Modeling Continuous Motion for 3D Point Cloud Object Tracking

Zhipeng Luo[1,2*], Gongjie Zhang[1], Changqing Zhou[3], Zhonghua Wu[3], Qingyi Tao[3], Lewei Lu[3], Shijian Lu[1†]

[1]S-Lab, Nanyang Technological University
[2]Black Sesame Technologies
[3]SenseTime Research
zhipeng001@e.ntu.edu.sg, shijian.lu@ntu.edu.sg

Abstract

The task of 3D single object tracking (SOT) with LiDAR point clouds is crucial for various applications, such as autonomous driving and robotics. However, existing approaches have primarily relied on appearance matching or motion modeling within only two successive frames, thereby overlooking the long-range continuous motion property of objects in 3D space. To address this issue, this paper presents a novel approach that views each tracklet as a continuous stream: at each timestamp, only the current frame is fed into the network to interact with multi-frame historical features stored in a memory bank, enabling efficient exploitation of sequential information. To achieve effective cross-frame message passing, a hybrid attention mechanism is designed to account for both long-range relation modeling and local geometric feature extraction. Furthermore, to enhance the utilization of multi-frame features for robust tracking, a contrastive sequence enhancement strategy is proposed, which uses ground truth tracklets to augment training sequences and promote discrimination against false positives in a contrastive manner. Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art method by significant margins on multiple benchmarks.

Introduction

The rapid advancement of LiDAR technology has sparked a growing interest in point cloud-based vision solutions over recent years. 3D single object tracking (SOT) based on point clouds is a fundamental task that holds enormous potential for various applications, including autonomous driving and robotics. Nevertheless, 3D SOT remains a challenging and open problem owing to the inherent properties of point clouds, such as point sparsity, partial observation, and lack of texture information. The development of effective 3D SOT solutions continues to be an active research focus.

Most existing 3D SOT approaches (Giancola, Zarzar, and Ghanem 2019; Qi et al. 2020; Zheng et al. 2021; Hui et al. 2021; Zhou et al. 2022; Hui et al. 2022) follow the prevalent paradigm of appearance matching (Fig. 1(a)), which originated from their 2D counterparts (Tao, Gavves, and Smeulders 2016; Guo et al. 2017; Li et al. 2019). Such matching-based methods perform feature matching between a cropped template and a search region to locate target objects but are prone to errors in cases of fast movement, occlusion, and misleading objects with similar appearances. Recently, a new motion-centric paradigm (Zheng et al. 2022) (Fig. 1(b)) has been proposed, which utilizes concatenated point clouds from two successive frames as input to preserve the motion connection. It performs motion estimation based on segmented foreground points to predict the relative motion

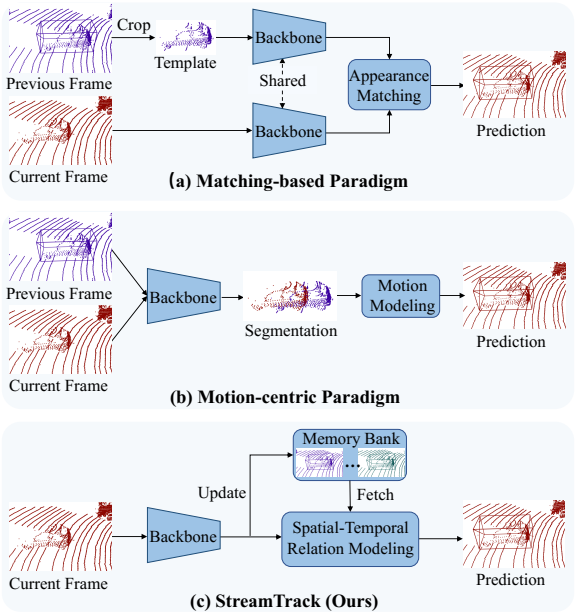

Figure 1: Comparison of 3D single object tracking paradigms. (a) The matching-based paradigm extracts features from a cropped template and a search region, and object localization is performed via appearance matching. (b) The motion-centric paradigm takes concatenated point cloud frames as input and estimates relative motion based on segmented objects. (c) Our proposed StreamTrack only takes the current frame as input, while historical features are fetched from a memory bank, allowing for the exploitation of multi-frame continuous motion for robust tracking.

*Work done at S-Lab, Nanyang Technological University
†Corresponding author

and achieves outstanding tracking performance. However, it has two limitations. First, it neglects the dynamics (*e.g.*, velocity and acceleration) contained in multi-frame historical movements, which could provide strong cues to future motion. Second, the segmentation stage eliminates background points, which might contain helpful contextual information for subsequent object localization. Erroneous segmentation could also affect tracking adversely.

