Supplementary Material of MuSHRoom: Multi-Sensor Hybrid Room Dataset for Joint 3D Reconstruction and Novel View Synthesis

Xuqian Ren,[1] Wenjia Wang,[2] Dingding Cai,[1] Tuuli Tuominen,[1] Juho Kannala,[3] Esa Rahtu[1]
[1]Tampere University, Finland [2]The University of Hong Kong, China [3]Aalto University, Finland
{xuqian.ren,dingding.cai,tuuli.tuominen,esa.rahtu}@tuni.fi wwj2022@connect.hku.hk
Juho.Kannala@aalto.fi

1. Dataset details

1.1. Visualization System during capture

We leverage a visualization system developed by Spectacular AI SDK [1] to inspect the integrity of the point cloud reconstructed from the captured RGB-D images in real-time when using Kinect. In Figure 1, we show an example of the visualization interface.

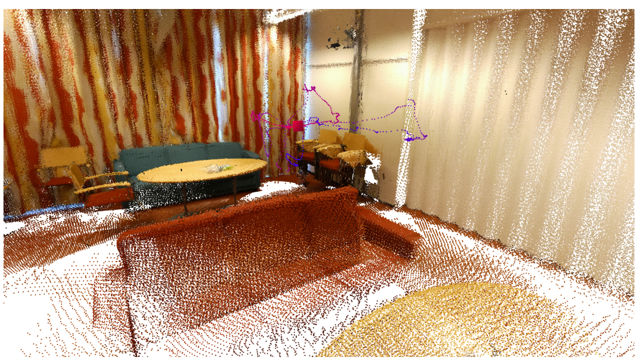

Figure 1. The visualization system. The colored area indicates successful capture, while the empty area is repeatedly scanned by checking the integrity of the point cloud.

1.2. The details of each room

In Figure 2, we show an example image of each room. The MuSHRoom is a room-scale dataset with various styles, colors, illumination, and objects, demonstrating real-world challenges. In Table 1, we show the details of the captured rooms, including the room names, scales, and camera settings.

When using COLMAP [8] to calculate the globally optimized pose for activity and koivu room captured with iPhone sequence, COLMAP failed to calculate accurate poses, so we walked around twice and captured a long sequence. We cut the long sequence from the middle and use the original Polycom of the first circle for training and the frames in the second circle for testing.

Scene	Scale (m)	Exposure time (μs)	White Balance (K)	Brightness	Gain
coffee room	$6.3 \times 5 \times 3.1$	41700	2830	128	130
computer room	$9.6 \times 6.1 \times 2.5$	33330	3100	128	255
classroom	$8.9 \times 7.2 \times 2.8$	33330	3300	128	88
honka	$6.1 \times 3.9 \times 2.3$	16670	3200	128	128
koivu	$10 \times 8 \times 2.5$	16670	4200	128	128
vr room	$5.1 \times 4.4 \times 2.8$	8300	3300	128	88
kokko	$6.7 \times 6.0 \times 2.5$	133330	4000	128	255
sauna	$9.9 \times 6.5 \times 2.4$	Auto	3300	Auto	Auto
activity	$12 \times 9 \times 2.5$	50000	3200	128	130
olohuone	$19 \times 6.4 \times 3$	Auto	3600	Auto	Auto

Table 1. The parameters of each room in our datasets. We introduce the room names, room scales and camera parameters.

2. Comparison methods

This section introduces the baseline methods we have compared with.

Volumetric Fusion. Volumetric Fusion [4] proposes to fuse depth from multiple views into the signed distance functions (SDF) and then extract the mesh model using marching cubes (MC) [6]. We use the implementation from Open3D [16], then further cluster the connected triangles and clean small clusters. Since the novel view images can only be synthesized from the textured mesh, the appearance has a large domain gap with the real images.

GO-Surf. Go-Surf [10] represents geometry and color features with the multi-resolution feature grid and decodes these representations into SDF and RGB with two shallow MLP networks. It improves speed and accuracy by simultaneously optimizing the feature volumes, decoders, and camera poses.

Nerfacto. Nerfstudio [9] is an end-to-end workflow that encapsulates various state-of-the-art NeRF techniques, which is friendly to user-collected real-world data. Nerfacto is one of the NeRF pipelines assembled by Nerfstudio that combines components from practical novel methods to balance

Figure 2. The MuSHRoom dataset. Our dataset contains ten rooms with different shapes, colors, illumination, and objects. We show an image example of each room captured by Kinect.

speed and quality. With a proposal sampler [3], scene contraction [3], and density field, Nerfacto can achieve immersive novel view synthesis quality even with real-world noisy data. However, the density field is optimized exclusively for visual consistency, which means that this model sacrifices geometry accuracy and creates occupancy regions to support the volumetric rendering even in parts of the space that are not occupied by the underlying surface. The surface will be predicted with the help of the density rather than accurate zero thickness surface [7]. When the surface is not sharp enough, the consistency of the predicted depth and normal from multiple views cannot be guaranteed, leading the 3D model extracted from the density field to become ambiguity [12] when representing mesh with truncated signed distance function (TSDF). In our evaluation, we use Poisson surface reconstruction [5] to extract the mesh model.

NeuS-facto. SDFStudio [13] is a unified framework that focuses on 3D reconstruction based on Nerfstudio, combined with recent techniques designed from implicit surface reconstruction. Similar to Nerfacto, we chose NeuS-facto with components of proposal network, multi-resolution feature grid, SDF output, and background modeling [14]. SDF output can largely improve the geometry accuracy, but it constrains the occupancy predicting flexibility, which impedes the learning of details in appearance during volume rendering [12].

3. Per-room Evaluation

3.0.1 Implementation Details:

For GO-Surf, Nerfacto, NeuS-facto, and our method, we train each model with 10k, 40k, 60k, and 70k iterations on NVIDIA RTX-2080Ti separately. We train each model without camera optimization. For Go-Surf, when training with iPhone collected data, we set ADAM optimizer with a learning rate of 1×10^{-1} for MLP decoders, and the weights for rgb, depth, sdf, fs loss are 10 times of the default one. When training Nerfacto, NeuS-facto and our method, we did not include camera pose optimization. Other settings are the same as the default setting reported in each paper. When synthesising pseudo images/depth for our data augmentation strategy, we set the interplation number n to be 3 for kinect device and 4 for iPhone device. During data augmentation, we render pseudo RGB images from Nerfacto, and pseudo depths from mesh reconstructed by NeuS-facto. These two methods produce the relatively best synthesis and mesh results in our comparison.

