International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02176-7

Cyclic Refiner: Object-Aware Temporal Representation Learning for Multi-view 3D Detection and Tracking

Mingzhe Guo[1] · Zhipeng Zhang[2] · Liping Jing[1] · Yuan He[2] · Ke Wang[2] · Heng Fan[3]

Received: 31 August 2023 / Accepted: 3 July 2024

Abstract

We propose a unified object-aware temporal learning framework for multi-view 3D detection and tracking tasks. Having observed that the efficacy of the temporal fusion strategy in recent multi-view perception methods may be weakened by distractors and background clutters in historical frames, we propose a cyclic learning mechanism to improve the robustness of multi-view representation learning. The essence is constructing a backward bridge to propagate information from model predictions (*e.g.,* object locations and sizes) to image and BEV features, which forms a circle with regular inference. After backward refinement, the responses of target-irrelevant regions in historical frames would be suppressed, decreasing the risk of polluting future frames and improving the object awareness ability of temporal fusion. We further tailor an object-aware association strategy for tracking based on the cyclic learning model. The cyclic learning model not only provides refined features, but also delivers finer clues (*e.g.,* scale level) for tracklet association. The proposed cycle learning method and association module together contribute a novel and unified multi-task framework. Experiments on nuScenes show that the proposed model achieves consistent performance gains over baselines of different designs (*i.e.,* dense query-based BEVFormer, sparse query-based SparseBEV and LSS-based BEVDet4D) on both detection and tracking evaluation. Codes and models will be released.

Keywords Cyclic refiner · Backward refinement · Object-aware representation · Temporal learning · Multi-view 3D detection and tracking

1 Introduction

Perception with multi-view cameras has received extensive attention in autonomous driving because of their complementarity in observing the physical world and the potential to replace expensive sensors like LiDAR. Recent advanced methods translate different perspective camera features to the bird's-eye-view (BEV) space (Huang et al., 2021; Li et al., 2022c, b; Fischer et al., 2022; Shi et al., 2022), which have demonstrated promising performances in 3D tasks. As autonomous driving naturally is a temporal task, features in past frames are usually used to enhance the representation learning of current timestamp Li et al. (2022c); Liu et al. (2022b); Huang and Huang (2022); Pang et al. (2022).

Revisiting recent related methods, we observe that the models are commonly constructed in a "sequential" manner, which forms a "Multi-view Images → Image/BEV Features

Communicated by Wanli Ouyang.

Mingzhe Guo and Zhipeng Zhang have contributed equally to this work.

✉ Zhipeng Zhang
zhipeng.zhang.cv@outlook.com

✉ Liping Jing
lpjing@bjtu.edu.cn

Mingzhe Guo
mingzheguo@bjtu.edu.cn

Yuan He
akinaheyuan@didiglobal.com

Ke Wang
kewang1@didiglobal.com

Heng Fan
heng.fan@unt.edu

[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China

[2] KargoBot, Beijing, China

[3] Department of Computer Science and Engineering, University of North Texas, Denton, USA

Published online: 16 July 2024

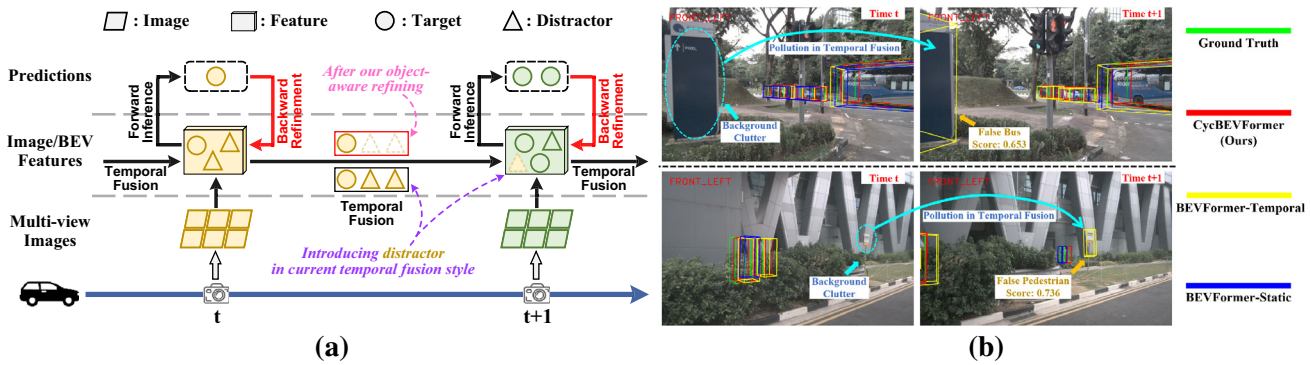

Fig. 1 **a** The illustration of our cyclic pipeline. After the first forward inference at time t, instead of directly propagating the distractor-contained features to the next frame through temporal fusion (black arrows), we exploit the posterior predictions as object-aware prior information to refine the former learned image and BEV features, i.e., "Backward Refinement" (red arrows). Then the refined features at time t are forwarded to the temporal fusion and second forward inference at time $t + 1$, which formulates a cyclic route to perform object-aware representation learning. **b** Visualization of background clutter pollution in temporal fusion. At time t, no detections are predicted over the background clutter (cyan circles). Then, at time $t + 1$, the temporal model (BEVFormer-Temporal) mistakenly produces false positives over the background clutter, yet the static model (BEVFormer-Static) surprisingly does not, illustrating that background with high semantics in previous frames may corrupt future features after temporal fusion

$\rightarrow$ Predictions" pipeline (see the black arrows shown in Fig. 1a). In this strategy, the "Image/BEV Features" are used for both forward inferences in the current frame and temporal fusion in the next frame. However, complex driving scenarios in the real world contain diverse distractors and background clutters (the triangles in Fig. 1a). Directly and simply using features from the previous frame in temporal fusion may introduce historical interferences and degrade the representation learning of future frames, which eventually leads to false positives (see the purple arrows in Fig. 1a and visualization in Fig. 1b). In contrast, cognitive science has proved that the human recognition system is more brilliant, which can introspect to backward refine the learned knowledge before the next reasoning Bechtel (2013); Price (1998).

Motivated by the above observation, for the first time, we attempt to learn the multi-view image and BEV representations in a cyclic manner. The essence is to treat the posterior predictions (*e.g.,* object locations and sizes) of a frame as the prior information to refine its Image/BEV representation (see the red arrows in Fig. 1a). In the training of deep networks, the gradients are used to optimize model parameters, which "implicitly" refine the learned representations. The information in the predictions can be considered as "fake gradients" in inference (without groundtruth) to "explicitly" reinforce the learned representations. As the sparse predictions contain compact object information, it is expected that the refined image and BEV features are more discriminative, and the responses of distractors are suppressed (the magenta arrows in Fig. 1a). Notably, the proposed "backward refinement" is conducted before temporal fusion, therefore, the representation learning of the next frame can benefit from the object-aware refinement process of the previous frame.

To this end, we propose an object-aware temporal learning framework for multi-view 3D detection and tracking. The core is the proposed Cyclic Refiner, which backwards the crucial information in model predictions to refine the input image and BEV features. Specifically, for the predicted objects, their corresponding features, which contain image ROI (region of interest) embedding, BEV embedding and head embedding,[1] are concatenated to predict masks for filtering distractors in image and BEV features. The mask can be considered as the combination of different 2D gaussian attention maps, where the peaks are the centers of objects and the attention values are generated by linearly mapping the concatenated features. Furthermore, it is aware that the object sizes in different categories are diverse in BEV spaces (*e.g.,* truck and cone), and even the same object occupies different spatial ranges in the image and BEV spaces because of camera projection. Therefore, it is necessary to encode object scale information into the filter mask, which prevents overlarge mask including background clutters or too-small mask missing target information. In our method, we assign each object a scale level to determine its spatial attention range in the filter mask. To realize that, we apply linear layers on the concatenated features to predict an object's scale levels in BEV and image spaces, respectively.

Interestingly, we observed that the multiple feature embeddings and scale level estimation benefit both detection and the downstream tracking task. The embeddings provide suf-

[1] Head embedding denotes the "object query" in DETR-based Carion et al. (2020) methods and the ROI (region of interest) pooling feature in other detection heads.

ficient appearance clues for association, while the scale level identifies objects with similar scales to reduce false matches. We thus propose multi-clue matching and cascaded scale-aware matching for object-aware association in tracking. In particular, the multiple appearance features (*i.e.,* image/BEV embeddings from refined image/BEV features and head embedding) are exploited to perform multi-clue matching by computing the similarity. Then cascaded scale-aware matching divides objects into different splits with the same scale level to associate with box IoUs separately, which prevents false matches caused by the overlap between large objects and nearby small ones in BEV space. Notably, we also propose the buffering strategy to provide reasonable box IoUs in BEV space, since the coverage scale of box predictions in BEV plane is smaller than that in image space.

We apply our cyclic refiner to three different detection methods (*i.e.,* dense-query-based BEVFormer Li et al. (2022c), sparse-query-based SparseBEV Liu et al. (2023) and LSS-based BEVDet4D Huang and Huang (2022)), and use SimpleTrack Pang et al. (2021) after the detectors to conduct the tracking baselines. Experimental results show that our unified framework achieves 1.7%/1.8%/2.9% mAP and 13.0%/13.9%/16.0% AMOTA improvements on the test splits of nuScenes detection and tracking datasets respectively, demonstrating the effectiveness and generality.

In summary, our main contributions are as follows: **(1)** We propose the cyclic refiner to learn object-aware image and BEV representations; **(2)** We propose the multi-clue matching and cascaded scale-aware matching for robust association; **(3)** We unify the BEV detection and tracking tasks with the proposed temporal representation learning framework; **(4)** Experiments show that the proposed model generally improves baselines of different design concepts (*i.e.,* query-based and LSS-based) with considerable performance gains on both BEV detection and tracking tasks.

2 Related Work

2.1 3D Detection with Multi-view Cameras

The modeling fashion in camera-based 3D object detection is transitioning from single view to multi-view because of the natural complementarity among different cameras. Recent state-of-the-art (SOTA) methods devote most efforts to learning a more discriminative BEV representation. One representative branch follows LSS Philion and Fidler (2020), which lifts 2D image features to 3D space via the predicted depth distributions Reading et al. (2021); Huang et al. (2021); Li et al. (2022a). The other burgeoning branch is query-based framework Wang et al. (2022b); Li et al. (2022c); Pang et al. (2022); Li et al. (2023); Liu et al. (2023), which projects each 3D sampling point in BEV space to multi-view 2D images for visual feature extraction. Specifically, the dense query-based methods Li et al. (2022c); Pang et al. (2022); Jiang et al. (2022) build explicit BEV features by arranging image features into corresponding locations of the BEV plane. In contrast, the sparse query-based works Wang et al. (2022b); Liu et al. (2022a, 2023) directly encode image features into the learnable queries. Since autonomous driving is a sequential task, recent works Huang and Huang (2022); Li et al. (2022c); Liu et al. (2022b); Pang et al. (2022) exploit temporal cues by aligning and fusing image/BEV features at different timestamps to enhance representation learning and detection capability. The noticeable drawback of this learning strategy is that the ability of temporal learning heavily depends on the feature quality of historical features. Once the historical features are polluted by distractors or background clutters, fusing them may even bring negative effects to the representation learning of the future frames. Therefore, the lack of post-processing historical features becomes the bottleneck of most current temporal learning methods. In this work, we propose the cyclic refiner to alleviate this issue by filtering target-irrelevant responses in historical features. In addition, FrustumFormer Wang et al. (2023a) and MV2D Wang et al. (2023b) also consider feature learning on target regions, which exploit the generated 2D proposals by off-the-shelf 2D detectors (*i.e.,* MaskRCNN He et al. (2017) and FasterRCNN Ren et al. (2015)) as priors for query generation. Differently, our method exploits inherent 3D model predictions for object-aware refining, which is a general design to improve 3D detectors while requiring little computation overhead.

2.2 3D Tracking with Multi-view Cameras

Multi-view 3D tracking is the downstream task after object detection, which aims to temporally associate the trajectories of each object in 3D space and record their unique identity. With similar task definition, camera-based 3D tracking usually adapts the design of 2D multi-object tracking methods Liang et al. (2022b, a) to perform association in 3D space. In particular, following the tracking-by-detection paradigm, many works Chaabane et al. (2021); Hu et al. (2022); Fischer et al. (2022); Shi et al. (2022) match the detections of a frame with previous tracklets by appearance similarity and spatial proximity. Considering the depth information of 3D scenes, SimpleTrack Pang et al. (2021) performs matching by computing the 3D IoU between objects. Notably, Kalman filter Welch and Bishop (1995) is exploited to predict current locations of previous tracklets for motion compensation. QD-3DT Hu et al. (2022) further improves the accuracy of motion prediction by learning temporal clues with LSTM Hochreiter and Schmidhuber (1997). Recently, MUTR3D Zhang et al. (2022) and SRCN3D Shi et al. (2022) verify the effectiveness of appearance-based association by matching with

head embeddings or 2D ROI features. Compared with object detection, tracking is more sensitive to distractors and background clutters, since association heavily relies on the quality of the learned features to estimate the similarity of each two objects. In this work, we show the generality of our cyclic refiner in both 3D detection and tracking tasks. With the tailored object-aware association, our method could perform robust 3D tracking.

