Supplementary Materials of Omni-Recon: Harnessing Image-based Rendering for General-Purpose Neural Radiance Fields

Yonggan Fu, Huaizhi Qu, Zhifan Ye, Chaojian Li,
Kevin Zhao, and Yingyan (Celine) Lin

Georgia Institute of Technology
{yfu314,zye327,cli851,kzhao14,celine.lin}@gatech.edu

1 Overview and Outline

In this supplementary material, we provide additional experiments, visualizations, and analysis as complements to the main content, outlined as follows:

- We extend our predict-then-blend strategy, introduced in Sec. 4 of our main paper, to the 2D CLIP-LSeg model [3] for language-driven 3D semantic segmentation in Sec. 2;
- We offer additional visualizations of the reconstructed meshes on various datasets and the scene editing effects in Sec. 3 and Sec. 4, respectively;
- We elaborate on more details about the dataset, training, and finetuning settings of our performed experiments in Sec. 5;
- We provide a detailed formulation of the subtraction attention used by our appearance transformer in Sec. 6;
- We conduct ablation studies on different components of our proposed NeRF model backbone in Sec. 7;
- We discuss the limitations and future directions of our work in Sec. 8.
- We discuss the potential negative societal impacts of our work in Sec. 9.

2 Zero-shot Language-driven 3D Semantic Segmentation

In addition to monocular models, we also apply our predict-then-blend strategy, introduced in Sec. 4 of our main paper, to the 2D CLIP-LSeg model [3] for language-driven 3D semantic segmentation, i.e., performing segmentation based on the relevance between pixel-wise embeddings and the target class's textual embedding. This is achieved by lifting the CLIP embeddings of the source views to novel views of the 3D scene. More specifically, we apply CLIP-LSeg [3] to each source view to acquire their embeddings in CLIP-LSeg's latent space and then reuse the blending weights and density to derive the embeddings of novel views through volumetric rendering. The visual effects, including the zero-shot performance with and without PET, on Replica and ScanNet are provided in Fig. 1. This set of experiments indicates that our strategy can effectively lift a diverse range of vision models into the 3D world in a zero-shot manner.

