Seeing in Double: Dual-Granularity BEV Segmentation via Mamba-Driven Alignment and Polar-Decoupled Experts

Jiaxin Cai[1], Rui Lin[1], Jingze Su[1], Qi Li[1], Wenjie Yang[1], Yuanlong Yu[1], Wenxi Liu[1]*

[1]College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China

Abstract

Bird's Eye View (BEV) representation has become pivotal for autonomous driving, yet existing polar coordinate-based approaches face two critical limitations: (1) distant semantic misprojection caused by radial resolution decay, and (2) region-specific geometric distortions from non-uniform polar discretization. To address these issues, we propose a novel framework addressing these challenges through three key innovations. First, we present a bilateral heterogeneous network constructs multi-granularity BEV spaces, efficiently exploiting dual-resolution visual information for distant detail preservation. Second, we employ an align-fusion strategy for multi-granularity feature aggregation. Specifically, the Mamba-Based Cross-Resolution Alignment module establishes semantic consistency for perspective features through shared state-space optimization. In the later stage, the Adaptive BEV Space Selector dynamically aggregates multi-granularity BEV features. Third, we introduce a Mixture of Radial-Angular Decoupled Experts, which employs polar-aware expert routing to disentangle radial compression and angular shear distortions through specialized geometric refinement. Comprehensive experiments on nuScenes and Lyft L5 demonstrate the state-of-the-art performance of our model across various resolution settings, visibility filtering, and perception ranges.

Introduction

Vision-based Bird's Eye View (BEV) representation has emerged as a cornerstone for autonomous driving perception systems. By transforming multi-camera images into a unified BEV space, it enables reasoning for detection (Li et al. 2022; Liu et al. 2023c), segmentation (Liu et al. 2024b; Cai et al. 2024), and planning (Hu et al. 2023).

In this paper, we focus on BEV segmentation from multiple cameras. Recent query-based BEV paradigms (Zhou and Krähenbühl 2022; Yang et al. 2023) have demonstrated remarkable progress by initializing a regular grid of BEV queries in Cartesian coordinates and iteratively refining them through cross-view feature sampling. However, such Cartesian-based initialization fundamentally conflicts with the wedge-shaped imaging geometry of cameras – a physical constraint where radially distributed 3D space is projected onto perspective views with non-orthogonal axes (Jiang et al. 2023; Liu et al. 2023b). To address this geometric mismatch, emerging methods (Chen, Vora, and Beijbom 2021; Chu et al. 2024) adopt polar coordinate BEV representations, achieving superior performance through radial-angular spatial alignment with camera imaging principles.

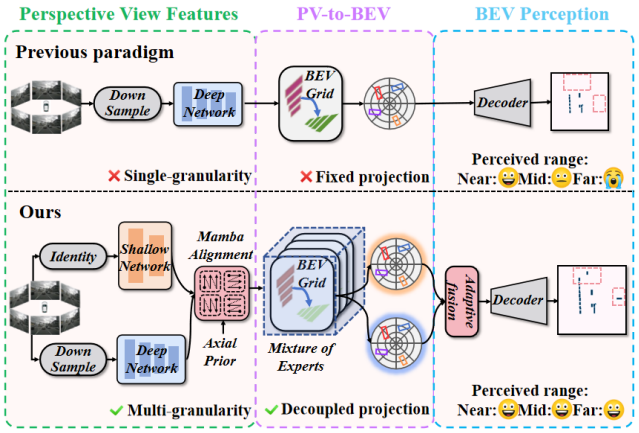


Figure 1: In contrast to prior BEV perception paradigms, our framework introduces a dual-resolution heterogeneous network that constructs aligned multi-granularity BEV spaces to mitigate distant semantic misprojection in polar coordinates. For PV-to-BEV transformation, we propose a decoupled radial-angular MoE projection, which adaptively counteracts geometric perturbations via subspace-specialized expert routing. Compared to existing methods, our approach achieves superior performance across diverse perception ranges. (Please zoom for details.)

Despite these advancements, two critical challenges persist. *i) Distant Semantic Misprojection:* In polar BEV grids, the radial resolution decreases with distance from the ego-vehicle, leading to dense query distributions in nearby regions and sparse distributions in distant areas. During the transformation from Perspective View (PV) to BEV, sparse distant BEV queries are required to sample from highly compressed image features. Since distant objects occupy only a minimal pixel width on the image plane, even minor misalignments in query projection positions can lead to a substantial loss of critical geometric information during the

*Corresponding author.

view transition. *ii) Region-specific Distortion:* Non-uniform polar discretization introduces geometric distortions that vary radially and angularly, such as compression effects at large radii and shear distortion at oblique angles. Existing methods (Liu et al. 2023b; Jiang et al. 2023) utilize uniform height prediction across all regions, which lack the adaptability to address these localized distortion patterns, necessitating region-aware geometric compensation. In Fig. 1, we propose a novel paradigm to address the above issues.

To address distant semantic misprojection, we propose a dual-resolution heterogeneous network that constructs multi-granularity BEV spaces. Unlike prior works that solely project low-resolution features into a single BEV contextual space, we introduce a parallel lightweight branch to generate BEV features enriched with high-resolution fine-grained cues. This branch retains critical distant spatial details (e.g., small objects at depth) while maintaining computational efficiency. For multi-granularity feature aggregation, we employ an align-fusion strategy. Specifically, to address feature misalignment in dual-resolution feature spaces, we propose a Mamba-Based Cross-Resolution Alignment (MCRA) module. The proposed MCRA module introduces a unified Mamba-based state space representation to jointly optimize cross-resolution features via collaborative parameter learning, enabling semantic consistency across resolutions through injecting axial priors. In the later stage, we incorporate an Adaptive BEV Space Selector (ABSS), which adaptively selects high-confidence regions from BEV feature and dynamically aggregates multi-granularity BEV feature spaces, enhancing robustness to scale variations.

