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

AAAI 2022

Reported by Nengqiang Xiang

Introduction

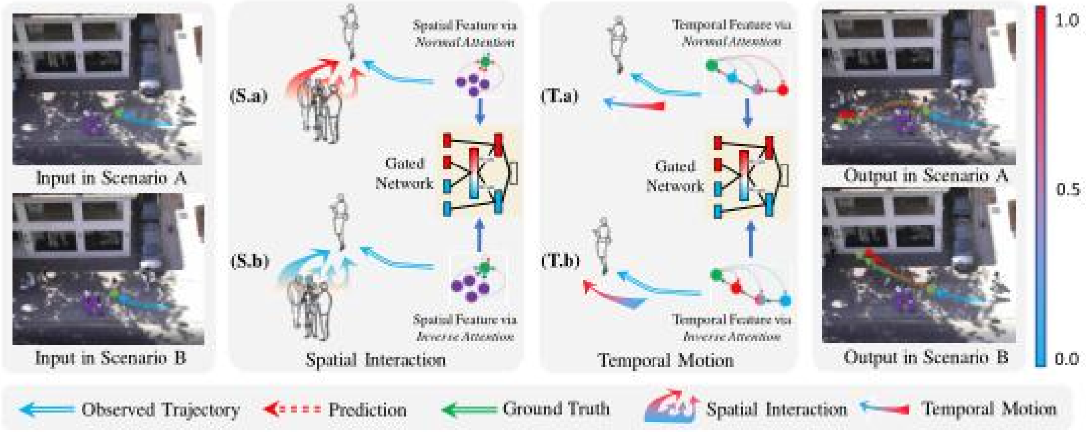


Figure 1: Real scenarios A and B represent frequent and peculiar modals with similar inputs and different outputs, respectively. **(Spatial Interaction)** Diverse spatial interaction, including frequent and peculiar spatial interaction, is generated by fusing complementary attention via the gated network. **(S.a)** Frequent interaction processed by normal attention. **(S.b)** Peculiar interaction processed by inverse attention. **(Temporal Motion)** Diverse temporal motions, including frequent and peculiar temporal motions, are generated by fusing complementary attention via the gated network. **(T.a)** Frequent motion processed by normal attention. **(T.b)** Peculiar motion processed by inverse attention.