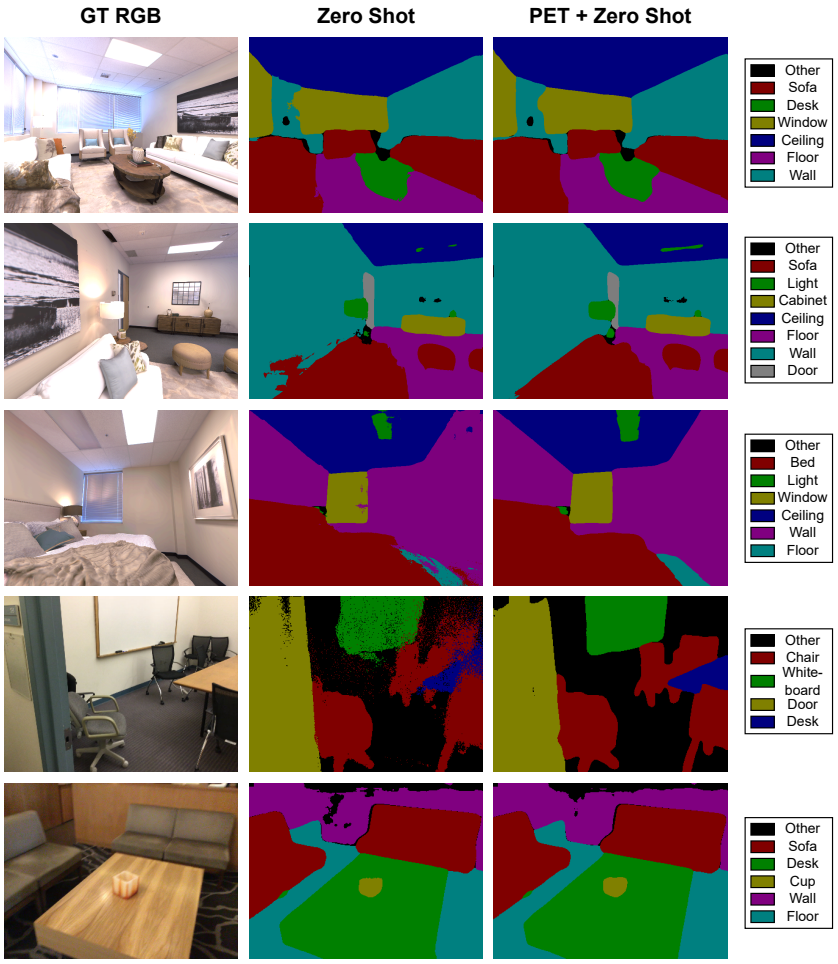

Fig. 1: Language-driven semantic segmentation by blending CLIP embeddings.

3 Additional Visualizations of The Reconstructed Meshes

We provide additional visualizations of the reconstructed mesh on DTU [1], NeRF-Synthetic [7], and NSVF-Synthetic [5] in Fig. 2, following the settings in Sec. 3.5 of our main paper. We observe that (1) On the test scenes from DTU, which have a similar domain w.r.t. the training scenes employed by both our model and the baselines, the mesh quality achieved by our Omni-Recon and the strongest baseline ReTR [4] are comparable (ours may have slightly fewer missing holes), and both are smoother than VolRecon [9]; (2) On more challenging scenes with larger domain gaps w.r.t. the employed training scenes, our Omni-Recon can achieve notably higher-quality meshes with better-maintained structures, fewer holes, and smoother surfaces, consistent with our observations in Sec. 3.5 of our main paper; (3) Current generalizable mesh reconstruction methods,

