AutoRecon: Automated 3D Object Discovery and Reconstruction

Yuang Wang Xingyi He Sida Peng Haotong Lin Hujun Bao Xiaowei Zhou†
State Key Lab of CAD&CG, Zhejiang University

Abstract

A fully automated object reconstruction pipeline is crucial for digital content creation. While the area of 3D reconstruction has witnessed profound developments, the removal of background to obtain a clean object model still relies on different forms of manual labor, such as bounding box labeling, mask annotations, and mesh manipulations. In this paper, we propose a novel framework named AutoRecon for the automated discovery and reconstruction of an object from multi-view images. We demonstrate that foreground objects can be robustly located and segmented from SfM point clouds by leveraging self-supervised 2D vision transformer features. Then, we reconstruct decomposed neural scene representations with dense supervision provided by the decomposed point clouds, resulting in accurate object reconstruction and segmentation. Experiments on the DTU, BlendedMVS and CO3D-V2 datasets demonstrate the effectiveness and robustness of AutoRecon. The code and supplementary material are available on the project page: https://zju3dv.github.io/autorecon/.

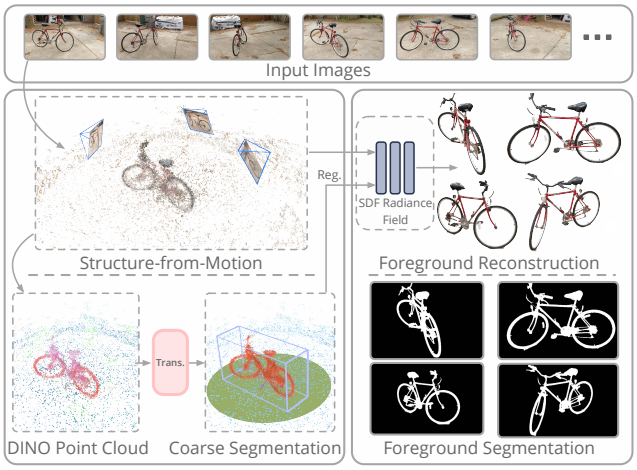

Figure 1. **Overview of our fully-automated pipeline and results.** Given an object-centric video, we achieve coarse decomposition by segmenting the salient foreground object from a semi-dense SfM point cloud, with pointwise-aggregated 2D DINO features [3]. Then we train a decomposed neural scene representation from multi-view images with the help of coarse decomposition results to reconstruct foreground objects and render multi-view consistent high-quality foreground masks.

1. Introduction

3D object reconstruction has long been investigated in computer vision. In this work, we focus on the specific setting of reconstructing a salient foreground object from multi-view images and automatically segmenting the object from the background without any annotation, which enables scalable 3D content creation for VR/AR and may open up the possibility to generate free 2D and 3D object annotations at a large scale for supervised-learning tasks.

Traditional multi-view stereo [8, 32] and recent neural scene reconstruction methods [40, 46] have attained impressive reconstruction quality. However, these methods cannot identify objects and the reconstructed object models are typically coupled with the surrounding background. A straightforward solution is utilizing the foreground object masks to obtain clean foreground object models. However, accurate 2D object masks are expensive to annotate, and salient object segmentation techniques [21, 34, 41] generally produce masks with limited granularity, thus degrading the reconstruction quality, especially for objects with thin structures. Recently, some methods [23, 30, 50] attempt to automatically decompose objects from 3D scenes given minimal human annotations, such as 3D object bounding boxes, scribbles or pixel labels. But the requirement of manual annotations limits the feasibility of more scalable 3D content creation.

In this paper, we propose a novel two-stage framework for the fully-automated 3D reconstruction of salient objects, as illustrated in Fig. 1. We first perform coarse decomposition to automatically segment the foreground SfM point cloud, and then reconstruct the foreground object geometry by learning an implicit neural scene representation under explicit supervision from the coarse decomposition. The key idea of our coarse decomposition is to leverage the semantic features

The authors are affiliated with the ZJU-SenseTime Joint Lab of 3D Vision.
†Corresponding author: Xiaowei Zhou.

provided by a self-supervised 2D Vision Transformer (ViT) [3]. Specifically, we aggregate multi-view ViT features from input images to the SfM point cloud and then segment salient foreground points with a point cloud segmentation Transformer. To train the Transformer on large-scale unlabeled data, we devise a pseudo-ground-truth generation pipeline based on Normalized Cut [33] and show its ability to produce accurate segmentations and 3D bounding boxes upon training. For object reconstruction, we learn a neural scene representation within the estimated foreground bounding box from multi-view images. Our main idea is to reconstruct a decomposed scene representation with the help of explicit regularization provided by the previously decomposed point cloud. Finally, we can extract a clean object model and obtain high-quality object masks with foreground-only rendering.

We conduct experiments on the CO3D [29], BlendedMVS [45], and DTU [12] datasets to validate the effectiveness of the proposed pipeline. The experimental results show that our approach can automatically and robustly recover accurate 3D object models and high-quality segmentation masks from RGB videos, even with cluttered backgrounds.

In summary, we make the following contributions:

- We propose a fully-automated framework for reconstructing background-free object models from multi-view images without any annotation.
- We propose a coarse-to-fine pipeline for scene decomposition by first decomposing the scene in the form of an SfM point cloud, which then guides the decomposition of a neural scene representation.
- We propose an SfM point cloud segmentation Transformer and devise an unsupervised pseudo-ground-truth generation pipeline for its training.
- We demonstrate the possibility of automatically creating object datasets with 3D models, 3D bounding boxes, and 2D segmentation masks.

