Temporal Pyramid Network with Spatial-Temporal Attention for Pedestrian Trajectory Prediction

Yuanman Li, *Member, IEEE,* Rongqin Liang, *Student Member, IEEE,* Wei Wei, *Senior Member, IEEE,* Wei Wang, *Member, IEEE,* Jiantao Zhou, *Senior Member, IEEE,* and Xia Li, *Member, IEEE*

Abstract—Understanding and predicting human motion behavior with social interactions have become an increasingly crucial problem for a vast number of applications, ranging from visual navigation of autonomous vehicles to activity prediction of intelligent video surveillance. Accurately forecasting crowd motion behavior is challenging due to the multimodal nature of trajectories and complex social interactions between humans. Recent algorithms model and predict the trajectory with a single resolution, making them difficult to exploit the long-range information and the short-range information of the motion behavior simultaneously. In this paper, we propose a temporal pyramid network for pedestrian trajectory prediction through a squeeze modulation and a dilation modulation. The hierarchical design of our framework allows to model the trajectory with multi-resolution, then can better capture the motion behavior at various tempos. By progressively combining the global context with the local one, we finally construct a coarse-to-fine hierarchical pedestrian trajectory prediction framework with multi-supervision. Further, we introduce a unified spatial-temporal attention mechanism to adaptively select important information of persons around in both spatial and temporal domains. We show that our attention strategy is intuitive and effective to encode the influence of social interactions. Experimental results on two benchmarks demonstrate the superiority of our proposed scheme.

Index Terms—Social Computing, Deep Learning, Social Interactions, Temporal Pyramid Network, Social Behavior, Trajectory Prediction, Spatial-Temporal Attention

1 Introduction

Modeling the behaviors of pedestrians is an essential step for many applications [1], [2], [3], [4], [5], [6], [7], [8], including socially-aware robots for visual navigation [1], [2], video surveillance systems to identify suspicious activities [3], [4], [5], and self-driving platforms for safe decision making [7]. Trajectory prediction as one of the most important future behavior modeling tasks, aims to predict possible future trajectories according to historical paths in the last few seconds [9], [10], [11], [12]. An accurate trajectory prediction can help autonomous systems plan ahead through complex social interaction scenarios.

Despite its importance, predicting the trajectory is very challenging due to the inherent properties of pedestrians. First, human motions are highly *multimodal*, which means that there could be several socially-acceptable and distinct future behaviors under the same historical trajectory. The resulting motion randomness is often hard to formalize considering different personal habits. Second, human motions are highly affected by the people around them, and the interactions between pedestrians could make them walk in parallel, walk within a group, or change direction/speed for collision avoidance. Jointly modeling the complex social behaviors is rather challenging in reality. Third, the scene is dynamic due to the fact that humans may frequently appear and disappear in a scene, and each scene often has a different number of humans. The dynamic property requires that algorithms are capable of handling a variable number of inputs.

The pioneering work to trajectory prediction mainly focused on the social behavior modeling by using handcrafted features [13], [14], [15], [16], [17], [18], [19], [20]. This type of approaches ignored the multimodal property of human behavior and was designed to predict a single future trajectory for each person. In recent years, many deep learning based algorithms were devised for pedestrian trajectory prediction and achieved noticeable performance improvements. The temporal attributes of pedestrian motions encouraged researchers to model trajectories using Recurrent Neural Networks (RNNs) [7], [10], [21], [22], [23], [24]. The outputs of recurrent models were made to interact with each other through a designed aggregation layer. In order to directly modeling interactions between pedestrians, some

- *This work was supported in part by the Natural Science Foundation of China under Grant 62001304, Grant 61871273 and Grant 61971476; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515110410; in part by the Macau Science and Technology Development Fund under Grant SKL-IOTSC-2018-2020, Grant 077/2018/A2 and Grant 0060/2019/A1; in part by the Research Committee at University of Macau under Grant MYRG2018-00029-FST and Grant MYRG2019-00023-FST. (Corresponding author: Xia Li)*
- *Yuanman Li, Rongqing Liang and Xia Li are with Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China (e-mail: yuanmanli@szu.edu.cn, 1810262064@email.szu.edu.cn, lixia@szu.edu.cn)*
- *Wei Wei is with Xi'an University of Technology, China (e-mail: weiwei@xaut.edu.cn).*
- *Wei wang is with School of Intelligent Systems Engineering, Sun Yat-sen University, China (e-mail: wangw328@mail.sysu.edu.cn)*
- *Jiantao Zhou is with the State Key Laboratory of Internet of Things for Smart City, and also with the Department of Computer and Information Science, University of Macau, Macau (e-mail: jtzhou@um.edu.mo)*

works proposed to use Graph Neural Networks (GNNs) to encode the social information [25], [26], [27], [28]. Though the topology of graphs seems like a straightforward way to represent social interactions, it is rather challenging since the dynamic property makes the topology of graphs changing in-between prediction steps. To predict socially plausible and diverse trajectories, Generative Adversarial Networks (GANs) have been employed to predict the distribution of the future trajectory [11], [29], [30], [31], [32]. For these algorithms, both generators and discriminators were based on RNNs, where social interactions between pedestrians were encoded by an aggregation layer, such as pooling.

Though trajectory prediction has been studied from many aspects, the existing algorithms still suffer from the following two limitations. First, all the existing methods encoded the input historical trajectory and decoded the output trajectory with a single resolution (*i.e.*, a fixed length of time steps), which failed to fully exploit the temporal relations of the motion behavior. We argue that simultaneously modeling the global context (*e.g.*, where the pedestrian plans to go, and where is the destination of the trajectory to be predicted) and the local context (*e.g.*, the direction and speed of the pedestrian at a certain time) with a single resolution is not very effective or rather difficult, if possible. Second, most of previous methods adopted an aggregation mechanism that takes the pedestrian trajectory modelled by recurrent units as inputs, such as pooling, to model the interactions between pedestrians through learning a global compact representation of the scene. However, such mechanisms are not intuitive and may fail to capture some important information in both temporal and spatial domains [12]. Further, it was shown that such approaches are ineffective in real scenarios [11].

To alleviate above two limitations, in this work, we propose a novel Temporal Pyramid Network with Spatial-Temporal Attention (TPNSTA) for pedestrian trajectory prediction. Our method achieves remarkable performance gains comparing to existing algorithms. Specifically, we devise a pyramid feature extractor composed of a squeeze module and a dilation one for multi-scale feature generation from a fixed length input trajectory. The pyramidal features are then fed into an RNN based Encoder-Decoder to generate coarse-to-fine future trajectories through progressively combining higher pyramid levels with lower ones. We further use a multi-supervision strategy to endow effective representations of all the pyramid levels. Our multi-scale prediction framework obeys human intuitive understandings, *i.e.*, gradually from a destination to a detailed path planning, thus could potentially simplify the inference process of the future trajectory. Besides, to produce socially-aware trajectories, we further design a spatial-temporal attention mechanism to encode the influence caused by pedestrians around. Finally, similar to Social-GAN [11], our network is trained in an adversarial manner to produce multiple *socially-acceptable* and diverse motion trajectories conforming to the multimodal behavior of pedestrians. It should be noted that pyramid representation as one of the most important hierarchical representation techniques has been extensively studied in computer vision, such as SIFT [33], HoG [34], FPN [35] and SPP [36]. Most of them were designed in spatial domain and only for detection or recognition tasks. To the best of our knowledge, this is the first attempt that models trajectories in a scene as temporal pyramids. As will be shown later, our method achieves the state-of-the-art performance on two benchmark datasets. The main contributions of our work can be summarized as follows:

1) A novel temporal pyramid network is proposed to capture the motion behaviors of pedestrians at various tempos. With our hierarchical design, both short-range and long-range motion behaviors can be effectively exploited. Through progressively combining the global context with the local one, our method allows coarse-to-fine trajectory modeling in a multi-supervised fashion.
2) We further propose a spatial-temporal attention mechanism to encode the influence of social interactions in both spatial and temporal domains. Different from previous aggregation strategies, our approach is more intuitive and effective to adaptively select important information from surrounding people.
3) The proposed hierarchical design and attention mechanism can be regarded as general modules, and easily extended to other pedestrian trajectory prediction frameworks, thus potentially bringing performance improvements.

The remainder of this paper is organized as follows. Section 2 gives a brief review of related works. Section 3 details our proposed TPNSTA for pedestrian trajectory prediction. Extensive experimental results are presented in Section 4, and we finally draw a conclusion in Section 5.

2 Related Works

2.1 Pedestrian Trajectory Prediction

Traditional pedestrian trajectory prediction algorithms heavily rely on the handcrafted rules to describe human motions [13], [14], [15], [16], [17], [18], [19], [20]. For example, Helbing *et al.* [13] devised a social forces model, employing dynamic systems to model both the attractive force towards a destination and repulsive force to avoid collision. Pellegrini *et al.* [20] designed a linear trajectory avoidance method through jointly modeling the interactions between different targets and the scene information. These methods, though demonstrated the importance of interaction modeling, were based on handcrafted rules, which are difficult to generalize in more complex new scenes.

