S2-Track: A Simple yet Strong Approach for End-to-End 3D Multi-Object Tracking

Tao Tang[‡ * 1] **Lijun Zhou**[* 2] **Pengkun Hao**[2] **Zihang He**[2] **Kalok Ho**[2] **Shuo Gu**[2] **Zhihui Hao**[2] **Haiyang Sun**[2] **Kun Zhan**[2] **Peng Jia**[2] **Xianpeng Lang**[2] **Xiaodan Liang**[† 1]

Abstract

3D multiple object tracking (MOT) plays a crucial role in autonomous driving perception. Recent end-to-end query-based trackers simultaneously detect and track objects, which have shown promising potential for the 3D MOT task. However, existing methods are still in the early stages of development and lack systematic improvements, failing to track objects in certain complex scenarios, like occlusions and the small size of target object's situations. In this paper, we first summarize the current end-to-end 3D MOT framework by decomposing it into three constituent parts: query initialization, query propagation, and query matching. Then we propose corresponding improvements, which lead to a strong yet simple tracker: S2-Track. Specifically, for query initialization, we present 2D-Prompted Query Initialization, which leverages predicted 2D object and depth information to prompt an initial estimate of the object's 3D location. For query propagation, we introduce an Uncertainty-aware Probabilistic Decoder to capture the uncertainty of complex environment in object prediction with probabilistic attention. For query matching, we propose a Hierarchical Query Denoising strategy to enhance training robustness and convergence. As a result, our S2-Track achieves state-of-the-art performance on nuScenes benchmark, i.e., 66.3% AMOTA on test split, surpassing the previous best end-to-end solution by a significant margin of 8.9% AMOTA. We achieve 1st place on the nuScenes tracking task leaderboard.

[*]Equal contribution , [‡] Work done during an internship at Li Auto Inc. [1]Shenzhen Campus of Sun Yat-sen University [2]Li Auto Inc. Correspondence to: Xiaodan Liang <liangxd9@mail.sysu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

3D multiple object tracking (MOT) (Li et al., 2023d; Doll et al., 2023; Pang et al., 2023; Qing et al., 2023; Yang et al., 2022; Li et al., 2023a; Wang et al., 2023a; Ding et al., 2024) is an essential component for the perception of autonomous driving systems. The ability to accurately and robustly track objects in dynamic environments is crucial for ensuring smooth and safe navigation and reasonable decision-making. Traditional 3D MOT methods (Yang et al., 2022; Chaabane et al., 2021; Zhou et al., 2020; Yin et al., 2021; Pang et al., 2022; Weng et al., 2020; Wojke et al., 2017; Guo et al., 2024) rely on detector outcomes followed by a post-processing module like data association and trajectory filtering, leading to a complex pipeline. To avoid human-crafted heuristic design in detection-based trackers, recent advancements in end-to-end query-based approaches have shown impressive potential in addressing the 3D MOT task by simultaneously detecting and tracking objects (Zeng et al., 2022; Zhang et al., 2022a; Doll et al., 2023; Li et al., 2023d; Pang et al., 2023; Ding et al., 2024). These methods have demonstrated promising results in terms of tracking performance and efficiency. However, current end-to-end trackers are still in the early stages of development and can not effectively handle the various complex driving scenarios with their naive solution.

In driving scenarios, the environment could be highly complex, often when driving in cities, with numerous objects such as vehicles and pedestrians interleaving across the scene, and exhibiting substantial variations in their motion patterns. Furthermore, tracked objects often cover a wide spatial tracking range and a long temporal tracking sequence. As a result, occlusion situations and the small size of target objects, frequently occur, which usually leads to some undetected or occluded objects losing track. These factors present significant challenges to current vanilla end-to-end query-based approaches for achieving accurate and robust 3D MOT. As shown in Figure 1 (a), the previous state-of-the-art end-to-end tracker, PF-Track (Pang et al., 2023), fails to track objects in challenging scenarios.

In this paper, we aim to comprehensively enhance the existing end-to-end 3D MOT framework to achieve robust and

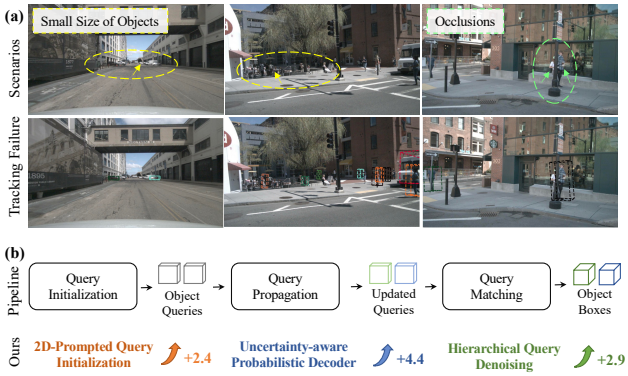

Figure 1: **(a)** In complex driving scenarios, there are various challenge factors, e.g., the occlusions and small size of target objects, which present significant challenges to achieving accurate tracking. The previous state-of-the-art end-to-end tracker, PF-Track (Pang et al., 2023), fails to track objects in certain complex scenarios. **(b)** Our S2-Track proposes three simple yet strong modules to enhance baseline comprehensively, leading to improved tracking performance.

accurate tracking results in complex driving environments. As illustrated in Figure 1 (b), we first delve into the current query-based framework and decompose it into three constituent parts: query initialization, query propagation, and query matching. Then we propose corresponding improvements, which lead to a **S**trong yet **S**imple tracker: S2-Track. Firstly, for *query initialization*, we present the 2D-Prompted Query Initialization module, which leverages predicted 2D object location and depth information to enhance the accuracy of initial object localization, leading to more reliable tracking results. Secondly, for *query propagation*, we introduce an Uncertainty-aware Probabilistic Decoder to capture and model the uncertainty of complex environments during object prediction. Specifically, we model attention scores as Gaussian distributions instead of deterministic outputs, to quantify the predictive uncertainty. Moreover, for *query matching*, we propose an Hierarchical Query Denoising strategy to further improve the training process. During the training stage, we add noises to ground-truth bounding boxes to form noised queries and selectively denoise queries based on their noised levels, enhancing robustness and convergence. Experimental results on the nuScenes benchmark demonstrate the effectiveness of our S2-Track framework. It achieves state-of-the-art performance with an impressive 66.3% AMOTA on the test split, surpassing the previous best end-to-end solution by a significant margin of 8.9% AMOTA. These results highlight our simple yet non-trivial improvements and showcase the potential of our framework in advancing the field of autonomous driving perception.

To summarize, our contributions are as follows:

- We delve into the current end-to-end 3D MOT framework and decompose it into three constitute modules, and propose a stronger yet simple framework, S2-Track, which enhances each module comprehensively.
- We propose three well-designed modules, 2D-Prompted Query Initialization, Uncertainty-aware Probabilistic Decoder, and Hierarchical Query Denoising to enhance the previous pipeline from multiple aspects, leading to improved tracking performance.
- We demonstrate the effectiveness of our S2-Track framework quantitatively and qualitatively through extensive experiments on nuScenes benchmark and achieve leading performance with a remarkable 66.3% AMOTA.

2. Related Work

Tracking by Detection. Multi-object tracking (MOT) in 3D scenes takes multi-view images from surrounding cameras or LiDAR point clouds to track multiple objects across frames (Marinello et al., 2022; Wang et al., 2023b; Qing et al., 2023; Fischer et al., 2022; Yang et al., 2022; Li et al., 2023a; Wang et al., 2023a; Sadjadpour et al., 2023; Guo et al., 2024). Taking advances in 3D object detection (Huang et al., 2021; Li et al., 2023c; 2022b; Liang et al., 2022; Liu et al., 2023b;a; Lin et al., 2023; Wang et al., 2022), most 3D MOT methods follow the *tracking by detection* paradigm (Yang et al., 2022; Chaabane et al., 2021; Zhou et al., 2020; Zhang et al., 2022b; Yin et al., 2021; Pang et al., 2022), where tracking is treated as a post-processing step after object detection. Take the detected objects at each frame, traditional 3D MOT usually uses motion models, e.g., Kalman filter (Weng et al., 2020; Wojke et al., 2017), to predict the status of corresponding trajectory and associate the candidate detections using 3D IoU (Pang et al., 2022; Weng et al., 2020) or L2 distance (Zhou et al., 2020; Yin et al., 2021).

