Complementary Attention Gated Network for Pedestrian Trajectory Prediction

Jinghai Duan[1], Le Wang[2*], Chengjiang Long[3]
Sanping Zhou[2], Fang Zheng[1], Liushuai Shi[1], 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
{caesardjh98,zhengfang,shiliushuai}@stu.xjtu.edu.cn, {lewang,spzhou}@xjtu.edu.cn, {cjfykx, ganghua}@gmail.com

Abstract

Pedestrian trajectory prediction is crucial in many practical applications due to the diversity of pedestrian movements, such as social interactions and individual motion behaviors. With similar observable trajectories and social environments, different pedestrians may make completely different future decisions. However, most existing methods only focus on the frequent modal of the trajectory and thus are difficult to generalize to the peculiar scenario, which leads to the decline of the multimodal fitting ability when facing similar scenarios. In this paper, we propose a complementary attention gated network (CAGN) for pedestrian trajectory prediction, in which a dual-path architecture including normal and inverse attention is proposed to capture both frequent and peculiar modals in spatial and temporal patterns, respectively. Specifically, a complementary block is proposed to guide normal and inverse attention, which are then be summed with learnable weights to get attention features by a gated network. Finally, multiple trajectory distributions are estimated based on the fused spatio-temporal attention features due to the multimodality of future trajectory. Experimental results on benchmark datasets, *i.e.*, the ETH, and the UCY, demonstrate that our method outperforms state-of-the-art methods by **13.8%** in Average Displacement Error (ADE) and **10.4%** in Final Displacement Error (FDE). Code will be available at https://github.com/jinghaiD/CAGN

Introduction

Pedestrian trajectory prediction aims to predict the future trajectory based on the observed trajectory. It plays an important role in many applications, such as automatic driving (Bai et al. 2015; Luo et al. 2018), visual recognition (Donahue et al. 2015; Hua et al. 2018; Hu, Long, and Xiao 2021; Long and Hua 2017; Islam, Long, and Radke 2021), anomaly detection (Liu et al. 2021), human motion prediction (Dang et al. 2021), and traffic early warning system (Luber et al. 2010; Yasuno, Yasuda, and Aoki 2004; Alahi, Ramanathan, and Fei-Fei 2014).

Although significant progress has been made recently, pedestrian trajectory prediction is still challenging due to the complex traffic scenarios. For example, pedestrians behind are prone to follow the trajectories of those in front (Yi, Li, and Wang 2016), people tend to turn at a random angle to avoid collisions (Sun, Jiang, and Lu 2020), and pedestrians who walk in groups perform the different from those who walk alone (Mohamed et al. 2020). In addition, different pedestrians show different behaviors when dealing with similar situations in specific scenarios as shown in Figure 1.

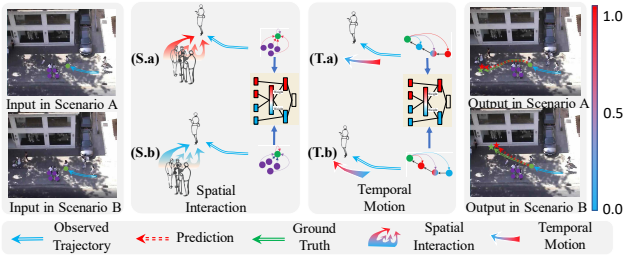

Figure 1: Scenarios A and B represent frequent and peculiar modals with similar inputs and different outputs. (Spatial Interaction) Diverse spatial interaction, including frequent and peculiar spatial interaction, are generated by fusing complementary attention via the gated network. (Temporal Motion) Diverse temporal motions, including frequent and peculiar temporal motions, are generated by fusing complementary attention via the gated network.

Recently, the attention mechanism (Vaswani et al. 2017) achieves excellent progress in capturing spatial interactions and temporal motion of trajectory sequence. Due to the data-driven learning strategy, prior attention-based methods (Kosaraju et al. 2019; Yu et al. 2020; Shi et al. 2021; Zheng et al. 2021) easily collapse into the frequent data modal representing major common scenarios. Since the behaviors of pedestrians are naturally random and diverse, frequent information will mislead the attention model to focus only on common scenarios and ignore other possibilities of pedestrian movements.

As an example of frequent and peculiar scenarios illustrated in Figure 1, Scenarios A and B represent frequent and peculiar modals with similar inputs and different outputs. Scenario A in spatial interaction as shown in **(S.a)** represents the frequent modal that the green pedestrian avoids the collision with the purple group pedestrians. In contrast, Scenario B as shown in **(S.b)** represents the peculiar modal that the

*Corresponding author.

