

OneBEV++: Towards Unifying Bird's-Eye-View Semantic Mapping with Panoramas

Jiale Wei, Zhifeng Teng, Fei Teng, Junwei Zheng, Ruiping Liu, Yufan Chen, Jie Hu, Kailun Yang*, Jiaming Zhang*, and Rainer Stiefelhagen

Abstract—Bird's-Eye-View (BEV) perception, crucial for holistic scene understanding, faces limitations with narrow field-of-view sensors (indoors) and complex multi-camera setups (outdoors). To address these limitations and enable omni-range perception with a single image, we propose a new pipeline called panoramic-to-BEV semantic mapping. We extend existing household-related and driving-scene datasets to create four benchmarks for Panoramic BEV in both indoor and outdoor settings. As ground-breaking solutions, the 360BEV (indoor) and OneBEV (outdoor) methods have shown promise in their respective domains. Moreover, we propose a unified framework **OneBEV++**, generating BEV semantic maps from a single panoramic image across all environments. Extensive experiments on our four panoramic datasets demonstrate that OneBEV++ outperforms previous methods with consistent accuracy gains and improved efficiency. By simplifying complexity while enhancing performance, OneBEV++ enables scalable applications across diverse environments from outdoor autonomous vehicles to indoor robots, paving the way for more holistic cross-environment autonomous systems. Data and code will be available at <https://github.com/JialeWei/OneBEV>.

Index Terms—Omni-range Image Processing, View Transformation, Semantic Mapping, Scene Understanding.

1 INTRODUCTION

BIRD'S-EYE-VIEW (BEV) semantic mapping involves generating a top-down view where each pixel is labeled with its semantic category, such as roads, pedestrians, and cars for outdoor scenes [1], [2], [3], or walls, floors, and furniture for indoor environments [4], [5]. This mapping is crucial for higher-level real-world tasks like navigation [6], [7] and scene understanding [8], [9], [10]. For autonomous driving, BEV maps are essential for path planning, obstacle detection, and understanding the road layout, facilitating safe and efficient navigation [1], [8], [10]. In indoor robotics, they assist robots in navigating complex spaces, locating objects, and performing tasks such as cleaning or delivery, enhancing operational efficiency [6], [11], [12].

Traditional methods for BEV semantic mapping often rely on pinhole-based cameras [1], [2], [3], [13], which have significant limitations. These methods typically require multiple cameras positioned around the vehicle or robot to capture the entire surrounding environment, leading to increased hardware complexity, synchronization issues, and computational overhead. Additionally, pinhole cameras suffer from limited field of view, often necessitating complex multi-view fusion algorithms to generate comprehensive BEV representations. Furthermore, these camera arrangements are often platform-specific, making cross-platform

deployment challenging and requiring substantial reconfiguration for different vehicle types or environments.

360° panoramic cameras, in contrast, offer a promising alternative by capturing the entire surrounding environment in a single image. These omnidirectional cameras provide complete visual information with a single sensor, reducing hardware complexity and eliminating synchronization issues across multiple cameras. The continuous nature of panoramic imagery simplifies the mapping pipeline by providing a unified visual input, which inherently offers richer contextual information for semantic interpretation. Despite these advantages, leveraging panoramic images for BEV semantic mapping introduces its own challenges, such as handling geometric distortions and transforming front-view features into a top-down perspective, tasks that demand innovative architectural solutions.

Building upon the foundational concepts and demonstrated successes of our previous conference publications 360BEV [5] and OneBEV [14], we introduce a novel and unified framework termed OneBEV++. The major objective of OneBEV++ is to create a single, more advanced solution for BEV semantic mapping using panoramic images that is effective across both indoor and outdoor environments. While previous methods like 360BEV (focused on indoor settings) and OneBEV (designed for outdoor environments) have shown promise in their respective domains, they were optimized for specific applications and lacked the versatility required for cross-domain deployment. OneBEV++ bridges this gap by introducing a unified architecture that leverages the strengths of both approaches while addressing their limitations, requiring no domain-specific redesign to transfer across environments.

Our framework introduces two key technical innovations that enable its cross-domain effectiveness. First, we develop Columnar Radial Projection (CRP), a specialized panoramic projection technique that effectively maps BEV coordinates to panoramic image columns. CRP establishes geometric relationships between the BEV space and front-view panorama through azimuthal angle

- This work was supported in part by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, and in part by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT and HOREKA@KIT partition.
- J. Wei, J. Zheng, R. Liu, Y. Chen, and R. Stiefelhagen are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.
- Z. Teng is with the XYZ Robotics GmbH, 41468, Neuss, Germany.
- F. Teng, J. Hu, K. Yang, and J. Zhang are with the School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082, China.
- *Corresponding author (E-Mail: jiamingzhang@hnu.edu.cn).

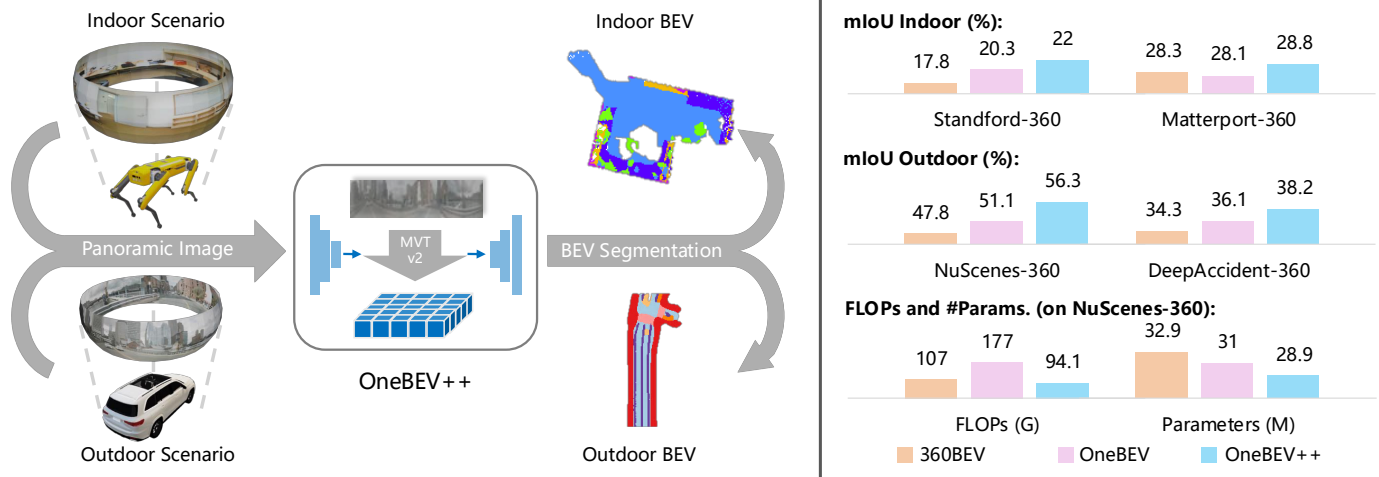


Fig. 1: Our OneBEV++ performs unified panoramic BEV semantic mapping across all scenes, outperforming specialized methods 360BEV (indoor) and OneBEV (outdoor). Our method achieves better performance (higher mIoU) with fewer FLOPs and parameters across four panoramic datasets: Stanford-360 and Matterport-360 (indoor), and NuScenes-360 and DeepAccident-360 (outdoor).

calculations, providing a robust foundation for feature transformation across diverse environmental scales and geometries. The approach inherently accommodates panoramic image characteristics while enabling adaptive feature sampling through learnable parameters. Second, we design a Feature Selection Block (FSB) that efficiently identifies and extracts the most salient features through strategic max-pooling operations before applying state space modeling. This mechanism enhances efficiency by focusing computational resources on critical visual cues, ensuring accurate BEV transformation with reduced overhead. These advancements are embedded within our MVTv2 (Mamba View Transformation version 2) module, which seamlessly converts panoramic features into the BEV domain.

Thanks to this advanced architecture design, our new framework obtains state-of-the-art performance on **Stanford-360** [5] and **Matterport-360** [5] for indoor scenes, **NuScenes-360** [14] and **DeepAccident-360** [14] for outdoor scenes, while reducing computational complexity (FLOPs) by 46% and parameter count by 7% compared to using separate specialized models. Our comprehensive experiments demonstrate that OneBEV++ not only outperforms previous specialized approaches in their respective domains but also maintains consistent performance across diverse environments, proving its robustness and versatility.

