PanoVGGT: Feed-Forward 3D Reconstruction from Panoramic Imagery

Supplementary Material

A. Training and Evaluation Details

Training Details. We adopt DINOv2 ViT-B/14 as the image encoder [5], configured with an input height of 336 pixels (respecting each dataset's native ERP aspect ratio), a patch size of 14, and an embedding dimension of 768. The proposed Geometry Aggregator consists of 24 alternating-attention blocks. Training is performed on a consolidated panoramic corpus comprising Matterport3D [4], Stanford2D3D [2], Structured3D [11], and PanoCity, with respective training and validation splits of $4800/2400/2400/16800$ and $480/240/1400/2400$.

A dynamic set-based dataloader is employed, where each training sample randomly selects 2–24 panoramas (validation uses 2–12). Color augmentations include random color jittering, grayscale conversion, and gamma adjustments. Geometric augmentations employ physically consistent $SO(3)$ spherical rotations via spherical resampling, ensuring alignment across RGB, depth, and pose modalities. We use AdamW for optimization with a peak learning rate of 5×10^{-5} and a weight decay of 0.05. Training follows a 5% linear warm-up schedule ($1\times10^{-8} \rightarrow 5\times10^{-5}$), followed by a 95% cosine decay ($5\times10^{-5} \rightarrow 1\times10^{-8}$). All remaining hyperparameters follow those reported in the main paper.

For fair comparison, we retrained a panoramic variant of π^3, denoted π^{3*}, by adapting its geometric head to equirectangular imagery while keeping datasets, augmentations, and optimization settings identical to ours, except that spherical-rotation augmentation was disabled for π^{3*} [9]. All other baselines rely on official implementations or publicly released pretrained weights under their default configurations (e.g., VGGT [7], MoGe [8], PanDA [3], UniK3D [6]).

Evaluation Details. All evaluations use input resolutions of 518×1036 pixels. For the zero-shot depth estimation benchmark on Pano3D [1], we use single-view inputs, as the dataset provides no cross-view geometric supervision. For 3D point cloud evaluation, we assess multi-view fusion quality by sampling 10 frames per scene for PanoCity and 3 frames per scene for both Stanford2D3D [2] and Matterport3D [4]. The network predicts point maps in both the camera frame and a unified world (anchor) frame; for evaluation, all points are transformed into the world coordinate system using the predicted camera poses. We maintain consistent evaluation protocols across all datasets, using the same voxel size, nearest-neighbor distance metric, and Accuracy/Completeness thresholds as defined in the main paper, with any exceptions noted beneath the respective tables.

We note that Stanford2D3D and Matterport3D were not originally designed for multi-view panoramic reconstruction and therefore contain scenes with limited or negligible overlap between panoramas even after re-splitting and post-processing [2, 4]. Despite the scarcity of tightly aligned panoramic 3D data, these datasets are retained for training, validation, and evaluation to broaden coverage and to test model robustness under varying levels of viewpoint overlap.

B. Additional Depth Results

We extend the depth estimation experiments by adding the retrained panoramic variant π^{3*} [9] to the evaluation. All methods follow identical preprocessing, data splits, and evaluation protocols as in the main paper to ensure direct comparability. Table 1 presents the updated monocular results on *Matterport3D* [4], *Stanford2D3D* [2], *Structured3D* [11], and *PanoCity*. Across all datasets, PanoVGGT maintains the best Abs Rel and δ<1.25 accuracy, while π^{3*}—although retrained on panoramic imagery—remains clearly inferior. This shows that simply adapting a pinhole-based architecture to panoramic inputs does not yield precise depth estimation, whereas the proposed PanoVGGT effectively learns panoramic geometry from diverse training data.

To further test cross-dataset generalization, we evaluate both models on the Pano3D (GibsonV2) benchmark [1, 10] and report the results in Tab. 2. Under the Scale-only configuration, π^{3*} achieves moderate accuracy but remains less stable than PanoVGGT, which attains lower errors across all metrics. With Scale+Shift alignment, π^{3*} improves slightly yet still lags behind in accuracy and structural consistency. These results confirm that direct retraining on panoramic data provides limited benefit, while our explicit panoramic design maintains robustness under distribution shift.

