

# MonoPatchNeRF: Improving Neural Radiance Fields With Patch-Based Monocular Guidance

## Supplementary Material

### 1. Training Details

We provide the detailed network architecture in Table 1. Our model contains a total of 16,385,392 parameters and is trained for 50,000 steps on ETH3D in 2.25 hours and 200,000 steps on TanksandTemples in 9 hours using a single Nvidia A40 GPU.

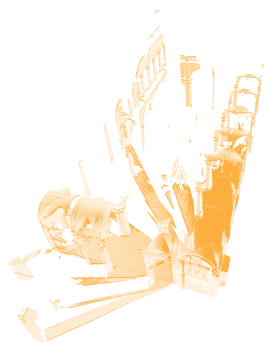
Table 1. **Network Architecture Details.** We have a spatial feature extractor that uses implementation of QFF [3] to extract features for each point. The extracted features are passed into the Density MLP to obtain per-point density and geometric features of length 15. The extracted geometry features are passed into Color MLP (along with the direction) and Surface Normal MLP to extract color  $\mathbf{c}$  and surface normal  $\mathbf{n}_\theta$  respectively.

Name	# Parameters	Input	Output Size
<b>Spatial Features</b>			
QFF [3]	32x80x80x80	$(x, y, z) \in \mathbb{R}^3$	32
<b>Density MLP</b>			
$D0_\theta$	32x16	QFF	16
$D1_\theta$	16x15	$ReLU(D0_\theta)$	15
$D2_\theta$	16x1	$ReLU(D0_\theta)$	$\sigma \in \mathbb{R}^1$
<b>Color MLP</b>			
$C0_\theta$	18x16	$D1_\theta + (\mathbf{d} \in \mathbb{R}^3)$	16
$C1_\theta$	16x3	$ReLU(C0_\theta)$	$\mathbf{c} \in \mathbb{R}^3$
<b>Surface Normal MLP</b>			
$S0_\theta$	15x16	$D1_\theta$	16
$S1_\theta$	16x3	$ReLU(S0_\theta)$	$\mathbf{n}_\theta \in \mathbb{R}^3$

### 2. Baseline Training Details

**MonoSDF:** We show how varying the initialization bias [1] of MonoSDF [17] affects its reconstruction quality. We used the author provided configs of TnT on ETH3D, but found that MonoSDF suffer from local minimum in some challenging scenes with original bias parameters due to the scene scale. We therefore contacted the MonoSDF authors and were advised to use a small bias for initialization. Figure 1 compares the reconstruction of MonoSDF given different bias parameters in a challenging scene *relief\_2* of ETH3D [8].

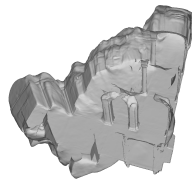
**Neuralangelo:** For Neuralangelo [4] experiment on ETH3D [8], we follow author provided setup on TanksAndTemples [2], but use a batch size of 4 instead of 16 to run on the same device settings. We additionally disable image embedding features, as we empirically found it to worsen the results. One visualization of results with different batch size is present in Figure 2. The  $F$ -score<sub>2cm</sub>,  $F$ -score<sub>5cm</sub> of high and low batch size results are 1.53, 11.5 and 1.46, 11.8.