Compared to the previous conference publications, we deliver the additional contributions:

- (1) We propose **OneBEV++**, a novel unified framework for omnidirectional to Bird's-Eye-View semantic mapping that operates effectively across both indoor and outdoor environments, overcoming the domain-specific limitations of previous approaches.
- (2) We extend and standardize four benchmark datasets for panoramic BEV mapping, *i.e.*, **Stanford-360** and **Matterport-360** for indoor scenes, and **NuScenes-360** and **DeepAccident-360** for outdoor scenes, providing comprehensive evaluation environments with consistent annotation protocols and evaluation metrics.
- (3) We introduce a versatile architecture anchored by the optimized **MVTv2** module, achieving top-tier performance with reduced computational demands, thus setting a new standard

for efficient and effective BEV perception.

- (4) We provide extensive analyses covering MVT and CRP component ablation, field-of-view impact and stitching robustness under diverse corruption types, offering detailed insight into the design choices and practical applicability of OneBEV++.

2 RELATED WORK

2.1 Panoramic Image Processing

Omni-range image understanding has emerged as a critical area of research due to its ability to provide comprehensive scene representations beyond the limitations of traditional narrow-view imaging. Panoramic images, which capture 360° scenes, offer a foundation for holistic environmental perception [15]. However, the equirectangular projection commonly used to represent these omnidirectional scenes introduces significant geometric distortions that pose unique challenges for downstream computer vision tasks. Research addressing these challenges has evolved across several fronts. Early efforts focused on adapting conventional computer vision tasks to panoramic imagery, with object detection approaches [16], [17] tackling the warping and scale variation issues in equirectangular projections. This was followed by advancements in semantic segmentation techniques [15], [18], [19], [20], [21], [22] that account for the continuous nature of panoramic views. Depth estimation methods [23], [24], [25], [26] and panoptic segmentation approaches [27], [28], [29] further expanded the capabilities of omnidirectional scene understanding. Despite these advancements, most existing approaches primarily operate within the front-view perspective, leaving the challenge of accurate feature transformation to bird's-eye view relatively unexplored. Our work addresses this specific gap by developing a framework that not only handles the distortion in front-view panoramas but also accurately transforms these omnidirectional features into a bird's-eye perspective representation.

2.2 Panoramic Scene Understanding

Scene understanding encompasses a diverse set of computer vision tasks aimed at comprehensively interpreting visual environments.

These tasks have traditionally developed along separate trajectories for indoor and outdoor settings due to their fundamentally different characteristics and challenges. In indoor environments, the relatively controlled conditions and structured geometry have facilitated specialized approaches. Room layout estimation methods [30], [31] leverage geometric constraints, while indoor semantic segmentation techniques [4], [32], [33] capitalize on well-defined object relationships. Indoor 3D reconstruction approaches [34], [35] further exploit the bounded nature of interior spaces to create detailed spatial representations. Outdoor scene understanding, particularly in the context of autonomous driving, presents contrasting challenges due to unbounded spaces, dynamic conditions, and greater environmental variability. Foundational datasets like KITTI [36] and Cityscapes [37] have spurred development in this domain. More recent approaches have evolved toward multi-task learning frameworks [1], [38] that integrate multiple perception objectives. Specialized methods addressing adverse conditions such as inclement weather [39], [40] and low-light scenarios [41], [42] have further expanded the robustness of outdoor perception systems. The dichotomy between indoor and outdoor approaches highlights a significant research gap: the lack of unified frameworks that can seamlessly operate across diverse environmental contexts. Recent architectures like [43], [44] have begun exploring more generalizable approaches, showing promising potential for cross-environment applications. Our work advances this direction by developing a unified omnidirectional scene understanding framework that effectively bridges indoor and outdoor domains while maintaining a shared representation foundation.

2.3 BEV Semantic Mapping

Bird's-eye view (BEV) semantic mapping has evolved through two primary paradigms for view transformation: *2D-3D* and *3D-2D* approaches [45], each offering distinct advantages for converting perspective images to overhead representations. The *3D-2D* paradigm originated with Inverse Perspective Mapping (IPM) [46], which utilized homography transformations to convert camera images to BEV. This approach has evolved significantly, with contemporary methods [1], [2], [47], [48], [49] employing multi-layer perceptrons (MLPs) or Transformer architectures to perform 3D-to-2D projection with camera parameters. [13], [50] advanced this field by implementing view projection without relying on explicit camera parameters. The *2D-3D* approach, introduced by LSS [51], established a novel framework for connecting BEV segmentation with 3D feature maps by estimating implicit pixel depth from 2D features. Subsequent works including BEVDet [3], M²BEV [52], and BEVFusion [38] built upon this foundation to enhance transformation accuracy through improved depth estimation techniques. While these approaches have shown promising results for multi-view camera setups, the transformation from panoramic imagery to BEV representations presents unique challenges. 360BEV [5] pioneered this direction for indoor environments using Transformer architecture and supervised depth information, followed by OneBEV [14], which introduced a distortion-aware Mamba [53]-based framework for outdoor panoramas without requiring ground-truth depth. Our work builds upon and extends this evolutionary trajectory by presenting a unified model that seamlessly handles both indoor and outdoor panoramic-to-BEV transformations. By addressing the distinct challenges of each domain while maintaining a common

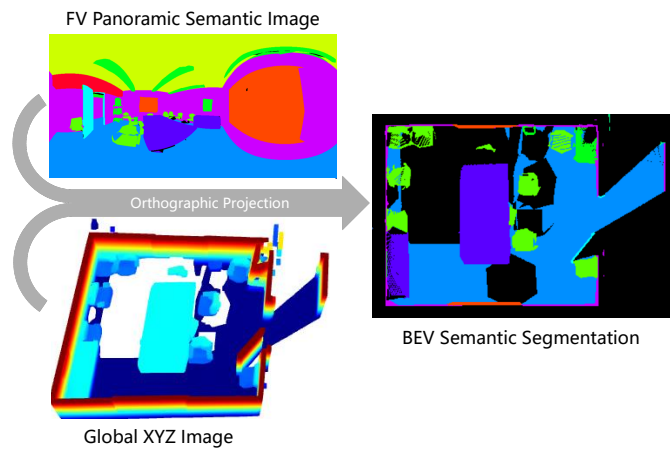


Fig. 2: **BEV semantic map generation** via orthographic projection, derived from the front-view semantic image and the global XYZ image into the BEV semantic map.

architectural framework, our approach represents a significant advancement toward versatile omnidirectional scene understanding.

3 PANORAMA-TO-BEV BENCHMARK

Building upon our previous works [5], [14], this section details the image processing pipeline for four omni-range datasets, covering both indoor and outdoor environments:

- (A) Indoor: **Stanford-360** in Sec. 3.1.
- (B) Indoor: **Matterport-360** in Sec. 3.2.
- (C) Outdoor: **NuScenes-360** in Sec. 3.3.
- (D) Outdoor: **DeepAccident-360** in Sec. 3.4.

Each dataset undergoes a structured processing pipeline that involves panoramic image construction and BEV annotation generation, ensuring consistency across different environments.

3.1 Stanford-360 Dataset Construction

Panorama Generation. The Stanford-360 dataset is built upon the existing Stanford2D3D dataset [54], which provides front-view panoramic images captured in diverse indoor environments. These images are directly incorporated into the new dataset.

BEV Annotation Generation. While the Stanford2D3D dataset includes front-view panoramic images, along with front-view semantic segmentation and global XYZ images (as depicted on the left side of Fig. 2), it lacks corresponding BEV semantic labels. To address this, we generate BEV semantic maps by orthogonally projecting the provided semantic segmentation and global XYZ information. The resulting BEV semantic map, representing the ground truth, is generated within the visible range (as shown on the right side of Fig. 2, where invisible areas are black and visible areas are colored).

3.2 Matterport-360 Dataset Construction

Panorama Generation. The Matterport-360 dataset is built upon Matterport3D [55], which originally consists of narrow field-of-view (FoV) pinhole camera images. To generate panoramic front-view images, we synthesize 360° panoramas by stitching 18 narrow-view images using rotation-translation matrices, as illustrated in Fig. 3. The corresponding semantic labels are similarly transformed to ensure spatial alignment with the panoramic projection.



Fig. 3: **Panoramic front-view generation** in the Matterport3D dataset. For each center point P_{ij} of the narrow-view images, rotation R and translation t relative to a reference center O are used to construct the panoramic front-view image.

BEV Annotation Generation. Applying the supplied front-view depth map, we build a global XYZ for each panoramic image, inspired by the global XYZ from the Stanford2D3D dataset. The construction proceeds in three steps: (1) computing spherical coordinates for each panoramic pixel, (2) converting them to 3D points in the camera coordinate system using the depth map, and (3) projecting these points onto the BEV plane via an orthographic projection matrix. The full derivation is given in Appendix ??.

