KD360-VoxelBEV: LiDAR and 360-degree Camera Cross Modality Knowledge Distillation for Bird's-Eye-View Segmentation

Wenke E*   Yixin Sun*   Jiaxu Liu   Hubert P. H. Shum
Amir Atapour-Abarghouei   Toby P. Breckon
Durham University, UK

Abstract

We present the first cross-modality distillation framework specifically tailored for single-panoramic-camera Bird's-Eye-View (BEV) segmentation. Our approach leverages a novel LiDAR image representation fused from range, intensity and ambient channels, together with a voxel-aligned view transformer that preserves spatial fidelity while enabling efficient BEV processing. During training, a high-capacity LiDAR and camera fusion Teacher network extracts both rich spatial and semantic features for cross-modality knowledge distillation into a lightweight Student network that relies solely on a single 360-degree panoramic camera image. Extensive experiments on the Dur360BEV dataset demonstrate that our teacher model significantly outperforms existing camera-based BEV segmentation methods, achieving a 25.6% IoU improvement. Meanwhile, the distilled Student network attains competitive performance with an 8.5% IoU gain and state-of-the-art inference speed of 31.2 FPS. Moreover, evaluations on KITTI-360 (two fisheye cameras) confirm that our distillation framework generalises to diverse camera setups, underscoring its feasibility and robustness. This approach reduces sensor complexity and deployment costs while providing a practical solution for efficient, low-cost BEV segmentation in real-world autonomous driving. The code is available at: `https://github.com/Tom-E-Durham/KD360-VoxelBEV`.

1. Introduction

Bird's-Eye-View (BEV) representations play a crucial role in autonomous driving, offering a top-down perspective that greatly simplifies sensor fusion [24, 30, 42, 49] and benefits downstream tasks such as detection [25, 28, 50], segmentation [3, 4, 15, 17, 33], motion forecasting [10, 51], and navigation planning [18, 36]. While most BEV generation approaches rely on multi-camera solutions [3, 4, 15, 25, 48],
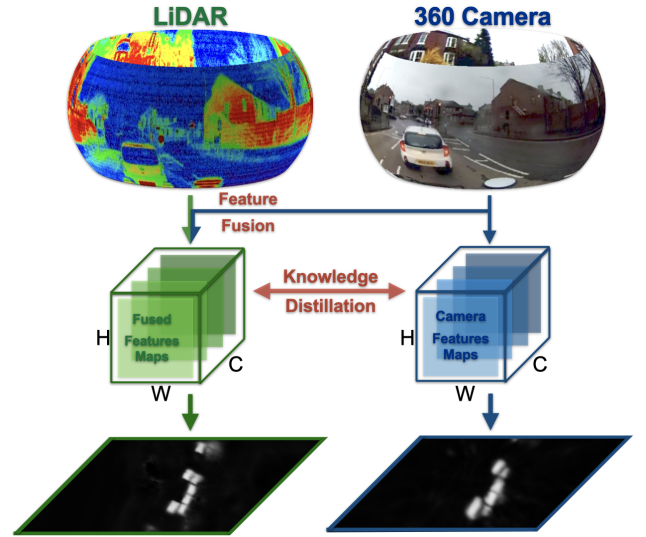
---

*Equal contribution.

Figure 1. Overview of cross-modality channel-wise knowledge distillation from fused LiDAR–camera features (Teacher) to a single 360-degree camera model (Student) for enhanced feature representation and scene understanding.

the simultaneous deployment of multiple sensors introduces significant hardware complexity and cost. In contrast, a single 360-degree camera must distribute its finite number of pixels across the entire panorama, leading to lower per-pixel density and reduced feature quality [9, 34, 45]. This resolution constraint poses additional challenges for downstream tasks like BEV segmentation, which rely on precise object boundaries and detailed spatial cues.

Recent advances show that scaling backbone depth can mitigate low-resolution inputs by extracting richer features [4, 15]. However, accuracy gains quickly plateau with model size, incurring prohibitive compute, latency, and on-vehicle energy costs for real-time deployment. In parallel, LiDAR supplies high-fidelity 3D measurements [37] and multimodal fusion (camera+LiDAR/radar) attains strong performance [26, 28]; yet keeping LiDAR active at inference introduces substantial cost, maintenance,

and size/weight/power burdens, limiting large-scale feasibility. This raises a practical question: can we exploit LiDAR during training while discarding it at deployment?

We address this with *KD360-VoxelBEV*, a cross-modality distillation framework that leverages LiDAR for training but relies only on a single camera at inference. Unlike existing cross-modality distillation works in BEV detection or HD map construction [5, 14, 20, 22, 46], our focus is on the under-explored single panoramic camera BEV segmentation setting, where the 360-degree equirectangular geometry and the low-resolution constraint introduce unique challenges that prior multi-camera methods do not address. We introduce a high-capacity Teacher network that fuses LiDAR and camera features in BEV space via a soft-gated mechanism, while our LiDAR representation and voxel-aligned projection ensure that LiDAR-derived features retain both geometric fidelity and compatibility with 2D backbones. In contrast to existing approaches [14, 52] that enforce direct alignment across different modalities and often yield suboptimal results, our module employs *Soft-Gated Fusion Module* (SGFM) to reduce representation discrepancies and further introduces an *Auxiliary Module* (AM) to provide additional distillation signals, thereby facilitating faster convergence of the Student. These fused features supervise a lightweight camera-only Student, enabling deployment-friendly BEV segmentation without sacrificing accuracy.

We primarily develop and evaluate *KD360-VoxelBEV* on Dur360BEV [9], the latest real-world 360° single-camera dataset with the highest-resolution LiDAR among existing benchmarks. To further test generalisation, we also conduct experiments on KITTI-360, which provides two fisheye cameras with full 360-degree coverage.

The main contributions of this work are as follows:

- We propose *KD360-VoxelBEV*, the first cross-modality distillation framework for BEV segmentation, where the SGFM integrates LiDAR and camera features in the Teacher, providing strong multimodal supervision to a lightweight panoramic camera-only Student.
- We introduce a novel voxel-aligned view transformer that preserves the geometric fidelity of LiDAR while remaining compatible with efficient 2D backbones, thereby combining the accuracy of voxel-based encoders [55] with the efficiency of pillar-based designs [23].
- We design a unified panoramic LiDAR image representation, co-parameterised with equirectangular camera imagery, which ensures consistent cross-modal alignment and efficient 2D processing. In conjunction with the voxel-aligned view transformer, this representation enables a strong BEV Teacher and supports robust cross-dataset distillation.

2. Related Work

We consider prior work in two related topic areas: BEV segmentation (Section 2.1) and cross-modality knowledge distillation (Section 2.2).

2.1. BEV Segmentation

BEV segmentation maps provide a view of the environment for tasks such as scene understanding, path planning and motion prediction [18]. Moreover, BEV representations fuse data from multiple sensors (e.g., LiDAR, camera, radar) [28, 35, 44, 53]. BEV methods differ mainly in their inputs: image-based approaches rely solely on camera data, while multi-modality methods integrate additional sensors.

2.1.1. Image-Based vs. Multi-Modality BEV Approaches

Most image-based BEV methods rely on multi-camera setups, typically employing six cameras, to provide a complete 360-degree surround view around the vehicle. This configuration offers high-definition views in all directions and provides dense semantic information. However, the multi-camera approach increases the memory requirements for data input and may introduce inter-view redundancy. In contrast, a single 360-degree camera input reduces the input data memory requirements while still capturing sufficient scene context. Other works fuse multiple sensors, such as LiDAR and radar, to enhance the performance of BEV segmentation by providing point cloud.

