NeurNCD: Novel Class Discovery via Implicit Neural Representation

Anonymous Submission

Abstract

This work uses implicit neural representations and RGB-D sensing to discover novel semantic classes in the open world. Towards this end, most prior works have relied on object descriptors that have been artificially created or explicitly build a 3D map to discover novel classes. However, these explicit representations are discrete and use a lot of memory and drastically limit the utilization of high-resolution scenes. To address these issues, we present NeurNCD, a novel method that discovers novel classes from multi-view RGB-D images via neural radiation fields. The proposed general system first segments each RGB-D frame to generate a set of convex sub-instance level segments, and then we draw inspiration from the quickly developing field of implicit neural representations, training an Embedding-NeRF model for each scene to fuse and render the semantic embedding extracted by the pre-trained semantic segmentation model, which is a key component of our approach as it not only can effectively aggregate 2D semantic features from multi-views but also effectively enable 2D-3D feature transfer and association. Following that, use the feature aggregation module to aggregate sub-instance level segments, semantic embedding, and entropy features together for Markov clustering, which can associate sub-instance level segments of the same class and help discover novel classes. Extensive experiments show that our method significantly outperforms the state-of-the-art approaches on the NYUv2 and Replica datasets. The code can be found in the supplementary material.

1 Introduction

With the rapid expansion of computer vision and robotics, a paradigm shift from “Supervised AI” to “Embodied AI” has occurred, implying that AI algorithms and agents can acquire knowledge through interactions with their surroundings, utilizing an egocentric perspective similar to that of humans. Obviously, a geometric and semantic comprehension of the scene is necessary to enable intelligent agents, such as indoor mobile robots, to plan context-sensitive activities in their surroundings. So scene understanding has become one of the core elements to realize embodied artificial intelligence. In prior work, supervised semantic segmentation [Gupta *et al.*, 2015; Chen *et al.*, 2018; Seichter *et al.*, 2021; Wang *et al.*, 2022b] methods have achieved tremendous success in image segmentation tasks; regardless of how large the scene is or how changeable the external conditions (texture, lighting), supervised deep learning can almost perfectly predict the category to which each pixel belongs. However, the significant progress gained by such algorithms is predicated on massive and costly data sets, which require a substantial amount of time and effort. Furthermore, when real-world circumstances diverge too much from their training data, their performance in applications is frequently poor, in other words, it is impossible to identify novel classes with accuracy. This is caused by current semantic segmentation methods making a closed-world assumption and is trained only to segment a limited number of semantic classes.

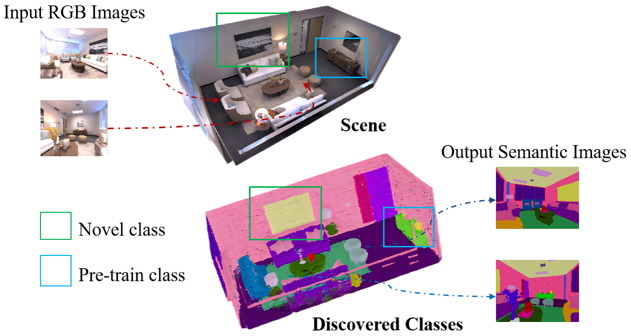

Figure 1: The proposed method discovers novel classes(e.g., pictures) in the neural radiation field.

Nakajima *et al.* were one of the first to demonstrate semantic scene understanding that can identify novel objects. Their method achieves the best results in discovering novel class tasks, but their results rely heavily on explicitly building a single segmented dense 3D map of the environment to find coherent regions to discover novel classes. This novel class discovery method of explicitly representing scenes and aggregating coherent regions’ semantic embeddings will result in artifacts like overlapping because of insufficient refinement, which uses a lot of memory and drastically limits the utilization of high-resolution scenes.

Implicit neural representations are an emerging paradigm to tackle problems in 3D scene understanding and also show great potential in fine modelling and rendering. we look beyond the horizon and explore a new question: ***Can we solve the discover novel classes issue in the context of implicit neural representation ?***

Zhi *et al.* proposed a method (Semantic-NeRF) for supervised semantic segmentation using neural radiation fields. They discovered that the label propagation and fusion capabilities of semantic-NeRF are astounding. They also showed how supervision from a single pixel per class may provide surprisingly high-quality rendered labels with a well-preserved global and fine structure. Obviously, their method is supervised to learn semantic features to complete the scene understanding task, and in the novel class discovery task, we can only discover new classes by learning semantic embedding, but no work has explored whether NeRF can implicitly aggregate and render semantic embeddings, and how to preserve the information contained in semantic embeddings during the rendering process of NeRF.

Inspired by previous work, we propose a novel method for discovering novel classes by aggregation and rendering semantic embedding using neural radiance fields. Unlike Semantic-NeRF, which has a closed world assumption and is trained only to segment a restricted number of semantic classes, our method uses RGB-D data to divide interior environments semantically without supervision and works in an "open set" manner to discover novel classes.

