



# NeurNCD: Novel Class Discovery via Implicit Neural Representation

Junming Wang\*

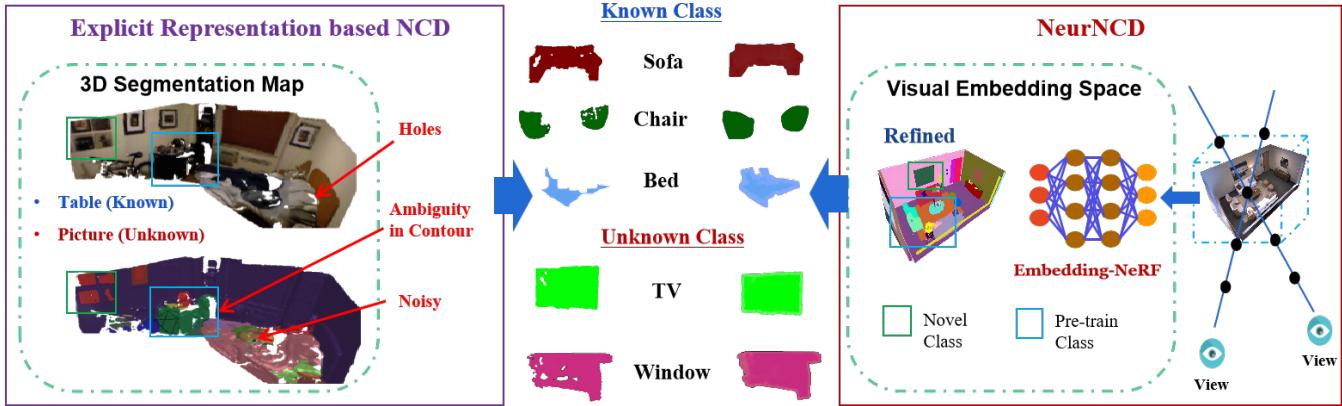
jmwang@cs.hku.hk

The University of Hong Kong  
Hong Kong SAR, China

Yi Shi

21120237@bjtu.edu.cn

Beijing Jiaotong University  
Beijing, China



**Figure 1:** NeurNCD leverages implicit neural representation, replacing traditional explicit 3D segmentation maps[19], to enhance the accuracy of novel class discovery. Specifically, the meticulously designed Embedding-NeRF model employs KL divergence, achieving the transfer and association of 2D-3D features while producing semantic embedding and entropy by aggregating information from multiple views. Then by integrating with other key components, i.e., feature query, feature modulation and clustering, to ultimately reconstruct refined, low-noise, and hole-free images and 3D structures.

## ABSTRACT

Discovering novel classes in open-world settings is crucial for real-world applications. Traditional explicit representations, such as object descriptors or 3D segmentation maps, are constrained by their discrete, hole-prone, and noisy nature, which hinders accurate novel class discovery. To address these challenges, we introduce NeurNCD, the first versatile and data-efficient framework for novel class discovery that employs the meticulously designed Embedding-NeRF model combined with KL divergence as a substitute for traditional explicit 3D segmentation maps to aggregate semantic embedding and entropy in visual embedding space. NeurNCD also integrates several key components, including feature query, feature modulation and clustering, facilitating efficient feature augmentation and information exchange between the pre-trained semantic segmentation network and implicit neural representations. As a

result, our framework achieves superior segmentation performance in both open and closed-world settings without relying on densely labelled datasets for supervised training or human interaction to generate sparse label supervision. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art approaches on the NYUv2 and Replica datasets.

## CCS CONCEPTS

- Computing methodologies → Image-based rendering; Computer graphics; Image manipulation.

## KEYWORDS

Neural Radiation Field, Visual Embedding Space, Novel Class Discovery, Feature Fusion, Novel View Synthesis

## ACM Reference Format:

Junming Wang and Yi Shi. 2024. NeurNCD: Novel Class Discovery via Implicit Neural Representation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24)*, June 10–14, 2024, Phuket, Thailand. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3652583.3658073>

## 1 INTRODUCTION

The swift advancements in computer vision and robotics have transitioned from "Supervised AI" to "Embodied AI," whereby AI algorithms and agents can learn through interactions with their environment, adopting a human-like egocentric perspective. However, most existing perception algorithms[3, 11, 22, 32, 33] operate in a

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '24, June 10–14, 2024, Phuket, Thailand

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0619-6/24/0...\$15.00  
<https://doi.org/10.1145/3652583.3658073>

closed-world setting and are trained to segment a limited number of semantic classes, which do not adequately address the needs of Embodied AI applications in constantly changing and open environments. In open-world perception scenarios, novel classes continually emerge, but a perception model trained on a limited number of semantic classes may either treat a novel class as background or misclassify it as one of the known objects[38]. Ideally, an Embodied AI system should discover and incrementally learn to recognize novel classes through interactions with the environment. This challenge, known as novel class discovery, has garnered significant interest within the research community and has crucial real-world implications.

