Supplementary Materials for Instance-Centric Spatio-Temporal Modeling for Online Vectorized Map Construction

Anonymous Authors

1 ABSTRACT

This document supplements our main submission "IC-Mapper: Instance-Centric Spatio-Temporal Modeling for Online Vectorized Map Construction". We first conducted further exploration and ablation studies on the design of each module, then summarized some important details during the model training process, and finally provided visual images to illustrate the entire process of online map construction.

2 ADVANCED EXPLORATION OF MODULE DESIGN AND FUNCTIONALITY

2.1 Reflections on End-to-End Map Tracking Tasks

The primary goal of the tracking task is to ascertain the matching relationships between map instances across consecutive frames, thereby laying the foundation for subsequent instance-centered map construction. To avoid complex post-processing, we have implemented an end-to-end detection and tracking network framework. Generally, joint training of multiple tasks on a shared network can lead to reduced accuracy in individual tasks. However, as demonstrated by the experimental results in the main paper, our designed temporal tracking module achieves good performance in both detection and tracking tasks. This section will compare the implementation details of different end-to-end tracking networks and further analyze the reasons behind their performance disparities.

2.1.1 Comparison of Different Temporal Modeling Design Approaches. Figure 1 provides the implementation details of three temporal modeling approaches. Figure 1(a) displays the temporal modeling method used by StreamMapNet, which maintains the features of the TopK scoring instances from the historical sequence and combines them with the initial queries of the current frame to perform the detection task. Since the TopK instance selection mechanism does not explicitly calculate the matching relationships between instances, this method primarily uses implicit temporal modeling to enhance detection performance. Figure 1(b) illustrates the supervised training method in MOTR, where once each instance is matched to the corresponding ground truth object, it is locked in for subsequent training, and only newly generated detection instances perform Hungarian matching with the remaining labels. Our proposed IC-Mapper combines these designs. As shown in Figure 1(c), we retain the TopK transmission mechanism to ensure detection performance and introduce an additional learnable association module to calculate the matching relationships between detection instances and historical tracking instances. This module is supervised using cross-entropy loss, and specific design details can be referred to in the main paper.

Table 1: Comparison of Detection and Tracking Metrics Between MOTR and Our IC-Mapper.

Method	MOTA	MOTP	ID-switch	mAP
StreamMapNet	-	-	-	34.1
MOTR-track	0.40	5.08	2.38	22.7
IC-Mapper(Ours)	0.55	3.18	1.82	35.0

2.1.2 Analysis of the Reasons for the Decline in MOTR Detection Performance. As shown in Table 1, using a similar end-to-end detection and tracking model, our proposed IC-Mapper surpasses the MOTR[2] approach in all detection and tracking metrics. Compared to the original StreamMapNet, our method also achieves an improvement of nearly one point in mAP.