The main contributions of this work are:

- A novel method leverages implicit neural rendering for the discovery of novel classes and unsupervised semantic segmentation. It neither relies on densely labelled datasets for supervised training nor requires human interaction to generate sparse label supervision.
- It is proposed that Embedding-NeRF be used to fuse and render semantic embedding in RGB-D images and to reduce the KL divergence between extracted and rendered semantic embeddings in order to provide high-quality, globally consistent semantic embeddings.
- Extensive experiments indicated that our method outperformed state-of-the-art approaches on NYUv2 and Replica datasets. The design of each component is supported by comprehensive experimental validation and extensive ablation investigations, which also show the usefulness of our system in applications like indoor scene semantic segmentation.

2 Related work

2.1 Semantic Segmentation

As a fundamental task in computer vision, semantic segmentation, which seeks to predict semantic labels for every pixel in an image, has received much attention. It divides an image into several cohesive, semantically relevant sections and plays an important role in visual scene understanding. In recent years, substantial progress has been achieved in the field of supervised semantic segmentation[Li *et al.*, 2021; Liu *et al.*, 2018], however, such work is "labour-intensive" and appears to be at a loss when confronted with new environments or unknown classes.

In order to remove the dependence on annotations, Unsupervised semantic segmentation has caught the interest of researchers because it can reduce the amount of pixel-level annotations needed for semantic segmentation while also discovering novel classes. Nakajima *et al.* were one of the first works to discover novel classes, they rely on superpixel segmentation, mapping, and clustering to identify object categories. Frey *et al.* shows a ready-to-deploy continuous learning approach for semantic segmentation that does not require any prior knowledge of the scene or any external supervision and can simultaneously retain the knowledge of previously seen environments while integrating new knowledge. In order to deploy the semantic segmentation model on the robot, Seichter *et al.* proposed ESANet, which is an efficient and robust RGB-D segmentation approach that can be optimized to a high degree using NVIDIA TensorRT [Vanholder, 2016]. They evaluated ESANet on the common indoor datasets NYUv2 and SUNRGB-D, and the results demonstrated that the method achieves state-of-the-art performance while enabling faster inference.

2.2 Radiance Field-based Scene Representations

Our work on discovering novel classes and unsupervised semantic segmentation build on neural radiance fields (NeRF) [Mildenhall *et al.*, 2021], which represent a scene using a multi-layer perceptron (MLP) that maps positions and directions to densities and radiances. The following work[Xu *et al.*, 2022; Müller *et al.*, 2022; Wang *et al.*, 2022a; Chen *et al.*, 2022; Fu *et al.*, 2022] improve NeRF for faster training and inference and more realistic rendering. Using MLP or explicit feature grids, these radiance field-based scene representations achieve unprecedented novel view synthesis effects. Considering Semantic Segmentation in the Neural Radiation Field, NeSF[Vora *et al.*, 2021], a method for simultaneous 3D scene reconstruction and semantic segmentation from posed 2D images, is demonstrated by Suhani Vora et al. Their approach, which is based on NeRF, is trained entirely on posed 2D RGB images and semantic mappings. Their method creates a dense semantic segmentation field during inference that can be queried directly in 3D or used to produce 2D semantic maps from novel camera postures, but their method is a supervised method the same as [Zhi *et al.*, 2021] and [Zhi *et al.*, 2022]. At the same time, their method only verified that NeRF has a strong ability in the low-dimensional image or semantic rendering, but did not study the performance of NeRF fusion and rendering of higher-dimensional image embeddings.

2.3 Novel Class Discovery and Clustering

For novel class discovery, Zhao *et al.* proposed a method to discover novel classes with the help of a saliency detection model and use an entropy-based uncertainty modelling and self-training (EUMS) framework to overcome noisy pseudo-labels, further improving the model's performance on the novel classes. But their method can only segment and discover a limited number of salient categories, while our