2.3 Online Update in 2D Visual Object Tracking

To handle the appearance variance of the target in 2D visual object tracking (VOT), many works Bhat et al. (2019); Danelljan et al. (2019); Cui et al. (2022) are dedicated to the online update mechanism. MOSSE Bolme et al. (2010) updates the learned filter by maximizing the response gap between the target and background. DiMP Bhat et al. (2019) introduces a predictor to online optimize the target model instructed by a discriminative loss. ATOM Danelljan et al. (2019) updates the classification layers with the proposed fast optimization method for efficiency. FCOT Cui et al. (2022) further verifies the effectiveness of online regression by merging the online model with the static one. In this work, we design the cyclic refiner to online update the image/BEV features, which eventually relieves the temporal error accumulation caused by inaccurate predictions (*e.g.,* false positives) in historical frames. This is clearly different from the mechanism in VOT aiming to improve matching robustness. Besides, the proposed method does not require any gradient backward based optimization, which thus shows a better trade-off between performance and efficiency in inference.

3 Approach

Our proposed object-aware temporal representation learning framework is detailed in this section. We first recap the detail of our core contribution, *i.e.,* cyclic refiner, in Sec. 3.1. Then the designed object-aware association method for tracking is described in Sec. 3.2. Finally, we conduct a unified detection and tracking framework based on the proposed cyclic refiner and association strategy in Sec. 3.3.

3.1 Cyclic Refiner

The essence of cyclic refiner is the proposed "backward refinement" mechanism, which creates a cycle between the image/BEV features and model predictions, together with regular forward inference. The representations produced by the cyclic refiner are used for temporal fusion.

As shown in Fig. 2, "backward refinement" first collects information from each predicted object O_i $(i = 1, 2, ..., N)$. In our method, both the representations and predicted values (*i.e.,* location and size) of each object are exploited for backward refinement. In particular, besides the apparent image features $\mathbf{F}_{img} \in \mathbb{R}^{H \times W \times C}$ and BEV features $\mathbf{F}_{bev} \in \mathbb{R}^{H' \times W' \times C}$, we also exploit the head features $\mathbf{F}_{head}$, which are sparse object queries ($\mathbb{R}^{N \times C}$) in DETR-based methods Carion et al. (2020) and dense 2D features in other detection heads, in the refinement module. With the center and object size predicted for each object, we extract the feature embeddings $\{\mathbf{e}_{img}, \mathbf{e}_{bev}, \mathbf{e}_{head} \in \mathbb{R}^{1 \times C}\}$ from $\{\mathbf{F}_{img}, \mathbf{F}_{bev}, \mathbf{F}_{head}\}$ with ROI pooling Ren et al. (2015). Notably, $\mathbf{e}_{head} = \mathbf{F}^{i}_{head}$ for DETR-based methods. Then we concatenate the three embeddings as the representation $\mathbf{e}_{cat} \in \mathbb{R}^{1 \times 3C}$ of an object. So far, the state of each object is represented as $O_i = \{\mathbf{e}_{cat}, p\}$, where p denotes the object location and size information.

Then "backward refinement" exploits the collected object information O to refine the image/BEV features (*i.e.,* $\mathbf{F}_{img}$ and $\mathbf{F}_{bev}$). In our design, each object's feature representation and posterior prediction are transferred to a filter mask, serving as the prior information of image/BEV features. The filter mask is used to decrease the responses of target-irrelevant regions, *e.g.,* distractors and background clutters. It contains four steps, **1)** Firstly, we generate an initial 2D weight mask for each object, where the location corresponds to the predicted object center. **2)** Secondly, we assign each object with a scale level by mapping $\mathbf{e}_{cat}$ to a one-hot vector, which determines the spatial scope of the 2D weight mask. The weights of the positions out of the spatial scope are set to zero. **3)** Thirdly, the weight distribution of the positions inside the spatial scope is predicted by linearly mapping $\mathbf{e}_{cat}$, which assigns higher weights for the discriminative areas of each object while suppressing target-irrelevant parts (*e.g.,* corners and scattered background). **4)** Finally, the weight masks of objects belonging to the same scale level l are combined to get the final filter mask $\mathbf{M}_l$. Figure 2 illustrates the process of generating filter masks in different scale levels. For simplicity, only three scale levels are visualized in Fig. 2.

We treat the predicted filter mask of a scale level M_l as the spatial attention, which is applied to both image and BEV features by element-wise multiplication. The masked features will be further processed by DCNs Dai et al. (2017) of different kernel sizes, improving the scale awareness of the learned representations. The refined features from different scale levels are concatenated and fused with a DCN layer, forming the object-aware features $\hat{\mathbf{F}}_{img} \in \mathbb{R}^{H \times W \times C}$ and $\hat{\mathbf{F}}_{bev} \in \mathbb{R}^{H' \times W' \times C}$. The original features (i.e., $\mathbf{F}_{img}$ and $\mathbf{F}_{bev}$) are also used for fusion to avoid losing informative details.

After refining the image and BEV features by the cyclic refiner at time t, the next step is to forward the refined object-aware representations $\hat{\mathbf{F}}^t = \{\hat{\mathbf{F}}^t_{img}, \hat{\mathbf{F}}^t_{bev}\}$ to the next

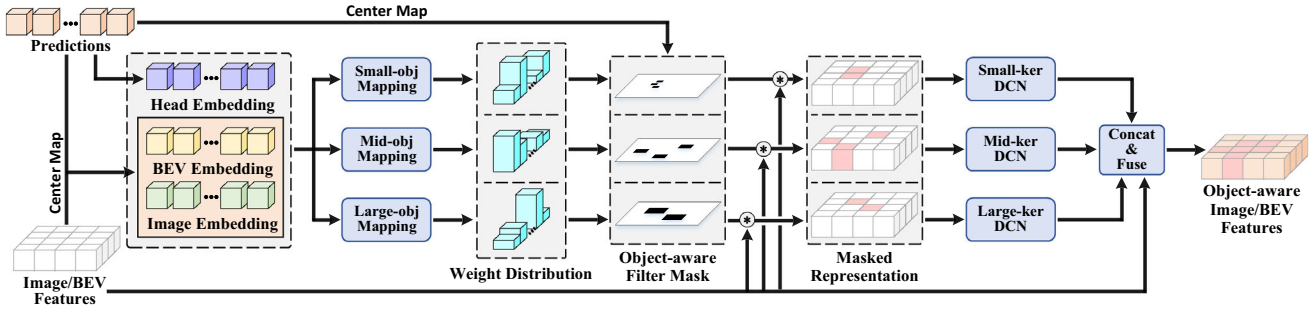

Fig. 2 Illustration of the "backward refinement" in the proposed cyclic refiner. Each predicted object determines its scale level (i.e., small, mid and large) and weight in the mask by linearly mapping the concatenated three feature embeddings. By applying the masks on the image/BEV features, it can filter target-irrelevant distractors and benefit representation learning

frame $t+1$. The fusion mechanism is not the contribution of our work, therefore, we simply follow the baseline methods (*i.e.,* BEVFormer Li et al. (2022c) and BEVDet4D Huang and Huang (2022)) to construct temporal fusion modules. Notably, the baselines use deformable attention to solely fuse BEV features in different timestamps, we inherit this design for the temporal fusion of image features. Here we describe how our refined object-aware features $\hat{\mathbf{F}}^t$ guide representation learning in the temporal module, as shown in Fig. 3.

Instead of simple concatenation, the temporal fusion exploits the object-aware prior knowledge of $\hat{\mathbf{F}}^t$ to further refine the learned features $\mathbf{F}^{t+1} = \{\mathbf{F}^{t+1}_{img}, \mathbf{F}^{t+1}_{bev}\}$ at time $t+1$, which benefits the successive representation learning of the forward inference. Specifically, the temporal object-aware prior of $\hat{\mathbf{F}}^t$ is employed to reconstruct $\mathbf{F}^{t+1}$ with the deformable attention Zhu et al. (2020). As illustrated in Fig. 3, the refined object-aware features $\hat{\mathbf{F}}^t$ concatenate with the features $\mathbf{F}^{t+1}$ to generate object-aware attention weights $\mathbf{A}$ and sampling offsets Δs. The sampling offsets Δs are then applied to the sampling grid s. to improve the sampling locations on target regions. The attention weights $\mathbf{A}$ are expected to perceive and assign higher values to informative areas while suppressing target-irrelevant ones. Finally, the features

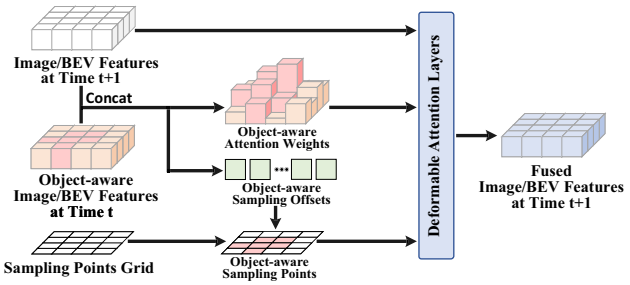

Fig. 3 Illustration of the temporal fusion with our object-aware representations. The refined image/BEV features at time t will concatenate with the learned features at time $t+1$ to generate object-aware attention weights and sampling offsets, guiding feature sampling on target-relevant regions in the deformable attention Zhu et al. (2020)

$\mathbf{F}^{t+1}$ are refined with the object-aware attention weights $\mathbf{A}$ and the shifted sampling points $s + \Delta s$. The calculation is defined as

$$\text{DeformAttn}(\mathbf{A}, \boldsymbol{p}, \Delta \boldsymbol{p}, \mathbf{F}^{t+1}) = \sum_{h=1}^{H} \mathbf{W}_h \Big[\sum_{k=1}^{K} \mathbf{A}_{hk} \cdot \mathbf{W}'_h \mathbf{F}^{t+1}(\boldsymbol{s} + \Delta \boldsymbol{s}_{hk}) \Big], \tag{1}$$

where h and k are the indexes of the attention head and sampled feature point, respectively. $\mathbf{W}'_h \in \mathbb{R}^{C_v \times C}$ and $\mathbf{W}_h \in \mathbb{R}^{C \times C_v}$ are the learnable weights ($C_v = C/H$ by default). $\Delta \boldsymbol{s}_{hk}$ and $\mathbf{A}_{hk}$ denote the sampling offset and attention weight of the kth sampling point in the hth attention head, respectively (please refer to Zhu et al. (2020) for more details). With the object-aware attention weights and sampling locations, the temporal fusion could propagate the refined target information to the forward inference of the next frame, which benefits the representation learning and prediction with the enhanced object awareness ability.

3.2 Object-aware Association

As mentioned, our ultimate purpose is to build a unified detection and tracking framework which can both benefit from the proposed cyclic refiner. Therefore, a tailored association method for tracking to fully take advantage of the refined image and BEV features is necessary.

As shown in Fig. 4, given the detections $\mathcal{D}_t$ of frame t and existing tracklets $\mathcal{T}$ (empty set for the first frame), our object-aware association (dubbed OAA) aims to match each detected object from $\mathcal{D}_t$ with its corresponding tracklet in $\mathcal{T}$. Notably, before the association, we adopt Kalman Filter Welch and Bishop (1995) to predict the location in the current frame for each tracklet in $\mathcal{T}$. The association contains two main steps, *i.e.,* Multi-clue Matching and Cascaded Scale-aware Matching.

