Title

End-to-End Joint Multi-View 3D Object Detection and Tracking via Learning to Associate

Author

Yi, Wenlong

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

End-to-End Joint Multi-View 3D Object Detection and Tracking via Learning to Associate

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Computer Science

by

Wenlong Yi

2024

ABSTRACT OF THE THESIS

End-to-End Joint Multi-View 3D Object Detection and Tracking via Learning to Associate

by

Wenlong Yi

Master of Science in Computer Science

University of California, Los Angeles, 2024

Professor Bolei Zhou, Chair

Associating objects accurately across cameras and frames is essential but challenging for vision-based perception in autonomous driving system. In a vanilla tracking-by-detection fashion, most prior works associate detected objects over views and time via a great many of heuristic matching rules. In this work, we propose a simple yet efficient method named EMAT, or End-to-end Multi-view Association Tracker, which jointly performs 3D detection and tracking from multi-camera and multi-frame images in an end-to-end manner. Our key design is to predict the object appearing information and affinity of each object in sequential frames from temporally fused object query embedding features, which extracted from a temporal fusion module designed for learning to associate. Without any post-process, 3D tracklets are built up across frames, along with 3D detection and velocity estimation. Additionally, we propose a novel strategy to boost object velocity estimation by the information of object appearing. Experiments on the large-scale nuScenes dataset demonstrate that our approach outperforms the 3D detection baseline we build upon, achieving superior camera-based 3D tracking and velocity estimation performance. Additionally, it surpasses traditional 3D tracking methods, showcasing its effectiveness in real-world scenarios.

The thesis of Wenlong Yi is approved.

Miryung Kim

Jonathan Chau-Yan Kao

Bolei Zhou, Committee Chair

University of California, Los Angeles

2024

To my family

Contents

List of Figures

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Bolei Zhou, for his invaluable guidance, support, and encouragement throughout my Master's journey. His insights and expertise have been instrumental in shaping my academic and professional growth.

I am also profoundly thankful to all the professors who have taught me during my Master's program. Their dedication and knowledge have inspired me and equipped me with the skills necessary to excel in the field of computer science.

Thank you all for your unwavering support and mentorship.

Chapter 1

Introduction

Camera-based 3D object detection and multi-object tracking (MOT) are crucial for visual perception system in the field of autonomous driving and robot. 3D object detection task aims to predict 3D bounding box, namely the location, orientation and dimension of objects in 3D space. 3D MOT requires detection of objects in each frames and associate those to build up tracklets. The tracking-by-detection paradigm plays a dominant role in 3D MOT for a long time. Normally, a 3D detector is firstly conducted to estimate 3D bounding boxes for individual frames, and then association module establishes correspondence between the objects in different frames to form trajectories. Usually formulating association step as a bipartite matching problem, recent works have explored several cues in recent years, like Intersection-over-Union (IoU) [2], motion modeling [1, 38], appearance [36, 38, 42] and etc. to association. Although simple and fast, these hand crafted procedures may fail easily in challenging scenarios, consequently hinder the performance of data association. Also following this paradigm, some recent learning-based methods are introduced to first find object proposals in different frames and then obtain the association in the feature space in a supervised learning manner, applied in 2D MOT [3, 32] and 3D MOT [7, 12, 38]. However, optimizing detection and association separately will cut off the transmission of the uncertainty and error gradient between detector and tracker [18].

Recently, multi-view 3D perception in bird's eye-view (BEV) has attracted considerable attention and shown promising performance and efficiency over those in Perspective View (PV), which based on the monocular 3D detector followed by hand craft cross-camera post-processing rules. At the same time, transformer-based framework demonstrates strong capabilities in several perception tasks like detection [6, 43] and segmentation [28, 39]. Further, multi-camera 3D perception methods based on transformer are proposed for autonomous driving scenarios [15, 17, 18, 20, 35, 41]. By leveraging multi-frame information [17, 18, 20], the temporal attributes of objects such as velocity can be better estimated than methods which only take single frame of multi-view images as input.

