M3SOT: Multi-Frame, Multi-Field, Multi-Space 3D Single Object Tracking

Jiaming Liu[1], Yue Wu[1*], Maoguo Gong[1], Qiguang Miao[1], Wenping Ma[1], Cai Xu[1], Can Qin[2]

[1]Xidian University, China
[2]Northeastern University, USA
{ljm@stu., ywu@, qgmiao@, wpma@mail., cxu@}xidian.edu.cn, gong@ieee.org, qin.ca@northeastern.edu

Abstract

3D Single Object Tracking (SOT) stands a forefront task of computer vision, proving essential for applications like autonomous driving. Sparse and occluded data in scene point clouds introduce variations in the appearance of tracked objects, adding complexity to the task. In this research, we unveil M3SOT, a novel 3D SOT framework, which synergizes multiple input frames (template sets), multiple receptive fields (continuous contexts), and multiple solution spaces (distinct tasks) in ONE model. Remarkably, M3SOT pioneers in modeling temporality, contexts, and tasks directly from point clouds, revisiting a perspective on the key factors influencing SOT. To this end, we design a transformer-based network centered on point cloud targets in the search area, aggregating diverse contextual representations and propagating target cues by employing historical frames. As M3SOT spans varied processing perspectives, we've streamlined the network—trimming its depth and optimizing its structure—to ensure a lightweight and efficient deployment for SOT applications. We posit that, backed by practical construction, M3SOT sidesteps the need for complex frameworks and auxiliary components to deliver sterling results. Extensive experiments on benchmarks such as KITTI, nuScenes, and Waymo Open Dataset demonstrate that M3SOT achieves state-of-the-art performance at 38 FPS. Our code and models are available at https://github.com/ywu0912/TeamCode.git.

Introduction

Visual object tracking is a basic task in computer vision, while single object tracking (SOT) is tracking a specific object in sequential data, considering only its initial pose. With the development of 3D sensors such as LiDAR, the acquisition of 3D data and the progress of 3D tasks become more active. In particular, great progress has been made in the 3D field based on point clouds (Wu et al. 2022, 2023b,d; Huang, Mei, and Zhang 2023; Liu et al. 2023c). Yet, SOT remains challenging due to the variation in object appearance and the sparseness caused by sensors with inherent limitations.

Existing 3D SOT methods can be summarized into two main paradigms, *i.e.*, Siamese network and spatio-temporal modeling. As a pioneering work, SC3D (Giancola, Zarzar, and Ghanem 2019) crops the target from the $(t-1)$-th frame and compares the target template with a large number of potential candidates in the t-th frame. P2B (Qi et al. 2020) optimizes this process by taking the cropped template and the search area as inputs, and propagating the cues to the search area by training again to predict the current bounding box. This idea has broad implications for subsequent research. Yet, the paradigm rooted in Siamese networks overlooks the background information from two sequential frames. Moreover, it fails when the target is potentially absent in the current frame. To address these issues, M2-Track (Zheng et al. 2022) presents a motion-centric approach, processing two point cloud frames as input and directly segmenting the target points from their respective backgrounds, eliminating the need for cropping. TAT (Lan, Jiang, and Xie 2022) ensures dependable target-specific feature propagation. It achieves this by sampling high-quality target templates derived from historical frames, applying template data across various timelines. However, these strategies predominantly operate on cropped subregions, which are fragmentary of essential contextual information in localization. Echoing the sentiments of CXTrack (Xu et al. 2023a), leveraging the contextual information surrounding the target for predicting its current bounding box is indeed an applicable move.

Hence, a logical proposition might be: *by integrating multi-frame input, motion modeling, and context extraction, would SOT performance enhance?* We refute this seemingly straightforward yet inelegant hypothesis. Through experimentation, we find that this approach places excessive strain on the network, potentially diminishing performance. Additionally, we re-examine the variables influencing SOT, un-

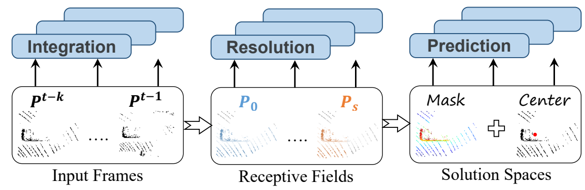

Figure 1: Illustration of the proposed M3SOT. M3SOT collects multi-frame point clouds for propagating target cues, and extracts spatio-temporal context information through multiple receptive fields. We set additional mask and center prediction tasks for the backbone at intermediate stages.

[*]Corresponding author.

earthing three pivotal insights that bolster tracking.

1) *Multiple input frames.* Directly using the template set composed of multi-frame point clouds (Lan, Jiang, and Xie 2022) and indirectly introducing motion modeling (Zheng et al. 2022) or updating the memory library (Xu et al. 2023b) is of great significance for tracking, as the unique temporal nature of SOT tasks can play a significant role. Inspired by these, our key idea is simple, *i.e.*, integrating past frames, gradually correcting errors and refining bounding boxes over time. Specifically, we employ a powerful attention mechanism to learn contexts from historical templates and then integrate them into the search area for rich information aggregation and precise object localization.

2) *Multiple receptive fields.* Fusion of multi-scale features is a well-known technique. For 3D SOT, most methods tend to use PointNet++ (Qi et al. 2017) or DGCNN (Wang et al. 2019) as the backbone for collecting multi-stage features. Yet, fusing these features is challenging, given the inherent tension between higher resolutions and expansive receptive fields. In response, we introduce a new multi-receptive field module with a transformer backbone designed to gather contextual information from multi-frame point clouds. Specifically, we obtain point cloud features representing the complete template through multi-stage computation-free range sampling and pointwise transformation. Our core insight is the conviction that predicting objects directly from sparse point features is both viable and effective (Chen et al. 2023).

3) *Multiple solution spaces.* Reviewing the previous SOT journey, we find that most methods rely only on the final localization head to discriminate the bounding box after pointwise transformation (Hui et al. 2021; Zheng et al. 2021; Liu et al. 2023b). This paradigm is agnostic to the intermediate stages of the network under training, since only the point features with maximum probability are finally acquired. For this reason, we revisit SOT, whose discriminative process should be asymptotic, *i.e.*, it can characterize the rough distribution of bounding boxes during the training process. To take full advantage of this cue, we set additional solution spaces in the intermediate stage for solving the mask and center of the predicted search area, with the former estimating the overall distribution of the bounding box and the latter pinpointing. Specifically, the designed transformer used to extract and transform point features has L stacked layers, where the output of each layer is supervised, while only the updated search area features of the last layer are forwarded to the localization head for the prediction.

