

SeqTrack3D: Exploring Sequence Information for Robust 3D Point Cloud Tracking

Yu Lin, Zhiheng Li, Yubo Cui, Zheng Fang*

Abstract—3D single object tracking (SOT) is an important and challenging task for the autonomous driving and mobile robotics. Most existing methods perform tracking between two consecutive frames while ignoring the motion patterns of the target over a series of frames, which would cause performance degradation in the scenes with sparse points. To break through this limitation, we introduce “*Sequence-to-Sequence*” tracking paradigm and a tracker named SeqTrack3D to capture target motion across continuous frames. Unlike previous methods that primarily adopted three strategies: matching two consecutive point clouds, predicting relative motion, or utilizing sequential point clouds to address feature degradation, our SeqTrack3D combines both historical point clouds and bounding box sequences. This novel method ensures robust tracking by leveraging location priors from historical boxes, even in scenes with sparse points. Extensive experiments conducted on large-scale datasets show that SeqTrack3D achieves new state-of-the-art performances, improving by 6.00% on NuScenes and 14.13% on Waymo dataset. The code will be made public at <https://github.com/aron-lin/seqtrack3d>.

I. INTRODUCTION

3D Single Object Tracking (SOT) is an important task in computer vision that aims to locate a given target in subsequent frames based on its initial state. Owing to the ability to capture the continuous state changes of a specific target, 3D SOT trackers find widespread application in fields such as robotics, autonomous driving, and surveillance systems.

Inspired by image-based 2D SOT methods [1]–[3], early 3D SOT trackers [4]–[10] usually follow the “*Two-to-One*” paradigm (in Fig. 1(a)), which captures the target by matching template points to a search region in the current frame. However, the sparse and textureless point clouds make it challenging to track fast-moving or occluded targets. To solve this problem, M²-Track [11] considers tracking as a motion prediction task, estimating the relative motion between two clusters of the foreground points across frames. Nonetheless, both matching-based [4]–[10] and motion-based [11] methods simplify tracking to a transient prediction problem (only two frames). This not only breaks the continuity of tracking but also ignores motion pattern of target reflected in previous frames, making tracker hard to handle sparse points cases.

Lately, some works [12], [13] try to adopt a “*Sequence-to-One*” approach to merge target-specific information from historical point clouds into the current frame (in Fig. 1(b)). However, in the scenes with extremely sparse points or an

This work was supported by the National Natural Science Foundation of China under Grants 62073066, the Fundamental Research Funds for Central Universities under Grant N2226001, and 111 Project under Grant B16009. (Corresponding author: Zheng Fang, e-mail: fangzheng@mail.neu.edu.cn)

The authors are all with the Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China;

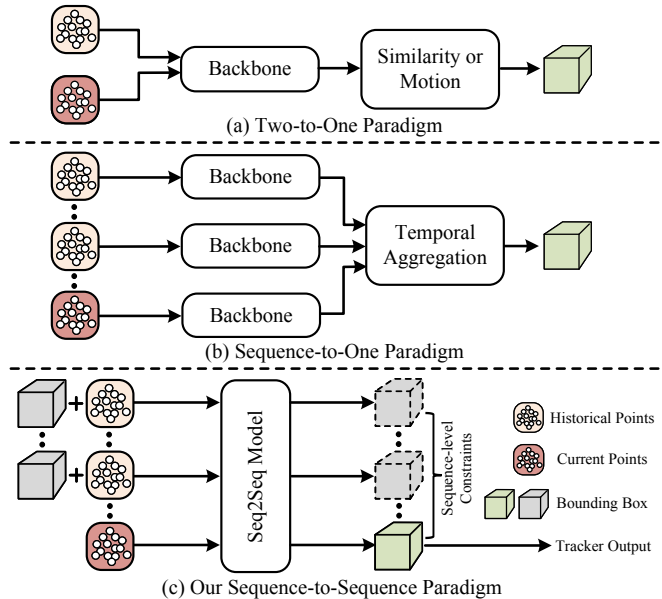


Fig. 1: **The comparison of tracking paradigms.** (a) Two-to-One paradigm exploits two frames to locate target through appearance matching or motion prediction. (b) Sequence-to-One approach uses point clouds in multi-frames to integrate the target information at different times. (c) Our Sequence-to-Sequence paradigm considers temporal clues of the target in points and box sequences to overcome sparse points cases.

invisible target, it is non-trivial to locate the target reliably through limited historical features. Compared to solely relying on point cloud sequences, we argue that the historical 3D bounding boxes (BBox) offer additional crucial insights. For instance, even when target points become sparse or are entirely absent, the current target location can still be inferred from the historical BBox trajectory which reflects the target states during a period.

Building upon the above idea, we propose a “*Sequence-to-Sequence*” (Seq2Seq) paradigm (in Fig. 1(c)), which captures the long-term motion pattern of the target in point clouds and box sequences. Specifically, our Seq2Seq framework takes a sequence of cropped points and historical BBoxes as input, performs intra- and inter-frame feature augmentation, and obtains an updated BBox sequence. The BBox in the current frame serves as the final output of tracker, while the other BBoxes enforce a sequence-level constraint to compel the network to explicitly learn the motion patterns.