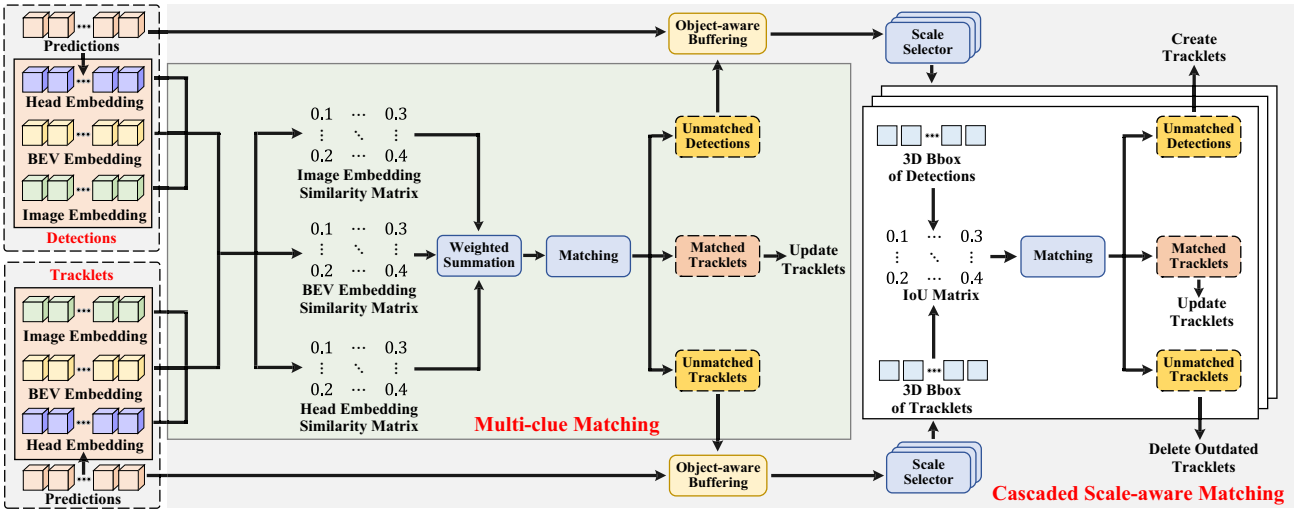

Fig. 4 Architecture of the proposed Object-aware Association **(OAA)**. The Multi-clue Matching first matches the new detections and existing tracklets with the weighted summation of three embedding similarity matrixes. Then the 3D boxes of unmatched detections and tracklets are buffered with the assigned scale level of cyclic refiner, which are fed into the Cascading Scale-aware Matching to perform hierarchical IoU matching from large-scale objects to small-scale ones

3.2.1 Multi-clue Matching

The similarity score between appearance embeddings of two objects is regarded as crucial evidence to judge if they are the same object in 2D MOT. We follow this design and adapt it for multi-view 3D tracking. Specifically, besides the commonly-used image ROI pooling embedding $\hat{\mathbf{e}}_{img}$, we introduce BEV and head embeddings (i.e., $\hat{\mathbf{e}}_{bev}$ and $\mathbf{e}_{head}$) as extra appearance clues, which form the appearance state $E = \{\hat{\mathbf{e}}_{img}, \hat{\mathbf{e}}_{bev}, \mathbf{e}_{head}\}$ for each object. Notably, $\{\hat{\mathbf{e}}_{img}, \hat{\mathbf{e}}_{bev}\}$ are sampled from the refined object-aware features $\hat{\mathbf{F}}$, which aims to perform accurate matches with refined target information. Given the existing tracklets $\mathcal{T} = \{\mathcal{T}_j = \{E^{\mathcal{T}_j}, p\}, j = 1, 2, ..., M\}$ and new detections $\mathcal{D} = \{\mathcal{D}_i = \{E^{\mathcal{D}_i}, p\}, i = 1, 2, ..., N\}$ (p denotes the object location and size information, defined in Sec. 3.1), the Multi-clue Matching computes similarities between $E^{\mathcal{T}}$ and $E^{\mathcal{D}}$ with the normalized inner product, which generates three similarity matrixes $\{\mathbf{C}_{img}, \mathbf{C}_{bev}, \mathbf{C}_{head}\}$. The weighted summation of $\{\mathbf{C}_{img}, \mathbf{C}_{bev}, \mathbf{C}_{head}\}$ is regarded as the cost matrix $\mathbf{C}$ in Hungarian Algorithm Kuhn (1955) to find the optimal bipartite matching. The calculation is formulated as

$$\begin{aligned}\mathbf{C} = w_{img} \cdot \left\langle \hat{\mathbf{e}}^{\mathcal{D}}_{img}, \hat{\mathbf{e}}^{\mathcal{T}}_{img} \right\rangle + w_{bev} \cdot \left\langle \hat{\mathbf{e}}^{\mathcal{D}}_{bev}, \hat{\mathbf{e}}^{\mathcal{T}}_{bev} \right\rangle \\ + w_{head} \cdot \left\langle \mathbf{e}^{\mathcal{D}}_{head}, \mathbf{e}^{\mathcal{T}}_{head} \right\rangle,\end{aligned} \tag{2}$$

where $w_{img}, w_{bev}, w_{head}$ are the weight coefficients and $\langle \cdot \rangle$ represents the operation of normalized inner product. The matched detections are used to update associated tracklets, and unmatched detections $\mathcal{D}_{remain}$ and tracklets $\mathcal{T}_{remain}$ are sent to the second Cascaded Scale-aware Matching.

3.2.2 Cascaded Scale-aware Matching

The second association is Cascaded Scale-aware Matching, which associates object by the box IoUs between $\mathcal{T}_{remain}$ and $\mathcal{D}_{remain}$. We noticed that the coverage scale of an object box in BEV space is smaller than that in image space, especially for the objects close to cameras, making it lack sufficient context clues for matching. Motivated by BIoU Yang et al. (2023), we use the buffering strategy to expand the matching space, which buffers (enlarges) the box B of each object with a ratio r. The operation is formulated as

$$B_{buffer} = (1 + r) \cdot B. \tag{3}$$

Notably, we set larger buffer ratio for the objects with small scale level (*i.e.,* predicted by the cyclic refiner in Sec. 3.1), since it is difficult to generate reasonable IoUs for small objects. Besides, it is also noticed that, after Kalman Filter propagation, large objects are more likely to cover nearby small objects in BEV space, which may cause false matches and track fragmentation. Therefore, we perform IoU association from large to small scales, and only allow the matches between close scale levels. Specifically, for the detections with scale level l, we select unmatched tracklets in scale levels $[l - 1, l, l + 1]$ to perform IoU matching.

After the two-step association, the unmatched outdated tracklets $\mathcal{T}_{re-remain}$ will be deleted from $\mathcal{T}$, and the remain-

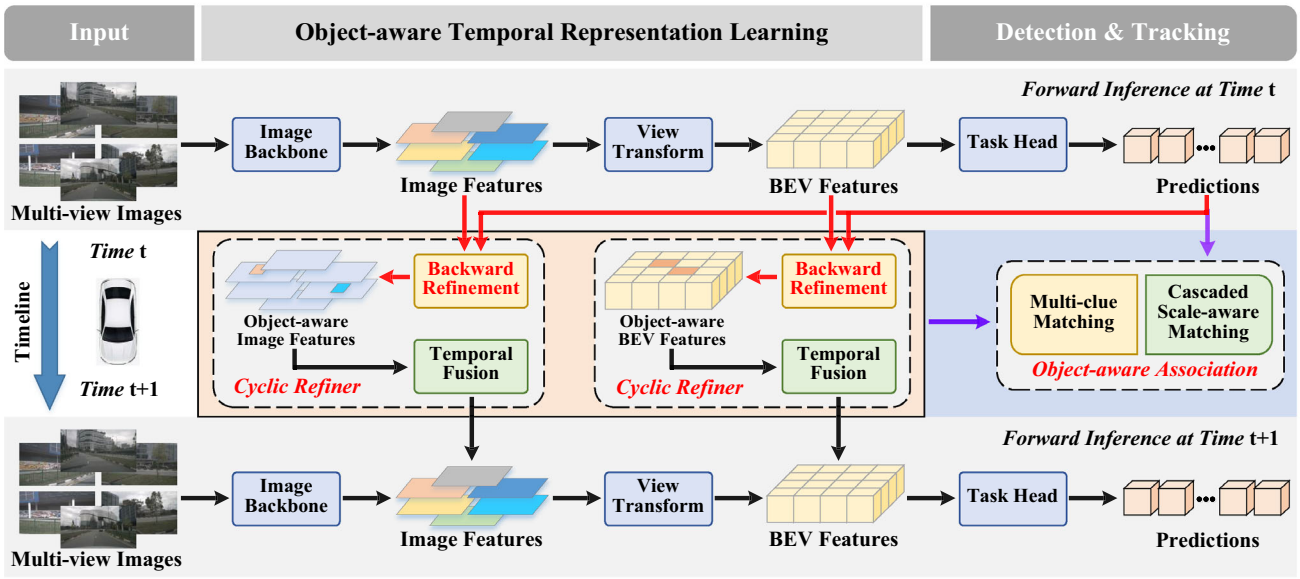

Fig. 5 Architecture of the proposed object-aware temporal learning framework for both 3D detection and tracking tasks. After the forward inference at time t (black arrows), the "backward refinement" of cyclic refiner exploits information in predictions to refine image and BEV features (red arrows). The refined features are then used for temporal fusion at time $t+1$. The proposed object-aware association exploits the refined features and predicted scale levels from cyclic refiner to perform object tracking (purple arrows) (Color figure online)

ing detections $\mathcal{D}_{remain}$ with scores higher than τ will be initialized as new tracklets.

In summary, the proposed object-aware association performs a two-step matching after the detection of each frame, as shown in Alg. 1. The pipeline consists of four stages: **1)** The Kalman Filter is applied to predict the location for each tracklet in $\mathcal{T}$ (line 1 to 3 in Alg. 1); **2)** The first association is performed between $\mathcal{D}_t$ and $\mathcal{T}$ with multi-clues, *i.e.,* head embedding $\mathbf{e}_{head}$, BEV embedding $\hat{\mathbf{e}}_{bev}$ and image ROI pooling embedding $\hat{\mathbf{e}}_{img}$ from the refined features. The three embeddings respectively generate the similarity matrixes $\{\mathbf{C}_{img}, \mathbf{C}_{bev}, \mathbf{C}_{head}\}$ with tracklets $\mathcal{T}$ by inner product. The summation of the three similarity matrixes is regarded as the cost matrix in Hungarian Algorithm Kuhn (1955) for matching. The associated tracklets will be updated with the newly detected objects, and the unmatched detections and tracklets are kept in $\mathcal{D}_{remain}$ and $\mathcal{T}_{remain}$, respectively (line 4 to 6 in Alg. 1); **3)** The second association matches objects between $\mathcal{T}_{remain}$ and $\mathcal{D}_{remain}$ based on box IoUs. The box of each object B is buffered with a ratio r (line 7 to 10 in Alg. 1). Then, IoU association is performed from large to small scales, which divides the matching group by the assigned scale level l (line 11 to 15 in Alg. 1). **4)** The unmatched tracklets $\mathcal{T}_{re-remain}$ are deleted from $\mathcal{T}$, and the remaining detections $\mathcal{D}_{remain}$ are filtered with the score threshold τ to generate new tracklets (line 16 to 22 in Alg 1).

3.3 Unified Detection and Tracking Framework

With the proposed Cyclic Refiner and Object-aware Association, we construct a unified temporal representation learning framework for both BEV detection and tracking, as shown in Fig. 5. Our framework consists of three main parts: Input, Object-aware Temporal Representation Learning module, Detection and Tracking heads. Taking multi-view images at time t as input, the image backbone first extracts image features. Then, the view-transformer transforms image features to the BEV representation, serving as the input of task-specific heads. Before the next forward inference at time $t+1$, the proposed cyclic refiner exploits the information in the predictions at time t to refine the image and BEV features. After that, the refined features are used for temporal fusion between t and $t+1$.

4 Experiments

In this section, we first recap the experimental setup in Sec. 4.1. Then, we respectively present the evaluation results (Sec. 4.2) in detection and tracking tasks. Finally, we detail the ablation studies (Sec. 4.3) and analysis (Sec. 4.4) to demonstrate the effectiveness of the proposed methods.

Algorithm 1: Pseudo-code of OAA.

Input: features of current frame; detections $\mathcal{D}_t$; existing tracklets $\mathcal{T}$; score threshold τ;
Output: updated tracklets $\mathcal{T}$

```
/* Predict New States of Tracklets */
for T_j in T do
    T_j ← KalmanFilter(T_j)
end
/* Multi-clue Matching */
Associate T and D_t using {ê_img, ê_bev, e_head}
D_remain ← remaining objects from D_t
T_remain ← remaining tracklets from T
/* Cascaded Scale-aware Matching */
for D_i in D_remain do
    l ← Scale Level(D_i) (Cyclic Refiner in Sec. 3.1)
    Buffer box scales B
end
/* Second Association */
for l in {large, mid, small} do
    D_select ← D_remain[l]
    T_select ← T_remain[l − 1, l, l + 1]
    Associate T_select and D_select using IoU
end
/* Delete Unmatched Tracklets */
T_re−remain ← remaining tracklets from T_remain
T ← T \ T_re−remain
/* Initialize New Tracklets */
for D_i in D_remain do
    if D_i.score > τ then
        T ← T ∪ {D_i}
    end
end
```

Return: $\mathcal{T}$

4.1 Experimental Setup

4.1.1 Dataset and Metrics

We conduct experiments on nuScen-es Caesar et al. (2020), which collects autonomous driving data from 1,000 scenes. The benchmark is composed of 40,157 samples and is divided into 28,130, 6,019, and 6,008 ones for training, validation, and testing, respectively. For the 3D detection task, we adopt mean average precision (mAP) and nuScenes Detection Score (NDS) as primary metrics, as well as five True Positive (TP) metrics, including mATE, mASE, mAOE, mAVE and mAAE. For the 3D tracking task, we follow the prior works Fischer et al. (2022); Zhang et al. (2022); Shi et al. (2022) to use average multi-object tracking accuracy (AMOTA) and average multi-object tracking precision (AMOTP) as the major evaluation criteria, along with RECALL, MOTA, IDS. Reports of our methods on all detection and tracking metrics will be publicly available on nuScenes leaderboard.

4.1.2 Implementation Details

To verify the effectiveness and generality of the proposed methods, we apply Cyclic Refiner and OAA on recent state-of-the-art BEVFormer Li et al. (2022c) (both Small and Base versions), SparseBEV Liu et al. (2023) and BEVDet4D Huang and Huang (2022). The unified detection and tracking frameworks, i.e., CycBEVFormer, CycSparseBEV and CycBEVDet4D, are evaluated on both 3D detection and tracking tasks. Notably, there is no need to fine-tune the model for tracking after the training of detection. Following the tracking-by-detection paradigm, we construct the tracker by applying our plug-and-play object-aware association to the trained detectors. The training and inference settings are the same as the three baseline methods. We recommend the readers to Li et al. (2022c) and Huang and Huang (2022) for more details. All our models are trained on 8 NVIDIA RTX 3090 GPUs and the inference is measured on one NVIDIA RTX 3090 GPU.