Based on the aforementioned observations, we aim to propose a solution that can effectively exploit multi-frame continuous motion for accurate and robust object tracking. A straightforward method is to extend the existing motion-centric paradigm by concatenating points from multiple frames to form the input. However, this approach would result in high computational overhead and a potential problem that objects could exceed the predefined search range. To address these issues, we propose a new framework for 3D SOT named *StreamTrack*. As shown in Fig. 1(c), we treat each tracking sequence as a *stream*: at each timestamp, only the current frame is used as input, while historical features are stored in a live memory bank. Multi-frame features undergo a spatial-temporal relation modeling process to generate tracking predictions in an end-to-end manner. To achieve effective cross-frame message passing, we design a hybrid attention mechanism that can handle both long-range relation modeling and local geometric feature extraction simultaneously. To further improve the utilization of multi-frame features for robust tracking, we incorporate a contrastive sequence enhancement strategy where ground truth tracklets are used to augment training sequences and promote discrimination against false positives in a contrastive manner. This effectively mitigates the issue of target-switch, where a wrong object is tracked during tracking. We evaluate our proposed approach on KITTI, nuScenes, and Waymo datasets, and the experimental results demonstrate that StreamTrack achieves new state-of-the-art performance on all benchmarks.

The contributions of this work are summarized below: **1)** We identify an overlooked aspect in existing 3D SOT paradigms and propose StreamTrack – a new paradigm that treats each tracking sequence as a stream and utilizes a memory bank for efficient exploitation of multi-frame continuous motion; **2)** We propose a hybrid attention mechanism that can handle both long-range relation modeling and local geometric feature extraction to achieve effective cross-frame message passing; **3)** We design a contrastive sequence enhancement scheme that further improves the utilization of multi-frame features for robust tracking; **4)** Experimental results on KITTI, nuScenes, and Waymo demonstrate that StreamTrack achieves new state-of-the-art performance while still being computationally efficient.

Related Work

3D Single Object Tracking. Given a point cloud sequence and the bounding box of an object in the first frame, the goal of 3D SOT is to locate the object in subsequent frames. The primary application of 3D SOT in autonomous driving is to enable a vehicle to follow a specific object (*e.g.*, a car). Most existing 3D SOT approaches follow the matching-based paradigm, which is inspired by the success of Siamese networks in 2D tracking (Tao, Gavves, and Smeulders 2016; Li et al. 2018, 2019). As the pioneering work, SC3D (Giancola, Zarzar, and Ghanem 2019) generates a series of target proposals to match with the template based on feature similarities and selects the proposal with the top similarity. P2B (Qi et al. 2020) uses a Region Proposal Network (Qi et al. 2019) for efficient proposal generation and employs Hough Voting to generate the tracking prediction. Motivated by the success of P2B, a series of follow-up work further improves the feature correlation operation or prediction generation with more sophisticated designs. For example, SA-P2B (Zhou et al. 2021) designs an auxiliary task to learn the structure of objects. BAT (Zheng et al. 2021) encodes structural information with Box Cloud for individual points. MLVSNet (Wang et al. 2021) enhances the feature aggregation with multi-level Hough Voting. V2B (Hui et al. 2021) performs Voxel-to-BEV transformation for object localization on the densified feature maps. Inspired by the success of Transformer (Vaswani et al. 2017) on computer vision tasks (Liu et al. 2021; Carion et al. 2020a), several studies (Zhou et al. 2022; Cui et al. 2021; Shan et al. 2021; Hui et al. 2022; Guo et al. 2022; Nie et al. 2023; Xu et al. 2023) incorporate Transformer for enhanced feature extraction and correlation modeling and achieve improved accuracy.

Despite its success, the matching-based paradigm breaks the motion connection between successive frames with template cropping and does not fully exploit the distortion-free property of point clouds. This makes it sensitive to distractors with similar geometric shapes (Zheng et al. 2022). Recently, a motion-centric tracker M^2-Track (Zheng et al. 2022) achieves outstanding performance by tackling the tracking problem from the perspective of relative motion. Our proposed method also resorts to motion modeling but differs from M^2-Track by exploiting multi-frame continuous motion and having an end-to-end design.

Contrastive Learning works under the principle that similar sample pairs should be close in a learned embedding space, while distinct ones should be well separated. It has been extensively studied in representation learning (Chen et al. 2020; He et al. 2020; Henaff 2020; Oord, Li, and Vinyals 2018; Wu et al. 2018; Grill et al. 2020) and achieved remarkable success in boosting the performance of downstream tasks. Several studies (Lang, Braun, and Valada 2021; Wu et al. 2022; Zhu et al. 2022; Yao et al. 2021) extend contrastive learning to the supervised setting to learn more robust feature representations to reduce misclassification or wrong instance associations. Inspired by the above works, we design a contrastive sequence enhancement strategy to improve the robustness of tracking. To our best knowledge, this is the first effort to utilize contrastive learning in 3D SOT approaches.

Methodology

As illustrated in Fig. 2, StreamTrack consists of three modules: **1)** memory-assisted feature extraction, **2)** spatial-temporal relation modeling, and **3)** query-based prediction. We describe them in detail in the remainder of this section.