Image-Based BEV: Recent work includes LSS [33], which employs an end-to-end pipeline that transforms multi-view camera images into a unified BEV representation by implicitly unprojecting 2D features to 3D space. FIERY [17] refines this by applying a multi-task framework with uncertainty weights and a probabilistic model for predicting future instance segmentation and motion in BEV. BEVFormer [25] and its updated version BEVFormerV2 [48] adopt transformer-based architectures to effectively fuse multi-view inputs, demonstrating the potential of more complex models in generating rich BEV representations. SimpleBEV [15] introduces a parameter-free feature pulling method to replace the need for predicting depth distribution, whilst PointBEV [4] presents a novel sparse paradigm to replace the dense grid sampling method, which offers flexible memory/performance trade-offs. Furthermore, GaussianBEV [3] applies an optimisation free 3D Gaussian generator to transform image feature map into 3D Gaussians. Related monocular depth estimation work leverages semantic segmentation priors to regularise single-view 3D reconstruction [1].

Multi-Modality BEV: Several approaches improve BEV segmentation by incorporating additional sensors beyond cameras. For instance, SimpleBEV [15] enhances performance by down-sampling radar and LiDAR data into grid

representations, which are then concatenated with RGB features. Similarly, BEVFusion [28] uses a LiDAR encoder to extract features that are flattened along the z-axis before merging with RGB features in BEV space. While these fusion methods boost segmentation performance, they introduce extra processing steps that increase computational load and may slow inference, posing challenges for real-time deployment. Additionally, PointBEV [4] departs from conventional fusion methods by proposing a coarse and fine training using flattened LiDAR data as sample points. These 2D BEV points then serve as sample locations for constructing 3D pillars of fixed height for sparse feature extraction. Nevertheless, the reliance on fixed-height pillars overlooks crucial vertical geometry, leading to a loss of LiDAR-specific spatial detail.

2.2. Cross-Modality Knowledge Distillation

Traditional knowledge distillation is originally designed for transferring information from a large teacher model to a compact student model with the same input data [16], whereas cross-modality knowledge distillation supports distillation between different input modalities. Previous cross-modality knowledge distillation model have been applied between data types: depth to RGB [12], sketch to photographic [13], and synthetic to real [40]. More recently, several works have adapted cross-modality distillation between LiDAR and cameras for 3D perception. BEVDistill [5] introduces a dense foreground-guided feature imitation mechanism along with sparse instance-wise distillation to transfer rich spatial information from LiDAR to multi-camera 3D object detection networks. Similarly, DistillBEV [46] employs region decomposition and adaptive scaling to achieve a more fine-grained cross-modal distillation, enhancing the alignment between LiDAR and camera modalities for BEV perception tasks. UniDistill [54] adopts a three-level approach—feature, relation, and response—allowing for flexible teacher-student pairs of LiDAR or camera networks.

One challenge of cross-modality knowledge distillation is that it is not always effective when the representation and distribution gap between modalities is large, as forcing the student to mimic the teacher often leads to suboptimal performance [19, 32]. Another challenge is that most existing approaches have focused on 3D object detection rather than BEV segmentation, and none address the 360-degree equirectangular format that naturally arises from a single panoramic camera. In this work, we bridge these gaps by proposing an SGFM-based cross-modality knowledge distillation framework with AM for BEV segmentation, introducing the first system that leverages LiDAR data to guide a single-camera model operating on full 360-degree equirectangular image input.

Consequently, we do not provide direct quantitative comparisons to existing cross-modality distillation frameworks [5, 14, 20, 22, 46, 54] as the existing literature focuses on different tasks (e.g., 3D object detection or map reconstruction), employ LiDAR point cloud encoders and assume multi-camera rigs, while our framework targets single panoramic camera BEV segmentation using a unified LiDAR–camera equirectangular image representation.

3. Methodology

Here, we introduce our novel LiDAR-driven distillation framework (Section 3.1), which includes a new LiDAR data representation, a voxel-aligned view transformer (Section 3.2), and a cross-modality knowledge distillation architecture (Section 3.3).
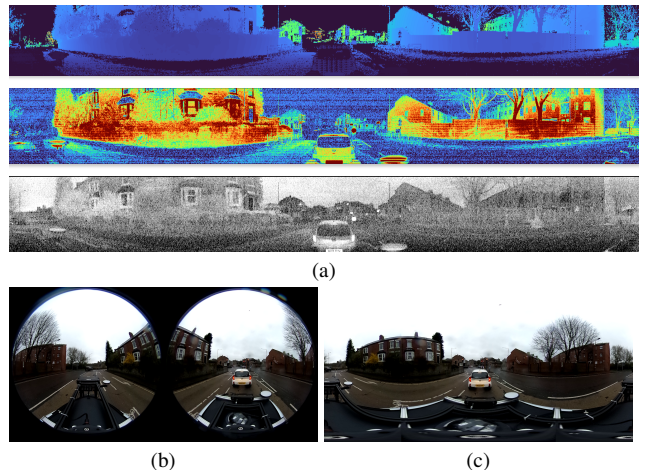

Figure 2. **Illustration of the Dur360BEV dataset [9].** (a) LiDAR data in equirectangular representation [Top: range image; Middle: intensity image; Bottom: ambient image]. (b) Dual-fisheye spherical image. (c) Equirectangular-projected 360-degree image.

3.1. Unified LiDAR Image Representation

We introduce a unified LiDAR image representation that encodes raw point clouds into a dense equirectangular grid of shape $(H \times W \times 3)$. Each pixel corresponds to a unique combination of azimuth and elevation for a full 360-degree coverage consistent with the sensor's field of view. The three channels store *range*, *intensity* and *ambient*, thereby capturing both geometric and radiometric information. This representation offers two key advantages. First, it integrates seamlessly into efficient image-based BEV segmentation architectures, fully exploiting the detailed spatial and contextual information provided by high-resolution LiDAR. Second, by expressing both LiDAR and RGB in a consistent spherical image space, it ensures natural alignment between modalities and thus reduces the distribution gap in cross-modality knowledge distillation (Figure 2).

Dur360BEV [9] natively provides all three fields for constructing this representation(as shown in Figure 2a),
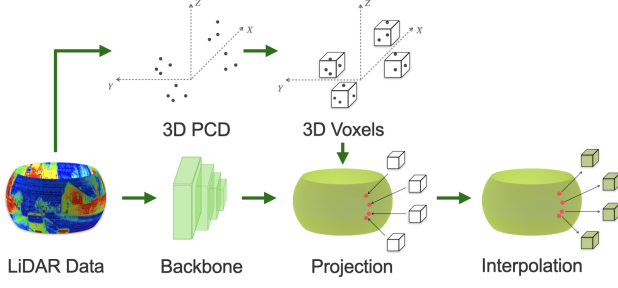
Figure 3. **Sparse Voxel Pulling Module (View Transformer).** 3D voxels derived from LiDAR point cloud are mapped to localised equirectangular LiDAR features, then bilinearly interpolated to form 3D BEV features.

whereas other datasets offer only a subset of modalities. The corresponding dataset-specific adjustments are described in Section 4.

