Boosting 3D Single Object Tracking with 2D Matching Distillation and 3D Pre-training

Qiangqiang Wu[1], Yan Xia*[2,3], Jia Wan[4], and Antoni B. Chan[1]

[1] Department of Computer Science, City University of Hong Kong
[2] Technical University of Munich
[3] Munich Center for Machine Learning (MCML)
[4] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
qiangqwu2-c@my.cityu.edu.hk, yan.xia@tum.de, jiawan1998@gmail.com, abchan@cityu.edu.hk

Abstract. 3D single object tracking (SOT) is an essential task in autonomous driving and robotics. However, learning robust 3D SOT trackers remains challenging due to the limited category-specific point cloud data and the inherent sparsity and incompleteness of LiDAR scans. To tackle these issues, we propose a unified 3D SOT framework that leverages 3D generative pre-training and learns robust 3D matching abilities from 2D pre-trained foundation trackers. Our framework features a consistent target-matching architecture with the widely used 2D trackers, facilitating the transfer of 2D matching knowledge. Specifically, we first propose a lightweight Target-Aware Projection (TAP) module, allowing the pre-trained 2D tracker to work well on the projected point clouds without further fine-tuning. We then propose a novel IoU-guided matching-distillation framework that utilizes the powerful 2D pre-trained trackers to guide 3D matching learning in the 3D tracker, i.e., the 3D template-to-search matching should be consistent with its corresponding 2D template-to-search matching obtained from 2D pre-trained trackers. Our designs are applied to two mainstream 3D SOT frameworks: memory-less Siamese and contextual memory-based approaches, which are respectively named SiamDisst and MemDisst. Extensive experiments show that SiamDisst and MemDisst achieve state-of-the-art performance on KITTI, Waymo Open Dataset and nuScenes benchmarks, while running at above real-time speed of 25 and 90 FPS on a RTX3090 GPU.

1 Introduction

3D Single Object Tracking (SOT) has emerged as a basic 3D task for numerous practical applications in autonomous driving [30,67], robotics [41,66] and virtual reality [1]. Given a target described as a 3D bounding box (bbox) in the first frame, 3D SOT aims to track the target across all the frames in the 3D scene by predicting the bbox pose and position, which can be considered as a 3D extension of the 2D SOT task [62] but using a different input modality of 3D point clouds.

* Corresponding Author

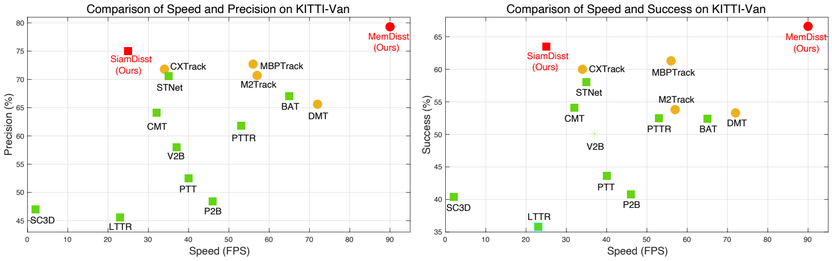
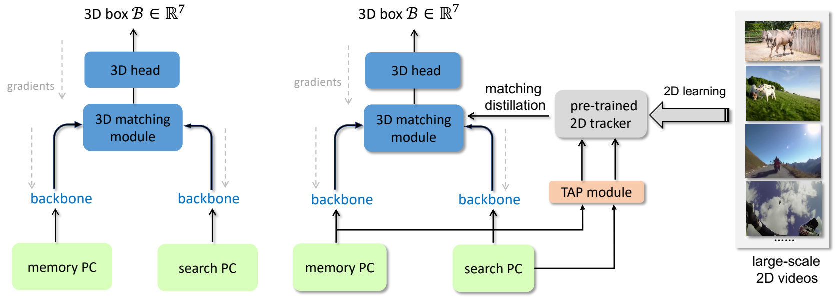


Fig. 1: Comparison of our SiamDisst and MemDisst to state-of-the-art 3D SOT on the Van category of the KITTI dataset [18]. Compared with Siamese-based trackers (plotted as *squares*), our SiamDisst achieves the best performance while running at the real-time speed of 25 FPS. Our lightweight MemDisst outperforms the other contextual memory and motion-based approaches (plotted as *circles*), setting new SOTA performance on KITTI-Van, and runs efficiently at 90 FPS. This demonstrates that our designs effectively boost 3D matching with limited category-specific training data (e.g., the KITTI-Van training set only contains 1994 examples).

As illustrated in Fig. 2, existing 3D SOT methods [38,62,73] mainly focus on learning robust 3D tracking models from sparse point cloud datasets [5,18]. For example, 3D siamese trackers [27,33,47,77] follow a siamese matching paradigm for accurate localization. To leverage more historical cues, MBPTrack [70] further extends this framework to a contextual memory-matching mechanism. Although the 3D matching models proposed by these methods play a key role in obtaining favorable performance [5,18], the sparsity and incompleteness of the point cloud training limits their abilities. Moreover, a paucity of *large-scale* category-specific 3D tracking datasets leads to insufficient training data for effectively learning the 3D matching modules in existing 3D trackers.

In contrast to the 3D SOT task, various large-scale annotated video datasets [29] are available for training 2D SOT methods. Existing 2D deep trackers can fully benefit from the feature backbones learned from 2D image [22] or video pre-training methods [62], enabling the learning of robust matching modules for 2D SOT on videos. The state-of-the-art trackers [38,62,73] with strong matching abilities show excellent tracking results even without using online memory mechanisms. With the remarkable success of 2D SOT, our work is motivated by the following questions: 1) *Inspired by 2D SOT, can we design a general and effective 3D SOT framework that benefits from 3D pre-training?* 2) *Can we transfer the well-learned matching ability of state-of-the-art 2D trackers to 3D matching in 3D SOT trackers, so as to facilitate training from limited category-specific point cloud data?*

The main challenge of achieving the transfer of the 2D matching ability lies in two aspects. First, there is a significant misalignment between the input modalities for 2D and 3D trackers, dense 2d images and sparse 3d point clouds, respectively. Note that in the 3D SOT setting, only the single modality of point



(a) Previous Siamese 3D SOT

(b) Our Proposed 3D SOT (i.e., SiamDisst and MemDisst)

Fig. 2: The comparison of 3D SOT frameworks (a) the previous Siamese 3D SOT trackers [27, 47, 77] focus on learning the 3D matching module from limited category-specific point cloud (PC) training data. (b) The proposed 3D SOT framework bridges the gap between 2D and 3D SOT by using our designed target-aware projection (TAP) module, which enables the 2D tracker [73] trained on large-scale 2D videos to operate well on the point clouds. The well-learned 2D matching pattern is further distilled to facilitate the 3D matching learning. Note that only the 3D-related modules (denoted as blue components) are optimized during the distillation training stage. The 2D tracker and TAP module are not used during online inference to increase speed.

cloud data is used as input, which makes it difficult to directly apply the current 2D SOT trackers. Second, how to effectively transfer the 2D matching knowledge in pre-trained 2D trackers to guide the learning of the 3D matching in 3D trackers is unclear given the differences in internal representations between 2D and 3D SOT, i.e., 2d dense feature maps and unordered 3D point embeddings.

