Leveraging Panoptic Prior for 3D Zero-shot Semantic Understanding within Language Embedded Radiance Fields

Yuzhou Ji[1 [0009−0009−3572−060X]], Xin Tan[1,2⋆ [0000−0001−9346−1196]], He Zhu[1 [0009−0006−0835−7153]], Wuyi Liu[1 [0009−0008−3209−1281]], Jiachen Xu[1 [0009−0004−6511−3599]], Yuan Xie[1,2 [0000−0001−6945−7437]], and Lizhuang Ma[1 [0000−0003−1653−4341]]

[1] School of Computer Science and Technology
East China Normal University, Shanghai 200062, China
[2] Chongqing Institute of East China Normal University, Chongqing 401120, China
{xtan,yxie,lzma}@cs.ecnu.edu.cn
102151024{69,73,92,94}@stu.ecnu.edu.cn

Abstract. Language Embedded Radiance Fields (LERF) achieves promising results in real-time dense relevancy maps within NeRF 3D scenes. Although LERF shows impressive zero-shot ability in many long-tail open-vocabulary queries, the quality of relevancy maps could degrade in certain camera angles especially novel views and may even fail to localize. In this work we propose a method to bring in prior knowledge as the guidance of building a multi-scale CLIP (Contrastive Language-Image Pretraining) feature pyramid, achieving better localization ability and 3D consistency without any harm to original zero-shot capability. Specifically, we use panoptic segmentation to preprocess training images and reconstruct multi-scale image pyramid with segmented tiles. Unlike some other works, we only use the continuous semantic meaning of image tiles for accurate CLIP features, instead of labels or IDs which are inconsistent across views. And the tiles are partially overridden based on location and scale, preserving also a large amount of non-prior knowledge. And in order to effectively compare the results with LERF, we designed a metric based on pixel relevancy, which could further support future research based on LERF representation. Additionally, we explore the possibility of grounding dense 3D consistent segmentation information within LERF during experiments, providing an inspiring train of thought about distilling 2D knowledge into 3D scenes for 3D manipulation.

Keywords: Neural Radiance Fields · CLIP feature · zero-shot learning · semantic 3D scene · cross-modal distillation.

1 Introduction

In recent years, the Neural Radiance Fields (NeRF)[26] has witnessed remarkable growth and development. This advancement has led to substantial improvements in both the

⋆ Corresponding author.

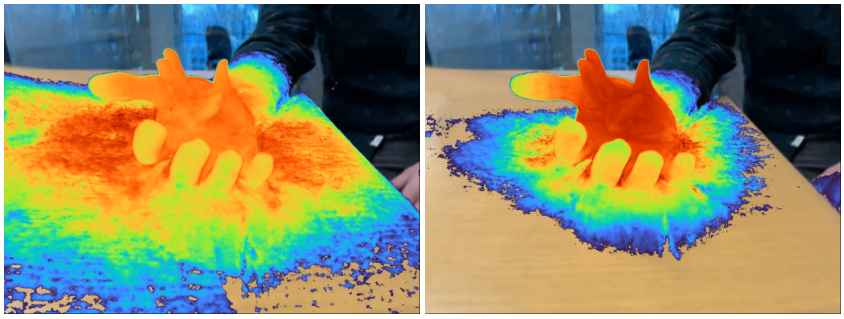

Fig. 1. LERF (left) vs ours (right). Query:"*porcelain hand*". While LERF has almost all high relevancy pixels misplaced, our method has a more accurate outcome.

precision of reconstructed novel views and the efficiency of the training process[2, 15, 29, 39], paving the way for NeRF to find promising applications in industries such as films, games, and Digital Twin production. However, NeRF is lack of language-level semantic information, making the training process rather difficult to respond to the human purpose, which becomes one of the key challenges of applying NeRF to achieve 3D contextual awareness.

Fortunately, a recent approach, Language Embedded Radiance Fields (LERF)[18], emerges and grounds semantic information within NeRF by directly training a CLIP[30] field, successfully utilizing CLIP's zero-shot ability to achieve the large-scale semantic understanding of NeRF reconstructed 3D scenes. While LERF is able to deal with a wide range of language proposals including abstract concepts ("*spill*"), text ("*Computer Vision*") and long-tail labels, some queries may render defective relevancy maps as shown in Figure 1.left where "*porcelain hand*" is queried but high relevancy pixels are rendered around the human hand. This problem may even get worse in views rendered using camera angles that are not in training data set, becoming a serious limitation of LERF.

One of the key causes of such limitation is that LERF utilizes no prior knowledge and directly implements CLIP for uniformly divided image tiles. Although LERF uses a multi-scale pyramid to ensure attention on objects of different sizes, in most cases these targets are either partially cut in smaller scales or not big enough to pass sufficient information through the CLIP encoder. This leads to fuzzy CLIP feature both inside and around objects, which gets worse after trilinear interpolation, causing defective outcomes.

