Self-Supervised Implicit 3D Reconstruction via RGB-D Scans

1st Hongji Yang
College of Computer Science
Nankai University
Tian Jin, China
hongjiyang@mail.nankai.edu.cn

2nd Jiao Liu
College of Computer Science
Nankai University
Tian Jin, China
jiaoliu@mail.nankai.edu.cn

3rd Shaoping Lu
College of Computer Science
Nankai University
Tian Jin, China
slu@nankai.edu.cn

4th Bo Ren*
College of Computer Science
Nankai University
Tian Jin, China
rb@nankai.edu.cn

***Abstract*—Recently, 3D reconstruction methods based on the neural radiance fields have demonstrated remarkable generative performance. However, these methods frequently tend to be resource hungry and are challenging to regulate large low-textured regions in typical indoor scenes. In this work, we analyze and integrate inherent semantic geometry cues for self-supervised 3D reconstruction training via a unified framework of volume rendering and signed distance implicit representations. In contrast to previous neural implicit methods, we simultaneously incorporate the pixel-aligned features and image patches for multi-view consistency, thereby enabling us to depict a large indoor scene from challenging scenarios with rich visual details and large smooth backgrounds. Extensive experiments and comparisons demonstrate that our proposed method has achieved state-of-the-art results by a large margin in various tasks (e.g. actual surface reconstruction, novel view synthesis, and learning a universal scheme in occlusion or distorted regions).**

***Index Terms*—Indoor reconstruction, implicit neural rendering, self-supervised guidance**

I. Introduction

Neural rendering techniques on 3D surface reconstruction are being hotly pursued and have made impressive progress in virtual reality, augmented reality, and scene animation. Recently, Neural Radiance Fields (NeRF) [1] and its variants [2]–[6] have achieved unprecedented view synthesizing fidelity by designing a compact neural volume representation. These methods focus on rendering photo-realistic images in a view-consistent manner and it is challenging to extract a zero-crossing area (i.e. the surface). Noticing that signed distance function (SDF) serves to recover subtle geometry details in a zero level set manner, neural implicit representation [7]–[10] have been proposed to avoid surface ambiguity caused by the uniform density accumulation. However, these methods assume that only the intersection of the ray near the surface votes for the appearance calculation and ignore the background estimation. Later, Wang et al. combined radiance field representation and neural SDF representation and proposed a new paradigm (i.e. NeuS [11]) to extract high-fidelity surfaces. This implicit SDF rendering scheme can be a dramatic enhancement in learning accurate surface intersections from RGB scans and corresponding object masks. However, reconstructing large cluttered indoor scenes with texture-less regions can be easily stuck in local optima and it remains a challenging task.

To address this issue, [12]–[14] present a new benchmark with Manhattan-world constraint or additional normal priors to enforce the training process. By incorporating monocular geometric cues, these methods recover relatively smooth boundaries between accurate 3D objects and free space. However, heavily relying on a well-designed layout manner and carefully pre-trained priors as supervision severely limits the portability of user engagement from unseen datasets.

In our approach, we conjecture that neural implicit representation will achieve better with fine-grained correspondence from semantic segmentation corpora. To this end, we assemble the homogeneous pixels counting up to 50 which implies texture-less regions, and leverage pixel gradient examination in cross-entropy loss to learn a hybrid guideline between the predicted semantic segmentation item and the depth change information. A comprehensive set of qualitative and quantitative experiments conducted on room-scale datasets ScanNet [15] and Replica [16] show that our method achieves high fidelity results without fine-tuning compared to state-of-the-art methods.

To summarize, the main contributions of this paper are:

- We introduce a novel self-supervised framework to effectively handle correlation patterns between depth maps and semantic labels in the task of room-scale 3D reconstruction given only RGB-D scans.
- We design a depth-check module to prevent overwhelming the training process, especially in challenging scenarios with radical depth changes.

II. Related work

A. Classical Multi-view Stereo Methods.

Some traditional multi-view stereo (MVS) methods estimate multi-plane images [17]–[20] and apply depth fusion [21] in a point cloud to reconstruct the scene. The key point of this feature matching procedure is to build correspondence [22] in a multi-view consistency manner. Voxel-based representation [23] is also a trend to extend MVS in exploiting photometric consistency. MVS algorithms perfectly adapt to inhomogeneous sampling patch and require low memory to

*Corresponding author.

