

SUPPLEMENTARY MATERIALS

3D RECONSTRUCTION WITH GENERALIZABLE NEURAL FIELDS USING SCENE PRIORS

Anonymous authors

Paper under double-blind review

A APPENDIX

In the supplementary document, we introduce more implementation details (Sec. A), comparison with RGB-D surface reconstruction on ScanNet (Sec. B.1, Sec. B.2 and Sec. B.3), single-view novel view synthesis (Sec. B.4) and qualitative results (Sec. B.5 and Sec. B.6). We specifically discussed the importance sampling methods w.r.t the two-stage generalizable prior training, which plays an important role in the surface representation, as described in Sec. 3.2 in the main paper. We provide additional experiments including (i) quantitative results of neural scene prior, *i.e.*, without per-scene optimization, (ii) quantitative comparisons with state-of-the-art RGB-D surface reconstruction approaches, (iii) quantitative comparisons to the MVS based approaches, (iv) single-view novel view synthesis, and (v) qualitative results on ScanNet and self-collected data. **Videos of full-size room reconstruction are included and recommended to watch.**

A IMPLEMENTATION DETAILS

A.1 GENERALIZABLE NEURAL SCENE PRIOR

The generalizable neural scene prior is trained on the training split of ScanNet Dai et al. (2017). We discuss some details of every component including geometry encoder, texture encoder, generalizable geometric prior module and generalizable geometric prior in this section.

Geometry and Texture Encoder. For the geometry encoder, we first sample 512 keypoints from all the points projected from 2D pixels, via Farthest Point Sampling (FPS) for each frame. For each surface point, we apply the K-Nearest-Neighbor algorithm to select 16 adjacent points. Then, we adopt two PointConv Wu et al. (2019) layers, to extract the geometry feature whose output channels are set to 64. To extract the RGB feature we use a U-Net Ronneberger et al. (2015) with ResNet34 He et al. (2016) as the backbone network. We further use an additional convolutional layer to output a per-point feature with the dimension as 32. All the encoder modules are jointly trained with the whole pipeline.

Generalizable Geometric Prior. Given an RGB-D image and its corresponding camera pose, we first randomly sample 256 rays from regions where depth values are valid, *e.g.*, non-zero. Then for each ray, we define a small truncation region near the ground-truth depth where 32 points are sampled uniformly. We then use two MLPs to map the geometry features to SDF values. The hyperparameters λ_{depth} , λ_{sdf} and λ_{eik} are set to 1.0, 1.0 and 0.5, respectively.

Generalizable Texture Prior. Initialized with the geometric prior, we learn the texture prior via the volumetric rendering loss Wang et al. (2021); Yariv et al. (2021). Different from the sampling strategy used in geometric prior learning, we restrict the importance sampling to the samples concentrated on the surface as described in Sec.3.4 of our main paper. In particular, we first sample 2048 rays from each RGB-D image where we uniformly sample 64 points in the predefined near-far region. Following Wang et al. (2021), then, we sample 32 points that are close to predicted surface. For rays with non-zero depth values, we further sample 16 points within the truncation region around the ray's depth. Therefore, 128 points are sampled along each ray. For each point, we utilize 2 MLPs in the texture decoder to estimate its RGB value. The hyperparameters λ_{depth} , λ_{sdf} , λ_{eik} and λ_{rgb} are set to 1.0, 1.0, 0.5 and 10.0, respectively.

Scene prior extraction and fusion. To leverage multiple views of RGB-D frames, with the scene prior networks, we can directly aggregate the keypoints along with their geometry and texture feature from these frames in the volumetric space. Then, the colored surface reconstruction can be decoded from the fused representation following the same procedure in Sec.3.1 and 3.2. No further learnable modules are required, in contrast, to [Chen et al. \(2021\)](#); [Zhang et al. \(2022\)](#); [Li et al. \(2022\)](#).