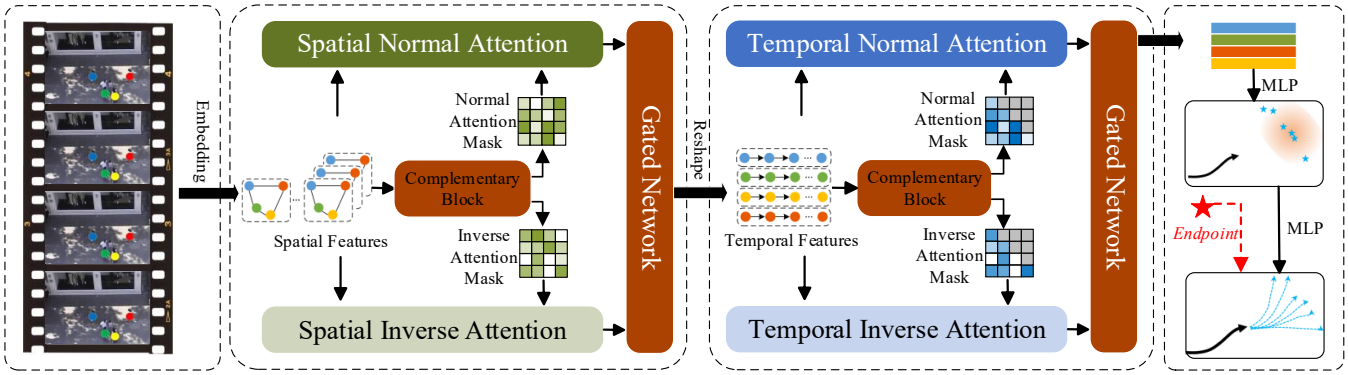

Figure 2: The framework of our CAGN. We generate a pair of complementary masks through the complementary block to guide the normal and inverse attention. Then, we leverage a gated network to adaptively fuse the dual-path spatio-temporal features. Finally, an MLP generates the Gaussian mixture distribution of the trajectory endpoint, from which multiple predicted trajectories are sampled. The red dashed line input indicates that the ground truth of the endpoint is used only in training.

green pedestrian does not interact or weakly interacts with the purple group though they have a similar historical trajectory. Similarly, Scenarios A and B in temporal motion show the frequent and peculiar temporal motions, respectively. Pedestrians will walk along the historical motion direction as shown in **(T.a)**. **(T.b)**shows that pedestrian chooses a different destination compared with Scenario A. Therefore, the pedestrian trajectory prediction will suffer from collapsing into frequent modal and fail to generalize to peculiar cases if we directly use the normal attention mechanism by data-driven methods both in spatial interaction and temporal motion. In order to generalize to peculiar modal and do not hurt the frequent modal, a better approach is to design a complementary attention counterpart and integrate the normal and inverse attention by a gated mechanism as illustrated in Figure 1 (**Spatial Interaction**) and (**Temporal Motion**). After finished, the gated network can balance these two types of modals and thus predict more accurate trajectories.

Motivated by this, we propose a novel complementary attention gated network named CAGN, which captures frequent and peculiar modals in spatial interaction and temporal motion. CAGN applies a dual-path architecture including normal attention and inverse attention, as presented in Figure 2, which is designed to capture both frequent and peculiar modals. Specifically, the spatio-temporal complementary attention learning module is proposed to learn normal and inverse spatial interaction and temporal motion through our proposed complementary block as illustrated in Figure 3. A normal attention mechanism is first used to generate a normal asymmetric attention matrix, which is then sent to a convolutional module to obtain the complementary mask. Thus, the normal and inverse attention can be obtained by the generated complementary mask. Moreover, a gated mechanism (Chung et al. 2015; Shazeer et al. 2017) is employed to generate diverse attention features by fusing the normal and inverse attention features with learnable weights. By combining the spatial and temporal diverse attention features, the results are finally fed into a mixed Gaussian model (Reynolds 2009; Zheng et al. 2021) to estimate the multi-distribution of future trajectory inspired by the multimodality of trajectory (Mangalam et al. 2020).

Extensive experimental results on the ETH (Pellegrini et al. 2009) and UCY(Lerner, Chrysanthou, and Lischinski 2007) datasets show that our method outperforms all the competing state-of-the-art methods. To the best of our knowledge, this is the first work that explicitly models the peculiar spatial interaction and temporal motion. In summary, our contributions are three-fold:

- We propose to model the frequent and peculiar spatial interaction and temporal motion to improve trajectory prediction.
- We design an adaptive complementary attention module that can not only focus on the frequent modal of data but also take account of the peculiar modal.
- Our method improves the state-of-the-art performance by 13.8% in Average Displacement Error (ADE) and 10.4% in Final Displacement Error (FDE) on ETH and UCY.

Related Work

Pedestrian Trajectory Prediction. Thanks to deep learning, pedestrian trajectory prediction achieves remarkable progress. Social-LSTM (Alahi et al. 2016) extracts the trajectory feature of each pedestrian through a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), and integrates hidden states of pedestrians in a certain area by a pooling mechanism. A series of deep learning-based methods (Lisotto, Coscia, and Ballan 2019; Hao, Du, and Reynolds 2018; Liang et al. 2019; Lee et al. 2017; Tao et al. 2020; Zhang et al. 2019) improve the accuracy of trajectory prediction by using visual information. SGAN (Gupta et al. 2018) uses a generative adversarial network (GAN) (Felsen, Lucey, and Ganguly 2018) to tackle multimodal trajectory prediction. RSBG (Sun, Jiang, and Lu 2020) introduces the prior knowledge of kinematics experts to learn pedestrian interaction. Some works employ the variational autoencoder (VAE) to model the future multimodal trajectory (Liang et al. 2020; Pang et al. 2021). PECNet (Mangalam et al. 2020) embeds the endpoint

into a latent space by the conditional variational autoencoder (CVAE) architecture. DisDis (Chen et al. 2021b) further studies the latent space and proposes to select more useful variable code sampling from the learnable latent space. With the application of graph network model in trajectory prediction (Yu et al. 2020; Mohamed et al. 2020; Ivanovic and Pavone 2019), Social-STGCNN (Mohamed et al. 2020) uses a spatio-temporal graph to represent spatial interaction and temporal trajectory. Social-BiGAT (Kosaraju et al. 2019) and STAR (Yu et al. 2020) use the graph model and attention mechanism (Vaswani et al. 2017) to model pedestrian interaction. SGCN (Shi et al. 2021) improves the graph attention model by a learnable sparse graph both in interaction and motion tendency. UNIN (Zheng et al. 2021) uses the graph attention model and Gaussian mixture model (GMM) to deal with multi-category trajectory prediction. What's more, TPNMS (Liang et al. 2021) proposes a pyramid structure to handle the relationship between global and local motion tendencies. DMRGCN (Bae and Jeon 2021) simulates complex social relationships through multi-scale separation and aggregation. In addition, a concurrent work (Chen et al. 2021a) proposes to address the deviation between training and testing scenarios.

