2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

BroadBEV: Collaborative LiDAR-camera Fusion for Broad-sighted Bird's Eye View Map Construction

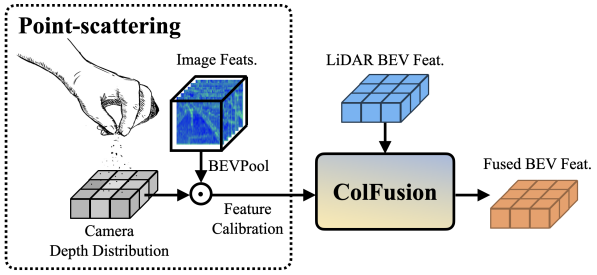
Minsu Kim[1†], Giseop Kim[2], Kyong Hwan Jin[3], Sunwook Choi[2*]

***Abstract*— A recent sensor fusion in a Bird's Eye View (BEV) space has shown its utility in various tasks such as 3D detection, map segmentation, etc. However, the approach struggles with inaccurate camera BEV estimation, and a perception of distant areas due to the sparsity of LiDAR points. In this paper, we propose a BEV fusion (*BroadBEV*) that aims to enhance camera BEV estimation for broad perception in the pre-defined BEV range, while simultaneously improving the completion of LiDAR's sparsity in the entire BEV space. Toward that end, we devise *Point-scattering* that scatters LiDAR BEV distribution to camera depth distribution. The method boosts the learning of depth estimation of the camera branch and induces accurate location of dense camera features in BEV space. For an effective BEV fusion between the spatially synchronized features, we suggest *ColFusion* that applies self-attention weights of LiDAR and camera BEV features to each other. Our extensive experiments demonstrate that the suggested methods enable a broad BEV perception with remarkable performance gains.**

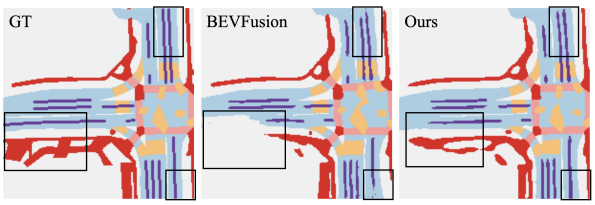
I. Introduction

Visual perception and understanding of the surrounding environment are crucial to implementing reliable robotic systems such as Simultaneous Localization and Mapping (SLAM), and Advanced Driver Assistance Systems (ADAS). Because the perception provides an ego-frame agent with detailed local features and structural information, various approaches to the perceptions have been actively studied including 3D detection and semantic segmentation. As a representation of latent variables for the tasks, Bird's Eye View (BEV) space has been frequently employed. BEV space is free from distortions of homogeneous coordinate systems and categorizes object shapes into a few classes. Thus, it provides a robust representation of elements in 3D space including cars, buildings, pedestrians, large-scale scenes, etc.

Recently, approaches using the fusion of multiple sensors' features in a shared BEV space [21], [25], [2], [17] have demonstrated effective representations even for images and points with noise and motion blur. Because each sensor has clearly distinct strong and weak points, complementary usages between sensors lead to the achievement of consistent and robust model performances. For example, in LiDAR-camera fusion, LiDAR detects accurate depth values of surrounding environments but has sparse scene contexts in far regions. In contrast, the camera obtains dense 2D signals



(a) Overview of BroadBEV.



(b) A Comparison of Broadness on Map Segmentation Results.

Fig. 1: BroadBEV scatters LiDAR points to camera depth distribution to guarantee geometric synchronization of cross-modality. After it evaluates a camera BEV feature, Our collaborative fusion (or ColFusion) ensembles the BEV features of LiDAR and cameras. As shown in the bottom comparison, our method enables a broad perception.

projected on a camera plane but loses corresponding depth information. Thus, an optimal fusion can be summarized as a method that improves a camera branch's depth learning and LiDAR branch's sparsity completion.

The existing implementations of BEV fusion usually lift and splat [31] images to transform 2D features into 3D space. The view transform is practical as it prevents BEV projection from long-tail artifacts caused by Inverse Perspective Mapping (IPM) [8]. However, because the lifting needs image depth distribution, it is inevitable to use monocular depth estimation. As a result, the requirement sometimes imposes inaccurate depths on a camera branch and negatively affects the performance. The limitation becomes a bottleneck for a model's perception of distant regions because it will fail to reasonably interpolate the large sparsity of LiDAR due to incorrect camera depths.