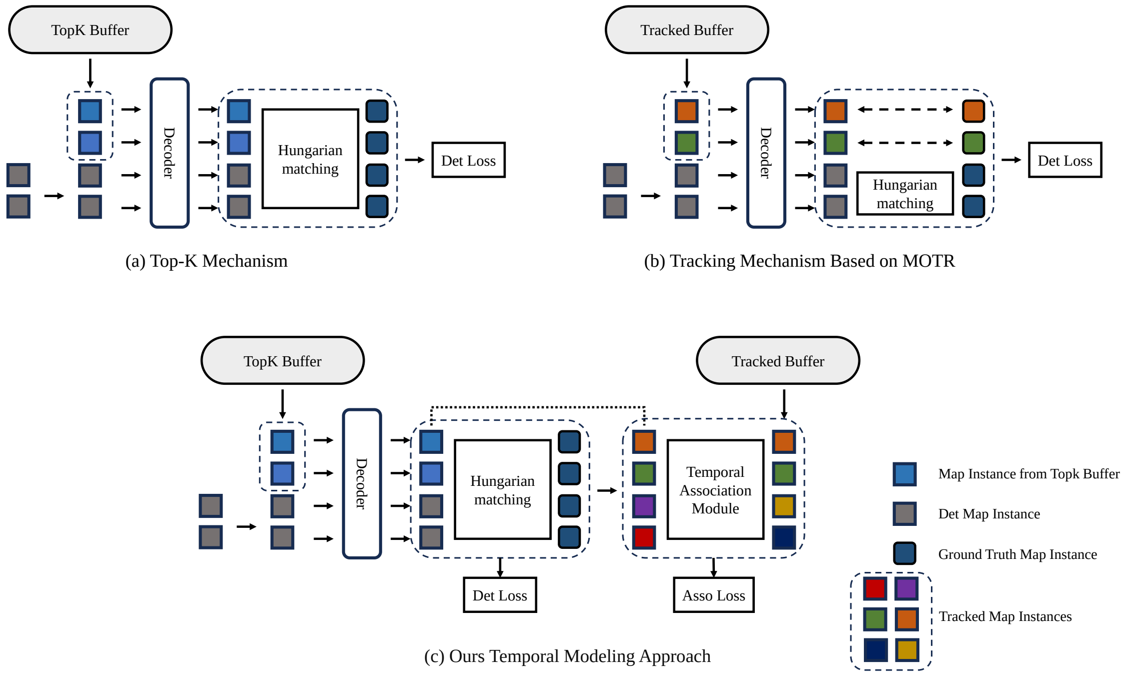
We attribute the performance disparity primarily to two reasons: 1. During the training phase, MOTR assigns fixed true labels based on IDs to tracking instances for supervised training. This causes early errors in matching to accumulate over subsequent frames, thereby affecting detection performance. 2. In detection tasks for dynamic targets, the same instance typically appears as the same box shape over time, showing little deformation. However, as shown in Figure 3, road elements often cover a large area, leading to significant deformation of the same instance across different frames, resulting in substantial variations. In MOTR, the network has to decode different feature shapes using the same tracking query, which increases the learning burden. In contrast, implicitly incorporating historical query features using the TopK method avoids this issue and maintains high detection accuracy. Building on this, we introduce an additional module to perform explicit tracking tasks, thus achieving good detection and tracking performance.

2.2 Further Exploration of Spatial Fusion Module Potential

The fusion module primarily integrates current detection results with historical maps, using spatial continuity of instances to enhance map feature detection accuracy within localized perception patches. In the main paper, we use the point set output from the current detector as the initial query, and compute cross-attention with the corresponding sampled point set from the historical map to achieve feature fusion. To further explore the potential of the fusion module, as shown in Figure 2, we introduce deformable attention computed between the BEV features generated by the detection module and the instance point set, and evaluate the detection accuracy before and after fusion.

$$\mathrm{SCA}(Q, F_{bev}) = \sum_{j=1}^{N_p} \mathrm{DA}(Q, \mathcal{P}(p, j), F_{bev}), \tag{1}$$

where Q represents the initial query encoding corresponding to the current instance, p is the position of j-th point of the instance,



(a) Top-K Mechanism

(b) Tracking Mechanism Based on MOTR

(c) Ours Temporal Modeling Approach

Figure 1: Three instance-level temporal modeling approaches. (a) In StreamMapNet[1], instance features are implicitly modeled temporally using a Topk mechanism. (b) In MOTR[2], a tracking buffer is maintained to store track instances, and tracking instances are bound with corresponding labels for calculating the detection loss. (c) We retain the Topk implicit temporal modeling from StreamMapNet and introduce a temporal association module to calculate the matching relationships between detection instances and tracking instances.

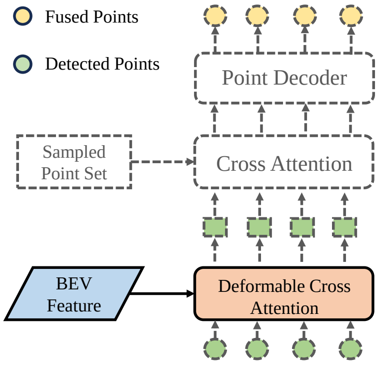


Figure 2: Building on the calculation of cross-attention with sampled point sets from historical maps, we additionally introduce deformable cross-attention using the BEV intermediate features from the detector and the current point set. This further enhances the effectiveness of the fusion module.

$\mathcal{P}$ represents a learnable mapping function applied to the initial coordinate point p, F_{bev} is the BEV feature generated by the detector, N_p is the number of reference points for each query, and DA means the operation of deformable attention.