Furthermore, as the core of BEV perception, the PV-to-BEV transformation on non-uniform polar discretizations suffers from severe region-specific geometric distortions existed in grid-wise height estimation. To address this concern, we propose a Mixture of Radial-Angular Decoupled Experts (MRADE), inspired by the adaptive routing principles of Mixture of Experts (MoE) architectures (Shazeer et al. 2017; Lepikhin et al. 2020). The MRADE dynamically assigns BEV queries to specialized experts based on polar coordinate dependencies, enabling localized geometric refinement. Specifically, we decouple experts into radial- and angular-dependent branches to align with distortion patterns inherent to polar coordinate systems. By coordinating complementary expertise through radial-angular collaborative learning, our framework achieves region-aware height distribution estimation, significantly improving feature sampling precision for distorted regions.

Overall, the contributions of this paper can be summarized as follows:

- For BEV segmentation, we propose a bilateral heterogeneous network architecture, which introduces an additional lightweight high-resolution branch specifically designed for capturing distant details. Our dual-resolution network design constructs multi-granularity BEV spaces that effectively and efficiently captures contextual cues of objects at various distances.
- For multi-granularity feature aggregation, we employ an align-fusion strategy. We first introduce the Mamba-Based Cross-Resolution Alignment module that jointly optimizes dual-resolution perspective features via shared state-space modeling, enabling linear-complexity heterogeneous semantic alignment. In the later stage, we leverage the adaptive BEV Space Selector to dynamically prioritizes and aggregates the most discriminative regions from aligned multi-granular BEV features.
- For PV-to-BEV transformation, We propose a Mixture of Radial-Angular Decoupled Experts to estimate regionally adaptive height distributions. By decoupling experts along radial and angular dimensions - each specializing in distortion patterns unique to their subspace - MRADE mitigates region-specific geometric perturbations in polar coordinate height estimation.
- Through comprehensive evaluations on the nuScenes and Lyft L5 benchmarks, our method surpasses previous state-of-the-art approaches across various resolution settings, visibility filtering, and perception ranges.

Related Work

Vision-based BEV Segmentation. BEV perception has received widespread attention in the field of autonomous driving (Liu et al. 2024a; Cai et al. 2025). However, these methods (Philion and Fidler 2020; Li et al. 2022; Liu et al. 2023a; Xu et al. 2024; Lu, Tsai, and Chen 2025) typically establish the BEV in a Cartesian coordinate system, which may not align well with the imaging geometry of cameras. To address this gap, some methods (Saha et al. 2022; Gong et al. 2022; Zhao et al. 2024; Chu et al. 2024) choose to construct polar coordinate representations for BEV perception. For instance, PolarFormer (Jiang et al. 2023) incorporates a polar alignment module to aggregate rays from multiple cameras. PolarBEV (Liu et al. 2023b) adopts a strategy of decomposing polar coordinate embeddings to build its representation. In contrast to previous methods that merely simplistically employ polar coordinate BEV representation, we conduct a thorough analysis of the inherent limitations embedded in polar grids and propose a novel framework.

State Space Models. State Space Models (SSMs), rooted in control theory, capture input-output dynamics through hidden state evolution (Gu, Goel, and Ré 2021). Mamba (Gu and Dao 2023) introduces the S6 module, achieving a simple structure with excellent efficiency in long-sequence relationship modeling. This breakthrough has led to many adaptations in computer vision (Hatamizadeh and Kautz 2024; He et al. 2025), including Vim (Zhu et al. 2024a) and VMamba (Liu et al. 2024c) that develop 2D scanning strategies for image data. Unlike these prior works, our approach considers utilizing shared Mamba state spaces across different resolutions and network branches. This allows us to achieve cross-scale semantic alignment through coordinated state evolution while maintaining geometric fidelity during subsequent polar coordinate transformations.

Mixture-of-Experts. Mixture-of-Experts (MoE) (Jacobs et al. 1991) can dynamically adjust its structure based on different inputs to accommodate complex data. Many successful examples of MoE have been demonstrated in different vision tasks, including image fusion (Cao et al. 2023; Zhu