To address the problem, we suggest a broad BEV fusion (BroadBEV). Specifically, it synchronizes the 3D geometry of sensors in BEV space to share LiDAR's consistent sensing to a camera branch. To implement the idea, we propose "*Point-scattering*", which scatters LiDAR BEV distribution to the estimated camera depth distribution. Based on the enhanced camera BEV features, our collaborative fusion (*ColFusion* in short) applies attention weights of LiDAR

*Corresponding author.
† Work done during an internship at NAVER LABS.
[1]Minsu Kim is with Korea Institute of Science and Technology (KIST) minshu.kim@kist.re.kr
[2]Giseop Kim and Sunwook Choi are with Vision Group, NAVER LABS {giseop.kim, sunwook.choi}@naverlabs.com
[3]Kyong Hwan Jin is with School of Electrical Engineering, Korea University, Seoul, Republic of Korea kyong_jin@korea.ac.kr

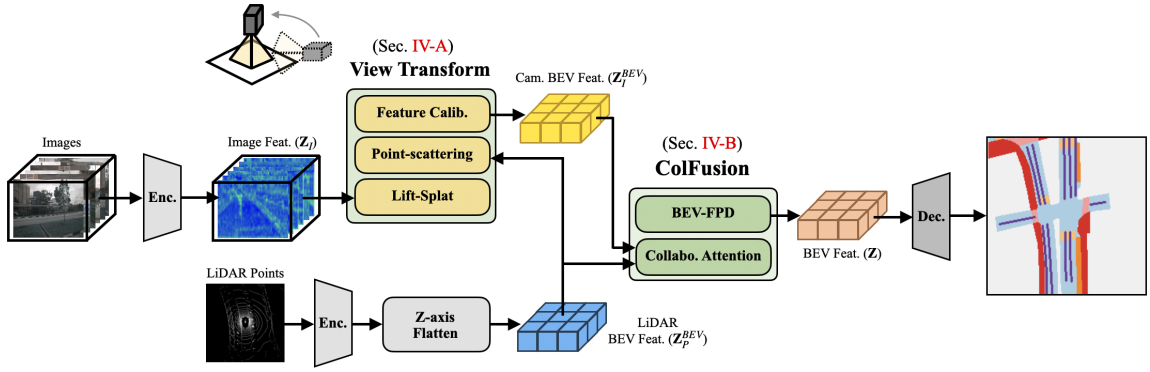


Fig. 2: **Overall Scheme of BroadBEV.** Our view transform takes in LiDAR BEV feature ($\mathbf{Z}_P^{BEV}$) to lift and splat [31] image features ($\mathbf{Z_I}$) with our proposed Point-scattering to camera depth distribution. Then it calibrates the camera BEV feature. After preparing camera and LiDAR BEV features, ColFusion which consists of collaborative attention and BEV-FPD [42] fuses them into a unified BEV feature ($\mathbf{Z}$). BroadBEV decodes the feature into a constructed map.

and camera BEV features to each other for boosting model robustness.

Our BroadBEV provides a broad-sighted BEV perception because our enhancement of the camera branch induces the improved completion of LiDAR's large sparsity in distant areas. We evaluate our approach to semantic map construction as the task requires the accurate positioning of context features with precise depths in a BEV space. The extensive experiments in Section V show that our method achieves the best performance. As far as we know, our BroadBEV is the first approach to explore a broad-sighted BEV fusion. Our contributions are summarized as:

- Our Point-scattering guarantees a synchronized geometry to the camera and LiDAR BEV branches. Our experiments in Section V validate that it contributes to the extraction of an enhanced camera BEV feature.
- For an effective fusion of BEV features with shared 3D geometry, we propose a novel collaborative fusion (or ColFusion) that shows robust perception under challenging rainy and night driving environments. In our ablation in Table IV, this method boosts model performance.
- BroadBEV provides a broad-sighted perception. With the strong point, our approach shows the best performances in extensive experiments.

II. Related Works

Camera BEV Representation. To encode perspective images to BEV features, various view transformation methods have been employed including Inverse Perspective Mapping (IPM) [34], [15], [55], [32], Lift-Splat [31] based explicit transformation [14], [13], [35], [11], [18], and query-based implicit transformation [28], [4], [44], [19], [47], [29], [37]. Because the methods have to estimate depth or learn an implicit BEV space with monocular camera features, they show weak generalization in challenging environments such as rainy days, and low-light conditions. To free those methods from the problem, they can adopt camera-LiDAR fusion. Because the sensor is relatively independent of environments, it can be a helpful guideline under challenging conditions.

LiDAR-camera Fusion. As the recent examples, the fusion of LiDAR and camera features have shown successful demonstrations in various tasks such as 3D detection [41], [1], [21], [43], [20], [46], [48], map construction [10], [12], [36], [19], [17], [27], [2], [42], [56], and both tasks [25], [40], [50]. Among them, BEVFusions [17], [21] introduce a shared BEV space for efficient fusions. Despite their promising perceptions, their BEV estimations in camera branches lack methods to share obtained sensing of each modality in the early stages. Although the models learn complementary fusion of BEV features, the spatially unsynchronized geometry between the sensor branches sometimes limits the range of model perceptions because the unlinked geometry causes worse interaction between modalities. In this work, we dig into a methodology for efficient sharing of modality sensings and a collaborative BEV fusion to effectively leverage geometry-synchronized sensors.

