Depth Any Panoramas: A Foundation Model for Panoramic Depth Estimation

This supplementary material provides expanded technical discussions, extended analyses, and additional visual results that further support the contributions presented in the main paper. The supplementary material is organized into the following nine sections:

1. Simulated Data Processing
2. Implementation Details
3. Analysis of Range Mask
4. Jitter Verification
5. Details and Analysis of the Discriminator
6. Different Scale Model
7. Analysis of Dense Fidelity Loss
8. More Qualitative Results
9. Future Work

1. Simulated Data Processing

To generate high-quality outdoor training data, we rely on our UE5-based AirSim360 simulation pipeline, which provides photorealistic rendering, precise geometric control, and a fully automated data acquisition framework. During rendering, the engine records physically accurate per-pixel depth by writing the Z-buffer values at the pre-render stage into the SceneDepth channel. These depth values are further exported as metric depth maps directly from the render target (RT) buffer, encoded in the alpha channel of 4-channel RGBA outputs to ensure bit-exact alignment with RGB panoramas.

To maintain consistency across omnidirectional viewpoints, we equip the virtual drone with a multi-view normal camera rig consisting of six synchronized perspective cameras. This rig captures the full $360^\circ \times 180^\circ$ field of view with unified photometric conditions. Since all cameras share the same illumination and rendering context, the collected panoramas exhibit consistent color, exposure, and shadow behavior—critical for stable training of large metric-depth models.

A key advantage of AirSim360 is its tight integration with Unreal Engine 5's global illumination system (Lumen). We dynamically adjust directional light intensity, sky atmosphere, volumetric fog, and GI parameters to mimic natural lighting transitions across time of day and weather conditions. These adjustments improve realism while preserving depth-lighting consistency. In addition, the built-in flight-control API enables fully automated camera trajectories. We script mid- and low-altitude drone flights that sweep through urban, suburban, and vegetation-rich environments, generating panoramas with diverse viewpoints, distances, and scene geometries.

The final rendered dataset includes pixel-aligned RGB–depth pairs, precise camera poses, and metadata for each panoramic frame. This simulated corpus serves as a high-quality, noise-free outdoor supervision signal used in Stage-1 training of the Scene-Invariant Labeler, providing strong geometric priors that generalize beyond synthetic scenes.

2. Implementation Details

All experiments are conducted on eight NVIDIA H20 GPUs (96GB memory each). We use PyTorch 2.3 with CUDA 12 and adopt Distributed Data Parallel (DDP) for all multi-GPU training. To improve domain balance, each batch samples indoor and outdoor panoramas with a ratio of 1:4.

Scene-Invariant Labeler (Stage 1). The Scene-Invariant Labeler is trained for 5 epochs, and we use AdamW with a backbone learning rate of 5×10^{-6} and a decoder learning rate of 5×10^{-5}, weight decay of 0.05, and $(\beta_1, \beta_2) = (0.9, 0.999)$. The global batch size is 32. Random horizontal flipping, color jittering, and $\pm 10\%$ horizontal shifts are applied as data augmentation.

Depth Quality Discriminator (Stage 2-A). The PatchGAN-based discriminator is trained for approximately 60 epochs. It takes the concatenation of RGB and normalized predicted depth as input. We use hinge adversarial loss with Adam optimizer and a learning rate of 1×10^{-4}. The global batch size is 64.

Realism-Invariant Labeler (Stage 2-B). After filtering the top 300k indoor and 300k outdoor pseudo-labeled samples using the trained discriminator, the Realism-Invariant Labeler is trained for 4 epochs. The optimizer, learning rate schedule, and augmentation strategy follow Stage 1, but with a global batch size of 64 to better utilize the large pseudo-labeled dataset.

DAP Foundation Model (Stage 3). For the final DAP training, we initialize the model with the Realism-Invariant Labeler weights and train for 2 epochs on all labeled data and the 1.9M refined pseudo-labeled panoramas.

We use AdamW with backbone and decoder learning rates of 5×10^{-6} and 5×10^{-5}, respectively, and a global batch size of 80. All loss components, including the geometry-centric losses (L_{normal}, L_{pts}) and sharpness-centric losses (L_{DF}, L_{grad}) are enabled throughout training. The loss weights follow the main paper with $\lambda_1 \dots \lambda_6 = \{1.0, 0.4, 5.0, 2.0, 2.0, 2.0\}$. The plug-and-play range mask head are jointly optimized.