In Sec. 3.1, we present the definition of "scale level", which determines how to group-wisely process the image/BEV features in the cyclic refiner, and serves as an important clue in the association strategy. For the image features, which usually have small spatial sizes (e.g., 15×25), we set scale level $L = 3$ to model the object attentive from hierarchical large, middle, and small levels, which also corresponds to the kernel sizes $\{5, 3, 1\}$ of DCNs, respectively. For the BEV features, which represent the whole driving scenarios and usually have a larger spatial size of 200×200, we set $L = 5$ to perform dedicated refining. The kernel sizes of DCNs are $\{9, 7, 5, 3, 1\}$ in this case. Furthermore, the scale levels in OAA follow the settings in BEV features, since the object size may vary in different camera views. More analyses are presented in Sec. 4.4.

4.2 State-of-the-art Comparison

4.2.1 NuScenes Detection Evaluation

Table 1 presents the performance comparison with state-of-the-art methods on both validation and test splits of nuScenes detection benchmark. The proposed CycSparseBEV and CycBEVFormer-Small outperform the baselines for 1.8%/4.3% mAP and 2.4%/3.9% NDS on the test split, respectively. On the indicator mAVE which reflects the ability of temporal modeling, our methods impressively surpass the baseline SparseBEV /BEVFormer-Small for 2.2%/4.7%, respectively. The results prove that filtering target-irrelevant distractors before temporal fusion is necessary to achieve better representation learning. On the validation split, our Cyc-

Table 1 Comparison with state-of-the-art detectors on nuScenes

Method	Backbone	val split							test split						
		mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D Wang et al. (2021)	R101	0.343	0.415	0.725	0.263	0.422	1.292	0.153	0.358	0.428	0.690	0.249	0.452	1.434	0.124
PGD Wang et al. (2022a)	R101	0.369	0.428	0.683	0.260	0.439	1.268	0.185	0.386	0.448	0.626	0.245	0.451	1.509	0.127
PETR Liu et al. (2022a)	R50	0.339	0.403	0.748	0.273	0.539	0.907	0.203	–	–	–	–	–	–	–
PETR Liu et al. (2022a)	R101	–	–	–	–	–	–	–	0.391	0.455	0.647	0.251	0.433	0.933	0.143
DETR3D Wang et al. (2022b)	R101	0.346	0.425	0.773	0.268	0.383	0.842	0.216	–	–	–	–	–	–	–
DETR3D Wang et al. (2022b)	V2-99	–	–	–	–	–	–	–	0.412	0.479	0.641	0.255	0.394	0.845	0.133
PolarFormer Jiang et al. (2022)	R101	0.396	0.458	0.700	0.269	0.375	0.839	0.245	0.415	0.470	0.657	0.263	0.405	0.911	0.139
BEVDet† Huang et al. (2021)	R50	0.312	0.392	0.691	0.272	0.523	0.909	0.247	–	–	–	–	–	–	–
BEVDet† Huang et al. (2021)	V2-99	–	–	–	–	–	–	–	0.424	0.488	0.524	0.242	0.373	0.950	0.148
MatrixVT Zhou et al. (2022)	R101	0.396	0.467	0.577	0.261	0.397	0.870	0.207	–	–	–	–	–	–	–
Fast-BEV Huang et al. (2023)	R101	0.402	0.531	0.582	0.278	0.304	0.328	0.209	–	–	–	–	–	–	–
M^2BEV Xie et al. (2022)	R101	0.417	0.470	0.647	0.275	0.377	0.834	0.245	0.429	0.474	0.583	0.254	0.376	1.053	0.190
BEVDepth† Li et al. (2022b)	R101	0.412	0.535	0.565	0.266	0.358	0.331	0.190	–	–	–	–	–	–	–
BEVStereo† Li et al. (2023)	R50	0.372	0.500	0.598	0.270	0.438	0.367	0.190	–	–	–	–	–	–	–
BEVStereo† Li et al. (2023)	V2-99	–	–	–	–	–	–	–	0.525	0.610	0.431	0.246	0.358	0.357	0.138
SOLOFusion† Pang et al. (2022)	R50	0.427	0.534	0.567	0.274	0.511	0.252	0.181	–	–	–	–	–	–	–
SOLOFusion† Pang et al. (2022)	ConvNeXt-B	–	–	–	–	–	–	–	0.483	0.582	0.503	0.264	0.381	0.246	0.207
FrustumFormer Wang et al. (2023a)	R101	0.457	0.546	0.624	0.265	0.362	0.380	0.191	0.478	0.561	0.575	0.257	0.402	0.411	0.132
MV2D Wang et al. (2023b)	R50	0.459	0.546	0.613	0.265	0.388	0.385	0.179	0.472	0.554	0.587	0.251	0.464	0.419	0.127
BEVDet4D Huang and Huang (2022)	R50	0.355	0.482	0.619	0.276	0.527	0.338	0.195	0.369	0.493	0.602	0.266	0.552	0.377	0.121
CycBEVDet4D	R50	0.374	0.537	0.639	0.230	0.316	0.322	0.193	0.398	0.545	0.596	0.258	0.400	0.356	0.133
BEVFormer-Small Li et al. (2022c)	R101	0.370	0.479	0.721	0.279	0.406	0.437	0.220	0.377	0.487	0.675	0.268	0.454	0.479	0.138
CycBEVFormer-Small	R101	0.398	0.505	0.650	0.276	0.379	0.439	0.201	0.420	0.526	0.607	0.262	0.405	0.432	0.133
BEVFormer-Base Li et al. (2022c)	R101	0.416	0.517	0.673	0.274	0.372	0.394	0.198	0.435	0.538	0.621	0.254	0.400	0.435	0.143
CycBEVFormer-Base	R101	0.433	0.532	0.639	0.270	0.318	0.416	0.201	0.452	0.549	0.575	0.255	0.405	0.407	0.131
SparseBEV Liu et al. (2023)	R50	0.448	0.558	0.581	0.271	0.373	0.247	0.190	0.466	0.568	0.555	0.254	0.443	0.269	0.128
CycSparseBEV	R50	0.467	0.582	0.533	0.261	0.297	0.230	0.188	0.484	0.592	0.500	0.246	0.389	0.247	0.128

For fair comparisons, we reproduce the baseline method under the same settings as our method. † indicates methods using CBGS training Zhu et al. (2019) which brings 4.5× training cost. No other tricks (e.g., CBGS, test time augmentation) are used during training and test in our methods

SparseBEV achieves 46.7% mAP, surpassing most recent SOTA detectors with the backbones of ResNet-50 or ResNet-101 He et al. (2016), without any bells and whistles (*e.g.,* test time augmentation). LSS-based BEVDet4D Huang and Huang (2022) also achieves consistent improvements with our Cyclic Refiner. The resulted object-aware detector, *i.e.,* CycBEVDet4D, outperforms the baseline for impressive 2.9% mAP and 5.2% NDS on the test split respectively, showing the effectiveness of object-aware temporal representation learning. All models and configs for the test split evaluation will be released.

4.2.2 NuScenes Tracking Evaluation

In Table 2, we report our performances on both validation and test splits of nuScenes tracking benchmark. The proposed CycSparseBEV achieves state-of-the-art performance in the camera-only tracking task and exceeds the baseline method by a large margin. Specifically, our CycSparseBEV achieves significant AMOTA improvements of 13.9%/14.6% on the test/val-idation splits over the baseline model. CycBEVFormer-Base and CycBEVDet4D also outperform the baselines for 13.0% /16.0% AMOTA on the test split respectively, showing the generalization of the proposed cyclic pipeline and object-aware association. Moreover, AMOTP is an important criterion in practical applications to evaluate the precision of a tracking system, which is crucial for safe autonomous driving. As shown in Table 2, our CycSparseBEV decreases the AMOTP for significant 45.3%/47.0% on the test and validation splits respectively, showing the robustness and reliability of our model. CycBEVFormer-Base and CycBEVDet4D also achieve considerable gains with the proposed cyclic refiner and object-aware association, *i.e.,* 1.452/1.618 $\rightarrow$ 1.055/1.317 AMOTP on the test split, showing the effectiveness and generality of our method.

4.3 Component-wise Ablation

This section presents the ablations on components of the proposed Cyclic Refiner and Object-aware Association.

4.3.1 Feature Refinement for Detection

We first analyze the influence of refining features with the proposed cyclic refiner on the detection task. The ablation experiments are conducted on CycBEVFormer-Small, and results are presented in Table 3. By directly applying the cyclic refiner on the image features of different views (i.e., "ImgRefine"), our model obtains mAP/NDS gains of 1.6% and 1.5%, respectively (②*v.s.*①). It shows that BEV representation also enjoys the bonus of the proposed cyclic refiner (i.e., "BEVRefine"), which improves 2.0% mAP compared with the baseline model (③*v.s.*①). When applying the proposed module on both image and BEV features, it can further bring 0.8%/1.5% gains on mAP/NDS (④*v.s.*③), respectively, which shows the effectiveness of the proposed object-aware temporal learning framework.

4.3.2 Feature Refinement and OAA for Tracking

We further analyze the influence of refining features with the cyclic refiner and conducting object-aware association (OAA) for tracking task based on CycBEVFormer-Small. Results are presented in Table 4. It shows that even without OAA, applying "ImgRefine" or "BEVRefine" still brings considerable performance gains compared with the baseline method BEVFormer (③,⑤,⑦*v.s.*①), which evidence that the proposed cyclic refiner empowers both detection and tracking tasks. Consistent with the detection task, refining both image and BEV features can achieve better performance, which surpasses the baseline for 3.1% AMOTA and 1.6% AMOTP, respectively (⑦*v.s.*①). Compared with the default association method SimpleTrack Pang et al. (2021), the proposed OAA shows superiority for achieving 2.3% AMOTA gains (②*v.s.*①). Notably, when applying both cyclic refiner and OAA on the baseline model, it brings considerable performance gains of 9.2% on AMOTA and 34.0% on AMOTP, respectively (⑧*v.s.*⑦). This demonstrates the complementarity of cyclic temporal learning and object-aware association for multi-view 3D tracking.

4.3.3 Backward Refinement and Temporal Fusion on Image Features

The "backward refinement" and temporal fusion are two crucial modules in our framework. As the baseline method BEVFormer only designs temporal fusion for BEV features, it is necessary to prove applying refinement and conducting temporal fusion on image features is crucial. The results on detection task are presented in Table 5, and tracking performances are reported in Table 6. Table 5 shows that without "backward refinement" and temporal fusion of image features, the detector obtains mAP score of 37.0% on the nuScenes detection `val` set. When refining the BEV features, it brings 1.1 points gains of mAP (②*v.s.*①). One interesting observation is that applying temporal fusion on image features without refinement degrades the performance for 0.9% mAP (③*v.s.*①), which in turn proves our claim that the distractors in historical features may interfere the representation learning. When simultaneously refining and applying temporal fusion on image and BEV features, it shows the best performance with 39.8% mAP and 50.5% NDS (④), which proves the effectiveness of our method. The experimental results on the tracking task (shown in Table 6) also demonstrate consistent conclusions.

Table 2 Comparison with state-of-the-art trackers on nuScenes dataset

Method	Backbone	val split					test split				
		AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	IDS↓	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	IDS↓
CenterTrack Zhou et al. (2020)	R101	–	–	–	–	–	0.046	1.543	0.233	0.043	3807
DEFT Chaabane et al. (2021)	R101	0.201	–	–	0.171	–	0.177	1.564	0.338	0.156	6901
QD-3DT Hu et al. (2022)	R101	0.247	1.507	0.405	0.221	5919	0.217	1.550	0.375	0.198	6856
MUTR3D Zhang et al. (2022)	R101	0.294	1.498	0.427	0.267	3822	0.270	1.494	0.411	0.245	6018
SRCN3D Shi et al. (2022)	R101	0.439	1.280	0.545	–	–	0.398	1.317	0.538	–	–
CC-3DT Fischer et al. (2022)	R101	0.429	1.257	0.534	0.385	2219	0.410	1.274	0.538	0.357	3334
BEVDet4D Huang and Huang (2022)	R50	0.261	1.516	0.398	0.253	6287	0.209	1.618	0.433	0.215	20,997
CycBEVDet4D*	R50	0.389	1.154	0.422	0.319	5466	0.369	1.317	0.432	0.310	3906
BEVFormer-Small Li et al. (2022c)	R101	0.274	1.506	0.456	0.249	8911	0.244	1.521	0.398	0.222	11,336
CycBEVFormer-Small*	R101	0.397	1.150	0.463	0.320	7239	0.372	1.175	0.469	0.299	8967
BEVFormer-Base Li et al. (2022c)	R101	0.337	1.426	0.496	0.316	9064	0.303	1.452	0.490	0.287	9755
CycBEVFormer-Base*	R101	0.469	1.002	0.457	0.354	3613	0.433	1.055	0.492	0.334	6621
SparseBEV Liu et al. (2023)	R50	0.376	1.261	0.545	0.346	2031	0.358	1.287	0.532	0.318	3422
CycSparseBEV*	R50	0.522	0.791	0.564	0.392	1419	0.497	0.834	0.561	0.365	2573

* Denotes using object-aware association (OAA)