III. Problem Formulation

We notate I and P as camera and LiDAR modalities. The descriptions are under the assumption that image features ($\mathbf{Z}_I \in \mathbb{R}^{N\times H\times W\times C_I}$) and a LiDAR BEV feature ($\mathbf{Z}_P^{BEV} \in \mathbb{R}^{H'\times W'\times C_P}$) are prepared by Liu *et al*'s method. [25]

Conventional Method. Most existing approaches [21], [25], [2], [27] obtain a unified BEV feature $\mathbf{Z} \in \mathbb{R}^{H'\times W'\times C_B}$ as

$$\mathbf{Z} = G_\theta(\mathbf{C}, \mathbf{D}_I, \mathbf{Z}_P^{BEV}), \text{ where } [\mathbf{C}; \mathbf{D}_I] = E_\omega(\mathbf{Z}_I), \quad (1)$$

G_θ is a fusing model that consists of BEV pooling [25] and deep nets such as CNN, transformer, and feature pyramid network (FPN) [22] with parameter θ. $\mathbf{C}$ and $\mathbf{D}_I$ respectively denote an image context and a depth distribution [31]. $E_\omega := \mathbb{R}^{C_I} \mapsto (\mathbb{R}^{C_I}, \mathbb{R}^{C_D})$ is a single CNN with parameter set $\{\omega_1, \omega_2\} \subset \omega \subset \theta$. ω_1 and ω_2 are for $\mathbf{C}$ and $\mathbf{D}_I$, respectively.

BroadBEV. As shown in Eq. (1), the lack of hint for estimation of depth distribution sometimes becomes a bottleneck for a camera branch. As a result, the design shows weakness at night and rainy scenes because of harmful elements for depth estimation like limited camera eyesight and raindrops with unnatural textures. To this end, our strategy is to reduce the domain gap between the camera and a LiDAR BEV space. Specifically, our Point-scattering provides LiDAR

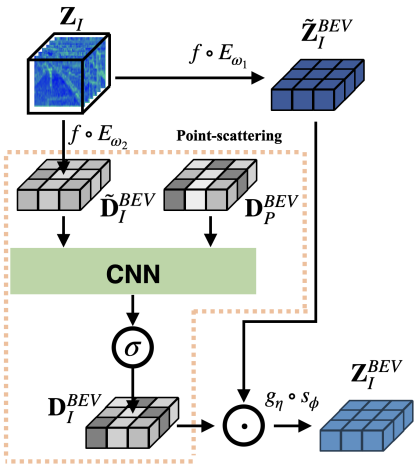
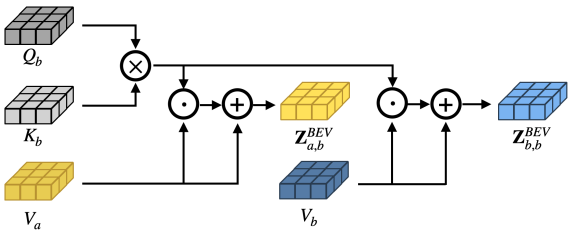
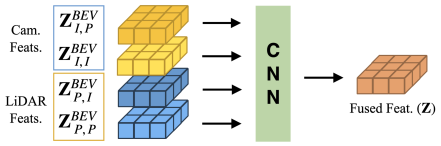


Fig. 3: **Our Point-scattering and View Transform.**



(a) Cross Modality Attention.



(b) Extraction of a Unified BEV Feature.

Fig. 4: **Our Collaborative BEV Fusion (ColFusion).**

BEV distribution $\mathbf{D}_P \in \mathbb{R}^{H' \times W'}$ to the camera depth distribution to guarantee a BEV space with synchronized geometry. The design addresses the concern that the scattering of LiDAR points to perspective camera planes causes poor compatibility with CNNs as shown by Wang *et al.* [43]. After BroadBEV prepares camera and LiDAR BEV features, our ColFusion ($\mathbf{G}_\theta$) fuses them with shared attention weights for a robust BEV representation in real environments. The overall formulation of our BroadBEV is represented as

$$\mathbf{Z} = G_\theta(\mathbf{C}, \mathbf{D}_I, \mathbf{D}_P^{BEV}, \mathbf{Z}_P^{BEV}), \quad (2)$$
$$\text{where } [\mathbf{C}; \mathbf{D}_I] = E_\omega(\mathbf{Z}_I), \ \mathbf{D}_P^{BEV} = \sigma(h_\nu(\mathbf{Z}_P^{BEV})),$$

h_ν is a single CNN layer with a parameter set ν, σ is a sigmoid operation. $\nu \subset \theta$ is a set of parameters for the estimation of LiDAR BEV distribution.

