Panoptic Compositional Feature Field for Editable Scene Rendering with Network-Inferred Labels via Metric Learning

Xinhua Cheng[†], Yanmin Wu[†], Mengxi Jia[‡], Qian Wang[†], Jian Zhang[†]
[†]Shenzhen Graduate School, Peking University, China
[‡]School of Software and Microelectronics, Peking University, China
chengxinhua@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

Abstract

*Despite neural implicit representations demonstrating impressive high-quality view synthesis capacity, decomposing such representations into objects for instance-level editing is still challenging. Recent works learn object-compositional representations supervised by ground truth instance annotations and produce promising scene editing results. However, ground truth annotations are manually labeled and expensive in practice, which limits their usage in real-world scenes. In this work, we attempt to learn an object-compositional neural implicit representation for editable scene rendering by leveraging labels inferred from the off-the-shelf 2D panoptic segmentation networks instead of the ground truth annotations. We propose a novel framework named **P**anoptic **C**ompositional **F**eature **F**ield (PCFF), which introduces an instance quadruplet metric learning to build a discriminating panoptic feature space for reliable scene editing. In addition, we propose semantic-related strategies to further exploit the correlations between semantic and appearance attributes for achieving better rendering results. Experiments on multiple scene datasets including ScanNet, Replica, and ToyDesk demonstrate that our proposed method achieves superior performance for novel view synthesis and produces convincing real-world scene editing results.*

1. Introduction

Virtually editing real-world scenes (*e.g.*, moving a chair in the room) in mixed reality applications on various devices is desired by users. Such expectation requires an effective 3D scene representation with the capacity of photo-realistic view rendering and promising scene decomposition. Recently, emerging neural implicit representations with volumetric rendering, especially neural radiance field (NeRF) [29] and its variants, show impressive results in novel view synthesis [1, 2] and scene reconstruction [13, 31, 43] tasks. However, decomposing a neural implicit representation into objects for scene editing is challenging due to the holistic scene is implicitly encoded as weights of connectionist networks like multi-layer perceptions (MLPs). To build object-compositional neural implicit representations for instance-level scene editing, several works [46, 47, 52] jointly encode the appearance and instance attributes with extra instance annotations.

Though existing object-compositional methods can extract convincing object representations from the scene representation for further editing, their successes rely heavily on ground truth instance annotations, which are labeled manually and expensive to obtain in real-world practice. An intuitive alternative solution is training object-compositional representations with labels inferred by 2D panoptic segmentation networks [5, 6, 19] instead of ground truth instance annotations. However, their methods are tough to leverage the network-inferred labels due to the significant 3D index inconsistency, and a detailed discussion is shown in Fig. 1. We note that **3D index consistency** is the instance indices of a specific object are same across multi-view labels. Due to network-inferred labels are individually predicted on each view image and objects order is uncertain in each prediction, the instance indices of a specific object in different view labels are usually index inconsistent from the perspective of 3D, *e.g.*, the index of the target chair is purple in the label of view #1 and is red in the label of view #2. Therefore, how to learn object-compositional neural implicit representations by leveraging network-inferred labels is critical for real-world application.

In this work, we propose a novel panoptic compositional feature field (PCFF), which integrates the deep metric learning [18, 27] into the learning of object-compositional representations to overcome the challenge of using 2D network predictions. Concretely, we employ metric learning to constrain the distances among projected panoptic features of pixels in each view separately, which circumvents the requirement of 3D index consistent labels and builds a discriminating panoptic feature space. Combined with the

This work was supported in part by the Shenzhen General Research Project JCYJ20220531093215035. *(Corresponding author: Jian Zhang.)*

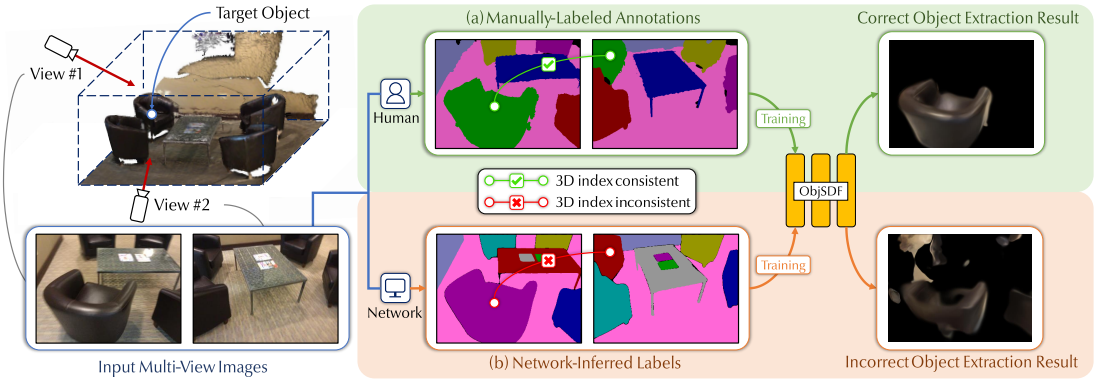

Figure 1. **Core challenge.** Existing methods (*e.g.*, ObjSDF [46]) require (a) manually-labeled ground truth annotations to train object-compositional representations, and the trained representation can extract the target object correctly. We note that (a) manually-labeled annotations are **3D index consistency**, *i.e.*, the instance indices of the target object are same across multi-view labels. However, when (b) network-inferred labels predicted by 2D panoptic segmentation networks are utilized for training their representations, the corresponding object extraction result is obviously incorrect. Due to (b) network-inferred labels are inferred by networks on each view image individually, these labels are usually index inconsistent from the perspective of 3D and tough to be used by existing methods.

feature spaces, we provide an easily query-based manner for scene decomposition, *i.e.*, given a user-specified pixel query in an arbitrary view, our trained PCFF extracts the target object by measuring the similarity between the projected feature of query pixel and corresponding features of each 3D point. Furthermore, two semantic-related learning strategies are proposed based on our observation of the correlations between semantic and appearance attributes for improving our rendering performance. The semantic-appearance hierarchical learning enforces our framework to encode appearances and semantics with MLPs of different depths, and the semantic-guided regional refinement impels the framework to focus on inaccurate regions with the guidance of semantic information entropy maps.