Humans are adept at recognizing novel classes by identifying consistent features within their surroundings. Drawing on this observation, Nakajima et al. proposed a method [19] for aggregating spatially consistent features by explicitly constructing a 3D segmentation map and clustering to discover novel classes (in Fig.1). Although this method demonstrates the feasibility of the novel class discovery task and presents a solution framework, it suffers from several limitations related to noise, hole-prone, overlapping, and mapping errors, which significantly impact the accuracy of discovery of novel classes. Furthermore, the method introduces manually designed update strategies for continuously updating semantic features and entropy, leading to additional hyperparameters and suboptimal strategies.

Recently, implicit neural representations have gained considerable attention due to their exceptional performance in 3D scene modelling from novel viewpoints. Thus, we investigate whether implicit neural representations can replace traditional explicit 3D segmentation maps to enhance novel class discovery accuracy. In contrast to a conventional point cloud or voxel-based 3D map, implicit neural representations utilize a multilayer perceptron (MLP) to represent a 3D map. They can reconstruct low-noise and hole-free 3D structures by aggregating information from different perspectives. Moreover, implicit neural representations have been employed to aggregate semantic labels[41], achieving state-of-the-art performance. However, this method relies on dense annotation and cannot discover novel classes.

The above limitations show that implementing implicit neural representations for discovering novel classes demands addressing two significant challenges. The first challenge arises from the fundamental differences between implicit and explicit representations. Explicit representations (e.g., point clouds or voxels) enable direct operations like clustering to discover novel classes, whereas implicit representations require information extraction through queries for the same purpose. Consequently, it is crucial to devise a novel class discovery framework that accommodates the unique operational mode of implicit representations while considering complex factors, i.e., query-based information extraction.

The second challenge involves developing suitable methods for semantic feature aggregation and an update strategy tailored specifically for novel class discovery. Although NeRF's inherent feature aggregation capabilities allow for direct implementation of semantic feature aggregation and updates through error backpropagation, such as Semantic NeRF[41] employs softmax loss for aggregating semantic labels. However, this approach is unsuitable for novel class discovery, as softmax loss generally categorizes unknown classes as

background classes[38]. This challenge highlights the necessity for designing an implicit representation specifically adapted for novel class discovery, potentially involving the exploration of new loss functions and the enhancement of feature extraction methodologies to overcome these limitations.

Tackling the challenges mentioned earlier, we present NeurNCD, the first versatile and data-efficient approach for novel class discovery using neural radiance fields and feature modulation, applicable to both "open set" and "closed set" scenarios.

In general, the main contributions of this work are:

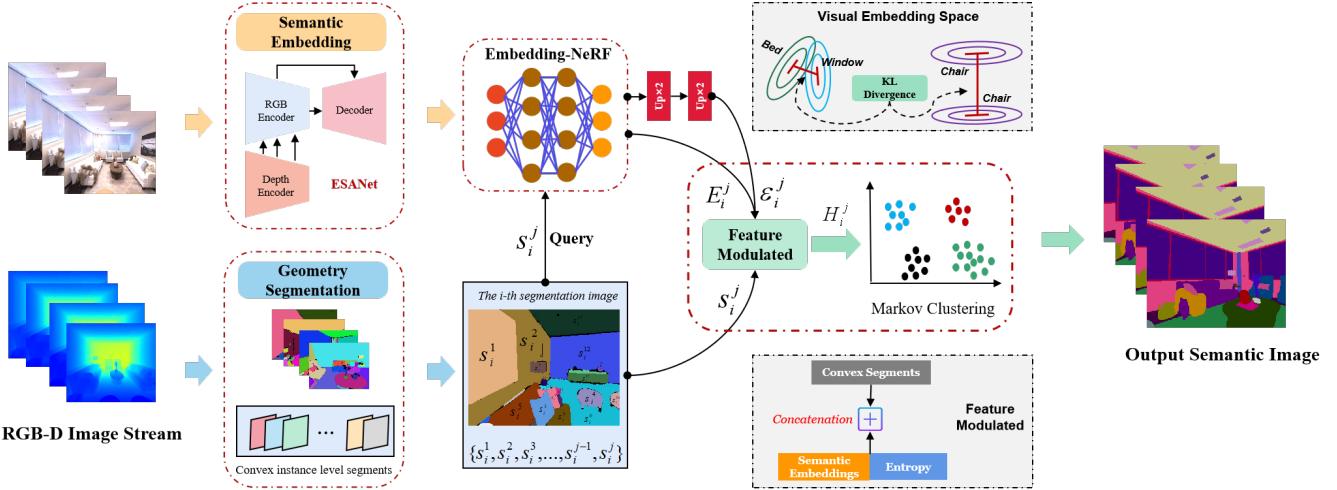
- We address the challenge of discovering novel classes in implicit neural representation tasks by proposing a novel framework named NeurNCD. This method is versatile and data-efficient, thanks to the advantages of implicit representation in terms of refinement, low noise and continuity, the accuracy of new class discovery is significantly improved.
- Our method introduces the key component, Embedding-NeRF, which is specifically designed for novel class discovery tasks, and enables the replacement of traditional explicitly constructed 3D segmentation maps. Through the minimization of the Kullback-Leibler (KL) divergence, it generates semantic embeddings and entropy, thus bestowing a considerable advantage in the realm of novel class discovery.
- Extensive experiments demonstrate that our method outperforms 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.