Recently, the data-driven based algorithms have received significant attention in the community. Among them, RNN and its variant LSTM have been widely adopted for pedestrian trajectory prediction [7], [10], [21], [22], [23], on the basis of their good performance in many sequence prediction tasks [37], [38], [39]. Social-LSTM [10] as one of the earliest deep learning based algorithms on pedestrian trajectory prediction, encoded the motion information of each pedestrian using a recurrent network. It further employed a social-pooling mechanism for interaction information aggregation. CIDNN [22] leveraged an LSTM to model the motion behaviors of all pedestrians, which considered different importance of persons to a target pedestrian in a crowd interaction module. The recent works such as PIF [7] and SR-LSTM [23] aimed to enhance the performance

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/TNSE.2021.3065019, IEEE Transactions on Network Science and Engineering

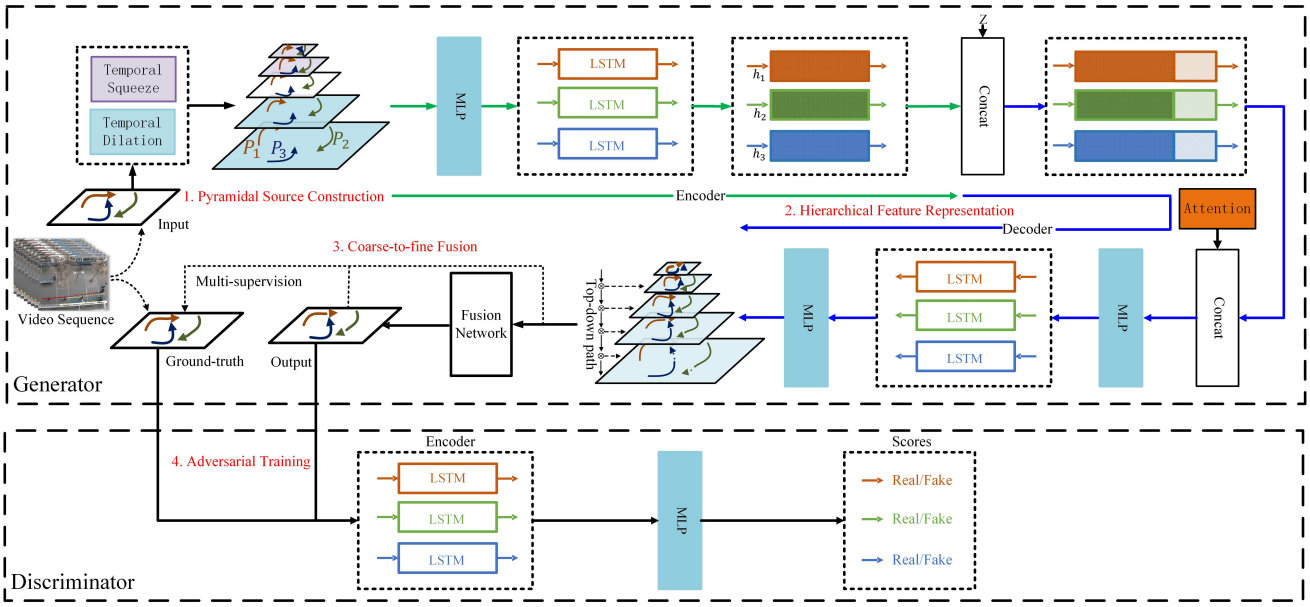


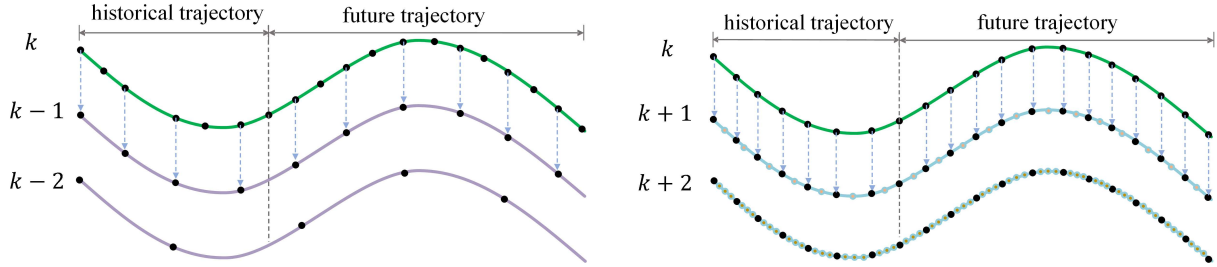
Fig. 1. The framework of our proposed algorithm. The network consists of a generator and a discriminator. The input of the generator is the historical trajectories of pedestrians, and the output is the corresponding predicted future trajectories. 1) First, the pyramidal source is constructed through the temporal squeeze modulation and the temporal dilation modulation; 2) then, an encoder-decoder network is adopted for hierarchical feature learning; 3) features are fed into a coarse-to-fine fusion network to generate the future trajectories with multi-supervision. Parameters are shared across all scales of the pyramid. The fusion network is presented in Fig. 3, and the attention block to model the influence of social interactions is detailed in Section 3.3.

of [10] by taking the scene context as side information, and adopting new pooling mechanisms. Though the RNN architecture endowed above methods to learn and predict pedestrian trajectories in a data-driven manner, they failed to capture the multimodal nature of human, *i.e.*, there could be multiple socially-acceptable and distinct possible future trajectories under the same historical trajectory.

The graph based approaches have recently been devised to model the social interactions between pedestrians [12], [25], [26], [27], [40], [41]. The work [25] formulated the task of pedestrian trajectory prediction using a spatio-temporal graph, where the nodes of the graph represented pedestrians in the crowd. The method [40] proposed a social graph network constructed on timely location and speed direction, to extract non-symmetric pairwise relationships and social interactions. Social-BiGAT [41] introduced a graph attention network to model the social interactions between pedestrians, where all pedestrians in a scene were allowed to interact. A similar graph attention network was also used in [26] to model spatial interactions. More recently, the work Social-STGCNN [12] proposed to model trajectories using the spatio-temporal graph convolution neural network, and achieved quite good performance. In order to produce multimodal pedestrian trajectories, some researchers suggested constructing the recurrent models with generative settings, which led to learning the distribution of the future trajectory rather than directly generating a deterministic path [11], [29], [30], [31], [32]. Social-GAN [11] is the pioneering trajectory prediction work incorporating the LSTM model with the generative adversarial networks (GANs) [42], permitting to produce multiple plausible trajectories, then capture the multimodal nature of the future paths. SoPhie [29] improved Social-GAN [11] through a scene feature extraction component, and it also employed an attention mechanism for social and physical constraints consideration. A similar work using Gated Recurrent Units (GRUs) instead of LSTMs was introduced in [32]. SafeCritic [30] synergized GANs with reinforcement learning, where the former generated "real" trajectories while the latter produced "safe" trajectories. MATF [31] encoded historical trajectories of multiple agents and the scene context as a tensor, then applied convolutional fusion to capture interactions through an adversarial loss. The aforementioned generative approaches achieved promising performance to reveal the multimodal nature of human. However, they all modeled the motion behaviors with a single resolution, making them difficult to capture the long-range and short-range behaviors simultaneously.

2.2 Pyramid Representations

Feature pyramids play a significantly important role in the field of computer vision [33], [34], [43], [44], [45], [46]. For example, the popular traditional hand-engineered feature extractors such as SIFT [33] and HoG [34], were designed to compute features in a multi-scale space, and have been used in numerous computer vision tasks. Lin *et. al.* [35] accommodated the idea of pyramid representation to deep convolutional neural networks, achieving quite promising performance in the detection task. A similar strategy was also adopted to construct the pyramidal pooling layer, which aimed to produce a fixed length representation regardless of the input image size [36]. Most of previous pyramid representation approaches were originally designed in spatial domain. More recently, some works proposed to extract hierarchical features in the temporal domain, and demonstrated its effectiveness in video signal processing tasks, such as action recognition [47] and scene classification [48].



(a) Temporal squeeze modulation (b) Temporal dilation modulation

Fig. 2. Illustration of the proposed (a) temporal squeeze modulation, and (b) temporal dilation modulation.

2.3 Attention Mechanism

There has been a long line of previous works [49], [50] employing the attention mechanism into neural networks. Bahdanau *et al.* [49] implemented a mechanism of attention for the machine translation task, modeling the dependencies between different parts regardless of their distance in the input or output sequences. Henceforward, many follow-up attention based natural language processing algorithms have been proposed, such as [51], [52]. Besides the machine translation task, the concept of attention mechanism has also been used in image generation [53] and image caption generation [54]. More recently, a few researchers have begun to apply attention mechanisms to pedestrian trajectory prediction, encoding the influence of social interactions among pedestrians [21], [23], [25], [55]. Fernando *et al.* [21] applied the "soft" and "hardwired" attention mechanisms to map the trajectory information from the local neighborhood to the future positions of the target pedestrian. Zhang *et al.* [23] established an attention based information selection strategy, where a pedestrian-wise attention mechanism and a motion gate were devised to jointly select important information from neighboring pedestrians. Nevertheless, most of the existing methods focused only on the spatial context for trajectory prediction, while the temporal information describing the historical walking speeds and directions was ignored. In this work, we propose a unified attention mechanism, where both the spatial context and the temporal context are considered.

