

NTO3D: Neural Target Object 3D Reconstruction with Segment Anything

Xiaobao Wei^{1,2,3} Renrui Zhang⁴ Jiarui Wu^{1,3} Jiaming Liu¹
 Ming Lu⁵ Yandong Guo⁶ Shanghang Zhang^{1†}

¹National Key Laboratory for Multimedia Information Processing, School of Computer Science,

Peking University ²Institute of Software, Chinese Academy of Sciences

³University of Chinese Academy of Sciences ⁴Shanghai Artificial Intelligence Laboratory

⁵Intel Labs China ⁶AI² Robotics

Abstract

Neural 3D reconstruction from multi-view images has recently attracted increasing attention from the community. Existing methods normally learn a neural field for the whole scene, while it is still under-explored how to reconstruct a target object indicated by users. Considering the Segment Anything Model (SAM) has shown effectiveness in segmenting any 2D images, in this paper, we propose NTO3D, a novel high-quality Neural Target Object 3D (NTO3D) reconstruction method, which leverages the benefits of both neural field and SAM. We first propose a novel strategy to lift the multi-view 2D segmentation masks of SAM into a unified 3D occupancy field. The 3D occupancy field is then projected into 2D space and generates the new prompts for SAM. This process is iterative until convergence to separate the target object from the scene. After this, we then lift the 2D features of the SAM encoder into a 3D feature field in order to improve the reconstruction quality of the target object. NTO3D lifts the 2D masks and features of SAM into the 3D neural field for high-quality neural target object 3D reconstruction. We conduct detailed experiments on several benchmark datasets to demonstrate the advantages of our method. The code will be available at: <https://github.com/ucwxb/NTO3D>.

1. Introduction

The neural field has made significant progress over the past few years and become one of the most popular 3D representations. The pioneering Neural Radiance Field (NeRF) [28] and its variants [2, 8, 30, 43, 59] learn coordinate-based neural networks to predict the density and color from multi-view images and use volume rendering to conduct novel view synthesis. NeuS [47] improves the 3D reconstruction quality of NeRF by representing a surface with a Signed

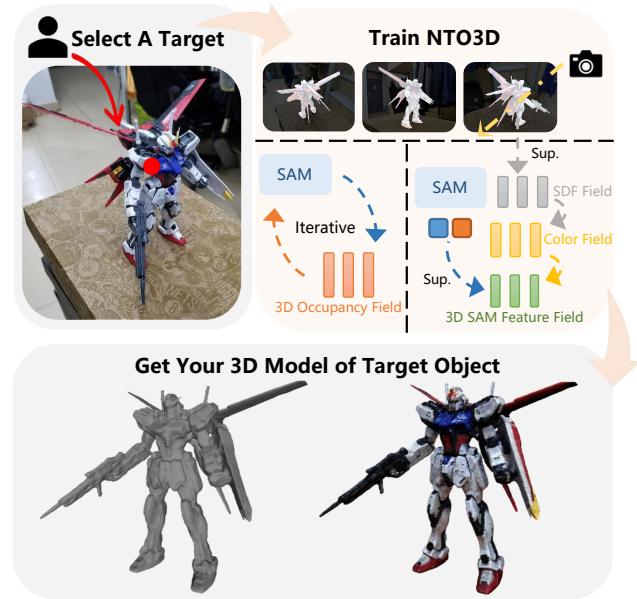


Figure 1. **Overview of NTO3D.** First, a user selects a reconstruction target in the scene. Then, our NTO3D utilizes a 3D occupancy field iteratively to merge the multi-view 2D segmentation masks into 3D space. NTO3D further lifts the features of the SAM encoder into a 3D SAM features field and optimizes the feature field together with other fields. Finally, the user can obtain a high-quality 3D reconstruction model of the target object with NTO3D.

Distance Function (SDF). They also developed a new volume rendering method to train the neural SDF representation. Many studies are proposed to improve the reconstruction quality and reduce the training cost [44, 52]. However, existing methods usually learn a neural field for the whole scene, ignoring the reconstruction of a target object in the scene, which can be indicated by end users on the fly.

Although traditional techniques such as in-hand scanning [51] have been proposed for target 3D object reconstruction, it is still non-trivial for neural 3D reconstruction

[†]Corresponding Author E-mail: shanghang@pku.edu.cn

methods since we need to obtain the multi-view consistent target object segmentation, which is labor-intensive and time-consuming.

Recently, the Segment Anything Model (SAM) [20] has shown great potential for zero-shot segmentation, which can be used to segment a target object out of the scene. However, with a single prompt, SAM can only obtain the 2D segmentation of a single-view image, other than multi-view images. In addition, how to leverage the features of SAM to improve the reconstruction quality is still under-explored to the best of our knowledge.

To address the above issue, we propose NTO3D, a novel high-quality Neural Target Object 3D (NTO3D) reconstruction method that fully leverages the benefits of both neural field and SAM. Specifically, to separate the target object from the neural field, we first train a 3D occupancy field to merge the multi-view 2D segmentation masks. Our 3D occupancy field is based on the following assumptions: (1) If a pixel is foreground, then at least one of the positions passed through the ray is foreground. (2) If a pixel is background, then all the positions passed through the ray are background. We design a corresponding loss based on the assumptions to optimize the 3D occupancy field, lifting the 2D masks to a unified 3D occupancy field. The 3D occupancy field is then projected into 2D space and generates the new prompts for SAM. This process is iterative until convergence to finally segment the target object out of the scene.

