MCBLT: Multi-Camera Multi-Object 3D Tracking in Long Videos

Supplementary Material

A. Experiments on Mutli-View 3D Detector

A.1. 3D Object Detector Comparisons

We conducted experiments on different settings for our 3D object detector in Tab. 7. We considered both BEVFormer [18] and BEVFormer v2 [37] with ResNet-50, ResNet-101, and V2-99 backbones. The maximum number of cameras for training is reported for each experiment due to the limitation of H100 GPU memory. All experiments were trained for 24 epochs with a learning rate of 2×10^{-4}.

Method	Backbone	Max # of Cameras	mAP
BEVFormer [18]	ResNet-50	16	83.14
BEVFormer [18]	ResNet-101	15	88.64
BEVFormer v2 [37]	ResNet-50	15	82.78
BEVFormer v2 [37]	ResNet-101	15	85.03
BEVFormer v2 [37]	V2-99	14	79.95

Table 7. 3D object detection results with different detectors on a customized AICity'24 validation set.

Compared with BEVFormer, BEVFormer v2 receives relatively lower mAP by adding the perspective supervision. Therefore, the perspective supervision may not be helpful for our MTMC application for model convergence. As for the V2-99 backbone, we need to decrease the number of camera views during the training to fit our GPU memory of around 80 GB. This will downgrade the detection performance significantly. In the future, we will improve the memory efficiency to make it possible to utilize larger image backbones.

A.2. Scene Re-Centering for BEVFormer

The definition of the BEV coordinate system is important for BEVFormer training. In the original autonomous driving settings, the origin is located on the ego-vehicle, which is the center of the area to be perceived. In our MTMC settings, we define the origins of the multi-camera scenes as the centers of the floor plans and transform the annotations and calibration matrices to the newly defined BEV coordinates. We call this step "re-centering".

In Tab. 8, we evaluate the model performance on the WildTrack dataset before and after this re-centering step. Before the re-centering, the origin was defined at the corner of a scene. We notice that re-centering can dramatically improve the detection performance by $+22.33$ mAP, especially for those objects farther from the origin.

Method	mAP
Baseline	66.03
+ re-centering	88.36
+ pre-training	92.03

Table 8. A comparison of detection results on the WildTrack dataset with *re-centering* and *pre-training*.

A.3. Pre-Training on AICity'24 Dataset

Since WildTrack is a small dataset with only 400 frames in total, we considered training BEVFormer with a pre-trained model on the AICity'24 dataset, which is a much larger dataset with various scenes. As shown in Tab. 8, this pre-training leads to $+3.67$ detection performance improvement on the WildTrack test set. This also illustrates the important role of large and well-annotated synthetic datasets in boosting the performance on limited real data.

B. Detection Association Algorithm

B.1. Algorithm Details

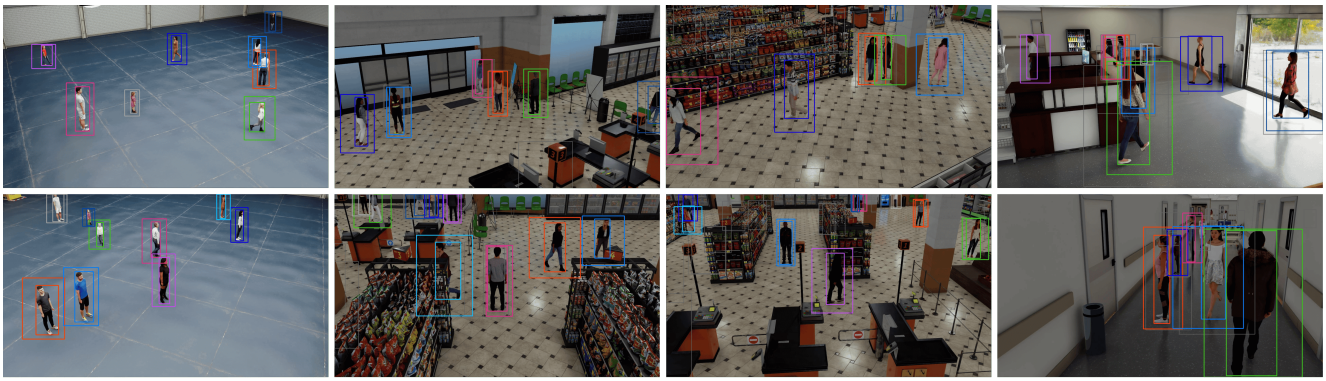
Algorithm 1: 2D-3D detection association

Input : 2D detection set $\mathcal{D}^v$ from camera v; 3D detection set $\mathcal{E}$ from BEVFormer with all camera views; projection matrix $\mathcal{P}^v$ of camera v.

