PanoHK360: A Large-Scale 8K Urban Panoramic Dataset and Benchmark for Depth Estimation

Anonymous Author(s)
Affiliation
Address
email

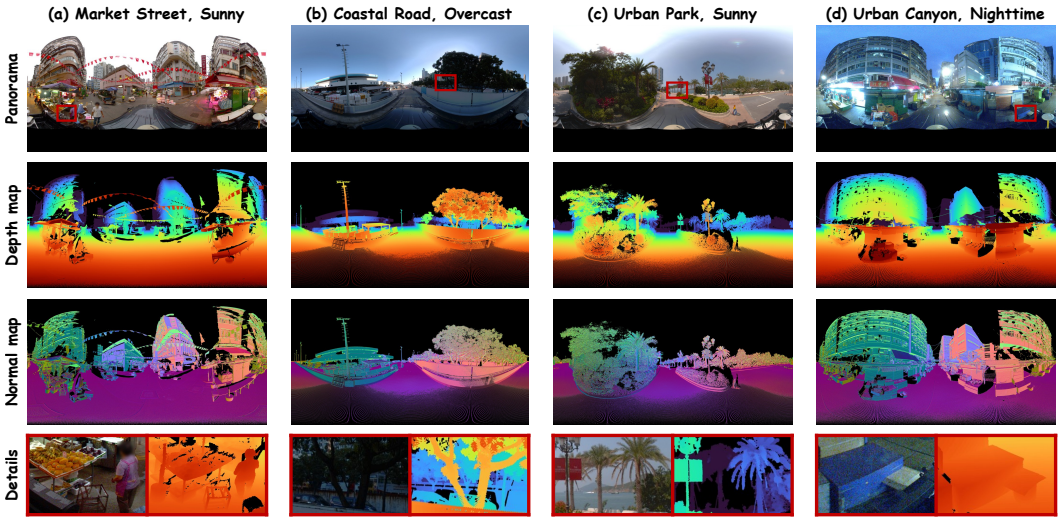


Figure 1: **Overview of PanoHK360.** A diverse, high-resolution panoramic RGB-D dataset of real-world urban scenes captured under widely varying conditions. Each column presents one representative scene category from the dataset. From top to bottom, the rows show the corresponding 8K equirectangular RGB panorama, the metric depth obtained from survey-grade LiDAR, and the surface normal computed directly from depth. We further provide raw LiDAR point clouds, accurate 6-DoF camera poses, and temporal capture sequences for downstream tasks.

Abstract

Panoramic 360° depth estimation underpins autonomous navigation, 3D scene reconstruction, and scene understanding. However, existing outdoor panoramic datasets remain limited in three aspects. Small-scale collections are insufficient for modern data-hungry architectures and tend to produce models that generalize poorly across cities. Synthetic datasets transfer poorly to real scenes due to persistent domain gaps in illumination, geometry, and material appearance. Pseudo-label datasets distilled from pretrained estimators inherit the errors of the teacher model. Their downstream accuracy is therefore capped at the performance ceiling of the teacher itself. To address these issues, we present PanoHK360, a city-scale outdoor panoramic RGB-D dataset of diverse urban scenes in Hong Kong. The data are collected by a vehicle-mounted rig that pairs 8K panoramic imaging with survey-grade LiDAR. To the best of our knowledge, PanoHK360 is the first panoramic depth dataset that simultaneously combines three defining properties: city-scale geographic coverage, 8K equirectangular resolution, and sensor-based metric depth.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

The depth is recovered directly from LiDAR rather than estimated by a pretrained network. The release further includes raw point clouds, 6-DoF camera poses, and temporal capture sequences. We further benchmark representative monocular 360° depth methods on PanoHK360 under a unified protocol with a location-disjoint training and validation split. The results reveal a substantial gap between current methods and the level of resolution and scene diversity required for real-world urban deployment. These findings establish PanoHK360 as a challenging testbed for future research.

1 Introduction

Omnidirectional perception has become increasingly essential for embodied agents operating in complex real-world environments. Unlike perspective cameras with a narrow field of view, equirectangular panoramas capture the full $360^\circ \times 180^\circ$ scene in a single frame. This format enables holistic scene understanding without explicit view stitching [2, 38]. Among the perception tasks built upon this format, panoramic depth estimation serves as a fundamental prerequisite for downstream applications such as autonomous driving, robotic navigation, mixed reality, and 3D scene reconstruction [11, 21].

Recent panoramic depth datasets have begun extending the task from indoor benchmarks [7] to outdoor environments. Pioneering works such as Depth360 [11] first formulated the outdoor task. Subsequent datasets such as PanoVILD [16] and PAIR360 [22] introduced LiDAR-aligned supervision, while synthetic alternatives such as OmniScape [29] enabled training at a large scale. Despite this progress, outdoor panoramic data remain bottlenecked by three key limitations. First, existing real-world collections rarely exceed tens of thousands of frames [11]. This scale is far below the data appetite of modern 360° depth networks and limits generalization across cities [6]. Second, purely simulated datasets [29] introduce a sim-to-real domain gap in illumination, geometry, and appearance that degrades transfer to real scenes. Third, labels distilled from pretrained estimators or sparse SfM and MVS reconstructions [16, 28] inherit the errors of the teacher model and cap downstream accuracy at the performance ceiling of the teacher itself.

We present PanoHK360, a city-scale outdoor panoramic RGB-D dataset collected across diverse urban scenes in Hong Kong. The data are captured by a vehicle-mounted rig that pairs an 8K 360° camera with a survey-grade LiDAR scanner. To the best of our knowledge, PanoHK360 is the first panoramic depth dataset to simultaneously offer three defining properties. These are city-scale coverage across multiple districts, 8K equirectangular resolution at 8000×4000 pixels, and sensor-based metric depth. The depth is recovered directly from LiDAR rather than predicted by a pretrained network. The release further includes raw point clouds, 6-DoF camera poses, and temporal sequences. We additionally introduce PanoHK360-Mini, a curated benchmark subset on which we evaluate representative monocular 360° depth methods under a unified and location-disjoint protocol. Across all baselines, the results reveal a substantial gap between current methods and the resolution and scene diversity required for real-world urban deployment. These findings position PanoHK360 as a challenging testbed for future research in high-resolution outdoor panoramic perception. A detailed comparison with existing panoramic RGB-D datasets is provided in Table 1.

Contributions. Our main contributions can be summarized as follows.

- We introduce PanoHK360, the first publicly available city-scale outdoor panoramic RGB-D dataset that simultaneously offers 8K equirectangular resolution, sensor-based metric depth from a survey-grade LiDAR scanner, and geographically diverse urban coverage, comprising approximately 2M RGB-D frames.
- We release a documented and reproducible automated pipeline that converts raw LiDAR point clouds into geometrically consistent panoramic depth maps. The pipeline handles spatial extraction, visibility-aware projection, and adaptive splatting in a unified workflow.
- We establish PanoHK360-Mini, a carefully curated benchmark subset, and evaluate representative monocular 360° depth estimation methods under a unified, location-disjoint protocol. These results jointly provide strong baselines and a quantified challenge for future work on outdoor panoramic perception.

Table 1: **Comparison with existing panoramic RGB-D datasets.** Entries reflect each dataset's original release; properties added by later third parties are not counted. ERP = equirectangular, FE = fisheye, Cyl = cylindrical. **Depth**: dense panoramic depth maps released. **Metric**: depth is recovered from sensor measurement rather than estimated by a learned model. **Sync.**: hardware-level synchronization between RGB and depth streams. **PC**: raw point clouds. PanoHK360 is the only entry that satisfies every property at city scale and 8K resolution.

Dataset	Proj.	Scenes	Real	Depth	Metric	Temp.	Pose	Sync.	PC	Size	Resolution
Stanford2D3DS [4]	ERP	indoor	✓	✓	✗	✗	✓	✓	✓	1.4K	4096×2048
Matterport3D [7]	ERP	indoor	✓	✓	✓	✗	✓	✓	✗	10.8K	2048×1024
KITTI-360 [25]	FE	outdoor	✓	✗	✓	✓	✓	✓	✓	300K	1400×1400
ToF360 [19]	ERP	indoor	✓	✓	✓	✗	✗	✓	✓	1K	5792×2896
Pano3D [3]	ERP	indoor	✗	✓	✗	✗	✓	✗	✗	42K	1024×512
JRDB [27]	Cyl/FE	out.&in.	✓	✗	✓	✓	✓	✓	✓	28K	3760×480 1280×720
Depth360 [11]	ERP	outdoor	✓	✓	✗	✗	✗	✗	✗	30K	1.03 Mpx
PanoVILD [16]	FE/ERP	outdoor	✓	✗	✓	✓	✓	✓	✓	15K	1600×1200
PAIR360 [22]	ERP/FE	outdoor	✓	✓	✗	✓	✓	✓	✓	88K	7680×3840 3840×2880
Deep360 [24]	ERP	outdoor	✗	✓	✗	✗	✓	✓	✗	12K	1024×512
Helvipad [37]	ERP	out.&in.	✓	✓	✓	✓	✗	✓	✓	40K	1920×512
PanoHK360 (ours)	ERP	outdoor	✓	✓	✓	✓	✓	✓	✓	**2M**	**8000×4000**

2 Related Work

2.1 Perspective RGB-D Datasets

Monocular depth estimation has long been propelled by perspective RGB-D benchmarks. These typically include indoor collections such as NYU Depth V2 [31] and ScanNet [9], as well as outdoor driving datasets such as KITTI [12], KITTI-360 [25], nuScenes [5], and Waymo [33]. Although these resources have collectively shaped the field over the past decade, they are designed around a narrow field of view. Reprojecting them into the equirectangular format inevitably introduces severe distortion that limits their utility for panoramic depth estimation.