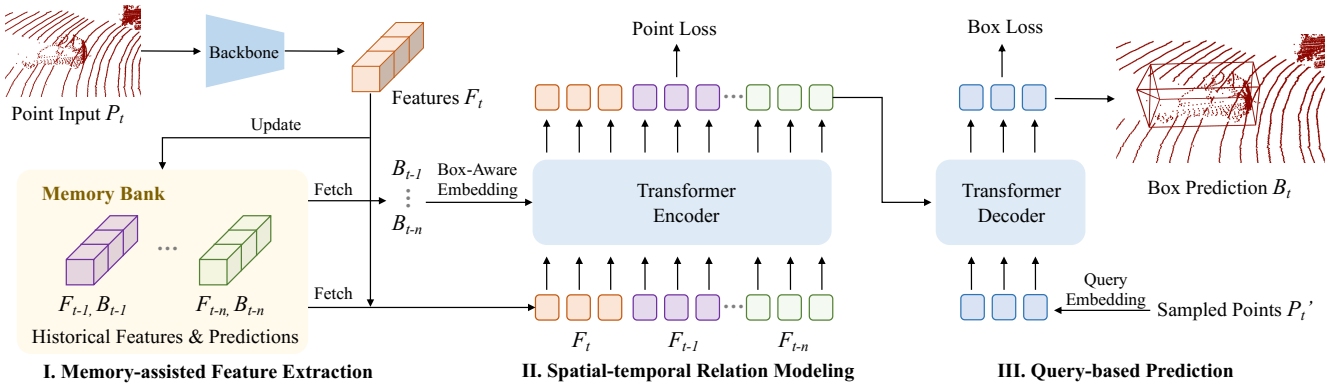

Figure 2: Overall architecture of StreamTrack. StreamTrack consists of three modules: memory-assisted feature extraction, spatial-temporal relation modeling, and query-based prediction. At timestamp t, StreamTrack only takes as input the current frame P_t, while historical features and box predictions are fetched from a memory bank for efficient computation. A Transformer encoder-decoder architecture is adopted for cross-frame message passing and the generation of tracking predictions.

Memory-assisted Feature Extraction

Existing 3D SOT methods typically rely on point clouds from two consecutive frames as input. However, multiple frames capture richer motion information, which can be exploited for more accurate and robust tracking. One possible approach to utilize multi-frame information is to concatenate points from a number of frames to form the input as in (Zheng et al. 2022). However, such an approach becomes computationally expensive as the number of points increases. Additionally, fast-moving objects may go beyond the predefined search range, as a small range is typically used to reduce search complexity and improve efficiency.

To address these issues, we propose a memory-assisted feature extraction scheme for the efficient utilization of multi-frame features. As shown in Fig. 2, given a point cloud frame $P_t \in \mathbb{R}^{N\times 3}$ at timestamp t, where N is the number of input points, we use a backbone model to extract the point features $F_t \in \mathbb{R}^{N'\times C}$, where N' is the number of points sampled by the backbone. We adopt PointNet++ (Qi et al. 2017) as our backbone as it is widely used in existing 3D SOT methods (Qi et al. 2020; Zheng et al. 2021; Zhou et al. 2022), although it is possible to use more sophisticated backbone networks to further improve the tracking performance. We employ a memory bank to store historical point features and box predictions of the past n frames, which are denoted by $\{F_{t-i}\}_{i=1}^{n}$ and $\{B_{t-i}\}_{i=1}^{n}$, respectively. At the end of each iteration, we update the memory bank with F_t and B_t and discard those from the earliest frame (F_{t-n} and B_{t-n}).

The memory bank design allows StreamTrack to bypass the repetitive computation of multi-frame point features and greatly improves the efficiency of our framework. Apart from only requiring the current frame as input at each timestamp, another major difference between our StreamTrack and M^2-Track is that M^2-Track utilizes a unified coordinate system for all input frames, whereas we shift the coordinate system for each frame to follow the movement of objects. Specifically, the canonical coordinate system defined by box prediction B_{t-1} is employed for input P_t. This design allows reusing historical features without enlarging the input range to cover long-range movements while preserving the critical relative motion.

Spatial-temporal Relation Modeling

The goal of this module is to model the cross-frame relation and propagate target information from past frames to the current frame for subsequent object localization. To deal with the complexity introduced by multi-frame features, we employ the attention mechanism of Transformer (Vaswani et al. 2017) to leverage its strong capability to model long-range dependencies. Specifically, we use an encoder consisting of L_{enc} stacked Transformer layers to encode point features and historical box locations. In the following text, we first describe the spatial-temporal relation modeling process using the original (vanilla) attention mechanism (Vaswani et al. 2017) and then introduce our proposed hybrid attention designed for more effective cross-frame feature exchange.

