
RPG360: Robust 360 Depth Estimation with Perspective Foundation Models and Graph Optimization

Dongki Jung Jaehoon Choi Yonghan Lee Dinesh Manocha
University of Maryland, College Park
{jdk9405, kevchoi, lyhan12, dmanocha}@umd.edu

In this supplementary material, we provide additional qualitative and quantitative results, along with details on how to apply our RPG360 to downstream 360° vision tasks.

A Accuracy and Inference Time depending on the Number of Iterations

Table 1: Performance depending on the number of iterations.

# Iterations	Chamfer ↓	a1 ↑	Time (sec)
(300, 150, 30)	0.34	0.86	4.59
(400, 40, 40)	0.34	0.85	4.44
(40, 40, 400)	0.37	0.83	8.71
(160, 160, 160)	0.35	0.85	6.11
(150, 75, 15)	0.35	0.85	2.56
(600, 300, 60)	0.33	0.86	8.54

We evaluated RPG360 using Metric3D v2 [9] on the Matterport3D dataset [5]. The inference time of our method varies with both the input image resolution and the number of optimization iterations. In this experiment, we used an input resolution of 512×1024 and employed a multi-scale strategy.

B Benchmark Selection for Comparison

As also noted in Depth Anywhere [24], many recent methods [1, 2, 29, 15, 16, 21, 28] have not fully released pre-trained models or provided their code and implementation details. Some of these works utilized other synthetic datasets such as Structured3D [30] or 3D60 [6], instead of Matterport3D [5] or Stanford2D3D [4]. Specifically, Elite360D [1] and HRDFuse [2] do not provide pre-trained weights, while EGFormer [29] has primarily been evaluated on Structured3D [30] and Pano3D [3]. S2Net [15] also does not offer pre-trained weights. OmniFusion includes pre-trained weights for Stanford2D3D [4] and 3D60 [32], but does not provide the weights for Matterport3D [5]. Additionally, its official implementation has some code issues related to these pre-trained weights. Regarding PanoFormer [21], Depth Anywhere [24] reports that its evaluation code and results are incorrect. Joint360 [28] only provides pre-trained weights trained on 3D60 [32] and Structured3D [30].

Thus, we compared our method with fully accessible benchmarks. For **learning-based methods**, SliceNet [19], UniFuse [11], ACDNet [31], and BiFuse++ [23] were selected as they have publicly available code. Additionally, to enable comparison with recent algorithms, we trained Elite360D [1] since its pre-trained weights were not provided. As Depth Anywhere [24] utilizes Depth Anything [27], it produces outputs in inverse depth. However, the official repository does not provide code for converting it back to standard depth. Thus, we implemented this conversion ourselves. For **optimization-based methods**, we used the official code for 360MonoDepth [20] with Poisson blending. Peng and Zhang [18] also proposed a combined approach that incorporates a perspective depth estimation module and an equirectangular projection (ERP) depth estimation module during

optimization. However, as this method inherently relies on the quality of ERP depth estimation results, we excluded it from the comparison to focus on evaluating performance robustness under cross-validation and zero-shot settings with 3D metrics.

The reported 2D metric performances vary across papers due to slight differences in experimental settings. When constructing the evaluation Table 3 in the main paper, we referenced the official results from the most recently published works: Elite360D [1] (learning-based) and Peng and Zhang [18] (optimization-based). For methods not covered in these papers, we referred to the original publications to obtain their reported values.

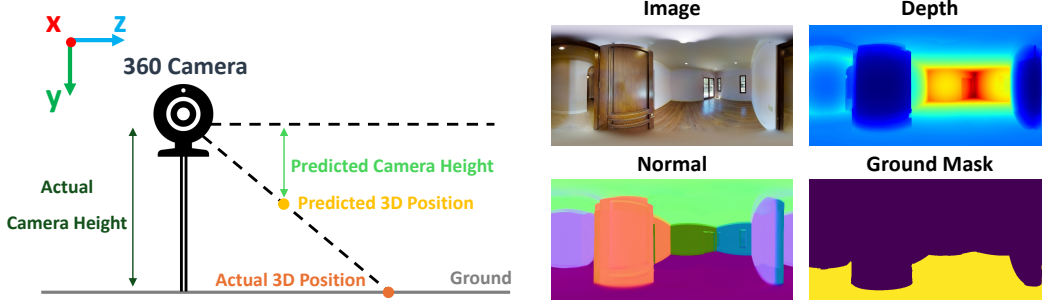


Figure 1: Visualization of metric scale estimation for downstream tasks. By leveraging the camera height assumption [26], we estimate the metric scale of the predicted depth map. From the predicted depth, we back-project the values into 3D points. Using these predicted 3D points, we compute the y-direction distance to define the predicted camera height. To identify ground regions, we utilize predicted normal maps under the assumption that the normal direction of ground surfaces is $(0, -1, 0)$. By incorporating cosine similarity, we extract ground regions and select the corresponding 3D points. Finally, using the similarity ratio, we restore the metric scale based on the actual camera height. For simplicity, we assume all images have an actual camera height of 1m.

