Social Interpretable Tree for Pedestrian Trajectory Prediction

Liushuai Shi[1], **Le Wang**[2*], **Chengjiang Long**[3]
Sanping Zhou[2], **Fang Zheng**[1], **Nanning Zheng**[2], **Gang Hua**[4]
[1]School of Software Engineering, Xi’an Jiaotong University
[2]Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University
[3]JD Finance America Corporation, [4]Wormpex AI Research
{shiliushuai,zhengfang}@stu.xjtu.edu.cn, {lewang,spzhou,nnzheng}@xjtu.edu.cn, {cjfykx, ganghua}@gmail.com

Abstract

Understanding the multiple socially-acceptable future behaviors is an essential task for many vision applications. In this paper, we propose a tree-based method, termed as Social Interpretable Tree (SIT), to address this multi-modal prediction task, where a *hand-crafted tree* is built depending on the prior information of observed trajectory to model multiple future trajectories. Specifically, a path in the tree from the root to leaf represents an individual possible future trajectory. SIT employs a coarse-to-fine optimization strategy, in which the tree is first built by high-order velocity to balance the complexity and coverage of the tree and then optimized greedily to encourage multimodality. Finally, a teacher-forcing refining operation is used to predict the final fine trajectory. Compared with prior methods which leverage implicit latent variables to represent possible future trajectories, the path in the tree can explicitly explain the rough moving behaviors (*e.g.*, go straight and then turn right), and thus provides better interpretability. Despite the hand-crafted tree, the experimental results on ETH-UCY and Stanford Drone datasets demonstrate that our method is capable of matching or exceeding the performance of state-of-the-art methods. Interestingly, the experiments show that the raw built tree without training outperforms many prior deep neural network based approaches. Meanwhile, our method presents sufficient flexibility in long-term prediction and different best-of-K predictions. *Code:* https://github.com/shuaishiliu/SIT

Introduction

Pedestrian trajectory prediction plays an essential role in many vision systems, *e.g.*, the automatic vehicle understands the future trajectory of the pedestrian to prevent the accident, and the monitoring system recognize the abnormal action in advance by predicting the future trajectory of human.

In a real traffic scenario illustrated in Figure 1 (first row), due to the intrinsic randomness of pedestrians’ moving and intangible various intent only based on observed trajectory (A), the future trajectory is largely uncertain and naturally multi-modal, which means there are multiple possible trajectories that pedestrian could take (B). One kind of approaches (Gupta et al. 2018; Mangalam et al. 2020) to model this multi-modal future trajectory embed them into an implicit latent space (second row) generated by the conditional variational autoencoder (CVAE) or generative adversarial network (GAN). Then, multiple latent variables sampled repeatedly from the generated latent space are used to represent multimodality. Despite the significant performance, the latent variables still suffer from uninterpretability, and such models (CVAE, GAN) are persecuted by the problem of model collapse (Arjovsky and Bottou 2017). What’s more, the sampling operation could result in performance variance due to the disturbance from randomness.

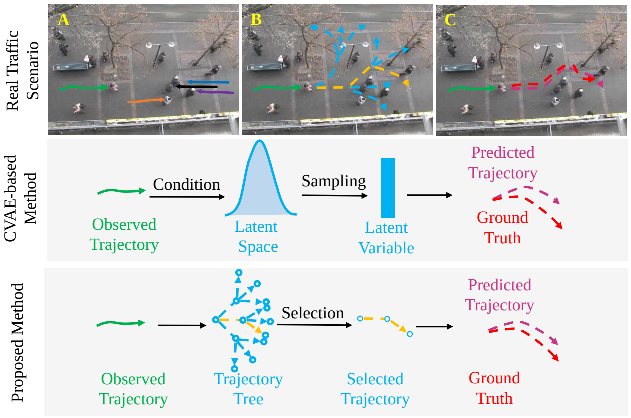

Figure 1: **Illustration of real traffic scenario and comparison of our method with CVAE-based method**. The first row illustrates the future trajectory is largely uncertain only referencing observed trajectory (A). The future trajectory is multi-modal and thus can be represented by a tree (B). The yellow closet path is refined to predict future trajectory (C). The second and third rows describe the process of CVAE-based method and our proposed method, respectively.

To cope with those problems, we propose to model the multi-modal future trajectory into a tree as shown in Figure 1 (third row), where the paths from the root to leaf in the tree could represent multimodality naturally and the closet path (yellow path) with ground truth is selected to obtain the final fine-grained predicted trajectory. Compared with the latent variable, the path in the tree could explain rough movement behaviors, *e.g.*, the yellow path expresses go straight and then turn right, and thus can provide well interpretabil-

[*]Corresponding author.

ity. Furthermore, the sampling operation is replaced with the "selection" to ensure obtaining stable results.

Inspired by the interpretable tree, we propose the Social Interpretable Tree (SIT) to predict multi-modal future trajectories. SIT first builds a future trajectory tree to generate plausible future trajectories according to the velocity of the observed trajectory. To obtain concise representation, the tree is specified with a ternary tree, in which the tree splits in three directions, *i.e.*, go straight, turn left and right with a specific angle, at each time step. The turn round and keep still can be viewed in the specific cases of go straight and turn left or right, respectively. Since the complexity of the tree grows exponentially as the depth increases, we propose to build a coarse trajectory tree (CTT) to balance the complexity and coverage of the tree. Instead of splitting time step by time step, the CTT splits through multi-time steps recursively and the high-order velocity in this temporal interval of observed trajectory is considered as the split direction. After obtaining the CTT, SIT optimizes it greedily to prevent the tree from collapsing to the average modal of data because there is only a single ground truth to refer to. Particularly, we convert the ground truth to coarse ground truth by high-order velocity and the generated coarse ground truth is used to optimize the closet path to it in the CTT. Finally, a teacher-forcing refining strategy is used to refine the top-1 coarse future trajectory scored and selected from the optimized CTT in training time, while the top-K coarse future trajectories are selected to obtain the final multi-modal future trajectories in inference time.