Tracking with Query. To overcome the independent nature of the detection and tracking and to implicitly solve the association between frames, the recent *tracking with query* paradigm models the tracking process with transformer queries (Zeng et al., 2022; Zhang et al., 2022a; Doll et al., 2023; Li et al., 2023d; Pang et al., 2023; Ding et al., 2024). MUTR3D (Zhang et al., 2022a) extends the object detection method DETR3D (Wang et al., 2022) for tracking by utilizing a 3D track query to jointly model object features across timestamps and multi-view. STAR-TRACK (Doll et al., 2023) proposes a latent motion model to account for the effects of ego and object motion on the latent appearance representation. DQTrack (Li et al., 2023d) separates object and trajectory representation using decoupled queries, allowing more accurate end-to-end 3D tracking.

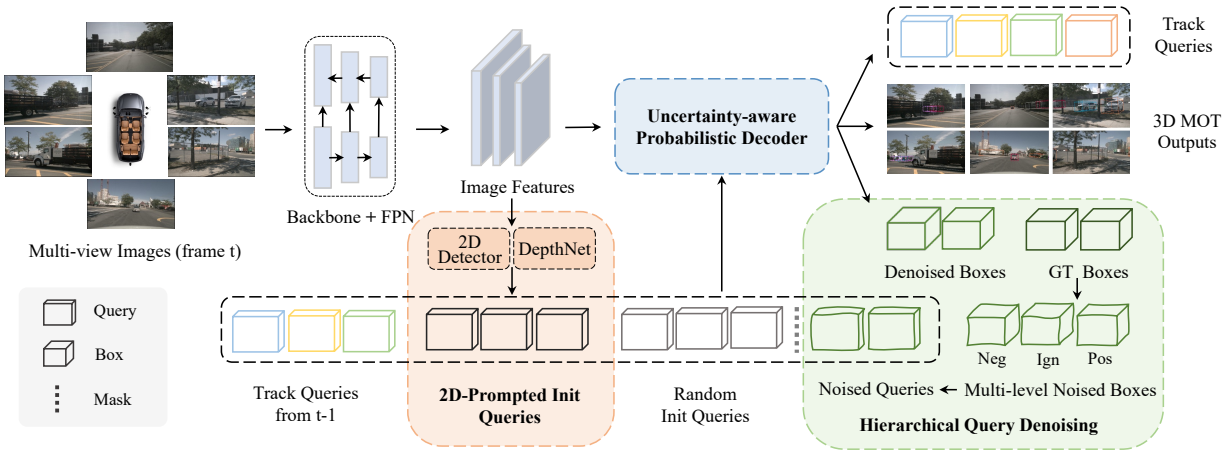

Figure 2: **S2-Track framework**. The proposed 2D-Prompted Query Initialization (PQI), Uncertainty-aware Probabilistic Decoder (UPD), and Hierarchical Query Denoising (HQD) are incorporated together to improve tracking performance. Neg: negative, Ign: ignore, Pos: positive, Mask: separate the normal queries and the denoising part to prevent information leakage.

PF-Track (Pang et al., 2023) extends the temporal horizon to provide a strong spatio-temporal object representation. Although these methods achieved impressive performance, when applied to complex scenarios, especially the occlusions and the small target objects, the tracking performance becomes unsatisfactory. In this work, we delve into the current end-to-end 3D MOT framework and decompose it into three modules, and further propose corresponding improvements to enhance each module comprehensively. As a result, these improvements lead to a strong yet simple framework, S2-Track, and consequently achieve leading performance.

3. S2-Track

In this section, we introduce our S2-Track in detail. We first give a brief problem definition and an overview of the framework in Figure 2. Then, we clarify our key contributions: 2D-Prompted Query Initialization, Uncertainty-aware Probabilistic Decoder, and Hierarchical Query Denoising.

3.1. Preliminaries

Overview. At each timestamp t, given c images from surrounding cameras, the tracking objective is to estimate a set of bounding boxes $\mathbf{b}_t^{id} \in \mathbf{B}_t$ with consistent id across frames. Under the *tracking with query* paradigm, an overview of proposed S2-Track is presented in Figure 2, which is conceptually simple: encoder and transformer decoder are adopted to encode input images and decode 3D MOT outputs with queries. From the perspective of the query, it can be divided into three stages: query initialization, query propagation with image features, and query matching with ground truth.

Queries initialization. Specifically, following previous studies (Pang et al., 2023; Ding et al., 2024), our S2-Track utilizes a set of object queries to tackle multi-object tracking from multi-view images. Each query $\mathbf{q}_t^i \in \mathbf{Q}_t$ represents a unique 3D object with a feature vector $\mathbf{f}_t^i$ and a 3D location $\mathbf{c}_t^i$, i.e., $\mathbf{q}_t^i = \{\mathbf{f}_t^i, \mathbf{c}_t^i\}$. The queries are randomly initialized as learnable embeddings, denoted as $\mathbf{Q}_{init}$. Here, we propose 2D-Prompted Query Initialization (orange module) to improve the query initialization with predicted 2D object location and depth information.

Queries propagation. During training, the object is tracked by updating its unique query. Specifically, the object queries $\mathbf{Q}_t = \{\mathbf{q}_t^i\}$ are propagated from the previous frame $t-1$ (colored squares) and numerous initial queries (gray squares in Figure 2):

$$\mathbf{Q}_t \leftarrow \mathbf{Prop}(\mathbf{Q}_{t-1}, \mathbf{Q}_{init}). \tag{1}$$

The queries from the previous frame represent tracked instances, while numerous initial queries aim to discover new objects. Here, we introduce an Uncertainty-aware Probabilistic Decoder (blue module) to model and capture the uncertainty of complex environments in object prediction.

Queries matching. Then, to predict 3D bounding boxes, decoder-only transformer architectures such as DETR3D (Wang et al., 2022) and PETR (Liu et al., 2023a), are utilized to decode image features $\mathbf{F}_t$ with object queries:

$$\mathbf{B}_t, \mathbf{Q}_t \leftarrow \mathbf{Decoder}(\mathbf{F}_t, \mathbf{Q}_t), \tag{2}$$

where $\mathbf{B}_t$ and $\mathbf{Q}_t$ are the detected 3D bounding boxes and updated query features respectively. Here, we present an Hierarchical Query Denoising strategy (green module) to enhance the model robustness and convergence.

In the following sections, we give detailed elaboration.

3.2. 2D-Prompted Query Initialization

In *tracking with query* frameworks, high-quality initial queries are crucial for achieving rapid convergence and improving tracking precision. This becomes particularly important when dealing with complex driving scenarios, such as occlusion and small-sized target objects. In previous methods that solely rely on random query initialization, the lack of reliable initial queries often results in some undetected or occluded objects losing track. To address this, we propose 2D-**P**rompted **Q**uery **I**nitialization (PQI) module which enhances the initialization of queries using learned certain priors obtained from network training. Specifically, after utilizing the shared image backbone and the feature pyramid network (FPN) layers to extract image features from each camera, we introduce additional auxiliary tasks, i.e., the 2D detection and the depth prediction, as follows:

$$\mathbf{B}_t^{2d}, \mathbf{D}_t \leftarrow \text{Networks}_{\text{auxiliary}}(\mathbf{F}_t), \tag{3}$$

where $\mathbf{B}_t^{2d}, \mathbf{D}_t$ denotes the 2D bounding boxes and the depth respectively. The 2D detection head follows YOLOX (Ge et al., 2021). The depth network combines multiple residual blocks and is supervised with the projected LiDAR points. The optimization objectives of the two auxiliary tasks are:

$$\mathcal{L}_{PQI} = \mathcal{L}_{Det}^{2D} + \mathcal{L}_{Depth}. \tag{4}$$

Then, we estimate 3D location $\mathbf{C}_t = \{\mathbf{c}_t^i\}$ through the coordinate transformation: $\mathbf{C}_t = T_{cam}^{lidar} K^{-1} \mathbf{D}_t [u, v, 1]^T$, where T_{cam}^{lidar} and K^{-1} represent the transformation matrix from the camera coordinate system to the lidar coordinate system and the camera's intrinsic parameters respectively, and (u, v) denotes the center location of the 2D boxes. Then we initialize our object queries with the preliminary 3D location, denoted as $\mathbf{Q}_{\text{PQI-init}}$ (dark gray squares in Figure 2). We also retain the random initialization $\mathbf{Q}_{\text{init}}$ to explore missing objects. To summarize, the query initialization and propagation process in Equation (1) can be improved as follows:

$$\mathbf{Q}_t \leftarrow \mathbf{Prop}(\mathbf{Q}_{t-1}, \mathbf{Q}_{\text{PQI-init}}, \mathbf{Q}_{\text{init}}). \tag{5}$$

We also provide qualitative results in Fig. 6 of Appendix, which shows the initial queries generated by our PQI module are accurately positioned within the regions of interest for the objects, resulting in more accurate tracking results.

