Appendix

A. Dur360BEV-mini

Similar to the official nuScenes-mini [1] 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 [1] Expriments

nuScenes [1] 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 [6] and SimpleBEV [4], where points inside "vehicle" bounding boxes are labeled positive and others negative.

We found that applying *KD360-VoxelBEV* to nuScenes [1] is problematic due to the image formation process. As shown in Fig. 1a, 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. 1c), 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.

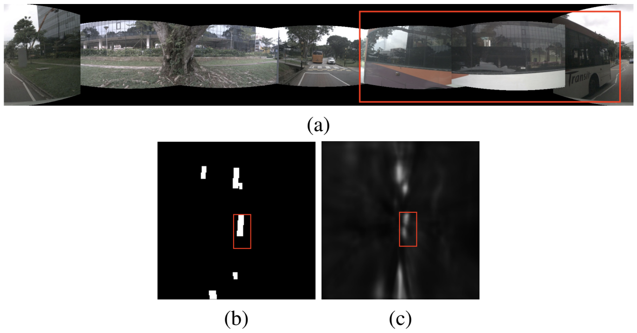

Figure 1. **An illustrative case of suboptimal results when applying *KD360-VoxelBEV* on nuScenes [1].** (a) Equiangular image converted from six cameras; (b) BEV car label; (c) distillation segmentation results. The red box highlights a bus, which is fragmented due to discontinuities between stitched 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 1. Ablation study on input modalities and SGFM using the Dur360BEV-mini [3]. 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 1, 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 [3] compared to KITTI-360 [5], 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 1, 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 2 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 [4]	38.9	25.7
PointBEV [2]	66.7	15.0
Dur360BEV (dense) [3]	38.5	25.3
Dur360BEV (coarse/fine) [3]	67.1	14.9
Ours (Teacher)	71.7	14.0
KD360-VoxelBEV	32.1	31.2

Table 2. 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 2b, 2c). This conversion yields a panoramic representation consistent with Dur360BEV, facilitating cross-dataset training and evaluation.

E.2. Annotation Process

The raw KITTI-360 [5] 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) 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 same procedure as in Dur360BEV to rasterize 3D bounding

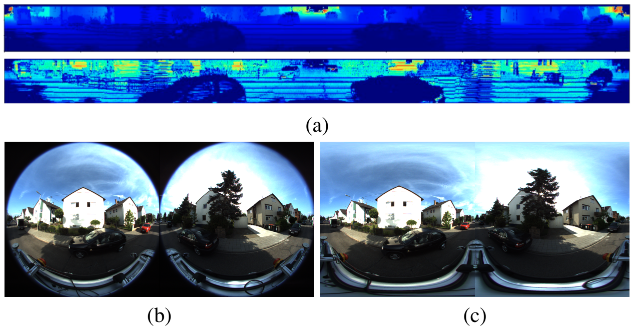

Figure 2. **Illustration of the KITTI-360 dataset [3].** (a) LiDAR data in equirectangular representation [Up: range image; Down: intensity image]. (b) Dual-fisheye spherical image. (c) Equirectangular-projected 360-degree image.

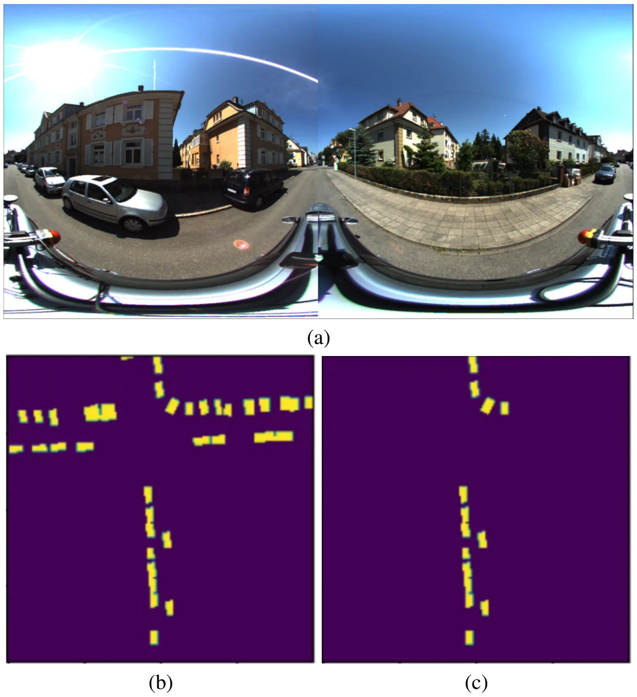

Figure 3. **Illustration of the annotation filtering on KITTI-360 [3].** (a) Equirectangular image. (b) BEV GT segmentation with distance-based strategy, where many vehicles fully occluded by buildings are still projected into the current frame. (c) BEV GT segmentation after per-frame LiDAR checks, which effectively removes such artifacts.

boxes into BEV ground-truth maps.

References

[1] 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. 1

[2] 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. 2

[3] 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

[4] 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

[5] 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. 1, 2

[6] 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