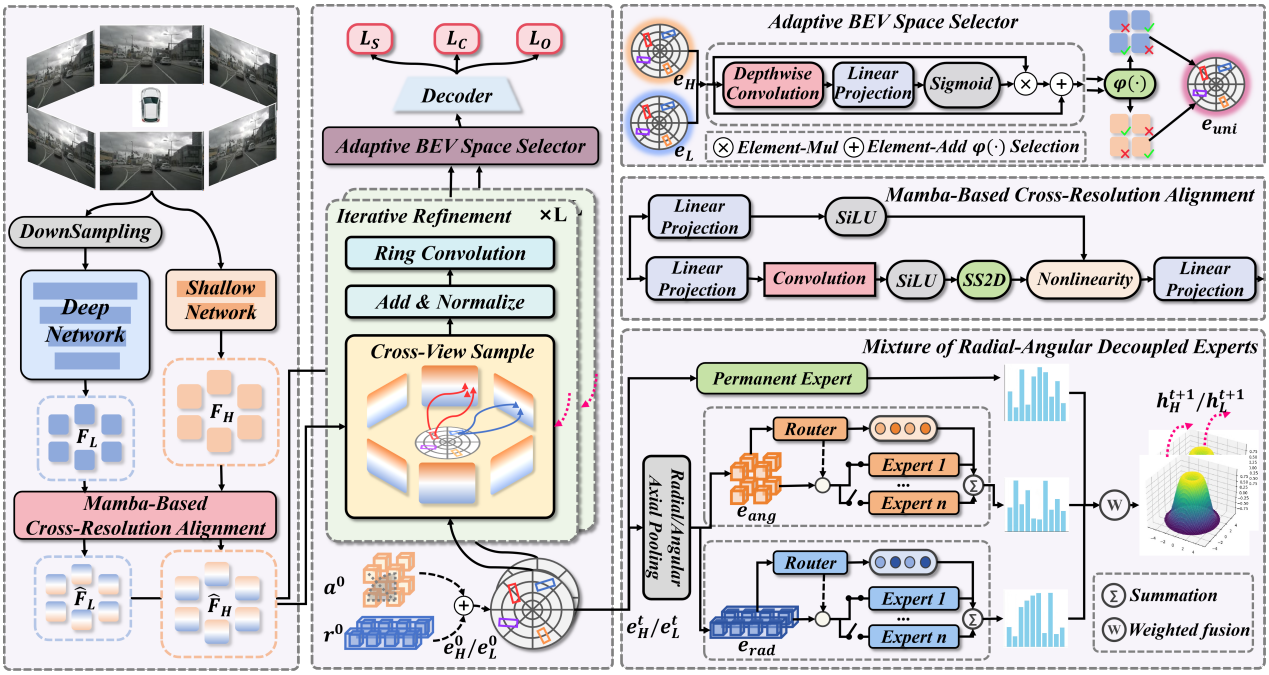


Figure 2: Given multi-view images, our framework first extracts multi-granularity features via bilateral dual-resolution networks. These features undergo Mamba-based Cross-Resolution Alignment (MCRA) for spatial-semantic consistency before being projected into polar-coordinate BEV spaces. For precise PV-to-BEV projection, Mixture of Radial-Angular Decoupled Experts (MRADE) are employed to estimate grid-wise heights through specialized expert routing. Finally, the Adaptive BEV Space Selector (ABSS) dynamically fuses dual BEV features by emphasizing regionally discriminative cues, producing a unified BEV representation for downstream segmentation tasks.

et al. 2024b), multi-task representations (Yang et al. 2024), novel view synthesis (Cong et al. 2023), and low-light image enhancement (Li et al. 2025). While existing methods typically employ MoE for general visual feature enhancement, we introduce a novel radial-angular decoupled MoE architecture specifically designed for polar coordinate-aware height distribution estimation.

Method

Overview

The proposed network architecture is illustrated in Fig. 2. Given multi-view images as input, our framework operates through four key stages: **1) Heterogeneous Feature Extraction:** Dual-resolution (deep/shallow) networks extract multi-view features, capturing complementary contextual semantics and fine-grained visual cues. **2) Cross-Resolution Semantic Alignment:** To bridge resolution-induced semantic gaps across heterogeneous encodings, the proposed Mamba-based Cross-Resolution Alignment (MCRA) module establishes unified feature correspondence. **3) Polar-coordinate BEV Feature Transformation:** Polar-coordinate BEV spaces are predefined for sampling multi-resolution features, where the proposed Radial-Angular Decoupled Experts (MRADE) estimates grid-wise heights via specialized experts while transforming perspective features into BEV space using predicted heights and camera parameters. **4) Adaptive BEV Space Aggregation:** A lightweight Adaptive BEV Space Selector (ABSS) learns to dynamically fuse dual BEV representations by emphasizing regionally discriminative features. The unified BEV representation is subsequently decoded for a comprehensive BEV segmentation task. Technical details of these above modules are elaborated in the following.

Dual-Resolution Heterogeneous Networks

Existing polar-based BEV perception methods rely on single-encoder architectures to extract multi-view features at a unified resolution. This imposes a fundamental challenge: lower-resolution features compromise polar rasterization's capacity to preserve distant small objects, while higher resolutions incur prohibitive computational costs. To address this issue, we propose a dual-resolution bilateral architecture. This architecture consists of a conventional deep feature encoder and a lightweight shallow encoder, which is specifically designed for high-resolution feature extraction. This design helps avoid over-compression of distant small targets while maintaining computational efficiency.

However, aggregating dual-resolution encoded features from perspective image spaces introduces camera intrinsic parameters mismatch between resolution streams, causing geometric ambiguity during view transformation. To address this, we propose projecting dual-granularity features into separate BEV spaces to preserve their distinct advantages. Specifically, we first harmonize heterogeneous fea-

tures in perspective space via implicit alignment, retaining their granularity-specific properties. The aligned features are then mapped into two complementary BEV spaces - one capturing fine-grained spatial patterns and the other global context. These representations are adaptively unified, forming a phased alignment-to-unification paradigm that ensures geometric coherence while maximizing complementary granularity strengths. The above steps will be detailed in the following section.