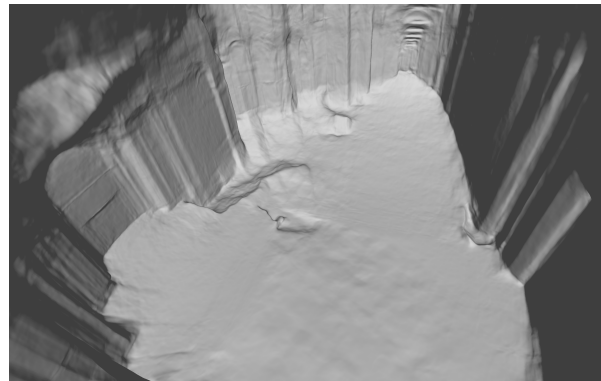
Ground Truth



Ours

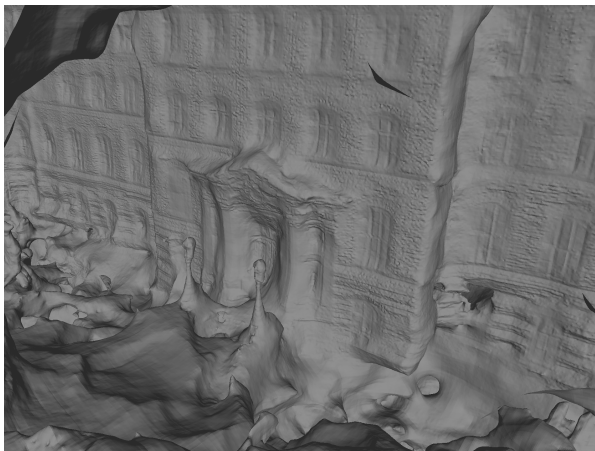


MonoSDF<sub>0.1</sub>

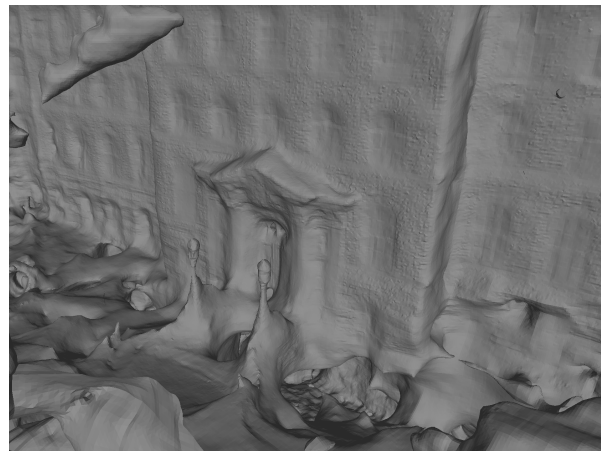


MonoSDF<sub>1.0</sub>

Figure 1. **Visualization of MonoSDF with different parameters in *Relief\_2*.** MonoSDF<sub>1.0</sub> denotes MonoSDF trained with the default bias parameter (1.0) in the code provided by the authors on the large scale ThT [2] evaluation. MonoSDF<sub>0.1</sub> denotes MonoSDF trained with the parameter suggested by the authors (0.1) for large-scale scenes specifically. MonoSDF [17] reconstructs better mesh given smaller bias, and falls into local minimum with original bias.



Batch size = 4



Batch size = 16

Figure 2. **Visualization of Neuralangelo with different batch size in *Facade*.**



### 3. Depth Evaluation

We provide evaluation of depth maps from radiance based methods and MVS methods in Tab. 2 following robust multi-view benchmark [9]. We report the Absolute Relative Error (rel) and Inlier Ratio ( $\tau$ ) with a threshold of 1.03, and split the evaluated methods based on the requirement for poses, depth ranges, and intrinsics. To better compare the performance, we additionally group NeRF models together. We show that though not as accurate as MVS methods, we predict more complete depths, as our rel is lower than all classical MVS methods.

Table 2. **Multi-view depth evaluation** with different settings: a) Classical approaches; b) with poses and depth range, without alignment; c) absolute scale evaluation with poses, without depth range and alignment; d) without poses and depth range, but with alignment; e) neural radiance field based models. ‘med’ means alignment based on median ground truth depth and the median predicted depth. Numbers with \* are from our results, while others are from DUST3R [12]. The best results for each setting are in **bold**.