3 Proposed Temporal Pyramid Network with Spatial-Temporal Attention (TPNSTA)

3.1 Problem Formulation

Given a set of N pedestrians with observed positions in a video sequence over a fixed duration, the trajectory prediction algorithm aims to jointly reason and forecast the future trajectories of all pedestrians in the upcoming time steps. Let (x_i^t, y_i^t) be the position of the i-th pedestrian at the time step t, where $i \in \{1, ..., N\}$. Denote $X_i^{(t_1:t_2)} = [(x_i^{t_1}, y_i^{t_1}), ..., (x_i^{t_2}, y_i^{t_2})]$ as the observed historic trajectory of the i-th pedestrian from the time step t_1 to t_2. Similarly, we define $Y_i^{(t_1:t_2)}$ as the future trajectory of the i-th pedestrian from the time t_1 to t_2. The trajectory prediction algorithm takes as input the previous trajectories with t_o time steps of all pedestrians in a scene, denoted by

$$\mathcal{X} = \{X_1^{(1:t_o)}, ..., X_N^{(1:t_o)}\}, \tag{1}$$

and aims to predict their trajectories in the next t_p time steps simultaneously. We use $\mathcal{Y}$ to represent the ground-truth future trajectories, i.e.,

$$\mathcal{Y} = \{Y_1^{(t_o+1:t_o+t_p)}, ..., Y_N^{(t_o+1:t_o+t_p)}\}. \tag{2}$$

For the sake of brevity, we hereafter will drop the superscript when there is no ambiguity, i.e., $X_i \triangleq X_i^{(1:t_o)}$ and $Y_i \triangleq Y_i^{(t_o+1:t_o+t_p)}$. We further use X and Y to represent a generic historic trajectory and the corresponding future trajectory, respectively.

3.2 Temporal Pyramid Network for Trajectory Prediction

Motivated by the great success of the pyramid representation [33], [34], [35], we propose a temporal pyramid framework tailored for pedestrian trajectory prediction. Compared with the existing algorithms, which model the trajectory with a single resolution, our method has many fundamental advantages. For instance, our temporal pyramid architecture is effective in exploiting the motion behaviors at various tempos, and the coarse-to-fine generation process could greatly facilitate the joint modeling of both global and local contexts with multi-resolution. Besides, benefiting from the LSTM network, all levels of pyramids share the same parameters. This allows our method to operate on a single-branch backbone network regardless how many levels are adopted, then avoid to increase the model complexity.

To ensure that the predicted pedestrian trajectories conform to the multimodal behavior of pedestrians, we design our model under the framework of GANs. As shown in Fig. 1, the proposed framework contains a generator and a discriminator, which are trained in opposition to each other. For better illustration, we decompose our framework into the following four components: 1) pyramidal source construction; 2) hierarchical feature representation; 3) coarse-to-fine fusion with multi-supervision and 4) adversarial training.

3.2.1 Pyramidal source construction

For each trajectory of the input $\mathcal{X}$, we propose to generate a set of L hierarchical features with multi-resolution, and then construct a feature pyramid, having increasingly richer temporal information from top to bottom. With the aid of the pyramid framework, our method can fully exploit the short range behavior and the long range behavior in a hierarchical

way. As depicted in Fig. 1, this process can be summarized as two procedures, i.e., 1) the temporal squeeze modulation, and 2) the temporal dilation modulation.

Temporal squeeze modulation: Assume that there are totally L scales of the temporal pyramid network for each trajectory. Denote the feature of the k-th scale as X_i^k, which is identical to X_i. The goal of the temporal squeeze modulation is to reduce the impact of the local context, and generate a set of features with increasingly stronger global context from $\mathcal{X}$. A simple way to this end is to consecutively sample features along the temporal dimension. We propose to gradually produce the top $k-1$ scales through uniformly sampling from the scale below with an interval factor of 2. In this work, we refer to the above process as the temporal squeeze modulation.

Fig. 2(a) illustrates the procedure of the temporal squeeze modulation. For the ℓ-scale ($\ell < k$), the feature can be represented as

$$X_i^\ell = [(\tilde{x}_i^1, \tilde{y}_i^1), ..., (\tilde{x}_i^{m_\ell}, \tilde{y}_i^{m_\ell})], \tag{3}$$

where $m_\ell = \lceil t_o / 2^{k-\ell} \rceil$, and

$$\tilde{x}_i^j = x_i^{1+2^{k-\ell}(j-1)}, \;\; \tilde{y}_i^j = y_i^{1+2^{k-\ell}(j-1)}. \tag{4}$$

We can see that the detailed short-range information of the motion is gradually weakened by the temporal squeeze modulation, which encourages the higher scales to capture more long-range motion behaviors of pedestrians.

Temporal dilation modulation: It should be noted that the observed trajectories are usually of short duration, then the number of scales generated by the temporal squeeze modulation is rather limited. As a consequence, the sole temporal squeeze modulation cannot fully capture the motion behaviors at various tempos. To handle this issue, we further introduce a complementary procedure called temporal dilation modulation, which is similar to the dilated convolution operator widely used in various vision tasks. The temporal dilation modulation could generate more dense trajectories for hierarchical feature representation, then exploit richer short-range information of the motion.

We propose to conduct the temporal dilation modulation through the trajectory interpolation. Note that pedestrians usually walk/run at varying speeds, accelerations and in different directions over time. In order to generate smooth dense trajectories, in this work, we adopt the cubic spline algorithm for the trajectory interpolation. For simplicity, we rewrite the observed trajectory of the i-th pedestrian as a series of time-position pairs

$$TX_i = \big[(1, (x_i^1, y_i^1)), ..., (t_o, (x_i^{t_o}, y_i^{t_o}))\big]. \tag{5}$$

We adopt the cubic spline algorithm to seek for a piecewise-cubic function $f(t) : \mathbb{R} \to \mathbb{R}^2$

$$f(t) = \begin{cases} f_1(t), & 1 \le t < 2 \\ \vdots & \\ f_{t_o-1}(t), & t_o - 1 \le t \le t_o \end{cases}, \tag{6}$$

where

$$f_k(t) = a_k + b_k(t-k) + c_k(t-k)^2 + d_k(t-k)^3 \tag{7}$$

represents the curve between the time steps k and $k+1$, and $a_k, b_k, c_k, d_k \in \mathbb{R}^2$ are parameters of the cubic spline. According to the cubic spline algorithm, given the trajectory X_i, there exists a unique set of parameters $\{a_k, b_k, c_k, d_k\}_{k=1,\cdots,t_o-1}$ such that the resulting trajectory curve passes through all the positions in X_i with continuous velocity and acceleration at each position.

Upon having $f(t)$, the feature X_i^ℓ at the ℓ-scale ($\ell > k$) can be calculated by interpolating positions of the in-between and unobserved time steps as shown in Fig. 2(b). Mathematically, we have

$$X_i^\ell = [f(1), f(1+\frac{1}{c}), f(1+\frac{2}{c}), ..., f(2), ..., f(t_o - \frac{1}{c}), f(t_o)], \tag{8}$$

where $c = 2^{\ell-k}$. The interpolated dense trajectories offer more local information for the lower scales, permitting to capture more short-range motion behaviors of pedestrians. With above two modulations, we finally construct the pyramidal source as shown in Fig. 1.

3.2.2 Hierarchical feature representation

Learning a good feature representation is crucial for the trajectory modeling [56]. In this work, we simply use a similar network architecture (without the pooling module) proposed in [11], as the backbone to extract hierarchical features from the constructed pyramid. As shown in Fig. 1, the backbone network consists of two components, *i.e.*, the encoder and the decoder.

At the encoder side, we first adopt a single layer MLP with ReLU activation to embed the position of each pedestrian as

$$e_i^t = MLP(x_i^t, y_i^t; \Theta_{me}), \tag{9}$$

where $t \le t_o$ and Θ_{me} represents the parameters of MLP. The embedded feature e_i^t is then fed into an LSTM block, which produces the hidden state at the time step t

$$h_i^t = LSTM(h_i^{t-1}, e_i^t; \Theta_{le}). \tag{10}$$

Note that the parameters of LSTM (Θ_{le}) are shared among all the pedestrians and the scales of the pyramid.

In order to generate multiple socially-acceptable trajectories, our model is designed under the framework of GANs. The generator of GANs aims to build a mapping function from a prior noise distribution $p_z(z)$ to the data space [42]. Similar to [11], at the decoder side, we concatenate a noise vector z sampled from the standard normal distribution to the hidden state $h_i^{t_o}$. We write

$$h_i^{t_o} := [h_i^{t_o}, z]. \tag{11}$$

Then for each z, we recurrently decode the hierarchical features of the trajectory as follows

$$\begin{aligned} \hat{e}_i^{t-1} &= MLP(\hat{x}_i^{t-1}, \hat{y}_i^{t-1}; \Theta_{md}) \\ h_i^t &= LSTM(h_i^{t-1}, \hat{e}_i^{t-1}; \Theta_{ld}), \\ (\hat{x}_i^t, \hat{y}_i^t) &= MLP(h_i^t; \Theta'_{md}) \end{aligned} \tag{12}$$

where $t \ge t_o + 1$, $(\hat{x}_i^{t_o}, \hat{y}_i^{t_o}) = (x_i^{t_o}, y_i^{t_o})$, and $\Theta_{md}, \Theta_{ld}, \Theta'_{md}$ are parameters to be learned.

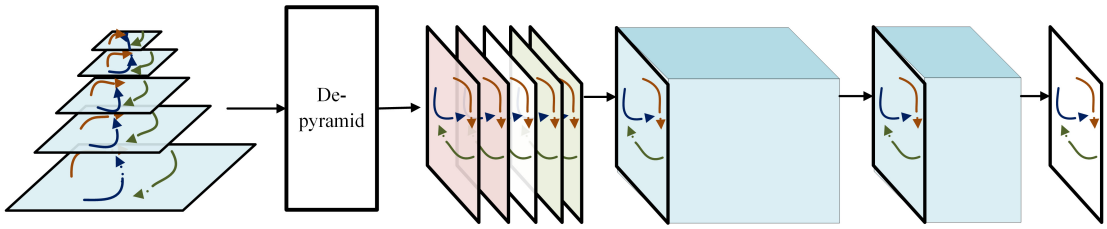


Fig. 3. The framework of the fusion network, where the convolutional layers have the 1×1 kernel size, and the number of channels are $8, 4$ and 1, respectively.

