Scalable Crowd-Sourced Global HD Map Construction via Collaborative Map Perception and Sparse Graph Fusion

Ruiyang Zhu[1†] Minkyoung Cho[1†] Shuqing Zeng[2] Fan Bai[2] Xiang Gao[2] Z. Morley Mao[1]
[1]University of Michigan [2]General Motors
{ryanzhu, minkycho, zmao}@umich.edu {shuqing.zeng, fan.bai, xiang.1.gao}@gm.com

Abstract

High-definition (HD) maps are vital for autonomous driving, providing fine-grained geometric and semantic information beyond the scope of onboard perception. However, automatically constructing accurate vectorized maps at scale using learning-based methods remains challenging, as individual vehicles observe only partial, localized environments. This motivates the need for collaborative HD map construction, where multiple vehicles contribute local observations to build a unified global map. While collaborative perception has been extensively studied through dense BEV fusion, existing methods are fundamentally ego-centric and operate within a fixed perception range, making them ill-suited for large-scale, open-world mapping. In this paper, we propose a graph-based sparse fusion framework for collaborative vectorized HD map construction. Vehicles build local HD maps collaboratively and encode them as sparse geometric graphs, which are fused by a sparse-to-sparse fusion algorithm that incrementally aligns and merges graphs across space and time. This design leverages multi-agent fine-grained features and enables scalable, memory-efficient fusion without relying on dense tensors. Experimental results show that our method constructs accurate global maps under sparse and asynchronous observations, outperforming baselines by over 10.3 mAP.

1. Introduction

High-definition (HD) maps play a critical role in autonomous driving by providing rich geometric and semantic information, such as lane boundaries and pedestrian crossings. Among various formats [4, 10–13, 16, 23], vectorized HD maps — which represent map elements as structured primitives such as polylines and polygons — have gained increasing popularity due to their compactness, interpretability, and alignment with the formats used in planning and simulation [14]. In addition, vectorized representations are more memory-efficient, easier to update, and better suited for real-time applications. However, building accurate vectorized HD maps at scale remains a major challenge. Due to limited sensing range [27], occlusions, and coverage gaps, individual vehicles cannot construct reliable maps on their own. This motivates the need for ***crowd-sourced global*** HD map construction, where a fleet of vehicles (also referred to as agents) contributes local observations to incrementally build a unified global map.

Previous efforts in collaborative perception and mapping have primarily relied on **dense feature fusion**, where multi-agent sensor data are encoded and projected into a shared bird's-eye-view (BEV) representation [6, 17, 23, 24, 28, 29, 31]. These methods aim to align perspective differences among agents by aggregating features on fixed-size BEV tensors. However, they are inherently **ego-centric**, centered around the field of view of a reference ego vehicle, and typically confined to a fixed local range (e.g., 60×30 m). Consequently, dense fusion methods struggle with **scalability**, and are ill-suited for global HD mapping that must efficiently handle asynchronous, large-scale, and partially overlapping observations. Therefore, building global HD maps efficiently requires sparse and scalable fusion mechanisms. Naively aggregating dense features from all agents introduces significant redundancy and inefficiency, and is prone to misalignment errors under temporal or spatial discrepancies. To enable robust global HD map construction, a new approach must: (1) incorporate multi-agent observations over extended space and time; (2) handle spatial-temporal misalignments and inconsistencies; and (3) avoid redundant computation and memory overhead.

To address these challenges, we propose `CrowdMap`. Our key insight is to move away from dense, ego-centric fusion and instead adopt a graph-based sparse fusion framework. Vehicles collaboratively build semi-global HD map elements as a sparse geometric graph, which captures both structure and semantics. These graphs can be incrementally fused across agents and across time, without requiring dense alignment or overlapping perception fields. The resulting system is inherently scalable to large maps and large fleets while maintaining precision and resource efficiency.

[†]Equal contribution.

