PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Möbius Spatial Augmentation

Supplementary Material

A. Datasets

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

Structured3D dataset [55]. It is an indoor synthetic dataset. We employ its training set, which consists of 18,298 samples, for our training purposes. In terms of data processing, we initially scale the depth map values by a factor of 0.001. Subsequently, we clip these values to a range of 0 to 10 meters. Finally, we apply depth normalization [21].

Deep360 dataset [26]. This dataset is synthetic and contains outdoor scenes, generated using the CARLA simulator [15]. It comprises pairs of fisheye images and depth maps. Following the official guidance of [26], we transform the fisheye format into the ERP representation. We clip the depth values to a range of 0 to 100 meters. We restrict larger depth values in the sky region to 100 meters. Subsequently, depth normalization is applied following [21].

ZInD dataset [14]. This is an indoor dataset with room layout annotations but lacking depth labels. We employ it for semi-supervised learning to enhance the scene diversity of indoor environments. We utilize its training set with 54034 samples for training.

360+x dataset [12]. This dataset encompasses both indoor and outdoor scenes, showcasing its diversity. For data processing, we uniformly extract frames from the high-resolution videos contained within the 360+x dataset. From approximately 200 videos, we extract a total of 47,956 frames. Subsequently, we observe that the performance of our PanDA is suboptimal in extremely dark regions. As a result, these scenes are omitted from the training set. Finally, we utilize SegFormer [47] to detect sky regions and assign a depth value of 1.0 to these areas, which represents the maximum value on the normalized depth map.

Matterport3D dataset [11]. It is used to validate the effectiveness of our PanDA in real-world scenes. The maximum depth value is set at 10 meters.

Stanford2D3D dataset [5]. It is also used to validate the effectiveness of our PanDA in real-world scenes. The maximum depth is set at 10 meters. Given that the top and bottom parts of panoramas in the Stanford2D3D dataset are missing, we fill in these missing areas by following the methods described in UniFuse [20].

Other datasets. Besides Deep360 [26], there is another synthetic dataset [7] containing outdoor scenes. However, it was not publicly available at the time of submission.

(a) Cube Map

(b) Tangent Patch

(c) Horizontal Slice

Figure 12. Illustration of the indexes of patches in different panoramic representations.

B. Metrics and Alignment

Metrics. We evaluate with two standard metrics: Absolute Relative Error (*AbsRel*) and Root Mean Squared Error (*RMSE*). Performance assessments are confined to valid regions where ground truth depth, denoted as D^*, is available. We denote the number of valid pixels by K. Additionally, we employ three percentage metrics, δ_i, for values

Representation	Patch Number	FoV	Resolution	Equator Region	Pole Region
ERP	1	$180° \times 360°$	504×1008	–	–
Cube map (CP)	6	$90° \times 90°$	252×252	{Front, Left, Right, Back}	{Top, Down}
Tangent patch (TP)	18	$80° \times 80°$	126×126	4^{th} to 15^{th}	1^{st} to 3^{rd}, and 16^{th} to 18^{th}
Horizontal slice (HS)	4	$45° \times 360°$	126×1008	2^{nd} and 3^{rd}	1^{st} and 4^{th}
Vertical slice (VS)	4	$180° \times 90°$	504×252	–	–

Table 9. The settings of panoramic representations.

