Standing Between Past and Future: Spatio-Temporal Modeling for Multi-Camera 3D Multi-Object Tracking

Ziqi Pang[1]*, Jie Li[2], Pavel Tokmakov[2], Dian Chen[2], Sergey Zagoruyko[3], Yu-Xiong Wang[1]†
University of Illinois Urbana-Champaign[1], Toyota Research Institute[2], Woven Planet Level-5[3]

Abstract

This work proposes an end-to-end multi-camera 3D multi-object tracking (MOT) framework. It emphasizes spatio-temporal continuity and integrates both past and future reasoning for tracked objects. Thus, we name it "Past-and-Future reasoning for Tracking" (PF-Track). Specifically, our method adopts the "tracking by attention" framework and represents tracked instances coherently over time with object queries. To explicitly use historical cues, our "Past Reasoning" module learns to refine the tracks and enhance the object features by cross-attending to queries from previous frames and other objects. The "Future Reasoning" module digests historical information and predicts robust future trajectories. In the case of long-term occlusions, our method maintains the object positions and enables re-association by integrating motion predictions. On the nuScenes dataset, our method improves AMOTA by a large margin and remarkably reduces ID-Switches by 90% compared to prior approaches, which is an order of magnitude less. The code and models are made available at https://github.com/TRI-ML/PF-Track.

1. Introduction

Reasoning about object trajectories in 3D is the cornerstone of autonomous navigation. While many LiDAR-based approaches exist [36, 58, 63], their applicability is limited by the cost and reliability of the sensor. Detecting, tracking, and forecasting object trajectories only with cameras is hence a critical problem. Significant progress has been achieved on these tasks separately, but they have been historically primarily studied in isolation and combined into a full-stack pipeline in an ad-hoc fashion.

In particular, 3D detection has attracted a lot of attention [20, 24, 25, 28, 53], but associating these detections over time has been mostly done independently from localization [19, 31, 43]. Recently, a few approaches to end-to-end detection and tracking have been proposed, but they operate on neighboring frames and fail to integrate longer-term spatio-temporal cues [7, 12, 33, 65]. In the prediction literature, on the other hand, it is common to assume the availability of ground truth object trajectories and HD-Maps [4, 8, 11, 59]. A few attempts for a more realistic evaluation have been made [16, 21], focusing only on the prediction performance.

In this paper, we argue that multi-object tracking can be dramatically improved by jointly optimizing the detection-tracking-prediction pipeline, especially in a camera-based system. We provide an intuitive example from our real-world experiment in Fig. 1. At first, the pedestrian is fully visible, but a model with only single-frame information makes a prediction with large deviation (frame $t = 0$ in Fig. 1). After this, integrating the temporal information from the past gradually corrects the error over time (frame $t = t_1$ in Fig. 1), by capitalizing on the notion of spatio-temporal continuity. Moreover, as the pedestrian becomes fully occluded (frame $t = t_2$ in Fig. 1), we can still predict their location by using the aggregated past informa-

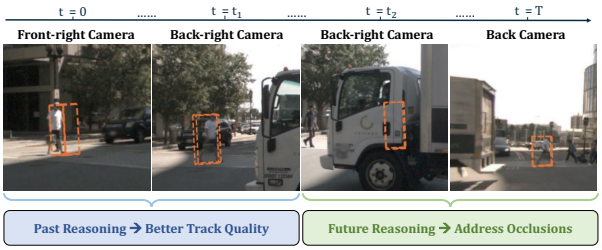

Figure 1. We visualize the output of our model by projecting predicted 3D bounding boxes onto images. In the beginning, image-based detection can be inaccurate ($t = 0$) due to depth ambiguity. With "Past Reasoning," the bounding box quality ($t = t_1$) gradually improves by leveraging historical information. With "Future Reasoning," our PF-Track predicts the long-term motions of objects and maintains their states even under occlusions ($t = t_2$) and camera switches. This enables re-association without explicit re-identification ($t = T$), as the object ID does not switch. Our PF-Track further combines past and future reasoning in a joint framework to improve spatio-temporal coherence.

*Work done while interning at Toyota Research Institute.
†Corresponding to Ziqi Pang at ziqip2@illinois.edu and Yu-Xiong Wang at yxw@illinois.edu.

tion to estimate a future trajectory. Finally, we can successfully track the pedestrian on re-appearance even on a different camera via long-term prediction, resulting in correct re-association (frame $t = T$ in Fig. 1). The above robust spatio-temporal reasoning is enabled by seamless, bi-directional integration of past and future information, which starkly contrasts with the mainstream pipelines for vision-based, multi-camera, 3D multi-object tracking (3D MOT).

To this end, we propose an end-to-end framework for joint 3D object detection, tracking, and trajectory prediction for the task of 3D MOT, as shown in Fig. 2, adopting the "tracking by attention" [34,64,65] paradigm. Compared to our closest baseline under the same paradigm [65], we are different in explicit past and future reasoning: a 3D object query consistently represents the object over time, propagates the spatio-temporal information of the object across frames, and generates the corresponding bounding boxes and future trajectories. To exploit spatio-temporal cues, our algorithm leverages simple attention operations to capture object dynamics and interactions, which are then used for track refinement and robust, long-term trajectory prediction. Finally, we close the loop by integrating predicted trajectories back into the tracking module to replace missing detections (*e.g.*, due to an occlusion). To highlight the capability of joint past and future reasoning, our method is named "Past-and-Future reasoning for Tracking" (PF-Track).

We provide a comprehensive evaluation of PF-Track on nuScenes [4] and demonstrate that joint modeling of past and future information provides clear benefits for object tracking. In particular, PF-Track decreases ID-Switches by over **90**% compared to previous multi-camera 3D MOT methods.

To summarize, our contributions are as follows.

1. We propose an end-to-end vision-only 3D MOT framework that utilizes object-level spatio-temporal reasoning for both past and future information.
2. Our framework improves the quality of tracks by cross-attending to features from the "**past**."
3. We propose a joint tracking and prediction pipeline, whose constituent part is "Future Reasoning", and demonstrate that tracking can explicitly benefit from long-term prediction into the "**future**."
4. Our method establishes new state-of-the-art on large-scale nuScenes dataset [4] with significant improvement for both AMOTA and ID-Switch.