In general, the previous methods mainly leverage social interaction to capture accurate pedestrian interaction for various scenarios. However, pedestrians exhibit different behaviors when facing similar traffic scenarios. Solely using the normal attention mechanism, they will focus on the frequent modal while ignore the peculiar modal.

Self-Attention Mechanism. Self-attention is the core idea of the transformer (Vaswani et al. 2017), which greatly improves the performance in many sequence prediction tasks, such as text generation (Yang et al. 2019), image captioning (Dong et al. 2021), and image denoising (Yu et al. 2021). Self-attention captures the dependencies between sequence elements and avoids the gradient disappearance of recurrent neural network. Since the trajectory can naturally represent a temporal sequence, self-attention can fit the trajectory prediction unexceptionally. Besides, spatial interaction can be represented by a sequence without temporal order.

However, due to the individual behaviors of different pedestrians, the normal self-attention mechanism is difficult to produce diverse attention results. In recent years, some works use the inverse method to obtain extra information discarded by the traditional attention model, *e.g.*, pedestrian counting (Si and Patel 2019; Liu et al. 2020). Inspired by them, we use the complementary method to obtain both normal and inverse attention.

Our Approach

Problem Formulation

Pedestrian trajectory prediction aims to predict the future positions of trajectories. Given the trajectory coordinates $\{(x_t^n, y_t^n)\}_{n=1,t=1}^{N,T_{\text{obs}}}$ of N pedestrians observed in the video over time T_{obs}, our goal is to predict the future trajectory coordinates $\{(x_t^n, y_t^n)\}_{n=1,t=T_{\text{obs}}+1}^{N,T_{\text{pred}}}$ of the N pedestrians from time $T_{\text{obs}}+1$ to T_{pred}.

Framework

As aforementioned, traditional attention-based methods only focus on the most frequent modal of the data, resulting in insufficient generalization to peculiar modal. To capture the frequent and peculiar modals simultaneously, we propose the complementary attention gated network (CAGN), as illustrated in Figure 2. Given $\{(x_t^n, y_t^n)\}_{n=1,t=1}^{N,T_{\text{obs}}}$, we first extract the D_e-dimensional trajectory embedding $F_{in} \in \mathbb{R}^{T_{obs}\times N\times D_e}$ by non-linear multi-layer perceptron (MLP). Then, we generate a pair of complementary social masks through the complementary block to learn the dual-path attention, resulting in focusing on both the diverse spatial interaction and the temporal tendency of pedestrians. Afterwards, we leverage a gated network to fuse the dual-path spatio-temporal features adaptively. Finally, an MLP is used to generate mixed Gaussian distribution of trajectory endpoints, and different predicted trajectories are obtained by sampling and interpolation from the distribution.

Spatial Learning via Complementary Attention

Complementary Block. To generate the peculiar modal as well as the frequent modal in spatial interaction, we first extract the multi-head interactions S between pedestrians by the multi-head attention mechanism (Vaswani et al. 2017) based on the embedding F_{in} as follows:

$$\begin{aligned} Q_i &= \phi\left(F_{in}, W_i^Q\right), \\ K_i &= \phi\left(F_{in}, W_i^K\right), \\ A_i &= \text{Softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_k}}\right), \\ S &= \text{Concact}(A_i), i = 1, 2, \ldots, H, \end{aligned} \tag{1}$$

where $\phi(\cdot,\cdot)$ denotes linear transformation, and i is the index of H heads. $Q_i \in \mathbb{R}^{T_{obs}\times N\times D_q}$ and $K_i \in \mathbb{R}^{T_{obs}\times N\times D_k}$ are the query and the key of the attention, respectively. W_i^Q and W_i^K are the weights of the linear transformation. $\sqrt{d_k} = \sqrt{D_k}$ is the scaled factor. $A_i \in \mathbb{R}^{T_{obs}\times N\times N}$ denotes the attention score of the i-th attention head. $S \in \mathbb{R}^{H\times T_{obs}\times N\times N}$ is the multi-head attention matrix over all time steps.

$S_{htij} \in S$ represents the quantified interaction value between the *i*-th and the *j*-th pedestrian on *h*-th head of *t*-th frame. Due to the different linear transformations of query and key, the attention matrix is intrinsically asymmetric. Since the multi-head attention values come from different subspaces (Vaswani et al. 2017), we fuse them by a convolution network at the head dimension and further map the convolutional results to $[0, 1]$ via a sigmoid function:

$$J = \delta\left(\text{Conv}\left(S, \mathcal{K}\right)\right), \tag{2}$$

where $\mathcal{K}$ denotes the 1×1 convolution kernels, δ is the sigmoid function, and $J \in \mathbb{R}^{T_{obs}\times 1\times N\times N}$ denotes the interaction logits based on normal attention.