3.3. Uncertainty-aware Probabilistic Decoder

In complex driving scenarios, the trajectories of multiple targets exhibit substantial variations in their temporal duration and sequence. Furthermore, the target objects themselves can vary in size, ranging from large trucks to small children. These diversities present significant challenges for current end-to-end 3D MOT methods, leading to uncertainties. The uncertainty issue refers to neural networks that do not deliver certainty estimates or suffer from underconfidence, as current methods struggle to capture the noise and variations inherent in the input data. Although the uncertainty issue has been recognized in certain fields (Gawlikowski et al., 2021; Subedar et al., 2019; Wang et al., 2021), e.g., action recognition (Guo et al., 2022) and camouflaged object detection (Yang et al., 2021), it has not been discussed or explored in the context of 3D MOT. Moreover, due to the complex driving scenarios and the unique challenge of tracking tasks, previous solutions for specific domains cannot be directly applied here.

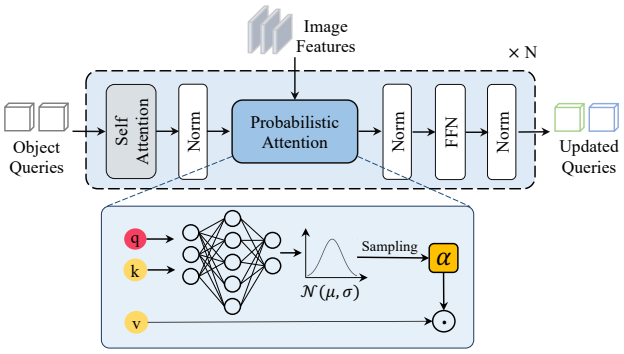

Figure 3: **Uncertainty-aware Probabilistic Decoder (UPD) architecture**. The traditional cross-attention is upgraded with probabilistic attention to quantifying the uncertainty. The probabilistic attention utilizes a multi-layer perception that takes the query q and key k as input to generate the mean and standard deviation, which are used to form a Gaussian distribution. Subsequently, the attention value α is sampled from the constructed Gaussian distribution.

To address this limitation, inspired by previous uncertainty quantify works (Guo et al., 2022; Pei et al., 2022; Blundell et al., 2015), we introduce **U**ncertainty-aware **P**robabilistic **D**ecoder (UPD) for 3D MOT. Current end-to-end methods employ decoders with conventional transformers using deterministic attention mechanisms, which compute determinate attention α between queries (Q) and keys (K) as $\alpha = \frac{Q \cdot K}{\sqrt{d}}$, where d is the dimension of queries and keys. The deterministic attention limits the ability to quantify uncertainty in predictions effectively. Instead, our UPD utilizes the probabilistic attention computation, which assumes attention α follows a Gaussian distribution: $\alpha_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij})$. Through the reparameterization trick (Kingma et al., 2015): $\alpha_{ij} = \mu_{ij} + \sigma_{ij}\epsilon, \epsilon \sim \mathcal{N}(0, 1)$. As illustrated in Figure 3, a multi-layer perception is adopted to fit the mean and standard deviation with q_i, k_j as input: $\mu_{ij}, \sigma_{ij}^2 = MLP(q_i, k_j)$. Thus, we introduce the uncertainty-aware probabilistic parameters μ_{ij} and σ_{ij} into the decoder, allowing uncertainty adaptation in the training

process. Practically, we utilize the negative log-likelihood loss to constrain the probabilistic attention as:

$$\mathcal{L}_{UPD} = \sum_{i,j} \log\left(\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}\right) \exp\left(-\frac{(\alpha_{ij} - \frac{q_i k_j}{\sqrt{d}})^2}{2\sigma_{ij}^2}\right). \quad (6)$$

As a summary, our UPD module captures and models the uncertainty of complex driving environments by representing attention scores as Gaussian distributions, which are more robust and capable of handling variations and noise in the 3D MOT, and the Equation (2) can be improved as:

$$\mathbf{B}_t, \mathbf{Q}_t \leftarrow \mathbf{Decoder}_{\text{UPD}}(\mathbf{F}_t, \mathbf{Q}_t). \quad (7)$$

3.4. Hierarchical Query Denoising

In complex 3D MOT scenarios, challenges such as occlusions and varying object sizes can hinder the learning and convergence of query-based methods. The slow convergence and suboptimal results from the instability of bipartite graph matching. To address this challenge, we draw inspiration from DN-DETR (Li et al., 2022a) and propose **H**ierarchical **Q**uery **D**enoising (HQD) training strategy, which incorporates query denoising to enhance training process for stable optimization. We perturb GT bounding boxes with noises into the decoder and train the model to reconstruct the original boxes, which effectively reduces graph matching difficulty and leads to faster convergence.

Specifically, we start by perturbing the ground truth boxes to generate noised queries. To enhance the model's ability to handle various complex driving scenarios, we define hierarchical challenging levels for the perturbed queries. We set lower and upper bound thresholds, denoted as β_{lower} and β_{upper} respectively, to help categorize the noised queries into three classes based on their challenging levels. We identified positive samples ("Pos" in Figure 2), i.e., low-challenging samples, when the 3D Intersection over Union (IoU) between a noised query and its corresponding ground truth exceeds the β_{upper} threshold, i.e., $IoU_{\mathbf{q}_t^i} > \beta_{\text{upper}}$. Conversely, negative samples ("Neg" in Figure 2), i.e., high-challenging samples, are defined when the 3D IoU falls below the β_{lower} threshold, i.e., $IoU_{\mathbf{q}_t^i} < \beta_{\text{lower}}$. Intermediate IoU values are disregarded ("Ign" in Figure 2), as they do not provide clear indications of any challenging factors, and can disrupt the normal query learning process as demonstrated in Table 5. The resulting set of noised queries that meet these requirements is denoted as $\mathbf{Q}_{\text{HQD-noised}}$, and then the decoder process can be expanded as:

$$\mathbf{B}_t^{HQD}, \mathbf{Q}_t^{HQD} \leftarrow \mathbf{Decoder}(\mathbf{F}_t, \mathbf{Q}_{\text{HQD-noised}}). \quad (8)$$

For optimization, the loss for positive samples and negative samples are calculated to form the target:

$$\mathcal{L}_{HQD} = \mathcal{L}_{box}^{pos} + \mathcal{L}_{cls}^{pos} + \mathcal{L}_{cls}^{neg}, \quad (9)$$

where $\mathcal{L}_{cls}^{pos}$ and $\mathcal{L}_{box}^{pos}$ are respectively focal loss (Lin et al., 2017) and L1 loss for the classification and box loss of $\mathbf{B}_t^U$, while $\mathcal{L}_{cls}^{neg}$ is focal loss to distinguish the background.

By incorporating the query denoising and handling the noised queries based on their noise/challenging levels, our HQD module enhances the training robustness and convergence, leading to more stable and accurate results.

3.5. Overall Optimization

To summarize, the overall optimization target of our S2-Track is formulated as:

$$\begin{aligned}\mathcal{L} =& \lambda_{tracking}\mathcal{L}_{tracking} + \lambda_{PQI}\mathcal{L}_{PQI} \\ &+ \lambda_{UPD}\mathcal{L}_{UPD} + \lambda_{HQD}\mathcal{L}_{HQD},\end{aligned} \quad (10)$$

where $\mathcal{L}_{tracking} = \mathcal{L}_{cls} + \mathcal{L}_{box}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{box}$ are classification loss and box loss for the tracked objects, $\mathcal{L}_{PQI}$, $\mathcal{L}_{UPD}$, and $\mathcal{L}_{HQD}$ are defined in Equation (4), Equation (6), and Equation (9), and λ indicate the weight balance coefficients that are all set to 1.0 by default.