## 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. In recent years, substantial progress has been achieved in the field of supervised semantic segmentation[14, 16], 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 [29]. 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.



**Figure 2: An overview of our method.** For input RGB-D image, the pre-trained semantic segmentation network  $f_\theta$  to extract semantic embedding, and then use our proposed Embedding-NeRF  $F_\theta(P)$  model to generate globally consistent semantic embedding and entropy. Meanwhile, we leverage the geometric segmentation  $\mathcal{F}$  to obtain a set of convex sub-instance-level segments  $\{s_i^1, s_i^2, s_i^3, \dots, s_i^{j-1}, s_i^j\}$ . These segments will then go to the output of Embedding-NeRF to query the corresponding semantic embedding  $E_i^j$  and entropy  $\varepsilon_i^j$ . Finally, in the feature modulated, the three parts were concatenated and clustered which can obtain the final semantic segmentation results (including known classes and novel classes).

## 2.2 Radiance Field-based Scene Representations

Our work on discovering novel classes and unsupervised semantic segmentation build on neural radiance fields (NeRF) [17], which represent a scene using a multi-layer perceptron (MLP) that maps positions and directions to densities and radiances. The following work [1, 5, 8, 10, 18, 28, 31, 35, 36] 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[30], 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 [41] and [42]. 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 research the performance of NeRF fusion and rendering of higher-dimensional image embeddings.

## 2.3 Novel Class Discovery and Clustering

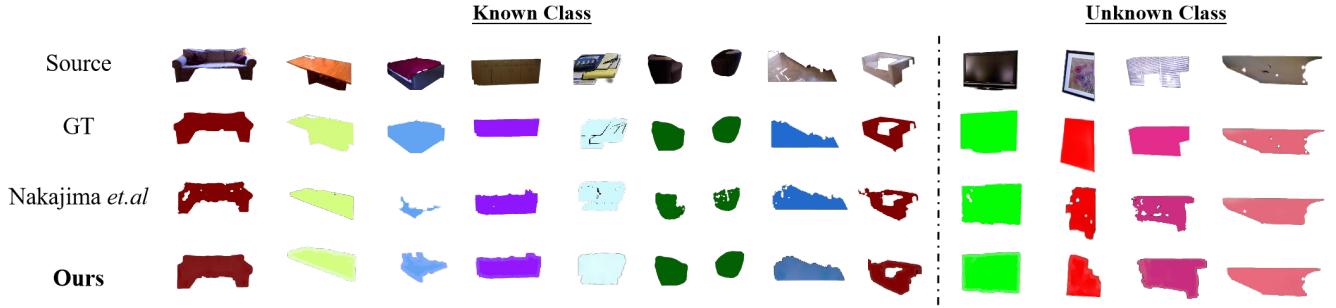
For novel class discovery, Zhao et al. proposed a method to discover novel classes with the help of a saliency detection model and use an entropy-based uncertainty modelling and self-training (EUMS) framework to overcome noisy pseudo-labels, further improving the model's performance on the novel classes. But their method can only segment and discover a limited number of salient categories, while our 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 [12]. 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 [39] performs effective random walks for unsupervised clustering without pre-defined cluster numbers.

## 3 METHOD

In this section, we present NeurNCD, a method specifically designed for novel class discovery comprised of several key components. An overview of our approach is shown in Figure 2.