After this, in order to improve the reconstruction quality of the target object, we further lift the features of the SAM encoder into a 3D feature field. We add a lightweight output head to the neural field for learning the 3D features of SAM and use volume rendering to render the 2D features. The rendered 2D features are directly supervised by the 2D features of SAM. By lifting the 2D features of SAM, our method can reconstruct a more accurate 3D model for the target object.

Our main contributions are summarized as follows:

- We propose NTO3D, a novel method that iteratively lifts the 2D masks of SAM into a unified 3D occupancy field, segmenting the target object out of the neural field. With our method, users can easily reconstruct any target objects by prompting in a single view.
- To boost the reconstruction quality, we further present a tactful strategy to lift the 2D features of SAM into a 3D feature field.
- We conduct detailed experiments on DTU, LLFF, and BlendedMVS datasets, where NTO3D surpasses the state-of-the-art reconstruction methods, demonstrating the advantages of our approach.

2. Related Work

Neural Implicit Representation. Neural implicit representation has recently become prevailing in computer vision and graphics. This representation utilizes coordinate-based neural networks to represent a field, which can encode continuous signals of arbitrary dimensions at arbitrary resolutions. Neural implicit representation has shown promising results in shape reconstruction [1, 6, 11, 24, 25, 34, 35, 58], novel view synthesis [16, 21, 23, 27, 37, 38, 41, 42, 46] and multi-view 3D reconstruction [15, 17, 22, 31, 57]. In particular, Neural Radiance Fields (NeRF) [28] learn a continuous volume density and radiance field from multi-view images. After training, it can render images from arbitrary views via volume rendering. To improve the surface reconstruction quality of NeRF, NeuS [47] utilizes a Signed Distance Function (SDF) to represent a surface. Voxsurf and NeRF2Mesh[44, 52] are proposed to reduce the training cost and improve the reconstruction quality. However, how to effectively reconstruct a target object with neural implicit representation is still under-explored, as it is difficult to obtain multi-view consistent target object segmentation. Though SA3D [5] achieve neural rendering of target objects with SAM, it fails to impose geometry constraints for neural reconstruction. In this paper, we proposed a unified 3D occupancy field to effectively segment a target object out of the neural field.

Image Segmentation. Great efforts have been made for different segmentation tasks such as semantic segmentation [7], instance segmentation [45], and panoptic segmentation [19]. Various models have also been developed for segmentation, including encoder-decoder structures [36], dilated convolutions [60], pyramid structures [63] and transformers [53]. Recently, the Segment Anything Model (SAM) [20] and its variants [48, 61] have demonstrated strong zero-shot generalization ability, enabling 2D segmentation for diverse real-world target objects. However, SAM is currently limited to 2D segmentation of a single image, which is insufficient for multi-view consistent target object segmentation. Moreover, utilizing features of SAM for improving the 3D reconstruction quality remains under-explored. In this study, we lift the 2D masks and features of SAM into the 3D neural field for high-quality neural target object 3D reconstruction.

3D Reconstruction. The problem of 3D reconstruction has been extensively studied in computer vision with numerous methods proposed for various applications. Traditional RGB-based methods usually rely on multi-view stereo techniques to predict the depth from posed images [9, 40]. Recent learning-based methods aggregate multi-view information to learn the 3D representation of the scene [54, 55]. With the development of neural implicit representation, current methods start to represent the 3D scene with various neural fields [27, 47]. Plenty of methods are also proposed for reconstructing various specific objects such as 3D faces [3, 4], bodies [64], and hands [49].

Apart from RGB cameras, many methods are proposed to use more cameras for 3D reconstruction. For example, KinectFusion [12] enables a user to rapidly create detailed 3D reconstruction by holding and moving a standard RGB-D camera. VoxelHashing [32] improves the regular grid data structure of KinectFusion with a simple spatial hashing scheme that compresses space. AutoRecon [50] leverages self-supervised 2D vision transformer features and reconstruct decomposed neural scene representations with decomposed point clouds, to achieve accurate object reconstruction and segmentation. Although 3D reconstruction is a well-studied problem, how to reconstruct a certain object indicated by users on the fly is still a difficult problem. By leveraging the benefits of both neural fields and SAM, our method enables users to easily reconstruct any target objects by prompting in a single view.

3. Method

In this section, we first briefly review neural object 3D reconstruction. Subsequently, we proceed to elaborate on the pipeline of the proposed Neural Target Object 3D (NTO3D). Finally, we further elucidate the novel designs incorporated in NTO3D.

3.1. Preliminaries

Recent neural object 3D reconstruction works such as NeRF [28] and NeuS [47] both learn coordinate-based neural networks to represent the scene. NeRF constructs a mapping function from spatial location $x \in \mathbb{R}^3$ and view direction $d \in \mathbb{R}^2$ to color $c \in \mathbb{R}^3$ and volume density σ . Different from NeRF, NeuS replaces the density field with a signed distance field. We can extract the geometry surface S of the scene by the zero-set of the SDF values $S = \{x \in \mathbb{R}^3 | f_{sdf}(x) = 0\}$, where f_{sdf} is the signed distance function. Based on the signed distance function, we can further calculate the opaque density ρ and opacity values α . Finally, the pixel color \hat{C} of a ray t can be computed by the classical volume rendering function:

$$\hat{C}(t) = \sum_{i=1}^n T(t_i)\alpha(t_i)(t_i) \quad (1)$$

where n is the number of sample points along one ray and T represents the discrete accumulated transmittances, which is defined as $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$.