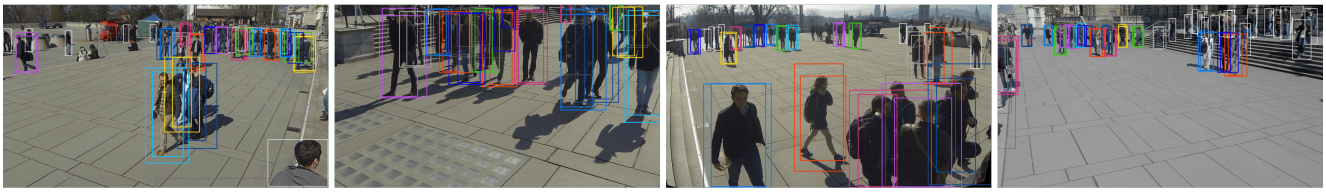
Output: Mapping of indices from $\mathcal{E}$ to $\mathcal{D}^v$.

$\mathcal{E} \leftarrow$ filter $\mathcal{E}$ by confidence score;
$\mathcal{E} \leftarrow \text{CircleNMS}(\mathcal{E}, \delta)$;
// optional, δ: NMS threshold
$\mathcal{E}^v \leftarrow \mathcal{P}^v(\mathcal{E})$; // projected 3D boxes
for *camera* v **to** V **do**
  Initialize the cost matrix $\mathbf{c}^v = [c_{ij}^v]$ as zeros;
  for $\mathbf{b}_i^{3D}$ **from** $\mathcal{E}^v$ **do**
    for $\mathbf{b}_j^{2D}$ **from** $\mathcal{D}^v$ **do**
      $c_{ij}^v \leftarrow$ compute cost by Eq. (4);
    end
  end
  Matches $m^v \leftarrow \text{Hungarian}(\mathbf{c}^v, \Delta)$;
  // Δ: cost threshold
end

The detailed 2D-3D detection association algorithm is shown in Algorithm 1. We set the threshold for CircleNMS to $\delta = 0.2$m and set the cost threshold to $\Delta = 150$.



(a) AICity'24 dataset



(b) WildTrack dataset

Figure 7. Visualization of 2D-3D detection association results.

B.2. Visualization

We visualized some 2D-3D detection association results on sample frames of the AICity'24 and WildTrack datasets in Fig. 7. The associated bounding boxes are in the same color, where the smaller ones are 2D detections and the larger ones are projected 3D detections from BEVFormer. Those 2D detections in white are not associated with any 3D detections.

B.3. Improvements with Detection Association

We compared the tracking performance of MCBLT with and without the proposed 2D-3D detection association algorithm in Tab. 9. The baseline result is based on the ReID features extracted from the large projected 3D bounding boxes shown in Fig. 3. With the noisy background or other objects included in the image crops, ReID feature quality will be significantly affected.

Method	IDF1	MOTA	MOTP	MT	ML
Baseline	63.2	73.4	93.7	24.0	4.0
+ det association	93.4	87.5	94.3	90.2	2.4

Table 9. A comparison of results on the WildTrack test set with our 2D-3D detection association algorithm.

C. ReID Feature Quality Analysis

We conducted ReID feature quality analysis on both the AICity'24 and WildTrack datasets. For the AICity'24 dataset, we sampled 500 characters with their 2D bounding boxes and object IDs from the ground truth across all scenes and cameras from the test set. The total object image crop count is 40,000. We filtered out 2D bounding boxes that are smaller than 5,000 pixels, as well as those whose aspect ratio (*i.e.*, width / height) is less than 0.15. Similarly, for the Wildtrack dataset, we sampled 330 characters from the sequence and applied the same filters bringing the total object crop count to 41,284.

Dataset	Rank-1	Rank-5	Rank-10	mAP
AICity'24	95.02	97.44	98.08	73.85
WildTrack	77.18	84.49	87.97	63.11

Table 10. Evaluation on our ReID feature quality.

We evaluated the ReID feature quality by the mean average precision (mAP), rank-1, rank-5, and rank-10 accuracies. The evaluation results are shown in Tab. 10. We found that the feature quality on the WildTrack dataset is worse than that on the AICity'24 dataset. This is because i) WildTrack is a real-world dataset with more noises and diverse illuminations from different camera views; ii) 2D bounding box annotations are not as accurate as the synthetic AICity'24 dataset. Nevertheless, our MCBLT achieved impressive results on WildTrack based on these ReID features.

D. Model Time Complexity Analysis

Although MTMC detection and tracking tasks do not usually require real-time performance and are tolerant to time delays, we record the running time of the proposed MCBLT pipeline in Tab. 11 to provide a rough impression of the complexity of the model. The model inference is conducted on one single NVIDIA A100 GPU, with 10 cameras in the scene. Our method achieves around 1.5 FPS end-to-end before any further model optimization. The 2D detection, ReID, and tracking models are very efficient and can operate in parallel with BEVFormer so that their running time is negligible.

Detection		Tracking
BEVFormer 1.6 FPS		SUSHI
DINO 65.0 FPS	SOLIDER 58.9 FPS	452.7 FPS

Table 11. MCBLT model efficiency analysis.

E. Overall Visualization

We also visualized MTMC detection and tracking results of MCBLT on the AICity'24 and WildTrack datasets. Please find the demos in the attachment.