3.2.3 Coarse-to-fine fusion with multi-supervision

With the above two components, one can generate a set of hierarchical features for each trajectory. As shown in Fig. 1, the coarse-to-fine pyramid organizes features in a top-down pathway, where the top coarse features with global context are progressively merged to the bottom fine features with rich local context. Denote the extracted feature of the ℓ-scale as $\hat{X}_i^\ell$, which is updated by merging the information of the above scale. We write

$$\hat{X}_i^\ell := \frac{1}{2}(\hat{X}_i^\ell \oplus \hat{X}_{i,\uparrow}^{\ell-1}), \tag{13}$$

where $\hat{X}_{i,\uparrow}^{\ell-1}$ means upsampling (experimentally, MLP was adopted) $\hat{X}_i^{\ell-1}$ by a factor of 2, and $\oplus$ serves as the element-wise addition. This process is iterated until the finest resolution feature is merged.

To ensure that both the long-range and short-range motion behaviors can be fully exploited, all the scales are supervised during training, where the corresponding loss function is formulated as

$$\mathcal{L}_s = \frac{1}{NL}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\lambda_\ell \|\hat{X}_i^\ell - Y_i^\ell\|_2^2. \tag{14}$$

Here Y_i^ℓ is the ground-truth pyramidal source of the future trajectory, which can be constructed from Y_i in the same way detailed in Section 3.2.1. The hyper-parameter λ_ℓ is inversely proportional to the feature length of $\hat{X}_i^\ell$, which we empirically set

$$\lambda_\ell = \frac{t_p}{m'_\ell}. \tag{15}$$

Here m'_ℓ represents the length of $\hat{X}_i^\ell$.

The final predicted trajectory is produced by a fusion layer as shown in Fig. 3. A de-pyramid layer is first adopted to sample the hierarchical features $\{\hat{X}_i^\ell\}_{\ell=1}^L$ to a fixed length t_p. In detail, we use an MLP to up-sample those features with the length shorter than t_p to a fixed length t_p, while uniformly sampling the features longer than t_p with an interval factor of 2 to the length t_p. The results are then concatenated as a tensor of size $L \times 2 \times t_p$, which is further processed through three convolutional layers to fuse information across the whole pyramid. The fusion layer finally generates the predicted trajectory $\hat{Y}_i$. We supervise the final output using the loss

$$\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N}\|\hat{Y}_i - Y_i\|_2^2. \tag{16}$$

3.2.4 Adversarial training

The discriminator D aims to enforce the generator G to capture the data distribution, then to generate more plausible trajectories. The architecture of the discriminator is shown in Fig. 1. Similar to [11], we design the discriminator as a neural network consisting of an LSTM component and an MLP component. The LSTM component takes as input the ground-truth trajectory $[X, Y]$ or the generated trajectory $[X, \hat{Y}]$. The last hidden state of the LSTM is fed into the MLP, which outputs the classification score. Let $\hat{Y} = G(z, X)$. The adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_{avd} &= \mathbb{E}_{X,Y\sim P_{data}(X,Y)}[\log D(X,Y)] \\ &+ \mathbb{E}_{X\sim P_{data}(X),z\sim P_z(z)}[\log(1 - D(X, G(z,X))].\end{aligned} \tag{17}$$

Finally, training the network is cast into a two-player min-max game with the following objective function

$$\min_G \max_D \mathcal{L}_{avd} + \mathcal{L}_s + \mathcal{L}_f. \tag{18}$$

Above problem can be solved by alternatively updating the generator G and the discriminator D.

3.3 Proposed Spatial-Temporal Attention Mechanism

In order to model the interactions in a crowding scene, one needs to share the information among all pedestrians. Traditional aggregation mechanisms, such as pooling [10], are not intuitive and also shown to be ineffective in real scenarios [41]. To avoid collision in a crowding environment, pedestrians not only consider the distances between people around, but also pay attention to other status, such as the walking speed and walking direction over the previous time steps. Obviously, such information contains both spatial context and temporal context.

In recent years, attention mechanism has been attracting great interest in many tasks. For instance, it was successfully employed in the machine translation task to emphasize those most related parts of the source sentence to the current decoding state [52], [57], [58]. Motivated by its success, in this work, we propose a unified attention mechanism in both the spatial domain and the temporal domain for pedestrian trajectory prediction. The framework of the proposed spatial-temporal attention is presented in Fig. 4.

3.3.1 Temporal attention

Temporal context can describe the walking speed and direction over the previous time steps of each person. Such information, obviously is very crucial for the natural navigation of pedestrians in a social interaction scene. For example, people often tend to slow down or stop when someone

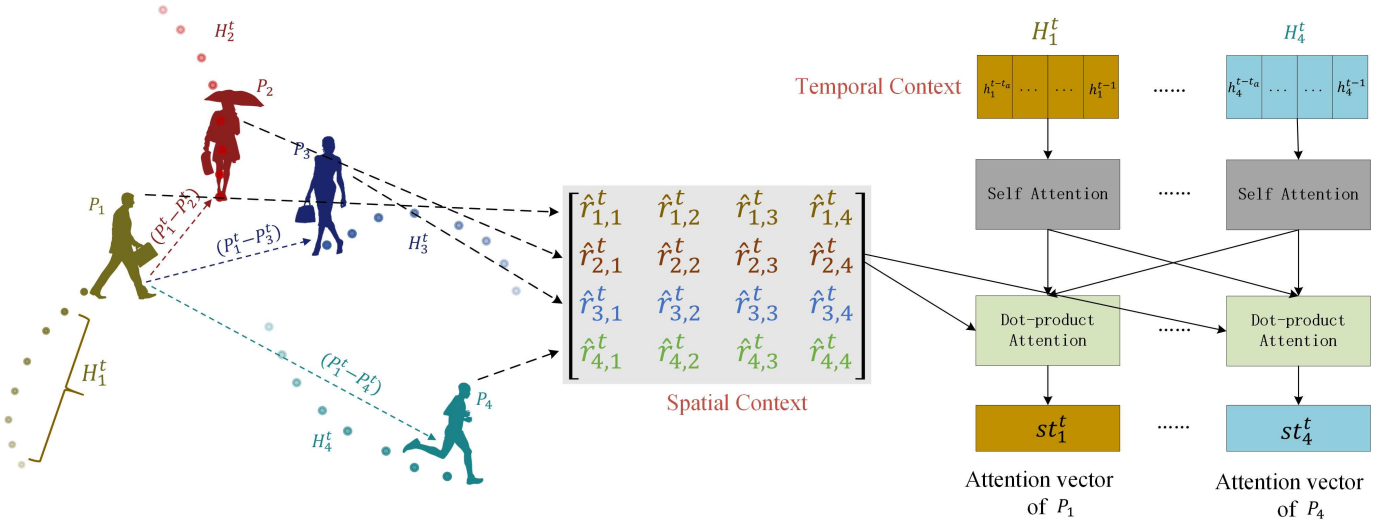


Fig. 4. Illustration of the spatial-temporal attention mechanism. The dots represent historical trajectories of pedestrians. p_i represents the pedestrian i, and t denotes a certain time step.

is running towards them at a fast speed. Similar to the sequence-to-sequence generation problem [57], the temporal context allows us to endow different degrees of attention to different parts of the historical trajectory. We use H_i^t to represent the temporal context, which consists of all the latest hidden states in a period of t_a time steps. Namely,

$$H_i^t = [h_i^{t-t_a}, h_i^{t-t_a+1}, ..., h_i^{t-1}] \in \mathbb{R}^{t_a \times d_k}, \tag{19}$$

where h_i^t is calculated by (10), serving as the hidden state of the i-th pedestrian at the time step t, and d_k is the dimension of each hidden state. In this work, we adopt the self-attention mechanism [58] for the temporal attention, which takes input as the whole hidden states, and outputs an attention weight vector, which can be formulated as

$$\alpha_i^t = \text{softmax}(w_2 \times tanh(W_1(H_i^t)^T)) \in \mathbb{R}^{1 \times t_a}. \tag{20}$$

Here W_1 is a weight matrix of size $d_r \times d_k$, and w_2 is a vector of length d_r, where d_r is a hyper-parameter. The parameters W_1 and w_2 are shared by all the pedestrians, and can be learned through back-propagation. The softmax function ensures that all the computed weights sum up to 1. The weight vector α_i^t adaptively emphasizes important information from the temporal domain. The final temporal attention vector is then computed as a weighted sum of previous hidden states

$$tp_i^t = \alpha_i^t H_i^t \in \mathbb{R}^{1 \times d_k}. \tag{21}$$

3.3.2 Spatial attention

The information of relative positions to other pedestrians is also a key factor to predict the future motion behaviors of pedestrians, such as keeping walking in parallel, changing direction for collision avoidance and so on. This motivates us to design another spatial attention mechanism, which makes the network emphasize more on the persons having tight connections with the target pedestrian.