We conduct extensive experiments on two popular benchmark pedestrian trajectory prediction datasets, *i.e.*, ETH-UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007) and Stanford Drone (Robicquet et al. 2016). Despite the hand-crafted tree, the experimental results demonstrate that: 1) the proposed SIT is capable of matching or exceeding the performance of state-of-the-art methods; 2) SIT contributes to breaking the stereotype of hand-crafted methods in pedestrian trajectory prediction. Without any training, the raw ternary tree can outperform many deep neural-based methods; 3) SIT shows effective interpretability to explain pedestrians' future moving behaviors; 4) SIT shows the sufficient flexibility in long-term prediction and different best-of-K predictions.

Related Work

Pedestrian Trajectory Prediction. Traditionally, pedestrian trajectory prediction has been studied by hand-crafted methods. Since the trajectory space of pedestrians is a 2D plane, many works (Antonini, Bierlaire, and Weber 2006; Robin et al. 2009; Ondřej et al. 2010) split this space into multiple subspaces and then calculate the probability of each subspace based on the strong prior information. Relying on prior knowledge, the hand-crafted methods show interpretability to explain the predicted trajectory. Unfortunately, they are restricted in specific scenarios and difficultly generalize to more complex scenarios due to over-depending on prior knowledge.

Recently, deep learning has been applied to visual recognition (Hu, Long, and Xiao 2021), action recognition (Islam, Long, and Radke 2021), image denoising (Yu et al. 2021), style transfer (Xu et al. 2021), shadow removal (Wei et al. 2019; Zhang et al. 2020; Chen et al. 2021), anomaly detection (Liu et al. 2021), human motion prediction (Dang et al. 2021), as well as image and video forgery detection research (Islam et al. 2020). Thanks to deep learning, pedestrian trajectory prediction achieves significant progress. As a temporal sequential learning task, many works (Alahi et al. 2016; Gupta et al. 2018; Bisagno, Zhang, and Conci 2018; Zhang et al. 2019) employ the recurrent neural networks (RNNs) or its variants (LSTM and GRU) to capture temporal dependencies and spatial interaction. Considering the interaction as a spatial graph (Sun, Jiang, and Lu 2020; Kosaraju et al. 2019; Ma et al. 2019; Mohamed et al. 2020; Shi et al. 2021; Bae and Jeon 2021), the graph convolutional network (GCN) (Kipf and Welling 2017) with a physical adjacency matrix and the attention mechanism (Vaswani et al. 2017) with a learnable adjacency matrix are used to integrate spatial interactive messages. Moreover, some works (Ivanovic and Pavone 2019; Sadeghian et al. 2019; Liang et al. 2019; Shafiee, Padir, and Elhamifar 2021) leverage the visual information to improve the prediction performance.

Due to the multimodality of future trajectory, most works focus on the generative model to predict multi-modal future trajectories. The CVAE-based methods (Lee et al. 2017; Ivanovic and Pavone 2019; Mangalam et al. 2020) and the GAN-based methods (Gupta et al. 2018; Sadeghian et al. 2019) map each possible future trajectory into a latent space in training time, and sample repeatedly from the latent space to obtain the multi-modal results in inference time. In contrast, our method builds a trajectory tree on more general rules (*i.e.*, go straight, turn left and right) to represent multi-modal future trajectories and then optimizes it to obtain the fine-grained predicted trajectory. Thus, it is not only suitable for various scenarios, but also can provide better interpretability, stable predicted results, and sufficient flexibility in different prediction settings, verified by the experimental results.

Tree in Trajectory. Tree-based algorithms in trajectory related tasks mainly focus on path planning (Svenstrup, Bak, and Andersen 2010), which aims to search an acceptable path to the given destination. LaValle *et al.* (LaValle et al. 1998) propose a typical sampling-based planning approach, which extends non-holonomic constraints and supports dynamic environments as well. Followed by that, many variants (Kuffner and LaValle 2000; Adiyatov and Varol 2013; Goerzen, Kong, and Mettler 2010) are proposed to improve the performance of path planning. There are also tree-based trajectory prediction methods that serve motion and path planning. Aoude *et al.* (Aoude et al. 2011) combine the closed-loop rapidly-exploring random tree (CL-RRT) (Kuwata et al. 2009) with a Gaussian mixture model for collision avoidance and conflict detection. Jurgenson *et al.* (Jurgenson, Groshev, and Tamar 2019) divide a path into multiple sub-goals and use a divide-and-conquer process to generate a complete trajectory. In contrast, pedestrian trajectory prediction is more challenging due to the absence of any future trajectory information.

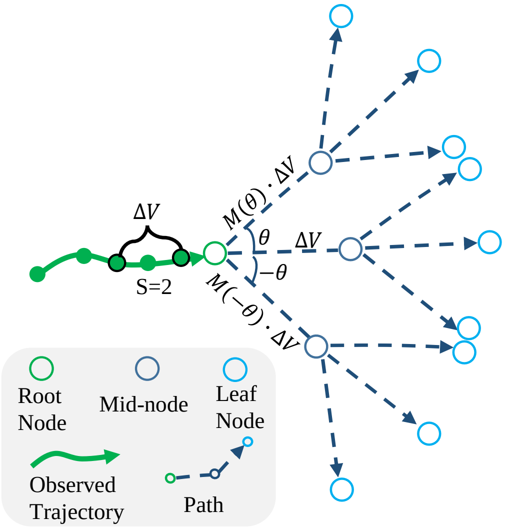

Figure 2: **An example of the generating of coarse trajectory tree with the depth $d = 2$ and interval $S = 2$.** The high-order velocity ΔV, *i.e.*, the displacement in a temporal interval S, is regarded as the forward split direction. The left split direction is gained by multiplication between the rotation matrix M with angle θ and ΔV, while the right split direction is obtained by multiplication between the rotation matrix M with angle $-\theta$ and ΔV. The coarse trajectory tree is generated recursively at each split, where each path from the root to leaf represents a coarse possible future trajectory.