4. Experiment

4.1. Experimental Setup

We conduct experiments on the large-scale nuScenes benchmark (Caesar et al., 2020) and follow the official evaluation metrics from nuScenes. Detailed Metric and Implementation are present in the Appendix.

4.2. State-of-the-art Comparison

Tracking on nuScenes val set. In Table 1, we compare our S2-Track with state-of-the-art methods on nuScenes val set. First, our method significantly outperforms existing algorithms across all tracking metrics, whether they are end-to-end or non-end-to-end methods. Specifically, S2-Track achieves impressive performance with 65.2% AMOTA and 0.924 AMOTP. When compared with the previous query-based tracker Sparse4D-v3, the performance gap is further enlarged to 8.5% AMOTA. Second, we validate the generality of S2-Track by applying different encoder backbones, i.e., V2-99 and ViT. Equipped with V2-99, the proposed framework achieves consistent gains with 8.7% AMOTA (S2-Track-F 0.566 vs PF-Track-F 0.479 vs ADA-Track 0.479). Moreover, when employing the larger ViT-L as our backbone, we further achieve leading performance.

Tracking on nuScenes test set. In Table 2, we compare our S2-Track with state-of-the-art camera-based methods on nuScenes test set. Our proposed S2-Track maintains an end-to-end tracking pipeline without heuristic post-processing and achieves leading performance with 66.3% AMOTA, surpassing the previous best solution Sparse4D-v3 by a significant margin of 8.9% AMOTA.

Table 1: **Comparisons with previous methods on the nuScenes *val* set**. Our S2-Track outperforms all existing camera-based 3D MOT methods in all metrics.

Method	Backbone	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	MOTP ↓	IDS↓
DEFTT [arxiv21] (Chaabane et al., 2021)	DLA-34	0.201	–	–	0.171	–	–
QD-3DT [TPAMI2022] (Hu et al., 2022)	DLA-34	0.242	1.518	39.9%	0.218	–	5646
MSGLMB[ICCAIS24] (Ma et al., 2024)	–	0.382	1.235	55.6%	0.399	–	2929
CC-3DT [CoRL2022] (Fischer et al., 2022)	R101	0.429	1.257	53.4%	0.385	–	2219
Cyclic [IJCV24] (Guo et al., 2024)	R101	0.469	1.002	45.7%	0.354	–	3613
QTrack [arxiv23] (Yang et al., 2022)	V2-99	0.511	1.090	58.5%	0.465	–	1144
RockTrack [arxiv24] (Li et al., 2024)	–	0.514	1.144	–	–	–	–
Tracking with Query							
MUTR3D [CVPR2022] (Zhang et al., 2022a)	R101	0.294	1.498	42.7%	0.267	0.709	3822
STAR-TRACK [RA-L2023] (Doll et al., 2023)	R101	0.379	1.358	50.1%	0.360	–	372
Sparse4D-v3 [arxiv23] (Lin et al., 2023)	R101	0.567	1.027	65.8%	0.515	0.621	557
DQTrack [ICCV23] (Li et al., 2023d)	V2-99	0.446	1.251	62.2%	–	–	1193
PF-Track-S [CVPR2023] (Pang et al., 2023)	V2-99	0.408	1.343	50.7%	0.376	–	166
HSTrack [arxiv24] (Lin et al., 2024)	V2-99	0.464	1.262	56.9%	0.423	–	204
PF-Track-F [CVPR2023] (Pang et al., 2023)	V2-99	0.479	1.227	59.0%	0.435	–	181
ADA-Track [CVPR2024] (Ding et al., 2024)	V2-99	0.479	1.246	60.2%	0.430	–	767
S2-Track-F (Ours)	V2-99	0.566	1.090	62.2%	0.498	0.657	174
S2-Track-S (Ours)	ViT-L	0.600	1.020	68.8%	0.538	0.614	167
S2-Track-F (Ours)	ViT-L	**0.652**	**0.924**	**72.2%**	**0.574**	**0.577**	**134**

"**S**" and "**F**" represent the settings of small-resolution and full-resolution respectively.

4.3. Analysis Study

Analysis and ablations of the proposed modules of S2-Track. In Table 3, we validate our proposed modules for our model's performance on nuScenes val set. It is clear that incorporating each module leads to performance gain in tracking. Specifically, the Uncertainty-aware Probabilistic Decoder (UPD) module significantly improves the baseline with 4.4% AMOTA, and the 2D-Prompted Query Initialization (PQI) and Hierarchical Query Denoising (HQD) modules obtained consist boost with 2.4% and 2.9% AMOTA respectively. Moreover, combining the three modules leads to further improvements.

Analysis of uncertainty. Quantifying uncertainty can be challenging for transformer-based models which do not inherently provide uncertainty estimates like Bayesian methods. Thus, we can only employ computationally intensive strategies to approximate uncertainty, such as the Monte Carlo Dropout (MC dropout) (Gal & Ghahramani, 2016) and Ensemble strategy (Ens.) (Lakshminarayanan et al., 2017). As shown in Table 3, the proposed UPD module successfully reduces uncertainty. Surprisingly, other modules also effectively reduce uncertainty, even though they were not designed to aim at uncertainty.

Analysis of complex situations. In Table 4, we analyze the performance of S2-Track under different driving conditions, i.e., different visibilities, different object sizes, and different distances. Our S2-Track achieves consistent improvements over PF-Track (Pang et al., 2023) under all visibility, object size, and distance settings. Furthermore, S2-Track brings larger improvements in more challenging situations, such as lower visibilities (+ 6.3% AMOTA), smaller objects (+ 5.5% AMOTA), and more distant objects (+ 4.1% AMOTA), both of which demonstrate our effectiveness in addressing diverse challenges in 3D MOT.

4.4. Ablation Study

4.4.1. Ablations on thresholds of HQD module.

In Table 5, we validate various settings of the threshold of HQD module. (1) In the first setting, we do not consider the noise thresholds, meaning that all noised queries undergo box optimization. (2) In the second setting, we do not selectively denoise queries based on hierarchical challenging levels; instead, a single threshold is used to classify all noised queries into positive and negative samples for optimization. (3) In our HQD module, we selectively denoise queries based on hierarchical challenging levels. The improved performance demonstrates our HQD, which optimizes samples based on different challenging levels, effectively handles various complex driving environments, and achieves superior results. We also investigate varying lower and upper bound thresholds, i.e., β_{lower} and β_{upper}.

Table 2: **Comparisons with state-of-the-art camera-based methods on the nuScenes *test* set.**

Method	Backbone	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	MOTP ↓	IDS↓
PermaTrack [ICCV2021] (Tokmakov et al., 2021)	DLA-34	0.066	1.491	18.9%	0.060	0.724	3598
DEFT [arxiv21] (Chaabane et al., 2021)	DLA-34	0.177	1.564	33.8%	0.156	0.770	6901
QD-3DT [TPAMI2022] (Hu et al., 2022)	DLA-34	0.217	1.550	37.5%	0.198	0.773	6856
CC-3DT [CoRL2022] (Fischer et al., 2022)	R101	0.410	1.274	53.8%	0.357	0.676	3334
Cyclic [IJCV24] (Guo et al., 2024)	R101	0.433	1.055	49.2%	0.334	–	6621
QTrack [arxiv23] (Yang et al., 2022)	V2-99	0.480	1.100	58.3%	0.431	0.597	1484
Tracking with Query							
MUTR3D [CVPR2022] (Zhang et al., 2022a)	R101	0.270	1.494	41.1%	0.245	0.709	6018
PF-Track-F [CVPR2023] (Pang et al., 2023)	V2-99	0.434	1.252	53.8%	0.378	0.674	**249**
STAR-TRACK [RA-L2023] (Doll et al., 2023)	V2-99	0.439	1.256	56.2%	0.406	0.664	607
DQTack [ICCV23] (Li et al., 2023d)	V2-99	0.523	1.096	62.2%	0.444	0.649	1204
Sparse4D-v3 [arxiv23] (Lin et al., 2023)	V2-99	0.574	0.970	66.9%	0.521	**0.525**	669
ADA-Track [CVPR2024] (Ding et al., 2024)	V2-99	0.456	1.237	55.9%	0.406	–	834
S2-Track-F (Ours)	V2-99	0.608	0.925	**75.8%**	0.547	0.559	963
S2-Track-F (Ours)	ViT-L	**0.663**	**0.815**	72.3%	**0.554**	0.530	844

"**F**" represent on the full-resolution settings.

