ST^{2}360D: Spatial-to-Temporal Consistency for Training-free 360 Monocular Depth Estimation —Supplementary Material—

Zidong Cao[1*] **Jinjing Zhu**[1*] **Hao Ai**[2*] **Lutao Jiang**[1] **Yuanhuiyi Lyu**[1] **Hui Xiong**[1,3†]
[1]Thrust of Artificial Intelligence, HKUST (Guangzhou), China
[2]University of Birmingham, UK
[3]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
caozidong1996@gmail.com; jinjingzhu.mail@gmail.com;
aihao199712@gmail.com; jianglutao98@gmail.com;
ryan.lyu.mail@gmail.com; xionghui@ust.hk

To better understand our ST^{2}360D, we include more details and visualizations in the supplementary material. The table below can guide you to the sections of interest.

Contents

[*]Equal contributions.
[†]Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

A Evaluation Protocol

A.1 Datasets

In this section, we explain the data processing of each dataset in detail.

Matterport3D dataset [1]. It is used to examine the effectiveness of ST^{2}360D in 1K resolution. The maximum depth is 10 meters. In ablation studies, we utilize the Matterport3D dataset with resolution 512×1024, following previous 360° depth estimation methods. Instead, in Tab. 1 of the main paper, we utilize 504×1008 to align with the implementation of PanDA [2] for a fair comparison. The testing set includes 2014 samples.

Stanford2D3D dataset [3]. It is also for validating the effectiveness of ST^{2}360D in 1K resolution. The maximum depth of this dataset is 10 meters. We fill the missing areas in the north and south polar regions, following [2, 4]. The testing set includes 373 samples.

Matterport3D-2K [5]. It is to validate the performance of ST^{2}360D in 2K resolution. The spatial resolutions of 360° RGB images and depth maps are 1024×2048. It is re-rendered by [5] from [1]. The testing set includes 1850 samples.

Replica360-2K [5]. It is also a dataset of 2K resolution. The spatial resolution of samples is 1024×2048. It is re-rendered by [5] from [6]. The testing set includes 130 samples.

Replica360-4K [5]. It is a dataset of 4K resolution, with spatial resolution 2048×4096. It is re-rendered by [5] from [6], including 130 samples in the testing set.

A.2 Metrics

To evaluate the depth estimation performance, we utilize the metrics: *AbsRel*, *RMSE*, *MAE*, *RMSE-log*, and three percentage metrics, δ_i, where $i \in \{1.25, 1.25^2, 1.25^3\}$, following [7, 5]. We only evaluate valid regions of ground truth depth. Formally, we denote ground truth depth as D^*, the number of valid pixels as K, and the predicted depth as D. The metrics can be formulated as follows:

- Absolute Relative Error (*AbsRel*):

$$\frac{1}{K}\sum_{i=1}^{K}\frac{||D(i)-D^*(i)||}{D^*(i)}. \tag{1}$$

- Mean Absolute Error (*MAE*):

$$\frac{1}{K}\sum_{i=1}^{K}||D(i)-D^*(i)||. \tag{2}$$

- Root Mean Square Error (*RMSE*):

$$\sqrt{\frac{1}{K}\sum_{i=1}^{K}||D(i)-D^*(i)||^2}. \tag{3}$$

- Root Mean Square Logarithmic Error (*RMSE-log*):

$$\sqrt{\frac{1}{K}\sum_{i=1}^{K}||\log10(D(i))-\log10(D^*(i))||^2}. \tag{4}$$

- δ_i, where $i \in \{1.25, 1.25^2, 1.25^3\}$:

$$\max\{\frac{D(p)}{D^*(p)}, \frac{D^*(p)}{D(p)}\} < i, \tag{5}$$

A.3 Scale-and-Shift Alignment

We employ scale-and-shift alignment, following [8, 2]. The scale and shift are calculated between the depth prediction D and the depth ground truth D^* in the valid regions. As our ST^{2}360D pipeline outputs 360° depth map, the scale-and-shift alignment is manually implemented in the depth space.

