Open-NeRF: Towards Open Vocabulary NeRF Decomposition

Hao Zhang, Fang Li, Narendra Ahuja
University of Illinois Urbana-Champaign
{haoz19, fangli3, n-ahuja}@illinois.edu

Abstract

In this paper, we address the challenge of decomposing Neural Radiance Fields (NeRF) into objects from an open vocabulary, a critical task for object manipulation in 3D reconstruction and view synthesis. Current techniques for NeRF decomposition involve a trade-off between the flexibility of processing open-vocabulary queries and the accuracy of 3D segmentation. We present, Open-vocabulary Embedded Neural Radiance Fields (Open-NeRF), that leverage large-scale, off-the-shelf, segmentation models like the Segment Anything Model (SAM) and introduce an integrate-and-distill paradigm with hierarchical embeddings to achieve both the flexibility of open-vocabulary querying and 3D segmentation accuracy. Open-NeRF first utilizes large-scale foundation models to generate hierarchical 2D mask proposals from varying viewpoints. These proposals are then aligned via tracking approaches and integrated within the 3D space and subsequently distilled into the 3D field. This process ensures consistent recognition and granularity of objects from different viewpoints, even in challenging scenarios involving occlusion and indistinct features. Our experimental results show that the proposed Open-NeRF[1] outperforms state-of-the-art methods such as LERF [16] and FFD [18] in open-vocabulary scenarios. Open-NeRF offers a promising solution to NeRF decomposition, guided by open-vocabulary queries, enabling novel applications in robotics and vision-language interaction in open-world 3D scenes. Please find the code at https://github.com/haoz19/Open-NeRF.

1. Introduction

Neural Radiance Fields (NeRFs) show great promise for high-quality 3D reconstruction and novel view synthesis from 2D image observations taken from various viewpoints (camera positions and viewing angles). However, to manipulate an object in the field, such as extracting, removing, or altering its color or texture, it is essential to first decompose the 3D field. Some methods [14, 37, 44] introduce an extra field component to learn semantic information or unique codes for all individual objects in 3D space from 2D supervision. However, these methods require annotations of ground-truth 2D masks for supervision, thus introducing additional a cost for labeling. Other methods, such as FFD [18] and N3F [36], address the issue using pre-trained image feature extractors like openclip-LSeg and DINO to distill 2D image features into the 3D field. However, as shown in Figure 2 (a), they are limited to closed-set scenarios, containing objects from predefined classes, e.g., in the COCO [21] dataset. Some methods, like LERF [16], ground language embeddings from off-the-shelf models like openclip into NeRF to locate open-vocabulary queries within 3D scenes. However, these methods lack accurate 3D segmentation, limiting their practical application.

Fortunately, with the emergence of large-scale off-the-shelf segmentation models such as the Segment Anything Model (SAM) [17], obtaining 2D mask proposals in open-world images has become possible. Intuitively, grounding SAM into the NeRF may help achieve both flexible open-vocabulary query processing and decent 3D segmentation accuracy. However, directly distilling the 2D knowledge obtained from the mask proposals generated by SAM into the 3D field results in undesirable outcomes due to a lack of consistency across different viewpoints. As shown in Figure 3, (a) the recognition confidence varies with camera parameters, because some objects may be difficult to identify from certain viewing angles due to a lack of distinctive features or occlusion; (b) the granularity of masks proposals of the same object for different camera parameters is not consistent, because SAM may generate a varying number of masks proposals for the same objects in different views.

In this paper, we present a novel approach, named Open-vocabulary Embedded Neural Radiance Fields (Open-NeRFs) that introduces an integrate-and-distill paradigm. It leverages hierarchical embeddings to address issues arising from direct distilling of 2D knowledge from SAM, for flexible real-time handling of various types of open-vocabulary queries while also providing high-quality 3D segmentation

[1] The support of the Office of Naval Research under grant N00014-20-1-2444 and of USDA National Institute of Food and Agriculture under grant 2020-67021-32799/1024178 are gratefully acknowledged.

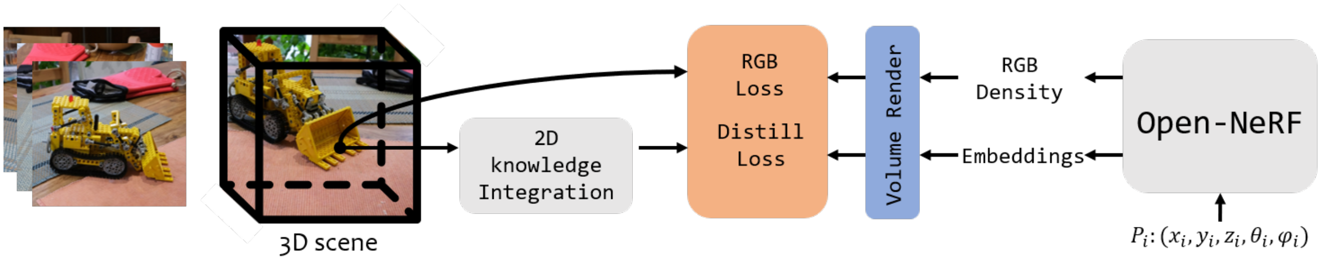


Figure 1. Overview of our Open-NeRF. We follow the integrate-and-distill paradigm, by first integrating the 2D knowledge and then distilling it into the 3D fields.

results. Instead of distilling the 2D semantic information or image embeddings from off-the-shelf models for different viewpoints to the 3D field independently, we use the following procedure (Figure 2): (1) Leverage off-the-shelf models such as SAM to generate hierarchical 2D mask proposals for images taken from multiple viewpoints; (2) Align every mask in the 2D images with its corresponding object in the 3D space; (3) Integrate the knowledge of all 2D mask proposals for the same object across all viewpoints; (4) Distill the integrated hierarchical knowledge to the 3D field. Through this integrate-and-distill approach, we enable recognition and granularity consistency of objects from different perspectives, even when they are partially occluded or are otherwise difficult to recognize. To more robustly assess the performance of Open-NeRF in open-vocabulary scenarios, we incorporate an additional set of ten real-world scenes for evaluation. Experimental results demonstrate the effectiveness of our proposed approach, surpassing state-of-the-art performance in challenging open-vocabulary scenarios.