This work was supported by the Natural Science Foundation of China (No. 62132012).

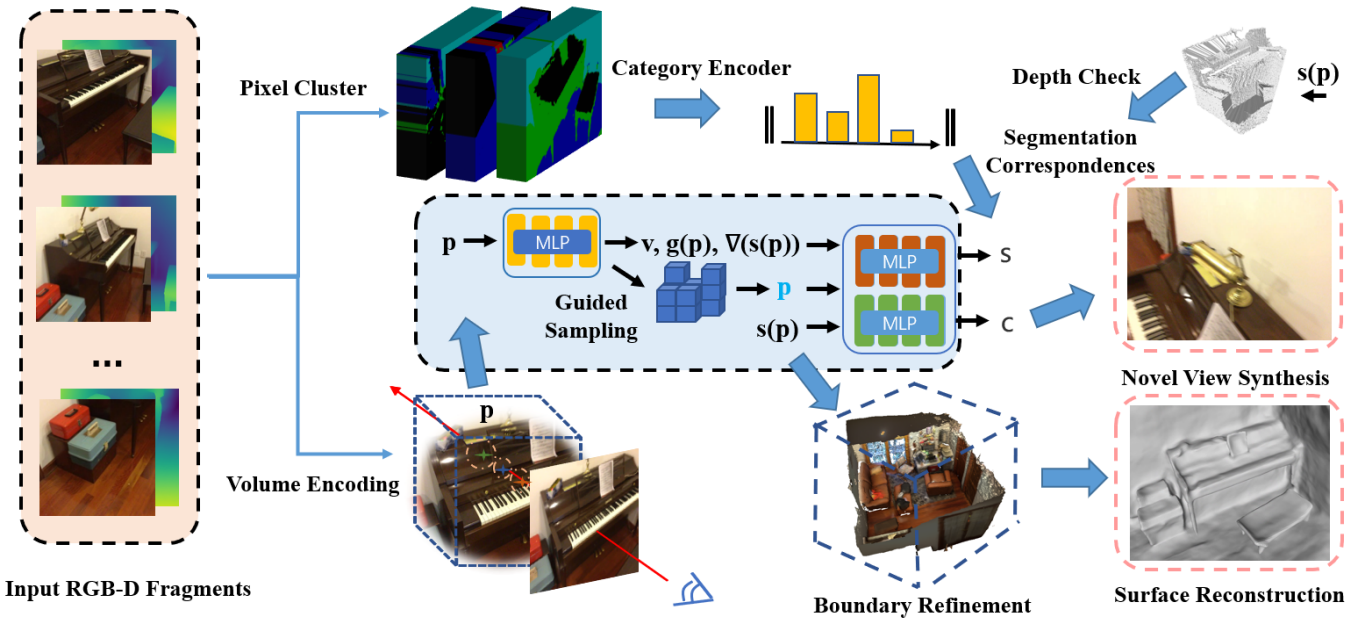


Fig. 1: **Overview of our optimization pipeline.** Given RGB-D scans, we sample a batch of rays and decouples the semantic information with depth-check in two separate side branch. The former branch fits the implicit Volume Rendering network with high-level semantic annotations extracted from the later category correspondence encoder, enabling color prediction via volume rendering (**Green** MLP blocks), semantic segmentation (**Bittersweet** MLP blocks) and surface reconstruction (**Dandelion** MLP blocks).

storage cost features. However, these methods may predict an infinity density of objects in case of trivial solutions due to their visual inconsistency nature in volume density formulation.

B. Neural Implicit Scene Reconstruction.

Recently, the seminal work NeRF [1] explores multi-layer perceptrons (MLPs) to regress 3D volumetric densities and colors with sinusoidal functions encoding. Neural rendering is suitable to predict the color map for novel views without any preliminary mask prefetching. However, without depth constraints, it's hard to resemble an accurate topological surface distribution, especially in real-world indoor scenes proposed of large texture-less regions. In follow-up work, NeRF has been extended with depth supervised [24]–[26] from smarter sampling sparse points by running structure-from-motion (SFM). It implies that reasons to predict closest-surface depth outweigh reasons to accumulate ordinary 3D points. Despite decent synthesis performance and flexible capture requirements, such an approach inevitably constructs imprecise depth completion when minimizing the KL divergence with estimated camera poses and intrinsics from poor COLMAP optimization.