We evaluate our method on multiple scene datasets including ScanNet [8], Replica [40], and ToyDesk [47]. Experiments demonstrate our method outperforms state-of-the-art methods in novel view synthesis. More importantly, PCFF successfully leverages 2D network-inferred labels to build object-compositional representations for object extraction and real-world scene editing, whereas existing methods fail to utilize labels without 3D index consistency. The main contributions of our work are summarized as:

- We propose a novel Panoptic Compositional Feature Field (PCFF) that learns object-compositional representations for editable scene rendering with network-inferred panoptic labels by building a discriminating feature space with the assistance of introduced instance quadruplet metric learning.
- We propose strategies including semantic-appearance hierarchical learning and semantic-guided regional refinement to properly exploit the correlations between semantic and appearance attributes and improve the rendering capacity of our method.
- Our method achieves superior novel view synthesis performance than state-of-the-art methods and produces convincing scene edits by using network-inferred panoptic labels on real-world scenes.

2. Related Works

Neural Implicit Representations. Recently proposed neural implicit representations such as DeepSDF [33] and Occupancy Networks [28] have attracted broad interest due to their powerful representation capability, which map 3D coordinates into continuous geometries and overcome the discretization issue of explicit representations. Without requiring ground truth geometry, NeRF [29] models the target 3D scene as a continuous volumetric field of density and color via a multi-layer perceptron from posed multi-view images and achieves remarkable photo-realistic rendering results. The following works extend NeRF to address its limitations in various aspects, including efficient rendering [2, 9, 25, 38, 41, 48], better generalization [3, 16, 44, 49], dynamic synthesis [10, 23, 35, 36, 45], appearance and shape editing [17, 26, 42], and scene stylization [7, 11, 15, 51]. In this work, we adopt NeRF as the basic neural implicit representation and explore how to learn object-compositional scene implicit representations with network-inferred labels predicted by 2D panoptic segmentation networks.

Object-compositional Implicit Representations. Although neural implicit representations especially NeRFs achieve significant accomplishment, decomposing NeRFs into object representations for further editing is still challenging due to the implicitly represent formulation. Most existing object-compositional implicit representations can be roughly summarized into two streams, including category-specific methods and scene-specific methods.

Several works explore category-specific object representations by leveraging large-scale images or videos of cor-

responding categories. Some attempts [14, 30, 50] build object-centric radiance fields and represent scenes as the composition of object representations for flexible scene editing. Other studies [12, 22, 32] learn multiple neural representations for scenes in the traffic domain to enable panoptic segmentation and object manipulation on novel view synthesizes. However, category-specific methods are struggled to represent objects in unseen categories due to the lacking of essential data as prior.

Related to ours, scene-specific methods directly encode the specific scene as a holistic implicit representation [21, 52] or a composition of representations [46,47] for supporting object extraction and scene editing. DFF [21] distills the knowledge of pre-trained 2D image feature extractors in NeRF and semantically selects regions by the text or image patch queries, whereas DFF does not support instance-level object selection. SemanticNeRF [52] augments NeRF to jointly encode appearance and semantics by simply adding a prediction head supervised by ground truth annotations, hence SemanticNeRF allows decomposing scene into objects according to the instance predictions of 3D points. ObjectNeRF [47] and ObjectSDF [46] are built upon NeRF and SDF respectively, they decompose the target scene as the representations of background and foreground objects by utilizing precise instance masks, leading to promising scene decomposition and reconstruction results. However, [46, 47, 52] are limited in practice due to their high dependence on ground truth instance annotations which is labeled manually and expensive to obtain. In contrast, we introduce metric learning to train object-compositional implicit representations by utilizing network-inferred labels, which is critical for wide real-world applications.

3. Method

Given a set of posed images of a specific scene and corresponding panoptic labels inferred by a pre-trained 2D panoptic segmentation network, we aim to learn an object-compositional NeRF with the capacity of view rendering and scene decomposition. The core challenge is that network-inferred labels are 3D index inconsistency, resulting in the direct prediction of instances adopted by existing methods being infeasible. We propose an effective framework for leveraging such labels by introducing metric learning to achieve convincing scene decomposition results.

In the following, we first give the background of NeRF and panoptic segmentation in Sec. 3.1. We then show the details of the proposed framework in Sec. 3.2, and we finally describe how to extract the representation of the target object with the query pixel for scene editing in Sec. 3.3.

3.1. Backgrounds

Neural Radiance Fields. Given multi-view images with known camera parameters, NeRF [29] uses a multi-layer perception (MLP) to implicitly represent the 3D scene as a continuous volumetric radiance field. Specifically, MLP F_Θ maps a spatial coordinate $\mathbf{x} = (x, y, z)$ and a view direction $\mathbf{d} = (\theta, \phi)$ to a view-independent density σ and view-dependent color $\mathbf{c} = (r, g, b)$. Given the camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with camera position $\mathbf{o}$ and depth $t \in [t_n, t_f]$, the projected color of $\mathbf{r}(t)$ is obtained by sampling N points along the ray and using volume rendering:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i, \tag{1}$$

where σ_i and $\mathbf{c}_i$ denote the density and color of sampled point $\mathbf{x}_i$ along the ray $\mathbf{r}$, $T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$ indicates the accumulated transmittance along the ray, and $\delta_i = ||\mathbf{x}_{i+1} - \mathbf{x}_i||_2$ is the distance between adjacent sample points. To supervise NeRF, RGB loss $\mathcal{L}_{rgb}$ is adopted and formulated as the squared error between the projected color $\hat{\mathbf{C}}(\mathbf{r})$ and the ground-truth color $\mathbf{C}(\mathbf{r})$:

$$\mathcal{L}_{rgb} = \sum_{\mathbf{r}\in\mathcal{R}} ||\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})||_2^2, \tag{2}$$

where $\mathcal{R}$ is the ray set of sampled pixels. We adopt the popular NeRF as the basic neural implicit representation due to its high-quality view synthesis capacity.