2.2 Panoramic and Omnidirectional Datasets

Indoor panoramic datasets. Early panoramic depth datasets focused primarily on indoor environments. Matterport3D [7] and Stanford2D3DS [4] pair real sensor scans with high-quality equirectangular imagery, while Structured3D [39] further scales the coverage with photo-realistic CAD-rendered panoramas. These benchmarks have anchored nearly a decade of panoramic depth research. However, the depth distributions in confined indoor spaces differ fundamentally from those in outdoor scenes, both in absolute range and in overall geometric layout. Consequently, models trained exclusively on indoor data transfer poorly to outdoor driving environments [22].

Outdoor panoramic datasets. Recent efforts have begun to extend the field to outdoor driving settings. Depth360 [11] first scaled outdoor panoramic depth estimation by harvesting in-the-wild web videos and deriving pseudo-labels through Structure-from-Motion [28] and Multi-View Stereo [36]. Sidestepping real-world capture altogether, OmniScape [29] renders synthetic driving panoramas at scale through CARLA-based simulation. Closer to real sensing, PanoVILD [16] and PAIR360 [22] pair panoramic imagery with 32-beam LiDAR and densify the resulting sparse returns using learned depth-completion networks. Despite this collective progress, each line of work trades one limitation for another. Specifically, a small geographic footprint, the sim-to-real domain gap, and label noise introduced by learned densification still persist across existing datasets. The community is therefore left without a city-scale, real-world outdoor panoramic resource with sensor-grounded metric supervision, which is precisely the gap that motivates PanoHK360.

2.3 360° Depth Estimation Methods

Early 360° depth methods adapt perspective architectures to the equirectangular grid through spherical convolutions [8], uniform spherical sampling [17], or hybrid cubemap and equirectangular representations [1]. More recent work further leverages vision foundation models for zero-shot panoramic depth estimation. PanDA [6] distills Depth Anything V2 [35] via LoRA fine-tuning on equirectangular data. DA^2 [23] introduces SphereViT together with a perspective-to-panoramic curation engine for scale-invariant distance estimation. DA360 [18] adapts Depth Anything V2 with a learnable shift parameter and circular padding. DAP [26] combines a plug-and-play range-mask head with a DINOv3 [32] backbone for absolute metric scale. Although these methods have considerably advanced architectural design, their training and evaluation predominantly rely on indoor or synthetic benchmarks that lack the geometric complexity, illumination diversity, and metric depth range characteristic of large-scale outdoor environments. Consequently, none of them has yet been evaluated on large-scale, real-world outdoor panoramic data with sensor-grounded metric ground truth. PanoHK360 is designed precisely to enable this evaluation setting.

3 The PanoHK360 Dataset

3.1 Motivation and Overview

Omnidirectional depth estimation provides geometric cues for 3D scene understanding. Existing benchmarks primarily focus on indoor environments or synthetic domains. Current outdoor collections are often limited in scale and annotated with non-metric pseudo-labels. However, generalization in complex urban scenes requires real-world data with metric annotations. This data scarcity bottlenecks the development of reliable outdoor spatial intelligence [11, 18].

Motivated by these challenges, we introduce PanoHK360, the first city-scale and high-resolution panoramic RGB-D dataset captured across diverse real-world urban environments in Hong Kong. Our benchmark integrates dense RGB observations with survey-grade metric depth and hardware-synchronized camera trajectories. The full release comprises approximately 2 million equirectangular frames at 8000×4000 resolution. We additionally provide PanoHK360-Mini as a curated 22,751-frame subset for quantitative evaluation. Figure 2 illustrates our extensive acquisition routes and the corresponding multimodal data alignments. This metric depth benchmark supports research in outdoor omnidirectional perception.

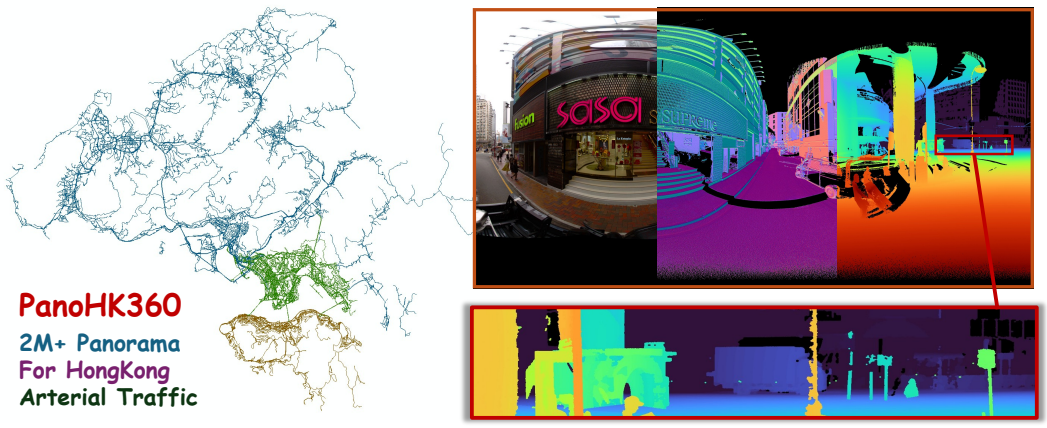


Figure 2: **Overview of PanoHK360. Left**: Data acquisition trajectories across diverse urban and arterial environments in Hong Kong. **Right**: An 8K RGB equirectangular panorama alongside its corresponding surface normal map and depth map, all geometrically aligned via calibrated intrinsic and extrinsic parameters. The dataset contains 2M+ panoramic frames and 50+ TB of point clouds.

3.2 Data Collection System

Our data acquisition platform integrates two perception sensors. We deploy a Teledyne FLIR Ladybug5+ panoramic camera and a RIEGL VUX-1UAV survey-grade LiDAR scanner. The camera module comprises six Sony IMX264 global shutter CMOS sensors. It provides a combined 30 MP resolution at 30 fps. This setup achieves approximately 90% spherical coverage of the urban environment. Hardware-level PPS and NMEA GPS inputs provide multi-sensor synchronization. Furthermore, the RIEGL VUX-1UAV scanner operates on the Time-of-Flight principle. It delivers 10 mm ranging accuracy at 1.2×10^6 measurements per second and records up to 15 target echoes per laser pulse. These hardware specifications provide high point density and support measurements through partially transparent structures such as vegetation.

3.3 Data Processing Pipeline

We introduce an automated pipeline to register raw LiDAR point clouds onto panoramic RGB frames. This pipeline is designed to address the scale, sparsity, and occlusion challenges inherent to large-scale outdoor data acquisition.

Geometric point cloud filtering. For each camera pose $\mathbf{t}_c \in \mathbb{R}^3$, we retrieve a local point cloud subset $\mathcal{P}$. We extract this subset using bounding-box indexing within a cropping radius R. We apply a 20.0 m height ceiling to remove aerial returns. Subsequently, we eliminate back-facing points using geometric constraints. We estimate surface normals $\mathbf{n}_i$ using Principal Component Analysis [14]. We enforce the following cosine similarity condition to isolate the front-facing subset $\mathcal{P}_{norm} \subseteq \mathcal{P}$.

$$\cos(\theta_n) = \frac{\mathbf{n}_i \cdot (\mathbf{t}_c - \mathbf{p}_i)}{\|\mathbf{n}_i\|_2 \, \|\mathbf{t}_c - \mathbf{p}_i\|_2} > \tau_n$$

Visibility-aware spherical projection. We resolve inter-point occlusions using spherical-inversion-based hidden-point removal [20]. We map each point onto a reference sphere of radius $r_{hpr} = 100\, D_{max}$ and retain only convex hull vertices, producing the visible point set $\mathcal{P}_{vis}$. We then transform $\mathcal{P}_{vis}$ into the camera coordinate frame via a rigid-body motion $\mathbf{T} \in \mathbb{SE}(3)$. We compute the metric depth ρ_i and spherical angles (θ_i, ϕ_i) for each transformed point. Finally, we project these coordinates onto the discrete $W \times H$ panoramic grid.

