Safety-aware Motion Prediction with Unseen Vehicles for Autonomous Driving
Supplementary Material

Xuanchi Ren[1*] Yang Tao[2*] Li Erran Li[3] Alexandre Alahi[4] Qifeng Chen[1]
[1]HKUST [2]Xi'an Jiaotong University [3]Alexa AI, Amazon [4]EPFL

A. Ablation Study for Attention

In this section, we perform an ablation study on the different designs of the self-attention unit in our model. We compare our design with two alternatives, the non-local design [1] and the one-stream design (one CNN). The CNN used in our model is composed of three 3×3 convolution layers with a stride of 1 and padding of 1, of which the activation function is ReLU. As shown in Table 1, our design significantly outperforms the other two designs. One possible reason is that the reception field after the dilated bottleneck is large enough, and thus the extracted feature contains both local and global features. Therefore, we do not need to perform non-local operations. Moreover, two-branch CNNs provide more parameters to learn the attention map. The only exception is the MSE, and the underlying reason may be the false-positive unseen vehicles our model predicted. The example of false-positive unseen vehicles can be observed in the fourth column in Figure 2.

Method	MR (%) ↓	Aggressiveness ↓	$UR_{0.3}$ (%) ↑	$UR_{0.5}$ (%) ↑	$UR_{0.7}$ (%) ↑	MSE↓
Ours w/ non-local	5.48	2.58	39.45	22.16	6.92	6.70
Ours w/ one-stream	5.30	2.60	40.01	22.87	9.70	**6.52**
Ours	**1.37**	**2.48**	**63.28**	**43.48**	**18.85**	10.61

Table 1. Ablation study for the self-attention unit. Our two-stream version outperforms the non-local version and one-stream version.

B. Key Hyper-parameter Tuning

We can vary γ_h to trade-off between safety and aggressiveness. As shown in Table 2, when γ_h becomes larger, the MR decreases, and the values of Aggressive and MSE increase, which shows that the prediction becomes more conservative. For the loss weight γ_u for the unseen loss, we empirically find it better to set $\gamma_u = \gamma_h$.

Method	MR (%) ↓	Aggressiveness ↓	$UR_{0.3}$ (%) ↑	$UR_{0.5}$ (%) ↑	$UR_{0.7}$ (%) ↑	MSE↓
γ_h = 0	18.00	1.39	36.98	20.73	6.76	6.55
γ_h = 10	7.62	1.45	49.95	29.96	11.69	9.25
γ_h = 100	6.89	1.26	53.14	32.63	11.86	9.48
γ_h = 1000	1.37	2.48	63.28	43.48	18.85	10.61
γ_h = 10000	0.63	2.54	60.35	50.76	23.24	12.60

Table 2. Ablation study for the weight for hard loss. Here we set $\gamma_u = \gamma_h$. When the weights become larger, the MR decreases, and the Aggressiveness and MSE increase, indicating the prediction tends to be more conservative.

C. More Qualitative Results

Here we provide more qualitative results on the nuScenes dataset as shown in Figure 1. We also provide qualitative results on the Lyft dataset as shown in Figure 2. We can observe that our prediction is more accurate than the baselines and includes

*Equal contribution

the predictions of unseen vehicles from these qualitative results. Besides, the earliest occupancy map captures multi-modal predicted motion in a single output. Furthermore, the motions predicted by our method tend to have a bit larger range, which shows that our model captures uncertainty of prediction. Therefore, our method can generalize well to complex urban environments by modeling motion prediction as an image-to-image translation problem.