In summary, the main contributions of this work are:

- We propose `CrowdMap`, a scalable, graph-based framework for collaborative HD map construction from partially overlapping multi-agent observations. To our knowledge, this is the first *multi-agent framework that collaboratively constructs the global vectorized HD map*.
- We develop a sparse-to-sparse graph encoding/decoding algorithm that efficiently integrates multi-agent and cross-time maps, with mechanisms to handle scalable spatial-temporal misalignment via simple yet effective overlapping detection and point-set registration.
- Different from prior HD map work that evaluates single-vehicle datasets such as nuScenes [3] and Argoverse [22], we introduce the multi-agent crowd-sourced global HD map construction task and evaluate the performance using real-world multi-vehicle datasets [26].

2. Method

2.1. Problem Definition and Solution Sketch

HD map is a collection of vectorized static map elements, including pedestrian crossings, lane dividers, *etc*. In this work, we aim to build a framework for **global HD map** construction, where the generated map spans large spatial regions far beyond the sensing range of a single vehicle.

Challenges in building global HD maps. Existing local HD map models suffer from *small perception range*. For example, MapTR [13] only constructs HD maps with a small range of 60 × 30 m around the ego vehicle. StreamMapNet [27] extends it to 100 × 50 m. However, these methods still focus on a relatively small area compared to constructing a global HD map. Building an end-to-end global HD map construction framework faces two main challenges: (1) Single-frame data from one vehicle is insufficient for building large regions; (2) Scaling transformer-based BEV models [14, 16] to global regions is computationally infeasible due to the size of dense feature maps.

Limitation of existing global HD map construction methods. Prior work [5, 20, 21, 25] on global HD map construction can be broadly categorized based on their fusion strategies: (1) *late-fusion:* Methods like PolyMerge [20] merge vectorized map instances (e.g., polylines) from multiple local frames using rule-based heuristics like proximity thresholds. However, this coarse-grained instance-level merging overlooks fine-grained features across agents or time, leading to suboptimal map quality. (2) *temporal intermediate-fusion:* Recent methods [5, 25, 27] incorporate temporal priors to aggregate historical map predictions from a single vehicle. While this approach enables global map construction by accumulating sequential data over time, it is limited to **single-agent** settings and cannot take advantage of crowdsourced data from multiple vehicles for broader and more comprehensive map construction.

Solution sketch. To address the above limitations, we propose a novel framework for collaborative global HD map construction, named `CrowdMap` (Fig. 1). The proposed framework consists of two main components: an enhanced semi-global collaborative map model and a sparse graph-fusion model. In the first stage, we partition the global HD map into semi-global tiles, where we train a collaborative HD map model to generate vectorized map outputs. In the second stage, we operate on the vectorized map outputs and develop a sparse graph fusion mechanism to fuse the semi-global map elements into the final global map.

2.2. Collaborative Semi-Global Map Model

Building HD maps at a global scale directly from BEV features is computationally infeasible due to memory and scalability constraints. To overcome this, we propose a collaborative semi-global map model that partitions the global region into spatial tiles and aggregates multi-agent observations within each tile. In contrast to existing online HD map models that rely on single-vehicle inputs with limited perception ranges, our approach fuses fine-grained features across multiple agents over larger areas, generating more complete maps with improved coverage and robustness.

BEV feature generation. Given multi-agent sensor scans (*e.g.*, cameras or LiDAR), each agent's data is processed independently through a CNN backbone, a Feature Pyramid Network (FPN) [15] fusion module, and a BEV encoder [19] to produce per-agent BEV feature maps.

Multi-agent BEV fusion. To fuse multi-agent features, we use GPS/IMU metadata to apply affine transformations that map each BEV feature to a global coordinate frame. Following [24], we apply spatial warping via bilinear interpolation and fuse the aligned features using a simple maxout operation [6] over corresponding spatial locations.

Map decoder. We build our map decoder using a variant of the Deformable DETR model [32]. Following MapTR [13], we use a set of learnable queries to interact with fused BEV features and directly predict vectorized map element instances within the semi-global region.

