Progressive Pretext Task Learning for Human Trajectory Prediction

Xiaotong Lin[1], Tianming Liang[1], Jianhuang Lai[1,2,3], and Jian-Fang Hu[1,2,3]*

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Guangdong Province Key Laboratory of Information Security Technology, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{linxt29, liangtm}@mail2.sysu.edu.cn, {stsljh, hujf5}@mail.sysu.edu.cn

Abstract. Human trajectory prediction is a practical task of predicting the future positions of pedestrians on the road, which typically covers all temporal ranges from short-term to long-term within a trajectory. However, existing works attempt to address the entire trajectory prediction with a singular, uniform training paradigm, neglecting the distinction between short-term and long-term dynamics in human trajectories. To overcome this limitation, we introduce a novel Progressive Pretext Task learning (PPT) framework, which progressively enhances the model's capacity of capturing short-term dynamics and long-term dependencies for the final entire trajectory prediction. Specifically, we elaborately design three stages of training tasks in the PPT framework. In the first stage, the model learns to comprehend the short-term dynamics through a stepwise next-position prediction task. In the second stage, the model is further enhanced to understand long-term dependencies through a destination prediction task. In the final stage, the model aims to address the entire future trajectory task by taking full advantage of the knowledge from previous stages. To alleviate the knowledge forgetting, we further apply a cross-task knowledge distillation. Additionally, we design a Transformer-based trajectory predictor, which is able to achieve highly efficient two-step reasoning by integrating a destination-driven prediction strategy and a group of learnable prompt embeddings. Extensive experiments on popular benchmarks have demonstrated that our proposed approach achieves state-of-the-art performance with high efficiency. Code is available at https://github.com/iSEE-Laboratory/PPT.

Keywords: Human Trajectory Prediction · Progressive Learning

1 Introduction

Human trajectory prediction has found extensive applications in various critical domains, such as autonomous driving [5, 20, 30, 35], surveillance systems [43], robotic navigation [6, 21] and planning [23, 38]. Given an observed human trajectory, the objective of human trajectory prediction is to precisely forecast the

* Corresponding Author.

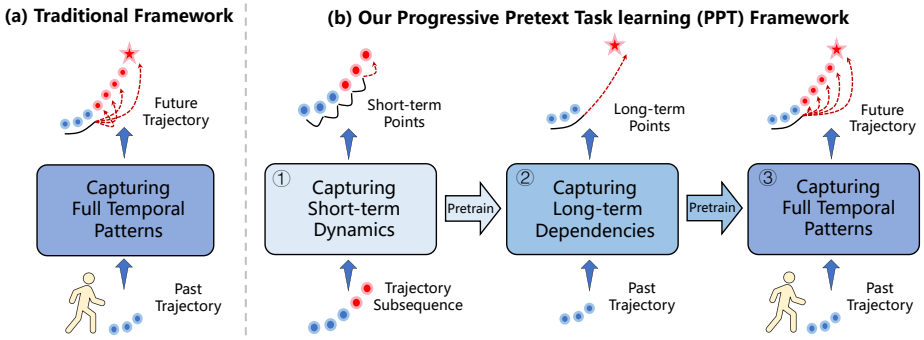

Fig. 1: Comparing our Progressive Pretext Task learning (PPT) framework with regular trajectory prediction. (a) Existing works tend to aggressively force the model to learn complicated full temporal patterns at once. (b) Our Framework employs three learning stages to progressively enhance the model for future trajectory prediction.

unobserved plausible future trajectories. This includes predicting positions from short-term to long-term future, covering all temporal ranges within a trajectory.

Generally, making predictions across different temporal distances relies on distinct aspects of understanding abilities. On one hand, short-term future prediction requires to recognize the local dynamic patterns from the immediate, fine-grained variations between timesteps. On the other hand, long-term future prediction aims to deduce the global tendencies by capturing long-term dependencies of trajectories. However, this distinction is always neglected by existing methods [12, 13, 27, 28, 36, 42, 50]. They attempt to address both short-term and long-term prediction with a singular, uniform training paradigm, often struggling in a suboptimal compromise between short-term and long-term performance.

To overcome this limitation, in this paper, we present a novel Progressive Pretext Task learning (PPT) framework, which progressively enables the model to capture the complicated short-term dynamics and long-term dependencies for the entire future trajectory prediction. To be specific, PPT comprises three stages of progressive training tasks. Task-I aims to equip the model with the basic capacity to comprehend short-term dynamics inherent in the trajectories, by predicting the next position given a trajectory of arbitrary length. Task-II intends to enhance the model to capture long-term dependencies by predicting the destinations of trajectories, where a diversity loss is employed to encourage intention diversity of a pedestrian. Once pretext Task-I and Task-II are completed, the model acquires the ability to capture both short-term dynamics and long-term dependencies within a trajectory. Given this, in Task-III, we take full advantage of the enhanced knowledge for more accurate prediction, by finetuning the well-pretrained model from Task-II for entire future trajectory prediction. Moreover, to preserve the knowledge acquired from previous pretext tasks and stabilize prediction performance, we introduce cross-stage knowledge distillation, transferring the knowledge of Task-I and Task-II into the models in Task-III.

In our PPT framework, we further devise a novel Transformer-based trajectory predictor. Compared to previous Transformer predictors [9,12,34,53,55] that autoregressively generate the future positions, our model is able to efficiently predict the trajectory of any length in only two steps: determining the destination firstly and then generating the rest future points all at once. Specifically, our model consists of a destination predictor and a trajectory predictor. The former predictor aims to capture long-term dependencies for predicting destinations, which are used to guide the latter one in generating the entire future trajectories. To achieve the efficient parallel generation of trajectory points, we introduce a series of learnable prompt embeddings to indicate the certain timesteps.