3.3 NuScenes-360 Dataset Construction

Panorama Generation. The NuScenes-360 dataset is generated from the original NuScenes dataset [56], which provides images from six narrow FoV cameras mounted on a vehicle. These cameras are positioned to ensure full 360° coverage around the vehicle, with five cameras having a 70° FoV and one rear camera featuring a wider 110° FoV, mounted at 55° intervals with overlapping views. Since the cameras have non-coincident optical centers, we stitch the images by projecting each panoramic pixel onto an observation sphere and mapping it back to the corresponding camera image plane via ray-plane intersection. The full geometric derivation is given in Appendix ??.

Due to the restricted overlapping areas among multiple cameras in datasets, cases may arise where the same object doesn't exist in neighboring images. Consequently, employing feature point-based approaches [57], [58], [59] for matching may be inefficient or perhaps ineffective. Our stitching technique, grounded on the camera's physical model, is rapid but fails to attain superior stitching quality. Furthermore, objects within the panoramic image might show distortion or warping, posing challenges to the neural network's robustness. Fig. 4 depicts several of these challenges.

BEV Annotation Generation. The original NuScenes dataset contains high-resolution BEV semantic segmentation maps. Following [38], we obtain the semantic annotation of a given location.

3.4 DeepAccident-360 Dataset Construction

Panorama Generation. We additionally present a synthetic dataset, DeepAccident [60], generated using CARLA [61] simulator. DeepAccident retains an identical structure with NuScenes, hence, we employ the same methodology to create front-view panoramas in DeepAccident-360 dataset.

BEV Annotation Generation. The original DeepAccident dataset also comprises BEV semantic segmentation map, which were transformed from its front-view semantic segmentation. We use directly for our new DeepAccident-360 dataset.

3.5 Dataset Statistics

After front-view panorama combination and BEV annotation generation, two indoor and two outdoor BEV datasets for panoramic semantic mapping are obtained. We provide comprehensive statistical data for these datasets and more details on the analysis and selection of specific BEV semantic classes in the Appendix ??.

4 THE UNIFIED ONEBEV++ FRAMEWORK

This section presents the OneBEV++ framework, a unified approach for Bird's-Eye-View (BEV) semantic mapping across both indoor and outdoor environments using a single panoramic image. First, we define the omni-directional to BEV semantic mapping task (Sec. 4.1). Next, we introduce the overall architecture and its key components (Sec. 4.2). Finally, we detail the view transformation module, focusing on the improved MVTv2 design (Sec. 4.5).

4.1 Task Definition

Given a single panoramic RGB image $I \in \mathbb{R}^{H \times W \times 3}$, OneBEV++ aims to generate a top-down semantic map $M \in \mathbb{R}^{H_q \times W_q \times C}$, where C represents the number of semantic classes. Unlike conventional BEV methods that rely on multiple pinhole cameras or LiDAR, OneBEV++ leverages a single panoramic image to infer BEV semantics, making it adaptable to diverse environments. This is achieved through a distortion-aware view transformation that maps features from the front-view panoramic domain to a structured BEV space. By extending BEV perception beyond domain-specific solutions, OneBEV++ provides a unified approach applicable to robotics, autonomous driving, and scene understanding.

4.2 Architecture Overview

OneBEV++ follows an end-to-end architecture similar to OneBEV [14], designed for BEV semantic segmentation using a single panoramic image. As illustrated in Fig. 5, the model consists of three main components: a feature extraction backbone, a view transformation module, and a BEV segmentation head. The pipeline starts by processing a front-view panoramic image through the VMamba-T [62] backbone to extract hierarchical feature representations. These extracted features are then passed to the MVTv2 module, an optimized variant of the original Mamba View Transformation (MVT), designed to enhance computational efficiency while maintaining accurate BEV feature transformation. MVTv2 selects key feature points from the front-view domain and maps them to the BEV space, ensuring robust spatial alignment across indoor and outdoor environments. Finally, a lightweight BEV segmentation head processes the transformed features to generate the top-down semantic map. By integrating MVTv2 while preserving the computational efficiency of the overall pipeline, OneBEV++ provides a scalable and lightweight framework for diverse BEV perception tasks.

4.3 BEV Queries

The foundation of BEV semantic mapping lies in a set of BEV queries $Q \in \mathbb{R}^{H_q \times W_q \times C_{Emb}}$ paired with their corresponding positional embeddings $Z \in \mathbb{R}^{H_q \times W_q \times C_{Emb}}$. These embeddings encode spatial priors, enabling the model to maintain geometric consistency throughout the transformation process. Both Q and Z operate within a structured spatial framework defined by the BEV plane dimensions H_q and W_q with an embedding size of C_{Emb} .

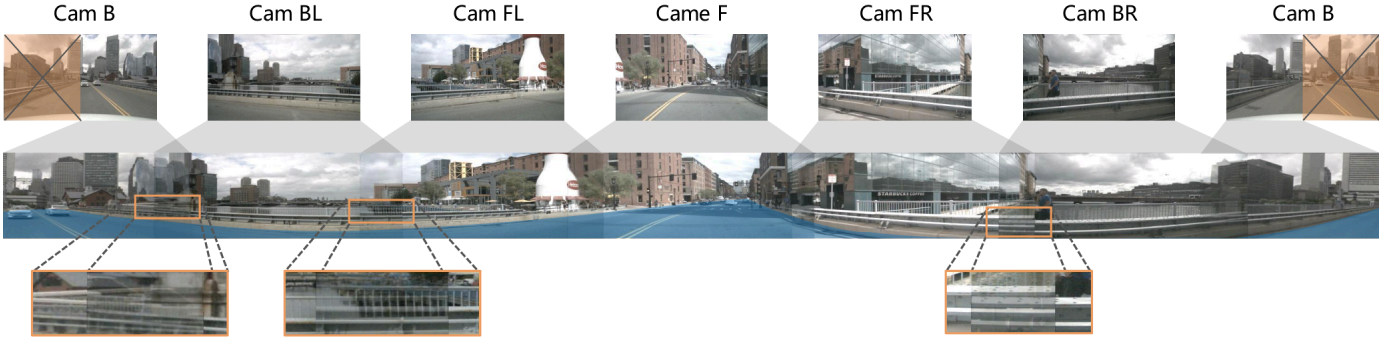


Fig. 4: **Reconstructed panorama and associated challenges.** The top row shows images from six cameras surrounding a car. The second row displays the multi-view images mapped into panorama. As seen in the panorama, some challenges exist. A straight road appears curved and incomplete, with scattered segments (blue shade). Due to stitching quality issues, seams and ghosting are present at the junctions of adjacent views (orange boxes), leading to object misalignment.