3.1. Metrics

Metrics for comparing reconstruction We compare the mesh reconstruction ability from both the accuracy and completeness aspects. As introduced in [10], we measure accuracy (Acc), completion (Comp), Chamfer distance (C-ℓ_1), normal consistency (NC), and F-score metrics when evaluating reconstruction results. Acc refers to what proportion of the predicted point cloud aligns with a reference point cloud with a certain threshold. Comp refers to how well a reconstructed mesh represents the full reference mesh. C- ℓ_1 distance measures the similarity between the predicted point cloud and the ground truth point cloud. It computes the average distance from a point in one point cloud to the nearest point in the other point cloud, measuring how closely two point clouds are in space. NC refers to the alignment of normals between two surfaces, representing the influence of the orientation of the surface. F-score used to balance the percison P and recall R by

$$F_{score} = \frac{2PR}{P+R} \tag{1}$$

Precision comes from the percentage of Acc within a

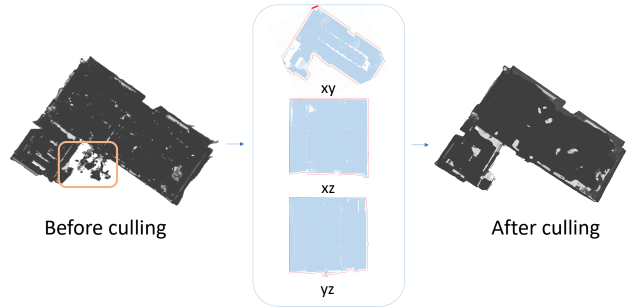

Figure 3. Effect of our culling protocol. On the left is a predicted mesh of the koivu room, reconstructed from a Kinect sequence and have culled using the prior culling protocol. Notably, there remains redundant mesh, as highlighted in the yellow square. However, when culled based on the contour of the reference mesh's projections, the mesh is cleanly trimmed.

threshold:

$$P\left(t_i\right)=\frac{1}{n}\sum_{j=1}^{n} I\left(Acc_j \leq t_i\right) \tag{2}$$

Recall comes from the percentage of Comp within a threshold:

$$R\left(t_i\right)=\frac{1}{n}\sum_{j=1}^{n} I\left(Comp_j \leq t_i\right) \tag{3}$$

In our comparison, we set the threshold t_i to 5cm.

Metrics for comparing novel view synthesis

We use PSNR, SSIM [11], and LPIPS [15] to mesh the pixel and feature distances between synthesized images and real images.

3.2. Mesh culling protocol

In the previous method [10], the mesh is culled based on several criteria: first subdivided to have the maximum edge length below 1.5cm and then culled by whether the parts are visible within the camera's frustum, and if there's valid depth in the corresponding region, and if they are occluded. However, we noticed that for some non-rectangular rooms without precise boundaries, not all redundant mesh parts are effectively culled. For instance, as depicted in Figure 3, the mesh takes on an "L" shape. The exterior mesh is not culled because it can be observed through a transparent window door. Therefore, meshes culled using the previous protocol can lead to imprecise comparison results. Here, we propose a new culling method that uses the boundary of the projection of the reference mesh to further cull the predicted mesh. In our culling protocol, after aligning the predicted mesh to the reference mesh, we project the reference mesh into the xy, xz, and yz planes. To avoid the boundaries being too close to the predicted mesh and causing some incorrect cuts, we first dilate the projections. Then we detect the contours of three projections and cut the parts of the predicted mesh that are outside of the contours. For meshes reconstructed from Kinect sequences, we applied both the previous culling protocol and our cutting method. This was due to their non-rectangular geometry and unbounded areas. For meshes derived from iPhone sequences, we only employed our culling protocol. This is because the granularity of these meshes is too coarse to accurately determine if regions are within the camera's frustum, occluded, or if they constitute valid mesh sections. We compare all regions culled by our protocol with the reference mesh.

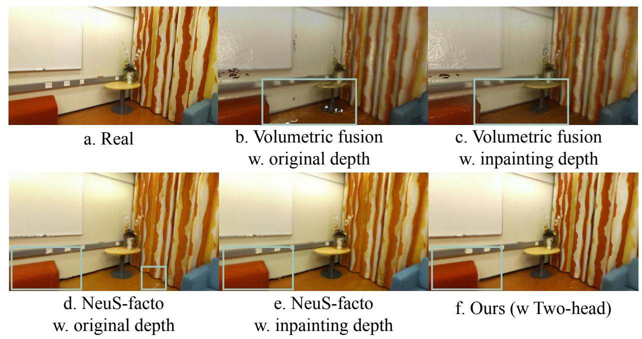

Figure 4. The ablation study of inpainting depth and two-head structure. We visualize the results of Volumetric fusion and Neus-facto methods with original depth and with inpainting depth. We also present the results with or without the two-head structure. The two-head structure can help Neus-facto fit the color better and avoid underfitting.

3.3. Per-room quantitative comparison result

In Table 3 and Table 4, we measure the reconstruction and rendering quality quantitatively for each room. Our method can obtain a good trade-off between reconstruction and rendering results. Note that we did not apply the data augmentation to the classroom, computer, and sauna room of Kinect sequences. Because the data augmentation requires accurate pseudo images and depths, the current NeuS-facto model still cannot render accurate pseudo depths and cannot further contribute to the final results.

We also try NeRF++ [14], Mip-NeRF [2] on the MuSH-Room dataset, but these pipelines cannot work on real-world dataset, which indicates the proposal sampling is very crucial for the noisy real-world data. The overall rendering and reconstruction quality of Kinect sequences are relatively better than the results of the iPhone. Potential estimation comes from Kinect can obtain more accurate depth map, which contribute to both the reconstruction and novel view synthesis. For methods that predict SDF for reconstruction, inaccurate depths are not only detrimental to the reconstruction, but also hinder the synthesis learning.

