MetaBEV: Solving Sensor Failures for 3D Detection and Map Segmentation

Chongjian Ge[1*] Junsong Chen[3,1*] Enze Xie[2†] Zhongdao Wang[2]
Lanqing Hong[2] Huchuan Lu[3] Zhenguo Li[2] Ping Luo[1,4†]
[1]The University of Hong Kong [2]Huawei Noah's Ark Lab
[3]Dalian University of Technology [4]Shanghai AI Laboratory
rhettgee@connect.hku.hk jschen@mail.dlut.edu.cn lhchuan@dlut.edu.cn
{xie.enze,wangzhongdao,honglanqing,li.zhenguo}@huawei.com pluo@cs.hku.hk
Project page: https://chongjiange.github.io/metabev.html

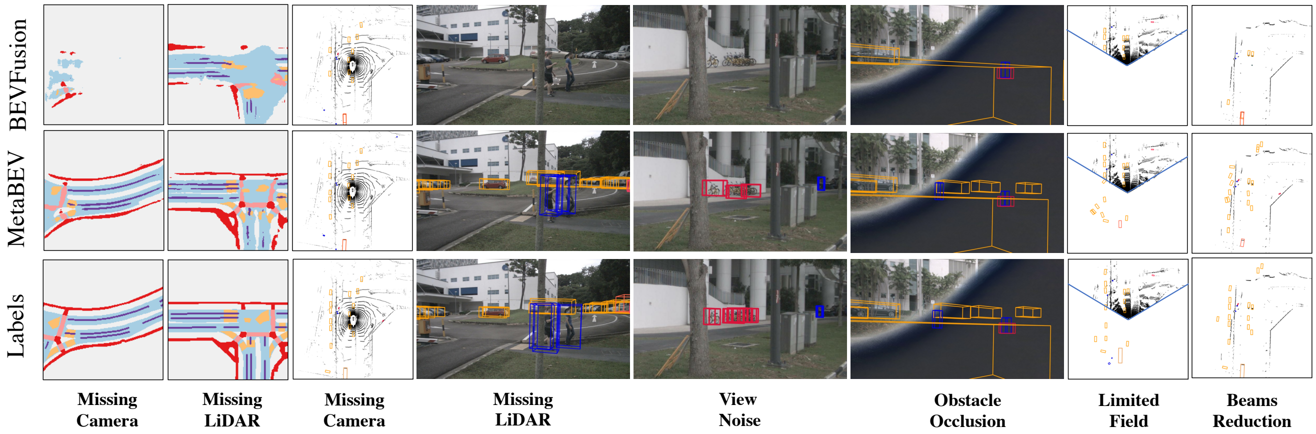


Figure 1. **Selected 3D prediction results under various sensor failures.** MetaBEV shows stronger robustness on both map segmentation and 3D detection compared to the representative multi-modal method, *e.g.*, BEVFusion [25]. Notably, in case of an entire sensor absent (e.g., Camera or LiDAR), MetaBEV provides satisfactory outcomes while existing methods fail to. **Top two rows**: 3D prediction results. **Bottom row**: the corresponding ground truths.

Abstract

Perception systems in modern autonomous driving vehicles typically take inputs from complementary multi-modal sensors, e.g.*, LiDAR and cameras. However, in real-world applications, sensor corruptions and failures lead to inferior performances, thus compromising autonomous safety. In this paper, we propose a robust framework, called MetaBEV, to address extreme real-world environments, involving overall six sensor corruptions and two extreme sensor-missing situations. In MetaBEV, signals from multiple sensors are first processed by modal-specific encoders. Subsequently, a set of dense BEV queries are initialized, termed meta-BEV. These queries are then processed iteratively by a BEV-Evolving decoder, which selectively aggregates deep features from either LiDAR, cameras, or both modalities. The updated BEV representations are further leveraged for multiple 3D prediction tasks. Additionally, we introduce a new M^2oE structure to alleviate the performance drop on distinct tasks in multi-task joint learning. Finally, MetaBEV is evaluated on the nuScenes dataset with 3D object detection and BEV map segmentation tasks. Experiments show MetaBEV outperforms prior arts by a large margin on both full and corrupted modalities. For instance, when the LiDAR signal is missing, MetaBEV improves* 35.5% *detection NDS and* 17.7% *segmentation mIoU upon the vanilla BEVFusion [25] model; and when the camera signal is absent, MetaBEV still achieves* 69.2% *NDS and* 53.7% *mIoU, which is even higher than previous works that perform on full-modalities. Moreover, MetaBEV performs moderately against previous methods in both canonical perception and multi-task learning settings, refreshing state-of-the-art nuScenes BEV map segmentation with* 70.4% *mIoU.*

[0]∗ Equal contribution.
[0]† Corresponding authors.

1. Introduction

Perceiving the surrounding environment is a fundamental capability of autonomous driving systems. In pursuit of higher perceptual accuracy, prior works make significant efforts in designing stronger task-specific modules [41, 16, 21], cultivating effective training paradigms [5], leveraging multi-modalities [25, 22, 7], etc. Among all these, the multi-sensors fusion strategy exhibits significant advantages in achieving stronger perception abilities [31, 39, 25, 1], thus being widely explored in both academia and industry. While the majority of works focus on achieving optimal performance on ideal multi-modal inputs for a single specific task, they unintentionally neglect how the designed models perform with sensor failures, which are commonly encountered and inevitable in real-world applications.