To address this problem, in this work, we propose to bring in prior knowledge obtained in image segmentation tasks to guide the preprocessing of image patches. Different from other works[21, 32], we first conduct panoptic segmentation on all training images and cut out image tiles for each segmentation label. This step will extract image parts with continuous semantics yet we only keep image information without segmentation labels to maintain the original zero-shot ability. Next, we go through all scales and override the preprocessed image tiles centered at certain segmentation with segmentation tiles of closest scales. After overriding the image patches now share no restriction

that tile sizes should be the same in one scale, endowing preprocessed ray origins with overridden scales to have the most possibly accurate CLIP features. This method can not only increase pixel relevancy among queried objects but also decrease the relevancy for pixels outside that belong to another segmentation. It is clearly shown that our method has better results (Figure 1.right).

Notably, our way of using prior knowledge does not weaken the original zero-shot ability nor significantly slow down the implementation process and can still query objects that segmentation models fail to extract. Meanwhile, because LERF renders query results by pixel relevancy, common metrics used in image segmentation and NeRF may not be suitable for quality representation. Thus we present a new metric (Sec 3.5) for effective evaluation upon queried relevancy maps. Besides, we also found it possible to directly distill segmentation information into CLIP field (See Sec 4.3 Ablation Study) during experiments, which could be a new research direction.

In summary, we make the following contributions:

- We put forward and prove it feasible to bring in prior knowledge to optimize LERF's language field without harming the original zero-shot ability.
- We propose a method to utilize panoptic segmentation information in re-constructing CLIP feature pyramid and achieve higher quality results across scenes.
- We present a new metric for evaluation based on queried pixel relevancy, which could serve many future works that use LERF representation.
- We found a novel way of distilling segmentation information into NeRF using CLIP feature instead of segmentation label ID during experiments, showing a possible direction of 3D segmentation.

2 Related Works

2.1 NeRF with Semantics

NeRF [26] is a deep learning model employed for three-dimensional scene reconstruction which can predict the color and density in novel views. NeRF's capacity for high-quality three-dimensional reconstruction and its flexibility have established it as a cornerstone in the field of three-dimensional vision. In subsequent work, Semantic NeRF [41] enhances semantic output by designing a network that jointly considers semantic and geometric shape. It can predict high-quality new viewpoint segmentation results using only a small number of keyframes in large scenes. CLIP-NeRF [36] incorporates CLIP embeddings into the network space, enabling the use of simple textual prompts or image manipulations with NeRF.

Panoptic Lifting [32] utilizes a pre-trained network to infer a two-dimensional panoramic mask, allowing the generation of a unified multi-view 3D panoramic representation. Embedding 2D image features into 3D space has proven to be feasible. FFD [20] addresses the semantic scene segmentation challenge of NeRF by refining the knowledge from an existing 2D image feature extractor into a parallel-optimized 3D feature field within the radiance field. This enables specific queries to be made. In [23], CLIP [30] and DINO's [5] open-vocabulary and textual knowledge are distilled into a neural radiance field for 3D segmentation, but query capabilities are limited. LERF [18], on the

other hand, integrates language into NeRF using CLIP, allowing for real-time zero-shot queries. However, LERF's performance at high resolutions falls short of expectations.

2.2 Panoptic Segmentation

Although existing works on panoptic segmentation have successfully identified and segmented objects within images [4, 38], these methods largely operate under a closed-vocabulary assumption. Approaches like Panoptic-DeepLab [7] and DETR [4] utilize models that are trained to recognize and segment only objects and instances from a predefined vocabulary. This becomes a significant limitation in 3D reconstruction tasks, where the types of objects and instances are not always known in advance.

For our 3D reconstruction, the diversity and unpredictability of real-world objects require a model capable of operating within an open vocabulary. Conventional methods [8, 9, 38] are constrained by their inability to adapt to such scenarios, underlining the need for a more versatile panoptic segmentation model.

2.3 Open-Vocabulary Object Detection

Semantic segmentation in 2D visual scenes has largely been limited to closed vocabularies. However, the emergence of open-vocabulary semantic segmentation [14, 16, 17] has opened new research directions, enabling the recognition and classification of entities not in the training set. DetPro [14] employs pre-trained models to learn continuous prompt representations for open-vocabulary object detection and outperforms ViLD [17] across multiple metrics. Such advancements in 2D environments have influenced 3D segmentation, inspiring models like LERF [18].

To translate the successes of 2D open-vocabulary semantic segmentation into the 3D realm, LERF employed CLIP. The CLIP model has been a cornerstone in many NLP applications, including VQGAN-CLIP[12] and ClipCap[28]which can perform caption generation and other applications[25, 35] which can detect instances. Its capability to map visual and textual embeddings in a shared space made it a suitable choice for object detection in 3D reconstructions.

Despite its advantages, CLIP has limitations in distinguishing closely related or nuanced categories[19]. It can also be computationally demanding and lacks precision in mask generation. These drawbacks necessitate the exploration of more specialized approaches for 3D object detection. In this context, recent developments in generative models have been promising. Techniques using GANs[3, 22, 40] and diffusion models[13, 33] for semantic segmentation[37] have shown effectiveness. Therefore, our work incorporates ODISE [37], which employs a text-to-image diffusion model and offers significant improvements in mask accuracy and object differentiation, making it a compelling alternative for open-vocabulary 3D scene reconstruction.

