Strategy

We train our network in an end-to-end fashion using the following loss function,

$$\mathcal{L} \triangleq \mathcal{L}_{mse} + \lambda \, \mathcal{L}_{reg} \tag{5}$$

where $\mathcal{L}_{mse}$ denotes the mean squared error and $\mathcal{L}_{reg}$ is a regularization term to predict consistent future trajectories with respect to the observed ones. In particular, the regularization is defined as the sum of Euclidean distances between each step of the predicted trajectory, $\mathcal{T}_f$, and a line fitted to the observed trajectory, $\mathcal{T}_o$.

We calculate $\mathcal{L}_{mse}$ by first sampling C future trajectories, then picking the top N trajectories closest to the ground-truth and finally computing the average of the mean squared error between these N trajectories and the ground-truth (in the experiments, we set $C = 20$ and $N = 5$). We empirically observed that this strategy allows our network to converge faster while having more accurate predictions.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our proposed method on two publicly available datasets of UCY [27] and ETH [35]. The ETH dataset has two scenes, where each scene has 750 pedestrians and is split into two sets (ETH and Hotel). The UCY dataset also has two scenes with 786 pedestrians and is split into three sets (ZARA1, ZARA2 and UCY). These datasets consist of videos from a static bird-eye view camera and kinematic trajectories of pedestrians and contain diverse types of pedestrians activities such as individual and group walking, crossing, group forming and scattering. We evaluate our method on all 5 sets of data.

Evaluation Metrics. Similar to existing works [1, 3, 13, 39, 41, 54], we use the following error metrics:

- Average Displacement Error (ADE): Average L2 distance between ground-truth and predicted trajectories over $[t_o + 1, t_f]$.
- Final Displacement Error (FDE): The distance between the ground-truth and predicted position at t_f.

Evaluation Method. We follow a similar evaluation method as in prior works [1, 13, 39]. We use a leave-one-out approach, where we train on 4 sets of data and test on the remaining set. We observe each training trajectory for 8 times-steps (3.2 seconds) and measure the prediction errors for 8 (3.2 seconds) and 12 (4.8 seconds) time-steps.

Baselines. We compare our method against several state of the art. 1) Social-LSTM, [1], which employs two LSTMs for encoding and decoding and social pooling to capture the effect of neighboring pedestrians on a target pedestrian's trajectory. 2) Trajectron++ [41], which uses a spatio-temporal graph input to encode the motion of agents based on their type $\in \{car, pedestrian\}$ using LSTM. To generate the predicted trajectory, Trajectron++ gathers all the encoded information from the graph data and visual information and decodes them using LSTM. 3) STAR [54], which employs one temporal and one spatial transformer to learn the crowd interactions.

4) Social GAN [13], which employs generative adversarial networks and pools social information from neighboring pedestrians. 5) SoPhie [39], which takes advantage of attentive GAN networks and uses both kinematics and static visual inputs. 6) Social ways [3], which employs info-GAN and attention pooling to generate multi-modal trajectories based on capturing information from both a target pedestrian and neighbors. 7) PECNET [31], which first predicts the end point of the future trajectory and using the endpoint generates the path. 8) BiGAT [23], which takes advantage of a Bicycle-GAN framework and graph representation to model social interactions.

Kinematic Encoder	
embedding	$2 \rightarrow 64$
3D Visual Encoder	
Θ_1	$K = [3,3,3]\,, S = [3,3,3]$
Θ_2	$2 \times K = [3,3,3]\,, S = [1,3,3]$
Θ_1	$K = [3,3,3]\,, S = [1,3,3]$
Ψ_1, Ψ_2, Ψ_3	$m = 16, n = 8$
Ψ_3(linear)	$1563 \rightarrow 256$
ρ_1, ρ_2, ρ_3	$K = [1,1,1]\,, S = [1,1,1]$
μ(linear)	$258 \rightarrow 16 \rightarrow 16$
Trajectory Decoder	
embedding	$2 \rightarrow 64$
decoder(linear)	$2 \times 256 \rightarrow 256 \rightarrow 4$

Table 1. Architecture details of our proposed sequence to sequence model, which consists of a kinematic encoder, 3D conditional visual encoder and a trajectory decoder.