To tackle the aforementioned challenges, we propose a general 3D SOT framework that can effectively utilize knowledge learned in 2D SOT, which is illustrated in Fig. 2(b). We first propose to align the different input modalities of 2D and 3D trackers via a learnable target-aware projection (TAP) module, which takes the raw point cloud as the input and projects them into the 2D image space. These projected images are suitable for 2D target matching and 2D bounding box regression by 2D *pre-trained* trackers, i.e., no fine-tuning of 2D trackers is required. Secondly, we propose a matching distillation approach that uses the knowledge in the high-quality 2D matching patterns from the 2D tracker as guidance for learning the matching patterns for 3D SOT. The proposed framework employs a standard Vision Transformer (ViT) as the matching module, which can naturally benefit from 3D pre-training approaches [42, 49] and also creates a design that is consistent with the 2D SOT trackers (e.g., OSTrack [73]) for more natural distillation of matching knowledge.

We implement our framework with both Siamese [26, 27] and contextual memory-based [70] matching paradigms, which are respectively named SiamDisst and MemDisst. Extensive experiments demonstrate that our SiamDisst achieves significant improvements over its baseline, and MemDisst sets new state-of-the-art performance on KITTI, while respectively running at a fast speed of 90 FPS on a single RTX3090 GPU.

In summary, the main contributions of our work are:

- We build a general SOT framework that employs a standard ViT as the target matching module, which is consistent with dominant 2D trackers [62, 73] for better knowledge transfer of matching patterns, and meanwhile can naturally benefit from 3D pre-training.
- We propose a lightweight target-aware projection (TAP) module to bridge the gap between the 2D and 3D tracking domains, which enables 2D pre-trained trackers to be operated on point cloud data without additional 2D tracker fine-tuning.
- We propose an IoU-guided matching distillation approach to guide 3D SOT learning, which facilitates transfer of knowledge from powerful 2D pre-trained foundation trackers to the 3D tracker. We show that the matching distillation facilitates the 3D model learning in the low-data regime.

2 Related Work

3D Single Object Tracking. Inspired by the great success in the 2D tracking community, much progress has been made in 3D SOT [43]. The pioneering work SC3D [19] proposes the first 3D Siamese network to compute the similarity between the template and 3D candidate proposals. However, the time-consuming matching pipeline in SC3D is not end-to-end trainable, thus degrading its performance on 3D tracking benchmarks [5, 18, 53]. Inspired by the 2D region proposal network (RPN) [20, 36], P2B [48] proposes a 3D RPN to generate high-quality 3D proposals, which achieves better tracking results while running at the real-time speed. BAT [77] extends P2B with a box-aware feature design, which provides a strong prior of the target shape. In contrast to previous approaches that directly regress 3D bboxes from a sparse point cloud, V2B [26] proposes a voxel-based localization head to localize the target in the Bird's Eye View (BEV).

With the development of transformers, 3D transformer SOT trackers (e.g., LTTR [8], PTTR [79] and CMT [21]) are designed for target-aware feature learning. Further extensions include iterative transformer blocks [27], historical cue modeling [33], contextual memory [69, 70], motion prediction [68, 78] and advanced architecture design [16, 64]. Despite these successes, learning robust 3D tracking models is still limited by the sparse and insufficient category-specific point cloud data. In this work, our aim is to leverage off-the-shelf 2D trackers trained on large-scale video data for enhancing 3D trackers, thus alleviating the current data bottleneck in 3D SOT.

2D Single Object Tracking. Traditional 2D SOT [10, 17, 23, 39, 59, 60] follow a correlation filter tracking framework due to its fast running speed and favorable tracking performance. With the development of the deep learning, much progress has been made in designing convolutional neural networks for video object tracking. Specifically, SiamFC [3] and SINT [50] employ a siamese neural network and treat tracking as a template matching problem. Based on SiamFC, many improvements have been made, including architecture design [14,35,36,76], unsupervised representation learning [56,61,62] and template updating [4,72,75]. Recent advances using transformers [7,9,38,63,71,73] achieve state-of-the-art performance on popular tracking benchmarks [13,25,32,65]. The great success in 2D tracking significantly inspires the development in the 3D tracking community. However, there is no previous work trying to bridge the gap between the 2D and 3D tracking areas. In this paper, we propose a lightweight target-aware projection (TAP) module to bridge the gap between 2D and 3D SOT, and use the 2D tracker to effectively guide the 3D learning.

Knowledge Distillation. Knowledge distillation (KD) aims to transfer the knowledge from a teacher model to a student model. Previous response-based KD approaches [24,54,55,80] focus on using the final output of the teacher model to supervise the training of the student model. To supervise the intermediate feature learning, various feature-based KD approaches [2,12,28,37,45,51] have been proposed. FitNet [51] first uses the intermediate representations learned by the teacher as hints to distill a student model. However, the intermediate supervision is unsuitable in our cross-modality distillation case, since the 2D and 3D representations have a large gap. Unlike the above approaches, relation-based KD [34,44] aims to leverage the relationships between various samples or feature maps, e.g., FPS [74] uses the inner products of feature maps for distillation. However, naively applying feature correlation as supervision is not helpful in 2D-to-3D tracking, which is due to the transfer noise and causes performance degradation (see Table 2). In this paper, we propose a novel IoU-guided matching distillation approach for better 2D knowledge transfer, which is also the first attempt to perform cross-modality transfer for 3D SOT.

3 Methodology

In this section, we first present our general 3D SOT framework with a standard ViT as the target relation modeling module, which could naturally benefit from 3D self-supervised pre-training and be consistent with 2D pre-trained trackers for better distillation. To use the 2D pre-trained model (frozen) for guiding 3D matching, we introduce a lightweight target-aware projection (TAP) module to bridge the domain gap between 2D and 3D tracking. Finally, we explore our IoU-guided matching distillation to boost the 3D tracking performance.

3.1 Unified 3D SOT Framework

Revisit of 3D SOT. Given the 3D bbox in the first frame, the goal of 3D SOT is to accurately predict 3D bboxes in the following frames. Existing Siamese matching approaches [26,47,68,77] mainly use the point cloud (PC) data $\hat{\mathbf{p}}_1 \in \mathbb{R}^{\hat{N}\times 3}$ cropped in the first frame as the template to predict the 3D bbox $\mathcal{B}_t \in \mathbb{R}^7$ from the PC $\mathbf{p}_t \in \mathbb{R}^{N_s\times 3}$ in the t-th search frame. $\hat{N}$ is the number of sampled target points within the annotated 3D bounding box, N_s is the number of points in each search frame, and $\mathcal{B}_t$ is parameterized by its center coordinates