Prior-guided pruning and sampling. To optimize a single scene, we discard the encoders and treat the volume feature representation as learnable to be optimized together with the decoders. To further speed up the optimization, we accelerate the feature query process of sampled points, i.e., instead of optimizing the unstructured keypoints, from which the feature extraction can be inefficient, we introduce the prior-guided voxel pruning to leverage the advantage of voxel-grid sampling and surface representation. Specifically, we initialize uniform grids in the volumetric space and then query each grid feature. Instead of optimizing a large number of uniform grids, we remove the redundant grids adaptively based on the geometric prior using the Algo. 1 described below. To concentrate the sampled ray points near the surface, we apply an importance sampling strategy, similar to that used in training the generalizable texture prior, to mask out those far away from the surface. Starting from a large threshold at the early training stage, we decrease it gradually with more training iterations to prune more unnecessary grids. A similar procedure is also applied to the coarsely sampled points to remove some useless points and help more points concentrate around the surface region. Notably, compared to the voxel-based approach [Wang et al. \(2022\)](#) having more than 4,000,000 uniform grids to be optimized, the number of learnable keypoints in our case is around 40,000 – a 100x reduction in computational complexity.

Algorithm 1: Prior-guided voxel pruning

Input: Grid feature $\{f_i\}_{i=1:N}$;
Grid position $\{x_i\}_{i=1:N}$;
Positional encoding $\gamma(\cdot)$;
Geometry decoder $s(\cdot)$;
Number of grids N ;
Number of iterations T
Output: Grid feature after prune $\{f_j\}_{j=1:M}$
Initialization: $\tau_0 = 0.16$
for $t = 1 : T$ **do**
 $\tau = \max(0.005, 0.8^{\frac{20t}{T}} \cdot \tau_0)$
 for $i = 1 : N$ **do**
 $s_i \leftarrow s(f_i, \gamma(x_i))$;
 if $|s_i| \geq \tau$ **then**
 Prune i -th grid
 end
 end
end

B ADDITIONAL EXPERIMENTS

B.1 COMPARISON WITH RGB-D SURFACE RECONSTRUCTION ON SCANNET

Clarification of Table 2 in main paper. Table 2 of the main paper presents the comparison of our method with ManhattanSDF ([Guo et al., 2022](#)) and MonoSDF ([Yu et al., 2022](#)) with the depth supervision. For the fairness, we keep every components of each method by involving an additional depth loss. Unlike some RGB-D surface reconstruction methods, we did not optimize the camera pose while during training. In the course of these modifications, both ManhattanSDF and MonoSDF were observed to have an architecture quite similar to NeuralRGBD [Azinović et al. \(2022\)](#). Given these circumstances, we are confident that comparing our approach with ManhattanSDF and MonoSDF on ScanNet is indeed fair and effective.

Comparison with Go-surf ([Wang et al., 2022](#)) and NeuralRGBD ([Azinović et al., 2022](#)). We compare our method with Go-surf and Neural RGB-D in Table 1. To have a fair comparison, instead

Method	# frames	opt. time	Prec \uparrow	Recall \uparrow	F-score \uparrow
Neural-RGBD Azinović et al. (2022)	400	240	0.932	0.918	0.925
Go-surf Wang et al. (2022)	400	35	0.946	0.956	0.950
Ours	400	15	0.947	0.962	0.954
Neural-RGBD Azinović et al. (2022)	40	240	0.837	0.855	0.846
Go-surf Wang et al. (2022)	40	35	0.842	0.861	0.851
Ours	40	15	0.858	0.866	0.862

Table 1: Quantitative comparisons for mesh reconstruction on ScanNet.

Method	per-scene optim	opt. (min)	Acc \downarrow	Comp \downarrow	Prec \uparrow	Recall \uparrow	F-score \uparrow
Manhattan SDF (Guo et al., 2022)	✓	640	0.072	0.068	0.621	0.586	0.602
MonoSDF (Yu et al., 2022)	✓	720	0.039	0.044	0.775	0.722	0.747
Ours-prior	✗	≤ 5	0.084	0.057	0.695	0.764	0.737

Table 2: Quantitative comparisons of neural scene prior on ScanNet. Both Manhattan SDF and MonoSDF require to optimize on a specific scene for several hours, while the proposed neural scene prior can achieve comparable performance without any optimization.

of optimizing camera poses and neural scene representation jointly, we fix the original camera poses as provided by ScanNet [Dai et al. \(2017\)](#). Follow the same setting in the main paper, we report the performance of different models training with dense and sparse training views. As shown in Table 1, our approach achieves better performance over all metrics. More importantly, although Go-surf achieves similar performance within relative similar time, it cannot produce any reasonable results without optimization as demonstrated by our approach.