Table 3 Influence of refining features with the proposed cyclic refiner

#	Method	ImgRefine	BEVRefine	mAP↑	NDS↑	mAVE↓	mAAE↓
①	Cycer-S			0.370	0.479	0.437	0.220
②	Cycer-S	✓		0.386	0.494	0.445	0.202
③	Cycer-S		✓	0.390	0.490	0.477	0.211
④	Cycer-S	✓	✓	0.398	0.505	0.439	0.201

Experiments are conducted with nuScenes detection `val` set. ImgRefine/BEVRefine denote refining image and BEV features, respectively. Cycer-S indicates CycBEVFormer-Small

Table 4 Influence of refining features and object-aware association (OAA) for tracking task

#	Method	ImgRefine	BEVRefine	OAA	AMOTA↑	AMOTP↓	MOTA↑
①	Cycer-S				0.274	1.506	0.249
②	Cycer-S			✓	0.297	1.492	0.282
③	Cycer-S	✓			0.295	1.482	0.281
④	Cycer-S	✓		✓	0.370	1.156	0.296
⑤	Cycer-S		✓		0.300	1.464	0.284
⑥	Cycer-S		✓	✓	0.376	1.149	0.296
⑦	Cycer-S	✓	✓		0.305	1.490	0.303
⑧	Cycer-S	✓	✓	✓	0.397	1.150	0.320

Experiments are conducted on nuScenes `val` set. The default association module without OAA is the standard SimpleTrack Pang et al. (2021)

Table 5 Influence of applying refinement and temporal fusion on image features for detection task

#	Method	Back	ImgTemp	mAP↑	NDS↑	mAVE↓	mAAE↓
①	Cycer-S			0.370	0.479	0.437	0.220
②	Cycer-S	✓		0.381	0.486	0.488	0.200
③	Cycer-S		✓	0.361	0.468	0.517	0.203
④	Cycer-S	✓	✓	0.398	0.505	0.439	0.201

Experiments are conducted on nuScenes `val` set. "Back" and "ImgTemp" denote "backward refinement" and temporal fusion for image features, respectively

Table 6 Influence of applying refinement and temporal fusion on image features for tracking task

#	Method	Back	ImgTemp	OAA	AMOTA↑	AMOTP↓	MOTA↑
①	Cycer-S			✓	0.297	1.492	0.282
②	Cycer-S	✓		✓	0.366	1.240	0.306
③	Cycer-S		✓	✓	0.327	1.454	0.310
④	Cycer-S	✓	✓	✓	0.397	1.150	0.320

Experiments are conducted on nuScenes `val` set. "Back" and "ImgTemp" denote "backward refinement" and temporal fusion for image features, respectively

Table 7 Ablation study for Object-aware Association on the nuScenes trcking `val` set

#	Method	MC	Buff	Cascade	AMOTA↑	AMOTP↓	MOTA↑
①	Cycer-S				0.305	1.490	0.303
②	Cycer-S	✓			0.356	1.287	0.303
③	Cycer-S		✓		0.349	1.293	0.301
④	Cycer-S			✓	0.365	1.295	0.308
⑤	Cycer-S	✓	✓		0.381	1.204	0.316
⑥	Cycer-S	✓		✓	0.384	1.189	0.317
⑦	Cycer-S		✓	✓	0.371	1.192	0.317
⑧	Cycer-S	✓	✓	✓	0.397	1.150	0.320

"MC", "Buff", and "Cascade" denote multiple clues, buffering strategy, and cascaded matching

4.3.4 Modules in Object-aware Association

We explore the influences of the modules in OAA, i.e., multi-clue matching (MC), buffering strategy (Buff), and cascaded scale-aware matching (Cascade), in Table 7. It shows that the three modules bring performance gains of 5.1%/4.4%/6/0% on AMOTA of the nuScenes tracking `val` set, respectively (②,③,④*v.s.*①). This proves the effectiveness and complementarity of our method. Notably, the Cascade solely brings the largest performance gains compared with the two other modules, which demonstrates associating objects in different size groups is necessary to perform robust multi-view 3D tracking. Applying all the three modules achieves the best performance with 39.7% AMOTA and 1.150 AMOTP, showing the cooperation of our designs contributes to an effective and robust tracker.

4.4 Further Analysis

4.4.1 Positive/Negative Influence of Temporal Fusion

As mentioned before, we argue that directly and simply using features from the previous frame in temporal fusion may introduce historical distractors and degrade the representation learning of future frames. We demonstrate this by ablating the positive/negative influence of temporal fusion based on the baseline method BEVFormer-Small Li et al. (2022c), as shown in Table 8. The improved performance on the commonly detected targets (*i.e.,* "Intersection Set") shows that temporal information could help to perceive the accurate position (*e.g.,* 0.762 mIoU of BEVFormer-Temporal). The newly detected objects on "***Difference Set***" of the **temporal** version further proves the effectiveness of temporal learning. However, the objects on "***Difference Set***" of the **static** version, which has been detected without temporal fusion, are surprisingly missed after introducing historical features. It evidences our claim that the background clutters of previous frames would distract the feature learning through temporal fusion, resulting in inferior performance (*e.g.,* 3.9% mAP loss). In comparison, our CycBEVFormer-Small in Table 8 suffers only 0.9% mAP loss, showing the effectiveness of the proposed "Backward Refinement" to relieve the negative influence of introducing distractors in temporal fusion.

4.4.2 FP/FN Number of Temporal/Static Version

Table 8 shows that the direct introduction of temporal fusion brings newly detected TP objects (*i.e.,* the "***Difference Set***" of the **temporal** version) while losing part detected ones (*i.e.,* the "***Difference Set***" of the **static** version) compared with the static version. A natural question is where the lost part goes and the newly detected part comes from? We further explore the temporal influence by detailing the FP/FN number under different score thresholds in Table 6. The FP curve shows that the direct temporal fusion causes more false positives compared with the static version (i.e., BEVFormer-Temporal *v.s.* BEVFormer-Static), which corresponds to the lost part. This demonstrates the historical background clutters would distract the target prediction. Notably, the difference of FP number between the temporal and static versions grows with the score thresholds, indicating that the target perception is more interfered by temporal distractors of high semantics (*e.g.,* threshold $>$ 0.6, as illustrated in Fig. 1b). In comparison, our CycBEVFormer-Temporal has fewer FPs with the proposed object-aware temporal learning, showing the effectiveness. The FN curve shows the positive influence of temporal fusion for decreasing the risk of missing targets, compared with the static version. Our method significantly improves the FN number of the baseline BEVFormer-Temporal under all score thresholds, evidencing the necessity of filtering target-irrelevant distractors before temporal fusion.

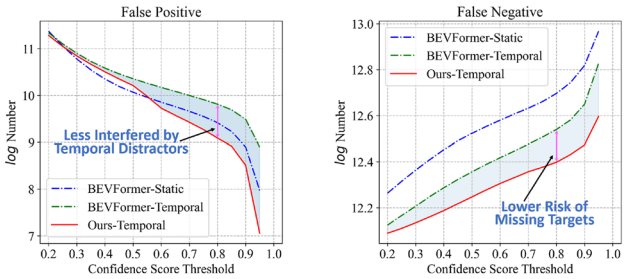

Fig. 6 Comparison of FP/FN number under different confidence score thresholds on nuScenes detection `val` set

4.4.3 Runtime Analysis

Running speed is a crucial metric for practical autonomous driving deployment. We present the runtime analysis to demonstrate that our method is well-balanced between the efficiency and effectiveness, as shown in Table 9. Based on BEVFormer-Base, our Cyclic Refiner only slows down the runtime for negligible 0.104 FPS while bringing considerable performance gains of 1.7% mAP scores (⑤*v.s.*④). Besides, our temporal method is also resource-friendly for achieving impressive 5.8% mAP improvements with a small storage cost of 6.3M compared with the static baseline (⑤*v.s.*③), which also proves the effectiveness of our object-aware representation learning. CycSparseBEV and CycBEVDet4D also enjoy the bonus of our method with small overload (⑦*v.s.*⑥, ⑨*v.s.*⑧), showing the generality of our method.

Table 8 Positive/Negative influence of temporal fusion

Method	Version	Detection Result							Overall (Insec. + ***Diff.***)		
		Intersection Set				***Difference Set***					
		mAP	NDS	Num.	mIoU	mAP	NDS	Num.	mAP↑	NDS↑	Num.↑
BEVFormer	Static	0.281	0.390	353,968	0.760	**0.039**	**0.015**	**38,337**	0.320	0.405	392,305
	Temporal	0.293	0.448		0.762	0.077	0.031	59,555	0.370	0.479	413,523
Ours	Static	0.312	0.402	380,932	0.761	**0.009**	**0.003**	**8254**	0.321	0.406	389,186
	Temporal	0.332	0.476		0.766	0.066	0.029	46,410	0.398	0.505	427,342

The static and temporal versions are compared by evaluating the commonly detected TP objects (*i.e.,* Intersection Set, dubbed Insec.) and exclusive ones (*i.e.,* ***Difference Set***, dubbed ***Diff.***) on the nuScenes detection `val` set. Num. denotes the number of detected TP objects and mIoU is the mean intersection over union

Table 9 Runtime and result comparison on nuScenes detection `val` set

#	Method	Backbone	HF	FPS	FLOPs	# Param.	mAP↑	NDS↑
①	BEVFormer-S	R101	1	2.738	622.3G	59.6M	0.370	0.479
②	Cycer-S	R101	1	2.601	630.5G	65.4M	0.398	0.505
③	BEVFormer-B	R101	0	n/a	1303.5G	68.7M	0.375	0.448
④	BEVFormer-B	R101	1	1.747	1324.9G	69.1M	0.416	0.517
⑤	Cycer-B	R101	1	1.643	1338.5G	75.0M	0.433	0.532
⑥	SparseBEV	R50	7	21.732	257.9G	44.6M	0.448	0.558
⑦	CycSparseBEV	R50	7	20.916	260.1G	46.9M	0.467	0.582
⑧	BEVDet4D	R50	8	2.014	1053.1G	57.8M	0.355	0.482
⑨	CycBEVDet4D	R50	8	1.819	1072.4G	63.7M	0.374	0.537

The inference is measured on a 3090 GPU. Cycer-S and Cycer-B indicate our CycBEVFormer-Small and CycBEVFormer-Base, respectively. "HF" denotes the number of used historical frames

Table 10 Ablation for image/BEV scale levels on nuScenes `val` set

#	Method	Img scale levels			BEV scale levels				mAP↑	NDS↑	AMOTA↑	AMOTP↓
		0	3	5	0	3	5	7				
①	Cycer-S	✓			✓				0.370	0.479	0.297	1.492
②	Cycer-S		✓		✓				0.386	0.494	0.370	1.156
③	Cycer-S			✓	✓				0.383	0.489	0.356	1.204
④	Cycer-S	✓				✓			0.386	0.489	0.366	1.142
⑤	Cycer-S	✓					✓		0.390	0.490	0.376	1.149
⑥	Cycer-S	✓						✓	0.381	0.486	0.344	1.237
⑦	Cycer-S		✓				✓		0.398	0.505	0.397	1.150

Cycer-S indicates CycBEVFormer-Small

Table 11 Ablation for detector and association method on the nuScenes tracking `val` set

#	Detector	Association	mAP↑	NDS↑	AMOTA↑	AMOTP↓
①	BEVFormer-S	SimpleTrak Pang et al. (2021)	0.370	0.479	0.274	1.506
②	BEVFormer-S	OAA	0.370	0.479	0.337	1.306
③	Cycer-S	SimpleTrak Pang et al. (2021)	0.398	0.505	0.305	1.490
④	Cycer-S	OAA	0.398	0.505	0.397	1.150

BEVFormer-S indicates the baseline BEVFormer-Small. OAA denotes the proposed Object-aware association

Table 12 Ablation for the object-aware strategy

#	Method	Object-aware strategy	mAP↑	NDS↑	AMOTA↑	AMOTP↓
①	BEVFormer-S	None	0.370	0.479	0.305	1.490
②	Cycer-S	Ocean Zhang et al. (2020)	0.367	0.386	0.286	1.516
③	Cycer-S	Cyclic Refiner	0.398	0.505	0.397	1.150

BEVFormer-S denotes the baseline BEVFormer-Small Li et al. (2022c)

4.4.4 Different Scale Levels for Image/BEV Feature

As mentioned above, the scale level is designed to divide each object into the matched group for customized modeling area, which prevents overlarge mask introducing background clutters or too small mask missing target details. We ablate the influence of different scale levels for refining image/BEV features based on CycBEVFormer-Small in Table 10. The results show that the scale level of 3 is optimal for image features to capture object-aware messages from different spatial scopes (②*v.s.*①,③). The version with a BEV scale level of 5 achieves superior performance (⑤*v.s.*④,⑥), which also evidences that the object sizes in the BEV feature are more diversified compared with the image space and require more fine-grained modeling. Notably, 3 image scale levels and 5 BEV scale levels contribute the best performance with 39.8% mAP and 50.5% NDS (⑦), showing the effectiveness of our proposed Cyclic Refiner.