IV. Method

This section provides details of our formulation in Eq. (2). We first describe our novel view transform that guarantees depth synchronization between cross-modality. After that, we present descriptions for a collaborative BEV fusion (or ColFusion). The overall scheme of BroadBEV is illustrated in Fig. 2. As shown there, our BroadBEV takes in N cameras' image features ($\mathbf{Z}_I$), and a LiDAR BEV feature ($\mathbf{Z}_P^{BEV}$). After view transform and fusion, a semantic map is constructed.

A. View Transform

Compared to a camera, LiDAR shows consistent sensing in challenging conditions like signals damaged by raindrops, and limited eyesight by low-light illuminations. Our view transform aims to contain the advantage in camera BEV features. To implement the idea, we lift and splat $\mathbf{D}_I$ and interact it with LiDAR BEV distribution $\mathbf{D}_P^{BEV}$ in BEV space. The overall pipeline is provided in Fig. 3.

Lift-Splat [31]. At first, we define two variables: a camera BEV feature without depth distribution ($\tilde{\mathbf{Z}}_I^{BEV}$), and a depth distribution without image features ($\tilde{\mathbf{D}}_I^{BEV} \in \mathbb{R}^1$). They are prepared as

$$\tilde{\mathbf{Z}}_I^{BEV} = f(\mathbf{C}, \mathbb{1}), \quad \tilde{\mathbf{D}}_I^{BEV} = f(\mathbb{1}, \mathbf{D}_I), \quad (3)$$

where $\mathbb{1}$ is a $H \times W$ tensor of ones. f denotes a BEV pooling, i.e., $f(\mathbf{Z}_I, \mathbf{D}_I)$ is the view transformation of Liu *et al.* [25].

Point-scattering. After preparing variables, BroadBEV scatters $\mathbf{D}_P$ to a depth distribution $\tilde{\mathbf{D}}_I^{BEV}$ as

$$\mathbf{D}_I^{BEV} = \sigma(h_\psi(\tilde{\mathbf{D}}_I^{BEV}, \mathbf{D}_P^{BEV})), \quad (4)$$

where $h_\psi := (\mathbb{R}^1, \mathbb{R}^1) \mapsto \mathbb{R}^{C_I}$ is a CNN layer, $\mathbf{D}_I^{BEV}$ is a refined camera depth distribution.

Feature Calibration. As a LiDAR BEV represents sparse shapes while a camera BEV contains dense image features, there is an inevitable domain gap between modalities. To address the problem, we complete the camera BEV feature $\mathbf{Z}_I^{BEV}$ with a self-calibrated convolution (s_ϕ) [23] as

$$\mathbf{Z}_I^{BEV} = g_\eta(s_\phi(\tilde{\mathbf{Z}}_I^{BEV} \cdot \mathbf{D}_I^{BEV})), \quad (5)$$

where g_η is an FPN [22].

B. Collaborative Fusion

After the preparation of BEV features ($\mathbf{Z}_I^{BEV}, \mathbf{Z}_P^{BEV}$), our devised collaborative fusion (or ColFusion) ensembles them. Specifically, ColFusion computes the self-attention weights of its modality branches and shares them with each other to effectively leverage the synchronized sensors. Then the model fuses the BEV features with deep nets. ColFusion's pipelines are illustrated in Fig. 4.

Attention of Cross Modality. When given two inputs of modality $a, b \in \{I, P\}$, we obtain a refined BEV feature $\mathbf{Z}_{b,a}^{BEV}$ from an attention weight ($\mathbf{A}_a$) and a BEV feature $\mathbf{Z}_b^{BEV}$. $\mathbf{A}_a$ is computed from a query (Q_a), and a key (K_a) [39] of $\mathbf{Z}_a^{BEV}$, and a value (V_b) is computed from $\mathbf{Z}_b^{BEV}$. The equation can be represented as

$$\mathbf{A}_a = \text{softmax}(\frac{Q_a \cdot K_a^T}{\sqrt{d_a}}), \quad \mathbf{Z}_{b,a}^{BEV} = \mathbf{A}_a \cdot V_b + \mathbf{Z}_b^{BEV}, \quad (6)$$

where d_k denotes a scale factor [39]. In other words, we compute four BEV features ($\mathbf{Z}_{I,I}^{BEV}, \mathbf{Z}_{I,P}^{BEV}, \mathbf{Z}_{P,I}^{BEV}, \mathbf{Z}_{P,P}^{BEV}$) with shared self-attention weights and the prepared BEV features.