Adaptive depth rendering. We render continuous depth maps using an adaptive splatting strategy [21]. We accumulate depth values in back-to-front order following the classic painter's algorithm. Furthermore, we calculate a row-adaptive fill radius $r_f(v_i)$ to close discretization gaps near the ground plane.

$$r_f(v_i) = 1 + \mathbb{I}(v_i > 0.7\, H)$$

A subsequent morphological closing operation fills small residual voids left by sparse projection and produces dense 8K metric depth maps aligned with the input panorama.

3.4 Dataset Statistics

Depth distribution. PanoHK360 exhibits a right-skewed depth distribution concentrated in the near range, with an interquartile range of [7.3, 19.6] m (Figure 3a). The pixel-weighted and frame-equal distributions remain nearly identical (Figure 3b), indicating consistent depth characteristics across frames. The per-frame mean depth is 16.4 m (95% CI [16.4, 16.5] m, Figure 3e), and the within-frame depth spread averages 13.4 m (Figure 3f).

Valid-pixel coverage. On average, 25.7% of pixels per frame contain valid LiDAR depth (Figure 3c). We render each LiDAR return as exactly one pixel without interpolation, which preserves the original measurements and avoids interpolation artefacts at the cost of small gaps between neighbouring points. Users requiring dense supervision can apply standard depth completion.

Spatial distribution. The valid-pixel coverage is highly non-uniform along the vertical axis from sky to ground (Figure 3d). On the 8000×4000 equirectangular grid, the 35% and 65% cumulative valid-mass boundaries fall at rows $v = 1817$ and $v = 2203$. These boundaries partition the panorama into a sky-dominated upper region, a horizon band containing most road-level structures, and a lower region dominated by ground returns.

Scene and illumination diversity. The benchmark subset spans diverse structural categories, including high-density urban areas with high-rise facades, densely vegetated corridors, and complex traffic intersections, captured across multiple times of day and weather conditions ranging from bright daylight to nighttime artificial lighting. Detailed geographic coverage across Hong Kong Island, the Kowloon Peninsula, and the New Territories is provided in Appendix B.3.

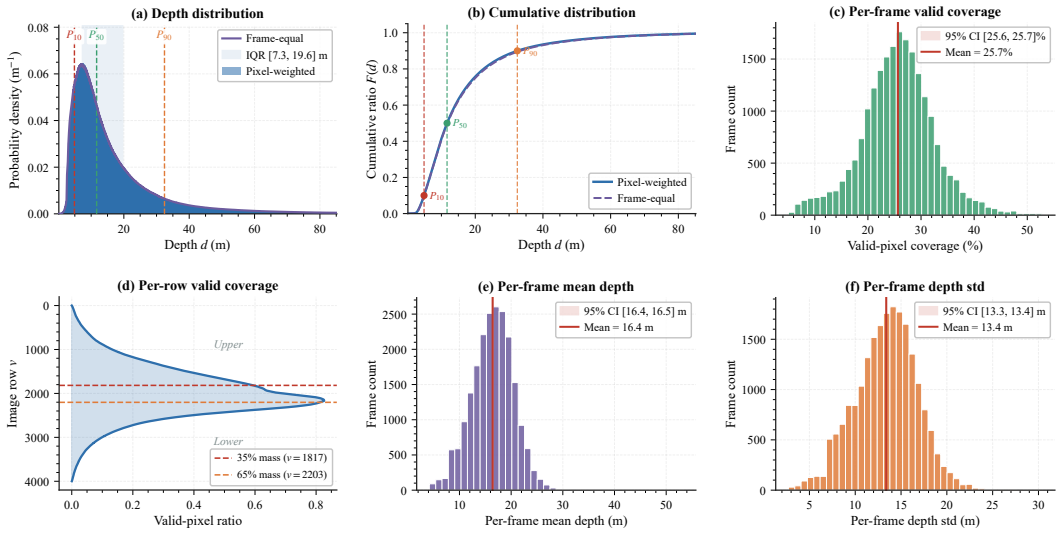


Figure 3: **Depth statistics of PanoHK360.** (a) Pixel-weighted and frame-equal depth PDFs with the x-axis truncated at P99.5. The distribution is right-skewed with IQR = [7.3, 19.6] m. (b) Corresponding CDFs marking the P10, P50, and P90 percentiles. (c) Per-frame valid-pixel coverage. The mean is 25.7% with a 95% CI of [25.6, 25.7]%. (d) Per-row valid-pixel ratio computed on the native 8000×4000 panoramic grid. We define $v = 0$ at the top and $v = 4000$ at the bottom. The data-driven boundaries $v = 1817$ and $v = 2203$ correspond to the 35% and 65% cumulative valid-mass quantiles, respectively. (e) Per-frame mean depth. The mean is 16.4 m with a 95% CI of [16.4, 16.5] m. (f) Per-frame depth standard deviation. The mean is 13.4 m with a 95% CI of [13.3, 13.4] m. This fully characterizes the within-frame depth spread.

3.5 Train / Validation Split

PanoHK360-Mini is a curated benchmark subset extracted from the full release of approximately 2 million frames. We manually select ten geographically distinct locations. These areas jointly cover high-density urban environments, vegetated corridors, and complex traffic intersections under diverse illumination conditions. To suppress temporal redundancy, we retain one frame out of every ten along each trajectory. A subsequent manual quality-filtering step removes frames with insufficient point cloud coverage. Full details of this curation procedure are provided in Appendix D.1.

The final benchmark comprises 22,751 panoramic frames, of which 21,352 are assigned to the training set and 1,399 to the validation set, yielding a 93.8%/6.2% split ratio. We adopt a *location-disjoint* partitioning strategy in which entire geographic locations are assigned to either split, rather than randomizing individual frames. This design prevents scene-level memorization and requires models to generalize to entirely unseen geographic layouts and illumination conditions. Following established conventions in panoramic depth estimation, the validation set also serves as the evaluation benchmark and is used for all quantitative comparisons in this work. The complete file lists are released as `train.txt` and `val.txt`.

4 Experiments

4.1 Experimental Setup

Baselines and training setup. We evaluate six representative monocular 360° depth estimation methods spanning diverse architectural paradigms: CNN-based multi-projection fusion (UniFuse [17],

BiFuse++ [34]), Transformer-based panoramic perception (PanoFormer [30]), efficient semantic- and distance-aware fusion (Elite360D [1]), cross-projection attention (Cross360 [15]), and foundation-model adaptation (PanDA [6]). All models are trained on the PanoHK360-Mini training split and evaluated on the held-out validation split. We resize input panoramas to 1024×512 and preserve the default optimizer, learning-rate schedule, and loss formulation of each original implementation. All experiments run on NVIDIA RTX 4090 GPUs.

Evaluation protocol and metrics. All metrics are computed only on pixels corresponding to the original un-interpolated LiDAR returns; regions densified by morphological closing (Section 3.3) are excluded. Following established conventions [10, 17], we report absolute relative error (AbsRel), squared relative error (SqRel), root mean squared error (RMSE), root mean squared log error (RMSElog), mean $\log_{10}$ error, and threshold accuracies $\delta_1, \delta_2, \delta_3$ at 1.25^k. Following [26], all evaluations are restricted to the $100\,\text{m}$ range.

4.2 Baseline Results

Table 2 reports quantitative results on the PanoHK360-Mini validation split. Elite360D achieves the best or tied-best scores on six of eight metrics, including all three threshold accuracies and the lowest RMSElog. PanDA attains the lowest AbsRel and RMSE, reflecting its strong absolute-scale calibration. Cross360 ties Elite360D on $\log_{10}$ and remains competitive on the threshold metrics. The three recent architectures consistently outperform the earlier baselines.

UniFuse and BiFuse++ are the weakest performers, with RMSE above $11.5\,\text{m}$ and δ_1 below 0.50. Both were originally tuned on indoor benchmarks, and their degradation on PanoHK360-Mini indicates that indoor-oriented networks do not transfer cleanly to outdoor urban panoramas. Across the leaderboard, no method attains $\delta_1 > 0.60$ and the lowest RMSE remains above $8\,\text{m}$, leaving substantial headroom for future architectural progress.

Table 2: **Quantitative results on the PanoHK360 validation split.** $\downarrow$ denotes lower is better. $\uparrow$ denotes higher is better. We explicitly highlight the best result in each column in **bold**.