Panoptic Segmentation. To generate a rich and complete scene segmentation, the panoptic segmentation (PS) task is proposed [20] to unify the typical semantic segmentation and instance segmentation task. PS networks map each pixel of the input image to a pair (s, z), where s represents the semantic class and z represents the instance index of the pixel. The semantic classes set $\mathcal{C}$ consists of *stuff* subset $\mathcal{C}_s$ (amorphous regions of similar texture or material, *e.g.*, wall and floor) and *things* subset $\mathcal{C}_t$ (countable objects, *e.g.*, chairs and tables). Concretely, all pixels labeled with the same s belong to the same instance and z is irrelevant when $s \in \mathcal{C}_s$, while all pixels labeled with the same (s, z) belong to the same instance when $s \in \mathcal{C}_t$.

We leverage panoptic segmentation networks [5, 6, 19] to make predictions on each view image individually, and network-inferred labels are divided into semantic and instance sub-labels with the PS definition. Since semantic labels are 3D index consistent when the adopted PS network is pre-trained with the same $\mathcal{C}$, we can jointly learn appearance and semantic attributes in a similar way. However, the instance labels are usually inconsistent due to the uncertain order of object predictions. Thus a metric-based constraint is designed for leveraging inferred instance labels.

3.2. Panoptic Compositional Feature Field

In this section, we give the details of our proposed framework named Panoptic Compositional Feature Field (PCFF), and the overview is shown in Fig. 2. Firstly, we describe the joint learning of color $\mathbf{c}$ and semantic $\mathbf{s}$ and proposed semantic-related strategies which explore the correlations

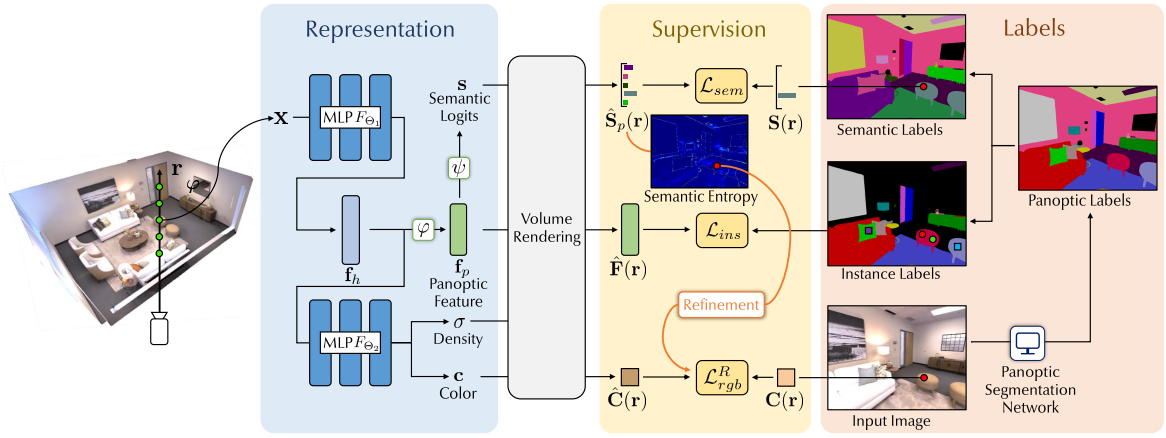

Figure 2. **Overview of the proposed PCFF framework**. PCFF adopts MLPs to map a 3D coordinate $\mathbf{x}$ and the view direction of ray $\mathbf{r}$ to density σ, color $\mathbf{c}$, semantic logits $\mathbf{s}$ and panoptic feature $\mathbf{f}_p$. For the image pixel of ray $\mathbf{r}$, we use differentiable volume rendering to obtain corresponding projected color $\hat{\mathbf{C}}(\mathbf{r})$, projected semantic probabilities $\hat{\mathbf{S}}_p(\mathbf{r})$ and projected panoptic feature $\hat{\mathbf{F}}(\mathbf{r})$. $\hat{\mathbf{C}}(\mathbf{r})$ and $\hat{\mathbf{S}}_p(\mathbf{r})$ are supervised with input images and semantic labels, and semantic-related strategies are proposed to explore the correlations between $\mathbf{c}$ and $\mathbf{s}$ for improving rendering performance. To leverage instance labels without 3D index consistency, we introduce instance metric constraint on $\hat{\mathbf{F}}(\mathbf{r})$ to build a discriminating feature space for scene decomposition and editing.

between $\mathbf{c}$ and $\mathbf{s}$ for improving our rendering capacity. Then, we show how to utilize metric learning to build discriminating feature spaces for scene decomposition at instance level with network-inferred labels.