The 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

Method

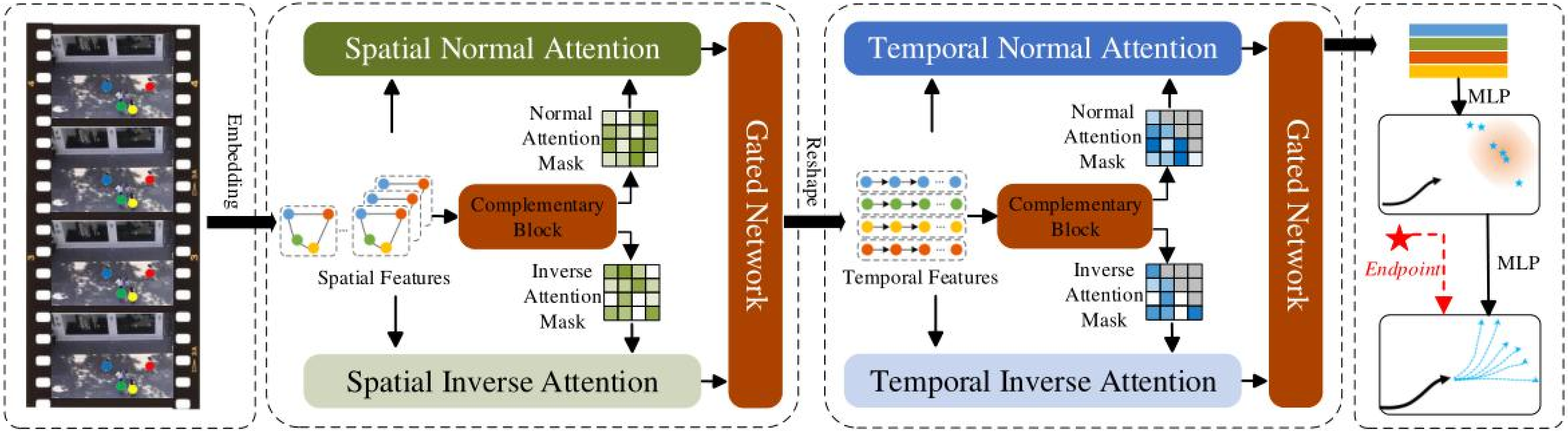