2.4 Zero-shot Learning in 3D

In the realm of three-dimensional semantic understanding, the majority of the literature has predominantly centered on point cloud representations. These endeavors have

unearthed the profound potential of imbuing point cloud data with zero-shot language capabilities [1, 6, 10, 11]. The point cloud, given its explicit and structured nature, provides a fertile ground for achieving such intricate semantic tasks. However, when transitioning to the domain of Neural Radiance Fields (NeRF) [26], this zero-shot prowess seems to dwindle. Notably, the inherent implicit representation of NeRF poses challenges for directly porting over the same zero-shot capabilities that thrived in point cloud spaces.

Yet, amidst these challenges, a few ventures, such as LERF [18], have boldly embarked on the journey to amalgamate the zero-shot language abilities within NeRF. LERF, in particular, was seminal in integrating language into NeRF using the prowess of CLIP, facilitating real-time zero-shot queries.

While these works deeply discovered the possibility of empowering point cloud with zero-shot language ability, few researches have tried such implementation in NeRF like LERF does because NeRF representation is usually less explicit than point cloud. Our work, building on this nascent foundation, further investigates how to achieve a more refined and comprehensive semantic understanding in NeRF. We leverage the rich semantic and visual knowledge offered by CLIP to bolster the NeRF framework's ability to cater to intricate semantic queries and representations.

2.5 Cross-modal Knowledge Distillation

Knowledge distillation across modalities, particularly from 2D image domains to 3D representations, has shown to be a promising avenue in enhancing the understanding of three-dimensional scene structures. In the realm of Neural Radiance Fields (NeRF), a noteworthy attempt in this direction is epitomized by Panoptic Lifting[32]. It proficiently infuses the insights gained from 2D segmentations into a 3D NeRF framework. However, like some other pioneering endeavors[21, 24, 31], Panoptic Lifting, while excelling at partitioning structures and accuracy, does not support open-vocabulary labels. Consequently, these methods fall short in facilitating zero-shot querying capabilities, which restricts their adaptability in dynamic and real-time applications.

Although these works may be very good at partitioning upon point clouds or pixels, they are weak in zero-shot ability. In our work, we managed to make significant improvements using 2D segmentation knowledge without any detriment to zero-shot capabilities.

3 Method

3.1 Overview

We tackle the challenge of fuzzy CLIP feature generation in LERF by bring in prior knowledge. Given a sequence of NeRF standard images, we first extract semantic prior knowledge from training images using panoptic segmentation and cut out image tiles with continuous semantics. While LERF having problems due to uniformly divided image tiles, we override the division tiles centered at known segmentation with corresponding preprocessed segmentation tiles. By restricting scale selection, each segmentation tile will be used at most once, preserving still sufficient non-prior information

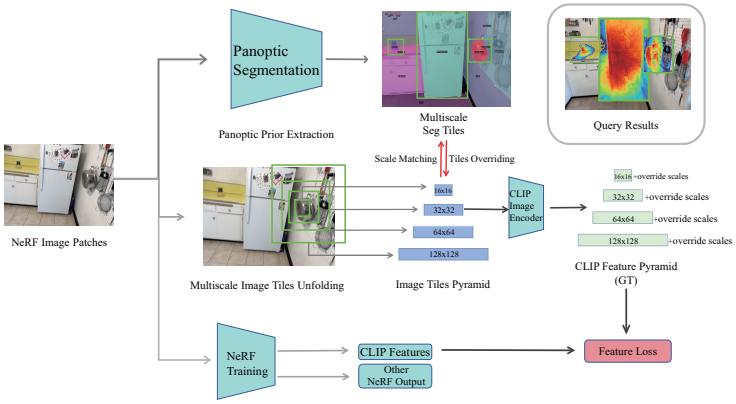

Fig. 2. Optimization Pipeline: *Top*: Extracting language prior using panoptic segmentation, which is for the refinement of image pyramid. *Mid*: Uniformly multi-scale image cropping to generate original image pyramid, and utilize segmented tiles provided above to do scale overriding as the reconstruction of image pyramid. Then send the image pyramid to CLIP Image Encoder to build the ground truth CLIP feature pyramid. *Buttom*: Basic NeRF method as backbone. For computing the feature loss of CLIP field, we use the reconstructed ground truth pyramid and rendered features. Loss functions remain the same as in LERF.

.

in multi-scale image patches. Therefore the outcome CLIP embedding could hold both precise and abundant information, providing more precise queried relevancy maps.

In the following sections we will detail our method (See Figure 2) and also provide a new metric for effectively evaluating the results.

3.2 Field Structure

LERF builds upon the foundational approach of NeRF, grounding CLIP embeddings into a 3D field within NeRF by learning a field of language embeddings over volumes centered at the sample point.

Given a viewing ray parameterized as

$$\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d} \tag{1}$$

where $\boldsymbol{o}$ is the ray's origin and $\boldsymbol{d}$ its direction, the integral[26]:

$$C(\boldsymbol{r}) = \int_{t_{\text{near}}}^{t_{\text{far}}} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \tag{2}$$

captures the accumulated color for the ray, considering its traversal through the volume. Here, $T(t)$ represents the accumulated transparency along the ray up to point t.