To implement Seq2Seq framework, we introduce a *SeqTrack3D* tracker. It first exploits a Transformer-based encoder to extract both key geometry and motion information from

the point clouds sequence. After that, the historical boxes guide the decoder to integrate target-specific information and produce embeddings for sequential BBoxes. In this way, the model can explicitly learn motion patterns of the target while projecting the point clouds sequence to the BBox sequence. Meanwhile, instead of utilizing complex head networks such as VoteNet [14] in [4]–[6] and CenterNet [15] in [13], [16], we directly predict target attributes (x, y, z, θ) , leading to a more efficient network.

Furthermore, we prove the effectiveness of our method on two large-scale datasets, NuScenes [17] and Waymo [18]. Experimental results show that SeqTrack3D outperforms the recent STTracker by 6.00% on NuScenes and outperforms the state-of-the-art (SOTA) M²-Track by 14.13% on Waymo while running at 38 FPS.

In summary, the contributions of our work are as follows:

- We introduce a novel Seq2Seq paradigm, which advocates using sequence-level information and constraint to model continuous object motion in 3D SOT task.
- Based on Seq2Seq paradigm, we propose a SeqTrack3D tracker that leverages an encoder-decoder structure to extract spatial-temporal information from point clouds and BBoxes sequences, resulting in remarkable performance improvements.
- The experimental results on the NuScenes and Waymo demonstrate that our approach significantly outperforms the previous state-of-the-art methods. The code will be released to the research community.

II. RELATED WORK

A. 3D Single Object Tracking

Early 3D SOT approaches mainly inherited siamese structure adopted in 2D SOT [1]–[3], [19], which can be rigorously described as the “Two-to-One” paradigm. For instance, pioneering SC3D [10] adopted a siamese structure to capture the target shape and generated a series of candidate boxes within the search region. The best proposal was then selected based on cosine similarity. However, heuristic sampling in SC3D hinders end-to-end training. After that, P2B [4] tried to utilize VoteNet [14] to improve target generation mechanism and achieved end-to-end training. In subsequent studies such as PTT [6], BAT [5], LTTR [7], SMAT [16], STNet [20], etc., researchers have attempted to enhance performance through improving the quality of similarity features and optimizing target generation strategies. Despite advances in matching-based methods [4]–[6], they are limited by the sparse and incomplete natures of point clouds. To alleviate this problem, M²-Track [11] introduced the motion-centric paradigm that located the target by estimating the relative motion between two frames and achieved remarkable performance improvement. Nevertheless, the above methods overlook that tracking is a complicated dynamic process and do not fully explore spatial-temporal contextual information.

Recently, some studies introduced the “Sequence-to-One” paradigm that exploits multi-frame point clouds to estimate the target box in the current. TAT [12] selected high-quality

historical templates and aggregated historical clues through the recurrent neural networks (RNN). However, the selective usage of previous frames interrupts the continuous motion information of target. Additionally, STTracker [13] encoded points sequence in the bird’s eye view (BEV) space and split feature maps into sequential patches to capture target motion through deformable attention [21]. Unfortunately, the target might be separated into different patches, thereby destroying appearance integrity. Thus, the approaches in [12], [13] still struggle to effectively exploit historical information to locate target accurately in challenging scenes.

B. Transformer-based Sequence Learning

The Transformer structure has developed into a universal approach for sequence learning and has also been adapted for hot topics in computer vision, such as classification and object detection [22]–[26]. Specifically, ViT [22] transformed image into sequential patches and captured global contextual information by attention mechanism. Swin Transformer [23] introduced window attention to reduce computational complexity. Meanwhile, DETR [24] utilized sequential patches for object detection, which has also been extended to the 3D task in DETR3D [25]. Similarly, Segformer [26] considered segmentation as sequence learning and utilized Transformer to learn long-range dependencies among pixels. Inspired by the above works, we employ Transformer to construct mapping relationship between point clouds and object sequences, resulting in robust tracking.

III. METHODOLOGY

A. Framework Overview

3D SOT aims to identify the target bounding box $\mathcal{B}_t = (x, y, z, w, l, h, \theta) \in \mathbb{R}^7$ within the current point cloud $\mathcal{P}_t \in \mathbb{R}^{W \times 3}$ based on a given initial BBox \mathcal{B}_0 . Here, (x, y, z) , (w, l, h) and θ represent the 3D center, size and orientation, respectively. $W \times 3$ denotes W points with 3D coordinates. Unlike in 2D images where target dimensions can be changed due to the different viewpoints, we assume that the size is constant in 3D space. Thus, following previous studies, we only estimate (x, y, z, θ) in every frame. Besides, different from approaches [12], [13] which focus on multi-frame point clouds information $\{\mathcal{P}_{t-n}\}_{n=0}^N$, we additionally introduce object sequence $\{\mathcal{B}_{t-n}\}_{n=1}^N$ and reformulate the tracking task as a Seq2Seq problem:

$$\mathcal{F}(\{\mathcal{P}_{t-n}\}_{n=0}^N, \{\mathcal{B}_{t-n}\}_{n=1}^N) \rightarrow \{\hat{\mathcal{B}}_{t-n}\}_{n=0}^N \quad (1)$$

where N denotes the length of time window, and $\hat{\mathcal{B}}_t$ is the estimated BBox. Note that we transform both the points and BBox sequence into a unified local coordinate system with ego pose.