Implementation Details. Table 1 shows the details of our proposed deep architecture (K denotes the kernel size and S denotes the stride size). Each sub-module element is augmented by ReLU activation function. Similar to other works, we employ two Vanilla LSTMs as the encoder and decoder with hidden size 256. We embed each of $\mathcal{T}_o$ and $\mathcal{T}_f$ through two linear layers and then pass it to the encoder and the decoder, respectively. Also, the output of the decoder is passed to an embedding layer, which generates mean and variance of a Gaussian distribution in each time-step.

We train the entire network in an end-to-end fashion using stochastic gradient descent optimizer using our proposed loss function in (5) with $\lambda = 0.5$. For a faster convergence, we initially applied the teacher force strategy to the 70% of batches and decreased the percentage linearly to 0% during the training. As stated before, during training, we sample the output trajectory 20 times (i.e. $C = 20$) and

ADE/FDE	$t_f - t_o$	University	Zara 1	Zara 2	Hotel	ETH	AVG
Social LSTM*	12	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.79 / 1.76	1.09 / 2.35	0.72 / 1.54
Social GAN	12	0.60 / 1.26	0.34 / 0.69	0.42 / 0.84	0.72 / 1.61	0.81 / 1.52	0.58 / 1.18
SoPhie	12	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.76 / 1.67	0.70 / 1.43	0.54 / 1.15
BiGAT	12	0.55 / 1.32	0.30 / 0.62	0.36 / 0.75	0.49 / 1.01	0.69 / 1.29	0.48 / 1.00
Social Ways	12	0.55 / 1.31	0.44 / 0.64	0.51 / 0.92	0.39 / 0.66	0.39 / **0.64**	0.46 / 0.83
PECNet	12	0.35 / 0.60	0.22 / 0.39	0.17 / 0.30	0.18 / 0.24	0.54 / 0.87	0.29 / 0.48
Star	12	0.31 / 0.62	0.26 / 0.55	0.22 / 0.46	0.17 / 0.36	**0.36** / 0.65	0.26 / 0.53
Trajectron++	12	0.22 / 0.43	0.17 / 0.32	**0.12 / 0.25**	0.12 / 0.19	0.43 / 0.86	**0.21** / 0.41
Introvert (ours)	12	**0.20 / 0.32**	**0.16 / 0.27**	0.16 / **0.25**	**0.11 / 0.17**	0.42 / 0.70	0.21 / **0.34**
Social LSTM*	8	0.41 / 0.84	0.27 / 0.56	0.33 / 0.70	0.49 / 1.01	0.73 / 1.48	0.45 / 0.91
Social GAN	8	0.36 / 0.75	0.21 / 0.42	0.27 / 0.54	0.48 / 0.95	0.61 / 1.22	0.39 / 0.78
Introvert (ours)	8	**0.16 / 0.24**	**0.12 / 0.19**	**0.14 / 0.19**	**0.09 / 0.12**	**0.32 / 0.49**	**0.17 / 0.25**

Table 2. The average/final displacement error (ADE/FDE) of all methods across all datasets. Models with * have deterministic outputs. Stochastic models sample 20 trajectories and report the best result. All models receive $t_o = 8$ observed time-steps and predict positions for $t_f - t_o = 12/8$ future time-steps.

use 5 samples with the lowest loss value to train the model.

Our proposed model is implemented in Pytorch on a server running Ubuntu 18.04 with an Intel Xeon Gold CPU and four NVIDIA Quadro RTX 6000 GPUs. Similar to all existing works, we used leave-one-out strategy for training and testing and trained our model over 200 epochs.

4.2. Experimental Results

4.2.1 Quantitative Analysis

Table 2 shows the average FDE and ADE results of different methods for both $t_f - t_o \in \{8, 12\}$ on all five datasets. From the results, we make the following conclusions:

– On the FDE metric, our method significantly improves the state of the art, where it outperforms existing algorithms on 4 out of 5 datasets. In particular, our method achieves 0.34 FDE average error over all the datasets compared to 0.41 obtained by the second best method (Trajectron++). As expected, the error for 12 time-step ahead is always larger than 8 time-step prediction, due to higher uncertainties and more drastic changes to trajectories.

– On the ADE metric, our method outperforms existing algorithms on 3 out of 5 datasets and similar to Trajectron++ achieves the lowest average ADE error of 0.21 over the datasets. Notice that ADE is in general an easier metric than FDE as the immediate future predictions are often close to position at t_o.