4.4.5 Different Detectors and Association Methods

As a common sense, the improvement of detectors usually consistently enhances the tracking robustness. This conclusion is also revealed by our experiment in Table 7. Yet, since the tracking module is a plug-and-play design to guarantee its generality and simplicity in our work, which is not jointly trained with the detector, it thus has no clear effect on the detection model. We summarize the individual performance of detection and tracking in Table 11. The results show that our cyclic refiner significantly improves the baseline detector for 2.8% mAP and 2.6% NDS (③ *v.s.* ①), proving the effectiveness of our design for the detection task. Following the tracking-by-detection paradigm, our OAA surpasses the baseline tracker (*i.e.,* BEVFormer-S + SimpleTrack Pang et al. (2021)) for 6.3% AMOTA and 0.2 AMOTP (② *v.s.* ①). This evidences the robustness of our OAA in complex driving scenarios. Notably, the model combining our Cycer-S and OAA (④) achieves better tracking performance, proving that better detectors usually contribute to stronger tracking capability.

4.4.6 Object-Aware Strategy Between Cycer and Ocean

The proposed cyclic refiner exploits the predictions of each frame to filter target-irrelevant distractors in the learned features for object-aware learning. Similarly, Ocean Zhang et al. (2020) in 2D object tracking also adopts the predicted box as the prior proposal, which extracts corresponding ROI features for classification. We then compare the two object-aware strategies based on BEVFormer-Small. Specifically, we collect $\mathbf{e}_{cat}$ (see Sec. 3.1) as the object-aware ROI feature in the Ocean implementation. The results in Table 12 show that the strategy in Ocean degrades the performance for 0.3% mAP and 9.3% NDS respectively (② *v.s.* ①). In contrast, our cyclic refiner significantly improves the baseline for 2.8% mAP and 2.6% NDS respectively (③ *v.s.* ①). The reason

Table 13 Evaluating different cases with proposed object-aware association on nuScenes tracking `val` set: (a) small objects matching with image embeddings, (b) occluded objects with BEV embeddings, (c) robust tracking with head embeddings, (d) enhanced IoU matching with buffering strategy, and (e) dense objects with cascaded scale-aware matching. The baseline tracker is CycBEVFormer-Small. "MC$_{img}$", "MC$_{bev}$", "MC$_{head}$" indicate multi-clue matching with image/BEV/head embeddings. "Buff" and "Cascade" denote buffering strategy and cascaded matching. "MT" and "IDS" mean the number of mostly tracked trajectories and identity switches, respectively

(a)

Settings	AMOTA↑	AMOTP↓
w/o. MC$_{img}$	0.287	1.455
w/. MC$_{img}$	0.346	1.244

(b)

Settings	AMOTA↑	AMOTP↓
w/o. MC$_{bev}$	0.249	1.406
w/. MC$_{bev}$	0.291	1.286

(c)

Settings	MT↑	IDS↓
w/o. MC$_{head}$	2603	10,428
w/. MC$_{head}$	2987	8911

(d)

Settings	AMOTA↑	AMOTP↓
w/o. Buff	0.384	1.189
w/. Buff	0.397	1.150

(e)

Settings	AMOTA↑	AMOTP↓
w/o. Cascade	0.324	1.312
w/. Cascade	0.394	1.177

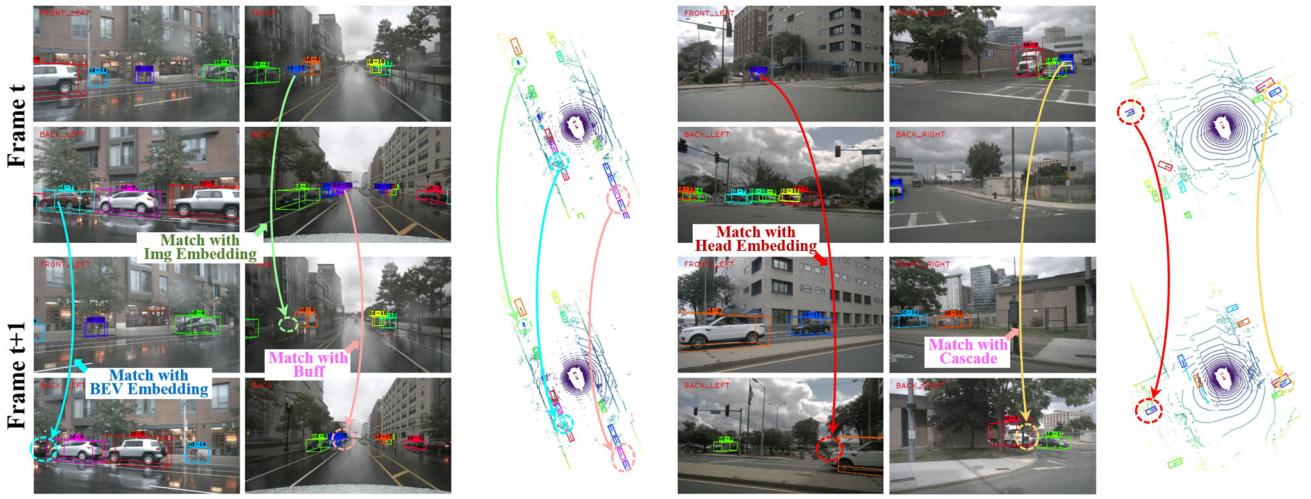

Fig. 7 Visualization of association with multi-clue matching, buffering strategy and cascaded scale-aware matching. The multiple clues include image/BEV/head embeddings

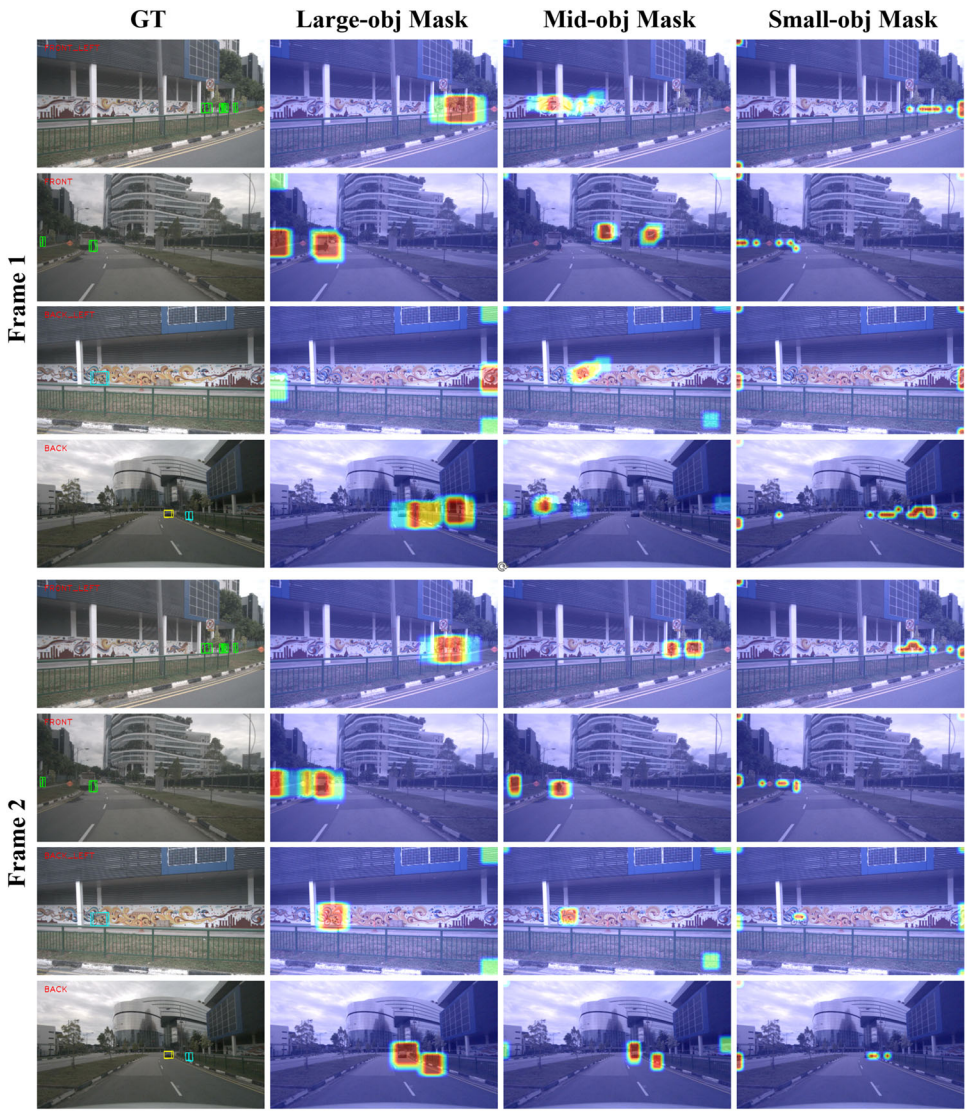

Fig. 8 Visualization of the object-aware masks in the multi-view images along time stamps (four rows of each frame represent "FRONT_LEFT", "FRONT", "BACK_LEFT" and "BACK" cameras respectively). The last three columns demonstrate the focus areas of the masks in different scale levels (i.e., large, mid and small, respectively). The objects are marked with colored 3D boxes in the first column

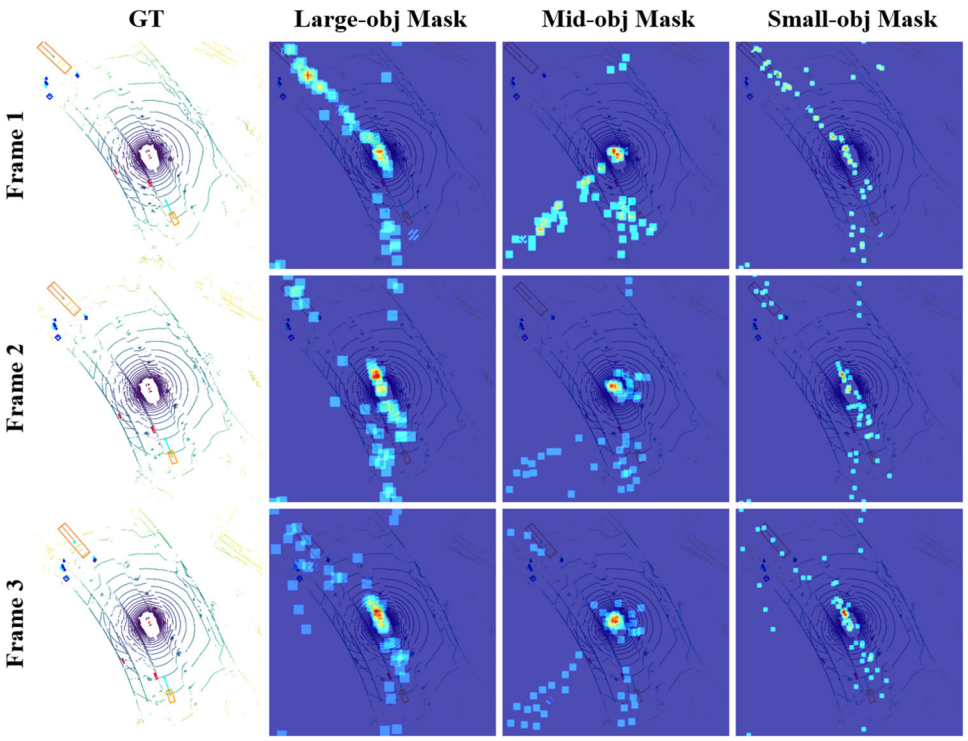

Fig. 9 Visualization of the object-aware masks in the bird's-eye-views along time stamps. The last three columns demonstrate the focus areas of the masks in different scale levels (i.e., large, mid and small, respectively). The objects are marked with colored boxes in the first row (Color figure online)

lies in the diversified distractors and variant object sizes in complex driving scenarios compared with that in 2D object tracking. As mentioned before, the distractors require careful discrimination (*e.g.,* mask prediction in our cyclic refiner). Yet, the Ocean strategy directly fuses the ROI features of the predictions without distinguishing the background clutters, leading to inaccurate results. Besides, the variant object sizes would also distract the extraction of target information, *e.g.,* a large box prediction for a small object. We then design the scale level to encode the scale information into the filter mask, which customs the spatial modeling size for each object. These demonstrate the superiority of our object-aware strategy in cyclic refiner.