2. Related Work

Multi-view 3D object reconstruction. The reconstruction of 3D objects from multi-view images has long been studied with broad applications. Aside from multi-view images, accurate object masks are needed to separate the object of interest from its surroundings and optionally provide additional geometric constraints. Multi-view Stereo (MVS) methods [32, 44] recover a background-free reconstruction by recovering frame-wise depth maps, followed by fusing depth only within object masks. Recently, neural reconstruction methods, built upon differentiable neural renderers and scene representations, have witnessed profound development. Surface-rendering-based methods [26, 47] get rid of 3D supervision, but they still rely on object masks as substitutive geometry constraints. The recent volume-rendering-based reconstruction methods [27, 40, 46] allow mask-free training but still require object masks supervision to produce background-free object models. Aside from object masks, existing methods also require manual annotation of the 3D spatial extent of the foreground object. Instead, we propose a fully-automated object reconstruction pipeline without any human labeling, which further improves the usability and scalability of 3D object reconstruction.

Decomposition of neural scene representations. Many recent works try to decompose neural scene representations (NSR). We categorize related works based on the annotations required. Explicit 3D geometric primitives provide simple but effective decompositions of different entities. NeRF++ [50] separates foreground and background with a sphere. 3D bounding boxes manually annotated or predicted by category-specific models are used for decomposed modeling of static and dynamic scenes [16, 24, 28]. Multi-view segmentation masks provide dense annotations for scene decomposition. It has been shown that semantic fields can be learned with multi-view semantic masks [15, 39, 51] for semantic scene decomposition. Moreover, decomposed object representations can also be built from multi-view object masks [42, 43]. To alleviate the annotation cost of multi-view segmentation, methods relying on human interactions [52] and different forms of sparse human annotations are proposed, such as scribbles [30] and seed points [23]. The decomposition is less stable as they rely on handcrafted non-dedicated features from various sources to distinguish the manually specified entities. Apart from learning discrete semantic labels, DFF [14] and N3F [38] distill 2D features into neural scene representations for query-based scene decomposition. However, they still require manually-provided queries and their query-based nature is more suitable for local editing and impedes applications requiring global reasoning upon a scene, such as the decomposition of a salient object. Different from existing approaches, our pipeline requires no annotation and facilitates global reasoning.

Unsupervised object discovery. Unsupervised object discovery (UOD) aims at the unsupervised learning of object concepts from a large-scale dataset. Recently, many works strive for UOD with compositional generative modeling [2, 6, 9]. Slot Attention [18] facilitates the inference of object-centric representations directly from images. This idea is further extended to 3D-aware modeling and inference with neural radiance fields or light fields [31, 35, 36, 48]. However, these works have only been shown to work on synthetic datasets, not applicable to complex real-world situations. Our method focuses on the recovery of decomposed single-object representations and is shown to work on real-world data such as casual video captures. Another recent trend makes use of self-supervised visual representations for unsupervised object discovery in various forms, such

as object localization [34], salient detection, and semantic segmentation [49]. TokenCut [41] and DSM [21] show promising results by localizing and segmenting salient objects with spectral clustering. However, their 2D nature leads to multi-view inconsistency and unstable results when applied to object-centric videos. To overcome these limitations, we propose to perform unsupervised object discovery from videos in 3D, which facilitates coherent salient object discovery upon a global representation, instead of many isolated inferences upon local 2D observations.

3. Preliminaries

In this section, we briefly review the following preliminaries: the self-supervised ViT features used to segment point clouds, the Normalized Cut algorithm employed to generate pseudo segmentation labels, and the neural surface reconstruction method NeuS utilized for object reconstruction.

Self-supervised ViTs. A Vision Transformer [5] flattens an $H \times W$ sized image $\mathbf{I}$ into a sequence of $P \times P$ sized 2D patches $\mathbf{I}_p$. Each image patch is embedded with a trainable linear projection and added with a positional embedding. A special learnable [CLS] token is usually prepended to the sequence of patches for modeling global and contextual information. The 1D sequence of token embeddings is fed to several Transformer encoder blocks composed of multi-head self-attention (MSA) and MLP layers:

$$\begin{aligned} \mathbf{z}^{\ell'} &= \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}^{\ell-1}\right)\right) + \mathbf{z}^{\ell-1}, \\ \mathbf{z}^{\ell} &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z}^{\ell'}\right)\right) + \mathbf{z}^{\ell'}, \end{aligned} \tag{1}$$

where $\mathbf{z}^{\ell}$ is the output of the ℓ-th Transformer encoder layer.

It has been shown in [3] that self-supervised ViT features contain explicit semantic information such as scene layout and object boundaries, which is not found in the supervised counterparts.

Normalized cut algorithm (NCut) [33]. Spectral clustering is a wildly used clustering technique that originated from graph partitioning. Given a set of data points x_i, spectral clustering builds an undirected graph $G = (V, E)$ and partitions it into two disjoint sets A, B. Each data point x_i corresponds to a vertex v_i, and the weight $w(i, j)$ of each graph edge represents the similarity or the connectivity between two data points. Normalized Cut (NCut) is a wildly used criterion for spectral clustering, which can be efficiently minimized by solving a generalized eigenvalue problem as shown in [33].