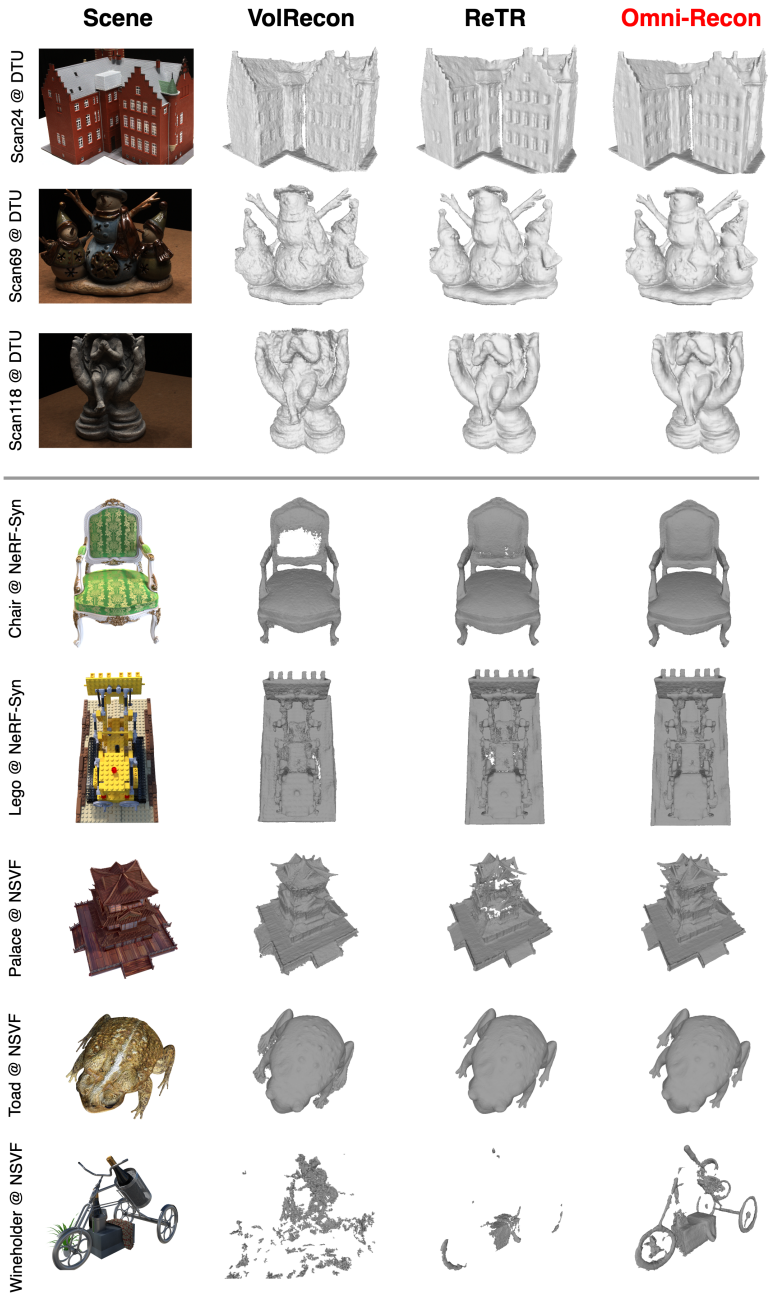

Fig. 2: Visualize reconstructed meshes of our Omni-Recon and the two strongest baselines [4,9]. Rows 1-3: Three test scenes (Scan24/69/118) from DTU [1]; Rows 4-8: Scenes from NeRF-Synthetic [7] and NSVF-Synthetic [5], which present relatively challenging cases due to domain shifts w.r.t. the training scenes. Zoom in for a better view.

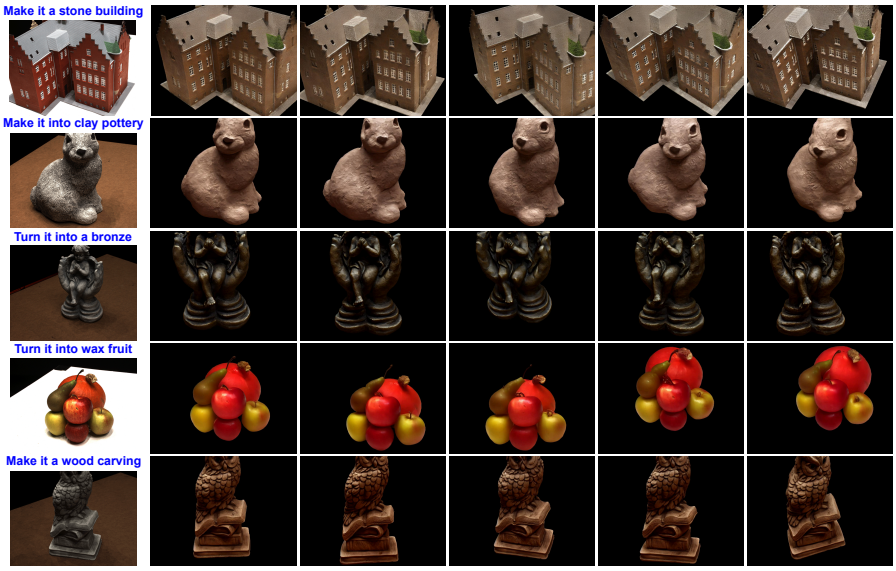

Fig. 3: More text-guided scene editing examples using our proposed pipeline.

including ours, still struggle with thin structures and complex fine-grained details in out-of-domain scenes. This indicates the importance of assessing generalizable reconstruction methods on out-of-domain scenes and calls for more advanced scene representations and image-based rendering pipelines.

4 Additional Visualizations of Text-guided Scene Editing

We provide additional examples of text-guided scene editing in Fig. 3, further showcasing the ability to ensure both instruction-following and 3D consistency. We anticipate that with more powerful backbone models pretrained on larger datasets, as mentioned in Sec. 8 of this supplementary material, the achievable editing effects could become more realistic and diverse.

5 More Detailed Experiment Settings

Datasets. For 3D reconstruction on the DTU dataset [1] in Sec. 3.5 of our main paper, we follow the train/test split described in [4, 6, 9], where 109 scenes are used for model training and 15 scenes are used for testing under a generalizable reconstruction setting. For evaluating the rendering quality on DTU in Sec. 3.5, we use 1/8 of the total views of each scene as test views and select four nearby views from the remaining ones as source views per rendering. For Nvdiffrast-based finetuning on each scene for photorealistic and real-time rendering in Sec. 5.4 of our main paper, 1/8 of the total views of each scene are held out for PSNR evaluation, and the rest are used for finetuning, similar to the strategy in [12].
TSDF fusion. We utilize TSDF fusion [8, 10] to reconstruct scene meshes from predicted source view depths, following [4, 9]. Specifically, for sparse view reconstruction in Sec. 3.5 of our main paper, we adopt 3 views for each test scene,