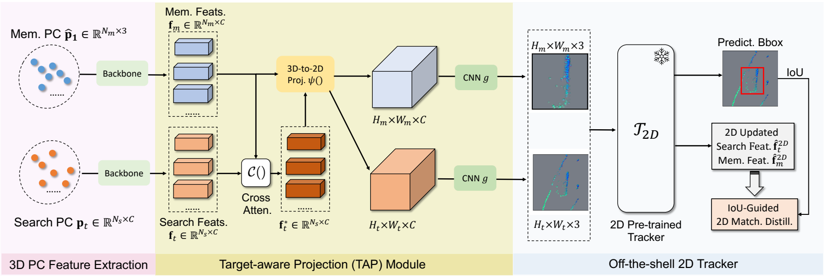

Mem. PC $\hat{\mathbf{p}}_1 \in \mathbb{R}^{N_m \times 3}$
Backbone
Mem. Feats.
$\mathbf{f}_m \in \mathbb{R}^{N_m \times C}$
3D-to-2D
Proj. $\psi()$
$H_m \times W_m \times C$
CNN g
$H_m \times W_m \times 3$
Predict. Bbox
IoU
$\mathcal{T}_{2D}$
Search PC $\mathbf{p}_t \in \mathbb{R}^{N_s \times C}$
Backbone
Search Feats.
$\mathbf{f}_t \in \mathbb{R}^{N_s \times C}$
$\mathcal{C}()$
Cross
Atten.
$\mathbf{f}_t^* \in \mathbb{R}^{N_s \times C}$
$H_t \times W_t \times C$
CNN g
$H_t \times W_t \times 3$
2D Pre-trained
Tracker
2D Updated
Search Feat. $\hat{\mathbf{f}}_t^{2D}$
Mem. Feat. $\hat{\mathbf{f}}_m^{2D}$
IoU-Guided
2D Match. Distill.
3D PC Feature Extraction
Target-aware Projection (TAP) Module
Off-the-shell 2D Tracker

Fig. 3: The pipeline converting 3D point clouds (PC) to 2D images using our target-aware projection (TAP) module, followed by the 2D pre-trained tracker.

(x, y, and z), orientation θ and box size (width, height and length). Recent approaches [69, 70] fully leverage contextual information and use more memory frames $\{\mathbf{p}_i\}_{i=t-k}^{t}$ to predict $\mathcal{B}_t$, where k is the number of memory frames. Instead of explicitly removing template points, a target mask set $\{\mathcal{U}_i\}_{i=t-k-1}^{t-1}$ is used to identify the points within the 3D bbox, where $\mathcal{U}_i \in \mathbb{R}^{N_s \times 1}$.

Target matching module. For simplicity and consistency, we use $\mathbf{m} \in \mathbb{R}^{N_m \times 3}$ to represent the unified memory used in the above approaches, where $N_m = \hat{N}$ in memory-less Siamese approaches and $N_m = N_s \times k$ in memory approaches. We then obtain the search point features $\mathbf{f}_t = \phi(\mathbf{p}_t) \in \mathbb{R}^{N_s \times C}$ and memory point features $\mathbf{f}_m = \phi(\mathbf{m}) \in \mathbb{R}^{N_m \times C}$ by applying a point feature extractor $\phi(\cdot)$ on $\mathbf{p}_t$ and $\mathbf{m}$ respectively. The current 3D SOT trackers mainly focus on designing a robust 3D matching module $\mathcal{M}(\mathbf{f}_s, \mathbf{f}_m)$ to perform the target matching between the search and memory point features. However,learning this matching module is challenging due to the sparsity of LIDAR point clouds and the lack of sufficient category-specific training data.

To tackle these issues, inspired by 2D SOT [62, 73], we propose to use a standard ViT [11] for 3D matching since: 1) the standard ViT can benefit from 3D self-supervised pre-training (e.g., [42,49]), which eases the need for large-scale 3D annotated data for 3D SOT training; 2) it has the consistent target relation modelling architecture (i.e,. ViT) with dominant 2D trackers, which could more naturally transfer its 2D matching knowledge. The 3D matching process with ViT is formulated as:

$$\hat{\mathbf{f}}_t = \mathcal{M}(\mathbf{f}_t, \mathbf{f}_m + \theta(\mathcal{U})), \tag{1}$$

where the output $\hat{\mathbf{f}}_t \in \mathbb{R}^{N_s \times C}$ is the updated search features output from the last layer of $\mathcal{M}$, $\theta(\cdot)$ is a learnable linear mapping that converts the 1-dim masks into the C-dim feature maps. For siamese tracking paradigms, we drop this operation since the template is well cropped. Finally, the matching output $\hat{\mathbf{f}}_t$ is fed to the prediction head [26, 27, 70] for predicting the 3D bbox $\mathcal{B}_t$. Briefly, we call the memory-less Siamese-based and memory-based approaches with our standard ViT matching module as SiamBase and MemBase (see Supp. for more details).

3.2 Target-Aware Projection Module

Although SiamBase and MemBase can naturally benefit from 3D self-supervised pre-training [42, 49], their performance is still limited due to the absence of large-scale 3D datasets (e.g., ShapeNet [6] only contains 55 object categories with 50k point clouds) for obtaining effective pre-training weights. To further improve the 3D tracking performance, we propose to employ a powerful 2D pre-trained tracker (denoted as $\mathcal{T}_{2D}$) as the distillation teacher to guide the 3D matching learning in SiamBase and MemBase. However, in 3D SOT, only the single modality of point cloud data is available, making it hard to directly utilize current 2D SOT trackers. To bridge the modality gap between 2D and 3D tracking, we thus propose a learnable target-aware projection (TAP) module in this subsection, illustrated in Fig. 3.

3D-to-2D Projection. Using the domain knowledge that the objects-of-interest for 3D tracking are typically on the ground plane (at similar fixed z-level heights), we project the search point cloud $\mathbf{p}_t$ from the orthogonal view along the z axis. For each point, we round down the (x, y) coordinates to obtain its location on the 2D xy plane, i.e., from the bird-eye view (BEV). We denote this 3D-to-2D projection as $\psi(\mathbf{p}_t)$.

Target-Aware Projection. Based on [58], the point features $\mathbf{f}_t$ can be mapped to the 2D plane based on the corresponding projection $\psi(\mathbf{p}_t)$. However, simply projecting search features to the 2D plane is not optimal since it does not consider any target information, which may not well support the subsequent 2D matching (see Table 2). Therefore, we use a cross-attention module $\mathcal{C}$ here to obtain more effective target-aware search features:

$$\mathbf{f}_t^* = \mathcal{C}(Q(\mathbf{f}_t), K(\mathbf{f}_m + \theta(\mathcal{U}) + E(\mathbf{m})), V(\mathbf{f}_m + \theta(\mathcal{U}) + E(\mathbf{m}))), \tag{2}$$

where Q, K and V are query, key and value embedding modules, respectively. E is the 3D positional embedding module. We project the target-aware search features $\mathbf{f}_t^* \in \mathbb{R}^{N_s \times C}$ into the 2D plane via $\psi(\mathbf{p}_t)$ and obtain the projected feature maps with C channels, which are further input to a lightweight CNN g [58] to reduce its channel dimension, yielding the projected search image:

$$I_t = g(\psi(\mathbf{p}_t)) \in \mathbb{R}^{H_t \times W_t \times 3}, \tag{3}$$

where H_t and W_t are height and width of the projected search image. The same operations are made on each memory frame and its corresponding features, thus we obtain k projected memory images $\{I_m^i\}_{i=t-k-1}^{t-1}$, where $I_m^i \in \mathbb{R}^{H_m \times W_m \times 3}$.