2. Related Work

LiDAR-based 3D MOT. The majority of prior works in 3D MOT leverage the LiDAR modality. Due to the recent advances in LiDAR-based 3D detection [23, 61], especially the reliable range information, most state-of-the-art 3D MOT algorithms adopt a "tracking-by-detection" paradigm [56]. Given single frame detection outputs, different approaches have been proposed to improve data association [36, 58, 63], motion propagation [9, 68], and life cycling [36, 51]. However, most of these works assume the localization accuracy of detection output. Therefore, data association is usually conducted based on location, optionally combined with abstracted object attributes (*e.g.*, 3D intersection over untion (3D IoU) [56], 3D generalized intersection over union (GIoU) [36], and L2 distance [61]). This bias causes the proposed systems to be fragile when migrated into the camera modality, where 3D detection suffers from higher localization uncertainty. Although the latest methods incorporate learning-based algorithms to improve association with high-fidelity features such as low-level features from point clouds [46] or intermediate features from cameras [9], these approaches are built on top of the LiDAR-based frameworks and share their dependence on localization quality.

Camera-based 2D MOT. Camera-based multi-object tracking in 2D is a classic task in computer vision. Dominated by "tracking by detection" paradigm [3], 2D MOT has seen more success in leveraging high-fidelity features [38, 49, 60, 66]. Earlier works like DeepSORT [60] leveraged intermediate features from a deep net to measure appearance similarity. FairMOT [66] employed an additional Re-ID branch to learn discriminative features in a detection network. TransMOT [38] proposed to incorporate spatio-temporal features using a graph network.

Camera-based 3D MOT. Camera-based 3D MOT has recently drawn more attention in autonomous driving applications thanks to advances in monocular depth estimation [13, 15, 17] and image-based 3D object detection [20, 24, 25, 28, 29, 37, 41, 52, 53]. Early methods adapt the 2D MOT algorithms and lift the 2D tracking result using monocular depth [50, 68]. More recent approaches employ additional 3D information in data association [19, 31, 43]. [31] proposes to leverage 3D reconstruction, and [19] augments the 2D Re-ID features with 3D attributes (*e.g.* depth and orientation). CC-3DT [12] merges the multi-view camera features for identical objects to improve the cross-time cross-view association. However, considering or correcting the high uncertainty and bias in camera-based 3D detection has been less explored. In this work, we leverage long-term object reasoning, especially past reasoning, to improve the quality of 3D bounding boxes.

Tracking by Attention. A rising trend in MOT is the "tracking by attention" paradigm [34, 47, 64, 65], inspired by the novel transformer-based detection architecture DETR [5]. MOTR [64] and Trackformer [34] extended the query-based detection framework in DETR [5] by propagating queries across different frames. In this paradigm,

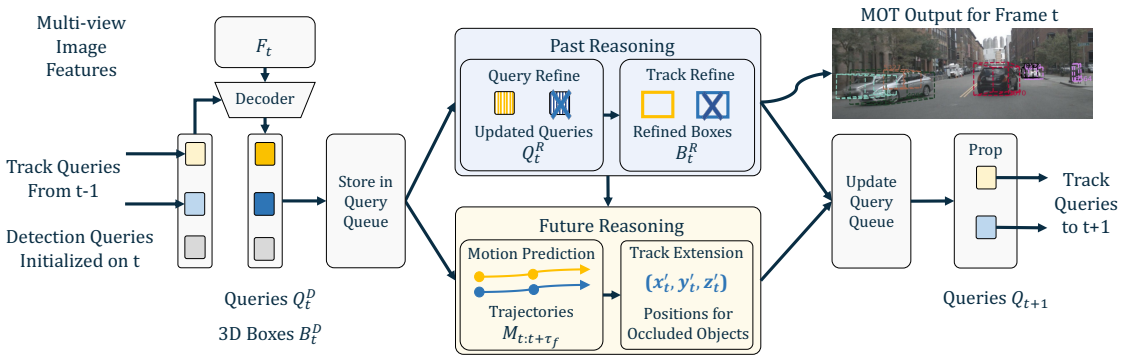

Figure 2. **PF-Track Framework.** PF-Track represents objects as queries, decodes image features, and predicts bounding boxes. To improve spatio-temporal coherence, we incorporate novel "Past Reasoning" and "Future Reasoning" modules. (1) "Past Reasoning" refines the features of queries and bounding boxes of tracks by exploiting the historical information in the query queue. (2) "Future Reasoning" improves the propagation of queries across frames by estimating long-term future trajectories. Furthermore, if an object is lost due to low confidence or occlusion (blue squares with ×), the "track extension" module can use a long-term trajectory to maintain its location. Finally, PF-Track incorporates past and future reasoning jointly for 3D MOT. (Best viewed in color, details in Sec. 3.1.)

the data association is replaced by "detection" in the current frame with a set of track queries. MUTR3D [65] proposes the first framework applying this paradigm to the 3D MOT domain. It uses a 3D track query to jointly model object features across timestamps and multi-views. Despite its improvement at the time, MUTR3D mostly follows the designs of 2D MOT methods and does not include special treatment to improve the localization quality of tracks and better propagate the queries to future frames. Our proposed algorithm also operates in the "tracking by attention" paradigm but extends the temporal horizon of existing methods. In particular, we demonstrate that joint past and future reasoning can improve the tracking framework by providing a strong spatio-temporal object representation.

Motion Prediction. Predicting agent trajectories is critical for self-driving [14, 22, 30, 35, 42, 45, 62]. The most common setting is to predict from clean tracks annotated by humans or auto-labeling [8, 11, 48, 59]. Numerous studies focus on end-to-end prediction from perception [1, 6, 18, 26, 32, 39, 40, 44, 54, 55, 57, 67], especially how to improve motion prediction directly from perception. However, our objective is different: Could a motion prediction model benefit 3D MOT? In the 2D setting, this problem has received only limited attention recently [10]. Our algorithm advances this research into a more challenging multi-camera, 3D scenario and does not require explicit re-identification.