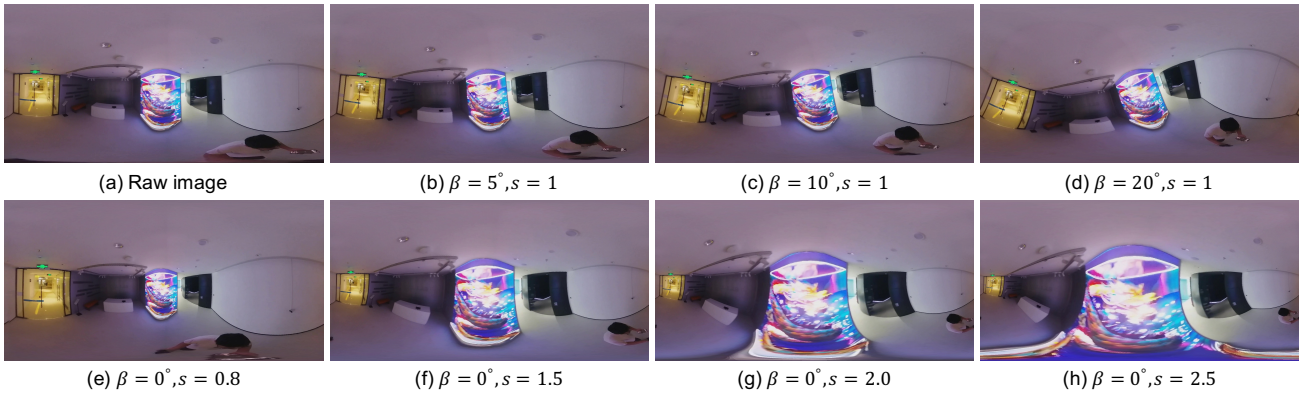


Figure 13. Visualization of panoramas under different transformations.

of $i \in \{1.25, 1.25^2, 1.25^3\}$. With the predicted depth D, metrics can be formulated as follows:

- Absolute Relative Error (*AbsRel*):

$$\frac{1}{K}\sum_{i=1}^{K}\frac{||D(i) - D^*(i)||}{D^*(i)}. \tag{8}$$

- Root Mean Square Error (*RMSE*):

$$\sqrt{\frac{1}{K}\sum_{i=1}^{K}||D(i) - D^*(i)||^2}. \tag{9}$$

- δ_i, the fraction of pixels where the relative error between the depth prediction D and ground truth depth D^* is less than the threshold i:

$$\max\{\frac{D(p)}{D^*(p)}, \frac{D^*(p)}{D(p)}\} < i. \tag{10}$$

Alignment. In the main paper, the reported results of PanDA-{S,B,L} in Tab. 5, 6 do not apply any alignment operation for a fair comparison. In addition, to assess the zero-shot performance of Depth Anything v1 and v2, Marigold, and our PanDA, we employ scale and shift alignment as described in [31]. The scale and shift adjustments of the depth predictions are manually aligned with the depth ground truth. In Tab. 1, 2, and Fig. 6, this alignment is performed in the disparity space. Conversely, in Tab. 4, 7, 8, the alignment occurs in the depth space.

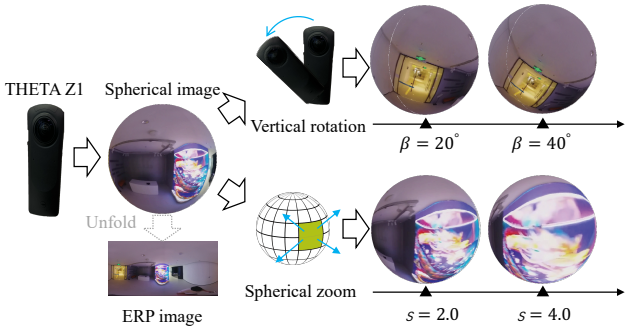


Figure 14. Illustration of the spherical spatial transformations.

C. Analysis

C.1. Different Panoramic Representations

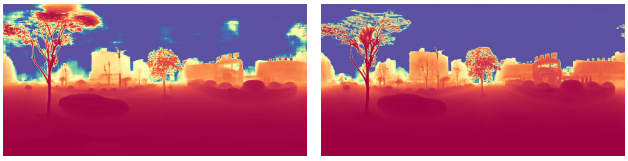
In Tab. 9, we detail the settings of the panoramic representations used, including the number of patches, field-of-view (FoV), spatial resolution, and the grouping of the equator and polar regions. The indices of patches for CP, TP, and HS are illustrated in Fig. 12.

C.2. Different Camera Positions

In Fig. 3, we utilize iPad Pro and the app "polycam" to scan and generate the point cloud of the scene.

C.3. Various Spatial Transformations

Meaning. As depicted in Fig. 14, given the 360° camera, such as with the THETA Z1, it is not always possible to en-



504x1008 input 1008x2016 input

Figure 15. Visual comparison between different input resolutions.

sure that spherical images are captured vertically. Another scenario occurs in virtual reality (VR) environments, where users have the freedom to adjust their viewing directions and zoom in on objects of interest for an immersive experience. In these cases, spherical transformations are crucial to meet the practical demands of real-world applications.