Methods	GT	GT	GT	Align	ETH3D	
	Pose	Range	Intrinsics		rel ↓	$\tau$ ↑
COLMAP [6, 7]	✓	×	✓	×	16.4	55.1
(a) COLMAP Dense [6, 7]	✓	×	✓	×	89.8	23.2
ACMMP* [13]	✓	×	✓	×	<b>16.0</b>	<b>91.6</b>
MVSNet [16]	✓	✓	✓	×	35.4	31.4
MVSNet Inv. Depth [16]	✓	✓	✓	×	21.6	35.6
(b) Vis-MVSSNet [18]	✓	✓	✓	×	<b>10.8</b>	<b>43.3</b>
MVS2D ScanNet [15]	✓	✓	✓	×	27.4	4.8
MVS2D DTU [15]	✓	✓	✓	×	99.0	11.6
DeMon [11]	✓	×	✓	×	19.0	16.2
DeepV2D KITTI [10]	✓	×	✓	×	30.1	9.4
DeepV2D ScanNet [10]	✓	×	✓	×	18.7	28.7
(c) MVSNet [16]	✓	×	✓	×	507.7	8.3
MVSNet Inv. Depth [16]	✓	×	✓	×	60.3	5.8
Vis-MVSNet [18]	✓	×	✓	×	51.5	17.4
MVS2D ScanNet [15]	✓	×	✓	×	30.7	14.4
MVS2D DTU [15]	✓	×	✓	×	78.0	0.0
Robust MVD Baseline [9]	✓	×	✓	×	<b>9.0</b>	<b>42.6</b>
DeMoN [11]	×	×	✓	t	17.4	15.4
DeepV2D KITTI [10]	×	×	✓	med	27.1	10.1
DeepV2D ScanNet [10]	×	×	✓	med	11.8	29.3
(d) <b>DUST3R 224-NoCroCo</b> [12]	×	×	×	med	9.51	40.07
<b>DUST3R 224</b> [12]	×	×	×	med	4.71	61.74
<b>DUST3R 512</b> [12]	×	×	×	med	<b>2.91</b>	<b>76.91</b>
RegNeRF* [5]	✓	×	✓	×	24.9	15.0
(e) FreeNeRF* [14]	✓	×	✓	×	194.0	7.3
Ours*	✓	×	✓	×	7.4	67.2
Ours (MVS-Depth)*	✓	×	✓	×	<b>7.2</b>	<b>83.5</b>

## 4. Sparse-view Tanks and Temples Comparison

We compare our method with Neuralangelo [4] on Tanks and Temples [2] with 1/5 views that are uniformly sampled from original views in Tab. 3. We follow Neuralangelo to preprocess the scene, and use the same parameters (include the batch size) as the original paper except for using 200,000 instead of 500,000 training steps due to reduced input images. Our method outperform Neuralangelo in all scenes, showing that our method works better for the challenging sparse view setup. See Fig 3 for qualitative comparison.

Table 3. **Comparison on TnT [2] with sparse input views.** We report the  $F$ -score of our method and Neuralangelo [4] in three large-scale TnT scenes following Neuralangelo preprocessing. Best results are in **bold**.

	Meetingroom	Courthouse	Barn
Ours	<b>14.0</b>	<b>16.0</b>	<b>30.8</b>
Neuralangelo	1.7	7.7	5.7