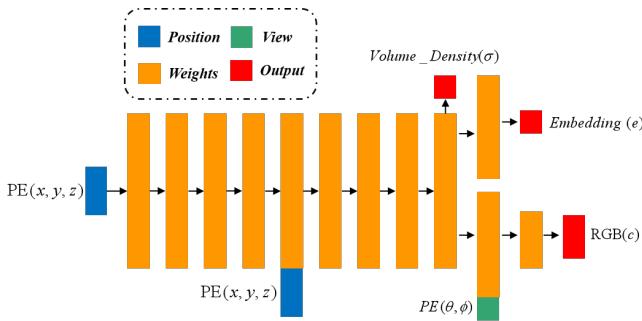
Firstly, to enhance the accuracy of novel class discovery and replace traditional explicit 3D segmentation maps, we delve into the visual embedding space to decode valuable features for semantic classes, namely, semantic embedding and entropy. To achieve this, we introduce the **Embedding-NeRF model**, which employs KL divergence[38] to facilitate the migration and aggregation of 2D-3D features and then generates the above features. Meanwhile, minimizing KL divergence loss achieves multiple goals, such as reducing distances between genuine prototypes, increasing distances among incorrect prototypes, and effectively capturing unknown class feature distributions – an advantage absent in cross-entropy loss. Moreover, Embedding-NeRF also generates entropy as an uncertainty measure for semantic classes, offering robust supplementary



**Figure 3: Quantitative results for known and unknown classes in the NYUv2 dataset. With the powerful feature 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.**

information for novel class discovery (Section 3.1). Upon acquiring globally consistent semantic embeddings and entropy, we adopt the **Geometric Segmentation** method from [19] to segment depth images, dividing both known and new classes into a set of convex sub-instance-level segments (Section 3.2). However, solely relying on segments’ geometric features is not enough for novel class discovery due to the absence of globally consistent semantic information. As a result, we fuse semantic embeddings and entropy with these segments to enhance the features. Specifically, we query the corresponding semantic embeddings and entropy for the segments from the Embedding-NeRF output and concatenate them during the **Feature Modulation**, ensuring that segments of the same class share identical semantic embeddings and entropy (Section 3.3). Ultimately, we achieve known class segmentation and novel class discovery using **Markov clustering** based on cosine similarity (Section 3.4).

### 3.1 Embedding-NeRF



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

In the context of novel class discovery tasks, semantic embedding and entropy [4, 15, 34] play a crucial role in enhancing the accuracy

and robustness of semantic classes. Consequently, we draw inspiration from the burgeoning field of implicit neural representation and propose the Embedding-NeRF model as a substitute for the traditional explicit 3D segmentation map, and generate these two features using this model.

**Extract Semantic Embedding as Input.** Utilizing a pre-trained semantic segmentation network  $f_\theta$  with parameters  $\theta$ , we extract the semantic embedding of each set of RGB-D images, denoted as  $\Theta = (I_i^{RGB}, I_i^D)$ , where  $I_i^{RGB}$  and  $I_i^D$  correspond to an RGB image and its associated Depth image, respectively. Specifically, applying the segmentation network to the RGB-D datasets, we remove the network’s classification layer and use the feature layer preceding the softmax layer as the input for Embedding-NeRF, denoted as  $E_i$ , a high-dimensional vector with dimensions of  $N \times H \times W \times S$ , where  $S = 37$ . Consequently, the semantic embedding extraction process can be expressed as:

$$E_i = f_\theta(\Theta) = f_\theta((I_i^{RGB}, I_i^D)) \quad (1)$$

**Embedding-NeRF.** NeRF [17] 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  $\{t_k\}_{K=1}^{K=1}$  between  $t_n$  and  $t_f$ , the approximated expected colour is given by:

$$\hat{C}(r) = \sum_{k=1}^K \hat{T}(t_k) \alpha(\sigma(t_k) \delta_k) c(t_k) \quad (2)$$

where

$$\hat{T}(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_k) \delta_k\right) \quad (3)$$

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.

$$F_\theta(P) = (c, e, \sigma) \quad (4)$$

**Table 1: Quantitative comparison on the NYUv2 dataset. Supervised methods, unsupervised methods versus our methods.**

Method	Classes in Training Dataset										Novel Classes				mIoU
	Bed	Book	Chair	Floor	Furn.	Obj.	Sofa	Table	Wall	Ceil.	Pict.	TV	Window		
[22]	49.62	25.08	40.67	49.85	53.74	21.11	42.55	43.36	55.62	-	-	-	-		
[20]	62.82	27.27	42.56	68.43	44.62	24.63	45.04	42.30	26.82	-	-	-	-		
[27] + 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	
[19]	64.22	22.28	41.79	67.38	56.15	28.61	49.31	40.95	63.18	29.30	<b>28.69</b>	52.20	53.92	46.05	
<b>NeurNCD</b>	<b>69.23</b>	<b>29.82</b>	<b>58.63</b>	<b>69.67</b>	<b>60.11</b>	<b>32.18</b>	<b>58.86</b>	<b>48.25</b>	<b>69.28</b>	<b>31.92</b>	25.59	<b>59.38</b>	<b>53.95</b>	<b>51.29</b>	

where  $F_\theta$  is a MLP parameterised by  $\theta$ ;  $c, e$  and  $\sigma$  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:

$$\hat{E}(r) = \sum_{k=1}^K \hat{T}(t_k) \alpha(\sigma(t_k) \delta_k) e(t_k) \quad (5)$$

where  $\hat{T}(t_k)$ ,  $\alpha(x)$  and  $\delta_k$  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  $L_e$ :

$$L_p = \sum_{r \in R} \left[ \|\hat{C}_c(r) - C(r)\|_2^2 + \|\hat{C}_f(r) - C(r)\|_2^2 \right] \quad (6)$$

$$L_e = D_{KL} \left( E(r) || \hat{E}(r) \right) = \sum_{n=1}^N E(r_n) \log \frac{E(r_n)}{\hat{E}(r_n)} \quad (7)$$