As a significant result, we achieve the framework unification and unleash the potential of 3D SOT. In short, we inherit the above three findings into a framework, M3SOT, as shown in Figure 1. M3SOT is reinvigorated in the loop with spatio-temporal cues in the input phase and contextual information and task reasoning in the intermediate phase. Benefiting from the information aggregation of historical templates, sufficient contextual information and additional hidden spaces, M3SOT can efficiently track specific targets even in the case of occlusion or missing. Extensive experiments show that M3SOT achieves state-of-the-art performance on three benchmarks while running at 38 FPS on a single NVIDIA RTX 3090 GPU.

Related Work

3D SOT. Recently, 3D point cloud-based tracking can effectively avoid problems such as reliance on RGB-D information and sensitivity to illumination changes and object size variations in the 2D image tracking domain. SC3D (Giancola, Zarzar, and Ghanem 2019) is the first 3D Siamese tracker based on shape completion that generates a large number of candidates in the search area and compares them with the cropped template, taking the most similar candidate as the tracking result. The pipeline relies on heuristic sampling and does not learn end-to-end, which is very time consuming. P2B (Qi et al. 2020) addresses the previous problem by first using feature augmentation to enhance the perception of the specific template in the search area, and then using VoteNet (Qi et al. 2019) to localize the specific object in the search area. Most of the subsequent work basically follows the Siamese model. MLVSNet (Wang et al. 2021) aggregates information in multiple stages to achieve more effective target localization. BAT (Zheng et al. 2021) introduces a box-aware module to enhance discriminative learning between object templates and search areas. V2B (Hui et al. 2021) proposes a voxel-to-BEV object localization network, which projects sparse point features into a dense BEV feature map to address the sparsity of point clouds.

3D SOT by Transformer. Transformer (Vaswani et al. 2017) captures long-term dependencies of input sequences by the attention mechanism. Recently, transformer is applied to 3D vision and achieves good performance (Wu et al. 2023a,e,c; Yuan et al. 2023; Liu et al. 2023a). LTTR (Cui et al. 2021), PTTR (Zhou et al. 2022), and STNet (Hui et al. 2022) introduce various attention mechanisms to 3D SOT tasks for better target-specific feature propagation. CXTrack (Xu et al. 2023a) uses adjacent frames and employs a target-centric transformer to propagate target cues into the current frame while exploring the contextual information around the target. This "tracking by attention" paradigm is on the rise, as it has been shown to be effective for interactive learning of templates and search areas. However, these methods only exploit the target cues in the latest frame while ignoring the rich information in the historical frames. Our proposed method is applied in this paradigm, but extends the temporal scope of existing methods. In particular, we demonstrate that joint past inference can provide robust representations of spatio-temporal objects to improve the tracking.

3D SOT by Temporality. Continuous temporal context with logical processes is meaningful for 3D cognition, especially for dynamic 3D SOT task. M2-Track (Zheng et al. 2022) models consecutive frames as a motion-centric paradigm. TAT (Lan, Jiang, and Xie 2022) samples high-quality templates from historical frames and aggregate target cues. CAT (Gao et al. 2023) aggregates the features of historical frames to enhance the representations of the templates. MBPTrack (Xu et al. 2023b) designs an external memory for historical frames, and propagates the tracked target clues from the memory to the current frame. Unlike them, we utilize the contextual information of historical frames to learn interactively with the current frame respectively. As our insight is simple and efficient: there is a comprehensive transformer and a task solver, just make sure the inputs are sufficient.

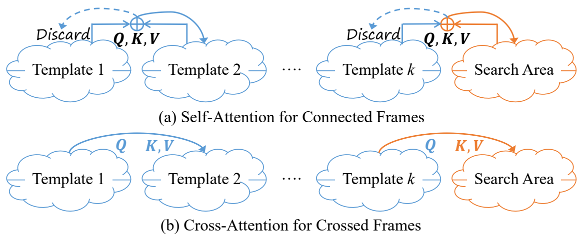

Figure 2: Illustration of frame-by-frame target propagation. We verify two generative paradigms (a) and (b) by attention.

Template set size	1	2	3	4
(a) self-attn (Precision)	82.1	74.1(↓8.0)	71.9(↓2.2)	71.6(↓0.3)
(b) cross-attn (Precision)	81.6	76.3(↓5.3)	74.5(↓1.8)	72.1(↓2.4)

Table 1: Generative paradigms for frame-wise propagation.

Pilot Study: Revisit Multi-Frame 3D SOT

Problem Formulation. In the 3D SOT task, given the initial bounding box (BBox) of the target in the first frame, the tracker aims at predicting the BBox of the target in the subsequent search area point cloud $\boldsymbol{P^t} \in \mathbb{R}^{N_t \times 3}$. It is generally assumed that the target size is fixed, and the rotation direction is just around the z-axis. Therefore, for each frame $\boldsymbol{P^t}$, the tracker only regresses the translational offsets $(\Delta x, \Delta y, \Delta z)$ and rotational angles $\Delta\theta$ from $\boldsymbol{P^{t-1}}$ to $\boldsymbol{P^t}$.

Further, the multi-frame 3D SOT extends the previous formulation, *i.e.*, $\boldsymbol{P^{t-1}}$ becomes $\boldsymbol{P^{t-K:t-1}}$. In addition, to represent the position and pose of the tracked target on the historical frames, we utilize the predicted targeting masks as auxiliary inputs. As a result, we reformulate the 3D SOT as

$$\texttt{Track}(\{\boldsymbol{P}, \boldsymbol{M}\}^{\boldsymbol{t-K:t-1}}, \boldsymbol{P^t}) \mapsto (\Delta x, \Delta y, \Delta z, \Delta\theta). \quad (1)$$

Since 3D SOT tracks the target in a dynamic sequence, it has timing. Therefore, we first discuss *whether timing can be reflected by frame-by-frame propagation?* In other words, the template set is passed progressively from the first frame to the next frame to the final search area frame. We design two generative paradigms to study it, as shown Figure 2.

(a) Self-attention. We concatenate consecutive frames into a new frame and perform self-attention to split the next frame taking the cue propagation from the previous frame.