Due to the fact that the number of pedestrians in a scene is dynamically changing over time, we cannot directly employ the self-attention mechanism in the spatial domain. In this work, we adopt the dot-product attention in [52] for the spatial context encoding, which is formulated as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{22}$$

where Q, K, V represent the matrices containing a set of queries, keys and values, respectively, and d is the dimension of each key vector. Please refer to [52] for more details. Let $p_i^t = (x_i^t, y_i^t)$ be the position of the i-th pedestrian at the time step t. The relative position between the i-th pedestrian and the j-the pedestrian is

$$r_{i,j}^t = p_i^t - p_j^t. \tag{23}$$

Each 2-dimensional relative position is embedded by a two-layer MLP with ReLU activation

$$\hat{r}_{i,j}^t = MLP(r_{i,j}, \Theta_{re}) \in \mathbb{R}^{d_r}, \tag{24}$$

where Θ_{re} contains the parameters of MLP. Then the spatial information of all pedestrians in a scene at the time step t can be formed as

$$D^t = \begin{bmatrix} \hat{r}_{1,1}^t & \cdots & \hat{r}_{1,N}^t \\ \vdots & \ddots & \vdots \\ \hat{r}_{N,1}^t & \cdots & \hat{r}_{N,N}^t \end{bmatrix} \in \mathbb{R}^{N \times d_r N}. \tag{25}$$

The i-th row of D^t denoted by $D_{i,:}^t$, is the spatial embedding of the i-th pedestrian, reflecting its relationship with the other pedestrians in the spatial domain. According to the dot-product attention strategy [52], we calculate the attention weight as

$$\hat{\alpha}_i^t = \text{softmax}(\frac{D_{i,:}^t (D^t)^T}{\sqrt{N}}) \in \mathbb{R}^{1 \times N}. \tag{26}$$

Intuitively, $\hat{\alpha}_i^t$ denotes the degree that the i-th pedestrian pays attention to the other pedestrians.

Combining with the temporal attention, we finally construct a unified spatial-temporal attention mechanism to

model the social interactions of pedestrians. Specifically, the spatial-temporal attention vector of the i-th pedestrian is

$$st_i^t = \hat{\alpha}_i^t \begin{bmatrix} tp_1^t \\ \vdots \\ tp_N^t \end{bmatrix} \in \mathbb{R}^{1\times d_k}. \tag{27}$$

We can observe that the attention vector st_i^t encodes both the spatial context and the temporal context.

3.3.3 Trajectory prediction with attention

Upon having the attention vector st_i^t, we use it as auxiliary information at the decoder side to help trajectory prediction. Specifically, incorporating with the attention vector st_i^t, we rewrite (11) as

$$\begin{aligned} v_i^{t_o} &= MLP([st_i^{t_o}, h_i^{t_o}]; \Theta_v) \\ h_i^{t_o} &= [v_i^{t_o}, z]. \end{aligned} \tag{28}$$

where Θ_v is the embedding weight. Besides, the second equation in (12) is modulated accordingly as

$$\begin{aligned} v_i^{t-1} &= MLP([st_i^{t-1}, h_i^{t-1}]; \Theta_v') \\ h_i^t &= LSTM(v_i^{t-1}, \hat{e}_i^{t-1}; \Theta_{ld}) \end{aligned} \tag{29}$$

Note that except (11) and (12), all the other procedures detailed in Section 3.2 remains unchanged after employing the proposed spatial-temporal attention technique.

4 Experiments

In this section, we evaluate the performance of our proposed TPNSTA, which is implemented using the PyTorch framework. All the experiments are conducted on a desktop running Ubuntu 18.04 with an NVIDIA TITAN Xp GPU. Our source code and trained models will be publicly available upon acceptance.

4.1 Implementation Details

In this subsection, we give the implementation details of our method TPNSTA. Based on a series of ablation experiments, we empirically set the number of pyramid scales as 5, and the dimensions of the hidden state for the encoder and the decoder as 32. Each input coordinate (x, y) is embedded as a 16-dimensional vector through an MLP for both the encoder and decoder. The length of the noise vector z is 8. We adopt Adam algorithm [59] to optimize the loss function (18) and train our network with the following hyper-parameter settings: batch size is 64; learning rate for the Generator is 1e-4; learning rate for the Discriminator is 1e-3; betas are 0.9 and 0.999; weight decay is 1e-4 and the number of epochs is 400.

4.2 Datasets and Metrics

Datasets: We evaluate our method on two benchmark public datasets, *i.e.*, ETH [20] and UCY [60]. Both contain videos and annotated trajectories with social interactions in real world environments. All the trajectories are converted to world coordinates. These two datasets consist of 5 video sequences: ETH, HOTEL, UNIV, ZARA1 and ZARA2 with 4 different scenes. There are totally 1536 pedestrians with thousands of trajectories containing challenging behaviors such as walking together, crossing each other, forming groups and dispersing.

Metrics: For the sake of fairness, we use the widely adopted leave-one-out approach evaluation methodology, *i.e.*, training on 4 datasets while testing on the remaining one. The number of observed time steps is 8 (3.2 seconds) of each person and the upcoming trajectory of 12 time steps (4.8 seconds) is used to predict. Following prior works [10], [11], [12], [31], we use two error metrics to evaluate the performance of different pedestrian trajectory prediction models.

1) *Average Displacement Error* (ADE): The average Euclidean distance between the ground-truth trajectory and the predicted one,

$$\text{ADE} = \frac{\sum_{i=1}^{N}\sum_{t=t_o+1}^{t_o+t_p}\left\|Y_i^{(t)} - \hat{Y}_i^{(t)}\right\|_2}{N \times t_p}. \tag{30}$$

2) *Final Displacement Error* (FDE): The Euclidean distance between the ground-truth destination and the predicted one at the final prediction time $t_o + t_p$,

$$\text{FDE} = \frac{\sum_{i=1}^{N}\left\|Y_i^{(t_o+t_p)} - \hat{Y}_i^{(t_o+t_p)}\right\|_2}{N}. \tag{31}$$

4.3 Baselines

We compare our method with following baseline approaches: ***Linear*** [10]: a linear regressor that predicts the next coordinates based on previous input points. ***S-LSTM*** [10]: a pedestrian trajectory prediction method based on LSTM and social pooling. ***S-GAN*** [11]: a model that employs GAN to generate multimodal pedestrian trajectories. It should be noted that our method TPNSTA degrades to S-GAN without the temporal pyramid module and the spatial-temporal attention module. ***S-GAN-P*** [11]: the Social-GAN with a global pooling module. Besides, we also compare our method with the most recent algorithms: ***PIF*** [7]: a multi-task method using both visual features and interaction information. ***SoPhie*** [29]: an improved GAN based model considering the physical constraints. ***SR-LSTM*** [23]: a state refinement method for LSTM based pedestrian trajectory prediction. ***STSGN*** [40]: a stochastic trajectory prediction model with social graph network. ***Social-BiGAT*** [41] and ***STGAT*** [26]: methods based on GAN and graph attention. ***Social-STGCNN*** [12]: an approach that models the social behavior of pedestrians using a graph. Similar to previous works [11], [12], [29], [40], we generate 20 samples based on the predicted distribution for the generative models, whereas producing one sample for the deterministic methods.

4.4 Quantitative Analysis

We compare our model TPNSTA with all the above methods in terms of ADE and FDE metrics. Table 1 summarizes the results of different algorithms, where the ***TPN*** denotes our method without the spatial-temporal attention mechanism. Besides the performance on each dataset, we also report the average results for each method in the last two columns. We can see that all the algorithms perform much better than the linear model as they can model more complex trajectories

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/TNSE.2021.3065019, IEEE Transactions on Network Science and Engineering

TABLE 1
ADE / FDE performance of several methods compared to TPNSTA.

Methods \ Datasets	ETH		Hotel		Univ		Zara1		Zara2		AVG	
	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
Linear [10]	1.33	2.94	0.39	0.72	0.82	1.59	0.62	1.21	0.77	1.48	0.79	1.59
S-LSTM [10]	1.09	2.35	0.79	1.76	0.67	1.40	0.47	1.00	0.56	1.17	0.72	1.54
S-GAN [11]	0.81	1.52	0.72	1.61	0.60	1.26	0.34	0.69	0.42	0.84	0.58	1.18
S-GAN-P [11]	0.87	1.62	0.67	1.37	0.76	1.52	0.35	0.68	0.42	0.84	0.61	1.21
PIF [7]	0.73	1.65	0.30	0.59	0.60	1.27	0.38	0.81	0.31	0.68	0.46	1.00
SoPhie [29]	0.70	1.43	0.76	1.67	0.54	1.24	**0.30**	0.63	0.38	0.78	0.54	1.15
STGAT [26]	0.65	1.12	0.35	0.66	0.52	1.10	0.34	0.69	0.29	0.60	0.43	0.83
SR-LSTM [23]	0.63	1.25	0.37	0.74	0.51	1.10	0.41	0.90	0.32	0.70	0.45	0.94
STSGN [40]	0.75	1.63	0.63	1.01	0.48	1.08	**0.30**	0.65	**0.26**	0.57	0.48	0.99
Social-BiGAT [41]	0.69	1.29	0.49	1.01	0.55	1.32	**0.30**	0.62	0.36	0.75	0.48	1.00
Social-STGCNN [12]	0.64	1.11	0.49	0.85	**0.44**	**0.79**	0.34	**0.53**	0.30	**0.48**	0.44	0.75
TPN	0.53	0.89	0.23	0.43	0.54	1.13	0.33	0.68	**0.26**	0.55	0.39	0.74
TPNSTA	**0.51**	**0.87**	**0.22**	**0.39**	0.52	1.09	0.34	0.68	**0.26**	0.54	**0.37**	**0.71**