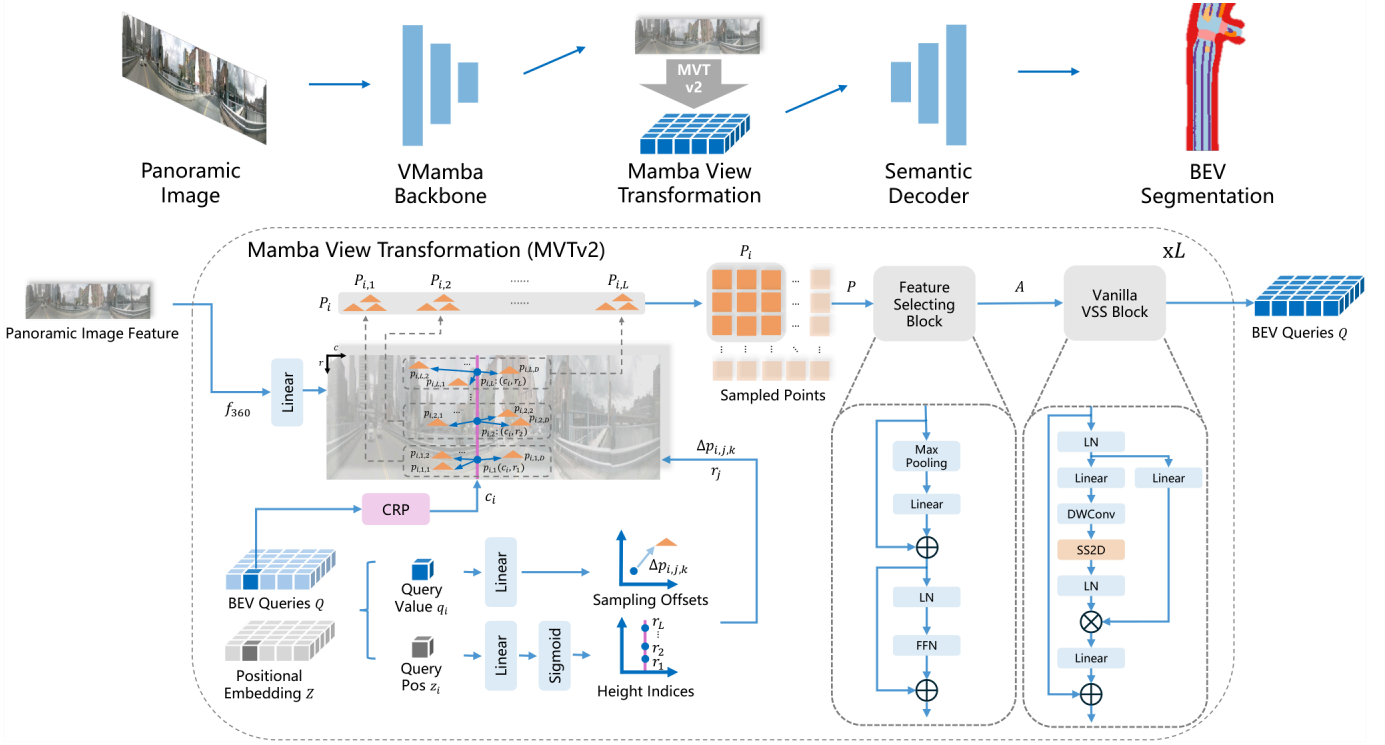


Fig. 5: **Architecture of OneBEV++.** The VMamba-T [62] backbone extracts features from the panorama. The optimized Mamba-based view transformation (MVTv2) module maps these features into BEV space through iterative refinement, followed by a lightweight decoder for semantic segmentation.

Each query q_i represents a discrete location in BEV space and is responsible for aggregating relevant features from the panoramic input.

4.4 Columnar Radial Projection

Building upon the Inverse Radial Projection (IRP) proposed in 360BEV [5], we introduce Columnar Radial Projection (CRP), a specialized panoramic projection technique tailored for scenarios lacking depth information. In depth-supervised settings, IRP enables precise point-to-point mapping between BEV and front-view images. However, in depth-free settings, direct correspondence is not always feasible. Instead, CRP establishes a point-to-column

mapping, ensuring that for any given BEV query q_i , a corresponding relative column position c_i in the panoramic view can be identified through geometric transformations. As illustrated in Fig. 6, for a BEV query point $q_i = (x_i, y_i)$ in the ground plane coordinate system, CRP computes its corresponding column index c_i in the panoramic image through the following transformations:

$$\begin{aligned} \phi_i &= \tan^{-1} \left(\frac{-x_i}{y_i} \right), \\ \phi_i &= \phi_i + 2\pi, \text{ if } \phi_i < 0, \\ c_i &= \frac{\phi_i}{2\pi}, \end{aligned} \quad (1)$$

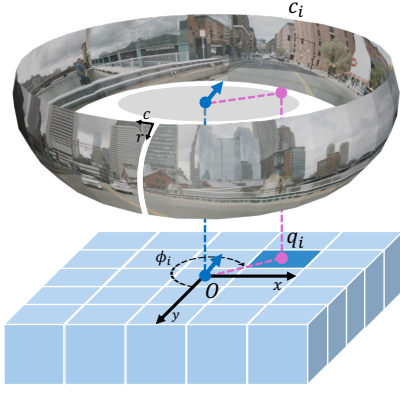


Fig. 6: **Columnar Radial Projection (CRP)** translates BEV features q_i into columnar representations c_i in a panoramic front view. **Blue arrows** indicate the ego-vehicle's driving direction, while the **magenta region** highlights radial-to-columnar mapping where points along a ray (varying radial distances) collapse to a single column.

where ϕ_i represents the azimuth angle between q_i and the y -axis. The magenta region illustrates how points at different radial distances map to the same column in the panoramic view, highlighting the column-wise nature of CRP. While this method introduces minor inaccuracies due to image stitching artifacts and geometric distortions, these limitations are mitigated by incorporating learnable parameters in Sec. 4.5. These parameters allow each BEV query q_i to adaptively refine its attention based on the CRP-predicted column, enabling more robust feature sampling.

4.5 View Transformation

As illustrated in the lower part of Fig. 5, the MVTv2 module efficiently transforms front-view panoramic features into BEV space through three key operations:

1. **Query-based Sampling:** MVTv2 processes each BEV query location by generating learnable sampling offsets $\Delta p \in \mathbb{R}^{H_q \times W_q \times D \times L \times 2}$ through linear projection of BEV queries Q . Here, L represents the number of sampling locations and D denotes the sampling points per location. The reference sampling points $p := (r, c)$ combine learnable height indices $r \in \mathbb{R}^{H_q \times W_q \times L \times 1}$ with CRP-derived column indices c . For panoramic features $f_{360} \in \mathbb{R}^{H_{360} \times W_{360} \times C_{Emb}}$, the sampling point locations are computed as:

$$P_i = W f_{360}(p_{i,j} + \Delta p_{i,j,k}), \quad (2)$$

where $W \in \mathbb{R}^{C_{Emb} \times C_{Emb}}$ denotes learnable projection matrices.

2. **Feature Selection and State Space Modeling:** The sampled features $P \in \mathbb{R}^{H_q \times W_q \times L \times D \times C_{Emb}}$ undergo two processing stages: First, a Feature Selection Block (FSB) employs *max-pooling* to identify and extract the most salient features. Subsequently, a Vanilla VSS Block (VVB) from VMamba [62] performs efficient spatial modeling in the BEV domain through state space computations.
3. **Iterative Enhancement:** The transformed features Q serve a dual purpose as both the BEV representation and query inputs for subsequent MVTv2 layers, enabling progressive refinement through additional front-view sampling iterations.

TABLE 1. **Comparison with the state-of-the-arts** on the outdoor datasets (NU-360: NuScenes-360, DA-360: DeepAccident-360).

Method	#Param.	FLOPs	NU-360 val	DA-360 val
BEVFusion [38]	53.7M	215G	42.2	23.8
BEVSegformer [13]	53.1M	112G	44.8	30.8
BEVFormer [1]	38.8M	213G	45.3	26.1
360BEV (ours)	32.9M	107.0G	47.8	34.3
OneBEV (ours)	31.0M	177.0G	51.1	36.1
OneBEV++ (ours)	28.9M	94.1G	56.3	38.2
w.r.t. 360BEV	(-4.0M)	(-12.9G)	(+8.5)	(+3.9)

This enhanced architecture preserves the transformation fidelity of the original MVT while substantially reducing computational complexity. The resulting BEV features provide a rich representation that effectively captures both spatial relationships and semantic content from the panoramic input, establishing a robust foundation for downstream segmentation tasks.

5 EXPERIMENT RESULTS

5.1 Implementation Details

Our implementation largely follows the settings of OneBEV [14], with modifications to accommodate the improvements in OneBEV++. The embedding dimension is set to 128. For the MVTv2 module, we configure it with 2 layers, 2 sampling locations, and 2 offset points per location, while keeping the inner dimension and depth of the vanilla VSS block at 128 and 1, respectively. OneBEV++ is trained using 4 A100 GPUs with a batch size of 2 per GPU. The initial learning rate is set to 5×10^{-5} across all datasets, following a training schedule of 50 epochs with linear warm-up and cosine annealing. We use the AdamW optimizer with a weight decay of 0.01 to enhance generalization. Input resolution differs between indoor and outdoor scenarios: indoor panoramic images are processed at 512×1024 , while outdoor images use 180×3200 . The BEV semantic map maintains a consistent output resolution of 200×200 pixels, with an adaptive perception range of $[-50m, 50m] \times [-50m, 50m]$ for outdoor environments and $[-5m, 5m] \times [-5m, 5m]$ for indoor spaces. For evaluation, we follow the OneBEV [14] protocol, using class-averaged mean Intersection-over-Union (mIoU) as our primary metric. The evaluation spans multiple datasets with varying numbers of semantic classes: Stanford-360 (13 classes), Matterport-360 (20 classes), NuScenes-360 (6 classes), and DeepAccident-360 (17 classes).

5.2 Outdoor OneBEV++ Results

In our previous work, OneBEV [14], we evaluated various state-of-the-art models on outdoor datasets, specifically the real-world NuScenes-360 and the synthetic DeepAccident-360, as summarized in Tab. 1. These datasets represent the original intended application of OneBEV, covering complex outdoor driving scenarios with diverse weather conditions, varying illumination, and dynamic traffic environments.