following [4, 6, 9]. For mesh extraction on NeRF-Synthetic and NSVF, which present larger domain shifts, we adopt a full-view reconstruction setting, where the mesh is acquired by fusing the depth from all views of each scene's training set. For real-time rendering in Sec. 5.4 of our main paper, the initial mesh is acquired by fusing the depth from $1/5$ of the total views of each scene. We adopt a voxel size of 1.5 mm for DTU [1] and a voxel size of 0.01 for NeRF-Synthetic [7] and NSVF-Synthetic [5], based on the scales of the scenes.

Training settings. During training, we employ $N = 4$ source views with a resolution of 640×512. The ray number per batch and the batch size are set to 1024 and 2, respectively, following [4, 9]. The initial learning rate is set to $1e-3$ and is decayed to $1e-6$ using a cosine learning rate schedule, following [4]. In accordance with [4, 9], we adopt hierarchical sampling in both training and testing, with 64 points for coarse sampling and 64 points for fine sampling. During testing, we set the image resolution to 800×600. In addition, for the global feature volume $\mathbf{V}$, we adopt a resolution of 128, following [4]. For the CNN encoder used to extract source features $\{\mathbf{F}_i\}_{i=1}^{N}$, we employ the multi-resolution feature extractor proposed by [4].

Nvdiffrast-based mesh finetuning settings. For the Nvdiffrast-based finetuning on DTU in Sec. 5.4 of our main paper, we train the model for a maximum of 5 minutes, which corresponds to 50 epochs. We adopt initial learning rates of 0.1 and 0.001 for the mesh and shader, respectively, which are decayed by a factor of 0.1 using a cosine learning rate schedule. Additionally, every 10 epochs, we prune redundant mesh faces that are invisible under all training camera poses, i.e., those that cannot intersect with any camera ray in the previous epoch.

Text-guided scene editing. For each target scene to be edited, we select 10 source views and apply the iterative editing and reconstruction pipeline discussed in Sec. 5.3 of our main paper. We stop the iterative process after 5 to 10 iterations, when the edited scenes converge with a good balance between 3D consistency and adherence to instructions.

6 Detailed Formulation of the Subtraction Attention

For the subtraction attention adopted by the appearance transformer $\mathbf{M}_{sdf}^{appr}$ mentioned in Sec. 3.3 of our main paper, we provide a more detailed description in this section. Specifically, the input query features $\mathbf{x} \in \mathbb{R}^{R \times S \times C}$ of $\mathbf{M}_{sdf}^{appr}$ are the output of the previous geometry transformer $\mathbf{M}_{\text{sdf}}^{\text{geo}}$, where R and S represent the number of rays and the number of sampled points along each ray, respectively. The key and value features are the appearance features $\{\mathbf{f}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times R \times S \times C}$. Subtraction attention, which is found to be more effective for geometric relationship reasoning [11, 14], computes the attention scores $\mathbf{A} \in \mathbb{R}^{N \times R \times S \times C}$ between the query features $\mathbf{x}$ and the key features $\{\mathbf{f}_i\}_{i=1}^{N}$ in a subtractive manner, where a broadcast is performed along the view dimension N. Next, the softmax-normalized attention scores are used to perform a weighted sum of the value features along the view dimension N, producing the final output with the shape $\mathbb{R}^{R \times S \times C}$. This process can be formulated as follows:

Table 1: The quantitative performance achieved by different design variants and our full design in sparse view mesh reconstruction in terms of Chamfer distance (the lower, the better) on 15 testing scenes from DTU, following Sec. 3.5 of our main paper.