To alleviate performance drop on sensor failures, previous works encounter two challenges as follows. **1) Features misalignment:** Existing fusion methods typically utilize CNNs and features concatenation for fusion [25, 22]. The pixel-level position correlation is consistently imposed, giving rise to multi-modal features misalignment, especially when geometric-related noises are introduced. This issue could be attributed to the intrinsic characteristics of CNNs, which exhibit limitations in long-range perception and adaptive attention to input features. **2) Heavily reliant on complete modalities:** Prior arts generate the fused BEV features using either query-indexing or channel-wise fusion manners. Query-indexing methods [29, 31, 32, 14] typically rely on LiDAR and 2D Camera features for mutual querying, while channel-wise fusion approaches [25, 22] are inevitable to involve element-wise operations (*e.g.* element-sum) for feature merging. Both fusion strategies are heavily dependent on complete modality inputs and lead to inferior perception performance encountering extreme sensor failures such as LiDAR-missing or Cameras-missing, thus being limited in the practical applications.

In this research, we propose MetaBEV to tackle the above features misalignment and full-modality dependence problems through modality-arbitrary and task-agnostic learning in the unified bird's-eye view (BEV) representation space [18]. We identify the major bottleneck in modality-dependent methods is the lack of designs that enable independent fusion of different modalities by the fusion module. Therefore, we present a modality-arbitrary BEV-Evolving decoder, which leverages cross-modal attention to correlate the learnable meta-BEV queries with either a single camera-BEV feature, LiDAR-BEV feature, or both to eliminate the bottleneck. Finally, we apply a few task-specific heads to support different 3D perception predictions.

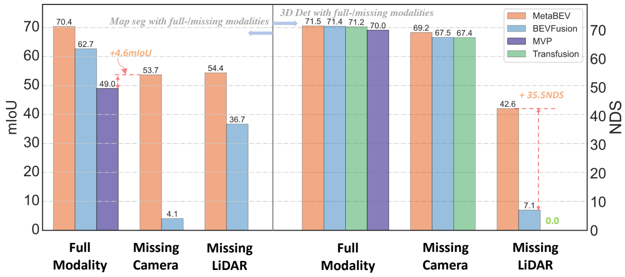


Figure 2. **MetaBEV shows stronger robustness on missing sensors.** We perform representative methods with full/missing modalities on both map segmentation (the left part) and 3D detection (the right part). Results show MetaBEV can mitigate the performance drop on input absence. Quantitatively, when facing missing cameras, MetaBEV still achieves 53.7% mIoU, which outperforms the representative method (i.e., BEVFusion [25]), by +49.6% mIoU, and is even better than MVP [39] performed on full-modalities. The superior performance could also be found on 3D detection tasks.

Except for the canonical perception (on no corrupted sensors), we also evaluate MetaBEV on overall six sensor corruptions (Limited Field (LF), Beams Reduction (BR), Missing Objects (MO), View Drop (VD), View Noise (VN), Obstacle Occlusion (OO)) and two sensor-missing scenarios (Missing LiDAR (ML) and Missing Camera (MC)). Compared with prior works, MetaBEV performs more robustly as depicted in Fig. 1. For example, it achieves 69.2% NDS and 42.6% NDS on detection when totally missing the cameras or LiDAR, respectively. For map segmentation, when total cameras absence occurs, MetaBEV could still achieve better performance compared to the work trained on multi-modalities (*i.e.*, 54.7% *v.s.* MVP's 49.0% mIoU [39]). Besides, the attention-based MetaBEV exhibits inherent robustness against multiple heavy corruptions with zero-shot and in-domain tests. For instance, as shown in Tab. 3, even missing 66.6% of LiDAR points, MetaBEV still achieves 55% NDS, outperforming the competitor by +11.7%.

Moreover, considering the limited computational resource in practice, using a single framework with shared parameters for different tasks is more efficient than using separate frameworks for multiple tasks. However, the task conflicts in the joint learning of detection and segmentation often lead to severe performance drop [36, 21, 25], and existing methods rarely analyze and design for multi-task learning (MTL). We incorporate MetaBEV with a flexible module based on Multi-Task Mixture of Experts (M^2oE) to demonstrate one possible solution for MTL and hope to stimulate further research in this area.

The appealing advantages of MetaBEV are concluded:

1. MetaBEV is a novel BEV perception framework for 3D object detection and BEV map segmentation, which can maintain resilient performance under arbitrary sensor input. Plenty of real-world sensor corruptions are formulated as well as methodically experimented with and analyzed to verify its robustness.

2. MetaBEV involves the M^{2}oE structures to alleviate tasks conflict when performing 3D detection and segmentation tasks with the same trained weights.

3. MetaBEV achieves state-of-the-art performance on nuScenes dataset [3]. It's the first method designed for both sensor failures and tasks conflict. We hope MetaBEV will facilitate future research.

2. Related Work

2.1. LiDAR-Camera Fusion

In 3D perception tasks, the effectiveness of multi-sensor fusion is notable. Therefore, researchers have focused on better combining geometric-centric point clouds and semantic-centric images. Existing methods mainly focus on three aspects: proposal-level [6], point-level [7, 1, 19], and feature-level [25, 22]. While attempting to combine multiple modalities, existing approaches typically adopt similar fusion strategies, including query-indexing fusion [1] and channel-wise fusion [25, 22]. Though showing effective in full-modality input, the above fusion manners usually associate multi-modal features tightly among points or pixels, causing features in these methods to be susceptible to spatial misalignment issues and suffering from system collapses due to sensor missing. In our work, we explore an approach that can handle arbitrary modality freely to mitigate the above limitations.

2.2. Sensors-Failure Perception

Sensor failures can significantly impact the accuracy of 3D perception, thereby jeopardizing the safety of autonomous driving. Therefore, conducting research on sensor failures is of utmost practical importance. Prior arts have made preliminary attempts to propose robust frameworks that address specific sensor failure scenarios on a case-by-case basis. These scenarios include camera-views drop [1], changes in illumination conditions [25, 7, 1], noisy inputs [20], and etc. More recently, [40] provided a comprehensive benchmark for verifying the robustness of methods mainly on sensor degradation. While previous works mainly focus on sensor corruptions, sensor absence is rarely noticed. In this work, we propose a novel perception framework that analyzes not only the aspects of sensor corruptions but also sensor absence comprehensively.