Table 3: **Analysis and ablations on the proposed modules of S2-Track.**

Method			Tracking						Uncertainty		
UPD	PQI	HQD	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	MOTP↓	IDS↓		s↓	σ↓
✗	✗	✗	0.394	1.363	51.4%	0.372	0.753	178	MC Dropout	1.99	0.108
✓	✗	✗	0.438	1.261	55.9%	0.419	0.681	175	MC Dropout	1.86	0.085
✗	✓	✗	0.418	1.251	54.6%	0.394	0.667	177	MC Dropout	1.91	0.097
✗	✗	✓	0.423	1.264	55.6%	0.397	0.671	183	MC Dropout	1.88	0.093
✓	✓	✓	**0.458**	**1.230**	**56.6%**	**0.433**	**0.664**	**172**	MC Dropout	**1.81**	**0.078**
PF-Track-S			0.408	1.343	50.7%	0.376	–	166	Ens.	1.96	0.100
S2-Track-S			**0.458**	**1.230**	**56.6%**	**0.433**	**0.664**	**172**	Ens.	**1.75**	**0.072**

s and σ donate entropy and standard deviation of uncertainty quantification, respectively.

The model achieves the best result when β_{lower} is set to 0.30 and β_{upper} is set to 0.70.

4.4.2. Ablations on network strides of PQI module.

In Table 6, we conduct an ablation study to analyze the effects of the network stride of PQI module. As described in Equation (3), our PQI module incorporates two additional auxiliary tasks based on the extracted feature $\mathbf{F}_t$. To reduce computational overhead, we directly reuse features from different layers with varying strides in the feature pyramid network, thereby alleviating the burden on the auxiliary task heads. We present the performance achieved with different strides for the image features. The experimental results reveal that the model performs optimally when using a stride of 16.

4.4.3. Ablations on different decoders.

In Table 7, we validate the generality of S2-Track by applying different decoders. Specifically, we employ two popular decoders: PETR (Liu et al., 2022) and DETR3D (Wang et al., 2022), which are widely adopted by query-based tracking methods. Our approach achieves excellent results with both decoders, yielding comparable performance. The results obtained with DETR3D are slightly higher, leading us to select DETR3D as our default decoder.

4.5. Qualitative Comparison

In Figure 4, we provide the qualitative results on nuScenes dataset. We compare our S2-Track with the previous state-of-the-art end-to-end tracker, PF-Track (Pang et al., 2023). In Figure 4 (a), PF-Track exhibits commendable tracking accuracy when the line of sight is clear at time t_i. However, as the occlusion gradually intensifies, the accumulated er-

Table 4: **Analysis of complex situations.**

Visibilities	0-40%	40%-60%	60%-100%
PF-Track	38.3	38.6	39.5
S2-Track	44.6 (+6.3)	44.8 (+6.2)	45.2 (+5.7)
Size	0-2m	2-3.5m	>3.5m
PF-Track	22.8	20.4	19.1
S2-Track	28.3 (+5.5)	25.5 (+5.1)	24.0 (+4.9)
Distance	>30m	20-30m	0-20m
PF-Track	7.5	38.9	64.6
S2-Track	11.6 (+4.1)	42.8 (+3.9)	67.0 (+2.4)

The metric is AMOTA↑.

Table 5: **Ablations on thresholds of HQD**. (1) Without thresholds, (2) without selecting by levels, (3) Our HQD.

Threshold	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	IDS↓
(1)	0.440	1.249	55.9%	0.419	181
(2)	0.445	1.239	56.5%	0.425	174
(3)	**0.458**	**1.230**	**56.6%**	**0.433**	**172**
Lower Bound Thresholds					
0.20	0.450	1.236	56.3%	0.429	179
0.30	**0.458**	**1.230**	**56.6%**	**0.433**	**172**
0.40	0.448	1.241	56.4%	0.427	173
Upper Bound Thresholds					
0.60	0.436	1.261	56.1%	0.419	204
0.65	0.455	1.243	**57.5%**	**0.434**	198
0.70	**0.458**	**1.230**	56.6%	0.433	**172**
0.75	0.444	1.253	55.9%	0.416	214

ror arising significantly escalates. Especially at t_{i+12}, the predicted bounding boxes for two pedestrians completely overlap. In contrast, our S2-Track consistently maintains a high level of tracking precision throughout the entire duration of continuous tracking.

In complex scenarios of Figure 4 (b), which are characterized by multiple complex factors such as occlusions in crowded and spacious environments, as well as the small size of vehicles and pedestrians, S2-Track achieves more precise tracking bounding boxes and successfully recognizes more tracked objects compared to PF-Track (Pang et al., 2023). Furthermore, the visualization of our attention scores demonstrates a higher concentration on the object center, indicating that our model pays more attention to the target objects. In contrast, PF-Track (Pang et al., 2023) exhibits low attention scores for challenging samples, failing to capture and track objects. The outstanding 3D MOT results of S2-Track demonstrate the effectiveness of our framework, which addresses the challenges across various complex tracking scenarios.

Table 6: **Ablations on network strides of PQI module**.

Stride	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	IDS↓
8	0.456	1.245	56.1%	0.422	178
16	**0.458**	**1.230**	**56.6%**	**0.433**	**172**
32	0.446	1.268	56.2%	0.419	214

Table 7: **Ablations on different decoders**.

Decoder	AMOTA↑	AMOTP↓	RECALL↑	MOTA↑	IDS↓
PETR	0.452	1.246	**57.1%**	0.429	196
DETR3D	**0.458**	**1.230**	56.6%	**0.433**	**172**

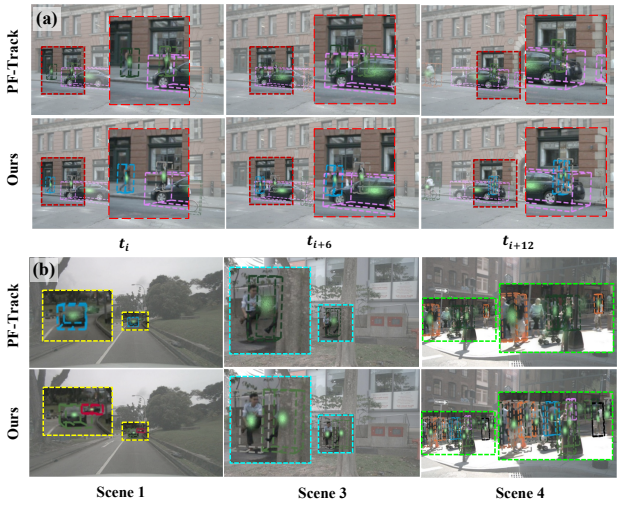

Figure 4: **Qualitative results on the nuScenes dataset**. **(a)** The tracking results for an occlusion scenario of two pedestrians of consecutive frames ($t_i - t_{i+12}$). **(b)** The tracking results on several challenging tracking scenes. Moreover, we plot the attention scores of object queries, which indicate how strongly the model focuses on the target objects. A higher concentration of color represents a higher attention score and a stronger confidence in the corresponding object.

5. Conclusion

In this paper, we improve the current end-to-end 3D MOT framework from multiple aspects. Specifically, for query initialization, we propose 2D-Prompted Query Initialization to improve object localization accuracy by incorporating predicted 2D location and depth cues. For query propagation, the Uncertainty-aware Probabilistic Decoder models the uncertainty of complex driving situations using probabilistic attention, providing a comprehensive understanding of predictive uncertainty. Then, for query matching, the Hierarchical Query Denoising strategy enhances training robustness and convergence. Experimental results on nuScenes benchmark demonstrate that S2-Track achieves state-of-the-art performance, i.e., 66.3% AMOTA on the test split, with a significant improvement of 8.9% AMOTA.