(b) Cross-attention. We transform the previous frame into a query matrix and perform cross-attention with the next frame to propagate cues to the next frame.

These two generative paradigms are negative for multi-frame 3D SOT (see Table 1, tested in KITTI Car). We conclude that the target clues in the template set cannot be propagated to the search area frame by frame, because the template sets originally have their own targets, and redundant propagation may make the search area get wrong signals.

Differently, our intuition is that discontinuous frames can be complementary. This is contrary to the above, as it is unnecessary to build potential movement in an unbalanced point cloud sequence. Recalling at the difficulties of 3D SOT, we argue that sparseness and occlusion are the most important factors. Therefore, we directly adopt the many-to-one matching scheme for 3D SOT, as shown in Figure 3.

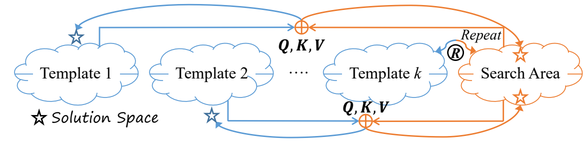

Figure 3: Illustration of M3SOT's many-to-one target propagation. Novel solution spaces for intermediate predictions.

Proposed Method: M3SOT

Overview. Based on Eq. (1), we propose M3SOT, a multi-frame, multi-field, multi-space SOT framework to fully utilize the spatial and temporal information of history frames. The overall framework is shown in Figure 4, where the template set and search area are first divided based on a 3D input sequence $\{\boldsymbol{P^i}\}_{i=1}^{t}$. A shared backbone with a hierarchical structure is used to extract local geometric features of the point cloud and aggregate them into point features. $\boldsymbol{F^i} \in \mathbb{R}^{N \times C}$ denotes the point cloud features in the i-th frame, and the corresponding targetness mask $\boldsymbol{M^i} \in \mathbb{R}^{N \times 1}$ is obtained from the first frame or estimated from past frames to identify the tracked targets in historical frames. To unify the computation, we design a targetness mask $\boldsymbol{M^t}$ initialized to 0.5 for the current frame due to the consistent initial state of the unknown points. Inspired by (Xu et al. 2023a), we concatenate the point features and targetness masks of the template set and the search area respectively to form a many-to-one pattern. Then, we design an interactive feature propagation module based on transformer to embed the geometric and mask information of the point cloud through GeoFormer and MaskFormer, respectively, and then feed them into SpaceFormer to perform spatial-separation tasks and feature transformations. Finally, the targeting proposals and confidences are predicted by the splited search area features with the support of localization head, and the proposal with the maximum confidence is identified as the bounding box.

Start from Inputs: Multi-frame SOT

Traditional Siamese trackers take as input a single template point cloud $\boldsymbol{P^{t-1}}$ and a search area point cloud $\boldsymbol{P^t}$, and match the closest target to the template within the search area. Differently, we focus on extracting rich temporal contextual information from a set of collected templates $\boldsymbol{P^{t-K:t-1}}$ for robust object localization. Considering the temporal context between the template set to the search area, we use simple and realistic closest frame sampling.

To exploit the potential of the template set to propagate cues to the search area, we construct multiple perceptual networks to predict the point-wise scores of the search area under the action of each historical template, and then select the point with the highest score for regression to the bounding box. Specifically, the historical templates and the search area are input to the network, where the former is used to provide reliable and necessary background information. Since different history templates are relevant to the search area in different degrees, the network can generate compatible regular terms for the weight of the search area with the support of shared weights.

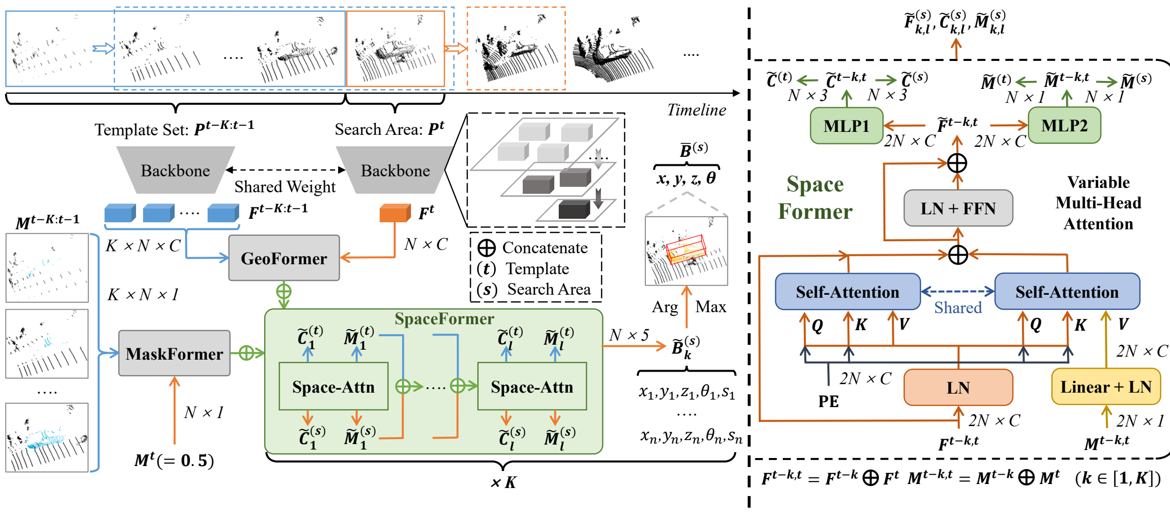

Figure 4: The overall framework of M3SOT. Given a point cloud sequence, M3SOT first employs a backbone with multiple receptive fields to extract the template set features of the previous k frames and the search area features of the current frame. Then, the geometric and mask features obtained by GeoFormer and MaskFormer are fed to a multi-task SpaceFormer to explore the spatio-temporal context of consecutive frames and propagate the target cues of each template into the search area. In addition, we design a multi-layer network with variable multi-head attention in SpaceFormer to predict masks and centers.

Refactor Point Features: Multi-field SOT

Given that the 3D SOT task is realized by a point-specific transformation of the search area, it is important how to inject more prior information into the points and enhance their discriminative ability. One fact is that the target-specific features provided by the template set are the most critical clues. In addition, it is also beneficial to add the perceptual information of the BBox to the point. BAT (Zheng et al. 2021) uses the 8 corners and 1 center of the point to the BBox as the additional information of the point. The difference is that we directly generate an additional mask set to record the probability of the point being in the BBox. These two factors can complement each other and are supervised differently.