2.3. Multi-Task Learning

Multi-task learning denotes performing multiple task predictions with one set of trained weights, which appeals to the research community for its practical values, such as the complementary performance, less computational cost, etc. However, multi-task learning can also be challenging, as the model must learn to balance various objectives of each task and avoid tasks conflict. In 3D perception, the tasks conflict has been proven by M^{2}BEV [36] and BEVFormer [21], which both perform joint training on camera-only 3D object detection and BEV map segmentation. It is empirically found that joint training leads to severe accuracy degradation on each single task. Prior arts do not address the issues effectively due to their unified models and usually solve the problems from the learning perspective. For example, UniAD [13] adopts multiple training stages to adjust the overall network module by module to avoid gradient conflicts in multi-task learning, which is not end-to-end and can not be optimized layer-by-layer to adjust gradients. Inspired by Mixture-of-Expert (MoE), which is specially designed for large language model in the Natural Language Processing [35, 17, 15, 37, 33] and self-supervised field [8, 11, 12, 10], we introduce a robust fusion module with a new M^{2}oE-FFN layer. The main purpose is to mitigate the gradient conflict between detection and segmentation to achieve a more balanced performance. MetaBEV is the first framework introducing MoE into 3D object detection and BEV map segmentation as a multi-modal, multi-task, and robust approach.

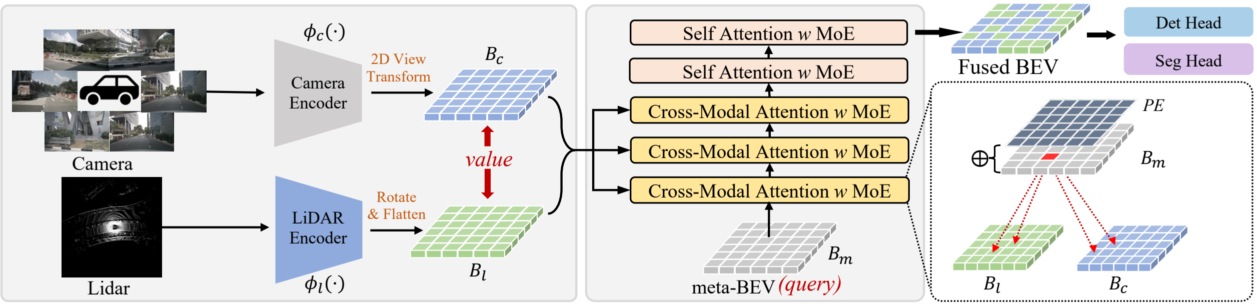
3. MetaBEV Method

We introduce a new baseline, which targets solving a range of sensor failures on 3D object detection and BEV map segmentation tasks. As opposed to existing perception methods that heavily rely on the complete sensor inputs, we connect different modalities through a parameterized *meta-BEV query* and perform cross-modal attention to integrate both semantic and geometry representations from cameras and LiDAR. As shown in Figure 3, the overall pipeline consists of a multi-modalities feature encoder, a BEV-Evolving decoder with cross-modal deformable attention, and task-specific heads. Architecture designs are detailed as follows.

3.1. BEV Feature Encoder Overview

MetaBEV generates fused features in the BEV space, as BEV representation provides a solution to combine image features with insufficient geometric knowledge and LiDAR features lacking semantic understanding in a unified domain, allowing for complementary fusion. Additionally, the regularity of the BEV feature facilitates the effective integration of various advanced task heads, which could benefit plenty of perception tasks.

Camera/LiDAR to BEV. We build our multi-modal feature encoder based on the state-of-the-art perception method BEVFusion [25], which takes multi-view-images and LiDAR-points pair as input and transforms the camera features into BEV space with depth prediction and geometric projection, respectively. Specifically, we adopt a camera backbone $\phi_c(\cdot)$ to generate N-view 2D image representations $F_c^i, i \in [1, 2, \cdots, N]$. Then, following LSS [27], we



(a) Feature Encoder **(b) BEV-Evolving Decoder** *Cross-Modal Deformable Attention*

Figure 3. **An overview of MetaBEV framework.** The multi-modal inputs are separately processed by the camera encoder $\phi_c(\cdot)$ and LiDAR encoder $\phi_l(\cdot)$ to produce the BEV representations B_c, B_l. To generate the fused BEV features, a BEV-Evolving decoder takes multi-modal BEV representations and an externally initialized meta-BEV feature (as a query feature) for correlation computation. Task-specific heads take the fused features for 3D detection.

scatter the dense feature F_c^i into discrete 3D space and compress the generated 3D voxel features $V_c \in \mathbb{R}^{D\times C\times X\times Y}$ into the camera BEV representations $B_c \in \mathbb{R}^{C\times X\times Y}$ with the pillar format [25], where D, C, (X, Y) denote the depth, dimension and spatial resolution of BEV feature, respectively. As for the LiDAR points, voxelization and sparse 3D convolutions $\phi_l(\cdot)$ are utilized for encoding LiDAR BEV representation B_l.

Subsequently, Previous works [25, 22] usually involve concatenating and CNN-based feature fusion, which may cause significant performance drop under sensor corruptions or even system collapse when sensor failures occur. As mentioned in Sec. 1, CNNs exhibit limited effectiveness in large-corruption of input modality (due to the local-aggregation characteristic) and are inappropriate to deal with the absent input modality caused by sensor missing (due to the weight-sharing modeling). We solve the above challenges by introducing *BEV-Evolving Decoder* as follows.

3.2. BEV-Evolving Decoder

The BEV-Evolving decoder consists of three key components: the cross-modal attention layers, the self-attention layers, and the plug-and-play M^{2}oE blocks. The above two components facilitate fusing arbitrary modalities, while the M^{2}oE blocks are designed for mitigating the tasks conflict.