Extensive experiments demonstrate that our framework achieves state-of-the-art results on various popular datasets, validating the superiority of our framework. Moreover, ablation studies are conducted to verify the effectiveness of each pretext task and other key components. Qualitatively, our framework can produce human trajectories that are more accurate and temporally acceptable.

Overall, our contributions are summarized as follows:

- We present PPT, a novel progressive pretext task learning framework to progressively enable the model to capture the complicated dependencies across various temporal ranges in human trajectories, including short-term dynamics and long-term dependencies, for the entire future trajectory prediction.
- We propose a Transformer-based trajectory predictor, which adopts a two-step destination-driven strategy and integrates a series of learnable prompts to achieve effective and efficient prediction.
- Extensive experiments on four commonly used datasets demonstrate that our framework can consistently outperform the current state-of-the-art methods.

2 Related Work

Human trajectory prediction aims to forecast the reasonable future path given an observed sequence of movements. Considering the indeterminate nature of human motion, this task is particularly challenging due to the necessity of predicting the precise coordinates of the positions over all timesteps, which requires addressing both the short-term dynamics and long-term dependencies.

2.1 Human Trajectory Prediction

Existing works can be briefly divided into two branches: one branch focuses on the utilization of scene maps [19,25,34,46,56], while the other aims to mine the movement patterns and interactions [13,17,26,28,36,39,42,47–50]. Considering the computational costs of modeling scene maps, in this paper, we follow the latter branch to explore a more effective approach for understanding temporal movement patterns within trajectories. To address this task, a lot of efforts have been made. For example, Gupta et al. [13] initially proposed to utilize a GAN-based [10] network, and train the model by directly aligning various

temporal positions in future trajectories with GT without differentiation. Gu et al. [12] employ a Transformer-based diffusion network and train the model to produce the entire future trajectory at once. However, these works overlook the differences between the learning patterns of short-term and long-term prediction, which cause suboptimal performance during joint optimization. While recent destination-based methods [26, 49, 57] attempt to alleviate this issue by initially predicting the destination with one predictor and then interpolating intermediate positions with another, they overlook the knowledge transfer between destination prediction and intermediate position prediction, which results in a significant gap between the destination predictor and the trajectory predictor. To overcome the limitations, in our paper, we devise a Progressive Pretext Task learning framework, which introduces two well-designed pretext tasks to incrementally enhance the model to capture both short-term dynamics and long-term dependencies for the entire future trajectory prediction.

2.2 Transformer-based Human Trajectory Prediction

In recent years, Transformer [41, 44] architectures have demonstrated impressive capability in capturing complex sequential dependencies. Considering its effectiveness, researchers [9, 12, 34, 37, 53, 55] have increasingly turned to Transformer for human trajectory prediction. For example, STAR [53] modeled the crowd as a graph and leveraged a graph-based Transformer to learn the spatiotemporal interaction of the crowd motion. Also, Tsao et al. [42] use a Transformer as a backbone model and propose some pretext tasks regarding cross-sequence modeling. However, they are always inefficient during inference since they generate the trajectory points in an autoregressive manner. Recently, MID and TUTR have attempted to explore the non-autoregressive Transformer in this task. Nevertheless, MID [12] relies on a diffusion model, which significantly increases the inference time. TUTR [37] ignores the temporal motion dynamics in the trajectory, leading to suboptimal performance. In this work, we propose a novel non-autoregressive Transformer to overcome the above limitations. Compared to TUTR, our model introduces a series of effective learnable prompts to represent unobserved positions, which significantly improves the prediction performance.

2.3 Progressive Pretraining

So far, progressive learning techniques have been explored in a wide range of tasks, including image generation [11, 14], image enhancement [7, 22], object detection [4, 8, 16, 29] and motion prediction [24, 40]. Specifically, Karras et al. [14] proposed to start with low-resolution images, and then progressively increase the resolution by adding layers to the networks. PGBIG [24] utilize multiple stages to progressively refine the initial guess of the future frames. Fu et al. [7] introduce a progressive learning strategy for low-light image enhancement. In the process of self-knowledge distillation, they gradually increase the proportion of low-light images as input to the student branch, aiming to progressively enhance the learning difficulty for the student. However, progressive pretraining

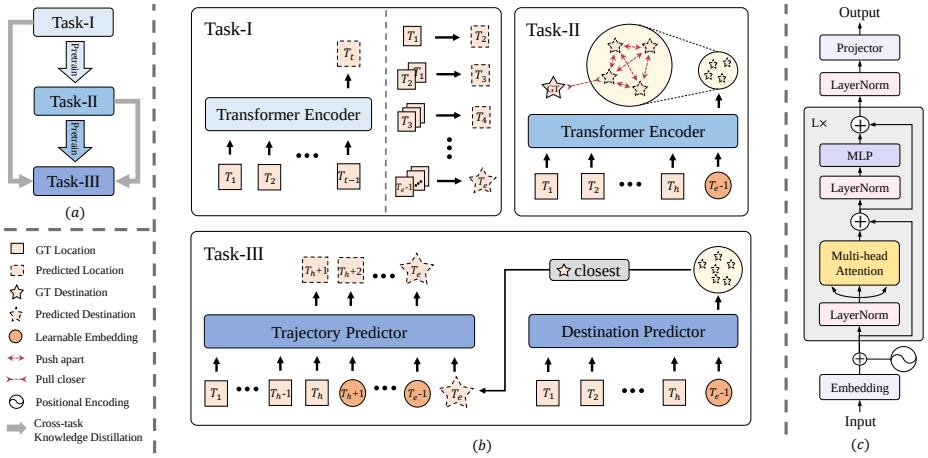

Fig. 2: Illustration of our overall Progressive Pretext Task learning (PPT) framework. (a) demonstrates our progressive training pipeline, where each training stage employs a corresponding task to incrementally enhance the model's capacity for the entire future trajectory prediction. A cross-task knowledge distillation is introduced to avoid knowledge forgetting. Specifically, as in (b), we sequentially perform the stepwise next-position prediction (Task-I), the leapfrog destination prediction (Task-II) and the complete trajectory prediction (Task-III). (c) shows our backbone model.

remains unexplored in the community of human trajectory prediction. To the best of our knowledge, our work is the first to explore progressive pretraining in human trajectory prediction, introducing two well-designed pretext tasks to incrementally enable the model to capture short-term dynamics and long-term dependencies for the entire future trajectory prediction.