Semantic-Appearance Hierarchical Learning. Inspired by the [22, 37, 52], our framework is developed from an intuitively semantic extension of NeRF, which uses a MLP F_Θ to map a spatial coordinate $\mathbf{x}$ and a view direction $\mathbf{d}$ into a view-dependent color $\mathbf{c}$, view-invariant density σ and semantic logits $\mathbf{s}\in\mathbb{R}^{|\mathcal{C}|}$, where $\mathcal{C}$ is the set of semantic classes. The semantic extension of NeRF is formulated as:

$$(\sigma, \mathbf{s}) = F_\Theta(\mathbf{x}),\ \mathbf{c} = F_\Theta(\mathbf{x}, \mathbf{d}). \tag{3}$$

Giving the camera ray $\mathbf{r}$ of a pixel, the projected semantic logits $\hat{\mathbf{S}}(\mathbf{r})$ is formulated as:

$$\hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{s}_i, \tag{4}$$

where $\mathbf{s}_i$ denotes the predicted semantic logits of point $\mathbf{x}_i$ along the ray $\mathbf{r}$. We forward the logits $\hat{\mathbf{S}}(\mathbf{r})$ into a softmax normalization layer to compute the multi-class probabilities $\hat{\mathbf{S}}_p(\mathbf{r})$. The semantics are supervised by a cross-entropy loss between $\hat{\mathbf{S}}_p(\mathbf{r})$ and the semantic label $\mathbf{S}(\mathbf{r})$:

$$\mathcal{L}_{sem} = -\sum_{\mathbf{r}\in\mathcal{R}}\sum_{j=1}^{|\mathcal{C}|} \mathbf{S}^j(\mathbf{r}) \log \hat{\mathbf{S}}_p^j(\mathbf{r}), \tag{5}$$

where $\mathcal{R}$ is the ray set of pixels. $\mathbf{S}^j(\mathbf{r}) \in \{0, 1\}$ is the j-th class label of the intersection pixel of ray $\mathbf{r}$ and image.

Although the joint encoding of $\mathbf{s}$ and $\mathbf{c}$ enhances the network for better utilizing the inherent multi-view consistency and improves the accuracy of semantic segmentation, the introduction of $\mathbf{s}$ incurs negative effects for the rendering capacity, as discussed in [52]. Due to the observation that $\mathbf{s}$ is an easier attribute to learn than $\mathbf{c}$, we assume that the performance degradation is caused by the tight coupling of $\mathbf{s}$ and $\mathbf{c}$ brought by the deeply shared MLPs. Hence, a semantic-appearance hierarchical learning strategy is proposed to encode $\mathbf{s}$ with shallower layers, which is simply implemented by modifying F_Θ into two sub-MLPs F_{Θ_1} and F_{Θ_2} (see Fig. 2) for hierarchical learning:

$$\mathbf{f}_h = F_{\Theta_1}(\mathbf{x}),\ \sigma = F_{\Theta_2}(\mathbf{f}_h),\ \mathbf{c} = F_{\Theta_2}(\mathbf{f}_h, \mathbf{d}), \tag{6}$$

where $\mathbf{f}_h$ is the intermediate hidden feature for the subsequent encoding of $\mathbf{s}$:

$$\mathbf{f}_p = \varphi(\mathbf{f}_h),\ \mathbf{s} = \psi(\mathbf{f}_p), \tag{7}$$

where $\varphi(\cdot), \psi(\cdot)$ are linear layers. $\mathbf{f}_p$ is named panoptic feature because it is both supervised by the semantic and instance losses, and $\mathbf{f}_p$ is further utilized as the selected criterion for the scene decomposition.

Semantic-Guided Regional Refinement. According to our observation, the framework is struggled to render photo-realistic results in regions with semantic prediction difficulty, *e.g.*, object boundaries or observe lacking parts. Thus we design a semantic-guided regional refinement strategy to impel the framework to focus on semantic inaccurate regions with the guidance of semantic information entropy maps (see Fig 2). Specifically, following [39], we formulate the semantic information entropy of ray $\mathbf{r}$ as $H(\mathbf{r})$ to reflect the semantic prediction difficulty:

$$H(\mathbf{r}) = -\sum_{j=1}^{|\mathcal{C}|} \hat{\mathbf{S}}_p^j(\mathbf{r}) \log \hat{\mathbf{S}}_p^j(\mathbf{r}), \tag{8}$$

where $\hat{\mathbf{S}}_p(\mathbf{r})$ is the projected multi-class probabilities of ray $\mathbf{r}$, $\mathcal{C}$ is the set of semantic classes, and $H(\mathbf{r}) \in [0, \log|\mathcal{C}|]$. Therefore, we adjust the weights of color loss (Eq. 2) under

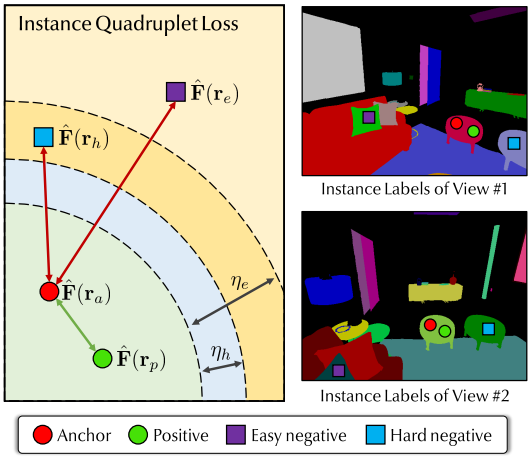

Figure 3. **Conception of the proposed instance quadruplet loss. Left:** For a pixel quadruplet sampled in the same view, the proposed loss closes the feature distance between the anchor and the positive while enlarging feature distances among the anchor and negatives. **Right:** The proposed loss constrains the features consistently in different views regardless of the specific indices.

the guidance of $H(\mathbf{r})$ for imposing stronger constraints on semantic complicated regions:

$$\mathcal{L}^R_{rgb} = \sum_{\mathbf{r}\in\mathcal{R}} (1 + \epsilon H(\mathbf{r}))||\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})||_2^2, \tag{9}$$

where ϵ is the hyper-parameter to control the strength.