Apart from NeRF method, LERF augments a language $F_{\text{lang}}(\boldsymbol{x}, s) \in R^d$ that takes an input position $\boldsymbol{x}$ and physical scale s, and outputs a view-independent d-dimensional language embedding.

The rendering of these embeddings into the 3D scene is executed as[18]:

$$\hat{\phi}_{\text{lang}} = \int_t w(t) F_{\text{lang}}(r(t), s(t)) dt \tag{3}$$

Here, $w(t)$ represents the rendering weights.

In order to build better CLIP field, we need to effectively supervise each rendered frustrum with an image crop of size s_{img} centered at ray origin. While CLIP operates over image patches, it is necessary that for each ray the supervising image crop should be given in multi-scale, ensuring objects of different sizes in one scene can all be allocated with appropriate attention and accurate CLIP embeddings. Meanwhile, because it is unaffordable to compute CLIP embedding for every ray with multiple scales, we can only pre-compute a image pyramid including part of ray origins and perform trilinear interpolation to compute other CLIP features. While normally the image pyramid used for trilinear interpolation is required to have the scale of each layer half the size of previous layer, we argue this to be a must because results of CLIP encoder hardly varies upon small size changes. Therefore, for image scales between s_{min} and s_{max}, we override certain image tiles at original pyramid levels with our calibrated tiles that have similar scales yet better CLIP feature after image encoding. Finally we can supervise CLIP fields of each scale using the ground truth embedding φ_{lang}^{gt} generated from this reconstructed pyramid, which will be further introduced later. During query, we will go through CLIP fields of all pre-defined scales and choose the best field to render relevancy map.

3.3 Semantic Prior Extraction

To bring in prior knowledge, we choose image segmentation for guidance. In LERF's pre-computing of image pyramid, many objects were either not entirely included in one small crop or too tiny to contribute enough information in big crops. These defective crops lead to lower relevancy after image encoder, and get vaguer for other image points when performing interpolation. In terms of this issue, we propose to first extract pixels containing complete and continuous semantic and then send for encoding of relevant ray origins, which no longer need to be at the center of encoding image tiles. Specifically, we perform image segmentation to pre-crop training images. Based on segmentation masks, for each segmentation index, we cut out its image tile from (*top*,*left*) to (*bottom*,*right*) of the targeted mask and perform zero padding (with extracted image tiles at center of padded images) to create a calibrated segmentation tile with the size of $\max(vertical_{int}, horizontal_{int})$. The reason why we choose image tiles instead of cut-out pixels of the same label and do zero padding for empty spaces is that we need to preserve the surrounding information for long-tail label queries. For example, if we only use the pixels of a coffee cup, CLIP encoder will simply ignore the surrounding information of coffee on the table and resulting in failure while querying "$spilled\ coffee$". Such knowledge will also be needed when computing the CLIP features of nearby ray origins.

After extraction, these generated tiles can provide better CLIP features concerning the segmented parts compared with many uniformly divided tiles which can not be fully

captured by CLIP image encoder. Moreover, this step is label-free which means we do not need segmentation labels. One of the difficulties of grounding language information within 3D scenes is the inconsistency of segmentation labels between images. For example, in the sequential images of the NeRF data set, a hand can be given labels from “hand” and ”person” to “handle” in spite of accurate segmentation, making it impossible to ground semantic information by training with label ids. But since we are now using CLIP, even if one object in different segmentations may have varied labels, the encoded vectors are still close in space, which solves the problem of 3D semantics inconsistency.

Explicitly, we require every pixel to have one segmentation label, instead of only focusing on foreground objects, thus panoptic segmentation is chosen. We find panoptic segmentation very effective because it also provides clear masks for different background and irrelevant parts. By sending these parts for supervising the CLIP features around interested objects, surrounding pixels will bring less interference compared with supervising these areas with uniformly divided image tiles, which could usually crop parts of foreground objects and result in incorrectly high relevancy concerning surrounding background parts of queried targets. In practice, we used ODISE[37] which supports open-vocabulary labels and works well in the tested scenes. The segmentation model used here can be replaced, but we discovered that many models may need much more refinement to provide enough segmentation tiles concerning LERF’s in-the-wild scenes.

3.4 CLIP Pyramid Reconstruction

The processed segmentation tiles are used for reconstructing the CLIP feature pyramid, but it is important to decide how much tiles built upon prior knowledge should be implemented concerning the base pyramid patches. If we override all tiles using only the patches extracted by image segmentation, the zero-shot ability will be largely affected and reduced to segmentation label sets, or even renders wrong outline (See Figure 6.right).

In order to improve accuracy without affecting the original zero-shot ability, we override uniformly cropped tiles using segmentation tiles with matched points and scales. First, the original tile should be centered at certain segmentation (some pixels are not given labels in segmentation if not belong to the label set). Second, the scale of this tile must be close to the corresponding segmentation tile. During processing, while the original tile has the scale s_i in s_{min} to s_{max}, we require the target segmentation tile to have a scale s_x that $s_{i-1}*1.1 < s_x \leq s_{i+1}*1.1$ or $s_x \geq s_{max}$. With this restriction, one segmentation tile can only override at most one tile at a certain scale. The overridden tiles can generate the most precise CLIP feature for center points and increase peripheral relevancy through trilinear interpolation, resulting in better relevancy map in this scale level. This brings an extra benefit that other overridden tiles centered at different segmentation now have CLIP features that are closer to other segmented stuff, meaning lower relevancy concerning current query and better rendered high relevancy outline. Moreover, for objects with other sizes or stuff ignored by the segmentation model, their CLIP features are still stored in other scale levels in the CLIP feature pyramid, and can also be queried.