3 Method

Problem Formulation. Human trajectory prediction aims to accurately predict the future trajectory based on the observation of past trajectory, with the key challenge of capturing both short-term dynamics and long-term dependencies. Formally, given a series of past observed trajectories presented as $\mathcal{S}^{T_1:T_h} = \{(x^{T_1}, y^{T_1}), ..., (x^{T_h}, y^{T_h})\}_{n=1}^{N}$ for N agents over time $T_1, T_2, ..., T_h$, the target of human trajectory prediction is to forecast the subsequent 2D positions for the unobserved future $\mathcal{S}^{T_h+1:T_h+T_f} = \{(x^{T_h+1}, y^{T_h+1}), ..., (x^{T_h+T_f}, y^{T_h+T_f})\}_{n=1}^{N}$. In the following, we denote $T_e = T_h + T_f$ as the entire trajectory length.

Overview. As shown in Figure 2, we propose a Progressive Pretext Task learning (PPT) framework for trajectory prediction, aiming to incrementally enhance the model's capacity to understand the past trajectory and predict the future trajectory. Specifically, our framework consists of three stages of progressive training tasks, as illustrated in Figure 2 (b). In Stage I, we pretrain our predictor on pretext Task-I, aiming to fully understand the short-term dynamics

of each trajectory, by predicting the next position of a trajectory of arbitrary length. In Stage II, we further train this predictor on pretext Task-II, intending to capture the long-term dependencies, by predicting the destination of a trajectory. Once Task-I and Task-II are completed, the model is capable of capturing both the short-term dynamics and long-term dependencies within the trajectory. Finally, in Stage III, we duplicate our model to obtain two predictors: one for destination prediction and the other for intermediate waypoint prediction. In this stage, we perform Task-III that enables the model to achieve the complete pedestrian trajectory prediction. For the sake of stable training, we further employ a cross-task knowledge distillation to avoid knowledge forgetting.

Backbone. In this work, we employ a Transformer encoder [44] as our backbone model, which is shown in Figure 2 (c). Given the 2D positions as input, *e.g.*, trajectory sequences $S^{T_m:T_n}$ from time T_m to T_n, we first use an embedding layer to convert them to input features. Then, these features with corresponding temporal position embeddings $\{T_m, T_m + 1, ..., T_n\}$ are passed through multiple Transformer layers, each consisting of pre-norm [45], multi-head attention, LayerNorm (LN) and an MLP. The model learns to understand the trajectory through the feature interactions between different positions, and outputs the interactive representation for each position. These outputs are fed into a final LN, followed by a linear projector to obtain the future 2D positions $\hat{S}^{T_m+1:T_n+1}$, which represents the predicted next-frame positions corresponding to each input location. Specially for Task-II and Task-III, we employ learnable prompt embeddings to represent the unobserved future positions in the trajectory, as illustrated in Figure 2 (b). Details will be presented in the following.

3.1 Task-I: Stepwise Next-position Prediction

Given an observed trajectory sequence of arbitrary length, the target of the first pretext task is to accurately predict the position of the next point. This task promotes the model to explore the motion patterns and understand the short-term dynamics of each pedestrian trajectory.

Specifically, for the trajectory sequence $\mathcal{S}^{T_1:T_e}$, we randomly sample its subsequence $\mathcal{S}^{T_1:T_{t-1}}$, and then feed it into the model θ to infer the next position $\mathcal{S}^{T_t}$, as shown in Figure 2 (b). The randomness can bring the effect of data augmentation. In practice, multiple random subsequences would be sampled from one trajectory for improving training efficiency, and this can be effectively implemented by leveraging the causal self-attention mask [32]. We use θ_I to indicate the model trained with Task-I.

While this task is simple and straightforward, it can effectively enable the model to identify the motion patterns and capture the short-term dynamics within the trajectory. As understanding the patterns and dynamics of the trajectory is an essential capacity for trajectory prediction, this knowledge can be transferred and further exploited to facilitate the prediction in later tasks.

3.2 Task-II: Leapfrog Destination Prediction

The target of the second pretext task is to predict the destination of a trajectory. This is challenging since it requires the model to speculate the pedestrian moving intention from the past trajectory and capture the long-term dependency between the final destination and the early trajectory.

Specifically, taking the past trajectory sequence $\mathcal{S}^{T_1:T_h}$ as input, Task-II continues to train θ_I for predicting the destination $\mathcal{S}^{T_e}$ of the entire trajectory. Considering the inherent indeterminate nature of human motion, we follow the previous works [13, 49, 50] to predict multiple destinations (*e.g.*, K) once. In practice, we feed the output feature at the destination to an MLP to regress K destinations. In order to ensure prediction accuracy, we incorporate a precision loss [13] to minimize the distance between the ground truth destination $\mathbf{E}$ and its closest predicted destination, which is formed as $L_{Precision} = \min_k L_2(\hat{\mathbf{E}}_k, \mathbf{E})$. Here, L_2 is the Euclidean distance function. Furthermore, to prevent K predicted destinations from falling into the same modality, we employ a diversity loss as [51, 54] to provide sufficient diversity. As shown in Figure 2 (b), we promote the pairwise distance between the trajectory destinations as follows:

$$L_{Diversity} = \frac{1}{K(K-1)} \sum_{i}^{K} \sum_{j \neq i}^{K} e^{-\frac{L_2^2(\hat{\mathbf{E}}_i, \hat{\mathbf{E}}_j)}{\sigma_s}}, \tag{1}$$

where σ_s is a scaling factor. With this diversity loss, the model can produce more diverse destinations, thus leading to more diverse trajectories.