We observe that using all points and masks directly as the only input results in two bad situations: 1) overloading the network and 2) poor and unstable results. This is due to the fact that not all points are equal, and the target points represent only a small fraction of the input points. Therefore, we design multiple receptive fields for the input point cloud, gradually decreasing the number and aggregating local information for points. Specifically, for the input point cloud $\boldsymbol{P_0}$, we generate new inputs $\boldsymbol{P_s}$ and $\boldsymbol{F_s}$ by a backbone with $\boldsymbol{S}$ range sampling and feature aggregation operations.

$$\boldsymbol{P_s} = RS(\boldsymbol{P_0}), \boldsymbol{F_s} = DGCNN(\boldsymbol{F_0}), \tag{2}$$

where RS requires no computation and retains the relationship between points, $DGCNN$ is used to extract point features with local aggregation (Wang et al. 2019).

Intuitively, deeper features are coarse but reliable since they gather more information through a larger receptive field. We generate the corresponding mask $\boldsymbol{M_s}$ through the sampled position indexes. Note that while range sampling may miss target points, background points aggregated with target points can yield robust predictions, which is an important inspiration for dealing with sparsity in point clouds.

Integrate a Hybrid Transformer: Multi-space SOT

To efficiently handle the template set and the search area, we aim to enhance both point features and localize a intra-frame target in the search area, while propagating target cues from historical frames to the current frame. Inspired by (Xu et al. 2023a), we propose a hybrid transformer that integrates multiple inputs and tasks with consideration of timing.

MaskFormer. To fully utilize the predicted results of history frames, we encode the point-box relationships of the template point clouds in a masked manner, *i.e.*, $\mathbf{ME}$. Note that the mask of the i-th point p_i is defined as

$$m_i^{(t)} = \begin{cases} 0, & p_i^{(t)} \text{ not in } \boldsymbol{B}^{(t)}, \\ 1, & p_i^{(t)} \text{ in } \boldsymbol{B}^{(t)}. \end{cases} \tag{3}$$

$\mathbf{ME}$ is similar to the positional encoding $\mathbf{PE}$, and N here denotes the number of sampled points. In addition, we set a mask initialized to 0.5 on the search area for computation.

GeoFormer. As the template set and the search area are processed by the backbone in a many-to-one manner, the extracted geometric features $\boldsymbol{F^{t-k,t}} = \boldsymbol{F^{t-k}} \oplus \boldsymbol{F^t}$ are sufficient to represent the overall information of the two, where k represents the k-th template in front of the search area.

SpaceFormer. To explore how point features predict bounding boxes, we feed $\boldsymbol{F^{t-k,t}}$ and $\boldsymbol{M^{t-k,t}}$ to the space-attention module in SpaceFormer, since both inputs are included, the cross-attention is potentially performed. This process is the cornerstone of delivering the target cues of the template set to the search area, as shown in Figure 4 (right).

Specifically, we first employ $LN(\cdot)$ (Ba, Kiros, and Hinton 2016) to normalize features, which is formulated as

$$\widetilde{\boldsymbol{F}}^{\boldsymbol{t-k,t}} = LN(\boldsymbol{F^{t-k,t}}). \tag{4}$$

Then, we build the basic components of attention: query $\boldsymbol{F_Q} \in \mathbb{R}^{2N \times C}$, key $\boldsymbol{F_K} \in \mathbb{R}^{2N \times C}$ and value $\boldsymbol{F_V} \in \mathbb{R}^{2N \times C}$, and add the positional encoding ($\mathbf{PE}$) to the query and key.

$$\boldsymbol{Q} = \boldsymbol{K} = \widetilde{\boldsymbol{F}}^{\boldsymbol{t-k,t}} + \mathbf{PE}, \boldsymbol{V} = \widetilde{\boldsymbol{F}}^{\boldsymbol{t-k,t}}. \tag{5}$$

Importantly, SpaceFormer employs a global multi-head self-attention module to model dependencies between point and mask features, formulated as

$$\begin{aligned}\widetilde{\boldsymbol{F}}^{\boldsymbol{t-k,t}} = \boldsymbol{F}^{\boldsymbol{t-k,t}} &+ MHA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \\ &+ MHA(\boldsymbol{Q}, \boldsymbol{K}, \mathbf{ME}),\end{aligned} \tag{6}$$

where MHA stands for multi-head attention, and the single-head attention with $d_h = C/H$ of the i-th in all subspaces being concatenated is $\boldsymbol{Q_i}$, $\boldsymbol{K_i}$, $\boldsymbol{V_i}$, calculated as

$$Attn(\boldsymbol{Q_i}, \boldsymbol{K_i}, \boldsymbol{V_i}) = softmax\left(\frac{\boldsymbol{Q_i}\boldsymbol{K_i}^T}{\sqrt{d_h}}\right)\boldsymbol{V_i}. \tag{7}$$

One reference is CXTrack which sets the number of layers to $L = 4$, the number of heads to $H = 1$. However, we argue that the gain this brings to multi-frame SOT is limited, since using the same configuration for different templates makes it difficult to model the network dynamically. Therefore, we propose a variable multi-attention mechanism that is simple and effective. Briefly, for different templates $\boldsymbol{P_k^{(t)}}$, we set the depth L of the network to be proportional to H for obtaining $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(t)}}$, and the same for $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}$.

For $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{t-k,t}}$ generated by different inputs at different layers, we separate them into $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(t)}}$ and $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}$. Our insight is that setting supervision on the outputs of each layer enables the targeting masks and centers to be consistently refined,

$$\widetilde{\boldsymbol{F}}_{\boldsymbol{l}}^{\boldsymbol{t-k,t}} = \widetilde{\boldsymbol{F}}_{\boldsymbol{l-1}}^{\boldsymbol{t-k,t}} + FFN(LN(\widetilde{\boldsymbol{F}}_{\boldsymbol{l-1}}^{\boldsymbol{t-k,t}})), \tag{8}$$

$$\widetilde{\boldsymbol{M}}_{\boldsymbol{k,l}}^{\boldsymbol{(t,s)}} = MLP_m(\widetilde{\boldsymbol{F}}_{\boldsymbol{l}}^{\boldsymbol{t-k,t}}), \widetilde{\boldsymbol{C}}_{\boldsymbol{k,l}}^{\boldsymbol{(t,s)}} = MLP_c(\widetilde{\boldsymbol{F}}_{\boldsymbol{l}}^{\boldsymbol{t-k,t}}), \tag{9}$$

where $FFN(x) = \mathrm{ReLU}\left(xW_1 + b_1\right)W_2 + b_2$.