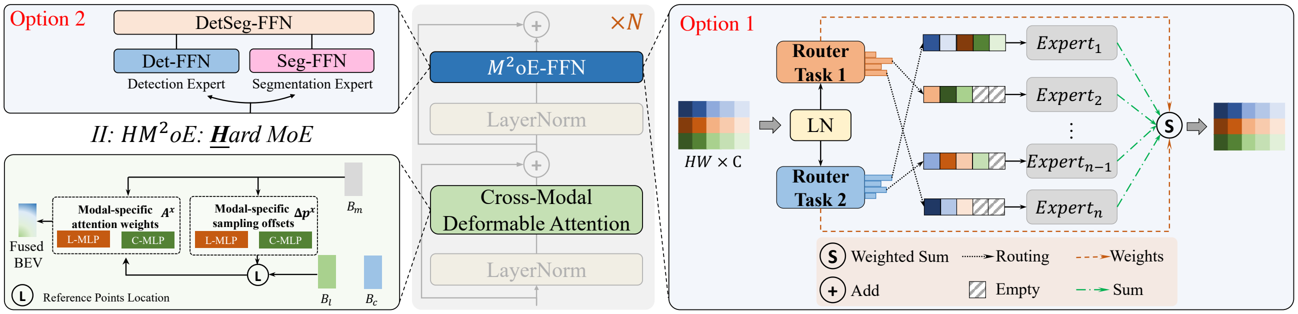
Cross-Modal Attention Layer. We first initialize a set of dense BEV queries, each representing the feature within a specific spatial grid, termed meta-BEV (B_m). B_m is added with position embeddings (PE) and then correlates with either a camera BEV feature (B_c), LiDAR BEV feature (B_l), or both. To improve the computational efficiency, we adopt deformable attention $\mathrm{DAttn}(\cdot)$ [43]. However, the original implementation $\mathrm{DAttn}(\cdot)$ is not suitable for processing arbitrary inputs, as it utilizes one unified MLP layer to sample reference points Δp and attention weights A. We then introduce model-specific MLP layers (i.e., C-MLP for cameras and L-MLP for LiDAR in Fig. 4 III) for the flexibility of fusion. Given any input BEV representations $x \in [B_c, B_l]$ as value features, we first generate model-specific sampling offsets Δp^x and attention weights A^x from query B_m. We then combine the pixel coordinates of B_m and the corresponding offsets Δp^x to locate the sampled value features. After re-scaling the sampled features by the attention weights A^x, meta-BEV is updated from informative sensor features. The overall process is formulated in the following,

$$\mathrm{DAttn}(B_m, p, x) = \sum_{m=1}^{M} W_m [\sum_{x\in[B_c\cup B_l]} \sum_{k=1}^{K} A_{mk}^x \cdot W_m^{'} x(p + \Delta p_{mk}^x)], \tag{1}$$

where m denotes the attention head, K denotes the number of sampled keys, and p denotes reference points. We use W_m and $W_m^{'}$ to represent the learnable projection matrixes. Note that in Eq. 1, the input feature x could be either B_c, B_l, or both. It enables MetaBEV to flexibly utilize multi-modal features for the deformable attention calculation during both training and testing.

The cross-attention mechanism performs the fusion process layer by layer, enabling the meta-BEV to iteratively "evolve" into fused features that capture both semantic and geometric information from the cameras and LiDAR modalities, respectively. Note that we sequentially incorporate N cross-modal attention layers to capture inter-correlations among the meta-BEV and diverse sensor features. Nonetheless, we have not yet explicitly modeled the intra-correlations, which refer to as the connection among different queries. Therefore, we introduce the self-attention layers to facilitate fused features capturing the intra-correlations. The details are discussed below.

Self-Attention Layer. To construct the self-attention layer, we downgrade the modal-specific MLP into the unified MLP layer. Besides, to sufficiently model intra-correlations



II: HM²oE: Hard MoE

III: Cross-Modal Deformable Attention

BEV-Evolving Block

I: RM²oE: Routing Multi-task Mixture-of-Experts

Figure 4. **Detailed illustration of a BEV-Evolving block.** In the cross-modal deformable attention layer, We incorporate model-specific MLP layers to enable flexibly calculating sampling offsets and attention weights for an arbitrary modality. In FFN layer, we incorporate two MoE options to alleviate tasks conflict in multi-task learning.

among queries, the input feature x in Eq. 1 is substituted by B_m, resulting in a self-attention calculation $\mathrm{DAttn}(B_m, p, B_m)$. We assemble only *M=2* self-attention layers in the BEV-Evolving Decoder, achieving a favorable trade-off between performance and computational efficiency. By modeling inter-modalities and intra-queries correlations, the fused BEV features are finally output for 3D predictions. It's experimentally found that the hybrid designs, combining both intra- and inter-correlations, provide comprehensive modeling on the fused BEV features, thus benefiting various tasks (as validated in Sec. 4.5).