Instance Quadruplet Metric Learning. We now describe our *core design* named instance quadruplet metric learning for leveraging the network-inferred labels. We emphasize that predicting instances on 3D points in a similar way with $\mathbf{s}$ will cause severe collisions in 3D space when inferred instance labels are 3D index inconsistency. Therefore, we introduce metric learning to close or enlarge the distances between the projected features of pixels sampled in the same view according to their labels (see Fig. 3 left), which leads to a discriminating feature space for the scene decomposition at instance level. Besides, our metric-based loss constrains features consistently in different views even if the instance indices are changed, *e.g.*, though the index of the left chair is red in view #1 and is green in view #2, the anchor feature should be closed to the positive feature in both views (see Fig. 3 right). Thus proposed metric-based constraint overcomes the 3D index inconsistency in network-inferred labels without degrading the inherent multi-view consistency brought by the joint encoding of $\mathbf{c}$ and $\mathbf{s}$.

Concretely, we denote that (s_a, z_a) as the semantic label and instance label of pixel a by the definition in Sec. 3.1. To implement the instance quadruplet loss, we construct a quadruplet for the given pixel a which belongs to a countable object ($s_a \in \mathcal{C}_t$), consisting of a itself as the anchor, a positive sample p ($s_a{=}s_p$, $z_a{=}z_p$), an easy negative sample e ($s_a{\neq}s_e$, $z_a{\neq}z_e$), and a hard negative sample h ($s_a{=}s_h$, $z_a{\neq}z_h$). We note that all pixels in a quadruplet are sampled in the same view. For the anchor pixel a which is the intersection of ray $\mathbf{r}_a$ and the view image, we calculate its projected panoptic feature $\hat{\mathbf{F}}(\mathbf{r}_a)$ as:

$$\hat{\mathbf{F}}(\mathbf{r}_a) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{f}_{p,i}, \tag{10}$$

where $\mathbf{f}_{p,i}$ denote the panoptic feature of point $\mathbf{x}_i$ along the ray $\mathbf{r}_a$. Due to the supervision of $\mathbf{s}$, the panoptic feature $\hat{\mathbf{F}}(\mathbf{r}_a)$ is associated with the semantic label s_a, and the distances among panoptic features of pixels with the same semantic label are naturally closer. To alleviate the degradation of rendering performance caused by the introduction of metric learning, we divide negative samples into easy or hard ones according to their corresponding semantic labels, and different margins η_e and η_h are assigned with the principle $\eta_e{>}\eta_h$ to establish more reasonable relationships. Therefore, the instance quadruplet loss is defined as:

$$\mathcal{L}_{ins} = \sum_{\mathbf{r}_a\in\mathcal{Z}} ([\eta_e + d_p - d_e]_+ + [\eta_h + d_p - d_h]_+),$$
$$s.t.\ \ d_k = \mathcal{D}(\hat{\mathbf{F}}(\mathbf{r}_a), \hat{\mathbf{F}}(\mathbf{r}_k)),\ \ k \in \{p, e, h\}, \tag{11}$$

where $\mathcal{Z}$ is the ray set of pixels that belong to countable objects ($s_a \in \mathcal{C}_t$) in the target view, $[\cdot]_+$ is the $\max(\cdot, 0)$ function, and $\mathcal{D}(\cdot,\cdot)$ is the distance function.

Joint Optimization. We jointly optimize our framework by the color loss $\mathcal{L}^R_{rgb}$ (Eq. 9), the semantic loss $\mathcal{L}_{sem}$ (Eq. 5), and instance quadruplet loss $\mathcal{L}_{ins}$ (Eq. 11) in the training stage. Overall, the total loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}^R_{rgb} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{ins}\mathcal{L}_{ins}, \tag{12}$$

where λ_{sem} and λ_{ins} are trade-off hyper-parameters to balance the magnitude of losses. We set $\lambda_{sem}{=}1{\times}10^{-3}$ and $\lambda_{ins}{=}5{\times}10^{-4}$ in experiments. We notice that the learnable layers consist of MLPs $F_{\Theta_1}, F_{\Theta_2}$ and linear layers φ, ψ.

3.3. Query-Based Object Selection

Extracting the representation of the target object from a trained PCFF is the key to instance-level scene editing, we thus propose a query-based manner to select the 3D points belonging to the desired object according to the feature and semantic similarity between the query and points. Concretely, given a user-specified pixel query q of the target object in an arbitrary view, for each point $\mathbf{x}$ in the 3D space, we calculate the feature similarity $\alpha_{\mathbf{x}}$ and the semantic similarity $\beta_{\mathbf{x}}$ between q and $\mathbf{x}$:

$$\alpha_{\mathbf{x}} = 1 - \mathcal{D}(\mathbf{f}_p, \hat{\mathbf{F}}(\mathbf{r}_q)), \tag{13}$$

$$\beta_{\mathbf{x}} = \text{sigmoid}(\mathbf{s}^{j'}),\ \ j' = \arg\max_j(\hat{\mathbf{S}}^j(\mathbf{r}_q)), \tag{14}$$

where $\mathbf{f}_p$ and $\mathbf{s}^{j'}$ are panoptic feature and j'-th semantic logit of point $\mathbf{x}$. $\hat{\mathbf{F}}(\mathbf{r}_q)$ and $\hat{\mathbf{S}}(\mathbf{r}_q)$ are projected panoptic feature and projected semantic logits of query pixel q, and $\mathcal{D}(\cdot,\cdot)$ is the cosine distance. We select points whose similarities are both higher than corresponding thresholds γ_f, γ_s as the representation of the target object for scene editing.

4. Experiments

The main purpose of our method is to learn object-compositional neural implicit representations with posed images and corresponding network-inferred labels predicted by 2D panoptic segmentation networks. We conduct experiments and demonstrate that existing scene-specific methods fail to utilize such labels. In contrast, our method shows robust rendering and editing performance whether using ground truth annotations or network-inferred labels.