While range and intensity are widely adopted in LiDAR-based perception frameworks, the ambient modality is less commonly explored. Unlike range, which encodes geometric distance, or intensity, which measures the strength of actively reflected laser pulses, ambient captures the background near-infrared radiation (800–2500 nm) received by the sensor. It therefore provides information complementary to range and intensity, reflecting scene properties under natural illumination conditions rather than active returns alone. This additional modality enriches our unified LiDAR image representation (Figure 2a) and offers robustness in scenarios with challenging lighting, such as nocturnal or adverse-weather environments, without replacing the geometric or reflective cues already available.

3.2. Voxel-Aligned View Transformer

We introduce a novel *voxel-aligned view transformer*, a feature projection module that maps equirectangular image features into the BEV space guided by LiDAR voxelisation (as shown in Figure 3). Unlike grid-based methods that predict on dense BEV grids [15] or pillar-based approaches that form fixed-height vertical columns from 2D BEV points [4], our design aligns voxels directly with the underlying LiDAR geometry. This yields a representation that combines the strengths of voxel- and pillar-based models. It preserves the spatial fidelity of voxel features while, once compressed into the BEV plane, remaining lightweight and directly amenable to standard 2D CNN backbones. In contrast to pillar-based methods that risk sampling outside the true object extent, our voxel alignment ensures precise and noise-free feature projection.

Formally, given an equirectangular feature map $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ extracted from the input image backbone, we voxelise the LiDAR point cloud $\mathbf{P} = \{p \mid p = (x, y, z) \in \mathbb{R}^3\}$ within a detection range $[-R, R]$ using voxel size (r_x, r_y, r_z). Each voxel centre is then mapped to image

coordinates, and bilinear interpolation retrieves the corresponding visual features, producing a voxel-level feature tensor $\mathbf{F} \in \mathbb{R}^{B \times C \times Z \times Y \times X}$. This operation leverages the 360-degree coverage of the equirectangular format to fully utilise the voxel grid without suffering from multi-camera overlaps or out-of-FoV artefacts [4, 15, 25, 48].

Finally, the voxel features are compressed along the vertical axis to yield a BEV feature map, which is further refined using a lightweight sparse U-Net to generate segmentation predictions.

In summary, our design combines the unified LiDAR image representation with the voxel-based view transformer. This integration fully exploits the geometric and radiometric data from LiDAR since the voxel projection preserves spatial accuracy, while the image-based representation ensures compatibility with efficient 2D CNN architectures.

3.3. Knowledge Distillation

Overview. As illustrated in Figure 4, our approach employs a soft-gated LiDAR–camera *Teacher* model (Section 3.3.2) to guide a cost-effective *Student* (Section 3.3.3) that operates solely on single-camera images. The *Teacher* uses SGFM (Section 3.3.1) to adaptively balance geometric information from LiDAR with appearance cues from the camera yielding richer multimodal representations. To further reduce the feature gap between *Teacher* and *Student*, we introduce an AM (Section 3.3.4) during training. These spatial and semantic insights are distilled into *Student* to allow the camera-only model to achieve improved BEV segmentation accuracy at test time without the hardware and computational overhead of LiDAR sensors.

3.3.1. Soft-Gated Fusion Module

The proposed SGFM is designed to adaptively combine LiDAR- and image-derived BEV features. It consists of three main steps: feature concatenation, gate generation and gated fusion, followed by a refinement stage. Let $\mathbf{F}^I \in \mathbb{R}^{C \times H \times W}$ and $\mathbf{F}^L \in \mathbb{R}^{C \times H \times W}$ denote the feature maps from the image and LiDAR branches, respectively. After channel-wise concatenation, we obtain the joint feature map $\mathbf{F}^{cat} \in \mathbb{R}^{2C \times H \times W}$. A 1×1 convolution with weights $\mathbf{W} \in \mathbb{R}^{C \times 2C \times 1 \times 1}$ is applied to $\mathbf{F}^{cat}$ to model cross-modal correlations and predict a gating coefficient map:

$$\mathbf{G}_{k,i,j} = \sigma \Big(\sum_{k'=1}^{2C} \mathbf{F}^{cat}_{k',i,j} \mathbf{W}_{k,k'} \Big),$$

$$\forall\, k \in [1, C],\; i \in [1, H],\; j \in [1, W]. \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function, ensuring $\mathbf{G} \in [0, 1]^{C \times H \times W}$. Here, $\mathbf{G}$ acts as a soft gate: higher values increase the contribution of image features, while lower values emphasize LiDAR features. The gated fusion result is
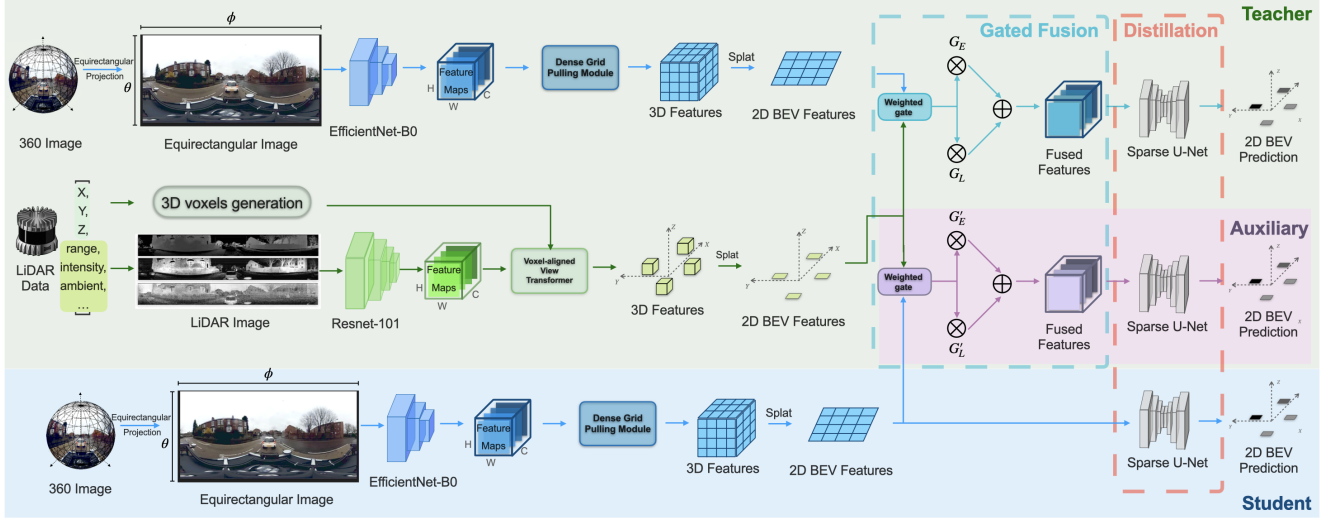
Figure 4. **Overview of the proposed *KD360-VoxelBEV* architecture. Teacher network (green):** equipped with the **SGFM (blue dashed block)**, which integrates LiDAR range, intensity, and ambient cues with 360-degree camera features to produce enriched BEV representations. **AM (pink):** fuses the Student and pre-trained LiDAR branch during training to reduce the feature gap between Teacher and Student, providing additional reliable guidance. **Student network (blue):** a camera-only BEV segmentation model that benefits from cross-modal distillation, achieving robust BEV predictions from a single 360-degree input image. **Distillation (red dashed block):** highlights the regions where multi-channel dense feature distillation is applied, specifically between Teacher–Student and Student–Auxiliary pairs. At inference, only the Student network is employed, ensuring lightweight and deployment-friendly BEV segmentation.

thus computed as

$$\mathbf{F}^{fuse} = \mathbf{G} \odot \mathbf{F}^{I} + (1 - \mathbf{G}) \odot \mathbf{F}^{L}, \qquad (2)$$

where $\odot$ denotes element-wise multiplication. Finally, $\mathbf{F}^{fuse}$ is refined by a 3×3 convolution followed by batch normalization and ReLU activation:

$$\mathbf{F}^{out} = \phi\big(\text{Conv}_{3\times3}([\mathbf{F}^{fuse}; \mathbf{F}^{fuse}])\big), \qquad (3)$$

where $\phi(\cdot)$ denotes the BN–ReLU operation. This formulation allows the network to learn spatially varying, channel-aware weights that balance LiDAR geometry and image semantics, thereby yielding more reliable multimodal BEV representations for distillation.