A.4 Compared Methods

Two 360° depth estimation methods [5, 7] are most relevant to our ST^{2}360D. However, since [7] does not release complete code and implementation details, we only report their quantitative results, instead of qualitative comparisons. Additionally, 360MonoDepth [5] employs MiDaS [8] as its perspective depth estimator. However, MiDaS is inferior to recent image-based depth foundation models [9, 10]. For a fair comparison, we replace MiDaS with Depth Anything v2 (DAv2) in the 360MonoDepth pipeline. VDE models demonstrate comparable performance to DAv2 in single-image depth estimation. Our experimental results, presented in Tab. 2 and Tab. 9 of the main paper, show that ST^{2}360D still outperforms 360MonoDepth (with DAv2) by leveraging the temporal consistency inherent in VDE models.

B Methodology

Visualization of LGS strategy. As shown in Fig. 1, we illustrate the generated scanning paths from our proposed LGS strategy, ranging from the level $l = 0$ to level $l = 2$.

Algorithm of latitude-aware traversing. To easily understand the proposed latitude-aware traversing in the LGS strategy, we present it with an algorithm block in Algorithm. 1.

Algorithm 1 Latitude-based Traversing

Input: A set of N Viewpoints $\mathbf{v} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, and K key latitudes $0 < \theta_1 < \cdots < \theta_K \leq 90°$
Latitude computation: For each viewpoint $\mathbf{v}_i$, compute $\Theta(\mathbf{v}_i)$
Slice partitioning: Partition viewpoints into latitude slices based on K key latitudes:
    $\mathbf{v}_{\text{slice}_0} = \{\mathbf{v}_i \mid \mathbf{v}_i \in \mathbf{v}, |\Theta(\mathbf{v}_i)| \leq \theta_1\}$
    $\mathbf{v}_{\text{slice}_k} = \{\mathbf{v}_i \mid \mathbf{v}_i \in \mathbf{v}, \theta_k < |\Theta(\mathbf{v}_i)| \leq \theta_{k+1}\}$, for $k = 1, \ldots, K-1$
    $\mathbf{v}_{\text{slice}_K} = \{\mathbf{v}_i \mid \mathbf{v}_i \in \mathbf{v}, \theta_K < |\Theta(\mathbf{v}_i)| \leq 90°\}$
Slice ordering: Arrange the slices in ascending latitude: $\mathbf{v}_{\text{slices}} \leftarrow [\mathbf{v}_{\text{slice}_0}, \ldots, \mathbf{v}_{\text{slice}_K}]$
Output: The ordered slices $\mathbf{v}_{\text{slices}}$