Method	Error Metrics $\downarrow$					Accuracy Metrics $\uparrow$		
	AbsRel	SqRel	RMSE	RMSElog	$\log_{10}$	δ_1	δ_2	δ_3
UniFuse [17]	0.65	14.12	11.50	0.55	0.18	0.46	0.64	0.76
PanoFormer [30]	0.26	**3.03**	10.15	0.48	0.14	0.54	0.74	0.85
BiFuse++ [34]	0.46	5.19	12.72	0.73	0.22	0.30	0.56	0.74
PanDA [6]	**0.24**	3.53	**8.40**	0.43	0.14	0.54	0.78	0.88
Elite360D [1]	0.28	3.20	8.77	**0.39**	**0.12**	**0.58**	**0.81**	**0.90**
Cross360 [15]	0.32	3.53	8.79	0.40	**0.12**	0.55	0.79	0.89

4.3 Zero-Shot Generalization Results

Table 3 reports the zero-shot depth estimation results on the PanoHK360 validation split. DA360† [18] achieves the highest relative accuracy, with an AbsRel of 0.50 and a δ_1 of 52.2%. In contrast, DAP [26] attains lower absolute errors, with an RMSE of $15.45\,\text{m}$ and a SqRel of 18.02. This advantage stems from the metric-scale output formulation adopted in DAP. Meanwhile, DA^{2†} [23] underperforms both counterparts across nearly all evaluated metrics, which is likely due to the absence of targeted real-world panoramic training data.

Across all evaluated methods, the best δ_1 remains at only 52.2%, and all RMSE values exceed $15.0\,\text{m}$. These margins reflect the difficulty of zero-shot generalization on PanoHK360, where models must simultaneously handle wide metric depth ranges and pronounced equirectangular distortion. These results suggest that reliable outdoor panoramic depth estimation requires dataset-specific training or dedicated fine-tuning.

Table 3: **Quantitative evaluation of zero-shot depth estimation methods on the PanoHK360 validation split.** † denotes methods that produce scale-invariant rather than metric depth. All reported metrics for these methods are computed after median-scale alignment. For such methods, **RMSE** and **SqRel** are included for reference only and carry no strict physical unit interpretation under scale alignment. δ_k denotes the percentage of pixels satisfying $\max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < 1.25^k$.

Method	Error ↓					Accuracy ↑		
	AbsRel	SqRel	RMSE	RMSElog	$\log_{10}$	δ_1	δ_2	δ_3
DA^{2†} [23]	0.56	20.79	18.88	0.59	0.18	0.44	0.67	0.79
DA360† [18]	**0.50**	19.50	15.57	**0.56**	**0.16**	**0.52**	**0.73**	**0.83**
DAP [26]	0.56	**18.02**	**15.45**	0.58	0.18	0.45	0.69	0.81

4.4 Qualitative Comparison

Figure 4 presents qualitative depth estimation results on representative scenes from the PanoHK360 validation split. We compare three trained baselines, namely UniFuse, Elite360D, and PanDA, against the ground-truth LiDAR depth. UniFuse consistently produces overly smooth predictions and fails to recover fine structural boundaries, particularly around complex building edges and vegetation. Elite360D and PanDA yield sharper near-range boundaries but exhibit notable failure modes in the upper hemisphere, where the absence of valid LiDAR returns leads to unconstrained predictions over open sky. Depth bleeding at equirectangular stitching boundaries and degraded geometric consistency in distant backgrounds remain common failure modes across all evaluated methods. These persistent artifacts indicate that PanoHK360 poses a substantial challenge as an outdoor panoramic benchmark.

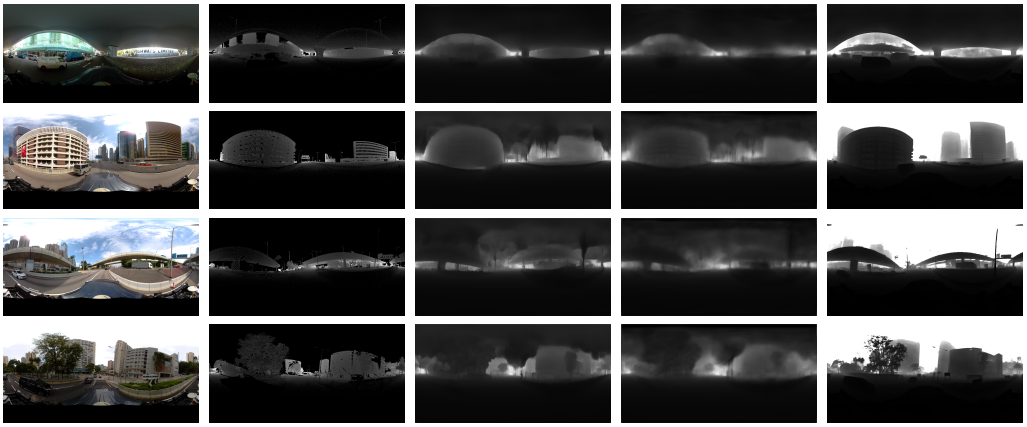


Figure 4: **Qualitative depth estimation results of trained baselines on the PanoHK360 validation split.** Each row corresponds to a distinct outdoor urban scene. The successive columns display the input equirectangular image, the ground-truth LiDAR depth (GT), and the predictions from UniFuse, Elite360D, and PanDA. Darker regions indicate farther distances, whereas brighter regions correspond to closer distances.

Figure 5 further presents qualitative depth estimation results on additional scenes from the same validation split. We evaluate three representative zero-shot baselines, namely DA^2 [23], DA360 [18], and DAP [26]. All three methods recover the coarse scene layout across diverse urban environments. However, DA^2 overestimates depth in sky and distant background regions, and consequently produces abrupt depth discontinuities at complex object boundaries. DA360 exhibits improved structural sharpness, although it introduces noticeable artifacts around thin structures such as tree canopies. DAP yields the most spatially consistent predictions but suffers from limited metric accuracy near the ground plane. These visual observations are consistent with the quantitative findings in Table 3 and reinforce the value of PanoHK360 as an outdoor panoramic benchmark.

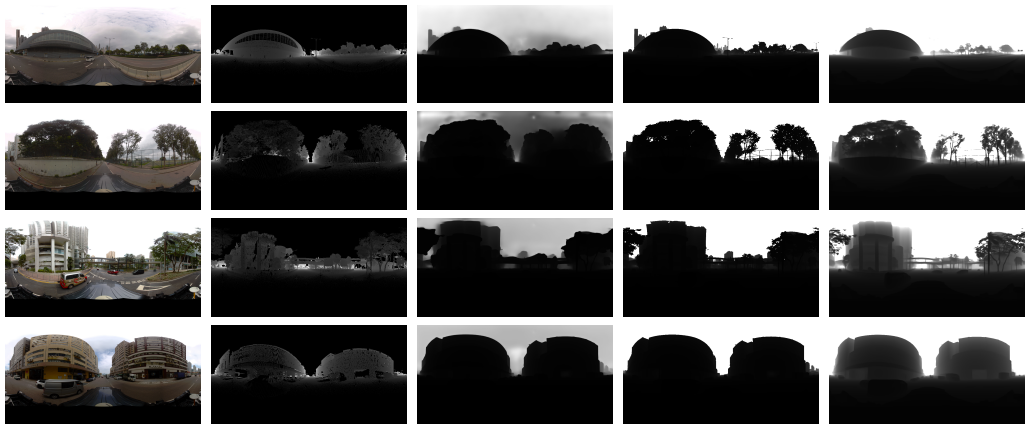


Figure 5: **Qualitative depth estimation results of zero-shot baselines on the PanoHK360 validation split.** Each row corresponds to a distinct outdoor urban scene. The successive columns display the input equirectangular image, the ground-truth LiDAR depth (GT), and the predictions from DA^2, DA360, and DAP. Warmer colors indicate closer distances, whereas cooler colors correspond to farther distances.

4.5 Challenges

Despite recent progress, outdoor metric depth prediction remains far from solved. A key bottleneck lies in the scarcity of large-scale, high-resolution panoramic datasets. Existing benchmarks share common structural limitations: they are predominantly built on indoor capture, synthetic rendering, or low-resolution imagery, with sparse LiDAR sampling or inaccurate pseudo-depth labels as the primary supervision source. Such a data deficit restricts real-world generalization. In this context, PanoHK360 introduces a higher level of scene diversity and geometric complexity, which exposes the vulnerability of current paradigms. Even recent foundation models exhibit pronounced performance degradation on PanoHK360.

The benchmark further reveals the gap between existing algorithms and real-world deployment. Outdoor panoramas exhibit pronounced projection distortion and wide depth ranges, and contain complex thin structures, highly reflective surfaces, and substantial illumination variations. Under these conditions, mainstream architectures struggle to produce metrically consistent predictions, particularly around fine structural boundaries and distant backgrounds. Moreover, owing to computational constraints, existing networks cannot process native 8K inputs and instead apply aggressive spatial downsampling, which discards a substantial portion of the high-frequency structural information critical for outdoor scene understanding. This limitation raises an open challenge: future architectures should be trained directly on full 8K panoramas in order to recover fine details such as thin wires, tree branches, and distant pedestrians.