2. Related Work

2.1. Implicit Neural Representation

In novel view synthesis, the recent introduction of NeRF [26] model releases the model from the problems of low-resolution geometry and photo-unrealistic rendering of novel views. Instead of explicit representations, NeRF uses an implicit 5D radiance field to represent a scene, which can produce more detailed realistic novel view renderings. Several NeRF-based methods [25, 40] also free the NeRF model from camera parameters. In addition, some methods [9, 15, 28] also equip NeRF with the capacity to render novel views in dynamic scenes. With implicit density and color prediction of samples on the rays cast from each camera center through the collected photos by volume rendering, NeRF-based models have been the most convincing ones for novel view synthesis. Several works [14, 37, 44] try to decompose NeRF field with its implicit representation of scene objects. Although exceptional performance is obtained on some synthetic and real indoor scene datasets, the high cost of 2D ground truth mask annotations makes them less useful in real-world scenarios.

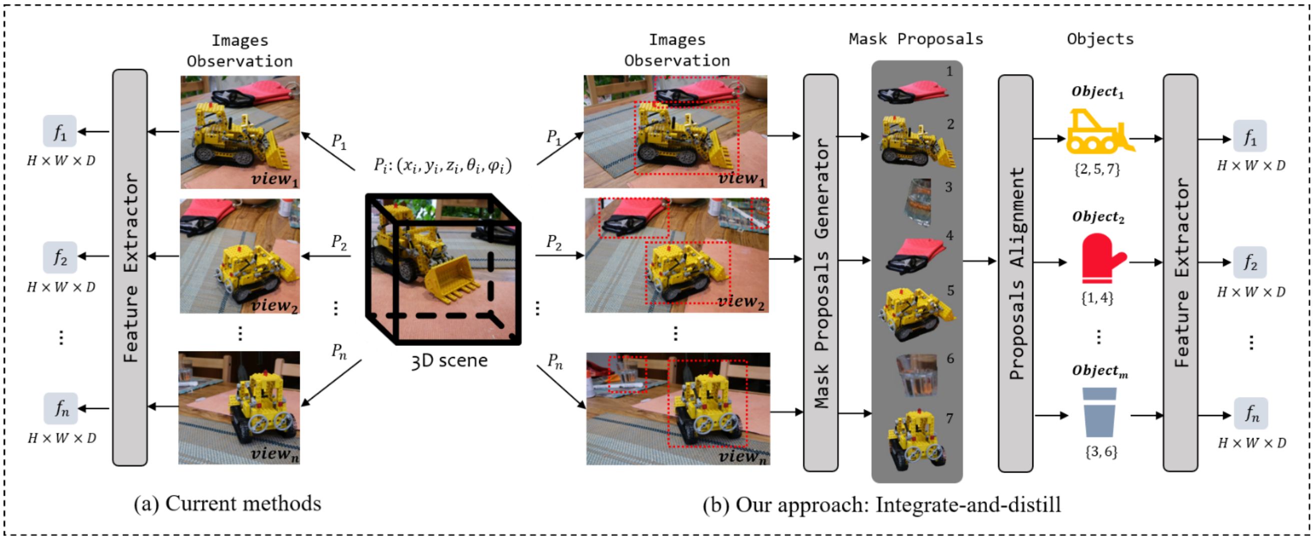
2.2. Open-Vocabulary 2D Segmentation

To obtain accurate 2D segmentation masks, traditional methods [4, 5, 8, 12, 20] achieve excellent results on annotated closed-set public datasets like COCO [21] and ADE20K [45], their performance on Out-Of-Distribution (OOD) objects is much poorer. Some open vocabulary segmentation models [2, 10, 19, 27, 41] inspired by open-clip [33] use multi-modal vision-and-language methods to segment OOD objects. They rely on calculating the pixel-wise similarity between the image embeddings and the language embeddings, obtained by image encoders [7, 29] and language encoders [33]. Although some works [**?**, 39, 42] make progress in the simple scenarios, the under-expected predictions in complex in-the-wild scenes still block their way to real-world applications

Recently, SAM [17], trained on large-scale datasets, has been shown to provide a good solution. SAM uses a transformer-based image encoder for feature extraction, a prompt encoder for query tokenization, and a mask decoder to output the segmentation results. [3, 13, 23, 31] show that a pre-trained SAM model can benefit several downstream tasks in diverse applications such as medical image segmentation and 3D decomposition. [11, 24, 30] point out that the model results are unsatisfactory on other tasks such as segmenting images of glasses and those taken from space.

2.3. 2D Features Field Distillation into NeRF.