Therefore, to avoid the weakness of traditional visual 3D MOT framework, we propose a simple but efficient framework EMAT (End-to-end Multi-view Association Tracker), which have the ability to detect and track 3D objects from multi-view and multi-frame images in an end-to-end manner. As shown in the bottom part of Figure 1.1, EMAT takes multi-view images I_{t-1} and I_t at both t-1 and t timestamps as input, and simultaneously performs 3D detection Det_t, data association $Asso_{t-1,t}$ and object velocity estimation $Vel_{t-1,t}$. To be specific, on the basis of off-the-shelf transformer-based multi-view 3D detector, such as DETR3D [35], PETR [21, 22], BEVFormer [20], we first design Association-oriented Temporal Fusion (ATF) module to aggregate features from object query embedding of different frames. Instead of simply stacking sequential features as most previous works, ATF module leverages attention mechanism for feature fusion. Second, to achieve the purpose of end-to-end training and unified optimization for 3D detection and MOT, these fused features are utilized to predict the object appearing information and affinity of each object in sequential frames, eventually to establish correspondence across time. Additionally, we propose Trustworthy Velocity Estimation (TVE) strategy, which further leverages object appearing information as an important cue to improve velocity estimation.

We summarize our main contributions as follows:

1. We propose a unified and general framework EMAT, to jointly optimize multi-view object detection and 3D multi-object tracking by learning to associate, where data association is finished without any post-process via estimating sequential objects appearing information and affinity.

2. We design ATF, a temporal fusion module to aggregate features from sequential object query embedding features for association.

3. We introduce a novel strategy, Trustworthy Velocity Estimation (TVE) to reduce object velocity error.

4. Our simple approach EMAT outperforms the 3D detection baseline we build upon, and surpasses traditional 3D tracking methods.

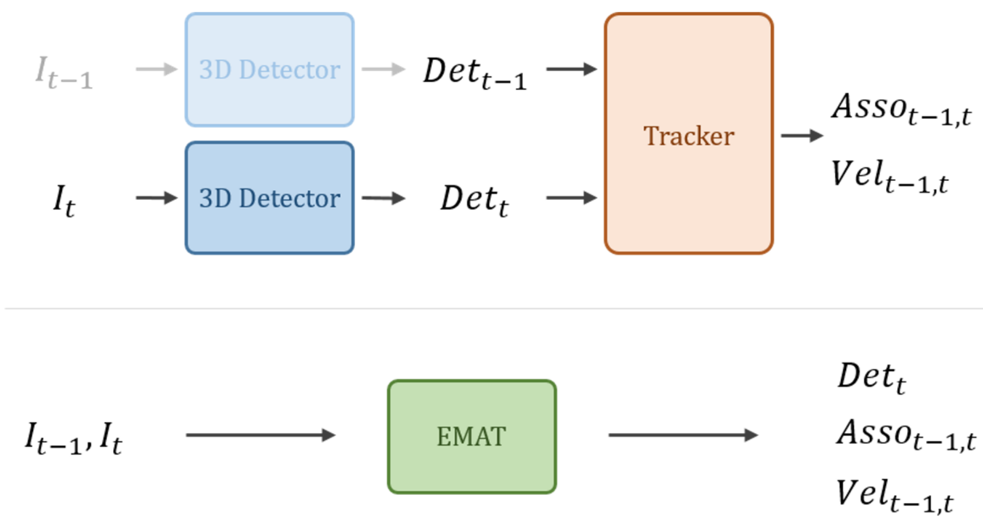

Figure 1.1: **Brief Comparison Between Tracking-by-Detection Pipeline with Our Proposed Framework EMAT.** Traditional tracking-by-detection paradigm first performs 3D detection frame by frame, and then a tracker is used for establishes correspondence between the objects in different frames to form trajectories, along with object velocity estimation. EMAT takes multi-view images input I_{t-1} and It at both t-1 and t timestamps, and simultaneously performs 3D detection Det_t, data association $Asso_{t-1,t}$ and object velocity estimation $Vel_{t-1,t}$. Best viewed in color.

Chapter 2

Related Work

Multi-View Camera-based 3D Object Detection For camera-based 3D object detection in the field of autonomous driving, early methods [4, 23, 24, 26, 29, 31] mainly estimate the 3D bounding boxes and categories in the single image view. To perform multi-view 3D detection, a naive solution is first to process different views separately with 3D detector and then utilize cross-camera postprocessing.

To make full use of information across cameras and take efficiency into account, recent works attempt to conduct 3D object detection in 3D world space given multi-view images as input. Based on view transformation methods from perspective view (PV) to bird's eye-view (BEV), most multi-view detection works can be divided into two main categories [25]. One branch of works [13,14,19,29] follow LSS [27], predicting depth distribution and generate pseudo point clouds with probability-weighted image features for BEV object detection. Another paradigm follows DETR3D [35]. Extending from DETR [35] and Deformable DETR [43], DETR3D defines sparse 3D object queries interacting with extracted multi-view image features by iterative transformer attention layers. PETR [21] and PETRV2 [22] further introduce 3D position-aware representations. BEVFormer [20] uses deformable transformer to aggregate image features and uses temporal attention to fuse the sequential BEV features. As temporal information plays an important role in predict the motion state of dynamic objects, more recent works get better 3D detection and velocity estimation performance by leveraging multiframe features [13, 18–20, 22]. Under this trend of exploiting temporal information, a joint 3D detection and tracking framework is proposed in this work.