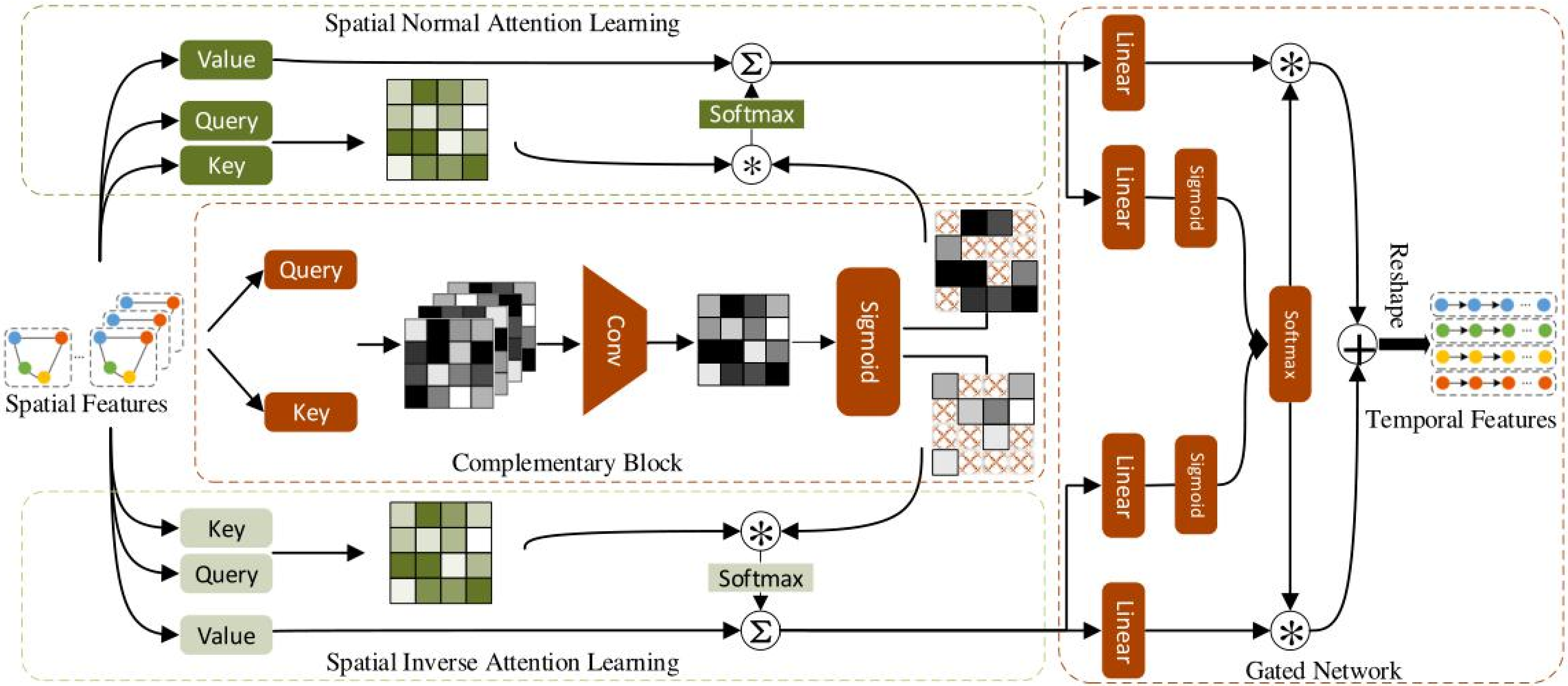
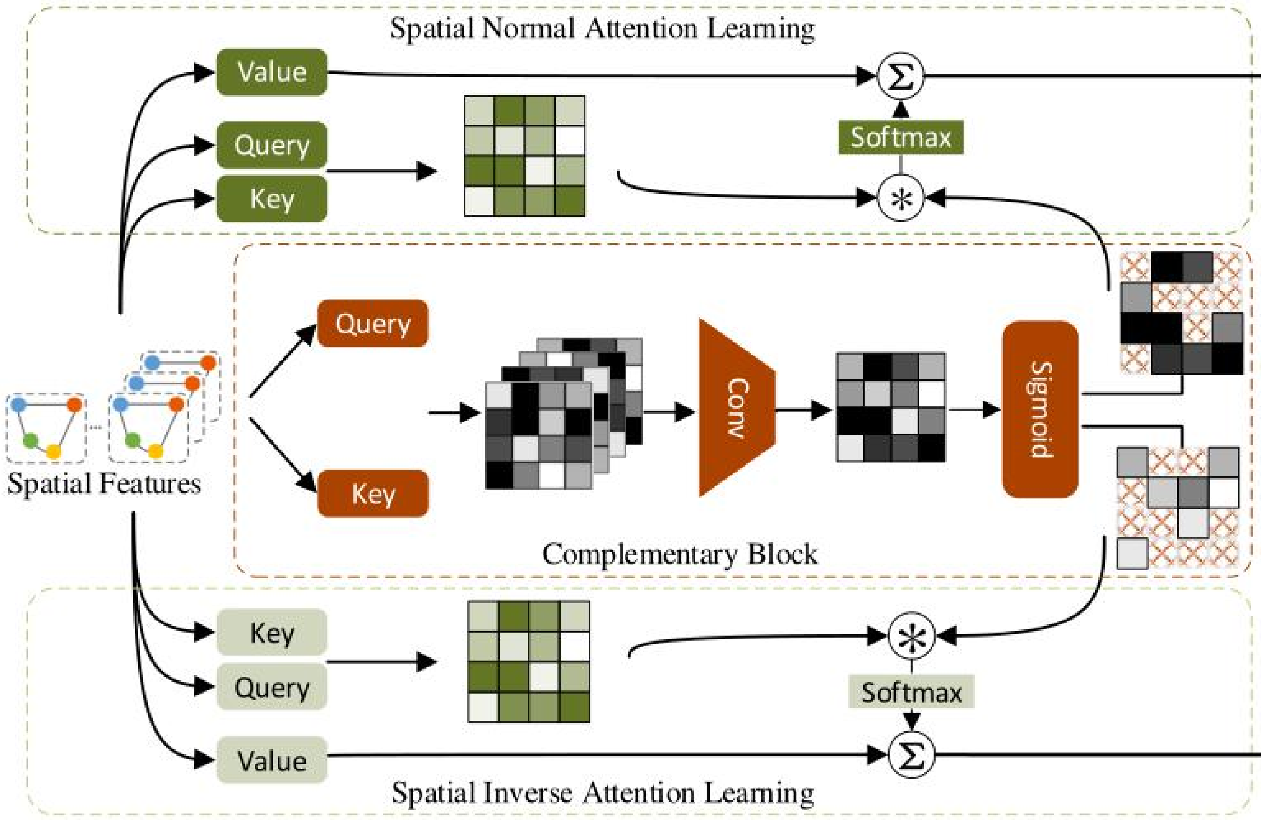