More visualization of spatial transformations. Additional visualization results of the Möbius transformation are presented in Fig. 13, including vertical rotations with different angles β and spherical zooms with different zoom levels s. It is obvious that the transformations introduce more curves, which complicates the task of panoramic depth estimation compared to panoramas captured vertically.

D. The Proposed Method

D.1. EPNL

For each panorama, we sample 32 patches. The horizontal position of the patch center is randomly selected from a range of 0 to W. For the vertical position, we use a Gaussian distribution to sample more patches around the equator region. The mean of this distribution is set at $\frac{H}{2}$, and its variance at $\frac{H}{6}$.

D.2. Spatial Resolution of Pseudo Depth Labels

As shown in Fig. 15, when generating pseudo depth labels for unlabeled panoramas, increasing the input resolution significantly reduces noise and enhances structural details.

D.3. MTSA

Overview. We illustrate the detailed process of the Möbius transformation for panoramas. The formulas are based on [10]. Differently, we take the equator center as the pole to zoom in on the objects at the equator. As illustrated in Fig. 16, a panorama u_i undergoes an initial projection from the plane to the sphere via spherical projection (SP). Subsequently, this spherical representation is projected onto the complex plane using stereographic projection (STP). In our conduction, the specific point on the complex plane is determined by the intersection of the equator point and a designated spherical point. The Möbius transformation is applied on the complex plane. Following this, we apply the inverse stereographic projection (STP^{-1}) and inverse spherical projection (SP^{-1}) to obtain the transformed panorama $\mathcal{M}(u_i)$.

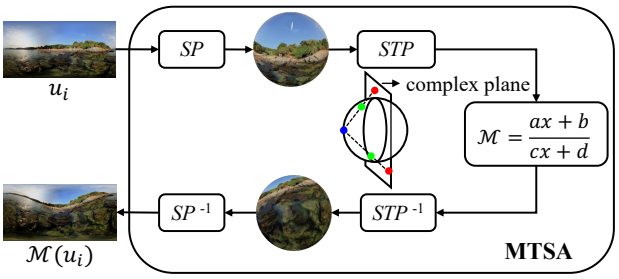


Figure 16. Illustration of the process of MTSA. •: Equator point; •: Spherical point; •: Complex plane point.

The Möbius transformation is conducted in the complex plane. To achieve it, a panorama with ERP representation is first projected from the plane to the sphere via spherical projection (SP). The plane coordinate is proportional to the angle coordinate (θ, ϕ) (where θ represents the longitude and ϕ represents the latitude), while the spherical coordinate can be defined as (x, y, z). In this case, SP can be defined as follows [10]:

$$\text{SP}: \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos\phi\cos\theta \\ \cos\phi\sin\theta \\ \sin\phi \end{pmatrix}. \tag{11}$$

Then, we project from the sphere to the complex plane with stereographic projection (STP). By defining the coordinate of the complex plane as $Z = (x', y')$ and selecting the equator center as the pole, the STP can be formulated as follows:

$$\text{STP}: x' = \frac{y}{1-x}\ ,\ y' = \frac{z}{1-x}. \tag{12}$$

In the complex plane, the Möbius transformation is conducted with the following formulation:

$$f(Z) = \frac{aZ+b}{cZ+d}, \tag{13}$$

where a, b, c, and d are complex numbers. In addition, a, b, c, and d should satisfy $ad - bc \neq 0$. For the vertical rotation with angle β, the parameters of Möbius transformations can be represented as follows [10]:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos\beta + j\sin\beta & 0 \\ 0 & 1 \end{pmatrix}. \tag{14}$$

For the zoom operation with level s, the parameters of Möbius transformations can be represented as follows [10]:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} s & 0 \\ 0 & 1 \end{pmatrix}. \tag{15}$$

The Möbius transformation obeys the matrix chain multiplication rule. After the Möbius transformation in the complex plane, we conduct inverse projections to project

from the complex plane to the sphere and the plane, respectively. The inverse projections can be formulated as follows:

$$\begin{aligned} \mathrm{STP}^{-1}: \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} \frac{-1+x'^2+y'^2}{1+x'^2+y'^2} \\ \frac{2x'}{1+x'^2+y'^2} \\ \frac{2y'}{1+x'^2+y'^2} \end{pmatrix}; \\ \mathrm{SP}^{-1}: \begin{pmatrix} \theta \\ \phi \end{pmatrix} &= \begin{pmatrix} \arctan(y/x) \\ \arcsin(z) \end{pmatrix}. \end{aligned} \tag{16}$$