3. Method: PF-Track

This section introduces our novel 3D multi-object tracking framework, shown in Fig. 2. It is centered around explicit past and future modeling of object trajectories in an end-to-end framework. We first provide an overview of the pipeline in Sec. 3.1, and then explain how to efficiently leverage "Past" (Sec. 3.2) and "Future" (Sec. 3.3) information. Finally, we summarize the losses used in our framework in Sec. 3.4.

3.1. PF-Track Pipeline

Our proposed PF-Track iteratively uses a set of object queries [34, 64, 65] to tackle multi-view, multi-object, 3D tracking. At each timestamp t, given K images $\mathbf{I}_t^k$ from surrounding cameras, the objective of 3D MOT is to generate object detections with consistent IDs across frames, denoted by $\mathbf{B}_t = \{\mathbf{b}_t^i\}$, where i is an object ID.

3D Object Queries. The entry point in our framework is to receive the object queries $\mathbf{Q}_t = \{\mathbf{q}_t^i\}$ propagated from the previous frame $t-1$ (yellow and blue squares in Fig. 2), which represent the tracked objects:

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

Such a query-based design naturally addresses the task of tracking as the queries carry the identity of objects over time. Apart from queries from the previous frame that represent tracked instances, we also add a fixed number of detection queries (gray squares in Fig. 2) to discover new objects. In practice, we use 500 detection queries initialized as learnable embeddings.

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 $\mathbf{c}_t^i$: $\mathbf{q}_t^i = \{\mathbf{f}_t^i, \mathbf{c}_t^i\}$. Here we highlight that the query position is an active participant in decoding the bounding boxes of objects below.

Decoder. To predict 3D bounding boxes and update queries with the latest image inputs, PF-Track adopts an attention-based detection architecture [5, 69] to decode image features $\mathbf{F}_t$ with object queries:

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

where $\mathbf{B}_t^D$ and $\mathbf{Q}_t^D$ are the detected 3D bounding boxes and updated query features, respectively. In the decoding process, the decoder lifts the 3D positions $\mathbf{c}_t^i$ of queries into positional embeddings to concentrate on the image regions relevant to the spatial locations of the objects. While the design of PF-Track is agnostic to query-based detection algorithms, we mainly adopt a current state-of-the-art 3D detector, PETR [28], for experiments.

Past and Future Reasoning for Refinement and Propagation. After decoding the queries and boxes from single-frame image features, PF-Track conducts past and future reasoning sequentially to (1) refine the current detections $\mathbf{B}_t^D$ into $\mathbf{B}_t^R$ and queries $\mathbf{Q}_t^D$ into $\mathbf{Q}_t^R$. (R is short for "refinement."); (2) propagate the queries to the next timestamp with the predicted motions.

"Past Reasoning" $\mathbf{PR}(\cdot)$ is the component that aggregates the information from previous frames to generate refined queries $\mathbf{Q}_t^R$ and refined bounding boxes $\mathbf{B}_t^R$:

$$\mathbf{Q}_t^R, \mathbf{B}_t^R \leftarrow \mathbf{PR}(\mathbf{Q}_t^D, \mathbf{B}_t^D, \mathbf{Q}_{t-\tau_h:t-1},). \tag{3}$$

In practice, the historical queries $\mathbf{Q}_{t-\tau_h:t-1}$ come from a **query queue** that maintains the queries from past τ_h frames (h for "history").

After past reasoning, the "Future Reasoning" module $\mathbf{FR}(\cdot)$ improves the coherence of object positions from the aspect of query propagation. It achieves this by forecasting the motions up to τ_f frames (f for "future") and transforms the positions of queries accordingly:

$$\mathbf{Q}_{t+1}, \mathbf{M}_{t:t+\tau_f} \leftarrow \mathbf{FR}(\mathbf{Q}_t^R, \mathbf{Q}_{t-\tau_h:t-1}). \tag{4}$$

Specifically, future reasoning extracts the object dynamics from historical query features to predict the trajectories $\mathbf{M}_{t:t+\tau_f}$. The single-step movement $\mathbf{M}_{t:t+1}$ is leveraged to propagate the current queries $\mathbf{Q}_t^R$ to the next timestamp, and long-term trajectories $\mathbf{M}_{t+1:t+\tau_f}$ are used for addressing occlusions. The "Track Extension" in Fig. 2 refers to occlusion reasoning through the predicted trajectories.

PF-Track iteratively executes the above procedures. The refined 3D bounding boxes $\mathbf{B}_t^R$ are the output for 3D MOT.

3.2. Past Reasoning

To address the uncertainty of detection in vision-only 3D localization, past reasoning focuses on two aspects: (1) enhancing the query features by attending to historical embeddings; (2) refining the tracks by adjusting the bounding boxes using the improved query features.

Query Refinement: from $\mathbf{Q}_t^D$ to $\mathbf{Q}_t^R$. We first apply attention across the time and instance axes to explicitly update the query features with historical information, as illustrated in Fig. 3. "Cross-frame" attention encourages the interplay

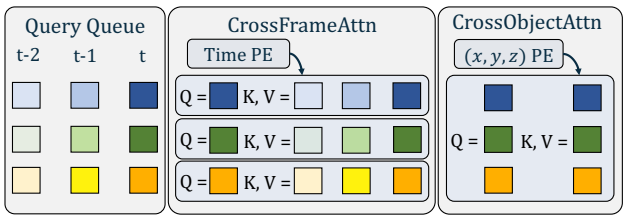

Figure 3. **Query Refinement.** "*Cross-frame*" and "*Cross-object*" attention modules process the query queue to capture the temporal and inter-object relationship, respectively. They apply the positional encoding for time t and spatial locations (x, y, z), respectively. (Best viewed in color.)

of features within a history window of τ_h frames per object:

$$\begin{aligned}\mathbf{f}_t^i \leftarrow \mathbf{CrossFrameAttn}(&\mathrm{Q} = \mathbf{f}_t^i, \\ &\mathrm{K} = \mathbf{f}_{t-\tau_h:t}^i, \mathrm{V} = \mathbf{f}_{t-\tau_h:t}^i, \\ &\mathrm{PE} = \mathbf{Pos}(t-\tau_h:t)),\end{aligned} \tag{5}$$

where, $\mathbf{Pos}(t-\tau_h:t)$ converts the timestamps into positional embedding, and the history frames with empty features are ignored for attention computation.

