# CoopTrack: Exploring End-to-End Learning for Efficient Cooperative Sequential Perception
# — Supplementary Material —

Jiaru Zhong[1,2] Jiahao Wang[1] Jiahui Xu[3] Xiaofan Li[4] Zaiqing Nie[1]* Haibao Yu[3,1]*
[1] Tsinghua University [2] The Hong Kong Polytechnic University
[3] The University of Hong Kong [4] Baidu Inc.
zhong.jiaru@outlook.com, zaiqing@air.tsinghua.edu.cn, yuhaibao94@gmail.com

## A. Appendix Overview

In the appendix, we present 1) additional implementation details in Sec. B; 2) extended experimental results in Sec. C, covering inference speed analysis, backbone architecture comparisons, ablation studies on roadside data contribution, communication latency tests, and pose estimation error impacts; 3) qualitative analyses of instance association and tracking results in Sec. D.

## B. More Details of Experiments

### B.1. Implementation Details

We use two versions of the backbone: ResNet50 and ResNet101. For ResNet50, we crop the images to 540 × 960 and set the BEV feature size to 50 × 50. For the larger one, we keep the input image size unchanged and set the BEV size to 200 × 200. For Griffin, we only trained the ResNet-50 version. To reduce memory consumption, we adopt a streaming video training approach [9, 11], which may slow down model convergence. Consequently, for V2X-Seq, the vehicle and infrastructure models are trained for 48 and 24 epochs respectively in the first stage, while the cooperative model undergoes 48 epochs of training. For Griffin, the vehicle and UAV models are each trained for 48 epochs in the first stage, followed by 48 epochs of training for the second-stage cooperative model.

### B.2. Baseline Settings

To demonstrate the superiority of our method, we compare it with several existing cooperative tracking methods: (1) No Fusion: This baseline uses only the ego vehicle's images as input and does not activate the cooperative module. (2) Late Fusion + AB3DMOT [12]: This method follows the tracking by cooperative detection paradigm, where detection results from multiple agents are fused, and the cooperative detection results are fed into the classic tracking method AB3DMOT [12]. For a fair comparison, we use BEVFormer [6] as the detector and implement late fusion following DAIR-V2X [14]. (3) BEV Feature Fusion + AB3DMOT [12]: This method also follows the tracking by cooperative detection paradigm but uses feature fusion for cooperative detection. Specifically, we use BEVFormer [6] as the detector, align the two BEV features spatially based on relative positions, and then fuse the concatenated BEV features using a multi-layer convolution neural network before feeding them into the decoder. (4) UniV2X [17]: This is the first end-to-end cooperative planning method, which includes modules for tracking, mapping, occupancy, and planning. Since we focus on the 3D MOT task, we retain only the Agent Fusion module and the tracking framework, referred to as UniV2X-Track, which belongs to the end-to-end cooperative tracking paradigm mentioned earlier. (5) Other SOTA methods, including V2X-ViT [13], where2comm [3], DiscoNet [5], CoAlign [8]. For a fair comparison, they use the same inputs and evaluation settings as ours. All methods, except for CoCa3D [4] based on depth estimation, are implemented using BEVFormer [6].

| Backbone | mAP↑ | AMOTA↑ |
|---|---|---|
| ResNet101 | 0.390 | 0.328 |
| ConvNeXt-S | 0.413 (+0.023) | 0.429 (+0.101) |

Table 1. **Performance of Different Backbones on the V2X-Seq.**

| Method | MDFE | CAA | GBA+Aggr. | Total |
|---|---|---|---|---|
| CoopTrack-ResNet50 | 42.01 | 2.08 | 9.95 | 121.88 |
| CoopTrack-ResNet101 | 131.91 | 2.04 | 8.68 | 207.99 |

Table 2. **Runtime of Key Modules of CoopTrack** *(unit: ms)*. MDFE stands for the multi-dimensional feature extraction module, CAA for the cross-agent alignment module, GBA for the graph-based association module, and Aggr. for feature aggregation.

*Corresponding authors.