C Evaluation Metrics

For the practical application of monocular depth estimation, emphasizing 3D structure awareness is becoming increasingly important [22, 17]. Therefore, we employ 3D metrics rather than 2D metrics to assess improvements in 3D structure and geometry.

Chamfer The Chamfer distance measures the discrepancy between the ground-truth 3D points Q and the predicted 3D points \hat{Q} ,

$$\sum_{\mathbf{q} \in Q} \min_{\hat{\mathbf{q}} \in \hat{Q}} \|\mathbf{q} - \hat{\mathbf{q}}\| + \sum_{\hat{\mathbf{q}} \in \hat{Q}} \min_{\mathbf{q} \in Q} \|\mathbf{q} - \hat{\mathbf{q}}\|. \quad (1)$$

F-Score and IoU The F-Score measures the harmonic mean of precision P and recall R , while IoU assesses the volumetric quality of a 3D reconstruction. Here, precision is the fraction of predicted points within distance δ of the ground-truth surface, and recall is defined vice versa. δ is set to 0.1, following [22, 17],

$$\begin{aligned} \text{F-Score} &= 2 * \frac{P \cdot R}{P + R}, \quad \text{IoU} = \frac{P \cdot R}{P + R - P \cdot R}, \\ \left\{ \begin{aligned} P &= \sum_{\hat{\mathbf{q}} \in \hat{Q}} \left[\min_{\mathbf{q} \in Q} \|\mathbf{q} - \hat{\mathbf{q}}\| < \delta \right], \\ R &= \sum_{\mathbf{q} \in Q} \left[\min_{\hat{\mathbf{q}} \in \hat{Q}} \|\mathbf{q} - \hat{\mathbf{q}}\| < \delta \right], \end{aligned} \right. \quad (2) \end{aligned}$$

where $[\cdot]$ is the Iverson bracket function.

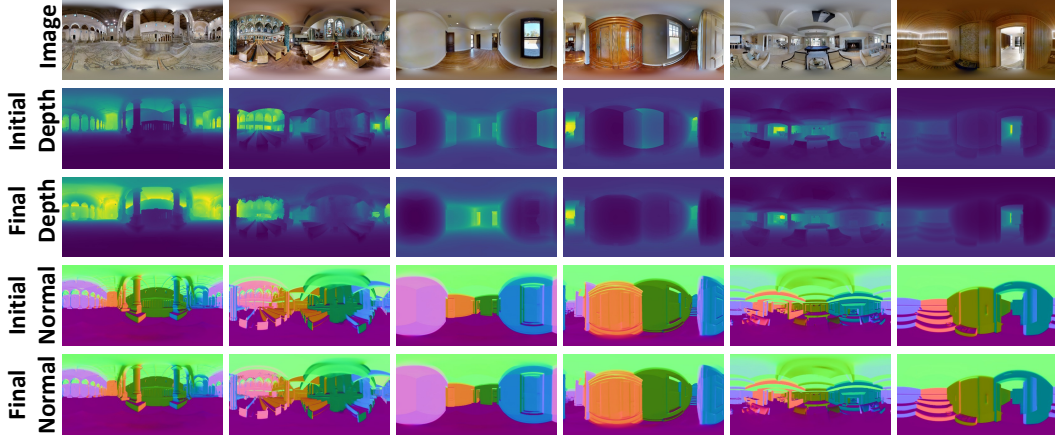


Figure 2: Qualitative results of depth and normal estimation on Matterport3D. The initial depth predicted by the perspective foundation model [9] exhibits depth scale inconsistency artifacts along the cubemap face boundaries. After applying our proposed method, RPG360, the depth estimation results become more consistent and 3D structure-aware.