As the peculiar attention is complementary to the normal attention, an all one matrix O is used to subtract J to obtain inverse attention logits J_{inverse}. A classification is then operated on J and J_{inverse} to generate the masks M_{normal} and

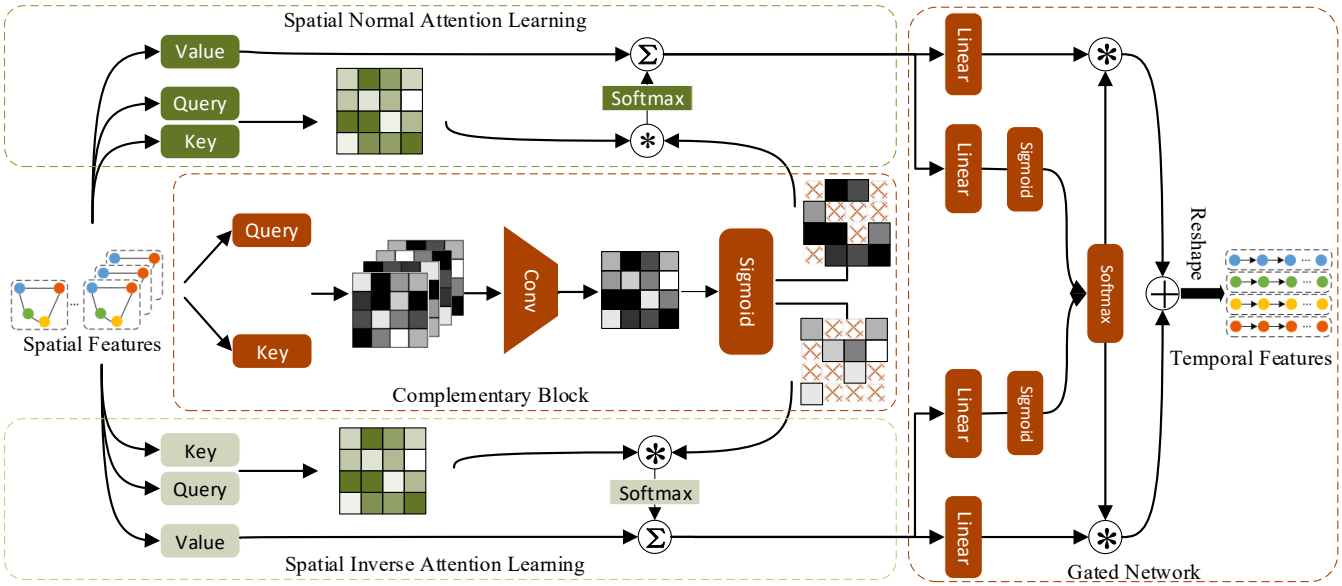

Figure 3: Spatial complementary attention gated network of our CAGN. Complementary block maps the interaction through multi-head attention and convolution neural network. By setting the threshold, a pair of complementary social masks are obtained to guide the dual-path attention to focus on frequent and peculiar interactions. Finally, the spatial features can be weighted-summed through the gated network.

M_{inverse}, *i.e.*,

$$\begin{aligned} M_{\text{normal}} &= \mathbb{I}\left\{J \leq \xi\right\}, \\ M_{\text{inverse}} &= \mathbb{I}\left\{(O-J) \leq \xi\right\}, \end{aligned} \tag{3}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function, and it equals 0 if the inequality holds, otherwise 1. ξ is the classification threshold.

Once obtained M_{normal} and M_{inverse}, they will be multiplied with the subsequent dual-path attention to obtain frequent and peculiar attention features.

Dual-Path Spatial Attention. In order to capture both the frequent and peculiar modals of trajectory, we design a dual-path attention structure to generate the normal and inverse spatial interaction guided by M_{normal} and M_{inverse}. Similar with the attention matrix S, we first employ the multi-head attention mechanism to calculate the attention score $\hat{A} \in \mathbb{R}^{H\times T_{obs}\times N\times N}$. Then the normal attention $T^{\text{spa}}_{\text{normal}} \in \mathbb{R}^{H\times T_{obs}\times N\times N}$ and peculiar attention $T^{\text{spa}}_{\text{inverse}} \in \mathbb{R}^{H\times T_{obs}\times N\times N}$ are generated by the $\hat{A}$ and masks:

$$\begin{aligned} T^{\text{spa}}_{\text{normal}} &= \text{Softmax}\left(\hat{A}\odot M_{\text{normal}}\right), \\ T^{\text{spa}}_{\text{inverse}} &= \text{Softmax}\left(\hat{A}\odot M_{\text{inverse}}\right), \end{aligned} \tag{4}$$

where $\odot$ denotes element-wise multiplication

Upon the frequent and temporal attention matrices, two D_f-dimensional attention features $F^{\text{spa}}_{\text{normal}} \in \mathbb{R}^{H\times T_{obs}\times N\times D_f}$ and $F^{\text{spa}}_{\text{inverse}} \in \mathbb{R}^{H\times T_{obs}\times N\times D_f}$ can be obtained by matrix multiplication like last step of self-attention.

Gated Network. With the normal and inverse attention features, it is expected the model can identify which modal of learned features is suitable for specific traffic scenarios. Therefore, we employ the gated mechanism to learn the weights fusing normal and inverse attention features. Motivated by this, the gated network integrates the $F^{\text{spa}}_{\text{normal}}$ and $F^{\text{spa}}_{\text{inverse}}$ to compute the final spatial feature as:

$$\begin{aligned} R^{\text{spa}}_{\text{normal}} &= \phi\left(F^{\text{spa}}_{\text{normal}}, W^{\text{r}}\right), \\ G^{\text{spa}}_{\text{normal}} &= \delta(\phi\left(F^{\text{spa}}_{\text{normal}}, W^{\text{g}}\right)), \end{aligned} \tag{5}$$

where W^r and W^g are learnable weights of linear projection. δ is the sigmoid function. $R^{\text{spa}}_{\text{normal}}$ and $G^{\text{spa}}_{\text{normal}}$ denote the intermediate features and the weights of gated mechanism respectively, and they have the same shape with $F^{\text{spa}}_{\text{normal}}$.