C. Additional Point-Cloud Results

We extend the point-cloud evaluation to two indoor panoramic datasets, Matterport3D [4] and Stanford2D3D [2]. To thoroughly investigate alternative 360° projections and address how perspective-only approaches perform on panoramic data, we introduce an additional baseline denoted as $\pi^{3\dagger}$. Following the dodecahedral projection protocol of MoGe, we project each equirectangular panorama into 12 overlapping pinhole views. Consequently, a standard 3-panorama input is expanded into 36 individual perspective images and evaluated

Table 1. Monocular depth estimation performance including the retrained panoramic variant π^{3}* [9] on *Matterport3D* [4], *Stanford2D3D* [2], *Structured3D* [11], and *PanoCity*. **Bold** = best, underline = second best.

Model	Input	Matterport3D (Indoor)		Stanford2D3D (Indoor)		Structured3D (Indoor)		PanoCity (Outdoor)	
		Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑
π^{3}* [9]	Monocular	0.0940	0.9142	0.0852	0.9291	0.0652	0.9649	0.0834	0.9161
PanoVGGT (Ours)	Monocular	0.0884	0.9157	**0.0711**	**0.9392**	0.0438	0.9728	0.0312	0.9713
PanoVGGT (Ours)	Multi-view	**0.0840**	**0.9266**	0.0778	0.9323	**0.0400**	**0.9870**	**0.0196**	**0.9812**

Table 2. Zero-shot depth results on *Pano3D (GibsonV2)* [1, 10] with π^{3}* (retrained panoramic variant) [9]. The top part reports results using Scale-only alignment; the bottom part uses Scale+Shift alignment.

Method	Abs Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑
π^{3}* [9]	0.1879	0.3743	0.9085	0.9717
PanoVGGT (ours)	**0.0869**	**0.3069**	**0.9223**	**0.9859**
π^{3}* [9]	0.1777	0.3639	0.9197	0.9759
PanoVGGT (ours)	**0.0833**	**0.3015**	**0.9299**	**0.9894**

using the pre-trained π^{3} model. We compare this against our proposed PanoVGGT, the original π^{3}, and its re-trained panoramic variant π^{3}*.

Each method is evaluated using Accuracy (Acc), Completion (Comp), and their average (Overall). Both mean and median scores are reported, where lower values indicate better performance. Tables 3 and 4 summarize the quantitative results. While processing dense pinhole splits ($\pi^{3\dagger}$) improves over the naive π^{3} baseline, it incurs significant computational overhead. Across both datasets, PanoVGGT operating directly on native equirectangular projections achieves the lowest average errors, demonstrating superior geometric accuracy and view-to-view consistency. All experiments are conducted with the same voxel size, nearest-neighbor distance metric, and Acc/Comp thresholds as adopted in the main paper, unless otherwise specified.

Table 3. Point-cloud results on *Matterport3D* [4]. † denotes the perspective-based model evaluated on panoramas using the dodecahedral projection protocol of MoGe.

Method	Acc ↓		Comp ↓		Overall ↓	
	Mean	Med	Mean	Med	Mean	Med
π^{3} [9]	0.4027	0.3853	0.9964	0.8803	0.6995	0.6328
π^{3}* [9]	0.2661	0.2199	0.3351	0.2378	0.3006	0.2288
$\pi^{3\dagger}$ [9]	0.1843	0.1420	0.2647	0.1857	0.2245	0.1638
PanoVGGT (global points)	**0.1743**	**0.1231**	**0.1530**	**0.0890**	**0.1636**	**0.1060**
PanoVGGT (local points)	0.1750	0.1242	**0.1530**	0.0932	0.1640	0.1087

Qualitative Visualizations. While qualitative results for Matterport3D [4] are provided in the main paper, Figs. 1 and 2 show additional multi-view point-cloud reconstructions on Stanford2D3D [2] and PanoCity. Each panel compares the original π^{3} [9], its retrained panoramic variant π^{3}* [9], the pinhole-split variant $\pi^{3\dagger}$ [9], and the proposed PanoVGGT. For the indoor Stanford2D3D dataset, two unordered panoramas are used as input, and reconstructions are displayed from a single viewpoint for clarity, with ceilings removed to expose interior structural details. For PanoCity, ten unordered panoramas form a long-trajectory input sequence, and only the PanoVGGT reconstructions are rendered from two viewpoints to illustrate large-scale outdoor geometry and fine structural details. These visualizations demonstrate cross-view geometric consistency and enable direct comparison of reconstruction quality across methods.

Table 4. Point-cloud results on *Stanford2D3D* [2]. † denotes the perspective-based model evaluated on panoramas using the dodecahedral projection protocol of MoGe.