The loss function for destination prediction in this task is demonstrated as:

$$L_{Des} = L_{Precision} + \lambda_d L_{Diversity}, \tag{2}$$

where λ_d balances the accuracy and diversity of the predicted destinations.

Notably, to align with the input of model θ_I, we assign a corresponding positional encoding to each position. However, due to the absence of the ground truth data for future trajectory, the $(T_e - 1)$-th position cannot be accessed as input to predict the T_e-th position (destination). Therefore, we introduce a learnable prompt embedding and append it after the past trajectory sequence, aiming to predict destinations in a leapfrog manner. We further set the positional encoding for this learnable embedding as $T_e - 1$ to maintain consistency with Task-I, indicating its prediction for the T_e-th position (destination), as in Figure 2 (b).

Through this leapfrog destination prediction task, the well-trained model θ_{II} can acquire the ability for long-term prediction, which can provide the guiding reference and the knowledge associated with the long-term dependencies for the entire future trajectory prediction.

3.3 Task-III: Comprehensive Trajectory Prediction

With the training on Task-I and Task-II, the model θ_{II} has the capacity of understanding short-term dynamics (acquired from Task-I) and capturing long-term dependencies within the future trajectory (acquired from Task-II). In the

final task, we take full advantage of this knowledge for the complete trajectory prediction task: predicting all the positions within the future trajectories.

To be specific, we replicate the model θ_{II} into a destination predictor and a trajectory predictor, as illustrated in Figure 2 (b). We employ the destination predictor to generate K candidate destinations, as mentioned in Task-II, and then feed the one closest to the ground truth (GT) into the trajectory predictor. The input sequence of the trajectory predictor can be divided into three parts: the observed trajectory from T_1 to T_h, the unobserved future trajectory from T_h+1 to T_e-1, and the pseudo destination at T_e. Specially for the unobserved future trajectory, we use learnable prompt embeddings as input. With these inputs, the trajectory predictor outputs the 2D positions for the entire future trajectory, *i.e.*, $\mathcal{S}^{T_h+1:T_e}$. During Task-III, we jointly train the destination predictor and trajectory predictor to regress the entire future trajectory.

To avoid the knowledge from previous pretext tasks being forgotten, we devise a cross-task knowledge distillation for additional regularization in Task-III. Specifically, we punish the output differences between θ_I and the trajectory predictor, as well as θ_{II} and the destination predictor, respectively, with the following loss functions:

$$\begin{aligned} L_{kd}^t &= ||\mathcal{F}_I^t - \mathcal{P}_t(\mathcal{F}_{III}^t)||_2, \\ L_{kd}^d &= ||\mathcal{F}_{II}^d - \mathcal{P}_d(\mathcal{F}_{III}^d)||_2, \end{aligned} \tag{3}$$

where $\mathcal{F}_i^t$ and $\mathcal{F}_i^d$ indicates the output features of future trajectory and destination obtained in i-th task, respectively. $\mathcal{P}_t$ and $\mathcal{P}_d$ denotes the linear projector.

Overall, the loss function in this stage is formulated as:

$$L_{Traj} = L_{Recon} + \lambda_{kd}^t L_{kd}^t + \lambda_{kd}^d L_{kd}^d, \tag{4}$$

where L_{Recon} is the L_2 distance between the predicted and ground truth future trajectory. The λ_{kd}^t and λ_{kd}^d are leveraged to control the trade-off between different loss terms.

3.4 Inference

After training on all three tasks that progressively enable the model to predict the entire future trajectory, we employ the well-trained destination predictor and trajectory predictor in the final stage for inference. Specifically, we first utilize the destination predictor to predict K destinations. Then, we take each of these destinations as the input to the trajectory predictor, guiding the generation of K future trajectories.

4 Experiments

In this section, we conduct extensive experiments on various popular pedestrian trajectory prediction benchmark datasets. The results show that our approach consistently outperforms the current state-of-the-art methods quantitatively and qualitatively. Further, ablation studies are provided to demonstrate the effectiveness of the key components in our proposed framework.

Table 1: Comparisons with the current state-of-the-art methods on the SDD dataset in minADE_{20} / minFDE_{20} (pixels) metric. Text in **bold** denotes the best results. Our method outperforms other approaches by a large margin.

Method	Social -GAN [13]	SOPHIE [34]	PECNet [26]	PCCSNet [39]	MemoNet [49]	Social-VAE [50]	MID [12]	LED [27]	TUTR [37]	PPT (Ours)
ADE ↓	27.23	16.27	9.96	8.62	8.56	8.10	7.61	8.48	7.76	**7.03**
FDE ↓	41.44	29.38	15.88	16.16	12.66	11.72	14.30	11.66	12.69	**10.65**

4.1 Experimental Setup

Datasets. Our proposed PPT framework is evaluated on four widely used public pedestrian datasets: Stanford Drone Dataset (SDD) [33], ETH [31]/UCY [18] dataset and Grand Central Station (GCS) [52] dataset. SDD is one of the most popular benchmarks which is a large-scale dataset recorded by drone cameras in bird's eye view. It contains trajectories of 5,232 pedestrians in eight different scenes. The ETH/UCY is a combination of two datasets with five different scenes. The ETH [31] dataset contains two scenes, ETH and HOTEL, with 750 pedestrians, and the UCY [18] is composed of three scenes with 786 pedestrians, including UNIV, ZARA1 and ZARA2. The GCS dataset captures a complex and densely populated scene within one of the largest and busiest train stations in the United States. This dataset includes trajectories of 12,684 pedestrians over a duration of approximately one hour.