3.5 Relevancy Evaluation Metric

Because the way LERF grounds language information into NeRF is rather novel and unique, commonly used metrics in NeRF or image segmentation such as PSNR and MIoU may not be appropriate for evaluating the quality of results based on LERF representations. While LERF do provide high-quality results that have segmentation-like outlines in certain views, most queries can only get a vague relevancy map. Therefore, we propose a metric based on LERF's pixel relevancy score to explain LERF's visual results with precise data.

Basically, in order to generate a better query result, we need both pixels inside the queried object to obtain high relevancy scores and outside pixels to have low relevancy scores (See also Figure 1 as an example). Normally, take $Inseg_{ove}$, $Outseg_{ove}$ as average pixel relevancy scores inside and outside queried objects, we can easily use $Ratio_{acc} = Inseg_{ove} \div Outseg_{ove}$ to represent the accuracy of pixel relevancy. However, this ratio cannot be restricted to a specific range and may be largely affected across scenes. Thus, we propose $D_{relevancy}$.

Take:

$$R_{gt}(x,y) = \begin{cases} 1, \text{if pixel } (x,y) \text{ inside query} \\ 0, \text{if pixel } (x,y) \text{ outside query} \end{cases} \tag{4}$$

as ground truth relevancy scores for all pixels, then

$$D_{relevancy} = \frac{\sum (R_{gt}(x,y) - R(x,y))^2}{img_{height} \times img_{width}}, \quad x,y \in [0, img_{height}), [0, img_{width}) \tag{5}$$

where $R(x,y)$represents the predicted pixel relevancy.

The reason why we define R_{gt} (x,y) to be 1 or 0 is because the best result should be only pixels belonged to queried object are presented with absolute relevancy and other irrelevant pixels having zero relevancy. As for the counted D-value, we process it using square value instead of absolute value because we want to filter small D-values and focus on large ones which are more definitive concerning the final visual results. By doing so, $D_{relevancy}$ ranges from 0 to 1 and can now effectively demonstrate the gap between ground truth and rendered results.

During experiments, we found $D_{relevancy}$ consistent with visual results that lower $D_{relevancy}$ maps usually have better visuality. We believe this metric may also be useful for the evaluation of future works based on LERF representation.

4 Experiments

4.1 Settings

We conduct extensive experiments upon the LERF datasets[18] with most variables including CLIP model (OpenClip ViT-B/16 model), NeRF method (Nerfacto) and trained steps consistent with LERF to examine the effects of introducing prior knowledge. We also use the same DINO[5] regulation which is proved to be essential in refining high

quality maps by LERF. We implement ODISE[37] as our open-vocabulary panoptic image segmentation model for extraction, with all labels chosen. Segmented tiles are zero-padded and original tiles are moved to the center of padded images to have better CLIP encoding. While the pixel relevancy and visual results change with rendering resolution, we test based on resolution 256×138, 512×298 and 994×738 (the resolution of training images)rendered using Nerfstudio[34]. In Table 2 image resolution is fixed to 512×298 and only pixels with relevancy more than 0.5 are highlighted. We use the metric mentioned in Sec 3.5 and text query tables provided by LERF. Each query is tested across 4-10 viewing angles depending on the original complete appearance frequency in training images. Camera FOV used in rendering is 50.