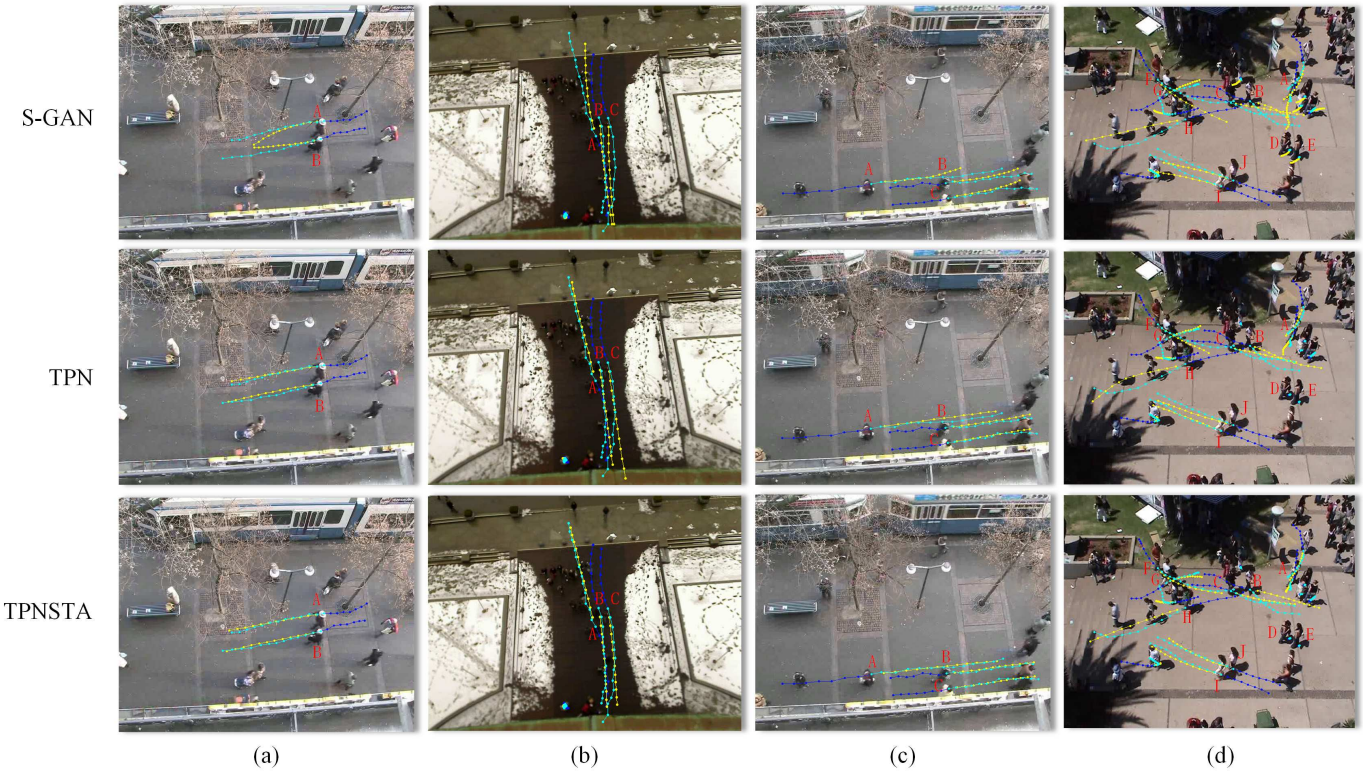


Fig. 5. Examples of predicted trajectories by different methods in four distinct scenarios. (a) walk in parallel; (b) meet from opposite directions; (c) follow people and (d) walk with complex social interactions. Blue line represents the historical trajectory; green line denotes the true future trajectory; yellow line shows the predicted future trajectory, and dots are the locations of pedestrians at different time steps. Better viewed in color.

and social interactions. Based on the results, we further draw the following conclusions:

- Overall, our method TPNSTA outperforms all the previous approaches in terms of the average ADE and FDE.
- Compared with the baseline approach S-GAN [11], TPNSTA achieves significant performance gains on ADE and FDE metrics. For example, S-GAN has an average error of 0.58 on ADE, and 1.18 on FDE, while our method TPNSTA has much lower ADE (0.37) and FDE (0.71), corresponding to 36% and 40% relative improvements, respectively. This demonstrates that our proposed temporal pyramid network and spatial-temporal attention mechanism indeed help for the pedestrian trajectory prediction.
- For the previous state-of-the-art method Social-STGCNN [12], TPNSTA still achieves noticeable performance gains. For instance, Social-STGCNN is with

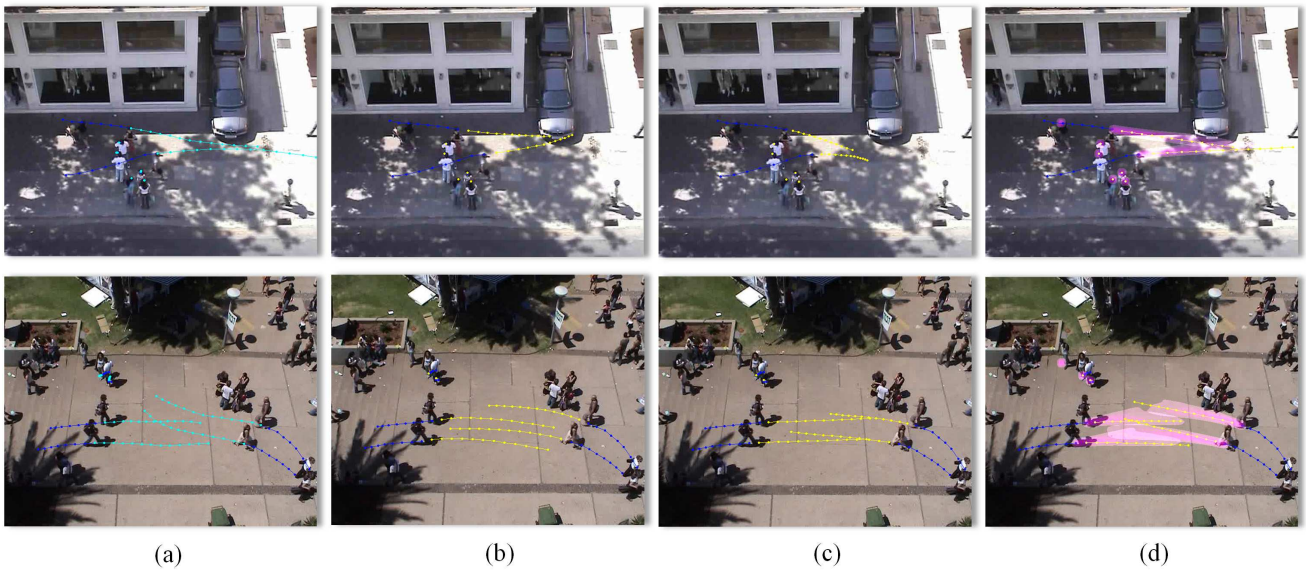


Fig. 6. Examples of diverse predictions. Blue line represents pedestrian historical trajectory, and yellow line denotes the predicted trajectories. (a) the ground-truth future trajectory, which is marked by the green line. (b, c) two examples of diverse predictions. (d) The density of the predicted trajectory, where the purple area is the visualization result of the predicted 20 pedestrian trajectories after mean filtering. Better viewed in color.

an error of 0.44 on ADE metric, while TPNSTA has an error of 0.37 under the same metric, decreasing the error over 16%. As for FDE metric, TPNSTA is also better than Social-STGCNN by over 5%.

- Even without using any side information, TPNSTA outperforms those methods utilizing the vision signal that contains scene context by a big margin, such as PIF [7], Sophie [29] and Social-BiGAT [41]. This implies that the performance of TPNSTA could potentially be further improved by considering the scene context.

4.5 Qualitative Analysis

In this subsection, we provide some examples to show how our TPNSTA successfully captures both the long-range and short-range motion behaviors of pedestrians, and how it fully exploits the social interactions between pedestrians. We qualitatively compare the prediction results between Social-GAN, our proposed method TPN without attention mechanism, and our proposed method TPNSTA with spatial-temporal attention.

4.5.1 Results in different interaction scenarios

We visualize examples from 4 scenarios in Figure 5, including walking in parallel, meeting from different directions, following people and walking with complex social interactions.

Walking in parallel When people are walking side by side, they usually have tight connection to each other, and their relative positions tend to be preserved and motion behaviors tend to change consistently during the next time steps. In the first scenario in Fig. 5, two target pedestrians A and B are walking in parallel. It can be noticed from Fig. 5(a) that S-GAN incorrectly predicts that these two pedestrians will walk across each other, and have a high possibility of collision. The predicted trajectories of persons A and B by S-GAN have a large deviation from the ground-truth trajectories marked by green lines. Compared with S-GAN, the predictions by our proposed TPN and TPNSTA all show that these two pedestrians will keep walking in parallel, and the predicted trajectories by TPN and TPNSTA have much smaller deviation from the ground-truth trajectories. This demonstrates the superiority of modeling motion behavior at various tempos using the proposed temporal pyramid network. Furthermore, with the mechanism of the spatial-temporal attention, TPNSTA performs better in maintaining a consistent behavior when people are walking side by side. This can be verified by the truth that the predicted trajectory by TPNSTA more closely matches with the ground-truth one than that of TPN.

Meeting from opposite directions People avoiding each other when moving in opposite direction is common in reality. Fig. 5(b) presents a scenario where two groups are meeting from opposite directions. We can see that the short-range behaviors of persons A, B and C are adjusted slightly to avoid collision. Compared with S-GAN, the trajectory of the person A predicted by TPN and TPNSTA is more accurate after meeting. Further, TPN and TPNSTA successfully predict that persons B and C will keep walking in parallel, while S-GAN forecasts that they will merge after a few time steps, which deviates from their true behaviors.

Following people When a person is following someone, he or she might want to draw attention to the person ahead, and maintain a safe distance between them. Fig. 5(c) shows a situation where the person A is walking behind the person B, and persons B and C are walking in parallel. We can see that S-GAN tends to decrease the speed of person A even when the distance between others is sufficiently large, making the predicted destination far behind the ground-truth one. Comparing with S-GAN, our proposed models TPN and TPNSTA can more accurately forecast the speed at each time step, and still preserve a safe distance to avoid collision.