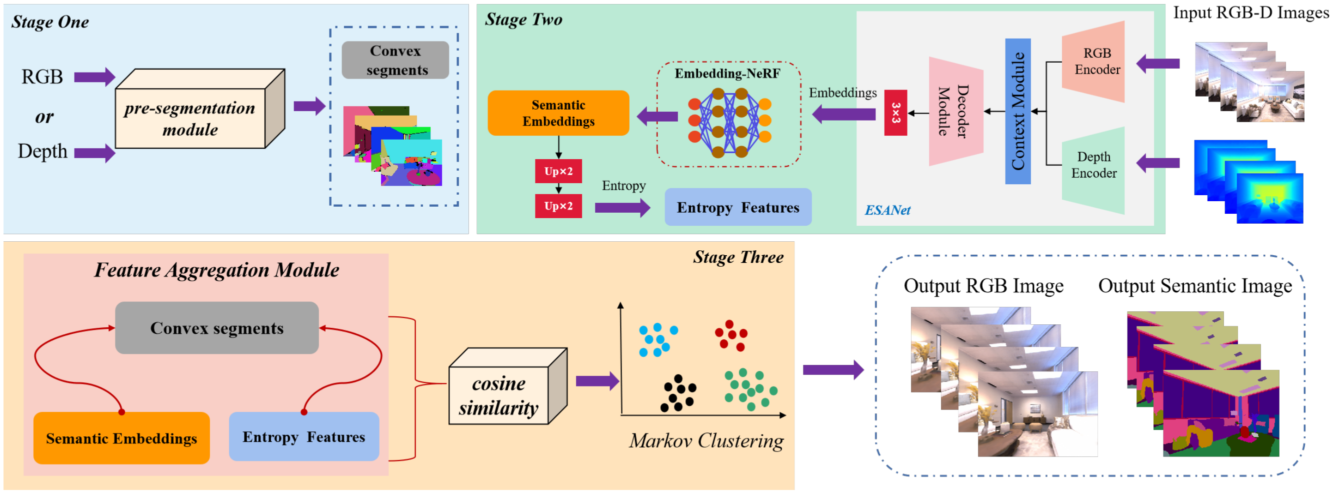

Figure 2: An overview of our method. **Stage One:** We first use the pre-segmentation module $\mathcal{F}$ to obtain convex sub-instance level segments s_i^j. The input of the module can be RGB or Depth, or RGB-D; **Stage Two:** using the pre-trained semantic segmentation model f_θ to obtain semantic embeddings and entropy features of multi-view images, then use our proposed Embedding-NeRF $F_\theta(P)$ to render these semantic embeddings; **Stage Three:** In the feature aggregation module, convex sub-instance level segments, entropy features, and semantic embeddings are associated for clustering to obtain the final semantic segmentation results (including known classes and novel classes)

method can segment all categories in the entire indoor scene with the help of implicit representations.

For classification, this can be understood as a two-part problem. First, high-dimensional descriptors for the items in question have to be found. Then, a clustering algorithm groups similar descriptors together. The established approach in representation learning is to learn a single good descriptor that can be clustered with KNN or k-means [Hamerly and Elkan, 2003]. K-means can be used with mini-batches and is differentiable, fast, and easy to implement. However, we argue that there are two big disadvantages: it requires a priori knowledge of the number of clusters k and only works in the space of a single descriptor. An alternative graph-based clustering algorithm like Markov clustering [Ye *et al.*, 2022] performs effective random walks for unsupervised clustering without pre-defined cluster numbers.

3 Method

In this section, we describe the proposed NeurNCD in detail. An overview of our approach is presented in Figure 2. In the first stage, we use a pre-segmentation module to generate sub-instance level segments (Second 3.1). In the second stage, we extract semantic embedding from RGB-D images under different viewpoints using a pre-trained semantic segmentation model: ESANet. We then propose an Embedding-NeRF model to fuse and render semantic embedding with spatial consistency characteristics and send the generated semantic embedding to the upsampling layer of ESANet to obtain entropy features. (Second 3.2). In the third stage, We first calculate the cosine similarity between sub-instance level segments after feature aggregation, and then associate known classes and discover novel classes through clustering (Second 3.3).

3.1 Pre-Segmentation Module

We provide a pre-segmentation module $\mathcal{F}$ for unsupervised discovery of novel classes, apply this module to the NYUv2 dataset and the Replica dataset, we denote the dataset as $\Theta = (I_i^{RGB}, I_i^D)$, obviously, I_i^{RGB} and I_i^D respectively represent an RGB image and its corresponding Depth image, where I_i^D is the input of the pre-segmentation module and i represents the current frame. Specifically, each incoming depth frame is divided into a set of sub-instance convex 3D segments using the geometry-based method described in [Furrer *et al.*, 2018], based on the idea that real-world objects have overall convex surface geometries. For example, a chair instance is a member of the chair class, which is further partitioned by $\mathcal{F}$ into chair legs, chair back, etc. At every depth image point, surface normals are first calculated. To determine the edges of the concave zone, angles between nearby normals are then compared, which are based on a local pixel neighbourhood and used to determine the local convexity of each pixel. The detection of significant depth discontinuities also makes use of large 3D distances between adjacent depth map vertices.

Lastly, the 3D distance measure and surface convexity are combined to generate a set of convex sub-instance level segments s_i^j in current frame i.

$$s_i^j = \mathcal{F}(I_i^D) \tag{1}$$

where j represent the j th sub-instance level segments. We denote the p th semantic class and q th instance in the i th frame as: $\mathcal{O}_i^p$ and $\mathcal{N}_i^q$, obviously, $s_i^j \in \mathcal{N}_i^q, \mathcal{N}_i^q \in \mathcal{O}_i^p$.

3.2 Embedding-NeRF and Feature Extraction

Embedding Extraction. we assume we are provided with a pre-trained semantic segmentation model f_θ extract the semantic embedding of each set of RGB-D images, which has

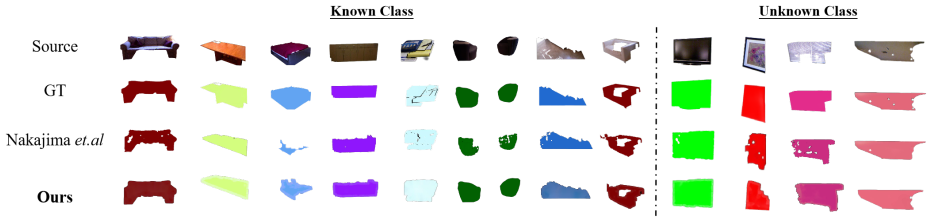

Figure 3: Quantitative results for known and unknown classes in the NYUv2 dataset. With the powerful label propagation and fusion capabilities of Embedding-NeRF, our method is very complete and smooth for each class segmentation, the baseline method relies on geometric segmentation results, and there are segmentation errors or incomplete phenomena.