Instead of predicting the ray distribution and depth uncertainty, we leverage implicit density accumulation strategy and L2 losses to achieve robust reconstruction, which can also avoid ambiguities in the presence of rich texture areas and generate visually-appealing images. Inspired by [27], we introduce a novel edge-aware smooth term that only small disparity semantic regions will contribute to the local density field, making it easier to perform density checks on top of volume rendering.

C. Prior Guided 3D Scene Optimization.

There have been many prior guided works that have emerged as a promising trend by acquiring geometry or appearance cues. [28], [29] predict dense 2D semantic labels and fuse them to 3D scene, which enables more effective learning for 3D representation. Manhattan-SDF [13] jointly optimize the scene appearance and geometry based on Manhattan-world assumption. Although achieving accurate parallel or orthogonal constraints to indoor texture-less regions, the process is not scalable for fitting small objects because of specified three semantic predictions (e.g. floor, wall, or the background regions). Instead of naive brute-force training, NeuRIS [14] incorporate estimated normal prior into geometry modeling and output an appealing 3D geometry. However, without actual depth range measurements, NeuRIS is tough to determine whether the sampled points along the rays are near the surface or at a large distance from the surface, thereby reasoning about an over-smooth object surface of tiny objects. Besides, due to the heavy burden of carefully training normal clues, this standalone method is not capable of on-site image captures.

To address the distortion and incompleteness problems, we leverage unsupervised semantic constraints in the early training stage and depth-check module in later surface completion.

III. METHODOLOGY

A. Overview.

In this paper, we propose an end-to-end framework guided by unsupervised semantic constraints while only given commercial RGB-D Kinect camera captures with corresponding intrinsic and camera poses. Though the pixel's color is related to the surface geometry, performing reconstruction with the same backbone and adaptive prediction heads makes it vulnerable to creating accurate level sets. Thus we first sample rays and map the object surface into a zero-level set. After identifying suitable geometry in coarse volume rendering, we utilize depth priors for precise measurement of the distances to objects and perform STEGO predictor [30] to get segmentation results that serve as supervision in a principled unified formulation for relatively smooth layout(floor, desk, sofa, etc.) generation. Fig 1 demonstrates an overview of our proposed approach.

Implicit Volume Rendering. Following NeRF, given a training set of indoor images, our model samples M points $\mathbf{p} = \{\mathbf{p}_i = \mathbf{o} + \mathbf{v}d_i | i = 1, ..., M\}$ from camera center $\mathbf{o}$ along the view direction $\mathbf{v}$. Then sampled points are mapped into signed distance $s(\mathbf{p})$, geometry features $g(\mathbf{p})$ and surface gradient $\nabla(s(\mathbf{p}))$ through a coarse MLP network $f_{\theta_d} : \mathbb{R}^3 \rightarrow \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^3$ which can be expressed as:

$$(s(\mathbf{p}), g(\mathbf{p}), \nabla(s(\mathbf{p}))) = f_{\theta_d}(\gamma(\mathbf{p})), \tag{1}$$

After coarse geometry prediction, signed distance $s(\mathbf{p})$ is invoked to model the density $\sigma(\mathbf{p})$ with learnable parameter β in a more tractable transformation.

$$\sigma(\mathbf{p}) = \begin{cases} \frac{1}{2\beta} \exp\left(\frac{s(\mathbf{p})}{\beta}\right) & \text{if } s(\mathbf{p}) \leq 0 \\ \frac{1}{\beta} - \frac{1}{2\beta} \exp\left(\frac{-s(\mathbf{p})}{\beta}\right) & \text{if } s(\mathbf{p}) > 0. \end{cases} \tag{2}$$

Akin to conditional volume rendering techniques, we denote the accumulated alpha weights $\mathcal{W}$ as follows:

$$\mathcal{W} = \sum_{i=1}^{n} T_i \alpha_i, \tag{3}$$

where $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ denotes the accumulated transmittance along the ray, $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ is the discrete alpha value at point p_i where δ_i is the distance between neighboring sampled points.