Our Method

Problem Formulation

Given a traffic scenario, x_i^t represents the spatial coordinate of pedestrian i at the time step t. To collect N pedestrians' coordinates from time step 1 to T_{obs}, we can obtain the observed trajectories denoted as $\boldsymbol{X} = \{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i = \{x_i^t,\}_{t=1}^{T_{\text{obs}}}$. Due to the multimodality of future trajectory, there are K socially-acceptable future trajectories denoted by $\mathbf{Y} = \{\boldsymbol{Y}_j\}_{j=1}^K$. The single ground truth $\hat{\boldsymbol{Y}} = \{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i = \{x_i^t\}_{t=T_{\text{obs+1}}}^{T_{\text{pred}}}$, $\hat{\boldsymbol{Y}} \in \mathbf{Y}$. In a real traffic scenario, the trajectory is not only affected by the pedestrians' intention but also the interaction between pedestrians at each time step t denoted by $\boldsymbol{S} = \{s_t\}_{t=1}^{T_{obs}}$. Briefly speaking, our objective has two parts. First, the model predicts all socially-acceptable future trajectories $\mathbf{Y}$ based on the observed trajectory $\boldsymbol{X}$ and interaction $\boldsymbol{S}$, and then selects the trajectories with high confidence to obtain the final multi-modal future trajectories.

As discussed above, the process of our method can be formulated mathematically as

$$p(\hat{\boldsymbol{Y}}|\boldsymbol{X},\boldsymbol{S}) = \sum_{\boldsymbol{Y}\in\mathbf{Y}} p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{S})p(\hat{\boldsymbol{Y}}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S}), \tag{1}$$

where $p(.|.)$ is a discrete conditional distribution because $\mathbf{Y}$ is represented by a tree.

Under this formulation, previous works embed $\mathbf{Y}$ into an implicit continuous space (Gupta et al. 2018; Mangalam et al. 2020) by specific generative models, where the selecting process is directly replaced with the sampling repeatedly from the learned latent space in reference time. In contrast, our method embeds $\mathbf{Y}$ into a discrete structured space, which is specifically represented by a ternary tree. Since the ternary tree is not influenced by the average modal of data, each trajectory (path) contained in the tree can keep its individual moving behavior, and thus provide well interpretability and meanwhile not fall into the frequent modal. Moreover, the operation of selection could generate stable predicted results compared with sampling repeatedly.

The overall framework of our method is illustrated in Figure 3. Specifically, the coarse trajectory tree is built firstly to generate the coarse discrete structured space $\mathbf{Y}_{\text{coarse}}$ and then encoded by an MLP to gain the tree encoding. Meanwhile, the observed trajectory and spatial interaction are encoded to obtain the observed encoding and interaction encoding one after another. Next, the tree and interaction are fused to score each path in the coarse trajectory tree by an attention mechanism, and then the obtained confidence vector $\mathbf{p}$ is optimized supervised by the label $\mathbf{q}$, which is obtained by the path measurement between each path and the coarse ground truth. Particularly, the coarse ground truth is generated by the high-order velocity of ground truth. Subsequently, the $\mathbf{Y}_{\text{coarse}}$ is optimized greedily by the path with the highest confidence to generate the coarse predicted trajectory supervised by the coarse ground truth. Finally, a refining operation is employed on the coarse predicted trajectory to gain the fine-grained trajectory with the teacher forcing, which means the refining operation uses coarse ground truth for refinement in training time, while the coarse predicted trajectories with top-k confidences are used to refine for multi-modal future trajectory prediction in reference time.

Trajectory Prediction with Tree

Coarse Trajectory Tree. The fundamental operation for our method is to build the coarse trajectory tree, which refers to the generating of a ternary tree as preceding discussion. The whole process is considered as a recursive split in three directions (ternary tree) at each time step. Due to the temporal dependency of trajectory, the velocity vector of the observed trajectory is used to get the direction of forwarding split (go straight). In particular, the directions of left split (turn left) and right split (turn right) are gained by positive and negative rotation of the velocity vector with a specific angle, respectively. As shown in Figure 2, given the location of observed trajectories $\mathbf{x}_i$ (green arrowed line) for the pedestrian i, we can obtain the corresponding velocities denoted $\mathbf{v}_i = \{v_i^t\}_{t=1}^{T_{\text{obs}}}$, by the displacement from one time step to next time step. Note that we assume the pedestrian keeps still at the first time step, namely the $\{v_i^1\}_{i=1}^N = \mathbf{0}$. Since the complexity of tree grows exponentially as the depth (d) increases, *e.g.*, assuming the predicted length $T = T_{\text{pred}} - T_{\text{obs}} = 12$, we will generate a ternary tree with the $d = 12$ and it has 3^{12} paths if the split is taken at each time step. To balance the complexity and the coverage of the tree, the tree splits multi-time steps once instead of split time step by time step. Therefore, we set a specific temporal interval ($1 \le S \le T$) and the high-order velocity (ΔV), gained by summing all