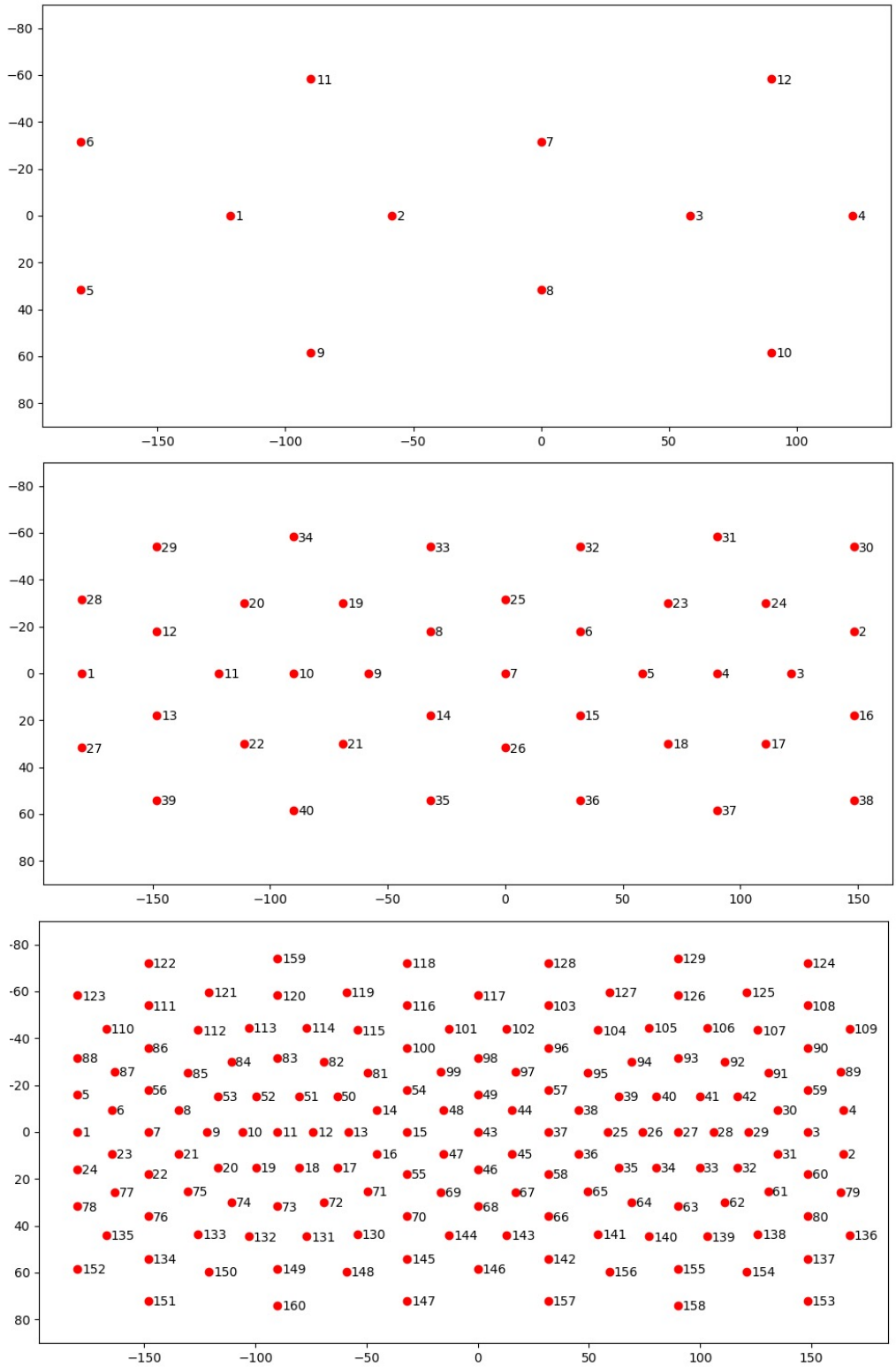


Figure 1: The illustration of the scanning paths of the LGS strategy in different levels. Top: level $l = 0$; Middle: $l = 1$; Bottom: $l = 2$. The indices are marked on the right of the corresponding viewpoints.

C Experiments

C.1 Effectiveness of Starting Longitude

In the main paper, we have discussed the effectiveness of starting latitudes in Sec. 3.3. As shown in Fig. 2, we further discuss the effectiveness of the starting longitudes. In our LGS strategy, the starting viewpoints are from the equator, whose latitude is 0°. With level $l = 2$, there are 16 viewpoints located at the equator. From Fig. 2, it can be found that choosing starting viewpoints with less absolute longitudes would obtain performance benefits slightly. Empirically, we choose the starting viewpoint with longitude -31°.

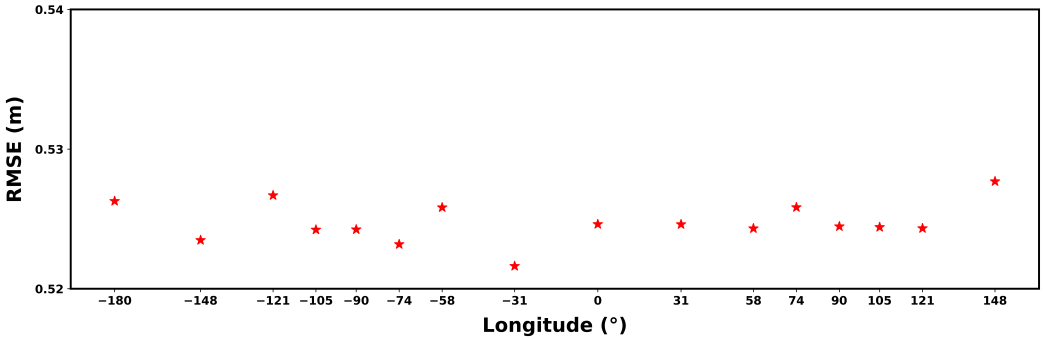


Figure 2: The effectiveness of the longitudes of starting viewpoints.