Then we reason about color c and categorical semantic s through a similarly designed fine MLP network $f_{\theta_{\{c,s\}}} : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^N \rightarrow \mathbb{R}^3$ which can be formalised as:

$$f_{\theta_{\{c,s\}}}(\gamma(\mathbf{p})) = (\mathbf{p}, \mathbf{v}, g(\mathbf{p}), \nabla(s(\mathbf{p}))). \tag{4}$$

The color map $\hat{C}$ and semantic map $\hat{S}$ along sampled rays are approximated as a weighted sum of every query point:

$$\{\hat{C}, \hat{S}\} = < \mathcal{W}, f_{\theta_{\{c,s\}}} >, \tag{5}$$

where $< ., . >$ means Frobenius inner product.

B. Semantic Constraints for Scene Representation.

Unlike the existing semantic guided work [13], we observe that in addition to background geometry (e.g. wall, floor, etc.), the horizontal planar like a large table in the council chamber or modern loveseat sofa contributes a lot to relative position among surrounding objects. An accurate 2D object mask is tough to generate in a cluttered scene, so we tame an encoder with five categories $\{\mathcal{O}_i \subset \mathbb{R}^3 | i = 1, \ldots, 5\}$ inside the scene to depict the disparity between semantic map prediction and semantic segmentation via maximum likelihood data clustering:

$$min \sum_{i=1}^{5} w_i y_i (log \frac{exp(\hat{S}_{\mathcal{O}_i, y_i})}{exp(\sum_{i=1}^{5} \hat{S}_{\mathcal{O}_i, y_i})}), \tag{6}$$

where w_i is area-imbalanced weight and y_i is the ground truth patch segmentation prediction, resolving ambiguities during the early coarse layout searching process.

Noticing that among homogeneous regions there exist an abrupt depth gap, which will definitely induce a very large misguidance. These depth uncertainties, also known as edge points, have larger gradient values in the image, while areas with continuous depth have smaller gradient values. Inspired but different from edge-aware smoothness item [31], we rearrange the semantic loss function as:

$$\mathcal{L}_s = \sum_{i=1}^{5} w_i y_i (log \frac{exp(\hat{S}_{\mathcal{O}_i, y_i} e^{|\nabla I_t^*|})}{exp(\sum_{i=1}^{5} \hat{S}_{\mathcal{O}_i, y_i} e^{|\nabla I_t^*|})}), \tag{7}$$

where I_t^* is sampled color input and we regress semantic constraints in the plausible areas where pixel gradient ∇I_t^* is below the threshold ϵ. For edge points, the pixel gradient will rush into an extreme value, which should not be taken into consideration and have a better constraint on the predicted depth map.

C. Optimization.

While enforcing color integration with a density field from sampled $\mathcal{R}$ 2D pixels, L1-norm photometric loss is defined as:

$$\mathcal{L}_c = \sum_{r \in R} \| \hat{C}(r) - C(r) \|_1. \tag{8}$$

Reconstructing 3D geometry from only 2D color input is an ill-posed problem, especially in poorly textured areas. To supervise the complex surface distance relationship, we utilize L2-norm depth loss to terminate viewing rays at the opaque object:

$$\mathcal{L}_d = \sum_{r \in R} \| < \mathcal{W}, d > - \bar{D}(r) \|^2, \tag{9}$$

where $< \mathcal{W}, d >$ is the rendered depth values, $\bar{D}(r)$ is the real depth from the depth sensor.

Due to the partially reflective or occluded conditions in volume rendering, the photometric loss may provide error clues in the region of similar appearance. Eikonal loss [34]

Method	ScanNet					Replica				
	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
COLMAP [32]	0.035	0.167	0.760	0.403	0.527	0.098	0.144	0.604	0.485	0.538
NeRF [1]	0.701	0.182	0.153	0.295	0.201	0.573	0.421	0.085	0.166	0.112
UNISURF [33]	0.486	0.172	0.195	0.338	0.247	0.399	0.386	0.298	0.335	0.315
NeuS [11]	0.107	0.126	0.524	0.465	0.493	0.312	0.167	0.406	0.437	0.421
VolSDF [10]	0.234	0.131	0.317	0.442	0.369	0.227	0.103	0.489	0.546	0.516
M-SDF [13]	0.049	0.060	0.747	0.643	0.691	0.112	0.096	0.588	0.602	0.595
Ours*(w/o $\mathcal{L}_s$)	0.085	0.102	0.557	0.502	0.528	0.187	0.126	0.521	0.535	0.528
Ours*(w/o $\mathcal{L}_d$)	0.117	0.094	0.492	0.538	0.514	0.085	0.072	0.622	0.637	0.629
Ours	0.054	0.037	0.736	0.764	0.758	0.042	0.064	0.729	0.665	0.695

TABLE I: **Quantitative comparisons over room-scale scenes of state-of-the-art methods on ScanNet and Replica.**The best, second and third scores are highlighted in red, blue and teal, respectively. Following [13], we select F-score as the most representative metric for 3D surface reconstruction.