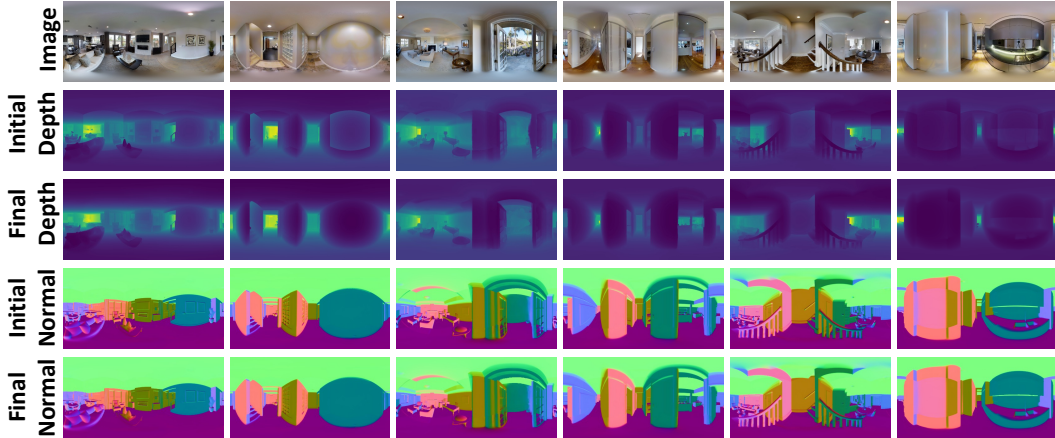


Figure 3: Qualitative results of depth and normal estimation on Matterport3D. The initial depth predicted by the perspective foundation model [9] exhibits depth scale inconsistency artifacts along the cubemap face boundaries. After applying our proposed method, RPG360, the depth estimation results become more consistent and 3D structure-aware.

D Metric Scale Depth for Downstream Tasks

To apply RPG360 to downstream 360° vision tasks, such as training feature matching [12] or Structure from Motion [13], we employ the Spherical-n-Point (SnP) method [8] using the predicted depth maps and correspondences. Utilizing the geometric relationships between multi-view images requires knowledge of a common global scale shared by all images. Therefore, we apply a simple camera height assumption [26], which estimates the metric depth scale using the similarity ratio between the predicted camera height and the actual camera height, as illustrated in Fig. 1.

E Evaluation of Pretrained Models Using 2D Metrics

Evaluation results of 2D-based metrics are shown in Table 2. We evaluate our method using different perspective foundation models. Similar to its performance on 3D metrics, Metric3D [9] achieves the best performance on the 2D metrics.

Table 2: Quantitative comparison on the Matterport3D test set using 2D-based evaluation metrics.

Method	Model	Lower is better				Higher is better		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
RPG360	Marigold [14]	0.45506	2.46920	1.50427	0.23184	0.43284	0.68761	0.82197
RPG360	GeoWizard [7]	0.42683	1.96726	1.29803	0.19848	0.50561	0.75861	0.87709
RPG360	Omnidata [6]	0.21483	0.65625	0.67233	0.10413	0.79576	0.93528	0.97297
RPG360	Metric3D [9]	0.20312	0.93469	0.66718	0.09575	0.85876	0.95323	0.97656

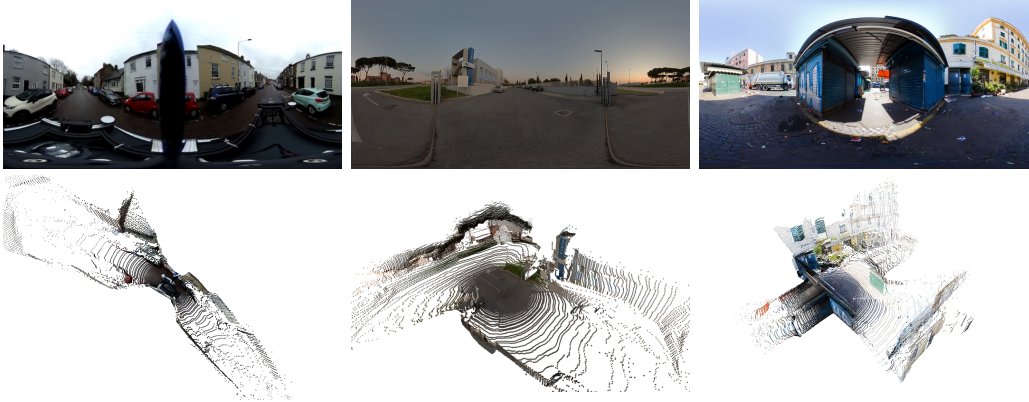


Figure 4: Qualitative results on Dur360BEV [25] and 360 images downloaded from the internet.