5 Conclusion

We have introduced PanoHK360, a city-scale outdoor panoramic RGB-D dataset captured across diverse urban environments in Hong Kong. The dataset comprises approximately 2M equirectangular RGB-D pairs at 8K resolution, together with PanoHK360-Mini, a curated benchmark of 22,751 frames with location-disjoint training and validation splits. All depth annotations are derived from a survey-grade LiDAR scanner through an open-source automated point-cloud-to-panorama processing pipeline that we release alongside the dataset. The benchmark evaluation reveals that state-of-the-art monocular and zero-shot methods struggle with two principal challenges: the pronounced geometric distortion introduced by equirectangular projection and the wide metric depth ranges of urban scenes. These difficulties underscore the challenging nature of PanoHK360. Looking ahead, the high-fidelity geometry, dense metric annotations, and continuous survey trajectories of PanoHK360 lay a foundation for three promising research directions: panoramic depth completion, surface normal estimation, and outdoor 3D Gaussian Splatting.

References

[1] Hao Ai and Ling Wang. Elite360D: Towards efficient 360 depth estimation via semantic- and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9926–9935, 2024.

[2] Hao Ai, Zidong Cao, and Ling Wang. A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *arXiv preprint arXiv:2502.10444*, 2025.

[3] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3D: A holistic benchmark and a solid baseline for 360° depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3727–3737, 2021.

[4] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-Semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[5] 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 (CVPR)*, 2020.

[6] Zidong Cao, Jin Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, and Ling Wang. PanDA: Towards panoramic depth anything with unlabeled panoramas and Möbius spatial augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017.

[8] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2366–2374, 2014.

[11] Qi Feng, Hubert P. H. Shum, and Shigeo Morishima. 360 depth estimation in the wild - the Depth360 dataset and the SegFuse network. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 664–673, 2022.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. doi: 10.1109/TPAMI.2012.213.

[14] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Surface reconstruction from unorganized points. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 71–78, 1992.

[15] Kun Huang et al. Cross360: 360° monocular depth estimation via cross projections across scales. *arXiv preprint arXiv:2601.17271*, 2026.

[16] Zeeshan Javed and Gon-Woo Kim. PanoVILD: A challenging panoramic vision, inertial and LiDAR dataset for simultaneous localization and mapping. *The Journal of Supercomputing*, 78 (6):8247–8267, 2022.

[17] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. UniFuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 6 (2):1519–1526, 2021.

[18] Hualie Jiang, Ziyang Song, Zhiqiang Lou, Rui Xu, and Minglang Tan. Depth anything in 360°: Towards scale invariance in the wild. *arXiv preprint arXiv:2512.22819*, 2025.

[19] Hideaki Kanayama, Mahdi Chamseddine, Suresh Guttikonda, So Okumura, Soichiro Yokota, Didier Stricker, and Jason Raphael Rambach. ToF-360 – a panoramic time-of-flight RGB-D dataset for single capture indoor semantic 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

[20] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3):24–es, 2007.

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023.

[22] Geunu Kim, Daeho Kim, Jaeyun Jang, and Hyoseok Hwang. PAIR360: A paired dataset of high-resolution 360° panoramic images and LiDAR scans. *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[23] Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. DA^2: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025.

[24] Ming Li, Xueqian Jin, Xuejiao Hu, Jingzhao Dai, Sidan Du, and Yang Li. MODE: Multi-view omnidirectional depth estimation with 360° cameras. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 197–213, 2022.

[25] 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 (TPAMI)*, 45(3):3292–3310, 2022.

[26] Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. *arXiv preprint arXiv:2512.16913*, 2025.

[27] Roberto Martín-Martín, Hamid Rezatofighi, Abhijeet Shenoi, Mihir Patel, JunYoung Gwak, Nathan Dass, Amir Sadeghian, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[28] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[29] Ahmed Rida Sekkat, Yoann Dupuis, Pascal Vasseur, and Paul Honeine. OmniScape: Efficient omnidirectional semantic segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[30] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[32] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni,

Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. BiFuse++: Self-supervised and efficient bi-projection fusion for 360° depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5448–5460, 2023. doi: 10.1109/TPAMI.2022.3203516.

[35] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[36] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018.

[37] Mehdi Zayene, Jannik Endres, Albias Havolli, Charles Corbière, Salim Cherkaoui, Alexandre Kontouli, and Alexandre Alahi. HELVIPAD: A real-world dataset for omnidirectional stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[38] Fan Zhang, Junsong Zhao, Yiwen Zhang, and Stefanie Zollmann. A survey on 360° images and videos in mixed reality: Algorithms and applications. *Journal of Computer Science and Technology*, 38(3):473–491, 2023.

[39] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

A Overview

This supplementary material is organized as follows.

- **Sec.** B details the data acquisition platform, multi-sensor synchronization, and geographic coverage across Hong Kong.
- **Sec.** C describes the automated pipeline that converts raw LiDAR point clouds into registered equirectangular depth maps.
- **Sec.** D presents the three-stage curation procedure of PanoHK360-Mini together with the directory layout of the released data.
- **Sec.** E provides a cross-resolution evaluation using Guided Filter upsampling, demonstrating that 8K ground truth serves as a fine-grained evaluation instrument beyond its role as training data.
- **Sec.** F discusses the limitations of the current release.
- **Sec.** G outlines future research directions enabled by PanoHK360.

Finally, Figure 9 presents additional qualitative depth comparisons on challenging corner cases drawn from the validation split.

B Data Acquisition and Hardware Setup

B.1 Data Collection Hardware

The data acquisition platform of PanoHK360 integrates two complementary sensors. We deploy a Teledyne FLIR Ladybug5+ spherical camera (Richmond, BC, Canada) and a RIEGL VUX-1UAV survey-grade LiDAR scanner (Horn, Austria). The two sensors are rigidly mounted on a vehicle-borne rig and synchronized through a shared GPS time reference. This setup establishes the geometric and temporal basis for the panoramic RGB-D pairs in PanoHK360. An overview of the hardware is shown in Figure 6.

B.1.1 Panoramic Camera: Teledyne FLIR Ladybug5+

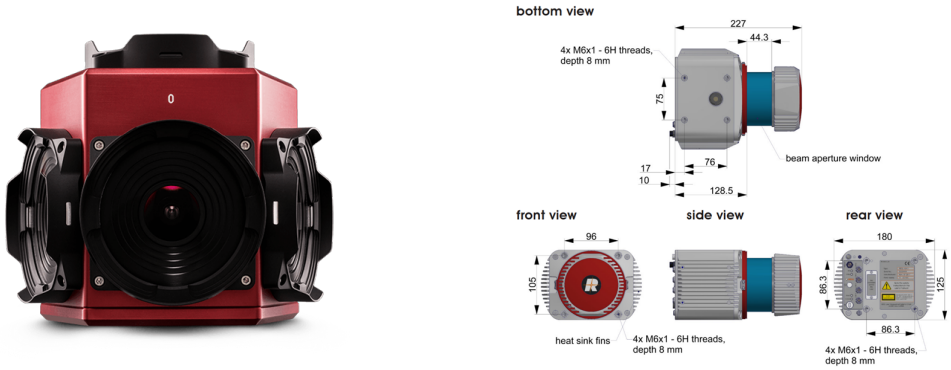
The panoramic imagery is captured by the Teledyne FLIR Ladybug5+, an industrial-grade 360° spherical imaging system. It comprises six global shutter CMOS sensors that produce a combined raw resolution of approximately 30 MP. The system covers close to 90% of the full sphere after stitching. The fixed lenses operate with factory-calibrated intrinsic and extrinsic parameters, supporting stable equirectangular reprojection across long acquisition trajectories. The camera supports hardware-level synchronization through PPS and NMEA GPS inputs. This input aligns image acquisition to the same temporal reference as the LiDAR scanner. In our deployment, the system operates at an effective 8K spherical resolution.

B.1.2 LiDAR Scanner: RIEGL VUX-1UAV

The metric depth supervision in PanoHK360 is grounded in measurements collected by the RIEGL VUX-1UAV, a survey-grade Time-of-Flight LiDAR scanner. The scanner provides a ranging accuracy of 10 mm and a precision of 5 mm. A high-speed rotating mirror produces a 360° field of view with regular and parallel scan lines. This configuration yields a uniform angular sampling pattern well suited to projection-based fusion with panoramic imagery. During acquisition, the pulse repetition rate is configured to deliver up to 1.2×10^6 measurements per second. This rate provides the point density needed to populate dense 8000×4000 depth maps after spherical projection. The scanner further records up to 15 target echoes per laser pulse, which improves the recovery of geometry behind partially transparent occluders such as foliage.

B.2 Hardware Synchronization

Reliable RGB-D supervision in PanoHK360 requires that every panoramic frame and every LiDAR point be expressed in a common spatio-temporal reference. We therefore align the two sensors at the hardware level, both in time and in space, before any data are released.