Similarly, $R^{\text{spa}}_{\text{inverse}}$ and $G^{\text{spa}}_{\text{inverse}}$ can be obtained in a same way. The final spatial interaction features $F^{\text{spa}} \in \mathbb{R}^{T_{\text{obs}}\times N\times D_{\text{final}}}$ are gained by the stacked intermediate features $\hat{R}^{\text{spa}}$ and normalized stacked the weighted score $\hat{G}^{\text{spa}}$ as:

$$\begin{aligned} \hat{R}^{\text{spa}} &= R^{\text{spa}}_{\text{normal}} \oplus R^{\text{spa}}_{\text{inverse}}, \\ \hat{G}^{\text{spa}} &= G^{\text{spa}}_{\text{normal}} \oplus G^{\text{spa}}_{\text{inverse}}, \\ F^{\text{spa}} &= \hat{R}^{\text{spa}} \odot \text{Softmax}(\hat{G}^{\text{spa}}), \end{aligned} \tag{6}$$

where $\oplus$ denotes the concatenated operation.

Temporal Learning via Complementary Attention

Following spatial learning via complementary attention, we learn the temporal motion features in a similar way. As illustrated in Figure 2, the output F^{spa} is fed into the temporal learning via attention module after reshaped into $\mathbb{R}^{N\times T_{\text{obs}}\times D_{\text{final}}}$ to obtain the corresponding intermediate products. The framework diagram of the temporal learning via complementary attention will be placed in the supplementary materials. Through the whole process, the spatio-

temporal features F^{st} with frequent and peculiar attention and are gained to generate the trajectory distribution.

Trajectory Prediction

Considering the multimodality of future trajectory, namely given an observed trajectory, pedestrians could take multiple possible future trajectories, we use a GMM to estimate the final trajectory distribution.

In the training process, the spatio-temporal features F^{st} are fed into a simple MLP network to generate the mixed Gaussian distribution of the trajectory endpoints and we use the ground truth of the endpoints and F^{st} to train a simple MLP network to generate other frames of the future trajectory. In the inference process, GMM is used to generate the predicted trajectory endpoints through sampling to replace the ground truth of the endpoints and an MLP accepts the F^{st} and the sampling endpoints to generate the complete predicted trajectory. Assuming that the weight of K Gaussian distributions in GMM model is $[w_1, w_2, \ldots, w_k]$, if we sample N endpoints, the points sampled by the k-th Gaussian distribution is $N * w_k$. Due to the strong representation ability of GMM, we can generate multi-modal trajectories. The model is trained end-to-end by minimizing the loss function $\mathcal{L}_{CAGN}$ as:

$$
\begin{aligned}
\mathcal{L}_{EP} &= -\log \sum_{k=1}^{K} \pi_k P\left((x_t, y_t) \mid \hat{\mu}_t, \hat{\sigma}_t, \hat{\rho}_t\right), t = T_{\text{pred}}, \\
\mathcal{L}_{AL} &= \frac{1}{T_{\text{pred}} - T_{\text{obs}} - 1} \sum_{t=T_{\text{obs}+1}}^{T_{\text{pred}-1}} \left((\hat{x}_t - x_t)^2 + (\hat{y}_t - y_t)^2\right), \\
\mathcal{L}_{CAGN} &= \mathcal{L}_{EP} + \mathcal{L}_{AL},
\end{aligned}
\tag{7}
$$

where $\mathcal{L}_{EP}$ means the negative log-likelihood loss for training the process of generating Gaussian mixture distribution. $\mathcal{L}_{AL}$ means the average trajectory L2 distance loss for training complete trajectory generation process. $\hat{\mu}_t^i$ is the mean, $\hat{\sigma}_t^i$ is the standard deviation, $\hat{\rho}_t^i$ is the correlation co-efficient, and π_k is the weight of the k-th Gaussian distribution.

Experiments

Datasets. To evaluate our method, we conduct extensive experiments on the ETH (Pellegrini et al. 2009) and UCY (Lerner, Chrysanthou, and Lischinski 2007) datasets. ETH includes ETH and HOTEL scenarios, and UCY includes UNIV, ZARA1, and ZARA2 scenarios. Following the recent method (Sun, Jiang, and Lu 2020), we use the "leave-one-out" strategy for training on four scenarios and testing on the rest ones. We observe the trajectory of 8 frames (3.2 seconds) and predict the trajectory of the next 12 frames (4.8 seconds).

Evaluation Metrics. Following common practice (Zheng et al. 2021), we employ average displacement error (ADE) and final displacement error (FDE) for evaluation. ADE calculates the average L2 distance between the ground truth and the predicted trajectory. FDE computes the L2 distance between the ground-truth at the last step and corresponding predicted trajectory.

Implementation Details. In our experiments, the embedding dimension D_e, D_f and D_{final} are set to 8, the number of head H of the complementary block is set to 4, and the head of the dual-path attention is set to 1. The dimension of MLP in endpoint prediction is set to 64-128-256-128-64. The threshold ξ is empirically set to 0.5, and the nonlinear activation function of MLP is ReLU. The Adam optimizer is used to train our model by 650 epochs with a learning rate of 0.0003, decaying by 0.1 with an interval of 50. During testing, 20 trajectories are sampled from the learned mixed Gaussian distribution according to the weights of multiple Gaussian distributions. The trajectory closest to the ground truth is used to calculate ADE and FDE.