Design	Mean	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122
Appear feat input	1.44	1.58	2.78	1.62	1.15	1.75	2.03	1.10	1.52	1.41	0.97	1.20	1.47	0.68	1.18	1.23
Vol reso=32	1.41	1.44	2.74	1.55	0.98	1.67	1.84	1.06	1.60	1.53	1.02	1.10	1.69	0.66	1.16	1.13
Vol reso=64	1.38	1.38	2.81	1.64	1.01	1.51	**1.64**	1.09	1.63	1.42	1.01	1.11	1.42	0.71	1.18	1.17
Vol reso=96	1.26	1.13	2.38	1.59	0.99	1.37	1.68	0.94	1.43	1.35	0.99	1.10	1.01	0.63	1.21	1.14
Full design	**1.13**	**0.91**	**2.13**	**1.52**	**0.93**	**1.09**	1.70	**0.84**	**1.29**	**1.20**	**0.83**	**1.04**	**0.81**	**0.55**	**1.05**	**1.05**

$$\mathbf{M}_{sdf}^{appr}(\mathbf{x}, \{\mathbf{f}_i\}_{i=1}^N) = \text{SubAttention}\left(\mathbf{q}=\mathbf{x}, \mathbf{k}=\mathbf{v}=\{\mathbf{f}_i\}_{i=1}^N\right) \tag{1}$$

$$= \sum_i \text{softmax}(\{\mathbf{A}_i\}_{i=1}^N)\mathbf{W}_v(\{\mathbf{f}_i\}_{i=1}^N), \tag{2}$$

$$\text{where } \mathbf{A}_i = \mathbf{W}_q(\mathbf{x}) - \mathbf{W}_k(\mathbf{f}_i) \tag{3}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the linear layers for generating the query, key, and value matrices, respectively.

Additionally, the other two attention modules, $\mathbf{M}_{\text{sdf}}^{\text{geo}}$ and $\mathbf{M}_{\text{sdf}}^{\text{occ}}$ in Sec. 3.3 of our main paper, employ standard attention across sampled points along the same ray, i.e., $\text{softmax}(\mathbf{W}_q(\mathbf{q})\mathbf{W}_k(\mathbf{k}))\mathbf{W}_v(\mathbf{v})$, given the query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$.

7 Ablation Studies on Our NeRF Backbone

In this section, we examine (1) which input features encode richer geometric clues that contribute more to accurate geometry estimation: the appearance features $\{\mathbf{f}_i\}_{i=1}^N$ or the feature volume V in Sec. 3.2 of the main paper, and (2) the contributions of the three transformer components, introduced in Sec. 3.3 of our main paper, to the final reconstruction accuracy.

7.1 Which Input Features Encode Richer Geometric Clues

Multiview stereo methods [2, 13] suggest that the 3D geometric information of a spatial point can be inferred by analyzing the disparities observed when the point is viewed from multiple perspectives. As such, we hypothesize that the feature volume $\mathbf{V}$, which contains the variance of features projected on each source view as introduced in Sec. 3.2 of our main paper, encodes richer geometric clues. Therefore, we utilize it as the input to the transformer-based geometry branch throughout our main paper. To validate this assumption, we perform an ablation study by comparing our design in the main paper with two variants: (1) the same model architecture but using the appearance features, which undergo max pooling along the source view dimension to align with the input dimensions of the following transformer modules, as the input to the geometry branch, and (2) the same model architecture but with reduced feature volume resolutions, resulting in less accurately encoded disparities.

Table 2: The quantitative performance achieved by our full design, compared to that with specific components removed, in sparse view mesh reconstruction in terms of Chamfer distance on 15 testing scenes from DTU, following Sec. 3.5 of our main paper.

Design	Mean	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122
no $\mathbf{M}_{\text{sdf}}^{\text{geo}}$	1.17	1.02	2.29	**1.50**	0.99	1.18	1.56	0.90	1.38	1.23	0.88	1.06	0.83	0.61	1.11	1.11
no $\mathbf{M}_{\text{sdf}}^{\text{appr}}$	1.20	1.09	2.45	1.64	0.99	1.24	**1.55**	0.87	1.40	1.23	0.88	1.08	**0.77**	0.58	1.09	1.12
no $\mathbf{M}_{\text{sdf}}^{\text{occ}}$	1.23	1.07	2.61	1.58	1.03	1.16	1.59	0.92	1.38	1.26	0.95	1.16	0.87	0.61	1.09	1.16
Full design	**1.13**	**0.91**	**2.13**	1.52	**0.93**	**1.09**	1.70	**0.84**	**1.29**	**1.20**	**0.83**	**1.04**	0.81	**0.55**	**1.05**	**1.05**

Observations and analysis. As shown in Tab. 1, we observe that (1) Using appearance features as inputs leads to a +0.31 increase in the Chamfer distance averaged across all test scenes, i.e., significantly degraded mesh quality, compared to our original design that uses the feature volume as inputs. This validates our hypothesis and indicates that the feature volume indeed encodes richer geometric clues compared to appearance features; (2) A lower feature volume resolution results in a higher Chamfer distance and greater mesh quality degradation, underscoring that a high-quality feature volume encoding accurate disparities across source views is crucial for accurate geometry reconstruction.