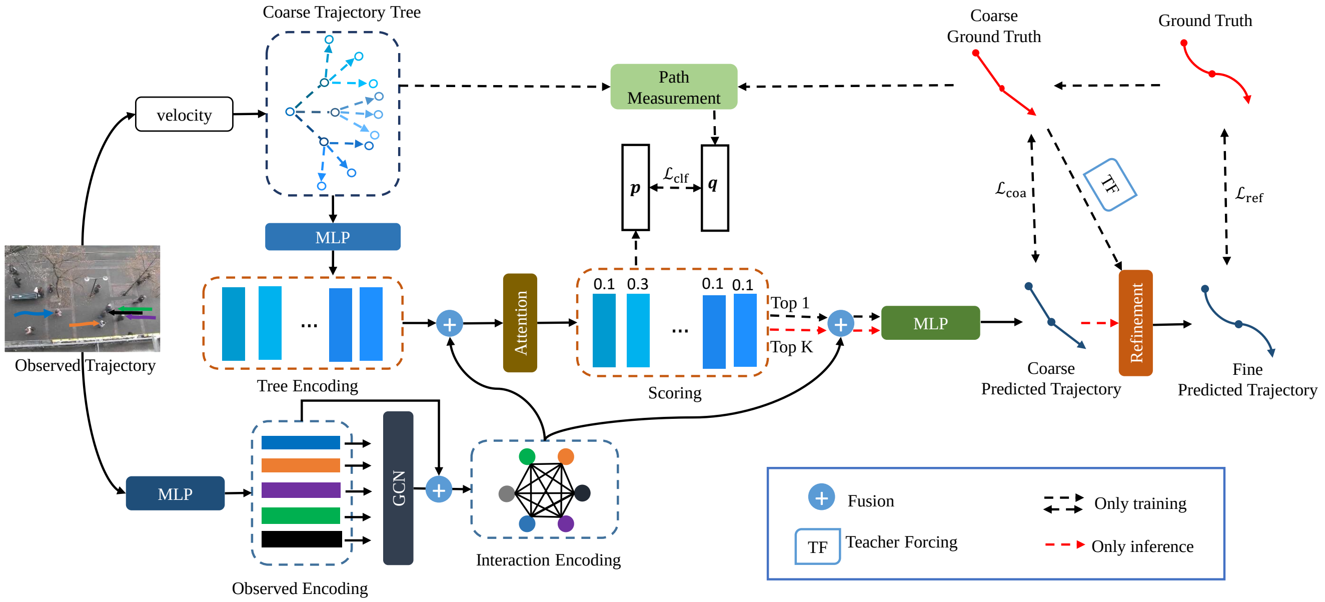

Figure 3: The overall framework of our method. The coarse trajectory tree is built firstly and then encoded by an MLP to gain the tree encoding. Another MLP and a GCN are used to obtain the observed encoding and interaction encoding one after another from the observed trajectory. Next, an attention mechanism is used to score each path between tree encoding and interaction encoding, and the results (confidence vector $\mathbf{p}$) is optimized by the index label ($\mathbf{q}$) of the distance between each path and coarse ground truth. Particularly, the coarse ground truth is generated by the high-order velocity of ground truth. Subsequently, the coarse predicted trajectory is obtained by the top-1 path and refined to predict fine future trajectory. Notably, the teacher forcing is used in refinement, which means the coarse ground truth is fed into the refinement in training time, while the coarse predicted trajectories are fed into the refinement to obtain multi-modal future trajectories in inference time.

velocity vectors in the last S of observed trajectory, is considered as the split direction of forwarding direction. The rotated angle (θ) is used to generate the directions of the left and right split. Finally, we will generate a ternary tree, *i.e.*, coarse trajectory tree, with the depth $d = \lceil T/S \rceil$ after a recursive process. Upon obtaining the coarse trajectory tree, each path from the root to leaf represents a coarse possible future trajectory, and thus the coarse discrete structured space $\mathbf{Y}_{\text{coarse}}$ could be composed of all paths in the coarse trajectory tree.

Trajectory Encoding. The future trajectory is not only affected by the internal motion information but the interactive states with other pedestrians. In this paper, we mainly focus on evaluating the effectiveness of the tree for pedestrian trajectory prediction. We use an simply multilayer perceptron (MLP) to encode the the observed trajectory $\boldsymbol{X}$ into observed encoding denoted $\mathbf{F}_{\text{x}}$. Another MLP is applied to encode the path of $\mathbf{Y}_{\text{coarse}}$ into tree encoding denoted $\mathbf{F}_{\text{tree}} = \{\mathbf{f}_i\}_{i=1}^{M}$, where the M is the size of $\mathbf{Y}_{\text{coarse}}$. In addition, a graph convolutional network (GCN) (Kipf and Welling 2017) implemented by the self-attention (Vaswani et al. 2017) without the positional encoding is used to model the interaction encoding denoted $\mathbf{F}_{\text{s}}$.

Scoring and Selection. After obtaining the coarse discrete structured space $\mathbf{Y}_{\text{coarse}}$, we need to optimize it to obtain more precise trajectory. To keep the interpretability of tree, we use a two-stage training strategy, in which the path in the $\mathbf{Y}_{\text{coarse}}$ is first scored and then those with high confidence are selected to optimize. To score the path, we model the attention scores between the interaction encoding $\mathbf{F}_{\text{s}}$ and the tree encoding $\mathbf{F}_{\text{tree}}$ as the confidence vector $\mathbf{p}$, *i.e.*,

$$\mathbf{p} = \text{Softmax}(\phi(\mathbf{F}_{\text{s}})\psi(\mathbf{F}_{\text{tree}})^{\text{T}}), \tag{2}$$

where ϕ and ψ are the linear projections, T is the transpose.

After that, since the closet path with ground truth can provide rough explanation about the moving behavior of ground truth, we expect it gains the highest confidence. Thus, a path measurement is employed to measure the distance between each path and the ground truth, and the location index in the coarse trajectory tree of the closet one is considered as the label $\mathbf{q}$ to supervise the scoring operation. Particularly, since the real trajectory of pedestrian is zigzag, we convert the ground truth $\hat{\boldsymbol{Y}}$ to its coarse version $\hat{\boldsymbol{Y}}_{\text{coarse}}$ to simplify the optimization. Namely, $\mathbf{q}$ is generated by measuring the distance between each path and $\hat{\boldsymbol{Y}}_{\text{coarse}}$. Similar with coarse trajectory tree, $\hat{\boldsymbol{Y}}_{\text{coarse}}$ is generated by dividing the ground truth into multiple equilong segments with temporal interval S and then connecting the break point as illustrated ground truth and coarse ground truth in Figure 3. The distance of path measurement is the mean of L-2 distance between each break point and corresponding point in the path. The loss function can be given by

$$\mathcal{L}_{\text{clf}} = \mathcal{L}_{\text{CE}}(\mathbf{p}, \mathbf{q}), \tag{3}$$

where the $\mathcal{L}_{\text{CE}}$ is the cross entropy loss.