Several works [18, 34, 36, 44] distill the knowledge of off-the-shelf, supervised, and self-supervised 2D image feature extractors into a 3D feature field. Specifically, FFD [18] and N3F [36] distill the pixel-level embedding vectors from DINO [2]and LSeg [19] into mplicit neural fields and could localize the corresponding objects matched with the given queries such like point-and-click selections, images patches, and texts. Besides, although these methods provide reasonable results, they support vocabulary queries from only predefined classes, from ADE20K citeade20k or



(a) Current methods

(b) Our approach: Integrate-and-distill

Figure 2. Comparison of existing methods and our proposed approach. (a) presents the methodology employed by existing methods like FFD [18] and N3F [36], which involve direct extraction of per-pixel image features using off-the-shelf feature extractors such as LSeg [19]. These methods then distill the extracted features into 3D fields. However, these approaches often yield subpar results in certain viewpoints due to occlusions or the absence of discriminative features. (b) showcases our approach, which follows an integrate-and-distill paradigm. Initially, we generate region proposals (bounding boxes) for all objects in the 3D scene. These proposals are then input into the Segment Anything Model (SAM) to generate corresponding mask proposals and align them for the same object. Finally, we fuse all the mask proposals pertaining to the same object and extract integrated embeddings.

COCO [21] datasets. Recently, Language Embedded Radiance Fields (LERF) [16] have been shown to perform better on a broad range of open-vocabulary queries, including combinations of concepts, colors, long-tail words, and text. However, LERF can provide only an approximate location of an object corresponding to a given vocabulary query; it cannot guarantee accurate segmentation. The current methods struggle to manage open-vocabulary NeRF decomposition well because they exhibit a trade-off between query processing flexibility and segmentation accuracy. Some methods prioritize accuracy at the expense of processing flexibility, making it difficult to handle open-vocabulary queries, while others can process open-vocabulary queries, but at the cost of generating imprecise segmentation results.

3. Preliminaries

3.1. Neural Radiance Fields

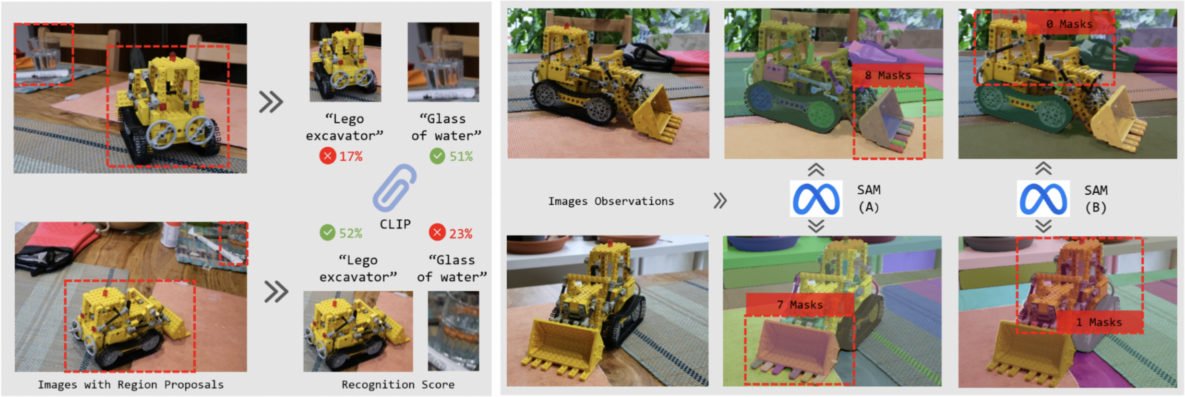
NeRF [26] utilizes Multilayer Perceptrons (MLPs) to estimate continuous 3D scene geometry and appearance, given a set of images acquired using given camera parameters. It operates with the input of 5D vectors, consisting of point coordinates and viewing directions, denoted by $p = (x, y, z, \theta, \phi)$, and forecasting volume density δ and color $c = (r, g, b)$ for point, p. This scene representation can be visually rendered and optimized using volume rendering techniques. For a pixel's camera ray, defined as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, with a depth t within bounds $[t_{near}, t_{far}]$, the camera position $\mathbf{o}$, and viewing direction $\mathbf{d}$, NeRF computes the color of a ray by rendering N sampled points $\{\mathbf{x}_n\}_{n=1}^{N}$ with respective depths $\{t_n\}_{n=1}^{N}$ as

$$\begin{aligned} \hat{\mathbf{C}}(\mathbf{r}) &= \sum_{n=1}^{N} \hat{T}(t_n)\, \alpha\left(\sigma(\mathbf{x}_n)\, \delta_n\right) \mathbf{c}(\mathbf{x}_n, \mathbf{d}), \\ \hat{T}(t_n) &= \exp\left(-\sum_{n'=1}^{n-1} \sigma(\mathbf{x}_{n'})\, \delta_{n'}\right), \end{aligned} \tag{1}$$

where $\alpha(x) = 1 - \exp(-x)$, and $\delta_n = t_{n+1} - t_n$ represents the distance between successive point samples. NeRFs are optimized exclusively on an image dataset and their respective camera parameters by minimizing rendering loss.