(a) Teledyne FLIR Ladybug5+ panoramic camera. (b) RIEGL VUX-1UAV survey-grade LiDAR scanner.

Figure 6: Data acquisition platform used in PanoHK360. (a) The Teledyne FLIR Ladybug5+ provides 360° spherical panoramic imagery at 8K resolution with hardware-level GPS/PPS synchronization. (b) The RIEGL VUX-1UAV provides survey-grade LiDAR point clouds at up to 1.2×10^6 measurements per second with $10\,\text{mm}$ ranging accuracy.

In the temporal domain, the Ladybug5+ camera and the RIEGL VUX-1UAV scanner are driven by a shared GPS time source. Synchronization is delivered through a Pulse-Per-Second (PPS) signal and an NMEA time string. The PPS edge defines a common clock tick, and each image exposure and LiDAR pulse is timestamped against this tick. As a result, every panoramic frame can be paired with the LiDAR returns acquired in the same time window. The synchronization residuals are reduced to the sub-millisecond regime. This temporal alignment is essential for vehicle-borne acquisition. At typical urban driving speeds, even small clock drifts would translate into pixel-level reprojection errors on the equirectangular grid.

In the spatial domain, the rigid mounting between the camera and the LiDAR is described by a fixed extrinsic transformation $\mathbf{T}_{C \leftarrow L} \in \mathbb{SE}(3)$. The transformation is initialized from the mechanical mount geometry. It is then refined through a target-based calibration procedure that minimizes the reprojection residual between known LiDAR-observed structures and their corresponding image features. The vehicle pose is further provided in a global geodetic frame by an integrated GNSS-INS unit. This unit allows panoramic frames captured at different times and along different trajectories to share a consistent world coordinate system. The combination of synchronized timestamps and calibrated extrinsics ensures geometric consistency across the entire dataset. This consistency applies to the projection from raw point clouds to panoramic depth maps in the processing pipeline (Section 3.3).

B.3 Geographic Coverage

PanoHK360 spans the three major regions of Hong Kong: Hong Kong Island, the Kowloon Peninsula, and the New Territories. The data were collected from multiple public areas across these regions. Coverage includes high-density commercial districts, residential neighbourhoods, arterial roads, expressways, and tunnel corridors. The complete acquisition trajectories for all three regions are illustrated in Figure 7.

Hong Kong Island provides dense vertical urban environments. Narrow streets are flanked by tall building facades that introduce strong occlusion, complex inter-reflections, and abrupt depth discontinuities at building boundaries. The Kowloon Peninsula contributes mid-rise mixed-use districts. These districts combine commercial corridors, transit hubs, and elevated infrastructure, producing frequent transitions between open intersections and partially enclosed scenes. The New Territories extend the coverage to suburban and semi-rural settings, including arterial roads, residential blocks, and densely vegetated corridors. These settings broaden the depth distribution toward longer ranges and introduce natural geometry such as trees and undulating terrain.

To further enrich the photometric diversity, acquisition is carried out at multiple times of day and under a range of weather conditions. The resulting trajectories therefore cover a wide spectrum of illumination regimes. These range from bright natural daylight to nighttime scenes dominated

by artificial lighting. The trajectories also include overcast and post-rain conditions with reduced contrast and pronounced specular reflections on wet road surfaces. This geographic and photometric variability exposes models to the breadth of conditions encountered in real-world urban deployment, rather than to any single canonical scene type.



Figure 7: Geographic coverage of PanoHK360 across Hong Kong. Acquisition trajectories span three major administrative regions: Hong Kong Island, the Kowloon Peninsula, and the New Territories, covering numerous distinct urban locations with diverse structural and illumination conditions.

C Automated Data Processing Pipeline

To generate panoramic RGB-D images from raw LiDAR point clouds in large-scale outdoor environments, we introduce an automated processing pipeline that projects each point cloud onto the equirectangular image plane and resolves inter-point occlusions via the painter's algorithm, a depth-sorting strategy for back-to-front rasterisation. Following the organization of Section 3.3, the pipeline is detailed below using the same component structure, with each component addressing a distinct challenge in converting raw point clouds into registered panoramic depth maps.

Geometric point cloud filtering. We first extract a local point cloud subset by means of spatial indexing. Directly loading large unordered point cloud files incurs substantial I/O overhead and risks memory overflow. We therefore construct a 3D axis-aligned bounding box index by parsing the header of each `.las` file. Given the translation vector $\mathbf{t}_c = [x_c, y_c, z_c]^T \in \mathbb{R}^3$ of the panoramic camera pose and a cropping radius R, the spatial bounds for local extraction are defined as

$$\mathcal{B} = [x_c - R,\ x_c + R] \times [y_c - R,\ y_c + R] \times [z_c - R,\ z_c + H_{\max}],$$

where $H_{\max} = 20.0\,\text{m}$ eliminates points above the effective scene height. The extracted local point cloud is then defined as

$$\mathcal{P} = \left\{ \mathbf{p}_i \in \mathbb{R}^3 \;\middle|\; \|\mathbf{p}_i - \mathbf{t}_c\|_2 \le R \ \wedge\ \mathbf{p}_i \in \mathcal{B} \right\}.$$

We then apply a normal-based filter that removes points whose surface normals are oriented away from the sensor viewpoint. For each point $\mathbf{p}_i \in \mathcal{P}$, a local surface normal $\mathbf{n}_i$ is estimated from its

k-nearest neighbourhood by means of Principal Component Analysis (PCA). Since visible surface points are oriented toward the sensor, we retain only those points whose normals are aligned with the line-of-sight direction from the point to the camera centre $\mathbf{t}_c$. Concretely, we compute the directional cosine

$$\cos(\theta_n) = \frac{\mathbf{n}_i \cdot (\mathbf{t}_c - \mathbf{p}_i)}{\|\mathbf{n}_i\|_2 \, \|\mathbf{t}_c - \mathbf{p}_i\|_2},$$

and retain only the points that satisfy $\cos(\theta_n) > \tau_n$, where τ_n is a predefined threshold. This criterion effectively removes back-facing surfaces and outliers from occluded regions. The surviving subset is denoted by $\mathcal{P}_{norm} \subseteq \mathcal{P}$.

Visibility-aware spherical projection. We perform viewpoint-dependent hidden point removal on $\mathcal{P}_{norm}$. In real urban scenes, direct spherical projection of $\mathcal{P}_{norm}$ maps occluded background points (e.g., those lying behind buildings) onto the panorama. To address this occlusion issue, we apply spherical inversion centred at $\mathbf{t}_c$, which maps each point onto a sphere of radius $r_{hpr} = 100\, D_{max}$, where $D_{max} = \sup_{\mathbf{p} \in \mathcal{P}_{norm}} \|\mathbf{p} - \mathbf{t}_c\|_2$. The convex hull of the inverted point set is then computed, and only its vertices are retained as the set of geometrically visible points, $\mathcal{P}_{vis} \subseteq \mathcal{P}_{norm}$.

The visible point set $\mathcal{P}_{vis}$ is then transformed into the local camera coordinate system and projected onto the equirectangular image plane. The world-to-camera transformation is represented as a rigid body motion $\mathbf{T} \in \mathbb{SE}(3)$. Given the viewing direction $\mathbf{F}$ and the up direction $\mathbf{U}$, the rotation matrix $\mathbf{R}_c \in \mathbb{SO}(3)$ is constructed as $\mathbf{R}_c = [\mathbf{r},\, \mathbf{F},\, \mathbf{U}']^T$, where $\mathbf{r} = (\mathbf{F} \times \mathbf{U})/\|\mathbf{F} \times \mathbf{U}\|_2$ and $\mathbf{U}' = \mathbf{r} \times \mathbf{F}$. Each point $\mathbf{p}_i \in \mathcal{P}_{vis}$ is subsequently mapped to its local coordinate

$$\mathbf{p}'_i = \mathbf{R}_c \, (\mathbf{p}_i - \mathbf{t}_c), \qquad \mathbf{p}'_i = [x'_i,\, y'_i,\, z'_i]^T.$$

The metric depth is then given by $\rho_i = \|\mathbf{p}'_i\|_2$. To obtain the corresponding spherical angles, we apply

$$\begin{bmatrix} \theta_i \\ \phi_i \end{bmatrix} = \begin{bmatrix} \arctan2(x'_i,\, y'_i) \\ \arcsin(z'_i/\rho_i) \end{bmatrix},$$

where $\theta_i \in [-\pi, \pi]$ denotes the yaw angle and $\phi_i \in [-\pi/2, \pi/2]$ denotes the pitch angle. These angles are subsequently mapped to discrete pixel indices (u_i, v_i) on the $W \times H$ panoramic image plane via

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \left\lfloor \begin{bmatrix} \big(\frac{\theta_i}{2\pi} + 0.5\big)(W-1) \\ \big(0.5 - \frac{\phi_i}{\pi}\big)(H-1) \end{bmatrix} \right\rfloor.$$

Adaptive depth rendering. We render the depth map by means of adaptive splatting with depth sorting. To ensure that foreground points correctly occlude background points and to fill holes caused by point cloud discretisation, we apply the painter's algorithm during rasterisation: the projected points are sorted in descending order of ρ_i and painted back to front. The fill radius r_f is defined as a function of the vertical pixel index v_i, so as to close fine gaps near the ground plane in large-scale outdoor scenes:

$$r_f(v_i) = 1 + \mathbb{I}(v_i > 0.7\, H).$$

A morphological closing operation is then applied to smooth the remaining gaps, yielding the registered RGB panorama together with the corresponding 16-bit depth map.