Acknowledgments

This work is supported by Scientific Research Innovation Capability Support Project for Young Faculty (No.ZYGXQNJSKYCXNLZCXM-I28), National Natural Science Foundation of China (NSFC) under Grants No.62476293, Shenzhen Science and Technology Program No.GJHZ20220913142600001, Nansha Key R&D Program under Grant No.2022ZD014, and General Embodied AI Center of Sun Yat-sen University.

Impact Statement

This paper presents a strong yet simple tracker, S2-Track, for 3D multiple object tracking. Since the tracking explored in this paper is for generic objects and does not pertain to specific human recognition, so we do not see potential privacy-related issues. At present, we primarily focus on the image representation of scenes. However, our framework can be expanded to incorporate other sensors, such as a fusion of LiDAR and cameras. We leave the extension of our method towards building such systems for future work. We hope that our work can inspire more future research in this field.

References

Bernardin, K. and Stiefelhagen, R. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10, 2008.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and 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.

Chaabane, M., Zhang, P., Beveridge, J. R., and O'Hara, S. Deft: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021.

Ding, S., Schneider, L., Cordts, M., and Gall, J. Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Doll, S., Hanselmann, N., Schneider, L., Schulz, R., Enzweiler, M., and Lensch, H. P. Star-track: Latent motion models for end-to-end 3d object tracking with adaptive spatio-temporal appearance representations. *IEEE Robotics and Automation Letters*, 2023.

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 preprint arXiv:2010.11929*, 2020.

Feng, C., Jie, Z., Zhong, Y., Chu, X., and Ma, L. Aedet: Azimuth-invariant multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21580–21588, 2023.

Fischer, T., Yang, Y.-H., Kumar, S., Sun, M., and Yu, F. Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. In *6th Annual Conference on Robot Learning*, 2022.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Guo, H., Wang, H., and Ji, Q. Uncertainty-guided probabilistic transformer for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20052–20061, 2022.

Guo, M., Zhang, Z., Jing, L., He, Y., Wang, K., and Fan, H. Cyclic refiner: Object-aware temporal representation learning for multi-view 3d detection and tracking. *International Journal of Computer Vision*, pp. 1–23, 2024.

Hu, H.-N., Yang, Y.-H., Fischer, T., Darrell, T., Yu, F., and Sun, M. Monocular quasi-dense 3d object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1992–2008, 2022.

Huang, J. and Huang, G. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

Huang, J., Huang, G., Zhu, Z., Ye, Y., and Du, D. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Jiang, X., Li, S., Liu, Y., Wang, S., Jia, F., Wang, T., Han, L., and Zhang, X. Far3d: Expanding the horizon

for surround-view 3d object detection. *arXiv preprint arXiv:2308.09616*, 2023.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lee, Y., Hwang, J.-w., Lee, S., Bae, Y., and Park, J. An energy and gpu-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

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

Li, X., Xie, T., Liu, D., Gao, J., Dai, K., Jiang, Z., Zhao, L., and Wang, K. Poly-mot: A polyhedral framework for 3d multi-object tracking. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9391–9398. IEEE, 2023a.

Li, X., Li, P., Zhao, L., Liu, D., Gao, J., Wu, X., Wu, Y., and Cui, D. Rocktrack: A 3d robust multi-camera-ken multi-object tracking framework. *arXiv preprint arXiv:2409.11749*, 2024.

Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., and Li, Z. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1486–1494, 2023b.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., and Li, Z. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023c.

Li, Y., Yu, Z., Philion, J., Anandkumar, A., Fidler, S., Jia, J., and Alvarez, J. End-to-end 3d tracking with decoupled queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18302–18311, 2023d.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., and Dai, J. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pp. 1–18. Springer, 2022b.

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., and Tang, Z. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

Lin, S., Kou, Y., Li, B., Hu, W., and Gao, J. Hstrack: Bootstrap end-to-end multi-camera 3d multi-object tracking with hybrid supervision. *arXiv preprint arXiv:2411.06780*, 2024.

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

Lin, X., Pei, Z., Lin, T., Huang, L., and Su, Z. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023.

Liu, Y., Wang, T., Zhang, X., and Sun, J. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pp. 531–548. Springer, 2022.

Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., and Zhang, X. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3262–3272, 2023a.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., and Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. IEEE, 2023b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ma, L. V., Hussain, M. I., Yow, K.-C., and Jeon, M. 3d multi-object tracking employing ms-glmb filter for autonomous driving. In *2024 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pp. ? IEEE, 2024.

Marinello, N., Proesmans, M., and Van Gool, L. Tripletrack: 3d object tracking using triplet embeddings and lstm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4500–4510, 2022.

Pang, Z., Li, Z., and Wang, N. Simpletrack: Understanding and rethinking 3d multi-object tracking. In *European Conference on Computer Vision*, pp. 680–696. Springer, 2022.

Pang, Z., Li, J., Tokmakov, P., Chen, D., Zagoruyko, S., and Wang, Y.-X. Standing between past and future:

Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17928–17938, 2023.

Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., and Zhan, W. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022.

Pei, J., Wang, C., and Szarvas, G. Transformer uncertainty estimation with hierarchical stochastic attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11147–11155, 2022.

Qing, L., Wang, T., Lin, D., and Pang, J. Dort: Modeling dynamic objects in recurrent for multi-camera 3d object detection and tracking. In *Conference on Robot Learning*, pp. 3749–3765. PMLR, 2023.

Sadjadpour, T., Li, J., Ambrus, R., and Bohg, J. Shasta: Modeling shape and spatio-temporal affinities for 3d multi-object tracking. *IEEE Robotics and Automation Letters*, 2023.

Shu, C., Deng, J., Yu, F., and Liu, Y. 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3580–3589, 2023.

Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O., and Huang, J. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6301–6310, 2019.

Tokmakov, P., Li, J., Burgard, W., and Gaidon, A. Learning to track with object permanence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10860–10869, 2021.

Wang, L., Zhang, X., Qin, W., Li, X., Gao, J., Yang, L., Li, Z., Li, J., Zhu, L., Wang, H., et al. Camo-mot: Combined appearance-motion optimization for 3d multi-object tracking with camera-lidar fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2023a.

Wang, S., Liu, Y., Wang, T., Li, Y., and Zhang, X. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3621–3631, October 2023b.

Wang, Y., Guizilini, V. C., Zhang, T., Wang, Y., Zhao, H., and Solomon, J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pp. 180–191. PMLR, 2022.

Wang, Z., Li, Y., Guo, Y., Fang, L., and Wang, S. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4568–4577, 2021.

Weng, X. and Kitani, K. A baseline for 3d multi-object tracking. *arXiv preprint arXiv:1907.03961*, 1(2):6, 2019.

Weng, X., Wang, J., Held, D., and Kitani, K. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10359–10366. IEEE, 2020.

Wojke, N., Bewley, A., and Paulus, D. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649. IEEE, 2017.

Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., and Fan, D.-P. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4146–4155, 2021.

Yang, J., Yu, E., Li, Z., Li, X., and Tao, W. Quality matters: Embracing quality clues for robust 3d multi-object tracking. *arXiv preprint arXiv:2208.10976*, 2022.

Yin, T., Zhou, X., and Krahenbuhl, P. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pp. 659–675. Springer, 2022.

Zhang, T., Chen, X., Wang, Y., Wang, Y., and Zhao, H. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2022a.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2022b.

Zhou, X., Koltun, V., and Krähenbühl, P. Tracking objects as points. In *European conference on computer vision*, pp. 474–490. Springer, 2020.

Zong, Z., Jiang, D., Song, G., Xue, Z., Su, J., Li, H., and Liu, Y. Temporal enhanced training of multi-view 3d object detector via historical object prediction. *arXiv preprint arXiv:2304.00967*, 2023.

Appendix

A. Additional Details

A.1. Depthnet Details

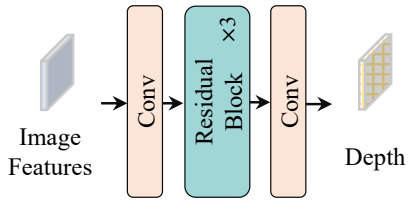

Figure 5: **Details of the depth network in PQI.**

The depth network of the proposed PQI module is composed of multiple residual blocks, and we provide an illustration in Figure 5.