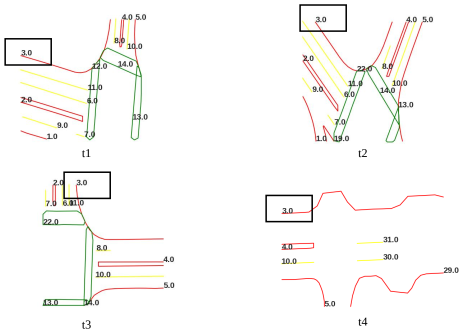


Figure 3: We extracted four frames of localized vector map ground truths from the same sequence, focusing on specific areas like the road boundary labeled with ID 3.0 in a black box. This example shows substantial variations in the shape of the same map instance when observed at different times and positions.

We initially conduct joint training of the detection and temporal association modules, then fix the parameters of the pre-trained model and introduce different fusion modules for fine-tuning training. As shown in the Table 2, introducing historical sampled point sets to the original model yields a slight improvement in detection accuracy, which is further enhanced by integrating BEV features. The accuracy gains from the fusion module are attributed to two

Table 2: Comparison of Detection results of different fusion module designs. 'w/o fusion represents the model pre-trained without fusion module; 'fusion (points)' means the fusion module only include the attention operation with sampled points, 'fusion (points+bev)' further introduces bev features in the fusion module.

Method	mAP
w/o fusion	34.95
fusion (points)	34.99
fusion (points+bev)	35.32

factors: first, the integration of spatial-dimensional features into the existing instance encoding introduces more priors; second, feature encoding and decoding are performed on the basis of instance point sets, which, combined with the previous instance-based encoding and decoding, constructs a detailed map feature modeling process from coarse to fine.

2.3 What affects the final mapping accuracy?

Since our task focuses on constructing a global vector map, the initial intent behind designing the temporal association and spatial fusion modules was to enhance the final mapping accuracy by achieving higher detection precision and better association accuracy. Here, we conduct further experiments to verify the impact of different detection and tracking accuracies on the final mapping task.

Table 3: The impact of different detector performances on the final mapping accuracy under the same tracker.

Detector mAP	**Tracker**	mCD
35.0	Post-track	5.35
34.6	Post-track	5.46

Table 4: The impact of different track performances on the final mapping accuracy under the same detector.

Detector mAP	**MOTP**	**mCD**
35.0	2.4	4.7
35.0	3.7	5.6

In Table 3, we use the same post-processing tracker to test the variation in final mapping accuracy (mCD) under different detection accuracies. In Table 4, we fix the accuracy of the detector and adjust the threshold to achieve different tracking performances, evaluating the final mapping accuracy. The experimental results indicate that both detection and tracking accuracies are positively correlated with the final mapping accuracy. This finding strongly supports the design of our end-to-end mapping framework and encourages the community to design sub-modules with higher detection and tracking performances to improve the final mapping metrics.

3 ABLATION STUDY OF CRITICAL EXPERIMENTAL DETAILS

The training method has a significant impact on model performance. This section summarizes several important experimental details to help the community better understand and expand upon this work.

3.1 The impact of geometric metrics in the temporal association module

We discovered that adjustments to geometric metrics during the joint training of detection and tracking affect the performance of both. We conducted an ablation study, as shown in the Table 5, where 'detach' indicates that the gradients of reference points used for calculating geometric metrics in the tracking module are truncated. The experimental results indicate that the geometric association component in the tracking module affects the accuracy of the detector. A geometric association module that allows gradient backpropagation can achieve higher tracking performance but at the expense of some detection accuracy.

Table 5: The impact of The impact of geometric metrics on the detection and tracking performance. '(detach)' indicates that the gradients of reference points used for calculating geometric metrics in the tracking module are truncated.

Method	**Tracker**			**mAP**
	MOTA↑	MOTP↓	ID-switch↓	
Ours	0.54	**3.17**	**1.81**	34.95
Ours(detach)	0.54	3.89	4.73	**35.13**

3.2 The impact of single-frame pre-training

In the original StreamMapNet[1], the detector is pre-trained using single-frame images, followed by training with multi-frame images. However, in our framework, we found that eliminating the single-frame pre-training phase after introducing the temporal association module can lead to greater accuracy gains. Specific experimental results are shown in the Table 6.

Table 6: Verify the impact of single-frame detection pre-training on the final detection accuracy. "w/" denotes with, and "w/o" denotes without.