3D Multi-Object Tracking Following the common tracking-by-detection paradigm, 2D multi-object tracking has been studied extensively in recent years. In autonomous driving, most modern 3D trackers still follow this pattern, where they perform 3D detection frame by frame and then associate them with previous tracklets via several hand crafted cues, and finally followed by a heuristic matching step, either using a greedy matching or Hungarian matching algorithm [16]. Many methods estimate motion state of tracklets through 3D Kalman filtering, then perform association using 3D IoU [37] or L2 distance [40]. Except for location of objects, more cues like velocity [40] and appearance features [7, 9, 12, 38] are explored for a more robust association. Apparently, there are still much left for improvement with regard to these approaches, where treating detection and association separately, with limited handcrafted cues and complicated post-process procedures. To this end, we jointly learn multi-camera detection and tracking in the 3D space by exploiting the framework of transformers [33] in this work.

Chapter 3

Method

In this section, we first elaborate overall architecture of EMAT, and then getting into the specifics of our designs. we introduce how to conduct end-to-end multi-view 3D detection and tracking by proposed EMAT framework in 3.2, how to aggregate temporal features for data association by proposed ATF module in 3.3 and how to boost object velocity estimation by proposed TVE strategy in 3.4.

3.1 Overall architecture

The overall framework of EMAT can be built upon off-the-shelf transformer-based multi-view 3D detector [15, 20–22, 35], where takes multi-view images as input, followed by a image feature extractor backbone and then multi-level image features and learnable object queries are interacted by iterative transformer attention layers to generate instance-level embedding feature and finally get 3D detection results via detection head.

As illustrated in Figure 3.1, we extend the basic transformer-based multi-view 3D detector to simultaneously detect and associate objects across time. Given the images captured by on-board cameras from N_v views at timestamp t, the multi-view images $I_t \in R^{H \times W \times C \times N_v}$ are fed to the backbone network (e.g. ResNet [11]) to obtain the image features. H, W, C denotes height, width and number of channels of images, respectively. With preserved instance-level embedding features E_{t-1} at the prior timestamp t-1, we aggregates temporal information from E_{t-1} and E_t at time t, by proposed association-oriented temporal fusion module. Then we build prediction heads upon these fused features for association and velocity estimation. We estimate the object appearing information and affinity across frames to produce association and generate tracking IDs directly. For velocity prediction, we propose trustworthy velocity estimation strategy to reduce object velocity error with objects appearing information.

Each sample of the video sequence will be fed into EMAT in chronological order during online inference. The instance embedding features E_t are preserved for the next.

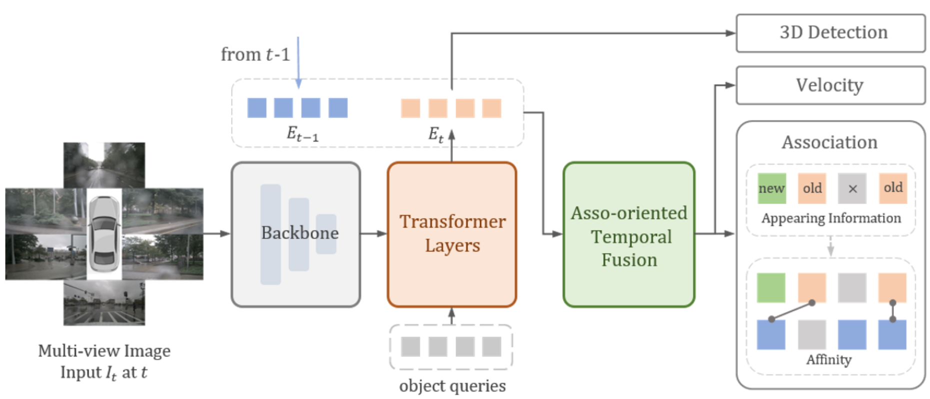

Figure 3.1: **Schematic illustration of our framework.** EMAT extends the basic transformer-based multi-view 3D detector to simultaneously detect and associate objects across frames. we aggregates temporal information from E_{t-1} and E_t by association-oriented temporal fusion module. We estimate the object appearing and affinity across frames to produce association $Asso_{t-1,t}$ and generate tracking IDs without any post-process. We propose trustworthy velocity estimation strategy to reduce object velocity estimation error.