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  $r$ .  $L_e$  is chosen as a KL-divergence loss [38] 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:

$$L = L_p + \lambda L_e \quad (8)$$

In addition, since entropy can reflect the uncertainty of each semantic class, which is of great benefit for novel class discovery, so we use Embedding-NeRF to generate it from the visual embedding space. Specifically, after the semantic embedding  $E_i$  obtained by fusion is sent to the two upsampling modules, the entropy  $\mathcal{U}_i^o$  are obtained. In current frame  $i$ , the entropy  $\varepsilon_i \in \mathbb{R}$  is computed as follows:

$$\varepsilon_i = - \sum_{o \in O} \mathcal{U}_i^o \log \mathcal{U}_i^o \quad (9)$$

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

### 3.2 Geometric Segmentation

When the two types of features were generated by Embedding-NeRF, inspired by the work of [9], We use geometric segmentation to segment the input depth image into a set of convex sub-instance-level segments, this process also can be regarded as an extract

translation/rotation-invariant and noise-robust geometric features for known classes and novel classes. We denote geometric segmentation as  $\mathcal{F}$  and apply it to the depth image, where  $I_i^D$  is the input of the method and  $i$  represents the current frame.

Specifically, each incoming depth frame is divided into a set of convex sub-instance-level segments using the geometry-based method described in [9], based on the idea that real-world objects have overall convex surface geometries. For example, a chair instance belonging to the chair class undergoes further segmentation into components such as chair legs and chair back. At every depth image, surface normals are initially calculated, followed by a comparison of angles between adjacent normals to identify the edges of concave zones. This process leverages local pixel neighbourhoods to ascertain each pixel's local convexity. Additionally, the detection of significant depth discontinuities capitalizes on the large 3D distances between neighbouring depth map vertices. Ultimately, the 3D distance measure and surface convexity amalgamate to generate a set of convex sub-instance level segments  $\{s_i^1, s_i^2, s_i^3, \dots, s_i^{j-1}, s_i^j\}$  in the current frame  $i$ .

$$\{s_i^1, s_i^2, s_i^3, \dots, s_i^{j-1}, s_i^j\} = \mathcal{F}(I_i^D) \quad (10)$$

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:  $O_i^p$  and  $N_i^q$ , obviously,  $s_i^j \in N_i^q, N_i^q \in O_i^p$ .

However, although the segments of all semantic classes are obtained, only relying on the geometric features of the segments themselves, it is not possible to complete class discovery through clustering. Next, we use these segments as carriers to query the corresponding semantic embedding  $E_i^j$  and entropy  $\varepsilon_i^j$  in the output of Embedding-NeRF. This process makes us not only obtained the segmentation fragments of each semantic class but also obtained the corresponding globally consistent features, which is of great benefit for novel class discovery.

### 3.3 Feature Modulation

After querying the corresponding features for sub-instance-level segments, we must incorporate these features into the segments to facilitate capturing more semantic information and class uncertainty information. The addition of these globally consistent features ameliorates issues of "over-segmentation" and "misclassification," ultimately yielding more accurate known and novel classes through clustering.

**Feature modulation.** Initially, we employ the  $j$ th sub-instance-level segment in the  $i$ th frame image to access the output of Embedding-NeRF, subsequently querying to obtain the semantic embedding and entropy attributable to this particular segment. Following this, in the feature modulation (in Algorithm 1), we concatenate the semantic embedding and entropy of the  $j$ th segment, denoted as  $E_i^j$  and  $\varepsilon_i^j$ , respectively, to generate the final comprehensive feature,  $\mathcal{H}_i^j$ , of the segment:

$$\mathcal{H}_i^j = E_i^j \oplus \varepsilon_i^j \quad (11)$$

where  $\oplus$  means concatenate operation,  $\mathcal{H}_i^j$  means a high-dimensional vector, its dimension is  $N \times H \times W \times (S + 1)$ .