3.3.2. Teacher: Soft-Gated LiDAR-Camera Fusion

Our Teacher network consists of LiDAR and 360-degree camera branches whose features are fused through a soft-gated mechanism) in BEV space, providing a unified multimodal representation for supervision. The LiDAR inherently provides precise 3D geometric information, while the 360-degree camera contributes complementary visual appearance cues. Inspired by [6], we adopt a late-fusion strategy at the BEV feature level, where modality-specific features are abstracted independently and subsequently fused after BEV conversion.

Conventional approaches that directly operate on raw point clouds with dedicated 3D backbones followed by

view transformation often suffer from distribution mismatch when combined with image features. To address this issue, we transform the LiDAR point cloud into a 3-channel equirectangular image representation, consisting of range, intensity, and ambient channels. This compact 2D representation enables the use of well-established convolutional backbones for effective feature extraction. The resulting feature maps are then projected into the BEV domain via our voxel-pulling module, which preserves fine-grained geometric details while maintaining global scene context. By combining these LiDAR features with those from the 360-degree image branch through a soft-gated fusion mechanism, the Teacher produces enriched BEV representations that serve as strong supervisory signals for distillation.

3.3.3. Student: Camera-Only

The Student model is lightweight and cost-effective, operating exclusively on a single 360-degree camera input. We adopt a compact 2D backbone to extract features and a dense view-transformation module (grid sampling) to map image features to BEV, mirroring the Teacher's projection to ensure feature-level alignment during distillation. This design allows both networks to produce spatially compatible BEV features. Although the Student lacks direct 3D geometry and thus tends to produce coarser BEV features, we mitigate this limitation via cross-modality distillation that aligns the Student's intermediate representations (and out-
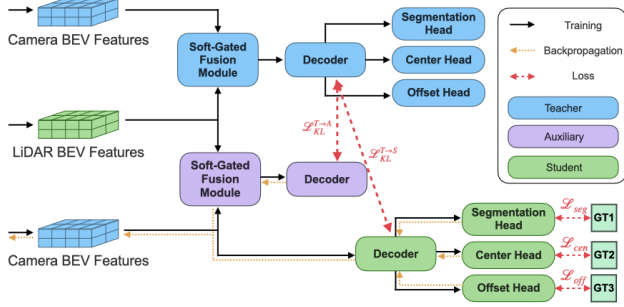
Figure 5. **Illustration of distillation and auxiliary details.** Feature maps from the decoder are used for channel-wise distillation, which is applied between Teacher and Student as well as between Auxiliary and Student, as indicated by the red dashed arrows.

puts) with those of the Teacher.

Following the Dur360BEV benchmark [8], the Student predicts three BEV heads: car segmentation, centerness, and offset. We supervise them with focal loss [39] (segmentation), balanced MSE [38] (centerness), and ℓ_1 loss (offset). These terms are combined into the Student's multi-task objective with learned weights via uncertainty-based weighting [21]; see Eq. (4).

$$\mathcal{L}_{\text{stu}} = \lambda_1 \, \mathcal{L}_{\text{seg}} + \lambda_2 \, \mathcal{L}_{\text{cen}} + \lambda_3 \, \mathcal{L}_{\text{off}}. \quad (4)$$

3.3.4. Auxiliary Module

The representation gap between the multimodal Teacher and the camera-only Student is substantial. Directly forcing the Student to match the Teacher often leads to unstable optimisation and suboptimal minima [32]. To mitigate this, we introduce a training-only auxiliary branch that fuses the Student's camera features with the Teacher's LiDAR features using the same soft-gated mechanism. The auxiliary branch is supervised by the Teacher and provides an additional distillation signal to the Student, as illustrated in Figure 5. This design supplies a smoother intermediate target that is closer to the Student's modality while preserving LiDAR geometry for faster convergence. The auxiliary branch is disabled at inference, adding no runtime overhead.

3.3.5. Channel-Wise Dense Feature Distillation

Inspired by channel-wise knowledge distillation for dense prediction [41], we perform feature-level distillation at the decoder output (Fig. 1). We investigate feature distillation at different network depths: early, middle and late stages, with detailed results reported in the ablation study (Section 4.5). We denote the fused BEV feature maps from the Teacher, Auxiliary, and Student as $\mathbf{F}^T$, $\mathbf{F}^A$, and $\mathbf{F}^S$ (each of size $C \times H \times W$). To compare features across modalities, we normalise each channel of the Teacher, Auxiliary, and Student feature maps into spatial probability distributions

using a softmax with temperature $\mathcal{T}$. For channel c this is defined as

$$\phi(\mathbf{F}_c) = \frac{\exp(\mathbf{F}_{c,i}/\mathcal{T})}{\sum_{i=1}^{H \cdot W} \exp(\mathbf{F}_{c,i}/\mathcal{T})}, \quad c = 1, \ldots, C. \quad (5)$$

where i indexes spatial locations and $\mathcal{T}$ controls the distribution smoothness.

The discrepancy between two normalised feature maps is measured using the Kullback–Leibler (KL) divergence. To aggregate across channels and compensate for temperature scaling, we define the KL-based distillation objectives for the Teacher–Student and Teacher–Auxiliary pairs as

$$\mathcal{L}_{\text{KL}}^{T \to S} = \frac{\mathcal{T}^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{H \cdot W} \phi(\mathbf{F}_{c,i}^T) \, \log \frac{\phi(\mathbf{F}_{c,i}^T)}{\phi(\mathbf{F}_{c,i}^S)}, \quad (6)$$

$$\mathcal{L}_{\text{KL}}^{T \to A} = \frac{\mathcal{T}^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{H \cdot W} \phi(\mathbf{F}_{c,i}^T) \, \log \frac{\phi(\mathbf{F}_{c,i}^T)}{\phi(\mathbf{F}_{c,i}^A)}. \quad (7)$$

Ablation results analysing the effect of applying KL at different stages in Section 4.5.1, and its advantage over alternative objectives in Section 4.5.2.

The overall channel-wise distillation loss is then expressed as

$$\mathcal{L}_{\text{KD}} = \alpha_1 \, \mathcal{L}_{\text{KL}}^{T \to S} + \alpha_2 \, \mathcal{L}_{\text{KL}}^{T \to A}, \quad (8)$$

The overall training objective combines the channel-wise distillation loss in Eq. (8) and the multi-task Student loss in Eq. (4):

$$\mathcal{L} = \mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{stu}}. \quad (9)$$

4. Experiments

We compare our approach with existing camera-based BEV segmentation methods. We benchmark against state-of-the-art methods on Dur360BEV [9] (Section 4.4) and further evaluate on KITTI-360 [27] with a two-fisheye setup (Section 4.4). We also present ablations (Section 4.5) on Dur360BEV to study distillation configurations and loss choices for *KD360-VoxelBEV*.