– Notice that most methods, including ours, have larger displacement errors on the University and ETH datasets. The larger errors on University is due to the higher crowd density in the dataset than others. In other words, as it involves more human-human interactions, it causes more challenging prediction of the future trajectory. In addition, the high density of the crowd forces a target pedestrian to choose between different options such as overtaking or following other pedestrians, making predictions more uncertain.

Notice, however, that our method achieves the lowest FDE (0.32 by ours compared to 0.43 by Trajectron++) and ADE (0.20 by ours compared to 0.22 by Trajectron++), showing the effectiveness of Introvert in capturing the multi-modal nature of future trajectories.

Also, the larger errors on ETH are due to lower frequency of the video frames and kinematic data compared to other datasets. Given that the trajectories would be for longer time periods and models need to predict farther in the future, the performances on ETH are generally lower than on other datasets.

– In the Hotel dataset, there is a large number of pedestrians who are waiting for train with limited motions. Therefore, most methods, including ours, obtain relatively small displacement errors due to successfully predicting small motions of the pedestrians. Our method also obtains the lowest FDE and ADE errors on this dataset.

– The fact that for the majority of datasets our method obtains lowest displacement errors, shows the effectiveness of our proposed framework in capturing human-human interactions using our 3D conditional visual encoder, without using kinematic data of neighboring pedestrians in the scene.

Fast Inference. One advantage of our method over existing works is being computationally efficient during inference (real-time). This comes from the fact that by using visual information directly, we do not need to run a human detection and tracking algorithm for all people in the scene and then connect RNNs using nearest neighbors graphs or attention, which are costly. To better demonstrate this, Table 3 shows the average inference time of Social Ways, Social GAN and our method (Introvert) on the datasets, where our method achieves 0.12 second inference time compared to 0.42 sec. and 0.82 sec. by others.

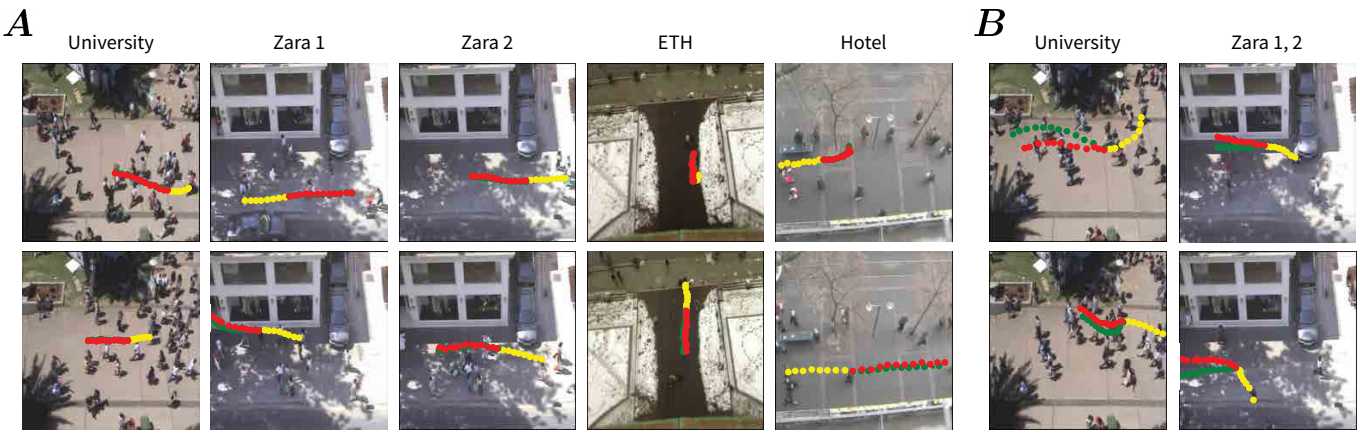

Figure 2. Qualitative examples of trajectory prediction by our method. Yellow, Red and Green dots correspond to observations, predictions and ground-truth, respectively. For each dataset in each column, we show trajectories of two pedestrians in two rows. The first 5 columns (A) demonstrate examples of successful predictions by our method, while the last two columns (B) show inaccurate predictions.

Method	Social Ways	Social GAN	**Introvert (ours)**
Time	0.817s	0.419s	**0.120s**

Table 3. Comparison of inference times.