Method	mAP
w/ single pre-train	32.92
w/o single pre-train	34.95

3.3 The impact of multi-stage training methods on results.

Overall, we adopted a two-stage training approach and experimented with two different training strategies. The first strategy involved jointly training the detection and tracking modules during the first phase. We used the Adam optimizer across 8 GPUs with an initial learning rate of 5e-4 for 24 epochs. In the second phase, we froze the parameters of the detection and tracking modules and fine-tuned the fusion module with an initial learning rate of 1e-4 for another 24 epochs. The second strategy involved training the detection and tracking modules in the first phase with an initial learning rate of 5e-4 for 16 epochs. In the second phase, we introduced the fusion module and conducted joint training with an initial learning rate of 3e-4 for 36 epochs. According to the detection accuracy metrics presented in Table 7, the second training strategy yields

Table 7: The impact of multi-stage training methods on results. '1' and '2' correspond to the two training strategies discussed in the article. (In this experimental setup, the gradients of the reference points in the temporal association module are truncated, resulting in a higher mAP metric.)

Training Strategies	mAP
1	35.13
2	36.31

greater benefits. However, this has not been proven to be the optimal training strategy. Training strategies for such multi-module, end-to-end networks require further exploration.

4 EXTENDED VISUALIZATION RESULTS

We demonstrated the sequence of input images along with the real-time process of map detection and updating as shown in Figure 4, 5, 6, 7.

REFERENCES

[1] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. 2024. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *WACV*. 7356–7365.

[2] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. 2022. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*. 659–675.