Evaluation Metrics. We employ the same data processing procedure and evaluation configuration as the previous works [3, 13, 26, 49]. For performance evaluation, we adopt the Average Displacement Error (ADE) and Final Displacement Error (FDE) as evaluation metrics, which measure the average position distance and the destination distance between the predicted trajectories and the ground truth (GT) trajectories, respectively. Considering the inherent uncertainty of the future and the indeterminate nature of human motion, we generate K=20 future trajectories for every past trajectory and calculate the minimum ADE and FDE (Best-of-20 strategy) performance as in the prior works [3, 12, 13, 26, 49]. For all datasets, we take the past 8 steps (3.2s) as the observed trajectory and predict the following future 12 steps (4.8s).

Implementing Details. In our implementation, the Transformer encoder in all stages comprises three layers, where the Transformer dimension is set to 128, and 8 attention heads are applied. The scaling factor σ_s in Equation (1) is assigned a value of 1, and the weight hyperparameter λ_d in Equation (2) is set to 100. We let $\lambda_{kd}^t = 5$ and $\lambda_{kd}^d = 0.5$ in Equation (4). To retain the knowledge acquired from Task-I to the fullest extent, in training Stage-II, we initially train a Multi-Layer Perceptron (MLP) for destination regression as a warm-up, and then jointly train the entire model. We employ the Adam optimizer [15] for all three training stages, with the learning rate set to {0.001, 0.0001, 0.0015} respectively. All of our experiments were conducted using PyTorch on a single RTX 3090 GPU.

Table 2: Comparisons with the current state-of-the-art methods on the ETH/UCY dataset in minADE_{20} / minFDE_{20} (meters) metric. Text in **bold** denotes the best results. Among all the methods, our proposed approach achieves the best performance.

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Social-GAN [13]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21
STAR [53]	0.36/0.65	0.17/0.36	0.31/0.62	0.29/0.52	0.22/0.46	0.26/0.53
PECNet [26]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
AgentFormer [55]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.24	0.23/0.39
PCCSNet [39]	**0.28**/0.54	**0.11**/0.19	0.29/0.60	0.21/0.44	0.15/0.34	0.21/0.42
MemoNet [49]	0.40/0.61	**0.11**/0.17	0.24/0.43	0.18/0.32	0.14/0.24	0.21/0.35
MID [12]	0.39/0.66	0.13/0.22	0.22/0.45	**0.17**/0.30	0.13/0.27	0.21/0.38
SocialVAE [50]	0.41/0.58	0.13/0.19	**0.21**/**0.36**	**0.17**/0.29	0.13/0.22	0.21/0.33
LED [27]	0.39/0.58	**0.11**/0.17	0.26/0.43	0.18/**0.26**	0.13/0.22	0.21/0.33
NPSN [3]	0.36/0.59	0.16/0.25	0.23/0.39	0.18/0.32	0.14/0.25	0.21/0.36
EigenTrajectory [2]	0.36/0.53	0.12/0.19	0.24/0.43	0.19/0.33	0.14/0.24	0.21/0.34
TUTR [37]	0.40/0.61	**0.11**/0.18	0.23/0.42	0.18/0.34	0.13/0.25	0.21/0.36
PPT (Ours)	0.36/**0.51**	**0.11**/**0.15**	0.22/0.40	**0.17**/0.30	**0.12**/**0.21**	**0.20**/**0.31**

Table 3: Comparisons with the current state-of-the-art methods on the GCS dataset in minADE_{20} / minFDE_{20} (pixels) metric. Text in **bold** denotes the best results. Our PPT method significantly outperforms other approaches.

Method	Social -GAN [13]	PECNet [26]	Social-STGCNN [28]	SGCN [36]	Agent-Former [55]	NPSN [3]	GP-Graph [1]	Eigen-Trajectory [2]	PPT (Ours)
ADE ↓	15.85	17.08	14.72	11.18	10.18	7.66	7.8	7.42	**6.20**
FDE ↓	32.57	29.30	23.87	20.65	16.91	13.41	13.7	12.49	**9.34**

4.2 Comparison with State-of-the-Art Methods

We quantitatively compare our proposed Progressive Pretext Task learning (PPT) framework with a wide range of current approaches on various datasets. The results show that our framework consistently achieves state-of-the-art (SOTA) performance, particularly surpassing the existing state-of-the-art methods by a significant margin, more than 0.58/1.01 and 1.22/3.15 in ADE/FDE metric on the SDD and GCS datasets, respectively.

On the Stanford Drone Dataset (SDD), we compare our framework with 8 existing methods, which is demonstrated in Table 1. As can be seen, our approach considerably improves the system performance, which reduces the ADE metric from 7.61 to 7.03 and reduces the FDE metric from 11.66 to 10.65 as compared to the current state-of-the-art methods. This illustrates the effectiveness of employing the three-stage progressive pretext tasks for learning short-term dynamics and long-term dependencies, incrementally equipping the model with the ability to predict the entire future trajectory.

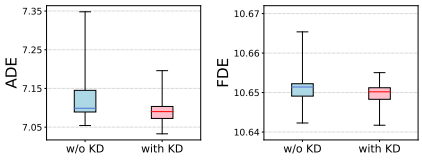

Fig. 3: Analysis on the utility of the cross-task knowledge distillation (KD). With cross-task knowledge distillation, the model can produce accurate future trajectories more consistently.

Table 4: The analysis of the pretext tasks on the SDD dataset. We investigate the performance of our framework with neither Task-I nor Task-II, with only Task-II, and with both Task-I and Task-II.

Task-I	Task-II	Task-III	ADE ↓	FDE ↓
✗	✗	✓	10.40	18.64
✗	✓	✓	7.71	11.42
✓	✓	✓	**7.03**	**10.65**

While on the ETH/UCY dataset, we compare our method with 10 existing approaches. As shown in Table 2, our progressive pretext task learning framework achieves the best prediction performance again, reducing the average FDE performance from 0.33 to 0.31 and the average ADE performance from 0.21 to 0.20 respectively, compared to the current state-of-the-art methods.

