ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888042

OARecon: Object-Aware Viewpoint Augmentation for Indoor Compositional Reconstruction

Yuanyuan Ding[1], Yiming Fei[2], Jiandang Yang[1], Xiaobin Wei[3], Jiajun Lv[1,*], Yong Liu[1,*]
[1] Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3] WASU Media & Network Co.Ltd., Hangzhou, China

Abstract—**Real-world scenes likely involve repetitive objects indicating that the reconstruction of the target object can be supplemented by the views of other identical objects. However, traditional 3D reconstruction methods do not take this a priori knowledge into account and fail to make full use of the available information. In this paper, we propose an object-aware viewpoint augmentation scheme for indoor compositional reconstruction. Within this scheme, a viewpoint supplementation strategy based on signed distance function and neural radiance fields is proposed to fully leverage the information from repetitive objects such that the occlusion problem is suppressed. Moreover, this scheme introduces monocular uncertainty priors and regional smoothness constraints to enhance the reconstruction accuracy of slender and thin structures and the smoothness of occluded background, respectively. Experimental results considering both synthetic and real-world scenes demonstrate that our method effectively improves the reconstruction quality of repetitive objects and background.**

Index Terms—**signed distance function, neural radiance fields, object-aware indoor reconstruction**

I. Introduction

Emerging neural implicit representation rendering methods have demonstrated considerable results in novel view synthesis [1] and 3D reconstruction [2], [3] recently. Neural Radiance Fields (NeRF) encode scene properties into a Multi-Layer Perceptron (MLP) via volumetric rendering, training the scene's volumetric radiance field by minimizing the difference between rendered images and real images [1]. However, due to the lack of direct geometric supervision, while these methods can implicitly learn 2D geometry during training, they perform poorly in extracting 3D meshes. To address this, some approaches [4]–[6] have been proposed to learn continuous Signed Distance Function (SDF) and color, enabling better reconstruction of complex 3D geometry. Some studies have provided additional information that aids in contextual understanding and scene navigation [7]–[9], thereby facilitating the learning of object-composed 3D scene representations from visible light images and semantic masks. However, these methods perform poorly in reconstructing the geometry of individual objects due to the insufficient view information included in the limited dataset. Additionally, the background often lacks smoothness in occluded areas, which means that even though the object is decoupled from the background, it still cannot fully support downstream scene editing applications.

ObjSDF++ [10] proposed an object-compositional NeRF 3D reconstruction model that encodes object semantics into neural implicit representations. This model not only renders RGB images from the SDF network but also generates 2D semantic maps. By calculating the opacity of each point along the light ray, it effectively identifies objects and their occlusion relationships, resulting in cleaner separations. However, this occlusion recognition is limited to visible areas. Even with multi-view images, parts of objects that are occluded or not visible due to sparse viewpoints, such as the seat under a table or the legs of a table behind a sofa, remain difficult to reconstruct. Therefore, additional strategies like object-guided assistance are required to help the network fully utilize known information.

In real-world scenes, different views of repetitive identical objects can provide supplementary information for the reconstruction of each of them. Following this idea, we can even reconstruct the occluded parts of repetitive objects with the same data setting. By leveraging these repetitions, the number of effective views used for the reconstruction of the same object increases. Considering the high-quality demands of downstream applications such as scene editing, additional issues need to be addressed to overcome the limitations of the current viewpoint supplementation based method: (i) for thin or slender structures, even with abundant viewpoint information, traditional reconstruction methods struggle to ensure reconstruction accuracy; (ii) the geometric smoothness of occluded backgrounds is difficult to guarantee in traditional reconstruction methods.