Finally, $\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}$, $\widetilde{\boldsymbol{C}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}$, and $\widetilde{\boldsymbol{M}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}$ in the last layer are forwarded to X-RPN (Xu et al. 2023a) to predict the BBox,

$$\widetilde{\boldsymbol{B}}_{\boldsymbol{k}}^{\boldsymbol{(s)}} = X\!-\!RPN(\widetilde{\boldsymbol{F}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}, \widetilde{\boldsymbol{C}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}, \widetilde{\boldsymbol{M}}_{\boldsymbol{k,l}}^{\boldsymbol{(s)}}). \tag{10}$$

Since there are $\boldsymbol{K}$ templates, there are $\boldsymbol{K}$ versions of the search area, and we concatenate them to predict the BBox $\bar{\boldsymbol{B}}^{\boldsymbol{(s)}}$ with the maximum confidence score.

Experiments

Experimental Settings

Datasets. We compare the proposed M3SOT with state-of-the-art methods on three large datasets: KITTI (Geiger, Lenz, and Urtasun 2012), nuScenes (Caesar et al. 2020), and Waymo OpenDataset (WOD) (Sun et al. 2020). Following (Hui et al. 2021; Pang, Li, and Wang 2021): For KITTI, we divide the training sequence into three parts, 0-16 for training, 17-18 for validation, and 19-20 for testing. For the more challenging nuScenes, we use its validation split to evaluate our model, which contains 150 scenarios. For WOD, we evaluate our method on 1121 tracklets, which is categorized into easy, medium, and difficult parts based on the sparsity.

Method	Car 6424	Pedestrian 6088	Van 1248	Cyclist 308	Mean 14068
SC3D	41.3/57.9	18.2/37.8	40.4/47.0	41.5/70.4	31.2/48.5
P2B	56.2/72.8	28.7/49.6	40.8/48.4	32.1/44.7	42.4/60.0
LTTR	65.0/77.1	33.2/56.8	35.8/45.6	66.2/89.9	48.7/65.8
MLVSNet	56.0/74.0	34.1/61.1	52.0/61.4	34.3/44.5	45.7/66.7
BAT	60.5/77.7	42.1/70.1	52.4/67.0	33.7/45.4	51.2/72.8
PTT	67.8/81.8	44.9/72.0	43.6/52.5	37.2/47.3	55.1/74.2
V2B	70.5/81.3	48.3/73.5	50.1/58.0	40.8/49.7	58.4/75.2
CMT	70.5/81.9	49.1/75.5	54.1/64.1	55.1/82.4	59.4/77.6
PTTR	65.2/77.4	50.9/81.6	52.5/61.8	65.1/90.5	58.4/77.8
STNet	72.1/84.0	49.9/77.2	58.0/70.6	73.5/93.7	61.3/80.1
TAT	72.2/83.3	57.4/84.4	58.9/69.2	74.2/93.9	64.7/82.8
M2-Track	65.5/80.8	61.5/88.2	53.8/70.7	73.2/93.5	62.9/83.4
CXTrack	69.1/81.6	67.0/91.5	60.0/71.8	74.2/**94.3**	67.5/85.3
MBPTrack	73.4/84.8	**68.6/93.9**	**61.3**/72.7	**76.7/94.3**	**70.3**/87.9
M3SOT	**75.9/87.4**	66.6/92.5	59.4/**74.7**	70.3/93.4	**70.3/88.6**

Table 2: Comparison with the SOTA methods on KITTI. “Mean” denotes the average results weighted by frame numbers. Bold and underline represent the best and second best results, respectively. Success/Precision are reported.

Implementation Details. We dilate the ground truth BBox by 2 meters to track possible objects in the area. DGCNN (Wang et al. 2019) with different configurations is used as the feature extractor, and X-RPN (Xu et al. 2023a) with the same parameters is used as the localization head.

Evaluation Metrics. We follow One Pass Evaluation (OPE) (Kristan et al. 2016). For both predicted and ground truth BBoxes, *Success* measures the intersection over union (IOU) between the two BBoxes from 0 to 1, while *Precision* measures the area under curve (AUC) for the distance between their centers from 0 to 2 meters.

Experimental Results

Evaluation on KITTI. We perform a comprehensive comparison of M3SOT with previous state-of-the-art methods on the KITTI dataset, including SC3D (Giancola, Zarzar, and Ghanem 2019), P2B (Qi et al. 2020), LTTR (Cui et al. 2021), MLVS-Net (Wang et al. 2021), BAT (Zheng et al. 2021), PTT (Shan et al. 2021), V2B (Hui et al. 2021), CMT (Guo et al. 2022), PTTR (Zhou et al. 2022), STNet (Hui et al. 2022), TAT (Lan, Jiang, and Xie 2022), M2-Track (Zheng et al. 2022), CXTrack (Xu et al. 2023a) and MBPTrack (Xu et al. 2023b). As shown in Table 2, M3SOT performs excellently overall. Note that, in order to standardize the training setup, the reported M3SOT is based on a template set of size 2. However, the dependence on the number of history frames varies across categories, see the subsequent ablation experiments. Compared to TAT and MBPTrack, which also utilize history frames, we tap the following advantages of M3SOT: 1) Unlike TAT, which considers complex sampling and aggregation operations for the template set, M3SOT only requires simple many-to-one matching; 2) Unlike MBPTrack, which focuses on changing only the BBox, M3SOT filters the BBox under the action of historical templates on the search area. As a result, the well-thought-out and elegant M3SOT is more suitable for 3D SOT.