Neural surface reconstruction with NeuS. NeuS [40] uses the zero-level set of a signed distance function (SDF) $f : \mathbb{R}^3 \to \mathbb{R}$ to represent a surface $\mathcal{S} = \left\{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\right\}$ and models appearance with a radiance field $\mathrm{c}(\mathbf{x}, \mathbf{v})$. The SDF-based radiance field is rendered via volume rendering. Given a ray $\{\mathbf{r}(t) = \mathbf{o} + t\mathbf{v} | \ t > 0\}$, where $\mathbf{o}$ denotes the camera center and $\mathbf{v}$ is the view direction, we can render its color $\hat{C}$ by

$$\hat{C} = \int_0^{\infty} \omega(t)\, \mathrm{c}(\mathbf{r}(t), \mathbf{v}) dt, \tag{2}$$

where $\omega(t)$ is an unbiased and occlusion-aware weight function as detailed in [40].

Notably, the spatial extent of the foreground object of interest needs to be manually annotated, which is scaled into a unit-sphere and represented with the SDF-based radiance field. The background region outside the sphere is represented with NeRF++ [50]. Since the object of interest can hardly be exclusively enclosed with a single sphere, the reconstructed object model includes background geometries, requiring manual post-processing for its removal.

4. Methods

An overview of our pipeline is illustrated in Fig. 1. Given an object-centric video, we aim to automatically decompose and reconstruct the salient foreground object whose high-quality 2D masks can be rendered from its reconstructed geometry. To achieve this goal, we propose a novel coarse-to-fine pipeline that decomposes a neural scene representation with the help of point cloud decomposition. Our coarse decomposition stage segments the foreground object from a scene-level SfM point cloud and estimates its compact 3D bounding box (Sec. 4.1). Then, a decomposed neural scene representation of the foreground object is recovered under explicit supervision of the coarse decomposition (Sec. 4.2).

4.1. Coarse decomposition of the salient object

To coarsely decompose the foreground object, we first reconstruct its SfM point cloud and fuse multi-view DINO [3] features on it. Then, the point cloud is segmented by our lightweight 3D segmentation Transformer, upon which a 3D bounding box of the salient foreground object is generated. Our coarse decomposition pipeline is shown in Fig. 2. Since we assume that no manual annotation is available, we devise an unsupervised point cloud segmentation pipeline to generate pseudo-ground-truth segmentations, as shown in Fig. 3. Upon training, the 3D segmentation Transformer outperforms our unsupervised pipeline and can be applied to point clouds at larger scales.

Neural point cloud reconstruction. We leverage the SfM point clouds for efficient coarse decomposition since SfM is usually performed prior to the dense reconstruction for camera pose recovery. Specifically, we use the recent semi-dense image matcher LoFTR [37] for SfM to reconstruct

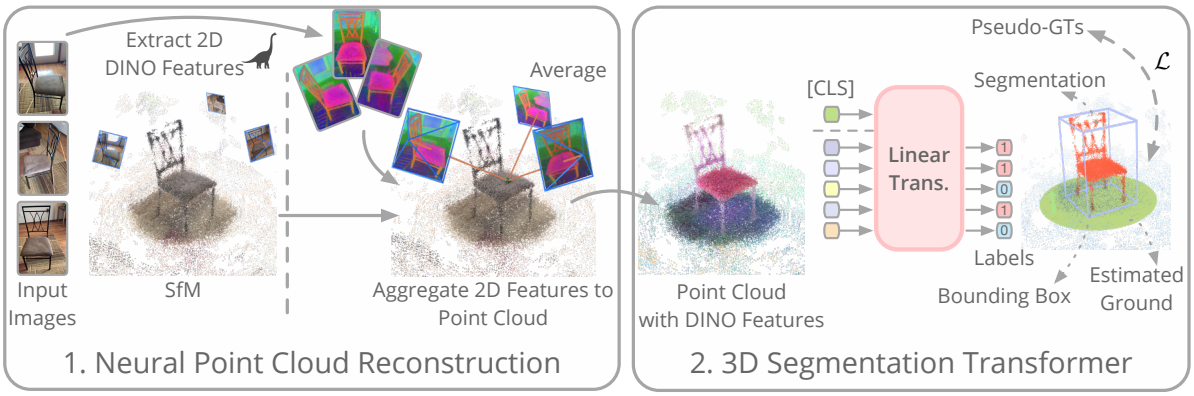

Figure 2. **Coarse Decomposition.** Given an object-centric image sequence, we first reconstruct the semi-dense Structure-from-Motion (SfM) point cloud and extract pointwise features by aggregating multi-view 2D DINO features, which are semantically rich as illustrated by the PCA-projected colors. Then, we segment the foreground object from the SfM point cloud with a lightweight 3D Transformer, which takes pointwise features (▢▢) and a global [CLS] feature (▢) as input and predicts pointwise labels (1 0). Finally, the 3D bounding box of the object and an optional ground plane are estimated from the decomposed point cloud.

semi-dense point clouds. It can recover the complete geometry of foreground objects, even the low-textured ones, which is discussed in [11]. This capability is appealing for robustly locating an object's complete spatial extent, which is less reliable with sparse keypoints-based SfM. To facilitate 3D segmentation, we lift self-supervised 2D ViT features to 3D. More specifically, frame-wise DINO-ViT features are extracted and aggregated onto the semi-dense point cloud, thanks to the explicit 3D-2D correspondences retained by SfM. We find that fusing multi-view features with a simple averaging operation is sufficient for our task. Additionally, frame-wise features of the [CLS] token globally describing each frame are also fused as a global description of our point cloud and further used as a segmentation prototype in our proposed 3D Transformer.