E. More Experimental Results

E.1. The parameters of MTSA

For the proposed MTSA, the default setting is that: the vertical rotation angle is uniformly sampled in $[-10°, 10°)$, denoted as $\mathcal{U}(-10°, 10°)$. Moreover, the zoom level is uniformly sampled in $[1, 1.5)$, denoted as $\mathcal{U}(1, 1.5)$. To further discuss the effect of the MTSA by introducing vertical rotation and spherical zoom into spatial augmentation, we conduct ablation studies for the range of sampling distribution in Tab. 10 and Tab. 11.

Methods	Original	Vertical angle θ		Zoom level s	
		10°	20°	2.0	3.0
$\mathcal{U}(-5°, 5°)$	**0.4896**	0.5309	0.6045	0.6140	0.7426
$\mathcal{U}(-10°, 10°)$	0.4915	0.5188	0.5706	0.6242	0.7461
$\mathcal{U}(-20°, 20°)$	0.4923	**0.5157**	0.5500	0.5948	0.7205
$\mathcal{U}(-30°, 30°)$	0.5000	0.5208	**0.5457**	**0.5939**	**0.7023**

Table 10. Examine the range of vertical rotation angles. We report *RMSE* metric on the Matterport3D dataset.

Methods	Original	Vertical angle θ		Zoom level s	
		10°	20°	2.0	3.0
$\mathcal{U}(1, 1.2)$	0.4943	**0.5158**	**0.5667**	0.7425	0.8870
$\mathcal{U}(1, 1.5)$	**0.4915**	0.5188	0.5706	**0.6242**	**0.7461**
$\mathcal{U}(1, 2)$	0.5250	0.5886	0.6890	0.8276	0.9649
$\mathcal{U}(1, 3)$	0.5187	0.5819	0.6841	0.8109	0.9494

Table 11. Examine the range of zoom levels. We report *RMSE* metric on the Matterport3D dataset.

Vertical rotation angle. As shown in Tab. 10, it can be found that a smaller angle distribution can benefit the depth estimation of the original panorama. Moreover, MTSA with a larger angle distribution benefits the depth prediction on panoramas with larger rotation angles, *e.g.*, 20°, and larger zoom levels, *e.g.*, 3.0. Transformations with larger vertical rotation angles and larger zoom levels would introduce severe curves to challenge the panoramic depth estimation. Our choice of $\mathcal{U}(-10°, 10°)$ is a balance between the performance of original and transformed ones.

Zoom level. As depicted in Tab. 11, we investigate the impact of various zoom level distributions. We observe that employing larger zoom level distributions, such as $\mathcal{U}(1, 2)$ and $\mathcal{U}(1, 3)$, can degrade the depth estimation performance for both original and transformed panoramas. We attribute this degradation to the severe distortions that hinder the model from learning effective structural information.

E.2. Pseudo Depth Labels

Different amounts of pseudo depth labels. To further examine the effect of pseudo depth labels in the SSL pipeline, we vary the amounts of pseudo depth labels, as illustrated in Tab. 12. The results show that the larger the amount of unlabeled data, the better the performance, especially under spherical transformations.

Num. of unlabeled data	Original	Vertical angle θ		Zoom level s	
		10°	20°	2.0	3.0
10199 (10%)	0.4998	0.5280	0.5912	0.6892	0.8248
20398 (20%)	0.4932	0.5252	0.5910	0.6919	0.8255
101990 (100%)	**0.4915**	**0.5188**	**0.5706**	**0.6242**	**0.7461**

Table 12. Vary the number of unlabeled data during SSL. We report *RMSE* metric on the Matterport3D dataset.

Only pseudo depth labels for training. To further investigate the effect of pseudo depth labels from the teacher model, in Tab. 13, we only utilize the unlabeled panoramas and the corresponding pseudo depth labels to train the student model. It is observed that training with only pseudo depth labels yields better performance compared to solely using synthetic depth ground truth. This improvement is likely due to several factors: 1) The amount of pseudo depth labels exceeds that of synthetic depth ground truth; 2) The unlabeled data consists of real-world samples; 3) The teacher model provides accurate pseudo labels that enhance student model training. However, both approaches show limited effectiveness in transformed panoramas.