Figure 2: The framework of our CAGN. We first generate a pair of complementary masks through the complementary block to guide the normal and inverse attention. After that, we leverage a gated network to adaptively fuse the dual-path spatio-temporal features represents frequent and peculiar scenarios. 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 process for generating other future trajectory positions except the endpoint.

Method


Spatial Normal Attention Learning
Value
Query
Key
Softmax
Spatial Features
Query
Key
Conv
Sigmoid
Complementary Block
Key
Query
Value
Softmax
Spatial Inverse Attention Learning
Linear
Linear
Sigmoid
Softmax
Linear
Sigmoid
Linear
Reshape
Temporal Features
Gated Network

Method

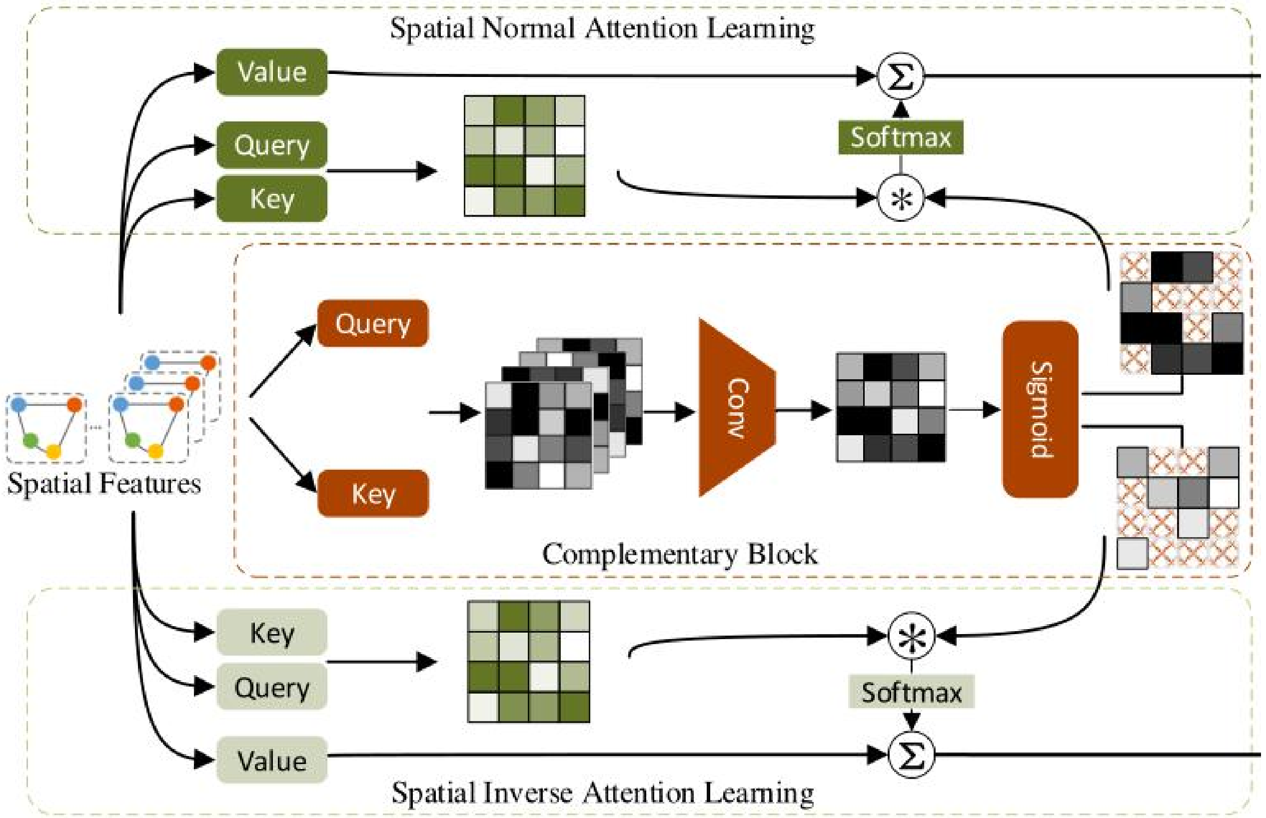


$$F_{in} = \mathrm{MLP}\left(\left\{ \left(x_t^n, y_t^n \right) \right\}_{n=1,t=1}^{N,T_{\text{obs}}} \right)$$