TAP training. The TAP module is trained via a simulated 2D tracking process on the projected search and target images, so that the projected images can obtain good 2D tracking performance. As illustrated in Fig. 3, the projected 2D search and memory images are input to a ViT-based pre-trained 2D tracker (e.g., OSTrack [73]) for 2D bounding box regression. Specifically, the input images are flattened into a sequence of 2D patches, projected into image embeddings, and then concatenated along the spatial dimension to input to the 2D ViT model for joint feature extraction and matching. The updated search features output

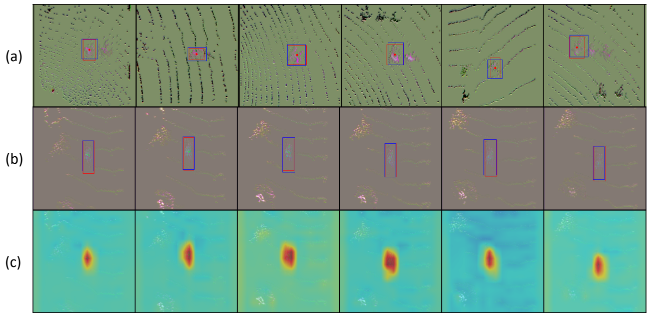


Fig. 4: Visualization of the projected search image I_t obtained by our TAP module on (a) KITTI-Pedestrian and (b) KITTI-Cyclist. (c): The attention maps are generated via template-to-search matching from the 2D tracker on (b) KITTI-Cyclist. The red box indicates the ground-truth (GT) 2D box projected from the GT 3D box. The blue box denotes the prediction of the 2D pre-trained OSTrack [73]. Based on the target-aware search features, the TAP module can effectively encode the target region with unique color and texture, enabling subsequent 2D matching.

from the last layer is input to the prediction head for 2D box regression, which can be generally formulated as:

$$\mathcal{B}_t^{2D}, \hat{\mathbf{f}}_t^{2D}, \hat{\mathbf{f}}_m^{2D} = \mathcal{T}_{2D}(\{I_m^i\}_{i=t-k-1}^{t-1}, I_t), \tag{4}$$

where $\mathcal{B}_t^{2D} \in \mathbb{R}^4$ is the predicted 2D bounding box in the projected search frame, and $(\hat{\mathbf{f}}_t^{2D}, \hat{\mathbf{f}}_m^{2D})$ are the updated 2D search and memory feature maps from the last layer of the 2D tracker. Here we use OSTrack as the 2D pre-trained tracker, and use its same 2D loss $\mathcal{L}_{2D}$ [62, 73] to evaluate the simulated 2D tracking result, thus providing supervision on the TAP module. The ground-truth 2D bounding box $\mathcal{G}_t^{2D}$ is obtained by projecting the corners of the ground-truth 3D box to the projected search image via the 3D-to-2D projection ψ.

After training the TAP module, the projection process is aligned with the 2D tracker. Note that we only update the parameters of the TAP module while keeping the parameters in the 2D pre-trained tracker $\mathcal{T}_{2D}$ fixed, which makes the TAP training more efficient, and less likely to overfit. We also visualize the projected 2D search images in Fig. 4. The learned TAP module can support accurate 2D matching in the projected search image space well via the pre-trained OSTrack model, which shows the effectiveness of the proposed TAP module.

3.3 IoU-guided Matching Distillation

We introduce how to distill the 2D matching knowledge in $\mathcal{T}_{2D}$ to guide the 3D matching learning in our 3D SOT trackers, SiamBase and MemBase. In (1), we obtain the updated 3D search features $\hat{\mathbf{f}}_t \in \mathbb{R}^{N_s \times C}$ and updated 3D memory

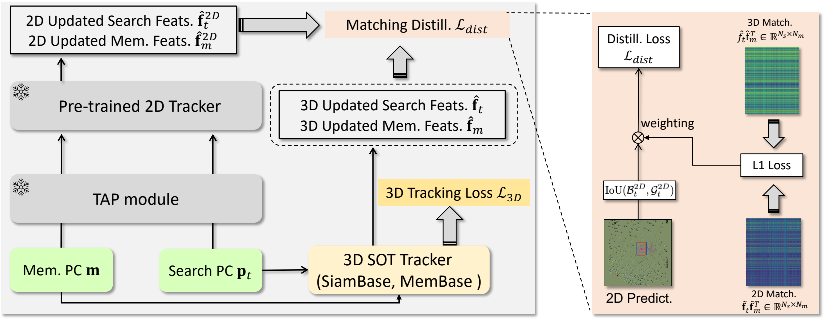

2D Updated Search Feats. $\hat{\mathbf{f}}_t^{2D}$
2D Updated Mem. Feats. $\hat{\mathbf{f}}_m^{2D}$
Matching Distill. $\mathcal{L}_{dist}$
Pre-trained 2D Tracker
3D Updated Search Feats. $\hat{\mathbf{f}}_t$
3D Updated Mem. Feats. $\hat{\mathbf{f}}_m$
TAP module
3D Tracking Loss $\mathcal{L}_{3D}$
Mem. PC $\mathbf{m}$
Search PC $\mathbf{p}_t$
3D SOT Tracker
(SiamBase, MemBase)
Distill. Loss $\mathcal{L}_{dist}$
3D Match.
$\hat{f}_t\hat{\mathbf{f}}_m^T \in \mathbb{R}^{N_s \times N_m}$
weighting
L1 Loss
$\text{IoU}(\mathcal{B}_t^{2D}, \mathcal{G}_t^{2D})$
2D Predict.
2D Match.
$\tilde{\mathbf{f}}_t\tilde{\mathbf{f}}_m^T \in \mathbb{R}^{N_s \times N_m}$

Fig. 5: Training pipeline of our SiamDisst or MemDisst via IoU-guided matching distillation. During the training stage, only the SiamDisst or MemDisst model is updated.

features $\hat{\mathbf{f}}_m \in \mathbb{R}^{N_m \times C}$ output from the last layer of $\mathcal{M}$. Similarly, the updated 2D search and memory feature maps, $\hat{\mathbf{f}}_t^{2D} \in \mathbb{R}^{\frac{H_t}{s} \times \frac{W_t}{s} \times C}$ and $\hat{\mathbf{f}}_m^{2D} \in \mathbb{R}^{\frac{H_m}{s} \times \frac{W_m}{s} \times k \times C}$, are obtained via (4) in the last layer of 2D ViT, where s denotes the patch size for patch embedding. The 3D and 2D features belong to different embedding spaces, and thus directly guiding the feature embeddings of the 3D tracker using the 2D tracker is not straightforward (see Supplementary). Instead, we use the matching correspondences (between feature maps) used by the 2D tracker to guide the matching correspondences (between feature vectors) of the 3D tracker.