---

#### Algorithm 1 Feature Modulation

---

```

1: Input: a set of Convex Instance-level Segment  $S$ , Semantic
   Embedding  $E$ , Entropy  $\varepsilon$ .
2: Output: Final segment feature  $\mathcal{H}$ 
3: function FEATURE_MODULATION( $S, E, \varepsilon$ )
4:    $N \leftarrow \text{length}(S)$ 
5:    $H, W \leftarrow \text{shape}(S[0])$ 
6:    $S_{emb} \leftarrow \text{shape}(E)[-1]$ 
7:    $\mathcal{H} \leftarrow \text{zeros}((N, H, W, S_{emb} + 1))$ 
8:   for  $i = 0$  to  $N - 1$  do
9:      $mask \leftarrow S[i]$ 
10:     $embedding \leftarrow E[mask]$ 
11:     $entropy \leftarrow \varepsilon[mask]$ 
12:     $combined \leftarrow \text{concat}((embedding, entropy), \text{axis} = -1)$ 
13:    for  $h = 0$  to  $H - 1$  do
14:      for  $w = 0$  to  $W - 1$  do
15:        if  $mask[h, w] == 1$  then
16:           $\mathcal{H}[i, h, w, : S_{emb} + 1] \leftarrow combined[h, w]$ 
17:        end if
18:      end for
19:    end for
20:  end for
21:   $\mathcal{H} \leftarrow \text{reshape}(\mathcal{H}, (N, -1))$ 
22:  return  $\mathcal{H}$ 
23: end function

```

---

### 3.4 Markov clustering

**Markov Clustering Based on Cosine Similarity.** we compute the cosine similarity [26] 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.

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

where  $m \neq n$ .

Through clustering, the process initially aggregates sub-instance level segments belonging to a single instance, effectively mitigating the "over-segmentation" issue resulting from geometric segmentation. Subsequently, instances corresponding to the same semantic class are clustered together, which not only resolves over-segmentation but also facilitates the discovery of novel classes.

Specifically, we employ the Markov clustering algorithm(MCL) [37] because of the flexible number of clusters and computational cost. As we were unable to locate all clustering parameters in [19], we hand-tune these parameters until achieving optimal results for the kitchen\_0004 scene in the NYUv2 dataset, subsequently employing these settings (inflation = 12) across all scenes.

## 4 EXPERIMENTAL EVALUATION

### 4.1 Datasets and Metrics

**NYUv2.** We evaluate our proposed method on the NYUv2 [23] dataset. Following the official guide, we preprocess the entire dataset with MATLAB and use Open3D [43] to compute the camera poses. We train a separate Embedding-NeRF model for each scene, generating semantic embeddings and entropy. The official split of 654 images is used for testing. Images are resized to a resolution of 320 × 240 pixels in all experiments.

**Replica.** Replica [25] is a reconstruction-based 3D dataset containing 18 high-fidelity scenes with dense geometry, HDR textures, and semantic annotations. Zhi et al. use the Habitat simulator [21] to render RGB images, depth images, and semantic labels from randomly generated 6-DOF trajectories, mimicking hand-held camera motions. We evaluate our method's performance in discovering novel classes using their open-source simulated dataset. Images are resized to a resolution of 320 × 240 pixels in all experiments.

**Metric.** We use pixel classification accuracy (Acc.) and mean intersection over union (mIoU) as our metrics.

### 4.2 Implementation Details

**Pre-trained Model.** Utilizing ESANet[22] with a ResNet34 NBt1D backbone as our semantic segmentation network, we train it on the SUN RGB-D dataset[24] for semantic embedding extraction, consisting of 5,285 RGB-D images. Specifically, by fine-tuning the dataset to include only 9 classes as known classes among the 13 classes defined in [6] and masking 4 classes as novel classes.

For the NYUv2 Datasets, we use pre-trained ESANET to extract semantic embeddings from the RGB-D images. The selected classes and the entire classes are shown in Table 1.

For the Replica dataset, we use the same pre-trained network to extract the semantic embedding. Since Semantic-NeRF is supervised and cannot discover novel classes. To ensure the fairness of the experiment, we compare NeurNCD with the sparse label propagation experiments in Semantic-NeRF. That is, in Semantic-NeRF, we apply single-click, 1%, 5% and 10% pixel annotations to the 9 known semantic classes we defined previously, and generate semantic segmentation results through weak supervision.

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 [13] with a learning rate of 5e-4 for 200,000 iterations. The training time is approximately 8 hours, consistent with Semantic-NeRF training time but faster than the original NeRF (time > 24 hours).