3.2. Segment Anything Model

As a large ViT-based model trained on an extensive visual corpus (SA-1B), SAM [17] exhibits impressive segmentation capabilities across diverse scenarios. There are two primary approaches to harness the capabilities of SAM: (1) Guided by prompts, and (2) Through automatic generation of mask proposals. Both begin by feeding input images into the image encoder. In the first approach, prompts, which may consist of points, a box, a mask, or text, are fed into the prompt encoder. The mask encoder receives the outputs from both the image and the prompt encoders as inputs and generates mask proposals corresponding to these prompts. In contrast, the second approach enables SAM to



(a) Recognition Inconsistency

(b) Granularity Inconsistency

Figure 3. Viewpoint Inconsistency Analysis. (a) Recognition inconsistency of openclip across different viewpoints. The back view of the Lego excavator presents a greater challenge for identification compared to the side view. Additionally, occlusion makes it difficult to identify the glass of water from certain viewpoints. (b) Granularity inconsistency of mask proposals generated by the Segment Anything Model (SAM) across viewpoints. Regions with different colors overlaid on the original image represent the corresponding mask proposals, while regions without colors indicate the absence of mask proposals. Under setting (A), SAM adheres to the default configuration, whereas under setting (B), the hyperparameter `points_per_side` is decreased to 8. SAM currently lacks the capability to automatically generate a region proposal that accurately encompasses the entire Lego excavator (complex objects) from certain viewpoints. Additionally, SAM may generate different numbers of mask proposals for the same parts in different viewpoints, resulting in granularity inconsistency.

generate mask proposals without any user prompts. This is accomplished by automatically generating a set of evenly distributed points at a user-specified interval within the images, which are then used as prompts.

Challenges in Utilizing SAM for Open-Vocabulary NeRF Decomposition. In the realm of open-vocabulary NeRF decomposition, the sole input typically comprises a collection of images alongside corresponding camera parameters. Remarkably, even in the absence of explicit camera parameters, methods like COLMAP [32] can readily extract this information from the input images. With such data at our disposal, the ideal model should not only be capable of synthesizing novel views but also possess the ability to decompose any objects within the scene through open vocabulary prompts. However, during the training phase, user prompts for SAM are not readily available, necessitating the employment of automatic methods. This strategy, unfortunately, presents its own set of challenges. As demonstrated in Figure 3 (b), SAM fails to generate mask proposals as anticipated. Moreover, the granularity of mask proposals across different viewpoints lacks consistency. These issues collectively suggest that directly distilling the knowledge from SAM to 3D fields may not be an optimal solution, underlining the inherent challenges of using SAM for open-vocabulary NeRF decomposition.

4. Open-vocabulary Embedded Neural Radiance Fields

In this section, we present our proposed method, Open-vocabulary Embedded Neural Radiance Fields (Open-NeRFs), which addresses the trade-offs in existing methods. Open-NeRFs can flexibly handle open vocabulary queries while generating precise segmentation results for specified objects in NeRF. First, we provide a detailed introduction to our proposed **integrate-and-distill paradigm**, which ensures recognition and granularity consistency across all viewpoints. Then, we describe the **hierarchical feature fusion** technique, which enables handling queries of different scales while preventing the loss of local detail. In addition, we explain how to perform **open-vocabulary querying** using Open-NeRF.

4.1. Integrate-And-Distill Paradigm

As shown in Figure 2, our approach follows the integrate-and-distill paradigm, which contains 3 main steps: (1) 2D knowledge extraction, (2) 2D knowledge integration across all viewpoints, and (3) integrated knowledge distillation to 3D fields.

4.1.1 Multi-view Knowledge Integration

Given a collection of images $\mathbf{I} = \{i_n\}_{n=1}^{N}$ with corresponding camera parameters, we first leverage the region proposal generator to produce K_n, $n \in \{1, ..., N\}$, region proposals $\mathbf{R} = \{r_k\}_{k=1}^{K_n}$, *i.e.*, bounding boxes, for image i_n. Then we utilize $\mathbf{R}$ as prompts for SAM to get K_n mask proposals $\mathbf{M} = \{m_k\}_{k=1}^{K_n}$. After that, we crop images along those mask proposals to get K_n sub-images i_k^s for i_n as shown in Figure 2 and obtain $\sum_{n=1}^{N} K_n$ sub-images from all N images. Noted that K_n varies between images because some objects are not visible in some viewpoints. Then we align all the sub-images of the same object from different viewpoints with the help of tracking models. For each object $\mathbf{O}_j$, there exists a set of sub-images $\mathbf{I}_j^s = \{i_{j,1}^s, ..., i_{j,L}^s\}$, where j is the index of objects within the 3D scene and L is the frequency of $\mathbf{O}_j$ being included in the images. We then feed $\mathbf{I}_j^s$ to the image encoder of openclip [33] and a norm operator to get a set of normalized image embeddings $E_j = \{e_{j,1}, ..., e_{j,L}\}$ for each object and integrate the knowledge from them by fusing E_j. One simple way of fusing is averaging all image embeddings from the same object: $E_j^f = \frac{1}{L}\sum_{l=1}^{L} e_{j,l}$. But this cannot ensure that the norm of E_j^f is 1, which will introduce errors when calculating similarity with the normalized text embeddings in the querying step. Therefore, we obtain fused embeddings by the following equation: $E_j^f = \sum_{l=1}^{L} e_{j,l}/|\sum_{l=1}^{L} e_{j,l}|$. With the per-object fused embeddings E_j^f and the mask proposals m_j for each object across images, we obtain per-pixel image embeddings $\mathbf{E} \in \mathbb{R}^{N \times H \times W}$ by assigning all pixels from the same mask proposal m_j the corresponding E_j^f, where (H, W) denote (height, width) of N input images.

4.1.2 Proposals Alignment

For aligning the mask proposals with different views, we leverage the tracking model to track each object and region of the background and provide a per-pixel prediction of their IDs: $\mathbf{I} \in \mathbb{R}^{H \times W}$ for every frame. Because we demonstrate that the segmentation results from the tracking model are not reliable across frames, we count the ID of each pixel in the mask proposal $\mathbf{M}$ and choose the one with the highest frequency within the entire mask proposal instead of directly using the segmentation results from the tracking model as mask proposals. Then the ID predictions are used to align the same object from different view points.