Method	Reconstruction quality					Rendering quality					
						Test within a single sequence			Test with a different sequence		
	Acc $\downarrow$	**Comp** $\downarrow$	**C-ℓ_1** $\downarrow$	**NC** $\uparrow$	**F-score** $\uparrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$
Volumetric Fusion (original depth)	0.0178	0.0218	0.0198	0.8514	0.9217	14.61	0.6920	0.3774	12.90	0.6634	0.4150
Volumetric Fusion (inpainting depth)	0.0207	0.0212	0.0210	0.8407	0.9143	14.92	0.6873	0.3950	13.84	0.6556	0.4170
NeuS-facto (original depth)	0.0145	0.0183	0.0164	0.9121	0.9565	21.08	0.7658	0.2198	**22.37**	**0.8483**	**0.1396**
NeuS-facto (inpainting depth)	**0.0136**	**0.0161**	**0.0149**	**0.9130**	**0.9655**	**21.21**	**0.7709**	**0.2132**	21.98	0.8465	0.1427

Table 2. The ablation study of inpainting depth and two-head structure. Test within a single sequence means we uniform sample test frames from a single sequence and train on all left frames. Test with a different sequence means we train on one sequence and test on another individual sequence.

3.4. Per-room qualitative comparison result

In this section we show more visualization comparison of each methods with both test within a single sequence and test with a different sequence methods. We show mesh comparison qualitatively in Figure 5 and Figure 6 for Kinect and iPhone sequences. Our method provides a relatively smoother and more completed mesh. The iPhone mesh is more coarse than Kinect mesh, except the mesh produced by Go-Surf [10] method, which shows this method is more robust to devices.

3.5. Ablation study

We further evaluate the effect of inpainting depth quantitatively and qualitatively and show the results in Table 2 and Figure 4. Volumetric fusion heavily relies on the completeness of the depth. Without inpainting the holes, the mesh will have parts missing where depth is invalid, as shown in Figure 4b. We also visualize the effect of the two-head structure in Figure 4. Without the two-head structure, the color of the object exhibits sub-optimal learning, as shown in the red sofa marked by the green square in Figure 4e and 4f, in which the color transitions from a rich red to a paler shade.

References

[1] Spectacular ai sdk. https://www.spectacularai.com, 2021. 1

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 3

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 2

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 1, 8, 9, 10, 11

[5] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. *Symposium on Geometry Processing,Symposium on Geometry Processing*, Jun 2006. 2

[6] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 1

[7] Marie-Julie Rakotosaona, Fabian Manhardt, Diego Martin Arroyo, Michael Niemeyer, Abhijit Kundu, and Federico Tombari. Nerfmeshing: Distilling neural radiance fields into geometrically-accurate 3d meshes. *arXiv preprint arXiv:2303.09431*, 2023. 2

[8] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, May 2016. 1

[9] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 1, 8, 9, 10, 11

[10] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *2022 International Conference on 3D Vision (3DV)*, pages 433–442. IEEE, 2022. 1, 2, 3, 4, 8, 9, 10, 11

[11] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3

[12] Yuting Xiao, Yiqun Zhao, Yanyu Xu, and Shenghua Gao. Resnerf: Geometry-guided residual neural radiance field for indoor scene novel view synthesis. *arXiv preprint arXiv:2211.16211*, 2022. 2

