MAESTRO: Task-Relevant Optimization via Adaptive Feature Enhancement and Suppression for Multi-task 3D Perception

Changwon Kang^{1*} Jisong Kim^{1*} Hongjae Shin^{2*} Junseo Park² Jun Won Choi^{2†}

¹Hanyang University and ²Seoul National University

{changwonkang, jskim}@spa.hanyang.ac.kr, {hjshin, jspark}@spa.snu.ac.kr, junwchoi@snu.ac.kr

Abstract

The goal of multi-task learning is to learn to conduct multiple tasks simultaneously based on a shared data representation. While this approach can improve learning efficiency, it may also cause performance degradation due to task conflicts that arise when optimizing the model for different objectives. To address this challenge, we introduce MAE-STRO, a structured framework designed to generate taskspecific features and mitigate feature interference in multitask 3D perception, including 3D object detection, bird'seve view (BEV) map segmentation, and 3D occupancy prediction. MAESTRO comprises three components: the Class-wise Prototype Generator (CPG), the Task-Specific Feature Generator (TSFG), and the Scene Prototype Aggregator (SPA). CPG groups class categories into foreground and background groups and generates group-wise prototypes. The foreground and background prototypes are assigned to the 3D object detection task and the map segmentation task, respectively, while both are assigned to the 3D occupancy prediction task. TSFG leverages these prototype groups to retain task-relevant features while suppressing irrelevant features, thereby enhancing the performance for each task. SPA enhances the prototype groups assigned for 3D occupancy prediction by utilizing the information produced by the 3D object detection head and the map segmentation head. Extensive experiments on the nuScenes and Occ3D benchmarks demonstrate that MAESTRO consistently outperforms existing methods across 3D object detection, BEV map segmentation, and 3D occupancy prediction tasks.

1. Introduction

Accurate and computationally-efficient perception is essential for safe autonomous driving, as autonomous vehicles

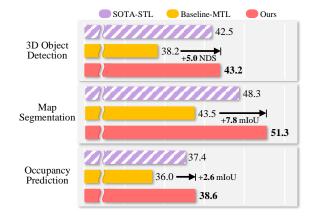


Figure 1. **Performance comparison across three perception tasks.** MAESTRO consistently outperforms both the Baseline multi-task-learning (Baseline-MTL) model and the state-of-the-art (SOTA) single-task-learning (SOTA-STL) models across all tasks.

must continuously understand their surroundings to enable reliable planning. To capture different aspects of the driving environment, various perception tasks have been explored. For example, 3D object detection models have been developed to identify and localize dynamic objects [13, 15, 20-22, 38, 41, 46, 52]. Bird's-Eye View (BEV) map segmentation models provide road-level information—such as lanes, road boundaries, and crosswalks—in the BEV domain [27, 30-32, 43, 56], offering crucial guidance on safely drivable areas. 3D occupancy prediction models discretize the environment into 3D voxels, assigning each voxel an occupancy state and semantic label to construct a dense and structured 3D representation of the scene [3, 12, 23, 36, 40, 53]. As a unified perception system, these models must be executed simultaneously and in real time to support safe and effective planning in autonomous driving.

While these perception models can be designed and optimized independently, training and deploying separate models for each task, as shown in Figure 2 (a), results in inefficiencies in both computation and data utilization. Since

^{*}Equal contribution.

[†]Corresponding author.

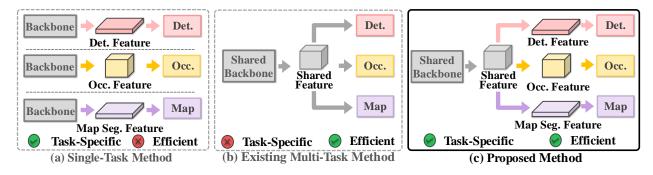


Figure 2. Comparison of Single-Task and Multi-Task 3D Perception Frameworks. (a) Single-task method: Each task utilizes a dedicated backbone and task-specific features, ensuring task specialization at the expense of high computational cost. (b) Existing multitask method: A shared backbone improves computational efficiency but lacks task-specific feature representation. (c) Proposed method: Task-specific features are generated from shared backbone features, achieving both efficiency and task-aware feature learning.

redundancy exists across tasks, it is reasonable to share a part of the architecture to exploit their commonalities. Multi-task learning (MTL) has been investigated as a means to design architectures that share common representations while also learning task-specific information across multiple tasks. In addition to reducing model complexity, MTL enables more efficient use of datasets by jointly training all tasks. This is particularly important for 3D perception tasks, which demand real-time performance and efficient resource utilization.

A straightforward approach to MTL is to share a backbone network that extracts features in the 3D domain for all tasks, as illustrated in Figure 2 (b). However, this naive approach often suffers from performance degradation due to task conflicts that arise when optimizing the shared backbone with different task-specific loss functions. The gradient signals from each loss may not be well-aligned, which limits the performance of individual tasks.

In 3D perception, different tasks focus on different semantic cues and spatial regions. For instance, 3D object detection models prioritize movable foreground objects, while BEV map segmentation models focus on static background structures. 3D occupancy estimation models, in contrast, may require attention to both foreground and background elements. These divergent behaviors can lead to conflicting supervisory signals, limiting the performance of MTL.