To address the aforementioned issues, we propose using pose transformations and virtual camera setups for complementary viewpoints of repetitive objects, incorporating monocular uncertainty priors and regional smoothness constraints to enhance the reconstruction accuracy of slender and thin structures and the smoothness of the occluded background, respectively [11], [12]. Based on the above analysis, we propose an editable indoor scene reconstruction enhancement scheme named **O**bject-**A**ware Viewpoint Augmentation for Indoor Compositional **Recon**struction (OARecon), aiming to improve the accuracy of indoor scene reconstruction while enabling flexible editing capabilities. The main contributions can be summarized as follows:

- By using pose transformations and virtual camera setups, we propose a method mapping from a single-view multi-object scene to a multi-view single-object setup, leveraging repetitive objects to learn geometry more accurately without additional data.
- An indoor compositional reconstruction scheme emphasizing on the reconstruction quality of both objects and background is proposed.

II. Methodology

A. Background

The key to learning neural implicit representations from multi-view images lies in volumetric rendering techniques [13]. For a ray $\mathbf{r}(t) = \mathbf{o} + t \cdot \mathbf{d}$, volumetric rendering calculates the target pixel color based on the scene density $\sigma(t)$ and the scene radiance $\mathbf{c}(t)$ at each 3D point along the ray

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))\, dt. \tag{1}$$

$*$ Corresponding authors. Email: yongliu@iipc.zju.edu.cn.

AI-based All-media Service Platform, Major Scientific and Technological Innovation Project of Hangzhou under Grant 2022AIZD0019.

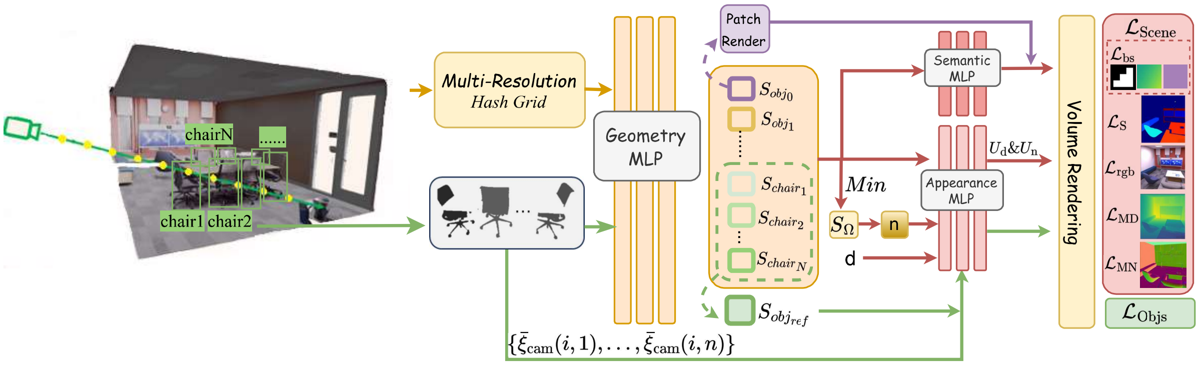

Fig. 1. Overview of **OARecon**. In this work, we propose a multi-view augmentation approach for repetitive objects in indoor scenes. For repetitive objects, Semantic guidance generates multi-instance patches and virtual camera views $\bar{\xi}_{cam}(i,1),\ldots,\bar{\xi}_{cam}(i,n)$ for image $\mathcal{I}(i)$. One instance is selected as a reference, and information from others optimizes $S_{obj_{ref}}$. The Geometry MLP (SDF MLP) outputs the SDF to represent object geometry. For all objects and the overall scene, we render semantics and appearance, optimizing the MLP network through volume rendering. Monocular uncertainty values U_{d} and U_{n} constrain the monocular prior for accurate, disentangled reconstruction. For the background Obj_0, We perform patch rendering and regularize the geometric smoothness of unobserved regions using $\mathcal{L}_{\text{bs}}$.

where t_n and t_f represent the near and far boundaries of the ray, respectively, and the scene transmittance function $T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}(v))\,dv\right)$ quantifies the energy loss as light passes through the scene.