Point cloud segmentation with Transformer. As shown in Fig. 2, the neural point cloud already contains discriminative semantic features separating foreground and background. Therefore, we suppose that a concise network with proper inductive bias and trained with limited supervision is enough to probe the salient object in our task. We build an efficient point cloud Transformer with only two Transformer encoder layers and linear attentions [13]. The global [CLS] token and pointwise tokens obtained by the previously built neural point cloud are added with positional encodings and transformed by the encoder. Then, the transformed [CLS] token is treated as an input-dependent segmentation prototype, which is correlated with pointwise features to produce a segmentation mask. Our design takes full advantage of the global information of [CLS] token to reason about the salient object and its global-local relationship with other pointwise features for segmentation. The use of pre-trained 2D ViT features alleviates the reliance on large-scale data for training Transformers.

Dataset generation with unsupervised segmentation. In order to generate training data for our 3D Transformer, we propose an unsupervised SfM segmentation pipeline, which can produce robust segmentations but is computationally more intensive. We propose to apply NCut on the previously built neural point clouds as it facilitates scene-level global reasoning. A large-scale dataset with pseudo-ground-truth segmentations can be automatically generated with the proposed pipeline. An overview is presented in Fig. 3.

To apply the NCut algorithm on our neural point cloud for 3D segmentation, we build a fully connected graph $G = (V, E)$ using the neural point cloud, where each graph vertex V_i corresponds to a 3D point. We combine feature similarities and spatial affinities when modeling the edge weights $w(i, j)$ between V_i and V_j. Though the DINO feature is semantically rich, it is hierarchical, and the salient object inferred is sometimes ambiguous in terms of its position in the part-whole hierarchy. We propose a grouped cosine similarity to avoid the saliency dominated by a certain object part, especially for objects with complex structures.

Formally, denote a group of multi-head attention features $\mathbf{Z}_i = \{\mathbf{z}_i^0, ..., \mathbf{z}_i^{h-1}\}$ from h heads of an MSA module, we compute the grouped cosine similarity S^* between $\mathbf{Z}_i$ and $\mathbf{Z}_j$:

$$\mathrm{S}^*(\mathbf{Z}_i, \mathbf{Z}_j) = \max_{k \in \{0,\dots,h-1\}} \mathrm{S}(\mathbf{z}_i^k, \mathbf{z}_j^k), \tag{3}$$

where S is the cosine similarity. The intuition is, taking the maximum similarity between a group of multi-head features assigns two points of high similarity if they are similar in any aspect, thus reducing the chances that the saliency is only dominated by a local part of an object. The foreground

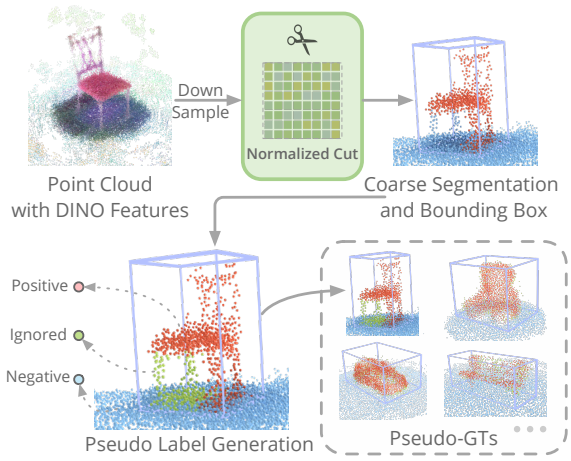

Figure 3. **Pseudo-ground-truth generation and label definition.** To train our point cloud segmentation Transformer with unlabeled data, we propose an unsupervised pipeline to generate pseudo labels. We segment the downsampled neural point cloud with Normalized Cut [33] (NCut) and estimate a bounding box for the foreground points. Taking the segmentation noise into account, we treat NCut's foreground segmentations as positive samples (), and background points outside the bounding box as negative ones (). Background segmentations located within the bounding box are regarded as segmentation noise () and thus ignored in training.

point cloud is then segmented with NCut on the graph defined above. An oriented 3D bounding box is subsequently inferred based on plane-aligned principle component analysis of the foreground point cloud. The pipeline is illustrated by Fig. 3. More details about our unsupervised segmentation pipeline are provided in the supplemental material.

4.2. Background-free salient object reconstruction

To reconstruct a background-free salient object model, we explicitly partition a scene with the coarse decomposition result and model each partition separately with neural scene representations [22, 40], which are trained upon multi-view posed images. Moreover, we incorporate multiple constraints into the optimization, which facilitates convergence and the decomposition of foreground objects from surroundings. Fig. 4 illustrates our foreground modeling.

Decomposed scene modeling. Previous methods [40, 47] scale the foreground object into a unit-sphere with manual annotation of its spatial extent and further rely on mask culling or manual manipulation of the reconstructed mesh to remove background geometries. Instead, we explicitly partition the scene into three parts with finer granularity without manual annotation, thanks to the estimated object bounding box in Sec. 4.1. More specifically, we use an SDF-based radiance field [40] to represent regions within the object bounding box and use a NeRF for regions outside. We use an additional tiny NeRF to model regions around the ground plane supporting the object, which can be located with the bottom plane of the object bounding box. Though there are overlaps between the inner region and the ground plane region, NeRF normally converges faster than the SDF-based radiance field and thus has an inductive bias to occupy the overlapped region. We use a foreground-object-aware version of the contraction function with L_∞ norm in MipNeRF-360 [1] to model unbounded scenes. More details are provided in the supplemental material.