4.1.3 Knowledge Integration through Open-Vocabulary Embedding Field

NeRF-based models employ neural radiance fields to determine view-independent volume density $\delta(x)$ and view-dependent color $c(x, d)$. Some variants of NeRF further expand upon this by incorporating an auxiliary decoder designed to predict additional properties of interest. For instance, Semantic-NeRF [44] utilizes an extra branch to estimate the probability distribution of closed-set semantic labels, DM-NeRF [37] learns a unique code for each object within predefined classes, and FFD [18] introduces a feature branch to generate a feature vector corresponding to the 3D coordinates.

Drawing upon these advancements, we propose an additional branch which we call the ”open-vocabulary embedding field”, designed to facilitate learning of open-vocabulary embeddings. During the multi-view knowledge integration phase, we produce per-pixel image embeddings $\mathbf{E}$ that encapsulate the integrated 2D knowledge aggregated from multiple perspectives. This knowledge is subsequently distilled into the open field. For any given 3D coordinate $\mathbf{x}_n$, where $n \in \{1, ..., N\}$ and N indicates the number of points on the ray $\mathbf{r}$, the open field yields a field embedding $\mathbf{e}(\mathbf{x_n})$, as illustrated in Figure 1. Analogous to volume rendering in NeRF as depicted in Eq.1, we extract the predicted per-pixel field embeddings through volume rendering along a set of rays. For each individual ray $\mathbf{r}$, the computation is performed as follows:

$$\hat{\mathbf{E}}(\mathbf{r}) = \sum_{n=1}^{N} \hat{T}\left(t_n\right) \alpha\left(\sigma\left(\mathbf{x}_n\right)\delta_n\right) \mathbf{e}\left(\mathbf{x}_n\right), \qquad (2)$$

Note that in contrast to the color function $\mathbf{c}$, the function $\mathbf{e}$, akin to the density function σ, is direction-independent, thereby ensuring that the semantics of a specific object remains invariant to viewing point changes.

We optimize $\mathbf{e}$ by minimizing the discrepancy between the rendered embedding $\hat{\mathbf{E}}(\mathbf{r})$ and the corresponding per-pixel image embeddings $\mathbf{E}[h, w]$, where h, w denote the pixel's position correlating to the ray $\mathbf{r}$. Consequently, the composite loss integrates both the photometric loss and the distillation loss: $L = L_p + \lambda L_e$, where, L_p signifies the Mean Squared Error (MSE) loss between the ground truth color and the rendered color, whereas L_e denotes the mean Huber loss between $\mathbf{E}[h, w]$ and $\hat{\mathbf{E}}(\mathbf{r})$. By default, λ is set to 0.1. To circumvent the potential disruption caused by external embedding branches to the original NeRF, we ensure the embedding field operates independently from the original NeRF, thereby avoiding mutual interference. As a result, L_e is solely employed for optimizing the embedding field.

4.1.4 Hierarchical Embedding

We present a novel approach Open-NeRF that capitalizes on hierarchical embeddings, enabling the decomposition of the 3D field in a hierarchical manner based on the scale of input queries. Our method involves the extraction of image embeddings at three distinct levels: (1) object level, (2) part level, and (3) background level. For the object and part level, we employ region proposal generation techniques at different granularity to identify and propose regions corresponding to complete objects. Following the methodology outlined in Sec 4.1, we extract features for each identified object or object part and embed them into the 3D field. During inference, Open-NeRF provides 2 relevancy predictions for object level and part level respectively, and automatically selects one following the rule: select the part level predictions only when there are at least N pixels with a