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

$$S \in \mathbb{R}^{H \times T_{obs} \times N \times N}$$

Method



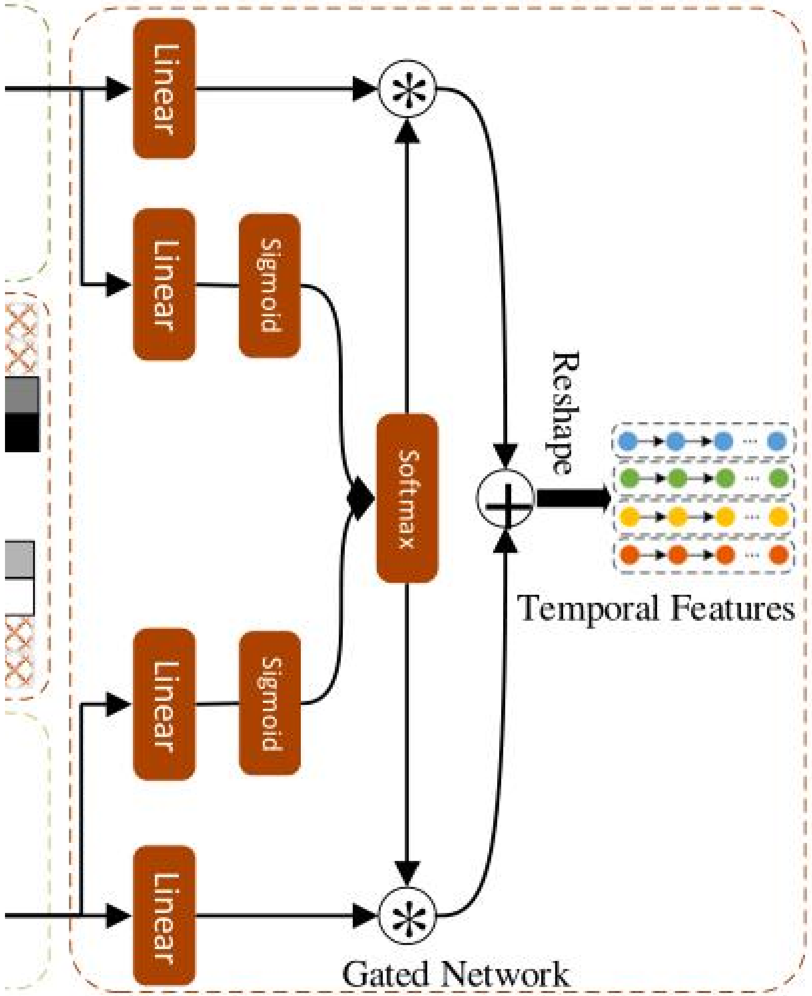
$$J = \delta\left(\mathrm{Conv}\left(S, \mathcal{K}\right)\right), \quad J \in \mathbb{R}^{T_{obs} \times 1 \times N \times N} \tag{2}$$

$$M_{\text{normal}} = \mathbb{I}\left\{J \le \xi\right\}, \quad M_{\text{inverse}} = \mathbb{I}\left\{(O - J) \le \xi\right\}, \tag{3}$$

Dual-Path Spatial Attention

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

Method

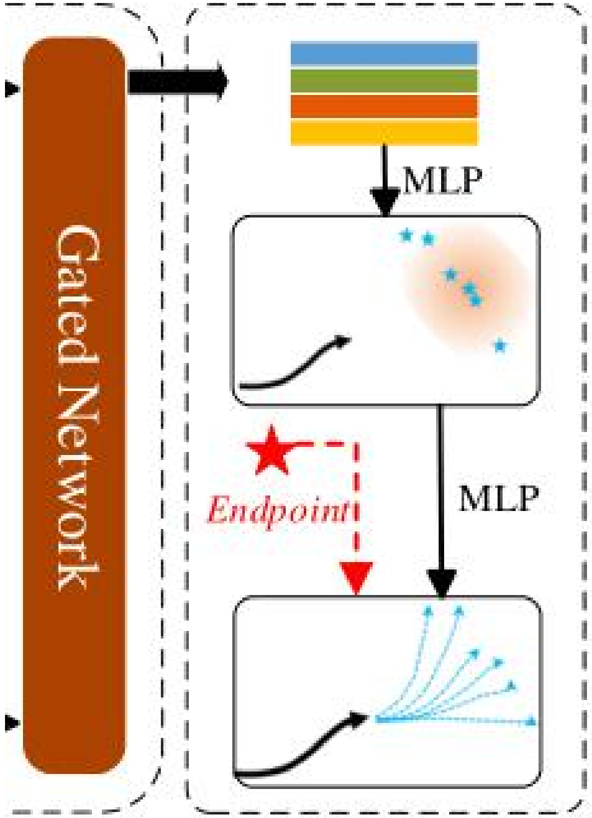


Gated Network

$$
\begin{aligned}
R_{\text{normal}}^{\text{spa}} &= \phi\left(F_{\text{normal}}^{\text{spa}}, W^{\text{r}}\right), \\
G_{\text{normal}}^{\text{spa}} &= \delta(\phi\left(F_{\text{normal}}^{\text{spa}}, W^{\text{g}}\right)),
\end{aligned} \tag{5}
$$