4.2.2 Qualitative Analysis

Figure 2 shows qualitative results of our method for trajectory prediction on several videos from UCY and ETH datasets. The two plots in each column show two different pedestrians (two walking scenarios) in the same dataset. In all cases, we show the first observed frame, denote the observed trajectory by yellow dots and the ground-truth and predictions by our method by green and red, respectively. The part A of Figure 2 shows 10 different success examples from the 5 datasets, where our method is able to accurately predict the future positions of pedestrians in the scene. For example, the top example from the University demonstrates a scenario of human-human interaction, where the target pedestrian slows down before reaching to a group of standing people, bypasses them from the left side and then speeds up. Notice that in such a crowded scene, our method is able to well capture interactions and predict the future positions. Also, the bottom example in Zara 1 demonstrate a success example of capturing human-space interaction in which our model accurately predicts that the target pedestrian will go through the door of the store in the left side of the scene. Also, for the top example in Hotel, our method correctly predicts that the target human who is entering the scene will avoid a tree and will turn left.

There are also scenarios in which our model could not predict the future positions accurately. The part B of Figure 2 shows four examples of such failures. The shared characteristic of these scenarios is the sudden change in the trajectory of the target human. Notice that even in these cases, our model is able to capture some general properties of the behavior of the target pedestrian, such as the walking orientation and velocity, but it does not provide an accurate prediction of the future positions. Although the predicted trajectory does not completely match the ground-truth, our method provides a feasible trajectory, which avoids moving and stationary obstacles in the scene. We believe this inaccuracy arises from the diverse multi-modal nature of future pedestrian paths, and is an avenue of further investigation.

Figure 3 shows the qualitative analysis of Introvert (ours), Social Ways [3], STAR [54] and SRLSTM [56], demonstrating that our method can more accurately predict future trajectories.

Figure 3. Trajectory prediction comparison between our method (Introvert) and other algorithms.

Visualization of Conditional Attentions. Next, we demonstrate the effectiveness of our conditional spatio-temporal attention model for successfully predicting the future trajectory of a target pedestrian. Figure 4 shows visualization of the conditional attention, where for each module (Ψ_2 and Ψ_3) we first calculate the element-wise product of dual attention matrices, then average them over the n spatial attention maps and plot the results. The figure includes three sample videos (corresponding to the three rows in the plot) from three different datasets. For each sample video in a row, we show two pedestrians (pedestrian 1: corresponding to the first three columns, pedestrian 2: corresponding to the second three columns) walking at the same time, so they share the same video input. We plot the original frame and our prediction in the first column, while the attentions generated by Ψ_2 and Ψ_3 modules have been shown in the second and third columns, respectively.