Then past reasoning applies "cross-object" attention to incorporate the context information and encourage more discriminative feature representation for each object. In particular, cross-object attention (Fig. 3, right) further updates the query features via

$$\begin{aligned}\mathbf{f}_t^{1:N_t} \leftarrow \mathbf{CrossObjectAttn}(&\mathrm{Q,K,V} = \mathbf{f}_t^{1:N_t}, \\ &\mathrm{PE} = \mathbf{Pos}(\mathbf{c}^{1:N_t})),\end{aligned} \tag{6}$$

where cross-object attention exchanges the features of N_t objects guided by their 3D positional embedding $\mathbf{Pos}(\mathbf{c}^{1:N_t})$. The final output $\mathbf{f}_t^{1:N_t}$ becomes the refined feature vectors in queries $\mathbf{Q}_t^R$.

As a brief remark, decoupling cross-frame and cross-object attention exhibits two advantages. Firstly, separating attention across frames (cross-frame) and objects (cross-object) enables us to design specialized positional encoding of time and locations for each of them. Secondly, it decreases the computational complexity from $\mathcal{O}(N_t^2\tau_h^2)$ for the global cross-attention to $\mathcal{O}(N_t^2 + N_t\tau_b^2)$, which is significantly less. Our design is also closely related to how motion prediction methods [14, 35] model spatial-temporal relationships.

Track Refinement: from $\mathbf{B}_t^D$ to $\mathbf{B}_t^R$. With the queries refined by historical information, past reasoning further uses track refinement to improve the 3D bounding box quality. As specified in Eqn. 7, we apply a multi-layer perceptron (MLP) to predict the updated properties of objects, including center residuals $(\Delta x, \Delta y, \Delta z)$, size (l, w, h), orientations (θ), velocities $(\mathbf{v})$, and scores (s):

$$(\Delta x, \Delta y, \Delta z, l, w, h, \theta, \mathbf{v}, s)^i = \mathbf{MLP}(\mathbf{f}_t^i). \tag{7}$$

These are then used to adjust the original boxes as follows:

$$\mathbf{b}_t^i = (\Delta x + x_t^i, \Delta y + y_t^i, \Delta z + z_t^i, l, w, h, \theta, \mathbf{v}, s), \quad (8)$$

resulting in $\mathbf{B}_t^R$, which is the final model output at frame t.

3.3. Future Reasoning

"Future Reasoning" concentrates on improving the propagation of queries across frames to benefit spatio-temporal coherence. It first learns a trajectory prediction, which is used for moving queries across adjacent frames. Then future reasoning exploits the predicted long-term trajectories for maintaining the positions of occluded or noisy tracks.

Motion Prediction. Trajectory prediction supervises the model's ability to capture object movements and is further beneficial for propagating query positions across timestamps. Similar to past reasoning, our future reasoning model adopts a simple attention-based architecture. Firstly, we generate the motion embeddings for τ_f timestamps $\mathbf{mf}_{t:t+\tau_f}^i$ with a cross-frame attention:

$$\begin{aligned} \mathbf{mf}_{t:t+\tau_f}^i \leftarrow \mathbf{CrossFrameAttn}(& \\ \mathrm{Q} &= \mathbf{mf}_{t:t+\tau_f}^i, \\ \mathrm{K}, \mathrm{V} &= \mathbf{f}_{t-\tau_h:t}^i, \mathbf{f}_{t-\tau_h:t}^i, \\ \mathrm{PE} &= \mathbf{Pos}(t-\tau_h : t+\tau_f)), \end{aligned} \quad (9)$$

where $\mathbf{mf}_{t:t+\tau_f}^i$ are initialized as zeros, and historical features $\mathbf{f}_{t-\tau_h:t}^i$ serve as the source of information. Then the movement at every timestamp is decoded by an MLP:

$$\mathbf{m}_{t:t+\tau_f}^i = \mathbf{MLP}(\mathbf{mf}_{t:T+\tau_f}^i), \quad (10)$$

and the object trajectory in the 3D space can be recovered by combining these frame-level outputs. Our architecture is inspired by SceneTransformer [35], which also employs a fully-attention-based architecture.

The predicted trajectories $\mathbf{m}_{t:t+\tau_f}^i$ have better fidelity compared to the velocities $\mathbf{v}$ predicted by the decoder in $\mathbf{B}_t^R$ and $\mathbf{B}_t^D$. Thus, we can propagate the positions of queries by adding a single step of the trajectory:

$$\mathbf{c}_{t+1}^i = \mathbf{c}_t^i + \mathbf{m}_{t:t+1}^i. \quad (11)$$

Track Extension. To handle occlusions or noisy observations, we propose to extend the tracks using the predicted trajectories. In particular, we replace missing or low-confidence detections with the output of our motion prediction module, which is initialized from confident observations. Previous 3D MOT approaches either terminate the tracks or prolong them with heuristic motion models (*e.g.* Kalman filters) under such conditions. However, these solutions both could lead to ID-Switches due to "early termination" [36] or false associations. In contrast, our learnable motion prediction module and track extension strategy are more accurate and robust.

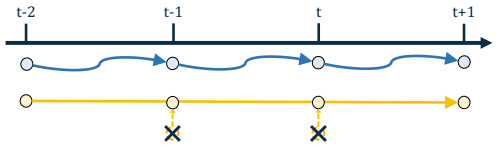

Figure 4. **Track extension.** PF-Track updates object positions and predicts future trajectories at every timestamp (top row). However, if the object cannot be confidently localized (*e.g.* due to occlusion or a noisy observation, bottom row at frames $t-1$ and t), our method will rely on the long-term trajectories predicted from confident timestamps (frame $t-2$) to infer the positions of this object and ignore the noisy observations (crossed-out circles).