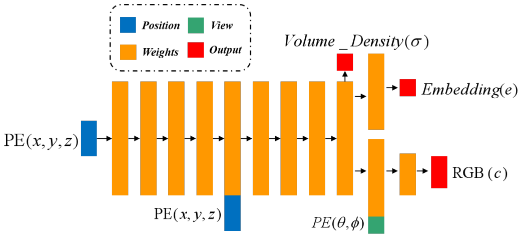

Figure 4: **Embedding-NeRF** 3D position (x, y, z) and viewing direction (θ, ϕ) are fed into the network after positional encoding (PE). Volume density σ and semantic embedding e are functions of 3D position while colours c additionally depend on viewing direction.

the parameters θ. The model is trained on the SUN RGB-D dataset, and only 9 of the classes are trained as known classes by fine-tuning the model, and the other 4 classes are used as novel classes. Apply the pre-trained model to the NYUv2 and Replica datasets, we use the feature layer before the semantic segmentation model softmax as the input of Embedding-NeRF, which is recorded here as: E_i, which is a high-dimensional vector of $N \times H \times W \times S$, in our paper, $S = 37$. So the semantic embedding extraction process can be expressed as:

$$E_i = f_\theta(\Theta) = f_\theta((I_i^{RGB}, I_i^D)) \tag{2}$$

Embedding-NeRF. NeRF [Mildenhall *et al.*, 2021] approximates volume rendering by numerical quadrature with hierarchical stratified sampling to determine the color of a single pixel. Within one hierarchy, if $r(t) = o + td$ is the ray emitted from the centre of projection of camera space through a given pixel, traversing between near and far bounds(t_n and t_f), then for selected K random quadrature points $\{t_k\}_K^{k=1}$ between t_n and t_f, the approximated expected colour is given by:

$$\hat{C}(r) = \sum_{k=1}^{K} \hat{T}(t_k)\alpha(\sigma(t_k)\delta_k)c(t_k) \tag{3}$$

where

$$\hat{T}(t_k) = exp(-\sum_{k'=1}^{k-1} \sigma(t_k)\delta_k) \tag{4}$$

where $\alpha(x) = 1 - exp(-x)$ and $\delta_k = t_{k+1} - t_k$ is the distance between adjacent sample points.

We now show how to extend NeRF to jointly encode appearance, geometry and embedding. As shown in Figure 4, we augment the original NeRF by appending an embedding renderer before injecting viewing directions into the MLP.

$$F_\theta(P) = (c, e, \sigma) \tag{5}$$

where F_θ is a MLP parameterised by θ; c, e and σ are the radiance, embedding logits and volume density at the 3D position $P = (x, y, z)$, respectively. The approximated expected embedding logits $\hat{E}(r)$ of a given pixel in the image plane can be written as:

$$\hat{E}(r) = \sum_{k=1}^{K} \hat{T}(t_k)\alpha(\sigma(t_k)\delta_k)e(t_k) \tag{6}$$

where $\hat{T}(t_k)$, $\alpha(x)$ and δ_k are consistent with the definitions in NeRF.

Embedding logits can then be transformed into multi-class probabilities through a softmax normalisation layer. We train the whole network from scratch under photometric loss L_p and embedding loss L_e:

$$L_p = \sum_{r \in R} \left[\left\| \hat{C}_c(r) - C(r) \right\|_2^2 + \left\| \hat{C}_f(r) - C(r) \right\|_2^2 \right] \tag{7}$$

$$L_e = D_{KL}\left(E(r) || \hat{E}(r)\right) = \sum_{n=1}^{N} E(r_n) \log \frac{E(r_n)}{\hat{E}(r_n)} \tag{8}$$

where R are the sampled rays within a training batch, and $C(r), \hat{C}_c(r)$ and $\hat{C}_f(r)$ are the ground truth, coarse volume predicted and fine volume predicted RGB colours for ray r,respectively.respectively. L_e is chosen as a KL-divergence loss to encourage the rendered embedding $\hat{E}(r)$ to be consistent with the embeddings extracted by the pre-trained model $E(r)$, whether these are ground-truth, noisy or partial observations. Hence the total training loss L is:

$$L = L_p + \lambda L_e \tag{9}$$

Entropy feature Extraction. In order to allow sub-instance level segments to have more features and achieve efficient clustering, so we also extracted the entropy feature, After the Embedding E_i obtained by fusion and rendering is sent to the last two upsampling modules of the ESANet network, the entropy features $\mathcal{U}_i^o$ are obtained. In current frame i, the entropy $\varepsilon_i \in \mathbb{R}$ is computed as follows:

$$\varepsilon_i = -\sum_{o \in \mathcal{O}} \mathcal{U}_i^o log \mathcal{U}_i^o \tag{10}$$

Table 1: Quantitative comparison on the NYUDv2 dataset. Supervised methods, unsupervised methods versus open set methods (ours).