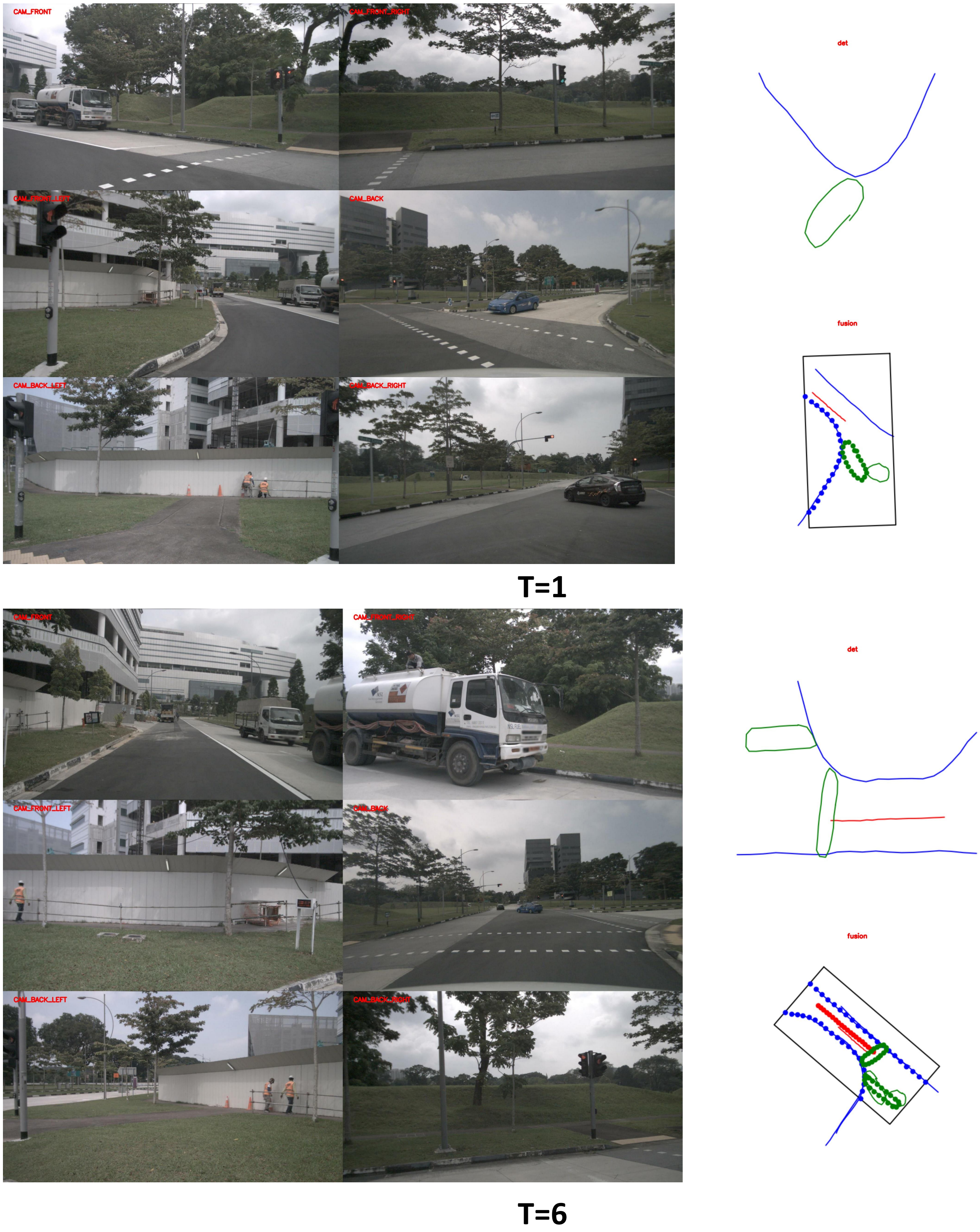


Figure 4: The visualization results in scene-0001 for frame 1 and 6, with the left side displaying the input image. On the right side, the top part shows the real-time detection results, while the bottom part illustrates the process of map updating by the fusion module.

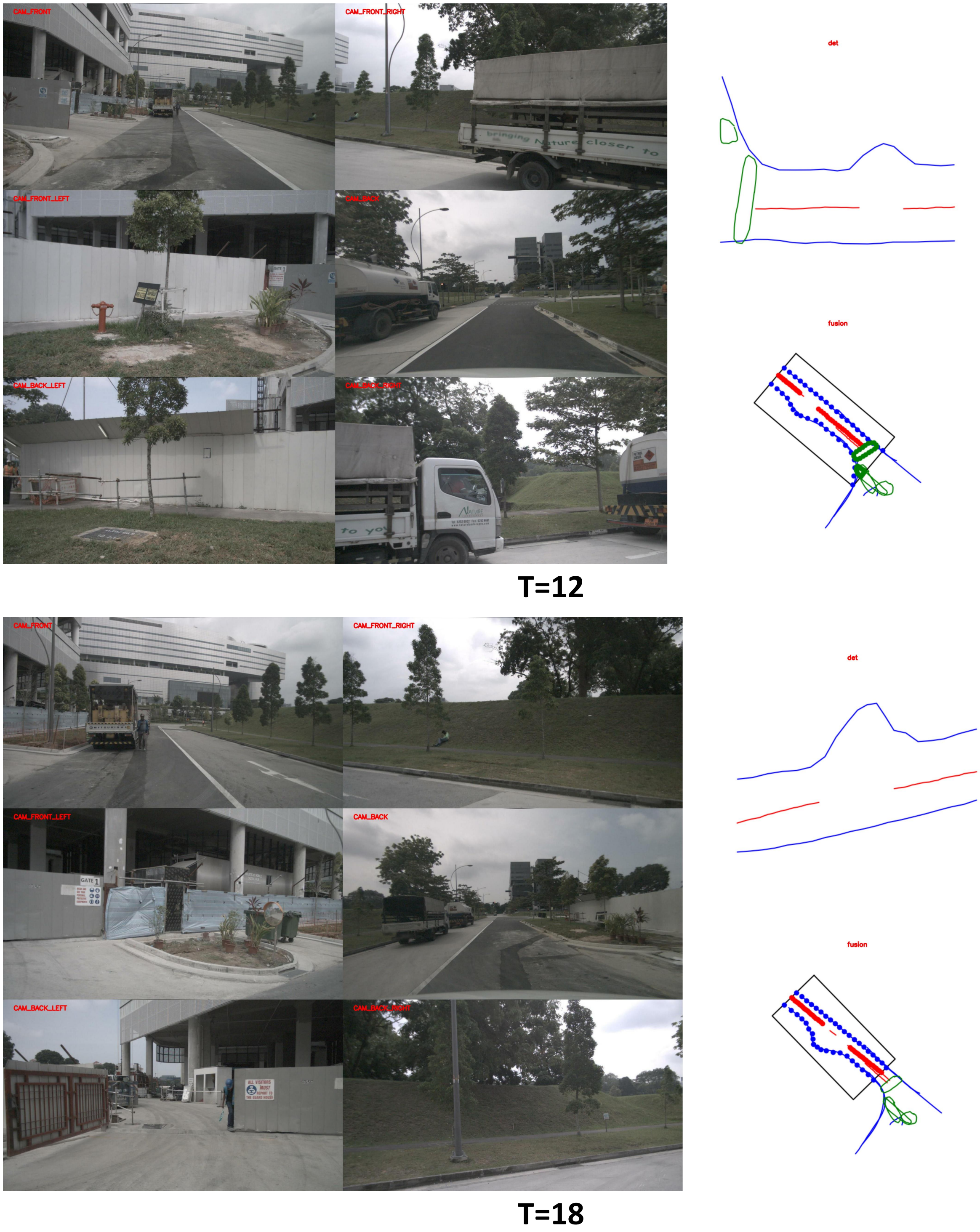


Figure 5: The visualization results in scene-0001 for frame 12 and 18, with the left side displaying the input image. On the right side, the top part shows the real-time detection results, while the bottom part illustrates the process of map updating by the fusion module.