While various MTL methods have been proposed to address these problems in [6, 19, 55, 57, 58], they still exhibit inferior performance compared to specialized single-task approaches [8, 18, 26, 27, 39, 57]. This persistent gap underscores the need for an effective MTL approach specifically tailored for 3D perception.

In this paper, we introduce MAESTRO (Multi-task Adaptive Feature Enhancement and Suppression for Task-Relevant Optimization), a framework designed to jointly optimize the performance of 3D perception tasks. As illustrated in Figure 2 (c), our method transforms shared back-

bone features into task-specific representations, explicitly designed to minimize interference between tasks. This improves the performance of our model for all tasks.

MAESTRO consists of three key modules: the Classwise Prototype Generator (CPG), the Task-Specific Feature Generator (TSFG), and the Scene Prototype Aggregator (SPA). CPG partitions the semantic categories into foreground and background groups and generates corresponding group-wise feature prototypes [35]. The foreground prototypes are allocated to the 3D object detection task, while the background prototypes are used for BEV map segmentation. Both sets of prototypes are utilized in the 3D occupancy prediction task. The resulting prototypes serve as semantic priors to guide the generation of task-specific features. TSFG then generates task-specific features by adaptively transforming the shared backbone features to suit each task. It effectively emphasizes taskrelevant information through a feature enhancement module, while suppressing irrelevant signals via a feature suppression module. By filtering out interfering components, this mechanism helps mitigate task conflicts and improve overall multi-task learning performance. SPA is specifically designed to enhance the performance of the 3D occupancy prediction task. It is motivated by the observation that information produced by 3D object detection and BEV map segmentation can be leveraged to benefit occupancy prediction. SPA integrates the prototypes obtained from the object detection and map segmentation heads into the prototypes assigned to the occupancy prediction module. This integration enriches the query features used for 3D occupancy prediction without compromising the performance of the other tasks.

We evaluate MAESTRO on the nuScenes benchmark [2]. As shown in Figure 1, our experiments demonstrate that MAESTRO achieves substantial performance improvements over the baseline MTL method. Remarkably, it even outperforms independently trained single-task models. This

performance gain is attributed to MAESTRO's ability to effectively leverage shared structures across tasks while minimizing inter-task interference. Moreover, MAESTRO significantly outperforms existing MTL methods for all three 3D perception tasks.

The main contributions of this work are summarized as follows:

- We propose MAESTRO, a novel multi-task learning (MTL) framework that effectively generates task-specific features while mitigating feature interference across multiple 3D perception tasks.
- We introduce a group-wise prototype generation method that clusters semantic classes and derives representative prototypes for each group. These prototypes are used to enhance task-relevant features and suppress interfering components from the shared backbone, thereby reducing task conflicts and improving overall performance.
- We exploit task dependencies for further performance improvement by leveraging the outputs from 3D object detection and BEV map segmentation models to enhance the performance of the 3D occupancy prediction.
- The source code will be made publicly available to facilitate future research and reproducibility.

2. Related Works

2.1. Camera-based Perception Tasks

3D object detection. Recent advances in multi-camera 3D object detection have centered around the transformation of image features into the BEV space, primarily following either query-based or depth-based projection approaches. Query-based methods generate learnable BEV queries and extract features from multi-view images via cross-attention mechanisms [15, 22, 38, 41, 46]. In contrast, depth-based methods—initially introduced by LSS [32]—explicitly project image features into BEV space by estimating depth distributions [13, 20, 21, 52]. More recent works such as SA-BEV [52] and DualBEV [20] improve upon this strategy by incorporating object-centric pooling, which enhances the relevance of extracted features while effectively suppressing background noise.

BEV map segmentation. BEV map segmentation methods adopt BEV transformation strategies. BEVSegFormer [31] employs deformable cross-attention to project image features into BEV space without requiring depth estimation or camera parameters. In contrast, explicit warping methods [27, 30, 32, 43] rely on camera intrinsics and estimated depth to perform geometric projections. M2BEV [43] further refines this approach by aligning features to the ego-vehicle frame, while BEV-Seg [30] improves projection quality through a dedicated parsing network and enhanced depth estimation.

3D occupancy prediction. MonoScene [3] pioneered 3D

occupancy prediction using a single image and a 3D U-Net architecture. BEV-based methods [12, 23, 51] simplify computation by collapsing the height dimension, while voxel-based approaches [17, 36, 40] adopt a coarse-to-fine strategy to produce dense and structured 3D predictions. TPVFormer [14] introduces a tri-perspective view representation, but its performance is limited by the loss of detailed 3D information. In contrast, OccFormer [53] advances voxel-based semantic occupancy prediction through a dual-path transformer architecture.

2.2. Multi-Task Learning

MTL methods are broadly categorized into optimizationbased and architecture-based approaches. Optimizationbased methods aim to balance task contributions through techniques such as loss weighting and gradient manipulation [4, 5, 9, 16, 25, 34]. Architecture-based methods can be further divided into encoder-focused and decoder-focused designs. Encoder-focused approaches [7, 29, 33] emphasize learning shared representations across tasks, while decoder-focused approaches [28, 44, 54] utilize a shared encoder followed by task-specific decoders to promote taskaware predictions. Recent decoder-focused models, such as InvPT [47] and InvPT++ [49], employ transformerbased decoders for enhanced task-specific feature extraction. Similarly, TaskExpert [48] leverages a Mixture-of-Experts (MoE) framework with spatial context-aware gating to dynamically route features to task-relevant experts.

