
Fin3R: Fine-Tuning Feed-Forward 3D Reconstruction Models via Monocular Knowledge Distillation

– *Supplementary Material* –

Anonymous Author(s)

Affiliation

Address

email

1 A Experiment Details

2 A.1 Training Details

3 In all experiments, we set both the rank and alpha of LoRA to 8.

4 **DUST3R.** Since DUST3R doesn't have a dedicated self-view head for canonical view estimation,
5 we use DUST3R's first viewpoint pointmap regression head for distillation. Training is performed
6 at a resolution of 512 width, with aspect ratios (e.g., 16:9, 4:3) randomly sampled for each batch.
7 During each epoch, we randomly sample 20,000 pairs from the SA-1B [4] dataset, 1,000 pairs from
8 the Hypersim [8] dataset, and 1,000 pairs from the TartanAir [16] dataset. The model is fine-tuned for
9 10 epochs. The learning rate is initialized at 1e-4 with a one-epoch warm-up phase and is gradually
10 decayed to a minimum of 1e-6. A batch size of 2 per GPU is used, and gradients are accumulated
11 over 8 iterations to achieve an effective batch size of 64.

12 **CUT3R/VGGT.** We compute the distillation loss using the self-view pointmap head for CUT3R
13 and the depth head for VGGT, following the same dataset configuration as in DUST3R fine-tuning.
14 CUT3R is trained at a resolution of 512 width, while VGGT is trained at a resolution of 518 width.
15 The model is fine-tuned for 10 epochs with an initial learning rate of 1e-4, which is warmed up for
16 one epoch and then gradually decayed to a minimum of 1e-6. Additionally, the sequence length is
17 dynamically selected between 2 and 8, with the product of batch size and sequence length fixed at 8.
18 The accumulation iteration is changed accordingly to ensure an effective total batch size of 64.

19 A.2 Evaluation Details

20 **Monocular Depth Estimation.** We follow the evaluation protocol from MoGe [14] to assess our
21 models. For DUST3R, we duplicate the input images and use the z value from the view-1 pointmap
22 head as the predicted depth. For CUT3R, depth is obtained from the z value of the self-view pointmap
23 head, and for VGGT, we use the output of the depth head. Since these models are trained at resolutions
24 of 512 width (or 518 width for VGGT), the original images are resized accordingly for evaluation.
25 Although this differs from the standard MoGe protocol, which evaluates at higher resolutions, we
26 ensure that both the base model and our fine-tuned models share the same settings. Furthermore, we
27 exclude evaluation datasets such as Sintel and Spring since DUST3R and VGGT are not designed for
28 dynamic scenes.

29 **Two-view Evaluation.** We extract two-view correspondences using the nearest neighbor matching
30 strategy from DUST3R, which leverages geometric distance and is well-suited for assessing our
31 enhanced geometry. We avoid using VGGT's tracking head for matching for two main reasons. First,
32 the current release of VGGT's tracking head does not perform as good as the version reported in

Table S1: Quantitative results for multi-view pose estimation on the CO3Dv2 dataset. "Ours" signifies the integration of our finetuning method. Best results in each session are highlighted in **bold**.

Methods	CO3Dv2		
	RRA@5	RTA@5	AUC@30
DUST3R [15]	80.49	75.22	81.03
DUST3R+Ours	85.75	78.02	82.83
CUT3R [13]	70.83	64.39	74.10
CUT3R+Ours	69.65	61.66	72.80
VGGT [10]	95.20	84.28	88.35
VGGT+Ours	95.47	84.18	88.77

the original paper¹. Second, in the Scannet-1500 relative pose estimation task, our geometry-based correspondence method outperforms the tracking-based approach described in the original VGGT paper. Furthermore, we plan to fine-tune the tracking head using our stronger encoder, which we believe can provide more accurate and robust features to further enhance tracking performance.