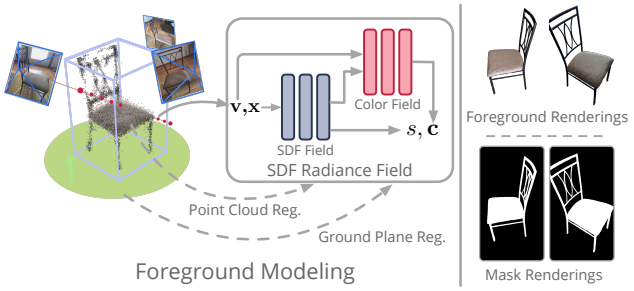

Figure 4. **Salient object reconstruction and 2D mask rendering.** We model the salient foreground object enclosed in the coarse bounding box with an SDF-based radiance field [40]. We use a decomposed scene representation consisting of separate fields for regions inside the bounding box, outside the bounding box, and near the ground plane. We regularize the optimization of the SDF-based radiance field with coarse decomposition results, i.e., the segmented foreground SfM point cloud and the estimated ground plane, for more robust foreground decomposition. After reconstruction, we can render high-quality multiview-consistent 2D object masks.

Explicit regularization with coarse decomposition. We empirically find that decomposed modeling alone cannot robustly separate a foreground object from its surroundings, especially for thin structures and regions in closed contact. Therefore, we leverage the geometric cues in the coarse decomposition results, including the segmented foreground SfM point cloud and the estimated ground plane, to provide extra regularization for training the SDF-based radiance field. Firstly, the unsigned distances $|f(\mathbf{x})|$ of SfM points $\mathbf{x} \in \mathbf{P}_g$ located on the estimated ground plane $\mathbf{P}_g$ are constrained to be larger than a lower bound $\theta(\mathbf{x})$:

$$\mathcal{L}_g = \frac{1}{N_g} \sum_{\mathbf{x} \in \mathbf{P}_g} \max(\theta(\mathbf{x}) - |f(\mathbf{x})|, 0), \qquad (4)$$
$$\theta(\mathbf{x}) = \mu(\mathbf{x}) + \lambda \cdot \sigma(\mathbf{x}),$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are means and standard deviations of unsigned distances between point $\mathbf{x}$ and its K neareast neighbors. This constraint prevents the foreground network from modeling the ground plane.

Moreover, the foreground SfM point cloud is regarded as a rough surface prior to regularize the signed distance

field, similar to Geo-NeuS [7]. This regularization can speed up convergence, alleviate the shape-radiance ambiguity and improve the reconstruction quality of thin structures. Instead of directly constraining the SDF value of each SfM point to zero like in [7], we take the noise of SfM point clouds into account. Specifically, we model the positional uncertainty $\tau(\mathbf{x})$ of each point $\mathbf{x} \in \mathbf{P}_{fg}$ from the foreground SfM point cloud $\mathbf{P}_{fg}$ by its distances to the neighboring points similar to $\theta(\mathbf{x})$ in Eq. (4). Then we constrain the unsigned distance $|f(\mathbf{x})|$ of $\mathbf{x}$ to be smaller than $\tau(\mathbf{x})$:

$$\mathcal{L}_{fg} = \frac{1}{N_{fg}} \sum_{\mathbf{x} \in \mathbf{P}_{fg}} \max(|f(\mathbf{x})| - \tau(\mathbf{x}), 0). \tag{5}$$

To further enhance high-quality foreground renderings with sharp boundaries, we add a beta distribution prior [19] $\mathcal{L}_{bin}$ on the accumulated weights $O(\mathbf{r})$ of each ray $\mathbf{r} \in \mathbf{R}_{fg}$ intersecting with the object bounding box. Finally, we use the eikonal term $\mathcal{L}_{eik}$ [10] on sampled foreground points. Our total loss is:

$$\mathcal{L} = \mathcal{L}_{color} + \alpha \mathcal{L}_{eik} + \beta \mathcal{L}_{g} + \gamma \mathcal{L}_{fg} + \zeta \mathcal{L}_{bin}. \tag{6}$$

Foreground rendering and salient object extraction. With the reconstructed SDF-based radiance field of the foreground object, we can easily render its multi-view consistent 2D masks and extract its mesh. As our reconstruction models the foreground object and background with different fields, we simply compute the accumulated weights of our foreground field along each ray intersecting the object bounding box as object masks, which are binarized with a threshold of 0.5. We use Marching Cubes [20] to extract the object mesh. We can obtain a background-free object 3D model without post-processing thanks to our decomposed scene modeling.

4.3. Implementation details

The input images used for SfM reconstruction are resized to a max area of 720,000. We extract frame-wise 2D features from the ViT-S/8 version of DINO-ViT. We use our data generation pipeline to process 880 object-centric videos from the CO3D-V2 dataset, which includes various categories. All chair objects are kept as a holdout set for validation. This leads to 800 objects for training and 80 objects for validation. We train our 3D segmentation Transformer for 20 epochs. We use multiresolution hash encoding [25] and separate proposal MLPs [1] in all fields of our scene representation. We train our scene representation for $60k$ iterations, which takes 2 hours on a single NVIDIA V100 GPU. All loss weights in Eq. (6) are set to 0.1. The explicit regularization terms are only applied during the initial $15k$ iterations with their loss weights gradually annealed to zero.