In implicit surface volumetric rendering based on the SDF, the geometry of the scene is represented by the SDF value $s(\mathbf{x})$ at each spatial point $\mathbf{x}$, which describes the distance from point x to the nearest surface. In practice, the SDF function is implemented via a MLP network f. The appearance of the scene, such as the view-dependent color $\mathbf{c}$, is defined by another MLP network g

$$\begin{aligned} f &: \mathbf{x}\in\mathbb{R}^3 \mapsto \left(s\in\mathbb{R}, \mathbf{f}\in\mathbb{R}^{256}\right) \\ g &: \left(\mathbf{x}\in\mathbb{R}^3, \mathbf{n}\in\mathbb{R}^3, \mathbf{d}\in\mathbb{S}^2, \mathbf{f}\in\mathbb{R}^{256}\right) \mapsto \mathbf{c}\in\mathbb{R}^3 \end{aligned} \tag{2}$$

where $\mathbf{f}$ represents a geometry feature vector, $\mathbf{n}$ is the normal at point $\mathbf{x}$, $\mathbf{d}$ is the viewing direction.

Following [5], we replaced NeRF's volumetric density σ with the SDF MLP output.

By combining semantic logits and object masks, we achieve composite reconstruction of multiple objects. To simplify the process, we treat the background as a separate object, similar to the approach in [10], and adopt the same network structure. In a scene with k objects, including n identical repetitive objects, the SDF MLP $f(\cdot)$ generates k SDF values S_j at each point, where $j = 0$ represents the background and $j = 1, 2, ..., k$ represents distinct objects (as shown in Fig. 1). The n repetitive objects are optimized through a dedicated repetitive object optimization network. The scene's SDF is $S_\Omega = \min(S_1, S_2, ..., S_k)$, used for ray sampling and volumetric rendering Eq. (1).

Additionally, the opacity of each object O_{obj} is computed through volumetric rendering, with the supervision of scene transmittance and instance segmentation masks, enabling opacity rendering to capture occlusion relationships.

$$\hat{O}_{obj}(j,\mathbf{r}) = \int_{t_n}^{t_f} T_\Omega(t)\sigma_{obj}(j,\mathbf{r}(t))\,dt, \quad j = 1,\ldots,k. \tag{3}$$

During training, opacity is obtained via volumetric rendering, and the cross-entropy loss applied to the opacity is backpropagated to the SDF values, facilitating the learning of composite geometry.

B. Multi-View Complementation for Repetitive Objects

While approximate occlusion-aware methods can effectively segment objects, in indoor scenes, the occlusion between objects often makes the reconstruction of unseen parts difficult. Moreover, for individual objects, the input from multi-view scenes is often sparse, leading to suboptimal reconstruction. As shown in Fig. 1, one effective method to address these issues is to use all views of other identical objects as information sources for the target object.

Given an RGB image $\mathcal{I}(i)$ captured by a camera with pose $\xi_{\text{cam}}(i)$, there are n identical objects with poses $T_{\text{obj}}(1),\ldots,T_{\text{obj}}(n)$. Taking one of the objects as reference Obj_ref, the relative pose transformation of the other objects with respect to Obj_ref is given by

$$T_{\text{relative}}(j) = T_{\text{obj\_ref}} \circ \left(T_{\text{obj}}(j)\right)^{-1}, \quad j = 1,2,\ldots,n \tag{4}$$

where $\circ$ is matrix multiplication for pose composition.

To ensure that the camera's viewpoint remains consistent when capturing other objects from a new position compared to when it captured the reference object, we need to generate the virtual camera pose based on the pose transformation between the reference object and the supplementary object. The virtual camera pose $\bar{\xi}_{cam}(i,j)$ can be calculated using the following equation:

$$\bar{\xi}_{cam}(i,j) = T_{\text{relative}}(j) \circ \xi_{cam}(i), \quad j = 1,2,\ldots,n. \tag{5}$$

The next step is to aggregate the information from repetitive objects to recover their 3D structure. We combine semantic information with raycasting to determine which object's coordinates to use, and generate the corresponding virtual camera poses $\xi_{\text{cam}}(i)$. Similar to the Scene SDF MLP, we parameterize the Objs SDF MLP. The key difference is that we model the geometry of the objects in local object space to ensure shape consistency across instances. Additionally, the geometry of repetitive objects is optimized within the repeated object optimization network to ensure accurate reconstruction of other instances.