Quantitative Evaluation

We compare our method with the state-of-the-art methods, *i.e.*, Social-LSTM (Alahi et al. 2016), Social-GAN-P (Gupta et al. 2018), RSBG (Sun, Jiang, and Lu 2020), STGAT (Huang et al. 2019), Social-BiGAT (Kosaraju et al. 2019), Social-STGCNN (Mohamed et al. 2020), STAR (Yu et al. 2020), PECNet (Mangalam et al. 2020), TPNMS (Liang et al. 2021), SGCN (Shi et al. 2021), DMRGCN (Bae and Jeon 2021). As shown in Table 1, compared with the previous best method PECNet (Mangalam et al. 2020), our method further improves the performance by 13.8% on ADE and 10.4% on FDE. Meanwhile, compared with the fully connected attention method STAR (Yu et al. 2020) and the sparse attention method SGCN (Shi et al. 2021), our method has an average performance increase of 35.9% on ADE and 43.4% on FDE.

Ablation Study

We conduct ablative experiments to verify the contribution of each component of our CAGN. To validate the effectiveness of our CAGN in spatial interaction and temporal motion, we first employ the spatio-temporal feature method widely used by previous methods such as Transformer (Yu et al. 2020) **(TF)**, LSTM (Alahi et al. 2016) **(LSTM)**, GCN (Mohamed et al. 2020) **(GCN)**, and SparseGCN (Shi et al. 2021) **(SGCN)** to replace dual-path complementary attention module in our framework **(CAGN)**. Then, we study the effect of the gated network by using the gated module and fusing the dual-path features with equal weights. Finally, we replace the GMM model with the Gaussian distribution used in Social-STGCNN (Mohamed et al. 2020) and SGCN (Shi et al. 2021).

As shown in Table 2, the comparison between (8) and (1), (2) and (3) shows that the proposed CAGN framework is better than previous methods in modeling spatial interaction. The temporal motion is also proved by comparing (8) with (4) and (5). The comparison between (8) and (6) shows that a gated network can combine the dual-path attention adaptively to improve the prediction. Meanwhile, the comparison between (8) and (7) indicates GMM is more suitable for pedestrian trajectory prediction. Since the traffic scenarios in the HOTEL are relatively simple and there are few changes in pedestrian movement, solely CAGN is able to model these simple scenarios. This leads to the improvements of both the gated network and GMM are limited.

Model	Venue	Year	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
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
Social-GAN-P	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
STGAT	ICCV	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
RSBG	CVPR	2020	0.80/1.53	0.33/0.64	0.59/1.25	0.40/0.86	0.30/0.65	0.48/0.99
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
STAR	ECCV	2020	0.56/1.11	0.26/0.50	0.52/1.15	0.41/0.90	0.31/0.71	0.41/0.87
PECNet	ECCV	2020	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/**0.30**	0.29/0.48
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
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
CAGN(Ours)	AAAI	2022	**0.41/0.65**	**0.13/0.23**	**0.32/0.54**	**0.21/0.38**	**0.16**/0.33	**0.25/0.43**

Table 1: Compare our CAGN with other state-of-the-art methods on ETH and UCY for ADE/FDE. Lower is better.

	Spatial	Temporal	Gate	GMM	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
(1)	GCN	CAGN	✓	✓	0.78/1.33	0.32/0.55	0.44/0.77	0.34/0.57	0.30/0.46	0.44/0.74
(2)	SGCN	CAGN	✓	✓	0.54/0.76	0.15/0.27	0.37/0.65	0.32/0.49	0.22/0.38	0.32/0.51
(3)	TF	CAGN	✓	✓	0.55/0.77	0.20/0.32	0.39/0.68	0.30/0.48	0.25/0.43	0.34/0.54
(4)	CAGN	LSTM	✓	✓	0.74/1.23	0.26/0.46	0.60/0.89	0.33/0.55	0.29/0.49	0.50/0.72
(5)	CAGN	TF	✓	✓	0.52/0.74	0.16/0.30	0.37/0.62	0.33/0.47	0.25/0.40	0.33/0.51
(6)	CAGN	CAGN	✗	✓	0.45/0.67	**0.13**/0.23	0.35/0.60	0.25/0.41	0.18/0.35	0.27/0.45
(7)	CAGN	CAGN	✓	✗	0.45/0.67	**0.13/0.22**	0.33/0.56	0.22/0.39	0.21/0.37	0.27/0.44
(8)	CAGN	CAGN	✓	✓	**0.41/0.65**	**0.13**/0.23	**0.32/0.54**	**0.21/0.38**	**0.16/0.33**	**0.25/0.43**

Table 2: Ablation study. We replace the temporal and spatial modules of our method with other methods in existing work.

Besides, we experimentally verify the influence of the order of spatial and temporal modules in supplementary materials.

Qualitative Evaluation

Trajectory Prediction. In order to clearly exhibit the improvement of our CAGN in prediction, we compare the visualization results between our proposed CAGN and the SGCN (Shi et al. 2021) in different scenarios, as shown in Figure 4. We choose the examples with the smallest ADE in the 20 predicted trajectories for comparison.

The visualization results on ETH and HOTEL show that our CAGN can better handle the encounter and walking together between pedestrians in relatively simple scenarios. In addition, since our CAGN leverages diverse attention and combines with GMM, the predicted results are more accurate and smoother than previous works. In particular, the FDE in ETH and HOTEL in Table 1 also demonstrates the improvement of our CAGN. For the more complex scenarios in UNIV and ZARA, both our CAGN and SGCN can produce promising prediction results for pedestrians with slow speed and small changes in direction. For large changes in direction, illustrated at the top right-hand corner of ZARA in Figure 4, our CAGN can capture this peculiar situation, *i.e.*, pedestrian turning, which validates the effectiveness of our CAGN. **Complementary Attention.** We visualize the proposed complementary attention process in Figure 5. In order to illustrate the strength and relationship of pedestrians' interaction under multiple conditions through the attention value, we use J and $O - J$ in the intermediate process as the visualization results.