M²oE Block. As shown in Fig. 4 I and II, following the previous setting of [28] that is designed for large language modeling via a mixture of experts layers (MoE), we incorporate the MLP layer in our BEV-Evolving block with MoE to introduce our M²oE blocks for multi-task learning. We first illustrate the RM²oE(as shown in Fig. 4 I) with the following formulation,

$$\mathrm{M}^2\mathrm{oE}(x) = \sum_{i=1}^{t} \mathcal{R}(x)_i \mathcal{E}_i(x), t \ll E \tag{2}$$

where $x \in \mathbb{R}^D$ is the input tokens to the RM²oE-FFN layer, $\mathcal{R} : \mathbb{R}^D \to \mathbb{R}^E$ is a routing function which assigns each token to its belonging experts and $\mathcal{E}_i : \mathbb{R}^D \to \mathbb{R}^D$ is the token processing in expert i. Both $\mathcal{E}_i$ and $\mathcal{R}$ are implemented by MLP layers, and E is a hyper-parameter as the total expert capacity. With sparse $\mathcal{R}$ where top t probability is selected, each token can only be routed to $t \ll E$ experts, while a lot of experts remain inactive. Under the extreme condition of $t = 1$, the additional calculation only comes from $\mathcal{R}$. For HM²oE, Fig. 4 II depicts a degenerate version when $E = task\ number$ and $\mathcal{R} = 1$. In this scenario, tokens bypass the router allocation process and pass through the task-specific FFN networks before being merged by the task-fusion network. This super-linear scaling of the MLP layer facilitates multi-task training and inference by enabling the router to select appropriate experts. During network optimization, such a process mitigates tasks conflict between 3D object detection and BEV map segmentation by separating the conflicting gradients (caused by diverse training objectives) with different experts.

3.3. Sensor Failures

A practical perception model is required to perform effectively even encountering corrupted or absent inputs. To this end, we define a series of sensor failures to simulate both sensor corruptions and complete sensor absence. For sensor corruptions, we include six types: 1) *Limited Field of LiDAR (LF)*, which occurs when LiDAR data can only be collected from a portion of the field of view due to incorrect collection or partial hardware damage [40]; 2) *Missing of Objects (MO)*, which appears when certain materials prevent some LiDAR points from being reflected [40]; 3) *Beams Reduction (BR)*, occurring due to limited power supply or sensor processing capabilities; 4) *View Drop (VD)* and 5) *View Noise (VN)*, resulting from camera faults; and 6) *Obstacle Occlusion (OO)*, that is a real-world phenomenon where objects are occluded from the cameras. Furthermore, we evaluate MetaBEV using extreme sensor absence scenarios, including *Missing Camera(MC)* and *Missing LiDAR(ML)*. As such, our evaluation takes into account both sensor corruptions and absence.

3.4. Switched Modality Training

The unique designs (i.e., modal-specific modules) in the BEV-Evolving block enable MetaBEV flexibly processing of either camera features, LiDAR features, or both. We propose a switched-modality training scheme to ensure precise predictions by using arbitrary modalities. This alternating strategy simulates real-world conditions during training, as MetaBEV randomly receives inputs from the aforementioned modalities with a predetermined probability. As a result, MetaBEV can conduct inferences on any input modality, thereby increasing its practicality in autonomous driv-

Methods	Modality	MTL	mAP(val)	NDS(val)	Drivable	Ped.Cross	Walkway	Stop Line	Carpark	Divider	Mean
M^{2}BEV[36]	C	✗	41.7	47.0	77.2	-	-	-	-	40.5	-
BEVFormer[21]	C	✗	41.6	51.7	80.1	-	-	-	-	25.7	-
BEVFusion[25]	C	✗	35.6	41.2	81.7	54.8	58.4	47.4	50.7	46.4	56.6
X-Align[2]	C	✗	-	-	82.4	55.6	59.3	49.6	53.8	47.4	58.0
MetaBEV-T	C	✗	49.4	49.7	**83.3**	**56.7**	**61.4**	**50.8**	**55.5**	**48.0**	**59.3**
MetaBEV-C			**55.5**	**60.4**							
PointPillars[16]	L	✗	52.3	61.3	72.0	43.1	53.1	29.7	27.7	37.5	43.8
CenterPoint[38]	L	✗	59.6	66.8	75.6	48.4	57.5	36.5	31.7	41.9	48.6
BEVFusion[25]	L	✗	64.7	69.3	75.6	48.4	57.5	36.4	31.7	41.9	48.6
MetaBEV-C	L	✗	62.5	68.6	**87.9**	**63.4**	**71.6**	**55.0**	**55.1**	**55.7**	**64.8**
MetaBEV-T			64.2	**69.3**							
PointPainting[31]	L+C	✗	65.8	69.6	75.9	48.5	57.1	36.9	34.5	41.9	49.1
MVP[39]	L+C	✗	66.1	70.0	76.1	48.7	57.0	36.9	33.0	42.2	49.0
TransFusion[1]	L+C	✗	67.3	71.2	-	-	-	-	-	-	-
BEVFusion[25]	L+C	✗	68.5	71.4	85.5	60.5	67.6	52.0	57.0	53.7	62.7
X-Align[2]	L+C	✗	-	-	86.8	65.2	70.0	58.3	57.1	58.2	65.7
MetaBEV-T	L+C	✗	68.0	**71.5**	**89.6**	**68.4**	**74.8**	**63.3**	**64.4**	**61.8**	**70.4**
BEVFusion†[25]	L+C	✓	-	69.7	-	-	-	-	-	-	54.0
BEVFusion‡[25]	L+C	✓	65.8	69.8	83.9	55.7	63.8	43.4	54.8	49.6	58.5
MetaBEV-MTL†	L+C	✓	65.6	69.5	**88.7**	**64.8**	**71.5**	**56.1**	**58.7**	**58.1**	**66.3**
MetaBEV-MTL‡	L+C	✓	65.4	**69.8**	**88.5**	**64.9**	**71.8**	**56.7**	**61.1**	**58.2**	**66.9**

Table 1. **Comparisons with SoTA methods on nuScenes val set.** We use -C and -T to denote equipping MetaBEV with the CenterPoint head [38] and Transfusion head [1]. MTL stands for testing multi-tasks with the same model. † and ‡ stand for separating or sharing the BEV feature encoder, respectively. MetaBEV outperforms the SoTA multi-modal fusion methods by +4.7% mIOU on nuScenes(val) BEV map segmentation and achieves comparable 3D object detection performance. MetaBEV also performs best in multi-task learning.

ing. Importantly, this approach requires only one set of pretrained weights for all the model's deployment.