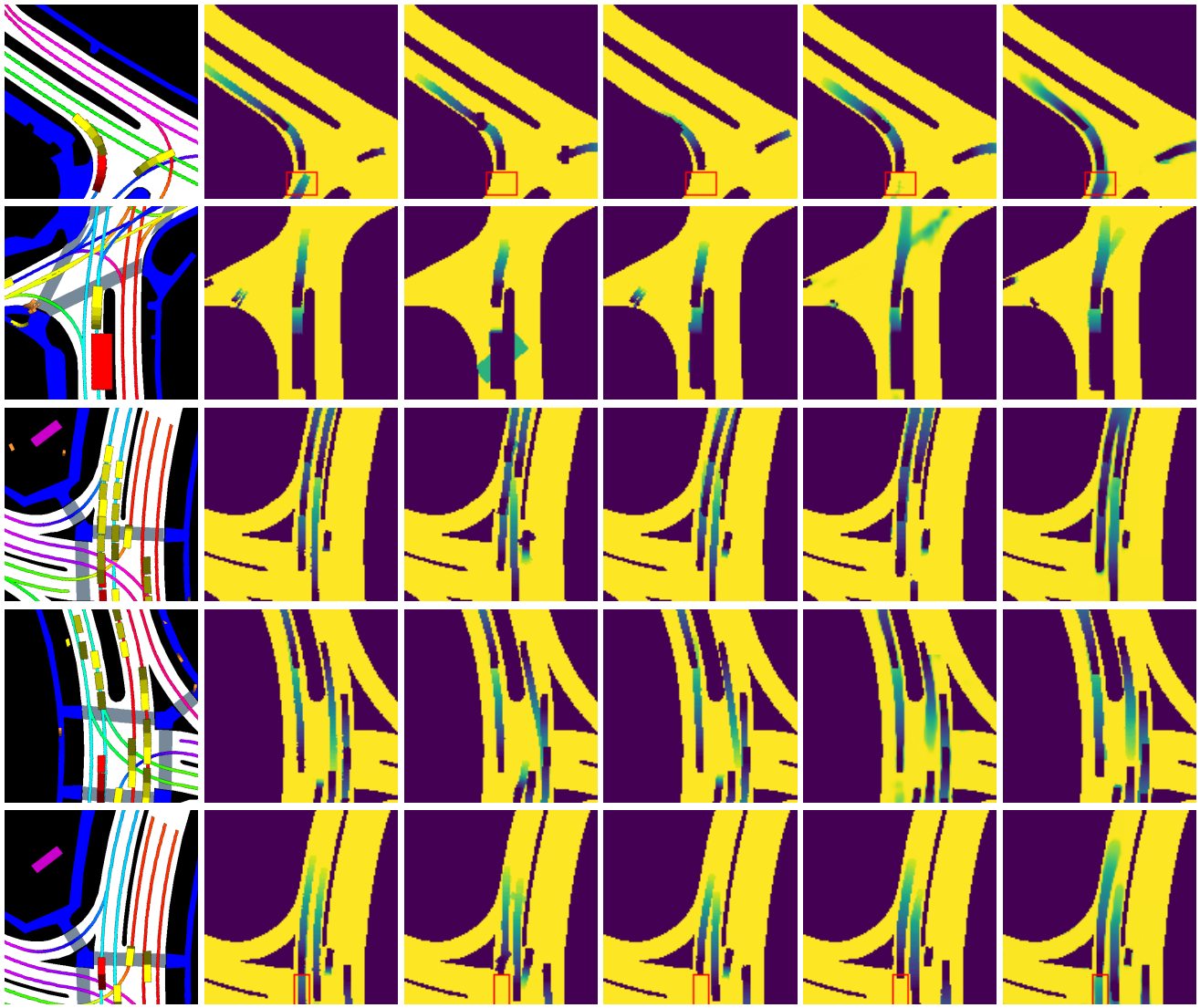

Figure 1. Visual comparisons between ours and other baselines on the **nuScenes** dataset. The unseen vehicles are annotated with red bounding boxes. All the prediction results are visualized with the earliest occupancy maps. Our predictions are earlier but as accurate as possible.