On the Grand Center Station dataset (GCS), we compare the proposed framework with 8 recent approaches. The results in Table 3 demonstrate that our progressive pretext task learning framework significantly outperforms the current state-of-the-art method, EigenTrajectory [2], by 16.4% and 25.2% in ADE and FDE metrics respectively, further verifying the superiority of our PPT framework for the future trajectory prediction.

4.3 Ablation Studies

We further conduct ablation studies on the SDD dataset to comprehensively analyze and study the influence of different components in our PPT framework, including the pretext tasks, the cross-task knowledge distillation, and the diversity loss leveraged in Task-II.

Effect of the Progressive Pretext Tasks. In Table 4, we evaluate the influence of the employed progressive pretext tasks, i.e., Task-I and Task-II, on the system performance. Specifically, we first train the model with all three prediction tasks and then sequentially remove Task-I and Task-II for comparisons. As can be observed, both the pretext tasks contribute positively to improving the system performance. Additionally, our experiment shows that with Task-I, the destination prediction performance in Task-II improves from 11.58 to 10.70 in FDE metric. We attribute these to the fact that: i) with Task-I, the model can effectively capture the short-term dynamics inherent in pedestrian trajectory modeling, which contributes a lot to the prediction accuracy. ii) The utilization of Task-II provides the guiding reference and the knowledge of long-term dependencies for the ultimate trajectory sequence prediction, thus significantly improving the prediction performance in both FDE and ADE metrics.

Analysis on the Cross-Task Knowledge Distillation. To examine the effectiveness of employing the cross-task knowledge distillation (KD), we compare the prediction performance of the models trained with and without KD.

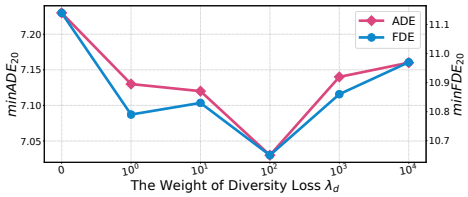

Fig. 4: ADE/FDE as a function of the weight λ_d in Equation 2. λ_d=100 provides the best performance.

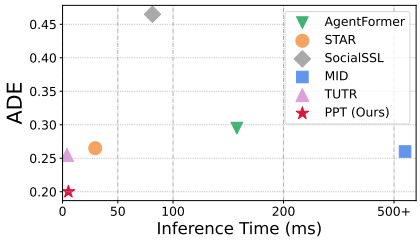

Fig. 5: Inference speed and prediction accuracy of Transformer-based models.

Over 20 independent runs (with different random seeds) are conducted for each model, and the experimental results are reported in the form of a boxplot in Figure 3. As shown, the model trained with KD achieves better prediction performance with smaller variance in both ADE and FDE metrics, suggesting the effectiveness of cross-task knowledge distillation in achieving prediction stability.

Weight for the Diversity Loss. Figure 4 illustrates the influence of different weight λ_d (in Equation 2) on the prediction performance. As can be observed, the system achieves the best performance when weight λ_d=100. Either too small or too large λ_d leads a performance degradation. This is because i) when λ_d is too small, the model tends to miss the intention modality of the pedestrians, leading to inefficient diversity and worse prediction performance; and ii) when λ_d is too large, the impact of the diversity loss gradually dominates the training process. Therefore, the model tends to sacrifice precision for minimizing the diversity loss, resulting in a decrease in prediction accuracy.

Efficiency of PPT. To verify the efficiency of our PPT, we first conduct a comparative analysis of its inference time against five existing Transformer-based approaches. As shown in Figure 5, 1) Leveraging the proposed learnable prompt embedding for efficient parallel generation, our predictor achieves an inference speed that significantly surpasses all autoregressive prediction models and remains comparable to the one-step prediction model TUTR [37] (5.28ms vs. 4.06ms). Also, 2) Trained through our Progressive Pretext Task learning framework, our predictor consistently outperforms all the existing Transformer-based methods in performance. Furthermore, we note that pretraining in earlier stages accelerates convergence in subsequent stages, thus making our PPT framework highly efficient in training time, e.g., 4.7 hours on the SDD dataset. All these results validate the high efficiency and strong effectiveness of our proposed model.

4.4 Qualitative Results

In this subsection, we provide some visualization results to verify our PPT framework and compare it with the current state-of-the-art approaches qualitatively.

Analysis on the Progressive Pretext Tasks. We carefully examine the future trajectories predicted by our framework trained with or without pretext Task-I and Task-II. As shown in Figure 6, on one hand, when pretrained

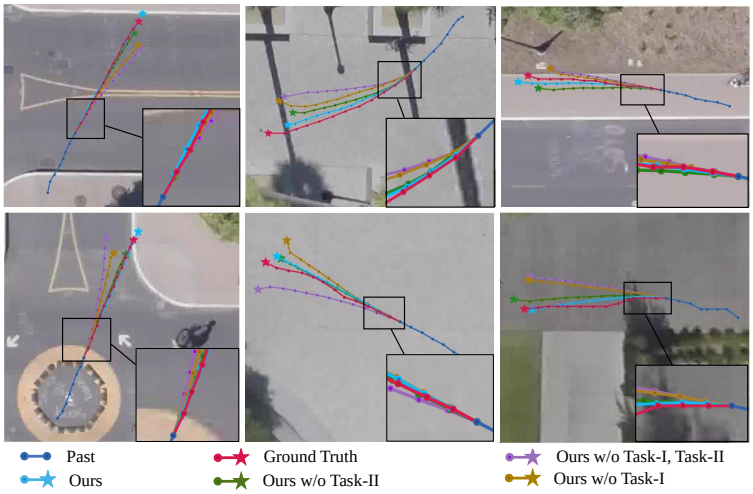

Fig. 6: Qualitative analysis on the pretext tasks. Our model trained with both Task-I and Task-II can produce more accurate and temporally acceptable trajectories.

with pretext Task-I, the model can produce more accurate near-future trajectories, validating the effectiveness of using pretext Task-I in capturing the short-term dynamics. On the other hand, better long-term prediction performance is achieved by using pretext Task-II, which suggests that the utilization of Task-II contributes a lot to capturing the long-term dependencies. Furthermore, with both pretext Task-I and Task-II, our framework can visually generate more accurate and more temporally acceptable future trajectories, demonstrating the effectiveness of each progressive pretext task in our PPT framework.