TABLE I: **Quantitative Comparisons on BEV Map Segmentation.**

Method	Backbone	Modality	Drivable	Ped. Cross.	Walkway	Stop Line	Carpark	Divider	mIoU
OFT [35]	ResNet-18	C	74.0	35.3	45.9	27.5	35.9	33.9	42.1
LSS [31]	ResNet-18	C	75.4	38.8	46.3	30.3	39.1	36.5	44.4
CVT [53]	EfficientNet-B4	C	74.3	36.8	39.9	25.8	35.0	29.4	40.2
M^{2}BEV [44]	ResNeXt-101	C	77.2	-	-	-	-	40.5	-
BEVFusion [25]	Swin-T	C	81.7	54.8	58.4	47.4	50.7	46.4	56.6
PointPillars [16]	VoxelNet	L	72.0	43.1	53.1	29.7	27.7	37.5	43.8
CenterPoint [49]	VoxelNet	L	75.6	48.4	57.5	36.5	31.7	41.9	48.6
PointPainting [40]	ResNet-101, PointPillars	C + L	75.9	48.5	57.1	36.9	34.5	41.9	49.1
MVP [50]	ResNet-101, VoxelNet	C + L	76.1	48.7	57.0	36.9	33.0	42.2	49.0
BEVFusion [25]	Swin-T, VoxelNet	C + L	85.5	60.5	67.6	52.0	57.0	53.7	62.7
X-Align [2]	Swin-T, VoxelNet	C + L	86.8	65.2	70.0	58.3	57.1	58.2	65.7
UniM^{2}AE [56]	Swin-T, SST	C + L	88.7	67.4	72.9	59.0	59.0	59.7	67.8
BroadBEV (Ours)	Swin-T, VoxelNet	C + L	**90.1**	**69.4**	**75.9**	**60.2**	**64.2**	**60.8**	**70.1**

TABLE II: **Quantitative Comparisons on HD Map Construction.**

Method	Backbone	Modality	Divider	Ped. Cross.	Boundary	mIoU
VPN [29]	EfficientNet-B0	C	36.5	15.8	35.6	29.3
Lift-Splat [31]	EfficientNet-B0	C	38.3	14.9	39.3	30.8
BEVSegFormer [30]	ResNet-101	C	51.1	32.6	50.0	44.6
BEVFormer [19]	ResNet-50	C	53.0	36.6	54.1	47.9
BEVerse [52]	Swin-T	C	56.1	44.9	58.7	53.2
UniFusion [33]	Swin-T	C	58.6	43.3	59.0	53.6
HDMapNet [17]	EfficientNet-B0	C	40.6	18.7	39.5	32.9
HDMapNet [17]	PointPillars	L	26.7	17.3	44.6	29.5
LiDAR2Map [42]	PointPillars	L	60.4	45.5	66.4	57.4
HDMapNet [17]	EfficientNet-B0, PointPillars	C + L	46.1	31.4	56.0	44.5
LiDAR2Map [42]	Swin-T, PointPillars	C + L	60.8	47.2	66.3	58.1
BroadBEV (Ours)	Swin-T, VoxelNet	C + L	**68.8**	**51.2**	**71.9**	**64.0**

TABLE III: **Comparison of Map Segmentation to state-of-the-arts under Rainy or Night Condition.**

Method	Rainy	Night	nuScenes
BEVFusion	55.9	43.6	62.7
X-Align	57.8	46.1	65.7
BroadBEV (Ours)	**63.7**	**50.8**	**70.1**

TABLE IV: **Ablation Study on Method Specifications.**

Baseline	w/ Scat.	w/ Fusion	Rainy	Night	nuScenes	Latency (ms)
✓			56.3	43.8	62.6	83
✓	✓		57.4	45.7	63.9	88
✓		✓	62.6	50.1	69.1	152
✓	✓	✓	**63.7**	**50.8**	**70.1**	158

BEV Representation. We unify the prepared 4 BEV features as a BEV representation $\mathbf{Z}$ as

$$\mathbf{Z} = g_{\varphi}(\sum_{a,\,b} \mathbf{W}_{a,\,b} \cdot \mathbf{Z}^{BEV}_{a,\,b}), \tag{7}$$

where $\mathbf{W}_{a,\,b}$ is a partial weight of CNN parameters. g_{φ} denotes BEV-FPD [42] which is a kind of FPN to encode the unified BEV feature. After BroadBEV obtains $\mathbf{Z}$, it decodes the feature with a task-specific head.

V. Experimental Results

Following the previous works, we evaluate BroadBEV on map segmentation and HD map construction in Table I and Table II, respectively. we use mean Intersection over Union (mIoU) as our evaluation metric in all the experiments including our ablation study and visualizations.

A. Configurations

Dataset. We use nuScenes [3], a large-scale dataset that contains 700, 150, and 150 scenes for training, validation, and testing. The collected samples include various data modalities such as perspective images from 6 cameras, 3D points from 1 LiDAR, and points with vectors from 5 Radars. We use camera images resized to 256×704 resolution, and LiDAR points voxelized to 0.1m grid resolution in our LiDAR BEV branch.

Implementation Details. We use common configurations on both BEV map segmentation and HD map construction tasks. We apply the same image and LiDAR data augmentation as the baseline work [25]. We use AdamW [26] with a weight decay of 10^{-2}. We train BroadBEV during 20 epochs on 4 NVIDIA A100 GPUs. We use 3 heads for multi-head attention of ColFusion. We employ Swin-T [24] (or VoxelNet [54]) as a backbone for the camera (or LiDAR) branch. Our implementation is built on top of MMDetection [5], [6].