[13] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. 2, 8, 9, 10, 11

[14] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020. 2, 3

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

Room	Device	Methods	Reconstruction quality					Rendering quality					
								Test within a single sequence			Test with a different sequence		
			Acc $\downarrow$	**Comp** $\downarrow$	**C-ℓ_1** $\downarrow$	**NC** $\uparrow$	**F-score** $\uparrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$
coffee room	Kinect	Volumetric Fusion	0.0207	0.0212	0.0210	0.8407	0.9143	14.92	0.6873	0.3950	13.84	0.6556	0.4170
		GO-Surf	0.0306	0.0369	0.0337	0.8659	0.8099	21.48	0.7698	0.2685	19.45	0.7189	0.2332
		Nerfacto	0.0349	0.2015	0.1182	0.7018	0.5870	22.99	0.7918	0.2245	22.95	0.8564	0.1276
		NeuS-facto	0.0136	0.0161	0.0149	0.9130	0.9655	21.21	0.7709	0.2132	21.98	0.8465	0.1427
		Ours	0.0139	0.0167	0.0153	0.9108	0.9625	22.95	0.7946	0.2396	22.99	0.8579	0.1354
	iPhone	Volumetric Fusion	0.0355	0.0187	0.0271	0.7990	0.8616	12.01	0.5546	0.4869	12.35	0.5849	0.4695
		GO-Surf	0.0278	0.0229	0.0254	0.8767	0.9248	18.88	0.6726	0.3659	18.42	0.6764	0.3940
		Nerfacto	0.0406	0.3443	0.1924	0.6573	0.4773	21.22	0.7884	0.2040	22.35	0.8136	0.2159
		NeuS-facto	0.0407	0.0231	0.0319	0.8604	0.8884	19.44	0.7250	0.3463	20.01	0.7230	0.3631
		Ous	0.0412	0.0234	0.0323	0.8653	0.8892	20.89	0.7564	0.2968	21.38	0.7636	0.2961
computer room	Kinect	Volumetric Fusion	0.0323	0.0402	0.0362	0.8402	0.8395	15.41	0.6764	0.4007	14.87	0.6867	0.4086
		GO-Surf	0.0281	0.0302	0.0291	0.9016	0.8972	22.40	0.7877	0.2615	22.44	0.8178	0.2305
		Nerfacto	0.0556	0.1008	0.0782	0.7314	0.5985	24.15	0.8270	0.2094	26.29	0.8849	0.1510
		NeuS-facto	0.0253	0.0229	0.0241	0.9033	0.9122	22.99	0.8080	0.2338	25.00	0.8561	0.1925
		Ours	0.0258	0.0233	0.0246	0.9018	0.9106	24.05	0.8262	0.2285	25.65	0.8722	0.1787
	iPhone	Volumetric Fusion	0.0560	0.0162	0.0361	0.8069	0.7925	12.29	0.6000	0.4543	11.96	0.5596	0.4635
		GO-Surf	0.0646	0.0407	0.0526	0.8566	0.6302	17.90	0.6768	0.3847	16.59	0.5820	0.4886
		Nerfacto	0.0623	0.1471	0.1047	0.6432	0.5684	21.94	0.7855	0.2393	21.86	0.7903	0.2159
		NeuS-facto	0.0551	0.0299	0.0425	0.8452	0.7854	16.59	0.6636	0.4197	15.83	0.6327	0.4321
		Ous	0.0474	0.0276	0.0375	0.8555	0.8172	20.10	0.7411	0.3530	18.10	0.6894	0.3649
activity	Kinect	Volumetric Fusion	0.0469	0.0431	0.0450	0.7633	0.7658	12.46	0.5750	0.4898	12.09	0.5455	0.5273
		GO-Surf	0.0401	0.0258	0.0330	0.8118	0.8816	18.99	0.6985	0.3823	19.42	0.7283	0.3120
		Nerfacto	0.0564	0.1043	0.0804	0.7089	0.6258	21.12	0.7535	0.2908	22.13	0.8112	0.2184
		NeuS-facto	0.0384	0.0320	0.0352	0.8150	0.8468	19.53	0.7258	0.3356	20.18	0.7707	0.2750
		Ours	0.0377	0.0332	0.0354	0.8151	0.8455	20.89	0.7509	0.3221	21.33	0.7919	0.2534
	iPhone	Volumetric Fusion	0.0688	0.0238	0.0463	0.7344	0.7479	11.27	0.4335	0.5815	11.08	0.4542	0.5731
		GO-Surf	0.1045	0.0347	0.0696	0.7592	0.6494	15.36	0.4892	0.5902	14.31	0.4928	0.6094
		Nerfacto	0.0709	0.1334	0.1022	0.6608	0.6430	18.27	0.6381	0.3617	17.05	0.6003	0.4137
		NeuS-facto	0.0938	0.0585	0.0761	0.7290	0.5719	15.61	0.5391	0.5654	14.23	0.5153	0.5743
		Ous	0.0837	0.0536	0.0687	0.7505	0.6108	16.64	0.5748	0.5026	15.30	0.5641	0.5274
kokko	Kinect	Volumetric Fusion	0.0426	0.0304	0.0365	0.7605	0.8316	14.51	0.6978	0.4051	12.36	0.6715	0.4362
		GO-Surf	0.0336	0.0311	0.0324	0.8173	0.8823	20.61	0.8096	0.2907	18.21	0.7975	0.2881
		Nerfacto	0.0567	0.1361	0.0964	0.6144	0.5255	22.24	0.8173	0.2479	20.34	0.8554	0.1972
		NeuS-facto	0.0265	0.0204	0.0235	0.8164	0.9300	20.94	0.8154	0.2654	19.18	0.8467	0.2138
		Ours	0.0272	0.0209	0.0240	0.8142	0.9249	22.25	0.8226	0.2619	20.47	0.8614	0.2086
	iPhone	Volumetric Fusion	0.0417	0.0149	0.0283	0.7688	0.8688	11.71	0.5719	0.5117	11.85	0.5564	0.5111
		GO-Surf	0.0513	0.0248	0.0381	0.8145	0.8863	17.11	0.6641	0.4564	18.58	0.6730	0.4440
		Nerfacto	0.0554	0.0769	0.0661	0.6156	0.6376	18.14	0.7492	0.2939	19.67	0.7839	0.2490
		NeuS-facto	0.0394	0.0219	0.0307	0.7946	0.8746	15.64	0.6806	0.4440	16.45	0.6624	0.4312
		Ous	0.0398	0.0248	0.0323	0.8019	0.8651	17.65	0.7280	0.3865	17.72	0.6814	0.3711
honka	Kinect	Volumetric Fusion	0.0265	0.0307	0.0286	0.8404	0.8759	13.30	0.6668	0.4357	12.20	0.6902	0.4233
		GO-Surf	0.0230	0.0189	0.0209	0.9006	0.9401	19.43	0.7871	0.2946	20.99	0.7976	0.2449
		Nerfacto	0.0476	0.1599	0.1038	0.6742	0.6076	21.00	0.7949	0.2374	22.88	0.8619	0.1533
		NeuS-facto	0.0185	0.0184	0.0184	0.9024	0.9403	19.40	0.7759	0.2658	22.45	0.8505	0.1616
		Ours	0.0191	0.0189	0.0190	0.9002	0.9354	20.86	0.8000	0.2584	23.59	0.8683	0.1527
	iPhone	Volumetric Fusion	0.0378	0.0134	0.0256	0.8184	0.8712	11.47	0.5243	0.5268	11.88	0.5479	0.5183
		GO-Surf	0.0414	0.0332	0.0373	0.8729	0.8627	17.35	0.6037	0.4502	17.10	0.6267	0.4529
		Nerfacto	0.0266	0.2674	0.1470	0.6725	0.6064	19.34	0.7414	0.2391	19.46	0.7605	0.2203
		NeuS-facto	0.0316	0.0195	0.0256	0.8803	0.9093	17.35	0.6671	0.3593	18.02	0.6810	0.3448
		Ous	0.0329	0.0258	0.0294	0.8820	0.8887	18.89	0.7054	0.3136	18.25	0.6978	0.3155

Table 3. The quantitative reconstruction and rendering comparison results of each room. The best results are highlighted in pink. The second best results are marked in yellow. Test within a single sequence means we uniformly sample test frames from a single sequence and train on all left frames. Test with a different sequence means we train on one sequence and test on another individual sequence. Our method provides a good trade-off between mesh and novel view synthesis results.