3D-to-2D feature indexing. Given a search point coordinate in the search point cloud $\mathbf{p}_t$, we index its corresponding 2D feature vector via the 3D-to-2D projection ψ. Specifically, we first obtain its 2D location in the projected search image. Then, we interpolate its 2D features $\hat{\mathbf{f}}_t^{2D}$ to its original spatial size $H_t \times W_t$ via bilinear interpolation, and finally index its feature vector via the projected 2D location. Based on the above 3D-to-2D feature indexing, we obtain the 2D search and memory features, denoted as $\tilde{\mathbf{f}}_t \in \mathbb{R}^{N_s \times C}$ and $\tilde{\mathbf{f}}_m \in \mathbb{R}^{N_m \times C}$, that correspond to the 3D search and memory point clouds.

Matching distillation. Note that $\tilde{\mathbf{f}}_t$ and $\tilde{\mathbf{f}}_m$ are obtained from the 2D ViT in the 2D tracker, which is well learned on large-scale 2D annotated videos, thus providing more robust matching results than our 3D counterparts (SimBase and MemBase) trained on limited category-specific point cloud data. Thus, we use the 2D memory-to-search matching pattern to guided the 3D matching process:

$$\mathcal{L}_{dist} = \text{IoU}(\mathcal{B}_t^{2D}, \mathcal{G}_t^{2D}) \cdot \mathcal{L}_1(\hat{\mathbf{f}}_t\hat{\mathbf{f}}_m^T, \tilde{\mathbf{f}}_t\tilde{\mathbf{f}}_m^T), \tag{5}$$

where $\hat{\mathbf{f}}_t\hat{\mathbf{f}}_m^T$ is the inner-product (matching) matrix between the 3D search and memory features, and likewise for $\tilde{\mathbf{f}}_t\tilde{\mathbf{f}}_m^T$, and $\mathcal{L}_1$ is the L1 loss that calculates the mean absolute error between two matrices. $\text{IoU}(\cdot,\cdot)$ calculates the intersection-over-union between the predicted 2D box $\mathcal{B}_t^{2D}$ and projected 2D ground-truth box $\mathcal{G}_t^{2D}$, which measures the reliability of our 2D tracker, i.e. if the 2D matching is reliable enough and predicts the 2D box accurately (i.e., with high IoU), this 2D matching pattern will be used more for guiding the 3D matching. We also use the typical 3D box regression losses $\mathcal{L}_{3D}$ in the siamese paradigm (i.e., STNet [27]) and memory-based paradigm (i.e., MBPTrack [70]) to supervise our SiamBase and MemBase, respectively. The overall training loss is:

$$\mathcal{L}_{overall} = \mathcal{L}_{3D} + \lambda \mathcal{L}_{dist}, \tag{6}$$

where λ is a hyperparameter. After matching distillation, we denote the updated 3D SOT trackers as SiamDisst and MemDisst. Note that the TAP module, the 2D pre-trained tracker and the matching distillation are only used during the offline training stage to improve the 3D matching ability. During online inference, these modules are removed, and only our 3D trackers SiamDisst and MemDisst are required, which run at the real-time speed of 25 and 90 FPS, respectively.

4 Experiments

Implementation details. For training the TAP module, as illustrated in Fig. 3, we sample memory and search PCs as the input, and project them into the 2D images for further matching in the 2D tracker. We use OSTrack [73] trained on large-scale 2D video datasets [13, 25, 40], as the 2D tracker, and freeze its weights, thus leading to efficient TAP training. The original 2D loss in OSTrack is used to evaluate the 2D predictions without modification. For TAP training, we use a batch size of 8 with a learning rate of 1e-4 and a total epoch of 40. For SiamDisst, we use the same hierarchical backbone network and 3D prediction head in STNet [27] for initial PC feature extraction and 3D prediction. For MemDisst, we use the same lightweight PointNet [46] with Point-MAE [42] as the backbone, and use the prediction head and memory mechanism in MBPTrack [70] for a fair comparison. We employ ViT-384-6 and ViT-96-3 as matching modules for SiamDisst and MemDisst respectively, which is illustrated in Table 3. The hyper-parameter λ is set to 1.0. The training and testing settings of SiamDisst and MemDisst are following their baselines STNet and MBPTrack. More details on implementation can be found in the Supplementary.

Datasets and Evaluation. We use three 3D datasets, KITTI [18], nuScenes [5] and Waymo Open Dataset (WOD) [53], for training and testing our proposed approaches. The KITTI dataset contains 21 training video sequences and 29 test sequences. Following previous work [19], the training set is split into three parts: 0-16 for training, 17-18 for validation and 19-20 for testing. Following [26,27,69], we use nuScenes and WOD to test the generalization ablity of our approaches. We follow the one one pass evaluation [31], which is widely used in previous 3D SOT works [19,26,27,70], to report *Precision* and *Success* metrics for evaluation.

4.1 Ablation Study

Component analysis. We evaluate the effect of the two main components used in our proposed framework: 3D pre-training and 2D matching distillation.

Variant	Baseline	3D Pre-training	2D Distillation	Success	Precision
SiamBase	✓			55.6	64.3
SiamBase		✓		60.3	69.3
SiamBase		✓	✓	**63.5**	**75.0**
MemBase	✓			61.3	72.7
MemBase		✓		63.5	76.5
MemBase		✓	✓	**66.6**	**79.3**

Table 1: Component analysis of SiamDisst and MemDisst. 'Baseline' indicates the variant initialized with the random weights and trained from scratch.

Variant	Succ./Prec.
SiamDisst	63.5/75.0
unlearnable TAP	61.1/72.3
TAP w/o target-aware interact.	61.2/73.1
2D distill. w/o IoU weight.	56.4/66.3

Pre-training	Succ./Prec.
Point-MAE [42]	63.5/75.0
ReCon [49]	62.9/73.6
2D MAE [22]	60.5/69.9
DropMAE [62]	61.9/74.0

Table 2: Ablation studies of SiamDisst on KITTI-VAN: (left) variants of the TAP module and distillation; (right) using different 3D pre-trained models.

The 3D pre-training means the usage of the default pre-trained Point-MAE [42] weights for 3D initialization. As shown in Table 1, SiamDisst and MemDisst with 3D pre-training achieve 5% and 3.8% improvements over the naive baselines regarding the precision metric. By applying the 2D matching distillation, SiamDisst and MemDissst can be significantly improved, leading to 5.7% and 2.8% precision gains, which shows the effectiveness of our approach.

TAP module. We study several variants of the proposed TAP module: including 1) unlearnable TAP: we remove the learnable projection module in the TAP, and use the z-axis value as the projected feature in the 2D image plane (repeated for each of the 3 image channels); 2) TAP module w/o target-aware design: we remove the cross attention module $\mathcal{C}$ and separately project the template and search images; Table 2 (left) presents the results. The unlearnable TAP obtains inferior performance, which is due to the low-quality projected images and more 2D tracking failures. Without the IoU weighting, the unlearnable TAP also has severely degraded performance (see Supplementary for more details). For the target-aware design, we can observe 2.1%/1.9% gains by using it, which shows that learning target-aware features is beneficial in the 2D projection stage.

IoU-guided matching distillation. In the IoU-guided matching distillation, we use the IoU to measure the reliability of the 2D prediction. To test its effectiveness, we remove this IoU weighting in (5) and show the comparison in Table 2 (left). The variant w/o the IoU weighting degrades the performance with large margins of 7.1%/8.7%, which is mainly because the 2D matching distillation may introduce some low-quality and inaccurate 2D matching patterns for the 3D matching learning, thus severely degrading the learning.