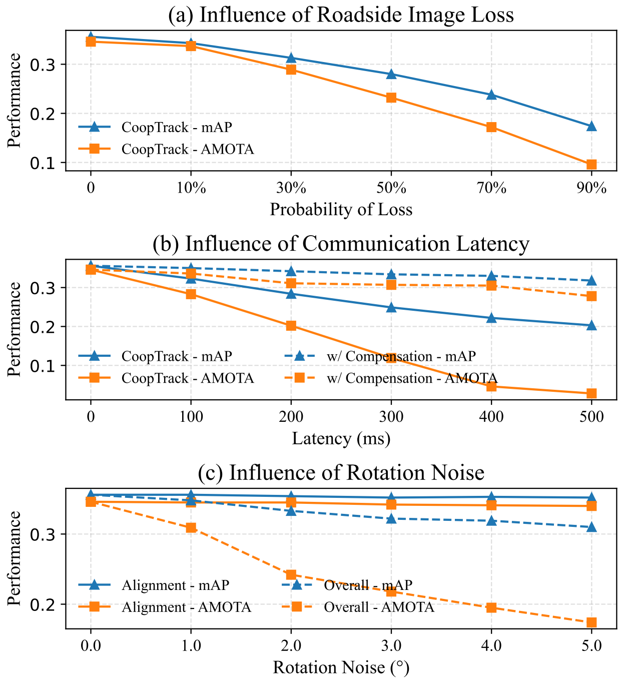


Figure 1. **Comparison of Performance in Different Conditions.**

### B.3. Evaluation Metrics

To assess performance, we utilize widely recognized metrics in 3D object detection and tracking [1], among which the primary indicators are Mean Average Precision (mAP) and Average Multi-Object Tracking Accuracy (AMOTA). Furthermore, to evaluate the transmission costs inherent in cooperative approaches, we employ Bytes per second (BPS) as another essential metric [14, 16]. To ensure fair comparison with existing methods, we follow the evaluation protocols established by UniV2X [17] and Griffin [10], reporting performance metrics exclusively for the vehicle category.

## C. More Experiments

### C.1. Recent Image Backbone

We upgrade the image feature extraction backbone by replacing ResNet101 [2] with the more recent ConvNeXt-Small [7] architecture. As demonstrated in our V2X-Seq experiments (see Tab. 1), this modification yields significant performance gains of +2.3% mAP and +10.1% AMOTA, confirming the scalability of our approach through backbone compatibility.

### C.2. Inference Speed

To evaluate computational efficiency, we measure the average inference time on a single NVIDIA RTX 3090 GPU across the V2X-Seq validation set, with detailed results presented in Table 2. It can be observed that the time consumption is primarily concentrated in the multi-dimensional feature extraction module, while cross-agent alignment and aggregation do not take much time. Consequently, with the ResNet50 backbone, CoopTrack achieves nearly real-time performance at approximately 10Hz.

### C.3. Ablation Study of Infrastructure Images

To investigate CoopTrack's dependence on roadside data, we evaluate the model without infrastructure image inputs. As shown in Fig. 1(a), while performance degrades without roadside images, the system still surpasses the No Fusion baseline (0.110 mAP and 0.087 AMOTA), demonstrating the inherent robustness of our approach. This suggests that while roadside information enhances perception accuracy, the framework maintains functional capability when operating independently.