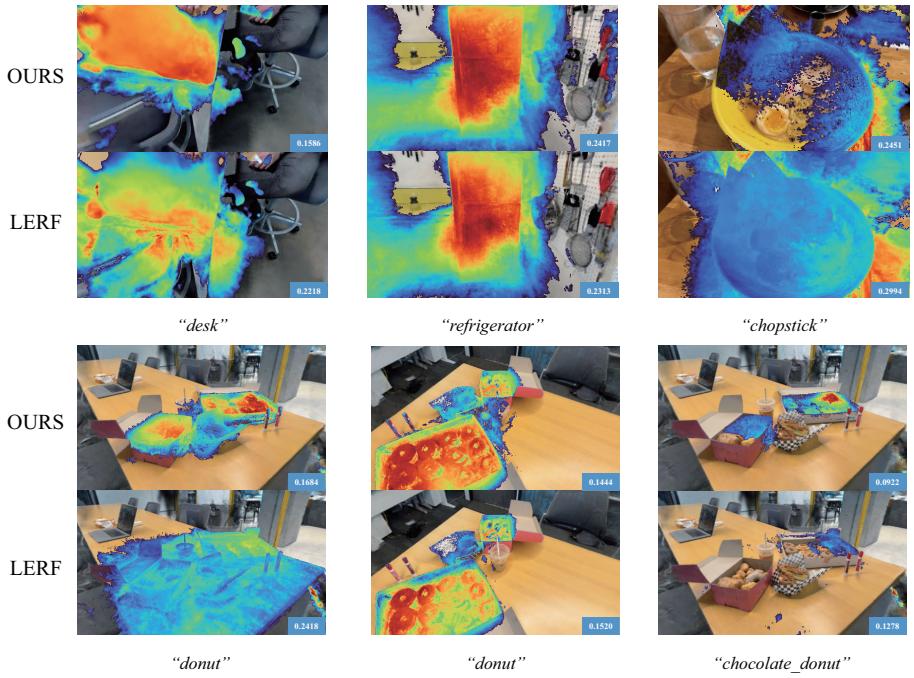

Fig. 3. Results in different scenes and views, with $D_{relevancy}$ in the bottom right corner. In most views, we have lower $D_{relevancy}$, and our method shows much better localization ability in LERF's failed maps. We also show a case where we have a bit higher $D_{relevancy}$ but a more even distribution of similarity and clearer outline in "refrigerator".

4.2 Qualitative Results

We implement the same visualization of relevancy score as in LERF for visual comparison. In Figure 3 and 4 we provide some outstanding results in different scenes that prove

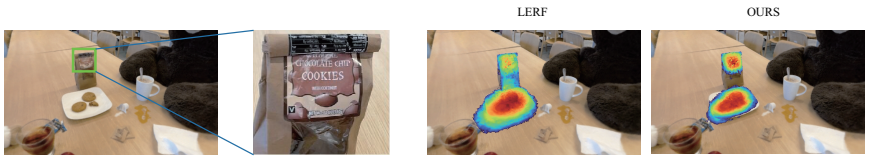

Fig. 4. Query "*cookies*" where exists both cookies on the plate and text "cookies" on the bag. Both should be given enough attention concerning single text query "*cookies*" after image encoding if trained correctly.

the effectiveness of our method in generating better localization maps. General data results across scenes are shown in Table 2, where we also provide average relevancy score for evaluation.

Table 1. Query Recall In LERF and our methods, we consider a label within certain view a success if at least nine out of ten highest relevancy (top 10 percent) pixels land inside the segmentation. In OWL-ViT we consider an object detection correct if the highest confidence prediction matches the label bounding box (if there are other predictions with similar confidence but fall upon the wrong object, we still consider it a mismatch). Each label is tested across 3-6 original views and OWL-ViT is tested based on training images.

DataSets	OWL-ViT	LERF	OURS
hand-hand	91.39%	89.62%	**93.24%**
waldo-kitchen	**93.87%**	91.82%	89.80%
ramen	88.44%	84.63%	**88.56%**
donuts	84.84%	87.83%	**90.81%**
teatime	87.71%	92.32%	**93.54%**
overall	90.52%	89.54%	**91.18%**

In most views, both our method and LERF can successfully localize queried targets, and our method has higher recall rate because we outperform LERF in its failed cases. We can also have higher detection ability compared with 2D open-vocabulary object detection model OWL-ViT[27] which also has zero shot capability (See Table 1). Sometimes, LERF's rendered quality can be really fuzzy and deranged in certain demanding views especially novel views. In novel view query "*porcelain hand*", when LERF fails to localize model hand, ours method provides high quality results. When given demanding tasks such as querying part of "*chopsticks*", "*table*" and "*donut*", our method also shows higher relevancy and better outline. In these views, we also have much lower $D_{relevancy}$, meaning closer relevancy grounding compared with ground truth segmentation. Our method succeeds across tested scenes in having lower average $D_{relevancy}$ and can handle many demanding views which LERF fails to localize, proving our improvements compared with LERF.

Moreover, when querying "*cookies*" in a demanding view (See Figure 4), our method shows not only higher accuracy in localizing cookies on the plate, but also gives the text

"cookies" on the bag equal attention, while LERF renders text "cookies" are less relevant. This means our method better grounds the text feature and can handle ambiguous queries. These results show that our method is more robust under general situations.

Table 2. Average $D_{relevancy}$ in different LERF datasets. Each label is tested across 5-8 views including novel views.

DataSets	LERF	OURS
hand-hand	0.2123	**0.1587**
waldo-kitchen	0.2460	**0.2436**
ramen	0.2423	**0.2240**
donuts	0.1375	**0.1333**
teatime	0.1307	**0.1201**

4.3 Ablation Study

In order to demonstrate why it is important that our method brings in prior knowledge only for partial reconstruction of CLIP feature pyramid across scales, we provide simple yet intuitive experiment results for understanding.