Comparison of 3D pre-trained weights. We use various pre-trained weights to initialize the ViT in our 3D tracker and test them in Table 2 (right). The

method	matching module	succ./prec.	FPS
SiamDisst	ViT-384-6	**63.5**/**75.0**	25
SiamDisst	ViT-96-3	60.5/71.7	30
MemDisst	ViT-384-6	**66.7**/79.1	78
MemDisst	ViT-96-3	66.6/**79.3**	90

Table 3: Comparisons of our approaches with various matching modules.

generative 3D Point-MAE pre-training provides the best result. Note that ReCon [49] adds the contrastive learning in the multi-modality generative pre-training, which may not be suitable in the pure point cloud tracking setting. We also test two typical 2D pre-trained models, i.e., MAE [22] and DropMAE [62] (see Supp. for details). Interestingly, our SiamDisst with DropMAE initialization achieves competitive results, which indicates that the temporal matching ability of DropMAE learned in 2D videos is also helpful for the 3D tracking task.

Lightweight or heavy matching module. We explore various matching modules used in SiamDisst and MemDisst in Table 3. Note that ViT-384-6 is built from the first 6 layers of ViT-Base [11] with 384 hidden dimensions and 6 heads. We also design a lightweight ViT-96-3 model with 4 heads for comparison. We initialize ViT-384-6 with the official Point-MAE weights and rerun Point-MAE pre-training to initialize ViT-96-3[5]. In Table 3, SiamDisst benefits more from the heavy ViT-384-6 while MemDisst achieves favorable performance even using the lightweight ViT-96-3. This is because SiamDisst has no effective online memory mechanism and heavily relies on its matching module. For MemDisst, the online memory provides more historical cues for matching, which eases the need for a heavy online matching module. Considering the speed-accuracy trade-off, we use ViT-384-6 and ViT-96-3 for SiamDisst and MemDisst, respectively.

4.2 State-of-the-art Comparison

We compare the proposed SiamDisst and MemDisst with state-of-the-art 3D trackers on the KITTI, nuScenes and WOD datasets. Specifically, for KITTI, comprehensive comparisons are made with two groups of trackers: 1) memory-less Siamese-based 3D trackers, SC3D [19], 3DSiamRPN [15], P2B [48], LTTR [8], MLVSNet [57], BAT [77], PTT [52], V2B [26], CMT [21], PTTR [79], and STNet [27]; and 2) contextual memory or motion prediction based 3D trackers, TAT [33], DMT [68], M2Track [78], CXTrack [69], and MBPTrack [70].

As shown in Table 4, our memory-less tracker SiamDissst outperforms the other Siamese 3D trackers by large margins in terms of both success and precision metrics. Specifically, the overall performance (Mean column) achieved by SiamDisst is 66.2%/83.9%, which has significant improvements (4.9%/2.8%) over the baseline tracker STNet and demonstrates the effectiveness of our proposed

[5] We follow the same pre-training hyper-parameters, steps and dataset (ShapeNet [6]) with Point-MAE [42] for pre-training.

Method	CM	MP	FPS	Car	Pedestrian	Van	Cyclist	Mean
				(6424)	(6088)	(1248)	(308)	(14068)
SC3D [19]			2	41.3/57.9	18.2/37.8	40.4/47.0	41.5/70.4	31.2/48.5
3DSiamRPN [15]			21	58.2/76.2	35.2/56.2	45.7/52.9	36.2/49.0	46.7/64.9
P2B [48]			46	56.2/72.8	28.7/49.6	40.8/48.4	32.1/44.7	42.4/60.0
LTTR [8]			23	65.0/77.1	33.2/56.8	35.8/45.6	66.2/89.9	48.7/65.8
MLVSNet [57]			**70**	56.0/74.0	34.1/61.1	52.0/61.4	34.3/44.5	45.7/66.7
BAT [77]			65	60.5/77.7	42.1/70.1	52.4/67.0	33.7/45.4	51.2/72.8
PTT [52]			40	67.8/81.8	44.9/72.0	43.6/52.5	37.2/47.3	55.1/74.2
V2B [26]			37	70.5/81.3	48.3/73.5	50.1/58.0	40.8/49.7	58.4/75.2
CMT [21]			32	70.5/81.9	49.1/75.5	54.1/64.1	55.1/82.4	59.4/77.6
PTTR [79]			53	65.2/77.4	50.9/81.6	52.5/61.8	65.1/90.5	57.9/78.1
STNet [27]			35	72.1/84.0	49.9/77.2	58.0/70.6	73.5/93.7	61.3/80.1
SiamDisst			25	**73.7/85.1**	**58.4/83.9**	**63.5/75.0**	**76.9/94.5**	**66.2/83.9**
TAT [33]	✓		-	72.2/83.3	57.4/84.4	58.9/69.2	74.2/93.9	64.7/82.8
DMT [68]		✓	72	66.4/79.4	48.1/77.9	53.3/65.6	70.4/92.6	55.1/75.8
M2Track [78]		✓	57	65.5/80.8	61.5/88.2	53.8/70.7	73.2/93.5	62.9/83.4
CXTrack [69]	✓		34	69.1/81.6	67.0/91.5	60.0/71.8	74.2/94.3	67.5/85.3
MBPTrack [70]	✓		56	73.4/84.8	68.6/93.9	61.3/72.7	76.7/94.3	70.3/87.9
MemDisst	✓		**90**	**74.1/85.6**	**69.1/94.1**	**66.6/79.3**	**77.2/94.7**	**71.3/88.9**

Table 4: 3D tracking results on KITTI [18] for (top) memory-less 3D trackers; (bottom) memory/motion-based 3D trackers. CM and MP represent the contextual memory and motion prediction. Trackers are evaluated by per-category Success/Precision metrics. 'Mean' indicates the overall result averaged over frames. The best result is shown in bold. For each category, the number in brackets indicates the test frame number.

IoU-guided matching distillation. Moreover, the proposed SiamDisst is more effective in identifying distractors during online tracking, e.g., obtaining best overall performance on the Pedestrian category. For MemDisst, it sets new state-of-the-art performance on all the categories while running at a fast speed of 90 FPS. The lightweight model used in MemDisst is well learned via the proposed learning framework, thus achieving promising tracking performance. Note that both SiamDisst and MemDisst performs well in the low data regime (e.g., the Van category only has 1994 training examples), which is mainly due to the effectiveness of transferring the matching knowledge from the 2D pre-trained tracker.