We visualize the high-level intuition in Fig. 4 and provide more details in the Supplementary (Sec. B.2). In Fig. 4, the long-term trajectories assist the propagation of the yellow instance (bottom row). When PF-Track encounters noisy observation or occlusion cases, it relies on the motion predictions from previous confident frames to simulate the movements of occluded objects. In extreme cases, our model is able to handle occlusion length of $\tau_f - 1$ frames. Our ablation study in Sec. 4.4 demonstrates that track extension can decrease ID-Switch by a large margin. To the best of our knowledge, we are *the first to incorporate long-term prediction into a query-based framework and address occlusion without explicit re-identification*.

3.4. Loss Functions

Our final loss function is defined as follows:

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{cls}}^D \mathcal{L}_{\text{cls}}^D + \lambda_{\text{box}}^D \mathcal{L}_{\text{box}}^D + \\ & \lambda_{\text{cls}}^R \mathcal{L}_{\text{cls}}^R + \lambda_{\text{box}}^R \mathcal{L}_{\text{box}}^R + \lambda_f \mathcal{L}_f \end{aligned} \quad (12)$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{cls}}^R$ are focal loss [27] with the coefficients of λ_{cls}^D and λ_{cls}^R. They supervise the classification scores of $\mathbf{B}_t^D$ and $\mathbf{B}_t^R$, respectively. $\mathcal{L}_{\text{box}}^D$ and $\mathcal{L}_{\text{box}}^R$ are both L1 loss applied to $\mathbf{B}_t^D$ and $\mathbf{B}_t^R$ for bounding box regression. Their coefficients are λ_{box}^D and λ_{box}^R. The motion prediction loss $\mathcal{L}_f$ is an L1 loss between the movements of predicted and ground truth trajectories, weighted by λ_f. The ground truth assignment couples a query with a consistent ground truth instance over time to encourage ID consistency. We discuss more details in the Supplementary (Sec. B.1).

4. Experiments

4.1. Datasets and Metrics

Datasets. We conduct experiments on the large-scale self-driving dataset nuScenes [4]. It contains 1,000 video se-

	AMOTA ↑	AMOTP ↓	RECALL ↑	MOTA ↑	IDS ↓
Validation Split					
DEFT [7]	0.201	N/A	N/A	0.171	N/A
QD3DT [19]	0.242	1.518	39.9%	0.218	5646
MUTR3D [65]	0.294	1.498	42.7%	0.267	3822
TripletTrack [33]	0.285	1.485	N/A	N/A	N/A
CC-3DT* [12]	0.429	1.257	53.4%	0.385	2219
PF-Track-S (Ours)	0.408	1.343	50.7%	0.376	**166**
PF-Track-F (Ours)	**0.479**	**1.227**	**59.0%**	**0.435**	181
Test Split					
CenterTrack [68]	0.046	1.543	23.3%	0.043	3807
PermaTrack [50]	0.066	1.491	18.9%	0.060	3598
DEFT [7]	0.177	1.564	33.8%	0.156	6901
QD3DT [19]	0.217	1.550	37.5%	0.198	6856
MUTR3D [65]	0.270	1.494	41.1%	0.245	6018
TripletTrack [33]	0.268	1.504	40.0%	0.245	1144
CC-3DT* [12]	0.410	1.274	**53.8%**	0.357	3334
PF-Track-F (Ours)	**0.434**	**1.252**	**53.8%**	**0.378**	**249**

Table 1. Comparison with state-of-the-art camera-based 3D MOT algorithms on nuScenes [4]. "**S**" and "**F**" denotes our model trained with small-resolution and full-resolution setting, respectively (clarified in Sec. 4.2). Our approach has a significant advantage on both AMOTA and ID-Switch (full-resolution), where ID-Switch is almost 90% less and *an order of magnitude smaller* compared to other methods. (*) indicates concurrent works.

quences with multiple modalities, including RGB images from 6 surrounding cameras, and point clouds from LiDAR and Radar. In this paper, we use camera sensors only. Every sequence spans roughly 20 seconds with keyframes annotated at 2Hz. The dataset provides 1.4M 3D bounding boxes covering 10 types of common objects on the road. For the tracking task, nuScenes selects a subset of 7 mobile categories, such as cars, pedestrians, and motorcycles, and excludes static objects like traffic cones.

Metrics. We strictly follow the official evaluation metrics for multi-object tracking tasks from nuScenes. It modifies CLEAR MOT metrics [2] by considering multiple recall thresholds. The main metric is "Average Multi-Object Tracking Accuracy" (AMOTA) [56]. Meanwhile, we also consider other analytical metrics such as "Identity Switches" (IDS) and "Average Multi-Object Tracking Precision" (AMOTP).

4.2. Implementation Details

Due to space limits, we clarify two training settings here and describe more implementation detail in the Supplementary (Sec. B). In our implementation, every training sample contains three adjacent frames from different timestamps. However, it requires extensive computation as every frame contains six high-resolution images. Therefore, we adopt two settings that downsample images to different resolutions, motivated by PETR [28].

Full-resolution. On every time frame, we crop the raw resolution images, 1600×900 to 1600×640, leaving the sky area out. However, training a multi-frame tracker on this resolution would not fit in a single A100 GPU. Thus, we first pretrain the backbone with single-frame detection for 24 epochs, following some previous works [50]. Then we fix the backbone and train the tracker on three-frame samples for another 24 epochs. We only use this setting for full model results indicated with "-F" in Tab. 1.

Small-resolution. We apply a small-resolution setting for all of our ablation analyses unless specified. In this setting, we downsample the cropped images to a resolution of 800×320. We first train a single-frame detection model for 12 epochs and then train the tracker on three-frame samples for another 12 epochs.