3.2 Objects Appearing and Affinity Estimation

In order to establish correspondence across time to produce data association $Asso_{t-1,t}$ and generate tracking IDs without any post-process, we estimate the dynamic objects appearing information and affinity from features produced by Association-oriented Temporal Fusion module described in the next subsection. Specifically, for each dynamic instance assigned ground truth at time t, we predict its appearing status by a binary category classification:

1. new-born object that only appears at time t.

2. old object that appears at both t-1 and t.

As new-born objects will be assigned a new tracking ID, we focus on relation of sequential old objects.

Then, for the set of old objects

$$B_t^{old} = \{b_t^{old}, i = 1, 2, ..., N_t^{old}\} \tag{3.1}$$

at timestamp t, we further predict the object affinity with filtered dynamic objects

$$B_{t-1}^{flt} = \{b_{t-1}^{flt}, i = 1, 2, ..., N_{flt}\} \in B_{t-1} \tag{3.2}$$

Where the B_{t-1} is the full set of objects at t-1, N_t^{old} is the number of old objects at time t, and N_{flt} is a pre-defined fixed number smaller than number of object query N_q and B_{t-1}^{flt} is the top-N_{flt} confidence objects within B_{t-1}. Based on the temporal fused features which embed rich instance-level information, we model the affinity estimation as a multi-category classification problem. We perform a N_{flt}-category classification for each old object $b_t^{old} \in B_t^{old}$ at time t. For the i^{th} old object ${b_t^{old}}_i$, we predict the corresponding index idx_i which denotes the index of filtered dynamic objects B_{t-1}^{flt}. The classification score normalized by softmax to 0-1, can be used as similarity for tracking, and the classification results can be used for association directly.

During inference, the appearing head distinguishes old or new-born objects. We keep the tracking IDs of old objects and assign new-born objects with new ones according to the prediction of affinity. Object trajectories are yielded with the assignment process frame by frame.

3.3 Association-oriented Temporal Fusion

Most previous works towards vision-based 3D detection try to exploit the temporal information by simply stacking spatial-aligned BEV features [13, 22] or pseudo point clouds [19] from several frames. Since all these temporal fusion strategies are designed to boost 3D detection and temporal attributes like velocity, the fused features may not be friendly for data association. To this end, instead of stacking features from different frames or adding a FFN on weights of cross-attention to producing affinity matrix [18], we design Association-oriented Temporal Fusion (ATF) module, that utilizes cross-attention to aggregate inter-frame features for learning to association.

To be specific, as illustrated in Figure 3, first, we combine implicit object query embedding feature Emb and explicit location information Loc decoded by regression branch to generate combined feature

$$CF = \mathbf{MLP}(Emb \oplus Loc) \tag{3.3}$$

by MLP layer, where weights are shared across frames and $\oplus$ denotes concatenation. Then, multi-head cross attention is adopt for aggregating objects information across time. In detail, we use the combined feature CF_t at time t as query, and the combined feature CF_{t-1} at time t - 1 as key and value for multi-head cross-attention:

$$FQ = \mathbf{MHA}(CF_t, CF_{t-1}, CF_{t-1}) \tag{3.4}$$

Where **MHA** denotes multi-head attention.

The learnable attention weight matrix AW explores the correspondence of detected objects in different frames and contains matching information [18]. So, both AW and output fused embedding feature FQ are concatenated for generating final feature

$$FA = \textbf{MLP}(AW \oplus FQ) \tag{3.5}$$

for learning to associate by another MLP layer.