Multi-View Pose Estimation. We evaluate our method primarily on the RealEstate10k dataset [19], following the procedure in VGGsFM [11] that involves randomly sampling 10 frames from each sequence for pose evaluation. Since some of the original YouTube links in RealEstate10k are unavailable, our evaluation is conducted on 1,756 out of the original 1,800 scenes.

Ablation Mix Dataset. For the ablation study, we replace the SA-1B dataset [5] with a mixed dataset composed of MegaDepth [6], CO3Dv2 [7], ARkitScene [1], Scannet++ [18], Scannet [3], VirtualKITTIv2 [2], BlendedMVS [17], and StaticThings3D [9]. Each dataset is equally weighted, providing coverage that is comparable to the DUST3R training set.

B Additional Experiments

We also conduct experiments on multi-view pose estimation using the CO3Dv2 dataset [7]. Following the evaluation protocol in PoseDiffusion [12], we select the first 10 frames from each sequence for evaluation. The results are presented in Table S1. Our fine-tuning improves DUST3R by refining the geometry-based correspondence. However, the performance of CUT3R on CO3Dv2 is negatively affected, and the impact on VGGT is marginal. We suspect this is primarily because CO3Dv2 is used to train the pose head, causing it to strongly memorize the dataset.

C Additional Visualizations

We provide additional visualizations on diverse, in-the-wild data in Figures S1, S2, and S3 to demonstrate how our fine-tuning method robustly enhances the original baseline. More visualization can be found in the Supplementary Video, which includes flythrough of the multi-view reconstruction results.