4. Experiments

In this section, we detail the implementations and experimental settings on both sensor failures and tasks conflict. The performances on 3D detection and BEV map segmentation are presented to validate the effectiveness, flexibility, and robustness of our MetaBEV.

4.1. Implementation Details

Network Architectures. Swin-T [24] and VoxelNet [41] are adopted as feature encoders for the cameras and LiDAR, respectively. In the BEV-Evolving decoder, we employ four cross-modal attention layers and two self-attention layers to produce a fused BEV. We initialize the meta-BEV with a resolution of 180 × 180 to capture fine-grained correlations. Subsequently, we apply an FPN layer [23] to generate multi-scale features from the fused BEV. Unless specified otherwise, we use a Transformer head for 3D detection [1, 4, 34] and a CNN head [25] for map segmentation.

Datasets and Evaluation Metrics. We evaluate MetaBEV on nuScenes [3], a large-scale multi-modal dataset for 3D detection and map segmentation. The dataset is split into 700/150/150 scenes for training/validation/testing. It contains data from multiple sensors, including six cameras, one LiDAR, and five radars. For camera inputs, each frame consists of six views of the surrounding environment at one specific timestamp. We resize the input views to 256 × 704 resolution and voxelize the point cloud to 0.075m and 0.1m for detection and segmentation, respectively. Our evaluation metrics align with [3]. For 3D detection, we utilize the standard nuScenes Detection Score (NDS) and mean Average Precision (mAP). For BEV map segmentation, we follow [25] to calculate the Mean Intersection over Union (mIoU) for map segmentation on the overall six categories.

Training configurations. We follow the image and LiDAR data augmentation strategies from MMDetection3D [9] to enlarge the training samples' diversities. AdamW [26] is utilized with a weight decay of 0.05 and a cyclical learning rate schedule [30] for optimization. We take overall 26 training epochs for 3D detection, and 20 epochs for BEV map segmentation, with CBGS [42] to balance the data sampling. MetaBEV is trained on 8 A100 GPUs. For the switched-modality training, the training ratios for camera/LiDAR/both are uniformly set to be 1/3, 1/3, and 1/3, respectively. Further details are available in the appendix.

4.2. Performance on full modalities

We evaluate MetaBEV on 3D object detection and BEV map segmentation, utilizing either single or full modality inputs. Note that the BEV feature map size in MetaBEV is 180 × 180, which is consistent with the BEV query utilized

Methods	Camera + Lidar			Missing Camera			Missing LiDAR		
	mAP	NDS	mIoU	mAP	NDS	mIoU	mAP	NDS	mIoU
TransFusion [1]	67.3	71.2	–	61.6	67.4	–	–	–	–
BEVFusion [25]	68.5	71.4	62.7	**61.8**	**67.5**	**4.1**	**0.5**	**7.1**	**36.7**
MetaBEV	68.0	71.5	70.4	63.6	69.2	53.7	39.0	42.6	54.4

Table 2. **Experimental comparisons on extreme sensor missing.** MetaBEV is able to totally drop the features from the missing modalities for inference, while others cannot. We attempt to replace the missing features with zero in other works so that they can output results, which are colored as **blue**. MetaBEV still consistently outperforms prior works when facing extreme sensor absence.

Methods	Evaluation	Limited Field **[-60,60]**		Missing Objects **rate=1.0**		Beam Reduction **4 beams**		View Drop **6 drops**		View Noise **6 noise**		Obstacle Occlusion **w occlusion**	
		NDS	mIoU	NDS	mIoU	NDS	mIoU	NDS	mIoU	NDS	mIoU	NDS	mIoU
BEVFusion [25]	zero-shot	41.6	47.2	62.1	61.8	58.3	55.3	67.5	4.1	66.9	25.8	68.6	45.8
MetaBEV		47.0	61.2	62.5	69.2	57.7	64.3	68.3	43.5	68.0	45.2	70.0	61.2
BEVFusion [25]	in-domain	42.6	55.5	65.2	61.9	51.2	59.0	68.3	52.8	68.3	53.1	69.7	59.3
MetaBEV		54.3	62.1	68.5	69.9	54.6	66.4	69.3	67.9	69.2	67.5	70.3	70.2

Table 3. **Experimental comparisons on sensor corruptions.** Texts in **blue** denote the corruption degree. More detailed results are available in the appendix. MetaBEV consistently outperforms BEVFusion on various sensor corruptions in zero-shot and in-domain tests.

in BEVFormer, thus being fairly compared. For 3D object detection, results in Tab. 1 show that MetaBEV achieves superior performance with the camera modality, and gets comparable performance to the state-of-the-art with both LiDAR and multi-modalities. For example, with CenterPoint-3D [38], MetaBEV-C achieves an impressive 60.4% NDS on nuScenes val set with camera inputs, for LiDAR and multi-modal inputs, MetaBEV gains 69.3% and 71.5%, respectively, which is comparable to several SoTA methods. We also notice that though MetaBEV exhibits a slightly lower mAP with full-modality inputs, it still achieves superior NDS with improved orientation estimation (27.2% *vs.* 30.4% mAOE) as well as distance evaluation (28.0% *vs.* 28.7% mATE) in such the case. Moreover, it is experimentally found that MetaBEV is effective in capturing fine-grained features for dense predictions, *e.g.*, semantic prediction. For single-modality inputs, with only several MLP heads for prediction, MetaBEV achieves a 59.3% mIoU on cameras, and an impressive 64.8% mIoU on LiDAR (+16.2% higher than BEVFusion [25]). Moreover, results in Tab. 1 show we set a new state-of-the-art (SOTA) performance of 70.4% mIoU on BEV map segmentation, outperforming the previous best model [2] with +4.7% mIoU and the second-best method [25] by +7.7% mIoU.