Vanilla Attention. For each Transformer layer, given point features $\{F_t, F_{t-1}, ..., F_{t-n}\}$, we first concatenate them to form the input $F \in \mathbb{R}^{(n+1)\times N'\times C}$. To incorporate the geometric locations of the points, we generate a position embedding PE, which is of the same shape as F, by mapping the 3D coordinates of the points with an MLP. To distinguish points from different frames, we also add a learnable temporal embedding to PE based on the temporal sequential order. Besides, the tracking process is conditioned on the prior knowledge of object locations in past frames. To incorporate past box locations, we generate a point mask $M \in \mathbb{R}^{(n+1)\times N'}$ to indicate the objectiveness of each point as in (Zheng et al. 2022). Concretely, $m_j^i \in M$ is defined as:

$$m_j^i = \begin{cases} 0 & \text{if } j \in [t-1, t-n] \text{ and } p_j^i \text{ is not in } B_j \\ 1 & \text{if } j \in [t-1, t-n] \text{ and } p_j^i \text{ is in } B_j \\ 0.5 & \text{if } j = t \end{cases} \quad (1)$$

where i indexes the points and j indexes the timestamps. Intuitively, m_j^i can be viewed as the probability of point

p_j^i belonging to the foreground. However, M does not accurately encode the box location and orientation especially when points are sparse. (Zheng et al. 2021) proposes to represent the point-to-box relation by including the distances from each point to the box center and 8 corners. We concatenate the said distances to M and obtain the box-aware point mask $M' \in \mathbb{R}^{(n+1)\times N'\times(1+9)}$, where the distance values for the current frame are set to zero due to unknown box location. Similarly, we map M' with an MLP to obtain a box-aware mask embedding *ME*. Finally, we add the position embedding to F to form query and key, while adding the mask embedding to F to obtain value, and perform the attention computation as defined in (Vaswani et al. 2017):

$$Q = K = F + PE; \quad V = F + ME \tag{2}$$

$$F' = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where F' denotes the attention output and d_k is the key dimension. A standard feedforward network (FFN) is further applied to generate the output of the Transformer layer.

Hybrid Attention. Albeit the strong global relation modeling capability of Transformer, it treats each point equally without paying specific attention to local geometric structures, which have been proven to be crucial in various point representation learning studies (Qi et al. 2017; Thomas et al. 2019; Zhao et al. 2021). To achieve more effective cross-frame message passing, we design a hybrid attention mechanism that incorporates such inductive bias into Transformer. As shown in Fig. 3, we introduce a local spatial attention operation in parallel with the regular global spatial-temporal attention to account for both local geometric feature extraction and long-range relation modeling. Specifically, we gather a local set of points $\{p_b||p_b - p_a| < r\}$ for each input point p_a with a predefined distance threshold r and replace K and V in the vanilla attention with the corresponding local point features. The outputs of both attention modules are then concatenated and merged with a Linear layer. The proposed hybrid attention enhances the learning of local geometric structures to improve the modeling of spatial-temporal relations.

Point Supervision. To promote cross-frame feature interaction and information propagation, we apply point-wise supervision on the encoder output. Specifically, for each encoder layer, we predict the point objectiveness s and point-to-box distances d as defined in the box-aware point mask generation process. The point loss $\mathcal{L}_{point}$ is formulated as:

$$\mathcal{L}_{point} = \sum_{l}^{L_{enc}} (\lambda_s \mathcal{L}_{CE}(s_l, \hat{s}) + \lambda_d \mathcal{L}_{smooth-l1}(d_l, \hat{d})) \tag{4}$$

where $\hat{(\cdot)}$ denotes the ground truth, l indexes the encoder layers, and $\mathcal{L}_{CE}$ reprensents the cross-entropy loss.

Query-based Prediction

Most existing 3D SOT methods (Qi et al. 2020; Zheng et al. 2021; Zhou et al. 2022; Guo et al. 2022; Hui et al. 2022) employ point-based RPNs to generate tracking predictions. Inspired by DETR (Carion et al. 2020b), we introduce a query-based prediction method in this work. As shown in Fig. 2, we first generate query embeddings based on the coordinates of the sampled points P_t' using an MLP. The query embeddings are then input to a decoder consisting of L_{dec} Transformer layers to interact with the encoder output for tracking prediction generation. Unlike RPN-based approaches that generate tracking predictions over the current frame, our approach allows each object query to interact with encoded features from all input frames.

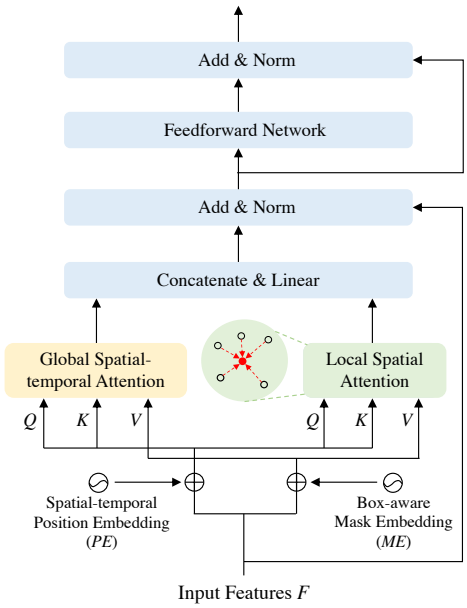

Figure 3: Architecture of the proposed hybrid attention. A local spatial attention module is introduced to work in parallel with global spatial-temporal attention to account for both local feature extraction and long-term relation modeling to achieve more effective cross-frame message passing.