C.2 More Choices of Key Latitudes

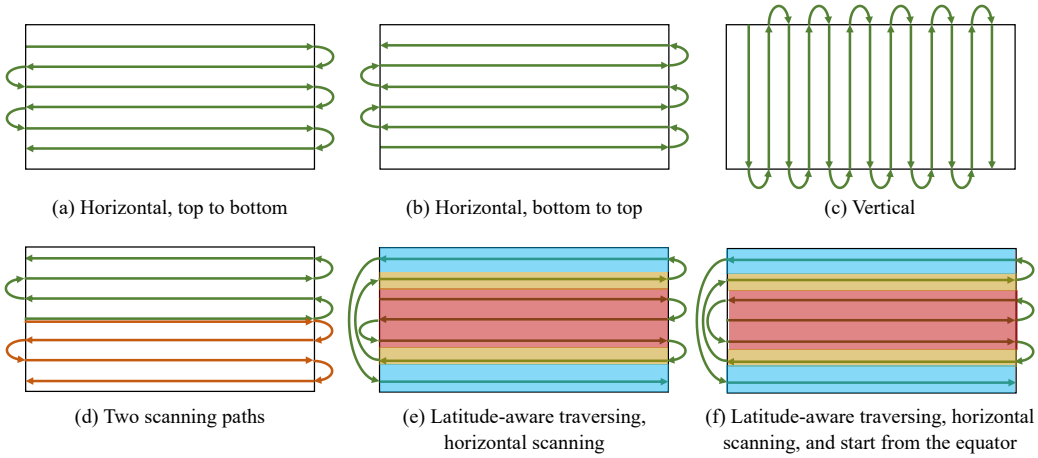
In the Tab. 6 of the main paper, we have done an ablation study about the choices of key latitudes in the LGS strategy. We then provide more choices of the key latitudes. From Tab. 1, it can be found that if only one key latitude is utilized, the smaller key latitude would be a better choice. Moreover, increasing the number of key latitudes from 2 to 4 can not yield further performance gains.

Table 1: More ablation studies on the choices of key latitudes in the LGS strategy.

Choice	{18}	{36}	{54}	{72}	{18, 36}	{18, 54}	{18, 72}	{36, 54}
RMSE ↓	0.5246	0.5236	0.5272	0.5315	**0.5216**	0.5226	0.5237	0.5232
Choice	{36, 72}	{54, 72}	{18, 36, 54}	{18, 36, 72}	{18, 54, 72}	{36, 54, 72}	{18, 36, 54, 72}	
RMSE ↓	0.5234	0.5272	0.5239	0.5218	0.5226	0.5232	0.5238	

C.3 More Scanning Strategies

In Fig. 3, to comprehensively demonstrate the effectiveness of our proposed LGS strategy, we compare it against additional possible scanning strategies. Strategies (a) and (b) are horizontal scanning approaches, one of which scans from top to bottom, while the other scans from bottom to top. In Tab. 2, these two strategies show no significant differences. Furthermore, strategy (c) employs a vertical scanning approach, and its performance decreases significantly. We attribute this to the smoother visual transition of horizontal scanning compared to vertical scanning. Next, starting from the equator, we divide the scanning path into two directions toward the top and bottom, respectively. From Tab. 2, it can be found that utilizing two paths from the equator achieves better performance than the horizontal and vertical scanning strategies by placing low-latitude regions early in the sequence. Moreover, with our proposed latitude-aware traversing strategy, the scanning direction within each slice can proceed either from top to bottom (e) or from low-latitude regions to high-latitude regions (f). In Tab. 2, using the latitude-aware traversing strategy further improves performance. Finally, our LGS strategy achieves the best performance compared to all these scanning strategies (a)-(f) by utilizing both the latitude-aware traversing and the spherical neighbor viewpoint searching strategies.



(a) Horizontal, top to bottom (b) Horizontal, bottom to top (c) Vertical

(d) Two scanning paths (e) Latitude-aware traversing, horizontal scanning (f) Latitude-aware traversing, horizontal scanning, and start from the equator

Figure 3: The illustration of more possible scanning strategies for comparison.

Table 2: Comparison among different scanning strategies. Highlighting: **best**, **second-best**.

Scan Index	*AbsRel* ↓	*RMSE* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
(a)	0.1776	0.5603	75.61	93.80	97.99
(b)	0.1821	0.5655	74.60	93.61	98.03
(c)	0.1945	0.6128	72.77	91.93	97.41
(d)	0.1587	0.5301	80.35	**95.43**	98.45
(e)	0.1580	0.5291	80.42	95.32	98.42
(f)	**0.1571**	**0.5287**	**80.66**	**95.43**	**98.46**
Ours	**0.1530**	**0.5216**	**81.37**	**95.62**	**98.52**

C.4 More Analysis of VDE models

VDA [11] has proposed a key frame alignment strategy for super-long videos. In Tab. 3, after removing the frame alignment, the performance has degraded slightly. It demonstrates that the temporal consistency encoded in VDE models is the most crucial component for our pipeline.