Method	Classes in Training Dataset									Novel Classes				**mIoU**
	Bed	Book	Chair	Floor	Furn.	Obj.	Sofa	Table	Wall	Ceil.	Pict.	TV	Window	
[Seichter *et al.*, 2021]	49.62	25.08	40.67	49.85	53.74	21.11	42.55	43.36	55.62	-	-	-	-	-
[Nakajima *et al.*, 2018]	62.82	27.27	42.56	68.43	44.62	24.63	45.04	42.30	26.82	-	-	-	-	-
[Tateno *et al.*, 2015] + 3D Map	62.80	23.96	33.10	63.41	50.58	27.28	58.68	40.23	54.53	31.42	19.37	43.98	31.30	41.59
[Nakajima *et al.*, 2019]	64.22	22.28	41.79	67.38	56.15	28.61	49.31	40.95	63.18	29.30	**28.69**	52.20	53.92	46.05
Our	**69.23**	**29.82**	**58.63**	**69.67**	**60.11**	**32.18**	**58.16**	**48.25**	**69.28**	**31.92**	25.59	**59.38**	**53.95**	**51.24**

where $\mathcal{U}_i^o \in \mathbb{R}$ is the probability for the o th class in i th frame.

3.3 Feature Aggregation and Clustering

Feature Aggregation module. We use the feature aggregation module to aggregate sub-instance level segments and their corresponding semantic embedding and entropy features. Specifically, we concatenate the semantic embedding E_i^j and the entropy feature ε_i^j of the j th sub-instance level segment in the i th frame to generate the corresponding sub-instance level segment feature $\mathcal{H}_i^j$, where j represents the j th sub-instance level segment:

$$\mathcal{H}_i^j = E_i^j \oplus \varepsilon_i^j \tag{11}$$

where $\oplus$ means concatenate operation, $\mathcal{H}_i^j$ means a high-dimensional vector, its dimension is $N \times H \times W \times (S+1)$
Clustering based on sub-instance level segment feature.we compute the cosine similarity [Tao *et al.*, 2019] between sub-instance level segments s_i^j base on it feature $\mathcal{H}_i^j$, the cosine similarity is a measure of similarity based on the cosine of the angle between two nonzero vectors of an inner product space.

$$Similarity(s_i^m, s_i^n) = \frac{\mathcal{H}_i^m \mathcal{H}_i^n}{\|\mathcal{H}_i^m\| \, \|\mathcal{H}_i^n\|} \tag{12}$$

where $m \neq n$.

Through clustering, the instances belonging to the same semantic class are associated after the sub-instance level segments of the same instance are first associated to remove the “over-segmentation” result brought on by the previous pre-segmentation model. We specifically employ the Markov clustering algorithm(MCL) [Xu and Wunsch, 2005] because of the flexible number of clusters and computational cost. Since we could not find all clustering parameters in [Nakajima *et al.*, 2019] paper, so we hand-tune the parameters until we get a good result for kitchen_0004 in NYUv2 Dataset and use these settings (inflation = 12) for all scenes.

4 Experimental Evaluation

4.1 Datasets and Metrics

NYUv2. We evaluate our proposed method on the NYUv2 [Silberman *et al.*, 2012] dataset. According to the official guide, we first preprocessed the entire dataset with matlab. At the same time, Open3D [Zhou *et al.*, 2018] was used to solve the camera pose of this dataset. We used all the images in each scene as the training set , train a separate Embedding-NeRF for each scene to fuse and render embeddings, meanwhile, use the official split: 654 images for test. In all the experiments we resize the images to a resolution of 320 × 240 pixels.

Replica. Replica [Straub *et al.*, 2019] is a reconstruction-based 3D dataset of 18 high-fidelity scenes with dense geometry, HDR textures and semantic annotations. Zhi *et al.* use the Habitat simulator [Savva *et al.*, 2019] to render RGB colour images, depth maps and semantic labels from randomly generated 6-DOF trajectories similar to hand-held camera motions. We validate the performance of our method on unsupervised semantic segmentation using their open source simulated dataset, and compared with their supervised Label propagation results using partial annotations of 1%, 5% of pixels per class within frames.

Metric. We adopt the widely-used pixel classification accuracy (Acc.) and mean intersection over union (mIoU) as our metric.

4.2 Implementation Details

Networks. We use ESANet [Seichter *et al.*, 2021] with a ResNet34 NBt1D backbones as our semantic segmentation network. We train the model using a subset of classes and evaluate the system using whole classes in order to assess the proposed system’s capacity for class discovery. This makes it possible to quantitatively analyze both seen and unknown classes. Using the SUN RGBD training dataset [Song *et al.*, 2015], which comprises of 5,285 RGB-D images, we train the ESANet.We finetune the model using pre-selected 9 classes among the 13 classes defined in [Couprie *et al.*, 2013]. The selected classes and the entire classes are shown in Table 1. In Section 4.4, the suggested approach and the methods for comparison [Nakajima *et al.*, 2018] and [Nakajima *et al.*, 2019] both employ the same trained model.