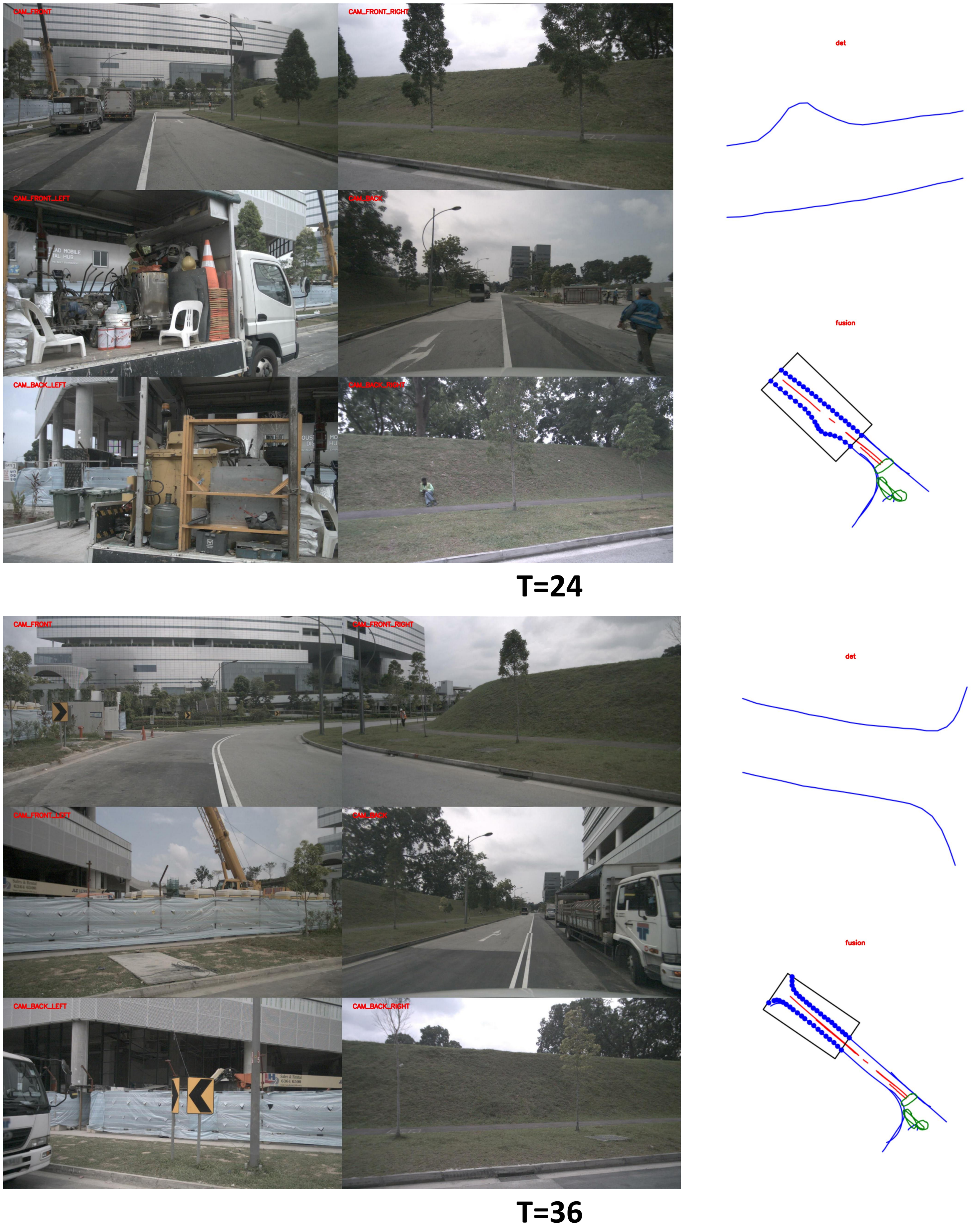


Figure 6: The visualization results in scene-0001 for frame 24 and 36, with the left side displaying the input image. On the right side, the top part shows the real-time detection results, while the bottom part illustrates the process of map updating by the fusion module.