Table 3: Check the effectiveness of frame alignment in VDA [11].

Methods	*AbsRel* ↓	*RMSE* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
w/o frame alignment	0.1530	0.5229	**81.63**	95.61	98.45
w frame alignment	**0.1530**	**0.5216**	81.37	**95.62**	**98.52**

C.5 Blending in Disparity and Depth Space

In 360MonoDepth [5], the alignment and blending processes are implemented in the disparity space. However, as depicted in Tab. 4, we find that blending in the depth space within our pipeline achieves better performance in 2 out of 7 metrics, with the *MAE* and *RMSE* showing significant performance gains. Across all seven metrics, blending in the depth space obtains a total of 13.16% improvement. Moreover, we observe that blending in the disparity space produces smoother visualization results in outdoor scenarios, particularly in sky regions.

C.6 Effectiveness of Perspective Patch Projection

In our ST^{2}360D pipeline, we project a 360° image into a series of perspective patches. As an alternative, we project a 360° image into slices directly from the ERP plane, as shown in Fig. 4. The spatial resolution of the 360° image is $H \times W$, where $W = 2H$. We set the spatial resolution of each slice as $H \times H$, which covers all latitudes. Moreover, by setting the horizontal stride as $W/64$, we can obtain 64 slices. These slices are then compiled into a sequence from left to right and fed into the

Table 4: Check the choices of blending space.

Blending space	*AbsRel* ↓	*MAE* ↓	*RMSE* ↓	*RMSE-log* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	Total
Disparity	**0.1233**	0.3183	0.6165	**0.0751**	**86.86**	**96.56**	**98.74**	
Depth	0.1403	**0.2894**	**0.4664**	0.0776	84.31	96.06	98.64	
Δ	-13.79%	+9.08%	+24.35%	-3.33%	-2.55%	-0.50%	-0.10%	+13.16%

VDE model to obtain slice-based video depth maps, which are finally merged back into a complete ERP depth map. As shown in Tab. 5, without the perspective patch projection, performance drops significantly. This is primarily because perspective patch projection reduces distortion, enabling the generated video frames to align better with the perspective VDE model.

Figure 4: The illustration of the projection process from a 360° image to horizontal slices.

Table 5: Check the effectiveness of perspective patch projection.

Projection methods	*AbsRel* ↓	*RMSE* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Horizontal slices	0.2555	0.7095	58.41	86.68	96.35
Perspective projection	**0.1530**	**0.5216**	**81.37**	**95.62**	**98.52**

C.7 Robustness of LGS Strategy in Outdoor Scenes

To verify the robustness of the proposed LGS strategy in outdoor scenarios, we describe the details of additional experiments below.

Outdoor datasets. We utilize two datasets containing outdoor scenes. The first one is the Deep360 dataset [12], including outdoor driving scenes generated via the CARLA simulator [13]. Its testing set contains 600 samples. The second dataset is derived from the Matterport3D dataset. We employ SegFormer [14] to detect sky regions in each sample of the Matterport3D dataset and collect samples where the detected sky regions constitute more than 10% of the entire ERP images. Subsequently, we manually select 100 samples by further filtering out wrongly detected samples.

Structural Information Across Latitudes in Outdoor Scenes. We quantitatively assess the distribution of structural information across latitudes. Specifically, we utilize the Sobel operator to detect edges of 360 ground-truth depth maps, and normalize edge maps to the range [0, 255]. Subsequently, we average all edge maps within each dataset and summarize the edge values within four distinct latitude slices: $[-90°, -45°)$, $[-45°, 0°)$, $[0°, 45°)$, and $[45°, 90°]$. The summarized edge values are normalized and presented in Tab. 6. The results indicate that the equator regions consistently exhibit rich structural information across both outdoor datasets. Furthermore, high-latitude regions can also contain significant structural information, particularly in the Deep360 dataset.