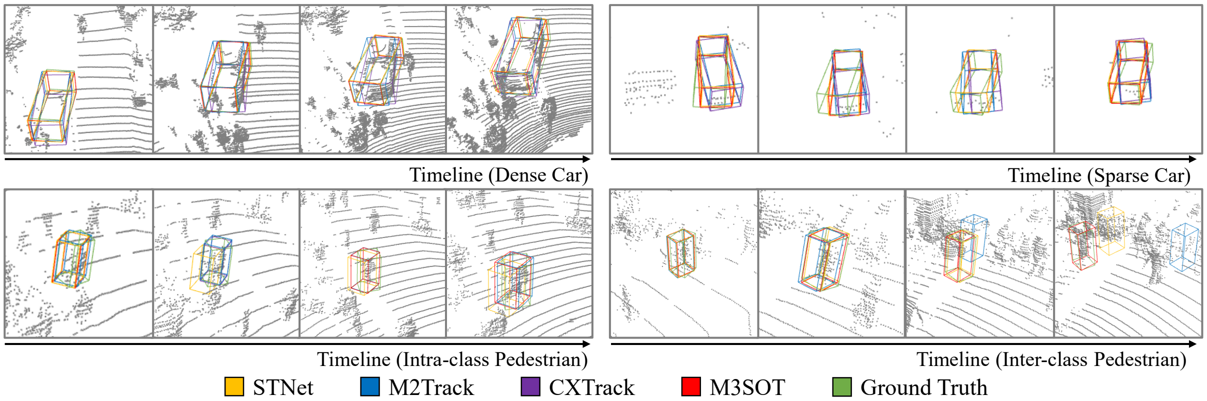

Figure 5: Visualization of tracking results from different methods on KITTI.

Method	Car 15578	Pedestrian 8019	Truck 3710	Bicycle 501	Mean 27808
SC3D	25.0/27.1	14.2/16.2	25.7/21.9	17.0/18.2	21.8/23.1
P2B	27.0/29.2	15.9/22.0	21.5/16.2	20.0/26.4	22.9/25.3
BAT	22.5/24.1	17.3/24.5	19.3/15.8	17.0/18.8	20.5/23.0
V2B	31.3/35.1	17.3/23.4	21.7/16.7	**22.2**/19.1	25.8/29.0
STNet	32.2/36.1	19.1/27.2	22.3/16.8	<u>21.2</u>/**29.2**	26.9/30.8
CXTrack	29.6/33.4	20.4/32.9	27.6/20.8	18.5/26.8	26.5/31.5
M3SOT-F1	**34.9/39.9**	<u>23.3</u>/25.6	**30.4/27.0**	16.5/22.6	**30.6**/33.7
M3SOT-F2	<u>34.2</u>/<u>38.6</u>	**24.6/37.8**	<u>29.6</u>/<u>25.5</u>	18.8/<u>27.9</u>	<u>30.5</u>/**36.4**
M3SOT-F3	33.7/38.3	22.2/<u>34.0</u>	26.4/23.0	18.7/25.8	29.1/<u>34.8</u>
M3SOT-F4	32.4/36.8	21.7/32.8	28.0/23.8	17.0/21.6	28.4/33.6

Table 3: Comparison with the SOTA methods on nuScenes.

We visualize the tracking results on KITTI, as shown in Figure 5. For Cars, the spatio-temporal context from historical frames allows M3SOT to produce discriminative semantic perceptions for the search area compared to non-multi-frame methods. For Pedestrians, most methods are prone to localize the wrong target due to the changing appearance of the target and distractors. However, due to the full use of temporal information, our M3SOT is able to accurately track the target in the presence of occlusions and appearance changes, and the aggregated information is more richer.

Evaluation on nuScenes and WOD. To validate the generalization ability of M3SOT, we follow (Hui et al. 2021; Pang, Li, and Wang 2021) and test the trained model on nuScenes and WOD. Note that the KITTI and WOD data are captured by 64-beam LiDAR, while the nuScenes data are captured by 32-beam LiDAR. Therefore, it is more challenging to generalize the trained model on the nuScenes dataset.

We set up four variants for M3SOT, each using models trained on the template set of sizes from 1 to 4. As shown in Table 3, our method achieves SOTA performance on the nuScenes, comprehensively outperforming previous methods. As a conclusion, M3SOT can not only generalize across different datasets, but also choose different configurations for different scenarios. Further, we visualize the impact of different template sets on the results in Figure 6 to explore how the cross-domain model aggregates features and predicts semantics in a new domain. It is observed that different

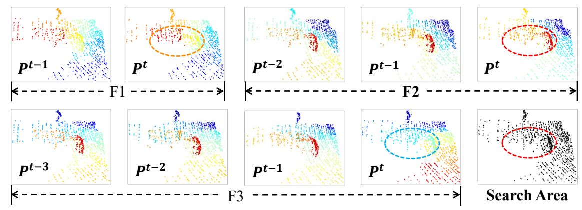

Figure 6: The effect of template set size on the search area.

template sets have different impacts on the search area. Like KITTI with different densities, the template point clouds in the F2 are sufficient to propagate valid and complementary target cues to the search area point cloud, and too many or too few templates are detrimental to feature propagation.

The comparison results on WOD are shown in Table 4. At different sparsity levels, our method is competitive than other methods, with an average gain of +0.6%/+0.9% compared to the recent MBPTrack. In any case, our M3SOT can not only accurately track a variety of targets, but also can be effectively generalized to unseen scenarios.

Ablation Studies

To verify the effectiveness of "M3" in M3SOT, we conduct ablation studies on the KITTI. In particular, "All" means that all categories are trained, not the weighted results above.

Multi-frame SOT. To explore the effect of template set size on the propagation of target cues in the search area, we report the results in Table 5. When the template set size is set to 1, only the previous frame $\boldsymbol{P}^{t-1}$ is used to train and test our model. M3SOT in this case can be regarded as a Siamese-based network. We argue that the targets of the Van category can be tracked well in this state, which is related to its less deformation and few intraclass interferers. A basic phenomenon is that different template set sizes have certain effects on the tracking performance of different categories. Based on this, we observe that performance starts to degrade when tracking more than 2 frames, since too many historical frames allow the network to collect redundant features and backfire output spaces. Compared with MBPTrack using 4 frames and TAT using 8 frames, our M3SOT can achieve