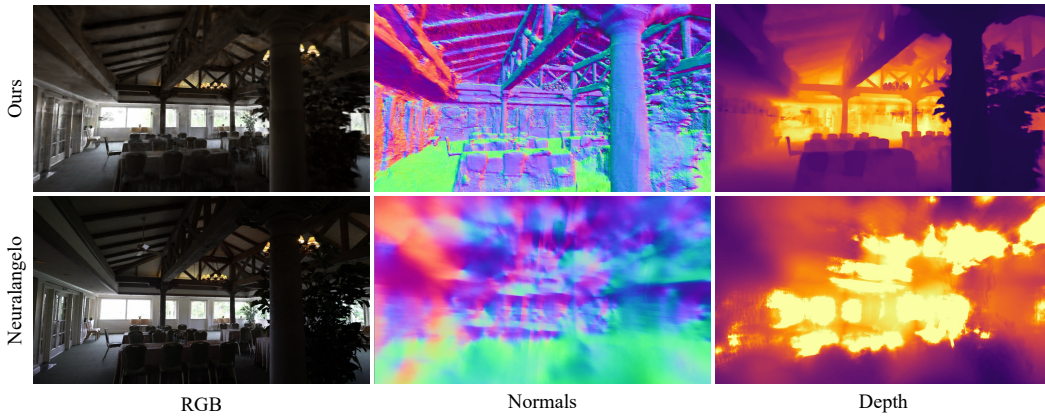


Figure 3. **Training view visualization on sparse view TnT.** We compare the rendered color images, depth map, and normals map from the training view on *Meetingroom* for our method and Neuralangelo [4].

## 5. Foreground Fattening

In patch-based MVS, Foreground fattening can happen due to the plane-based propagation of depth candidates and a patch-wise planar assumption in computing photometric scores. Our method does not suffer from foreground fattening because we do not make planar assumptions (current depth estimates are projected into other views to compute photometric scores), and the rendering loss discourages such artifacts. Figure 4 shows the alignment of RGB image and the depth images. The image and the depth are aligned precisely, indicating that our method does not suffer from the foreground fattening.



Figure 4. **Image and Depth overlay visualization.** From the left to right, we overlay the RGB image and the rendered depth map with varying fade thresholds. We show that our method does not experience foreground fattening as the images and the depths are precisely overlapped.

## 6. Additional Qualitative Results

**Images:** We present one additional comparison like our teaser figure in Figure 5. Additional visualization of mesh and novel view synthesis are shown in Figure 6. We also provide additional visualizations of our method on subsets of TanksAndTemples [2] advanced scenes in Figure 7 and Figure 8, and on ETH3D [8] in Figure 9 and Figure 10.

**Videos:** We present a free-view rendering of scenes *Relief\_2*, *Facade*, and *Kicker* on ETH3D [8] with trajectory interpolated from training poses on the attached html file. The html file also contains comparisons with RegNeRF [5], Neuralangelo[4], and MonoSDF [17]. There are only 31, 76, 31 training views for the scenes, but the rendering is realistic and the video is smooth. We also provide real-time video rendering of novel views for the TanksandTemples [2] scenes in the additionally attached files. All our results, except for ones annotated with (Ours-MVS-Depth) in Figure 7 and Table 1 of the main paper, are from our model with monocular cues.