Specifically, Figure 5 (a) indicates our CAGN is capable of handling diversity and randomness, namely it does not introduce redundant randomness into a simple trajectory path because our gated module can adaptively integrate different randomness. It can be observed our CAGN prevents the noise generated by redundant randomness from negatively affecting the prediction results in simple scenes.

In Figure 5 (b) and (c), the red pedestrians have similar historical trajectories and social interactions with the other four pedestrians. The inverse attention matrix is set to pay more attention to its own motion while ignoring the influence of the others. Therefore, whether the pedestrian's choice is to avoid others temporarily or change the long-term movement trend normally, our CAGN can still capture the corresponding tendencies in multiple possible endpoints.

In general, our CAGN is not only capable of strengthening the modeling of diverse scenarios, but also ensuring the accuracy of trajectory prediction in simple scenarios through the gated module.

Gated Network. We shows the average weights of the Normal/Inverse Gate of each module for the structure of S-T-S-T in Table 3, in which the results show that the gate values of two Spatial/Temporal modules tend to be the same for the

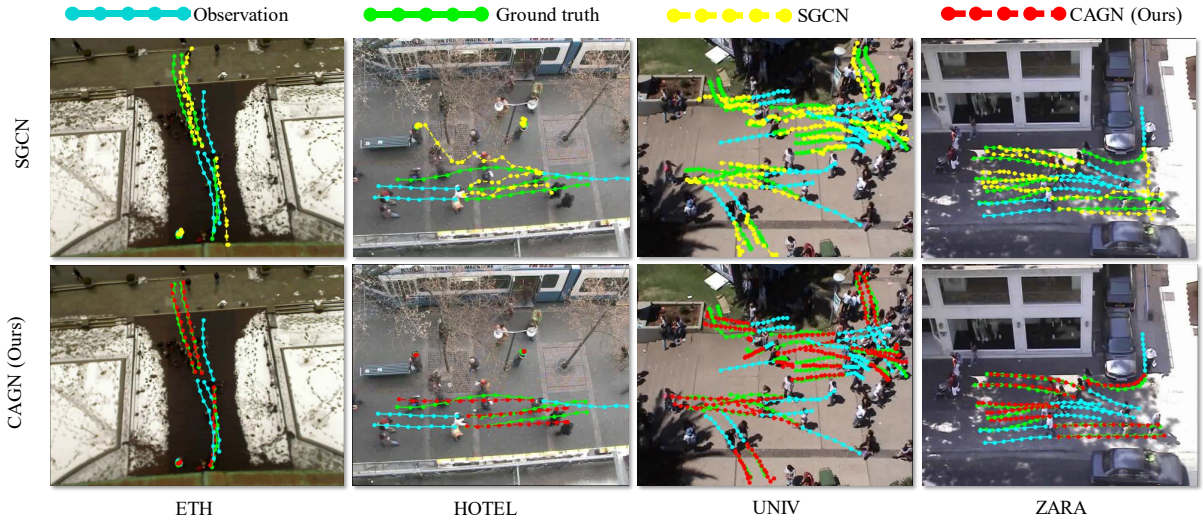

Figure 4: Visualization about performance. We visualize the pedestrian trajectory prediction, selecting the best result in 20 samples in different scenarios.

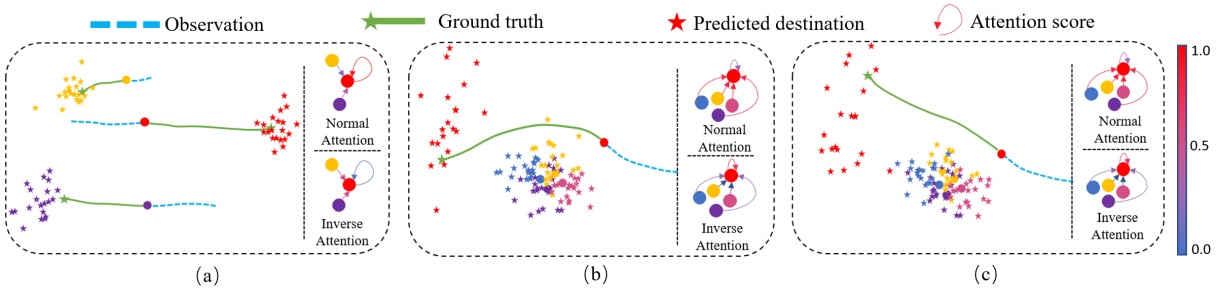

Figure 5: Visualization about inverse attention. We visualize the values of normal and inverse attention to show how our CAGN guides the model to learn different pedestrian movements.

Table 3	ETH	HOTEL	UNIV	ZARA1	ZARA2
First Spatial	0.51/0.49	0.48/0.52	0.48/0.52	0.48/0.52	0.58/0.42
Second Spatial	0.52/0.48	0.48/0.52	0.49/0.51	0.49/0.51	0.53/0.47
First Temporal	0.46/0.54	0.53/0.47	0.49/0.51	0.52/0.48	0.49/0.51
Second Temporal	0.46/0.54	0.51/0.49	0.44/0.56	0.46/0.54	0.50/0.50

Table 3: The average weights of the Normal/Inverse Gate

same dataset. Besides, the average gate value can not well show the adaptive selection for every person, so we further visualize the distribution of gate value Figure 6.