Box Supervision. For each decoder layer, a classification head and a regression head are applied to generate class predictions c and box predictions b. Box matching is then performed to match the predictions to the ground truth box $\hat{b}$. The matching process is similar to the set-to-set matching in DETR except that we aim to find one single prediction that is best matched to the ground truth. We define the matching cost function with a semantic term and a geometric term:

$$\mathcal{L}_{match} = -\lambda_{cls} c - \lambda_{giou} \mathcal{L}_{giou}(b, \hat{b}) \tag{5}$$

where $\mathcal{L}_{giou}$ measures the box overlap using GIoU (Rezatofighi et al. 2019). We then select the prediction with the lowest matching cost as the positive prediction and set the classification target as one, while the rest are regarded as negative predictions with classification targets of zeros. Finally, the box loss can be formulated by:

$$\mathcal{L}_{box} = \sum_{l}^{L_{dec}} (\lambda_{cls} \mathcal{L}_{focal}(c_l, \hat{c}_l) + \lambda_{reg} \mathcal{L}_{reg}(b_l^+, \hat{b})) \tag{6}$$

where $\mathcal{L}_{focal}$ denotes the focal loss (Lin et al. 2017), and the regression loss $\mathcal{L}_{reg}$ consists of smooth-L1 and GIoU terms,

which is detailed in the Appendix. The regression loss is only applied to the best-matched positive prediction b_l^+.

Contrastive Sequence Enhancement. It has been identified that distractors (nearby objects with similar appearances) pose a significant challenge to 3D SOT as point clouds inherently provide limited appearance cues (Zheng et al. 2022). We observe that there might not exist an abundance of distractors in the training sequences since typically only a small search region is considered, especially under data-constrained scenarios. As a result, tracking methods may not be fully trained to discriminate against negative targets. To this end, we propose a *sequence enhancement* scheme by purposely attaching additional tracklets to training sequences to serve as negative samples. As illustrated in Fig. 4, with a probability ρ, we randomly sample a tracklet of the same category from the training data and attach it to the input frames. The added tracklet is placed near the target tracklet with a random relative velocity to simulate object movements. Note that although the proposed sequence enhancement shares some similarities with the commonly used 'GT-AUG' (Yan, Mao, and Li 2018) in 3D detection, they differ in the following aspects. First, 'GT-AUG' is usually applied to a single frame, while sequence enhancement works with sequential frames. Second, 'GT-AUG' introduces additional positive targets, whereas sequence enhancement attaches negative samples to enhance discrimination.

Motivated by the success of contrastive learning (Chen et al. 2020; He et al. 2020; Henaff 2020), we introduce an auxiliary contrastive loss to explicitly enforce the separation between positive and negative targets in the feature embedding space. Specifically, we add an additional GT query as in (Li et al. 2022) by generating its query embedding based on the center location of the ground truth bounding box. Note that the inclusion of the GT query does not have a significant impact on the tracking performance (see Appendix). The prediction generated by the GT query naturally forms a positive pair with the best-matched positive prediction and forms negative pairs with the remaining negative predictions. We follow the practice in MoCo (He et al. 2020) to generate GT feature embedding f_g with a momentum decoder obtained via exponential moving average, and the prediction embeddings are generated by projecting the decoder outputs with an MLP. The auxiliary contrastive loss is calculated using the InfoNCE (Oord, Li, and Vinyals 2018) loss over all decoder layers:

$$\mathcal{L}_{aux} = \sum_{l}^{L_{dec}} -\log \frac{\exp(f_g^l \cdot f_+^l / \tau)}{\sum_{i=1}^{N'} \exp(f_g^l \cdot f_i^l / \tau)} \quad (7)$$

where f_+^l denotes the feature embedding of the matched positive prediction, and τ is a temperature hyper-parameter (Wu et al. 2018). Note that contrastive sequence enhancement is only applied during training so that it brings no extra computation to the inference process.

Training Loss

We define the total loss function as the linear sum of the point loss, the box loss, and the auxiliary contrastive loss:

$$\mathcal{L}_{total} = \lambda_{point}\mathcal{L}_{point} + \lambda_{box}\mathcal{L}_{box} + \lambda_{aux}\mathcal{L}_{aux} \quad (8)$$

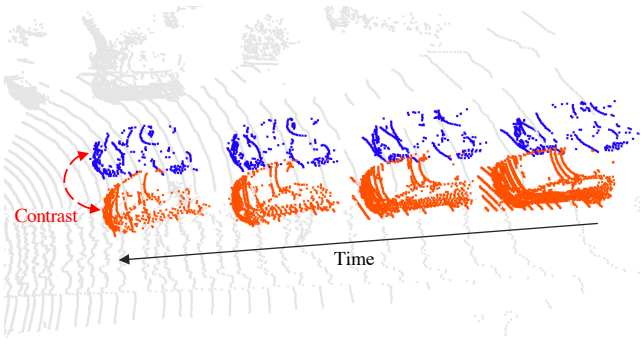

Figure 4: Illustration of contrastive sequence enhancement. Blue points denote the original target object in a tracking sequence, and orange points represent the added tracklet which serves as a negative sample. An auxiliary contrastive loss is applied to further promote discrimination.