2.3. Sparse-to-Sparse Graph Fusion

After generating vectorized map elements within semi-global regions, we apply a sparse-to-sparse graph fusion mechanism to construct the final global HD map. This approach avoids large BEV features and dense decoding by directly operating on aggregated multi-agent map elements, enabling scalable global map construction. The graph fusion module follows an encoder-decoder architecture.

Graph encoder. The graph encoder encodes the vectorized map elements produced by the semi-global map model. Each vectorized map element is represented as a node, and spatial or semantic relationships between elements are captured as edges in the graph. In addition to geometric structure, each node stores relevant metadata (e.g., the coordinates of its constituent points and the confidence score of

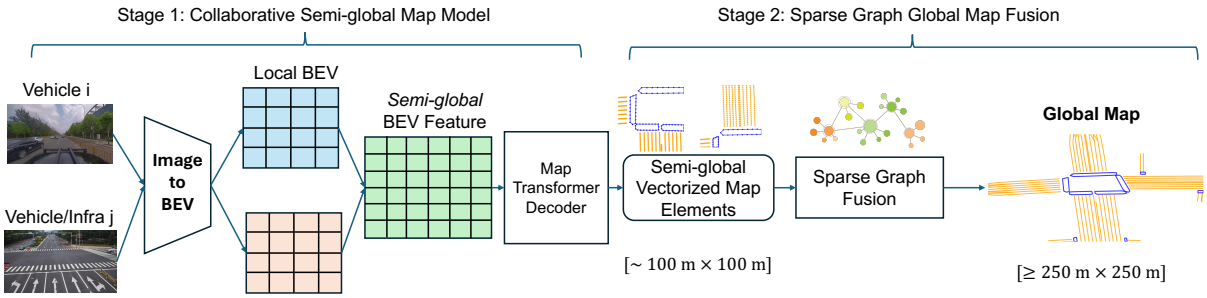

Figure 1. Overall architecture of `CrowdMap`.

its predicted label) which enables efficient identification of overlapping or conflicting elements at the node level.

Graph decoder. Unlike prior map decoding stages focused on generating accurate local HD maps, our approach introduces an inter-map decoding stage to mitigate discrepancies between partially overlapping maps from multiple agents. Leveraging graph-based representations—where each map element is modeled as a node—we aggregate nodes from neighboring agents and identify overlapping or conflicting elements based on spatial proximity. Overlaps are determined using Chamfer Distance [2, 12], with pairs considered overlapping if the distance falls below a predefined threshold. This enables selective refinement of semantically or geometrically inconsistent regions, avoiding full map regeneration and supporting efficient fusion of vectorized HD maps with structural interpretability.

To enhance the quality and consistency of the fused HD map, we introduce a lightweight, type-aware refinement module. After constructing the graph-based global map, we apply rigid point-set registration to address spatial-temporal misalignments in the 2D coordinates of points within map elements, inspired by ADoPT [7]'s demonstration that point registration can guide the measurement of temporal consistency. Since different element types respond differently to viewpoint variation, registration is performed in a type-specific manner. We adopt the Coherent Point Drift (CPD) rigid registration [18], which is well suited to our structured 2D primitives (polylines and polygons), aligning each overlapping element to the most plausible one—typically the segment with the highest confidence score. In contrast to the prior state-of-the-art method [20], which lacks a fine-grained understanding of viewpoint discrepancies between vehicles and relies on heuristics, our approach introduces an explicit graph-based inter-map decoding stage that structurally and semantically resolves misalignments between overlapping elements, ultimately enabling more consistent and high-fidelity collaborative HD map construction.