Thanks to the Mamba View Transformation (MVT) module, the original OneBEV (31.0M parameters, 177.0G FLOPs) exhibited excellent performance on both outdoor benchmarks, achieving an mIoU of 51.1% on NuScenes-360 and 36.1% on DeepAccident-360. These results outperformed our earlier baseline model, 360BEV [5], by substantial margins of 3.3% and 1.8%,

TABLE 2. **Comparison with the state-of-the-arts** on the indoor datasets (**S-360**: Stanford-360, **MP-360**: Matterport-360).

Method	#Param.	FLOPs	S-360	MP-360	
			val	val	test
BEVFusion [38]	53.7M	208.0G	7.8	21.0	18.3
BEVSegformer [13]	53.1M	100.0G	16.1	21.6	18.4
BEVFormer [1]	38.8M	206G	17.5	20.1	17.4
360BEV (ours)	32.9M	99.1G	17.8	28.3	24.0
OneBEV (ours)	31.0M	171G	20.3	28.1	23.6
OneBEV++ (ours)	28.9M	88G	22.0	28.8	24.3
<i>w.r.t. 360BEV</i>	(-4.0M)	(-11.1G)	(+4.2)	(+0.5)	(+0.3)

respectively, confirming the effectiveness of the OneBEV architecture in outdoor BEV semantic segmentation tasks. Notably, on the NuScenes-360 dataset, OneBEV surpassed all existing leading methods, such as BEVFusion [38] (42.2%), BEVSegformer [13] (44.8%), and BEVFormer [1] (45.3%), while also achieving highly competitive results on the more challenging DeepAccident-360 benchmark with a larger semantic category set.

Encouraged by these promising outcomes, we further refined the architecture by introducing OneBEV++, an optimized variant featuring an enhanced view transformation module (MVTv2). Remarkably, despite significantly reduced model complexity (28.9M parameters, 94.1G FLOPs), OneBEV++ achieved notable improvements on both outdoor datasets. Specifically, it reached a new best mIoU of 56.3% on NuScenes-360, surpassing the original OneBEV by 5.2% and the 360BEV baseline by 8.5%. Likewise, on the DeepAccident-360 dataset, OneBEV++ attained an mIoU of 38.2%, outperforming OneBEV by 2.1% and 360BEV by 3.9%.

These results demonstrate that OneBEV++ successfully fulfills its design objective: delivering lightweight efficiency without compromising accuracy for outdoor BEV semantic segmentation tasks. Its consistent and substantial performance improvements, coupled with reduced computational cost, not only validate our architectural innovations but also position OneBEV++ as a new leading method for efficient outdoor BEV semantic segmentation.

5.3 Indoor OneBEV++ Results

In this work, we further expand the scope of our architecture by exploring its performance on indoor tasks. Tab. 2 compares our OneBEV [14] model family with existing methods on the indoor Stanford-360 and Matterport-360 datasets. Previously, OneBEV had only been validated on large-scale outdoor datasets. Here, we adapt it to indoor environments for the first time. The original indoor-adapted OneBEV (31.0M parameters, 171G FLOPs) achieves an mIoU of 20.3% on Stanford-360, exceeding the 360BEV baseline by 2.5%. However, its performance slightly lags behind on the validation and test sets of the more diverse and challenging Matterport-360 dataset, indicating opportunities for further enhancement.

Addressing this, we introduce OneBEV++ for indoor contexts, further reducing model size while boosting segmentation accuracy. Specifically, OneBEV++ reaches an improved mIoU of 22.0% on Stanford-360, surpassing the baseline by 3.1%, and demonstrates moderate gains on Matterport-360 with validation and test mIoU scores of 28.8% and 24.3%, respectively, representing improvements of 0.5% and 0.3% compared to the baseline.

These findings reveal that OneBEV++ effectively transfers its hallmark advantages of lightweight efficiency and strong accuracy from outdoor to indoor scenarios. Through targeted enhancements

to the MVT architecture, OneBEV++ improves geometric understanding and feature representation, achieving robust performance gains in both relatively simpler indoor environments (Stanford-360) and more varied, complex scenarios (Matterport-360).

5.4 Analysis

TABLE 3. **Analysis of OneBEV++ components** on Standford-360 dataset. ‘VT Type’ denotes the method of view transformation. ‘VVB’ denotes whether to use Vanilla VSS Block. ‘Pooling Type’ denotes the pooling type (*max*: max-pooling, *avg*: average-pooling) in Feature Selection Block.

Backbone	VT Type	VVB	Pooling Type	mIoU
VMamba-T [62]	CrossAtt [13]	-	-	15.2
VMamba-T [62]	MVT	✓	-	18.9 (+3.7)
VMamba-T [62]	MVTv2	✓	<i>max</i>	22.0 (+6.8)
VMamba-T [62]	MVTv2	-	<i>max</i>	21.0
VMamba-T [62]	MVTv2	✓	<i>max</i>	22.0 (+1.0)
VMamba-T [62]	MVTv2	✓	<i>avg</i>	19.7
VMamba-T [62]	MVTv2	✓	<i>max</i>	22.0 (+2.3)
VMamba-T [62]	MVTv2	✓	<i>max</i>	22.0
GroupMamba-T [63]	MVTv2	✓	<i>max</i>	22.2 (+0.2)

Analysis of MVT. To better understand our proposed vision transformation method (MVT), we further analyze its components and hyperparameter choices, as shown in Tab. 3. Through rows (1-3), we can observe that when using VMamba-T [62] as the backbone, our proposed MVT method [14] achieves a 3.7% performance improvement compared to multi-camera deformable cross-attention methods [13] when ablating the vision transformation module. This demonstrates the advantages of using Mamba-based vision transformation approaches. With the introduction of the next-generation MVTv2, performance is further enhanced by a relative increase of 6.8%.

Looking specifically at the components within MVTv2, from rows (4-5) we ablate the Vanilla VSS Block (VVB), which is the core Mamba-based method that uses scanning mechanisms to replace cross-attention mechanisms for feature fusion. When we exclude this module from MVTv2, our model’s performance degrades, indicating that VVB is indispensable within the entire MVTv2 framework and plays a crucial role in fine-grained feature fusion.

Examining rows 6-7 in the table, we employ different pooling methods - *average-pooling* and *max-pooling* respectively. We find that using *max-pooling* makes the model more robust, as it forces the model to select only the most salient features. As illustrated in Fig. 7, we also visualize the differences between MVT and MVTv2 in terms of sampling points on the front view for the point on the BEV plane. The red point ● on the BEV semantically correspond to the label ‘floor’. In the bottom-left of Fig. 7, we can see that MVTv2 more accurately identifies the corresponding regions in the front view. This improvement benefits from our proposed CRP method, which reduces and concentrates the corresponding point ranges near column vectors. We highlight in blue ● the most salient feature selected through *max-pooling*, and we can observe that this feature point precisely falls within the ground region. However, when we observe in the upper left of Fig. 7, MVT encounters more difficulty in finding relevant features.

The experimental results demonstrate that each component of our proposed MVTv2 contributes meaningfully to the overall performance, with the VVB providing essential fine-grained feature

fusion capabilities and *max-pooling* enhancing feature selection robustness through saliency-based filtering. Finally, examining rows 8-9, we compare different backbone architectures, where GroupMamba-T [63] shows a marginal improvement of 0.2% mIoU over VMamba-T, indicating that while the backbone choice has some impact, the core MVTv2 architecture provides the primary performance gains.

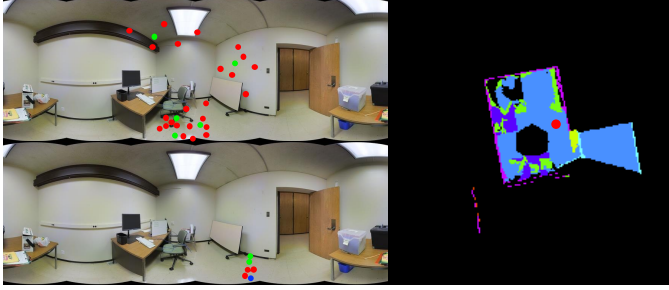


Fig. 7: **Deformable offsets visualization in MVT (OneBEV) and MVTv2 (OneBEV++).** Left column: Deformable offsets for MVT (top) and the proposed MVTv2 (bottom). The green points indicate sampling locations, red points represent deformable offsets, and blue point denotes salient point identified via *max-pooling* and selected as final features. Right column: Ground-truth BEV semantic map, with red points marking semantic locations to be computed.

Analysis of CRP. To quantify the contribution of each CRP component, we ablate four variants on Stanford-360 by selectively disabling parts of the ColumnSampling module: (V1) full CRP with column prior, vertical index, and learned offset; (V2) offset removed; (V3) column prior only; (V4) fully learned sampling without geometric prior (i.e., OneBEV’s original sampling). We additionally report an Inverse Radial Projection (IRP) oracle from 360BEV [5] that uses ground-truth depth, serving as an upper bound. Results are listed in Tab. 4.