To effectively guide the learning of each object's surface in the scene, we use instance segmentation masks to supervise object opacity. In the repetitive object optimization network, all objects in the scene, including the background, are treated as a single instantiated object and input into the semantic MLP, except for the repetitive objects. However, the occlusion-aware opacity rendering struggles with floating artifacts in unseen regions. Consider SDF properties: for a point $\mathbf{x}$ inside an object t, $S_t(\mathbf{x}){<}0$, while for other objects, $S_j(\mathbf{x}) \geq |S_t(\mathbf{x})|$ where $j = 1, 2, \ldots, k$ and $j \neq t$. Therefore, the following loss function is considered:

$$\mathcal{L}_{\text{dis}} = \mathbb{E}_{\mathbf{P}} \left[\sum_{j=1}^{k} \max\left(0, S_{\Omega}(\mathbf{x}) - S_j(\mathbf{x})\right) \right]. \tag{6}$$

C. Slender Structure and Background Optimization

[14] has demonstrated that monocular geometric priors play a crucial role in indoor scenes. In practice, it has been observed that although the pre-trained model Omnidata [15] provides monocular priors, the results are not always entirely accurate, especially when restoring the slender structures of objects, where monocular priors are often more important than color supervision. As pointed out in [11], if the monocular prior from a certain viewpoint significantly differs from other viewpoints, it is likely inaccurate. Therefore, this paper models the uncertainty of the prior as a view-dependent network representation and estimates the uncertainty values of depth and normal through volumetric rendering

$$g : (\mathbf{x}, \mathbf{n}, \mathbf{d}, \mathbf{f}) \mapsto (c, u_d \in \mathbb{R}, u_n \in \mathbb{R}^3). \tag{7}$$

Depth uncertainty is handled using a masked depth loss function based on the Laplace distribution, which masks areas with high uncertainty in the monocular priors

$$\mathcal{L}_{\text{MDepth}} = \log(U_{\text{d}} + \epsilon) + \frac{|D_{\text{pred}} - D|}{|\omega + \epsilon|}. \tag{8}$$

Similar to the depth loss, the normal priors are transformed into the world coordinate system and undergo uncertainty regularization computation

$$\mathcal{L}_{\text{Mnormal}} = \log(|U_{\text{n}}| + \epsilon) + \left(\frac{\|N_{\text{pred}} - N\|_2}{|\omega| + \epsilon} \right)^2. \tag{9}$$

In Eq. (8) and Eq. (9), U_d and U_n are the predicted uncertainties, the subscript $pred$ represents the network's predicted values, D and N are the pseudo-GT for depth and normal obtained from the pre-trained network, respectively, ϵ is a small value (e.g. $1e{-}8$) to prevent division by zero errors. When $U \leq \tau$, $\omega = U$, otherwise the gradients of $\frac{|D_{\text{pred}} - D|}{|\omega + \epsilon|}$ and $(\frac{\|N_{\text{pred}} - N\|_2}{|\omega| + \epsilon})^2$ corresponding to ω are detached from the computation graph of PyTorch.

In indoor scenes, occluded parts of the background are invisible in all images, leading to points behind the rays that cannot be directly optimized, resulting in random holes and artifacts [12]. Since it is impossible to obtain the true values of the occluded areas, we perform smoothing of the rendered depth and normal on the background surface by randomly sampling a $P \times P$ patch region in the given image. Along the rays in the patch, the depth and normal estimates are computed using only the background SDF. These estimates are then converted into semantic information, leaving only the background and other class mask estimates.

The depth and normal smoothing loss $\mathcal{L}_{\text{bs}}$ is based on the same principle and together form the background smoothing loss.

