Supplementary Material for BE-STI: Spatial-Temporal Integrated Network for Class-agnostic Motion Prediction with Bidirectional Enhancement

Yunlong Wang[1,2*†], Hongyu Pan[1†], Jun Zhu[1], Yu-Huan Wu[1,3*], Xin Zhan[1], Kun Jiang[2‡], Diange Yang[2‡]
[1]Alibaba DAMO Academy, [2]Tsinghua University, [3]Nankai University

1. Introduction

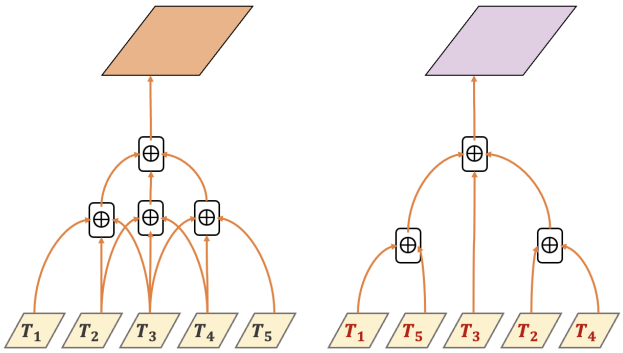
In this document, we present supplementary materials about ablation studies omitted from the main paper. In Sec. 2, the contribution of our proposed SeTE module is demonstrated in detail. In Sec. 3, detailed discussion on the toy example is presented.

2. Contribution of SeTE

In our main paper, we propose a spatial-enhanced temporal encoder (SeTE) to capture motion clues with the features of non-adjacent frames. As shown in Fig. 1b, we first apply a convolution layer with kernel size $2 \times 3 \times 3$ on the first and last frames. Then the same layer is applied on the second and penultimate frames. After that a $3 \times 3 \times 3$ convolution layer is applied to handle the generated feature maps together with the middle frame to capture the global feature.

Traditional temporal encoder (TTE) is usually built up with two stacked $3 \times 3 \times 3$ convolution layers [1, 2], which is applied to extract the global features of adjacent frames first and then merge them together, as shown in Fig. 1a.

We apply TTE and SeTE separately in our BE-STI framework to perform motion prediction. The performance of these two modules is shown in Tab. 1. As we can see, the proposed SeTE outperforms TTE by a large margin. Specifically, SeTE can reduce the mean prediction error by 0.0136m and 0.0020m for fast and slow moving objects, respectively. The experimental results show that SeTE contributes a lot to the motion prediction task and further demonstrate the effectiveness of spatial distinct features of non-adjacent frames.



(a) TTE (b) SeTE

Figure 1. Comparison between traditional temporal encoder (TTE) and our proposed spatial-enhanced temporal encoder (SeTE).

3. Toy Example

To intuitively show the influence of semantic information on motion prediction, we present a toy example in our main paper, where we feed MotionNet [2] with additional semantic segmentation ground truth together with the pseudo-images as input. Since it is hard to say whether the semantic information contributes to the motion task or the additional information makes the whole task easier, we try to answer this question by conducting additional experiments here. Considering that if the motion prediction task becomes easier with additional semantic information input, even the parameter number is slightly reduced, we can still achieve comparable performance with the full model. On the contrary, we can say the task is still hard while it benefits from the semantic information.

We conduct four tests on MotionNet [2] with different input and parameter numbers: (a) full model of MotionNet without additional input; (b) full model of MotionNet with additional semantic segmentation ground truth input; (c) MotionNet with 75% parameters and semantic segmentation ground truth input; (d) MotionNet with 50% parameters and semantic segmentation ground truth input. The performance of these four models is shown in Tab. 2. Comparing (b) with (a), we can see that with additional semantic information input, the motion prediction error of MotionNet [2] can be greatly reduced. Comparing (c), (d) with (b), the mean prediction error is increased by $0.027m$ and $0.0893m$ for fast moving objects when we reduce the parameters to 75% and 50%, respectively. Thus we can conclude that the motion prediction task itself is still hard with additional se-

*This work was done when Yunlong Wang and Yu-Huan Wu were research interns at Alibaba DAMO Academy.
†Equal contribution.
‡Corresponding author.

Module	Static		Speed ≤ 5 m / s		Speed > 5 m / s	
	Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓
TTE	**0.0242**	**0**	0.2395	0.0952	0.9214	**0.6147**
TeSE	0.0244	**0**	**0.2375**	**0.0950**	**0.9078**	0.6262

Table 1. Performance comparison of traditional temporal encoder and spatial-enhanced temporal encoder on motion prediction.

Method	Model	Parameters	Static		Speed ≤ 5 m / s		Speed > 5 m / s	
			Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓
(a)	MotionNet [2]	100%	0.0201	**0**	0.2292	0.0952	0.9454	0.6180
(b)	MotionNet+ GT_{seg}	100%	**0.0015**	**0**	**0.2139**	**0.0944**	**0.7990**	**0.6160**
(c)	MotionNet+ GT_{seg}	75%	0.0038	**0**	0.2141	0.0950	0.8260	0.6316
(d)	MotionNet+ GT_{seg}	50%	0.0037	**0**	0.2392	0.0957	0.8883	0.6696

Table 2. Performance of MotionNet [2] with different input data and parameter numbers.

mantic information input, while the performance improvement demonstrates that the semantic understanding of the scene can benefit motion prediction task.

References

[1] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2021. 1

[2] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11385–11395, 2020. 1, 2