Mamba-Based Cross-Resolution Alignment

Given high-resolution features F_H and low-resolution features F_L originated from dual-resolution networks, the MCRA module is introduced to handle heterogeneous features through a shared Mamba-like architecture (Gu and Dao 2023), which implicitly aligns the features and eliminates their inherent semantic biases. This design enforces both feature streams to evolve within a shared state space while preserving their resolution-specific characteristics, enabling our model to simultaneously resolve semantic discrepancies while maintaining geometric fidelity for polar coordinate transformation. Specifically, inspired by (Liu et al. 2024c), we incorporate a 2D-selective-scanning (SS2D) operation along the axial dimension to address this problem. This technique expands image patches along four directions, generating four unique feature sequences to establish axis-aware semantic correlations. For our task, this Mamba-based module inherently aligns with the radial-angular structure of polar BEV space, naturally preserving directional patterns and providing structural priors for subsequent PV-to-BEV transformation. Formally, given the high- or low-resolution feature F_k ($k = \{H, L\}$), the entire procedure of our proposed module can be expressed as below:

$$\begin{aligned} \bar{F}_k &= \mathrm{LN}(\mathrm{SS2D}(\mathrm{SiLU}(\mathrm{Conv}(\mathrm{Linear}(F_k))))), \\ G &= \mathrm{SiLU}(\mathrm{Linear}(F_k)), \\ \hat{F}_k &= \mathrm{Linear}(G * \bar{F}_k) + F_k, \end{aligned} \tag{1}$$

where $\mathrm{Conv}(\cdot)$ represents convolution operation, $\mathrm{LN}(\cdot)$ is layer normalization, and $\mathrm{SiLU}(\cdot)$ stands for SiLU activation function. An input-dependent gating mechanism is incorporated to selectively identify and utilize the most advantageous representations produced by the SS2D process.

PV-to-BEV Transformation

Given perspective features from high- and low-resolution features that are aligned through the previous module, we transform them into heterogeneous BEV representations separately in this stage. However, the major challenge of PV-to-BEV transformation in polar coordinate space lies in the regional geometric distortions in polar coordinates, often resulting in inaccurate height estimation.

Here, we employ a height-based iterative view projection scheme for PV-to-BEV view transformation. The BEV space can be discretized into grids, and thus the description of our method in the following is based on the operation on each grid.

$$h^{t+1} \leftarrow \Phi(e^t) + h^t, \tag{2}$$

$$e^{t+1} \leftarrow \mathcal{T}(h^{t+1}, \hat{F}), t \in \{1, ..., T\} \tag{3}$$

where h^t refers to the estimated height at the t-th iteration for a certain grid, and e^t denotes its learnable polar-coordinate BEV embeddings. The iteration process is initialized with the predefined height h_0 and an initial embedding e^0 (Note: Instead of directly initializing e^0, we initialize radial r^0 and angular a^0 embeddings separately and sum them via broadcasting to obtain e^0). During the iteration, $\Phi(\cdot)$ refines the estimated height h^t at the previous timestep t and $\mathcal{T}(\cdot)$ leverages the estimated height to update the polar-coordinate BEV embeddings from the feature $\hat{F}$. The computational steps described above and in the following apply independently and identically to high- and low-resolution processing; we omit the subscripts $\{H, L\}$ here for brevity.

Mixture of Radial-Angular Decoupled Experts. The core of the view transformation lies in $\Phi(\cdot)$ which is accomplished as the proposed Mixture of Radial-Angular Decoupled Experts (MRADE) module. Motivated by the geometric properties of polar embeddings, MRADE employs a decoupled mixture-of-experts architecture where distinct experts specialize in either radial or angular subspaces. Each expert independently predicts height distributions within designated radial-angular regimes, effectively mitigating region-specific interference caused by geometric variations.

As illustrated in Fig. 2, the radial-specific embeddings e_{rad} is disentangled from the grid embedding e (the superscript ignored for simplicity) through radial average pooling, and then directed to a radial routing layer, yielding specialization weights W_{rad} that prioritize experts for distinct radial regimes, thereby minimizing cross-region interference. An analogous procedure applies to the angular-specific embeddings e_{ang}. To ensure consistency across varying perspectives, we introduce an additional permanent expert E_{per}, which is responsible for extracting universal knowledge from the entire scene. Thus, the MRADE module can be briefly described as below:

$$\Phi = \lambda_1 W_{rad} \odot \mathbf{E}_{rad} + \lambda_2 W_{ang} \odot \mathbf{E}_{ang} + \lambda_3 \mathbf{E}_{per}, \tag{4}$$

where $\mathbf{E}_{rad}$, $\mathbf{E}_{ang}$, and $\mathbf{E}_{per}$ represent the groups of experts specialized in radial, angular, and global domains. λ_1, λ_2, and λ_3 are the balancing factors for different experts. W_{rad} and W_{ang} refer to the weights of prioritizing experts, which are computed via routing layers and determine each expert's contribution to the final height prediction. In practice, we formulate the selection of an expert E from experts $\mathbf{E}_{rad}$ or $\mathbf{E}_{ang}$ as conditional probability distribution given a certain embedding e, as below:

$$P(E|e) = \mathrm{Softmax}(\alpha e + \mathcal{N}(0,1)\,\mathrm{Softplus}(\beta e)), \tag{5}$$

where α refers to the parameters of the routing layer. $\mathcal{N}(0, 1)$ represents the standard normal distribution. β denotes a noise term that injects stochasticity into expert selection, enabling diverse expert exploration in training. Thus, the weights of prioritizing experts can be sampled from the above distribution, i.e.,

$$W_{rad} = \mathrm{TopK}(P(E|e_{rad}, E \in \mathbf{E}_{rad})), \tag{6}$$

$$W_{ang} = \mathrm{TopK}(P(E|e_{ang}, E \in \mathbf{E}_{ang})). \tag{7}$$

View Transformation. According to Eq. 2, we can obtain the refined estimated height (i.e., h^{t+1}). Next, with the estimated height, we can further update the polar-coordinate