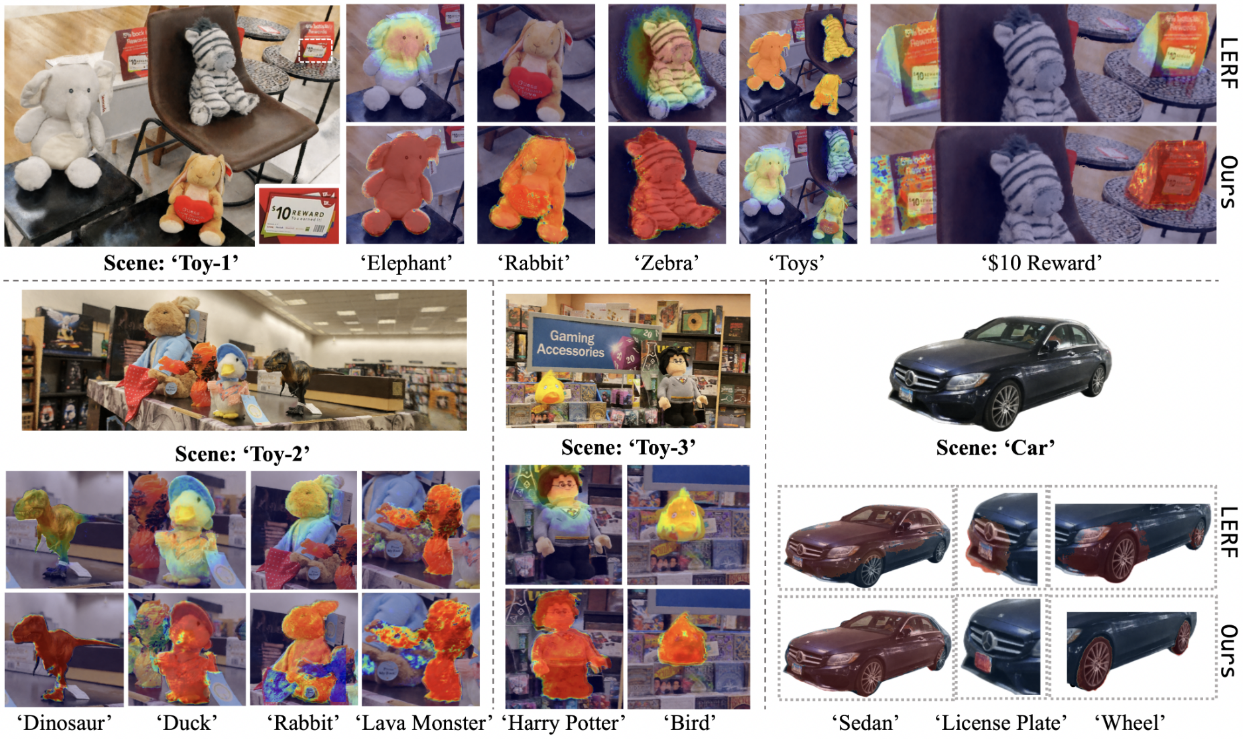


Figure 4. Relevance scores obtained by Open-NeRF and LERF in the scenes: *'Toy-1,2,3'* and *'Car'*.

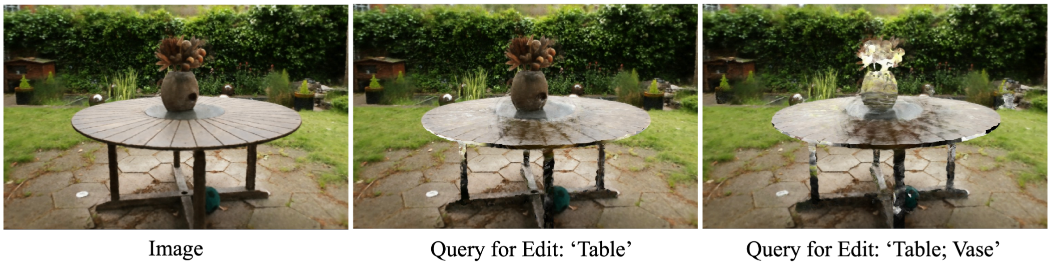
higher relevancy score than the max relevancy score of object level, and we set $N = 100$ by default. However, we encounter challenges when processing background regions, such as *grass, road,* and *sky*, as they typically encompass a substantial portion of the entire image. Consequently, conventional region proposal generators struggle to effectively handle such backgrounds. To address this, we adopt an alternative approach for the background component. Instead of employing the region proposal paradigm, we directly utilize a per-pixel feature extractor, such as LSeg [19]. Due to the comprehensive coverage of background categories in the training set of LSeg, it yields decent results for background. However, for open-world objects, it struggles to deliver accurate segmentation outcomes, as demonstrated in the appendix. In summary, our proposed method involves the extraction of per-pixel embeddings for each object and object part. We assign these embeddings to the pixels within the corresponding mask proposals. For pixels that remain unassigned, we utilize the embeddings generated by openclip-LSeg [19] directly. By combining these hierarchical embeddings, we achieve a hierarchical representation of the 3D field in Open-NeRF.

4.2. Open-vocabulary Querying

To enable open-vocabulary querying in NeRF decomposition, a model must be capable of predicting relevancy scores for given coordinates based on the open vocabulary queries received. A threshold can then be employed on the relevancy score to facilitate decomposition. Given a query $\mathbf{q}$, we initially utilize the text encoder of openclip to obtain the normalized text embedding $\mathbf{e}_t(\mathbf{q})$ for $\mathbf{q}$. As mentioned earlier, Open-NeRF generates hierarchical embeddings $\mathbf{e}(x_n)$ for each coordinate x_n. Consequently, we calculate the relevancy score $S_r(x_n)$ using the following equation: $S_r(x_n) = \mathbf{e}(x_n) \cdot \mathbf{e}_t(\mathbf{q})$. For NeRF decomposition, a threshold can be employed to determine whether some points belong to the object based on the given query. To generate 2D segmentation results for novel views, we render the field embeddings along the ray $\mathbf{r}$ following Eq.2, resulting in the rendered embedding $\mathbf{e}_\mathbf{r}(\mathbf{r})$. Then, we calculate the relevancy score for the pixel corresponding to $\mathbf{r}$ through the dot product between $\mathbf{e}_\mathbf{r}(\mathbf{r})$ and the text embedding $\mathbf{e}_t(\mathbf{q})$. As discussed in the preceding section, we learn embeddings at the object level, part level, and background level. Hence, during querying, the model can select the optimal level by comparing the maximum relevancy scores computed using the embeddings from these three levels.

5. Experiment

In this section, we thoroughly evaluate the capabilities of Open-NeRF and perform a comprehensive comparison with the current state-of-the-art method, LERF [16], as well as methods based on openclip-LSeg, such as FFD [18].



Image

Query for Edit: 'Table'

Query for Edit: 'Table; Vase'

Figure 5. Query-guided texture modifications based on the NeRF decomposition results from our proposed approach: Open-NeRF. The results are achieved by only giving a query for the target object and a query for the modification. For the editing step, we follow the CLIP-NeRF [38].