In the context of 3D perception, MTL methods predominantly adopt decoder-focused architectures [8, 27, 57], utilizing a shared encoder with multiple decoders to address various scene understanding tasks. However, empirical findings from BEVFormer [22] and M2BEV [43] reveal that naive joint training often leads to degraded task performance due to feature interference. To address this, SOGDet [57] introduces a weighted feature-sharing mechanism for effective modality fusion, while HENet [42] dynamically adjusts feature resolutions to better align with task requirements. Although these methods alleviate task conflict and improve overall performance, they still struggle to fully capture task-specific feature representations.

3. Proposed Methods

3.1. Overview

The overall structure of MAESTRO is illustrated in Figure 3. 2D feature maps are independently extracted from multi-view camera images and subsequently lifted into the 3D domain using the Lift-Splat-Shoot (LSS) method [32], resulting in a voxel representation $F_s \in \mathbb{R}^{C \times X \times Y \times Z}$ for downstream tasks. These voxel features serve as shared backbone features across all tasks. The CPG first groups class categories into foreground and background sets. It

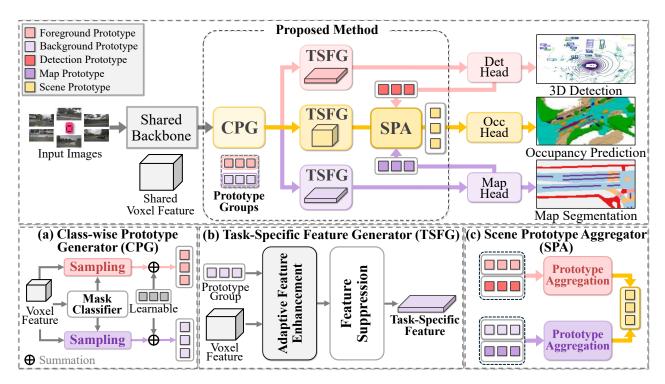


Figure 3. **Overall architecture of MAESTRO.** Multi-view images are processed by a shared backbone to generate a structured 3D voxel representation. The CPG generates foreground and background prototype groups, which guide the TSFG in refining task-relevant features for 3D object detection, BEV map segmentation, and 3D occupancy prediction. Task-oriented prototypes derived from the 3D object detection and BEV map segmentation heads are then integrated with the prototype groups via the SPA, forming Scene Prototypes. These Scene prototypes are subsequently processed by the occupancy decoder to produce the final 3D occupancy predictions.

then applies a mask classifier to produce a binary voxel mask for each class. Based on these masks, CPG generates group-wise prototype features by sampling and pooling masked voxel features from F_s . These prototype groups are assigned to their respective tasks, ensuring that each group is semantically aligned with its corresponding task. Next, the TSFG refines the shared backbone features using the task-assigned prototype groups via Adaptive Feature Enhancement and Feature Suppression modules. This process yields task-specific features that are optimized for each perception task. These features are subsequently passed through the 3D object detection, BEV map segmentation, and 3D occupancy prediction heads. The SPA further refines the prototype group associated with 3D occupancy prediction by incorporating outputs from the 3D object detection and BEV map segmentation heads. Finally, these adapted prototypes are then used to initialize the queries for the 3D occupancy decoder.

3.2. Class-wise Prototype Generator (CPG)

Naive feature sharing in MTL often results in representations that lack clear semantic separation across tasks. To address this limitation, we introduce the CPG whose structure is depicted in Figure 3 (a). CPG first organizes semantic categories into foreground and background groups. For example, the foreground group includes object classes such as car, truck, pedestrian, and bicycle, while the background group comprises semantic categories like drivable surface, sidewalk, manmade structures, and vegetation. The CPG then generates prototype features for each group. Let $\mathcal{K} = \{1,\ldots,K\}$ denote the set of all K semantic categories. CPG groups them into foreground and background groups \mathcal{K}_{fg} and \mathcal{K}_{bg} . CPG applies a lightweight mask classifier to F_s to compute voxel-wise semantic confidence scores $S_v \in \mathbb{R}^{K \times X \times Y \times Z}$. The most confident class is assigned to each voxel, forming class-wise masks $B_k \in \mathbb{R}^{X \times Y \times Z}$ as

$$B_k(i,j,l) = \mathbb{I}\left(\underset{k' \in \{1,\dots,K\}}{\operatorname{argmax}} S_v(i,j,l) = k\right), \quad (1)$$

where (i,j,l) represents voxel coordinates, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the voxel is assigned to class k and 0 otherwise. The prototype $P_k \in \mathbb{R}^C$ for the k-th class is obtained by applying average pooling to the masked features as $P_k = \operatorname{AvgPool}(F_s \otimes B_k)$, where \otimes denotes element-wise multiplication. If no voxels belong to class k, P_k is set to zero for consistency. The group-wise prototypes are generated by gathering the

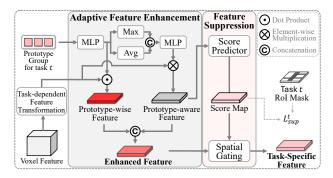


Figure 4. **Detailed structure of TSFG.** TSFG refines task-relevant features by leveraging the prototype group, followed by feature suppression to mitigate task-irrelevant information, resulting in task-specific features for each task.