Evaluation Metrics. AbsRel and RMSE are given by:

$$\text{AbsRel} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{|\hat{D}_p^{\text{med}} - D_p^{\star}|}{D_p^{\star}} \tag{1}$$

$$\text{RMSE} = \frac{1}{|\Omega|} \sqrt{\sum_{p \in \Omega} (\hat{D}_p^{\text{med}} - D_p^{\star})^2}, \tag{2}$$

where Ω is the set of valid pixels. δ_1 denotes the proportion of pixels satisfying $\text{Max}(D_p^{\star}/\hat{D}_p^{\text{med}}, \hat{D}_p^{\text{med}}/D_p^{\star}) < 1.25$.

3. Analysis of Range Mask

The range mask serves as a scene-specific verification mechanism that complements the depth prediction. Although the model aims to produce fully accurate dense depth, inevitable noise and missed estimations still occur, especially in extreme-distance regions. Over-estimation beyond the physical range often corresponds to invalid or sky pixels, and the independently predicted range mask helps correct such errors by providing a second-stage validity check. This design improves boundary sensitivity, reduces overshooting, and prevents depth leakage into invalid regions.

To better illustrate this behavior, we provide qualitative visualizations at four truncation levels: 10 m, 20 m, 50 m, and 100 m. The 10 m and 20 m masks mainly capture near-field structures such as walls, vegetation, and obstacles, offering precise constraints for indoor scenes and cluttered close-range geometry. In contrast, the 100 m mask reflects long-range depth patterns and is particularly important for open-world outdoor panoramas, where large unbounded sky regions would otherwise introduce instability. The 100 m truncation enables the model to consistently suppress invalid far-distance responses and to better distinguish sky from distant geometry.

Table 1 further quantifies these observations. The 10 m mask leads to better performance on indoor benchmarks (Stanford2D3D and Matterport3D), where near-range geometry dominates most of the panoramas. Conversely, the 100 m mask provides the best results on outdoor datasets (DAP-2M-Labeled and Deep360), where long-range visibility and sky suppression play a critical role. These results confirm that the range mask effectively adapts to different scene scales and serves as a robust depth validation mechanism across diverse environments.

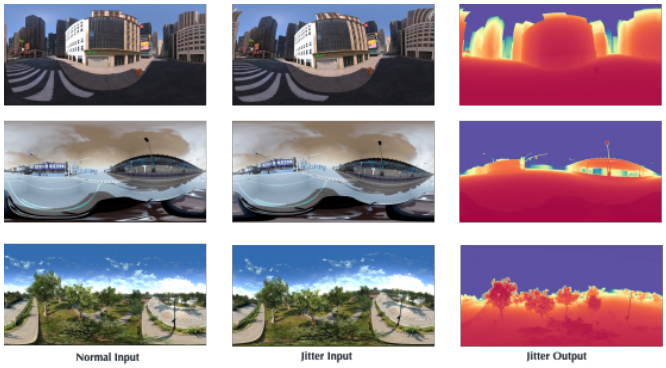

Figure 1. Quantitative evaluation under jitter perturbations.

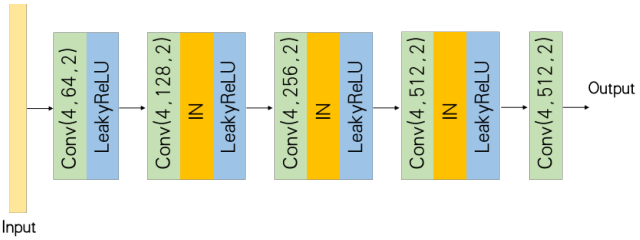

Figure 2. Architecture of the proposed discriminator.

4. Jitter Verification

To assess the robustness of our model under realistic capture instability, we conduct a jitter verification study based on controlled roll and pitch perturbations. In addition to the standard training pipeline, we synthetically generate a perturbed dataset using random rotational offsets sampled from

$$\text{roll} \sim \mathcal{U}(-5^{\circ}, 5^{\circ}), \quad \text{pitch} \sim \mathcal{U}(-10^{\circ}, 10^{\circ}),$$

which closely mimics the subtle shake and drift commonly observed in handheld, head-mounted, or drone-mounted panoramic cameras. These perturbations induce mild but realistic rotational distortions in the equirectangular domain, providing a faithful simulation of sensor noise and platform vibrations.

Our model can be further fine-tuned on this jitter-augmented dataset, and both the quantitative results (Table 2) and qualitative comparisons (Fig. 1) show that the model remains stable under such disturbances. The predictions retain consistent global structure and exhibit no flickering, boundary tearing, or sky-ground instability, even under nontrivial motion perturbations. Incorporating the jitter-augmented data does not degrade performance on the standard evaluation benchmarks, and the fine-tuned model maintains nearly identical accuracy to the version trained on the original dataset.

Table 1. Ablation on the range mask head. The 10 m mask used in indoor datasets dominated by near-range geometry, while the 100 m mask used on outdoor datasets that contain large-scale open spaces.