Method	Vehicle				Pedestrian				Mean
	Easy 67832	Medium 61252	Hard 56647	Mean 185731	Easy 85280	Medium 82253	Hard 74219	Mean 241752	427483
P2B	57.1/65.4	52.0/60.7	47.9/58.5	52.6/61.7	18.1/30.8	17.8/30.0	17.7/29.3	17.9/30.1	33.0/43.8
BAT	61.0/68.3	53.3/60.9	48.9/57.8	54.7/62.7	19.3/32.6	17.8/29.8	17.2/28.3	18.2/30.3	34.1/44.4
V2B	64.5/71.5	55.1/63.2	52.0/62.0	57.6/65.9	27.9/43.9	22.5/36.2	20.1/33.1	23.7/37.9	38.4/50.1
STNet	65.9/72.7	57.5/66.0	54.6/64.7	59.7/68.0	29.2/45.3	24.7/38.2	22.2/35.8	25.5/39.9	40.4/52.1
TAT	66.0/72.6	56.6/64.2	52.9/62.5	58.9/66.7	32.1/49.5	25.6/40.3	21.8/35.9	26.7/42.2	40.7/52.8
CXTrack	63.9/71.1	54.2/62.7	52.1/63.7	57.1/66.1	35.4/55.3	29.7/47.9	26.3/44.4	30.7/49.4	42.2/56.7
M2Track	68.1/75.3	58.6/66.6	55.4/64.9	61.1/69.3	35.5/54.2	30.7/48.4	29.3/45.9	32.0/49.7	44.6/58.2
MBPTrack	68.5/77.1	58.4/68.1	57.6/69.7	61.9/71.9	**37.5/57.0**	**33.0/51.9**	30.0/48.8	**33.7/52.7**	46.0/61.0
M3SOT	**70.4/79.6**	**60.7/70.6**	**61.5/73.3**	**64.5/74.7**	36.3/56.2	31.6/50.7	**30.1/48.9**	32.8/52.1	**46.6/61.9**

Table 4: Comparison with the SOTA methods on Waymo Open Dataset.

Frame	Car	Pedestrian	Van	Cyclist	All
$K = 1$	72.8/84.7	66.2/91.1	**63.1/78.0**	69.6/92.5	**70.1**/88.7
$\boldsymbol{K = 2}$	**75.9/87.4**	**66.6/92.5**	59.4/74.7	70.3/93.4	70.0/**88.9**
$K = 3$	73.5/85.3	64.8/90.8	58.9/74.4	**72.0/93.6**	68.7/87.7
$K = 4$	71.8/83.1	62.9/88.3	62.5/76.2	71.7/93.3	68.5/86.9

Table 5: Ablation studies: template set sizes.

Ratio	Car	Pedestrian	Van	Cyclist	All
[1]	59.2/73.0	51.7/89.2	28.4/28.6	33.3/82.7	54.8/76.6
[2]	60.6/75.8	52.1/89.1	50.4/56.8	22.5/42.9	57.3/77.5
[2,4]	70.4/81.8	53.9/87.5	53.4/60.4	39.7/85.0	64.5/83.7
[2,4,8]	**75.9/87.4**	**66.6/92.5**	**59.4/74.7**	**70.3/93.4**	**70.0/88.9**
[2,4,8,16]	71.6/83.5	58.2/88.9	58.3/66.2	22.3/38.1	64.4/83.5

Table 6: Ablation studies: feature generation ways.

M	C	Car	Pedestrian	Van	Cyclist	All
		24.4/32.3	36.7/67.0	16.7/21.0	63.6/90.1	25.7/41.0
✓		72.6/83.8	65.3/90.5	57.5/69.5	70.2/93.1	69.9/88.0
	✓	74.5/85.7	61.5/87.7	57.6/69.4	**72.3/93.5**	**70.1**/88.3
✓	✓	**75.9/87.4**	**66.6/92.5**	**59.4/74.7**	70.3/93.4	70.0/**88.9**

Table 7: Ablation studies: intermediate space tasks.

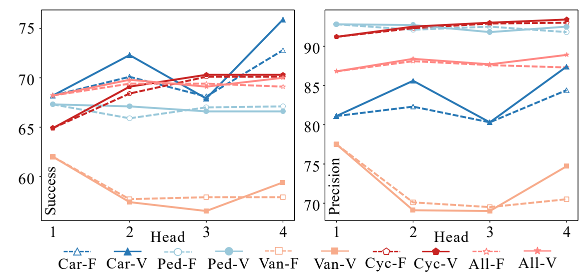

Figure 7: Ablation studies: variable multi-head attention *vs.* fixed multi-head attention.

superior results at a lower cost (with default 2 frames).

Multi-field SOT. We set five receptive fields for the input point cloud according to different point sampling and feature aggregation: "Ratio: [1]" means that all points are selected, and the feature aggregation operation of k-nearest neighbors is performed; the other four settings are performed in equal proportions 1/2 sampling and pointwise aggregation operations. In Table 6, we can observe that if all or half of the points after aggregated local features are used as input, the effect is not satisfactory. We argue that in the process of predicting unstable objects in sparse point clouds, the majority of points existing in shallow networks is noisy, and a large number of background points and target points cannot be distinguished. Besides, it is impossible to achieve good results only by extracting point features that may be less than the target. Therefore, we choose a three-layer ratio, which can be used as the optimal carrier to represent targets.

Multi-space SOT. We also investigate the effectiveness of mask prediction and center regression in SpaceFormer, which are used as intermediate tasks and supervised. Following CXTrack, we replace the mask and center layers present in the prediction space with classical transformer layers to integrate the targeting information, and analyze the effect of direct regression to the bounding boxes. As shown in Table 7, if both are eliminated, the performance is quite bad, even heavier than CXTrack's results. We infer that the point features with multiple frames and multiple fields are extremely rich and need to be analyzed progressively through a hierarchical network, requiring a shallow-to-depth process. In addition, the absence of mask or center makes M3SOT only slightly degraded, reflecting the inclusiveness of our method to them.

SpaceFormer. For the proposed SpaceFormer using two-stream self-attention, we study the advantages of variable multi-head attention for each category. As shown in Figure 7, compared to fixed multi-head attention, our method is more beneficial to SOT as the number of heads increases.

Conclusion

We discuss a comprehensive framework to serve 3D SOT. The proposed M3SOT consists of multi-frame, multi-field, multi-space, which is a tracking task-oriented pipeline. We analyze the necessity of each module in detail and reveal how to construct tasks to handle the SOT problem. Extensive experiments validate all aspects of the proposed method.

Limitations and Future Work. We reduce the network load by being task-oriented, however, coordinating such an integrated framework is not easy. We believe that potentially better configurations exist for different scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62036006, 62276200, 62103314), the Natural Science Basic Research Plan in Shaanxi Province of China (2022JM-327) and the CAAI-Huawei MINDSPORE Academic Open Fund. We acknowledge the support of MindSpore, CANN and Ascend AI Processor used for this research.