prototypes for each group and adding learnable embeddings, i.e., $P_{FG} = \{P_k + E_k\}_{k \in \mathcal{K}_{fg}} \in \mathbb{R}^{N_{fg} \times C}$ and $P_{BG} = \{P_k + E_k\}_{k \in \mathcal{K}_{bg}} \in \mathbb{R}^{N_{bg} \times C}$, where N_{fg} and N_{bg} represent the number of foreground and background classes, and E_k is a learnable embedding. Finally, the prototype groups P_{FG} and P_{BG} are assigned to the appropriate task. Considering the objective of 3D perception tasks, the group assignment can be determined as $G_{\mathrm{Det}} = P_{\mathrm{FG}}$, $G_{\mathrm{Map}} = P_{\mathrm{BG}}$, and $G_{\mathrm{Occ}} = P_{\mathrm{FG}} \cup P_{\mathrm{BG}}$, where G_{Det} , G_{Map} , and G_{Occ} are the prototype groups assigned to 3D object detection, BEV map segmentation, and 3D occupancy prediction tasks, respectively. These task-specific prototype groups G_{Det} , G_{Map} , and G_{Occ} are used for adaptive feature enhancement in TSFG.

3.3. Task-Specific Feature Generator (TSFG)

The structure of the TSFG is shown in Figure 4. TSFG transforms the shared backbone features into task-specific features tailored to each of the three 3D perception tasks. The TSFG module operates through three sequential stages: (1) Task-dependent Feature Transformation, (2) Adaptive Feature Enhancement, and (3) Feature Suppression.

Task-dependent Feature Transformation. Taskdependent Feature Transformation module first transforms the shared voxel features F_s into representations suited for each task. For BEV-oriented tasks, such as 3D object detection and BEV map segmentation, it generates features in the BEV domain. In contrast, for voxel-oriented tasks like 3D occupancy prediction, it produces features in the voxel domain. Specifically, for BEV-oriented tasks, the height axis Z of F_s is collapsed into the channel dimension, and a series of 2D convolutional layers is applied to produce BEV features $F_t^{\text{BEV}} \in \mathbb{R}^{C \times X \times Y}$, where $t \in \{\text{Det}, \text{Map}\}$. For voxel-oriented tasks, F_s is processed with a sequence of 3D convolutional layers to generate voxel features $F_{\mathrm{Occ}}^{\mathrm{voxel}} \in \mathbb{R}^{C \times X \times Y \times Z}$. These transformed features are then passed to the Adaptive Feature Enhancement module.

Adaptive Feature Enhancement. Adaptive Feature Enhancement module enhances the transformed features F_t^{BEV} or F_t^{voxel} using the corresponding prototype group G_t , where $t \in \{\text{Det}, \text{Map}, \text{Occ}\}$. After projecting the prototypes in G_t through an MLP, it generates the prototypewise features by taking a dot-product operation between each prototype group and the transformed features. Denote the prototype-wise features as $F_{\text{wise}}^{\text{BEV}} \in \mathbb{R}^{N_{G_t} \times X \times Y}$ for BEV-oriented tasks and $F_{\text{wise}}^{\text{voxel}} \in \mathbb{R}^{N_{G_t} \times X \times Y \times Z}$ for voxel-oriented tasks, where N_{wise} voxel-oriented tasks, where N_{G_t} is the number of prototypes in G_t . In effect, these prototype-wise features activate spatial regions that are semantically aligned with G_t . In parallel, Adaptive Feature Enhancement module generates the prototype-aware features via channel attention operation. First, the global context features are extracted from the prototype group G_t through max and average pooling. These pooled features are concatenated and processed by an MLP, producing a channel scaling vector $\gamma_t^{scale} \in$ \mathbb{R}^C . This scaling vector modulates the transformed features F_t^{BEV} or F_t^{voxel} via element-wise multiplication, resulting in prototype-aware features $F_{\mathrm{aware}}^{\mathrm{BEV}} \in \mathbb{R}^{C \times X \times Y}$ or $F_{\mathrm{aware}}^{\mathrm{voxel}} \in \mathbb{R}^{C \times X \times Y \times Z}$. Finally, the prototype-wise features and the prototype-aware features are concatenated and passed through convolution layers to yield the enhanced feature F_t , i.e.,

$$\tilde{F}_{t} = \begin{cases} \operatorname{Conv2D}(\operatorname{Cat}[F_{\text{wise}}^{\text{BEV}}; F_{\text{aware}}^{\text{BEV}}]), & t \in \{\text{Det, Map}\}, \\ \operatorname{Conv3D}(\operatorname{Cat}[F_{\text{wise}}^{\text{voxel}}; F_{\text{aware}}^{\text{voxel}}]), & t = \operatorname{Occ}, \end{cases}$$
(2)

where $\operatorname{Cat}[\cdot;\cdot]$ indicates channel-wise concatenation. The enhanced features and the prototype-aware features are passed to the subsequent suppression module.