We include more details such as hyper-parameter values in the Appendix due to space constraints.

Experiments

Experiment Setups

Datasets. We conduct extensive evaluations on three widely used datasets: KITTI (Geiger, Lenz, and Urtasun 2012), nuScense (Caesar et al. 2020), and Waymo Open Dataset (Sun et al. 2020). For KITTI, we follow the data split defined in (Giancola, Zarzar, and Ghanem 2019). The nuScenes and Waymo datasets are of significantly larger scales as compared to KITTI. We follow the implementation of (Zheng et al. 2022) for these two datasets, except we randomly sample 10% of the tracklets for training on the Waymo dataset due to its overwhelming sample size. Testing is conducted over all test samples to ensure fair comparisons.

Evaluation Metrics. We follow existing studies (Giancola, Zarzar, and Ghanem 2019; Qi et al. 2020) and use the One Pass Evaluation (Kristan et al. 2016) to measure the *Success* and *Precision* of tracking predictions. *Success* is computed from the intersection over union (IOU) of the predicted bounding box and the ground truth box, while *Precision* is defined as the area under the curve (AUC) for the distance between two box centers from 0 to 2 meters.

Benchmarking Results

Results on KITTI. On the KITTI dataset, we compare with state-of-the-art methods SC3D (Giancola, Zarzar, and Ghanem 2019), P2B (Qi et al. 2020), MLVSNet (Wang et al. 2021), BAT (Zheng et al. 2021), PTTR (Zhou et al. 2022), V2B (Hui et al. 2021), CMT (Zheng et al. 2022), and CXTrack (Xu et al. 2023). As shown in Tab. 2, the proposed StreamTrack outperforms existing methods by notable margins in terms of average success and precision, while achieving top performance for most categories. Notably, matching-based methods (*e.g.*, STNet (Hui et al. 2022) and CMT (Guo et al. 2022)) tend to perform well on Car and Van, while motion-centric method M^2-Track is competitive on Pedestrian and Cyclist. We conjecture that cars and vans are rigid

	Dataset	nuScense						Waymo		
	Category	Car	Pedestrian	Truck	Trailer	Bus	Mean	Vehicle	Pedestrian	Mean
	Frame Count	64,159	33,227	13,587	3,352	2,953	117,278	1,057,651	510,533	1,568,184
Success	SC3D	22.31	11.29	30.67	35.28	29.35	20.70	-	-	-
	P2B	38.81	28.39	42.95	48.96	32.95	36.48	28.32	15.60	24.18
	BAT	40.73	28.83	45.34	52.59	35.44	38.10	35.62	22.05	31.20
	M^2-Track	55.85	32.10	57.36	57.61	51.39	49.23	43.62	42.10	43.13
	StreamTrack	**62.05**	**38.43**	**64.67**	**66.67**	**60.66**	**55.75**	**60.23**	**47.07**	**55.95**
Precision	SC3D	21.93	12.65	27.73	28.12	24.08	20.20	-	-	-
	P2B	42.18	52.24	41.59	40.05	27.41	45.08	35.41	29.56	33.51
	BAT	43.29	53.32	42.58	44.89	28.01	45.71	44.15	36.79	41.75
	M^2-Track	65.09	60.92	59.54	58.26	51.44	62.73	61.64	67.31	63.48
	StreamTrack	**70.81**	**68.58**	**66.60**	**64.27**	**59.74**	**69.22**	**72.61**	**70.44**	**71.90**

Table 1: Performance comparison on nuScense and Waymo. *Mean* performance is weighted by the number of frames.

Category	Car	Pedestrian	Van	Cyclist	Mean
Frame Count	6424	6088	1248	308	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
MLVSNet	56.0/74.0	34.1/61.1	52.0/61.4	34.3/44.5	45.7/66.7
BAT	65.4/78.9	45.7/74.5	52.4/67.0	33.7/45.4	55.0/75.2
PTTR	65.2/77.4	50.9/81.6	52.5/61.8	65.1/90.5	57.9/78.1
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
GLT-T	68.2/82.1	52.4/78.8	52.6/62.9	68.9/92.1	60.1/79.3
STNet	72.1/**84.0**	49.9/77.2	58.0/70.6	73.5/93.7	61.3/80.1
M^2-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
StreamTrack	**72.6**/83.7	**70.5/94.7**	**61.0/76.9**	**78.1/94.6**	**70.8/88.1**

Table 2: Performance comparison on the KITTI dataset. Success / Precision are reported.

in shape and relatively sizeable, which makes them suitable for appearance matching. In contrast, humans are non-rigid and often appear in crowds, which poses challenges to the matching process. On the other hand, although M^2-Track is more robust to distractors due to its motion-centric property, its use of simple operations (*e.g.*, MLP and max-pooling) for feature extraction limits the capability of learning geometric structures. Our proposed StreamTrack employs hybrid attention for effective geometric feature extraction and leverages multi-frame continuous motion for robust tracking, thus achieving balanced performance.