D Dataset Composition and Structure

D.1 Curation Details for PanoHK360-Mini

This appendix describes the three-stage curation procedure used to construct PanoHK360-Mini from the full PanoHK360 release, together with the per-location breakdown of the train/validation split.

Location selection. From the full acquisition trajectories spanning Hong Kong Island, the Kowloon Peninsula, and the New Territories, we manually select ten geographically distinct locations chosen to jointly cover the three structural categories defined in Section 3.4 (high-density urban areas with enclosing high-rise facades, vegetated regions, and complex traffic intersections) under varied illumination conditions (clear, overcast, and nighttime). This curation prioritises scene distinctiveness over raw frame count, ensuring that PanoHK360-Mini captures the full structural and photometric diversity of the full release while remaining tractable for benchmark-scale training and evaluation.

Temporal subsampling. Vehicle-mounted panoramic acquisition produces highly redundant consecutive frames, with adjacent panoramas typically separated by less than one metre of camera motion. To suppress this temporal correlation and prevent inflated benchmark scores caused by near-duplicate samples leaking across train and validation, we retain one frame out of every ten along each trajectory, yielding a tenfold reduction in temporal redundancy while preserving full trajectory coverage.

Quality filtering. Frames produced by the LiDAR-to-panorama rendering pipeline (Section 3.3) occasionally exhibit insufficient point cloud coverage, for example in regions where the LiDAR field of view is partially obstructed or where the local point density is too sparse to support reliable hidden-point removal. We manually inspect the temporally subsampled candidate pool and discard frames whose depth maps exhibit large unannotated regions or visible geometric inconsistency, retaining only those whose valid-pixel coverage and lower-hemisphere completeness are consistent with the dataset-wide statistics reported in Section 3.4.

Per-location breakdown. The ten selected locations are distributed across all three administrative regions of Hong Kong, with multiple locations sampled from each region to ensure intra-region diversity. Each location is assigned exclusively to either the training or validation split under the location-disjoint partitioning described in Section 3.5, with eight locations contributing to training and the remaining two reserved for validation. Across the ten locations, the structural and illumination categories are jointly balanced to avoid systematic correlation between scene type and split assignment.

D.2 Dataset Structure and File Format

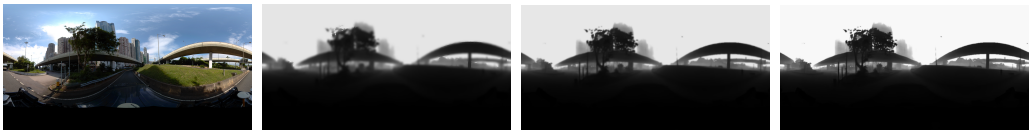
The PanoHK360 dataset is organised into a unified directory hierarchy with parallel `train/` and `val/` splits, each containing ten subdirectories corresponding to distinct data modalities. Raw panoramic RGB images are stored as JPEG files in `pano_raw/`, while metric depth maps are provided in three variants, each available in both `.npy` and 16-bit `.png` formats encoding millimetre-precision depth values: `depth_raw/` contains the direct LiDAR projections, `depth_filtered/` provides occlusion-filtered depth maps produced by the Hidden-Point Removal (HPR) stage of the processing pipeline, and `depth_completed/` stores dense depth maps generated by depth completion. Surface normal maps derived from the ground-truth depth are stored in `normal_map/`. Camera intrinsics, extrinsics, and spatio-temporal calibration parameters are provided in `poses/` as CSV files, with corresponding GNSS/IMU trajectories and per-frame timestamps in `spatial_temporal/` likewise in CSV format. Cropped regions of the panoramic images are stored in `pano_roi/`, while the corresponding point cloud segments used for depth projection are available in `pcd_roi/`. Raw LiDAR point cloud files are retained in `lidar_raw/` for reproducibility of the full processing pipeline. The complete directory structure is as follows:

```
PanoHK360/
├── train/
│   ├── pano_raw/ ..... Original panoramic RGB images (.jpg)
│   ├── depth_raw/ ..... Raw depth maps from LiDAR projection (.npy, .png)
│   ├── depth_filtered/ ..... Occlusion-filtered depth maps via HPR (.npy, .png)
│   ├── depth_completed/ ..... Dense depth maps via depth completion (.npy, .png)
│   ├── normal_map/ ..... Surface normal maps from ground-truth depth (.png)
│   ├── poses/ ..... Camera intrinsics, extrinsics & calibration (.json)
│   ├── spatial_temporal/ ..... GNSS/IMU trajectories & per-frame timestamps (.csv)
│   ├── pano_roi/ ..... Cropped regions of panoramic images (.jpg)
│   ├── pcd_roi/ ..... Point cloud segments for depth projection (.las)
│   └── lidar_raw/ ..... Raw LiDAR point cloud files (.las)
├── val/
│   └── ..... Identical structure to train/
├── train.txt ..... Training split file list
└── val.txt ..... Validation split file list
```

E Cross-Resolution Depth Evaluation via Guided Filtering

To demonstrate that the 8K ground truth of PanoHK360 provides a fine-grained evaluation scale for boundary recovery, we adopt a training-free Guided Filter [13] to upsample native 1024×512 depth predictions to 2K, 4K, and 8K, guided by the corresponding high-resolution RGB panoramas.

As shown in Figure 8, the Guided Filter progressively transfers RGB edge priors to higher resolutions. Thin structures and depth discontinuities become increasingly well aligned from 2K to 8K, confirming that the native 8K ground truth serves as a discriminative evaluation instrument that exposes per-method boundary recovery capacity otherwise invisible at standard low resolutions.



RGB Input **2K (GF)** **4K (GF)** **8K (GF)**

Figure 8: **Cross-resolution visual comparison of Guided Filter upsampling.** Starting from a native 1024×512 depth prediction, the Guided Filter [13] progressively transfers edge information from the high-resolution RGB panorama (leftmost column) to 2K, 4K, and 8K. Boundary alignment improves with resolution, confirming that the native 8K ground truth of PanoHK360 provides a fine-grained scale for evaluating per-method boundary recovery capacity.

F Intended Uses and Limitations

F.1 Intended Uses

PanoHK360 is primarily intended for training and evaluation of monocular 360° depth estimation models in real outdoor environments. The 8K resolution and survey-grade metric ground truth additionally support panoramic depth completion, surface normal estimation, and outdoor 3D Gaussian Splatting. These are tasks for which large-scale outdoor panoramic data are currently scarce.

F.2 Limitations

All data are collected within Hong Kong, a high-density sub-tropical metropolis. Models trained on PanoHK360 may require domain adaptation for deployment in cities with different building typologies or vegetation patterns. The LiDAR scanner produces no returns for the open sky hemisphere, leaving approximately 74.3% of pixels per frame without valid depth annotation. Evaluation masks and loss functions should account for this inherent sparsity. Additional limitations include temporal misalignment around dynamic objects and the absence of semantic annotations.

G Future Directions

The high-fidelity geometry, dense metric annotations, and continuous survey trajectories of PanoHK360 establish a foundation for a focused set of visually grounded research directions. First, the sparse LiDAR projections and completed depth products can support panoramic depth completion under real outdoor geometry. Second, surface normal maps derived from the metric LiDAR depth provide direct supervision for normal estimation on challenging structures such as building facades, road surfaces, and vegetation. Third, the aligned RGB-D panoramas and survey trajectories provide a testbed for outdoor 3D Gaussian Splatting under equirectangular imagery and metric scene geometry.