The above model is trained on a single 3090Ti GPU with 24GB memory. The batch size of rays is set to 1024 and the neural network using the Adam optimiser [Kingma and Ba, 2014] with a learning rate of 5e-4 for 200,000 iterations.

4.3 Baselines

As there are no previous works that use the neural radiation field to tackle the discovery of novel classes and unsupervised semantic segmentation problems, we compare our proposed method to the three most closely related approaches.

The first work was put forth by Nakajima *et al.*, who were among the pioneers in showing how semantic scene understanding can recognize novel things. We implement the method using the framework of [Blum *et al.*, 2022]. Since we could not find all clustering parameters in [Nakajima *et al.*, 2019], we use the parameter optimisation from [Blum *et al.*,

Table 2: Quantitative comparison of our unsupervised semantic segmentation results with two methods for supervised semantic segmentation via sparse annotation

Method	Label Propagation	Metrics		
	# Labelling per Class	mIoU	Avg Acc	Total Acc
Semantic NeRF [Zhi *et al.*, 2021]	1 %	68.2	82.7	84.5
	5 %	72.5	87.3	88.1
	100 %	96.15	97.65	98.92
iLabel [Zhi *et al.*, 2022]	20 click	48.0	-	-
	40 click	64.0	-	-
	60 click	72.0	-	-
Our	-	**73.8**	**88.4**	**89.7**

2022] on the inflation and η parameter of the MCL clustering for every scene.

Our method also has a breakthrough in unsupervised semantic segmentation, so we compared it with two supervised semantic segmentation methods in the neural radiation field: Semantic-NeRF and iLabel, although it is not an unsupervised segmentation method in the strict sense, it well reflects the characteristics of NeRF information fusion and dissemination, which is the main reason why we propose the Embedding-NeRF for novel class discovery and unsupervised semantic segmentation. The generality of our method is further verified by quantitative comparison and analysis of our fully unsupervised results with the results obtained under the different numbers of clicks in this method [Zhi *et al.*, 2022].

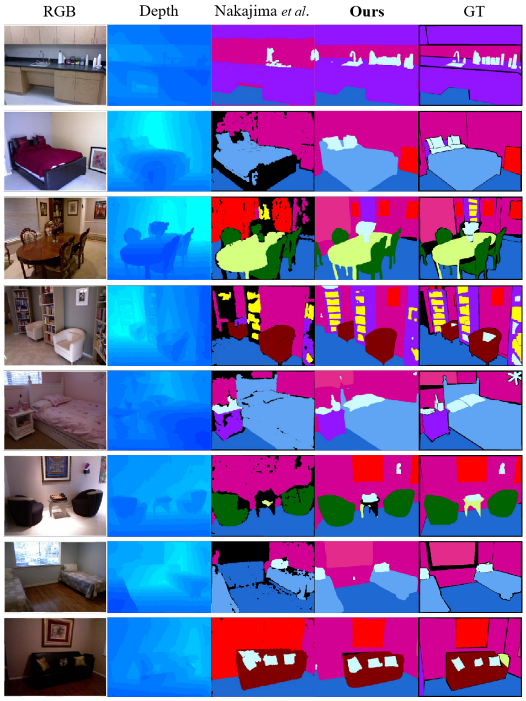

Figure 5: Novel class discovery results on the NYUv2 dataset. The third column is the result obtained by the method proposed by Nakajima *et al.* and the fourth column is the result obtained by our method. By rendering semantic embedding with spatial consistency, our results have fewer outliers and noise points.

4.4 Results

We use experiments to statistically and subjectively show how well the suggested strategy performs. Utilizing the test set of the NYUv2 dataset [Silberman *et al.*, 2012] for quantitative comparison, we calculate the intersection over union (IoU) and display the results in Table 1.

In Table 1, we compare the proposed method with two fully supervised methods and two unsupervised methods. We used two fully supervised approaches for 2D images: one is traditional semantic segmentation [Seichter *et al.*, 2021] and another is SLAM mapping for semantic segmentation [Nakajima *et al.*, 2018]. Meanwhile, we chose one state-of-the-art semantic mapping technique for the unsupervised methods [Nakajima *et al.*, 2019]. In this work, it is compared to a previous incremental 3D geometric segmentation method [Tateno *et al.*, 2015] from which the pre-segmentation module of our work was derived, so we also compared it to our work.

Obviously, these fully supervised methods can only predict the 9 classes in the training dataset and are unable to discover novel classes. Against the other two unsupervised methods, our method outperforms this work by a large margin compared to [Tateno *et al.*, 2015] for both known and novel classes; according to the quantitative results, the mIoU of our method is about one time higher than that method (from 41.59 to 51.24). Additionally, we observe that the [Nakajima *et al.*, 2019] method can identify certain novel classes but heavily relies on feature extraction and updating results and is unable to combine multi-view visual features, leading to incorrect segmentation and producing noise and outliers.