Index	Past QR	Past TR	Future Pred	Future Ext	AMOTA↑	AMOTP↓	IDS↓
1					0.368	1.421	507
2	✓				0.378	1.414	453
3	✓	✓			0.380	1.408	400
4			✓		0.374	1.402	469
5			✓	✓	0.391	1.360	**155**
6	✓	✓	✓	✓	**0.408**	**1.343**	166

Table 2. **Ablation of PF-Track Modules.** For past reasoning, "QR" and "TR" denote "query refinement" and "track refinement" in Sec. 3.2. For future reasoning, "Pred" and "Ext" denote "motion prediction" and "track extension" in Sec. 3.3. Past and future reasoning improve 3D MOT independently, and PF-Track achieves top results by combining them in an end-to-end framework.

4.3. State-of-the-art Comparison on nuScenes

In Tab. 1, we compare our model performance with the other published camera-based 3D MOT algorithms on nuScenes. Our approach establishes a new state-of-the-art with significant improvements on every metric. Our AMOTA improves more than 7% on the test set and 12% on the validation set over the previous methods, including a very strong concurrent work [12]. It is worth noting that with more established tracks (higher recall), *our ID-Switch number is only* 10% *of previous methods eliminating more than* 90% *of the ID-switching errors*. This result indicates the strong association ability of our algorithm attributed to leveraging both past and future reasoning. The advantage of our model holds *even when trained in the low-resolution setting*, whereas most of the previous works use full-resolution.

4.4. Ablation Studies

Efficacy of Past and Future Reasoning. In Tab. 2, we analyze the importance of individual modules for our model's performance using the validation set of nuScenes. In particular, we evaluate the following variants. (1) **Baseline.** Our baseline is a "tracking by attention" model without explicit spatio-temporal reasoning (row 1). It is a strong baseline

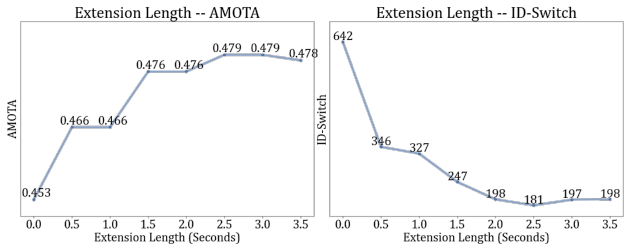

Figure 5. **Track extension assists MOT.** By using the predicted trajectories to maintain the states for low-confidence tracks, we significantly improve AMOTA and decrease ID-Switch.

and outperforms prior work in Tab. 1. (2) **Past Reasoning.** We first analyze the effect of query refinement (row 2), which explicitly incorporates the historical queries via cross-attention. As illustrated, it improves the overall tracking quality. We then exploit the enhanced feature to refine the 3D bounding boxes of tracks (row 3). It decreases ID-Switch and AMOTP, which indicates that track refinement is useful for 3D MOT. (3) **Future Reasoning.** Next, we demonstrate that learning to predict object motion and propagate positions (row 4) is beneficial for modeling object dynamics and leads to improved tracking performance. In addition, using the long-term trajectory predictions to replace low-confidence localizations (row 5) results in a 67% drop in ID-Switches. (4) **Joint Past and Future Reasoning.** Finally, combining past and future reasoning into an end-to-end framework shown in Fig. 2 (row 6) allows our model to achieve top performance. This result confirms that past and future reasoning are mutually beneficial for 3D MOT.

Length of Track Extension. In Fig. 5, we analyze the effect of track extension length on AMOTA and ID-Switches using our best-performing full-resolution model on the validation split of nuScenes. Compared to not using the extension strategy (0.0s), prolonging the tracks strongly improves the performance up to 2 seconds. Then the metrics saturate because only a few objects reappear after such a long period. *Please note that these improvements are achieved without explicit re-identification.*

Length of Prediction in Future Reasoning. Next, we analyze how the prediction length changes the 3D MOT performance on nuScenes validation split in Tab. 3. Concretely, we train three different full-fledged models with the prediction length of 2.0, 3.0, and 4.0 seconds. AMOTA and ID-Switch indicate that 4.0s (8 frames) has a slight advantage over 2.0s (4 frames) and 3.0s (6 frames). This result indicates that learning trajectory forecasting with longer horizon benefits our 3D MOT framework.

Comparison with "Tracking by Detection" Baselines. In Tab. 4, we compare the performance between our end-to-end framework and previous "tracking by detection" algorithms [36,56,61], which are strong baselines for LiDAR-based 3D MOT. For these experiments, we also use the validation split of nuScenes. For a fair comparison, we evaluate these methods with PETR [28] detections and tune their hyper-parameters for AMOTA (shown in the table with "Ψ," details are provided in the Supplementary, Sec. B.4). The results clearly demonstrate the advantages of our end-to-end approach compared to more traditional, modular frameworks, with improvements being especially significant on the ID-Switch metric.

Analysis on Prediction from Query Features. While our paper focuses on multi-object tracking, we additionally pro-

Length	Extention	AMOTA ↑	AMOTP ↓	IDS ↓
2.0s	✗	0.392	1.376	604
2.0s	✓	0.402	1.342	217
3.0s	✗	0.392	1.372	540
3.0s	✓	0.402	**1.340**	208
4.0s	✗	0.391	1.387	471
4.0s	✓	**0.408**	1.343	**166**

Table 3. **Length of motion prediction**. "Extension" means using "track extension." We train three models with the prediction horizon of 2.0s, 3.0s, and 4.0s. According to AMOTA and IDS, learning a longer prediction benefits tracking.

	AMOTA ↑	AMOTP ↓	IDS ↓
AB3DMOT [56]	0.292	1.333	2419
AB3DMOT [56]$^{\Psi}$	0.329	1.388	2677
CenterPoint [61]	0.233	**1.270**	2715
CenterPoint [61]$^{\Psi}$	0.383	1.329	3082
SimpleTrack [36]	0.320	1.295	1606
SimpleTrack [36]$^{\Psi}$	0.402	1.324	2053
PF-Track (Ours)	**0.408**	1.343	**166**

Table 4. **Comparison with "tracking by detection."** We apply strong baselines in 3D MOT to PETR [28]: AB3DMOT [56], CenterPoint [61], and SimpleTrack [36]. "Ψ" means that we tune the hyper-parameters of these methods to fit PETR detections, rather than using their original configuration. Our end-to-end approach has significant advantages.