TABLE 4. **CRP component ablation on Stanford-360 (mIoU %).** Column prior: geometric azimuth-to-column mapping. Vertical index: learned height sampling. Sampling offset: learned local refinement.

Variant	mIoU	Δ
V1: Full CRP (col. prior + vert. index + offset)	22.02	–
V2: No sampling offset	13.73	–8.29
V3: Column prior only	13.59	–8.43
V4: Fully learned, no CRP (OneBEV Sampling)	20.33	–1.69
IRP with oracle depth	45.80	+23.78

The learned sampling offset is the most impactful component: removing it (V2 vs. V1) drops mIoU by 8.29 points. This validates the coarse-to-fine design where CRP locates the correct panorama column for each BEV query while the offset refines the sampling position within that column.

The vertical index alone has marginal effect: V2 (column prior + vertical index, 13.73) and V3 (column prior only, 13.59) differ by only 0.14 mIoU. Its value emerges only when paired with the offset, indicating that height-aware sampling primarily provides an initialization seed for the learned refinement.

The geometric prior in CRP consistently outperforms purely learned sampling: V1 (22.02) exceeds V4 (20.33, OneBEV’s origi-

nal sampling) by +1.69 mIoU. Although the offset accounts for the largest accuracy gain, the column prior steers optimization toward geometrically meaningful features and leads to better convergence than unconstrained sampling.

The IRP oracle with ground-truth depth reaches 45.80 mIoU, 23.78 points above CRP. This gap reflects the inherent cost of depth-free projection. In practice, however, ground-truth depth is seldom available for panoramic images. CRP trades this accuracy ceiling for broad applicability: it operates on any panoramic image without depth sensors. When depth information is available, the framework can switch to IRP by replacing the column reference with a full 2D reference point, leaving the rest of the pipeline unchanged.

Analysis of FoV. To measure how much BEV performance benefits from wider input coverage, we retrain BEVSegFormer [13], 360BEV [5], and OneBEV++ at 60°, 180°, and 360° on NuScenes-360. At each FoV, the panoramic image is center-cropped to the target angular range around the forward direction, and the BEV ground truth is masked to the corresponding visible wedge (Fig. 8). Because the BEV evaluation region changes with FoV, raw mIoU values are not directly comparable across columns.

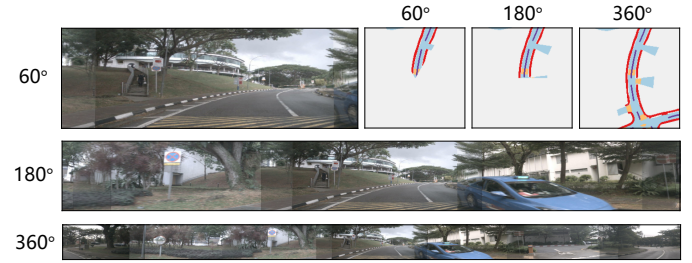


Fig. 8: **FoV ablation visualization on NuScenes-360.** The panoramic input is restricted to a forward-facing sector and the BEV supervision is masked accordingly.

TABLE 5. **FoV ablation on NuScenes-360 (mIoU %).** Each model is retrained at the target FoV. Δ : gain from 60° to 360°.

Method	60°	180°	360°	$\Delta_{60 \rightarrow 360}$
BEVSegFormer	20.09	19.77	44.80	+24.71
360BEV	27.57	30.08	47.80	+20.23
OneBEV++	32.41	34.50	56.30	+23.89

As shown in Tab. 5, 360° input consistently yields the highest mIoU across all three methods. OneBEV++ gains +23.89 mIoU when expanding from 60° to 360°, confirming that panoramic context carries information that a narrow forward view cannot recover. OneBEV++ also leads at every FoV level (32.41, 34.50, 56.30), indicating that the CRP+MVTv2 architecture generalizes across different input coverages. The 60° setting approximates a single front-view camera; the large gap to 360° underscores the practical value of panoramic input for BEV mapping.

Analysis of Stitching Robustness. Since our outdoor panoramas are assembled from multi-camera images, stitching artifacts are inevitable. We therefore design a robustness benchmark on NuScenes-360 in which OneBEV++ and OneBEV are trained on clean data and evaluated under four corruption types at four severity levels (S1–S4), applied at inference time only. Results are listed in Tab. 6.

TABLE 6. **Stitching robustness on NuScenes-360 (mIoU %).** Models trained on clean data; perturbations applied at test time.

Corruption	OneBEV++ (clean: 56.30)				OneBEV (clean: 51.11)			
	S1	S2	S3	S4	S1	S2	S3	S4
Seam Shift	56.37	56.29	54.82	53.20	50.90	50.84	50.40	49.19
Yaw Bias	56.39	55.66	51.77	47.18	50.66	50.66	48.66	44.87
Exposure Mis.	56.26	56.22	56.17	56.04	51.08	51.04	50.99	50.91
Missing View	43.66	38.43	29.46	15.42	41.08	34.76	26.08	13.63

The corruptions split into two severity profiles. *Appearance-level perturbations*—Seam Shift (3–15 px horizontal misalignment at camera boundaries) and Exposure Mismatch ($\pm 20\%$ – $\pm 80\%$ per-segment brightness/contrast variation)—have negligible impact: at S4, OneBEV++ degrades by only 5.5% and 0.5%, respectively. This resilience likely stems from the convolutional backbone and MVTv2 operating on local feature patterns that tolerate small spatial shifts and moderate photometric changes.

Geometry-level perturbations are more severe. Yaw Bias ($\pm 2^\circ$ – $\pm 10^\circ$ per-segment angular offsets simulating calibration drift) causes progressive degradation (-16.2% at S4), yet OneBEV++ still reaches 47.18 mIoU under $\pm 10^\circ$ perturbation, exceeding the clean-data performance of all baselines. This suggests that the CRP sampling grid can partially absorb moderate geometric distortions. Missing View (30° – 120° zeroed-out region simulating camera failure) is the most damaging corruption (-72.6% at S4), as expected when a third of the panoramic input is removed.

Notably, OneBEV++ and OneBEV show nearly identical *relative* drops under Missing View (-72.6% vs. -73.3%), indicating that this sensitivity is inherent to the panoramic BEV paradigm rather than a weakness of CRP. Across all corruption types and severity levels, OneBEV++ retains its absolute mIoU advantage over OneBEV (approximately +5 mIoU), confirming that the CRP + MVTv2 gains do not come at the cost of robustness.

Analysis of Efficiency. Tab. 7 provides a detailed comparison of inference efficiency across various BEV semantic segmentation models, including our OneBEV family (OneBEV and OneBEV++). Key metrics include parameters (Param.), Floating Point Operations (FLOPs), Frames Per Second (FPS), and mean Intersection-over-Union (mIoU).

The OneBEV family excels in performance and efficiency. OneBEV++ achieves an impressive mIoU of 22.0 on the Stanford-360 dataset with only 28.9M parameters and 88G FLOPs, surpassing all other models. This marks a significant improvement over OneBEV [14] (31M parameters, 171G FLOPs, 20.3 mIoU), reflecting our ongoing optimization efforts for enhanced accuracy and efficiency.

Compared to other leading models like BEVFusion [38], BEVSegformer [13], BEVFormer [1], and 360BEV [5], the OneBEV family offers a superior balance of performance and efficiency. For instance, BEVFusion and BEVFormer have massive computational costs (207G and 206G FLOPs, respectively) but deliver much lower accuracy (7.8 and 17.3 mIoU). While 360BEV achieves a higher FPS (47.75), its accuracy is limited (17.8 mIoU). OneBEV++, in contrast, delivers a significantly higher mIoU (22.0) while maintaining a strong 28.15 FPS, making it ideal for real-time applications that demand both high accuracy and efficient processing.