3. Preliminary Experiment Results

3.1. Experiment Setup

Datasets and evaluation. We evaluate our method on the real-world multi-vehicle dataset DAIR-V2X [26] by employing the standard processing setting of previous local HD map methods [12–14]. The DAIR-V2X dataset includes one vehicle and one infrastructure agent, each equipped with a front-view RGB camera, and provides 2D global vectorized HD map annotations across six intersections. Since only front camera data is available, we restrict the local perception range to the front (positive x-axis) of each agent. Specifically, we define the perception range as 0–30 meters for the vehicle agent and 0–50 meters for the infrastructure agent along the x-axis, with a shared lateral field of view of $[-15, 15]$ meters along the y-axis. For the collaborative semi-global model, we set the range to 120 × 60 m and randomly sample 10 images from the vehicle and infrastructure data scans for each data batch in training. We use the mAP metric [13, 14] over $[0.5, 1.0, 1.5m]$ chamfer distance thresholds for evaluation.

Training details. We train our model with 4 NVIDIA A100 GPUs with a batch size of 8. We adopt ResNet50 [9] as the image backbone and adopt AdamW optimizer for training.

Baseline solutions. We compare against baselines built on the MapTR [14] local HD map model, combined with different late-fusion strategies: PolyMerge [20], Fréchet Distance Clustering (FDC) [1], and Heatmap Buffer Merge (HBM) [8]. Since DAIR-V2X includes two distinct agents (vehicle/infrastructure), we train separate MapTR models for each, using their respective camera parameters.

3.2. Quantitative Results

Local and semi-global HD map performance. First, we evaluate the local HD map prediction on the multi-agent dataset DAIR-V2X, which has not been done by previous work. Table 1 compares the local HD map prediction accuracy of both the vehicle agent and infrastructure agent. We notice that existing online HD map solutions work well for short-range perception (*e.g.*, 30 × 30 m). However, as shown in Table 2, the performance of single-vehicle local HD map perception deteriorates significantly when the range increases to a semi-global tile of 120 × 60 m. Our prototype `CrowdMap` aggregates multi-agent information spatially and temporally, improving AP performance by more than 45% compared to MapTR local.

Global HD map performance after merging. Table 3 presents the global HD map accuracy at two intersections in the DAIR-V2X dataset. Our proof-of-concept implementa-

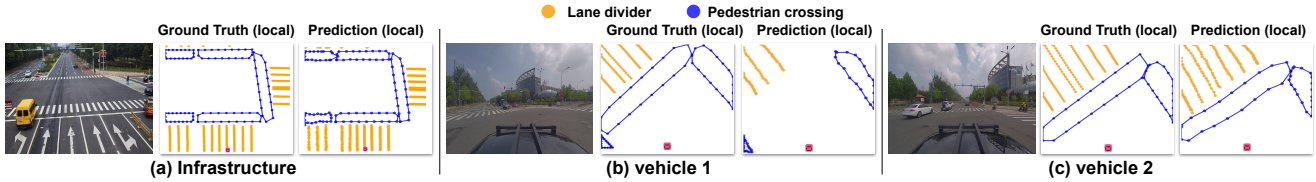

Figure 2. Visualization of local vectorized HD map construction using infrastructure and vehicle sensors.

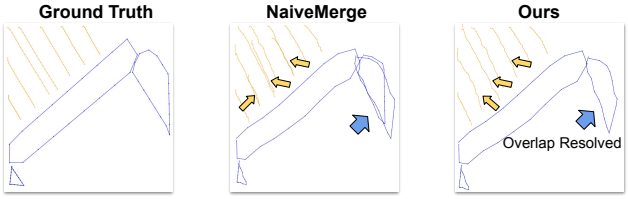

Figure 3. Sparse graph fusion of the local HD maps of vehicle 1 and 2 (see Fig. 2). Colored arrows highlight misalignments caused by viewpoint discrepancies across vehicles in overlapping objects.

Table 1. Performance of local vectorized HD map prediction.

Range	Method	Agent/Method	AP_{ped}	AP_{div}	mAP
30 × 30 m	MapTR	Vehicle-local	0.745	0.909	0.827
50 × 50 m	MapTR	Infra-local	0.772	0.785	0.778

Table 2. Performance on semi-global HD map prediction.