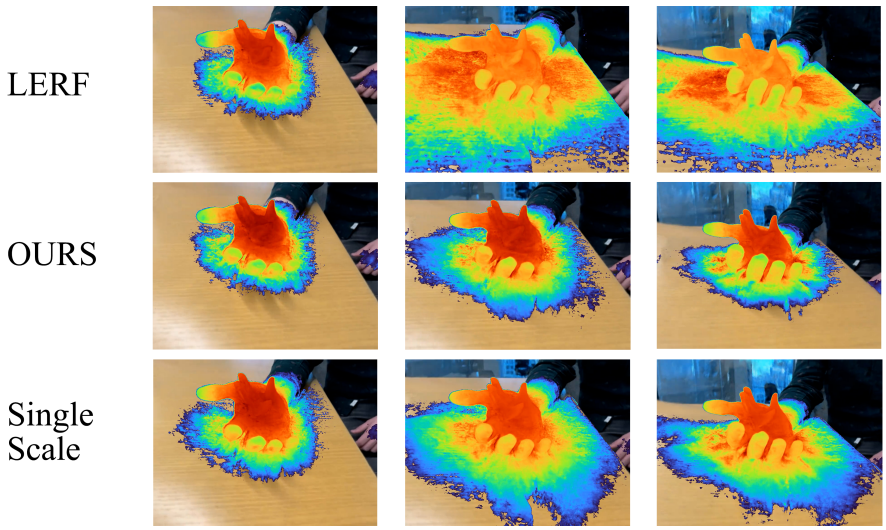

Fig. 5. Comparison of three methods. In the first training view three methods share similar results, then in the next training view LERF fails to localize and single scale result starts to show obvious retrogress but can still generally localize. In the last novel view our method still shows strong localization ability while LERF and single scale results keep getting worse.

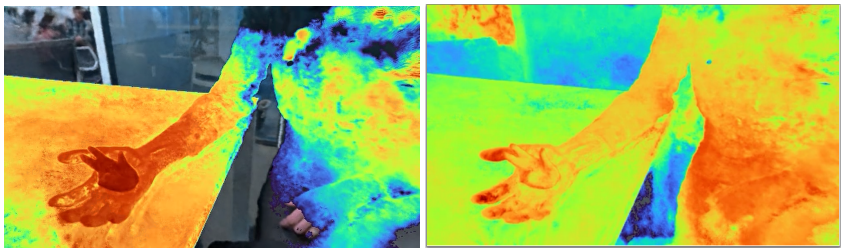

Fig. 6. Full segmentation supervision outcome.Left: Query "*porcelain hand*" shows high relevancy and clear outline. Right: Query "*hand*" yet high relevancy outline follows whole person. This is because segmentation label is "person".

Single scale overriding We override all image tiles of the the biggest scale in the image pyramid with segmented tiles and keep other scales of the image pyramid unchanged, so that the scales of segmented tiles are being ignored and the tiles are used directly for overriding in the same scale of image pyramid. Although this method can also localize "*porcelain hand*" in the previous viewing angle (See Figure 5), in most views the segmentation information is ignored or even wrongfully used. This is because many segmentation tiles could be much smaller than the biggest scale and their mixed use with correct scale tiles in interpolation could mislead the computed CLIP feature, causing worse relevancy. This method may result in even lower $D_{relevancy}$ than LERF in many views.

Full segmentation supervision In this experiment, we removed the tiles overriding procedure along with the multiscale CLIP feature pyramid. Instead, we supervise the CLIP features of all training rays by calculating the CLIP embeddings for each ray. The image tiles sent to CLIP image encoder are preprocessed segmentation tiles where the rays originate so that all training rays are supervised using prior knowledge obtained in image segmentation. This requires calculating CLIP outcomes for massive training rays and can be very time-consuming. We trained 9000 out of 30,000 steps within 12 hours when we already have usable relevancy maps. See outputs in Figure 6.

Although this could render the highest relevancy that gets close to segmentation outcome in certain views (Figure 6.left), the zero-shot ability is destroyed and reduced to segmentation labels. The rendered results are now highly consistent with image segmentation results and may fail to query part of segmented tiles (Figure 6.right), not to mention the unbearable training time.

To sum up, based on the above experiments, we believe our method is a rather suitable way of grounding image segmentation prior knowledge within LERF to have better open vocabulary query results in reconstructed scenes. Meanwhile, it is still worth noticing that full segmentation supervision successfully grounds high quality image segmentation information within NeRF reconstructed scenes, which some other works are struggling with. We believe this result proves CLIP's outstanding ability in such tasks and can serve as a very important tool for future works concerning 3D segmentation should they find a way to boost training efficiency.

5 Limitations

Our method shares the same "bag-of-words" behavior shown in CLIP and LERF that for example, takes "not red" similar to "red". Such proposals need more demanding language understanding and may not be correctly encoded. Future works could try to improve the performance of embedding open-vocabulary queries with directed focus on similar negative proposals.

Although we do train a 3D-consistent field, the image encoding and rendering are all based on 2D image, thus it is not available to query using spatial information. It may need to encode contextual information to solve this problem. Meanwhile, we still require queried objects to be captured with appropriate sizes across a couple of views, ensuring its attention and accurate CLIP embedding for high quality results. We suggest future works employ cross view attention and image augmentation concerning this limitation.

We have also limitations concerning segmentation models. For instance, ODISE fails to segment "expo markers" on the table in many views, causing us to lose that part of prior knowledge. A model capable of more precise segmentation could bring better results.