Based on Eq. 1, we introduce a SeqTrack3D tracker that can construct spatiotemporal correlations of point sequences and exploit continuous motion cues in object sequences. As shown in Fig. 2, given a sequence $\{\mathcal{P}_{t-n}\}_{n=0}^N$, SeqTrack3D first employs a weight-shared backbone to aggregate multi-frame features and predict a coarse BBox \mathcal{B}_t , which is then used to form an object sequence $\{\mathcal{B}_{t-n}\}_{n=0}^N$ with historical

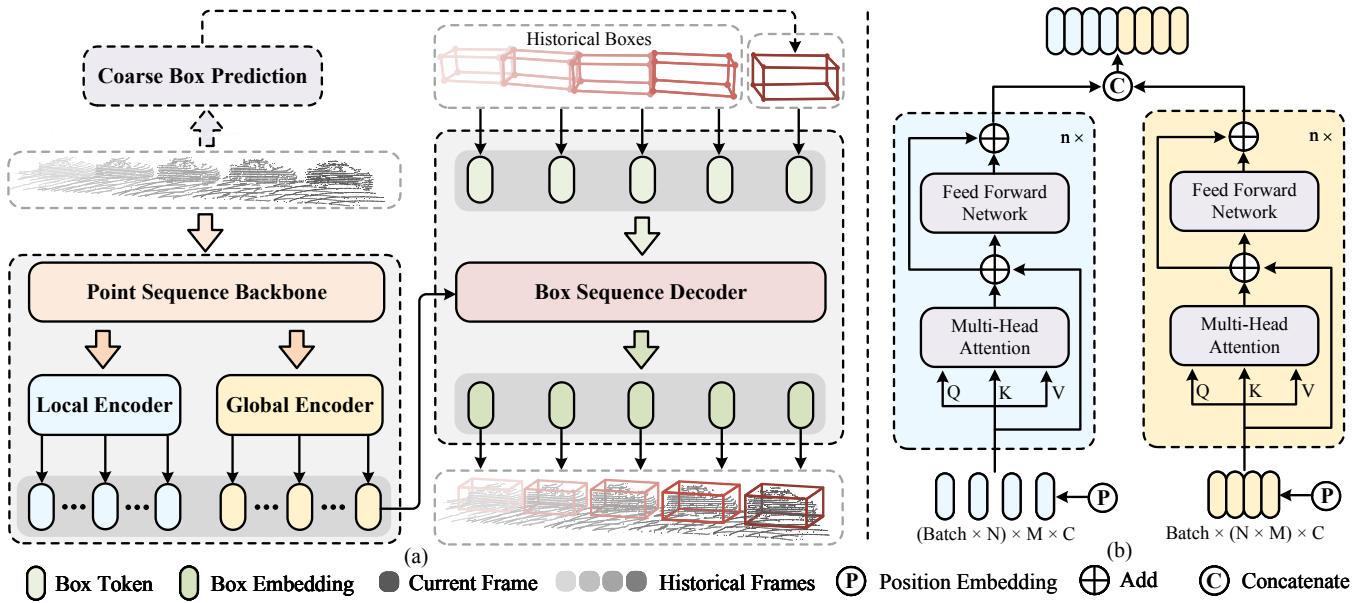


Fig. 2: (a) Overview of SeqTrack3D tracker. The encoder establishes spatial-temporal relations for point sequence. Guided by object sequence with target prior, the decoder generates box embeddings and utilizes them to predict sequential bounding boxes. (b) Details of the local-global encoder that encodes point sequence in a decoupled manner.

boxes. Next, we present a local-global encoder responsible for encoding inter- and intra-frame contextual information, thereby obtaining sequential features $\{F_{t-n}\}_{n=0}^N$. Finally, the decoder utilizes $\{F_{t-n}\}_{n=0}^N$ to produce final BBox sequence $\{\hat{\mathcal{B}}_{t-n}\}_{n=0}^N$ with guidance of object sequence $\{\mathcal{B}_{t-n}\}_{n=0}^N$.