[16] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 1

Room	Device	Methods	Reconstruction quality					Rendering quality					
								Test within a single sequence			Test with a different sequence		
			Acc $\downarrow$	**Comp** $\downarrow$	**C-ℓ_1** $\downarrow$	**NC** $\uparrow$	**F-score** $\uparrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$
classroom	Kinect	Volumetric Fusion	0.0365	0.0288	0.0327	0.8233	0.8398	14.56	0.7024	0.3721	13.54	0.6751	0.4106
		GO-Surf	0.0310	0.0221	0.0265	0.8836	0.8998	22.85	0.8109	0.2419	22.22	0.7921	0.2017
		Nerfacto	0.0567	0.1697	0.1132	0.6968	0.5461	24.96	0.8420	0.1989	25.80	0.8920	0.1277
		NeuS-facto	0.0296	0.0221	0.0259	0.8785	0.8879	23.10	0.8238	0.2458	24.16	0.8696	0.1552
		Ours	0.0298	0.0223	0.0260	0.8789	0.8853	24.66	0.8401	0.2270	25.44	0.8812	0.1471
	iPhone	Volumetric Fusion	0.0592	0.0209	0.0401	0.7736	0.7457	11.32	0.5425	0.5362	11.43	0.5337	0.5219
		GO-Surf	0.0511	0.0308	0.0410	0.8563	0.7725	17.40	0.6042	0.5191	19.27	0.6482	0.4793
		Nerfacto	0.0519	0.1311	0.0915	0.6976	0.6339	20.63	0.7095	0.3463	22.54	0.7900	0.2339
		NeuS-facto	0.0596	0.0446	0.0521	0.8047	0.6781	17.72	0.6559	0.5018	18.52	0.6833	0.4602
		Ours	0.0546	0.0402	0.0474	0.8130	0.7088	19.63	0.6901	0.4464	19.23	0.6729	0.3824
koivu	Kinect	Volumetric Fusion	0.0664	0.0344	0.0504	0.7953	0.7979	12.97	0.6039	0.4701	11.94	0.6087	0.4795
		GO-Surf	0.0815	0.0398	0.0606	0.8501	0.8046	20.31	0.7393	0.2873	19.31	0.7273	0.3268
		Nerfacto	0.0927	0.0924	0.0925	0.6793	0.5664	21.49	0.7696	0.2466	20.18	0.7571	0.2663
		NeuS-facto	0.0749	0.0231	0.0490	0.8506	0.8485	19.99	0.7460	0.2590	19.51	0.7397	0.2732
		Ours	0.0737	0.0228	0.0482	0.8507	0.8491	21.63	0.7707	0.2649	20.19	0.7553	0.2841
	iPhone	Volumetric Fusion	0.0843	0.0286	0.0564	0.7629	0.7108	11.82	0.5597	0.4896	11.89	0.5335	0.5313
		GO-Surf	0.0906	0.0320	0.0613	0.8106	0.7116	16.42	0.6263	0.4140	15.50	0.5589	0.5198
		Nerfacto	0.1004	0.1053	0.1028	0.6452	0.5358	20.68	0.7862	0.2150	19.07	0.7057	0.3112
		NeuS-facto	0.1021	0.0417	0.0719	0.7808	0.6454	17.13	0.6983	0.3764	16.04	0.6365	0.4552
		Ours	0.0924	0.0418	0.0671	0.7984	0.6758	19.94	0.7491	0.3059	18.04	0.6753	0.3938
vr room	Kinect	Volumetric Fusion	0.0199	0.0206	0.0202	0.8580	0.9003	14.32	0.7514	0.3708	14.37	0.7632	0.3785
		GO-Surf	0.0291	0.0808	0.0550	0.8706	0.7273	22.56	0.8429	0.2750	20.05	0.8112	0.3058
		Nerfacto	0.0501	0.2916	0.1709	0.6608	0.4312	24.25	0.8565	0.2038	24.90	0.9009	0.1384
		NeuS-facto	0.0152	0.0186	0.0169	0.9126	0.9376	22.65	0.8502	0.2351	23.55	0.8928	0.1658
		Ours	0.0151	0.0185	0.0168	0.9135	0.9378	26.46	0.9081	0.1546	24.69	0.8995	0.1589
	iPhone	Volumetric Fusion	0.0381	0.0158	0.0270	0.8177	0.8347	11.57	0.5975	0.5157	11.44	0.6022	0.5151
		GO-Surf	0.0281	0.0222	0.0252	0.8932	0.8795	19.64	0.6920	0.4356	19.69	0.7133	0.4240
		Nerfacto	0.0427	0.0705	0.0566	0.7026	0.6719	22.62	0.8254	0.2244	22.17	0.8095	0.2570
		NeuS-facto	0.0330	0.0219	0.0274	0.8818	0.8535	20.60	0.7705	0.3568	18.63	0.7055	0.3965
		Ours	0.0336	0.0226	0.0281	0.8831	0.8483	22.04	0.7937	0.3088	19.45	0.7178	0.3477
sauna	Kinect	Volumetric Fusion	0.0240	0.0395	0.0317	0.8461	0.8790	13.60	0.6623	0.4127	13.22	0.6549	0.4232
		GO-Surf	0.0233	0.0394	0.0314	0.9003	0.9124	19.32	0.7551	0.3788	18.35	0.7622	0.3539
		Nerfacto	0.0590	0.1231	0.0910	0.6892	0.6433	19.81	0.7688	0.3329	19.07	0.7833	0.2690
		NeuS-facto	0.0185	0.0349	0.0267	0.9010	0.9235	18.53	0.7553	0.3510	18.69	0.7884	0.2817
		Ours	0.0181	0.0348	0.0265	0.9052	0.9227	20.21	0.7750	0.3242	19.11	0.7896	0.2848
	iPhone	Volumetric Fusion	0.0387	0.0203	0.0295	0.8312	0.8537	12.16	0.5739	0.5277	11.99	0.5874	0.5135
		GO-Surf	0.0672	0.0279	0.0476	0.8699	0.8206	18.93	0.6581	0.4927	18.76	0.6736	0.4967
		Nerfacto	0.0609	0.0933	0.0771	0.6696	0.6440	21.19	0.7602	0.3131	21.43	0.7762	0.3156
		NeuS-facto	0.0630	0.0436	0.0533	0.8357	0.7241	17.37	0.6791	0.5354	16.78	0.6585	0.5330
		Ours	0.0648	0.0509	0.0579	0.8480	0.6784	19.82	0.7126	0.4700	18.61	0.6805	0.4778
olohuone	Kinect	Volumetric Fusion	0.0381	0.0521	0.0451	0.7911	0.7949	12.34	0.6049	0.4563	12.61	0.5572	0.5191
		GO-Surf	0.0345	0.0419	0.0382	0.8621	0.8648	17.63	0.7074	0.4143	19.67	0.7403	0.3148
		Nerfacto	0.0599	0.1056	0.0827	0.7216	0.5836	18.77	0.7497	0.2871	21.41	0.8538	0.1734
		NeuS-facto	0.0336	0.0446	0.0391	0.8453	0.8324	17.77	0.7281	0.3550	20.34	0.8240	0.2320
		Ours	0.0346	0.0466	0.0406	0.8454	0.8214	18.62	0.7522	0.3271	21.05	0.8452	0.2078
	iPhone	Volumetric Fusion	0.0610	0.0343	0.0477	0.7540	0.7632	11.78	0.5613	0.4957	12.12	0.5656	0.4771
		GO-Surf	0.1033	0.0361	0.0697	0.7914	0.6802	16.05	0.6052	0.5116	18.21	0.6018	0.5231
		Nerfacto	0.0798	0.0805	0.0801	0.6970	0.5545	21.31	0.7709	0.2589	21.63	0.7945	0.2377
		NeuS-facto	0.1405	0.1484	0.1445	0.7228	0.2688	13.61	0.6170	0.5961	13.80	0.6058	0.6174
		Ours	0.1389	0.1374	0.1381	0.7332	0.2988	17.27	0.6787	0.5140	16.78	0.6759	0.5079