B.2 MODEL EFFICIENCY

We take Go-surf [Wang et al. \(2022\)](#), which is so far one of the most efficient offline scene reconstruction approach, as the reference. Compared to it achieving an average run-time of 35 mins per scene, our Neural Scene Prior network takes only **5 mins** (note that the Neural Scene Prior is a feed forward network). The full pipeline leveraging the per-scene optimization stage takes an average run-time of 15 mins, which is still obviously more efficient. More importantly, our model takes a surface representation that facilitate scaling up to larger scenes, compared to architectures in Go-surf that based on dense voxel. Comprehensive comparison of run-times can be found in the Table 1 of the main paper.

B.3 COMPARISON WITH MVS-BASED METHODS

We show quantitative comparisons of our method with the state of the art approaches on surface reconstruction in Table 3. Different from what the Table 1 reported in the main paper, we mainly compare with the MVS-based methods here. For a fair comparison, we follow the evaluation script used in [Zou et al. \(2022\)](#) for computing 3D metrics on the ScanNet testing set. The top part of Table 3 includes offline methods while the middle one contains online methods with the fusion strategy. The bottom part of the table shows the methods that are finetuned on individual scenes. Compared to most MVS-based works that use a fusion strategy, our method achieves much better results in terms of F-score and normal consistency. Moreover, our method outperforms MonoNeuralFusion [Zou et al. \(2022\)](#), which also performs finetuning for individual scenes, by a large margin.

B.4 NOVEL VIEW SYNTHESIS

Novel View Synthesis. We show more qualitative results on novel view synthesis on ScanNet [Dai et al. \(2017\)](#) in Fig. 1 following the same setting described in the main paper. Both NerfingMVS ([Wei et al., 2021](#)) and Go-surf ([Wang et al., 2022](#)) fail on scenes with complex geometry and large camera motion (bottom two rows). The generalized representation enables the volumetric rendering to focus on more informative regions during optimization and improves its performance for rendering RGB images of novel views.

Single-view Novel View Synthesis. We demonstrate NFP enables high-quality novel view synthesis from single-view input (Fig. 2, mid), which has been rarely explored especially at on the scene-level, and potentially enable interesting applications, *e.g.*, on mobile devices.