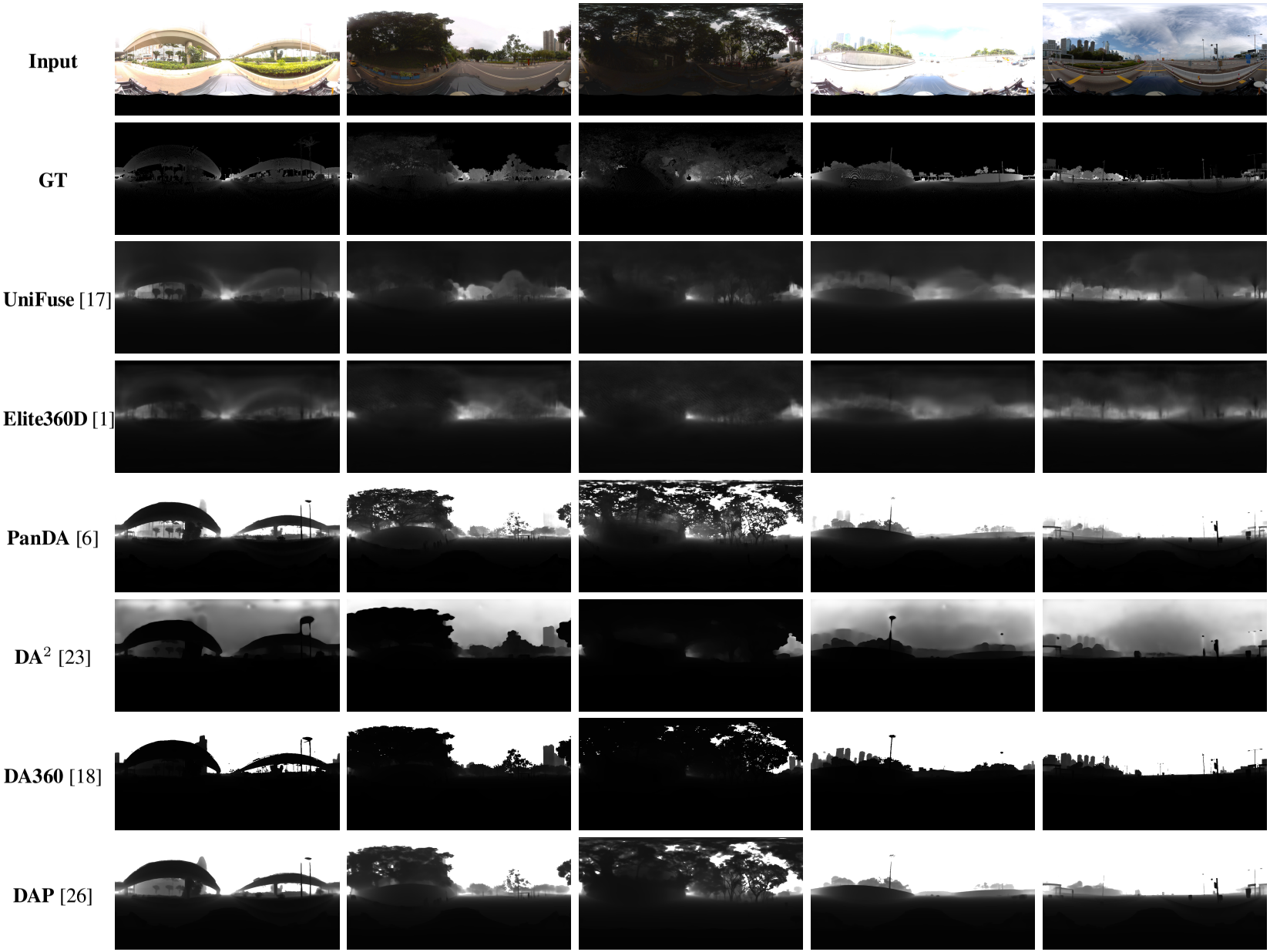


Figure 9: **Qualitative depth estimation results on challenging corner cases from the PanoHK360 validation split.** We deliberately select five atypical scenes that stress-test the evaluated methods, including frames with strong overexposure, dense foliage occlusion, and largely empty open spaces with sparse structural cues. Each column corresponds to a distinct outdoor urban scene, and each row corresponds to one method. From top to bottom, the rows show the input equirectangular image, the ground-truth LiDAR depth (GT), three trained baselines (UniFuse, Elite360D, PanDA), and three zero-shot baselines (DA^2, DA360, DAP). Darker regions indicate farther distances, whereas brighter regions correspond to closer distances.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 accurately reflect the paper's three main contributions: (i) a 2M-frame, 8K-resolution outdoor panoramic RGB-D dataset with survey-grade metric LiDAR annotations (Section 3), (ii) an open-source automated point-cloud-to-panorama processing pipeline, and (iii) baseline evaluations with representative monocular 360° depth estimation methods (Section 4). All claims are substantiated by the dataset statistics, pipeline description, and experimental results presented in the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix F.2 and Section 5, including: (i) geographic coverage restricted to Hong Kong, which may limit generalization to cities with different building typologies or vegetation patterns; (ii) approximately 74.3% of pixels per frame lack valid LiDAR depth due to sky regions, elevated structures, and the single-pixel back-projection scheme (Section 3.4); and (iii) temporal misalignment around dynamic objects and the absence of semantic segmentation annotations, identified as limitations of the current release.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and complete proofs?

   Answer: [N/A]

   Justification: This is a dataset paper; no theoretical proofs are presented.

4. **Experimental reproducibility**

   Question: Does the paper fully describe the experimental setup used to validate its findings?

   Answer: [Yes]

   Justification: The Experimental Setup subsection of Section 4 fully specifies the six evaluated architectures (UniFuse, BiFuse++, PanoFormer, Elite360D, Cross360, PanDA), training and validation splits (`train.txt` / `val.txt`), input resolution (1024×512), evaluation depth range $[0, 100]$ m, and all reported metrics (AbsRel, SqRel, RMSE, RMSElog, $\log_{10}$, δ_1–δ_3). The automated LiDAR-to-panorama processing pipeline is described in the Data Processing Pipeline subsection of Section 3.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Answer: [Yes]

   Justification: Dataset and processing code will be released upon publication.

6. **Experimental statistical significance**

   Question: Does the paper report error bars suitably and correct?

   Answer: [Yes]

   Justification: Dataset statistics including valid-pixel coverage and per-frame mean depth are reported with 95% confidence intervals in Section 3.4. Baseline results in Tables 2 and 3 are deterministic evaluations on fixed validation and test splits, consistent with standard practice in dataset benchmark papers.

7. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce?

Answer: [Yes]

Justification: All baseline models in Section 4 were trained and evaluated on NVIDIA RTX 4090 GPUs, with input panoramas resized to 1024×512. We do not modify the network architecture or default hyperparameters of any baseline; all training configurations follow the original implementations.

8. **Code of ethics**

Question: Does the research conform with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: All data were collected from public roads and publicly accessible urban spaces in Hong Kong, consistent with established practices in autonomous driving datasets (*e.g.*, KITTI [12], nuScenes [5]). To protect personal privacy, all human faces and vehicle license plates appearing in the released panoramas have been automatically detected and blurred prior to release; no personally identifiable information is intentionally retained. In addition, although the dataset only covers public roads and publicly accessible areas, we provide a takedown mechanism for any content inadvertently captured from private premises. A public contact channel will be released upon publication, through which individuals or property owners may request the removal or further anonymisation of specific frames or regions; such requests will be processed in subsequent dataset releases. The dataset collection and release procedure is consistent with the NeurIPS Code of Ethics.

9. **Broader impacts**

Question: Does the paper discuss both potential positive and negative societal impacts?

Answer: [Yes]

Justification: Positive societal impacts include improved geometric perception from high-resolution panoramic RGB-D data, as discussed in Section 5. A potential negative impact is the misuse of high-fidelity urban geometry for unauthorized surveillance; however, the dataset contains no identity-linked information and covers only publicly accessible road environments.

10. **Safeguards**

Question: Does the paper describe safeguards for responsible data release?

Answer: [Yes]

Justification: The dataset will be released under a license restricting use to non-commercial research purposes. As the data consist solely of public-road panoramas and LiDAR point clouds without personally identifiable attributes, no additional anonymization pipeline beyond standard practice is required.

11. **Licenses for existing assets**

Question: Are existing assets properly credited with licenses?

Answer: [Yes]

Justification: All baseline methods referenced in Section 2 and Section 4 (UniFuse [17], BiFuse++ [34], PanoFormer [30], Elite360D [1], Cross360 [15], PanDA [6], DA360 [18], DAP [26], Depth Anything V2 [35]) are cited with their original publications. No proprietary datasets or third-party code are incorporated into the PanoHK360 processing pipeline.

12. **New assets**

Question: Are new assets well documented?

Answer: [Yes]

Justification: PanoHK360 is fully documented in Section 3, covering data collection hardware, the automated processing pipeline, dataset statistics (Section 3.4), and the location-disjoint train/val split protocol (Section 3.5). The dataset will be released with README documentation, split files (`train.txt`, `val.txt`), and pipeline source code upon publication.

13. **Crowdsourcing and human subjects**

Question: Does the paper include full details of any crowdsourcing or human subjects study?

Answer: [N/A]

Justification: No crowdsourcing or human subjects involved.

14. **IRB approvals**

Question: Does the paper describe IRB approvals for research with human subjects?

Answer: [N/A]

Justification: No human subjects research.

15. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important component of the core methods?

Answer: [N/A]

Justification: LLMs were not used in core methodology.