B. BEV Map Segmentation

Settings. We follow the evaluation protocol of [25] under 100m$\times$100m with 0.5m egocentric BEV grid resolution. The employed metric measures mIoU for each class and then chooses the highest value over varying thresholds as the semantic map classes of urban environments may be semantically overlapped like car-parking and drivable areas.
Overall Comparisons. We report comparisons of BroadBEV to the state-of-the-arts in Table I, and Table III. The former focuses on per-class comparisons of method performances, and the latter provides the effectiveness of approaches under rainy, night conditions. In Table I, we investigate the employed backbones of all methods including ResNet [9], ResNext [45], EfficientNet [38], Swin Transformer [24], VoxelNet [54], PointPillars [16], and SST [7]. As shown in

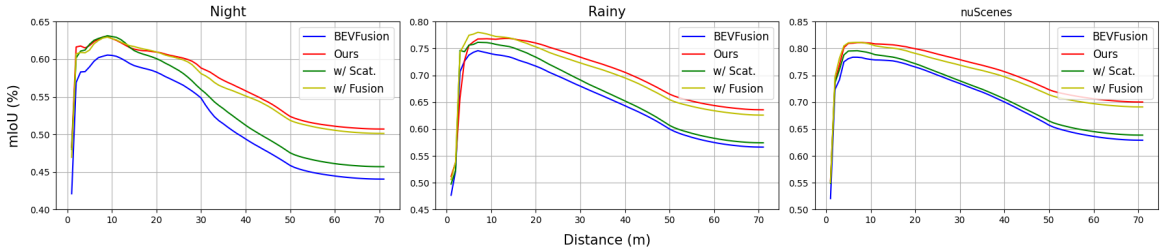


Fig. 5: **Averaged Performances for Varying Distances from Ego Vehicle.**

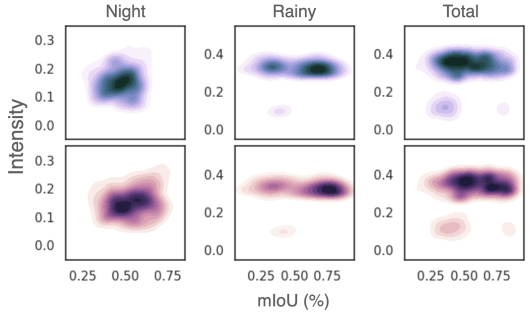


Fig. 6: **Performance Distribution under Varying Illuminations.** Blue and pink plots correspond to BEVFusion and Ours, respectively. Given average image intensities, ours achieves improved map constructions showing right-shifted distribution than BEVFusion.

the table, our method outperforms existing approaches without bells and whistles in every class. Especially, considering that map segmentation is more dependent on dense features than sparse LiDAR features, the remarkable achievements in the classes imply that BroadBEV provides improved fusion with enhanced camera BEV features. In Table III, our methods show favorable performance gains in rainy and night conditions. The results support that our methods contribute to model robustness and performance consistency.

Ablation Study. We explore the effectiveness of Point-scattering and ColFusion to check their importance. In Table IV, "w/ Fusion" and "w/ Scat." mean ColFusion and Point-scattering respectively. We denote '✓' if a method is activated. As in the table, our fusion largely boosts perception performance. In Fig. 7, we evaluate the models of Table IV to intuitively check module-by-module contributions. As shown in the camera images, the driving environment is under humid night which causes a performance drop. As demonstrated in the green ROIs, Point-scattering improves the perception broadness. Due to the large sparsity of LiDAR points in distant regions, the perception in the areas is dominantly inferred by camera branches. Despite this fact, BroadBEV shows favorable map segmentation owing to the Point-scattering's successful boosting of the camera branch weights in the training stage. Because the results show our accurate locating of scene contexts without any Lidar priors in the green boxes, they further validate our learning of camera depth refinement for completion of entire spatial coverage. In addition, ColFusion enhances the perception of distant regions as in the blue ROIs. Thanks to its collaborative modality fusion, the method further boosts BroadBEV's broadness.

Computation Time. In Table IV, we compare the computing cost of BroadBEV to our variations under single A100 GPU. Note that "baseline" has the same design as BEVFusion [25]. As in the table, ColFusion imposes dominant costs (about 70ms of 158ms) because of the heavy attention mechanism. We expect that a replacement of the attention with a light operation as in PoolFormer [51] will address the problem.

Distance Dependency. In Fig. 5, we explore a detailed comparison of our perception broadness to BEVFusion and our variated models used in our ablation study. The horizontal x-axis and vertical y-axis respectively mean $[1,x]$ and $[1,y]$ range of meter distances from an egocentric vehicle and the corresponding mIoU. In this experiment, we see that our method regularizes the negative correlation between map segmentation and distance. Especially, considering that LiDAR mostly preserves its sensing at night, the performance gain in "Night" plot validates the contribution of LiDAR Point-scattering for robust perception under low-light illumination. Furthermore, our performance gains in distant regions imply that BroadBEV provides a broad-sighted BEV representation.