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Cui, Y.; Fang, Z.; Shan, J.; Gu, Z.; and Zhou, S. 2021. 3D object tracking with transformer. In *British Machine Vision Conference*.

Gao, J.; Yan, X.; Zhao, W.; Lyu, Z.; Liao, Y.; and Zheng, C. 2023. Spatio-Temporal Contextual Learning for Single Object Tracking on Point Clouds. *IEEE Transactions on Neural Networks and Learning Systems*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Giancola, S.; Zarzar, J.; and Ghanem, B. 2019. Leveraging shape completion for 3d siamese tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1359–1368.

Guo, Z.; Mao, Y.; Zhou, W.; Wang, M.; and Li, H. 2022. CMT: Context-Matching-Guided Transformer for 3D Tracking in Point Clouds. In *Proceedings of the European Conference on Computer Vision*, 95–111.

Huang, X.; Mei, G.; and Zhang, J. 2023. Cross-source point cloud registration: Challenges, progress and prospects. *Neurocomputing*, 126383.

Hui, L.; Wang, L.; Cheng, M.; Xie, J.; and Yang, J. 2021. 3D siamese voxel-to-bev tracker for sparse point clouds. *Advances in Neural Information Processing Systems*, 34.

Hui, L.; Wang, L.; Tang, L.; Lan, K.; Xie, J.; and Yang, J. 2022. 3D Siamese Transformer Network for Single Object Tracking on Point Clouds. In *Proceedings of the European Conference on Computer Vision*, 293–310.

Kristan, M.; Matas, J.; Leonardis, A.; Vojíř, T.; Pflugfelder, R.; Fernandez, G.; Nebehay, G.; Porikli, F.; and Čehovin, L. 2016. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11): 2137–2155.

Lan, K.; Jiang, H.; and Xie, J. 2022. Temporal-aware Siamese Tracker: Integrate Temporal Context for 3D Object Tracking. In *Proceedings of the Asian Conference on Computer Vision*, 399–414.

Liu, J.; Wu, Y.; Gong, M.; Liu, Z.; Miao, Q.; and Ma, W. 2023a. Inter-Modal Masked Autoencoder for Self-Supervised Learning on Point Clouds. *IEEE Transactions on Multimedia*.

Liu, J.; Wu, Y.; Gong, M.; Miao, Q.; Ma, W.; and Xie, F. 2023b. Instance-Guided Point Cloud Single Object Tracking with Inception Transformer. *IEEE Transactions on Instrumentation and Measurement*.

Liu, J.; Wu, Y.; Gong, M.; Miao, Q.; Ma, W.; and Xu, C. 2023c. Exploring Dual Representations in Large-Scale Point Clouds: A Simple Weakly Supervised Semantic Segmentation Framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2371–2380.

Pang, Z.; Li, Z.; and Wang, N. 2021. Model-free vehicle tracking and state estimation in point cloud sequences. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8075–8082.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9277–9286.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 5099–5108.

Qi, H.; Feng, C.; Cao, Z.; Zhao, F.; and Xiao, Y. 2020. P2b: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6329–6338.

Shan, J.; Zhou, S.; Fang, Z.; and Cui, Y. 2021. PTT: Point-track-transformer module for 3D single object tracking in point clouds. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1310–1316.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2446–2454.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5): 1–12.

Wang, Z.; Xie, Q.; Lai, Y.-K.; Wu, J.; Long, K.; and Wang, J. 2021. Mlvsnet: Multi-level voting siamese network for 3d visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3101–3110.

Wu, Y.; Hu, X.; Zhang, Y.; Gong, M.; Ma, W.; and Miao, Q. 2023a. SACF-Net: Skip-attention Based Correspondence Filtering Network for Point Cloud Registration. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wu, Y.; Liu, J.; Gong, M.; Gong, P.; Fan, X.; Qin, A.; Miao, Q.; and Ma, W. 2023b. Self-Supervised Intra-Modal and Cross-Modal Contrastive Learning for Point Cloud Understanding. *IEEE Transactions on Multimedia*.

Wu, Y.; Liu, J.; Gong, M.; Liu, Z.; Miao, Q.; and Ma, W. 2023c. MPCT: Multiscale Point Cloud Transformer with a Residual Network. *IEEE Transactions on Multimedia*.

Wu, Y.; Liu, J.; Gong, M.; Ma, W.; and Miao, Q. 2022. Centralized Motion-Aware Enhancement for Single Object Tracking on Point Clouds. In *IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 186–192.

Wu, Y.; Liu, J.; Yuan, Y.; Hu, X.; Fan, X.; Tu, K.; Gong, M.; Miao, Q.; and Ma, W. 2023d. Correspondence-Free Point Cloud Registration Via Feature Interaction and Dual Branch [Application Notes]. *IEEE Computational Intelligence Magazine*, 18(4): 66–79.

Wu, Y.; Zhang, Y.; Ma, W.; Gong, M.; Fan, X.; Zhang, M.; Qin, A.; and Miao, Q. 2023e. RORNet: Partial-to-Partial Registration Network With Reliable Overlapping Representations. *IEEE IEEE Transactions on Neural Networks and Learning Systems*.

Xu, T.-X.; Guo, Y.-C.; Lai, Y.-K.; and Zhang, S.-H. 2023a. CXTrack: Improving 3D point cloud tracking with contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1084–1093.

Xu, T.-X.; Guo, Y.-C.; Lai, Y.-K.; and Zhang, S.-H. 2023b. MBPTrack: Improving 3D Point Cloud Tracking with Memory Networks and Box Priors. In *Proceedings of the International Conference on Computer Vision*.

Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Ma, W.; and Miao, Q. 2023. EGST: Enhanced Geometric Structure Transformer for Point Cloud Registration. *IEEE Transactions on Visualization & Computer Graphics*, (01): 1–13.

Zheng, C.; Yan, X.; Gao, J.; Zhao, W.; Zhang, W.; Li, Z.; and Cui, S. 2021. Box-aware feature enhancement for single object tracking on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13199–13208.

Zheng, C.; Yan, X.; Zhang, H.; Wang, B.; Cheng, S.; Cui, S.; and Li, Z. 2022. Beyond 3d siamese tracking: A motion-centric paradigm for 3d single object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8111–8120.

Zhou, C.; Luo, Z.; Luo, Y.; Liu, T.; Pan, L.; Cai, Z.; Zhao, H.; and Lu, S. 2022. Pttr: Relational 3d point cloud object tracking with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8531–8540.