Supplementary Material for VGGT-360

Jiayi Yuan[1], Haobo Jiang[2], De Wen Soh[1], Na Zhao*[1]

[1] Singapore University of Technology and Design, [2] Nanyang Technological University

jiayi_yuan@mymail.sutd.edu.sg, haobo.jiang@ntu.edu.sg, dewen_soh@sutd.edu.sg, na_zhao@sutd.edu.sg

This supplementary document provides additional experimental analyses and visualizations to further support the claims presented in the main paper. Specifically, it includes: (i) extended ablation studies for our VGGT-360 (Section 1); (ii) additional qualitative results across various backbones, including VGGT [9], Fastvggt [7], and π^3 [11] (Section 2); (iii) further visual comparisons against state-of-the-art (SOTA) methods (Section 3); and (iv) a discussion of limitations and future directions (Section 4).

1. Additional Ablation Studies

Effect of Different Structure-Detection Operators. We investigate how different structure-detection operators influence the computation of both the uncertainty score and the structure-saliency confidence map in VGGT-360. Our method relies on gradient-derived structural cues to assess geometric ambiguity across perspective base views. While the main paper adopts the Sobel operator for its balance of simplicity and stability, here we evaluate several widely used alternatives, including Canny [2], Gabor filters [4], and Scharr [6]. For each operator, we compute the corresponding structure responses and apply the same normalization procedure as in Eq. (1) of the main paper. As shown in Table 1, all operators deliver comparable performance, indicating that VGGT-360 is largely insensitive to the choice of structure extractor, while Sobel provides the best trade-off between accuracy and efficiency.

Ablation on the Number of Neighbor Views Per Selected Base View. To determine how many supplemental neighbor views are needed for each selected base view $v_b^* \in \mathcal{B}^*$, we evaluate configurations that generate one, two (default), or three neighbors around each v_b^*. These neighbors are created by applying small yaw–pitch perturbations, allowing VGGT to observe the uncertain region from slightly varied viewpoints. As shown in Table 1, $\mathcal{N}_{nv}$=2 achieves the best balance between accuracy and runtime.

* Corresponding author.

Table 1. Ablation studies of our VGGT-360 on Stanford2D3D [1].

Method	Abs Rel↓	RMSE↓	Time
Effect of Different Structure-Detection Operators			
Baseline + Sobel	0.067	0.316	1.50s
Baseline + Canny	0.068	0.319	1.47s
Baseline + Gabor Filters	0.066	0.314	1.63s
Baseline + Scharr	0.068	0.319	1.51s
Effect of Neighbor View Count per Base View			
$\mathcal{N}_{nv} = 1$	0.072	0.323	1.22s
$\mathcal{N}_{nv} = 2$	0.067	0.316	1.50s
$\mathcal{N}_{nv} = 3$	0.067	0.315	1.71s
Effect of Edge-Band Width			
$m = 0$	0.069	0.322	1.50s
$m = 0.03$	0.068	0.320	1.50s
$m = 0.05$	0.067	0.316	1.50s
$m = 0.08$	0.068	0.318	1.50s

Ablation on Edge-Band Width. In our structure-saliency enhanced attention module, the parameter m controls the width of the edge-band prior $\mathbf{E}$ used to emphasize pixels near the boundaries of each perspective view. A larger m expands the band of pixels that receive boosted confidence, while a smaller m restricts the enhancement to a narrower border region. To understand its influence, we evaluate the performance under multiple choices of m, including m=0 (no edge-band), m=0.03, m=0.05 (default), and m=0.1. As shown in Table 1, setting m=0.05 achieves the best overall performance, indicating that an appropriately sized edge-band prior provides the optimal balance between boundary-focused enhancement and avoiding excessive amplification of non-salient regions.