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

Method



Trajectory Prediction

$$\mathcal{L}_{EP} = -\log \sum_{k=1}^{K} \pi_k P\left((x_t, y_t) \mid \hat{\mu}_t, \hat{\sigma}_t, \hat{\rho}_t\right), t = T_{\text{pred}},$$

$$\mathcal{L}_{AL} = \frac{1}{T_{\text{pred}} - T_{\text{obs}} - 1} \sum_{t=T_{\text{obs}+1}}^{T_{\text{pred}-1}} \left((\hat{x}_t - x_t)^2 + (\hat{y}_t - y_t)^2\right),$$

$$\mathcal{L}_{CAGN} = \mathcal{L}_{EP} + \mathcal{L}_{AL}, \tag{7}$$

Experiments

Model	Venue	Year	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Linear Regression	CVPR	2016	1.33/2.94	0.39/0.72	0.82/1.59	0.62/1.21	0.77/1.48	0.79/1.59
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
SoPhie	CVPR	2019	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.51/1.15
PIF	CVPR	2019	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00
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.

Experiments

	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.

Experiments

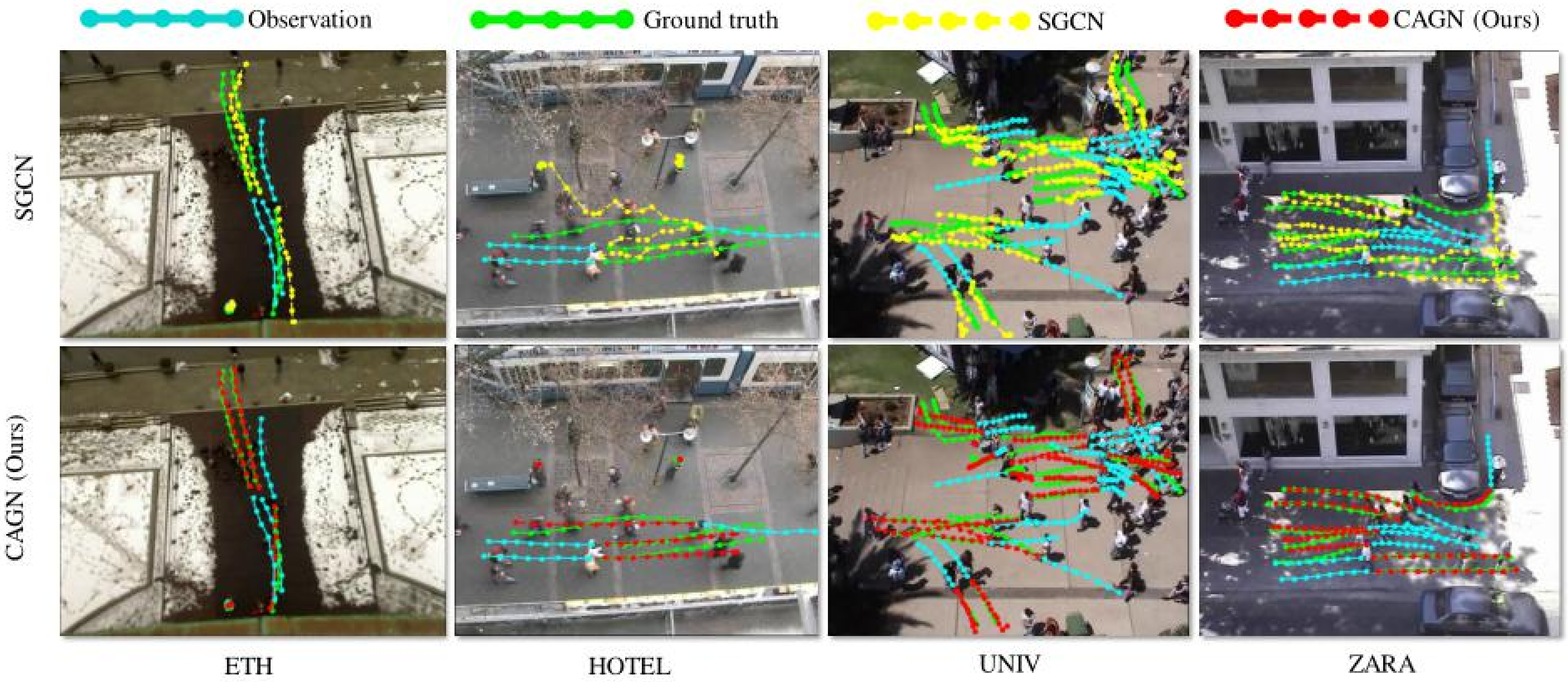


Figure 4: Visualization about performance. We visualize the pedestrian trajectory prediction, selecting the best result in 20 samples in different scenarios.