3.2. Overall Pipeline

As shown in Fig. 2, our method consists of two stages. In the first stage, we train the neural field of the scene on multi-view images. The users can then indicate the prompts of a target object in a single view. We use SAM to obtain the segmentation of this view and initialize the 3D occupancy field. The 3D occupancy field can generate coarse masks for

other views and further aggregate to the prompts of SAM. Precise masks of other views can be generated by SAM to refine the initialized 3D occupancy field. This process is iterative until the 3D occupancy field converges.

In the second stage, after segmenting the target object out of the scene, we can obtain the precise 2D masks of the target object in all views. We then leverage the features of SAM to improve the reconstruction quality of the target object, by distilling the 2D features of the SAM encoder into the 3D feature field. We will present the two stages in the following sections.

3.3. Stage-1: Segmentation by 3D Occupancy Field

In this section, we introduce a 3D occupancy field to lift 2D segmentation masks from different views into 3D space as shown in Fig. 3. The 3D occupancy field can be used to identify foreground and background voxels and generate a unified 3D segmentation mask of the target object. By constructing the 3D occupancy field, we can also obtain multi-view consistent 2D masks within a short time.

Given a coarse 2D mask rendered by the 3D occupancy field, we use SAM to refine it. Although SAM supports a variety of prompts as input, including points, boxes, masks, and text, we use points and boxes as prompts, which works better and saves computational memory. We compute the K-means clustering centers and minimum bounding rectangles of a 2D coarse mask as prompts. Then SAM encodes image I and prompts $P = (Point, Box)$ as features and decodes features into more precise masks M_{SAM} . The process can be formulated as:

$$M_{SAM} = Dec_M(Enc_I(I), Enc_P(P)) \quad (2)$$

where Enc_I , Enc_P , and Dec_M are the image encoder, prompts encoder, and mask decoder in SAM, respectively. M_{SAM} will be used as the label to supervise the learning of the 3D occupancy field.

Given a ray $r(t) = o + t \cdot d$ with camera position o and view direction d , the corresponding 2D segmentation is defined as $M(r)$. In the box of the scene, the majority of voxels belong to the background, with only a small portion belonging to the foreground. Therefore, the relations between the pixel-level segmentation M and the 3D occupancy field M_o are based on the following two assumptions: (1) If a pixel belongs to the foreground, then at least one of the positions passed through the ray is foreground. (2) If a pixel belongs to the background, then all the positions passed through the ray are background. With the above assumptions, we can simply apply maximization operation to formulate the relations between n points along the ray r :

$$M(r) = f_s(\max(\{M_o(r(t_i)) \cdot \omega_{r(t_i)} | i \in \{1, \dots, n\}\})) \quad (3)$$

where f_s indicates the sigmoid function and $\omega_{r(t_i)}$ represents point-wise weights, which are in stop-gradients from

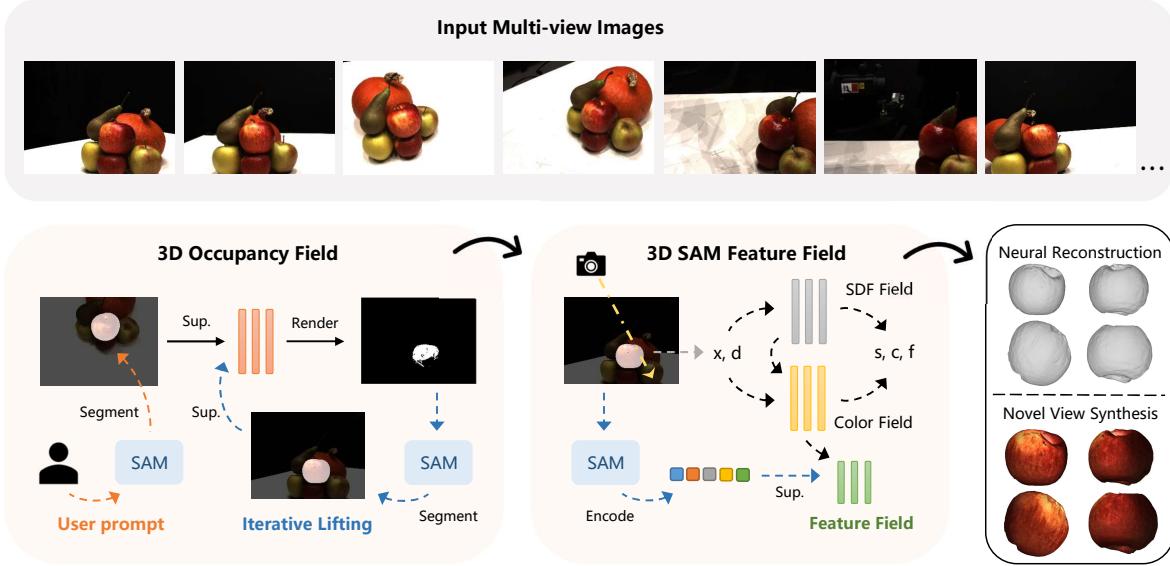


Figure 2. The overall pipeline of NTO3D. First, the user specifies the target object to be reconstructed and sends prompts to SAM for segmentation on the initial view. With multi-view images as input, we train the 3D occupancy field iteratively to lift cross-view masks into 3D space. When the 3D occupancy field converges to high-quality masks of the target objects, we finetune the pre-trained neural field based on the masked images and distill SAM encoder features into 3D space to obtain better reconstruction quality.