4.1. Datasets

Dur360BEV [9] offers single panoramic-camera imagery paired with 128-channel LiDAR. We use the Extended version, featuring longer routes, more diverse scenes, and higher-quality annotations over the Initial release, with 13.9k frames for training and 1.5k for validation. Ground-truth BEV segmentation maps are derived from 3D bounding boxes, following the protocol of SimpleBEV [15] and the official Dur360BEV setup [9]. **KITTI-360** [27] provides two side-mounted fisheye cameras covering 360-degree and a 64-channel LiDAR. Since the dataset does

not include BEV annotations, we generate ground-truth BEV segmentation maps following the same protocol as in Dur360BEV [9](details in Appendix E). We adopt the split from [11] and downsample the size to 20k frames for training and 1k for validation due to computational constraint.

We also considered other datasets, such as Mapillary [31] and nuScenes [2]. The former provides panoramic imagery with fused point clouds but lacks per-frame LiDAR scans and dynamic objects, while the latter offers six surrounding-view cameras with 32-beam LiDAR but requires fusing the perspective images into an equirectangular view, introducing visible seams and artifacts [47]. These limitations make both datasets unsuitable for our BEV segmentation framework, see details in Appendix B.

4.2. Evaluation Metrics

We report Intersection over Union (IoU) across three square-shaped ranges: 100m×100m, 50m×50m, and 20m×20m. Following the Dur360BEV benchmark [9], we also adopt the efficiency ratio (ER), defined as IoU divided by model parameters, to characterise the trade-off between accuracy and model complexity. In addition, we report frames per second (FPS) to measure inference speed.

4.3. Experimental Setup

In our *voxel-aligned view transformer* module, we use voxels of size 0.5m in each dimension, and both the Teacher and the Student networks are configured within a detection range of $[100m \times 8m \times 100m]$ (in x,y,z order, where y is height) yielding a 200×200 BEV map covering $100m \times 100m$ in the x–z plane. Following prior BEV segmentation works [4, 9, 15], we adopt EfficientNet and ResNet as our backbones. Since our contribution focuses on the cross-modality distillation framework rather than backbone design, we intentionally use these widely-adopted baselines to ensure fair comparison and isolate the improvements brought by our distillation method.

Both Teacher and Student models are trained independently on an NVIDIA A100 80G GPU. For all main experiments, we train each network for up to 12k iterations with a batch size of 6, a learning rate of 5×10^{-4}, and a weight decay of 1×10^{-5}, using AdamW [29] with a 1-cycle learning rate schedule [43]. For inference, we evaluate all models on an NVIDIA RTX 3080 GPU and detailed configurations are provided in the Appendix D.

4.4. State-of-the-Art Comparison

Dur360BEV. Table 1 reports a quantitative comparison with existing BEV segmentation methods, and the corresponding qualitative results are illustrated in Figure 6. The comparison reveals three key findings:

(i) By incorporating LiDAR through the proposed SGFM, the Teacher's performance increases to 58.3%

Model	Modalities	Backbone	IoU$_{100}$	IoU$_{50}$	IoU$_{20}$	ER	FPS
SimpleBEV [15]	C	RN-101	31.1	37.0	38.5	0.74	<u>25.7</u>
PointBEV [4]	C	EN-b4	31.5	38.9	39.7	3.75	15.0
Dur360BEV (dense) [9]	C	RN-101	32.7	40.4	42.0	0.78	25.3
Dur360BEV (coarse/fine) [9]	C	EN-b4	32.6	40.3	41.6	<u>3.88</u>	14.9
Ours (Teacher)	LC	RN-101/EN-b0	**58.3**	**64.3**	**68.0**	0.85	14.0
Ours (Student)	C	EN-b0	32.2	41.8	58.4	3.04	**31.2**
KD360-VoxelBEV, distill w/o AM	LC $\rightarrow$ C	EN-b0	39.4	49.3	64.2	3.72	**31.2**
KD360-VoxelBEV, distill w/ AM	LC $\rightarrow$ C	EN-b0	<u>41.2</u>	<u>51.7</u>	<u>67.1</u>	**3.89**	**31.2**

Table 1. Comparison results (%) on **Dur360BEV** dataset [9] with metrics: IoU ($\uparrow$), ER ($\uparrow$) and FPS ($\uparrow$). **Bold** and <u>underlined</u> numbers denote the best (Top-1) and second best (Top-2), respectively.

Model	Modalities	Backbone	IoU$_{100}$	IoU$_{50}$	IoU$_{20}$
Ours (Teacher)	LC	RN-101/EN-b0	53.9	65.8	76.4
Ours (Student)	C	EN-b0	24.1	37.9	63.8
KD360-VoxelBEV	LC $\rightarrow$ C	EN-b0	32.3	46.6	69.6

Table 2. Comparison results (%) on **KITTI-360** dataset [27].

IoU$_{100}$, confirming the effectiveness of multimodal fusion. The impact of different LiDAR channel combinations (Range, Ambient, and Intensity), as well as the effect of combining LiDAR with camera, are further analysed in the ablation study (see Appendix A). Results demonstrate that the configuration combining all three LiDAR channels with the camera achieves the best performance.

(ii) With the same lightweight EfficientNet-B0 backbone, our distillation framework improves the Student performance to 39.4% IoU$_{100}$, surpassing all existing cameraonly BEV segmentation methods. Moreover, the introduction of the AM further narrows the representation gap, pushing the distilled Student to 41.2% IoU$_{100}$ while achieving the highest efficiency ratio. This demonstrates an excellent trade-off between accuracy and computational cost, making the model highly suitable for deployment in resourceconstrained scenarios.

KITTI360. From Table 2, since no prior methods are tailored for the KITTI-360 car segmentation task, we directly transfer our distillation configuration and evaluate it on this dataset. The results show that knowledge distillation continues to exploit multimodal advantages even under limited sensor settings, where LiDAR provides only two channels (range and intensity) and the cameras are wide-FoV fisheye. Compared with the Student baseline, our *KD360-VoxelBEV* improves performance by +8.2% IoU$_{100}$, demonstrating the robustness and generalisability of our approach across datasets with different modality characteristics.

4.5. Ablation Studies

We conduct a series of ablation experiments to better understand the factors driving the performance of *KD360-VoxelBEV*.