Comparison with others. Figure 7 visualizes the future trajectories in the scenes of ETH/UCY datasets predicted by four different approaches, including PCCSNet [39], SocialVAE [50], MemoNet [49] and our PPT framework. The last column illustrates the best of 20 predictions generated by these approaches. The results indicate that among all methods, the future trajectories predicted by our PPT best fit the ground truth future trajectories, validating the superiority of our proposed framework visually. In a more detailed analysis, the first four columns demonstrate the 20 future trajectories predicted by these four methods correspondingly. We observe that compared to other methods, our PPT framework exhibits greater variability in destination predictions while simultaneously maintaining prediction accuracy, thereby generating more accurate and diverse future trajectories. Furthermore, when provided with a destination, a pedestrian typically exhibits a relatively uniform pace toward this destination. As shown, our method can produce future trajectories that are more in line with this motion pattern, compared to MemoNet [49]. This verifies the effectiveness of learning and understanding the temporal dynamics, particularly short-term dynamics and long-term dependencies, in our PPT framework for human trajectory modeling.

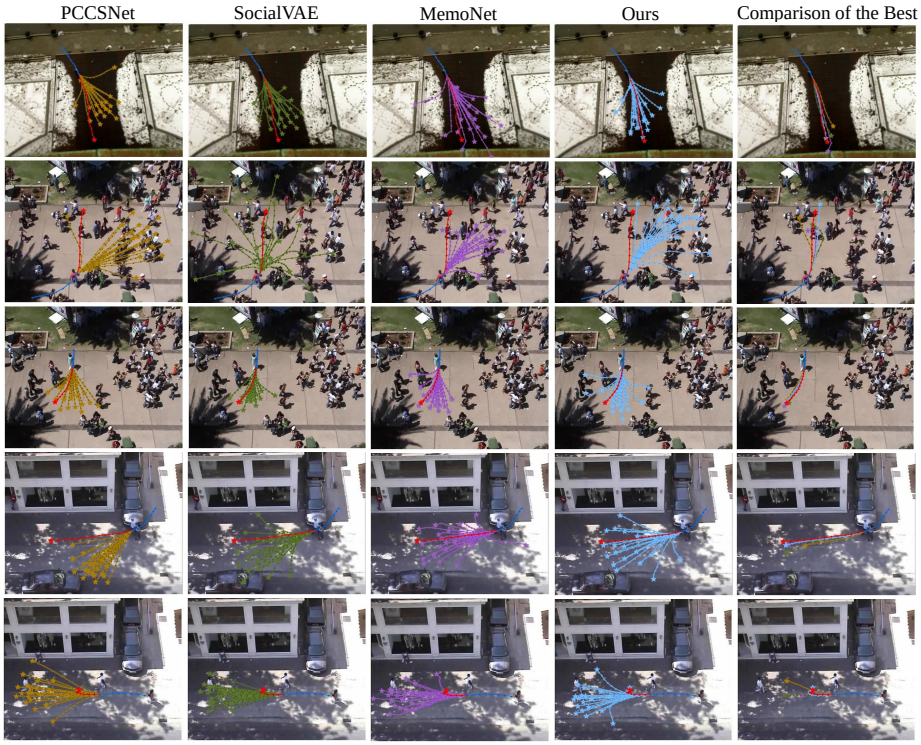

Fig. 7: The visualization of predicted trajectories on the ETH/UCY Dataset. Each row shows a sample in different scenes. The first four columns illustrate the 20 trajectories predicted by PCCSNet [39], SocialVAE [50], MemoNet [49], and our PPT framework. The last column demonstrates the best of 20 predictions produced by these approaches. Trajectories in red represent ground truth (GT) future trajectories.

5 Conclusion

In this paper, we present a novel progressive pretext task learning (PPT) framework to formulate pedestrian trajectory prediction, addressing the limitations of previous works by effectively capturing short-term dynamics and long-term dependencies within trajectories. The PPT consists of three stages of progressive training tasks to enhance the model's capacity. Task-I aims to equip the model with the basic ability to comprehend short-term dynamics inherent in the trajectories. Task-II intends to enhance the model to capture long-term dependencies. In Task-III, we finetune the model for the entire future trajectory prediction, exploiting the previously acquired knowledge. A cross-task knowledge distillation is introduced to preserve the knowledge from previous pretext tasks. Further, we design a Transformer-based predictor to complement our framework, which achieves great efficiency with a two-step inference. Extensive experiments are conducted to demonstrate the superiority of our elaborately devised framework.

Acknowledgements. This work was supported partially by the NSFC (U21A20471, U22A2095, 62076260, 61772570), Guangdong Natural Science Funds Project (2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B1515020009), Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026), and Guangzhou Science and Technology Plan Project (202201011134).