B. Coarse Box Prediction

In SeqTrack3D, we aim to project point cloud sequences to BBox sequences. Borrowing from object detection methods such as [21], [24], [25], we predefine a box query for every object at different time steps to establish query-object pairs. These queries are designed to extract target-specific features from point sequence $\{\mathcal{P}_{t-n}\}_{n=0}^N$. For objects in historical frames, we adopt their historical BBoxes to generate queries. Similarly, for the current frame, we also wish to produce a query in the same approach and thus opt to estimate a coarse BBox. For efficiency, we draw inspiration from M²-Track [11]. Firstly, we utilize a PointNet [27] to generate the multi-frame point features $H_{t-N}^t \in \mathbb{R}^{N \times W \times C}$, where W denotes the number of points in a single frame and C is the number of channels. Then, the features are fed into multi-layer perceptron (MLP) to generate a foreground point mask $\mathcal{M} \in \mathbb{R}^{N \times W}$. Finally, H_{t-N}^t is weighted by mask \mathcal{M} and undergoes max-pooling with another MLP to obtain a coarse bounding box \mathcal{B}_t in the local coordinate system. The above process can be noted as:

$$\mathcal{B}_t = \text{MLP}(\text{Pooling}(H_{t-N}^t \times \mathcal{M})), \mathcal{B}_t \in \mathbb{R}^4 \quad (2)$$

C. Local-Global Feature Encoding

Previous approaches like [12], [13] employ a single Transformer structure to extract the spatial and temporal features at the same time. Although the above methods are effective at capturing intricate dependencies, they could be problematic since spatial and temporal features possess distinct attributes.

In particular, spatial features focus on relative positions and partial shapes, while temporal features reflect target motion patterns. Therefore, the tightly-coupled manner in [12], [13] may cause the network to be biased during optimization and overlook useful cues for target tracking.

Inspired by video processing in [28]–[30], we mitigate the above issue via “divide-and-conquer” strategy and propose a local-global encoder that decouples the encoding of spatial and temporal features. The local encoder concentrates on perceiving target location in each frame, while global encoder constructs inter-frame relation for target dynamics over time.

In this part, we first exploit a modified PointNet++ [31] as a point sequence backbone for point feature extraction and generate $F_n \in \mathbb{R}^{M \times C}$ ($n \in \{t-N, \dots, t\}$). Contrasting with original PointNet++, our modified version with fewer set abstraction layers aims to ensure efficiency by reducing the number of point features, which also enables affordable subsequent attention calculations. Thereafter, we feed point features F_n into local encoder for further encoding. Following 3DETR [32], we employ LayerNorm [33] to get features $E_n \in \mathbb{R}^{M \times C}$:

$$E_n = \text{LayerNorm}(F_n) \quad (3)$$

Then, we project E_n into *query*, *key*, and *value* embeddings, which is formulated as:

$$Q_n = E_n W_q, K_n = E_n W_k, V_n = E_n W_v \quad (4)$$

where W_q, W_k, W_v are the linear projection matrices. After obtaining Q_n, K_n, V_n , we use a multi-head attention mechanism that can be described as:

$$\text{head} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

$$\text{MHA}(Q_n, K_n, V_n) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o \quad (6)$$

$$\hat{F}_n = F_n + \text{Dropout}(\text{MHA}(Q_n, K_n, V_n)) \quad (7)$$

Here, h means the number of heads, and MHA represents the attention mechanism across multiple subspaces that enhances the expressive capacity of model and increases sensitivity to different aspects of input data. Later, the features are fed into a feed-forward network (FFN) for further refinement:

$$\tilde{F}_n = \hat{F}_n + \text{Dropout}(\text{FFN}(\text{Layernorm}(\hat{F}_n))) \quad (8)$$

In the above procedure, the local encoder only calculates the attention map within each individual frame, leading to geometric features $\{\tilde{F}_{t-N}, \dots, \tilde{F}_t\}$ encoded in each frame independently. On the other hand, the global encoder adopts the same design as local encoder, but the difference is that we merge features $\{F_{t-N}, \dots, F_t\}$ into $E_m \in \mathbb{R}^{NM \times C}$, which enables attention maps to span entire sequence. Then, E_m is processed by multiple attention layers to learn inter-frame relations and embed motion features into $\{\tilde{G}_{t-N}, \dots, \tilde{G}_t\}$.

Finally, we combine the outputs from both local and global encoders to produce a sequential feature $\tilde{F}_{gl} \in \mathbb{R}^{2MN \times C}$. In this way, the network can take into account local geometric structures and global dynamics in a balanced manner.