Methods	Original	Vertical angle θ		Zoom level s	
		10°	20°	2.0	3.0
$\mathcal{L}_S$	0.5109	0.5711	0.6804	0.8381	**0.9793**
$\mathcal{L}_P$	**0.5031**	**0.5584**	**0.6557**	**0.8358**	0.9870

Table 13. The effect of only utilizing the pseudo depth labels to train the student model. We report *RMSE* metric on the Matterport3D dataset.

E.3. Few-shot Learning for Fine-Tuning

The student model has been trained using both synthetic data and large-scale unlabeled data. We explore whether a small amount of real-world panoramic depth ground truth is sufficient to fine-tune our PanDA for real-world scenes. In this context, Tab. 14 demonstrates the results from uniformly sampling the Matterport3D dataset [11] at percentages of 1%, 5%, 10%, and 25%. It is observed that with just 5% of the samples, our PanDA can be fine-tuned to

Percentage	*AbsRel* ↓	*RMSE* ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
1%	0.1340	0.5303	83.26	96.01	98.94
5%	0.1099	0.4356	89.48	97.93	99.31
10%	0.1002	0.4236	90.77	98.03	99.40
25%	0.0946	0.3967	91.91	98.26	99.47
100%	**0.0922**	**0.3950**	**92.26**	**98.30**	**99.47**

Table 14. Utilizing small parts of the training set of the Matterport dataset for fine-tuning.

achieve competitive results with existing SOTA panoramic monocular depth estimation methods. Additionally, at 25%, the performance closely approximates that achieved by using all the depth ground truth in the training set of Matterport3D [11].

E.4. LoRA Rank

By default, the rank parameter in LoRA is set as 4. In Tab. 15, it can be found that different choices of the rank parameter have a limited effect on the depth estimation performance.

Rank	2	4	8
AbsRel ↓	0.1049	**0.1036**	0.1047
RMSE ↓	**0.4531**	0.4539	0.4583

Table 15. The effect of LoRA rank parameter. We report *RMSE* metric on the Matterport3D dataset.

E.5. The effect of Sampling Regions in EPNL

In Tab. 16, by changing the sampling regions from equator region to polar regions, the performance degrades. We ascribe it as the polar regions contain less structural information. Sampling on the polar regions provides less structural guidance.

Methods	Matterport3D		Stanford2D3D	
	AbsRel ↓	*RMSE* ↓	*AbsRel* ↓	*RMSE* ↓
Sampling in Poles	0.1489	0.5403	0.1274	0.3542
Sampling in Equator	**0.1256**	**0.5062**	**0.1109**	**0.3401**

Table 16. Change to poles (Latitude [-90°,-30°] ∪ [30°,90°]).

E.6. The visualization Issue of DAMs

As illustrated in Fig. 17, some structural details can be neglected if we visualize the depth estimation result of DAM as a whole. This is because the global normalization before visualization would squeeze the details of local regions. Therefore, for a fair comparison, we showcase the local areas of the DAM prediction with local normalization.

E.7. Point Cloud Results

In Fig. 18, the point clouds generated from our depth predictions can recover reasonable structures of the scene, such as the chairs in the classroom and outdoor buildings.

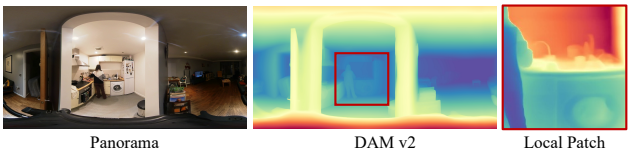


Figure 17. Illustration of the visualization issue of DAMs.

Figure 18. Visualization of point clouds generated from the depth estimation results of our PanDA.

E.8. Model Complexity

With only LoRA added, the parameters of PanDA are similar to DAM v2. As for inference speed, processing a 504×1008 panorama requires 49/90/234ms with PanDA-{S,B,L}, respectively. The running speeds are tested by averaging 100 times on an A40 GPU.