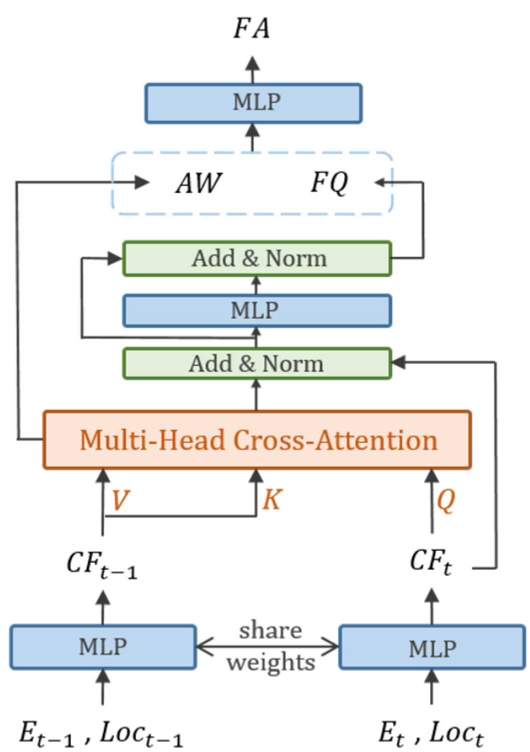

Figure 3.2: **Association-Oriented Temporal Fusion Module.** Using implicit object query embedding feature and explicit location information as input, multi-head cross-attention is adopt for aggregating objects information across frames. Attention weight matrix and fused feature are both used for generating final feature.

3.4 Trustworthy Velocity Estimation

Since velocity is a temporal attribute of dynamic objects, information from multi-frame is a necessity for valid velocity estimation. However, when an object is new-born, that is, only appearing at timestamp t instead of t-1, this is no temporal cues for velocity estimation of this object at timestamp t. Moreover, forcibly learning velocity of objects which only appear at a single frame brings into much noises for training, leading to lower velocity estimation performance. Therefore, we propose Trustworthy Velocity Estimation (TVE) strategy, which make sure the velocity estimation branch focus on objects with temporal cues, leading to reasonable and trustworthy velocity estimation.

In contrast to previous works, which widely use standard L_1 loss for all assigned object samples when training, we consider object appearing information for velocity learning. Specifically, we treat new-born and old objects differently through the velocity loss of TVE L_{vel}, defined as:

$$L_{vel} = \lambda_{new} \cdot l_{vel}^{new} + \lambda_{old} \cdot l_{vel}^{old} \tag{3.6}$$

where l_{vel}^{new}, l_{vel}^{old} present velocity loss generated by newborn objects and old objects, respectively. We set λ_{new} far smaller than λ_{old} to softly ignore new-born objects and ensure the velocity learning to focus on objects with temporal cues, avoiding velocity training noises by new-born objects. We adopt standard L_1 loss for l_{vel}^{new} and l_{vel}^{old}.

3.5 Training Loss

Following prior transformer-based works [6, 35], a set-to-set loss, of which the assignment is solved by Hungarian algorithm, is adopt for all of loss items in EMAT. The multi-task losses can be divided into three parts: multi-view 3D object detection loss L_{det}, object velocity loss L_{vel} introduced in 3.4 and association loss L_{asso}. Follows DETR3D [35], the multi-view 3D object detection loss consists of two major parts, a focal loss [30] for object category classification and a L_1 loss for object bounding box parameters regression, of which including location, dimension and orientation in 3D space. For association loss L_{asso}, we apply cross entropy loss L_{CE} for both object appearing and affinity. The object appearing loss is formulated as:

$$\mathcal{L}_{app} = \sum_{i=1}^{N_p} L_{CE}(\hat{c}_i, c_i) \tag{3.7}$$

where $\hat{c}$ is the predicted appearing categories in new-born, old, c is the ground truth categories, and N_p is the number of assigned positive samples by Hungarian algorithm [16] as most set-to-set detection methods [6, 20, 35, 43]. For each old object $b_t^{old} \in B_t^{old}$ at time t, we perform a N_{flt}-category classification. The affinity loss is formulated as:

$$\mathcal{L}_{aff} = \sum_{i=1}^{N_i^{old}} L_{CE}(i\hat{d}x_i, idx_i) \tag{3.8}$$

where idx_i denotes the index of filtered dynamic objects B_{t-1}^{flt} that the i_th old object $b_t^{old}i$ is corresponding to.

To summarize, EMAT is supervised with the combination of these loss items:

$$\mathcal{L} = \lambda_{det}\mathcal{L}_{det} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{app}\mathcal{L}_{app} + \lambda_{aff}\mathcal{L}_{aff} \tag{3.9}$$

where λ_{det}, λ_{vel}, λ_{app} and λ_{aff} are the loss weight of detection, velocity, appearing and affinity, individually.

Chapter 4

Experiments

4.1 Dataset and Metrics

We validate our proposed method on nuScenes dataset [5], a challenging large-scale public dataset in the field of autonomous driving. It contains 1000 scenes and is officially divided into 700/150/150 for training/ validation/testing set, respectively. Each sample consists of images from 6 cameras, and is fully annotated with 3D bounding boxes of 10 categories at 2hz. 7 of 10 object categories are also provided with tracking annotation.