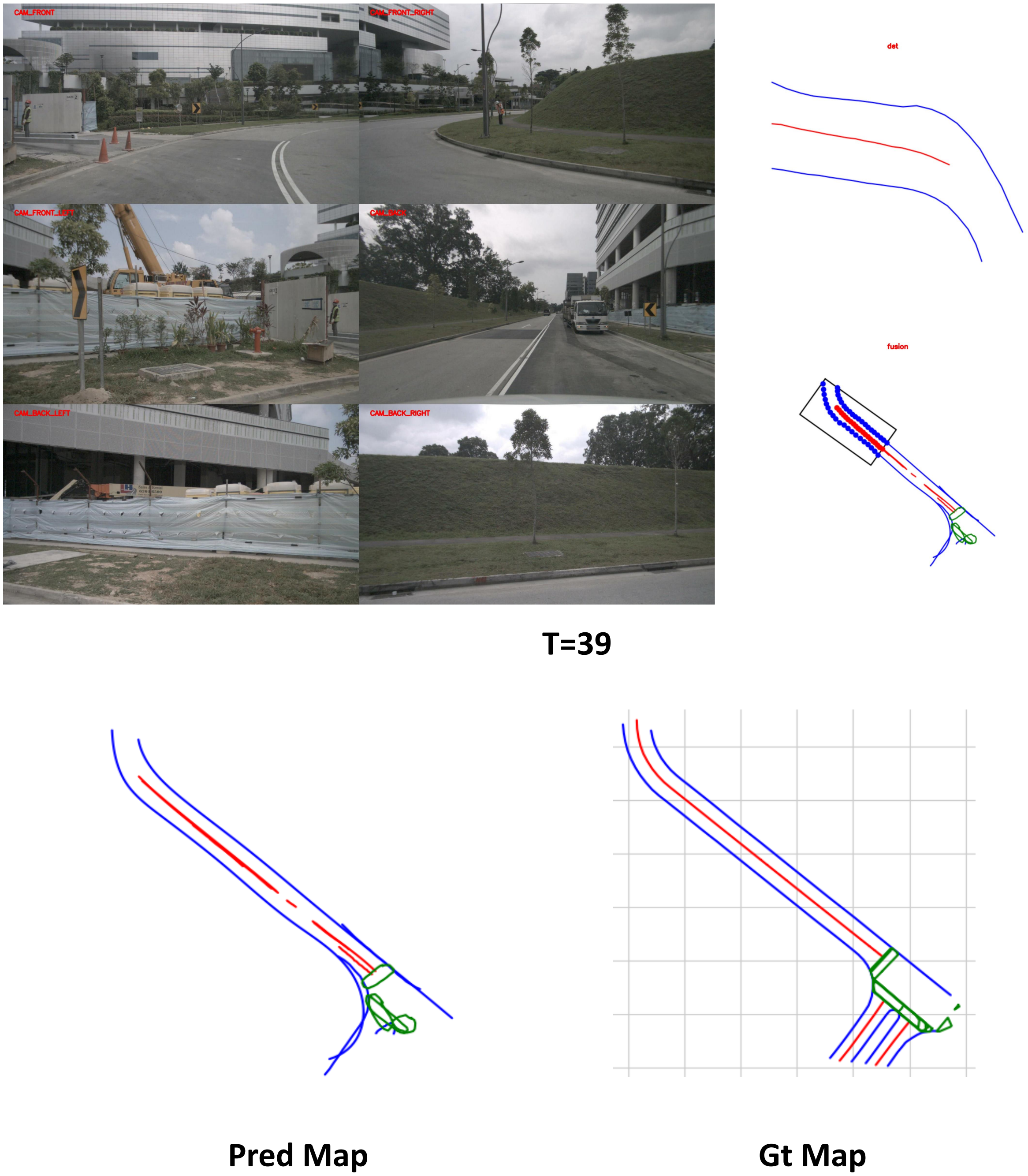


Figure 7: The visualization results in scene-0001 for frame 39, with the left side displaying the input image. On the right side, the top part shows the real-time detection results, while the bottom part illustrates the process of map updating by the fusion module. The bottom section displays a comparison between the final map reconstruction result for the entire scene-0001 and the true global map labels.