Method	Acc ↓		Comp ↓		Overall ↓	
	Mean	Med	Mean	Med	Mean	Med
π^{3} [9]	0.5087	0.4667	0.9394	0.9103	0.7241	0.6885
π^{3}* [9]	0.2590	0.2021	0.3143	0.2245	0.2866	0.2133
$\pi^{3\dagger}$ [9]	0.2359	0.1942	0.3668	0.2870	0.3013	0.2406
PanoVGGT (global points)	**0.2087**	**0.1752**	0.2624	**0.1943**	0.2355	**0.1848**
PanoVGGT (local points)	0.2109	0.1786	**0.2590**	**0.1943**	**0.2349**	0.1865

The original π^{3} [9], developed for pinhole imagery, directly interprets equirectangular panoramas as pinhole inputs and produces geometrically invalid reconstructions with severe warping and structural distortion. The retrained π^{3}* [9] partially adapts but still struggles to produce clean geometry and accurate pose estimates: it often fails to converge on PanoCity and yields incomplete or inconsistent structures on indoor scenes. While the dodecahedral projection approach ($\pi^{3\dagger}$) mitigates severe warping by processing local pinhole views, it introduces visible structural discontinuities among the decoupled patches within each panorama. More severely, it yields inaccurate relative poses between different panoramic views, causing the reconstructed point clouds of the same scene to be misaligned and fail to overlap correctly. In contrast, PanoVGGT reconstructs globally consistent 3D scenes with correct geometry and seamless alignment across unordered multi-view panoramas.

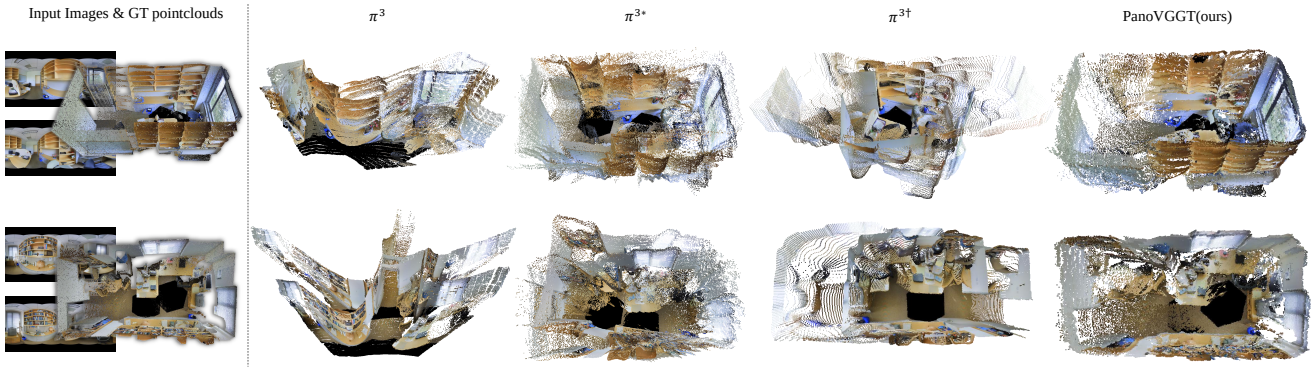


Figure 1. Multi-view point-cloud reconstructions on the Stanford2D3D dataset [2] using two unordered panoramic inputs. From left to right: π^3 [9], π^{3*} [9], $\pi^{3\dagger}$ [9], and PanoVGGT. Our method achieves higher geometric accuracy and cross-view consistency than the baselines.

D. Additional Visualization of PanoCity

Figure 3 provides an overview of the five key data modalities in the PanoCity dataset, arranged in the order of data acquisition and processing. The first row shows representative urban scenes from three different cities or regions (City-1, City-2, City-3), together with their approximate coverage areas (15km^2, 113km^2, 6km^2). These scenes span diverse urban contexts, including dense downtown blocks, residential districts, and major arterial roads. The second row presents example equirectangular panoramic RGB images sampled along the acquisition routes, capturing a wide range of lighting conditions, weather variations, and complex architectural structures. The third row visualizes the corresponding ground-truth 3D point clouds for these scenes, which provide accurate geometric reference for training and evaluation. The fourth row shows the associated ground-truth panoramic depth maps derived from the point clouds; depth is color-coded (e.g., from purple to yellow), making near objects such as trees and roads and far objects such as tall buildings clearly distinguishable in depth. The fifth row illustrates the camera trajectories in bird's-eye view, highlighting long-range paths with large baselines and substantial viewpoint changes. Together, these visualizations demonstrate that PanoCity offers large-scale, metrically calibrated panoramic RGB–depth–point-cloud data with realistic urban diversity, making it a suitable testbed for robust panoramic 3D reconstruction and depth estimation.