A.2. Experimental Setup

A.2.1. Dataset.

We conduct experiments on the popular nuScenes benchmark (Caesar et al., 2020), which is a large-scale autonomous-driving dataset for 3D detection and tracking, consisting of 700, 150, and 150 scenes for training, validation, and testing, respectively. Each frame contains one point cloud and six calibrated images from the surrounding cameras with a full 360-degree field of view. It provides 3D tracking bounding boxes from 7 categories for the tracking task.

A.2.2. Metrics.

We follow the official evaluation metrics from nuScenes. For the 3D tracking task, we report Average Multi-object Tracking Accuracy (AMOTA) (Weng & Kitani, 2019), Average Multi-object Tracking Precision (AMOTP), and the modified CLEAR MOT metrics (Bernardin & Stiefelhagen, 2008), e.g., MOTA, MOTP, and IDS. For a detailed understanding, please refer to (Caesar et al., 2020; Bernardin & Stiefelhagen, 2008).

A.2.3. Implementation details.

In this paper, we assess the generalization capability of our S2-Track through experiments using different encoders, e.g., V2-99 (Lee et al., 2019) and ViT (Dosovitskiy et al., 2020), and different decoders e.g., PETR (Liu et al., 2022) and DETR3D (Wang et al., 2022). All experiments are conducted on 8 NVIDIA A100-80GB GPUs. For each training sample, it contains three consecutive adjacent frames each with contains six surrounding images, and we use a fixed number of 500 initial queries for each sample. We adopt the AdamW optimizer (Loshchilov & Hutter, 2017) for network training, with the initial learning rate setting of 0.01 and the cosine weight decay set to 0.001. By default, the thresholds β_{lower} and β_{upper} are set to 0.3 and 0.7, and the weight coefficients λ that are all set to 1.0, respectively.

Due to the limited computation resources, we follow PF-Track (Pang et al., 2023) to apply two resolution settings, full-resolution and small-resolution. For full-resolution ("-F"), we crop the origin 1600×900 image to 1600×640. For small-resolution ("-S"), we scale down the cropped image to 800 x 320 in a further step. We pre-train the image backbone with single-frame detection task for 12 epochs (small-resolution setting) and 24 epochs (full-resolution setting) respectively, and further train the end-to-end tracker with consecutive frames (set to be 3 frames) for another 12 epochs (small-resolution) and 24 epochs (full-resolution). All the ablation studies are conducted on the small-resolution setting with V2-99 backbone.

We ulitze the 3D location to initialize queries by following steps: 1) normalize input coordinates to the $[0, 2\pi]$ range; 2) generate frequency bands using exponential temperature scaling; 3) compute sine/cosine components for each dimension (X, Y, Z); 4) concatenate the encoded dimensions; 5) project the concatenated features through two linear layers with ReLU activation. We will include these implementations in the revision.

A.2.4. ANALYSIS STUDY DETAILS.

The analysis studies in Tab. 4 are conducted on the small-resolution setting with V2-99 backbone. The experiments are performed on the nuScenes val set, in which we focus on specific challenging conditions to select clips for evaluation. We utilize the attributes of the bounding boxes provided by the nuScenes dataset, e.g., the visibility labels, and then calculate the average for each clip, finally group the results according to different ranges of the attributes. The categorization process involved the following criteria: *1. Different visibilities*: the dataset is divided based on the visibility attribute of the objects. Visibility ranges are considered as 0-40%, 40-60%, and 60-100%. *2. Different object sizes*: the dataset is divided into three groups based on the average object size: objects with a size greater than or equal to 3.5 meters, objects smaller than 3.5 meters and greater than or equal to 2 meters, and objects smaller than 2 meters. *3. Different object distances*: the dataset is split based on different distance ranges, namely 0-20 meters, 20-30 meters, and 30 meters and above. By applying these categorizing and calculations, subsets of data were selected from the clips in the validation dataset to evaluate their performance based on the specified challenging conditions.

B. Additional Discussion

B.1. PQI vs. 3DPPE

Previous 3DPPE (Shu et al., 2023) also involves depth priors in a query-based framework, it differs from S2-Track in several aspects. First, 3DPPE focuses on 3D object detection, whereas we tackle 3D MOT. Second, 3DPPE introduces 3D point positional encoding, while our PQI is designed for query initialization. Moreover, we also retain randomly initialized queries to explore missing objects. We will add this discussion into the revision.

B.2. Analysis of other modules also effectively reduce the uncertainty

Our PQI module leverages learned certain priors, i.e., 2D object location and depth information, to enhance the initialization of queries, thus effectively reducing the uncertainty in query initialization and resulting in more accurate object localization and tracking. The HQD strategy introduces different levels of noise to the queries and then applies a denoising process, allowing the model to encounter varying magnitudes of noise (i.e., uncertainty) during training. This effectively helps the model reduce uncertainty during query matching, leading to more stable and accurate tracking performance. Although the motivation of these two modules is not uncertainty, they both help the model reduce uncertainty during query initialization and matching. Moreover, they are incorporated together with the UPD module, which aims to reduce uncertainty during query propagation.

C. Additional Results

C.1. Inference Latency

We present the inference latency measure in Table 8. Our proposed S2-Track, demonstrates greater efficiency compared to previous end-to-end methods, i.e., MUTR3D (Zhang et al., 2022a) and DQTrack (Li et al., 2023d). In comparison to the state-of-the-art PF-Track (Pang et al., 2023), our S2-Track introduces little additional latency, and yet this trade-off results in a 5.0% improvement in AMOTA over PF-Track. Further efficiency enhancement in tracking is a promising direction for future research.

Table 8: **Inference latency**. Frame Per Second (FPS) is evaluated on a single NVIDIA A100 GPU from input images to tracking results.

Method	FPS
PF-Track-S (Pang et al., 2023)	9.2
DQTrack (Li et al., 2023d)	6.0
MUTR3D (Zhang et al., 2022a)	6.0
S2-Track-S (Ours)	7.5

C.2. Detection Results

C.2.1. METRICS.

For the 3D detection task, we follow the official evaluation metrics from nuScenes (Caesar et al., 2020), and report nuScenes Detection Score (NDS), mean Average Prediction (mAP), and five True Positive (TP) metrics including mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribute Error (mAAE).

Table 9: **Results of 3D detection on nuScenes *val* dataset**.

Method	Backbone	mAP ↑	NDS↑	mATE↓	mASE↓	mAOE ↓	mAVE↓	mAAE ↓
Detection-only								
PETR v2 (Liu et al., 2023a)	R101	0.421	0.524	0.681	0.267	0.357	0.377	0.186
BEVDet4D (Huang & Huang, 2022)	Swin-B	0.426	0.552	0.560	0.254	0.317	0.289	0.186
Cyclic (Guo et al., 2024)	R101	0.433	0.532	0.639	0.270	0.318	0.416	0.201
BEVDepth (Li et al., 2023c)	CNX-B	0.462	0.558	0.540	0.254	0.353	0.379	0.200
AeDet (Feng et al., 2023)	CNX-B	0.483	0.581	0.494	0.261	0.324	0.337	0.195
SOLOFusion (Park et al., 2022)	R101	0.483	0.582	0.503	0.264	0.381	0.246	0.207
StreamPETR (Wang et al., 2023b)	R101	0.504	0.592	0.569	0.262	0.315	0.257	0.199
HoP (Zong et al., 2023)	R101	0.454	0.558	0.565	0.265	0.327	0.337	0.194
Far3D (Jiang et al., 2023)	R101	**0.510**	**0.594**	0.551	0.258	0.372	0.238	0.195
Join Tracking and Detection								
MUTR3D (Zhang et al., 2022a)	R101	0.349	0.434	–	–	–	–	–
DQTrack (Li et al., 2023d)	V2-99	0.410	0.503	–	–	–	–	–
PF-Track-F (Pang et al., 2023)	V2-99	0.399	0.390	0.727	0.268	1.722	0.887	0.211
HSTrack (Lin et al., 2024)	V2-99	0.418	0.510	–	–	–	–	–
Sparse4D-v3 (Lin et al., 2023)	R101	0.537	0.623	0.511	0.255	0.306	0.194	0.192
S2-Track-F (Ours)	ViT-L	**0.589**	**0.655**	0.495	0.250	0.249	0.2100	0.1883

"**F**" represent on the full-resolution settings.