Method	Metric	Scene: *Toy*-1					Scene: *Kitchen*			Scene: *Garden*					
		Elephant	Rabbit	Zebra	Reward	Mean	Lego	Glasses	Mean	Vase	Table	Grass	Foot Ball	Umbrella	Mean
LERF	AUPRC	84.7	20.1	76.3	84.9	66.5	71.3	40.5	55.9	74.1	78.3	65.6	75.1	58.3	70.3
	FPR$_{95}$	5.4	40.3	1.9	0.5	12.0	10.2	33.1	21.7	8.4	7.3	18.3	8.1	43.1	17.1
Ours	AUPRC	96.1	88.7	97.4	89.1	92.8 ↑26.3	83.1	78.2	80.7 ↑24.8	85.8	90.4	70.2	91.3	62.8	80.1 ↑9.8
	FPR$_{95}$	1.1	3.6	0.07	0.2	1.2 ↑10.8	6.1	7.4	6.8 ↑14.9	4.9	3.3	8.5	2.1	21.7	8.1 ↑9
Ours w/o Integration	AUPRC	95.1	86.3	96.9	88.2	91.6	70.1	63.1	66.6	79.3	87.1	68.4	87.7	60.9	76.7
	FPR$_{95}$	1.7	4.1	0.11	0.23	1.54	10.3	18.1	14.2	6.8	5.3	9.2	3.4	20.1	9.0

Method	Metric	Scene: Toy-2					Scene: Toy-3				Scene: Car			
		Dinosaur	Duck	Rabbit	Lava Monster	Mean	Harry Potter	Bird	'Gaming Accessories'	Mean	Sedan	License Plate	Wheel	Mean
LEFR	AUPRC	78.2	95.8	79.2	21.3	68.6	60.1	96.5	53.8	70.1	73.5	45.7	53.2	57.5
	FPR$_{95}$	3.1	1.4	2.3	47.0	13.45	27.3	1.2	33.1	20.5	22.6	33.3	29.4	28.4
Ours	AUPRC	91.3	97.1	88.7	86.5	90.9 ↑22.3	91.2	95.3	85.2	93.3 ↑23.2	93.5	85.2	90.3	89.7 ↑32.2
	FPR$_{95}$	2.2	0.7	3.4	3.8	2.5 ↑10.9	1.6	1.3	5.8	1.5 ↑19	1.1	2.3	1.8	1.7 ↑26.7

Table 1. Quantitative Results in Scenes *Toy*-1, 2, 3, *Kitchen*, *Garden*, and *Car*. *Kitchen* and *Garden* scenes are provided by Mip-nerf 360 [1].

Our evaluation focuses on showcasing the proficiency of Open-NeRF in processing open-vocabulary queries. To achieve this, we conduct experiments not only on established datasets [1] but also on a collection of 5 diverse in-the-wild scenes encompassing office environments, markets, bookstores, and natural landscapes, featuring numerous long-tail objects. Our collected datasets will be introduced in detail in the appendix. All codes, datasets, and more results including videos will be released soon.

5.1. Implementation Details

We present the implementation of Open-NeRF within the Nerfstudio framework [35], building upon the advancements made by both LERF [16] and the Nerfacto [43] method. For efficient decomposition operations in future stages, we adopt the proposed sampling strategy employed by Nerfacto. Our approach leverages the Openclip [33] model, specifically the `ViT-B-32-quickgelu` version, which has been trained on the extensive LAION-400M dataset. To generate region proposals, we employ GroundingDINO [22] as our region proposal generator, utilizing a robust `Swin-B` backbone. Prior to its use in our framework, GroundingDINO has been pre-trained on multiple datasets, including COCO [21], O365, GoldG, Cap4M, OpenImage, ODinW-35, and RefCOCO. For mask generation, we rely on the capabilities of the Segment Anything Model (SAM) [17], utilizing the `sam-vit-h-4b8939` checkpoint. To ensure consistent alignment of mask proposals belonging to the same object across various viewpoints, we incorporate the state-of-the-art SAM-Track [6] technique. This allows us to establish accurate correspondences between masks, enhancing the overall quality of the decomposition process in our Open-NeRF implementation.

5.2. Comparison with Current Methods

In this section, we exhaustively compare both qualitative and quantitative results of our method with state-of-the-art methods, LERF [16], on multiple open-world scenes. Methods based on open clip-LSeg barely provide decent results on scenes with novel objects, which is shown in the appendix.

Qualitative Results. Figure 4 present a comprehensive assessment of the NeRF decomposition capabilities of our novel algorithms, Open-NeRF and LERF, across a variety of scenes. Each sub-image within these figures showcases the corresponding relevancy score, computed via the dot product of the rendered embeddings, generated through Open-NeRF, and the embeddings of the provided

Method	Metric	Scene: *Desktop*								
		MacBook	Magic Cube	iPhone	VR Glasses	Wireless Mouse	Speaker	Wallet	iPad	Mean
LERF	AUPRC	88.5	91.3	83.3	73.6	34.1	84.6	88.3	76.1	–
	FPR	3.4	1.6	4.9	11.7	42	5.7	3.1	8.4	–
Ours	AUPRC	89.2	92.6	94.5	91.2	93.2	84.1	93.9	83.7	–
	FPR	2.7	1.3	0.3	2.1	1.3	5.3	0.5	3.7	–
		UNO	Apple Watch	Airpods Case	Calculator	Lipstick	Toy	Airpods Box	Yellow Box	
LERF	AUPRC	83.6	67.8	90.1	81.3	73.9	76.1	84.8	82.6	78.8
	FPR	5.1	21.3	3.2	7.7	13.7	19.5	3.4	7.8	10.15
Ours	AUPRC	87.7	85.1	90.1	89.3	80.9	80.8	95.1	90.2	88.9 $_{\uparrow 10.1}$
	FPR	4.1	4.3	3.2	3.8	9.2	10.7	0.05	0.7	3.32 $_{\uparrow 6.8}$

Table 2. Quantitative comparison of Open-NeRF and LERF in Scenes *Desktop*.

text queries.