Ablation studies of LGS strategy in outdoor scenes. To evaluate the robustness of the proposed LGS strategy, we conduct ablation studies on the two outdoor datasets. The results are presented in Tab. 7. These results demonstrate that the LGS strategy consistently improves performance in outdoor scenes.

Table 6: The distribution of structural information in outdoor scenes.

Dataset	Samples	$[-90°, -45°)$	$[-45°, 0°)$	$[0°, 45°)$	$[45°, 90°]$
Deep360	600	38%	**41%**	19%	2%
Matterport3D	100	13%	35%	**37%**	15%

Table 7: Ablation studies of the LGS strategy in outdoor scenes.

Dataset		*AbsRel*↓	*RMSE*↓
Matterport3D	w/o LGS	0.2725	0.9498
	w/ LGS	**0.2512**	**0.8977**
Deep360	w/o LGS	2.3860	1.5312
	w/ LGS	**2.0369**	**1.3986**

C.8 The depth quality at edge regions

To evaluate the sharpness of estimated depth maps at edge regions, we employ the "*Boundary F1 score*" proposed in Depth Pro [15]. We conduct experiments on the Replica360-2K dataset.

We compare our method with PanDA [2], which fine-tunes Depth Anything v2 on synthetic 360° depth datasets. Given the significant influence of input resolution on boundary accuracy, we compare our method with PanDA at two input resolutions (the training resolution of PanDA is 504×1008). Our method takes perspective frames with 518×518 resolution. The results in Tab. 8 demonstrate that our ST^{2}360D with mean blending outperforms PanDA even when using a higher input resolution of 1008×2016. Employing Poisson Blending can further improve the boundary accuracy of our method. We think it is because our training-free pipeline is capable of preserving the high depth precision inherent in video depth estimation models, which have been trained on a diverse set of high-quality perspective depth datasets.

Table 8: Comparison of depth quality at edges.

Method	Input Resolution	Boundary F1↑
PanDA-Small	504×1008	0.0833
PanDA-Small	1008×2016	0.1911
Ours (VDA-Small)^{M}	518×518	0.2134
Ours (VDA-Small)^{P}	518×518	**0.2380**

D Visualization Results

In Fig. 5, we provide more visualization results of our ST^{2}360D in diverse scenarios, including both indoor and outdoor scenarios. Mean blending fails in texture-less regions, such as the sky. The 360° images are from Matterport3D dataset [1], Stanford2D3D dataset [3], 360+x [16] dataset, PanDA [2], and Ntire 2023 challenge on 360° images [17]. Moreover, in Fig. 6, we visualize the colored point clouds generated from our ST^{2}360D. Finally, in Fig. 7, we present the comparison results on the Matterport3D dataset along with error maps. We use the version of PanDA that has not been fine-tuned on the Matterport3D dataset. The results demonstrate that our training-free ST^{2}360D can produce clearer structural details than UniFuse and PanoFormer, while achieving performance comparable to PanDA.