Feature Suppression. While the previous process emphasizes task-relevant information using guidance from a prototype group, the Feature Suppression module mitigates irrelevant components to mitigate task interference. It applies a lightweight CNN to the prototype-aware features, generating suppression scores, which ideally approach one within the Region of Interest (RoI) region and zero elsewhere. Here, the RoI region is specified by RoI masks, which indicate bounding boxes in the BEV domain for 3D object detection, semantically active areas for BEV map segmentation, and occupied voxels for 3D occupancy prediction. The Suppression Score Map $S_t^{\rm supp}$ is computed from the prototype-aware features as

$$S_t^{\text{supp}} = \begin{cases} f_{score}^{2D}(F_{\text{aware}}^{\text{BEV}}) \in \mathbb{R}^{X \times Y}, & t \in \{\text{Det, Map}\}, \\ f_{score}^{3D}(F_{\text{aware}}^{\text{voxel}}) \in \mathbb{R}^{X \times Y \times Z}, & t = \text{Occ}, \end{cases}$$
(3)

where f_{score}^{2D} and f_{score}^{3D} denote the score predictors for BEV and voxel-oriented tasks, respectively. Note that the generation of the Suppression Score Map is supervised using the

ground truth. The resulting Suppression Score Map S_t^{supp} is multiplied to the enhanced feature \tilde{F}_t through a gating operation. This operation yields the task-specific features as $F_t^{\mathrm{TS}} = \tilde{F}_t \otimes S_t^{\mathrm{supp}}$, where \otimes represents element-wise multiplication. This design effectively filters out task-irrelevant components from the input features. These resulting F_t^{TS} are subsequently utilized by task-dedicated heads and the SPA.

3.4. Scene Prototype Aggregator (SPA)

SPA operates in two stages. First, it generates task-oriented prototypes based on the predicted outputs from the object detection and map segmentation heads. Then, these task-oriented prototypes are integrated into the original prototype group, resulting in the scene prototypes that serve as initial queries for the 3D occupancy decoder. (see Figure 3 (c).)

Task-oriented prototype generation. SPA produces task-oriented prototypes using the predicted bounding boxes from the object detection head and the segmentation masks from the BEV map segmentation head. Specifically, detection prototypes $P_{\mathrm{Det}} \in \mathbb{R}^{N_{\mathrm{Det}} \times C}$ are generated by applying RoIAlign [11] to the task-specific features $F_{\mathrm{Det}}^{\mathrm{TS}}$, followed by class-wise average pooling over the predicted bounding boxes, where N_{Det} denotes the number of detection classes. In parallel, map prototypes $P_{\mathrm{Map}} \in \mathbb{R}^{N_{\mathrm{Map}} \times C}$ are computed by masked average pooling over the task-specific features $F_{\mathrm{Map}}^{\mathrm{TS}}$, using the predicted masks, where N_{map} is the number of classes defined in the BEV map segmentation task. Both P_{Det} and P_{Map} constitute task-oriented prototypes.

Prototype aggregation. SPA employs explicit semantic aggregation rules to align heterogeneous semantic label spaces across tasks. First, the prototypes belonging to categories that have a one-to-one semantic correspondence between the prototype groups G_t and the task-oriented prototypes are aggregated via direct class-wise summation. In contrast, prototypes associated with multiple fine-grained labels are first averaged and subsequently summed with their corresponding prototypes in G_t . This rule-based aggregation yields scene prototypes, which are used to initialize the occupancy queries in the occupancy decoder. By leveraging complementary semantics from 3D object detection and BEV map segmentation tasks, SPA enhances occupancy prediction performance while preserving the effectiveness of the other tasks.

3.5. Training Loss

The total loss function L_{total} is given by

$$L_{total} = L_{depth} + L_{CPG} + L_{Sup} + L_{det} + L_{map} + L_{occ},$$
 (4)

where L_{depth} is for depth estimation, L_{CPG} is for classwise mask classification in CPG, L_{Sup} is the suppression loss summed across tasks used for supervising suppression

score maps, and L_{det} , L_{map} , and L_{occ} correspond to the task-specific losses for 3D object detection, BEV map segmentation, and 3D occupancy prediction, respectively. Dice and Lovasz losses [1] are used in L_{CPG} while Focal loss [24] is utilized in L_{Sup} . The detection loss L_{det} is formulated based on CenterPoint [50] for object detection. The map segmentation loss L_{map} is adopted from BEVFusion [27], and the occupancy prediction loss L_{occ} is based on the loss function employed in OccFormer [53].

4. Experiments

4.1. Experimental Settings

Datasets and metrics. Our experiments were conducted on the nuScenes [2] dataset. We particularly used the annotation provided in Occ3D [36] for 3D occupancy prediction. For 3D object detection evaluation, we adopt the official metrics: mean Average Precision (mAP) and nuScenes Detection Score (NDS). BEV map segmentation performance is evaluated following [27, 56], using mean Intersection-over-Union (mIoU) across six predefined background classes. The 3D semantic occupancy prediction performance is evaluated using voxel-level mIoU across 17 semantic categories. The details of the dataset and the performance metrics are provided in the Supplementary Material. Implementation details. We utilized ResNet-50 [10] for the image backbone. Our model was trained for 24 epochs. without employing class-balanced grouping and sampling (CBGS) [59]. The AdamW optimizer is used with a learning rate of 1×10^{-4} and weight decay of 0.01, with a batch size of 8 distributed across 4 RTX 3090 GPUs. Baseline-STL refers to models independently optimized for each task, while Baseline-MTL denotes a naive multi-task learning approach without the proposed modules, as illustrated in Figure 2 (b). We also compare the performance of MAE-STRO with those of existing STL and MTL methods. We include BEVFusion [27] and PanoOcc [39] as MTL baselines, as they are the officially available multi-task methods on the nuScenes benchmark. Further training details are provided in the Supplementary Material.