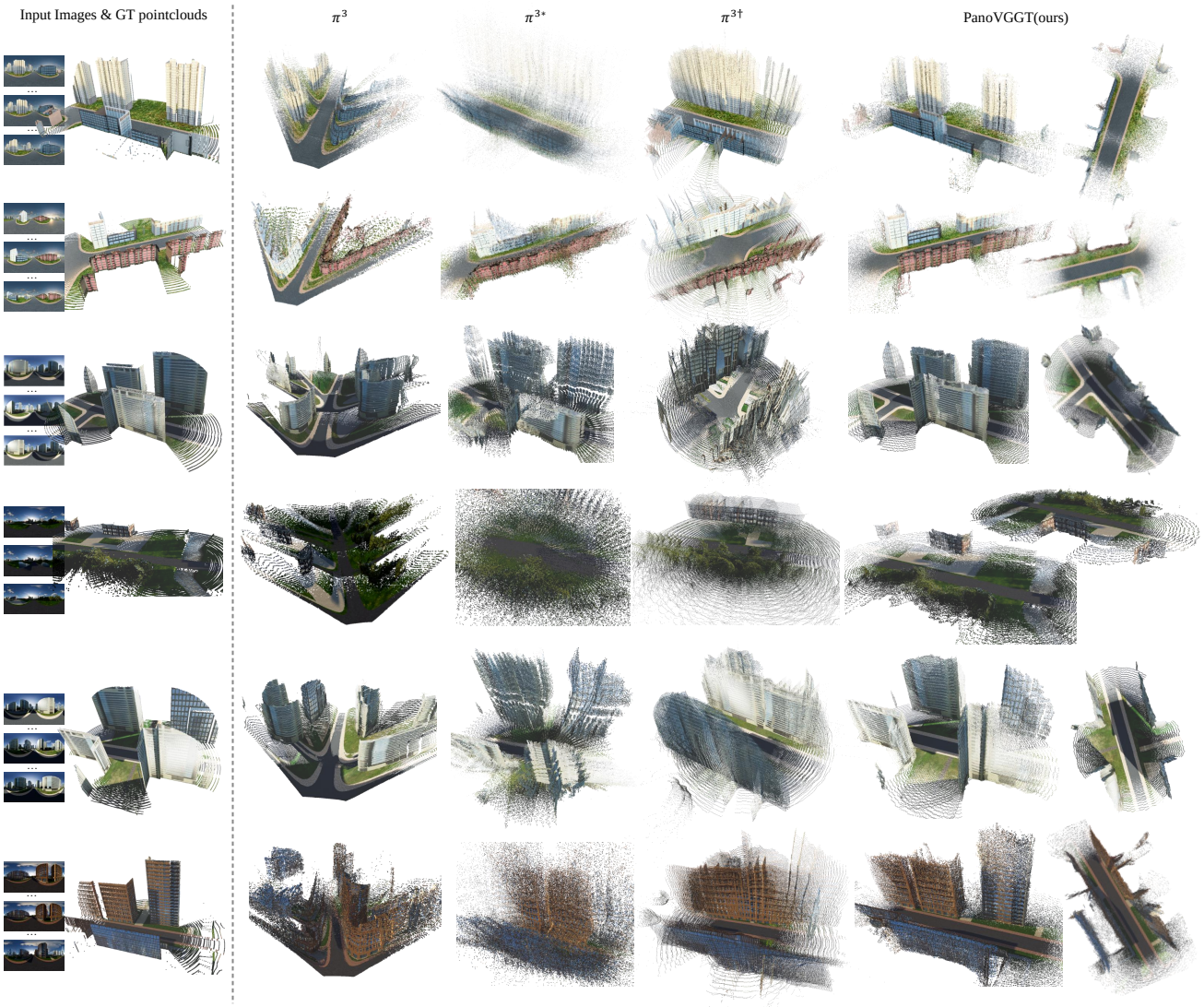
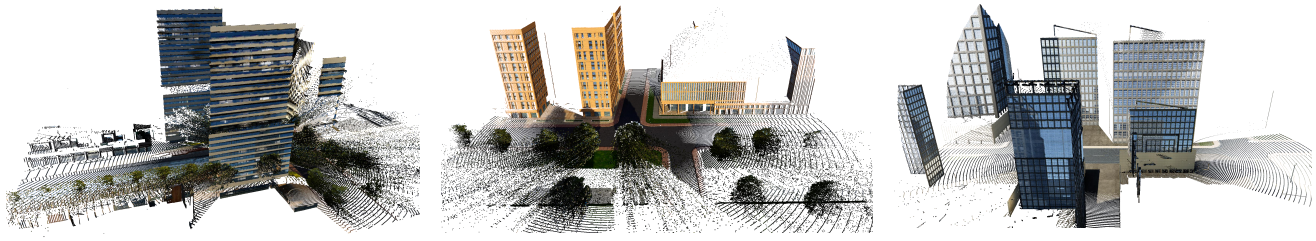