TABLE II: Quality metrics for novel view synthesis on Scannet datasets. Best results are **highlighted**.

Model	Metrics
	PSNR↑ / SSIM↑ / LPIPS↓
NeRF	28.29 / 0.922 / 0.119
NeuS	27.15 / 0.824 / 0.146
Ours*(full model)	**28.44 / 0.933 / 0.067**

is utilized to regularize the sampled points, thereby enabling the network to generate a clear boundary condition:

$$\mathcal{L}_{\text{eik}} = \sum_{p \in P} (\|\nabla f_\theta(p)\|_2 - 1)^2, \tag{10}$$

where P denote uniformly distributed points in the bounding box.

Leaving uncontrollable 3D surface normal cues, we take an initial step by leveraging a surface normal suppression component, which means the set of near-surface points should ensure spatial coherence and smooth consistency:

$$\mathcal{L}_{\text{reg}} = \sum_{p \in P} (\|\nabla f_\theta(p) - \nabla f_\theta(p+\epsilon)\|_2)^2, \tag{11}$$

where ϵ is the perturbation near the predicted surface.

In summary, the overall training loss L is defined as a weighted sum of the following six loss terms.

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_d + \lambda_2 \mathcal{L}_s + \lambda_3 \mathcal{L}_{reg} + \lambda_4 \mathcal{L}_{eik}, \tag{12}$$

where λ_1, λ_2, λ_3, λ_4 are the trade-off weight hyperparameters.

IV. EXPERIMENTS

A. Datasets.

In our experiments, we investigate our network on two room-level datasets (e.g. ScanNet(v2) [15] and Replica [16]). ScanNet provides more than 1500 room-scale scenes with corresponding high-quality reconstructed mesh. It takes about 1.2T to download the whole datasets indicating that it is computationally prohibitive for training all scenarios. We select 5 scenes of different categories for testing. Similarly, we perform our method on 6 scenes with high resolution and fine-grained surfaces from Replica. According to the video length of each scene, 100∼700 sampled views with accompanying camera parameters are applied to the network training.

B. Implementation Details.

Building on NeuS [11], our method is implemented in Pytorch. The SDF backbone network and appearance (semantic) module are modeled by an 8-layer MLP with a Softplus activation function and a 4-layer MLP with a Relu activation respectively. We utilize the ADAM as the optimizer with a learning rate of 2e-4 and randomly sample a batch size of 512 rays. Based on 64 query points in the coarse volume, we add extra 64 fine surface-guided sampled points which generate a reasonable search range. The first training process takes 100k iterations with λ_2 initialized to 1.0, while decay over training to 0.1 to refrain from radiance ambiguity problem. Our model can be trained for about 2.5 hours using a single GeForce RTX 3090 GPU with 6GB GPU memory.

C. Quantitative Results.

As shown in Table I, our method outperforms existing state-of-the-art baselines to a large extent. More precisely, we improve over the previous SOTA method Manhattan-SDF (abbreviated as M-SDF) by an average margin of 0.067 in F-score, which means higher accuracy and better completeness. For the metric Acc. (Accuracy), COLMAP and M-SDF achieve slightly higher results than our approach due to inconsistency elimination process and blindness for recovering fine-grained object regions.

Further, we also conduct ablation studies on the contributions of the two main components(e.g. depth constraints and semantic constraints). In contrast to training the neural implicit network without these two items, our method is able to optimize the depth prediction via semantic segmentation constraints which promise harmonization between depicting the space distribution and recovering subtle color details.