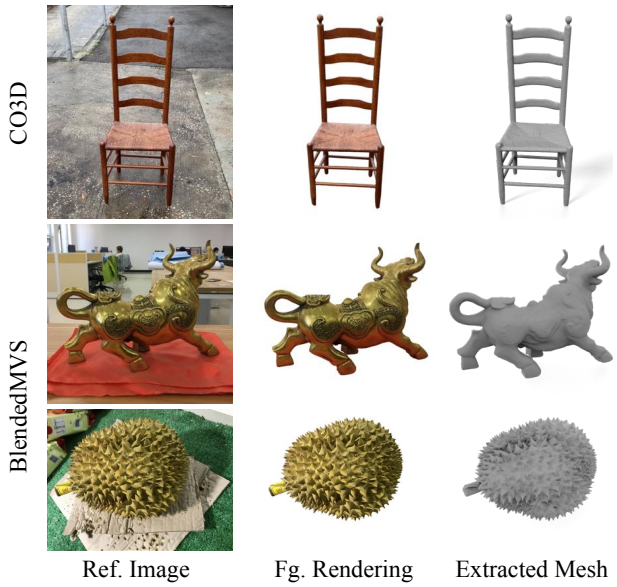

Figure 5. **Background-free salient object reconstruction results.**

5. Experiments

5.1. Datasets

We evaluate the proposed method on CO3D-V2 [29], BlendedMVS [45] and DTU [12] datasets.

CO3D contains 19,000 video sequences of objects from 50 MS-COCO categories. Many objects in CO3D contain thin structures, which are challenging for detection and segmentation from SfM point clouds. We use the CO3D dataset to evaluate 3D salient object detection and 2D segmentation to demonstrate the capabilities of our method on challenging objects and casual captures. To evaluate 3D detection, we manually annotate ground-truth 3D bounding boxes of 80 objects from the chair category based on the given MVS point clouds. Moreover, we annotate detailed 2D foreground masks of 5 objects to evaluate 2D segmentation.

BlendedMVS and DTU datasets are widely used for 3D reconstruction. We use these datasets to evaluate 3D salient object detection, reconstruction, and 2D segmentation. Since meshes provided by BlendedMVS come with backgrounds, we manually segment foreground meshes and render multi-view masks of 5 objects for evaluation. Foreground object meshes are also used for producing ground truth 3D bounding boxes. When evaluating reconstruction results, all meshes are preprocessed with object masks.

5.2. 3D salient object detection

In this part, we evaluate our coarse decomposition results based on the 3D bounding boxes inferred from segmented foreground point clouds. More details about generating bounding boxes can be found in supplemental material.

	CO3D		BlendedMVS		DTU	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
TokenCut + Seg. Agg.	0.816	0.204	0.875	0.625	0.500	0.167
Ours NCut (ablation)	0.867	0.306	**1.00**	0.75	**1.00**	0.667
Ours Transformer	**0.908**	**0.581**	**1.00**	**1.00**	0.833	**0.833**

Table 1. **Quntitative results of 3D salient object detection.** Our method is compared with baselines using the average precision (AP) of 3D bounding box IoU with different thresholds.

Baselines. To the best of our knowledge, there is no existing baseline that holds the same setting as our coarse decomposition pipeline, which detects 3D salient objects from SfM point clouds without manual annotation for model training. Therefore, we devise two straightforward pipelines for comparison to demonstrate the effectiveness of our design. The first baseline is TokenCut + segmentation aggregation (*TokenCut + Seg. Agg.*). We first use TokenCut [41] for 2D salient object segmentation on each image and then aggregate multi-view segmentation confidences to the SfM point cloud by averaging. Finally, we segment the SfM point cloud with a threshold of 0.5 to determine the salient object's 3D region. Another baseline is our neural point cloud + NCut-based segmentation (*Ours NCut*), which is used to generate pseudo-GTs for training *Ours Transformer*.

Evaluation metrics. We use the Intersection-over-Union (IoU) metric with thresholds of 0.5 and 0.7 to evaluate the bounding box accuracy. The average percision (AP) is used for comparison.

Results. As shown in Table 1, our approach substantially achieves better salient object detection performances on all datasets, especially on the challenging CO3D [29] dataset. Instead of individually segmenting 2D images as in *TokenCut + Seg. Agg.*, our strategy of aggregating multi-view 2D features and performing segmentation on 3D facilitates global reasoning of the salient object and eliminates multi-view inconsistency. The proposed *Ours Transformer* also outperforms *Ours NCut* baseline on most datasets and metrics although trained on pseudo-GTs generated by *Ours NCut*. We attribute this improvement to *Our Transformer*'s ability to accept point clouds with a higher density as inputs, its ability to capture global dependencies, and the extra learning on the dataset generated by *Ours NCut*.

5.3. Object reconstruction and 2D segmentation

We evaluate the reconstructed object geometry and 2D foreground mask renderings to demonstrate the capability of our approach to reconstruct and segment complex objects.

Baselines. For 3D reconstruction, we compare our method with the neural surface reconstruction baseline NeuS [40]. As for the evaluation of 2D segmentation, the proposed method is compared with following baselines in two categories: 1) single-view image segmentation baseline TokenCut [41], which performs 2D salient object segmentation on each image and does not consider multi-view information. 2) multi-view image segmentation baseline SemanticNeRF [51], which fuses noisy masks with a neural field and produces high-quality masks with neural rendering. Specifically, we use the segmentations from TokenCut as inputs for SemanticNeRF and evaluate its mask renderings.