To test the generalization performance of our proposed trackers, we use our SiamDisst and MemDisst trained on KITTI to evaluate on the WOD and nuScenes datasets. As seen in Table 5, SiamDisst performs best on both WOD and nuScenes among all the Siamese-based 3D trackers, e.g., STNet, V2B and BAT. This is mainly because our SiamDisst learns a more effective target matching module via the 2D matching distillation, thus leading to more accurate tracking results. Compared with MBPTrack, our MemDisst obtains comparable performance on WOD and outperforms it on nuScenes. Meanwhile, MemDisst runs efficiently at the above real-time speed of 89 FPS, which is 53% and 162% faster than MBPTrack and CXTrack, respectively. This lightweight MemDisst tracker

	Method	WOD [53]			nuScenes [5]					FPS
	Category	Vehicle	Pedestrian	Mean	Car	Pedestrian	Truck	Bicycle	Mean	-
	(No. Frames)	(185731)	(241752)	(427483)	(15578)	(8019)	(3710)	(501)	(27808)	-
Success	P2B	52.6	17.9	33.0	27.0	15.9	21.5	20.0	22.9	46
	SC3D	-	-	-	25.0	14.2	25.7	17.0	21.8	2
	BAT	54.7	18.2	34.1	22.5	17.3	19.3	17.0	20.5	**65***
	V2B	57.6	23.7	38.4	31.3	17.3	21.7	22.2	25.8	37
	STNet	59.7	25.5	40.4	32.2	19.1	22.3	21.2	26.9	35
	SiamDisst	**61.2**	**28.5**	**42.7**	**33.4**	**19.8**	**25.0**	17.7	**28.1**	25
	CXTrack	57.1	30.7	42.2	29.6	**20.4**	27.6	18.5	26.5	34
	M2Track	61.1	32.0	44.6	-	-	-	-	-	57
	MBPTrack	**61.9**	**33.7**	**46.0**	33.6	19.8	23.9	**20.0**	28.1	60*
	MemDisst	**61.9**	33.6	45.9	**34.0**	20.0	**28.1**	18.1	**28.9**	**89***
	Category	Vehicle	Pedestrian	Mean	Car	Pedestrian	Truck	Bicycle	Mean	FPS
Precision	P2B	61.7	30.1	43.8	29.2	22.0	16.2	26.4	25.3	46
	SC3D	-	-	-	27.1	16.2	21.9	18.2	23.1	2
	BAT	62.7	30.3	44.4	24.1	24.5	15.8	18.8	23.0	**65***
	V2B	65.9	37.9	50.1	35.1	23.4	16.7	19.1	29.0	37
	STNet	68.0	39.9	52.1	36.1	27.2	16.8	**29.2**	30.8	35
	SiamDisst	**70.5**	**44.5**	**55.8**	**37.3**	**30.4**	**20.1**	23.7	**32.8**	25
	CXTrack	66.1	49.4	56.7	33.4	**32.9**	20.8	26.8	31.5	34
	M2Track	69.3	49.7	58.2	-	-	-	-	-	57
	MBPTrack	**71.9**	52.7	61.0	37.6	32.7	20.7	**29.2**	33.8	60*
	MemDisst	71.8	**52.8**	**61.1**	**38.0**	32.2	**23.9**	24.1	**34.2**	**89***

Table 5: 3D tracking results on Waymo Open Dataset (WOD) [53] and nuScenes [5]. For a fair comparison, we compare our SiamDisst and MemDisst with memory-less Siamese-based trackers, and more complicated trackers with contexture memory or motion prediction, respectively. The best result in each comparison is shown in bold. * indicates the speed is re-evaluated on WOD for a fair comparison.

shows its potential to serve as a simple and strong baseline in the 3D tracking community.

5 Conclusion

This paper improves the power of 3D point cloud tracking from two aspects. First, we design a novel 3D SOT framework that could fully benefit from the 3D self-supervised pre-training, which shows that the self-supervised pre-training weights (e.g., Point-MAE [42] and Recon [49]) are effective to improve the 3D tracking performance, especially in the low data regime (e.g., KITTI-Van and KITTI-Cyclist). Second, we bridge the gap between 2D and 3D tracking domains by leveraging a target-aware projection module, which enables to use a powerful 2D pre-trained tracker to facilitate the learning of 3D matching in existing 3D SOT trackers [27, 70] via matching distillation. We show that our proposed SiamDisst and MemDisst trackers can achieve state-of-the-art performance on KITTI and Waymo Open datasets, while running at real-time speed.

Acknowledgment

This work was supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005665).

References

1. Anthes, C., García-Hernández, R.J., Wiedemann, M., Kranzlmüller, D.: State of the art of virtual reality technology. In: 2016 IEEE aerospace conference. pp. 1–19. IEEE (2016) 1
2. Ben-Baruch, E., Karklinsky, M., Biton, Y., Ben-Cohen, A., Lawen, H., Zamir, N.: It's all in the head: Representation knowledge distillation through classifier sharing. In: arXiv:2201.06945 (2022) 5
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision. pp. 850–865. Springer (2016) 4
4. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Leanring discriminative model prediction for tracking. In: ICCV. pp. 6182–6191 (2019) 5
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020) 2, 4, 10, 14
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 7, 12
7. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: CVPR. pp. 8126–8135 (2021) 5
8. Cui, Y., Fang, Z., Shan, J., Gu, Z., Zhou, S.: 3d object tracking with transformer. arXiv preprint arXiv:2110.14921 (2021) 4, 12, 13
9. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: CVPR (2022) 5
10. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: CVPR. pp. 21–26 (2017) 4
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929 (2010) 6, 12
12. Duong, C.N., Luu, K., Quach, K.G., Shrinkteanet, N.L.: Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. In: arXiv:1905.10620 (2019) 5
13. Fan, H., Lin, L., Yang, F.: Lasot: A high-quality benchmark for large-scale single object tracking. In: CVPR. pp. 5374–5383 (2019) 5, 10
14. Fan, H., Ling, H.: Siamese cascaded region proposal networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7952–7961 (2019) 5
15. Fang, Z., Zhou, S., Cui, Y., Scherer, S.: 3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud. IEEE Sensors Journal **21**(4), 4995–5011 (2020) 12, 13

16. Feng, S., Liang, P., Gao, J., Cheng, E.: Multi-correlation siamese transformer network with dense connection for 3d single object tracking. IEEE Robotics and Automation Letters (2023) 4
17. Galoogahi, H., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: ICCV (2017) 4
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 2, 4, 10, 13
19. Giancola, S., Zarzar, J., Ghanem, B.: Leveraging shape completion for 3d siamese tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1359–1368 (2019) 4, 10, 12, 13
20. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015) 4
21. Guo, Z., Mao, Y., Zhou, W., Wang, M., Li, H.: Cmt: Context-matching-guided transformer for 3d tracking in point clouds. In: European Conference on Computer Vision. pp. 95–111. Springer (2022) 4, 12, 13
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 2, 11, 12
23. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 583–596 (2015) 4
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: arXiv:1503.02531 (2015) 5
25. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 5, 10
26. Hui, L., Wang, L., Cheng, M., Xie, J., Yang, J.: 3d siamese voxel-to-bev tracker for sparse point clouds. Advances in Neural Information Processing Systems **34**, 28714–28727 (2021) 3, 4, 5, 6, 10, 12, 13
27. Hui, L., Wang, L., Tang, L., Lan, K., Xie, J., Yang, J.: 3d siamese transformer network for single object tracking on point clouds. In: European Conference on Computer Vision. pp. 293–310. Springer (2022) 2, 3, 4, 6, 10, 12, 13, 14
28. Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J.: Knowledge distillation via route constrained optimization. In: IEEE/CVF International Conference on Computer Vision (2019) 5
29. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 2
30. Kiran, B.R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A.A., Yogamani, S., Pérez, P.: Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems (2021) 1
31. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Cehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016) 10
32. Kristan, M., Matas, J., Danelljan, M.: The first visual object tracking segmentation vots2023 challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 5