Experiments

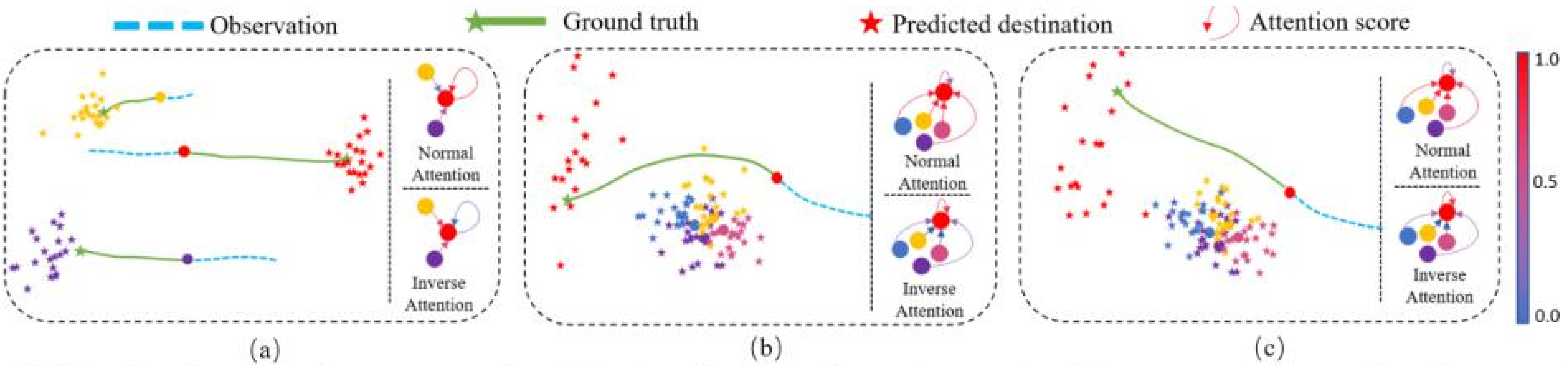


Figure 5: Visualization about inverse attention. We visualize the values of normal and inverse attention to show how our CAGN guides the model to learn different pedestrian movements.

Experiments

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

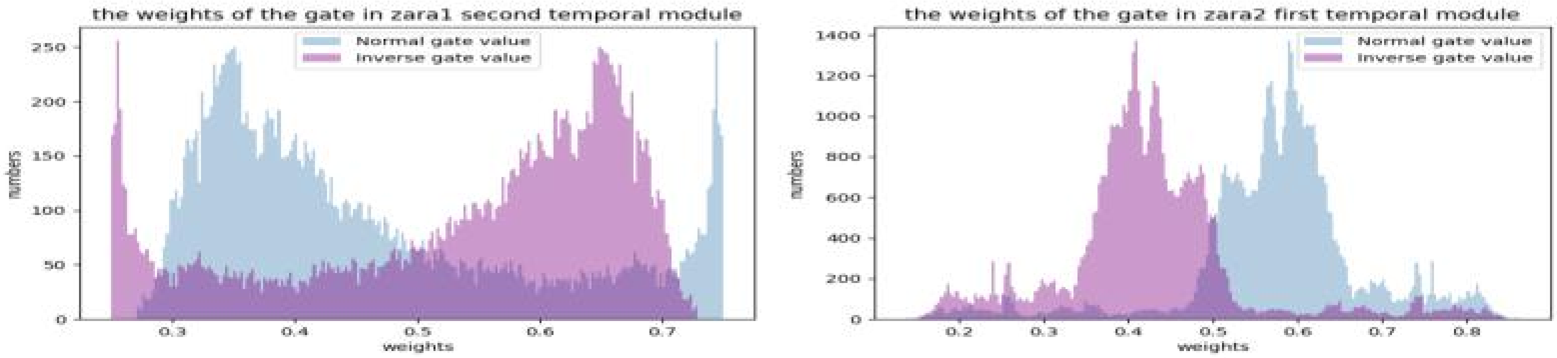


Figure 6: The distribution of gate value.

Thanks