Table 4. The quantitative reconstruction and rendering comparison results of each room. The best results are highlighted in pink. The second best results are marked in yellow. Test within a single sequence means we uniformly sample test frames from a single sequence and train on all left frames. Test with a different sequence means we train on one sequence and test on another individual sequence. Our method provides a good trade-off between mesh and novel view synthesis results.

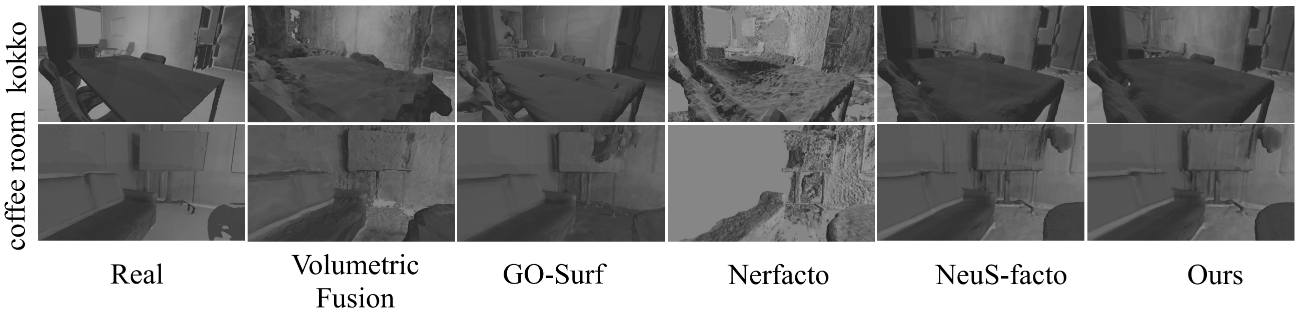

Figure 5. We compare the mesh reconstruction quality of Kinect sequences qualitatively. Please zoom in to see the details.

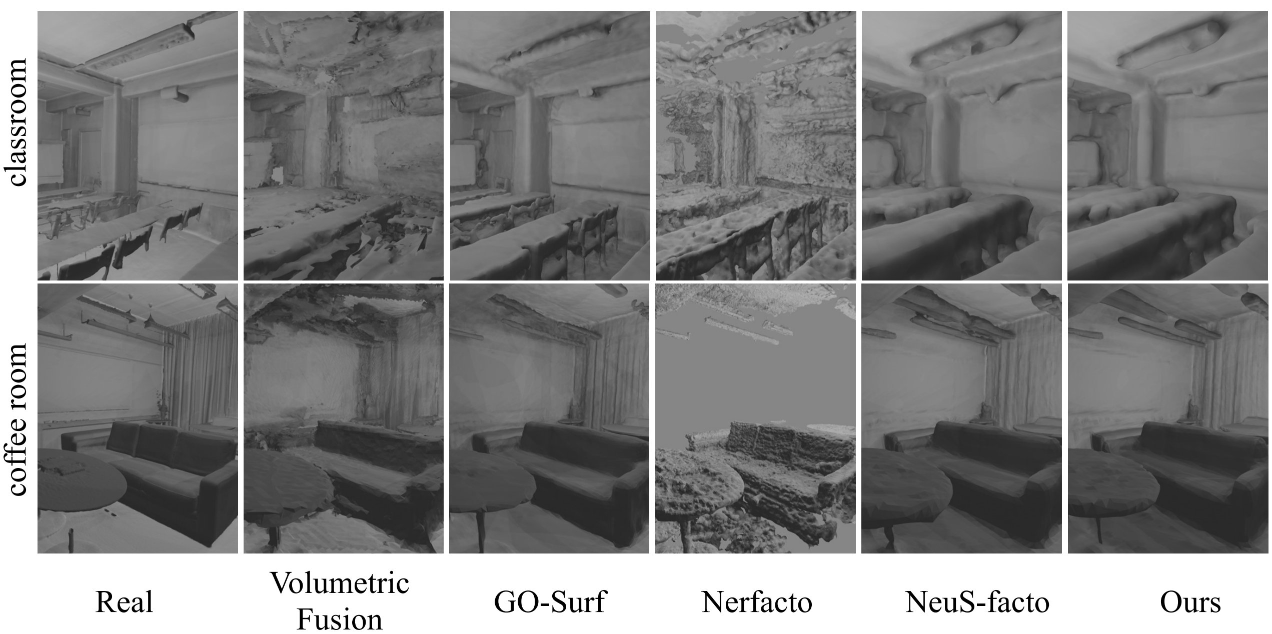

Figure 6. We compare the mesh reconstruction quality of iPhone sequences qualitatively. Please zoom in to see the details.