Fig. 9 highlights the OneBEV family’s strong learning behav-

TABLE 7. **Comparison of inference efficiency.**

Model	VT	Param.	FLOPs	FPS	mIoU
BEVFusion [38]	CrossAtt [13]	53.7M	207G	34.13	7.8
BEVSegformer [13]	CrossAtt [13]	53.1M	100G	31.19	16.0
BEVFormer [1]	CrossAtt [13]	38.9M	206G	33.90	17.3
360BEV (ours)	CrossAtt [13]	32.9M	99G	47.75	17.8
OneBEV (ours)	MVT [14]	31M	171G	13.36	20.3
OneBEV++ (ours)	MVTv2	28.9M	88G	28.15	22.0

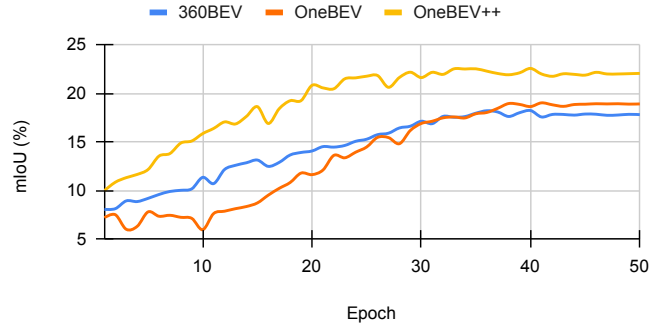


Fig. 9: **Training convergence curves** of 360BEV, OneBEV, and OneBEV++ on the Stanford-360 Dataset. OneBEV++ achieves faster convergence and superior performance.

ior. OneBEV initially lags behind 360BEV but ultimately exceeds it in final performance. OneBEV++ further excels with faster convergence and consistently higher accuracy, outperforming both OneBEV and 360BEV.

In conclusion, the OneBEV family, led by OneBEV++, sets a new standard for BEV semantic segmentation by reducing model size and computational cost while achieving state-of-the-art performance, making it highly effective for real-world applications requiring both accuracy and efficiency.

6 CONCLUSION

In this work, we introduce OneBEV++, a unified and efficient framework for panoramic BEV semantic mapping, capable of handling both indoor and outdoor environments with a single panoramic image input and no domain-specific adaptation. By leveraging a novel Columnar Radial Projection (CRP) and a Feature Selection Block (FSB), integrated into the optimized MVTv2 architecture, our framework achieves significant performance improvements with reduced computational complexity. Extensive experiments across four benchmark datasets, Stanford-360, Matterport-360, NuScenes-360, and DeepAccident-360, demonstrate that OneBEV++ consistently outperforms existing specialized methods in accuracy while simultaneously offering substantial reductions in model size and FLOPs. These results highlight the versatility and robustness of OneBEV++, paving the way for scalable and practical deployment in diverse real-world applications, from autonomous driving to indoor robotics.

REFERENCES

- [1] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEV-Former: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022. 1, 3, 6, 7, 9
- [2] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for 3d detection and map segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8721–8731. 1, 3
- [3] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021. 1, 3
- [4] C. Chen, J. Zhang, K. Yang, K. Peng, and R. Stiefelhofen, "Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping From Egocentric Images to Allocentric Semantics With Vision Transformers," in *WACV*, 2023. 1, 3
- [5] Z. Teng, J. Zhang, K. Yang, K. Peng, H. Shi, S. Reiß, K. Cao, and R. Stiefelhofen, "360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View," in *WACV*, 2024. 1, 2, 3, 5, 6, 8, 9
- [6] J. Jiang, Y. Yang, Y. Deng, C. Ma, and J. Zhang, "BEVNav: Robot autonomous navigation via spatial-temporal contrastive learning in bird's-eye view," *IEEE RAL*, 2024. 1
- [7] F. Cao, S. Wang, X. Chen, T. Wang, and L. Liu, "BEV-LSLAM: A novel and compact bev lidar slam for outdoor environment," *IEEE RAL*, 2025. 1
- [8] J. Schramm, N. Vödisch, K. Petek, B. R. Kiran, S. Yogamani, W. Burgard, and A. Valada, "BEVCar: Camera-radar fusion for bev map and object segmentation," in *IROS*. IEEE, 2024. 1
- [9] H. Shi, C. Pang, J. Zhang, K. Yang, Y. Wu, H. Ni, Y. Lin, R. Stiefelhofen, and K. Wang, "CoBEV: Elevating roadside 3d object detection with depth and height complementarity," *IEEE Transactions on Image Processing*, 2024. 1
- [10] S. Xie, L. Kong, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Benchmarking and improving bird's eye view perception robustness in autonomous driving," *IEEE T-PAMI*, 2025. 1
- [11] A. Singh, "Vision-radar fusion for robotics bev detections: A survey," in *IEEE IV*. IEEE, 2023. 1
- [12] Q. Zhang, Z. Zhang, W. Cui, J. Sun, J. Cao, Y. Guo, G. Han, W. Zhao, J. Wang, C. Sun, L. Zhang, H. Cheng, Y. Chen, L. Wang, J. Tang, and R. Xu, "Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots," *arXiv preprint arXiv:2503.09010*, 2025. 1
- [13] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *WACV*, 2023. 1, 3, 6, 7, 8, 9
- [14] J. Wei, J. Zheng, R. Liu, J. Hu, J. Zhang, and R. Stiefelhofen, "OneBEV: Using One Panoramic Image for Bird's-Eye-View Semantic Mapping," in *ACCV*, 2024, pp. 583–596. 1, 2, 3, 4, 6, 7, 9
- [15] J. Zhang, K. Yang, H. Shi, S. Reiß, K. Peng, C. Ma, H. Fu, P. H. Torr, K. Wang, and R. Stiefelhofen, "Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [16] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *2018 24th international conference on pattern recognition (icpr)*. IEEE, 2018, pp. 2190–2195. 2
- [17] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360 object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12959–12966. 2
- [18] S. Orhan and Y. Bastanlar, "Semantic segmentation of outdoor panoramic images," *Signal, Image and Video Processing*, vol. 16, no. 3, pp. 643–650, 2022. 2
- [19] Y. Xu, K. Wang, K. Yang, D. Sun, and J. Fu, "Semantic segmentation of panoramic images using a synthetic dataset," in *Artificial Intelligence and Machine Learning in Defense Applications*, vol. 11169. SPIE, 2019, pp. 90–104. 2
- [20] K. Yang, X. Hu, and R. Stiefelhofen, "Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 1866–1881, 2021. 2
- [21] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. Stiefelhofen, "Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16917–16927. 2
- [22] J. Zheng, R. Liu, Y. Chen, K. Peng, C. Wu, K. Yang, J. Zhang, and R. Stiefelhofen, "Open panoramic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 164–182. 2
- [23] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11536–11545. 2
- [24] H. Jiang, Z. Sheng, S. Zhu, Z. Dong, and R. Huang, "Unifuse: Unidirectional fusion for 360 panorama depth estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1519–1526, 2021. 2
- [25] M. Rey-Area, M. Yuan, and C. Richardt, "360monodepth: High-resolution 360deg monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3762–3772. 2
- [26] C.-H. Peng and J. Zhang, "High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3116–3125. 2
- [27] A. Jaus, K. Yang, and R. Stiefelhofen, "Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 1421–1427. 2
- [28] —, "Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4438–4453, 2023. 2
- [29] J. Mei, A. Z. Zhu, X. Yan, H. Yan, S. Qiao, L.-C. Chen, and H. Kretschmar, "Waymo open dataset: Panoramic video panoptic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 53–72. 2
- [30] C. Sun, M. Sun, and H.-T. Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2573–2582. 3
- [31] H.-a. Gao, B. Tian, P. Li, X. Chen, H. Zhao, G. Zhou, Y. Chen, and H. Zha, "From semi-supervised to omni-supervised room layout estimation using point clouds," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2803–2810. 3
- [32] V. Cartillier, Z. Ren, N. Jain, S. Lee, I. Essa, and D. Batra, "Semantic mapnet: Building allocentric semantic maps and representations from egocentric views," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 964–972. 3
- [33] H. Liu, R. Liu, K. Yang, J. Zhang, K. Peng, and R. Stiefelhofen, "Hida: Towards holistic indoor understanding for the visually impaired via semantic instance segmentation with a wearable solid-state lidar sensor," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1780–1790. 3
- [34] Z. Zhu, S. Peng, V. Larsson, Z. Cui, M. R. Oswald, A. Geiger, and M. Pollefeys, "Nicer-slam: Neural implicit scene encoding for rgb slam," in *International Conference on 3D Vision (3DV)*, March 2024. 3
- [35] H. Zhou, Z. Guo, Y. Ren, S. Liu, L. Zhang, K. Zhang, and M. Li, "Modslam: monocular dense mapping for unbounded 3d scene reconstruction," *IEEE Robotics and Automation Letters*, 2024. 3
- [36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 3
- [37] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. 3
- [38] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781. 3, 4, 6, 7, 9
- [39] M. Kondapally, K. N. Kumar, and C. K. Mohan, "Object detection in transitional weather conditions for autonomous vehicles," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8. 3
- [40] X. Yan, J. Yang, X. Zhu, Y. Liang, and H. Huang, "Denoising framework based on multiframe continuous point clouds for autonomous driving lidar in snowy weather," *IEEE Sensors Journal*, vol. 24, no. 7, pp. 10515–10527, 2024. 3
- [41] Z. Guo, S. Perminov, M. Konenkov, and D. Tsetserukou, "Hawkdive: A transformer-driven visual perception system for autonomous driving in night scene," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2598–2603. 3
- [42] J. Li, B. Li, Z. Tu, X. Liu, Q. Guo, F. Juefei-Xu, R. Xu, and H. Yu, "Light the night: A multi-condition diffusion framework for unpaired