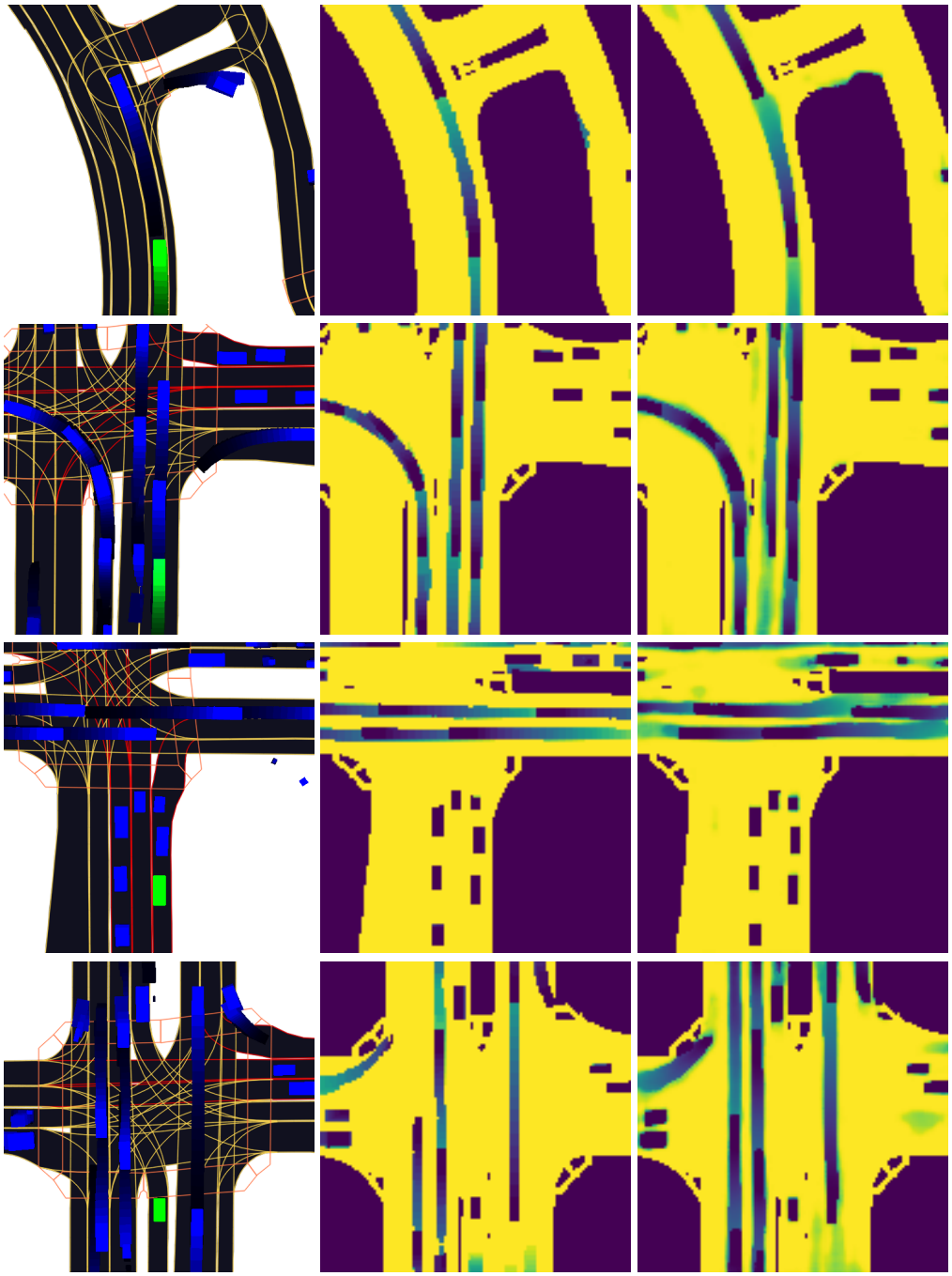

Figure 2. Visualization on the **Lyft** dataset. The inputs are rasterized using the official API of the Lyft dataset. The fourth row is a failure case with predicted false positives of unseen vehicles.

D. Representation Definition

D.1. Representation Illustration

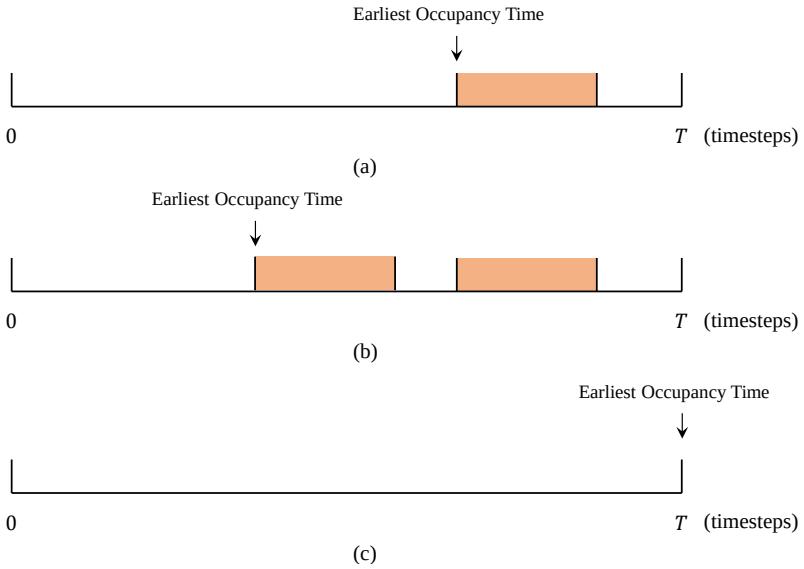

Figure 3. Illustration for the earliest occupancy time at a specific location. The orange region indicates that this location is occupied. (a) This location is once occupied. (b) This location is occupied multiple times. (c) This location is never occupied.

As mentioned in Section 3.2, we formulate the earliest occupancy map $E(x, y)$ as

$$E(x, y) = \min(\{\Delta t | O_{t+\Delta t}(x, y) = 1\} \cup \{T\}), \forall (x, y) \in I, \tag{1}$$

where $t + \Delta t$ is a timestep between t and $t + T$. Here we explain the earliest occupancy time using Figure 3 in different cases.