For example, the formula for the background smoothing loss based on the rendered depth is as follows

$$\mathcal{L}(\hat{D}) = \sum_{d=0}^{3} \sum_{m,n=0}^{P-1-2^d} \hat{M}(r_{m,n}) \odot \Big(\Big| \Delta_x \hat{D}(r_{m,n}, 2^d) \Big| + \Big| \Delta_y \hat{D}(r_{m,n}, 2^d) \Big| \Big). \tag{10}$$

where $\Delta_x \hat{D}$ and $\Delta_y \hat{D}$ represent the depth differences between neighboring pixels, d controls the patch spacing for applying smoothness, and m, n denote the pixel indices within the patch. The mask estimation $\hat{M}$ is used to filter out non-background regions through Hadamard product $\odot$.

D. Training Objective Details

Based on the method from [14], we incorporated depth and normal consistency loss, as well as the approximate occlusion-aware constraints from [10] to improve the geometric structure and object separation. Additionally, the SDF network is regularized with the Eikonal term $\mathcal{L}_{\text{E}}$ [16].

The instance mask M_{dup} for the repetitive object viewpoint supplementation network contains only two categories: repetitive instances and background. Therefore, the loss function for this network is defined as

$$\mathcal{L}_{\text{Objs}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{M\_dup}} + \lambda_d \mathcal{L}_{\text{D}} + \lambda_n \mathcal{L}_{\text{N}} + \lambda_1 \mathcal{L}_{\text{E}} + \lambda_2 \mathcal{L}_{\text{dis}}. \tag{11}$$

To improve the reconstruction of other objects in the scene, the loss function for the scene composition network is defined with the inclusion of monocular prior uncertainty and background smoothness constraints, as follows

$$\mathcal{L}_{\text{Scene}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{M}} + \lambda_{MD} \mathcal{L}_{\text{MD}} + \lambda_{MN} \mathcal{L}_{\text{MN}} + \lambda_1 \mathcal{L}_{\text{E}} + \lambda_2 \mathcal{L}_{\text{dis}} + \lambda_{bs} \mathcal{L}_{\text{bs}}. \tag{12}$$

III. Experiments

A. Experimental Setup

Datasets. We conducted experiments on two datasets: 1) **Replica [17]**: A synthetic dataset that depicts realistic indoor scenes and provides precise camera poses and clear object mask information. In the experiments, we used 4 scenes from this dataset, each containing multiple repetitive chairs; 2) **ScanNet [18]**: A widely used real-world dataset that has been extensively applied in previous works [10], [19]–[21].

Metrics. For the evaluation of reconstruction performance, this paper reports Chamfer Distance and F-score on the Replica dataset, following the methods from [10], [22]. The evaluation metrics are divided into two categories: repetitive objects and the overall scene. In the ScanNet dataset, we specifically observe the reconstruction performance of slender structures and background.

Baseline. We primarily compare with ObjSDF++ [10], which is the SOTA for this task.

B. Implementation Details

The method in this paper is implemented by PyTorch, using the Adam optimizer with a learning rate of 5e$-$4, the same as the baseline. A total of 200k iterations were performed, applying an error-bounded sampling algorithm, with 1024 rays sampled per iteration. The method runs on a single 24G 3090Ti GPU. The SDF MLP consists of two layers of 256-channel neural networks, with weight initialization following the approach in [5], [6], [10]. For the loss functions mentioned in Eq. (11) and Eq. (12), set the weight of RGB reconstruction and Mask loss as 1, $\lambda_1 = \lambda_d = \lambda_{bs} = 0.1$ in our experiments. For other loss terms, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $\lambda_d = 0.1$, $\lambda_n = 0.05$, $\lambda_{MD} = 0.006$, $\lambda_{MN} = 0.0025$, $\lambda_{bs} = 0.01$, and P of patch's size is set as 32 from [12]. The color prediction part consists of four layers of 256-channel networks. Geometric initialization adopts the method from [14], which initializes the reconstruction as a unit sphere.