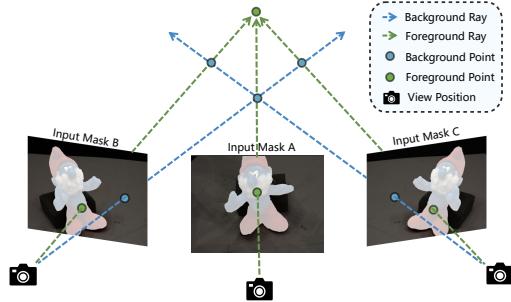


Figure 3. The illustration of the 3D occupancy field. Implicit interaction between multiple rays to decide which point is foreground or background. For a background ray, all points on it belong to the background. For a foreground ray, at least one point on it is foreground.

the SDF field. Then we can train the 3D occupancy field by the binary cross-entropy loss:

$$L_o = L_{BCE}(M_{SAM}, M) \quad (4)$$

By minimizing the above loss function, we can transfer the foreground segmentation of SAM to a unified 3D occupancy field. The above process is iterative until the 3D occupancy field converges. As shown in Fig. 4, at the beginning of the iteration, the rendered 2D coarse masks may exhibit defects since the 3D occupancy field does not converge. However, the refined 2D precise masks are mostly correct due to the proper prompting of SAM. After several

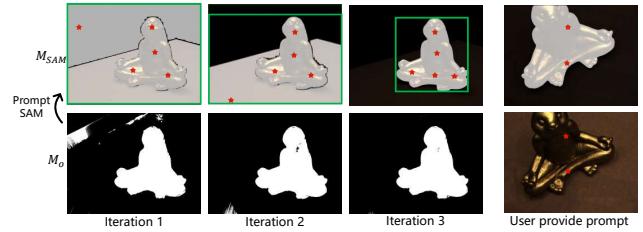


Figure 4. Mask iteratively lifting illumination, in which M_o represents masks generated by 3D occupancy field and M_{SAM} indicates masks provided by SAM base on prompts. Given users' prompts of specific objects, the 3D occupancy field renders a coarse mask in another view, which leads to bad prompts for SAM and defective masks. But the 3D occupancy field lifts 2D masks from all views into 3D space and efficiently corrects its false judgments of voxels in other views. With the iterative training, M_o and M_{SAM} begin to shrink and finally converge to the same.

iterations, the 3D occupancy field can correct erroneous predictions, ensuring multi-view consistent 2D masks.

3.4. Stage-2: Refinement by 3D Feature Field

As a foundation segmentation model, SAM possesses the ability to surpass most previous segmentation models and contains abundant knowledge. To better leverage the features of SAM, we add a lightweight output branch to the neural field, lifting the features of the SAM encoder into 3D space. With the encoder of SAM, we can obtain feature maps $Enc_I(I)$ with 256×256 resolutions corresponding

to the input images. Then this branch takes the 3D point position, geometry features, and color features as input and outputs the features $f(t_i)$ of each 3D point. Similar to color, volume rendering can be used to render the 3D feature field into the 2D image as:

$$\hat{F}(t) = \sum_{i=1}^n T(t_i)\alpha(t_i)f(t_i) \quad (5)$$

To optimize the 3D SAM feature field f , we adopt the L1 loss between the rendered features $\hat{F}(t)$ and SAM encoder features $Enc_I(I)$:

$$L_f = \frac{1}{R} \sum_r \left\| \hat{F}(r) - Enc_I(I) \right\|_1 \quad (6)$$

As for the color branch and SDF branch, we follow the previous work and adopt photometric loss and Eikonal loss [11] to supervise their training respectively, which can be defined as:

$$\begin{aligned} L_c &= \frac{1}{R} \sum_r \left\| \hat{C}(r) - C(r) \right\|_2^2 \\ L_{eik} &= \frac{1}{R \cdot N_s} \sum_{r,i} (|n| - 1)^2 \end{aligned} \quad (7)$$

Finally, we minimize the weighted sum of the above losses:

$$L_{total} = L_c + \lambda_{eik} L_{eik} + \lambda_f L_f + \lambda_v L_v \quad (8)$$

4. Experiments

We conduct extensive experiments to evaluate the effectiveness of the proposed method for neural 3D target object reconstruction. We first describe the experimental settings and then compare our method with the SOTA approaches on the DTU dataset [13]. We also provide the qualitative analysis on the LLFF dataset [26]. We further conduct a comprehensive ablation study to evaluate the contribution of each component. Due to the space limitation, we provide more reconstruction results on the BlendedMVS dataset [56] in the supplementary material.

4.1. Experimental Settings

Datasets. For **Quantitative comparison**, following the previous works [47], we evaluate our proposed method on the selected 15 scenes from the DTU dataset. Specifically, there are 64 or 49 images with 1600×1200 image resolution in each scene. Since the DTU dataset provides ground truth foreground masks, we first compare the segmentation generated by NTO3D and baselines. Then we evaluate the rendering and reconstruction quality from two aspects: training with mask supervision (w/ mask) and without mask supervision (w/o mask). For **Qualitative comparison**, following

the previous work [27], we show visualizations on 9 challenging scenes from the LLFF dataset. There are 20 to 62 images with a fixed image resolution of 1008×756 in each scene, and we randomly select 1/8 of the entire images to construct the test set. Since LLFF does not contain ground truth foreground masks, we first manually annotate a target object in the scene and compare the visual quality.