4.5.1. Effect of Distillation Stages and AM:

To evaluate the effectiveness of knowledge distillation, we consider three distinct stages (shown in Figure 7): (i) af-
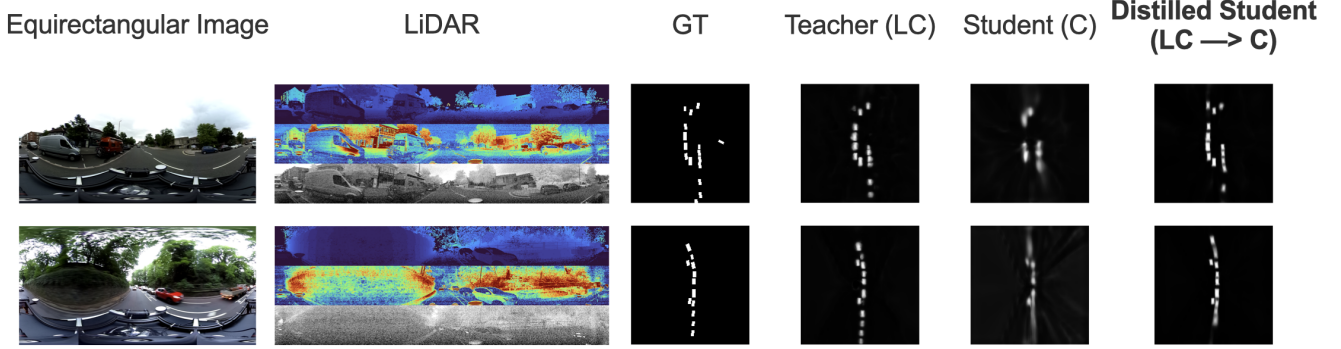
Figure 6. Qualitative BEV segmentation results on the Dur360BEV dataset [9].

Stage1	Stage2	Stage3	AM	IoU_{100}	IoU_{50}	IoU_{20}
✓	–	–	w/o	20.96	31.53	47.05
–	✓	–	w/o	20.58	30.88	49.48
–	–	✓	w/o	39.37	49.34	64.21
✓	–	–	w/	30.50	40.33	51.33
–	✓	–	w/	32.22	42.11	56.42
–	–	✓	w/	41.24	51.69	67.08

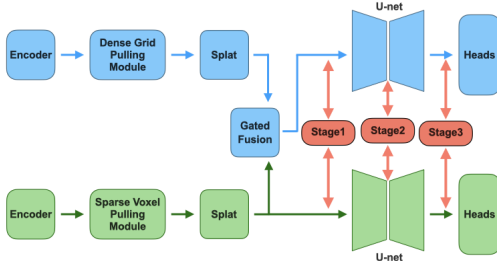Table 3. Ablation study of KD model with different stage setups.

Figure 7. Illustration of distillation stages for ablation study (Auxiliary module excluded). Blue blocks denote Teacher modules, and green blocks denote Student modules.

ter feature fusion (stage1; or, without gated fusion, the corresponding location in the LiDAR branch), (ii) the feature map at the output of the U-Net encoder (stage2), and (iii) the feature map at the output of the U-Net decoder (stage3). At each stage, we apply the channel-wise distillation loss in Eq. (6) between Teacher and Student, optionally combined with the auxiliary distillation loss in Eq. (7).

Overall, as shown in Table 3, applying distillation at later stages yields larger performance gains, since early convolutional features are highly modality-specific and may introduce noise or instability when distilled, whereas deeper semantic representations are more modality-invariant and thus transfer more effectively across modalities. In addition, incorporating the AM further mitigates the representation gap between Teacher and Student, leading to consistent improvements across all IoU thresholds.
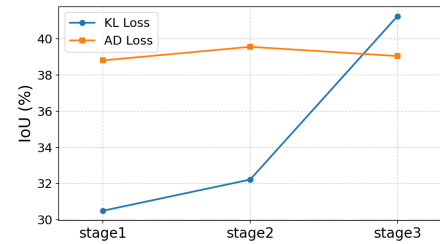

Figure 8. Comparison of distillation losses at different stages.

4.5.2. Distillation Losses:

We evaluate two different distillation objectives: the Kullback–Leibler (KL) divergence loss [16] and the Affinity Distillation (AD) loss [7]. A comparison of their performance is presented in Fig. 8. As shown, the AD loss yields more stable performance across all stages, whereas the KL loss underperforms AD in the earlier stages but catches up at deeper layers. This trend can be attributed to the fact that KL divergence is more effective at aligning high-level semantic distributions than low-level spatial correlations.

5. Conclusion

We present KD360-VoxelBEV, a cross-modality distillation framework tailored to BEV segmentation from a single panoramic camera, where a soft-gated LiDAR–camera fusion Teacher supervises a lightweight camera-only Student. Our framework combines a voxel-aligned view transformer with a unified panoramic LiDAR image representation, preserving geometric fidelity while allowing for efficient BEV reasoning with 2D backbones. Experiments on Dur360BEV [9] and KITTI-360 [27] show that the Teacher substantially outperforms existing camera-based BEV segmentation methods and that the distilled Student matches or surpasses state-of-the-art camera-only baselines [4, 9, 15] while running in real time. In future work, we plan to validate the framework on a broader range of datasets and investigate its integration with transformer-based architectures to further enhance representation capacity.

References

[1] Amir Atapour-Abarghouei and Toby P. Breckon. Monocular segment-wise depth: Monocular depth estimation based on a semantic segmentation prior. In *2019 IEEE International Conference on Image Processing*, pages 4295–4299, 2019. 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 7, 12

[3] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. GaussianBeV: 3D Gaussian Representation meets Perception Models for BeV Segmentation. *arXiv preprint arXiv:2407.14108*, 2024. 1, 2

[4] Loick Chambon, Eloi Zablocki, Mickaël Chen, Florent Bartoccioni, Patrick Pérez, and Matthieu Cord. PointBeV: A Sparse Approach for BeV Predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15195–15204, 2024. 1, 2, 3, 4, 7, 8, 13

[5] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection. *arXiv preprint arXiv:2211.09386*, 2022. 2, 3

[6] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-Sensitive Deconvolution Networks With Gated Fusion for RGB-D Indoor Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[7] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. MonoDistill: Learning Spatial Features for Monocular 3D Object Detection. *arXiv preprint arXiv:2201.10830*, 2022. 8

[8] Viet Hung Duong, Duc Quyen Nguyen, Thien Van Luong, Huan Vu, and Tien Cuong Nguyen. Robust Data Augmentation and Ensemble Method for Object Detection in Fisheye Camera Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024. 6

[9] Wenke E, Chao Yuan, Li Li, Yixin Sun, Yona Falinie A. Gaus, Amir Atapour-Abarghouei, and Toby P. Breckon. Dur360BEV: A Real-world Single 360-Degree Camera Dataset and Benchmark for Bird-Eye View Mapping in Autonomous Driving, 2025. 1, 2, 3, 6, 7, 8, 12, 13

[10] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, 2021. 1

[11] Nikhil Gosala and Abhinav Valada. Bird's-Eye-View Panoptic Segmentation Using Monocular Frontal View Images. *IEEE Robotics and Automation Letters*, 7(2):1968–1975, 2022. 7

[12] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross Modal Distillation for Supervision Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[13] Xiaoguang Han, Chang Gao, and Yizhou Yu. DeepSketch2Face: A Deep Learning Based Sketching System for 3D Face and Caricature Modeling. *ACM Transactions on graphics (TOG)*, 36(4):1–12, 2017. 3

[14] Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang. MapDistill: Boosting Efficient Camera-Based HD Map Construction via Camera-LiDAR Fusion Model Distillation. In *European Conference on Computer Vision*, pages 166–183. Springer, 2024. 2, 3

[15] Adam W. Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-BEV: What Really Matters for Multi-Sensor BEV Perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765, 2023. 1, 2, 4, 6, 7, 8, 12, 13

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015. 3, 8

[17] Anthony Hu, Zak Murez, Nikhil Mohan, Sofia Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: Future Instance Prediction in Bird's-Eye View from Surround Monocular Cameras. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15253–15262, Montreal, QC, Canada, 2021. IEEE. 1, 2

[18] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-Oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1, 2