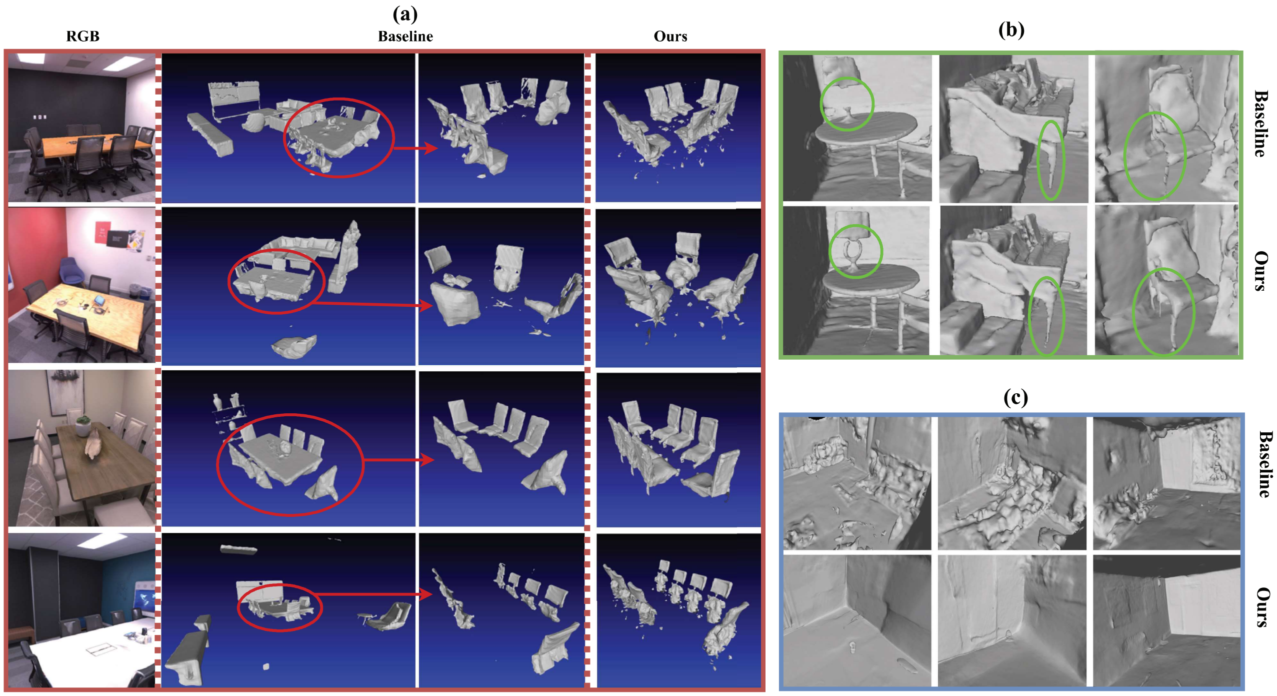

Fig. 2. **Visual effect comparison on Replica and Scannet**. (a) OARecon utilizes the complementary viewpoint information from repetitive chairs to improve the reconstruction quality of each chair such that the occlusion problem has been alleviated. (b) OARecon adopts monocular uncertainty priors such that the slender and thin structures can be reconstructed while ObjSDF++ fails to reconstruct them. (c) The use of regional smoothness constraints improves the smoothness of occluded background.

TABLE I
THE QUANTITATIVE AVERAGE RESULTS FROM REPLICA SCENES EVALUATED ON REPETITIVE OBJECT RECONSTRUCTION.