Implementation Details. We follow the implementation details specified in NeuS [47] and Instant-NSR [62]. we adopt the network architecture of Instant-NSR, which consists of two MLPs and the multi-resolution hash table to encode SDF and color, respectively. We utilize the Adam optimizer [18] with $(\beta_1, \beta_2) = (0.9, 0.999)$ to update our neural networks, and the learning rates warm up from 0 to 1×10^{-3} in the first 5k iterations and then controlled by the linear decay scheme to the latest learning rate of 1×10^{-5} . We set the number of rays to 4096 and sample 80 points for each ray. We first train the neural field of the scene and then learn the 3D occupancy field with 1k interval iterations. Finally, we finetune NTO3D for 20k iterations along with the 3D feature field. All experiments are conducted on NVIDIA 3090 GPUs. More implementation details are reported in the supplementary material.

4.2. Quantitative Comparison

In this section, we first report the comparison of generated segmentation masks. We compare against the segmentation masks generated by SPIn-NeRF [29], ISRF [10] and SA3D [5]. We also compare a simple baseline named Prompts Projection: We directly project prompts from a view to other views and use SAM to generate the 2D masks based on the projected prompts. Then we report the comparisons of neural rendering and reconstruction. We compare our method with IDR [57], NeRF [27], COLMAP [39], UNISURF [33], and NeuS [47]. We also compare a simple baseline named Post-processing Segmentation: We first reconstruct the whole scene and then use ground truth 2D masks provided by DTU to segment the desired object. It should be noted that NTO3D is designed for neural reconstruction similar to NeuS based on the SDF field, while ISRF, SPIn-NeRF and SA3D [5] are all designed for neural rendering based on the density field. To make a fair comparison, we first project the 2D foreground masks of ISRF, SPIn-NeRF and SA3D to 3D space and segment the 3D target object based on the reconstruction of NeuS.

Segmentation Comparison. As shown in Tab. 1, after iterative training, NTO3D achieves high segmentation mask quality. On one hand, the precise mask generated by SAM guides the 3D occupancy field in distinguishing foreground and background voxels effectively. On the other hand, the 3D occupancy field aggregates the multi-view 2D segmentation masks, enabling the generation of cross-view prompts even with only a single view annotated by users.

Table 1. Quantitative comparisons on masks generated by our iterative training. M_o represents masks generated by 3D occupancy field. M_{SAM} represents masks generated by SAM with prompts from 3D occupancy field. For baselines, M_{Proj} , M_{ISRF} , M_{DINO} and M_{SA3D} indicate masks from simply prompts projection, ISRF, DINO used by SPIn-NeRF and SA3D. Mean represents the average IoU.

Scan ID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
M_{Proj}	0.541	0.598	0.512	0.495	0.555	0.582	0.509	0.317	0.407	0.687	0.631	0.547	0.521	0.498	0.432	0.522
M_{ISRF}	0.846	0.765	0.715	0.558	0.717	0.787	0.687	0.692	0.231	0.578	0.761	0.687	0.641	0.690	0.549	0.660
M_{DINO}	0.941	0.961	0.952	0.972	0.945	0.959	0.943	0.897	0.934	0.932	0.937	0.976	0.937	0.957	0.925	0.945
M_{SA3D}	0.852	0.946	0.934	0.954	0.921	0.897	0.936	0.841	0.947	0.942	0.940	0.909	0.919	0.961	0.946	0.923
M_o	0.935	0.951	0.963	0.975	0.982	0.983	0.963	0.974	0.979	0.962	0.955	0.943	0.925	0.937	0.967	0.960
M_{SAM}	0.962	0.986	0.991	0.991	0.991	0.986	0.971	0.984	0.987	0.992	0.965	0.990	0.990	0.968	0.984	0.983

Table 2. Quantitative comparisons with other methods on the task of novel view synthesis. Mean represents the average value of PSNR and SSIM.

Scan ID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
<i>Train w/o mask setting</i>																
PSNR(NeuS)	23.46	26.01	28.34	22.41	14.99	16.66	24.09	15.37	19.87	15.10	26.74	24.73	28.91	35.92	35.20	23.85
PSNR(NetRF)	21.13	23.00	18.91	22.97	25.55	12.7	22.76	25.19	28.54	35.29	18.31	17.91	21.44	21.48	22.68	22.52
SSIM(NeuS)	0.876	0.611	0.911	0.903	0.897	0.744	0.844	0.840	0.856	0.826	0.838	0.871	0.914	0.968	0.972	0.858
SSIM(NetRF)	0.791	0.860	0.817	0.853	0.921	0.798	0.829	0.938	0.932	0.960	0.757	0.742	0.826	0.781	0.821	0.842
<i>Train w/ mask setting</i>																
PSNR(NeuS)	28.38	23.75	30.47	29.52	29.83	32.29	29.01	32.28	28.39	29.43	32.26	36.15	30.58	36.22	33.60	30.81
PSNR(Ours)	28.71	26.96	30.92	31.87	34.11	32.92	31.61	36.98	28.07	33.10	35.98	36.85	32.64	37.69	37.55	33.06
SSIM(NeuS)	0.887	0.891	0.917	0.958	0.951	0.969	0.935	0.969	0.920	0.924	0.954	0.970	0.942	0.969	0.968	0.942
SSIM(Ours)	0.897	0.913	0.925	0.968	0.958	0.956	0.948	0.969	0.920	0.956	0.979	0.973	0.967	0.969	0.980	0.952