2. Visualizations Across Different Backbones

To verify the generality of VGGT-360, we further evaluate our training-free modules on multiple 3D founda-

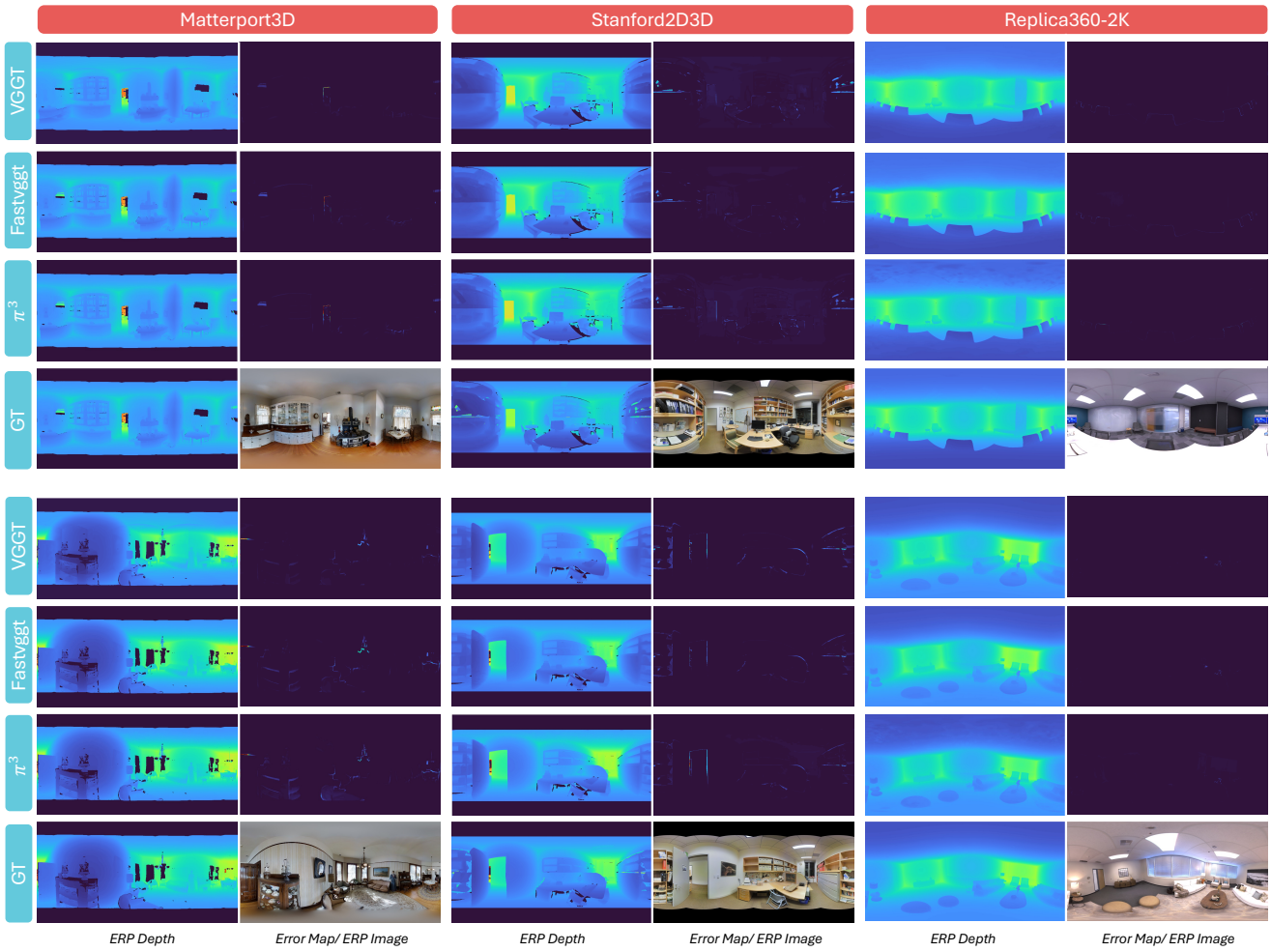

Figure 1. Qualitative comparisons using different backbone models, including VGGT [9], Fastvggt [7], and π^3 [11].

tion backbones, including VGGT [9], Fastvggt [7], and π^3 [11]. As shown in Fig. 1, despite the architectural differences among these models, VGGT-360 consistently produces sharp boundaries and coherent geometry in challenging panoramic regions such as geometry-sparse walls and fine structural details. These results demonstrate both the effectiveness of our 3D-aware design and the broad applicability of our method across different backbones.

3. Additional Visualizations with the SoTA

To complement the qualitative comparison shown in the main paper, we include additional visualizations across Matterport3D [3], Stanford2D3D [1], and Replica360-2K [5, 8] datasets. As illustrated in Fig. 2, VGGT-360 continues to outperform both the supervised state-of-the-art method Depth-Anywhere [10] and the training-free baseline 360MD [5]. Our predictions exhibit more stable geometry, cleaner structural boundaries, and fewer distortions across diverse indoor layouts. The error maps consistently show that VGGT-360 yields the lowest reconstruction error among all methods, confirming that the improvements observed in the main paper hold across a wide range of additional panoramic scenes.