Greedy Optimization. Due to the single provided ground truth, the model will collapse into frequent modal of data if we force multiple paths to approach the ground truth. To obtain multi-modal future trajectory, we optimize the path greedily, which means the path with highest confidence is used to optimize the $\mathbf{Y}_{\text{coarse}}$. Specifically, the tree encoding $\mathbf{f}_{*}$ of the path with highest confidence fused with the interaction encoding $\mathbf{F}_{\text{s}}$ is fed into an MLP to obtain the coarse predicted trajectory $\boldsymbol{Y}^{'}$. The objective function of the greedy optimization is shown by

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{reg}}(\boldsymbol{Y}^{'}, \hat{\boldsymbol{Y}}_{\text{coarse}}), \tag{4}$$

where $\mathcal{L}_{\text{reg}}$ is the Huber loss.

Trajectory Refining. The final step of our method refine the coarse predicted trajectory $\boldsymbol{Y}^{'}$ to obtain fine-grained trajectory $\boldsymbol{Y}_{\text{fine}}$. To ensure the optimization of refining correctly especially in the early stage of training, we use a teacher-forcing (Williams and Zipser 1989) strategy in training time. Namely, the closet coarse trajectory $\boldsymbol{Y}^{'}$ is replaced with the coarse ground truth $\boldsymbol{Y}_{\text{coarse}}$ to regress the final fine-grained trajectory. The loss function of trajectory refining is represented by

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{reg}}(\boldsymbol{Y}_{\text{coarse}}, \boldsymbol{Y}), \tag{5}$$

where $\mathcal{L}_{\text{reg}}$ is the Huber loss, $\boldsymbol{Y}$ is the ground truth.

Training and Inference. We train the proposed method in an end-to-end way. The total loss is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{coarse}} + \lambda_2 \mathcal{L}_{\text{clf}} + \lambda_3 \mathcal{L}_{\text{ref}}, \tag{6}$$

where λ_1, λ_2 and λ_3 are used to balance the training process.

At the inference step, we select the top-K predicted trajectories according to the attention scores, and those are refined to obtain the final K fine-grained trajectories, *i.e.*, multi-modal future trajectories.

Implementation Details. To implement the proposed method, all encoding modules are implemented by a 3-layer MLP with the PRelu non-linearity. We split the coarse trajectory tree three times and generates 27 paths. Other hyper-parameters of this tree are recorded in supplementary materials due to space limitation. The Adam optimizer is used to train the proposed method by 350 epochs with a learning rate of 0.001, decaying by 0.5 with an interval of 50. All the coefficients λ_1, λ_2, and λ_3 of total loss are set to 1.

Experimental Analysis

Datasets. To evaluate the effectiveness of our method, we conduct extensive experiments on two widely used datasets, *i.e.*, ETH-UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007) and Stanford Drone Dataset (SDD) (Robicquet et al. 2016), in pedestrian trajectory prediction. ETH-UCY includes five scenes: ETH, HOTEL UNIV, ZARA1 and ZARA2, and the coordinate of trajectory is recorded in world coordinate system with the meter as the unit. SDD contains 20 scenes and the coordinate of trajectory is recorded in pixel coordinate system with the pixel as the unit. For ETH-UCY, we follow the leave-one-out strategy (Shi et al. 2021) for training and evaluation, which the model is trained on four scenes and evaluated on the rest of the scene. For SDD, we use prior train-test split (Mangalam et al. 2020) for evaluation.

Following the common setting (Shi et al. 2021), we segment the trajectory sequences into $8s$ trajectory segments by sliding, in which the observed trajectory is $3.2s$ and the future trajectory is the rest $4.8s$, with a time step of $0.4s$.

Metrics. Following the common practice (Gupta et al. 2018), we adopt two widely used metrics to evaluate the performance of the predicted trajectory. Average Displacement Error (ADE) computes the average L-2 distance between predicted trajectory location and the ground truth location. Final Displacement Error (FDE) calculates the L-2 distance between the predicted trajectory at the last time step location and the corresponding group truth location. To measure the ADE and FDE of our method, we follow the previously commonly used measurement that selects K predicted trajectories and reports the performance of closet trajectory.

Quantitative Analysis

We conduct extensive experiments to evaluate the effectiveness of SIT in prediction accuracy, raw tree prediction, long-term prediction, and different best-of-K predictions. More experiments are reported in supplementary materials due to space limitation.

Performance on ETH-UCY. The results are given in Table 1, which are evaluated by ADE and FDE. Although our proposed SIT is based on a hand-crafted tree, the results indicate that our SIT outperforms all the competing methods on both ADE and FDE on average. Specifically, for ADE, our SIT surpasses the previous best method STAR (Yu et al. 2020) by 11.5% on average. For FDE, our SIT outperforms the previous best method PECNet (Mangalam et al. 2020) by a margin of 20.8% on average. The performances on both ADE and FDE underline the effectiveness of the tree in pedestrian trajectory prediction.