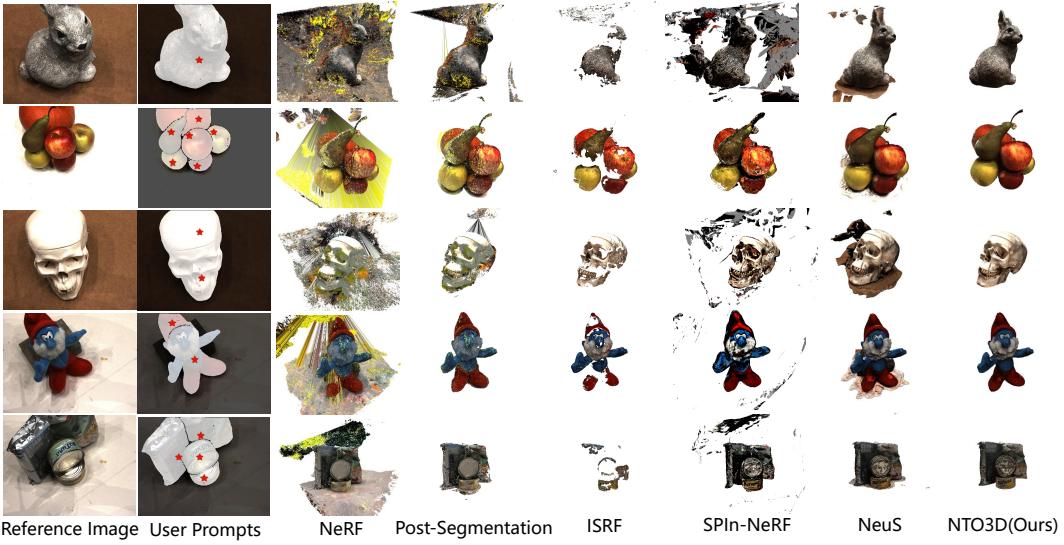


Figure 5. Qualitative comparison on DTU. Best viewed in colors.

Rendering Comparison. We compare NTO3D with the previous SOTA volume rendering approaches. We held out 10% of the images in the DTU dataset as the testing set and the others as the training set. During training, we split the baselines into two settings: train w/o ground truth masks and w/ ground truth masks. For NTO3D, we use masks generated by SAM after iteration lifting for training. We compare the rendering quality on the testing set with masks

regarding PSNR and SSIM. During the test, we calculate the metrics between prediction and masked ground truth. As shown in Tab. 2, baselines trained without object masks render lower quality than those trained with object masks. Our method shows significant improvement in PSNR and SSIM, demonstrating that our method can aggregate the multi-view 2D segmentation masks to improve the novel view synthesis quality of a target object.

Table 3. Chamfer distances comparisons with other methods on the DTU dataset. COLMAP results are achieved by trim=0.

Scan ID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
<i>Train w/o mask setting</i>																
Post-Segmentation	0.98	1.20	1.58	2.07	2.09	2.94	1.02	2.56	2.64	1.14	1.42	1.21	1.01	1.36	1.24	1.63
COLMAP	0.81	2.05	0.73	1.22	1.79	1.58	1.02	3.05	1.40	2.05	1.00	1.32	0.49	0.78	1.17	1.36
NeRF	1.90	1.60	1.85	0.58	2.28	1.27	1.47	1.67	2.05	1.07	0.88	2.53	1.06	1.15	0.96	1.49
UNISURF	1.32	1.36	1.72	0.44	1.35	0.79	0.80	1.49	1.37	0.89	0.59	1.47	0.46	0.59	0.62	1.02
NeuS	1.00	1.37	0.93	0.43	1.10	0.65	0.57	1.48	1.09	0.83	0.52	1.20	0.35	0.49	0.54	0.84
<i>Train w/ mask setting</i>																
Prompts Projection	1.21	1.14	1.47	2.37	1.60	2.01	1.27	1.03	3.28	1.54	1.47	1.20	1.03	1.54	1.21	1.56
ISRF	1.02	1.34	0.86	1.45	1.36	1.00	1.12	1.58	2.07	1.20	0.98	1.36	1.21	0.98	0.87	1.23
SPIn-NeRF	0.97	1.12	0.76	0.96	1.77	1.02	1.25	1.50	2.05	0.98	0.67	1.25	0.96	0.59	0.51	1.02
SA3D	0.95	1.09	0.82	1.03	1.63	1.23	1.14	1.41	2.13	0.95	0.73	1.29	0.99	0.72	0.56	1.11
IDR	1.63	1.87	0.63	0.48	1.04	0.79	0.77	1.33	1.16	0.76	0.67	0.90	0.42	0.51	0.53	0.90
NeuS	0.83	0.98	0.56	0.37	1.13	0.59	0.60	1.45	0.95	0.78	0.52	1.43	0.36	0.45	0.45	0.77
Ours	0.82	1.14	0.60	0.35	1.01	0.53	0.63	1.31	0.86	0.73	0.51	1.15	0.45	0.42	0.46	0.73