D. Box Sequence Generating

The intention of this section is to conduct the decoder to generate a more accurate BBox sequence via $\{\mathcal{B}_{t-n}\}_{n=0}^N$ and sequential feature \tilde{F}_{gl} . To bridge the gap between BBox and point representations, we convert the BBox parameters $P = (x, y, z, w, l, h, \theta)$ into the BBox corners $C \in \mathbb{R}^{L \times K}$, where $K = (x_c, y_c, z_c, t)$ contains spatial location and timestamp, and L is the number of corners. Then, we project $\{C_{t-n}\}_{n=0}^N$ to box token $T \in \mathbb{R}^{N \times L \times C'}$ with location prior of target:

$$\{C_{t-n}\}_{n=0}^N = \psi(\{P_{t-n}\}_{n=0}^N), T = \phi(\{C_{t-n}\}_{n=0}^N) \quad (9)$$

where N, C' mean the number of boxes and channel dimensions. $\psi(\cdot)$ projects BBox parameters to BBox corners, and $\phi(\cdot)$ performs a linear mapping. Subsequently, BBox tokens and point cloud features are transformed into *query*, *key*, and *value* embeddings for decoder input as follows:

$$Q' = TW'_q, K' = \tilde{F}_{lg}W'_k, V' = \tilde{F}_{lg}W'_v \quad (10)$$

Then, the decoder utilizes an attention mechanism consistent with encoder to generate BBox Embedding $D \in \mathbb{R}^{N \times L \times C'}$, which collects target-specific information from local-global features \tilde{F}_{gl} . Finally, we flatten D to the size of $N \times LC'$ and feed it through the MLP to derive the sequential BBox parameters $\{\hat{\mathcal{B}}_{t-n}\}_{n=0}^N$. These sequential parameters can act as constraints during training, while $\hat{\mathcal{B}}_t$ becomes final output for tracking in the inference phase.

E. Implementation Details

Loss Functions. The loss function is described by Eq. 11. Specifically, \mathcal{L}_{mk} represents the mask loss for points, which is computed using a cross entropy loss. \mathcal{L}_{cb} denotes the coarse BBox loss. Both \mathcal{L}_p and \mathcal{L}_c account for the loss of previous and current bounding boxes in the predicted object sequence, respectively. The BBox loss consists of two components: the

angle loss and the center loss. Both of these components use the Huber Loss [34] for computation. To prevent the network from overly relying on historical BBox that will accumulate errors, we set the weight ratio $\gamma_1 : \gamma_2 = 1 : 10$.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{mk} + \lambda_2 \mathcal{L}_{cb} + \lambda_3 (\gamma_1 \mathcal{L}_p + \gamma_2 \mathcal{L}_c) \quad (11)$$

Training & Inference. Our SeqTrack3D supports end-to-end training. To facilitate this, it is crucial to ensure uninterrupted gradient flow during the implementation of Eq. 9. Notably, we adopt random offsets to the ground truth during training to simulate the historical boxes with errors during the testing phase. In data processing, each point cloud is attached with a timestamp and mask. The mask is set to 0 or 1 to distinguish foreground points, while the points in the current frame are assigned 0.5 as their status remains unknown. The number of attention layers and heads h in local-global encoder is set to 3 and 4, respectively. We train the model for 40 epochs on NVIDIA RTX 4090 GPUs using an Adam optimizer with an initial learning rate of 1×10^{-4} .

IV. EXPERIMENTS

Datasets & Evaluation. We conduct extensive experiments on NuScenes and Waymo datasets. NuScenes uses 32-beam LiDAR and annotates around 1.4M instances, while Waymo employs 64-beam LiDAR. Compared with 10 Hz annotation frequency in Waymo, NuScenes is more challenging due to low frequency of only 2 Hz. Furthermore, we utilize the One Pass Evaluation (OPE) protocol [36] to evaluate our SeqTrack3D through two primary metrics: *Success* and *Precision*. *Success* quantifies the overlap between the predicted and ground-truth BBoxes via the Intersection Over Union (IOU), computing its Area Under Curve (AUC) when the overlap exceeds a threshold ranging from 0 to 1. *Precision* evaluates localization accuracy of target based on the distance between BBox centers, using an AUC for distances ranging from 0 to 2 meters.

A. Quantitative Results.

Comparison with SOTA methods. As displayed in Table. I, we compare SeqTrack3D against previous SOTA methods, including SC3D [10], P2B [4], BAT [5], GLT-T [9], CX-Track [35], PTTR [8], M²-Track [11] and STTracker [13]. The results exhibit that our method achieves significant superiority in all categories. Specifically, on the NuScenes dataset, our method outperforms the *Sequence-to-One* method [13] with improvements of 6.00%/2.26% in *Success/Precision*. This validates the advantage of the proposed *Sequence-to-Sequence* paradigm and the effectiveness of our decoupling strategy in space-time encoding. Furthermore, SeqTrack3D surpasses M²-Track [11] on Waymo by 14.13%/8.38% with a notable increase of 19.94%/11.84% in the vehicle category. Thus, we think that compared to estimating frame-to-frame motion in [11], our paradigm captures motion patterns over multiple frames and achieve significant performance gain.

Robustness to Sparsity. In order to present a finer-grained robustness comparison, we plotted the *Success* metric of different methods for the Car category under various degrees of

TABLE I: Comparison of SeqTrack3D against state-of-the-art methods on NuScenes and Waymo Open Dataset.