D. Qualitative Results.

Fig 2 demonstrates qualitative geometry reconstructed by our approach and other rendering-based methods. Note that NeRF lacks sufficient constraint on the surface and produces noisy meshes. Neus utilizes SDF representation, which can reconstruct a rough shape (see bed and single sofa). SOTA method M-SDF integrates Manhattan-world constraint to constrain smoothness in weak textures regions. Compared with the ground truth, our approach successfully reconstructs subtle

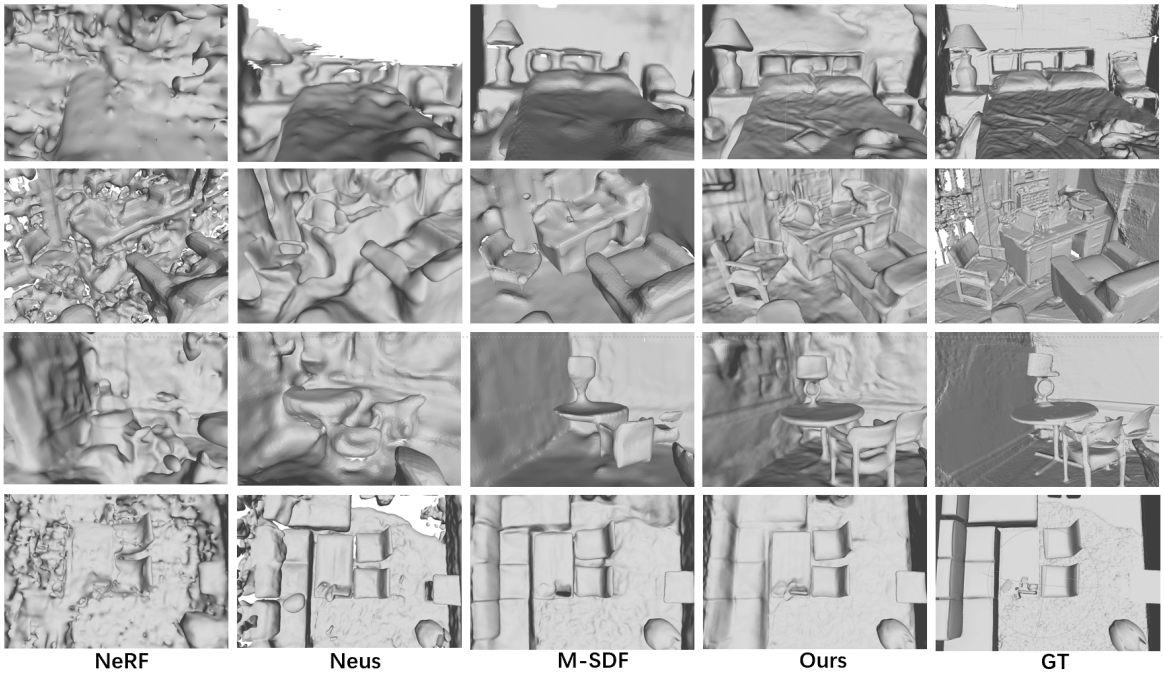
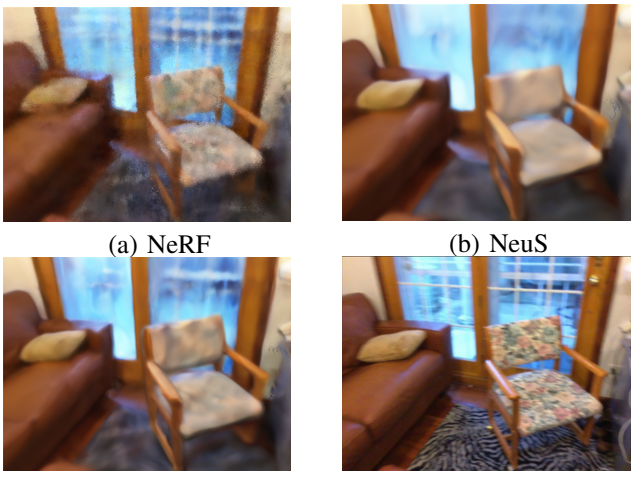


Fig. 2: **Visual comparisons on rendered meshes.** We compare our model with typical learning-based methods NeRF, Neus, M-SDF. NeRF and Neus tend to eliminate polygon parts and produce redundant results. M-SDF achieves smooth reconstruction in planar regions but fails to reconstruct fine texture details. In contrast, our method is capable of maintaining high completeness and accuracy.

texture details such as bed sheets in a bedroom scene and laptops jumbled on the worktable.