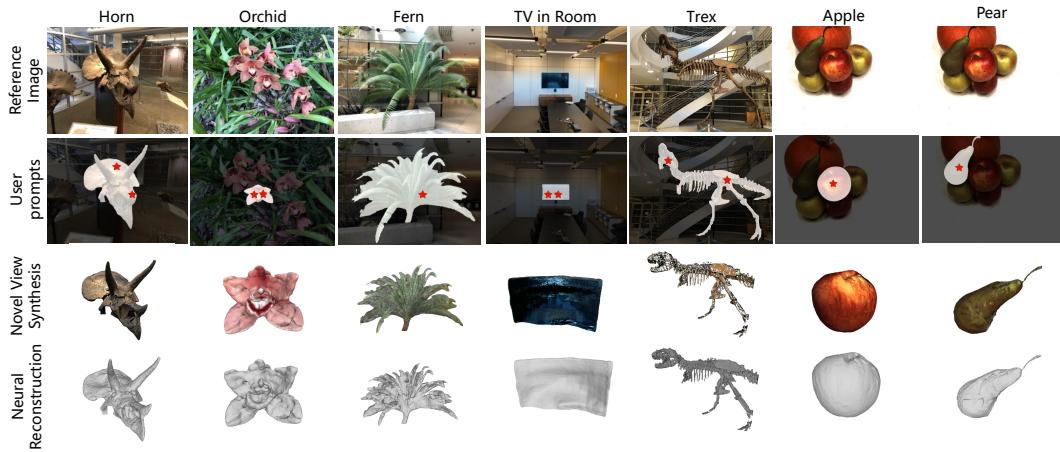


Figure 6. Qualitative comparison on LLFF and fruit scene in DTU. These scenes have the following characteristics: the foreground objects consist of multiple independent objects, and the background is more complex. Best view in colors.

Reconstruction Comparison. We also measure the reconstruction quality with the Chamfer distances and compare NTO3D with other methods, as shown in Tab. 3. Similar to the rendering comparison, training without masks introduces more background noises into the neural field. Simply projecting point prompts to other views leads to bad performance. This is because prompts projection heavily relies on the quality of depth, while the reconstructed depth of NeuS is not accurate enough. Therefore, SAM cannot correctly segment the target object. With the help of the 3D occupancy field and the 3D feature field, our approach reduces the Chamfer distance to 0.73 and outperforms the baseline methods. The results demonstrate that neural reconstruction can benefit from the NTO3D pipeline.

4.3. Qualitative Analysis

We conduct the qualitative comparisons on DTU and LLFF datasets. As shown in Fig. 5 and Fig. 6, we provide reference images and prompts indicated by users on the initial

view. Due to the space limitation, please refer to supplementary materials for qualitative results on BlendedMVS, which contains more complex and larger scenes.

On DTU datasets, we compare the visualization quality of NTO3D with other methods. NeRF shows the worst visualization quality since it reconstructs the whole scene. Post-Segmentation also leads to bad performance since reprojecting 2D segmentation masks provided by DTU back into 3D does not completely remove the background. ISRF offers a tool for interactive segmentation, but masks extracted by ISRF are sensitive to a bunch of parameters. Inappropriate parameters result in lower-quality masks for ISRF. SPIn-NeRF uses DINO for segmentation, which treats multi-view inputs as a video sequence. However, the segmentation accuracy of SPIn-NeRF is worse than NTO3D. SA3D uses an inverse rendering for mask generation but still exhibits lower segmentation accuracy compared to NTO3D, which results in worse chamfer distance. Although NeuS contains a background model, which helps it focus on the

foreground objects, its reconstruction results still inevitably show the background near the target objects. Thanks to the proposed 3D occupancy field, NTO3D can generate high-quality masks of foreground objects without tedious annotation on all views. Additionally, we can witness that the NTO3D reconstructs higher surface quality with the help of 3D SAM feature fields.

On LLFF datasets, we choose one object for each scene as the target object. We can see that whether the selected object is significant or not in the scene, NTO3D can segment the target object based on the user’s prompts and obtain impressive reconstruction quality. Besides, we also provide the results of one object among several foreground objects in the last two columns. This further demonstrates that with the help of NTO3D, we are able to reconstruct any target objects of the scene.

Table 4. **Ablation Study of NTO3D.** CD indicates the Chamfer distance.

Variant	PSNR↑	SSIM↑	CD↓
Instant-NSR	29.94	0.8914	0.82
+ 3D Occupancy Field	32.44	0.9319	0.76
+ 3D SAM Feature Field	30.76	0.9136	0.78
NTO3D (our)	33.06	0.9520	0.73

4.4. Ablation study

Evaluation of Each Component. We study the effectiveness of the proposed 3D occupancy field and 3D feature field. The experiments are done on the DTU dataset and average the results of all scenes. As shown in Tab. 4, with the aid of the 3D occupancy field, our method avoids the influence of the background and focuses on the reconstruction of target objects. Since the SAM encoder contains abundant knowledge, the 3D feature field helps to boost the reconstruction quality. With the proposed two contributions, NTO3D can efficiently segment and reconstruct the target object indicated by users in the scene.

Overlap Ratio Between Views. We further explore the effects of the max overlap ratio between views to study the choice of the initial view. We conduct the experiments on a random scene of the DTU dataset and calculate the Max View Distance and the Max Overlap Ratio between the first manually annotated view and other views. As shown in Tab. 5, the final reconstruction accuracy is approximately the same for different initial views with different overlap ratios. Since the proposed 3D occupancy field iteratively obtains foreground-background segmentation in 3D space, NTO3D can avoid confusion from different views. Once the initial view appropriately segments the target object, NTO3D can always successfully reconstruct target objects.

Different Stages of Distilling SAM Features. We conduct the experiments of distilling SAM features at different stages: 1). pretraining before the first step of our method

Table 5. Influence of overlap ratio on reconstruction quality.