Open-NeRF demonstrates superior performance, accurately segmenting both common and novel objects regardless of viewpoint. LERF, on the other hand, delivers only approximate results. A distinguishing feature of Open-NeRF is its capacity to process open vocabulary as shown in the appendix, locating objects based on a variety of attributes, including product name, brand, color, material, and any text inscriptions. For example, a 'Purse', either through direct reference or by using its brand, such as 'Gucci'. Similarly, it accurately identifies a 'Candy Box' when given the descriptor 'Yellow Box', effectively excluding boxes of other colors.

Moreover, Open-NeRF displays versatility in grouping objects. As evidenced in Figure 4, it can identify multiple objects as a group or independently, contingent on the provided query. For instance, it can locate three toys simultaneously when given the query 'Toys', or individually, by providing a more detailed descriptor for each toy. Also, Open-NeRF can locate the whole car or parts of the car, *i.e.*, wheels, license plate, and mirror. This demonstrates Open-NeRF's adeptness at handling diverse scales and complex object queries.

Quantitative Results. To numerically evaluate the capacity of the model on open-vocabulary NeRF decomposition, we employ three distinct metrics on multiple random novel views: (a) Pixel-wise Area Under the Precision-Recall Curve (AUPRC), (b) Pixel-wise False Positive Rate when True Positive Rate equals 95% (FPR_{95}), and (c) Area Under the Receiver Operating Characteristic Curve (AUROC). Table 1 provides quantitative results of our proposed methods Open-NeRF and LERF [16] in 6 scenes. The results clearly indicate that Open-NeRF consistently surpasses LERF across almost all metrics in every test scenario. The superiority of Open-NeRF is particularly notable in the context of long-tail objects or queries. For instance, we achieve 86.5% on AUPRC for 'Lava Monster', surpassing LERF's performance of 21.3% by a significant margin. Similarly, our achievement of 91.2% on AUPRC for 'Harry Potter' showcases a 31.1% increase compared to LERF. In the appendix, we show the visualization results of our proposed method compared with LERF [16] in the scene: *Desktop* and here we provide the quantitative results as shown in Table.2. Our proposed approach Open-NeRF surpasses LERF in all three metrics. LERF struggles to provide decent results for objects such as 'Wireless Mouse' and 'Apple Watch', where they obtain 42% and 21.3% in FPR_{95}, while Open-NeRF achieves 1.3% and 4.3% in FPR_{95} for the same objects.

Ablations. In addition, we conducted an ablation study on the integration procedure, the results of which are presented in Table.1. Incorporating the integration procedure leads to a marked improvement in performance, especially in scenes featuring occluded objects or objects that are illegible from certain perspectives. Notable examples include the 'Glass of Water', which achieves 78.2% on AUPRC with 2D knowledge integration while achieving 63.1% on AUPRC without it, and the 'Lego Excavator' in the *'Kitchen'* scene, which shows 13% improvement on AUPRC after adding the integration procedure. This further underscores the effectiveness of Open-NeRF and its robustness in handling complex and challenging scenarios.

6. Limitations & Conclusion

In closing, Open-vocabulary Embedded Neural Radiance Fields (Open-NeRFs), provide a solution to the challenges inherent in open-vocabulary NeRF decomposition. With the innovative integrate-and-distill paradigm and hierarchical embedding, Open-NeRFs facilitate consistent recognition and granularity, irrespective of differing viewpoints by leveraging multiple off-the-shelf models. This approach outstrips current state-of-the-art methods in open-vocabulary scenarios, thereby showcasing its potential for practical applications in fields such as robotics and vision-language interaction with 3D scenes. Importantly, as Open-NeRF is constructed upon the foundation of models like Openclip and SAM, its performance is intrinsically linked to its capabilities. Therefore, as these foundational models continue to improve, we can anticipate a corresponding enhancement in the performance of Open-NeRF.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 7

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *arXiv preprint arXiv:2304.12308*, 2023. 2

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[6] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2

[9] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[11] Dongsheng Han, Chaoning Zhang, Yu Qiao, Maryam Qamar, Yuna Jung, SeungKyu Lee, Sung-Ho Bae, and Choong Seon Hong. Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected. *arXiv preprint arXiv:2305.00278*, 2023. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[13] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023. 2

[14] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. *arXiv preprint arXiv:2304.04395*, 2023. 1, 2

[15] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 2

[16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023. 1, 3, 6, 7, 8

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 7

[18] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 1, 2, 3, 5, 6

[19] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 3, 6

[20] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2

[21] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 3, 7

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 7

[23] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[24] Zhiheng Ma, Xiaopeng Hong, and Qinnan Shangguan. Can sam count anything? an empirical study on sam counting. *arXiv preprint arXiv:2304.10817*, 2023. 2

[25] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[28] 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*, pages 10318–10327, 2021. 2

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2

[30] Simiao Ren, Francesco Luzi, Saad Lahrichi, Kaleb Kassaw, Leslie M Collins, Kyle Bradbury, and Jordan M Malof. Segment anything, from space? *arXiv preprint arXiv:2304.13000*, 2023. 2

[31] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 2

[32] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 4

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 5, 7

[34] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 2

[35] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 7

[36] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453. IEEE, 2022. 1, 2, 3

[37] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 1, 2, 5

[38] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields, 2022. 7

[39] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2

[40] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[41] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[42] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zeroshot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 3, 2021. 2

[43] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 7

[44] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 1, 2, 5

[45] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2