	Dataset	NuScenes						Waymo Open Dataset		
	Category Frame Number	Car 64,159	Pedestrian 33,227	Truck 13,587	Trailer 3,352	Bus 2,953	Mean 117,278	Vehicle 1,057,651	Pedestrian 510,533	Mean 1,568,184
Success	SC3D [10]	22.31	11.29	30.67	35.28	29.35	20.70	-	-	-
	P2B [4]	38.81	28.39	42.95	48.96	32.95	36.48	28.32	15.60	24.18
	BAT [5]	40.73	28.83	45.34	52.59	35.44	38.10	35.62	22.05	31.20
	GLT-T [9]	48.52	31.74	52.74	57.60	44.55	44.42	-	-	-
	CXTrack [35]	48.92	31.67	51.40	60.64	40.11	44.43	-	-	-
	PTTR [8]	51.89	29.90	45.30	45.87	43.14	44.50	-	-	-
	M ² -Track [11]	55.85	32.10	57.36	57.61	51.39	49.23	43.62	42.10	43.13
	STTracker [13]	56.11	37.58	54.29	48.13	36.31	49.92	-	-	-
SeqTrack3D (Ours)	62.55	39.94	60.97	68.37	54.33	55.92	63.56	44.21	57.26	
<i>Improvement</i>	<i>↑6.44</i>	<i>↑2.36</i>	<i>↑3.61</i>	<i>↑7.73</i>	<i>↑2.94</i>	<i>↑6.00</i>	<i>↑19.94</i>	<i>↑2.11</i>	<i>↑14.13</i>	
Precision	SC3D [10]	21.93	12.65	27.73	28.12	24.08	20.20	-	-	-
	P2B [4]	43.18	52.24	41.59	40.05	27.41	45.08	35.41	29.56	33.51
	BAT [5]	43.29	53.32	42.58	44.89	28.01	45.71	44.15	36.79	41.75
	GLT-T [9]	54.29	56.49	51.43	52.01	40.69	54.33	-	-	-
	CXTrack [35]	55.61	56.64	50.93	54.44	35.83	54.83	-	-	-
	PTTR [8]	58.61	45.09	44.74	38.36	37.74	52.07	-	-	-
	M ² -Track [11]	65.09	60.92	59.54	58.26	51.44	62.73	61.64	67.31	63.48
	STTracker [13]	69.07	68.36	60.71	55.40	36.07	66.68	-	-	-
SeqTrack3D (Ours)	71.46	68.57	63.04	61.76	53.52	68.94	73.48	68.52	71.86	
<i>Improvement</i>	<i>↑2.39</i>	<i>↑0.21</i>	<i>↑2.33</i>	<i>↑3.57</i>	<i>↑2.08</i>	<i>↑2.26</i>	<i>↑11.84</i>	<i>↑1.21</i>	<i>↑8.38</i>	

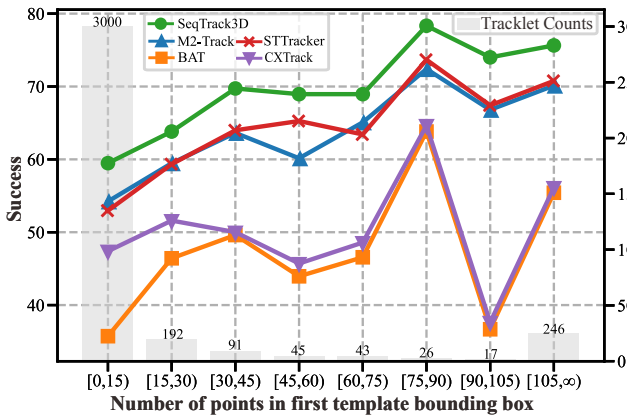


Fig. 3: Tracking performance across varying numbers of template points in the first frame.

point sparsity in Fig. 3. In most sparse cases, where the initial frame contains only 0-15 points, our SeqTrack3D surpasses other methods, which validates the obvious advantage of our paradigm when dealing with sparse scenes.

Inference Speed. Our model is designed to extract position-related information from historical point clouds and BBoxes to model a continuous motion. Although performing feature extraction on multiple point clouds does increase the computational load, we opt for a lightweight backbone, even though a more powerful structure such as DGCNN [37] is available. As a result, our choice strikes a balance between speed and performance, enabling it to maintain competitive runtime, as shown in Table. II.

B. Qualitative Analysis.

To better understand the behaviour of our model across different scenes, we illustrate the complete tracking trajectories in Fig. 4. In the case of sparse points (on the left), CXTrack is hard to extract robust point features and results in early localization failure, while M²-Track depend on

TABLE II: Comparison of inference speed.