6 Conclusions

We present a method of introducing prior knowledge obtained image segmentation into LERF and outperforms LERF with higher quality relevancy maps without any reduction in zero-shot ability. We found it very effective to store both prior and non-prior knowledge together for high recall, yet also possible to supervise using dense segmentation information through CLIP encoder to achieve accurate object outline in NeRF scenes. We provide this very baseline for future works concerning semantic NeRF scenes to balance between recall and precision. Our metric will also hopefully serve later research in the evaluation of relevancy maps.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No.62302167, U23A20343), Shanghai Sailing Program under Grant (23YF1410500), Natural Science Foundation of Chongqing (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9892–9902 (2022), https://api.semanticscholar.org/CorpusID:247187696
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5835–5844 (2021), https://api.semanticscholar.org/CorpusID:232352655

3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. ArXiv **abs/1809.11096** (2018), https://api.semanticscholar.org/CorpusID:52889459
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. ArXiv **abs/2005.12872** (2020), https://api.semanticscholar.org/CorpusID:218889832
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (October 2021)
6. Chen, R., Zhu, X., Chen, N., Li, W., Ma, Y., Yang, R., Wang, W.: Bridging language and geometric primitives for zero-shot point cloud segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5380–5388 (2023)
7. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12472–12482 (2019), https://api.semanticscholar.org/CorpusID:208248153
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1280–1289 (2021), https://api.semanticscholar.org/CorpusID:244799297
9. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Neural Information Processing Systems (2021), https://api.semanticscholar.org/CorpusID:235829267
10. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Mitigating the hubness problem for zero-shot learning of 3d objects. In: British Machine Vision Conference (2019), https://api.semanticscholar.org/CorpusID:196622565
11. Cheraghian, A., Rahman, S., Chowdhury, T.F., Campbell, D., Petersson, L.: Zero-shot learning on 3d point cloud objects and beyond. International Journal of Computer Vision **130**, 2364 – 2384 (2021), https://api.semanticscholar.org/CorpusID:233210533
12. Crowson, K., Biderman, S.R., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: Vqgan-clip: Open domain image generation and editing with natural language guidance. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:248239727
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. ArXiv **abs/2105.05233** (2021), https://api.semanticscholar.org/CorpusID:234357997
14. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.C.: Learning to prompt for open-vocabulary object detection with vision-language model. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14064–14073 (2022), https://api.semanticscholar.org/CorpusID:247778949
15. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.P.C.: Fastnerf: High-fidelity neural rendering at 200fps. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 14326–14335 (2021), https://api.semanticscholar.org/CorpusID:232270138
16. Ge, Y., Xu, J., Zhao, B.N., Joshi, N., Itti, L., Vineet, V.: Beyond generation: Harnessing text to image models for object detection and segmentation (2023)
17. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2021), https://api.semanticscholar.org/CorpusID:238744187
18. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields (2023)
19. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields (2023)

20. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation (2022)
21. Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L.J., Tagliasacchi, A., Dellaert, F., Funkhouser, T.A.: Panoptic neural fields: A semantic object-aware neural scene representation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12861–12871 (2022), https://api.semanticscholar.org/CorpusID:248572506
22. Li, D., Ling, H., Kim, S.W., Kreis, K., Barriuso, A., Fidler, S., Torralba, A.: Big-datasetgan: Synthesizing imagenet with pixel-wise annotations. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 21298–21308 (2022), https://api.semanticscholar.org/CorpusID:245906460
23. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., Saddik, A.E., Theobalt, C., Xing, E., Lu, S.: 3d open-vocabulary segmentation with foundation models (2023)
24. Liu, Y.C., Huang, Y.K., Chiang, H.Y., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. arXiv preprint arXiv:2104.04687 (2021)
25. Lu, Y.T., Liu, S., Thiagarajan, J.J., Sakla, W.A., Anirudh, R.: On-the-fly object detection using stylegan with clip guidance. ArXiv **abs/2210.16742** (2022), https://api.semanticscholar.org/CorpusID:253237985
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020)
27. Minderer, M., Gritsenko, A.A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. ArXiv **abs/2205.06230** (2022), https://api.semanticscholar.org/CorpusID:248721818
28. Mokady, R.: Clipcap: Clip prefix for image captioning. ArXiv **abs/2111.09734** (2021), https://api.semanticscholar.org/CorpusID:244346239
29. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics **41**(4), 1–15 (jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145%2F3528223.3530127
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
31. Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9881–9891 (2022), https://api.semanticscholar.org/CorpusID:247793124
32. Siddiqui, Y., Porzi, L., Buló, S.R., Müller, N., NieSSner, M., Dai, A., Kontschieder, P.: Panoptic lifting for 3d scene understanding with neural fields (2022)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ArXiv **abs/2010.02502** (2020), https://api.semanticscholar.org/CorpusID:222140788
34. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A.: Nerfstudio: A modular framework for neural radiance field development. ACM SIGGRAPH 2023 Conference Proceedings (2023), https://api.semanticscholar.org/CorpusID:256662551
35. Teng, Z., Duan, Y., Liu, Y., Zhang, B., Fan, J.: Global to local: Clip-lstm-based object detection from remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022), https://api.semanticscholar.org/CorpusID:234104424

36. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields (2022)
37. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., Mello, S.D.: Open-vocabulary panoptic segmentation with text-to-image diffusion models (2023)
38. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M.D., Zhu, Y., Adam, H., Yuille, A.L., Chen, L.C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2550–2560 (2022), https://api.semanticscholar.org/CorpusID:249890221
39. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields (2020)
40. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10140–10150 (2021), https://api.semanticscholar.org/CorpusID:233231510
41. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation (2021)