7.2 Contributions of Each Transformer Component

We evaluate the contributions of each transformer component $\mathbf{M}_{\text{sdf}}^{\text{geo}}$, $\mathbf{M}_{\text{sdf}}^{\text{appr}}$, and $\mathbf{M}_{\text{sdf}}^{\text{occ}}$ by removing each one individually and benchmarking against our original full design presented in the main paper.

Observations and analysis. As shown in Tab. 2, we can observe that (1) Generally, removing only one component from our backbone can still ensure decent reconstruction quality, as compared to the baselines in Tab. 1 of our main paper; (2) Removing $\mathbf{M}_{\text{sdf}}^{\text{occ}}$ results in the largest increase in Chamfer distance (i.e., the most significant mesh quality degradation) among the three components, which we conjecture is because the self-attention in $\mathbf{M}_{\text{sdf}}^{\text{occ}}$ mainly undertakes the responsibility in modeling occlusion effects across sampled points along the same ray and thus is more crucial. In comparison, removing $\mathbf{M}_{\text{sdf}}^{\text{geo}}$ results in the least mesh quality degradation, likely because the geometry features have already been utilized as the input to the geometry branch; (3) Enabling all components leads to the highest quantitative mesh quality, validating the effectiveness of our design in Sec. 3.3 of our main paper.

8 Limitations and Future Work

Although our method can achieve SOTA performance in generalizable mesh reconstruction and scene segmentation, we have identified two limitations: (1) Our work encounters an inevitable trade-off between geometry reconstruction and rendering quality. Emphasizing the former in training could challenge the learning of the latter given the limited capacity of the scene representation. In future work, we aim to push forward this trade-off by training larger models on

larger-scale datasets with accurate depth and camera poses in addition to DTU; (2) While our method surpasses previous SOTA methods [4,9] in reconstructing more challenging scenes, as demonstrated in Sec. 3.5 of our main paper as well as in Sec. 3 of this supplementary material, existing generalizable 3D reconstruction methods generally struggle with thin structures and fine-grained details in scenes with larger domain gaps compared to the training data. This motivates us to develop more advanced scene representations, training schemes, and data sampling strategies that focus more on these challenging scenes when training on larger-scale datasets. We believe the insights we provide could spark future innovations in more advanced generalizable 3D reconstruction pipelines.

9 Potential Negative Societal impacts

Similar to general 3D reconstruction and editing solutions, our work has two potential negative societal impacts: (1) the misuse of 3D editing to create highly realistic yet fake 3D assets that contribute to the spread of misinformation; and (2) the recreation of proprietary 2D/3D assets through reconstruction/editing, which may infringe on intellectual property rights. Both concerns necessitate the development of techniques for reliable and safe 3D reconstruction/editing, such as detection technologies to distinguish real from synthetic content and embedding 3D watermarks in NeRF/3D assets for copyright protection.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**, 153–168 (2016)
2. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
3. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
4. Liang, Y., He, H., Chen, Y.: Retr: Modeling rendering via transformer for generalizable neural surface reconstruction. Advances in Neural Information Processing Systems **36** (2024)
5. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)
6. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: European Conference on Computer Vision. pp. 210–227. Springer (2022)
7. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)

8. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE international symposium on mixed and augmented reality. pp. 127–136. Ieee (2011)
9. Ren, Y., Zhang, T., Pollefeys, M., Süsstrunk, S., Wang, F.: Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16685–16695 (2023)
10. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15598–15607 (2021)
11. Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al.: Is attention all nerf needs? arXiv preprint arXiv:2207.13298 (2022)
12. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
13. 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). pp. 767–783 (2018)
14. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)