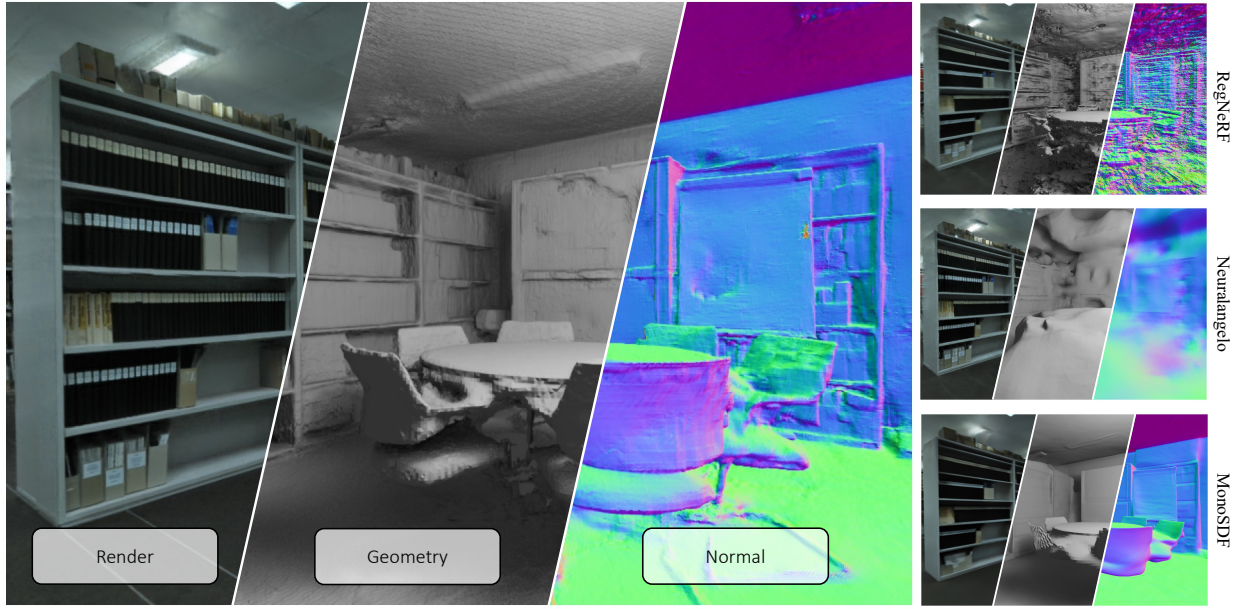


Figure 5. **Additional qualitative comparison on kicker.** We provide additional test view comparisons with baselines [4, 5, 14, 17].