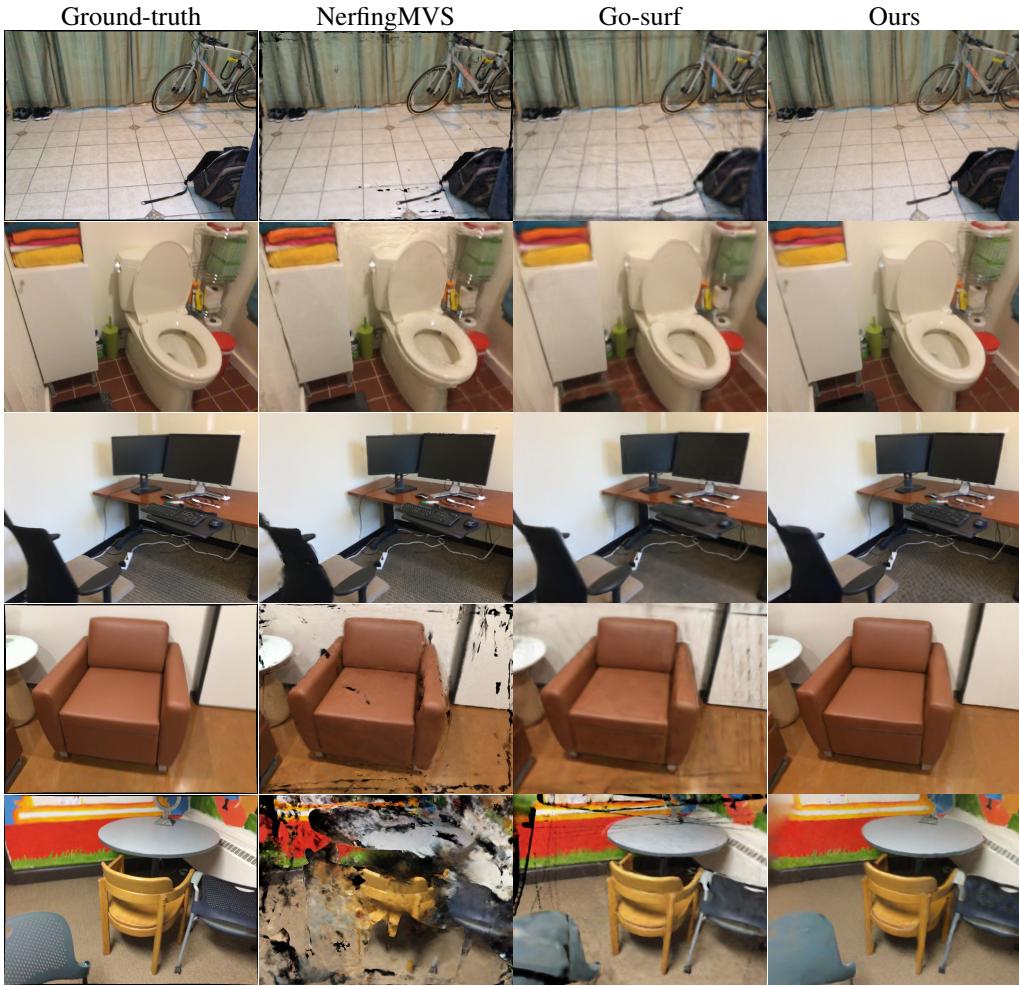


Figure 1: Qualitative comparison for novel view synthesis on ScanNet.



Figure 2: **Qualitative results for single-view novel view synthesis.** The left column shows the training source view, and the appearance reconstruction of the novel view are reported in the second column. The ground-truth images are listed at the last column as reference. **Better viewed when zoomed in.**

	Acc ↓	Comp ↓	Chamfer ↓	Precision ↑	Recall ↑	F-score ↑	NC ↑
FastMVSNet Yu & Gao (2020)	0.052	0.103	0.077	0.652	0.538	0.588	0.701
PointMVSNet Chen et al. (2019)	0.048	0.115	0.082	0.677	0.536	0.595	0.695
Atlas Murez et al. (2020)	0.072	0.078	0.075	0.675	0.609	0.638	0.819
GPMVS Hou et al. (2019)	0.058	0.078	0.068	0.621	0.543	0.578	0.715
DeepVideoMVS Duzceker et al. (2021)	0.066	0.082	0.074	0.590	0.535	0.560	0.765
TransformerFusion Azinović et al. (2022)	0.055	0.083	0.069	0.728	0.600	0.655	-
NeuralRecon Sun et al. (2021)	0.038	0.123	0.080	0.769	0.506	0.608	0.816
MonoNeuralFusion Zou et al. (2022)	0.039	0.094	0.067	0.775	0.604	0.677	0.842
Ours	0.086	0.068	0.077	0.917	0.889	0.875	0.878

Table 3: Quantitative comparisons of mesh reconstruction on ScanNet.

B.5 QUALITATIVE RESULTS OF MESH RECONSTRUCTION

We show qualitative comparisons of our method with other baselines in Fig. 3. It demonstrates that the reconstructed mesh results of our approach are consistently more coherent and detailed than others. In addition, we show the qualitative results of textured mesh for different scenes that obtained via neural scene prior in Fig. 4. **More video demos of texture mesh reconstruction can be found in the supplementary video.**

B.6 MESH RECONSTRUCTION ON THE LARGE-SCALE SCENE

Our results demonstrate that the neural scene prior we propose can generalize well to large-scale scenes, as shown in Fig 5. In contrast to the previous four scenes, we selected a larger room from ScanNet Dai et al. (2017) and applied our pre-trained model directly. The left figure in Fig 5 displays the mesh reconstruction obtained from the neural scene prior. Remarkably, our approach successfully recovers the geometry structure of the entire room with very sparse views (60 frames), without requiring any optimization process. Furthermore, by optimizing the prior on this scene for only 20 minutes on a single NVIDIA V100 GPU, we were able to achieve high-quality mesh reconstruction.

B.7 MESH RECONSTRUCTION ON THE SELF-CAPTURED SCENE

To further demonstrate the robustness of the neural scene prior, we evaluate the pretrained model on a self-captured living room and the reconstructed mesh w./w.o texture are shown in Fig. 6. Impressively, even without per-scene optimization, the proposed neural scene prior is capable of feasibly reconstructing a textured mesh.

C LIMITATION

The proposed neural scene prior could extract the geometric and texture prior for arbitrary scenes, but it does require the sparse RGB-D images as the input. To adapt this neural scene prior for RGB images, one possibility would be to initially create a sparse point cloud using Structure from Motion (SfM) on RGB images. However, as of our submission time, we haven’t yet experimented with this particular setup. Exploring this pathway in future research could certainly yield intriguing findings.

D REPRODUCIBILITY STATEMENT

All experiments in this paper are reproducible. We are committed to releasing the source codes once accepted.