4.4.7 Mechanism of Object-aware Association

We explore the working mechanism of the proposed multi-clue matching (MC), buffering strategy (Buff) and cascaded scale-aware matching (Cascade) with qualitative and quantitative analyses, as shown in Fig. 7 and Table 13. **(a) Matching with image embeddings MC**$_{img}$. For **small objects** that are 32×32 pixels or less (we follow the setting of COCO), it's difficult to acquire sufficient appearance clues from BEV features and motion information of minimal IoU between frames (see "CAM_FRONT" of the left case in Fig. 7). In contrast, the image ROI features (*i.e.,* image embeddings) are more informative under these circumstances. We prove our claim in Table 13a), which shows that the image embeddings impressively improve the baseline for 5.9% AMOTA and 0.211 AMOTP on small objects. **(b) Matching with BEV embeddings MC**$_{bev}$. Compared with 2D multi-view images, BEV provides a more general and clear 3D object description. When the objects are partially occluded in the images (*e.g.,* the half car in "CAM_BACK_LEFT" of the left case in Fig. 7), BEV features can provide more robust 3D appearance and shape clues for association. We prove this claim by exploring the influence of MC$_{bev}$ on the **occluded objects** (*i.e.,* visual levels ≤ 3 in nuScenes). In specific, Table 13b) shows that the BEV embeddings help to improve the tracking performance for 4.2% AMOTA and 0.12 AMOTP. **(c) Matching with head embeddings MC**$_{head}$. As the input features for classification and regression, the head embeddings contain more discriminative information regardless of the observation view compared with the image/BEV embeddings. This helps to improve the robustness of tracking. For the silver car in the right case of Fig. 7 that shifts from "CAM_FRONT_LEFT" at t to "CAM_BACK_LEFT" at $t+1$, it is capable of accurately associating the target with the head embeddings. The results in Table 13c) show that MC$_{head}$ helps to track extra 384 trajectories and reduce 1517 ID switches, evidencing our explanation. **(d) Matching with buffering strategy.** As mentioned before, we propose the buffering strategy to ensure reasonable box IoUs in BEV space, since the coverage scale of each box prediction in BEV plane is smaller than that in image space. The cases in "CAM_FRONT" of Fig. 7 illustrate the buffered 3D boxes help to generate accurate matches. Table 13d) shows that the proposed strategy generally improves the track-

ing performance of all objects for 1.3% AMOTA, proving its effectiveness. **(e) Matching with cascaded scale-aware strategy.** Large objects are more likely to cover nearby small objects in BEV space, which may cause false matches and track fragmentation (*e.g.,* the white car in the right case of Fig. 7). We propose Cascade to perform separate associations among objects with different scale levels. We evaluate the tracking performance on the objects that have a near target neighbor within 2 m. As shown in Table 13e), our cascade design brings performance gains of 7.0% AMOTA and 0.135 AMOTP.

4.4.8 Object-Aware Perception by Cyclic Refiner

The object-aware perception ability is the purpose of our designed "backward refinement" in the proposed cyclic refiner. It aims at increasing the responses of target regions and filtering distractors. We visualize the generated masks for refining multi-view image/BEV features in Fig. 8 and Fig. 9, respectively. The results show that the 2D target areas are captured and highlighted by the predicted masks of different scale levels, especially in the "FRONT", "FRONT_LEFT" and "BACK" camera images. The large-scale masks usually cover target-around areas to extract discriminative context messages, while the mid-scale and small-scale masks are adept at modeling fine-grained target information. Notably, few masks in the first frame are interfered by the background areas (*e.g.,* the "Mid-obj Mask") for not acquiring object-aware temporal messages. The misclassification is improved in the second frame by our Cyclic Refiner, demonstrating its effectiveness. Compared with the 2D images, the built BEV is highly abstract and determines the accuracy and robustness of final predictions. As shown in Fig. 9, the object-aware masks for refining BEV embed almost cover all the target areas, which are delivered into different scale levels to refine the learned representations. Similar to the image, the masks of the first frame cannot enjoy the object-aware temporal information and divert part focuses on the target-irrelevant areas. The distraction is relieved in later frames by exploiting the prior object-aware knowledge, which could help to well perceive and locate the objects. This evidences the necessity of solving the pollution of temporal fusion by target-irrelevant distractors and the effectiveness of our "backward refinement" for object-aware representation learning.

4.4.9 Background Clutter Suppression by the Cyclic Refiner

As mentioned, error accumulation is an inevitable problem in temporal fusion, but it is usually unconsciously ignored by recent works. Our cyclic refiner is indeed designed to relieve the temporal error accumulation caused by false positives (FPs) and background clutters (see Fig. 1). This is clearly different from previous temporal methods, *e.g.,* our baseline BEVFormer, which directly fuses features from the previous frames. In particular, we alleviate this issue by exploiting the object-aware mask prediction in cyclic refiner to suppress possible FPs, as shown in Fig. 10. The visualization illustrates that most FPs in the top 300 predictions from 900 object queries are suppressed by the predicted mask, which prevents polluting future features in temporal fusion. Notably, there are some hard examples of high scores mistakenly classified as object regions (*e.g.,* the false pedestrians in mid/small-obj masks of Frame 1) that are caused by the lack of effective temporal clues in the first frame. Then for the next frame, with the refined historical features that even contain several FP areas, our cyclic refiner is capable of collecting sufficient discriminative information to suppress the hard FPs (see Frame 2 in Fig. 10). This demonstrates the effectiveness of our cyclic refiner for relieving temporal error accumulation.

4.4.10 Object-Aware Temporal Learning

With the proposed cyclic pipeline, the refined features by "backward refinement" are forwarded to the next frame, which benefits the representation learning of future frames. To verify the effectiveness of temporal learning, we first visualize the sampling points with top 100 attention scores in the view-transformer (see BEVFormer Li et al. (2022c) for more details), as shown in Fig. 11. Compared with the baseline BEVFormer (the second row), our cyclic refiner (the third row) could exploit the refined target information of the last frame and force the attention to more accurately concentrate on the target-relevant areas. Notably, the sampling points are more and more centralized as time goes by, since the efficacy of our object-aware learning will gradually accumulate after longer temporal fusion. The sampling points with top 100 attention scores for the BEV feature in the task head are presented in Fig. 12. Compared with the information-intensive 2D images, the objects in BEV are relatively sparse, which raises more challenges to locate the target area and learn an object-aware representation. The results show that our sampling points (the third row) with the proposed "backward refinement" are more focused on target-relevant areas, in comparison with the baseline (the second row). Benefiting from the cyclic pipeline which constantly updates and refines the object-aware temporal information, our method could generate more centralized sampling on the targets in the later frames, evidencing the effectiveness.

4.4.11 Visualization of Detection and Tracking

Fig. 13 visualizes the detection results of the baseline and our model. By exploiting the refined object-aware representation in Frame 7, our method can transfer the prior knowledge to future frames and successfully predict the locations of occluded objects in Frame 8. In contrast, BEVFormer

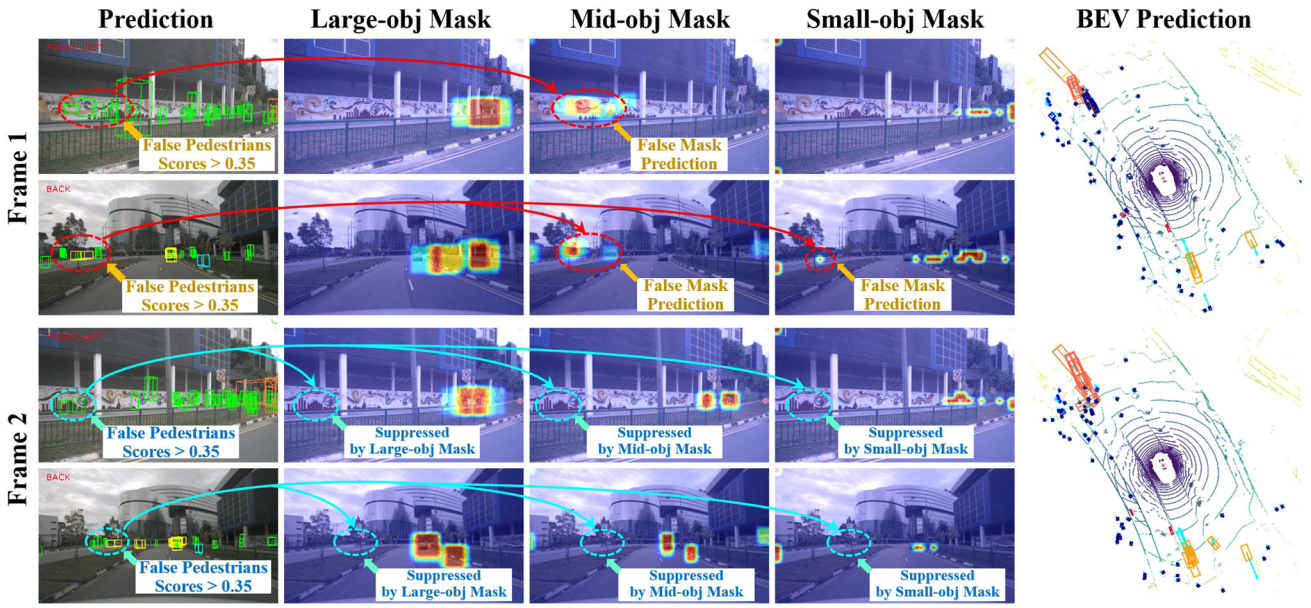

Fig. 10 Visualization of background clutter suppression by mask prediction along time stamps. For the total 900 predictions of each frame, we select the ones with the top 300 confidence scores for the cyclic refiner

Fig. 11 Visualization of the feature sampling points in the "FRONT" and "BACK" cameras. From left to right, the points with the top 100 attention scores are highlighted in the frames of different time stamps. Compared with the baseline method, our CycBEVFormer can concentrate on target regions

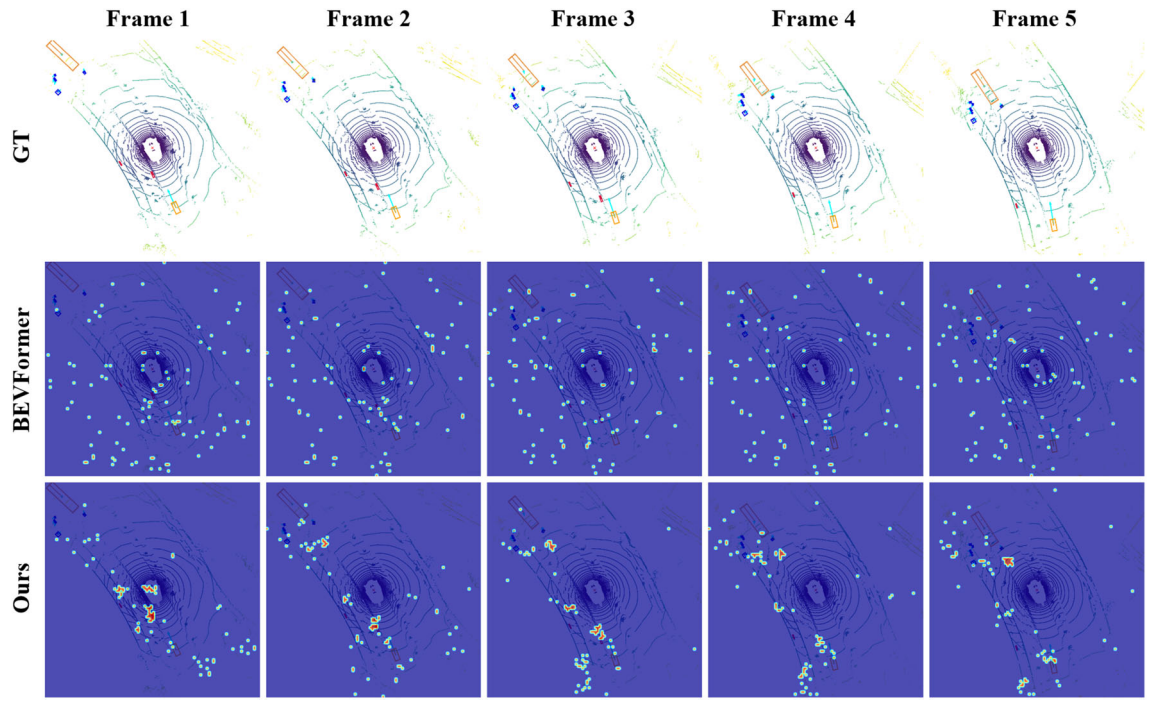

Fig. 12 Visualization of the sampling points in the bird's-eye-views. From left to right, the points with the top 100 attention scores are highlighted in the BEVs of different time stamps. Compared to the baseline BEVFormer (the second row), our CycBEVFormer could exploit the object-aware temporal information to enhance the target-perception ability in representation learning (the third row). The objects are marked with colored boxes in the first row (Color figure online)

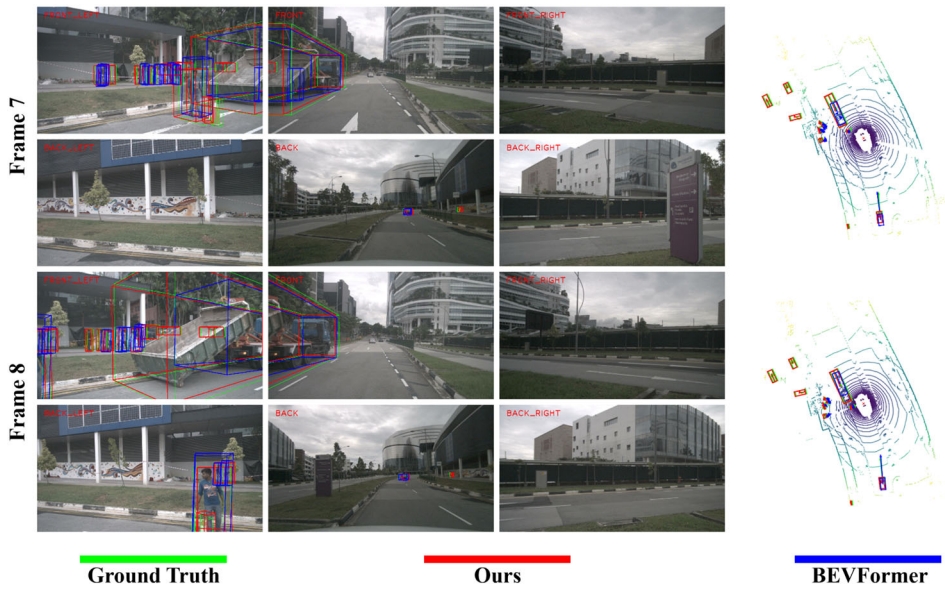

Fig. 13 Qualitative comparison between our CycBEVFormer (red) and the baseline method Li et al. (2022c) (blue) on detection task. Results show that our model achieves better recall after object-aware temporal fusion, especially in the cases that are not addressed by single frame detection (*e.g.,* occlusion) (Color figure online)