Range	Method	Agent/Method	AP_{ped}	AP_{div}	mAP
120 × 60 m	MapTR	Vehicle-local	0.182	0.103	0.143
120 × 60 m	`CrowdMap`	Semi-global	**0.599**	**0.682**	**0.641**

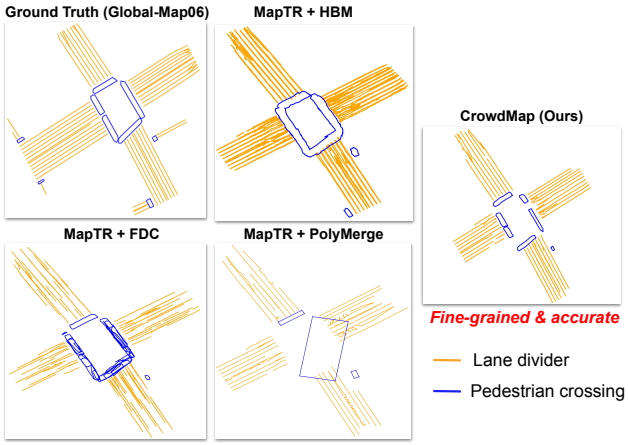

Figure 4. Global HD map construction results on DAIR-V2X.

tion of `CrowdMap` outperforms all baseline fusion methods by more than 10.0 absolute mAP, demonstrating the effectiveness of our approach. These results highlight the potential of our learning-based pipeline as a promising direction for scalable and accurate global HD map construction.

3.3. Visualization Results

Local map prediction results. Fig 2 illustrates examples of local HD map prediction from vehicle and infrastructure views. The local vectorized HD map construction quality is quite accurate using STOA approaches [14]. However, the issue arises in producing a consistent and accurate global map by directly fusing local observations.

Table 3. Performance comparison on global vectorized HD map construction.

Range	Method	Map Name	AP_{ped}	AP_{div}	mAP
300 × 300 m	MapTR + PolyMerge	Map-06	0.148	0.085	0.117
	MapTR + FDC	Map-06	0.202	0.027	0.114
	MapTR + HBM	Map-06	0.159	0.148	0.154
	`CrowdMap`	Map-06	**0.332**	**0.188**	**0.260**
250 × 250 m	MapTR + PolyMerge	Map-13	0.618	0.081	0.399
	MapTR + FDC	Map-13	0.516	0.018	0.264
	MapTR + HBM	Map-13	0.575	0.063	0.319
	`CrowdMap`	Map-13	**0.818**	**0.179**	**0.499**

Global map construction results. Fig 4 shows an example of global HD map construction results using different fusion methods. As shown, `CrowdMap` significantly outperforms baseline methods by producing a more fine-grained and accurate representation. This improvement stems from accurate point-set registration, which effectively resolves misalignments caused by discrepancies in viewpoints across different vehicles when handling overlapping or conflicting map elements, as illustrated in Fig. 3. None of the baseline late-fusion methods produces an accurate global HD map by fusing multi-agent local map observations. For instance, we find that PolyMerge [20] is fast and creates smooth results for most lane dividers, but it can easily be influenced by outliers and partial observations of pedestrian crossing, making the merged pedestrian crossing inaccurate. All three baseline methods fail to generate accurate global maps, indicating the inefficiency of existing late-fusion techniques for collaborative global map construction.

4. Discussion and Conclusion

In this paper, we experiment with building global HD maps with vision transformer-based local HD map models plus late-fusion techniques. Our analysis reveals significant accuracy degradation when applying late fusion on independently predicted local maps, highlighting the limitations of conventional approaches in collaborative multi-agent settings for global HD map construction. We propose `CrowdMap`, a novel end-to-end framework that leverages crowdsourced multi-agent data and performs sparse graph-based fusion to generate consistent and scalable global HD maps. We plan to optimize our framework and expand evaluation to more multi-vehicle datasets [30] in future work.