4.1. Implementation details

We implement our method with PyTorch [34] and all experiments are performed on an NVIDIA RTX 3090 GPU. We adopt MMDetection [4] for implementing 2D panoptic segmentation networks and networks are pre-trained on COCO [24] dataset. We use a batch size of 1024 rays unless otherwise stated, and 64 coarse points and 128 fine points are sampled for each ray. We train our method using Adam optimizer with default hyper-parameters for 200k iterations, and the learning rate begins at 5×10^{-4} and decays exponentially to 5×10^{-5}. The margin values η_h and η_e are set to 0.3 and 0.5, and the refine hyper-parameter ϵ is set to 0.1.

4.2. Datasets

ScanNet [8] is a large-scale RGB-D dataset of 1513 real-world indoor scenes with annotations including camera poses, panoptic segmentation, and surface reconstructions. We choose 3 scenes for evaluation, and there are 300/100 images with camera poses and panoptic annotations are sampled in each scene for training/testing.

Replica [40] is a reconstruction-based 3D dataset of 18 high-fidelity scenes with dense geometry, HDR textures, and panoptic annotations. We choose 6 one-room scenes for evaluation. Following the train/test data split in [52], we sample 180/180 images with camera poses and panoptic annotations in each scene for training/testing.

ToyDesk [47] is a real-world dataset including 2 scenes of a desk by placing several toys with different layouts and images are 360° captured by looking at the desk center. We follow the standard train/test data split [47] which sample 80% frames for training and use the rest for testing.

4.3. Learning with Network-Inferred Labels

In this section, we conduct experiments to demonstrate that our method can learn object-compositional implicit representations with network-inferred labels while other methods are failed, which is shown in Fig. 4. Due to ObjectNeRF [47] needs ground truth bounding boxes of target objects for extraction, we choose SemanticNeRF [52] and ObjectSDF [46] here for comparison. Mask2Former [5] is adopted as the 2D panoptic segmentation network to predict labels, which are significantly index inconsistent as shown in Fig. 4(b). Fig. 4(c) and (d) respectively show

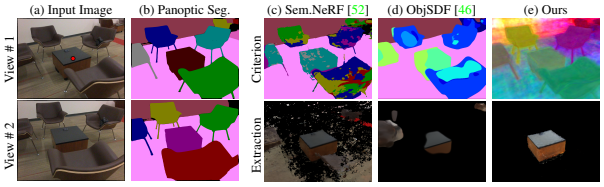

Figure 4. Qualitative comparison with object-compositional methods on object extraction. Our method successfully extracts the target object when trained with network-inferred labels while others are failed. The red dot in View #1 is our query pixel.

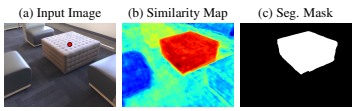

Figure 5. We utilize (b) feature similarity map of query to generate (c) segmentation masks for evaluation.

Methods	IoU ↑
SemanticNeRF [52]	0.357
ObjectSDF [46]	0.486
PCFF (*Ours*)	**0.891**

Table 1. Segmentation accuracy comparison of target objects on ScanNet.

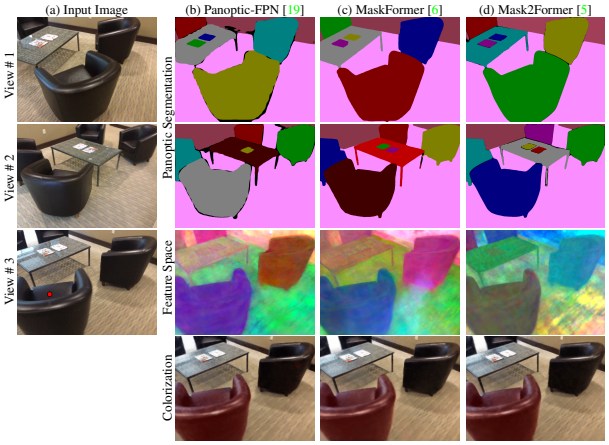

Figure 6. Qualitative comparison of editing capacity when training our method with labels predicted by various panoptic segmentation networks. Our method robustly generates discrimating feature spaces and produces scene edits.

the selection criterion (*i.e.*, segmentation masks) and object extraction results of SemanticNeRF and ObjSDF. Due to their explicit instance predictions suffering severe collision in 3D space when supervised by 3D index inconsistent labels, these methods generate mistaken instance masks and lead to unacceptable object extraction results. Different from their prediction-based criterion, we select the target object according to the panoptic feature space supervised by the proposed instance quadruplet loss. We roughly visualize our panoptic feature space by using 3-dimensional PCA components of the features as RGB [21] for reference and understanding in Fig. 4(e). Our panoptic feature space is discriminating at instance level, where features of different instances are presented in different colors. Therefore, we correctly extract the object by measuring the feature similarities between the query pixel and each 3D point.

To further verify our framework can leverage network-inferred labels while others cannot, we report the averaged

Methods	ScanNet				Replica			ToyDesk		
	Time ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SemanticNeRF [52]	∼8h	26.11	0.796	0.378	31.74	0.921	0.183	22.01	0.775	0.448
ObjectNeRF [47]	∼16h	26.29	0.785	0.350	31.68	0.925	0.159	21.93	0.718	0.431
ObjectSDF [46]	∼20h	26.50	0.798	0.387	27.25	0.858	0.281	21.32	0.794	0.432
PCFF $(Ours)$	∼9h	26.45	0.807	0.355	32.28	0.932	0.163	22.42	0.781	0.435
PCFF* $(Ours)$	∼15h	26.58	0.811	0.346	32.77	0.937	0.150	22.45	0.783	0.431

Table 2. Comparison of rendering performance with state-of-the-art object-compositional methods on ScanNet, Replica, and Toydesk datasets. All methods are supervised by ground truth annotations. * denotes we use a larger batch size to train our method for a fair comparison in training time with ObjectNeRF and ObjectSDF.