BEV embeddings for the next iteration, according to Eq. 3. We normalize h (the superscript t omitted for clarity) to the range of $[0, 1]$ with sigmoid function, which can further be upscaled to the predefined height range to obtain z in the world space. Then, the homogeneous coordinate (x, y, z) of a specific grid in the BEV space can be calculated, where $x = \rho\cos(\theta)$ and $y = \rho\sin(\theta)$. We project the coordinate back to the perspective space using camera's intrinsics and extrinsics. Finally, we transform features from BEV to PV with the projected coordinates and extract the corresponding feature, which can expressed as:

$$\begin{aligned} e &= \mathcal{T}(\hat{F}(f(x, y, z))), \\ &= \mathcal{T}(\hat{F}(x^p, y^p)) = A \cdot M \cdot \hat{F}(x^p, y^p), \end{aligned} \tag{8}$$

where $f(\cdot)$ represents the BEV-to-PV coordinate transformation. $\hat{F}(x^p, y^p)$ refers to the features sampled at the 2D coordinate (x^p, y^p) in the image space. In addition, M is a binary mask for masking out projected points that exceed the image boundary and A denotes the predicted attention weight. After that, the output embedding e is sent into a ring convolution layer and added to the previous BEV embedding for updating.

Adaptive BEV Space Selector

After view transformation, we obtain heterogeneous BEV embeddings e_H and e_L from high- and low-resolution images, respectively. During this stage, we aggregate the two embeddings to generate a unified BEV representation. Specifically, e_H retains fine-grained details from high-resolution imagery, while e_L encapsulates global contextual semantics. This dichotomy results in e_H exhibiting heightened sensitivity to small, distant objects but reduced efficacy for large, proximal objects, whereas e_H demonstrates inverse characteristics.

Unlike direct fusion approaches that may introduce noise in their respective weak regions, we propose a simple yet effective method by reframing this as an embedding selection problem. Our strategy selectively leverages the advantageous regions of each embedding space while avoiding noise-prone areas. To achieve this, we design a learnable scoring function $f(\cdot)$ to dynamically evaluate the informativeness of spatial regions across both BEV embeddings. The scoring function is formally defined as follows:

$$s_k = f(e_k) = \mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{DWConv}(e_k))), \tag{9}$$

where s_k represents the informative score mask and $k \in \{H, L\}$. $\mathrm{DWConv}(\cdot)$ denotes a 3×3 depthwise convolution. We apply the sigmoid function to constrain the values to the range [0, 1]. Upon obtaining score maps, we perform cross-space comparison for the heterogeneous BEV embedding spaces. To yield the unified BEV representation e_{uni} for each grid, we select the feature from $\{e_H, e_L\}$ with the highest confidence score for the corresponding area, prioritizing the most discriminative regions across multi-granularity BEV features, which can be expressed as below:

$$e_{uni} = \varphi(s_k \cdot e_k + e_k), k \in \{H, L\} \tag{10}$$

where $\varphi(\cdot)$ denotes the operator of selecting the maximum value. This selective fusion mechanism ensures optimal integration of fine-grained details and contextual semantics. Finally, e_{uni} is then fed into the decoder.

Objective Function

We follow the segmentation head design of the previous methods (Liu et al. 2023b; Hu et al. 2021). We train our model using the segmentation loss, the offset loss, and the centerness loss (Hu et al. 2021; Liu et al. 2023b; Chambon et al. 2024). Polar predictions are mapped into rectangular predictions for loss calculation based on the rectangular ground truths.

Experiments

Dataset, Metrics, and Implementation Details

Dataset. We evaluate our proposed framework on the challenging nuScenes (Caesar et al. 2020) and Lyft L5 (Houston et al. 2021) datasets. NuScenes contains 1000 scenes split into 750-150-150 scenes for training, validation, and test. Lyft L5 contains 180 scenes, each 25-45 seconds in length, annotated at 5Hz, following the split in (Hu et al. 2021).

Evaluation Metrics. We utilize the Intersection-over-Union (IoU) score between the predicted results and the ground-truth BEV labels as the main performance measure. Following (Chambon et al. 2024), we use a 100m $\times$ 100m grid map with 50cm resolution, resulting in a 200 $\times$ 200 grid map.

Implementation Details. We implement our framework in PyTorch on a workstation equipped with two NVIDIA RTX 3090 GPUs. Following Simple-BEV (Harley et al. 2023), we adopt standard image augmentation techniques including random scaling, random cropping, and random flipping. The model is trained for up to 35/65 epochs for different resolution settings, using the AdamW optimizer with an initial learning rate of 4e-3 and a one-cycle learning rate scheduler. The batch size is set to 6. For the deep branch, we employ EfficientNet-B4 (Tan and Le 2019) as encoder, while the shallow branch employs a lightweight STDC network (Fan et al. 2021), where only the first four stages are utilized. Both branches are initialized with the weights pretrained on ImageNet (Deng et al. 2009). Following PolarBEV (Liu et al. 2023b), we choose the polar grid resolution of 400 × 100 and PV-to-BEV iteration of 2 as the default setting. Under the 224×448 resolution setting, the deep branch processes images at the original resolution, whereas the shallow branch receives the inputs at $2\times$ resolution. For the 448×800 resolution setup, the shallow branch operates on $1.5\times$ resolution inputs. The balancing weights λ_1, λ_2, and λ_3 are empirically set to 0.25, 0.25, and 0.5, respectively.