Walking with complex social interactions A more com-

TABLE 2
The average ADE and FDE performance of variants.

Models \ Modules	TP Layer	Temporal Attention	Spatial Attention	AVG (ADE/FDE)
S-GAN	×	×	×	0.58/1.18
TPN	✓	×	×	0.39/0.74
TPNTA	✓	✓	×	0.37/0.73
TPNSTA	✓	✓	✓	**0.37/0.71**

plex scenario is shown in Fig. 5(d), where many groups and individuals are walking in a crowd. The complex interactions drive people using various ways to avoid collision while continuing towards their destinations. It can be seen that many generated trajectories by S-GAN have large deviations from the ground-truth ones, e.g., the persons A, F, H and I. Besides, we notice that S-GAN fails to adjust the behaviors of persons A, B and C, and then causes a collision at the end of the predicted trajectories. As for the methods TPN and TPNSTA, the predicted trajectories match better with the ground-truth trajectories. For instance, persons I and J are maintained walking in parallel; persons D and E are still standing around chatting. Furthermore, we observe that the speed of person A is clearly slowed down by TPN and TPNSTA, to avoid collision with persons B and C. For the trajectories of persons F and H, TPNSTA achieves much better prediction accuracy than TPN. This indicates the importance of our proposed spatial-temporal attention mechanism in scenarios of complex social interactions.

4.5.2 Results of diverse predictions

GAN based architecture permits to capture the distribution of the future trajectory, then our model is capable of producing multiple plausible and diverse trajectories conforming to the multimodal behavior of pedestrians. In Fig. 6, we show some examples of diverse predictions by sampling the noise vector z from the standard normal distribution. We can see from Fig. 6(b) and Fig. 6(c) that our model generates two socially acceptable and distinct trajectories with different z, including changing the direction and speed. For instance, the top image of Fig. 6(b) shows that the person is walking toward the car, where the direction is different from the true trajectory but the predicted path is still acceptable. Similar phenomenon can also be observed from the bottom image of Fig. 6(b). Besides, images presented in Fig. 6(c) show that z can also affect the speed of pedestrians. In Fig. 6(d), we draw the density of the predicted trajectory by 20 randomly generated samples then conducting mean filtering operation. The purple area constructs a plausible area that each pedestrian may pass. The position of darker color indicates a higher probability that the person will pass through. Furthermore, in Fig. 6(d), we also plot the best predicted trajectory from 20 samples for each scenario, and we can see that it closely matches with the true trajectory shown in Fig. 6(a).

4.6 Ablation Experiments

In this subsection, we conduct ablation experiments to investigate how our proposed temporal pyramid network and the spatial-temporal attention module impact the trajectory prediction.

4.6.1 Variants of our architecture

In Table 2, we systematically evaluate our method through a series of ablation experiments, where we consider the following variants of our method:
S-GAN: the method without the temporal pyramid module and the spatial-temporal attention module. With this setting, our method degrades to S-GAN;
TPN: the method only considers the temporal pyramid module without the attention mechanism;
TPNTA: the method considers the temporal pyramid module and the temporal attention;
TPNSTA: the method considers the temporal pyramid module and the spatial-temporal attention.

From the results reported in Table 2, we have the following conclusions. First, benefiting from the temporal pyramid architecture, our model can effectively model the global context and the local context of trajectories. Comparing TPN with S-GAN, we can see that TPN significantly reduces the ADE from 0.58 to 0.39, and the FDE from 1.18 to 0.74, which indicates that modeling the trajectory with multi-resolution in the temporal domain can more effectively capture the motion behaviors. Second, temporal information is an important factor to affect the future motion behaviors. After incorporating the temporal attention module, we observe that TPNTA can further improve the prediction accuracy in terms of both the ADE and the FDE metrics. Third, the best results are achieved by TPNSTA, which adopts both the temporal pyramid module and spatial-temporal attention. This demonstrates that both spatial context and temporal context are useful for trajectory modeling.

4.6.2 Spatial-temporal attention analysis

We further conduct experiments to find out what have been learned by the spatial-temporal attention module, and how they can help in trajectory prediction. Fig. 7 visualizes the spatial-temporal attention weights in different interaction scenarios, including 'merge', 'meet' and 'parallel walk'. Fig. 7(a) presents the temporal attention weights over previous 8 time steps and Fig. 7(b) gives the spatial attention weights at the current time. Fig. 7(c) shows the scene and our prediction results.

Fig. 7 indicates that our proposed spatial-temporal attention can effectively exploit the information from both temporal and spatial domains. Specifically, from Fig. 7(a), we can notice that the weights of the temporal attention tend to be small for the time far away from the predicting time. This phenomenon obeys our intuition: the most recent motion behavior of people should play a more important role for our navigation. Furthermore, we can also see that the temporal information modeled by our method is dominated by the contexts of the last two time steps, which implies that future trajectory of the target pedestrian is mainly influenced by a short-range previous motion behavior (about 0.8 second) of himself/herself and other pedestrians. As for the weights of the spatial attention presented in Fig. 7(b), it can be easily figured out that the largest attention weights always locate on the diagonal. This result, combined with the above temporal attention weights analysis, reveals that

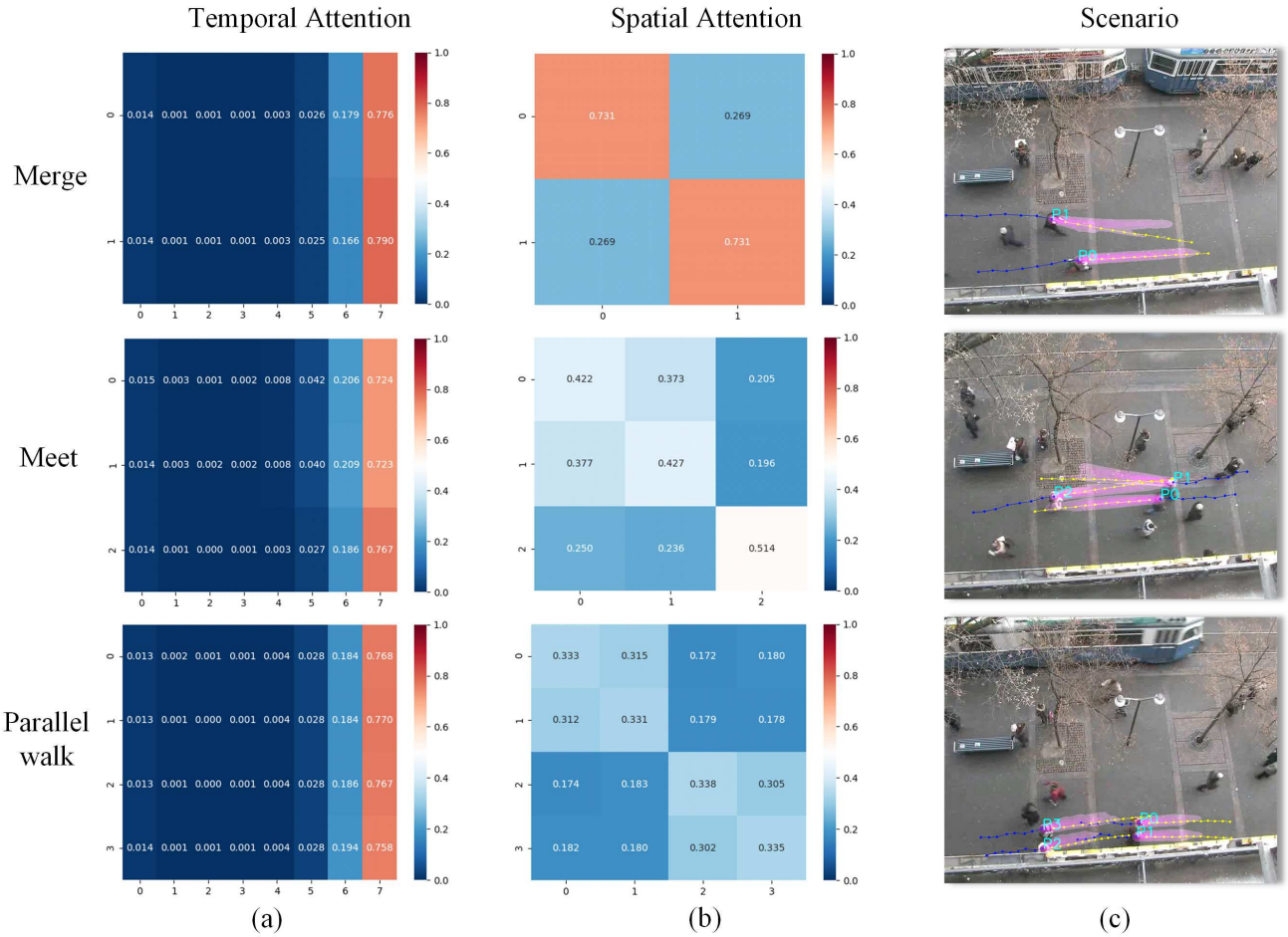


Fig. 7. Examples of spatial-temporal attention weights in different interaction scenarios. Rows represent different interaction scenarios. The first column shows the heatmap of the temporal attention weights for previous 8 time steps; the second column draws the heatmap of the spatial attention weights, and the third column visualizes the corresponding scene and the predicted results of our model. The pedestrian historical trajectory is marked in blue, while the best predicted future trajectory is marked in yellow. Purple areas show the visualization results of the predicted 20 pedestrian trajectories after mean filtering.

the future trajectory of the target pedestrian largely depends on the motion status of itself. We also notice that the off-diagonal weights have high correlations with the relative positions between pedestrians. Specifically, pedestrians are inclined to have large attention weights when they are close in the scene. This phenomenon is desirable, since large attention weights between adjacent pedestrians can help avoid collision when they are going to merge, or maintain consistent behaviors when they are walking side by side.