Method	ADE ↓ (@4.0s)	FDE ↓ (@4.0s)
LSTM [8]	2.32	2.87
VectorNet [14]	2.01	2.48
Velocity	2.10	2.64
PF-Track (Ours)	**1.88**	**2.38**

Table 5. **Motion prediction from features or abstract states.** We build a motion prediction benchmark from the true-positive tracks of PF-Track on the nuScenes validation split, and then train LSTM [8] and VectorNet [14] from the 3D positions of tracks. The "Velocity" row is the result under the assumption of a constant velocity motion model. The results indicate that predicting from features provides richer information for better trajectory quality.

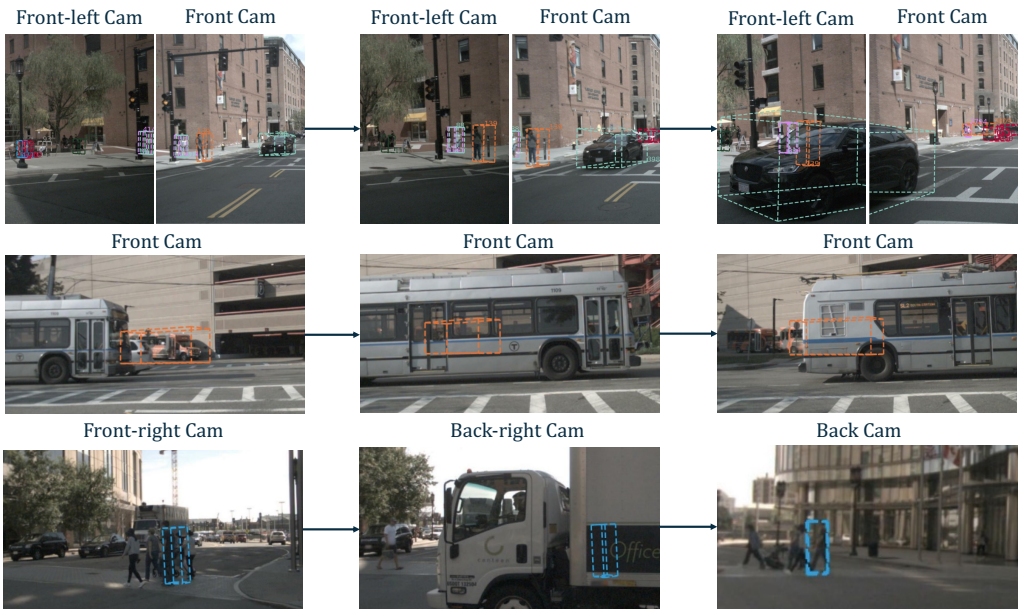

Figure 6. **Qualitative results for 3D MOT.** (1) In the top row, we provide image-level 3D MOT results. The figures highlight the consistency across images, such as the vehicles crossing the front-left and front cameras. (2) In the middle and bottom rows, we provide two dedicated examples for addressing large and small objects' occlusions.

vide an analysis of prediction performance. We show that predicting end-to-end from object features is advantageous over predicting from low-level object states, such as center positions. Specifically, we train two motion prediction baselines, VectorNet [14] and LSTM [8], using the true-positive tracks from PF-Track following previous studies [32], and report their results in the top rows of Tab. 5. As our method does not use HD-Maps, for a fair comparison, we exclude the parts of motion prediction algorithms that handle HD-Maps in these experiments. In addition, we report another baseline which uses the velocities predicted by our model's decoder for trajectory prediction, assuming a constant velocity motion model (third row in Tab. 5). The evaluation metrics are "average displacement error" (ADE) and "final displacement error" (FDE), which are better with lower values. More details are in the Supplementary (Sec. B.5).

Tab. 5 compares the performance between the end-to-end PF-Track and the baselines described above on the validation split of nuScenes. With lower ADE and FDE, PF-Track has better trajectory quality. Our conclusions agree with previous studies [16, 32, 55]. More specifically, LSTM is a shallow model and unable to capture meaningful dynamics from noisy tracks; the stronger VectorNet model can perform better than the other baselines, but it is still worse than forecasting trajectories in an end-to-end framework, as proposed in our method.

4.5. Qualitative Results

We visualize the 3D MOT results in Fig. 6 by projecting 3D bounding boxes onto images. The colors of bounding boxes are randomly selected from a pool of seven colors according to their IDs, so that each object has a consistent color over time.

In the top row, we provide an overall visualization of multi-camera 3D MOT, focusing on front-left and front cameras. As clearly shown, PF-Track tracks objects coherently, especially for the pedestrians and vehicles shown on two separate cameras. In the bottom two rows of Fig. 6, we illustrate two examples of addressing occlusions. For both large (bus) and small (pedestrian) objects, our method propagates their positions during the occluded frames and successfully re-associates them on de-occlusion frames even on a different camera. We highlight that *this is achieved without an explicit re-identification module*.

5. Conclusions

This paper proposes a query-based end-to-end method for multi-camera 3D MOT that enhances spatio-temporal coherence. By past reasoning, our framework enhances the query features and track quality with historical information. By future reasoning, the predicted trajectories better propagate the queries across adjacent frames and occluded long-term periods. We also demonstrate that joint past and future reasoning further strengthens the tracker's ability. Extensive evaluation of the large-scale nuScenes dataset demonstrates that our method is effective in providing coherent tracks.