Our method, which depends on Embedding-NeRF's fusion and renders capabilities, can correct and fill in the "incomplete classes" and "outlier classes" created by incorrect geometric segmentation. From the quantitative results, compared with this state-of-the-art method, the mIoU obtained by our method is improved from 46.05 to 51.24. At the same time, in the known class part, 6 classes in our results have achieved significant improvement (bed, book, chair, sofa, table, wall). In the unknown class part, there are 3 classes whose results exceed the state-of-the-art methods. But the picture class has not been significantly improved. This is because our pre-segmentation module only uses the depth image, and the Nakajima *et al.* method uses depth and colour for segmentation in the segmentation module, thus our method achieves slightly lower results than theirs on classes with poor

Table 3: Ablation study for our method on NYUv2 dataset.

Components					classes in training dataset									novel classes				mIoU
PSM	PSSM	EF	SE	MCL	bed	book	chair	floor	furn.	obj.	sofa	table	wall	ceil.	pict.	TV	wind.	
✓					48.77	20.76	35.23	49.89	45.25	20.86	40.18	38.22	51.80	27.76	15.93	45.88	39.23	36.91
	✓				51.67	26.19	41.55	50.79	54.24	22.18	43.75	44.86	57.88	-	-	-	-	-
✓	✓	✓		✓	53.55	27.10	44.79	55.93	59.07	25.12	45.98	45.19	59.25	31.89	19.64	49.24	42.82	43.04
✓	✓	✓	✓	✓	**69.23**	**29.82**	**58.63**	**69.67**	**60.11**	**32.18**	**58.16**	**48.25**	**69.28**	**31.92**	**25.59**	**59.38**	**53.95**	**51.24**

geometric features.

In Table 2, we use the sparse label annotation propagation experiments in Semantic-NeRF [Zhi *et al.*, 2021] and iLabel [Zhi *et al.*, 2022] as a baseline to verify the advancement of our method on the Replica dataset. We first pretrained ESANet using the SUN-RGBD dataset and achieved 41.98% mIoU on the validation sets of 40 classes. The pretrained model is then used to extract the semantic embedding of each scene in the Replica dataset, which are then fused and rendered using Embedding-NeRF. At the same time, the rendered embeddings are sent to the upsampling layer of ESANet to obtain entropy features. Finally, through the feature aggregation module and clustering associated sub-instance level segments belonging to the same class, But unlike semantic-NeRF and iLabel's supervised or weakly supervised semantic segmentation method, our method is completely unsupervised. In Semantic-NeRF, using partial annotations of 1% or 5% or 100% of pixels per class within frames, one achieves semantic segmentation, respectively. In iLabel, compared with the semantic segmentation results after 20, 40, and 60 interactive clicks, the semantic segmentation results are significantly better than the baseline, obtained using state-of-the-art results.

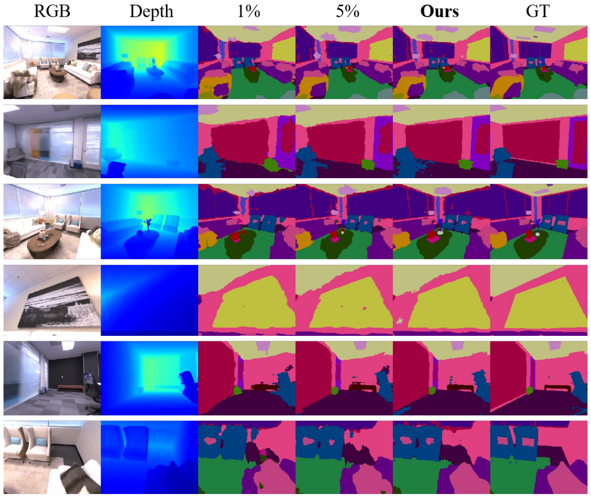

Figure 6: Unsupervised Semantic Segmentation Results. Columns 1 and 2 are the input RGB-D values, The third and fourth columns are the results of sparse annotation 1%, 5%, column 5 is the result of our method, and column 6 is the semantic ground truth.

4.5 Ablation Studies

In this section, we evaluate the effects of components such as the pre-segmentation module (PSM), pre-trained semantic segmentation model (PSSM), semantic embedding (SE), entropy features (EF), and Markov clustering (MCL). The results of the ablation study are shown in Table 3.

By comparing the two cases of PSM and PSSM, it can be seen that the segmentation effect of the pre-trained model on known categories is significantly better than that of pure geometric segmentation, but unknown categories cannot be found and simple geometric segmentation can found in unknown category, but it is based on the assumption that objects in the real world show overall convex surface geometry, so for objects with poor convexity, the segmentation results are also poor, and geometric segmentation also has over-segmentation, which depends heavily on clustering algorithms to associate segments of the same category.

In order to verify the contribution of entropy features and semantic embedding in the feature aggregation module to our method, we set up two cases of EF and EF+SE in our method for ablation comparison experiments. The results show that the results produced by only adding entropy features to segments are significantly lower. The result after adding semantic embedding is that the entropy feature is discrete and does not have spatial consistency. At the same time, it can only assign features to known classes and has no obvious contribution to the discovery of unknown classes. The semantic embedding obtained by Embedding-NeRF has spatial consistency and continuity.