D.2. Discussion on Representation

In the main paper, we introduce and discuss the **earliest occupancy map** in the case of predicting short-term future T timesteps. For each location, the duration between the earliest occupancy time and the end of the prediction horizon can be considered occupied, which adds a constraint to the planner. When the Missing Rate for the earliest occupancy map is nearly zero, the planning will be safe. However, this constraint may be too strict for the planner. Ideally, the occupancy map should be conservative to account for safety and leave a large feasible solution space for the planner at the same time.

Latest free map. To relax the constraint, we present a symmetric motion representation of earliest occupancy map $E(x, y)$, **latest free map** $L(x, y)$. For each position, the duration of occupancy can be calculated from the two maps. For those positions not being occupied in the future, we set $E(x, y) = T$; thus, we also set $L(x, y) = T$, which means that the duration being occupied is 0. For those positions predicted to be occupied, $L(x, y)$ indicates the first time they will no longer be occupied. Thus, we formulate the latest free map $L(x, y)$ as

$$L(x, y) = \begin{cases} T, E(x, y) = T \\ \min\{\Delta t | O_{t+T-\Delta t}(x, y) = 1\}, \quad E(x, y) < T \end{cases} \tag{2}$$

where $t + T - \Delta t$ is a timestep between t and $t + T$. Similar to the constraint on prediction of $P_E(x, y)$[1], the prediction $P_L(x, y)$ should be later than the ground truth $L(x, y)$ but as accurate as possible. Therefore, we also can formulate it by

[1]To differentiate prediction of latest free map, we use $P_E(x, y)$ to represent $P(x, y)$ in our main paper.

defining a hard loss $\mathcal{L}_h^L$ and a soft loss $\mathcal{L}_s^L$:

$$\begin{aligned}\mathcal{L}_h^L &= \sum_{(x,y)\in I} \mathbb{1}(P_L(x,y) < L(x,y)).\\ \mathcal{L}_s^L &= \sum_{(x,y)\in I} P_L(x,y).\end{aligned} \tag{3}$$

We use the hard loss to penalize predictions earlier than the ground truth. Note that the hard loss constrains the prediction $P_L(x,y)$ to be lower bounded by ground truth $L(x,y)$, we thus add a soft loss to make the prediction close to the ground truth. Similar to $E(x,y)$, we adopt a MSE term to stabilize the training,

$$\mathcal{L}_{rec}^L = \sum_{(x,y)\in I} \|P_L(x,y) - L(x,y)\|^2. \tag{4}$$

Combining **earliest occupancy map** and **latest free map**, the timesteps between these two maps can be considered as being occupied for each location. How to use these two losses will be discussed in the following sections. (added to the original one)

D.3. Additional Experiment

Method	MR* (%) ↓	MSE_1↓	MSE_2 ↓
Ours	4.27	6.48	3.98

Table 3. Result for modified model to predict both earliest occupancy map and latest free map.

Model. For the model, we modified the customized U-Net in the main paper. We add one additional output channel. The first channel is for the predicted earliest occupancy map $P_E(x,y)$, and the second channel is for the predicted latest free map $P_L(x,y)$. The full objective function for the model is:

$$\mathcal{L} = \gamma_h \mathcal{L}_h + \mathcal{L}_s + \mathcal{L}_{rec} + \gamma_h \mathcal{L}_h^L + \mathcal{L}_s^L + \mathcal{L}_{rec}^L, \tag{5}$$

where the weight for hard losses $\gamma_h = 1000$. In this section, to simplify the problem, we do not consider the unseen vehicles.

Metric. For the metric, we modified the **Missing Rate** (MR) in the main paper to **Missing Rate*** (MR*). MR* indicates the percentage of the predicted earliest occupancy map that is later than the ground truth or the predicted latest free map that is earlier than the ground truth. Furthermore, we also adopt **MSE** metric to evaluate the accuracy. We use $\textbf{MSE}_1$ to indicate the distance between the predicted earliest occupancy map and ground truth earliest occupancy map and use $\textbf{MSE}_2$ to indicate the distance between predicted latest free map and ground truth latest free map.

Evaluation. The result is shown in Table 3. Considering there is no suitable baseline for this new task, we only evaluate our model to show it can complete this more complex task. The missing rate is higher than only predicting the earliest case; however, it is still very small. The MSE metric for both the maps is small, which means the range between these two maps is tight, and the predictions are accurate.

References

[1] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.