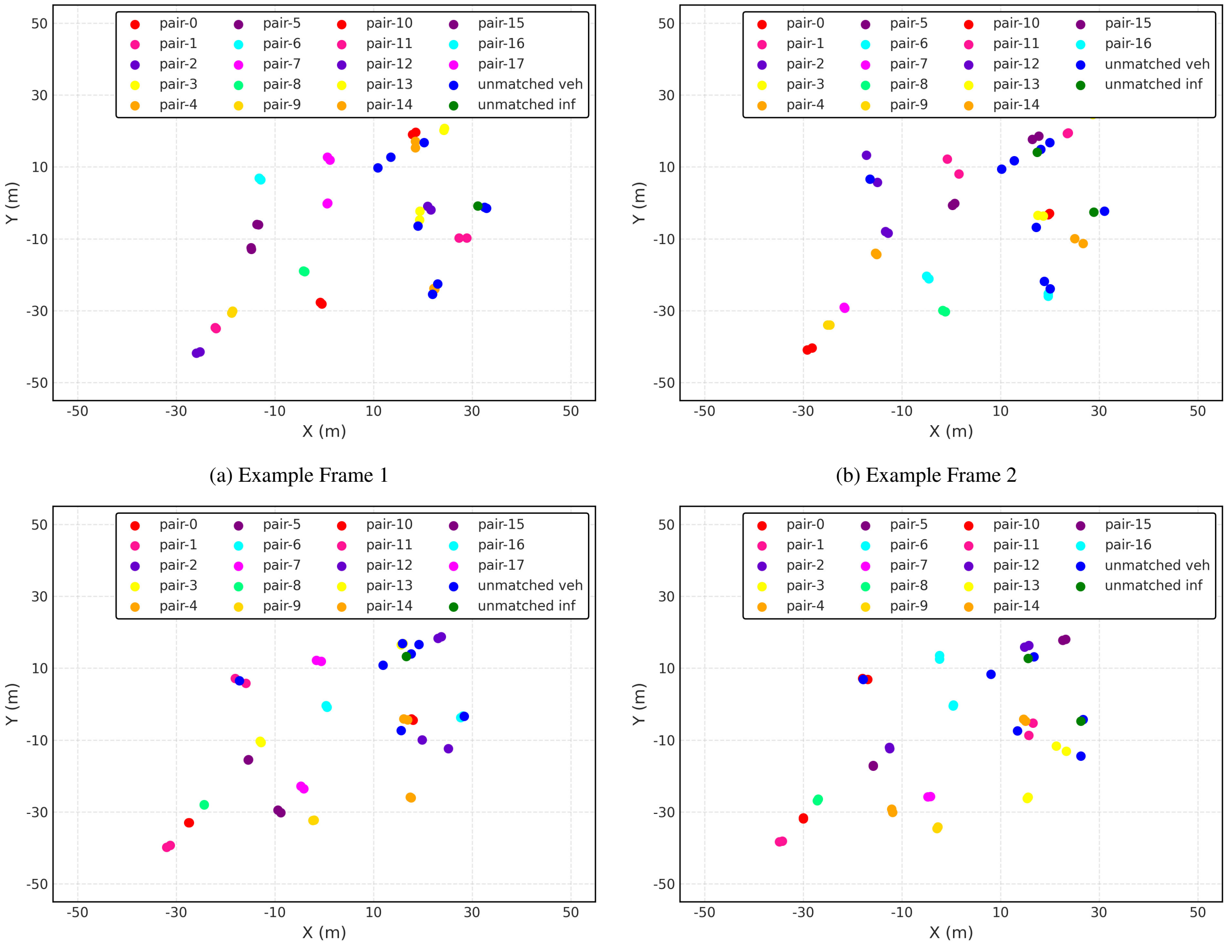
### C.4. Influence of Communication Latency

As a critical challenge in real-world cooperative perception systems, communication latency induces spatiotemporal misalignment between cooperative data and ego-vehicle observations, significantly degrading perception performance [15]. To analyze its impact on CoopTrack, we simulate delayed infrastructure-to-vehicle communication by introducing artificially lagged roadside data. As shown in Fig. 1(b), experimental results demonstrate that under 500ms latency, the system exhibits substantial performance degradation of 15.3% in mAP and 31.8% in AMOTA, highlighting the importance of latency mitigation for practical deployment.

To mitigate this issue, we introduce the feature flow prediction module [15] that leverages historical query states to learn temporal dynamics and calibrate incoming infrastructure queries based on their timestamps. As shown in Fig. 1(b), this module significantly enhances CoopTrack's robustness to latency, limiting performance degradation under 500ms delay to just 3.8% in mAP and 6.8% in AMOTA, a marked improvement over the uncompensated baseline. While this method reflects substantial progress, further research is needed to develop advanced temporal modeling techniques and adaptive compensation strategies tailored to variable latency.

### C.5. Impact of Rotation Noise

Cross-agent feature fusion relies on accurate relative poses between agents, where pose estimation errors can cause spatial misalignment and degrade cooperative perception performance. To investigate this effect, we follow V2X-ViT [13]'s methodology by injecting noise into rotation matrices under two settings: (1) noise applied solely to the cross agent alignment module's inputs, and (2) global noise applied throughout the framework. As illustrated in Fig. 1(c), global noise causes significant performance degradation, while introducing noise solely to the alignment module results in marginal decline. This reveals that although the