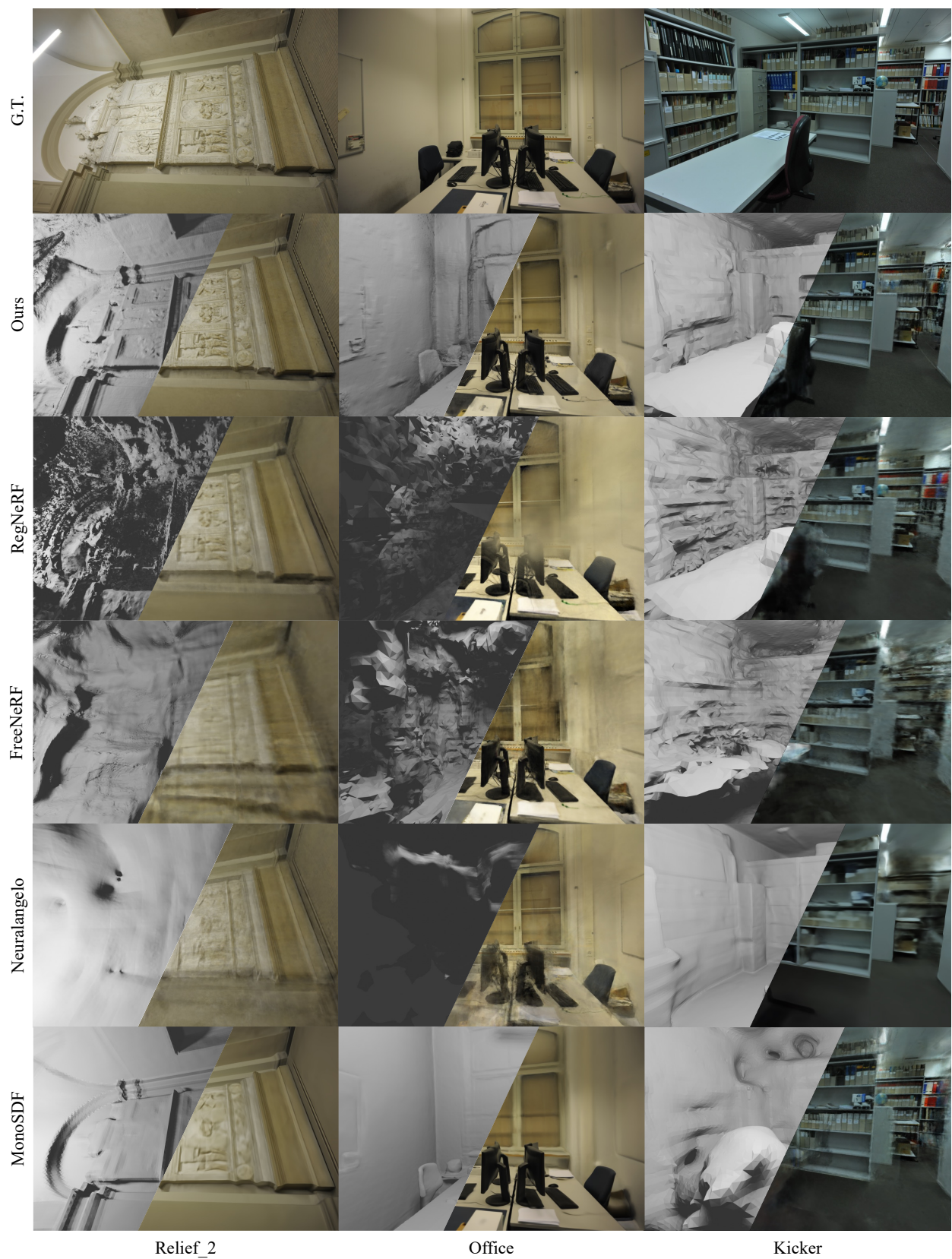


Figure 6. **Additional comparison of novel view images and meshes on ETH3D.** We provide additional comparisons with baselines [4, 5, 14, 17].





Figure 7. **Point clouds visualization for TnT advanced scenes [2].** We visualize interior and far-away views for point clouds to have a better visualization of the reconstructed geometry. Our method reconstructs complete and accurate point clouds.





Auditorium



Ballroom



Courtroom



Museum

Figure 8. Free-form views rendering for TnT advanced scenes [2].





Figure 9. **Point clouds visualization for ETH3D.** We visualize point clouds for all scenes on ETH3D [8] except *Facade*. Our method reconstructs complete and accurate point clouds.