Replica-Category	Method	Repeated Object Reconstruction	Average	
		F-Score	F-Score	Chamfer-L1
Scan3-Chair	**ObjSDF++**	**92.07** / **87.92** / 76.68 / **81.01** / 66.83 / 46.26 / 44.31 / 54.17	68.66	0.06185
	Ours	89.61 / 76.48 / **80.23** / 68.80 / **74.78** / **74.10** / **73.07** / **66.71**	**75.47**	**0.03782**
Scan6-Chair	**ObjSDF++**	39.63 / 79.61 / 75.47 / 84.07	69.6	0.062
	Ours	**78.45** / **83.23** / **80.08** / **86.03**	**81.85**	**0.03001**
Scan7-Chair	**ObjSDF++**	**92.70** / 76.65 / 85.20 / 93.01 / 64.31 / 44.28 / 37.88 / **85.32**	72.52	0.06145
	Ours	88.09 / **91.46** / **88.02** / **96.62** / **93.22** / **89.14** / **86.12** / 81.16	**89.23**	**0.02418**
Scan8-Chair	**ObjSDF++**	61.90 / 61.06 / 60.22 / 63.87 / 61.48 / 55.20 / 59.34 / 57.89	60.12	0.08075
	Ours	**66.81** / **69.79** / **72.17** / **66.93** / **68.61** / **66.84** / **72.83** / **71.91**	**69.49**	**0.04478**

C. *Performance Comparison*

As shown in Fig. 2, OARecon scheme outperforms the ObjSDF++ method in the reconstruction quality of repetitive objects, slender objects and occluded background.

Additionally, we developed a quantitative evaluation method for each decoupled object in the scene (including the background). TABLE. I provides corresponding quantitative results for different objects from different datasets. Chamfer-L1 and F-score represent the results for the reconstruction of occluded background regions and fully complete objects, where Chamfer-L1 $= \frac{\text{Accuracy}+\text{Completeness}}{2}$

To conclude, the OARecon scheme enhances the reconstruction of repetitive objects, occluded background, and slender structures for indoor scenes, whether in synthetic or real-world data sets, and its quantitative reconstruction metrics outperform ObjSDF++ in most cases.

IV. CONCLUSION

In this paper, a viewpoint supplementation strategy for repetitive objects in indoor scenes is proposed to improve the reconstruction accuracy of each repetitive object, which may be occluded. Building on this, considering the needs of downstream applications such as scene editing, monocular uncertainty priors and regional smoothness constraints are adopted to improve the reconstruction of slender structures and occluded backgrounds, respectively. Based on this design, our scheme named OARecon tackles the challenges of large-scale indoor scene reconstruction tasks. Experiments on the Replica and ScanNet benchmark data sets demonstrate the superiority of our method. In the future, we will consider using generative models to complete the view information of objects that are completely unseen in the dataset to further improve the reconstruction quality. In the future, we plan to integrate object pose estimation into the network, enabling simultaneous pose estimation during reconstruction.

References

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] Michael Oechsle, Songyou Peng, and Andreas Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5589–5599.

[3] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong, "Gram: Generative radiance manifolds for 3d-aware image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10673–10683.

[4] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[5] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman, "Volume rendering of neural implicit surfaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4805–4815, 2021.

[6] 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.

[7] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black, "Resolving 3d human pose ambiguities with 3d scene constraints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2282–2292.

[8] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang, "Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 55–64.

[9] Kejie Li, Hamid Rezatofighi, and Ian Reid, "Moltr: Multiple object localization, tracking and reconstruction from monocular rgb videos," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3341–3348, 2021.

[10] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai, "Objectsdf++: Improved object-compositional neural implicit surfaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21764–21774.

[11] Yuting Xiao, Jingwei Xu, Zehao Yu, and Shenghua Gao, "Debsdf: Delving into the details and bias of neural indoor scene reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2024.

[12] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu, "Rico: Regularizing the unobservable for indoor compositional reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17761–17771.

[13] James T Kajiya and Brian P Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[14] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *Advances in neural information processing systems*, vol. 35, pp. 25018–25032, 2022.

[15] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10786–10796.

[16] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman, "Implicit geometric regularization for learning shapes," *arXiv preprint arXiv:2002.10099*, 2020.

[17] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al., "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[19] 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*, 2022, pp. 5511–5520.

[20] 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*. Springer, 2022, pp. 139–155.

[21] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng, "Object-compositional neural implicit surfaces," in *European Conference on Computer Vision*. Springer, 2022, pp. 197–213.

[22] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison, "vmap: Vectorised object mapping for neural field slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 952–961.