Mask Depth	DAP-2M-Labeled (Outdoor)		Deep360 (Outdoor)		Stanford2D3D (Indoor)		Matterport3D (Indoor)	
	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
10	-	-	-	-	0.1084	0.8576	0.1352	0.8217
✗	-	-	-	-	0.1113	0.8553	0.1379	0.8198
100	0.0793	0.9353	0.0862	0.8719	-	-	-	-
✗	0.0832	0.9042	0.0938	0.8411	-	-	-	-

Table 2. Jitter robustness results.

Deep360 (Normal)		Deep360 (Jitter)		Stanford2D3D (Normal)		Stanford2D3D (Jitter)	
AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
0.0662	0.9521	0.0681	0.9482	0.0929	0.9118	0.0994	0.9047

Table 3. Performance of different versions of DAP.

Methods		Stanford2D3D (Indoor)			Matterport3D (Indoor)			Deep360 (Outdoor)		
		AbsRel(↓)	RMSE(↓)	δ_1(↑)	AbsRel(↓)	RMSE(↓)	δ_1(↑)	AbsRel(↓)	RMSE(↓)	δ_1(↑)
Metric	DAP-S	0.1226	0.4374	0.8862	0.1497	0.7931	0.8296	0.1028	7.792	0.9079
	DAP-B	0.1021	0.4086	0.8992	0.1319	0.7724	0.8427	0.0842	6.521	0.9377
	DAP-L	0.0921	0.3820	0.9135	0.1186	0.7510	0.8518	0.0659	5.224	0.9525

5. Details and Analysis of the Discriminator

The architecture of our discriminator follows a PatchGAN design, as illustrated in Fig. 2. During training, we treat the depth prediction from Scene-Invariant Labeler as the fake sample and the corresponding ground-truth depth as the real sample. The discriminator is trained to distinguish between these two distributions and provide localized realism feedback to the depth estimator.

To mitigate overfitting and reduce the impact of domain discrepancies, we train two separate discriminators: one dedicated to indoor scenes and one to outdoor scenes. Each discriminator is optimized independently for 60 epochs. After training, we apply each discriminator to all real-world unlabeled data and compute a realism score for every depth map. We then sort the samples based on this score and select the top 300,000 indoor samples and top 300,000 outdoor samples. These high-confidence pseudo-labels are subsequently used to improve the robustness and overall quality of the final DAP model.

6. Different Scale Model

We provide three model variants of DAP to support different computational budgets and deployment scenarios: DAP-S, DAP-B, and DAP-L. All three configurations share the same architectural design and training procedure, differing only the DINOv3 encoder version. As shown in the Table 3, the DAP-S model offers the best efficiency and is suitable for low-latency or resource-limited applications. The DAP-L model delivers the highest performance, benefiting from stronger global context modeling and improved representation of long-range geometry.

7. Analysis of Dense Fidelity Loss

For Dense Fidelity Loss $\mathcal{L}_{DF}$, we adopt a 12-view perspective projection rather than the standard 6-face cubemap in order to obtain more stable and distortion-free depth predictions. A cubemap assigns a very wide field-of-view (90°) to each face, which often results in discontinuities across face boundaries, and reduced sampling density near the poles. These issues lead to unstable depth estimation, especially in open-world outdoor scenes and sky regions.

In contrast, our 12-view projection uses smaller-FoV perspective cameras that produce more regular and less distorted image patches, following the design principles observed in MoGe [1]. The denser directional sampling provides smoother transitions across view boundaries, better