4. Limitations and Future Directions

Limitations. The performance and runtime of VGGT-360 are fundamentally bounded by the capability of the underlying 3D foundation model. Although our proposed modules improve structural consistency and stabilize panoramic reasoning, the method still inherits the representational limits of the backbone, which remain the primary factors affecting final depth quality. In addition, the computational cost of VGGT-360 scales with the inference speed of the backbone, implying that faster or panoramic-specialized 3D models would directly translate into improved efficiency.

Future Directions. A promising direction is to extend

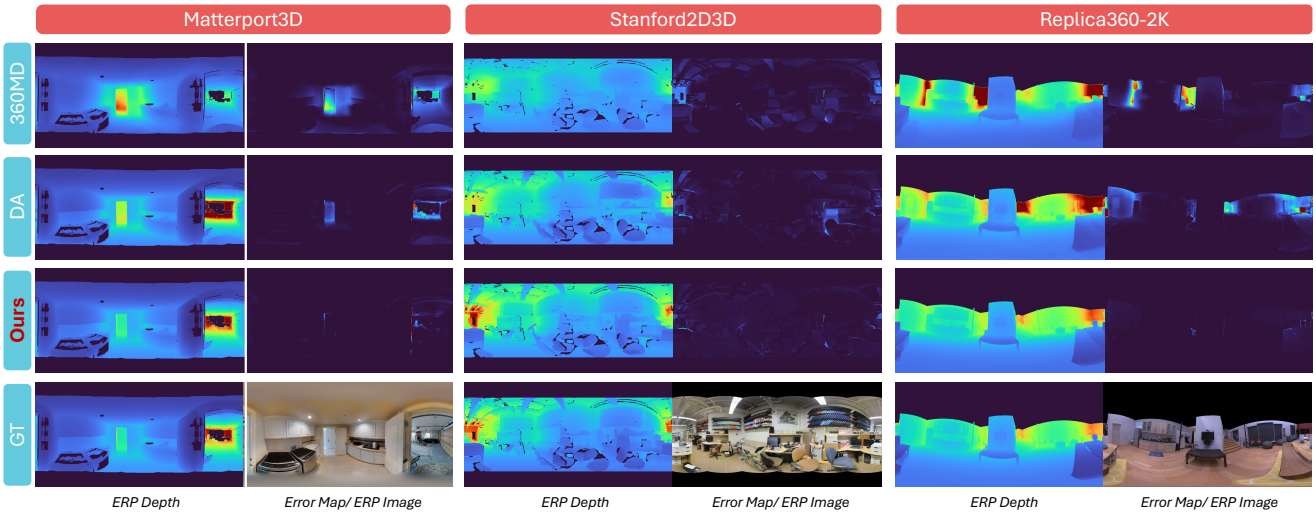

Figure 2. Qualitative comparisons with the state-of-the-art supervised method Depth-Anywhere (DA) [10] and the training-free method 360MD [5] across three indoor datasets: Matterport3D [3], Stanford2D3D [1], and Replica360-2K [5, 8].

VGGT-360 toward a unified panoramic perception framework that goes beyond depth estimation. The proposed adaptive view-generation strategy and structure-aware 3D reasoning can naturally support a wider range of 360° tasks, including panoramic semantic segmentation, surface normal prediction, and multi-modal 3D scene understanding. Unifying these tasks within a single geometry-consistent framework would enable more comprehensive omnidirectional scene interpretation and benefit downstream applications in robotics, VR, and large-scale 3D reconstruction.

References

[1] 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

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009. 1

[3] 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. *3DV*, 2017. 2, 3

[4] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, 93(26):429–441, 1946. 1

[5] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *CVPR*, pages 3762–3772, 2022. 2, 3

[6] Hanno Scharr. Optimal operators in digital image processing. 2000. 1

[7] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 1, 2

[8] 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. 2, 3

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 1, 2

[10] Ning-Hsu Albert Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *NeurIPS*, 37: 127739–127764, 2024. 2, 3

[11] 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. *arXiv preprint arXiv:2507.13347*, 2025. 1, 2