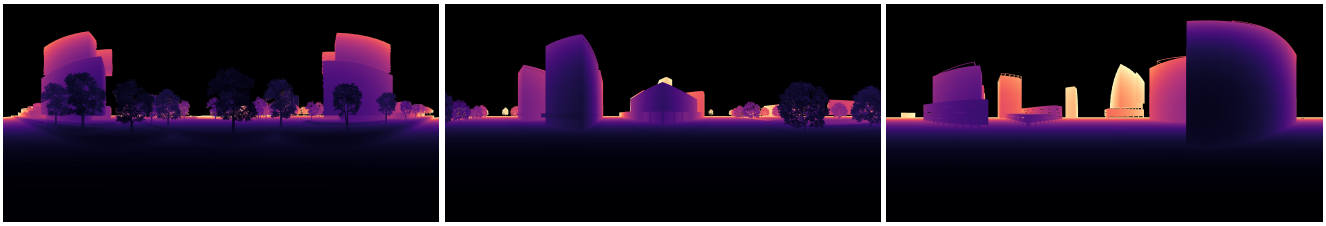
Figure 2. Multi-view point-cloud reconstructions on the PanoCity dataset using ten unordered panoramic inputs. From left to right: π^3 [9], π^{3*} [9], $\pi^{3\dagger}$ [9], and PanoVGGT. The baseline methods struggle to learn accurate geometry on this long-trajectory setup, whereas PanoVGGT reconstructs large-scale outdoor scenes with coherent structure and accurate alignment across views.

Representative urban scenes

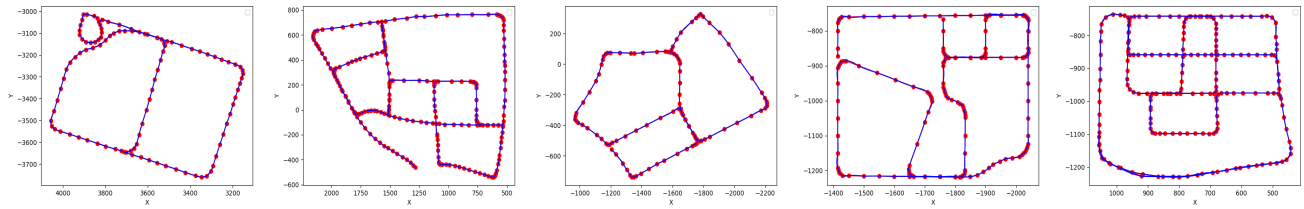
Example panoramic images



Corresponding Ground Truth Point Clouds



Corresponding Ground Truth Panoramic Depth Maps



Camera trajectories

Figure 3. Overview of the five core modalities in the PanoCity dataset. Row 1: representative urban scenes from three captured cities or regions (City-1, City-2, City-3), annotated with their coverage areas (15km^2, 113km^2, 6km^2). Row 2: example panoramic RGB images collected along the acquisition routes, showing diverse lighting, weather, and façade structures. Row 3: corresponding ground-truth 3D point clouds for the same scenes. Row 4: ground-truth panoramic depth maps derived from the point clouds, with depth encoded by color. Row 5: camera trajectories in bird's-eye view, illustrating long-range paths with wide baselines and varied viewpoints.

References

[1] 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. 1, 2

[2] 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. 1, 2, 3

[3] 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. 1

[4] 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. 1, 2

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[6] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025. 1

[7] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1

[8] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 1

[9] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 1, 2, 3, 4

[10] Fei Xia, Chengshu Li, Kevin Chen, William B Shen, Roberto Martın-Martın, Noriaki Hirose, Amir R Zamir, Li Fei-Fei, and Silvio Savarese. Gibson env v2: Embodied simulation environments for interactive navigation. *Stanford University, Tech. Rep.*, 2019. 1, 2

[11] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535. Springer, 2020. 1, 2