4.2. Comparison with State-of-the-art Results

Table 1 presents the performance of MAESTRO on the nuScenes validation set. We observe that the Baseline-MTL model performs worse than the Baseline-STL model due to task conflicts inherent in naive MTL. In contrast, MAESTRO, which incorporates the proposed techniques, demonstrates substantial performance improvements over Baseline-STL across all three tasks: a 2.6% increase in mAP for 3D object detection, a 3.8% increase in mIoU for BEV map segmentation, and a 2.1% increase in mIoU for 3D occupancy prediction. The improvements are even more pronounced when compared to Baseline-MTL. MAE-

Method	Venue	Image Backbone	Image Size	mAP	NDS	mIoU (Map)	mIoU (Occ)	Latency (ms)
BEVDet [13]	arXiv'21	ResNet-50	256×704	29.8	37.9	-	-	51.5
DETR3D [37]	CoRL'22	ResNet-50	800×1333	30.3	37.4	-	-	-
BEVDepth * [21]	AAAI'23	ResNet-50	256×704	33.7	41.4	-	-	-
BEVFormer v2 [46]	CVPR'23	ResNet-50	640×1600	35.1	41.4	-	-	171.9
DualBEV † [20]	ECCV'24	ResNet-50	256×704	35.2	42.5	-	-	65.1
CVT [56]	CVPR'22	ResNet-50	256×704	-	-	37.7	-	33.9
LSS [32]	ECCV'20	ResNet-50	256×704	-	-	41.0	-	72.4
BEVFusion [27]	ICRA'23	ResNet-50	256×704	-	-	47.1	-	45.6
DifFUSER [18]	ECCV'24	ResNet-50	256×704	-	-	48.3	-	92.2
MonoScene [3]	CVPR'22	ResNet-101	928×1600	-	-	-	6.1	830.1
OccFormer [53]	ICCV'23	ResNet-50	928×1600	-	-	-	21.9	349.0
TPVFormer [14]	CVPR'23	ResNet-101	928×1600	-	-	-	27.8	320.8
Vampire [45]	AAAI'24	ResNet-101	256×704	-	-	-	28.3	349.2
CTF-Occ [36]	NIPS'24	ResNet-101	928×1600	-	-	-	28.5	-
SurroundOcc [40]	ICCV'23	ResNet-101	800×1333	-	-	-	34.6	355.6
FB-Occ [23]	ICCV'23	ResNet-50	256×704	-	-	-	37.4	129.7
BEVFusion [27]	ICRA'23	ResNet-50	256×704	33.6	39.2	44.0	-	-
PanoOcc [39]	CVPR'24	ResNet-50	864×1600	29.5	34.8	-	31.8	124.7
Baseline -STL	-	ResNet-50	256×704	33.8	41.7	47.5	36.5	405.9
Baseline-MTL	-	ResNet-50	256×704	32.7	38.2	43.5	36.0	219.6
Ours-MTL	-	ResNet-50	256×704	36.4	43.2	51.3	38.6	250.3

Table 1. Performance comparison evaluated on the nuScenes validation set. "-" indicates tasks not supported by the method. All BEV map segmentation and MTL results were reproduced using official code with ResNet-50 backbone. "*" denotes the performance reported by DualBEV [20], † indicates the method using CBGS for training. Our MTL approach achieves state-of-the-art performance across all tasks.

CPG	Det.	TSFG Map.	Occ.	SPA	mAP	NDS	mIoU (Map)	mIoU (Occ)
					31.3	32.3	40.4	33.6
\checkmark					31.3	32.4	40.3	34.6
\checkmark	✓				32.6	34.1	39.1	34.1
\checkmark		✓			30.8	31.9	44.0	34.3
\checkmark			✓		31.1	32.2	41.2	35.9
\checkmark	✓	✓	✓		32.5	34.1	44.1	35.9
✓	√	√	√	√	32.6	34.3	44.2	36.9

Table 2. Ablation study for evaluating the main components in MAESTRO.

STRO significantly reduces latency by 155.6 ms compared to Baseline-STL, with only a marginal increase of 30.7 ms over Baseline-MTL. We also compare MAESTRO against existing STL and MTL methods. Notably, MAESTRO significantly outperforms the previous best-performing STL models for each individual task. Furthermore, compared to existing MTL approaches, MAESTRO achieves state-of-the-art performance.

4.3. Ablation Studies

We conducted ablation studies to evaluate the effectiveness of the proposed ideas. Our ablation study was conducted on 1/4 of the nuScenes training set for 24 epochs with all other settings unchanged.