F. Limitation and Future Work

Due to the scarcity of panoramic depth labels in diverse scenes, our teacher model is trained on limited scenes compared with the depth datasets for perspective images. To enhance the zero-shot capability of our model, future work will focus on collecting panoramas paired with depth labels across a broader range of environments, including both synthetic and real-world scenes.

References

[1] Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *CVPR*, 2024. 7

[2] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13273–13282, 2023. 2, 3, 4, 7

[3] Hao Ai, Zidong Cao, and Lin Wang. A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *arXiv preprint arXiv:2502.10444*, 2025. 1

[4] 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 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3727–3737, 2021. 2, 3

[5] 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. 2, 3, 6, 7, 8, 9

[6] Jongbeom Baek, Gyeongnyeon Kim, and Seungryong Kim. Semi-supervised learning with mutual distillation for monocular depth estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4562–4569. IEEE, 2022. 3

[7] Jay Bhanushali, Praneeth Chakravarthula, and Manivannan Muniyandi. Omnihorizon: In-the-wild outdoors depth and normal estimation from synthetic omnidirectional dataset. *arXiv preprint arXiv:2212.05040*, 2022. 9

[8] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 3

[9] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 2, 3

[10] Zidong Cao, Hao Ai, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Lin Wang. Omnizoomer: Learning to move and zoom in on sphere at high-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12897–12907, 2023. 2, 4, 11

[11] 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, 3, 4, 6, 8, 9, 12, 13

[12] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multimodal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19373–19382, 2024. 5, 9

[13] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018. 3

[14] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2133–2143, 2021. 5, 9

[15] 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. 9

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5

[17] Qi Feng, Hubert PH Shum, and Shigeo Morishima. 360 depth estimation in the wild-the depth360 dataset and the segfuse network. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 664–673. IEEE, 2022. 2

[18] Christopher Geyer and Kostas Daniilidis. Conformal rectification of omnidirectional stereo pairs. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, pages 73–73. IEEE, 2003. 4

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[20] 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. 2, 3, 4, 7, 9

[21] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 3, 5, 6, 7, 9

[22] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[25] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017. 3

[26] 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. 2, 5, 8, 9

[27] Meng Li, Senbo Wang, Weihao Yuan, Weichao Shen, Zhe Sheng, and Zilong Dong. S2net: Accurate panorama depth

estimation on spherical surface. *IEEE Robotics and Automation Letters*, 8(2):1053–1060, 2023. 7

[28] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. *CoRR*, abs/2203.00838, 2022. 2, 3, 4

[29] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 3

[30] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, 2021. 3, 4

[31] 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. 2, 3, 5, 10

[32] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic segmentation with active semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5966–5977, 2023. 3

[33] 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. 2, 3

[34] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 6

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 6

[36] Saul Schleimer and Henry Segerman. Squares that look round: transforming spherical images. *arXiv preprint arXiv:1605.01396*, 2016. 4

[37] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15804–15815, 2021. 2

[38] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022. 3, 7

[39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 6

[40] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3

[41] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2573–2582, 2020. 3, 7

[42] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020. 3

[43] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *CVPR*, pages 459–468. Computer Vision Foundation / IEEE, 2020. 3, 7

[44] 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, 2022. 3, 7

[45] Ning-Hsu Albert Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Advances in Neural Information Processing Systems*, 37:127739–127764, 2024. 3, 7

[46] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022. 3

[47] 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. 6, 9

[48] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3060–3069, 2021. 3

[49] 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. 1, 2, 3, 6, 7

[50] 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. 1, 2, 3, 5, 6, 7

[51] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022. 3

[52] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d:

Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 5, 8

[53] Haozheng Yu, Lu He, Bing Jian, Weiwei Feng, and Shan Liu. Panelnet: Understanding 360 indoor environment via panel representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 878–887, 2023. 3

[54] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. *arXiv preprint arXiv:2304.07803*, 2023. 3, 7

[55] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2, 5, 8, 9

[56] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3561–3571, 2023. 3

[57] Yitao Zhu, Zhenrong Shen, Zihao Zhao, Sheng Wang, Xin Wang, Xiangyu Zhao, Dinggang Shen, and Qian Wang. Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 2, 5

[58] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3653–3661, 2022. 7

[59] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 2, 3