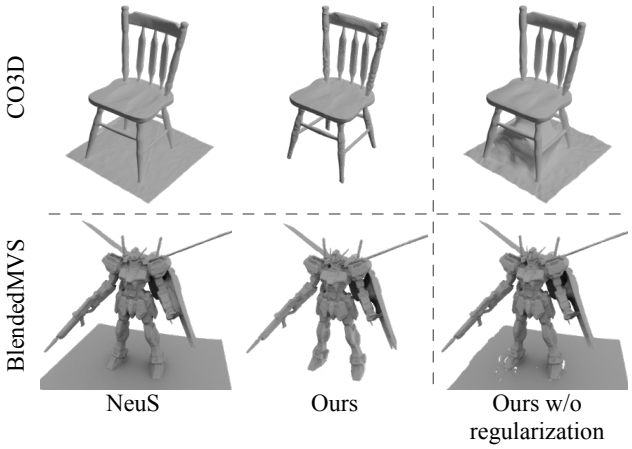

Figure 6. **Qualitative results of salient object reconstruction.** Our method is compared with NeuS on the CO3D and BlendedMVS datasets. We present results with and without our explicit regularization to illustrate its effectiveness.

Scan ID	1	2	3	4	5	6	Mean
NeuS (w/ annotated fg. region)	**0.390**	0.216	0.245	0.223	**0.345**	0.271	**0.282**
Ours (fully-automated)	0.411	**0.200**	**0.240**	**0.218**	0.379	**0.264**	0.285

Table 2. **Quantitative results on the BlendedMVS dataset.** We normalize the GT mesh so that its longest side equals one. Results on the Chamfer l_2 distance are presented as percentages.

Evaluation metrics. We evaluate 3D reconstruction on the Chamfer l_2 distance. Mask IoU and Boundary IoU [4] metrics are used to evaluate 2D segmentation, with the former reflecting overall segmentation quality and the latter focusing on boundary quality. The definitions of these metrics can be found in the supplemental material.

Results. For foreground object reconstruction, qualitative results are shown in Figs. 5 and 6, and quantitative results on the BlendedMVS dataset are presented in Table 2. The proposed fully-automated pipeline achieves comparable or better reconstruction quality compared with NeuS, which is provided with manually-annotated object regions and requires manual post-processing for background removal. Our pipeline eliminates these tedious labors and thus demonstrates the potential to automatically create large-scale datasets.

Our method also achieves better 2D segmentation accuracy on most of the evaluated scans, as shown in Table 3

	CO3D						BlendedMVS						DTU					
Scan ID	1	2	3	4	5	Mean	1	2	3	4	5	Mean	1	2	3	4	5	Mean
	Mask IoU																	
Ours	**0.933**	**0.951**	0.958	0.962	0.934	**0.947**	0.959	**0.987**	0.916	**0.936**	**0.977**	**0.955**	**0.931**	**0.969**	**0.961**	**0.959**	0.903	**0.945**
TokenCut (single-view)	0.784	0.888	**0.976**	0.975	**0.966**	0.918	0.785	0.904	0.919	0.855	0.943	0.881	0.829	0.921	0.905	0.955	0.971	0.916
TokenCut + SemanticNeRF	0.825	0.861	0.952	**0.980**	0.914	0.906	**0.972**	0.906	**0.924**	0.877	0.941	0.924	0.828	0.921	0.907	0.957	**0.975**	0.918
	Boundary IoU																	
Ours	**0.912**	**0.937**	0.839	0.771	0.843	**0.860**	**0.816**	**0.914**	**0.767**	**0.896**	**0.817**	**0.842**	**0.628**	**0.842**	**0.752**	**0.707**	0.613	**0.877**
TokenCut (single-view)	0.635	0.832	**0.877**	**0.839**	**0.887**	0.814	0.493	0.562	0.664	0.688	0.695	0.620	0.572	0.693	0.525	0.636	0.803	0.646
TokenCut + SemanticNeRF	0.701	0.819	0.847	0.822	0.769	0.792	0.512	0.578	0.699	0.730	0.642	0.632	0.539	0.633	0.522	0.661	**0.836**	0.638

Table 3. **Quantitative results of 2D segmentation.** We compare our foreground mask renderings with baselines on Mask IoU and Boundary IoU metrics on multiple datasets, including CO3D, BlendedMVS, and DTU. AutoRecon outperforms baselines on most of the scans.

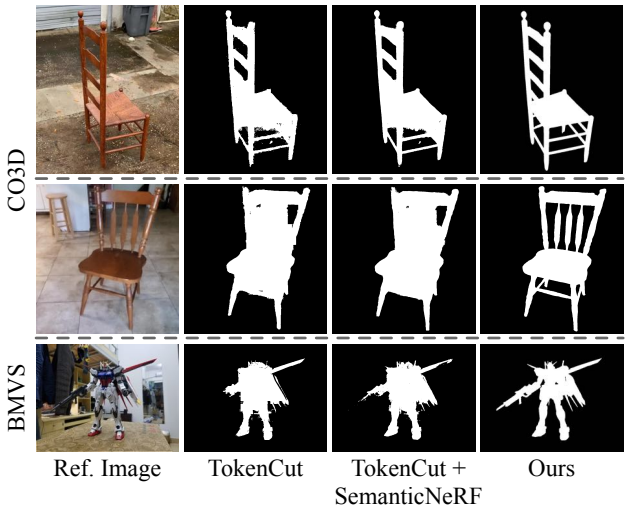

Figure 7. **Qualitative results of 2D segmentation.** We show foreground segmentation on the challenging chair category in CO3D and an object in BlendedMVS with complex geometry.

and visualized in Fig. 7. The results of the 2D salient object segmentation baseline TokenCut lack multi-view consistency and are noisy on scans with complex backgrounds. SemanticNeRF can bring improvement to the initial TokenCut segmentations on some scans. The proposed method can handle complex objects and backgrounds and outperforms these baselines significantly on the Boundary IoU metric, which demonstrates the capability of producing high-quality segmentations.