Illumination Dependency. In Fig. 6, we explore the relationship between the fusion approaches and illumination. We visualize distributions of scattered mIoUs for all validation samples. The densely (or sparsely) distributed regions are colored with darker (or more light) tones. To evaluate the illumination of a sample, we transform RGB to YUV and average Y channel. The horizontal x-axis and vertical y-axis respectively mean mIoU and averaged Y values (or intensity). In the plots, our Point-scattering and ColFusion provide the improved performance distributions that are located on the higher mIoU regions. Especially, the "Night" plot shows a promising demonstration implying that our fusion resolves the limited generalization of a camera at night.

C. HD Map Construction

HD map is a pre-constructed information hub that consists of descriptions of huge 3D volumes like buildings with many indoor, and urban environments. It includes 3D bounding boxes, semantic labels, vectorized road lanes, etc. Our experiment on this task follows the previous works [17], [42] that aim for auto-labeling or maintenance of semantic road labels in HD maps by constructing locally constructed maps.

Settings. Because the HD map contains densely distributed 3D points, its label prediction is deployed on a dense 30m×60m volume with 0.15m BEV grid resolution. Following the metric of the previous works [17], [42], we use mIoU

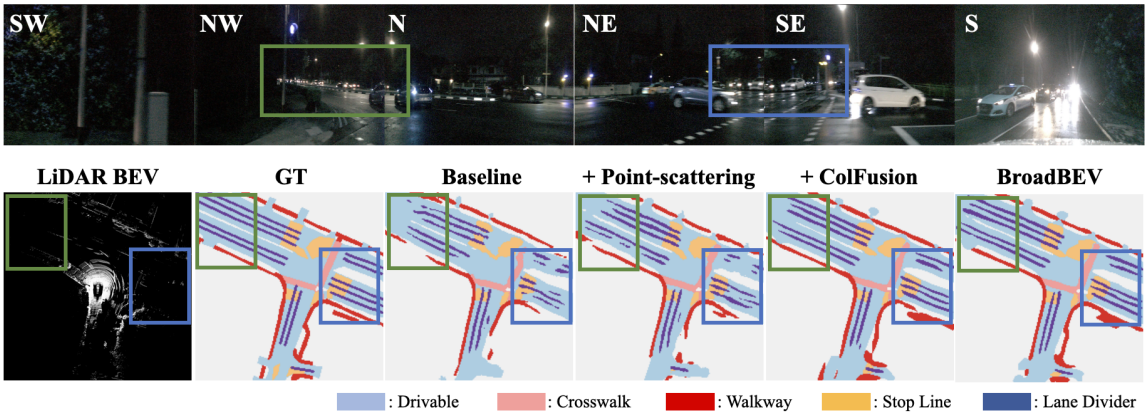
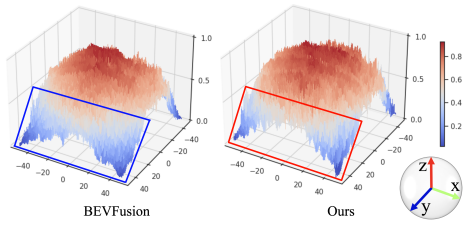
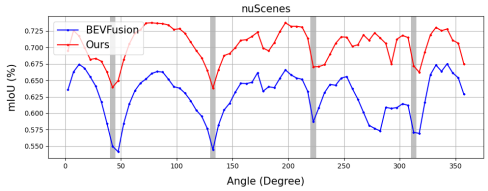


Fig. 7: **Ablation Study on Broadness.** The upper row displays RGB images. The lower row contains LiDAR BEV and map construction of our variated models. "**Baseline**", "**+ Point-scattering**", and "**+ ColFusion**" respectively correspond to "Baseline", "w/ Scat.", and "w/ Fusion" models in Table IV. Green ROIs show Point-scattering's contribution to broad-sighted perception. Blue ROIs demonstrate ColFusion's enhancement of perception performance in distant regions.



(a) Observation on mIoU Distribution at BEV Edges.



(b) Performance Undershoot caused by BEV Vertices.

Fig. 8: **Visualization of BEV Edges and Vertices.** (a) x and y axes indicate distances from an egocentric vehicle frame. z-axis is the mIoU averaged over all samples. (b) Gray shadings are the angles containing the BEV vertices.

without the thresholding of the map segmentation task.

Overall Comparisons. The comparisons of BroadBEV to the existing HD map constructors are provided in Table II. As indicated in the table, our model shows state-of-the-art performance with 5.9% (Divider: 8.0% ↑, Ped.: 4.0% ↑, Bnd.: 5.6% ↑) mIoU gain compared to the LiDAR2Map. Because HD maps require high-frequency detailed local descriptions for the surrounding environment, our performance gain implies BroadBEV's superior extraction of local features.