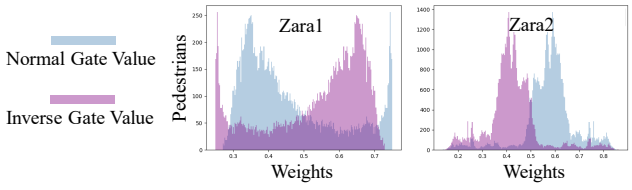

Figure 6: The distribution of gate value.

Conclusion

We propose a complementary attention gated network for trajectory prediction, which models both the frequent and peculiar attention in spatial attention and temporal motion by a complementary attention mechanism. Subsequently, the learned normal and inverse attention are fused by learnable weights to obtain diverse attention. Extensive experimental results show our method achieves better performance than competing methods. It is expected that our proposed complementary attention can also apply in other diversified prediction tasks besides of pedestrian trajectory prediction.

Acknowledgements

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

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.

Alahi, A.; Ramanathan, V.; and Fei-Fei, L. 2014. Socially-aware large-scale crowd forecasting. In *CVPR*, 2203–2210.

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

Bai, H.; Cai, S.; Ye, N.; Hsu, D.; and Lee, W. S. 2015. Intention-aware online POMDP planning for autonomous driving in a crowd. In *ICRA*, 454–460.

Chen, G.; Li, J.; Lu, J.; and Zhou, J. 2021a. Human Trajectory Prediction via Counterfactual Analysis. In *ICCV*.

Chen, G.; Li, J.; Zhou, N.; Ren, L.; and Lu, J. 2021b. Personalized Trajectory Prediction via Distribution Discrimination. In *ICCV*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2015. Gated feedback recurrent neural networks. In *ICML*, 2067–2075.

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*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2625–2634.

Dong, X.; Long, C.; Xu, W.; and Xiao, C. 2021. Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning. In *ACM MM*.

Felsen, P.; Lucey, P.; and Ganguly, S. 2018. Where Will They Go? Predicting Fine-Grained Adversarial Multi-agent Motion Using Conditional Variational Autoencoders. In *ECCV*, 732–747.

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.

Hao, X.; Du, Q. H.; and Reynolds, M. 2018. SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. In *WACV*, 1186–1194.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9(8): 1735–1780.

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.

Hua, G.; Long, C.; Yang, M.; and Gao, Y. 2018. Collaborative Active Visual Recognition from Crowds: A Distributed Ensemble Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3): 582–594.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *ICCV*, 6272–6281.

Islam, A.; Long, C.; and Radke, R. 2021. A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization. In *AAAI*, 1637–1645.

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

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.

Lee, N.; Choi, W.; Vernaza, P.; Choy, B. C.; Torr, H. P.; and Chandraker, M. 2017. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *CVPR*, 336–345.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer Graphics Forum*, volume 26, 655–664.

Liang, J.; Jiang, L.; Murphy, K.; Yu, T.; and Hauptmann, A. 2020. The garden of forking paths: Towards multi-future trajectory prediction. In *CVPR*, 10508–10518.

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.; Zhou, J.; Zou, W.; et al. 2021. Temporal Pyramid Network for Pedestrian Trajectory Prediction with Multi-Supervision. In *AAAI*, 2029–2037.

Lisotto, M.; Coscia, P.; and Ballan, L. 2019. Social and Scene-Aware Trajectory Prediction in Crowded Spaces. In *ICCVW*, 0–0.

Liu, Y. B.; Jia, R. S.; Liu, Q. M.; Xu, Z. F.; and Sun, H. M. 2020. Crowd counting via an inverse attention residual network. *Journal of Electronic Imaging*, 29(3): 1.

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*.

Long, C.; and Hua, G. 2017. Correlational Gaussian Processes for Cross-domain Visual Recognition. In *CVPR*, 118–126.

Luber, M.; Stork, J. A.; Tipaldi, G. D.; and Arras, K. O. 2010. People tracking with human motion predictions from social forces. In *ICRA*, 464–469.

Luo, Y.; Cai, P.; Bera, A.; Hsu, D.; Lee, W. S.; and Manocha, D. 2018. PORCA: Modeling and planning for autonomous driving among many pedestrians. *IEEE Robotics and Automation Letters*, 3(4): 3418–3425.

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.

Pang, B.; Zhao, T.; Xie, X.; and Wu, Y. N. 2021. Trajectory Prediction with Latent Belief Energy-Based Model. In *CVPR*, 11814–11824.

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.

Reynolds, D. A. 2009. Gaussian Mixture Models. *Encyclopedia of Biometrics*, 741: 659–663.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.

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

Si, V. A., Nd Agi; and Patel, V. M. 2019. Inverse Attention Guided Deep Crowd Counting Network. In *AVSS*, 1–8.

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

Tao, C.; Jiang, Q.; Duan, L.; and Luo, P. 2020. Dynamic and Static Context-Aware LSTM for Multi-agent Motion Prediction. In *ECCV*, 547–563.

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.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 5753–5763.

Yasuno, M.; Yasuda, N.; and Aoki, M. 2004. Pedestrian detection and tracking in far infrared images. In *CVPR*, 125–125.

Yi, S.; Li, H.; and Wang, X. 2016. Pedestrian behavior understanding and prediction with deep neural networks. In *ECCV*, 263–279.

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, 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.

Zheng, F.; Wang, L.; Zhou, S.; Tang, W.; Niu, Z.; Zheng, N.; and Hua, G. 2021. Unlimited Neighborhood Interaction for Heterogeneous Trajectory Prediction. In *ICCV*.