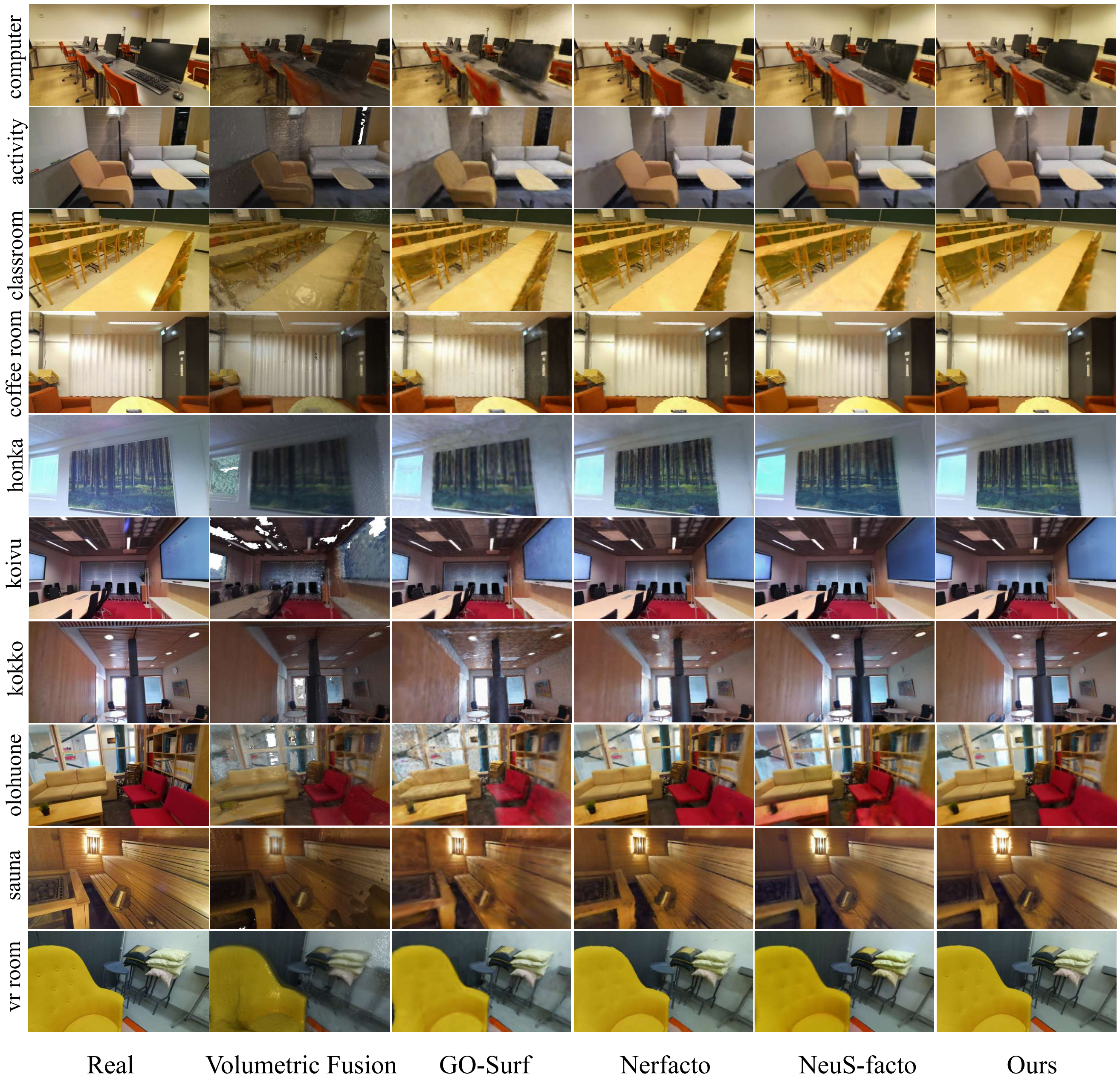

Figure 7. We compare the rendering quality of Kinect sequences with our test within a sequence method qualitatively. The color saturation level and fine-grained content of our method are comparable to the results of Nerfacto [9]. Volumetric Fusion [4] rendering results have a large content gap with the real images. GO-Surf [10] produces images lacking fine-grained details. Visualization results of NeuS-facto [13] still have some ripples, and colors are underfitting to some extent. Please zoom in to see the details.

Figure 8. We compare the rendering quality of Kinect sequences with our test with a different sequence method qualitatively. The color saturation level and fine-grained content of our method are comparable to the results of Nerfacto [9]. Volumetric Fusion [4] rendering results have a large content gap with the real images. GO-Surf [10] produces images lacking fine-grained details. Visualization results of NeuS-facto [13] still have some ripples, and colors are underfitting to some extent. Please zoom in to see the details.