Initial View	1	5%	10%	20%	30%
Max View Distance	103.63°	122.87°	108.5°	110.09°	101.49°
Max Overlap Ratio	0.471	0.149	0.500	0.250	0.322
Chamfer Distance↓	0.731	0.742	0.735	0.740	0.738

(denoted by pretrain). 2). iterative training for 3D occupancy field (denoted by iterative). 3). finetuning on the target object (our default setting, denoted by finetune). We conduct the experiments on a random scene of the DTU dataset and report the mIoU and chamfer distance. As shown in Tab. 6, distilling SAM features at pretraining and iterative training performs slightly worse than distilling at finetuning. Without the foreground mask, SAM features contain the background information, which might slightly degrade the performance of target object reconstruction. In addition, encoding images with SAM requires additional computational cost, therefore it is unnecessary to distill SAM features at early stages.

Table 6. Different stages of distilling SAM features.

Experiment Setting	pretrain	iterative	finetune
mIoU	0.981	0.980	0.991
Chamfer Distance	0.360	0.358	0.353

5. Limitation

Although SAM has a powerful segmentation ability, when facing challenging scenes, it is also difficult to segment the target object [14]. If SAM fails to segment target objects with available prompts, our method will fail to learn the 3D occupancy field. A possible solution is to fine-tune SAM on challenging scenes with parameter-efficient-finetuning (PEFT) techniques. Despite the limitations, NTO3D has demonstrated the potential of combining large foundation models and neural fields.

6. Conclusion

To reconstruct a certain object indicated by users on-the-fly and boost the reconstruction quality, the paper applies the Segment Anything Model to help 3D object reconstruction. The proposed method Neural Target Object 3D Reconstruction (NTO3D) first leverages a 3D occupancy field to lift the multi-view 2D segmentation masks generated by SAM. With the help of 3D occupancy field, NTO3D is able to segment target objects and eliminate background interference. To boost the reconstruction quality, we further propose a 3D SAM Feather Field to lift pixel-level features into voxel space. Finally, we conduct several experiments on several datasets and demonstrate NTO3D can obtain better reconstruction quality. Please refer to supplementary materials for more results.

Acknowledgement. Shanghang Zhang is supported by the National Science and Technology Major Project of China (No. 2022ZD0117801).

References

- [1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 2
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1
- [3] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Craig Gotsman, and Robert W Sumner. High-quality passive facial performance capture using anchor frames. 2
- [4] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9, 2010. 2
- [5] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023. 2, 5
- [6] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2019. 2
- [7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2
- [8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1
- [9] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2
- [10] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and P.J. Narayanan. Interactive Segmentation of Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5
- [11] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 2, 5
- [12] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 3
- [13] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 5
- [14] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023. 8
- [15] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [16] Srinivas Kaza et al. *Differentiable volume rendering using signed distance functions*. PhD thesis, Massachusetts Institute of Technology, 2019. 2
- [17] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. *arXiv preprint arXiv:2103.11571*, 2021. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2
- [21] Lingjie Liu, Jitao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33, 2020. 2
- [22] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2
- [23] Stephen Lombardi, Tomas Simon, Jason Saragh, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):65, 2019. 2
- [24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2
- [25] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [26] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 5
- [27] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 5

- [28] 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*, 65(1):99–106, 2021. 1, 2, 3
- [29] Ashkan Mirzaei, Tristan Amentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023. 5
- [30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1
- [31] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2
- [32] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 3
- [33] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *arXiv preprint arXiv:2104.10078*, 2021. 5
- [34] 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*, pages 165–174, 2019. 2
- [35] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *ArXiv*, abs/2003.04618, 2020. 2
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015. 2
- [37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, 2019. 2
- [38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2
- [39] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 5
- [40] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 519–528. IEEE, 2006. 2
- [41] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [42] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1119–1130, 2019. 2
- [43] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 1
- [44] Jiaxiang Tang, Hang Zhou, Xiaokang Chen, Tianshu Hu, Er-rui Ding, Jingdong Wang, and Gang Zeng. Delicate textured mesh recovery from nerf via adaptive surface refinement. *arXiv preprint arXiv:2303.02091*, 2023. 1, 2
- [45] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, pages 282–298. Springer, 2020. 2
- [46] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. *arXiv preprint arXiv:2010.04595*, 2020. 2
- [47] 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, 2, 3, 5
- [48] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Segpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2
- [49] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. Video-based hand manipulation capture through composite motion control. *ACM Transactions on Graphics (TOG)*, 32(4):1–14, 2013. 2
- [50] Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun Bao, and Xiaowei Zhou. Autorecon: Automated 3d object discovery and reconstruction. In *CVPR*, 2023. 3
- [51] Thibaut Weise, Thomas Wismer, Bastian Leibe, and Luc Van Gool. In-hand scanning with online loop closure. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1630–1637. IEEE, 2009. 1
- [52] Tong Wu, Jiaqi Wang, Xingang Pan, Xudong Xu, Christian Theobalt, Ziwei Liu, and Dahua Lin. Voxurf: Voxel-based efficient and accurate neural surface reconstruction. *arXiv preprint arXiv:2208.12697*, 2022. 1, 2
- [53] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

- [54] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2
- [55] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2
- [56] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 5
- [57] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 5
- [58] Wang Yifan, Shihao Wu, Cengiz Oztireli, and Olga Sorkine-Hornung. Iso-points: Optimizing neural implicit surfaces with hybrid representations. *arXiv preprint arXiv:2012.06434*, 2020. 2
- [59] Alex Yu, Rui long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1
- [60] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2
- [61] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 2
- [62] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *arXiv preprint arXiv:2209.08468*, 2022. 5
- [63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2
- [64] Hao Zhu, Yebin Liu, Jingtao Fan, Qionghai Dai, and Xun Cao. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):760–770, 2016. 2