5 Conclusion

In this paper, we have proposed a novel pyramid architecture for pedestrian trajectory prediction, which outperforms state-of-the-art methods on two benchmark datasets. First, we have devised a temporal pyramid network through squeeze and dilation modulations, which encodes and decodes the trajectory at multiple resolutions. This enables our method to capture both short-range and long-range motion behaviors of pedestrians. By resorting to a coarse-to-fine fusion strategy and the multi-supervision, our method can progressively merge high-scale global context with low-scale local context, finally resulting in an accurate trajectory prediction. Second, we have further introduced an attention mechanism to model the impact of social interactions between pedestrians. Our attention technique is conducted in both the spatial and temporal domains, which is more intuitive and effective compared with previous methods. Finally, with a GAN based framework, our method can generate multiple socially-acceptable trajectories conditioned on the same historical trajectory, obeying the multimodal property of pedestrians. Both quantitative and qualitative experimental results demonstrate the promising performance of our method under various situations.

References

[1] P. T. Szemes, H. Hashimoto, and P. Korondi, "Pedestrian-behavior-based mobile agent control in intelligent space," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 6, pp. 2250–2257, 2005.

[2] C. Ruch, J. Gachter, J. Hakenberg, and E. Frazzoli, "The +1 method: Model-free adaptive repositioning policies for robotic multi-agent systems," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–1, 2020.

[3] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric bayesian activity mining and analysis from surveillance video," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2089–2102, 2016.

[4] L. Lin, Y. Lu, Y. Pan, and X. Chen, "Integrating graph partitioning and matching for trajectory analysis in video surveillance," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4844–4857, 2012.

[5] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, 2009.

[6] M. Wu, H. Ling, N. Bi, S. Gao, Q. Hu, H. Sheng, and J. Yu, "Visual tracking with multiview trajectory prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 8355–8367, 2020.

[7] 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 Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 5725–5734.

[8] X. Liu, H. Song, and A. Liu, "Intelligent uavs trajectory optimization from space-time for data collection in social networks," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–1, 2020.

[9] R. Liang, Y. Li, X. Li, yi tang, J. Zhou, and W. Zou, "Temporal pyramid network for pedestrian trajectory prediction with multi-supervision," *arXiv preprint arXiv: 2012.01884*, 2020.

[10] 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 Recogn.*, 2016, pp. 961–971.

[11] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 2255–2264.

[12] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 14 424–14 432.

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

[14] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Trans. Graphics*, vol. 25, no. 3, pp. 1160–1168, 2006.

[15] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transportation Research Part B: Methodological*, vol. 40, no. 8, pp. 667–687, 2006.

[16] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2007.

[17] M. K. C. Tay and C. Laugier, "Modelling smooth paths using gaussian processes," in *Field and Service Robotics*. Springer, 2008, pp. 381–390.

[18] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 2203–2210.

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

[20] 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 Int. Conf. Comput. Vis.*, 2009, pp. 261–268.

[21] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neu. netw*, vol. 108, pp. 466–478, 2018.

[22] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 5275–5284.

[23] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 12 085–12 094.

[24] A. Al-Molegi, M. Jabreel, and B. Ghaleb, "STF-RNN: Space time features-based recurrent neural network for predicting people next location," in *Proc. IEEE Sym. Series Comput. Intel.*, 2016, pp. 1–7.

[25] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE int. Conf. Robot. and Auto.*, 2018, pp. 1–7.

[26] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6272–6281.

[27] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2375–2384.

[28] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Art. Intel.*, vol. 33, 2019, pp. 6120–6127.

[29] 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 Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 1349–1358.

[30] T. van der Heiden, N. S. Nagaraja, C. Weiss, and E. Gavves, "Safecritic: Collision-aware trajectory prediction," *arXiv preprint arXiv:1910.06673*, 2019.

[31] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 12 126–12 134.

[32] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *Proc. IEEE Int. Conf. Intel. Robots and Sys.*, 2019.

[33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Comp. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2005, pp. 886–893.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017, pp. 2117–2125.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[37] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2980–2988.

[38] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. conf. mach. learn.*, 2014, pp. 1764–1772.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 3156–3164.

[40] L. Zhang, Q. She, and P. Guo, "Stochastic trajectory prediction with social graph network," *arXiv preprint arXiv:1907.10233*, 2019.

[41] V. Kosaraju, A. Sadeghian, R. Martn-Martn, I. Reid, S. H. Rezatofighi, and S. Savarese, "Social-BiGAT: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *arXiv preprint arXiv:1907.03395*, 2019.

[42] 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.

[43] Y. Li, J. Zhou, and A. Cheng, "SIFT keypoint removal via directed graph construction for color images," *IEEE Trans. on Inf. Forensics and Security*, vol. 12, no. 12, pp. 2971–2985, 2017.

[44] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang, and H. Liu, "Feature pyramid reconfiguration with consistent loss for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5041–5051, 2019.

[45] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, 2018.

[46] Y. Li and J. Zhou, "Fast and effective image copy-move forgery detection via hierarchical feature point matching," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1307–1322, 2019.

[47] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 591–600.

[48] Y. Huang, X. Cao, X. Zhen, and J. Han, "Attentive temporal pyramid network for dynamic scene classification," in *Proc. AAAI Conf. Art. Intel.*, vol. 33, 2019, pp. 8497–8504.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[50] S. D. Roy, G. Lotan, and W. Zeng, "The attention automaton: sensing collective user interests in social network communities," *IEEE Trans. Netw. Sci. Eng.*, vol. 2, no. 1, pp. 40–52, 2015.

[51] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[53] 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.

[54] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE Trans. Image Process.*, vol. 29, pp. 7615–7628, 2020.
[55] J. Amirian, J.-B. Hayet, and J. Pettré, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, 2019.
[56] Y. Li, J. Zhou, X. Zheng, J. Tian, and Y. Y. Tang, "Robust subspace clustering with independent and piecewise identically distributed noise modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 8720–8729.
[57] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[58] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[60] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer graphics forum*, vol. 26, no. 3, pp. 655–664, 2007.

Yuanman Li (Member, IEEE) received the B.Eng. degree in software engineering from Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree in computer science from University of Macau, Macau, 2018. From 2018 to 2019, he was a Post-doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. He is currently an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include data representation, multimedia security and forensics, computer vision and machine learning.

Rongqin Liang (Student Member, IEEE) received the B.Eng. degree in communication engineering from Wuyi University, Guangdong, China, in 2018. He is currently a Postgraduate student at the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, and will continue to study for a Ph.D. degree. His current research interests include computer vision and deep learning.

Wei Wei (Senior Member, IEEE) is an Associate Professor in the School of Computer Science and Engineering at the Xi'an University of Technology in China. He has received his Ph.D from Xian Jiaotong University. His research interests include Internet of Things, wireless sensor networks, image processing, mobile computing, distributed computing, pervasive computing, smart city, artificial intelligence, sensor data clouds, etc. He has over 200 papers published or accepted by international conferences and journals (e.g., IEEE Transactions on Services Computing, IEEE Transactions on Industrial Informatics, IEEE Transactions on Computational Social Systems, IEEE Communications Magazine, IEEE Transactions on Parallel and Distributed Systems, IEEE Internet of Things). He is a TPC member of many conferences and regular reviewer of IEEE Communications Magazine, IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Image Processing, IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, IEEE Transactions on Industrial Informatics, IEEE Transactions on Industrial Electronics.

Wei Wang (Member, IEEE) is currently an Associate Professor with School of Intelligent Systems Engineering, Sun Yat-sen University, China. Before this, he had been the UM Macao Research Fellow at University of Macau, Macau SAR. He received PhD degree in software engineering from Dalian University of Technology in 2018. His research interests include computational social science, data mining, internet of things, and artificial intelligence. He has authored/co-authored over 50 scientific papers in international journals and conferences, e.g., IEEE Transactions on Computational Social Systems, IEEE Internet of Things, IEEE Transactions on Industrial Informatics, IEEE Transactions on Big Data, IEEE Transactions of Emerging Topics in Computing, IEEE Transactions on Human-Machine Systems, The Web Conference, etc. He is the Leading Guest Editor of International Journal of Distributed Sensor Networks, Computer Communication, and Wireless Communication & Mobile Computing. He is the guest editor of ACM Transactions on Internet Technology, IEEE Journal of Biomedical and Health Informatics, and IEEE Sensor Journal. He is a PC member of International Conference on Smart Internet of Things 2019 and regular reviewer of IEEE Communications Magazine, Future Generation Computer Systems, IEEE Transactions on Industrial Informatics, IEEE Transactions on Big Data, and IEEE Transactions of Emerging Topics in Computing.

Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, in 2009. He held various research positions with the University of Illinois at Urbana-Champaign, The Hong Kong University of Science and Technology, and the McMaster University. He is currently an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence and big data. He holds four granted U.S. patents and two granted Chinese patents. He has coauthored two articles that received the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Xia Li (Member, IEEE) received her B.S. and M.S. in electronic engineering and SIP (signal and information processing) from Xidian University in 1989 and 1992 respectively. She was later conferred a Ph.D. in Department of information engineering by the Chinese University of Hong Kong in 1997. Currently, she is a member of the Guangdong Key Laboratory of Intelligent Information Processing. Her research interests include intelligent computing and its applications, image processing and pattern recognition.