(a) Example Frame 1

(b) Example Frame 2

(c) Example Frame 3

(d) Example Frame 4

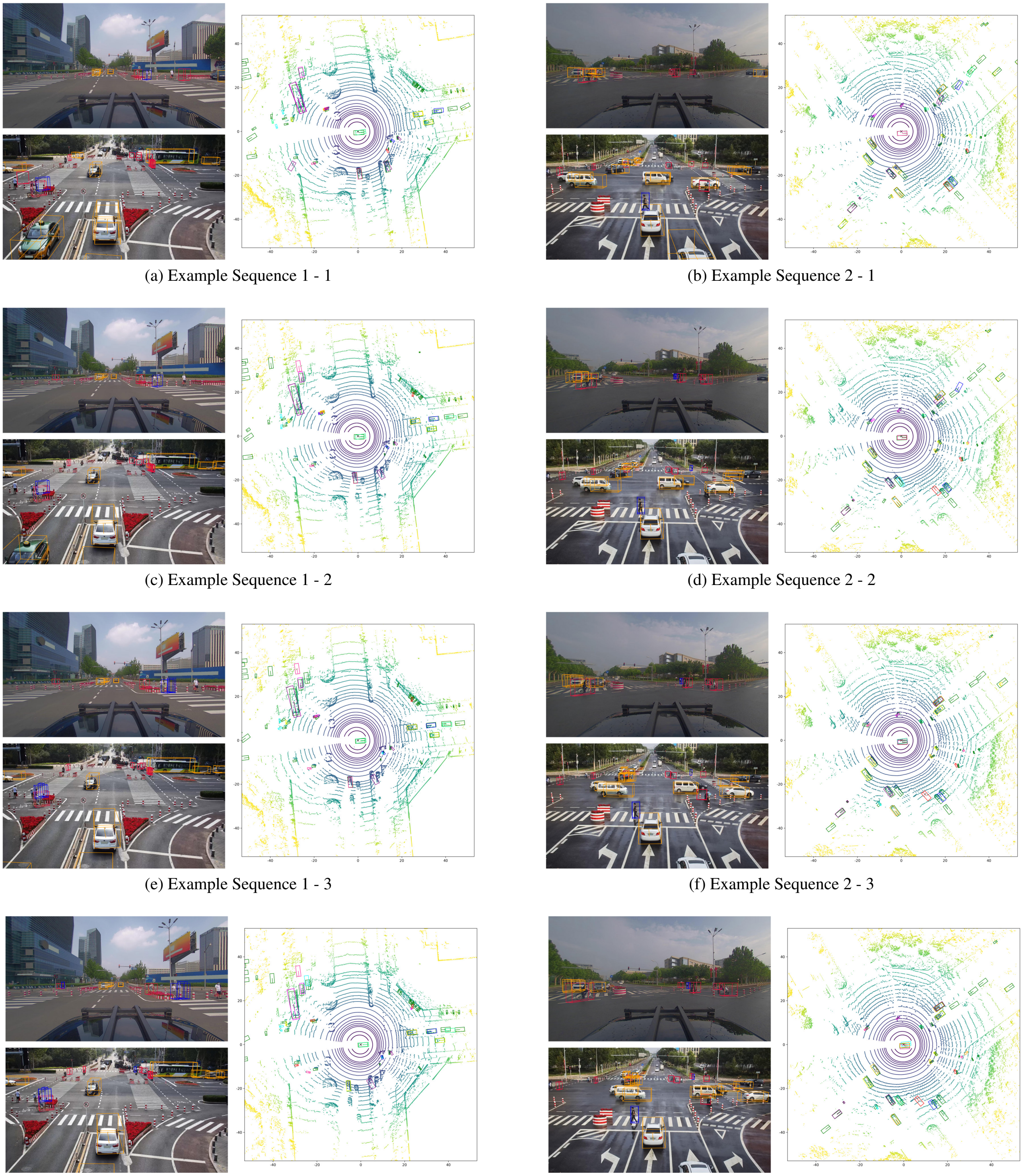
Figure 2. **Visualization of CoopTrack's association results on the V2X-Seq Dataset.** We visualize the association results of several key frames in a sequence by plotting instances within the ego-vehicle-centric BEV coordinate frame, where each instance is represented by its reference point. Matched instance pairs, including a vehicle instance and an infrastructure instance, are uniquely color-coded according to the ID, while unmatched vehicle and roadside instances are marked distinctly in blue and green, respectively. Note that due to the close proximity of instances, there is overlap in the figure.

alignment module takes pose parameters as input, it learns additional implicit information from features during training to achieve robust multi-agent feature alignment in latent space. The observed system-level sensitivity primarily stems from reference point perturbations rather than the alignment mechanism itself. Reducing the impact of pose noise remains an important direction for future research.