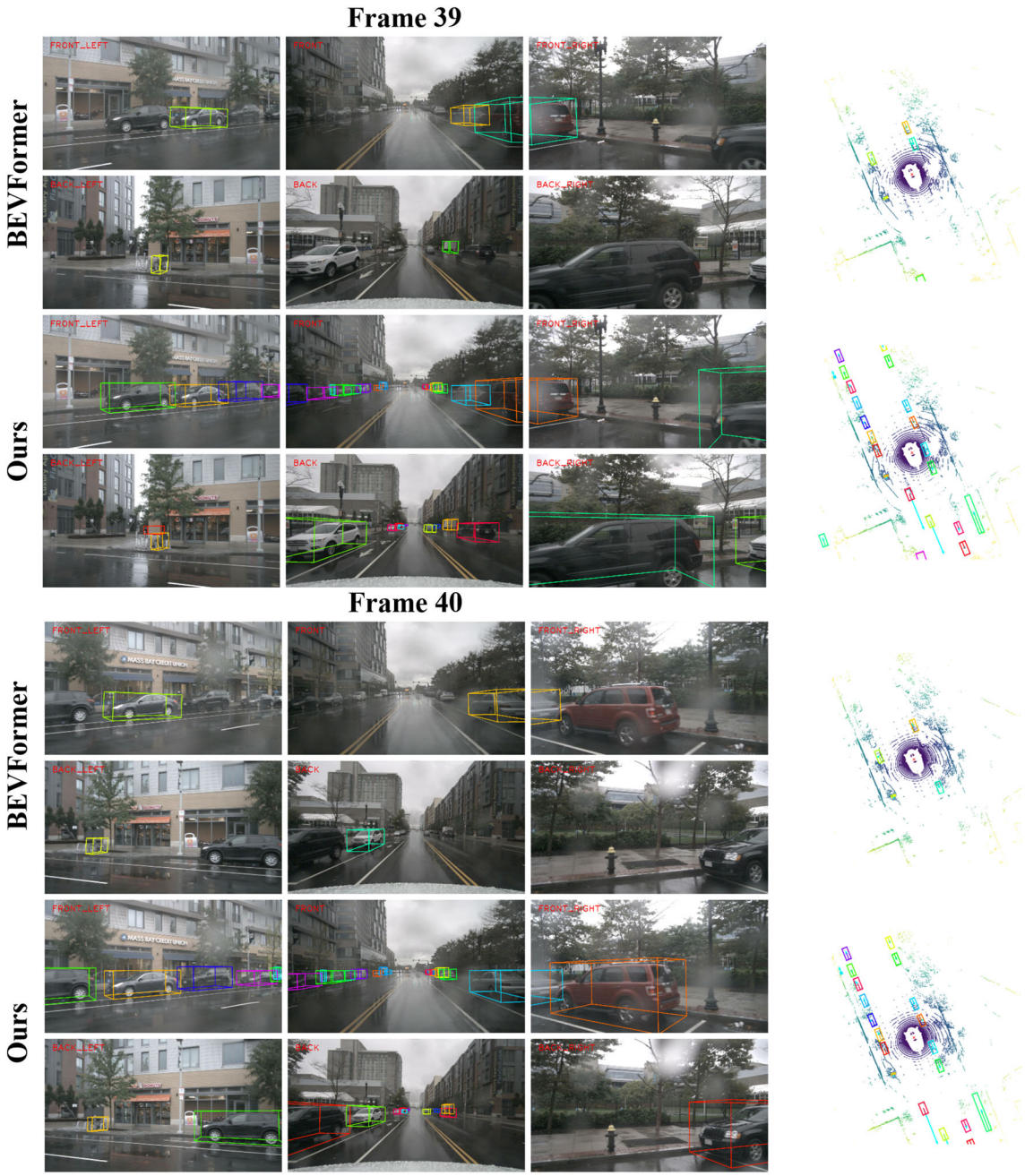

Fig. 14 Qualitative comparison between our method (bottom) and the baseline BEVFormer Li et al. (2022c) (top) on tracking task. We plot the box of each tracking object in both multi-view cameras and BEV, which is marked with the color corresponding to the identical tracking id. The comparison shows that our CycBEVFormer could perform robust tracking under complex scenarios (e.g., varied object sizes, occlusion and similar interferences) (Color figure online)

directly fuses features from the previous frame without filtering the distractors, which decreases the effectiveness of temporal learning and eventually causes false positives. This further proves the effectiveness of the proposed object-aware temporal learning framework in enhancing feature quality. Figure 14 shows the tracking results between our CycBEVFormer and the baseline. Compared with the detection task, tracking in complex driving scenarios is more sensitive to distractors, which may lead to false matches and increased fragmentations. Therefore, suppressing interferences with object-aware information is necessary to perform accurate and robust tracking. By exploiting the refined object-aware representations and assigned scale levels in cyclic refiner, our method can match each object with the multiple appearance

clues and scale-aware buffering strategy, which successfully maintains the identical tracking id under deformation, occlusion and interference with similar objects (the bottom row). In contrast, BEVFormer directly fuses features from the previous frame containing target-irrelevant distractors, which decreases the effectiveness of temporal learning and eventually causes tracking loss and id switches (the top row). This further proves the effectiveness of the proposed cyclic temporal learning framework and tailored object-aware association for multi-view 3D tracking.

4.4.12 Cyclic Refiner for Small Targets

Small objects are common in complex driving scenarios, which are hardly captured for the relatively small sizes in BEV space. One may wonder how our cyclic refiner models the features of small objects on the BEV plane of low resolution. Our cyclic refiner improves the recall of small targets from three aspects: **(1) Multiple feature sources.** As mentioned in Sec. 3.1, we collect object information from image/BEV features and head embeddings. For the small targets that contain minimal BEV features, the object information could be supplemented with the corresponding image features of more pixels. Besides, the compact head embeddings, which are responsible for object classification and regression, also provide target-relevant messages. With the three feature sources, our cyclic refiner can collect sufficient clues of small targets for object-aware temporal learning. **(2) Adaptive scale estimation.** For each object, we assign a scale level to determine the spatial modeling scope, which contributes to extracting features of small targets more effectively. Besides, the scale level also controls the kernel size of DCNs to model the masked features, further improving the effectiveness of feature sampling for small objects. **(3) Temporal fusion.** After object-aware modeling by our cyclic refiner, the refined features are fed into the next frame for temporal fusion. The contained object information guides to generate sampling points on target areas. Besides, Fig. 9 shows that the surrounding background clutters of small targets can be effectively suppressed by the predicted small-obj masks, further benefiting detection of small objects.

5 Conclusion

In this work, we aim to build a unified BEV detection and tracking framework by learning object-aware representations. The essence is to backward the information in model predictions to refine the afore-learned image and BEV features for temporal fusion. Tailored to the proposed cyclic learning pipeline, we design the object-aware association strategy to boost 3D tracking. Experimental results show that our method achieves consistent performance gains over different baselines on both detection and tracking tasks. Detection is the basic perception task in autonomous driving. Tracking is closer to downstream tasks, *i.e.,* planning and control. We hope our work can drive more interest in the research of designing unified and effective BEV detection and tracking framework.

Funding The funding was provided by National Natural Science Foundation of China (Grant No. 62176020), Key Technologies Research and Development Program (Grant No. 2020AAA0106800), Natural Science Foundation of Beijing Municipality (Grant No. Z180006), CAAI-Huawei MindSpore Open Fund and Chinese Academy of Sciences (Grant No. OEIP-O-202004), Key Laboratory of Road Traffic Safety Ministry of Public Security (Grant No. RCS2023K006).

Data Availability The datasets generated and/or analyzed during the current study are available in the original reference, *i.e.,* nuScenes Caesar et al. (2020) https://www.nuscenes.org/nuscenes.

References

Bechtel, W. (2013). *Philosophy of mind: An overview for cognitive science*. London: Psychology Press.

Bhat, G., Danelljan, M., Gool, L.V., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In: ICCV.

Bolme, D.S., Beveridge, J.R., Draper, B.A., & Lui, Y.M. (2010). Visual object tracking using adaptive correlation filters. In: CVPR.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In: CVPR.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In: ECCV.

Chaabane, M., Zhang, P., Beveridge, J.R., & O'Hara, S. (2021). Deft: Detection embeddings for tracking. arXiv.

Cui, Y., Jiang, C., Wang, L., & Wu, G. (2022). Fully convolutional online tracking. Computer Vision and Image Understanding.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In: ICCV.

Danelljan, M., Bhat, G., Khan, F.S., & Felsberg, M. (2019). Atom: Accurate tracking by overlap maximization. In: CVPR.

Fischer, T., Yang, Y.H., Kumar, S., et al. (2022). Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. NeurIPS.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In: ICCV.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: CVPR.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation

Hu, H.N., Yang, Y.H., Fischer, T., Darrell, T., Yu, F., & Sun, M. (2022). Monocular quasi-dense 3d object tracking. TPAMI.

Huang, B., Li, Y., Xie, E., Liang, F., Wang, L., Shen, M., Liu, F., Wang, T., Luo, P., & Shao, J. (2023). Fast-bev: Towards real-time on-vehicle bird's-eye view perception. arXiv.

Huang, J., & Huang, G. (2022). Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv.

Huang, J., Huang, G., Zhu, Z., & Du, D. (2021). Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv.

Jiang, Y., Zhang, L., Miao, Z., Zhu, X., Gao, J., Hu, W., & Jiang, Y.G. (2022). Polarformer: Multi-camera 3d object detection with polar transformers. arXiv.

Kuhn, H.W. (1955). The hungarian method for the assignment problem. Naval research logistics quarterly.

Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., & Li, Z. (2023). Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In: AAAI.

Li, Y., Chen, Y., Qi, X., et al. (2022). Unifying voxel-based representation with transformer for 3d object detection. arXiv.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z. (2022). Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., & Dai, J. (2022). Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV.

Liang, C., Zhang, Z., Zhou, X., Li, B., & Hu, W. (2022). One more check: making "fake background" be tracked again. In: AAAI.

Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., & Hu, W. (2022). Rethinking the competition between detection and reid in multiobject tracking. TIP.

Liu, H., Teng, Y., Lu, T., Wang, H., & Wang, L. (2023). Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In: ICCV.

Liu, Y., Wang, T., Zhang, X., & Sun, J. (2022). Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV.

Liu, Y., Yan, J., Jia, F., Li, S., Gao, Q., Wang, T., Zhang, X., & Sun, J. (2022). Petrv2: A unified framework for 3d perception from multi-camera images. arXiv.

Pang, Z., Li, Z., & Wang, N. (2021). Simpletrack: Understanding and rethinking 3d multi-object tracking. arXiv.

Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., & Zhan, W. (2022). Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv.

Philion, J., & Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV.

Price, C.J. (1998). The functional anatomy of word comprehension and production. Trends in cognitive sciences.

Reading, C., Harakeh, A., Chae, J., & Waslander, S.L. (2021). Categorical depth distribution network for monocular 3d object detection. In: CVPR.

Ren, S., He, K., Girshick, R., et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS.

Shi, Y., Shen, J., Sun, Y., Wang, Y., Li, J., Sun, S., Jiang, K., & Yang, D. (2022). Srcn3d: Sparse r-cnn 3d surround-view camera object detection and tracking for autonomous driving. arXiv.

Wang, T., Xinge, Z., Pang, J., & Lin, D. (2022). Probabilistic and geometric depth: Detecting objects in perspective. In: CORL.

Wang, T., Zhu, X., Pang, J., & Lin, D. (2021). Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: ICCV.

Wang, Y., Chen, Y., & Zhang, Z. (2023). Frustumformer: Adaptive instance-aware resampling for multi-view 3d detection. In: CVPR.

Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., & Solomon, J. (2022). Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: CORL.

Wang, Z., Huang, Z., Fu, J., Wang, N., & Liu, S. (2023). Object as query: Lifting any 2d object detector to 3d detection. In: ICCV.

Welch, G., Bishop, G., et al. (1995). *An introduction to the kalman filter*. NC, USA: Chapel Hill.

Xie, E., Yu, Z., Zhou, D., et al. (2022). M2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. arXiv.

Yang, F., Odashima, S., Masui, S., & Jiang, S. (2023). Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In: WACV.

Zhang, T., Chen, X., Wang, Y., et al. (2022). Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In: CVPR.

Zhang, Z., Peng, H., Fu, J., Li, B., & Hu, W. (2020). Ocean: Object-aware anchor-free tracking. In: ECCV.

Zhou, H., Ge, Z., Li, Z., & Zhang, X. (2022). Matrixvt: Efficient multi-camera to bev transformation for 3d perception. arXiv.

Zhou, X., Koltun, V., & Krähenbühl, P. (2020). Tracking objects as points. In: ECCV.

Zhu, B., Jiang, Z., Zhou, X., Li, Z., & Yu, G. (2019). Class-balanced grouping and sampling for point cloud 3d object detection. arXiv.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv.