

# Unsupervised Discovery of Object Radiance Fields

Hong-Xing Yu   Leonidas J. Guibas   Jiajun Wu  
Stanford University

## Abstract

We study the problem of inferring an object-centric scene representation from a single image, aiming to derive a representation that explains the image formation process, captures the scene’s 3D nature, and is learned without supervision. Most existing methods on scene decomposition lack one or more of these characteristics, due to the fundamental challenge in integrating the complex 3D-to-2D image formation process into powerful inference schemes like deep networks. In this paper, we propose unsupervised discovery of Object Radiance Fields (uORF), integrating recent progresses in neural 3D scene representations and rendering with deep inference networks for unsupervised 3D scene decomposition. Trained on multi-view RGB images without annotations, uORF learns to decompose complex scenes with diverse, textured background from a single image. We show that uORF performs well on unsupervised 3D scene segmentation, novel view synthesis, and scene editing on three datasets.\*

## 1 Introduction

Building factorized, object-centric scene representations is a fundamental ability in human vision and a constant topic of interest in computer vision and machine learning. We identify that such representations should bear three characteristics: they should be learned without supervision or prior knowledge about object categories, and therefore applicable to environments where object categories are unknown; they should explain the image formation process, addressing questions like ‘what if the object is not there?’; they should be 3D-aware, capturing geometric and physical object properties for navigation, interaction, and manipulation.

For decades, researchers have attempted to solve the problems from various angles. Inspiring as they are, these methods each lack in one or more of the three aspects. Computer vision research on unsupervised object discovery has achieved great success on deriving object segments from real images, but it doesn’t capture the image formation process, nor is it 3D-aware [46, 65]. Recent work on deep probabilistic inference for visual scene decomposition is unsupervised and generative [2, 10, 15, 28, 30], though most still formulate the problem as 2D segmentation and work on simple scenes of geometric primitives, ignoring the complex 3D nature of realistic visual scenes. A few recent papers on ‘scene de-rendering’ have attempted to reconstruct 3D, object-centric representations by leveraging the forward rendering procedure [61, 41]; they are however supervised, relying on annotations of specific object and scene categories, such as cars and road scenes.

The fundamental challenge that prevents these systems from acquiring all three desired characteristics is that the image formation process from 3D to 2D is complex and non-differentiable (e.g., due to occlusion); thus, for a long time, it has been unclear how it may be integrated with powerful inference schemes, such as deep neural networks. But most recently, progresses in differentiable and neural rendering [55, 23] have demonstrated that their continuous nature works well with gradient-based inference models, capturing high-fidelity 3D scenes. In particular, Neural Radiance Fields (NeRFs) [35] recover a 3D scene from a set of RGB images via differentiable volume rendering. Such encouraging advances in generative modeling suggest a promising route for inferring 3D, generative, and object-centric scene representations without supervision.

\*Project website: <https://kovenyu.com/uorf/>

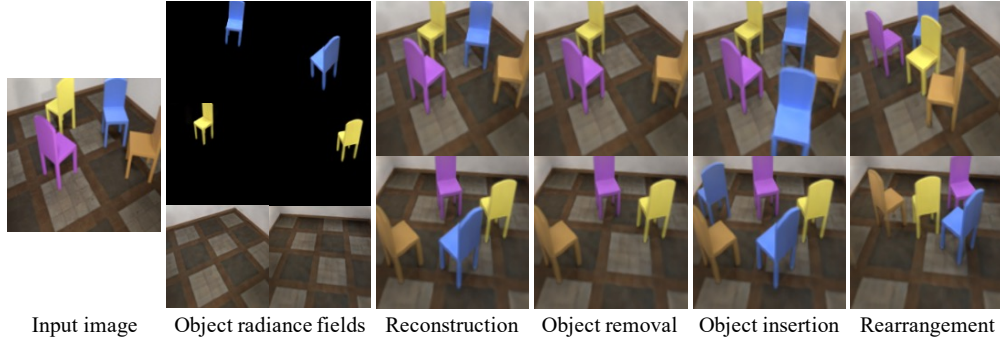


Figure 1: Illustration of unsupervised discovery of Object Radiance Fields. We aim to infer factorized object and background radiance fields from a single view.

In this paper, we propose unsupervised discovery of Object Radiance Fields (uORF), integrating conditional NeRFs as 3D object representations with deep inference networks for unsupervised 3D scene decomposition. uORF infers a set of object-centric latent codes through a slot-based encoder from a single image [30]. Each latent code is decoded into an object radiance field; thus, uORF represents a 3D scene as a composition of object radiance fields (Figure 1). During training, such radiance fields are neurally rendered in multiple views, with reconstruction losses in pixel space as training supervision; during testing, uORF infers the set of object radiance fields from a single image. Again, learning uORF does not require explicit supervision of 3D geometry or object segmentation, but only multi-view RGB images of training scenes.

The integration of NeRFs allows us to work with more realistic scenes with complex, diverse background environments, beyond simple scenes with the same textureless clean background, such as those in CLEVR [21] and multi-dSprites [15], as considered by most current unsupervised scene decomposition methods. We further make two innovations to improve uORF’s performance. First, as background geometry and appearance can be quite different from foreground objects, we design uORF with explicit modeling of both components. This background-aware design not only facilitates learning on complex scenes, but also allows single-image scene manipulation including moving individual objects and changing background. Second, as volume rendering requires massive queries to render a single pixel for the recomposed scene, a practical challenge of learning uORF lies in the computational inefficiency. We tackle this issue by proposing a novel progressive coarse-to-fine training which improves representation quality while remaining affordable computational cost.

We evaluate uORF on both scene representation learning (e.g., 3D segmentation) and scene generation (e.g., novel view synthesis, scene manipulation). Our evaluation is on three photo-realistic datasets with a gradually increasing complexity: first, CLEVR-like scenes with primitives foreground shapes; second, room scenes with complex chair shapes and textured backgrounds; third, more diverse room scenes with various foreground shapes and backgrounds. Our results show that uORF learns factorized representations that can segment 3D scenes into objects with fine shape details (e.g., thin chair legs) and backgrounds with well-recovered appearance details (e.g., irregular textures of a wooden floor). We also show that the learned representations allow 3D scene manipulation including moving objects and changing background appearances. We will release all code and data.

## 2 Related Work

**Co-segmentation and object discovery.** Our work is closely related to traditional computer vision methods on object discovery, which aims to locate (visually similar) objects in a collection of images. These methods typically model objects as visual words and adopted methods from topic modeling to localize objects [48, 51, 52], or cluster and group image patches [14, 22, 47, 57, 46, 7]. Recent works have integrated the clustering-based strategy with deep learning [27, 58]. Nevertheless, they do not explain image formation process nor are they 3D-aware.

**Unsupervised object-centric scene decomposition.** Our method is also closely related to recent work on deep probabilistic inference for scene decomposition. Most works formulate the problem as compositional generative models, in which a visual scene is represented by a set of latent codes that either correspond to localized object-centric patches [11, 8, 24, 28, 19] or scene mixture components [2, 15, 16, 17, 10]. The scene mixture models generate full-sized images for each latent code and blend them via attentional masks [2] in iterative variational inference frameworks. Recently, Locatello et

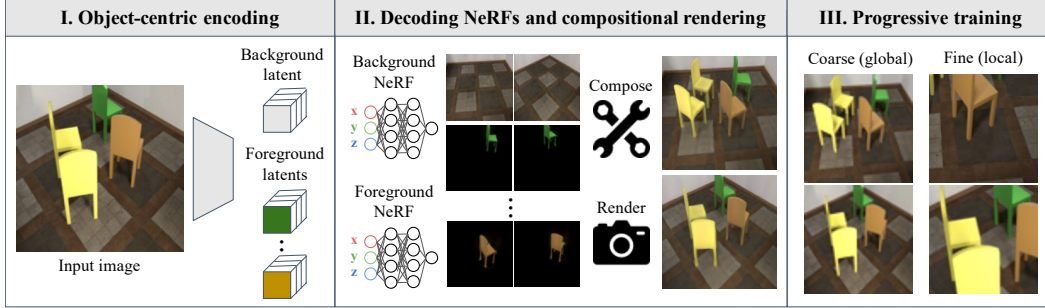


Figure 2: Overview of our model

al. [30] proposed the Slot Attention module to simplify the inference by a slot-based encoder. We adopt a similar slot-based encoder architecture [30], but ours explicitly models background environment to deal with complex scenes. Besides these inference models, Monnier et al. formulated scene decomposition as layered image decomposition and demonstrated it on real images [36]. However, these methods do not account for the 3D nature of scenes.

Very recently, a few works also focus on unsupervised 3D scene decomposition. Elich et al. [9] infer object shapes [42] from a single scene image, but they require pretraining on groundtruth shapes. Chen et al. [5] extend Generative Query Network [12] to decompose 3D scenes, but they require multi-view images during inference. The closest to our work is a concurrent work by Stelzner et al. [53] which also utilizes a slot-based encoder and NeRFs as 3D representations. However, [53] relies on groundtruth multi-view dense depth in addition to images in training. Moreover, we explicitly model the separation of objects and background to address various complex shapes and textured backgrounds, while they only demonstrate scenes with a single textureless background.

**Scene de-rendering.** A few recent works have shown reconstructing 3D object-centric representations by incorporating forward image rendering process [60, 61, 26, 41]. Yao et al. [61] de-render an image into semantic segments and geometric object attributes, which enable 3D scene manipulation. Most recently, Ost et al. propose Neural Scene Graph to represent dynamic scenes into a scene graph where each node encodes object-centric information. However, these methods rely on manual annotations of specific objects (such as cars) and scene categories (such as street scenes).

**Neural scene representations and rendering.** Our method is related to recent progresses in neural continuous scene representations [42, 34, 50] and neural rendering [55]. Neural scene representations parameterize 3D scenes with a deep network [50]. Combined with differentiable neural rendering techniques [23, 55], they can be learned from only 2D images [40, 50]. In particular, Neural Radiance Fields (NeRFs) [35] have shown impressive novel view synthesis from a set of densely captured images. Related follow-up works include those that infer NeRFs from a single image [63, 25, 45] and those that incorporate NeRFs into generative models [49, 39, 3]. Different from these works which cope with single objects or holistic scenes, we learn object NeRFs via decomposing a multi-object scene without segmentation annotations. Compositional generative modeling of 3D scenes is also related to our work [38, 18]. GIRAFFE [38] adversarially generates latent codes to condition object NeRFs and thus compose 3D scenes. While they target at compositional scene synthesis, we instead focus on object discovery (inference).

### 3 Approach

Our goal is to infer from a single image a set of object-centric 3D representations to recover the underlying scene. We show an illustration in Figure 2. We assume that an underlying 3D scene is composed of a background  $O_0$  and  $K$  foreground objects  $\{O_i\}_{i=1}^K$ , where we represent them by neural radiance fields [35] conditioned on latent codes. The latent codes  $\{z_i\}_{i=0}^K$  are inferred from an RGB image by a slot-based encoder (Figure 2-I). After being decoded, all foreground objects and background  $\{O_i\}_{i=0}^K$  can then be recomposed and re-rendered from arbitrary camera views (Figure 2-II). We train our model by comparing the re-rendered images to reference RGB images (Figure 2-III) without needing 3D geometry or segmentation annotations. We describe each of our model components in the following subsections, and leave implementation details in supplement.

#### 3.1 Object-centric Encoding

Our encoder infers latent object-centric representations from a single image. As shown in Figure 3, it consists of a convolutional net to extract features and a background-aware slot attention module to produce latent codes from the feature maps.

**Convolutional feature extraction.** The convolutional net extracts features from the input image for the slot attention module. Because we want the model to generalize to decompose unseen images, it is natural to represent foreground objects position and pose in the viewer coordinate system. As identified in previous studies [54], this facilitates the learning of 3D object position and helps generalization. In order for the object-centric representations to include such information in the viewer coordinate system, we inform the encoder of position information by feeding pixel coordinates and viewer-space ray directions as additional input channels.

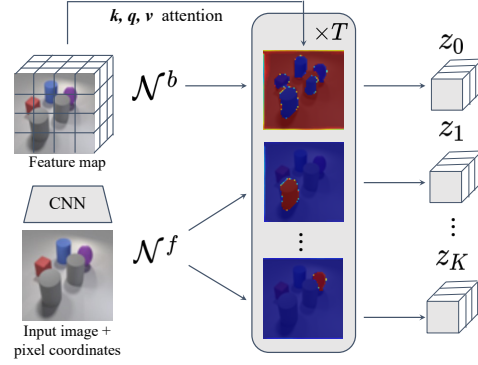


Figure 3: Our object-centric encoder.

**Background-aware slot attention.** Given the feature maps extracted from a convolutional network, we adopt the Slot Attention module [30] to produce a set of permutation-invariant latent codes in the same representational space. Each latent code binds to a specific group of the convolutional features to explain an object. However, in 3D scenes, the geometry and appearance of the background are usually highly different from those of foreground objects. Modeling them indistinguishably often leads to object representations entangled with blurry background segments [2, 30], which impedes applications such as scene manipulation and re-composition. Therefore, we propose modeling the separation of foreground objects and background explicitly.

To do this, we extend the slot-based encoder to allow a single slot to lie in a different latent space than the other slots to specialize for the background features. We show pseudo-code of our background-aware slot attention in the supplementary material. We also refer the readers to Locatello et.al. [30] for more details and insight of the slot attention module. In the following we describe a single iteration of our background-aware slot attention module.

We flatten convolutional feature maps into a set of  $N$  input feature vectors,  $\text{inputs} \in \mathbb{R}^{N \times D}$ . The latent representations (i.e., slots) are initialized by sampling from two learnable Gaussians, i.e.,  $\text{slot}^b \sim \mathcal{N}(\mu^b, \text{diag}(\sigma^b)) \in \mathbb{R}^{1 \times D}$  and  $\text{slots}^f \sim \mathcal{N}(\mu^f, \text{diag}(\sigma^f)) \in \mathbb{R}^{K \times D}$  for background and foreground objects, respectively. All slots are then competing to explain the inputs via a dot-product softmax-based attention [1, 31, 56]:

$$\text{attn}_{i,j} := \frac{\exp(M_{i,j})}{\sum_l \exp(M_{i,l})}, \quad \text{where } M := \frac{1}{\sqrt{D}} k(\text{inputs}) \cdot \begin{bmatrix} q^b(\text{slot}^b) \\ q^f(\text{slots}^f) \end{bmatrix}^T \in \mathbb{R}^{N \times (K+1)}. \quad (1)$$

Here  $k$  and  $q^b/q^f$  are learnable linear mappings  $\mathbb{R}^{D \rightarrow D}$  for computing dot-product similarity [31], and  $\sqrt{D}$  is a fixed softmax temperature [56]. The softmax normalization introduces competition among all slots for explaining the feature vectors by enforcing the attention coefficients for each input feature vector to add up to one. The background slot is expected to capture the modality of background features and explain all of them, allowing foreground slots to focus only on the objects without explaining background segments (Figure 3). With the attention coefficients, input values are aggregated via a weighted mean pooling updates<sup>b</sup> :=  $W^{bT} \cdot v^b(\text{inputs}) \in \mathbb{R}^{1 \times D}$ , where  $W_{i,1}^b := \text{attn}_{i,1} / (\sum_{l=1}^N \text{attn}_{l,1})$ , and updates<sup>f</sup> :=  $W^{fT} \cdot v^f(\text{inputs}) \in \mathbb{R}^{K \times D}$ , where  $W_{i,j}^f := \text{attn}_{i,j+1} / (\sum_{l=1}^N \text{attn}_{l,j+1})$ .

All slots are then updated using the aggregated values via a learnable updating rule parameterized by a Gated Recurrent Unit (GRU) [6]. Notice that the updating is applied independently for each slot with shared parameters (except for the background slot due to its different feature modality). The final latent codes  $\{z_i\}_{i=0}^K$  are the slots after being updated for  $T = 3$  iterations.

### 3.2 Compositional Neural Rendering

We use the latent codes  $\{z_i\}_{i=0}^K$  to condition neural radiance fields (NeRFs) [35] to represent the 3D objects. A NeRF is a continuous mapping  $g : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$  from spatial location  $\mathbf{x}$  and viewing direction  $\mathbf{d}$  to emitted color  $\mathbf{c}$  and volume density  $\sigma$  used for volume rendering [32]. This mapping

is parameterized by an MLP network. We use a conditional NeRF  $g(\cdot|\mathbf{z})$  that acts like an implicit decoder for each object. Specifically, we represent the background  $O_0$  by  $g^b(\cdot|\mathbf{z}_0)$  and the foreground objects  $\{O_i\}_{i=1}^K$  by another conditional NeRF network  $g^f(\cdot|\mathbf{z}_i)$ .

To compose individual objects and background into the holistic scene, we consider a scene mixture model and use density-weighted mean to combine all components:  $\bar{\sigma} = \sum_{i=0}^K w_i \sigma_i$ ,  $\bar{\mathbf{c}} = \sum_{i=0}^K w_i \mathbf{c}_i$ , where  $w_i = \sigma_i / \sum_{j=0}^K \sigma_j$ . Here  $\bar{\sigma}$  and  $\bar{\mathbf{c}}$  are the combined density and color, respectively. The color  $C(\mathbf{r})$  of a camera ray  $\mathbf{r}(t) = \mathbf{o} + \mathbf{d}(t)$  is then estimated via numerical integration of volume rendering, using  $S$  discrete combined samples along a ray [32]:  $C(\mathbf{r}) = \sum_{i=1}^S T_i [1 - \exp(-\bar{\sigma}_i \delta_i)] \bar{\mathbf{c}}_i$ , where  $T_i = \exp(-\sum_{j=1}^{i-1} \bar{\sigma}_j \delta_j)$ . Here  $\delta_j$  is the distance between adjacent samples along a ray.

### 3.3 Model Learning

**Loss functions.** During training, we render multiple views from a recomposed scene NeRF for supervision. Our training loss function comprises of a reconstruction loss, a perceptual loss, and an adversarial loss:  $\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{percept}} \mathcal{L}_{\text{percept}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}$ , where  $\lambda$  are weights. The reconstruction loss is  $\mathcal{L}_{\text{recon}} = \|\mathbf{I} - \hat{\mathbf{I}}\|^2$ , where  $\mathbf{I}$  and  $\hat{\mathbf{I}}$  denote the groundtruth image and rendered image, respectively.

Since we estimate 3D radiance fields from a single view, there can be uncertainties about the appearance from other views (e.g., the back view). For example, regarding visual appearance of objects, inaccurate global lighting estimation leads to uncertainties in brightness and shadows from occluded views even if the object shapes can be well estimated. To address this, we incorporate a perceptual loss [20] which is tolerant to mild appearance changes. The perceptual loss is defined by  $\|\mathcal{L}_{\text{percept}} = p(\mathbf{I}) - p(\hat{\mathbf{I}})\|^2$  where  $p$  is a deep feature extractor.

In addition to appearance, there can be even higher uncertainties in estimating object shapes from a single view, which is a multi-modal distribution. In this case, the unimodal reconstruction loss leads to blurry results (“mean shape”). We mitigate this issue by adding an adversarial loss which can deal with multi-modal distributions:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}[f(D(\hat{\mathbf{I}}))] + \mathbb{E}[f(-D(\mathbf{I})) + \lambda_R \|\nabla D(\mathbf{I})\|^2], \quad \text{where} \quad f(t) = -\log(1 + \exp(-t)). \quad (2)$$

Here we adopt the non-saturating loss with R1 regularization [33].

**Coarse-to-fine Progressive Training.** A practical challenge in training compositional NeRFs lies in the computational cost of neural volume rendering, as it requires massive queries to render a single pixel. While there have been attempts on fast inference [29, 43, 37, 13, 44, 62], high space complexity in training remains a challenge. Further, because our perceptual and adversarial losses depend on image patches, the system has to render a large enough patch (instead of a single pixel) at the same time, which further increases its space demand.

To allow training on a higher resolution, we propose a coarse-to-fine progressive training. In a coarse training stage, we bilinearly downsample image supervision to a base resolution, and train uORF on these downsampled images. Although the coarsely trained model can already decompose the 3D scenes and recover rough object radiance fields, fine details (e.g., thin legs of chairs) might be missing. Thus, in a following fine training stage, we refine our model by training on patches randomly cropped from images of the higher target resolution. Specifically, the fine training stage can be easily implemented by replacing the holistic downsampled images with patches of the same base resolution. We include more details in the supplementary material.

## 4 Experiments

We evaluate uORF on both scene representation (via 3D segmentation) and scene generation (via novel view synthesis and scene manipulation) on three photo-realistic datasets.

### 4.1 Data

We build three photo-realistic synthetic datasets with gradually increasing complexity. For each scene in the dataset, we point the camera to the scene center and render four images with a randomly chosen azimuth angle and a fixed elevation angle.

**CLEVR-567.** The first dataset includes scenes of 5–7 CLEVR objects [21], with a random position and orientation and a clean background. Foreground object shapes include three geometric primitives (i.e., cubes, spheres and cylinders). Since there is intrinsic ambiguity in estimating specularities from a



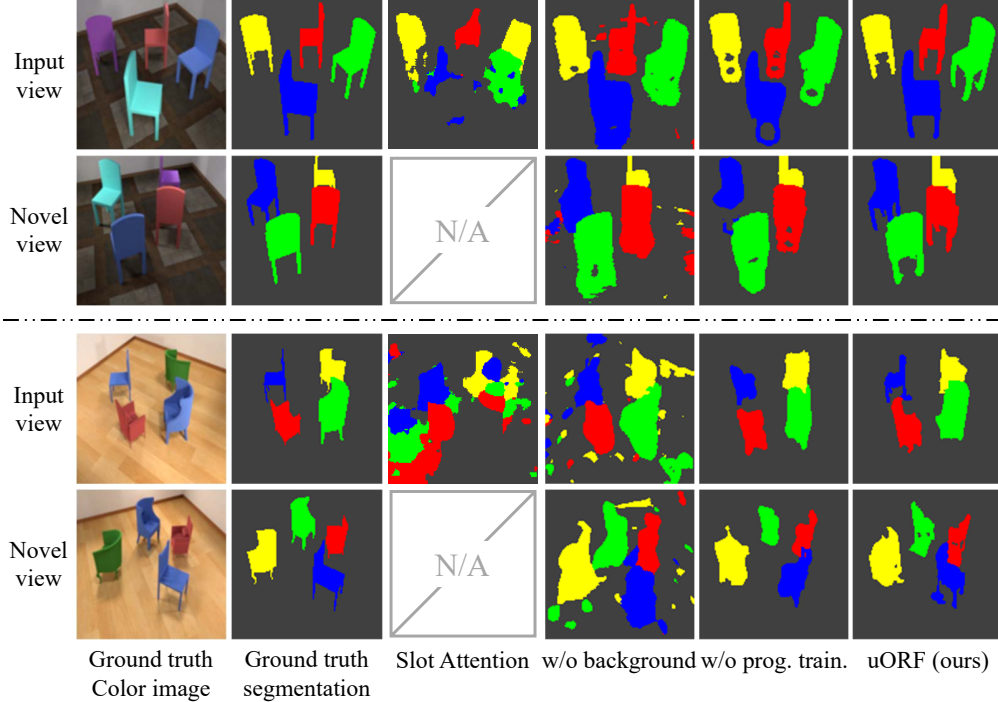


Figure 4: Examples on unsupervised 3D scene segmentation. Novel view images are for reference but not input.

single image, we use only the largely diffuse “Rubber” material. There are 1,000 scenes for training and 500 for testing.

**Room-Chair.** The second dataset includes scenes of 3 to 4 chairs of the same shape in a room with three different textured backgrounds. There are 1,000 scenes for training and 500 for testing.

**Room-Diverse.** The third dataset includes scenes of diverse foreground object shapes and background appearances. Each scene includes 4 different chairs, whose shape is randomly sampled from 1,200 ShapeNet chair shapes [4], and the background is sampled from 50 floor textures from the web. There are 5,000 scenes for training and 500 for testing.

## 4.2 3D Scene Segmentation

We first evaluate uORF’s factorized 3D scene representations via 3D scene segmentation.

**Baselines.** Because there is no previous work focusing on the same setting as uORF, we compare to a 2D state-of-the-art scene decomposition model Slot Attention [30] for unsupervised scene segmentation wherever possible. In addition, we compare to two ablated versions of uORF. First, we remove our background-aware modeling but keep the same number of slots. Second, we ablate our progressive training such that the training procedure only contains the coarse training stage. We refer to ablated models as “uORF (w/o background)” and “uORF (w/o prog. train.)”, respectively.

**Metrics.** We adopt the widely-used Adjusted Rand Index (ARI) as our metric. To evaluate 3D scene segmentation, we consider three kinds of ARIs: (1) For direct comparison to 2D methods, we compute ARI on reconstructed images. (2) To reflect the 3D nature, we also compute ARI on synthesized novel views, denoted as “NV-ARI”. Note that each scene includes 4 views, and only one is used as input, and the other three are treated as novel views for this metric. (3) In line with previous 2D methods, we also report foreground ARI (Fg-ARI), computed only on foreground regions indicated by groundtruth masks. Yet, we note that Fg-ARI is not as accurate as ARI to reflect the segmentation quality, because background segments assigned to foreground slots are treated correct.

**Results.** We show results on Table 1 and Figure 4. For all segmentation metrics, we show mean and standard deviation for three runs. uORF outperforms all methods in terms of ARI and NV-ARI. From Figure 4, it is clear that uORF is able to discover the 3D objects from a single image, and uORF can better depict the object outlines while entangling less background segments. These results validate that uORF can learn well-factorized 3D object-centric scene representations.

Models	CLEVR-567			Room-Chair			Room-Diverse		
	ARI $\uparrow$	NV-ARI $\uparrow$	Fg-ARI $\uparrow$	ARI $\uparrow$	NV-ARI $\uparrow$	Fg-ARI $\uparrow$	ARI $\uparrow$	NV-ARI $\uparrow$	Fg-ARI $\uparrow$
Slot Attention [30]	3.5 $\pm$ 0.7	-	<b>93.2</b> $\pm$ 1.5	38.4 $\pm$ 18.4	-	40.2 $\pm$ 4.5	17.4 $\pm$ 11.3	-	43.8 $\pm$ 11.7
uORF (w/o background)	11.7 $\pm$ 4.6	10.5 $\pm$ 3.6	86.4 $\pm$ 2.8	42.3 $\pm$ 10.6	40.4 $\pm$ 9.2	<b>93.3</b> $\pm$ 1.9	24.0 $\pm$ 9.9	21.0 $\pm$ 8.1	<b>78.9</b> $\pm$ 3.1
uORF (w/o prog. train.)	83.7 $\pm$ 0.8	81.1 $\pm$ 0.7	84.2 $\pm$ 0.5	65.4 $\pm$ 2.6	62.3 $\pm$ 2.5	81.0 $\pm$ 3.0	63.7 $\pm$ 1.7	53.8 $\pm$ 1.4	66.9 $\pm$ 4.1
uORF (ours)	<b>86.3</b> $\pm$ 0.1	<b>83.8</b> $\pm$ 0.3	87.4 $\pm$ 0.8	<b>78.8</b> $\pm$ 2.6	<b>74.3</b> $\pm$ 1.9	88.8 $\pm$ 2.7	<b>65.6</b> $\pm$ 1.0	<b>56.9</b> $\pm$ 0.2	67.9 $\pm$ 1.7

Table 1: Comparison on scene segmentation results. ‘‘Fg-ARI’’ refers to ARI evaluated with only foreground pixels. ‘‘NV-ARI’’ refers to ARI evaluated on novel views. Slot Attention [30] is a state-of-the-art 2D method.

Models	CLEVR-567			Room-Chair			Room-Diverse		
	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
NeRF-AE	0.1288	0.8658	27.16	0.1166	0.8265	28.13	0.2458	0.6688	24.80
uORF (w/o background)	0.0919	0.8924	28.93	0.1671	0.7852	27.86	0.2231	0.6924	25.90
uORF (w/o prog. train.)	0.1044	0.8894	28.84	0.1573	0.8287	28.33	0.2123	0.6760	25.19
uORF (ours)	<b>0.0859</b>	<b>0.8971</b>	<b>29.28</b>	<b>0.0821</b>	<b>0.8722</b>	<b>29.60</b>	<b>0.1729</b>	<b>0.7094</b>	<b>25.96</b>

Table 2: Comparison on novel view synthesis from a single image.

### 4.3 Novel View Synthesis

We then show that uORF is 3D-aware and generative via evaluation on novel view synthesis.

**Setup.** For each test scene, we randomly pick one image as input and the remaining three images as groundtruth for novel view synthesis. As Slot Attention is purely in 2D and does not support novel view synthesis, we compare to a variant of NeRF [35], equipped with an encoder similar to uORF, termed as ‘‘NeRF-AE’’. For fair comparison, we increase the latent bottleneck dimension for NeRF-AE to guarantee approximately the same computational cost, and we use the same training strategy and losses as uORF. Thus, NeRF-AE can also be seen as a monolithic alternative model to uORF. We also compare with the same ablated models, ‘‘uORF (w/o background)’’ and ‘‘uORF (w/o prog. train.)’’, as in scene segmentation. We use the perceptual metric LPIPS [64], together with SSIM [59] and PSNR, as our evaluation metrics.

**Results.** Quantitative results are in Table 2 and qualitative results are in Figure 5 (more qualitative results can be found in supplementary materials). Quantitatively, uORF outperforms all compared methods on all metrics. From the qualitative comparison in Figure 5, we highlight three advantages of uORF. First, compared with NeRF-AE, which has a monolithic latent structure for the entire scene, uORF better preserves the features of each object: for example, see how NeRF-AE fuses object colors in the first two rows, while uORF does not. This shows the advantage of factorized scene representations to structurally describe a visual scene. Second, compared with uORF (w/o background), one can clearly see how our background-aware modeling helps recovering background appearances: uORF can accurately recover background appearance of the Room-Chair example, while uORF (w/o background) does not. It also facilitates learning on complex scenes with diverse, textured background: uORF can learn to roughly recover object shapes in the Room-Diverse example. Third, compared with uORF (w/o prog. train.), we highlight that the fine training on image patches indeed improves both visual quality and representation quality: the full uORF tries to recover sharp edges of cubes, while uORF (w/o prog. train.) cannot distinguish cube from sphere.

Overall, the novel view synthesis results suggest that uORF can learn to represent 3D scenes with reasonable fidelity, even with the presence of complex foreground object shapes, such as chairs and different textured backgrounds.

### 4.4 Scene Design and Editing

Being object-centric and 3D-aware, uORF is able to edit 3D scene radiance fields inferred from a single view, and generate novel scene images.

**Setup.** We test uORF’s ability to edit scenes and synthesize novel images on the Room-Chair dataset. We consider both moving foreground objects and changing background appearance. For object moving, we randomly pick one object in a test scene and move it to a random position. We render 4 images for each of the 500 test scenes. For background changing, we replace the current background texture to a different one and also render 4 images for evaluation. To indicate the new background, we re-pick and re-put foreground objects such that the resultant background indicator image is different from the groundtruth image.

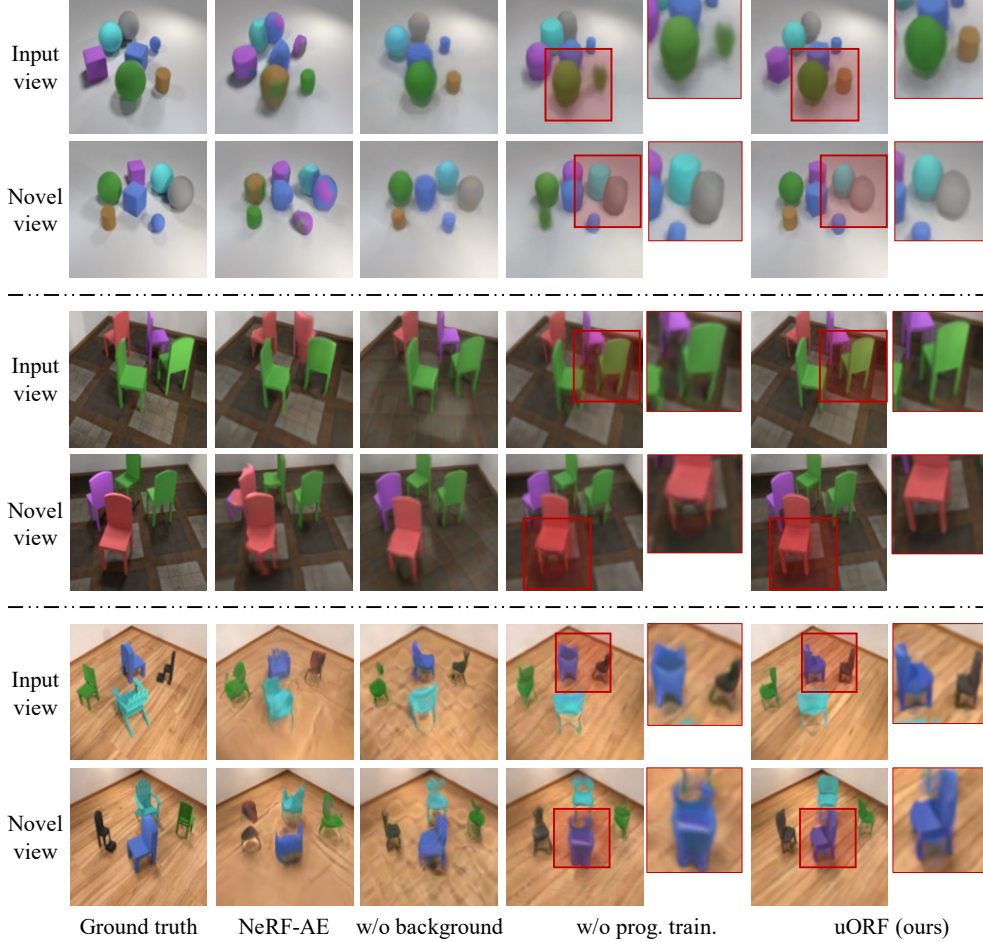


Figure 5: Qualitative results on scene decomposition and novel view synthesis. Within every two rows, the first is reconstruction and the second is a novel view.

For uORF and Slot Attention [30], we use groundtruth masks of the input view to determine which slot should be manipulated by picking the one with largest mask IoU. For NeRF-AE [35], we back-project the masks to frustums to determine the 3D regions to be moved/replaced. As in view synthesis, we use LPIPS, SSIM, and PSNR as our metrics.

**Results.** We show quantitative results in Table 3 and examples in Figure 6. Again, uORF outperforms all compared methods on all metrics. As Figure 6 depicts, images synthesized by uORF show least artifacts and highest quality and fidelity.

Models	Moving objects			Changing background		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
NeRF-AE	0.2451	0.7284	23.18	0.2185	0.7132	25.42
Slot Attention [30]	0.3941	0.7134	23.06	0.3689	0.7283	23.94
uORF (w/o background)	0.2206	0.7448	24.55	0.1879	0.7719	26.68
uORF (w/o prog. train.)	0.1583	0.8313	28.19	0.1586	0.8306	28.27
uORF (ours)	<b>0.0855</b>	<b>0.8711</b>	<b>29.26</b>	<b>0.0822</b>	<b>0.8729</b>	<b>29.53</b>

#### 4.5 Generalization and Analysis

Table 3: Comparison on scene manipulation.

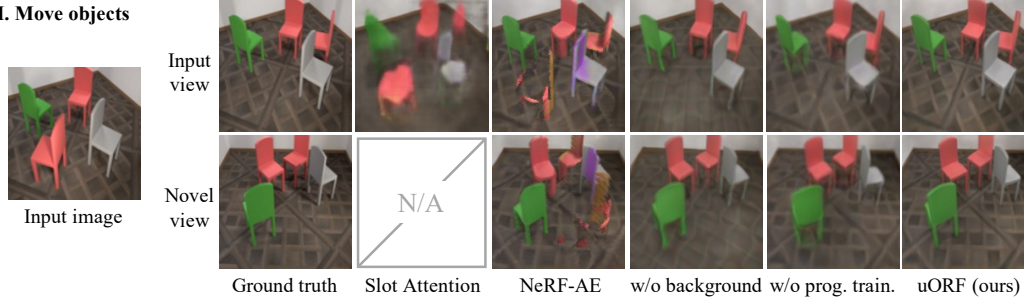
Finally we explore the generalization ability of uORF. We consider generalization on unseen, challenging spatial arrangement of objects, as well as generalization on unseen object appearances.

**Generalizing to challenging spatial arrangements.** We build a new test dataset, packed-CLEVR-11, where each scene has 11 objects that are closely packed into a cluster. Therefore, each scene bears an unseen number of objects in an unseen challenging arrangement. We test models trained on CLEVR-567, report results in Table 4 and the supplementary material. Despite uORF never sees such object arrangements, it still achieves a reasonable performance and outperforms baselines.

**Generalizing to new combination of shape and color.** For unseen object appearances, we consider generalization in a systematic way such that the model can deal with unseen combination of object color and shape. Thus, we build a new training set similar to CLEVR-567, but we remove red



### I. Move objects



### II. Change background

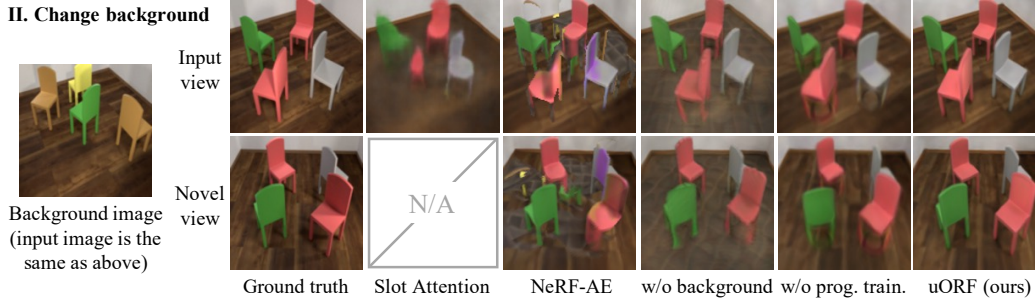


Figure 6: Qualitative results on single-image 3D scene manipulation. The first two rows are for moving object and the second two rows are for changing background.

Models	ARI $\uparrow$	LPIPS $\downarrow$	Models	ARI $\uparrow$	NV-ARI $\uparrow$	Loss functions	ARI $\uparrow$	LPIPS $\downarrow$
Slot Attention	5.7 $\pm$ 0.3	-	Slot Attention	2.2 $\pm$ 0.6	-	Rec.	59.1 $\pm$ 0.5	0.3610
NeRF-AE	-	0.2201	uORF (ours)	87.4 $\pm$ 0.4	85.0 $\pm$ 0.3	Rec. + Percept.	65.2 $\pm$ 0.8	0.2156
uORF (ours)	<b>83.2<math>\pm</math>0.6</b>	<b>0.1540</b>	uORF (oracle)	<b>87.5<math>\pm</math>0.3</b>	<b>85.5<math>\pm</math>0.3</b>	Rec. + Adv.	60.4 $\pm$ 2.2	0.2288
						Rec. + Percept. + Adv.	<b>65.6<math>\pm</math>1.0</b>	<b>0.1729</b>

Table 4: Generalization to novel combinations of color and shape.

Table 5: Generalization to unseen combinations of color and shape.

Table 6: Ablation study for losses on the Room-Diverse dataset.

cylinders and blue spheres from the object candidate pool. Then we test trained models on another dataset with only red cylinders and blue spheres in the candidate pool. We show results in Table 5 and examples in the supplement material. We see that although uORF has never seen any of the test set objects, it achieves similar results to the one trained on a normal CLEVR-567 dataset (denoted as “uORF (oracle)”). This suggests uORF’s ability for systematic generalization to unseen combinations of object color and shape.

**Evaluating loss functions.** uORF uses perceptual and adversarial losses to combat intrinsic uncertainties in single-image inference of 3D representations. We show ablation results on novel view synthesis in Table 6 and the supplementary material. Both losses significantly improve image quality.

## 5 Conclusion

In this work, we propose unsupervised discovery of Object Radiance Fields (uORF), which learns to infer object-centric 3D radiance fields from a single image of complex multi-object scenes. We demonstrate uORF’s ability on 3D scene segmentation and scene generation. Our positive results suggest a promising direction to integrate neural rendering into deep probabilistic inference scheme, allowing learning factorized 3D object-centric scene representations from only RGB images.

**Limitation and Broader Impact.** Learning object-centric scene representations is a long-standing topic in vision and it finds various applications in downstream tasks. We represent a 3D scene as a composition of simple radiance fields, which only models object appearances and entangles their physical properties that may be crucial to downstream tasks in a non-interpretable way. However, we envision that careful designs in more structured 3D object representations for specific downstream applications could help improve transparency and human interpretability in model prediction and behavior, allowing both better performances and secure, fair usage. In our code release, we will explicitly specify allowable uses of our system with appropriate licenses. We will use techniques such as watermarking to identify and label visual contents generated by our system.

**Acknowledgments.** This work is supported by an Amazon Research Award (ARA), the Samsung Global Research Outreach (GRO) Program, Toyota Research Institute, Autodesk, a Qualcomm Innovation Fellowship (QIF), and Stanford Institute for Human-Centered AI (HAI).

## References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 4
- [2] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 1, 2, 4
- [3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. *arXiv preprint arXiv:2012.00926*, 2020. 3
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6
- [5] Chang Chen, Fei Deng, and Sungjin Ahn. Learning to infer 3d object models from images. *arXiv preprint arXiv:2006.06130*, 2020. 3
- [6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4
- [7] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. 2
- [8] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2
- [9] Cathrin Elich, Martin R Oswald, Marc Pollefeys, and Jörg Stückler. Semi-supervised learning of multi-object 3d scene representations. *arXiv preprint arXiv:2010.04030*, 2020. 3
- [10] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019. 1, 2
- [11] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *arXiv preprint arXiv:1603.08575*, 2016. 2
- [12] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 2018. 3
- [13] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. <https://arxiv.org/abs/2103.10380>, 2021. 5
- [14] Kristen Grauman and Trevor Darrell. Unsupervised learning of categories from sets of partially matching image features. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006. 2
- [15] 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 *International Conference on Machine Learning*, 2019. 1, 2
- [16] Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hotloo Hao, Jürgen Schmidhuber, and Harri Valpola. Tagger: Deep unsupervised perceptual grouping. *arXiv preprint arXiv:1606.06724*, 2016. 2
- [17] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *arXiv preprint arXiv:1708.03498*, 2017. 2
- [18] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. 3
- [19] Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. Scalable object-oriented sequential generative models. *Unknown Journal*, 2019. 2
- [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 2016. 5

- [21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 2, 5
- [22] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 2
- [23] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020. 1, 3
- [24] Adam R Kosior, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: Generative modelling of moving objects. *arXiv preprint arXiv:1806.01794*, 2018. 2
- [25] Adam R Kosior, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J Rezende. Nerf-vae: A geometry aware 3d scene generative model. *arXiv preprint arXiv:2104.00587*, 2021. 3
- [26] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3
- [27] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8519–8528, 2019. 2
- [28] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020. 1, 2
- [29] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5
- [30] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020. 1, 2, 3, 4, 6, 7, 8
- [31] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 4
- [32] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1995. 4, 5
- [33] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 5
- [34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3
- [35] 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. *arXiv preprint arXiv:2003.08934*, 2020. 1, 3, 4, 7, 8
- [36] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised layered image decomposition into object prototypes. *arXiv preprint arXiv:2104.14575*, 2021. 3
- [37] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Chakravarty R. Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of neural radiance fields using depth oracle networks. *arXiv preprint arXiv: 2103.03231*, 2021. 5
- [38] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. *arXiv preprint arXiv:2011.12100*, 2020. 3
- [39] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. *arXiv preprint arXiv:2103.17269*, 2021. 3
- [40] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3
- [41] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 1, 3
- [42] 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 Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

- [43] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. DeRF: Decomposed radiance fields. <https://arxiv.org/abs/2011.12490>, 2020. 5
- [44] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *arXiv preprint arXiv: 2103.13744*, 2021. 5
- [45] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021. 3
- [46] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013. 1, 2
- [47] Jose C Rubio, Joan Serrat, Antonio López, and Nikos Paragios. Unsupervised co-segmentation through region matching. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2
- [48] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006. 2
- [49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 3
- [50] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1121–1132, 2019. 3
- [51] Josef Sivic, Bryan C Russell, Alexei A Efros, Andrew Zisserman, and William T Freeman. Discovering objects and their location in images. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005. 2
- [52] Josef Sivic, Bryan C Russell, Andrew Zisserman, William T Freeman, and Alexei A Efros. Unsupervised discovery of visual object class hierarchies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2
- [53] Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv preprint arXiv:2104.01148*, 2021. 3
- [54] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 4
- [55] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *Computer Graphics Forum*, 39(2):701–727, 2020. 1, 3
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4
- [57] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR 2011*, 2011. 2
- [58] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. 2
- [59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7
- [60] Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3
- [61] Shunyu Yao, Tzu Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, William T Freeman, and Joshua B Tenenbaum. 3d-aware scene manipulation via inverse graphics. *arXiv preprint arXiv:1808.09351*, 2018. 1, 3
- [62] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. *arXiv preprint arXiv:2103.14024*, 2021. 5
- [63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *arXiv preprint arXiv:2012.02190*, 2020. 3
- [64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7
- [65] Jun-Yan Zhu, Jiajun Wu, Yichen Wei, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1