Component analysis. Table 2 presents the performance of the proposed method as each component is incrementally

Method	mAP	NDS	mIoU (Map)	mIoU (Occ)
Baseline-MTL with CPG	31.3		40.3	34.6
+ Prototype-Wise Features	31.9	32.7	42.1	35.2
+ Prototype-Aware Features	32.2	33.6	43.4	35.7
+ Feature Suppression	32.5	34.1	44.1	35.9

Table 3. Ablation study for evaluating the components of TSFG.

added. The baseline model adopts a naive MTL approach that shares features across all tasks without any of the proposed enhancements, as illustrated in Figure 2(b). Integrating the CPG into the baseline yields a 1.0% mIoU improvement in occupancy prediction without impacting the performance for other tasks. This demonstrates the benefit of incorporating semantic prototypes. Applying the TSFG to a single task enhances the performance of that task but results in performance degradation in others due to task imbalance. When TSFG is applied to all tasks, it effectively mitigates task interference, leading to improvements of 1.2% mAP in 3D object detection, 3.8% mIoU in BEV map segmentation, and 1.3% mIoU in 3D occupancy prediction. Finally, the inclusion of the SPA further boosts occupancy prediction by an additional 1.0% mIoU without compromising the performance of the other tasks.

Impact of adaptive feature enhancement and suppression in TSFG. Table 3 presents an ablation study analyzing the impact of each component within TSFG. Starting from the Baseline-MTL with CPG, we progressively add the key components of TSFG to evaluate their individual contribu-

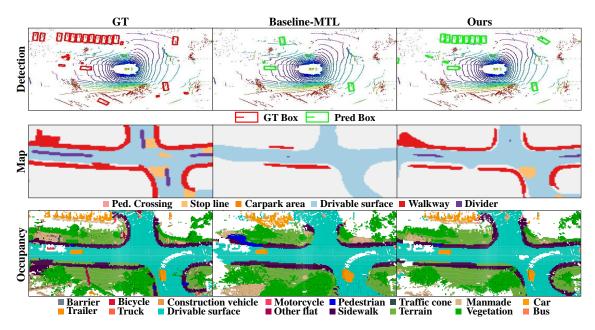


Figure 5. Qualitative results on the nuScenes validation set. From left to right, we show the qualitative results of Ground Truth, Baseline-MTL, and our MAESTRO.

Method	mIoU (Occ)		
MAESTRO	36.9		
w/o Map Prototypes w/o Det Prototypes	36.3 (-0.6) 35.9 (-1.0)		

Table 4. Ablation study for evaluating components of prototype aggregation used in SPA.

tions. Adding the Prototype-Wise Features results in a performance gain of 0.6% mAP in 3D object detection, 1.8% mIoU in BEV map segmentation, and 0.6% mIoU in 3D occupancy prediction. Incorporating the Prototype-Aware Features leads to an additional gain of 0.3% mAP, 1.3% mIoU, and 0.5% mIoU for the respective tasks. Finally, adding the Feature Suppression module further improves performance, with a 0.3% mAP gain in 3D object detection, and 0.7% and 0.2% mIoU gains in BEV map segmentation and 3D occupancy prediction, respectively.

Effect of task-oriented prototypes for prototype aggregation in SPA. Table 4 demonstrates the impact of task-oriented prototypes by progressively removing each component from the full MAESTRO model. Removing the map prototypes leads to a 0.6% decrease in 3D occupancy prediction mIoU, highlighting their role in capturing background semantics. Further excluding detection prototypes causes an additional drop of 0.4% mIoU, emphasizing their importance for foreground representation. These results confirm that jointly aggregating both types of prototypes significantly enhances 3D occupancy prediction performance.

4.4. Qualitative Results

Figure 5 presents a qualitative comparison between Baseline-MTL, our method, and the ground truth across the three perception tasks. As shown, our approach yields more accurate predictions across all tasks compared to Baseline-MTL. Additional qualitative results are provided in the Supplementary Material.

5. Conclusions

In this paper, we presented MAESTRO, a unified framework for multi-task 3D perception that explicitly mitigates feature interference while enhancing task efficiency. MAE-STRO introduces the CPG, which groups semantic categories into foreground and background sets and generates group-wise prototypes that serve as semantic priors for taskspecific feature enhancement. The TSFG leverages these priors to selectively enhance task-relevant features while suppressing irrelevant regions, enabling effective and discriminative feature learning for each task. Additionally, the SPA enriches the prototype groups used for 3D occupancy prediction by integrating complementary information from the 3D object detection and map segmentation heads. This leads to improved occupancy performance without degrading the accuracy of the other tasks. Extensive evaluations on the nuScenes and Occ3D benchmarks demonstrate that MAESTRO achieves state-of-the-art performance across object detection, BEV map segmentation, and occupancy prediction tasks, validating the effectiveness and robustness of our structured multi-task learning approach.