As Figure 4 shows, each conditional attention module

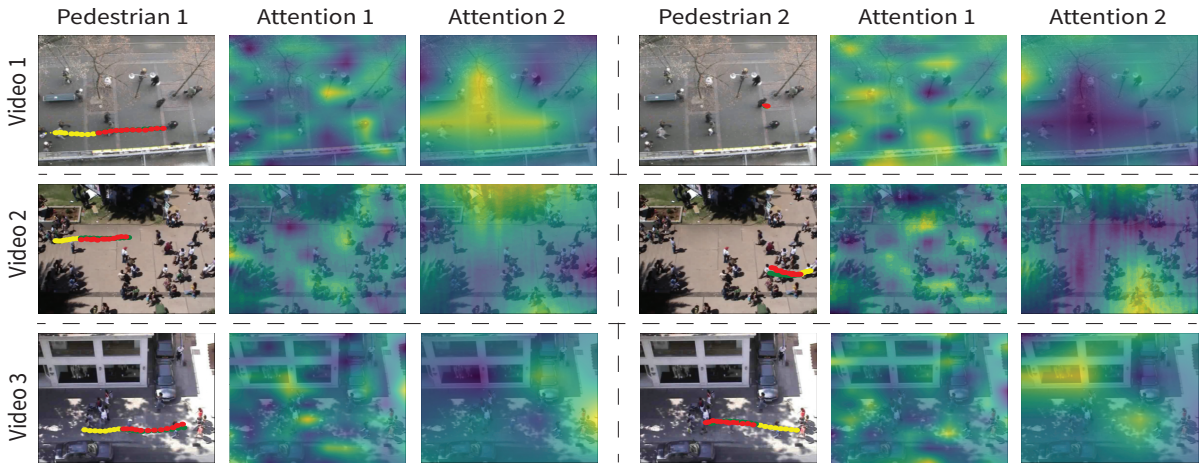

Figure 4. Visualization of conditional spatio-temporal attentions modules (Attention 1 = Ψ_2 , Attention 2 = Ψ_3) for the first observed frame. Each row corresponds to a different video from a different dataset. In each row, for two different pedestrians moving at the same time (first three columns for pedestrian 1 and second three columns for pedestrian 2), we show the observed/ground-truth/predicted trajectory in the first column and the output of the first and second conditional attentions (Ψ_2 and Ψ_2) in the second and third columns, respectively.

captures different levels of visual abstractions. Attention 1 (corresponding to Ψ_2) attends to the pedestrians and objects, while Attention 2 (corresponding to Ψ_3) attends to more distant visual primitives in the scene, such as locations in front of the target pedestrian, buildings and cars in distance. Notice also that the attention maps for different pedestrians in the same scene are completely different, thanks to our conditional model. For example, in video 3, for the first pedestrian who is moving from left to right, the Attention 2 focuses on the car and distant humans in front of the target pedestrian. On the other hand, for the second pedestrian moving from right to left, the Attention 2 focuses on the building and the humans in front of the target person.

4.2.3 Effect of Different Components

Number of Attention Maps. As mentioned before, the conditional attention module, Ψ_i, has a dual attention mechanism with n spatial attention mappings and m channels for each attention map. Figure 5 (left) shows the effect of the number of spatial attentions $n \in \{2, 4, 8, 16\}$ for fixed $m = 8$. Notice that our method performs robustly for different values of n.

Regularization Parameter. Our loss function in (5) is composed of the mean-squared loss and a regularization term, $\mathcal{L}_{reg}$, which controls the smoothness of the future trajectories compared to the observed ones. Figure 5 (right) shows the effect of the regularization parameter $\lambda \in \{0, 0.25, 0.50, 0.75, 1\}$ on the ADE performance. Our model obtains lower errors for $\lambda \in \{0.5, 0.75\}$ and the performance generally degrades for larger values of λ. This comes from the fact that larger regularization prevents the model from capturing sudden changes in trajectory of target pedestrians (we used $\lambda = 0.5$ for our main experiments).

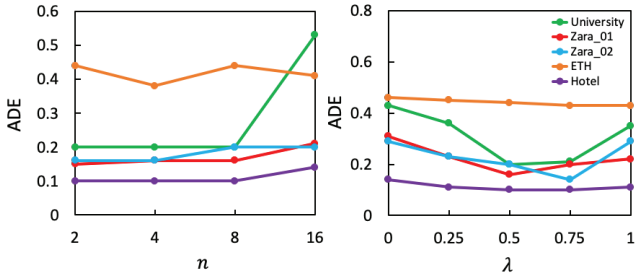

Figure 5. Left: Effect of the number of spatial attention maps (n) on the performance. Right: Effect of regularization parameter (λ).

5. Conclusions

We presented Introvert, a method for pedestrian trajectory prediction using conditional 3D visual attention mechanism on dynamic scene context. We showed that Introvert captures both human-space and human-human interactions by generating distinctive spatio-temporal attentions for each pedestrian. As we discussed, our computational cost is independent of the crowd density by taking advantage of generating flexible yet fixed-size visual primitives and their attentions. We also benchmark the performance of Introvert on ADE and FDE metrics across UCY and ETH datasets, showing that it improves the state-of-the-art performance.

Acknowledgements

This work is partially supported by DARPA Young Faculty Award (D18AP00050), ONR (N000141812132) and ARO (W911NF1810300).