¹<https://github.com/facebookresearch/vggt/issues/83>

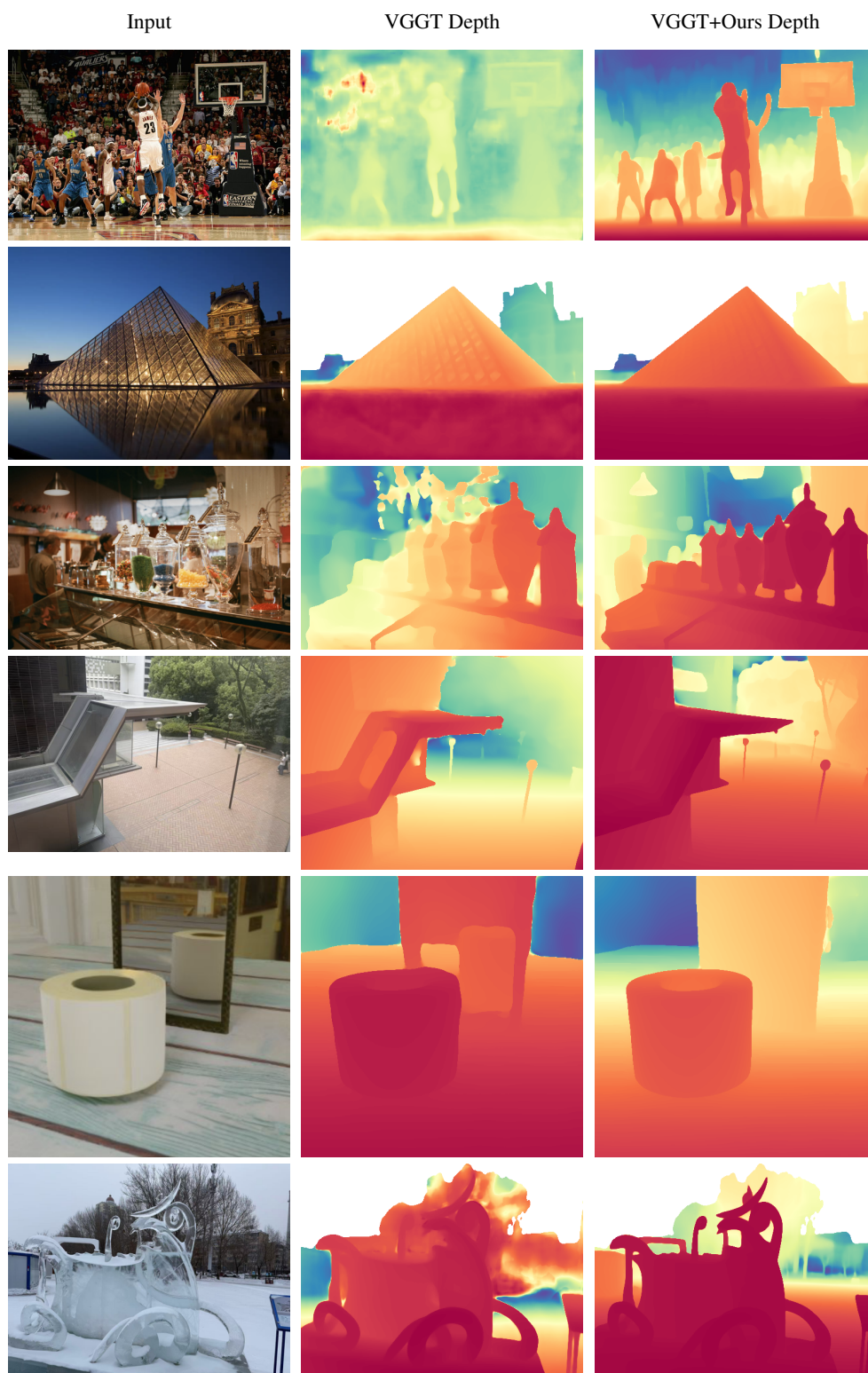


Figure S1: Additional Visualization of Depth Estimation Results: Input Images, Baseline (VGGT Depth), and Improved Method (VGGT+Ours Depth)