### 4.3 Baselines

As there are no previous works that use the neural radiation field to tackle the discovery of novel class problems, we compare our proposed method to the four most closely related approaches, i.e.,

**Table 2: Assessment of Quantitative Outcomes for 9 Known and 4 Unknown Classes within the Replica Dataset: Semantic-NeRF and iLabel facilitate supervised segmentation of known classes via sparse label propagation, whereas NeurNCD not only segments known classes but also discovers novel classes.**

Method	Label Propagation	known class mIoU	Novel class mIoU	Avg Acc	Total Acc
Semantic NeRF [41]	Single Click	50.1	-	84.7	80.5
	1 %	68.2	-	82.7	84.5
	5 %	76.5	-	86.3	87.1
	10 %	80.9	-	88.3	89.1
iLabel [42]	20 click	48.0	-	-	-
	40 click	64.0	-	-	-
	60 click	72.0	-	-	-
	80 click	78.0	-	-	-
Our	-	<b>81.3</b>	<b>50.6</b>	<b>89.1</b>	<b>89.7</b>



**Figure 5: 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.**

two explicit representation methods[19, 27] and two supervised implicit representation methods[41, 42].

The explicit representation work was put forth by Nakajima et al. we implement the method using the framework of [2]. Since

we could not find all clustering parameters in [19], we use the parameter optimisation from [2] on the inflation and  $\eta$  parameter of the MCL clustering for every scene.

Although no work has explored the important problem of discovering new classes in the open world with the help of neural radiation fields, in Semantic-NeRF[41] and iLabel[42], Zhi et al. method for semantic segmentation via the sparse label propagation validated the potential of neural radiation fields to discover new classes, therefore, we also compared these two methods.

#### 4.4 Results

Our experiments aim to demonstrate the effectiveness of the proposed method both statistically and subjectively. Firstly, we use the NYUv2 dataset for quantitative comparison, calculating the intersection over union (IoU) and presenting the results in Table 1.

Table 1 compares our method with two fully supervised methods and two unsupervised methods. Specifically, the fully supervised methods encompass a conventional semantic segmentation[22] and SLAM mapping for semantic segmentation[20], while the unsupervised approaches include a state-of-the-art semantic mapping technique[19] and a prior incremental 3D geometric segmentation method [27], which served as inspiration for our work’s geometric segmentation.

It becomes apparent that fully supervised methods are limited to predicting the nine classes in the training dataset and are incapable of uncovering novel classes. In stark contrast, our method significantly surpasses the other unsupervised techniques for both known and novel classes, achieving a mean IoU of 51.29. Although the unsupervised methods[19, 27] can identify certain novel classes, their dependence on feature extraction and updating hampers their capacity to amalgamate multi-view visual features, thereby leading to incorrect segmentation and the generation of noise and outliers.

By capitalizing on Embedding-NeRF’s feature fusion capabilities, our approach rectifies and supplements the "incomplete classes" and "outlier classes" arising from imprecise geometric segmentation. Quantitatively, our method enhances the mean IoU from 46.05 to 51.29 compared to the state-of-the-art technique. In the known class section, all classes in our results display significant improvement, whereas, in the unknown class section, three classes outperform

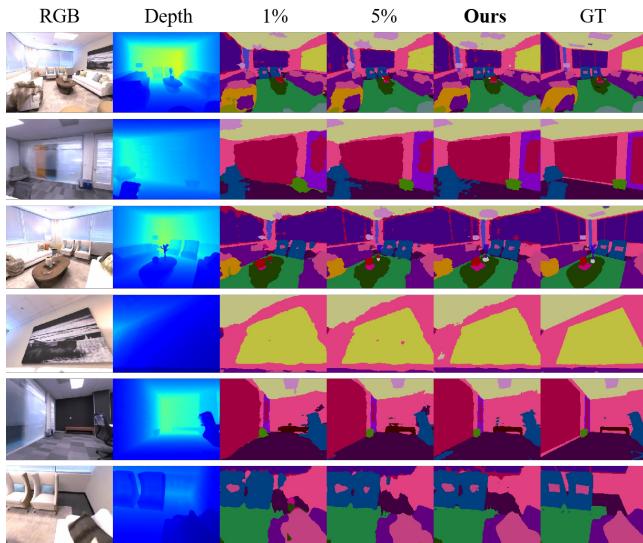
**Table 3: Ablation study for our method on NYUv2 dataset.**

Components					classes in training dataset									novel classes				mIoU
GS	PSSN	EP	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	58.86	48.25	69.28	31.92	25.59	59.38	53.95	51.29

state-of-the-art methods, albeit the picture class has no notable enhancement. This discrepancy stems from our geometric segmentation reliance on depth images alone, as opposed to the method in [19], which employs both depth and colour for segmentation, resulting in a marginally inferior performance for classes with suboptimal geometric features.

In Table 2, we compare our approach with sparse label annotation propagation experiments from Semantic-NeRF [41] and iLabel [42] as baselines to showcase the improvements in discovering novel classes in the Replica dataset. The pre-trained semantic segmentation network, as well as the known and new class settings, remain consistent with the NYUv2 dataset. Unlike Semantic-NeRF and iLabel’s supervised semantic segmentation methods, our approach not only excels in known class segmentation but also discovers novel classes.