## D. Qualitative Analyses

### D.1. Association Results

As shown in Fig. 2, we visualize instance association results for specific frames in a sequence within the ego-vehicle-centric BEV coordinate frame, where each instance is represented by its reference point. Successfully matched instance pairs, including both vehicle and infrastructure instances, are assigned unique IDs and labeled with colors from a predefined palette based on their IDs. In contrast, unmatched vehicle instances and roadside instances are displayed in blue and green, respectively. Note that due to the close proximity of instances, there is overlap in the figure. These results demonstrate that our association approach achieves pairing without relying solely on Euclidean distance, exhibiting strong robustness. For example, consider a pair of instances located within the 20–30 meter range along the x-axis and near -10 meters along the y-axis (pair-11 in Fig. 2(a), pair-14 in Fig. 2(b), pair-12 in Fig. 2(c), and pair-13 in Fig. 2(d)). Due to inevitable inaccuracies in the reference points of vehicle and roadside instances, their relative distance varies and reaches a maximum in Fig. 2(c) yet stable association is maintained. This is because our association module comprehensively consid-



(a) Example Sequence 1 - 1

(b) Example Sequence 2 - 1

(c) Example Sequence 1 - 2

(d) Example Sequence 2 - 2

(e) Example Sequence 1 - 3

(f) Example Sequence 2 - 3

(g) Example Sequence 1 - 4

(h) Example Sequence 2 - 4

Figure 3. **Visualization of CoopTrack's tracking results on the V2X-Seq Dataset.** The two columns display two different sequences from the validation split of the dataset, with four rows representing consecutive time steps. Each subfigure is divided into three parts: the top-left shows results from the vehicle perspective, the bottom-left shows results from the roadside perspective, and the right side presents the Bird's Eye View (BEV) visualization. In the forward-looking perspective, 3D bounding boxes are color-coded by object category: orange for vehicles, red for cyclists, and blue for pedestrians. In the BEV, the LiDAR point cloud is visualized for better presentation. For bounding boxes, we use green to represent the ground truth, while the colors of tracking results are randomly selected from a pool of colors based on their IDs, ensuring that each object maintains a consistent color over time.

ers multi-dimensional instance features and relative positional relationships, thereby providing more reliable information for downstream aggregation module.

## D.2. Tracking Results

As illustrated in Fig. 3, we visualize the tracking results of our proposed CoopTrack on two representative sequences from V2X-Seq [16], aiming to intuitively demonstrate the superiority of our approach. Each subfigure comprises three components: the vehicle-side input image positioned at the top-left, the roadside input image at the bottom-left, and the tracking results in the BEV view on the right. In the images, colors correspond to object categories: orange denotes vehicles, red denotes cyclists, and blue denotes pedestrians. In the BEV view, green bounding boxes represent ground truth, while colored boxes show tracking results with colors assigned by IDs to demonstrate temporal consistency. In the two sequences provided, we can observe that, thanks to the cooperative information from the roadside, the ego-vehicle can continuously track instances behind and to the sides of it, achieving comprehensive perception results. This highlights the significant advantage of cooperative perception over single-vehicle perception. Furthermore, despite relying solely on images and lacking depth information input, CoopTrack has also achieved relatively precise localization, demonstrating the accurate tracking capability of our method in complex traffic scenarios.

## References

[1] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 2

[2] 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. 2

[3] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022. 1

[4] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2023. 1

[5] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 1

[6] 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 lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[8] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4812–4818. IEEE, 2023. 1

[9] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 1

[10] Jiahao Wang, Xiangyu Cao, Jiaru Zhong, Yuner Zhang, Haibao Yu, Lei He, and Shaobing Xu. Griffin: Aerial-ground cooperative detection and tracking dataset and benchmark. *arXiv preprint arXiv:2503.06983*, 2025. 2

[11] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3631, 2023. 1

[12] 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. 1

[13] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 1, 2

[14] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 1, 2

[15] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36:34493–34503, 2023. 2

[16] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 2, 5

[17] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025. 1, 2