[19] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2024. 3

[20] Yuzhe Ji, Yijie Chen, Liuqing Yang, Rui Ding, Meng Yang, and Xinhu Zheng. VeXKD: The Versatile Integration of Cross-Modal Fusion and Knowledge Distillation for 3D Perception. *Advances in Neural Information Processing Systems*, 37:125608–125634, 2024. 2, 3

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, 2018. 6

[22] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X^3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13343–13353, 2023. 2, 3

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 12697–12705, 2019. 2

[24] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1477–1485, 2023. 1

[25] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *Computer Vision – ECCV 2022*, pages 1–18. Springer Nature Switzerland, Cham, 2022. 1, 2, 4

[26] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 1

[27] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 3292–3310, 2023. 6, 7, 8, 12, 13

[28] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 2, 3

[29] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[30] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. BEV-Guided Multi-Modality Fusion for Driving Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023. 1

[31] Mapillary. Mapillary metropolis: City-scale multimodality benchmarking dataset for 3d semantic map making. https://www.mapillary.com/dataset/metropolis, 2021. Accessed: 2025-09-17. 7

[32] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020. 3, 6

[33] Jonah Philion and Sanja Fidler. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *Computer Vision – ECCV 2020*, pages 194–210. Springer International Publishing, Cham, 2020. 1, 2, 12

[34] Yeqiang Qian, Ming Yang, and John M. Dolan. Survey on Fish-Eye Cameras and Their Applications in Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22755–22771, 2022. 1

[35] Yiran Qin, Chaoqun Wang, Zijian Kang, Ningning Ma, Zhen Li, and Ruimao Zhang. SupFusion: Supervised LiDAR-Camera Fusion for 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22014–22024, 2023. 2

[36] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. UniFusion: Unified Multi-View Fusion Transformer for Spatial-Temporal Representation in Bird's-Eye-View. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8690–8699, 2023. 1

[37] Thinal Raj, Fazida Hanim Hashim, Aqilah Baseri Huddin, Mohd Faisal Ibrahim, and Aini Hussain. A Survey on LiDAR Scanning Mechanisms. *Electronics*, 9(5):741, 2020. 1

[38] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for Imbalanced Visual Regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7926–7935, 2022. 6

[39] T-YLPG Ross and GKHP Dollár. Focal Loss for Dense Object Detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 6

[40] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[41] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-Wise Knowledge Distillation for Dense Prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5311–5320, 2021. 6

[42] Apoorv Singh. Vision-RADAR fusion for Robotics BEV Detections: A Survey. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2023. 1

[43] Leslie N Smith and Nicholay Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 7

[44] Yixin Sun, Li Li, Amir Atapour-Abarghouei, Toby P Breckon, et al. Tomd: A trail-based off-road multi-modal dataset for traversable pathway segmentation under challenging illumination conditions. *arXiv preprint arXiv:2506.21630*, 2025. 2

[45] Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. 360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 373–382, 2024. 1

[46] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. DistillBEV: Boosting Multi-Camera 3D Object Detection with Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8637–8646, 2023. 2, 3

[47] Jiale Wei, Junwei Zheng, Ruiping Liu, Jie Hu, Jiaming Zhang, and Rainer Stiefelhagen. OneBEV: Using One Panoramic Image for Bird, Aos-Eye-View Semantic Mapping. In *Proceedings of the Asian Conference on Computer Vision*, pages 583–596, 2024. 7

[48] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao,

Lewei Lu, Jie Zhou, and Jifeng Dai. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision, 2022. 1, 2, 4

[49] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. FusionAD: Multi-modality Fusion for Prediction and Planning Tasks of Autonomous Driving. *arXiv preprint arXiv:2308.01006*, 2023. 1

[50] Xin Ye, Burhaneddin Yaman, Sheng Cheng, Feng Tao, Abhirup Mallik, and Liu Ren. BEVDiffuser: Plug-and-Play Diffusion Model for BEV Denoising with Ground-Truth Guidance. *arXiv preprint arXiv:2502.19694*, 2025. 1

[51] Lu Zhang, Peiliang Li, Sikang Liu, and Shaojie Shen. SIMPL: A Simple and Efficient Multi-agent Motion Prediction Baseline for Autonomous Driving. *IEEE Robotics and Automation Letters*, 2024. 1

[52] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Zhe Chen, Jing Zhang, and Dacheng Tao. SimDistill: Simulated Multi-Modal Distillation for BEV 3D Object Detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7460–7468, 2024. 2

[53] Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15470–15480, 2024. 2

[54] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5116–5125, 2023. 3

[55] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2

Appendix

A. Dur360BEV-mini

Similar to the official nuScenes-mini [2] split (1% of the full dataset, 10 scenes in total), we construct a mini version of Dur360BEV-Extended to facilitate efficient input modality ablation studies and debugging. Specifically, we uniformly downsample the Extended split by a factor of 10, resulting in 1,350 training and 150 validation frames. This reduced set preserves the distribution of the original dataset while enabling much faster experimentation. Unless otherwise stated, all reported main results are obtained on the full Dur360BEV-Extended dataset.

B. nuScenes [2] Expriments

nuScenes [2] provides six surrounding-view cameras and a 32-channel LiDAR. We adopt the official train/val split (28.1k/6k samples) and generate BEV ground truth following the protocol in Lift-Splat [33] and SimpleBEV [15], where points inside "vehicle" bounding boxes are labeled positive and others negative.

We found that applying *KD360-VoxelBEV* to nuScenes [2] is problematic due to the image formation process. As shown in Fig. 9a, the equiangular projection is stitched from six individual cameras rather than captured by a true 360° sensor. Consequently, objects spanning across multiple cameras often become misaligned at the image borders, leading to duplicated or fragmented appearances. In addition, differences in exposure and illumination across cameras introduce visible seams and inconsistencies. These artifacts make it difficult for the model to extract coherent features, and as illustrated by the prediction in Fig. 9c, a single car may be broken into several disconnected parts. This highlights the limitation of using stitched multi-camera images in our distillation model, which relies on consistent 360° visual input.
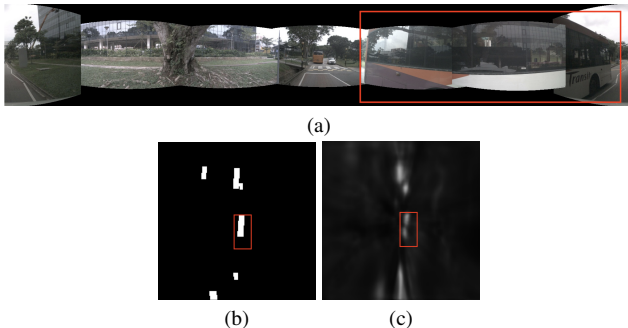

(a)

(b)　　　　(c)

Figure 9. **An illustrative case of suboptimal results when applying *KD360-VoxelBEV* on nuScenes [2].** (a) Equiangular image converted from six cameras; (b) BEV car label; (c) distillation segmentation results. The red box highlights a bus, which is fragmented due to discontinuities between stitched camera views.

Input Modality	IoU_{100}	IoU_{50}	IoU_{20}
Single-Modality Baselines			
C (Camera Only)	31.6	37.9	39.6
LiDAR (RI)	50.9	56.8	62.0
LiDAR (AI)	51.7	58.5	67.0
LiDAR (RA)	53.4	59.2	65.1
LiDAR (RAI)	54.0	61.0	67.2
Multi-Modality Fusion			
LiDAR (RAI) + C (w/ SGFM)	**58.8**	**63.4**	**70.1**

Table 4. Ablation study on input modalities and SGFM using the Dur360BEV-mini [9]. Metrics: IoU (↑). **R**: Range, **A**: Ambient, **I**: Intensity, **C**: 360° Camera. The proposed SGFM effectively fuses LiDAR and Camera features to outperform unimodal baselines.