4.3. Performance on corruption

We proceed to evaluate MetaBEV under various sensor failure cases, including complete sensor missing and various sensor corruptions. For the sensor missing, we train MetaBEV with the switched modality scheme, as described in Sec. 3.4. Previous fusion-based approaches, such as BEVFusion [25] and TransFusion [1], heavily rely on multi-modal features and cannot explicitly handle the absence of features caused by sensor missing. Nevertheless, we attempt to replace missing modalities with zero-initialized features for them, allowing the prior methods for predictions. Results in Tab. 2 show that MetaBEV demonstrates a stronger anti-corruption ability. Specifically, in case of missing LiDAR, MetaBEV improves +35.5% detection NDS and +17.7% segmentation mAP upon BEVFusion [25]; when encountering missing cameras, MetaBEV surpasses BEVFusion [25] by +1.7% NDS and 49.5% mIoU. Notably, even when the cameras are missing, MetaBEV still achieves 53.7% mIoU, surpassing the SoTA LiDAR-Only methods by 5.1% (i.e., CenterPoint [38]) and 9.9% (i.e., PointPillars [16]); and when the LiDAR is missing, MetaBEV even outperforms the camera-specific BEVFusion by +1.4% NDS (42.6% *v.s.* 41.2%).

On the other hand, we conducted both zero-shot and in-domain tests on sensor corruptions. To evaluate the model's abilities in performing with unseen corrupted or missing data, we performed the zero-shot test by directly evaluating a trained model on such data. For the in-domain test, we train MetaBEV on corrupted data with random degrees, and then conduct evaluations on noise data with one specific degree. Tab.3 shows MetaBEV outperforms BEVFusion [25] on *11/12 corruptions* on zero-shot evaluation. Moreover, when trained with randomly corrupted data, MetaBEV consistently outperformed BEVFusion. The results indicate that the fusion module, which is composed of CNNs (*e.g.*, BEVFusion), has limited representational capabilities for corrupted data, while our approach leverages the attention mechanism for efficient modeling.

		mAP	NDS	mIoU
Detection only		67.6	71.0	-
Segmentation only		-	-	70.4
H-MoE	R-MoE			
✗	✗	64.8	69.4	64.7
✓	✗	**65.6** (+0.8)	69.5 (+0.1)	66.3 (+1.6)
✗	✓	65.4 (+0.6)	**69.8** (+0.4)	**66.9** (+2.2)

Table 4. **We experiment on two kinds of MoE structures for Multi-Task Learning**, and both show significant advantages.

Method	#Params.(M)	mAP	NDS	#Params.(M)	mIoU
BEVFusion	40.8	68.5	71.4	43.1	62.7
BEVFusion (Heavy)	54.1	67.2	70.1	56.1	65.9
MetaBEV	53.5	68.0	71.5	55.9	70.4

Table 5. Comparisons on 3D detection and BEV-Map segmentation.

4.4. Performance in Multi-Task Learning

Tab. 4 presents the results of our multi-task learning (MTL) experiments for 3D object detection and BEV map segmentation. We begin with the framework designed for independent single-task learning with distinct task-specific heads and loaded a set of weights trained on multi-modal detection as pretrain. The model achieves 69.4% NDS and 64.7% mIoU for detection and segmentation, respectively, which is already a state-of-the-art performance. In light of this, we implement an MoE structure to help alleviate the gradient-conflict problem in MTL. Intuitively, we introduce a routing-based MoE structure, referred to as RM^{2}oE, to replace the vanilla MLP layer. This implementation results in a remarkable improvement of 0.4% NDS in detection and 2.2% mIoU in segmentation. Considering the additional computation required by the routing mechanisms, we propose to further experiment on a simplified version HM^{2}oE where the number of experts is adjusted to match that of the tasks, and the routing function is removed. We find it still achieves a notable improvement of 0.1% NDS and 1.6% mIoU. Upon the evaluation results, both implementations mitigate performance degradation in MTL to some extent.

4.5. Ablation Studies

Setup. We present the ablation studies for analyzing the network architectures and training schemes. All the ablation studies are performed on nuScenes validation set, with the training configurations set as default in Sec. 4.1.

Network Configurations. We first explore the optimal architectures of the BEV-Evolving Decoder. Specifically, we analyze three components, including the combination of the layers, reference points number (p in Eq.1), and expert number (E in Eq. 2). Our findings in Tab. 6a indicate that a few layers of cross-attention are sufficient to produce effective predictions while stacking more cross-attention layers does not necessarily enhance the model's capacities (68.3% and 68.2% NDS for 6 and 8 cross-attention layers, respectively). The same findings also suit the reference points number, where increasing the sampling points does not necessarily lead to better performance, as shown in Tab. 6b. On the contrary, we discovered that simply adding two layers of self-attention after the cross-attention layers significantly improve the prediction performance by +1.1% NDS in Tab. 6a. As mentioned in Sec. 3.2, the hybrid designs capture both intra- and inter-correlations simultaneously, resulting in complementary modeling. Moreover, in Tab. 6c, we discover that a larger model capacity (*i.e.*, incorporating more experts) can effectively alleviate the performance drop in 3D detection and map segmentation. Specifically, MetaBEV equipped with 2/8 expert (*i.e.* activating the 2 experts with the highest scores from a total of 8 experts.) achieves the best performance compared to other configurations (66.9% mIoU and 69.8% NDS), while HM^{2}oE is unaffected due to the fixed number of experts being equivalent to the number of tasks.