Results on nuScenes and Waymo. We compare StreamTrack with methods evaluated under the same setting on nuScenes and Waymo. As shown in Tab. 1, StreamTrack achieves new state-of-the-art performance for all categories and outperforms compared methods by clear margins. On the Waymo dataset, despite using only 10% of the training samples, our proposed method still outperforms existing methods trained with the full train set. The nuScenes dataset is known for its low point density as it is collected using 32-beam LiDARs as compared to 64-beam for other datasets, while the Waymo dataset captures complex traffic scenes with numerous objects. The outstanding performance of StreamTrack on both datasets demonstrates its strong capability of tracking objects under challenging scenarios.

Inference Speed. StreamTrack achieves an inference speed of 40.7 FPS when running on a single NVIDIA V100 GPU, which is on par with existing matching-based methods (*e.g.*, P2B and BAT). Please refer to the Appendix for more information.

Ablation Study

Effectiveness of Continuous Motion Modeling. The key motivation of StreamTrack is to exploit multi-frame continuous motion for more informed and robust tracking. We conduct experiments to study the impact on performance when information from different numbers of frames (including n historical frames and the current frame) is used to generate the tracking prediction. For a more comprehensive evaluation, we also extend M^2-Track (Zheng et al. 2022) to the multi-frame setting by concatenating points from multiple frames to form its input. As shown in Fig. 6, a significant improvement of 4.8% in success is observed for StreamTrack when the number of frames increases from 2 to 3, and the performance stabilizes when the frame number is further increased. We hypothesize that two historical frames could already provide strong motion cues (*e.g.*, velocity and acceleration) to aid the tracking process, while further increasing the frame count leads to marginal gains and extra complexity. In Fig. 5, we visualize a tracking sequence to further demonstrate the effectiveness of exploiting multi-frame information in improving the robustness of tracking. Based on the experimental results, we use two historical frames ($n = 2$) in our default setting considering the efficiency aspect. On the other hand, the performance of M^2-Track only improves marginally when multiple frames are used and a downward trend is observed when the number of frames further increases. This could be attributed to the simple architecture of M^2-Track, which is not designed to handle the complex multi-frame feature interaction.

Effectiveness of Model Components. We conduct comprehensive ablation studies to investigate the effectiveness of the building components of StreamTrack. As shown in Tab. 3, removing hybrid attention (comparing #1 and #7) leads to a decrease of 1.9% in mean success and preci-

#	Hybrid	Point Sup	Query Pred	Seq Enhance	Contrast	Car	Pedestrian	Van	Cyclist	Mean
1		✓	✓	✓	✓	70.5 / 81.2	69.1 / 94.3	57.4 / 70.7	76.5 / 93.4	68.9 / 86.2
2	✓		✓	✓	✓	70.4 / 81.0	62.4 / 90.7	58.7 / 72.3	72.2 / 93.6	65.9 / 84.7
3	✓	✓		✓		70.1 / 81.0	65.1 / 92.3	59.0 / 74.3	75.4 / 94.0	67.1 / 85.6
4	✓	✓	✓			70.1 / 80.6	65.2 / 90.6	52.1 / 63.0	75.1 / 94.3	66.5 / 83.7
5	✓	✓	✓	✓		70.6 / 81.3	67.8 / 93.1	57.4 / 69.0	78.6 / 95.1	68.4 / 85.6
6	✓	✓	✓		✓	70.9 / 81.2	67.6 / 93.5	53.2 / 63.5	77.2 / 94.7	68.0 / 85.2
7	✓	✓	✓	✓	✓	72.6 / 83.7	70.5 / 94.7	61.0 / 76.9	78.1 / 94.6	70.8 / 88.1

Table 3: Ablation studies on model components. 'Hybrid' denotes hybrid attention. 'Point Sup' denotes point supervision. 'Query Pred' denotes our proposed query-based prediction paradigm. When 'Query Pred' is disabled, we replace the decoder with the prediction head in (Zheng et al. 2022). 'Seq Enhance' denotes sequence enhancement, and 'Contrast' denotes auxiliary contrastive loss. Success / Precision are reported.

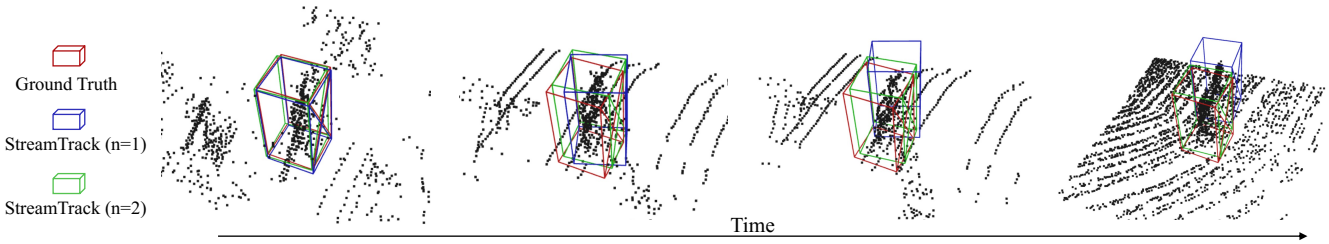

Figure 5: Visualization of tracking predictions on a Pedestrian sequence in which distractors exist. When $n = 1$, StreamTrack only relies on one historical frame, which is similar to the existing motion-centric paradigm (Zheng et al. 2022). It demonstrates that the exploitation of multi-frame continuous motion improves the tracking robustness effectively.