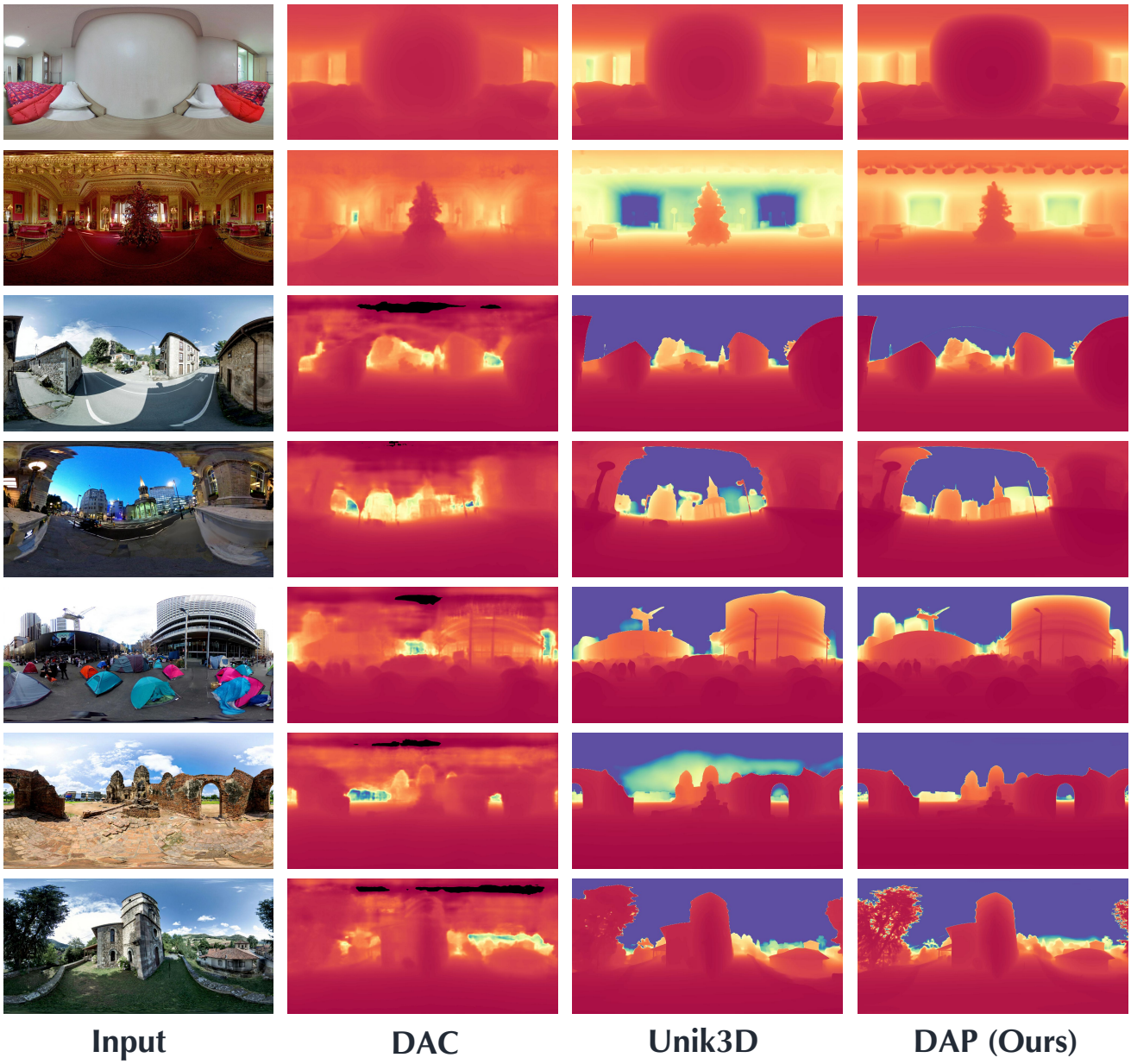

Figure 3. Qualitative comparison.

coverage around the poles, and more reliable far-distance geometry. This is particularly beneficial for handling large unbounded regions such as the sky, where cubemap representations tend to introduce artifacts or inconsistent depth.

For quantitative validation, we also compare the performance of the $\mathcal{L}_{DF}$ against the standard 6-view $\mathcal{L}_{Cube}$ in Table 4. The results show that the 12-view formulation consistently reduces both AbsRel and RMSE while improving δ_1, indicating that denser multi-directional sampling leads to more stable and geometrically reliable depth supervision.

Table 4. Comparison between 6-view cubemap $\mathcal{L}_{Cube}$ and our $\mathcal{L}_{DF}$ on Stanford2D3D.

Method	AbsRel(↓)	RMSE(↓)	δ_1(↑)
✗	0.1112	0.448	0.8509
$\mathcal{L}_{Cube}$	0.1096	0.444	0.8532
$\mathcal{L}_{DF}$	0.1084	0.442	0.8576

8. More Qualitative Results

We provide additional qualitative results to further illustrate the behavior of our model across diverse scenes in Figure 3. These examples include indoor and outdoor panoramas, complex vegetation, and large-scale urban layouts. The visualizations demonstrate that our model produces stable, structurally coherent depth across a wide variety of real-world scenarios, capturing both fine-grained details and global scene geometry. These supplementary results also highlight reliable sky handling, smooth object boundaries, and consistent performance under challenging illumination and viewpoint conditions.

9. Future Work

Although DAP already delivers strong performance and robust metric predictions, there remain many promising directions for future exploration. On the one hand, applying panoramic metric depth to downstream tasks such as autonomous drone navigation and obstacle avoidance, panoramic 3D reconstruction, or depth-conditioned panoramic image/video generation may further unlock its practical potential. On the other hand, extending the paradigm to temporal models (e.g., video-based panoramic depth estimation with motion cues) or incorporating multi-sensor fusion could improve stability and reliability in dynamic real-world environments.

References

[1] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025. 3