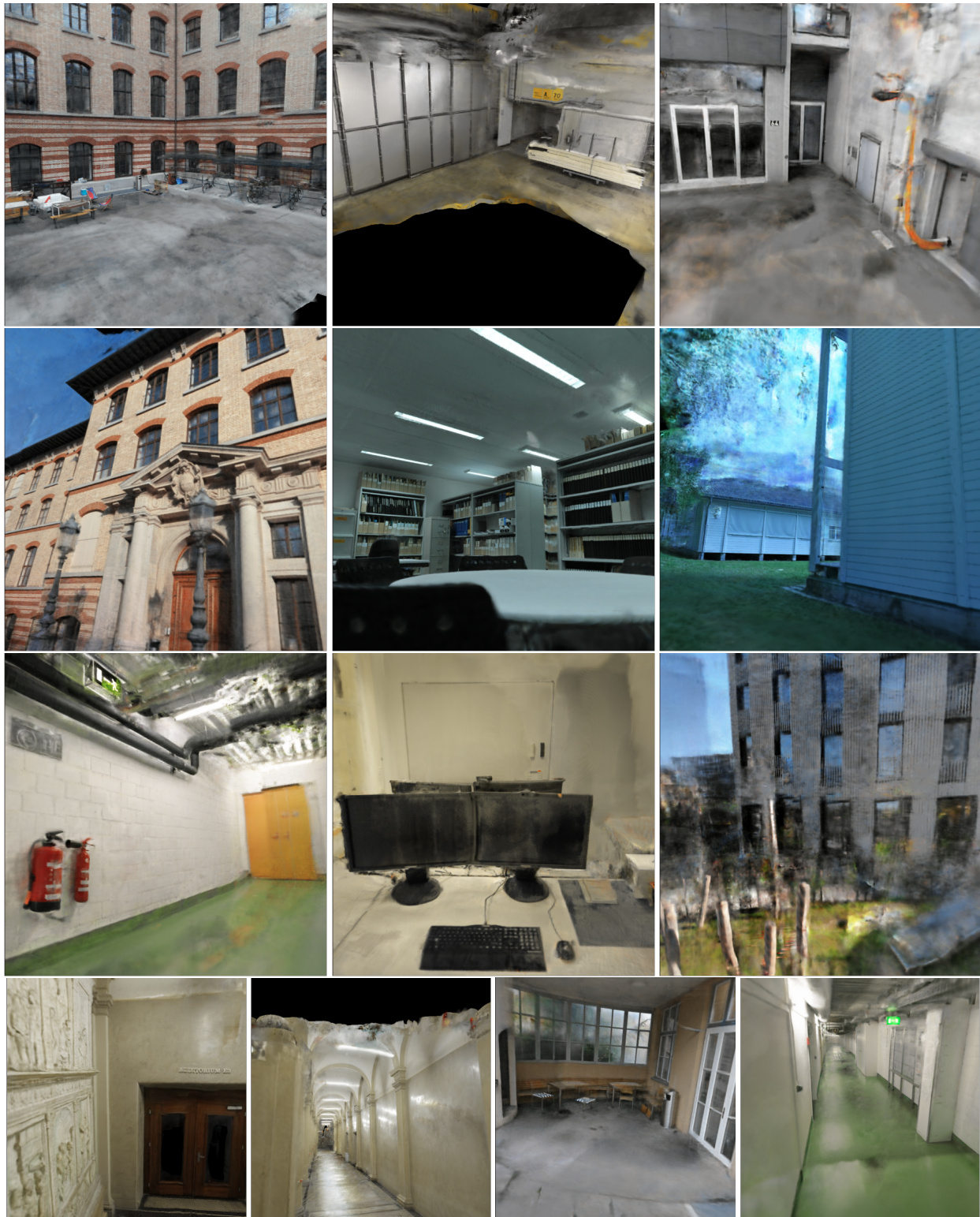


Figure 10. **Free-form views rendering for ETH3D [8].** We render novel views using our free-form viewer for each scenes in ETH3D. We show that our novel view rendering retains high-quality detailed textures in novel views.

## References

- [1] Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2565–2574 (2020) [1](#)
- [2] Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)* **36**(4), 1–13 (2017) [1](#), [2](#), [4](#), [5](#), [7](#), [8](#)
- [3] Lee, J.Y., Wu, Y., Zou, C., Wang, S., Hoiem, D.: Qff: Quantized fourier features for neural field representations. *arXiv preprint arXiv:2212.00914* (2022) [1](#)
- [4] Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) [1](#), [4](#), [5](#), [6](#)
- [5] Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022) [3](#), [5](#), [6](#)
- [6] Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) [3](#)
- [7] Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016) [3](#)
- [8] Schops, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3260–3269 (2017) [1](#), [5](#), [9](#), [10](#)
- [9] Schröppel, P., Bechtold, J., Amiranashvili, A., Brox, T.: A benchmark and a baseline for robust multi-view depth estimation. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 637–645 (2022) [3](#)
- [10] Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. In: International Conference on Learning Representations (ICLR) (2020) [3](#)
- [11] Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: DeMoN: Depth and motion network for learning monocular stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5622–5631 (2017) [3](#)
- [12] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) [3](#)
- [13] Xu, Q., Kong, W., Tao, W., Pollefeys, M.: Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022) [3](#)
- [14] Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8254–8263 (2023) [3](#), [5](#), [6](#)
- [15] Yang, Z., Ren, Z., Shan, Q., Huang, Q.: Mvs2d: Efficient multiview stereo via attention-driven 2d convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8564–8574 (2022) [3](#)
- [16] Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) [3](#)
- [17] Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems* **35**, 25018–25032 (2022) [1](#), [2](#), [5](#), [6](#)
- [18] Zhang, J., Li, S., Luo, Z., Fang, T., Yao, Y.: Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision (IJCV)* **131**(1), 199–214 (2023) [3](#)