5.4. Ablation studies

We conduct experiments to validate the effectiveness of our point cloud segmentation Transformer for the coarse decomposition and regularization terms for training our decomposed neural scene representation. More ablation studies are provided in the supplementary material.

Segmentation Transformer for coarse decomposition. We show the effectiveness of our 3D segmentation Transformer over our NCut-based pipeline from the higher 3D detection AP on multiple datasets, as shown in Table 1. Results show that although trained with pseudo-labels, the 3D detection accuracy improves significantly, especially on the CO3D dataset. Moreover, *Ours Transformer* runs $\sim 100\times$ times faster than *Ours NCut* to segment a point cloud with $10k$ points and is applicable to large-scale point clouds.

Explicit regularization for training decomposed neural scene representation. The qualitative results in Fig. 6 demonstrate the effectiveness of explicit regularization in disentangling foreground objects from their surroundings. Regularization provided by the coarse decomposition also alleviates the shape-radiance ambiguity as shown in the chair example.

6. Conclusion

We present a novel pipeline for fully-automated object discovery and reconstruction from multi-view images, without any human annotation. Experiments conducted on multiple real-world datasets show the effectiveness of our method in building high-quality background-free object models. We also demonstrate the capability of our pipeline in producing high-quality segmentation masks, which are directly applicable to 2D supervised learning.

Limitations and future work. Problems faced by neural reconstruction methods remain in our pipeline, like sensitivity to shadows and transient occluders and degraded results on thin-structured and non-Lambertian objects. Storing multi-view ViT features is memory-intensive, which we expect to be alleviated by distance-preserving compression techniques. The reconstruction quality of SfM point clouds can be further improved with refinement methods like [11, 17], which can further improve the quality of surface reconstruction and potentially eliminate reconstruction ambiguities. Our automated object reconstruction pipeline can be used to create large-scale 3D object datasets for graphics and perception tasks, such as training 2D segmentation networks and 3D generative models.

Acknowledgement. This work was supported by NSFC (No. 62172364), the ZJU-SenseTime Joint Lab of 3D Vision, and the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 5, 6

[2] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv:1901.11390*, 2019. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, 2021. 1, 2, 3

[4] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[6] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: generative scene inference and sampling with object-centric latent representations. In *ICLR*, 2020. 2

[7] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. In *NeurIPS*, 2022. 6

[8] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 1

[9] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*, 2019. 2

[10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 6

[11] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 4, 8

[12] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 2, 6

[13] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 4

[14] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022. 2

[15] Amit Pal Singh Kohli, Vincent Sitzmann, and Gordon Wetzstein. Semantic implicit neural scene representations with semi-supervised training. In *3DV*, 2020. 2

[16] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *CVPR*, 2022. 2

[17] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021. 8

[18] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 2

[19] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM TOG*, 2019. 6

[20] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH Comput. Graph.*, 1987. 6

[21] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, 2022. 1, 3

[22] 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*, 2021. 5

[23] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *ECCV*, 2022. 1, 2

[24] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *CVPR*, 2022. 2

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*, 2022. 6

[26] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 2

[27] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 2

[28] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2

[29] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 2, 6, 7

[30] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In *CVPR*, 2022. 1, 2

[31] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. In *NeurIPS*, 2022. 2

[32] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*. 2016. 1, 2

[33] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE TPAMI*, 2000. 2, 3, 5

[34] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 1, 3

[35] Cameron Smith, Hong-Xing Yu, Sergey Zakharov, Fredo Durand, Joshua B. Tenenbaum, Jiajun Wu, and Vincent Sitzmann. Unsupervised Discovery and Composition of Object Light Fields. *arXiv:2205.03923*, 2022. 2

[36] Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv:2104.01148*, 2021. 2

[37] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3

[38] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D Distillation of Self-Supervised 2D Image Representations. In *3DV*, 2022. 2

[39] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. NeSF: Neural Semantic Fields for Generalizable Semantic Segmentation of 3D Scenes. *arXiv:2111.13260*, 2021. 2

[40] 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 *NeurIPS*, 2021. 1, 2, 3, 5, 7

[41] Yangtao Wang, Xi Shen, Shell Hu, Yuan Yuan, James Crowley, and Dominique Vaufreydaz. Self-Supervised Transformers for Unsupervised Object Discovery using Normalized Cut. In *CVPR*, 2022. 1, 3, 7

[42] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-Compositional Neural Implicit Surfaces. In *ECCV*, 2022. 2

[43] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering. In *ICCV*, 2021. 2

[44] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2

[45] 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 *CVPR*, 2020. 2, 6

[46] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 1, 2

[47] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeuIPS*, 2020. 2, 5

[48] Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *ICLR*, 2022. 2

[49] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *ICLR*, 2023. 3

[50] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv:2010.07492*, 2020. 1, 2, 3

[51] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2, 7

[52] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Revealing objects in neural fields. *IEEE Robot. Autom. Lett.*, 2022. 2