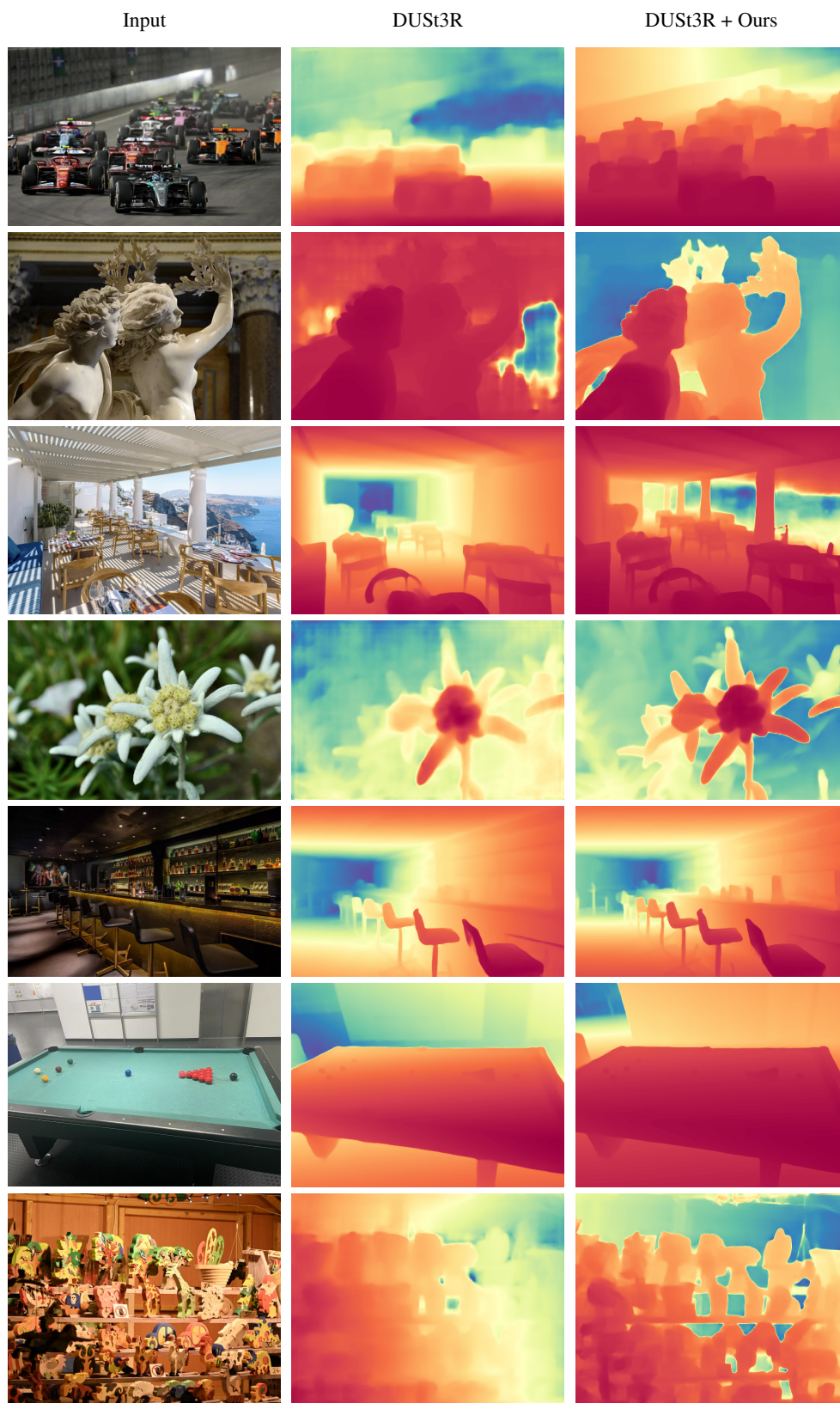


Figure S2: Additional Visualization of Depth Estimation Results: Input Images, Baseline (DUST3R Depth), and Improved Method (DUST3R+Ours Depth)