References

1. Bae, I., Jeon, H.G.: A set of control points conditioned pedestrian trajectory prediction. Proceedings of the AAAI Conference on Artificial Intelligence (2023)
2. Bae, I., Oh, J., Jeon, H.G.: Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
3. Bae, I., Park, J.H., Jeon, H.G.: Non-probability sampling network for stochastic human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6477–6487 (2022)
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
5. Choi, C., Choi, J.H., Li, J., Malla, S.: Shared cross-modal trajectory prediction for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 244–253 (2021)
6. Foka, A.F., Trahanias, P.E.: Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. International Journal of Social Robotics **2**, 79–94 (2010)
7. Fu, H., Zheng, W., Meng, X., Wang, X., Wang, C., Ma, H.: You do not need additional priors or regularizers in retinex-based low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18125–18134 (2023)
8. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE international conference on computer vision. pp. 1134–1142 (2015)
9. Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. In: 2020 25th international conference on pattern recognition (ICPR). pp. 10335–10342. IEEE (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: International conference on machine learning. pp. 1462–1471. PMLR (2015)
12. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17113–17122 (2022)
13. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2255–2264 (2018)

14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Komodakis, N., Gidaris, S.: Attend refine repeat: Active box proposal generation via in-out localization. In: BMVC (2016)
17. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., Savarese, S.: Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. Advances in Neural Information Processing Systems **32** (2019)
18. Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., Savarese, S.: Learning an image-based motion context for multiple people tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3542–3549 (2014)
19. Lee, M., Sohn, S.S., Moon, S., Yoon, S., Kapadia, M., Pavlovic, V.: Muse-vae: multi-scale vae for environment-aware long term trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2221–2230 (2022)
20. Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., et al.: Towards fully autonomous driving: Systems and algorithms. In: 2011 IEEE intelligent vehicles symposium (IV). pp. 163–168. IEEE (2011)
21. Li, L.L., Yang, B., Liang, M., Zeng, W., Ren, M., Segal, S., Urtasun, R.: End-to-end contextual perception and prediction with interaction transformer. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5784–5791. IEEE (2020)
22. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8094–8103 (2023)
23. Luo, Y., Cai, P., Bera, A., Hsu, D., Lee, W.S., Manocha, D.: Porca: Modeling and planning for autonomous driving among many pedestrians. IEEE Robotics and Automation Letters **3**(4), 3418–3425 (2018)
24. Ma, T., Nie, Y., Long, C., Zhang, Q., Li, G.: Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6437–6446 (2022)
25. Mangalam, K., An, Y., Girase, H., Malik, J.: From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15233–15242 (2021)
26. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 759–776. Springer (2020)
27. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5517–5526 (2023)
28. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14424–14432 (2020)

29. Najibi, M., Rastegari, M., Davis, L.S.: G-cnn: an iterative grid based object detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2369–2377 (2016)
30. Park, S.H., Lee, G., Seo, J., Bhat, M., Kang, M., Francis, J., Jadhav, A., Liang, P.P., Morency, L.P.: Diverse and admissible trajectory forecasting through multimodal context understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 282–298. Springer (2020)
31. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th international conference on computer vision. pp. 261–268. IEEE (2009)
32. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
33. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 549–565. Springer (2016)
34. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1349–1358 (2019)
35. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 683–700. Springer (2020)
36. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8994–9003 (2021)
37. Shi, L., Wang, L., Zhou, S., Hua, G.: Trajectory unified transformer for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9675–9684 (2023)
38. Song, H., Ding, W., Chen, Y., Shen, S., Wang, M.Y., Chen, Q.: Pip: Planning-informed trajectory prediction for autonomous driving. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 598–614. Springer (2020)
39. Sun, J., Li, Y., Fang, H.S., Lu, C.: Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13250–13259 (2021)
40. Tang, J., Sun, J., Lin, X., Zheng, W.S., Hu, J.F., et al.: Temporal continual learning with prior compensation for human motion prediction. Advances in Neural Information Processing Systems **36** (2024)
41. Tang, J., Wang, J., Hu, J.F.: Predicting human poses via recurrent attention network. Visual Intelligence **1**(1), 18 (2023)
42. Tsao, L.W., Wang, Y.K., Lin, H.S., Shuai, H.H., Wong, L.K., Cheng, W.H.: Social-ssl: Self-supervised cross-sequence representation learning based on transformers for multi-agent trajectory prediction. In: European Conference on Computer Vision. pp. 234–250. Springer (2022)
43. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. IEE Proceedings-Vision, Image and Signal Processing **152**(2), 192–204 (2005)

44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S.: Learning deep transformer models for machine translation. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL (2019)
46. Wong, C., Xia, B., Hong, Z., Peng, Q., Yuan, W., Cao, Q., Yang, Y., You, X.: View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In: European Conference on Computer Vision. pp. 682–700. Springer (2022)
47. Wong, C., Xia, B., Peng, Q., Yuan, W., You, X.: Msn: multi-style network for trajectory prediction. IEEE Transactions on Intelligent Transportation Systems **24**(9), 9751–9766 (2023)
48. Xie, J., Zhang, S., Xia, B., Xiao, Z., Jiang, H., Zhou, S., Qin, Z., Chen, H.: Pedestrian trajectory prediction based on social interactions learning with random weights. IEEE Transactions on Multimedia (2024)
49. Xu, C., Mao, W., Zhang, W., Chen, S.: Remember intentions: Retrospective-memory-based trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2022)
50. Xu, P., Hayet, J.B., Karamouzas, I.: Socialvae: Human trajectory prediction using timewise latents. In: European Conference on Computer Vision. pp. 511–528. Springer (2022)
51. Xu, S., Wang, Y.X., Gui, L.Y.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: European Conference on Computer Vision. pp. 251–269. Springer (2022)
52. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3488–3496 (2015)
53. Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S.: Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 507–523. Springer (2020)
54. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020)
55. Yuan, Y., Weng, X., Ou, Y., Kitani, K.M.: Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9813–9823 (2021)
56. Yue, J., Manocha, D., Wang, H.: Human trajectory prediction via neural social physics. In: European Conference on Computer Vision. pp. 376–394. Springer (2022)
57. Zhao, H., Wildes, R.P.: Where are you heading? dynamic trajectory prediction with expert goal examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7629–7638 (2021)