Method	P2B	BAT	CXTrack	PTTR
FPS	45.5	57.0	34.0	51.0
Method	GLT-T	M ² -Track	STTracker	SeqTrack3D
FPS	30.0	57.0	23.6	38.0

relative motion prediction to persist for a while. Despite this, once the target is continuously occluded and invisible, M²-Track also tends to fail. In contrast, our model could actively use prior knowledge from past states, which ensures a high probability of recapturing target when it reappears. In the dense scene (on the top right), M²-Track capture a specific target with a relatively tight fit. However, benefiting from utilizing sequence data, our model can produce a trajectory that aligns more consistently with the ground truth. Different from the vehicles that typically maintain a certain distance from each other, the pedestrians are often close together and easily mislead tracker to the incorrect object (on the bottom right). Nevertheless, due to considering the historical states, our model could suppress sudden wrong shifts that do not conform to motion patterns and prevent the box from being erroneously assigned to another object.

C. Ablation Study.

In this section, to thoroughly explore the effect of different components and settings in the proposed SeqTrack3D, we perform a series of ablation studies on Car and Pedestrian categories of NuScenes, which are the predominant classes. Specifically, these experiments mainly focus on three aspects: the contributions of individual model components, the impact of varying time window length, and the influence of explicitly constraining predicted historical BBoxes during training. For evaluation purposes, we train on 1/5 of the training set and assess the complete validation set.

Model Components. Our SeqTrack3D mainly consists of an encoder and decoder. In order to investigate the interplay be-

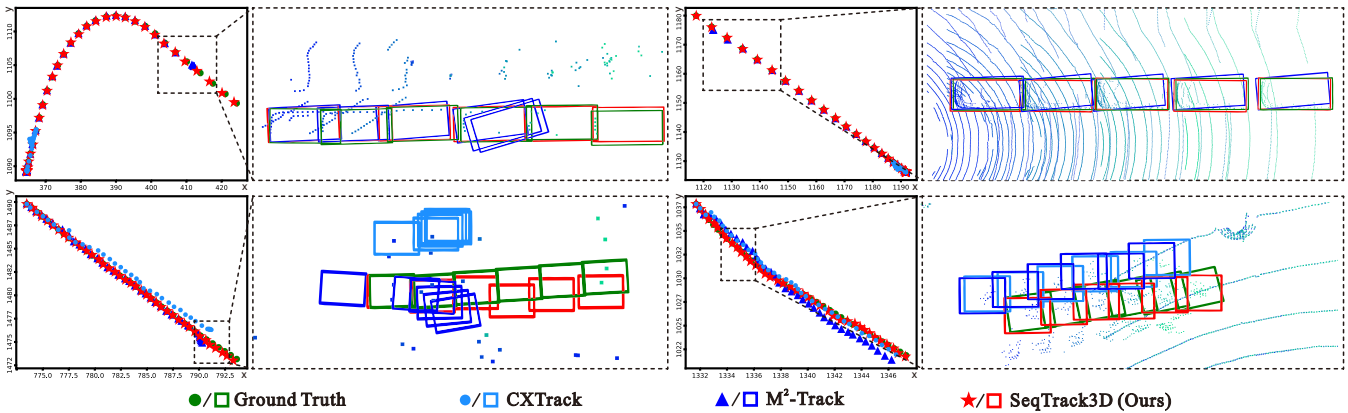


Fig. 4: **Visualization of tracking results on NuScenes.** Complete tracking trajectories are projected onto the X-Y plane of the global coordinate system. Left: Point sparsity cases; Right: Dense cases; Top: Car category; Bottom: Pedestrian category.

TABLE III: **Ablation study of components in SeqTrack3D.** “C” is Coarse Box Prediction. “L” and “G” refer to Local and Global Encoder. “D” means Decoder.

Component	Car		Pedestrian	
	Success	Precision	Success	Precision
C	51.42	59.66	30.89	57.65
C + G + D	52.17	62.60	31.83	59.06
C + L + D	54.30	64.57	32.05	59.22
C + L + G + D	56.73	66.01	32.80	61.53

TABLE IV: **Ablation study on the time window length.**

Time Length N	Car		Pedestrian	
	Success	Precision	Success	Precision
1 + 1	51.60	61.30	31.94	60.57
1 + 3 (Ours)	56.73	66.01	32.80	61.53
1 + 5	54.57	62.98	32.48	62.72
1 + 7	51.15	60.07	33.44	62.70

tween them and comprehend their importance in overall performance, we conduct experiments focusing on each model component. In Table. III, when using optimal time window length $N = 4$, we observe that omitting the local and global features will significantly damage model performance. Then, a moderate progress is noticed when only global encoder is employed to capture inter-frame dynamics. It is also worth noting that solely using local encoder would be better than global one. We speculate that the intra-frame geometry cue is important for accurate bounding box prediction. Finally, combining global and local information can achieve optimal result, which proves that decoupling local and global feature extraction allows the tracker to learn both geometric and dynamic aspects of target in point sequence.

Time Window Length. The time window length controls how far back in time our model draws knowledge from. It’s intuitive to consider that earlier historical frames contain older states of a target, but too old frames might not always contribute positively to the prediction in the current frame. Additionally, longer time windows during training can make it challenging to simulate conditions in the testing phase. Based on these considerations, we conduct experiments on the time window length to determine the optimal setting for

TABLE V: **Ablation study on the object-wise constraints for box sequence.** X+Y indicates constraints applied to X current and Y previous boxes.