Do More Parameters Guarantee Improved Model Accuracy? We note that incorporating deformable attention does introduce additional parameters to MetaBEV; however, this increment is minimal and acceptable (i.e., 54M *v.s.* 41M) in relation to its functional capabilities (addressing sensor failures). Besides, compared with BEVFusion, though MetaBEV achieves modest advancements on 3D detection (+0.1% NDS), the improvement on map segmentation is remarkably significant (+7.7% mIoU), which indicates the deformable attention facilitates more in per-pixel perception. To conduct a fair comparison, we consistently introduce more fusion layers to BEVFusion. Results in Tab. 5 demonstrate that increasing parameters of BEVFusion yields an improved performance in map segmentation, while a reduction in 3D detection accuracy. The decreased performance for 3D detection may be attributed to the challenging convergence for the larger parameters. Furthermore, it is observed that BEVFusion performs poorly in both 3D detection and map segmentation when compared to MetaBEV with the same parameter amount. This finding highlights the fact that having more model parameters does not guarantee improved model accuracy.

Switched Modality Training. For the training strategy, using switched modality scheme to simulate the sensors missing, though simple, is the key to enabling MetaBEV with robust performance on completely sensor-failure scenarios. In Tab. 6d, comparing with the vanilla full-modality training approach, our proposed training strategy yields significant improvements on 3D detection by +32.8% NDS and +1.0% NDS, and improvements on map segmentation by +17.9% mIoU and +16.7% mIoU under LiDAR-missing

Layers	mAP/NDS
4 cross	61.7/67.0
6 cross	63.0/68.3
8 cross	62.5/68.2
2 self+4 cross	**64.2/69.3**

(a) The number of self/Cross-layer

Points	mAP/NDS
4	62.3/67.9
8	**62.7/68.2**
12	62.3/68.0
16	62.3/68.1

(b) The number of reference points

Experts	mAP/NDS	mIoU
1/2 Expert	65.8/69.8	63.3
1/4 Expert	64.9/69.5	65.9
1/8 Expert	65.4/69.7	66.4
2/8 Expert	**65.4/69.8**	**66.9**

(c) Expert number

Task	Metrics	Vanilla training			Switched Modality Training		
		Lidar-Missing	Camera-Missing	Full	Lidar-Missing	Camera-Missing	Full
3D detection	NDS	9.8	68.2	71.0	42.6 **(+32.8)**	69.2 **(+1.0)**	71.5 **(+0.5)**
	mAP	2.9	62.4	67.4	39.0 **(+36.1)**	63.6 **(+1.2)**	68.0 **(+0.6)**
BEV-Map Segmentation	mIoU	36.5	27.0	70.4	54.4 **(+17.9)**	53.7 **(+16.7)**	68.5 (-1.9)

(d) Evaluation of the NuScenes val dataset when Cameras or LiDAR totally fail. We test the performance with the switched modality training on both 3D detection and BEV-map segmentation tasks.

Table 6. **Ablation studies for the model architectures and training strategy.** Default settings are marked in gray in (a), (b), and (c).

and camera-missing scenarios, respectively. More intriguingly, we find that the switched modality training could even enhance the model performance on full modalities, rising from 71.0% NDS to 71.5% NDS on 3D detection.

5. Conclusion

In this paper, we present a novel framework, named MetaBEV, for the purpose of solving sensor failures in the bird's-eye view (BEV) 3D detection and map segmentation. Our method integrates modal-specific layers into the cross-modal attention layer to enhance the fusion process, achieving appealing performance on full-modality inputs, and meanwhile MetaBEV effectively alleviates the significant performance degradation that is often caused by corrupted or missing sensor signals. We also introduce M^{2}oE to handle potential conflicts between tasks. We hope that MetaBEV will provide a more focused and effective approach for investigating sensor-failure scenarios in the autonomous driving field.

Limitations. Though we adopted the deformable attention [43] for efficiency, it unavoidably leads to a slight increase in network parameters compared to the lightweight solutions [25, 22]. Despite this, the benefits of mitigating sensor failures are highly desirable, and the additional computational overhead may be deemed acceptable. We acknowledge that lightweight networks could be an avenue for future research, potentially involving techniques such as network pruning, token reduction, etc.

Acknowledgement. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *IEEE/CVF Winter Conference on Applications of Computer Vision*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, 2020.

[5] Runjian Chen, Yao Mu, Runsen Xu, Wenqi Shao, Chenhan Jiang, Hang Xu, Zhenguo Li, and Ping Luo. Coˆ 3: Cooperative unsupervised 3d representation learning for autonomous driving. *arXiv preprint arXiv:2206.04028*, 2022.

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[7] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

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

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[11] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attention via transformers in self-supervised visual representation learning. *Advances in Neural Information Processing Systems*, 2021.

[12] Chongjian Ge, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, and Ping Luo. Soft neighbors are positive supporters in contrastive visual representation learning. *arXiv preprint arXiv:2303.17142*, 2023.

[13] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[14] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision*, 2020.

[15] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 2021.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *International Conference on Learning Representations*, 2020.

[18] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*, 2022.

[19] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[20] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[21] 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*, 2022.

[22] 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. *arXiv preprint arXiv:2205.13790*, 2022.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[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 *IEEE/CVF International Conference on Computer Vision*, 2021.

[25] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multitask multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

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

[27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, 2020.

[28] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[29] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvxnet: Multimodal voxelnet for 3d object detection. In *International Conference on Robotics and Automation*, 2019.

[30] Leslie N Smith. Cyclical learning rates for training neural networks. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2017.

[31] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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

[33] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562, 2020.

[34] 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*, 2022.

[35] Barret Zoph William Fedus and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Computer Research Repository*, 2021.

[36] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. Mˆ 2bev: Multi-camera joint 3d detection and segmentation

with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022.

[37] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 2019.

[38] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[39] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 2021.

[40] Kaicheng Yu, Tang Tao, Hongwei Xie, Zhiwei Lin, Zhongwei Wu, Zhongyu Xia, Tingting Liang, Haiyang Sun, Jiong Deng, Dayang Hao, et al. Benchmarking the robustness of lidar-camera fusion for 3d object detection. *arXiv preprint arXiv:2205.14951*, 2022.

[41] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[42] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019.

[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.