E USE OF EXISTING ASSETS.

As mentioned in the NeurIPS 2023 checklist, we describe the existing assets we used in our paper and the corresponding license of these assets.

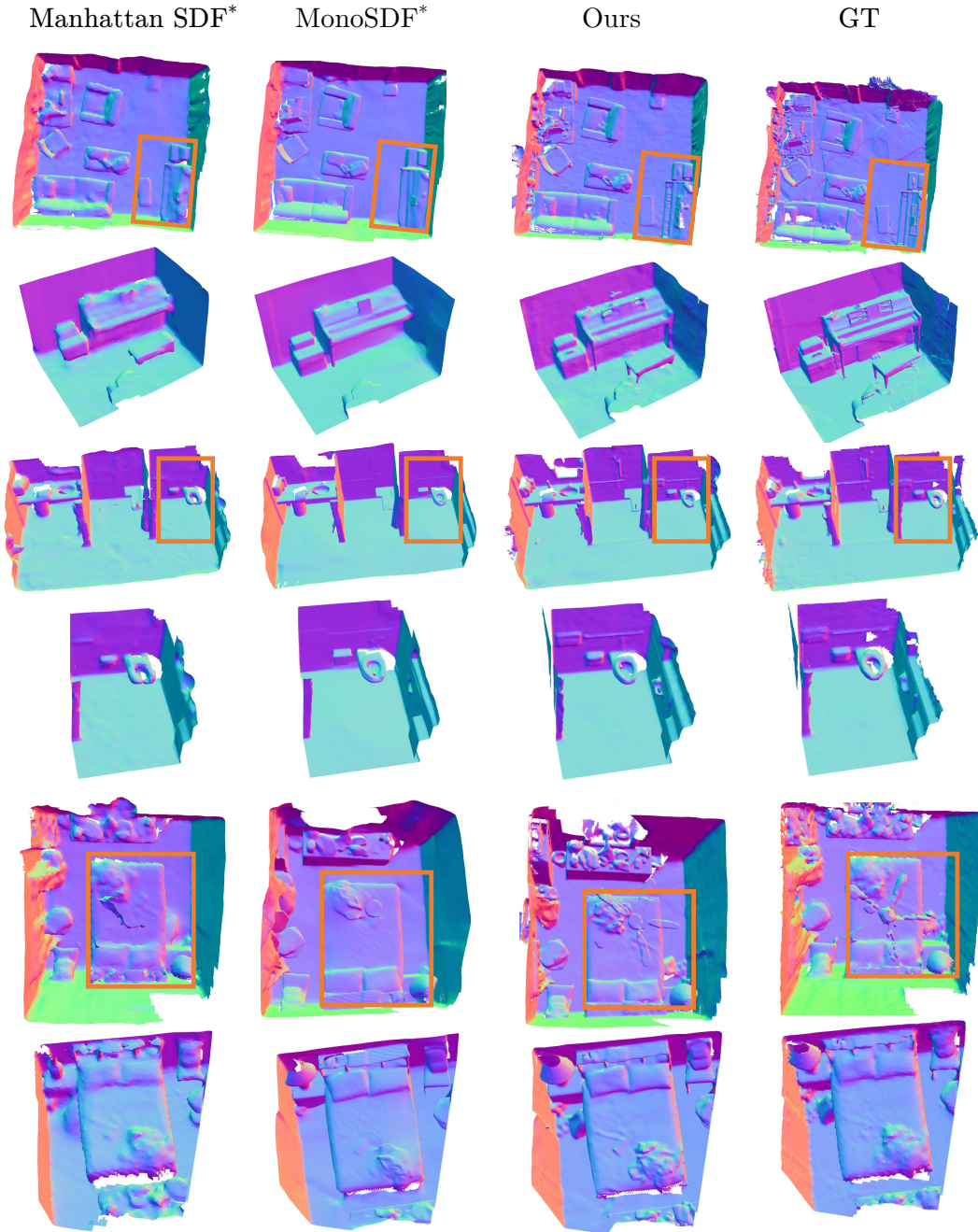


Figure 3: **Qualitative comparisons of mesh reconstruction on ScanNet.** Selected local regions are highlighted by the orange bounding box. **Better viewed when zoomed in.**

Datasets Most of experiments are conducted on ScanNet dataset and 10 synthetic scenes collected by Dai et al. (2017) and Azinović et al. (2022) which are released on their official website and public to everyone for non-commercial use.