Constraints Length	Car		Pedestrian	
	Success	Precision	Success	Precision
1 + 0	53.09	63.03	30.17	58.62
1 + 1	55.22	64.58	30.52	59.21
1 + 2	55.83	65.40	31.12	59.49
1 + 3 (Ours)	56.73	66.01	32.80	61.53

SeqTrack3D. As illustrated in Table. IV, the Car improves *Success* by 5.13% and *Precision* by 4.71% when time length increases from 1+1 to 1+3, indicating the positive impact of incorporating historical data. However, when the number of frames increases to 1+7, the tracking performance almost degrades to be equivalent to 1+1 frames. We believe that the reasons are two-fold: First, the random offsets for longer historical BBox sequences during training might make it difficult to simulate test conditions; Second, more cumulative errors contained in the historical boxes will also contribute to performance drop.

Constraints Length. Even though using BBox in the current frame as the final output, we still compute the loss of BBox in historical frame to impose constraints. This encourages the model to learn the motion patterns of target from previous priors. As demonstrated in Table. V, we change the number of constraints on historical boxes while fixing time window length at $N = 4$. The results indicate when the number of constraints equals to the sequence length, our method reaches optimal performance.

V. CONCLUSIONS

In this paper, we propose a Seq2Seq tracking framework to address 3D SOT task. By leveraging Transformer to process sequence data, our SeqTrack3D could capture dynamic traits of target from point cloud and BBox sequences and result in a remarkable performance improvement. However, we notice a degradation as the length of point sequence increases. In future work, we will further explore methods to effectively capture spatial and temporal information in long sequences.

REFERENCES

- [1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *European Conference on Computer Vision Workshops* (G. Hua and H. Jégou, eds.), pp. 850–865, 2016.
- [2] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8971–8980, 2018.
- [3] Z. Zhang and H. Peng, "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.
- [4] H. Qi, C. Feng, Z. Cao, F. Zhao, and Y. Xiao, "P2b: Point-to-box network for 3d object tracking in point clouds," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [5] C. Zheng, X. Yan, J. Gao, W. Zhao, W. Zhang, Z. Li, and S. Cui, "Box-aware feature enhancement for single object tracking on point clouds," 2021.
- [6] J. Shan, S. Zhou, Z. Fang, and Y. Cui, "Ptt: Point-track-transformer module for 3d single object tracking in point clouds," 2021.
- [7] Y. Cui, Z. Fang, J. Shan, Z. Gu, and S. Zhou, "3d object tracking with transformer," 2021.
- [8] C. Zhou, Z. Luo, Y. Luo, T. Liu, L. Pan, Z. Cai, H. Zhao, and S. Lu, "Ptr: Relational 3d point cloud object tracking with transformer," in *CVPR*, pp. 8531–8540, 2022.
- [9] J. Nie, Z. He, Y. Yang, M. Gao, and J. Zhang, "Glt-t: Global-local transformer voting for 3d single object tracking in point clouds," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1957–1965, 2023.
- [10] S. Giancola, J. Zarzar, and B. Ghanem, "Leveraging shape completion for 3d siamese tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [11] C. Zheng, X. Yan, H. Zhang, B. Wang, S. Cheng, S. Cui, and Z. Li, "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.
- [12] K. Lan, H. Jiang, and J. Xie, "Temporal-aware siamese tracker: Integrate temporal context for 3d object tracking," in *Proceedings of the Asian Conference on Computer Vision*, pp. 399–414, 2022.
- [13] Y. Cui, Z. Li, and Z. Fang, "Stracker: Spatio-temporal tracker for 3d single object tracking," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4967–4974, 2023.
- [14] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [15] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.
- [16] Y. Cui, J. Shan, Z. Gu, Z. Li, and Z. Fang, "Exploiting more information in sparse point cloud for 3d single object tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11926–11933, 2022.
- [17] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.
- [19] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "Siamcar: Siamese fully convolutional classification and regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6269–6277, 2020.
- [20] L. Hui, L. Wang, L. Tang, K. Lan, J. Xie, and J. Yang, "3d siamese transformer network for single object tracking on point clouds," *arXiv preprint arXiv:2207.11995*, 2022.
- [21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.
- [24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, pp. 213–229, 2020.
- [25] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.
- [26] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021.
- [27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [28] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, vol. 2, p. 4, 2021.
- [29] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.
- [30] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3163–3172, 2021.
- [31] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [32] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2906–2917, 2021.
- [33] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [35] T.-X. Xu, Y.-C. Guo, Y.-K. Lai, and S.-H. Zhang, "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.
- [36] M. Kristan, J. Matas, A. Leonardis, T. Vojtš, R. Pflugfelder, G. Fernandez, G. Nebel, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2137–2155, 2016.
- [37] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.