C. Effect of Input Modalities and SGFM

To enable faster experimentation, we perform this study on Dur360BEV-mini, while all main results are reported on the full dataset. The details of the Dur360BEV-mini are provided in Appendix A. We follow the same experimental configuration as in the main experiments, but train on the Dur360BEV-mini dataset up to 4k iterations to account for the reduced amount of input data.

Impact of LiDAR Channels. We first analyse the contribution of the three LiDAR channels (range, intensity, ambient) by selectively enabling or disabling them. As shown in Table 4, ambient information proves highly important: removing it (LiDAR (RI)) causes a notable performance drop compared to the full configuration (LiDAR (RAI)), decreasing IoU_{100} from 54.0% to 50.9%. The inclusion of ambient data consistently enhances BEV segmentation alongside range and intensity. This complementary cue, which captures environmental illumination beyond geometry and reflectivity, may explain why our distillation framework achieves stronger gains on Dur360BEV [9] compared to KITTI-360 [27], which lacks ambient measurements.

Effectiveness of SGFM. We further validate the effectiveness of our Soft-Gated Fusion Module (SGFM) by comparing single-modality baselines against the fused system. As reported in Table 4, the Camera-only (C) baseline achieves 31.6% IoU_{100}, limited by the lack of explicit depth information. The LiDAR-only (RAI) model performs significantly better at 54.0% IoU_{100} due to precise geometric sensing. However, by integrating both modalities via SGFM, our Teacher model (LiDAR (RAI) + C) reaches 58.8% IoU_{100}, outperforming the strongest single-modality baseline by +4.8%. This improvement indicates that SGFM effectively leverages the complementary nature of the two sensors—combining the rich semantic and texture cues from the 360° camera with the accurate spatial geometry from LiDAR—to produce a more robust BEV representation.

D. Inference Time Measurement Details

We benchmark inference time for all models on the Dur360BEV dataset using a single NVIDIA RTX 3080 GPU. All measurements are conducted at a fixed input resolution of 1024×2048 under PyTorch with CUDA/cuDNN enabled. Each experiment is repeated twenty times, and the average is reported.

Table 5 summarizes the latency, throughput (FPS), and batch size used for inference. While the teacher model exhibits the slowest inference, KD360-VoxelBEV achieves the best trade-off between speed and efficiency, significantly surpassing existing methods and demonstrating both the effectiveness of knowledge distillation and its suitability for real-world applications.

Model	Latency (ms) ↓	FPS ↑
SimpleBEV [15]	38.9	25.7
PointBEV [4]	66.7	15.0
Dur360BEV (dense) [9]	38.5	25.3
Dur360BEV (coarse/fine) [9]	67.1	14.9
Ours (Teacher)	71.7	14.0
KD360-VoxelBEV	32.1	31.2

Table 5. Inference time comparison of different models on the Dur360BEV dataset using an NVIDIA RTX 3080 GPU.

E. KITTI-360

E.1. Image Processing

KITTI-360 provides raw images from two side-mounted wide-FoV fisheye cameras. We project each fisheye image into spherical coordinates using the official calibration parameters and camera FoV, and then reproject to an equirectangular format. In this space, pixels are parameterised by azimuth and elevation angles, covering the full 360-degree field of view (see Figure 10b, 10c). This conversion yields a panoramic representation consistent with Dur360BEV, facilitating cross-dataset training and evaluation.

E.2. Annotation Process

The raw KITTI-360 [27] annotations consist of two categories of 3D bounding boxes: static and dynamic objects. Static objects are labeled on accumulated point clouds, while dynamic objects are annotated on individual frames. Directly using these annotations introduces two issues: (1) static boxes from accumulated scans may include objects that are either fully occluded in a given frame or located beyond the effective LiDAR/camera detection range, and (2) dynamic boxes must be associated with the correct frame using timestamps.

To address this, we explored several filtering strategies. Our initial attempt using the timestamp range (start/end)
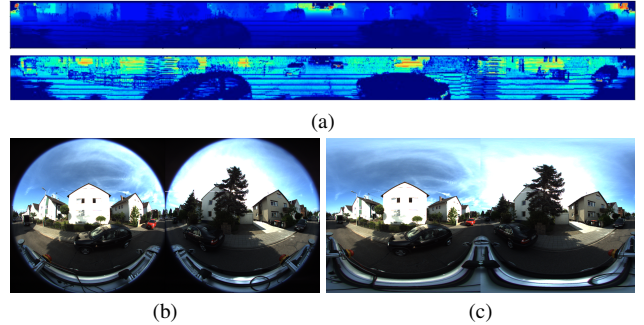

Figure 10. **Illustration of the KITTI-360 dataset [9].** (a) LiDAR data in equirectangular representation [Up: range image; Down: intensity image]. (b) Dual-fisheye spherical image. (c) Equirectangular-projected 360-degree image.

to filter static boxes resulted in missing objects on repeated routes, while a distance-based strategy occasionally retained static boxes outside the current FoV. Our final approach introduces per-frame LiDAR checks: static boxes are retained only if they contain points in the corresponding frame's point cloud, thereby ensuring consistency with the instantaneous scene visibility. Dynamic boxes are preserved using frame IDs. After this filtering, we follow the
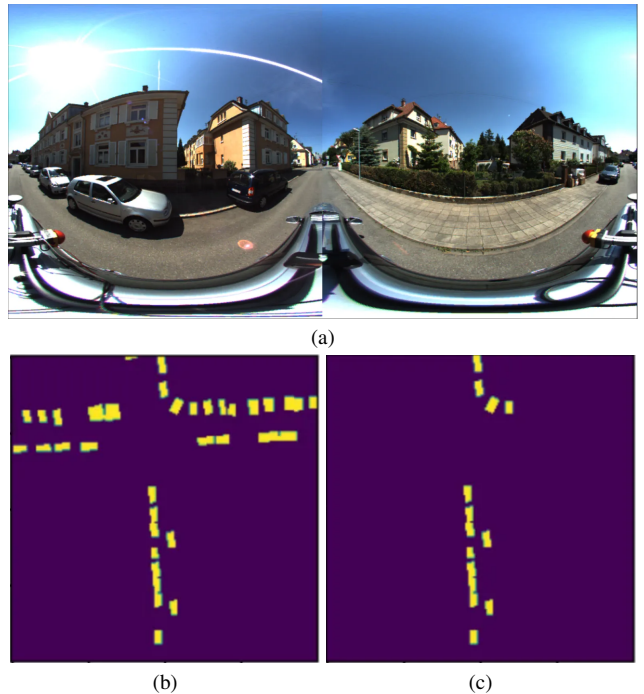

Figure 11. **Illustration of the annotation filtering on KITTI-360 [9].** (a) Equirectangular image. (b) BEV GT segmentation with distance-based strategy, where many vehicles fully occluded by buildings are still projected into the current frame. (c) BEV GT segmentation after per-frame LiDAR checks, which effectively removes such artifacts.

same procedure as in Dur360BEV to rasterize 3D bounding boxes into BEV ground-truth maps.