Comparison with State-of-the-arts

NuScenes Dataset. We benchmark our model against the following state-of-the-art BEV segmentation methods: FIERY (Hu et al. 2021), CVT (Zhou and Krähenbühl 2022), LaRa (Bartoccioni et al. 2023), BEVFormer (Li et al. 2022), PolarBEV (Liu et al. 2023b), BAEFormer (Pan et al. 2023), Simple-BEV (Harley et al. 2023), PointBeV (Chambon et al. 2024), GaussianLSS (Lu, Tsai, and Chen 2025). As depicted

Vehicle segm. IoU (↑)		No visibility filtering		Visibility filtering	
Method	Backbone	224×480	448×800	224×480	448×800
FIERY static	EN-b4	35.8	-	39.8	-
CVT	EN-b4	31.4	32.5	36.0	37.7
LaRa	EN-b4	35.4	-	38.9	-
BEVFormer	RN-50	35.8	39.0	42.0	45.5
PolarBEV	EN-B4	37.6	41.2	41.3	45.6
BAEFormer	EN-b4	36.0	37.8	38.9	41.0
Simple-BEV	RN-50	36.9	40.9	43.0	46.6
PointBeV	EN-b4	38.7	42.1	44.0	47.6
PointBeV	RN-50	38.1	41.7	43.7	47.0
GaussianLSS	EN-b4	38.3	40.6	42.8	46.1
Ours	EN-b4	**42.4**	**44.1**	**46.1**	**48.6**

Table 1: **BEV vehicle segmentation on nuScenes.** Evaluation on the validation set for different resolutions and filtering based on the vehicle's visibility. *No visibility filtering*: all the annotated vehicles are considered. *Visibility filtering*: only the vehicles having the visibility $> 40\%$ are considered. 'EN-b4' and 'RN-50' stand for EfficientNet-b4 and ResNet-50, respectively.

Pedestrian segm.	IoU (↑)
LSS (Philion and Fidler 2020)	15.0
FIERY (Hu et al. 2021)	17.2
ST-P3 (Hu et al. 2022)	14.5
TBP-former static (Fang et al. 2023)	17.2
PointBeV (Chambon et al. 2024)	18.5
GaussianLSS (Lu, Tsai, and Chen 2025)	18.0
Ours	**21.3**

Table 2: **BEV pedestrian segmentation on nuScenes.** Metrics are IoU (↑) with visibility filtering, evaluated on the validation set at 224×480 resolution.

in Table 1, our proposed model consistently outperforms existing approaches, achieving SOTA performance across varying resolutions and visibility filtering settings. Notably, under low-resolution scenarios, our method demonstrates significant improvements over GaussianLSS (+4.1% and +3.3% for different settings), attributing to the dual-resolution heterogeneous network architecture that better integrates fine-grained visual details with contextual semantics. Remarkably, our low-resolution model even matches or surpasses the high-resolution variant of GaussianLSS. We further expand our evaluations to pedestrian segmentation (Table 2). Without architectural modifications, our method also achieves superior results on this task, demonstrating its robustness in perceiving small objects. More qualitative results are detailed in the *Supplementary Material*.

Lyft L5 Dataset. To further validate scalability, we evaluate our model on the Lyft L5 dataset under both long- and short-range perception settings (Table 3). Our approach outperforms prior methods in both configurations, with particularly substantial gains in large-scale perception scenarios. This highlights our model's enhanced long-range perception capabilities—a critical advantage for autonomous driving applications. These experiments validate our framework's generalization across diverse perception ranges.

Vehicle segm. IoU (↑)	Long	Short
FIERY (Hu et al. 2021)	36.7	59.4
BEVFormer (EN-b4) (Li et al. 2022)	44.5	69.9
BEVFormer (RN-50) (Li et al. 2022)	43.2	68.8
Simple-BEV (EN-b4) (Harley et al. 2023)	44.5	70.4
Simple-BEV (RN-50) (Harley et al. 2023)	43.2	70.7
PointBeV (EN-b4) (Chambon et al. 2024)	45.4	72.6
PointBeV (RN-50) (Chambon et al. 2024)	44.5	72.3
Ours (EN-b4)	**48.3**	**73.3**

Table 3: **BEV vehicle segmentation on Lyft L5**. The comparison methods are evaluated in terms of IoU (↑), which are trained at 224×480 resolution and different ranges - 30m $\times$ 30m (Short) and 100m $\times$ 100m (Long).

Dual-Resolution	MCRA	MRADE	ABSS	IoU (↑)
				42.0
✓				43.9
		✓		43.1
✓	✓			44.9
✓		✓		44.5
✓	✓	✓		45.6
✓		✓	✓	45.3
✓	✓	✓	✓	**46.1**

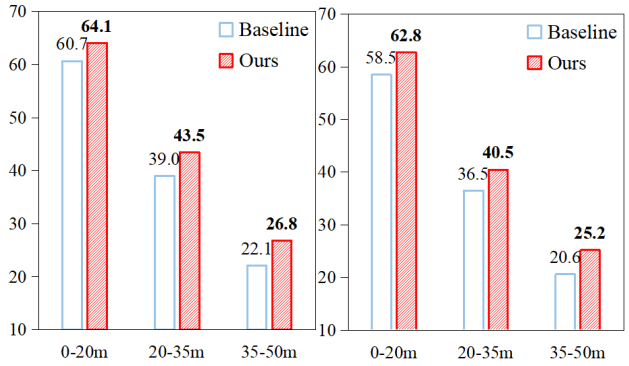
Table 4: Effectiveness of each component in our framework.

Ablation Study

We conduct a comprehensive ablation study, which is conducted on the nuScenes dataset under the visibility filtering configuration with a resolution of 224×448.