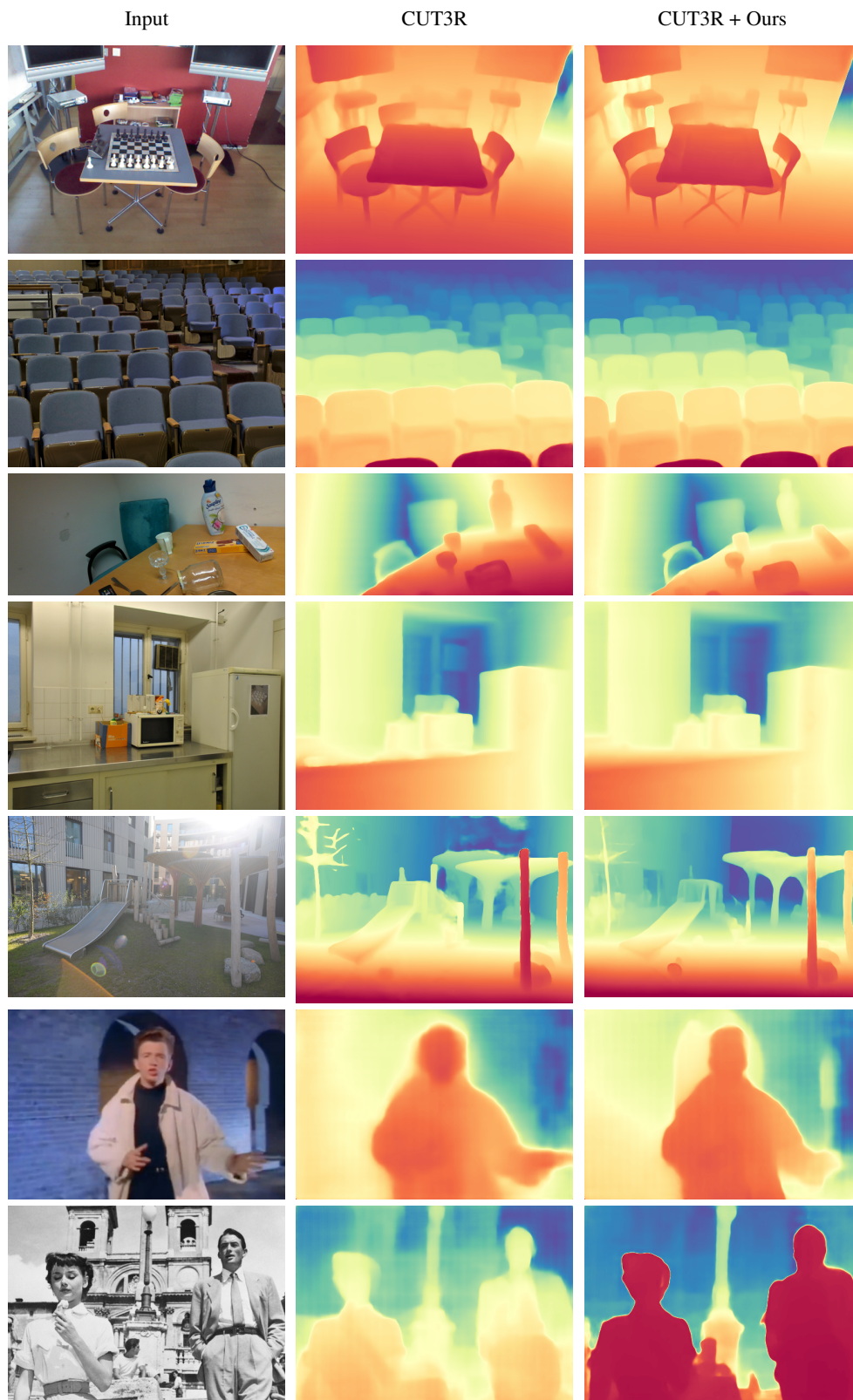


Figure S3: Additional Visualization of Depth Estimation Results: Input Images, Baseline (CUT3R Depth), and Improved Method (CUT3R+Ours Depth)

57 **D Proof for Long-Sequence Scale Uncertainty**

58 Let a 3D point in the world coordinate system be

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

59 with a multiplicative scale uncertainty modeled as

$$\hat{\mathbf{p}} = (1 + \delta)\mathbf{p}, \quad \delta \sim \mathcal{N}(0, \sigma^2),$$

60 where δ is a small perturbation. Under a rigid transformation characterized by rotation R and
61 translation T , the unperturbed point in the second view is given by

$$\mathbf{p}_2 = \mathbf{R}\mathbf{p} + \mathbf{T}.$$

62 When the uncertainty is introduced, the perturbed second-view point becomes

$$\hat{\mathbf{p}}_2 = \mathbf{R}\hat{\mathbf{p}} + \mathbf{T} = \mathbf{R}[(1 + \delta)\mathbf{p}] + \mathbf{T} = \mathbf{p}_2 + \delta(\mathbf{R}\mathbf{p}).$$

63 We define the rotated coordinates by writing

$$\mathbf{R}\mathbf{p} = \begin{bmatrix} \alpha \\ * \\ \beta \end{bmatrix},$$

64 where, due to the relationship $\mathbf{p}_2 = \mathbf{R}\mathbf{p} + \mathbf{T}$, the first and third components satisfy:

$$\alpha = X_2 - T_x, \quad \beta = Z_2 - T_z,$$

65 with $\mathbf{p}_2 \triangleq \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix}$.

66 Assuming a pinhole camera model with focal length f , the unperturbed horizontal image coordinate
67 is given by

$$u = \frac{f X_2}{Z_2}.$$

68 For the perturbed coordinates we express

$$X_2^\delta = X_2 + \delta \alpha, \quad Z_2^\delta = Z_2 + \delta \beta.$$

69 Thus, the image coordinate under perturbation is

$$u(\delta) = \frac{f(X_2 + \delta \alpha)}{Z_2 + \delta \beta}.$$

70 Our goal is to analyze the induced projection error,

$$\Delta u \triangleq u(\delta) - u,$$

71 without using a Taylor expansion. We begin by forming the exact difference:

$$\Delta u = \frac{f(X_2 + \delta \alpha)}{Z_2 + \delta \beta} - \frac{f X_2}{Z_2}.$$