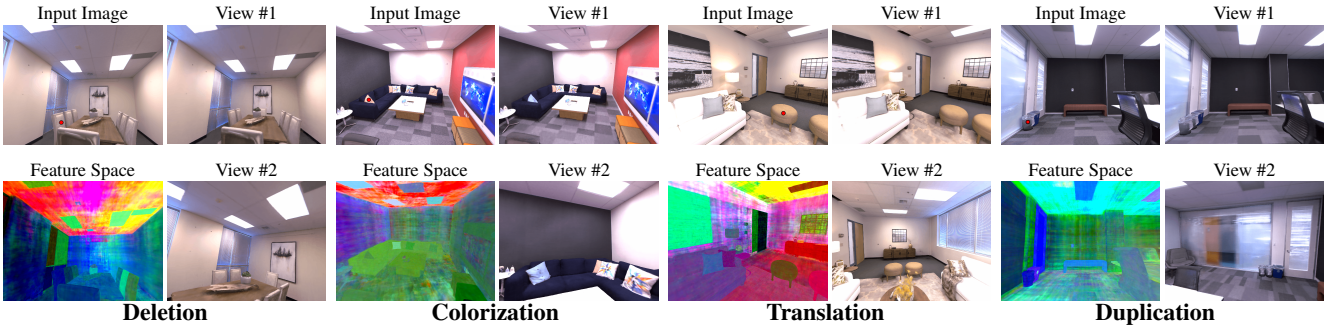

Figure 7. Query-based edits of target objects in Replica scenes. The red dots in input images are query pixels.

segmentation IoU of target objects on ScanNet in Tab 1. Different from SemanticNeRF and ObjSDF, our method cannot produce segmentation masks directly. Thus we select the target object by the query pixel in each scene, and segmentation masks are roughly generated according to the projected feature similarity maps in each view (see in Fig. 5). The comparison results show their explicit instance prediction leads to low performance, while our metric-based constraint can address the 3D index inconsistency.

We then show the robustness of our method when facing labels predicted by various 2D panoptic segmentation networks including Panoptic-FPN [19], MaskFormer [6] and Mask2Former [5], and the qualitative comparison is shown in Fig. 6. Although network-inferred labels are inconsistent and inaccurate from the perspective of 3D, our method robustly builds discriminating feature spaces and produces remarkable colorization results with the query pixel. We observe that the editing quality is related to the segmentation accuracy, *e.g.*, the colorization of the target chair in Fig. 6 (b) is slightly leaked to the adjacent chair caused by the prediction inaccuracy of Panoptic-FPN while the colorization in Fig. 6 (c) and (d) are better, indicating that our method can be benefited from the further development of panoptic segmentation networks.

4.4. Scene Rendering and Editing

To evaluate the scene rendering capacity of our method, we follow the standard metric in [29] by using PSNR, SSIM and LPIPS to measure the rendering quality. The scene-specific object-compositional methods including SemanticNeRF [52], ObjectNeRF [47] and ObjectSDF [46] are compared, and the ground truth instance annotations are used in this experiment because compared methods are tough to utilize network-inferred labels. As shown in Tab. 2, our method achieves comparable rendering performance on all scene datasets and requires significantly less time for training. Our time efficiency is mainly because that our method encodes the scene as a holistic representation, whereas ObjectNeRF and ObjectSDF build separate representations for the background and all foreground objects. Hence, we train PCFF with a larger (2048) batch size of rays to increase the training time to 15 hours for making a fair comparison in training time with ObjectNeRF and ObjectSDF, and PCFF outperforms compared methods in most metrics in this condition. We emphasize that our method is designed for leveraging 2D network-inferred labels. However, thanks to our proposed semantic-related strategies, our method shows remarkable rendering performance.

Furthermore, we show various edits including deletion, colorization, translation, and duplication in multiple Replica scenes in Fig. 7 to demonstrate that our method can produce convincing multi-view consistent scene edits, and the corresponding panoptic feature spaces are PCA-based visualized (see Sec. 4.3) for reference. All chosen scenes include many similar objects (*e.g.*, chairs) with the same semantic. However, our discriminating feature spaces drive our method to select accurate object representations for editing without affecting adjacent similar instances.

Methods	Components			ScanNet 0192		Replica office4	
	SAHL	SGRR	IQML	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
(a)				26.84	0.360	31.52	0.187
(b)	✓			26.86	0.356	31.68	0.177
(c)	✓	✓		27.02	0.348	31.86	0.173
(d)	✓	✓	✓	26.96	0.350	31.76	0.182

Table 3. Ablation for proposed components. Though IQML causes a decrease in rendering capacity, we notice that IQML is necessary for leveraging 2D network-inferred labels.

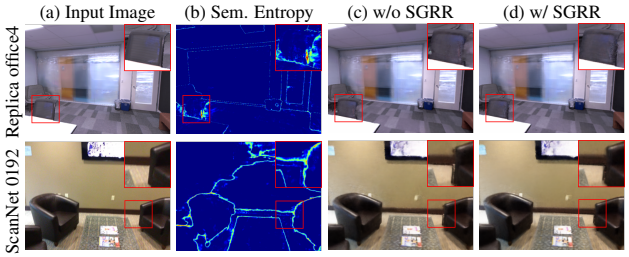

Figure 8. Qualitative comparison of proposed SGRR. With the guidance of semantic entropy maps, our method refines the rendering in semantic inaccurate regions such as object boundaries.

4.5. Ablation Study

We conduct ablation studies on two complicated scenes including Replica office4 and ScanNet 0192. These two scenes both contain multiple instances (chairs) with the same semantic, which can properly demonstrate the effectiveness of proposed components, especially the instance quadruplet metric learning.