At the same time, in the feature aggregation module, features and embeddings can be assigned to both known and unknown classes. Therefore, after clustering, you can use the sub-instance level segments with the same features and embeddings to better complete the task of segmenting known classes and discovering novel classes.

5 Conclusion

In this work, we propose a novel method for novel class discovery and unsupervised semantic segmentation using implicit neural rendering. Different from the previous work of explicitly building a 3D map and looking for relevant regions to discover new classes, we propose Embedding-NeRF, the first method of saving and aggregating semantic embedding in NeRF renderings for the discovery of novel classes. It directly renders and generates semantic embedding with spatial consistency, which can help in the discovery of novel classes. We also concatenate and aggregate the sub-instance level segments, semantic embedding, and entropy features using the feature aggregation module, and the sub-instance level segments that belong to the same class are then connected using clustering, the whole process is unsupervised. At the same time, our method not only associates known semantic classes but also discovers novel classes and achieves state-of-the-art results on two public datasets, NYUv2 and Replica.

Acknowledgments

We thank all anonymous reviewers for their helpful comments.

References

[Blum *et al.*, 2022] Hermann Blum, Marcus G Müller, Abel Gawel, Roland Siegwart, and Cesar Cadena. Scim: Simultaneous clustering, inference, and mapping for open-world semantic scene understanding. *arXiv preprint arXiv:2206.10670*, 2022.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Chen *et al.*, 2022] Zheng Chen, Chen Wang, Yuan-Chen Guo, and Song-Hai Zhang. Structnerf: Neural radiance fields for indoor scenes with structural hints. *arXiv preprint arXiv:2209.05277*, 2022.

[Couprie *et al.*, 2013] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. 2013.

[Frey *et al.*, 2022] Jonas Frey, Hermann Blum, Francesco Milano, Roland Siegwart, and Cesar Cadena. Continual adaptation of semantic segmentation using complementary 2d-3d data representations. *IEEE Robotics and Automation Letters*, 7(4):11665–11672, 2022.

[Fu *et al.*, 2022] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022.

[Furrer *et al.*, 2018] Fadri Furrer, Tonci Novkovic, Marius Fehr, Abel Gawel, Margarita Grinvald, Torsten Sattler, Roland Siegwart, and Juan Nieto. Incremental object database: Building 3d models from multiple partial observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6835–6842. IEEE, 2018.

[Gupta *et al.*, 2015] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149, 2015.

[Hamerly and Elkan, 2003] Greg Hamerly and Charles Elkan. Learning the k in k-means. *Advances in neural information processing systems*, 16, 2003.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014.

[Li *et al.*, 2021] Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Liu *et al.*, 2018] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang 0001, and Ming-Ming Cheng. Del: Deep embedding learning for efficient image segmentation. In *IJCAI*, volume 864, page 870, 2018.

[Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[Müller *et al.*, 2022] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[Nakajima *et al.*, 2018] Yoshikatsu Nakajima, Keisuke Tateno, Federico Tombari, and Hideo Saito. Fast and accurate semantic mapping through geometric-based incremental segmentation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 385–392. IEEE, 2018.

[Nakajima *et al.*, 2019] Yoshikatsu Nakajima, Byeongkeun Kang, Hideo Saito, and Kris Kitani. Incremental class discovery for semantic segmentation with rgbd sensing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 972–981, 2019.

[Savva *et al.*, 2019] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

[Seichter *et al.*, 2021] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[Song *et al.*, 2015] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

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

[Tao *et al.*, 2019] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. Multi-view saliency-guided clustering for image cosegmentation. *IEEE Transactions on Image Processing*, 28(9):4634–4645, 2019.

[Tateno *et al.*, 2015] Keisuke Tateno, Federico Tombari, and Nassir Navab. Real-time and scalable incremental segmentation on dense slam. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4465–4472. IEEE, 2015.

[Vanholder, 2016] Han Vanholder. Efficient inference with tensorrt. In *GPU Technology Conference*, volume 1, page 2, 2016.

[Vora *et al.*, 2021] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.

[Wang *et al.*, 2022a] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. *arXiv preprint arXiv:2203.17261*, 2022.

[Wang *et al.*, 2022b] Xiaoyang Wang, Jimin Xiao, Bingfeng Zhang, and Limin Yu. Card: Semi-supervised semantic segmentation via class-agnostic relation based denoising. In *Proc. IJCAI*, pages 1451–1457, 2022.

[Xu and Wunsch, 2005] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[Xu *et al.*, 2022] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

[Ye *et al.*, 2022] Minxiang Ye, Yifei Zhang, Shiqiang Zhu, Anhuan Xie, and Dan Zhang. Deep markov clustering for panoptic segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2380–2384. IEEE, 2022.

[Zhao *et al.*, 2022] Yuyang Zhao, Zhun Zhong, Nicu Sebe, and Gim Hee Lee. Novel class discovery in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2022.

[Zhi *et al.*, 2021] 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.

[Zhi *et al.*, 2022] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Revealing objects in neural fields. *IEEE Robotics and Automation Letters*, 2022.

[Zhou *et al.*, 2018] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.