For 3D object detection task, the evaluation is performed on the full 360° panorama. Followed the official evaluation protocol, we report mean Average Precision (mAP) and nuScenes Detection Score (NDS). NDS is computed by five metrics, including mATE, mASE, mAOE, mAVE, and mAAE defined for measuring translation, scale, orientation, velocity, and attribute errors of true positive objects, respectively. We report Mean Average Velocity Error (mAVE) to evaluate velocity error. For 3D MOT task, also consistent with office metrics, we evaluate performance by two major metrics, Average multi-object tracking accuracy (AMOTA) and average multi-object tracking precision (AMOTP), by averaging over the multi-object tracking accuracy (MOTA) and multi-object tracking precision (MOTP) metric at different recall thresholds, respectively. We refer readers to the official paper [5] of nuScenes dataset for more details.

4.2 Experiment setting

We build up our method based on the transformer-based multi-view 3D detector BEVFormer [20] with the ResNet-DCN [10, 11] backbone. The image feature extractor use pretrained model initialized from FCOS3D [34] checkpoint. The number of object queries N_q and filtered boxes N_{flt} are set to 900 and 300. We set λ_{det} = 1.0, λ_{vel} = 0.05, λ_{app} = 0.05 and λ_{aff} = 0.25. By default, we train our model for 24 epochs, with a starting learning rate of 2e-4. The batchsize is 8 on 4 GPUs, with 1 sample per GPU. For ablation study, we use half of the whole training set and ResNet-50-DCN as backbone for memory efficiency.

4.3 Comparison with Baseline

Without relying on any external data, we use only the training set for generating the results on the testing set. The results are obtained using a single model without any testing time augmentation. As shown in Tab 4.1, our method achieves significant improvements over traditional methods on the nuScenes camera 3D tracking benchmark, reaching an AMOTA of 37.0% and an AMOTP of 1.12m. Notably, it surpasses the best traditional method by a substantial margin of 9.3% in AMOTA.

We also compare this work with other vision-based methods which report both 3D detection and tracking performance on the nuScenes test set. As listed in Tab 4.2, our work outperforms the second place on the tracking metrics AMOTA and AMOTP, particularly on AMOTA by a large margin of 9.3%, and on velocity metric mAVE by a significant improvement of performance from 0.845 m/s to 0.326 m/s, while remaining comparable 3D detection performance evaluated by mAP and NDS.

Method	AMOTA(%) ↑	AMOTP(m) ↓
CenterTrack [40]	4.6	1.54
DEFT [7]	17.7	1.56
Time3D [18]	21.4	1.36
QD-3DT [12]	21.7	1.55
MUTR3D [41]	27.0	1.49
PolarDETR [8]	27.3	1.39
EMAT	**37.0**	**1.12**

Table 4.1: **Comparison with camera-based tracking methods on the nuScenes test set.** Our method demonstrates competitive performance for the camera-based 3D MOT task, achieving an AMOTA that surpasses traditional methods by a notable margin of 9.3% on the test split, without relying on any external data for training.

Method	mAP(%) ↑	NDS(%) ↑	mAVE(m/s) ↓	AMOTA(%) ↑	AMOTP(m) ↓
MonoDIS	30.4	38.4	1.553	1.8	1.79
Time3D	31.2	39.4	1.523	21.4	1.36
PolarDETR	**43.1**	49.3	0.845	27.3	1.19
EMAT	42.2	**52.9**	**0.326**	**37.0**	**1.12**

Table 4.2: **Comparison with camera-based detection and tracking methods on nuScenes test set.** EMAT outperforms the second place for 3D tracking task and velocity estimation by a large margin, with comparable 3D object detection performance.

4.4 Ablation Study

Association-oriented Temporal Fusion Module. First, we examine the effect of proposed ATF module by comparing the training loss. Figure 4.1 compares several settings of different fusion methods and involved features for learning to association. Exp. A is the baseline fusion setting, where using concatenated embedding features. Exp. B utilizes cross-attention to aggregate object query embedding features to produce fused feature for association. Instead of final fused features, Exp. C treats the learnable attention weight of cross-attention as the feature for association. Exp. D fuses the features of Exp. B and Exp. C, but considers no explicit location information. By adopting ATF module, we get lower training loss than other settings, revealing the effectiveness