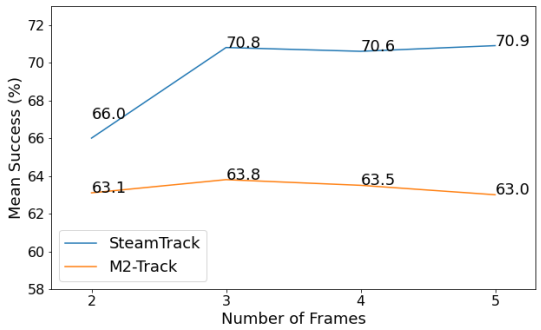

Figure 6: Mean Success vs Number of Frames that are used for predictions, evaluated over the KITTI dataset.

sion, which validates that the local attention design complements the vanilla global attention mechanism. #2 studies the impact of point supervision, and we observe a notable decrease of 4.9% and 3.4% in mean success and precision compared to our default setting (#7) when the point loss on encoder predictions is removed. In particular, the Pedestrian and Cyclist categories suffer from larger drops, which implies that low-level supervision is more important for objects of smaller sizes. To investigate the effectiveness of the proposed query-based prediction, we replace the decoder of StreamTrack with the prediction head in M^2-Track. Note that the contrastive loss is dependent on the query-based design so it is not included. By comparing #3 and #5, it can be seen that our query-based prediction achieves improved overall performance, which shows the advantage of utilizing global multi-frame features for prediction generation in our design. From #4 to #7, it can be observed that both sequence enhancement and the auxiliary contrastive loss have a positive impact on the tracking performance. Moreover, they appear to be complementary to each other, as the performance is further improved when both are applied. Remarkably, sequence enhancement introduces a substantial performance improvement (+5.3% in success) to the Van category, which has a limited number of training samples. This shows the potential of the sequence enhancement technique under data-constrained settings.

Conclusion

This paper presents StreamTrack, a new framework for 3D single object tracking (SOT) that considers each tracking sequence as a continuous stream and leverages the multi-frame motion information for more robust tracking. Our approach employs a memory-based feature extraction method to efficiently utilize multi-frame features and introduces hybrid attention to model spatial-temporal relations more effectively. We have also proposed a contrastive sequence enhancement strategy to improve the utilization of sequential information for false positive reduction. Our experimental results demonstrate that StreamTrack achieves state-of-the-art performance. We hope our work can serve as a baseline for utilizing sequential information in 3D SOT and inspire future research in this field.

Acknowledgements

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

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.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020a. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020b. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cui, Y.; Fang, Z.; Shan, J.; Gu, Z.; and Zhou, S. 2021. 3D Object Tracking with Transformer. *arXiv preprint arXiv:2110.14921*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

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.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; and Wang, S. 2017. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE international conference on computer vision*, 1763–1771.

Guo, Z.; Mao, Y.; Zhou, W.; Wang, M.; and Li, H. 2022. CMT: Context-Matching-Guided Transformer for 3D Tracking in Point Clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, 95–111. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, 4182–4192. PMLR.

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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, 293–310. Springer.

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.

Lang, C.; Braun, A.; and Valada, A. 2021. Contrastive object detection using knowledge graph embeddings. *arXiv preprint arXiv:2112.11366*.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4282–4291.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8971–8980.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13619–13627.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Nie, J.; He, Z.; Yang, Y.; Gao, M.; and Zhang, J. 2023. GLT-T: Global-Local Transformer Voting for 3D Single Object Tracking in Point Clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1957–1965.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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 International Conference on Computer Vision*, 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. *arXiv preprint arXiv:1706.02413*.

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.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Shan, J.; Zhou, S.; Fang, Z.; and Cui, Y. 2021. PTT: Point-Track-Transformer Module for 3D Single Object Tracking in Point Clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1310–1316. IEEE.

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.

Tao, R.; Gavves, E.; and Smeulders, A. W. 2016. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1420–1429.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

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, 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 International Conference on Computer Vision*, 3101–3110.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022. In defense of online models for video instance segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, 588–605. Springer.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

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

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yao, L.; Pi, R.; Xu, H.; Zhang, W.; Li, Z.; and Zhang, T. 2021. G-DetKD: towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3591–3600.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

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 International Conference on Computer Vision*, 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.

Zhou, X.; Wang, L.; Yuan, Z.; Xu, K.; and Ma, Y. 2021. Structure aware 3D single object tracking of point cloud. *Journal of Electronic Imaging*, 30(4): 043010.

Zhu, B.; Wang, Z.; Shi, S.; Xu, H.; Hong, L.; and Li, H. 2022. ConQueR: Query Contrast Voxel-DETR for 3D Object Detection. *arXiv preprint arXiv:2212.07289*.