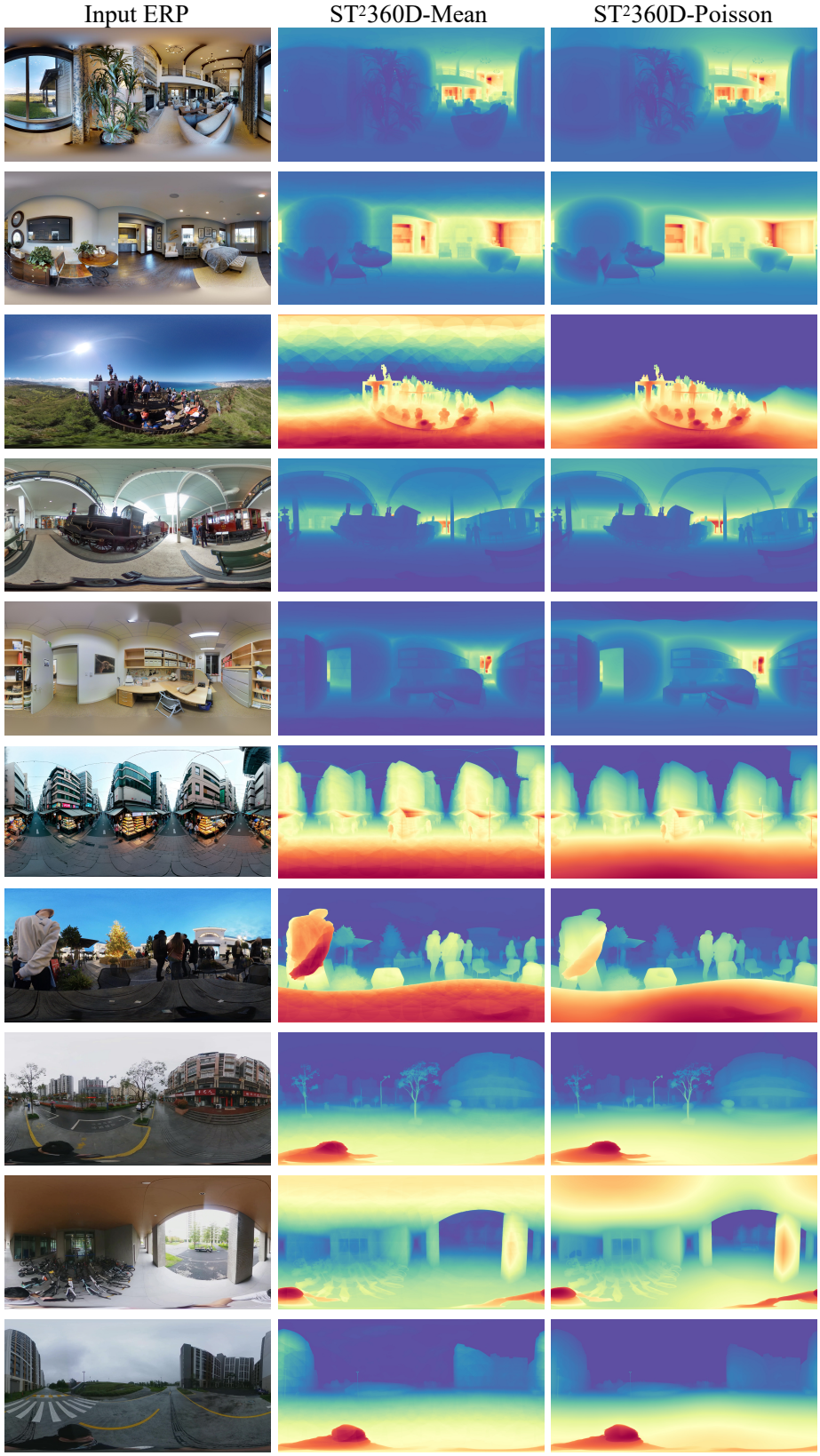


Figure 5: More visualization results of our ST^{2}360D using both Mean blending and Poisson blending.

Figure 6: Visualization results of point clouds generated from our ST^{2}360D.

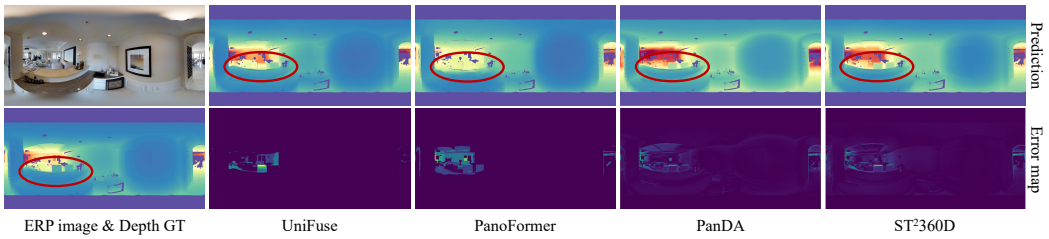


Figure 7: Visualization results on the Matterport3D dataset with error maps.

References

[1] 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. *arXiv preprint arXiv:1709.06158*, 2017.

[2] Zidong Cao, Jinjing Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, and Lin Wang. Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 982–992, 2025.

[3] I Armeni. Joint 2d-3d semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[4] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 6:1519–1526, 2021.

[5] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3772, 2022.

[6] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[7] Chi-Han Peng and Jiayao Zhang. High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3116–3125, 2023.

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[11] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025.

[12] Ming Li, Xueqian Jin, Xuejiao Hu, Jingzhao Dai, Sidan Du, and Yang Li. Mode: Multi-view omnidirectional depth estimation with 360 cameras. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022.

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[14] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[15] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations*.

[16] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multi-modal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19373–19382, 2024.

[17] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, et al. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1731–1745, 2023.