Dual-Resolution Architecture. As shown in Table 4, we evaluate the effect of different architectural components on model performance. The proposed dual-resolution heterogeneous network delivers significant performance gains, improving baseline from 42.0% to 43.9% in terms of IoU. Notably, even integrated with the baseline MRADE module (the 3rd & 5th rows), this dual-resolution heterogeneous design achieves an additional 1.4% improvement. To further verify its capability in segmenting distant and small objects, we conduct stratified analysis across three distance ranges: near (0-20m), medium (20-35m), and far (35-50m). As shown in Fig. 3, under the No Visibility Filtering configuration, our method surpasses the baseline by +3.4% in near-range regions, with more pronounced improvements of +4.5% and +4.7% for medium- and far-range regions, respectively. It demonstrates our superior capacity to enhance long-range perception, particularly critical for autonomous driving scenarios requiring early hazard detection.

MCRA. As shown in Table 4, ablating MCRA reduces the IoU metric from 46.1% to 45.3%, validating its critical role in aligning dual-resolution perspective features. To delve into our module design, we replace the MCRA with alternative alignment mechanisms (Table 5). The spatial-channel



(a) Visibility filtering (b) No visibility filtering

Figure 3: Comparison for varying distances under different visibility settings. (Please zoom for details.)

Feature Alignment Structure	IoU (↑)
Spatial-channel attention (Zhang et al. 2023)	45.0
Cross-attention (Vaswani et al. 2017)	45.5
Window Cross-attention (Liu et al. 2021)	45.3
MCRA (independent state space)	45.8
MCRA	**46.1**

Table 5: Comparison of different feature alignment structures.

attention mechanism yields suboptimal performance, as its local receptive field fails to establish global feature correspondences. While cross-attention and its variant achieve global alignment, they ignore the axial geometric constraints between perspective features and polar coordinates. Furthermore, deploying independent state spaces in MCRA is not suitable for feature alignment. In contrast, our mamba-driven MCRA leverages a shared state space to implicitly align heterogeneous features while preserving polar geometric fidelity, achieving optimal results.

MRADE. As shown in Table 4, integrating the MRADE module with the baseline model improves the IoU from 42.0% to 43.1%, demonstrating its effectiveness. A comprehensive analysis of different height estimation architectures (Table 6) reveals that fixed height bins lack the flexibility to adaptively predict object heights in 3D space. While the single height predictor dynamically estimates height distributions, it shows suboptimal performance due to limited robustness against polar coordinate regional perturbations. Through component ablation studies of MRADE, we observe performance degradation when removing the permanent expert responsible for extracting unified knowledge. Notably, after expert decoupling where dedicated experts independently predict height distributions within specific radial-angular regions, the model achieves optimal performance through polar coordinate-adaptive feature learning.

ABSS. Our ablation studies on the ABSS module reveal its positive impact on model performance. As shown in Table 4, integrating ABSS without MCRA elevates IoU from 44.5% to 45.3%. More visualization and results of the ablation experiments are detailed in the *Supplementary Material*.

Height Estimation Structure	IoU (↑)
Single height predictor (Liu et al. 2023b)	45.2
Fixed height bins (Harley et al. 2023)	44.7
MRADE (w/o $\mathbf{E}_{per}$)	45.5
MRADE (w/o decoupling experts)	45.6
MRADE	**46.1**

Table 6: Comparison of different height estimation structures.

Method	FPS	Params(M)	Mem.(G)	Train epoch	nuScenes(IoU)
PointBeV	15	10.4	1.21	100	38.7 / 44.0
Ours	14	10.5	1.58	35	42.4 / 46.1

Table 7: Overall model complexity and efficiency.

Model Efficiency

We benchmark the model efficiency of our model against the previous SOTA PointBeV (Chambon et al. 2024) on an NVIDIA RTX 3090 under a resolution of 224×448. As shown in Table 7, notably, despite incorporating dual network branches, our framework maintains competitive inference speeds, parameters, and memory usage, which are comparable to existing approaches. This efficiency stems from our proposed lightweight components and the use of only shallow layers in the lightweight network for high-resolution image processing, thus striking an effective balance between accuracy and computational efficiency. Moreover, our method has fewer training epochs and higher model performance. More details on the efficiency analysis of each component are provided in the *Supplementary Material*.

Conclusion

This paper presents a novel polar-based BEV segmentation framework mitigating inherent limitations in BEV representations, with key contributions: (1) a dual-resolution architecture generating complementary coarse/fine-grained BEVs via distinct feature pathways; (2) a Mamba-driven alignment module preserving resolution-specific features while correcting spatial misalignment via structured state-space modeling; (3) an attention-based adaptive selector dynamically focusing on salient regions across resolutions; and (4) radial-angular decoupled experts enabling region-aware height prediction through polar-coordinate routing. Superior performance on public benchmarks demonstrates the effectiveness of our model.

Limitations. While our framework improves polar coordinate BEV representations, the inherent distance-dependent density pattern (denser near ego vehicle, sparser at distance) might slightly affect distant object perception. Future work will explore adaptive BEV grids with dynamic resolution allocation to better balance near-far observational priorities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U21A20471, U21A20472, 62072110) and funds for the Innovation of Policing Science and Technology. Fujian Province (No. 2024Y0061).

References

Bartoccioni, F.; Zablocki, É.; Bursuc, A.; Pérez, P.; Cord, M.; and Alahari, K. 2023. Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation. In *Conference on Robot Learning*, 1663–1672. PMLR.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cai, J.; Li, Q.; Shen, Y.; Pan, J.; and Liu, W. 2024. Efficient Semantic Segmentation for Compressed Video. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4266–4272. IEEE.