Acknowledgement. This work was supported in part by Toyota Research Institute, NSF Grant 2106825, NIFA Award 2020-67021-32799, and the NCSA Fellows program.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. In *ECCV*, 2022. 3

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008. 6

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 2

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2, 5, 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[6] Sergio Casas, Abbas Sadat, and Raquel Urtasun. MP3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 3

[7] Mohamed Chaabane, Peter Zhang, J. Ross Beveridge, and Stephen O'Hara. DEFT: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021. 1, 6

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, 2019. 1, 3, 7, 8

[9] Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving. In *ICRA*, 2021. 2

[10] Patrick Dendorfer, Vladimir Yugay, Aljoša Ošep, and Laura Leal-Taixé. Quo Vadis: Is trajectory forecasting the key towards long-term multi-object tracking? In *NeurIPS*, 2022. 3

[11] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 1, 3

[12] Tobias Fischer, Yung-Hsu Yang, Suryansh Kumar, Min Sun, and Fisher Yu. CC-3DT: Panoramic 3D object tracking via cross-camera fusion. In *CoRL*, 2022. 1, 2, 6

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[14] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In *CVPR*, 2020. 3, 4, 7, 8

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2

[16] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. ViP3D: End-to-end visual trajectory prediction via 3D agent queries. *arXiv preprint arXiv:2208.01582*, 2022. 1, 8

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 2

[18] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, 2021. 3

[19] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3D object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 6

[20] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2

[21] Boris Ivanovic, Kuan-Hui Lee, Pavel Tokmakov, Blake Wulfe, Rowan McAllister, Adrien Gaidon, and Marco Pavone. Heterogeneous-agent trajectory forecasting incorporating class uncertainty. *arXiv preprint arXiv:2104.12446*, 2021. 1

[22] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *ICCV*, 2019. 3

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 2

[24] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 1, 2

[25] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 2

[26] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. PnPNet: End-to-end perception and prediction with tracking in the loop. In *CVPR*, 2020. 3

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[28] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *ECCV*, 2022. 1, 2, 4, 6, 7

[29] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETRv2: A unified framework for 3D perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 2

[30] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *CVPR*, 2021. 3

[31] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *Robotics and Automation Letters*, 2020. 1, 2

[32] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net. In *CVPR*, 2018. 3, 8

[33] Nicola Marinello, Marc Proesmans, and Luc Van Gool. TripletTrack: 3D object tracking using triplet embeddings and lstm. In *CVPR*, 2022. 1, 6

[34] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *CVPR*, 2022. 2, 3

[35] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified multi-task model for behavior prediction and planning. In *ICLR*, 2021. 3, 4, 5

[36] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3D multi-object tracking. *arXiv preprint arXiv:2111.09621*, 2021. 1, 2, 5, 7

[37] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-LiDAR needed for monocular 3D object detection? In *ICCV*, 2021. 2

[38] Chu Peng, Wang Jiang, You Quanzeng, Ling Haibin, and Liu Zicheng. TransMOT: Spatial-temporal graph transformer for multiple object tracking. In *CVPR*, 2021. 2

[39] Neehar Peri, Jonathon Luiten, Mengtian Li, Aljoša Ošep, and Laura Leal-Taixé. Forecasting from LiDAR via future object detection. In *CVPR*, 2022. 3

[40] John Phillips, Julieta Martinez, Ioan Andrei Bârsan, Sergio Casas, Abbas Sadat, and Raquel Urtasun. Deep multi-task learning for joint localization, perception, and prediction. In *CVPR*, 2021. 3

[41] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distributionnetwork for monocular 3D object detection. *CVPR*, 2021. 2

[42] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, 2020. 3

[43] Samuel Scheidegger, Joachim Benjaminsson, Emil Rosenberg, Amrit Krishnan, and Karl Granström. Mono-camera 3D multi-object tracking using deep learning detections and pmbm filtering. In *IEEE Intelligent Vehicles Symposium*, 2018. 1, 2

[44] Meet Shah, Zhiling Huang, Ankit Laddha, Matthew Langford, Blake Barber, Sidney Zhang, Carlos Vallespi-Gonzalez, and Raquel Urtasun. LiRaNet: End-to-end trajectory prediction using spatio-temporal radar fusion. *arXiv preprint arXiv:2010.00731*, 2020. 3

[45] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. In *NeurIPS*, 2022. 3

[46] Colton Stearns, Davis Rempe, Jie Li, Rares Ambrus, Sergey Zakharov, Vitor Guizilini, Yanchao Yang, and Leonidas J Guibas. SpOT: Spatiotemporal modeling for 3D object tracking. In *ECCV*, 2022. 2

[47] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. TransTrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[48] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3

[49] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017. 2

[50] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *ICCV*, 2021. 2, 6

[51] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal tracker: Tracklet never dies. *arXiv preprint arXiv:2111.13672*, 2021. 2

[52] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *ICCV*, 2021. 2

[53] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *CoRL*, 2022. 1, 2

[54] Xinshuo Weng, Boris Ivanovic, Kris Kitani, and Marco Pavone. Whose track is it anyway? Improving robustness to tracking errors with affinity-based trajectory prediction. In *CVPR*, 2022. 3

[55] Xinshuo Weng, Boris Ivanovic, and Marco Pavone. MTP: Multi-hypothesis tracking and prediction for reduced error propagation. In *IEEE Intelligent Vehicles Symposium*, 2022. 3, 8

[56] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *IROS*, 2020. 2, 6, 7

[57] Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, and Nicholas Rhinehart. Inverting the pose forecasting pipeline with SPF2: Sequential pointcloud forecasting for sequential pose forecasting. In *CoRL*, 2021. 3

[58] Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris M Kitani. GNN3DMOT: Graph neural network for 3D multi-object tracking with 2D-3D multi-feature learning. In *CVPR*, 2020. 1, 2

[59] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2022. 1, 3

[60] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2

[61] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *CVPR*, 2021. 2, 7

[62] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *ICCV*, 2021. 3

[63] Jan-Nico Zaech, Alexander Liniger, Dengxin Dai, Martin Danelljan, and Luc Van Gool. Learnable online graph representations for 3D multi-object tracking. *Robotics and Automation Letters*, 2022. 1, 2

[64] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022. 2, 3

[65] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. MUTR3D: A multi-camera tracking framework via 3D-to-2D queries. In *CVPRW*, 2022. 1, 2, 3, 6

[66] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 2021. 2

[67] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 3

[68] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 2, 6

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3