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 5

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

[3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 5, 7

[4] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006. 1

[5] Huikun Bi, Zhong Fang, Tianlu Mao, Zhaoqi Wang, and Zhigang Deng. Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10383–10392, 2019. 2

[6] Huikun Bi, Ruisi Zhang, Tianlu Mao, Zhigang Deng, and Zhaoqi Wang. How can i see my future? fvtraj: Using first-person view for pedestrian trajectory prediction. 2

[7] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *Advances in neural information processing systems*, pages 352–361, 2018. 4

[8] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 921–930, 2019. 2

[9] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012. 1

[10] Wongun Choi and Silvio Savarese. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013. 1

[11] Alex G Cunningham, Enric Galceran, Dhanvin Mehta, Gonzalo Ferrer, Ryan M Eustice, and Edwin Olson. Mpdm: Multi-policy decision-making from autonomous driving to social robot navigation. In *Control Strategies for Advanced Driver Assistance Systems and Autonomous Driving Functions*, pages 201–223. Springer, 2019. 1

[12] Stuart Eiffert, Kunming Li, Mao Shan, Stewart Worrall, Salah Sukkarieh, and Eduardo Nebot. Probabilistic crowd gan: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network. *IEEE Robotics and Automation Letters*, 5(4):5026–5033, 2020. 2

[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 1, 2, 5

[14] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 1, 2

[15] Noriaki Hirose, Amir Sadeghian, Patrick Goebel, and Silvio Savarese. To go or not to go? a near unsupervised learning approach for robot navigation. *arXiv preprint arXiv:1709.05439*, 2017. 1

[16] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6319–6328, 2020. 2

[17] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6272–6281, 2019. 2

[18] D. Huynh and E. Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[19] D. Huynh and E. Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[20] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2375–2384, 2019. 1, 2

[21] Vasiliy Karasev, Alper Ayvaci, Bernd Heisele, and Stefano Soatto. Intent-aware long-term prediction of pedestrian motion. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2543–2549. IEEE, 2016. 1

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

[23] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, pages 137–146, 2019. 1, 2, 5

[24] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014. 1

[25] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 120–127. IEEE, 2011. 1

[26] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting

agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017. 1, 2

[27] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 2, 5

[28] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from 3d simulation for pedestrian trajectory prediction in unseen cameras. *arXiv preprint arXiv:2004.02022*, 2020. 2

[29] Buyu Liu, Francesco Pittaluga, Manmohan Chandraker, et al. Smart: Simultaneous multi-agent recurrent trajectory prediction. *arXiv preprint arXiv:2007.13078*, 2020. 2

[30] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *2010 IEEE International Conference on Robotics and Automation*, pages 464–469. IEEE, 2010. 1

[31] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *arXiv preprint arXiv:2004.02025*, 2020. 2, 5

[32] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009. 2

[33] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 1, 2

[34] Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R Jadhav, Paul Pu Liang, and Louis-Philippe Morency. Diverse and admissible trajectory forecasting through multimodal context understanding. *arXiv preprint arXiv:2003.03212*, 2020. 2

[35] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 2, 5

[36] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010. 1, 2

[37] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John 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*, pages 6262–6271, 2019. 2

[38] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[39] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio 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*, pages 1349–1358, 2019. 1, 2, 5

[40] Amir Sadeghian, Ferdinand Legros, Maxime Voisin, Ricky Vesel, Alexandre Alahi, and Silvio Savarese. Car-net: Clairvoyant attentive recurrent network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018. 1, 2

[41] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. *arXiv preprint arXiv:2001.03093*, 2020. 1, 2, 5

[42] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015. 1

[43] Olly Styles, Tanaya Guha, Victor Sanchez, and Alex Kot. Multi-camera trajectory forecasting: Pedestrian trajectory prediction in a network of cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1016–1017, 2020. 2

[44] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7416–7425, 2020. 2

[45] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020. 2

[46] Meng Keat Christopher Tay and Christian Laugier. Modelling smooth paths using gaussian processes. In *Field and Service Robotics*, pages 381–390. Springer, 2008. 1

[47] Luca Anthony Thiede and Pratik Prabhanjan Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9954–9963, 2019. 2

[48] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 797–803. IEEE, 2010. 1

[49] Adrien Treuille, Seth Cooper, and Zoran Popović. Continuum crowds. *ACM Transactions on Graphics (TOG)*, 25(3):1160–1168, 2006. 1

[50] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3302–3309, 2014. 1

[51] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007. 1

[52] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011. 2

[53] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015. 1

[54] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. *arXiv preprint arXiv:2005.08514*, 2020. 2, 5, 7

[55] Lidan Zhang, Qi She, and Ping Guo. Stochastic trajectory prediction with social graph network. *arXiv preprint arXiv:1907.10233*, 2019. 2

[56] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019. 2, 7