Prediction with Raw Tree. We conduct a specific experiment to testify the tree is suitable for pedestrian trajectory prediction even without any training. The raw tree (*i.e.*, coarse trajectory tree) is built only based on the prior information, *i.e.*, velocity, and it is directly used to compare with other deep learning-based methods as shown in Table 1. The raw tree is tested with different depth d, which is set to $0, 1, 2,$ and 3, respectively. Note that $d = 0$ means the pedestrians keep going straight along the direction of the last time step of the observed trajectory. Since the tree is ternary, we can obtain 3^d trajectories for each d. The experimental results demonstrate our raw tree can match the deep learning-based methods, *i.e.*, Social-STGCNN (Mohamed et al. 2020) and TPNMS (Liang et al. 2021). Interestingly, the raw tree with $d = 0$, *i.e.*, only a trajectory keeps going straight, exceeds SGAN (Gupta et al. 2018) which uses best-of-20 to report metrics. This phenomenon indicate our raw tree could cover effective space of future trajectory even it is built by hand. Based on a general rule, (*i.e.*, go straight, tree left and turn right), the raw tree can generate effective trajectory that is more suitable for various scenarios and thus obtains better performance.

Performance on SDD. As shown in Table 2, our method outperforms previous state-of-the-art methods (Mangalam et al. 2020) on both ADE and FDE. It shows higher feasibility in pedestrian trajectory prediction. What's more, we

Model	Venue	Year	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Vanilla LSTM	CVPR	2016	1.09/2.41	0.86/1.91	0.61/1.31	0.41/0.88	0.52/1.11	0.70/1.52
Social LSTM	CVPR	2016	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Desire	CVPR	2017	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	CVPR	2019	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.51/1.15
GAT	NeurIPS	2019	0.68/1.29	0.68/1.40	0.57/1.29	0.29/0.60	0.37/0.75	0.52/1.07
Social-BIGAT	NeurIPS	2019	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
STAR	ECCV	2020	**0.36**/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
PECNet	ECCV	2020	0.47/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
SGCN	CVPR	2021	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
DMRGCN	AAAI	2021	0.60/1.09	0.21/0.30	0.35/0.63	0.29/0.47	0.25/0.41	0.34/0.58
SGAN	CVPR	2018	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21
Social-STGCNN	CVPR	2020	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
TPNMS	AAAI	2021	0.52/0.89	0.22/0.39	0.55/1.13	0.35/0.70	0.27/0.56	0.38/0.73
SIT (Ours)	AAAI	2022	0.39/**0.61**	**0.13/0.22**	**0.29/0.49**	**0.19/0.31**	**0.15/0.29**	**0.23/0.38**
Raw Tree (d=0)	AAAI	2022	0.99/2.23	0.32/0.61	0.52/1.16	0.43/0.96	0.32/0.72	0.51/1.13
Raw Tree (d=1)	AAAI	2022	0.91/2.00	0.27/0.51	0.43/0.94	0.35/0.75	0.26/0.56	0.44/0.95
Raw Tree (d=2)	AAAI	2022	0.86/1.85	0.25/0.46	0.41/0.90	0.31/0.65	0.23/0.51	0.41/0.87
Raw Tree (d=3)	AAAI	2022	0.82/1.64	0.24/0.40	0.38/0.77	0.29/0.53	0.22/0.43	0.39/0.75

Table 1: Comparison with baselines on the ETH-UCY using ADE/FDE, which are measured in meters. The lower the better. The models SGAN, Social-STGCNN and TPNMS are the closet methods compared with our raw tree.

Model	Venue	Year	ADE/FDE
Social-LSTM	CVPR	2016	31.19/56.97
SGAN	CVPR	2018	27.23/41.44
MATF	CVPR	2019	22.59/33.53
Desire	CVPR	2017	19.25/34.05
Sophie	CVPR	2019	16.27/29.38
SimAug	ECCV	2020	10.27/19.71
PECNet	ECCV	2020	9.96/15.88
SIT (Ours)	AAAI	2022	**8.59/15.27**

Table 2: Comparison with baselines on Stanford Drone using ADE/FDE. The metrics are measured in pixels.

also conduct extra experiments in multiple aspects to evaluate the effectiveness of our SIT on SDD. Please see the supplementary material for details.

Long-term Prediction. We conduct experiments on long-term prediction, which input the observed trajectory with normal (3.2s and 8 time steps) length, the longer future trajectory will be predicted. In this experiment, we set the longer future trajectory to 6.4s (16 time steps), 8.0s (20 times steps), and 9.6s (24 time steps), respectively. To indicate the flexibility of our SIT on long-term prediction by comparing against other baselines, we reproduce the LSTM and GAN-based method SGAN (Gupta et al. 2018), the graph-based method Social-STGCNN (Mohamed et al. 2020) and the CVAE-based method PECNet (Mangalam et al. 2020) by their official released codes[‡] to predict long-term future trajectory, respectively. Note that the PECNet is reproduced by the data loader of Social-STGCNN because the data loader of PECNet can not change the predicted length. As shown in Table 3, our SIT outperforms all competing methods on all long-term predicted lengths. Notably, the improvements are gradually increasing as the predicted length elongates. The underlying reason could be that the built tree provides effective "candidates" (paths) that are convenient for the optimization of deep neural network.

T_{pred}	Models	ADE	FDE
16	SGAN	2.16	3.96
	Social-STGCNN	0.54	1.05
	PECNet	2.89	2.63
	SIT	**0.49**	**1.01**
20	SGAN	2.40	4.52
	Social-STGCNN	0.71	1.30
	PECNet	3.02	2.55
	SIT	**0.55**	**1.12**
24	SGAN	2.79	4.66
	Social-STGCNN	0.92	1.76
	PECNet	3.16	2.53
	SIT	**0.68**	**1.22**

Table 3: Long-term prediction on ETH-UCY using average ADE and FDE.