Detection on nuScenes benchmark. As our S2-Track can jointly optimize tracking and detection, we present the detection results on nuScenes test set and val set in Table 10 and Table 9 respectively, which demonstrate consistent improvements of our method in the detection task. As a framework design for tracking, our model achieves comparable results (62.7% mAP and 68.0% NDS in the test set) with the concurrent leading detection methods, e.g., HoP and Far3D, and outperforms previous end-to-end detection and tracking model Sparse4D-v3 by a significant margin of 5.7% mAP and 2.4% NDS in the test set.

D. Additional Visualizations

D.1. Qualitative Results of PQI

We present qualitative results of our PQI module in Figure 6. Our PQI module leverages learned certain priors, i.e., the predicted 2D object location and depth information to formulate initial queries. As shown in Figure 6, the initial queries generated by our PQI module are accurately positioned within the regions of interest for the objects. This indicates that our module effectively bootstraps latent query states of the objects, leading to improved object localization and tracking results.

D.2. More Qualitative Results

We provide additional qualitative results in Figure 7. We compare our S2-Track with the previous state-of-the-art end-to-end tracker, PF-Track (Pang et al., 2023), on various complex scenarios. In challenging 3D MOT scenarios, characterized by multiple challenging factors such as occlusions and small target objects, S2-Track successfully predicts a greater number of tracked bounding boxes with higher localization precision compared to PF-Track (Pang et al., 2023). Furthermore, the

Table 10: **Results of 3D detection on nuScenes *test* dataset**. † indicates the results obtained from our testing using the provided official model.

Method	Backbone	mAP ↑	NDS↑	mATE↓	mASE↓	mAOE ↓	mAVE↓	mAAE ↓
Detection-only								
BEVDet4D (Huang & Huang, 2022)	Swin-B	0.451	0.569	0.511	0.241	0.386	0.301	0.121
Cyclic (Guo et al., 2024)	R101	0.452	0.549	0.575	0.255	0.405	0.407	0.131
PETR v2 (Liu et al., 2023a)	V2-99	0.506	0.592	0.536	0.243	0.359	0.349	0.120
BEVDepth (Li et al., 2023c)	CNX-B	0.520	0.609	0.445	0.243	0.352	0.347	0.127
BEVStereo (Li et al., 2023b)	V2-99	0.525	0.610	0.431	0.246	0.358	0.357	0.138
SOLOFusion (Park et al., 2022)	CNX-B	0.540	0.619	0.453	0.257	0.376	0.276	0.148
AeDet (Feng et al., 2023)	CNX-B	0.531	0.620	0.439	0.247	0.344	0.292	0.130
StreamPETR (Wang et al., 2023b)	ViT-L	0.620	0.676	0.470	0.241	0.258	0.236	0.134
HoP (Zong et al., 2023)	ViT-L	0.624	0.685	0.367	0.249	0.353	0.171	0.131
Far3D (Jiang et al., 2023)	ViT-L	**0.635**	**0.687**	0.432	0.237	0.278	0.227	0.130
Join Tracking and Detection								
PF-Track-F† (Pang et al., 2023)	V2-99	0.397	0.387	0.688	0.262	1.800	1.079	0.165
Sparse4D-v3 (Lin et al., 2023)	V2-99	0.570	0.656	0.412	0.236	0.312	0.210	0.117
S2-Track-F (Ours)	ViT-L	**0.627**	**0.680**	0.434	0.237	0.311	0.216	0.130

"**F**" represent on the full-resolution settings.

visualization of our attention scores demonstrates a higher concentration on the center of the objects, indicating that our model pays more attention to the target objects. This observation also highlights the effectiveness of our proposed UPD module with probabilistic attention in modeling the uncertainty associated with the object prediction.

D.3. Video Demo

In addition to the figures, we have also attached a video demo in the supplementary materials, which consists of hundreds of tracking frames that provide a more comprehensive evaluation of our proposed approach.

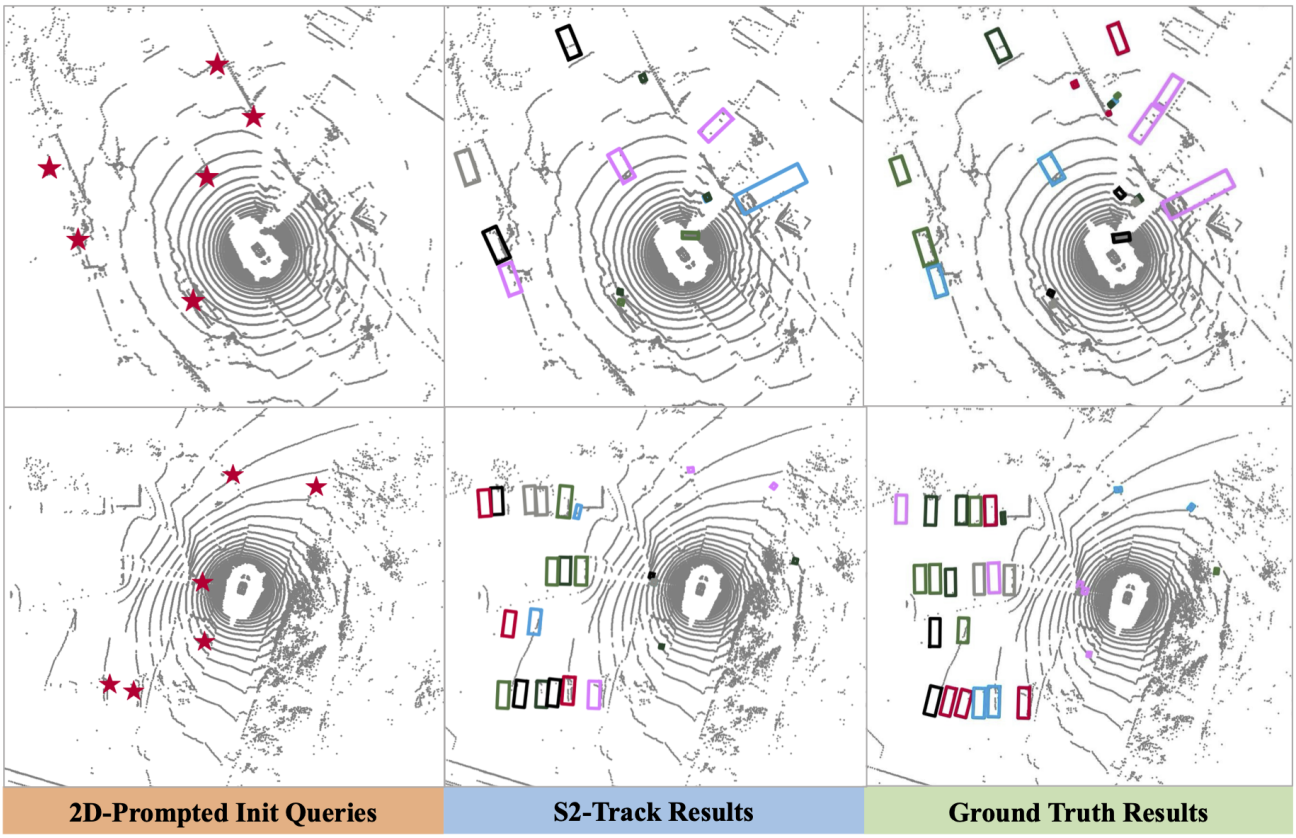

Figure 6: **Qualitative results of our PQI**. The initial queries generated by our PQI module accurately locate the regions of interest for the objects, resulting in more accurate tracking results.

Figure 7: **More qualitative results on the nuScenes dataset**. The tracking results on several challenging tracking scenarios, including the small size of the target objects and the occlusions. Moreover, we plot the attention scores of object queries, which indicate how strongly the model focuses on the target objects. A higher concentration of color represents a higher attention score and a stronger confidence in the corresponding object.