- low-light enhancement in autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 205–15 215. **3**
- [43] J. Hu, J. Zheng, J. Wei, J. Zhang, and R. Stiefelhagen, “Deformable mamba for wide field of view segmentation,” *arXiv preprint arXiv:2411.16481*, 2024. **3**
- [44] J. Strader, N. Hughes, W. Chen, A. Speranzon, and L. Carlone, “Indoor and outdoor 3d scene graph generation via language-enabled spatial ontologies,” *IEEE Robotics and Automation Letters*, 2024. **3**
- [45] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng *et al.*, “Delving into the devils of bird’s-eye-view perception: A review, evaluation and recipe,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. **3**
- [46] H. A. Mallot, H. H. Bühlhoff, J. Little, and S. Bohrer, “Inverse perspective mapping simplifies optical flow computation and obstacle detection,” *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991. **3**
- [47] B. Zhou and P. Krähenbühl, “Cross-view transformers for real-time map-view semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769. **3**
- [48] A. Saha, O. Mendez, C. Russell, and R. Bowden, “Translating images into maps,” in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 9200–9206. **3**
- [49] H. Wang, H. Tang, S. Shi, A. Li, Z. Li, B. Schiele, and L. Wang, “Unitr: A unified and efficient multi-modal transformer for bird’s-eye-view representation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6792–6802. **3**
- [50] C. Pan, Y. He, J. Peng, Q. Zhang, W. Sui, and Z. Zhang, “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*, 2023, pp. 9590–9599. **3**
- [51] J. Philion and S. 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*. Springer, 2020, pp. 194–210. **3**
- [52] E. Xie, Z. Yu, D. Zhou, J. Philion, A. Anandkumar, S. Fidler, P. Luo, and J. M. Alvarez, “M² bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation,” *arXiv preprint arXiv:2204.05088*, 2022. **3**
- [53] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023. **3**
- [54] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, “Joint 2D-3D-semantic data for indoor scene understanding,” *arXiv preprint arXiv:1702.01105*, 2017. **3**
- [55] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from RGB-D data in indoor environments,” in *3DV*, 2017. **3**
- [56] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631. **4**
- [57] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, pp. 91–110, 2004. **4**
- [58] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 404–417. **4**
- [59] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571. **4**
- [60] T. Wang, S. Kim, J. Wenxuan, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, “Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5599–5606. **4**
- [61] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “Carla: An open urban driving simulator,” in *Conference on robot learning*. PMLR, 2017, pp. 1–16. **4**
- [62] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, “Vmamba: Visual state space model,” *arXiv preprint arXiv:2401.10166*, 2024. **4, 5, 6, 7**
- [63] A. Shaker, S. T. Wasim, S. Khan, J. Gall, and F. S. Khan, “Groupmamba: Efficient group-based visual state space model,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 912–14 922. **7, 8**



Jiale Wei received his B.Sc. degree in Mechanical Engineering and Automation from Changshu Institute of Technology in 2018. He received his M.Sc. degree in Mechatronics and Information Technology from Karlsruhe Institute of Technology (KIT) in 2024 and is currently a research assistant at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at the Institute for Anthropomatics and Robotics (IAR) at KIT. His research interests include computer vision, autonomous driving, and AI4Health.



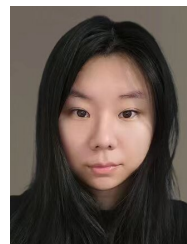
Zhifeng Teng received his B.Sc. degree in Mechanical Engineering and Automation from Zhejiang University in 2016. He received his M.Sc. degree in Mechatronics and Information Technology from Karlsruhe Institute of Technology (KIT) in 2023 and is currently a Robotics System Engineer at XYZ Robotics GmbH. He completed internships at Robert Bosch GmbH and worked at SolarLab Aiko. His research interests include computer vision, deep learning, and logistics robotic systems.



Fei Teng received his B.Sc degree in Process Equipment and Control Engineering from China University of Mining and Technology (CUMT) in 2015. He received his M.Sc degree in Computer Science from Karlsruhe Institute of Technology (KIT) in 2023. He is now working toward a Ph.D. degree in Control Science and Engineering at Hunan University, Changsha, China. His research fields include scene understanding and intelligent vehicles.



Junwei Zheng received his B.Sc degree from Guangdong University of Technology (GDUT) in 2016. He received his M.Sc degree from Karlsruhe Institute of Technology (KIT) in 2022 and is currently a research assistant at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at the Institute for Anthropomatics and Robotics (IAR) at KIT. He completed his internship at Fraunhofer Institute for Applied and Integrated Security (Fraunhofer-AISEC), Germany. His research interests include visual navigation/localization, semantic segmentation, and real-time applications.



Ruiping Liu received his B.Sc degree in Mechatronics from Tongji University in 2018. She received her M.Sc degree in Mechatronics and Information Technology from Karlsruhe Institute of Technology (KIT) in 2022 and is currently a research assistant at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at the Institute for Anthropomatics and Robotics (IAR) at KIT. She completed internships at Shanghai ShenZhou Vehicle Energy & Environmental Protection Co. and Bosch Rexroth. Her research interests include computer vision, deep learning, semantic segmentation, and real-time applications.



Yufan Chen received his B.Sc degree in Communication Engineering from Beijing Jiaotong University in 2018. He received his M.Sc. degree in Electrical Engineering and Information Technology from University of Stuttgart in 2022 and is currently a research assistant at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at the Institute for Anthropomatics and Robotics (IAR) at Karlsruhe Institute of Technology (KIT). He completed internships at ZTE Co., China, and Sony R&D Center Stuttgart, Germany. His research interests include computer vision, document analysis, image Information retrieval, and neural network robustness.



Jie Hu received his B.Sc. degree in mechatronics engineering from the Harbin Institute of Technology (HIT), China, in 2019. He received his M.Sc. degree in Mechatronics and Information Technology from the Karlsruhe Institute of Technology (KIT) in Germany in 2025. He is currently a research assistant with the Computer Vision for Human-Computer Interaction (CV:HCI) Lab and the Institute of Measurement and Control Systems (MRT) at KIT. His research interests include computer vision, semantic segmentation,

and autonomous driving.



Kailun Yang received his dual B.S. degrees in Measurement Technology and Instrument from Beijing Institute of Technology (BIT) and Economics from Peking University (PKU) in 2014. He received his Ph.D. degree in Information Sensing and Instrumentation from the State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University (ZJU) in 2019. He performed a Ph.D. internship with the Robotics and eSafety (RobeSafe) research group at the University of Alcalá (UAH) from 2017 to 2018.

He was a postdoctoral researcher with the Computer Vision for Human-Computer Interaction (CV:HCI) lab at Karlsruhe Institute of Technology (KIT) from 2019 to 2023. He is currently a professor at the School of Robotics and the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University (HNU). For more information, visit his website: <https://yangkailun.com>.



Jiaming Zhang (Member, IEEE) received his M.Sc. and Ph.D. (summa cum laude) degrees from the Karlsruhe Institute of Technology (KIT), Germany, in 2020 and 2023. He was a Postdoctoral Researcher at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at KIT. He was a visiting researcher at the Torr Vision Group (TVG) at the University of Oxford in 2023, and an Academic Guest at the Computer Vision and Geometry (CVG) Group at ETH Zurich in 2024. He is currently a Professor with

the School of Artificial Intelligence and Robotics at Hunan University (HNU). His research interests lie at the intersection of computer vision and assistive technologies, building intelligent and inclusive systems to assist humans, particularly those with disabilities.



Rainer Stiefelhagen received his Diplom (Dipl.-Inform) and Doctoral degree (Dr.-Ing.) from the Universität Karlsruhe (TH) in 1996 and 2002, respectively. He is currently a full professor for "Information technology systems for visually impaired students" at the Karlsruhe Institute of Technology (KIT), where he directs the Computer Vision for Human-Computer Interaction Lab at the Institute for Anthropomatics and Robotics as well as Center for Digital Accessibility and Assistive Technology. His research

interests include computer vision methods for visual perception of humans and their activities, in order to facilitate perceptive multimodal interfaces, humanoid robots, smart environments, multimedia analysis, and assistive technology for persons with visual impairments.