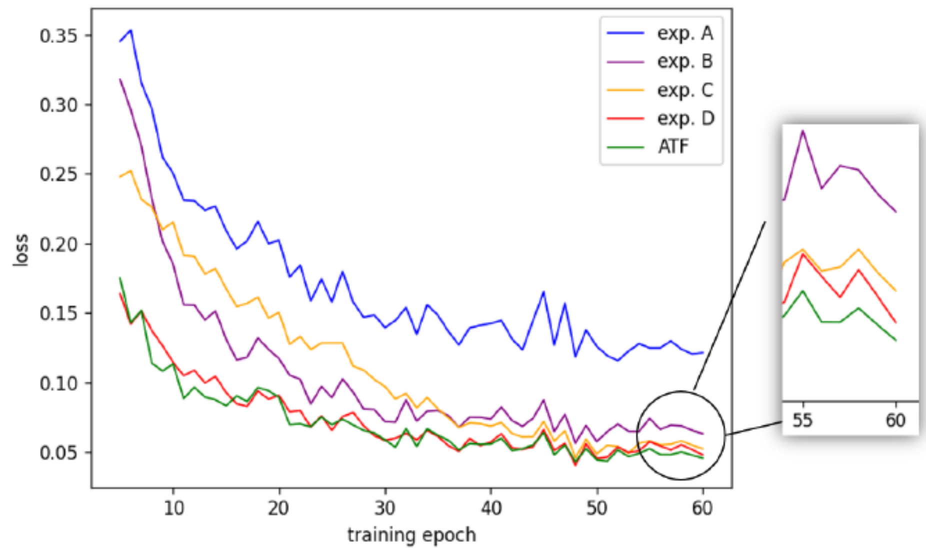

Figure 4.1: **Ablation on ATF module.** ATF module (plotted in green) achieve lower training loss than other settings. Best viewed in color.

of this module.

Trustworthy Velocity Estimation Strategy. Second, we study the effect of TVE strategy. As shown in Tab 4.3, training our models with proposed TVE strategy is beneficial for object velocity estimation performance. TVE reducing the velocity error metric mAVE from 0.6704 m/s to 0.6211 m/s when using two frames for training, and from 0.5269 m/s to 0.4918 m/s for four frames. Besides, the 3D detection metric mAP and NDS are also slightly improved. By softly ignore new-born objects and focusing on objects with temporal cues, more accurate and valid cues are learned to estimate velocity.

4.5 Qualitative Results

Figure 4.2 shows some qualitative detection and tracking results in BEV space for 3 different scenes of 4 seconds clips. EMAT produces stable tracking IDs across time for dynamic objects on the full 360° panorama by six cameras.