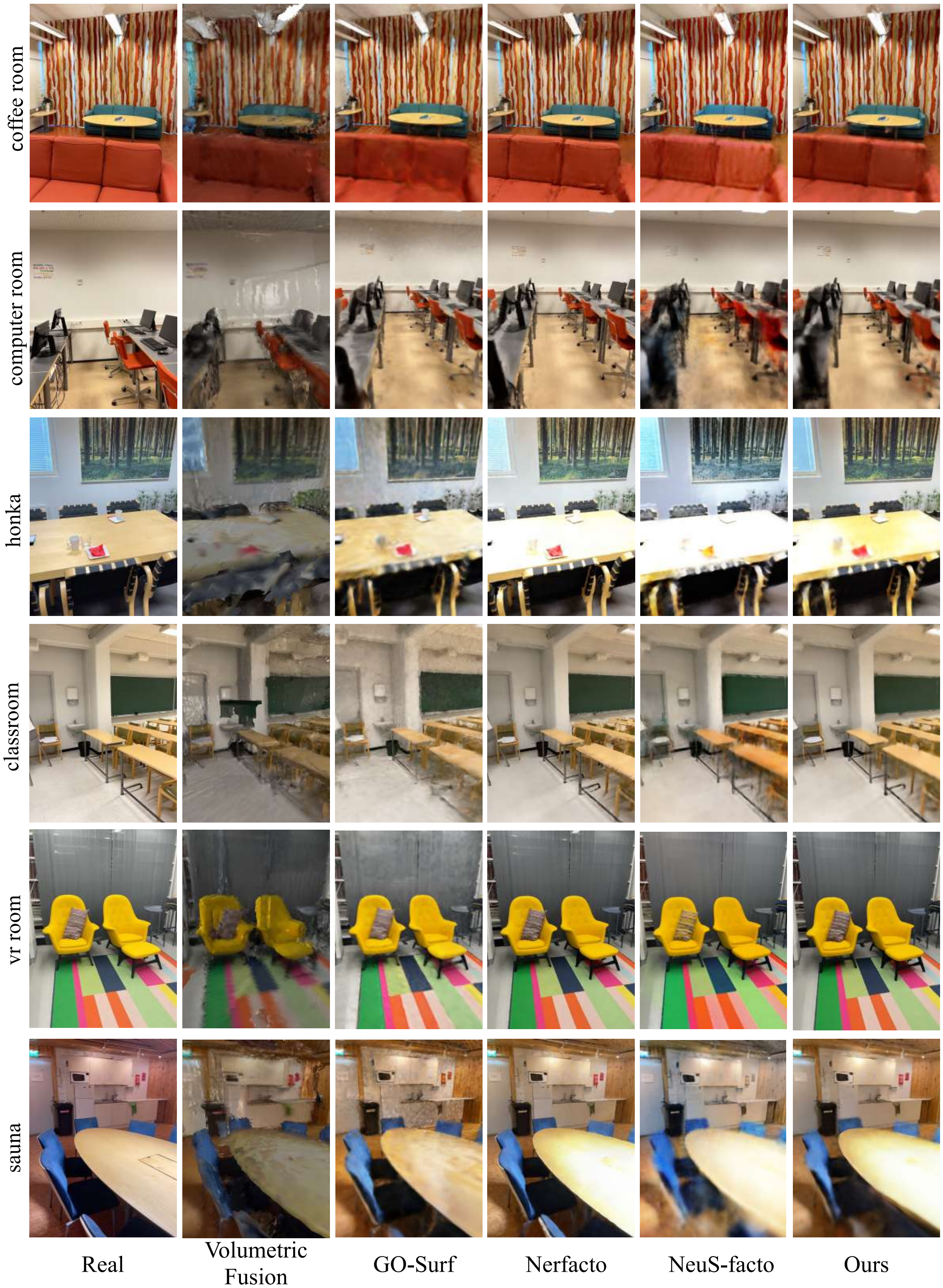

Figure 9. We compare the rendering quality of iPhone sequences with the test within a sequence method qualitatively. Nerfacto [9] method provides the most detailed and photorealistic results. Volumetric Fusion [4] rendering results have a large content gap with the real images. GO-Surf [10] produces images lacking fine-grained details. NeuS-facto [13] results are much more blurry. Our method improves the NeuS-facto from color and fine-grained details but still has a distance when compared with Nerfacto results. The blurry results also show the inaccurate depth in iPhone sequences can be detrimental to the synthesis quality. Please zoom in to see the details.

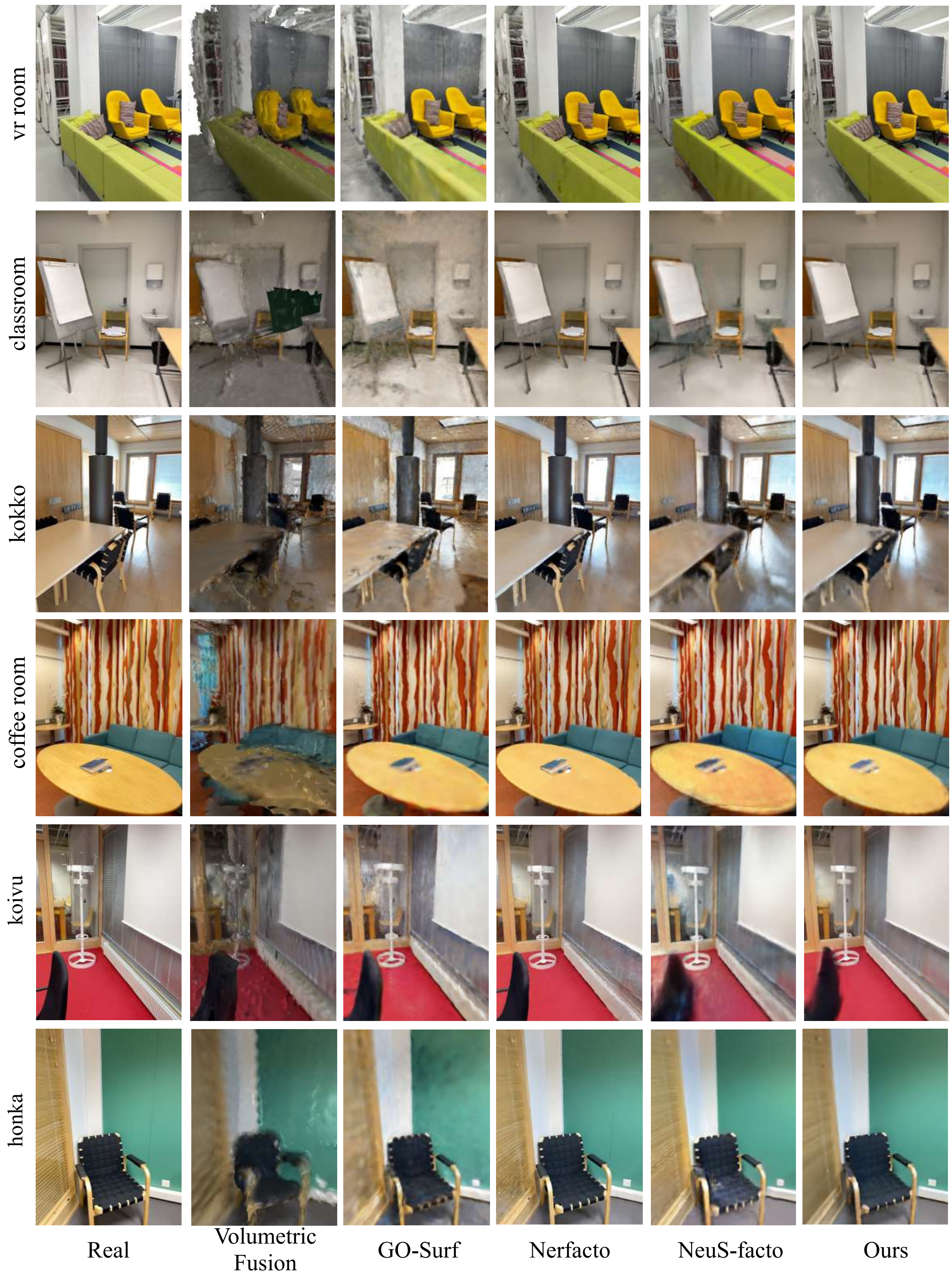

Figure 10. We compare the rendering quality of iPhone sequences with the test with a different sequence method qualitatively. Nerfacto [9] method provides the most detailed and photorealistic results. Volumetric Fusion [4] rendering results have a large content gap with the real images. GO-Surf [10] produces images lacking fine-grained details. NeuS-facto [13] results are much more blurry. Our method improves the NeuS-facto from color and fine-grained details but still has a distance when compared with Nerfacto results. The blurry results also show the inaccurate depth in iPhone sequences can be detrimental to the synthesis quality. Please zoom in to see the details.