References

- [1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 6
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of* the IEEE/CVF conference on computer vision and pattern recognition, pages 11621–11631, 2020. 2, 6
- [3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 3, 7
- [4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 3
- [5] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. Advances in Neural Information Processing Systems, 33:2039–2050, 2020.
- [6] Wonhyeok Choi, Mingyu Shin, Hyukzae Lee, Jaehoon Cho, Jaehyeon Park, and Sunghoon Im. Multi-task learning for real-time autonomous driving leveraging task-adaptive attention generator. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 14732–14739. IEEE, 2024. 2
- [7] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings* of the IEEE/CVF Conference on computer vision and pattern recognition, pages 11543–11552, 2020. 3
- [8] Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *Proceedings of the IEEE/CVF Interna*tional Conference on Computer Vision, pages 8721–8731, 2023. 2, 3
- [9] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on com*puter vision (ECCV), pages 270–287, 2018. 3
- [10] 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. 6
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [12] Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fas-

- tocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 16425–16431. IEEE, 2024. 1, 3
- [13] 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. 1, 3, 7
- [14] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 3, 7
- [15] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multicamera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 1042–1050, 2023. 1, 3
- [16] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on* computer vision and pattern recognition, pages 7482–7491, 2018. 3
- [17] Jungho Kim, Changwon Kang, Dongyoung Lee, Sehwan Choi, and Jun Won Choi. Protoocc: Accurate, efficient 3d occupancy prediction using dual branch encoder-prototype query decoder. arXiv preprint arXiv:2412.08774, 2024. 3
- [18] Duy-Tho Le, Hengcan Shi, Jianfei Cai, and Hamid Rezatofighi. Diffusion model for robust multi-sensor fusion in 3d object detection and bev segmentation. In *European Conference on Computer Vision*, pages 232–249. Springer, 2024. 2, 7
- [19] Chunliang Li, Wencheng Han, Junbo Yin, Sanyuan Zhao, and Jianbing Shen. Repvf: A unified vector fields representation for multi-task 3d perception. In *European Conference* on Computer Vision, pages 273–292. Springer, 2024. 2
- [20] Peidong Li, Wancheng Shen, Qihao Huang, and Dixiao Cui. Dualbev: Unifying dual view transformation with probabilistic correspondences. In *European Conference on Computer Vision*, pages 286–302. Springer, 2024. 1, 3, 7
- [21] 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*, pages 1477–1485, 2023. 3, 7
- [22] 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. 1, 3
- [23] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492, 2023. 1, 3, 7
- [24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Pro-

- ceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 6
- [25] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. Advances in Neural Information Processing Systems, 34:18878–18890, 2021. 3
- [26] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2
- [27] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multitask 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. 1, 2, 3, 6, 7
- [28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S Yu. Learning multiple tasks with multilinear relationship networks. Advances in neural information processing systems, 30, 2017. 3
- [29] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016. 3
- [30] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. arXiv preprint arXiv:2006.11436, 2020. 1, 3
- [31] 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. 3
- [32] 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. 1, 3, 7
- [33] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *Proceedings of the AAAI conference on artificial intelli*gence, pages 4822–4829, 2019. 3
- [34] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 3
- [35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017.
- [36] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems, 36:64318–64330, 2023. 1, 3, 6, 7
- [37] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d:

- 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 7
- [38] Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Frustum-former: Adaptive instance-aware resampling for multi-view 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5096–5105, 2023. 1, 3
- [39] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17158–17168, 2024. 2, 6, 7
- [40] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 3, 7
- [41] Yiming Wu, Ruixiang Li, Zequn Qin, Xinhai Zhao, and Xi Li. Heightformer: Explicit height modeling without extra data for camera-only 3d object detection in bird's eye view. *IEEE Transactions on Image Processing*, 2024. 1, 3
- [42] Zhongyu Xia, Zhiwei Lin, Xinhao Wang, Yongtao Wang, Yun Xing, Shengxiang Qi, Nan Dong, and Ming-Hsuan Yang. Henet: Hybrid encoding for end-to-end multi-task 3d perception from multi-view cameras. In *European Conference on Computer Vision*, pages 376–392. Springer, 2024.
- [43] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M² bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. arXiv preprint arXiv:2204.05088, 2022. 1, 3
- [44] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 675–684, 2018. 3
- [45] Junkai Xu, Liang Peng, Haoran Cheng, Linxuan Xia, Qi Zhou, Dan Deng, Wei Qian, Wenxiao Wang, and Deng Cai. Regulating intermediate 3d features for vision-centric autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6306–6314, 2024. 7
- [46] 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. 1, 3, 7
- [47] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *European Conference on Computer Vision*, pages 514–530. Springer, 2022.
- [48] Hanrong Ye and Dan Xu. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21828–21837, 2023. 3

- [49] Hanrong Ye and Dan Xu. Invpt++: Inverted pyramid multitask transformer for visual scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
- [50] 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. 6
- [51] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zong-dai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arXiv preprint arXiv:2311.12058, 2023. 3
- [52] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3348–3357, 2023. 1, 3
- [53] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1, 3, 6, 7
- [54] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, pages 94–108. Springer, 2014. 3
- [55] Xiao Zhao, Xukun Zhang, Dingkang Yang, Mingyang Sun, Mingcheng Li, Shunli Wang, and Lihua Zhang. Maskbev: Towards a unified framework for bev detection and map segmentation. In *Proceedings of the 32nd ACM International* Conference on Multimedia, pages 2652–2661, 2024. 2
- [56] 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. 1, 6, 7
- [57] Qiu Zhou, Jinming Cao, Hanchao Leng, Yifang Yin, Yu Kun, and Roger Zimmermann. Sogdet: Semantic-occupancy guided multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7668–7676, 2024. 2, 3
- [58] Zixiang Zhou, Dongqiangzi Ye, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarformer: A unified transformer-based multi-task network for lidar perception. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 14740–14747. IEEE, 2024. 2
- [59] 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. 6