72 By combining the terms over a common denominator, we have:

$$\Delta u = f \left(\frac{(X_2 + \delta \alpha)Z_2 - X_2(Z_2 + \delta \beta)}{Z_2(Z_2 + \delta \beta)} \right).$$

73 Expanding the numerator yields:

$$(X_2 + \delta \alpha)Z_2 - X_2(Z_2 + \delta \beta) = X_2Z_2 + \delta \alpha Z_2 - X_2Z_2 - \delta X_2 \beta = \delta (\alpha Z_2 - X_2 \beta).$$

74 Thus, the error simplifies to:

$$\Delta u = \delta f \frac{\alpha Z_2 - X_2 \beta}{Z_2(Z_2 + \delta \beta)}.$$

75 Substituting the expressions $\alpha = X_2 - T_x$ and $\beta = Z_2 - T_z$, we obtain:

$$\alpha Z_2 - X_2 \beta = (X_2 - T_x)Z_2 - X_2(Z_2 - T_z) = X_2 T_z - T_x Z_2.$$

76 Therefore, the error becomes:

$$\Delta u = \delta f \frac{X_2 T_z - T_x Z_2}{Z_2 (Z_2 + \delta (Z_2 - T_z))},$$

77 since $\beta = Z_2 - T_z$.

78 To gain further insight into the dependency on depth Z_2 , let us assume that along object boundaries
79 the ratio X_2/Z_2 remains approximately constant, i.e.,

$$X_2 \approx c Z_2,$$

80 for some constant c . Under this assumption, the numerator approximates as

$$X_2 T_z - T_x Z_2 \approx Z_2 (c T_z - T_x).$$

81 Substituting this back, we get:

$$\Delta u \approx \delta f \frac{Z_2 (c T_z - T_x)}{Z_2 (Z_2 + \delta (Z_2 - T_z))} = \delta f \frac{c T_z - T_x}{Z_2 + \delta (Z_2 - T_z)}.$$

82 For small δ , the term $\delta (Z_2 - T_z)$ in the denominator is negligible compared to Z_2 . That is,

$$Z_2 + \delta (Z_2 - T_z) \approx Z_2.$$

83 Thus, we arrive at the simplified expression:

$$\Delta u \approx \delta \frac{f (c T_z - T_x)}{Z_2}.$$

84 This result shows that the projection error Δu is inversely proportional to Z_2 , meaning that fore-
85 ground points (with small Z_2) experience larger epipolar displacements due to scale uncertainty—a
86 phenomenon we term foreground erosion. Moreover, our analysis demonstrates that, except for the
87 first view, the normalization process amplifies minor scale errors in the foreground; this amplification
88 results in substantial epipolar displacement and the erosion of fine details in these regions.

89 E Limitations

90 Although our method enhances fine geometric details, its boundary accuracy remains inferior to
91 that of MoGe [14], which produces sharper results. A mixed-teacher strategy or a better dedicated
92 distillation design may offer further improvements. Additionally, the current model supports only a
93 512/518 resolution, and scaling to higher resolutions remains a challenge for future work.

94 References

- 95 [1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas
96 Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a
97 diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In
98 *NeurIPS*, 2021. 2
- 99 [2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 2
- 100 [3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias
101 Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2
- 102 [4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson,
103 Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In
104 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026,
105 2023. 1

- [5] 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
- [6] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2
- [7] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 2
- [8] 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 *International Conference on Computer Vision (ICCV) 2021*, 2021. 1
- [9] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *2022 International Conference on 3D Vision (3DV)*, pages 637–645. IEEE, 2022. 2
- [10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2
- [11] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 2
- [12] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2
- [13] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state, 2025. 2
- [14] 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. *arXiv preprint arXiv:2410.19115*, 2024. 1, 7
- [15] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2
- [16] Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 1
- [17] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [18] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2
- [19] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2