(a) NeRF (b) NeuS

(c) Ours (d) GT

Fig. 3: **Qualitative comparison for novel view synthesis task on the Scannet dataset.** Our method is well designed to handle geometry dissimilarity in NeRF and image content degradation in Neus.

E. Novel View Synthesis Results.

We also conduct an in-depth analysis of novel view synthesis. To demonstrate the robustness of the networks, 6 long-distance views with rich-colored furniture are set over 5 test scenes. Fig 3 demonstrates a rendered test frame from a randomly selected camera pose. NeRF is specifically designed for novel view synthesis and only utilizes a color constraint which inevitably generalizes discretization artifacts and misinterpretation of the geometry. In contrast, NeuS integrate the implicit volumetric SDF into the radiance field. Though surface rendering is guaranteed, we can observe that rendering images from NeuS lack high-frequency details due to the dispersion of training goals. Thus, our full model gathers category information from partitioning clustering which helps alleviate the radiance ambiguity issue and sufficiently constrain the color degeneration.

As shown in Table 2, our method with full model outperforms the reconstruction method NeuS by a large margin for over 1dB. Even in comparison with synthesizing method NeRF, our approach also achieve competitive results, especially in the perceptual distance (e.g. LPIPS).

V. CONCLUSIONS

In this work, we propose a room-scale scene reconstruction method using a self-supervised implicit neural network. We follow the protocol of previous prior-guided methods but distinguish it from theirs by acquiring deep features in an unsupervised manner which ensures flexibility of training and surface completeness in case of abrupt depth changes with RGB-D inputs. Our method yields a significant improvement in the task of surface extraction and novel view synthesis on both the Scannet and Replica datasets. In future work, we will apply key point supervision into our model for meticulous training of neural implicit surfaces.

REFERENCES

[1] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*, 2020. 1, 2, 4

[2] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[3] 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 *IEEE International Conference on Computer Vision*, 2021. 1

[4] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[5] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu, "Editable free-viewpoint video using a layered neural representation," *ACM Transactions on Graphics (TOG)*, 2021. 1

[6] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, 2022. 1

[7] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser, "Learning shape templates with structured implicit functions," in *IEEE International Conference on Computer Vision*, 2019. 1

[8] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[9] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," in *Conference and Workshop on Neural Information Processing Systems*, 2020. 1

[10] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman, "Volume rendering of neural implicit surfaces," in *Conference and Workshop on Neural Information Processing Systems*, 2021. 1, 4

[11] 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," in *Conference and Workshop on Neural Information Processing Systems*, 2021. 1, 4

[12] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao, "Neuralrecon: Real-time coherent 3d reconstruction from monocular video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[13] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 4

[14] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang, "Neuris: Neural reconstruction of indoor scenes using normal priors," in *European Conference on Computer Vision*, 2022. 1, 2

[15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 4

[16] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe, "The Replica dataset: A digital replica of indoor spaces," *arXiv*, 2019. 1, 4

[17] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[18] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *IEEE International Conference on Computer Vision*, 2021. 1

[19] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys, "Itermvs: iterative probability estimation for efficient multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[20] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu, "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[21] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *European Conference on Computer Vision*, 2018. 1

[22] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, 2009. 1

[23] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger, "Raynet: Learning volumetric 3d reconstruction with ray potentials," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[24] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[25] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[26] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola, "Nerf-supervision: Learning dense object descriptors from neural radiance fields," in *International Conference on Robotics and Automation*, 2022. 2

[27] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[28] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru, "Virtual multi-view fusion for 3d semantic segmentation," in *European Conference on Computer Vision*, 2020. 2

[29] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison, "In-place scene labelling and understanding with implicit scene representation," in *IEEE International Conference on Computer Vision*, 2021. 2

[30] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *International Conference on Learning Representations*, 2021. 3

[31] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[32] Johannes Lutz Schönberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4

[33] Michael Oechsle, Songyou Peng, and Andreas Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *IEEE International Conference on Computer Vision*, 2021. 4

[34] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman, "Implicit geometric regularization for learning shapes," in *Proceedings of Machine Learning and Systems 2020*. 2020. 3