VI. Discussion

In this section, we analyze and discuss a limitation of BEV perception. The results are investigated using all the samples of the nuScenes validation set.

Performance Undershoot in BEV Vertices. As in Fig. 8a, BEV features provide relatively poor representations on the vertices. To explore how the limitation negatively affects map segmentation, we investigate performance distribution for varying scanning angles in Fig. 8b. From an egocentric vehicle frame, we divide a map into 72 segments with 5° field-of-view (FOV), i.e., the y value of each point denotes an averaged mIoU in an FOV segment. As depicted in the figure, the regions containing BEV vertices (gray regions) show significant performance drops. In fact, for objects at very far distances, LiDAR often fails to detect them. Furthermore, the mostly vanished shape of them in camera planes is one of the challenging limitations for learning BEV representation. Although our broadness addressed performance degradation, still the problem is a big hurdle to cross.

TABLE V: **Frustum Depth Range and Performance.**

Method	Range	mIoU
Variation	(-70.0m, 70.0m)	70.2
BroadBEV	(-60.0m, 60.0m)	70.1

Larger Camera Frustum Extending camera frustums [31] relaxes the undershoot. However, the approach fails to resolve the limitation. In Table V, we explore two BroadBEV models under 60m (BroadBEV) and 70m (Variation) frustum depth configurations. As in the table, "Variation" shows a slightly better segmentation but could not resolve the problem. Future works addressing this issue will be interesting.

VII. Conclusion

Our proposed method BroadBEV provides a broad-sighted bird's eye view representation. The synchronized geometry between cross-modality contributes to the precise location of dense contexts in the proper position of a BEV space. In addition, our ColFusion ensembles LiDAR and camera BEV features collaboratively with a novel attention mechanism-based fusion. Compared to the recent model, X-Align, BroadBEV shows noticeable improvements in overall driving conditions including rainy, night (or low-light), and whole data scenarios with 5.9%, 4.7%, and 4.4% mIoU gains, respectively. Furthermore, our model additionally validated its robustness on HD map construction, which requires high-frequency detailed features. In the task, we achieved 5.9% mIoU gain compared to LiDAR2Map.

REFERENCES

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.

[2] Shubhankar Borse, Marvin Klingner, Varun Ravi Kumar, Hong Cai, Abdulaziz Almuzairee, Senthil Yogamani, and Fatih Porikli. X-align: Cross-modal cross-view alignment for bird's-eye-view segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3287–3297, 2023.

[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.

[4] 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.

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[7] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022.

[8] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[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.

[10] Noureldin Hendy, Cooper Sloan, Feng Tian, Pengfei Duan, Nick Charchut, Yuesong Xie, Chuang Wang, and James Philbin. Fishing net: Future inference of semantic heatmaps in grids. *arXiv preprint arXiv:2006.09917*, 2020.

[11] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[12] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021.

[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] Youngseok Kim and Dongsuk Kum. Deep learning based vehicle position and orientation estimation via inverse perspective mapping image. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 317–323. IEEE, 2019.

[16] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[17] 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.

[18] 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.

[19] 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.

[20] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 641–656, 2018.

[21] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

[22] 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.

[23] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10096–10105, 2020.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[25] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023.

[28] Andrea Palazzi, Guido Borghi, Davide Abati, Simone Calderara, and Rita Cucchiara. Learning to map vehicles into bird's eye view. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I 19*, pages 233–243. Springer, 2017.

[29] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020.

[30] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5935–5943, 2023.

[31] 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.

[32] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13218–13228, 2023.

[33] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. Uniformer: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. *arXiv preprint arXiv:2207.08536*, 2022.

[34] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

[35] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018.

[36] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International conference on robotics and automation (ICRA)*, pages 9200–9206. IEEE, 2022.

[37] Juyeb Shin, Francois Rameau, Hyeonjun Jeong, and Dongsuk Kum. Instagram: Instance-level graph modeling for vectorized hd map learning. *arXiv preprint arXiv:2301.04470*, 2023.

[38] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[41] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[42] Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5186–5195, 2023.

[43] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019.

[44] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M^2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022.

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[46] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 244–253, 2018.

[47] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023.

[48] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022.

[49] 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.

[50] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[51] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[52] 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.

[53] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022.

[54] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

[55] Minghan Zhu, Songan Zhang, Yuanxin Zhong, Pingping Lu, Huei Peng, and John Lenneman. Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3814–3821. IEEE, 2021.

[56] Jian Zou, Tianyu Huang, Guanglei Yang, Zhenhua Guo, and Wangmeng Zuo. Unim^{2}ae: Multi-modal masked autoencoders with unified 3d representation for 3d perception in autonomous driving. *arXiv preprint arXiv:2308.10421*, 2023.