Proposed Components. We analyze the effectiveness of proposed components including semantic-appearance hierarchical learning (SAHL), semantic-guided regional refinement (SGRR), and instance quadruplet metric learning (IQML) in Tab. 3. We first construct the semantic extension of NeRF as the baseline for comparison and the rendering performance is shown in Tab. 3(a). The effectiveness of SAHL and SGRR is shown in Tab. 3(b) and (c), which demonstrate our presented semantic-related strategies can improve the evaluation metrics on both scenes, indicating that our exploitation of the correlation between appearance and semantic attributes is reasonable. Furthermore, visualize results of employing SGRR in Fig. 8 shows that SDRR refines the rendering on semantic inaccurate regions under the guidance of entropy maps. Tab. 3(d) shows that IQML decreases the rendering performance due to the introduction of the additional constraint on feature space, especially on the high-fidelity scene such as Replica office4. However, we emphasize that IQML is the core component for leveraging 2D network-inferred labels and the slight rendering capacity degradation is acceptable. The necessity is further shown in Fig. 9, IQML discriminates the panoptic feature space at instance level and enables our method to remove the target chair without affecting others.

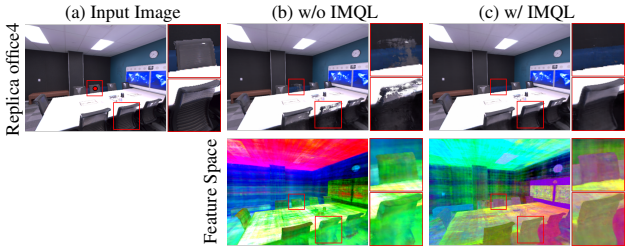

Figure 9. Qualitative comparison of proposed IQML. Constrained by IQML, the panoptic features are obviously grouped at instance level, which enables our method for scene editing.

Schemes	Margins		ScanNet 0192		Replica office4	
	η_h	η_e	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Triplet	0.5		26.89	0.355	31.65	0.191
Quadruplet	0.5	0.3	26.88	0.354	31.67	0.187
Quadruplet	0.3	0.5	**26.96**	**0.350**	**31.76**	**0.182**

Table 4. Ablation of the instance metric schemes. Compared with the triplet or the inverse quadruplet schemes, our quadruplet design achieves higher rendering performance.

Instance Quadruplet Loss. We construct two baselines for comparison to investigate the impact of the different instance metric schemes. The triplet baseline treats easy and hard negative samples equally in metric constraint and the inverse quadruplet baseline assigns a smaller margin δ_e for easy negatives. The ablative results in Tab. 4 show that our quadruplet constraint which assigns a smaller margin δ_h for hard negatives establishes reasonable relationships among samples and benefits the rendering performance. The qualitative results in Fig. 9 (b) show that features of pixels with the same semantic are naturally closer caused by the semantic supervision, validating that employing a less restrictive constraint for hard negatives is efficient.

5. Conclusion

We observe that existing object-compositional neural implicit representations are limited in real-world applications due to their requirement of manually-labeled ground truth instance annotations with 3D index consistency. To learn object-compositional representations with labels inferred by 2D panoptic segmentation networks, we propose a novel framework named panoptic compositional feature field (PCFF) for editable scene rendering by building a discriminating space of projected panoptic features supervised by the designed instance quadruplet metric learning. Besides, semantic-related strategies are proposed based on the correlation between appearance and semantic attributes to improve our rendering capacity. Experiments conducted on three scene datasets demonstrate our method achieves remarkable rendering performance and produces convincing scene edits with network-inferred labels.

References

[1] 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 (ICCV)*, pages 5855–5864, 2021. 1

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14124–14133, 2021. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[5] 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 (CVPR)*, 2022. 1, 3, 6, 7

[6] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3, 6, 7

[7] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1475–1484, 2022. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 2, 6

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2022. 2

[10] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314, 2021. 2

[11] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[12] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. 3

[13] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5511–5520, 2022. 1

[14] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. 3

[15] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18342–18352, 2022. 2

[16] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 2

[17] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12949–12958, 2021. 2

[18] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019. 1

[19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 6, 7

[20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019. 3

[21] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 6

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12871–12881, 2022. 3, 4

[23] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6498–6508, 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6

[25] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15651–15663, 2020. 2

[26] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5773–5783, 2021. 2

[27] Jiwen Lu, Junlin Hu, and Jie Zhou. Deep metric learning for visual understanding: An overview of recent advances. *IEEE Signal Processing Magazine (SPM)*, 34(6):76–84, 2017. 1

[28] 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 (CVPR)*, pages 4460–4470, 2019. 2

[29] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 1, 2, 3, 7

[30] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 3

[31] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5589–5599, 2021. 1

[32] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2856–2865, 2021. 3

[33] 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 (CVPR)*, pages 165–174, 2019. 2

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. volume 32, 2019. 6

[35] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 2

[36] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021. 2

[37] Shengyi Qian, Alexander Kirillov, Nikhila Ravi, Devendra Singh Chaplot, Justin Johnson, David F Fouhey, and Georgia Gkioxari. Recognizing scenes from novel viewpoints. *arXiv preprint arXiv:2112.01520*, 2021. 4

[38] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14335–14345, 2021. 2

[39] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 4

[40] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6

[41] 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 (CVPR)*, pages 5459–5469, 2022. 2

[42] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844, 2022. 2

[43] 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 *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[44] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2021. 2

[45] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[46] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6, 7

[47] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for ed-

itable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13779–13788, 2021. 1, 2, 3, 6, 7

[48] Alex Yu, Ruilong 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 (ICCV)*, pages 5752–5761, 2021. 2

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2

[50] Hong-Xing Yu, Leonidas Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[51] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[52] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labeling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15838–15847, 2021. 1, 3, 4, 6, 7