33. Lan, K., Jiang, H., Xie, J.: Temporal-aware siamese tracker: Integrate temporal context for 3d object tracking. In: Proceedings of the Asian Conference on Computer Vision. pp. 399–414 (2022) 2, 4, 12, 13
34. Lee, S.H., Kim, D.H., Song, B.C.: Self-supervised knowledge distillation using singular value decomposition. In: In Proceedings of the European Conference on Computer Vision (ECCV) (2018) 5
35. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4282–4291 (2019) 5
36. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8971–8980 (2018) 4, 5
37. Li, J., Guo, Z., Li, H., Han, S., Baek, J.w., Yang, M., Yang, R., Suh, S.: Rethinking feature-based knowledge distillation for face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 5
38. Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking pp. 16743–16754 (2022) 2, 5
39. Liu, Y., Liang, Y., Wu, Q., Zhang, L., Wang, H.: A new framework for multiple deep correlation filters based object tracking. In: ICASSP (2022) 4
40. Muller, M., Bibi, A., S, G.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV. pp. 300–317 (2018) 10
41. Osep, A., Mehner, W., Mathias, M., Leibe, B.: Combined image-and world-space tracking in traffic scenes. In: IEEE International Conference on Robotics and Automation. pp. 1988–1995. IEEE (2017) 1
42. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: European conference on computer vision. pp. 604–621 (2022) 3, 6, 7, 10, 11, 12, 14
43. Pang, Z., Li, Z., Wang, N.: Model-free vehicle tracking and state estimation in point cloud sequences. In: IROS (2021) 4
44. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Self-supervised knowledge distillation using singular value decomposition. In: IEEE/CVF International Conference on Computer Vision (2019) 5
45. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. In: IEEE/CVF International Conference on Computer Vision (2019) 5
46. Qi, C., Su, H., Mo, K., Guibas, L.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE conference on computer vision and pattern recognition (2017) 10
47. Qi, H., Feng, C., Cao, Z., Zhao, F., Xiao, Y.: 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. pp. 6329–6338 (2020) 2, 3, 5
48. Qi, H., Feng, C., Cao, Z., Zhao, F., Xiao, Y.: 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. pp. 6329–6338 (2020) 4, 12, 13
49. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining (2023) 3, 6, 7, 11, 12, 14
50. Ran, T., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1420–1429 (2016) 4

51. Romero, A., Ballas, N.: Fitnets: Hints for thin deep nets. In: arXiv:1412.6550 (2014) 5
52. Shan, J., Zhou, S., Fang, Z., Cui, Y.: Ptt: Point-track-transformer module for 3d single object tracking in point clouds. arXiv preprint arXiv:2108.06455 (2021) 12, 13
53. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 2446–2454 (2020) 4, 10, 14
54. Tao, F., Wang, M.: Response-based distillation for incremental object detection. In: arXiv:2110.13471 (2021) 5
55. Tao, F., Wang, M., Yuan, H.: Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 5
56. Wang, N., Song, Y., Ma, C.: Unsupervised deep tracking. In: CVPR. pp. 3708–1317 (2019) 5
57. Wang, Z., Xie, Q., Lai, Y.K., Wu, J., Long, K., Wang, J.: Mlvsnet: Multi-level voting siamese network for 3d visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3101–3110 (2021) 12, 13
58. Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J.: P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. Advances in neural information processing systems **35**, 14388–14402 (2022) 7
59. Wu, Q., Chan, A.: Meta-graph adaptation for visual object tracking. In: ICME (2021) 4
60. Wu, Q., Yan, Y., Liang, Y., Liu, Y., Wang, H.: Dsnet: Deep and shallow feature learning for efficient visual tracking. In: ACCV. pp. 119–134 (2018) 4
61. Wu, Q., Wan, J., Chan, A.B.: Progressive unsupervised learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2993–3002 (2021) 5
62. Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B.: Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14561–14571 (2023) 1, 2, 4, 5, 6, 8, 11, 12
63. Wu, Q., Yang, T., Wu, W., Chan, A.B.: Scalable video object segmentation with simplified framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 5
64. Wu, Q., Sun, C., Wang, J.: Multi-level structure-enhanced network for 3d single object tracking in sparse point clouds. IEEE Robotics and Automation Letters **8**(1), 9–16 (2022) 4
65. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR. pp. 2411–2418 (2013) 5
66. Xia, Y., Gladkova, M., Wang, R., Li, Q., Stilla, U., Henriques, J.F., Cremers, D.: Casspr: Cross attention single scan place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8461–8472 (2023) 1
67. Xia, Y., Shi, L., Ding, Z., Henriques, J.F., Cremers, D.: Text2loc: 3d point cloud localization from natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14958–14967 (2024) 1
68. Xia, Y., Wu, Q., Li, W., Chan, A.B., Stilla, U.: A lightweight and detector-free 3d single object tracker on point clouds. IEEE Transactions on Intelligent Transportation Systems (2023) 4, 5, 12, 13

69. Xu, T.X., Guo, Y.C., Lai, Y.K., Zhang, S.H.: Cxtrack: Improving 3d point cloud tracking with contextual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1084–1093 (2023) 4, 6, 10, 12, 13
70. Xu, T.X., Guo, Y.C., Lai, Y.K., Zhang, S.H.: Mbptrack: Improving 3d point cloud tracking with memory networks and box priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9911–9920 (October 2023) 2, 3, 4, 6, 10, 12, 13, 14
71. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: ICCV. pp. 10448–10457 (2021) 5
72. Yang, T., Chan, A.B.: Learning dynamic memory networks for object tracking. In: Proceedings of the European Conference on Computer Vision. pp. 152–167 (2018) 5
73. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European Conference on Computer Vision. pp. 341–357 (2022) 2, 3, 4, 5, 6, 7, 8, 10
74. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: IEEE conference on computer vision and pattern recognition (2017) 5
75. Zhang, L., Gonzalez-Garcia, A., Weijer, J.v.d., Danelljan, M., Khan, F.S.: Learning the model update for siamese trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4010–4019 (2019) 5
76. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4591–4600 (2019) 5
77. Zheng, C., Yan, X., Gao, J., Zhao, W., Zhang, W., Li, Z., Cui, S.: Box-aware feature enhancement for single object tracking on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13199–13208 (2021) 2, 3, 4, 5, 12, 13
78. Zheng, C., Yan, X., Zhang, H., Wang, B., Cheng, S., Cui, S., Li, Z.: 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. pp. 8111–8120 (2022) 4, 12, 13
79. Zhou, C., Luo, Z., Luo, Y., Liu, T., Pan, L., Cai, Z., Zhao, H., Lu, S.: Pttr: Relational 3d point cloud object tracking with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8531–8540 (2022) 4, 12, 13
80. Zhou, H., Song, L., Chen, J., Zhou, Y., Guoli: Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In: arXiv:2102.00650 (2021) 5