Different best-of-K predictions. Due to the multi-modal of future trajectory, related works use the best-of-K to report the quantified metrics. Namely, K (usually $K = 20$) future trajectories are predicted, while only the closet trajectory is used to report. To further testify the flexibility of our SIT, we conduct experiments on different best-of-K predictions, where we set $K = 15, 10,$ and 5, respectively. We also reproduce the state-of-the-art method PECNet for comparison. Since PECNet does not provide pretrained mode, we compare with it by above reproduced model. are presented in Table 4. It indicates that our SIT achieves signifi-

[‡] SGAN:https://github.com/agrimgupta92/sgan,
PECNet:https://github.com/HarshayuGirase/PECNet,
Social-STGCNN:https://github.com/abduallahmohamed/Social-STGCNN

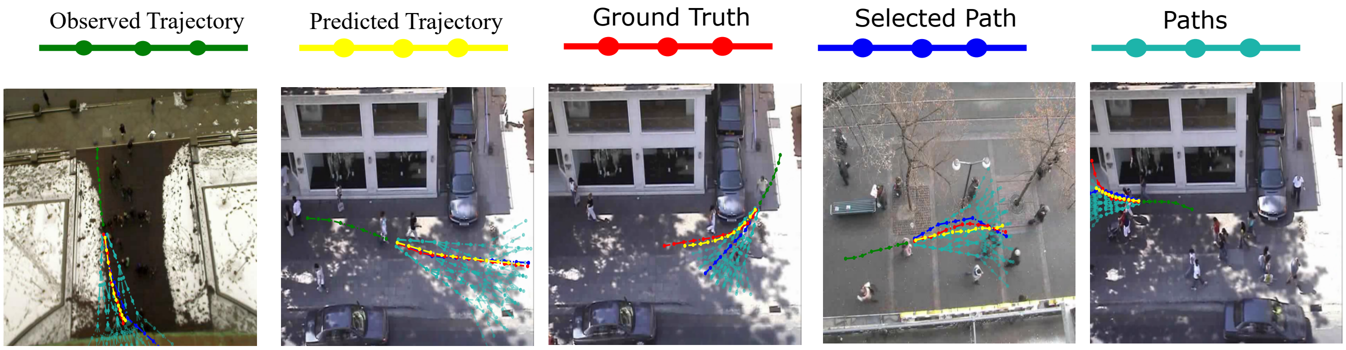

Figure 4: Visualization of selected path and refined trajectory.

K	Models	ADE	FDE
5	PECNet	0.64	1.21
	SIT	**0.43**	**0.74**
10	PECNet	0.64	1.14
	SIT	**0.30**	**0.54**
15	PECNet	0.64	1.11
	SIT	**0.27**	**0.48**

Table 4: Different best-of-K predictions on ETY-UCY using average ADE and FDE.

cant performance on all experimental settings. Interestingly, our SIT with a small K, *e.g.*, SIT-5, still outperforms Social-STGCNN (Mohamed et al. 2020) with $K = 20$, as compared between Table 1 and Table 4.

Ablation study

We conduct ablative experiments to isolate the performance contribution of each component of our method. The relevant components include the teacher forcing (TF), classification task (CLF), coarse ground truth (CGT), and the interaction encoding (IE). Specifically, to represent their effectiveness, the TF is verified by replacing the coarse ground truth with the path with the highest confidence (1), the CLF is verified by removing the loss function $\mathcal{L}_{\text{clf}}$ (2), the CGT is verified by removing the TF and CLF together (3), and the IE is verified by removing the attention encoding (4). Table 6 presents the full method (5) achieves the best performance, which clearly validates the effectiveness of each component.

Qualitative Analysis

Interpretability. The interpretability of our SIT mainly refers to the path in the tree that can provide a good explanation of future moving behaviors, *e.g.*, go straight and then turn right. To show our SIT is capable of selecting the closet path with ground truth, we make statistics in the testing set to record the rate that selecting the closet path in different best-of-K predictions. The closet path is selected by the minimum FDE with the coarse ground truth. Relevant results of minimum ADE are shown in supplementary. As shown in Table 5, our SIT shows significant accuracy (88.47% to 97.38%) in standard best-of-20 prediction. For the lower accuracy of top-1, the reason is SIT encourages multi-modal prediction, which does not welcome the single prediction.

K	ETH	HOTEL	UNIV	ZARA1	ZARA2
1	28.72%	41.59%	9.81%	14.29%	9.18%
5	58.56%	58.68%	36.29%	59.52%	39.17%
10	80.11%	73.59%	59.20%	79.49%	62.50%
15	88.39%	87.36%	76.46%	90.19%	79.92%
20	93.92%	94.39%	88.47%	97.38%	89.73%

Table 5: Top-K accuracy of selecting the closet path of tree.

	TF	CLF	CGT	IE	ADE	FDE
(1)	✗	✓	✓	✓	0.26	0.45
(2)	✗	✗	✓	✓	0.34	0.74
(3)	✗	✗	✗	✓	0.48	0.96
(4)	✓	✓	✓	✗	0.25	0.44
(5)	✓	✓	✓	✓	**0.23**	**0.38**

Table 6: Ablation study on ETH-UCY using average ADE and FDE.

Visualization. To further illustrate the interpretability of our SIT, we visualize the selected path in real traffic scenarios. As presented in Figure 4, from the left to right, the images represent the pedestrians go straight and then turn left, keep going straight, keep turn right, go straight and then turn right and keep turn right. Our SIT can select the path with similar behaviors of the ground truth and then refines it to gain a precisely predicted trajectory. For more visualizations please see supplementary.

Conclusion

We propose a simple yet effective tree-based method, named Social Interpretable Tree (SIT) to predict the multi-modal future trajectories. Compared with previous methods that embed the multi-modal future trajectories into a continuous latent space, we embed them into a discrete structured space, *i.e.*, a ternary tree. In our method, a coarse trajectory tree is first built and then a coarse-to-fine strategy is used to obtain the final multi-modal future trajectories. Experimental results on ETH-UCY and Stanford Drone Dataset validate the effectiveness of our SIT in standard prediction, long-term prediction, different best-of-K predictions, and interpretability. Furthermore, the raw tree without any training outperforms even many deep learning-based methods.