F Additional Qualitative Results

As illustrated in Fig. 2 and 3, we present additional qualitative results for RPG360. The proposed graph optimization method effectively aligns the scale across independently predicted depth maps from each cubemap face. We also include further qualitative results on additional datasets. Figure 4 shows 3D point predictions on the Dur360BEV dataset [25] and 360 images downloaded from the internet.¹ The Dur360BEV dataset employs two fisheye cameras facing opposite directions, from which we obtain 360 images using its official conversion code. However, this conversion introduces noticeable distortions. The quality of the resulting 360 images directly affects the initial predictions of perspective foundation models, which in turn influences the performance of our method. Because our method is a general optimization framework, improvements in future perspective foundation models are expected to further enhance its overall performance.

G Thorough Discussion on RPG360

In this section, we provide a thorough discussion associated with our study.

G.1 Potential Societal Impacts

The results of our work demonstrate the effectiveness of our approach in improving depth estimation for omnidirectional images. Our graph-based optimization method shows that perspective foundation models can be effectively leveraged for omnidirectional depth estimation. We believe this work will serve as a foundation for future research in omnidirectional vision tasks. However, the datasets required to train perspective foundation models involve licensing restrictions, and it is therefore essential to ensure that they are used strictly for research purposes.

¹

Luca Biada, <https://www.flickr.com/photos/pedroscreeamervsky/6873256488/>, CC BY 2.0 DEED
Luca Biada, <https://www.flickr.com/photos/pedroscreeamervsky/6839817342/>, CC BY 2.0 DEED

G.2 Licenses for existing assets

We compare our work with SliceNet² [19], UniFuse³ [11] (MIT License), ACDNet⁴ [31], BiFuse++⁵ [23] (MIT License), Elite360D⁶ [1], Depth Anywhere⁷ [24] (Apache License 2.0), and 360MonoDepth⁸ [20] (MIT License). We use the Matterport3D⁹ [5] (non-commercial academic use only), Stanford2D3D¹⁰ [4] (non-commercial academic use only), and 360Loc¹¹ [10] datasets.

References

- [1] Hao Ai and Lin 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*, pages 9926–9935, 2024.
- [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.
- [3] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiro 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.
- [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] 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.
- [6] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.
- [7] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.
- [8] Hao Guan and William AP Smith. Structure-from-motion in spherical video using the von mises-fisher distribution. *IEEE Transactions on Image Processing*, 26(2):711–723, 2016.
- [9] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024.
- [10] Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22314–22324, 2024.
- [11] 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(2):1519–1526, 2021.
- [12] Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, and Suyong Yeon. Edm: Equirectangular projection-oriented dense kernelized feature matching. *arXiv preprint arXiv:2502.20685*, 2025.

²<https://github.com/crs4/SliceNet>

³<https://github.com/alibaba/UniFuse-Unidirectional-Fusion>

⁴<https://github.com/zcq15/ACDNet>

⁵<https://github.com/fuenwang/BiFusev2>

⁶<https://github.com/haoai-1997/Elite360D>

⁷<https://github.com/albert100121/Depth-Anywhere>

⁸<https://github.com/manurare/360monodepth>

⁹<https://github.com/niessner/Matterport>

¹⁰<https://github.com/alexsax/2D-3D-Semantics>

¹¹<https://github.com/HuajianUP/360Loc>

- [13] Dongki Jung, Jaehoon Choi, Yonghan Lee, and Dinesh Manocha. Im360: Textured mesh reconstruction for large-scale indoor mapping with 360° cameras, 2025. URL <https://arxiv.org/abs/2502.12545>.
- [14] 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.
- [15] 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.
- [16] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2801–2810, 2022.
- [17] Evin Pınar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. arXiv preprint arXiv:2203.08122, 2022.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In European Conference on Computer Vision, pages 195–211. Springer, 2022.
- [22] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. arXiv preprint arXiv:2208.01489, 2022.
- [23] 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.
- [24] Ning-Hsu Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. arXiv preprint arXiv:2406.12849, 2024.
- [25] E Wenke, Chao Yuan, Li Li, Yixin Sun, Yona Falinie A Gaus, Amir Atapour-Abarghouei, and Toby P Breckon. Dur360bev: A real-world 360-degree single camera dataset and benchmark for bird-eye view mapping in autonomous driving. In 2025 IEEE International Conference on Robotics and Automation (ICRA), pages 3737–3744. IEEE, 2025.
- [26] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2330–2337. IEEE, 2020.
- [27] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10371–10381, 2024.
- [28] Ilwi Yun, Hyuk-Jae Lee, and Chae Eun Rhee. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3224–3233, 2022.
- [29] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6101–6112, 2023.

- [30] 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.
- [31] 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, volume 36, pages 3653–3661, 2022.
- [32] 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.