Code. Our code is built upon the Pytorch [Paszke et al. \(2019\)](#). And we leverages the code from the released codes by nerfstudio [Tancik et al. \(2023\)](#) under the Apache License.



Figure 4: **Qualitative results of Neural Scene Prior on ScanNet.**

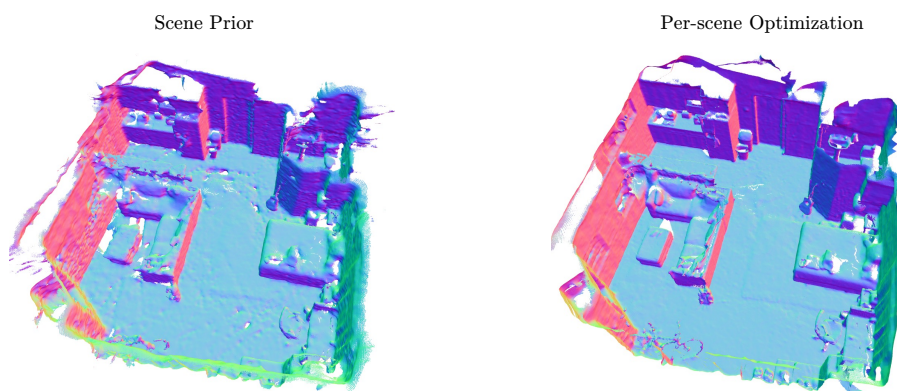


Figure 5: **Mesh reconstruction results on the large-scale scene.**

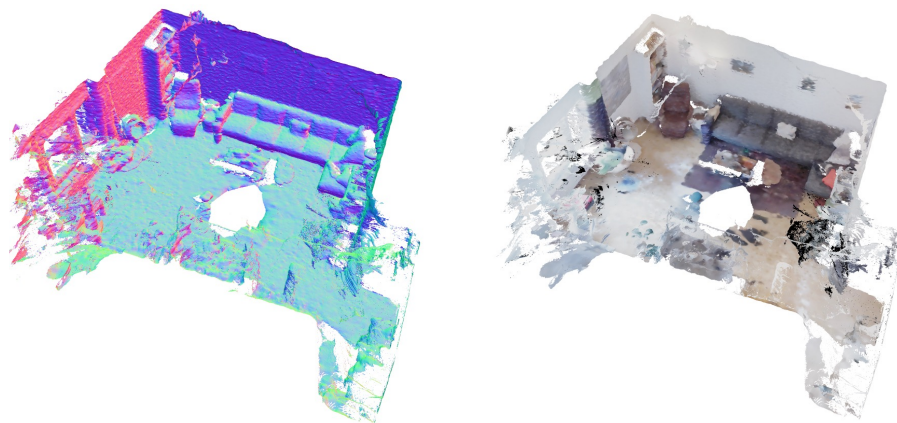


Figure 6: **Mesh reconstruction results on the self-collected scene without any optimization.**

F PERSONAL DATA AND HUMAN SUBJECTS

The dataset does not include the facial or other identifiable information of humans.

G ETHICAL CONCERNS.

The datasets used are standard benchmark proposed in previous works. Despite applying supervised learning, there may still be potential bias in our model trained with these datasets.

REFERENCES

- Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6290–6301, 2022. 2, 3, 5, 6
- Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnr: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021. 2
- Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, pp. 1538–1547, 2019. 5
- Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017. 1, 3, 5, 6
- Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15324–15333, 2021. 5
- Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5511–5520, 2022. 2, 3
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1
- Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *ICCV*, pp. 2651–2660, 2019. 5
- Kejie Li, Yansong Tang, Victor Adrian Prisacariu, and Philip HS Torr. Bnv-fusion: Dense 3d reconstruction using bi-level neural volume fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6166–6175, 2022. 2
- Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 5
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015. 1
- Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. 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. 5

- Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 6
- Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. *arXiv preprint arXiv:2206.14735*, 2022. 2, 3
- Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1
- Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5610–5619, 2021. 3
- Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630, 2019. 1
- Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1
- Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1949–1958, 2020. 5
- Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2, 3
- Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5449–5458, 2022. 2
- Zi-Xin Zou, Shi-Sheng Huang, Yan-Pei Cao, Tai-Jiang Mu, Ying Shan, and Hongbo Fu. Mononeuralfusion: Online monocular neural 3d reconstruction with geometric priors. *arXiv preprint arXiv:2209.15153*, 2022. 3, 5