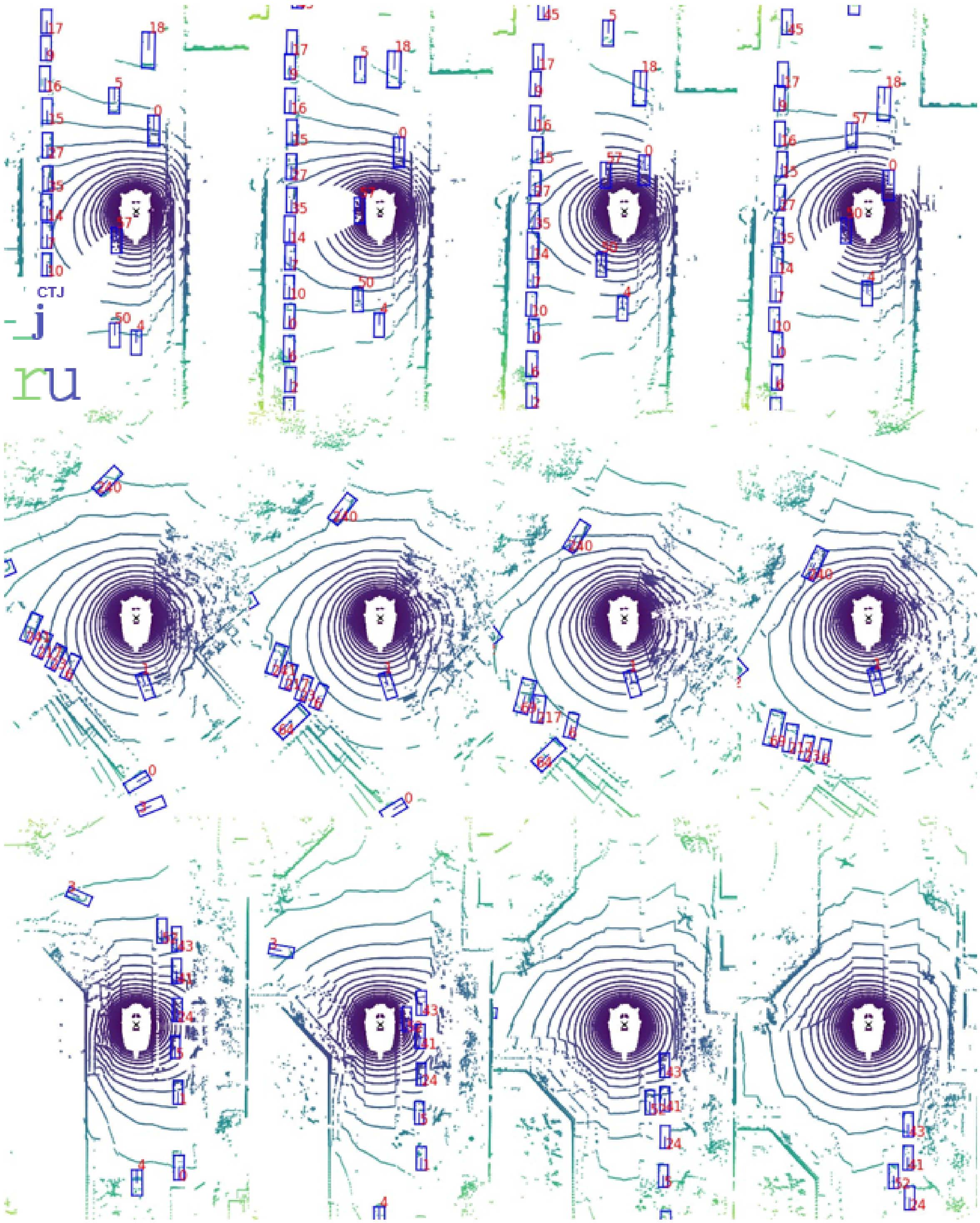

Figure 4.2: **Qualitative Results.** We visualize the 3D box in the BEV space and tracking ID on 4 consecutive frames with 1 FPS in each row. Stable tracking IDs for objects in BEV space are shown by these examples. LiDAR point clouds are only used for the purpose of visualization.

Exp. Setting	num. frame	mAVE(m/s) $\downarrow$	mAP(%) $\uparrow$	NDS(%) $\uparrow$
w/o TVE	2	0.6704	33.60	42.54
w/ TVE	2	0.6211	33.85	43.12
w/o TVE	4	0.5269	33.82	44.37
w/ TVE	4	**0.4918**	**34.15**	**45.06**

Table 4.3: **Ablation on TVE strategy.** "num. frame" denotes the number of frames during training. Training our models with the proposed TVE strategy is beneficial for object velocity estimation for both 2 and 4 frames used for training.

Chapter 5

Conclusion

In this work, we design a novel end-to-end multicamera joint 3D detection and MOT framework EMAT. On the nuScenes dataset, we outperform the 3D detection baseline we build upon, achieving superior camera-based 3D tracking and velocity estimation performance, along with notable improvements over traditional 3D tracking methods, showcasing its effectiveness in real-world scenarios. We have shown that the association-oriented temporal fusion module ATF effectively aggregates features from different frames, and proposed trustworthy velocity estimation strategy TVE reduces velocity error. Compared with traditional tracking-by-detection paradigm with hand-designed associating strategies and cumbersome post-processes, our proposed approach may inspire researchers to jointly optimize 3D detection and MOT task, which is of considerable practical significance especially in vision-based autonomous driving field.

Bibliography

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[2] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[3] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020.

[4] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9287–9296, 2019.

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

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

[8] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022.

[9] Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 14227–14233. IEEE, 2021.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] 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*, 45(2):1992–2008, 2022.

[13] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

[14] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[15] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023.

[16] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[17] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

[18] Peixuan Li and Jieyu Jin. Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3885–3894, 2022.

[19] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485, 2023.

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[21] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[22] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023.

[23] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 996–997, 2020.

[24] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6851–6860, 2019.

[25] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[26] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.

[28] Zipeng Qin, Jianbo Liu, Xiaolin Zhang, Maoqing Tian, Aojun Zhou, Shuai Yi, and Hongsheng Li. Pyramid fusion transformer for semantic segmentation. *IEEE Transactions on Multimedia*, 2024.

[29] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021.

[30] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

[31] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019.

[32] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):104–119, 2019.

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[34] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.

[35] 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 *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[36] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European conference on computer vision*, pages 107–122. Springer, 2020.

[37] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020.

[38] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[40] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[41] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2022.

[42] 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*, 129:3069–3087, 2021.

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.