Cai, J.; Su, J.; Li, Q.; Yang, W.; Wang, S.; Zhao, T.; He, S.; and Liu, W. 2025. Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+ X Semantic Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10587–10598.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23555–23564.

Chambon, L.; Zablocki, E.; Chen, M.; Bartoccioni, F.; Pérez, P.; and Cord, M. 2024. PointBeV: A Sparse Approach for BeV Predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15195–15204.

Chen, Q.; Vora, S.; and Beijbom, O. 2021. Polarstream: Streaming object detection and segmentation with polar pillars. *Advances in Neural Information Processing Systems*, 34: 26871–26883.

Chu, X.; Deng, J.; You, G.; Duan, Y.; Li, Y.; and Zhang, Y. 2024. RayFormer: Improving Query-Based Multi-Camera 3D Object Detection via Ray-Centric Strategies. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4620–4629.

Cong, W.; Liang, H.; Wang, P.; Fan, Z.; Chen, T.; Varma, M.; Wang, Y.; and Wang, Z. 2023. Enhancing nerf akin to enhancing llms: Generalizable nerf transformer with mixture-of-view-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3193–3204.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; and Wei, X. 2021. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9716–9725.

Fang, S.; Wang, Z.; Zhong, Y.; Ge, J.; and Chen, S. 2023. TBP-Former: Learning Temporal Bird's-Eye-View Pyramid for Joint Perception and Prediction in Vision-Centric Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1368–1378.

Gong, S.; Ye, X.; Tan, X.; Wang, J.; Ding, E.; Zhou, Y.; and Bai, X. 2022. Gitnet: Geometric prior-based transformation for birds-eye-view segmentation. In *European conference on computer vision*, 396–411. Springer.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Harley, A. W.; Fang, Z.; Li, J.; Ambrus, R.; and Fragkiadaki, K. 2023. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2759–2765. IEEE.

Hatamizadeh, A.; and Kautz, J. 2024. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*.

He, H.; Bai, Y.; Zhang, J.; He, Q.; Chen, H.; Gan, Z.; Wang, C.; Li, X.; Tian, G.; and Xie, L. 2025. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37: 71162–71187.

Houston, J.; Zuidhof, G.; Bergamini, L.; Ye, Y.; Chen, L.; Jain, A.; Omari, S.; Iglovikov, V.; and Ondruska, P. 2021. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, 409–418. PMLR.

Hu, A.; Murez, Z.; Mohan, N.; Dudas, S.; Hawke, J.; Badrinarayanan, V.; Cipolla, R.; and Kendall, A. 2021. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15273–15282.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jiang, Y.; Zhang, L.; Miao, Z.; Zhu, X.; Gao, J.; Hu, W.; and Jiang, Y.-G. 2023. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, 1042–1050.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard:

Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, Y.; Niu, Y.; Xu, H.; Da, H.; Xu, R.; and Liu, W. 2025. IPCMoE: Integrating Perceptual Cues with Mixture-of-Experts for Joint Low-Light Image Enhancement and Deblurring. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7644–7652.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liu, W.; Cai, J.; Li, Q.; Liao, C.; Cao, J.; He, S.; and Yu, Y. 2024a. Learning nighttime semantic segmentation the hard way. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7): 1–23.

Liu, W.; Li, Q.; Yang, W.; Cai, J.; Yu, Y.; Ma, Y.; He, S.; and Pan, J. 2024b. Monocular BEV Perception of Road Scenes Via Front-to-Top View Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024c. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023a. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Liu, Z.; Chen, S.; Guo, X.; Wang, X.; Cheng, T.; Zhu, H.; Zhang, Q.; Liu, W.; and Zhang, Y. 2023b. Vision-based uneven bev representation learning with polar rasterization and surface estimation. In *Conference on Robot Learning*, 437–446. PMLR.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023c. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, 2774–2781. IEEE.

Lu, S.-W.; Tsai, Y.-H.; and Chen, Y.-T. 2025. Toward Real-world BEV Perception: Depth Uncertainty Estimation via Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17124–17133.

Pan, C.; He, Y.; Peng, J.; Zhang, Q.; Sui, W.; and Zhang, Z. 2023. BAEFormer: Bi-Directional and Early Interaction Transformers for Bird's Eye View Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9590–9599.

Philion, J.; and Fidler, S. 2020. 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*, 194–210. Springer.

Saha, A.; Mendez, O.; Russell, C.; and Bowden, R. 2022. Translating images into maps. In *2022 International conference on robotics and automation (ICRA)*, 9200–9206. IEEE.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, H.; Ke, X.; Li, Y.; Xu, R.; Wu, H.; Lin, X.; and Guo, W. 2024. Vision-Language Action Knowledge Learning for Semantic-Aware Action Quality Assessment. In *European Conference on Computer Vision*, 423–440.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. 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*, 17830–17839.

Yang, Y.; Jiang, P.-T.; Hou, Q.; Zhang, H.; Chen, J.; and Li, B. 2024. Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27927–27937.

Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; and Stiefelhagen, R. 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694.

Zhao, T.; Chen, Y.; Wu, Y.; Liu, T.; Du, B.; Xiao, P.; Qiu, S.; Yang, H.; Li, G.; Yang, Y.; et al. 2024. Improving Bird's Eye View Semantic Segmentation by Task Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15512–15521.

Zhou, B.; and Krähenbühl, P. 2022. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13760–13769.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024a. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, 62429–62442.

Zhu, P.; Sun, Y.; Cao, B.; and Hu, Q. 2024b. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7099–7108.