References

[1] Philippe C Besse, Brendan Guillouet, Jean-Michel Loubes, and François Royer. Review and perspective for distance-based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 17(11):3306–3317, 2016. 3

[2] Gunilla Borgefors. Distance transformations in digital images. *Computer vision, graphics, and image processing*, 34 (3):344–371, 1986. 3

[3] 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. 2

[4] Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Trans4map: Revisiting holistic bird's-eye-view mapping from egocentric images to allocentric semantics with vision transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4013–4022, 2023. 1

[5] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024. 2

[6] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019. 1, 2

[7] Minkyoung Cho, Yulong Cao, Zixiang Zhou, and Z Morley Mao. Adopt: Lidar spoofing attack detection based on point-level temporal consistency. *arXiv preprint arXiv:2310.14504*, 2023. 3

[8] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10 (2):112–122, 1973. 3

[9] 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. 3

[10] Cherie Ho, Jiaye Zou, Omar Alama, Sai Mitheran Jagadesh Kumar, Benjamin Chiang, Taneesh Gupta, Chen Wang, Nikhil Keetha, Katia Sycara, and Sebastian Scherer. Map it anywhere (mia): Empowering bird's eye view mapping using large-scale public data. *arXiv preprint arXiv:2407.08726*, 2024. 1

[11] Zhou Jiang, Zhenxin Zhu, Pengfei Li, Huan-ang Gao, Tianyuan Yuan, Yongliang Shi, Hang Zhao, and Hao Zhao. P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. *IEEE Robotics and Automation Letters*, 2024.

[12] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022. 3

[13] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *International Conference on Learning Representations*, 2023. 1, 2, 3

[14] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, pages 1–23, 2024. 1, 2, 3, 4

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[16] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023. 1, 2

[17] 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

[18] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010. 3

[19] 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. 2

[20] Mohamed Sayed, Stepan Perminov, and Dzmitry Tsetserukou. Polymerge: A novel technique aimed at dynamic hd map updates leveraging polylines. In *2023 21st International Conference on Advanced Robotics (ICAR)*, pages 94–99. IEEE, 2023. 2, 3, 4

[21] Anqi Shi, Yuze Cai, Xiangyu Chen, Jian Pu, Zeyu Fu, and Hong Lu. Globalmapnet: An online framework for vectorized global hd map construction. *arXiv preprint arXiv:2409.10063*, 2024. 2

[22] 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. *arXiv preprint arXiv:2301.00493*, 2023. 2

[23] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022. 1

[24] 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

[25] Jing Yang, Sen Yang, Xiao Tan, and Hanli Wang. Histrackmap: Global vectorized high-definition map construction via history map tracking. *arXiv preprint arXiv:2503.07168*, 2025. 2

[26] 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. 2, 3

[27] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024. 1, 2

[28] Qingzhao Zhang, Xumiao Zhang, Ruiyang Zhu, Fan Bai, Mohammad Naserian, and Z Morley Mao. Robust real-time multi-vehicle collaboration on asynchronous sensors. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023. 1

[29] Qingzhao Zhang, Shuowei Jin, Ruiyang Zhu, Jiachen Sun, Xumiao Zhang, Qi Alfred Chen, and Z Morley Mao. On data fabrication in collaborative vehicular perception: Attacks and countermeasures. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 6309–6326, 2024. 1

[30] Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z. Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, Jacob Pham, Xin Xia, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. *arXiv preprint arXiv:2412.01812*, 2024. 4

[31] Ruiyang Zhu, Xiao Zhu, Anlan Zhang, Xumiao Zhang, Jiachen Sun, Feng Qian, Hang Qiu, Z Morley Mao, and Myungjin Lee. Boosting collaborative vehicular perception on the edge with vehicle-to-vehicle communication. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pages 141–154, 2024. 1

[32] 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. 2