Acknowledgments

This work was supported partly by National Key R&D Program of China under Grant 2018AAA0101400, NSFC under Grants 62088102, 61976171, and 62106192, China Postdoctoral Science Foundation under Grant 2020M683490, Natural Science Foundation of Shaanxi Province under Grant 2021JQ-054, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

Adiyatov, O.; and Varol, H. A. 2013. Rapidly-exploring random tree based memory efficient motion planning. In *ICRA*, 354–359.

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 961–971.

Antonini, G.; Bierlaire, M.; and Weber, M. 2006. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8): 667–687.

Aoude, G.; Joseph, J.; Roy, N.; and How, J. 2011. Mobile agent trajectory prediction using Bayesian nonparametric reachability trees. In *Infotech@Aerospace*.

Arjovsky, M.; and Bottou, L. 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.

Bae, I.; and Jeon, H.-G. 2021. Disentangled Multi-Relational Graph Convolutional Network for Pedestrian Trajectory Prediction. In *AAAI*, 911–919.

Bisagno, N.; Zhang, B.; and Conci, N. 2018. Group lstm: Group trajectory prediction in crowded scenarios. In *ECCVW*, 213–225.

Chen, Z.; Long, C.; Zhang, L.; and Xiao, C. 2021. CANet: A Context-Aware Network for Shadow Removal. In *ICCV*.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *ICCV*.

Goerzen, C.; Kong, Z.; and Mettler, B. 2010. A survey of motion planning algorithms from the perspective of autonomous UAV guidance. *Journal of Intelligent and Robotic Systems*, 57(1): 65–100.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2255–2264.

Hu, T.; Long, C.; and Xiao, C. 2021. A Novel Visual Representation on Text Using Diverse Conditional GAN for Visual Recognition. *IEEE Transactions on Image Processing*, 30: 3499–3512.

Islam, A.; Long, C.; Basharat, A.; and Hoogs, A. 2020. DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image Copy-move Forgery Detection and Localization. In *CVPR*.

Islam, A.; Long, C.; and Radke, R. 2021. A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization. In *AAAI*.

Ivanovic, B.; and Pavone, M. 2019. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *ICCV*, 2375–2384.

Jurgenson, T.; Groshev, E.; and Tamar, A. 2019. Sub-Goal Trees–a Framework for Goal-Directed Trajectory Prediction and Optimization. *arXiv preprint arXiv:1906.05329*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 137–146.

Kuffner, J. J.; and LaValle, S. M. 2000. RRT-connect: An efficient approach to single-query path planning. In *ICRA*, 995–1001.

Kuwata, Y.; Teo, J.; Fiore, G.; Karaman, S.; Frazzoli, E.; and How, J. P. 2009. Real-time motion planning with applications to autonomous urban driving. *IEEE Transactions on Control Systems Technology*, 17(5): 1105–1118.

LaValle, S. M.; et al. 1998. Rapidly-exploring random trees: A new tool for path planning. *The Annual Research Report*.

Lee, N.; Choi, W.; Vernaza, P.; Choy, C. B.; Torr, P. H.; and Chandraker, M. 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, 336–345.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. *Computer Graphics Forum*, 26(3): 655–664.

Liang, J.; Jiang, L.; Niebles, J. C.; Hauptmann, A. G.; and Fei-Fei, L. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *CVPR*, 5725–5734.

Liang, R.; Li, Y.; Li, X.; Tang, Y.; Zhou, J.; and Zou, W. 2021. Temporal Pyramid Network for Pedestrian Trajectory Prediction with Multi-Supervision. In *AAAI*, 2029–2037.

Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *ICCV*.

Ma, Y.; Zhu, X.; Zhang, S.; Yang, R.; Wang, W.; and Manocha, D. 2019. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, 6120–6127.

Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.-H.; Adeli, E.; Malik, J.; and Gaidon, A. 2020. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *ECCV*, 759–776.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *CVPR*, 14424–14432.

Ondřej, J.; Pettré, J.; Olivier, A.-H.; and Donikian, S. 2010. A synthetic-vision based steering approach for crowd simulation. *ACM Transactions on Graphics*, 29(4): 1–9.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 261–268.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 549–565.

Robin, T.; Antonini, G.; Bierlaire, M.; and Cruz, J. 2009. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1): 36–56.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *CVPR*, 1349–1358.

Shafiee, N.; Padir, T.; and Elhamifar, E. 2021. Introvert: Human Trajectory Prediction via Conditional 3D Attention. In *CVPR*, 16815–16825.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction. In *CVPR*, 8994–9003.

Sun, J.; Jiang, Q.; and Lu, C. 2020. Recursive Social Behavior Graph for Trajectory Prediction. In *CVPR*, 660–669.

Svenstrup, M.; Bak, T.; and Andersen, H. J. 2010. Trajectory planning for robots in dynamic human environments. In *IROS*, 4293–4298.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wei, J.; Long, C.; Zou, H.; and Xiao, C. 2019. Shadow Inpainting and Removal Using Generative Adversarial Networks with Slice Convolutions. *Computer Graphics Forum*, 38(7): 381–392.

Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2): 270–280.

Xu, W.; Long, C.; Wang, R.; and Wang, G. 2021. DRB-GAN: A Dynamic ResBlock Generative Adversarial Network for Artistic Style Transfer. In *ICCV*.

Yu, C.; Ma, X.; Ren, J.; Zhao, H.; and Yi, S. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *ECCV*, 507–523.

Yu, J.; Nie, Y.; Long, C.; Xu, W.; Zhang, Q.; and Li, G. 2021. Monte Carlo Denoising via Auxiliary Feature Guided Self-Attention. *ACM Transactions on Graphics*, 40(6).

Zhang, L.; Long, C.; Zhang, X.; and Xiao, C. 2020. RIS-GAN: Explore Residual and Illumination with Generative Adversarial Networks for Shadow Removal. In *AAAI*.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *CVPR*, 12085–12094.