Semantic-NeRF achieves semantic segmentation using partial annotations consisting of Single Click or 1%, 5%, or 10% of pixels per class within frames, while iLabel provides semantic segmentation results after 20, 40, 60, and 80 interactive clicks. Our method’s segmentation of the nine known classes ( $mIoU=81.3$ ) significantly outperforms the baseline ( $mIoU=80.9$  and  $mIoU=78.0$ ).



**Figure 6: Results on the Replica dataset. 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 assess the influence of various components, such as the geometric segmentation (GS), pre-trained semantic segmentation network(PSSN), semantic embedding (SE), entropy (EP), and Markov clustering (MCL), with ablation study results presented in Table 3.

Comparing GS and PSSN, we observe that the pre-trained model’s segmentation performance on known categories significantly surpasses that of pure geometric segmentation. However, the pre-trained model fails to identify unknown categories. While geometric segmentation can detect unknown categories, it assumes that real-world objects exhibit overall convex surface geometry, leading to suboptimal segmentation results for objects with poor convexity and causing over-segmentation. This issue necessitates a reliance on clustering algorithms to associate segments of the same category.

Examining the contributions of EP and SE in feature modulation, we conduct ablation comparison experiments for two cases: EP and EP +SE. Results reveal that incorporating only entropy features into segments yields inferior outcomes while adding semantic embedding improves performance. This is because Entropy, being discrete and lacking spatial consistency, can only assign characteristics to known classes, offering minimal aid in discovering unknown classes. Conversely, the semantic embedding derived from Embedding-NeRF demonstrates spatial consistency and continuity.

Moreover, the feature modulation module facilitates the assignment of entropy and embeddings to both known and unknown classes. Consequently, following clustering, sub-instance-level segments with matching features and embeddings can effectively accomplish the segmentation of known classes and the discovery of novel classes.

## 5 CONCLUSION

We investigate discovering novel classes in open-world settings and propose a powerful solution named NeurNCD. We develop exquisite designs like Embedding-NeRF with KL divergence, feature query and modulation, and unsupervised clustering. Whether compared with traditional explicit representation methods or with supervised implicit representation methods, Our method shows superior quantitative and qualitative results in both known class segmentation and novel class discovery.

## REFERENCES

- [1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5470–5479.
- [2] Hermann Blum, Marcus G Müller, Abel Gawel, Roland Siegwart, and Cesar Cadena. 2023. SCIM: Simultaneous Clustering, Inference, and Mapping for Open-World Semantic Scene Understanding. In *Robotics Research*. Springer, 119–135.
- [3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.
- [4] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Springer, 561–577.
- [5] Zheng Chen, Chen Wang, Yuan-Chen Guo, and Song-Hai Zhang. 2022. Struct-NeRF: Neural Radiance Fields for Indoor Scenes with Structural Hints. *arXiv preprint arXiv:2209.05277* (2022).
- [6] Camille Couprise, Clément Farabet, Laurent Najman, and Yann LeCun. 2013. Indoor semantic segmentation using depth information. (2013).
- [7] Jonas Frey, Hermann Blum, Francesco Milano, Roland Siegwart, and Cesar Cadena. 2022. Continual Adaptation of Semantic Segmentation Using Complementary 2D-3D Data Representations. *IEEE Robotics and Automation Letters* 7, 4 (2022), 11665–11672. <https://doi.org/10.1109/LRA.2022.3203812>
- [8] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. 2022. Panoptic NeRF: 3D-to-2D Label Transfer for Panoptic Urban Scene Segmentation. *arXiv preprint arXiv:2203.15224* (2022).
- [9] Fadri Furrer, Tonci Novkovic, Marius Fehr, Abel Gawel, Margarita Grinvald, Torsten Sattler, Roland Siegwart, and Juan Nieto. 2018. Incremental object database: Building 3d models from multiple partial observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6835–6842.
- [10] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. 2022. Neural 3D Scene Reconstruction with the Manhattan-world Assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5511–5520.
- [11] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2015. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision* 112, 2 (2015), 133–149.
- [12] Greg Hamerly and Charles Elkan. 2003. Learning the k in k-means. *Advances in neural information processing systems* 16 (2003).
- [13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. (2014).
- [14] Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. 2021. CTNet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [15] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. 2022. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838* (2022).
- [16] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang 0001, and Ming-Ming Cheng. 2018. Del: Deep embedding learning for efficient image segmentation.. In *IJCAI*, Vol. 864. 870.
- [17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- [18] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989* (2022).
- [19] Yoshikatsu Nakajima, Byeongkeun Kang, Hideo Saito, and Kris Kitani. 2019. Incremental class discovery for semantic segmentation with RGBD sensing. In *Proceedings of the IEEE/CVF international conference on computer vision*. 972–981.
- [20] Yoshikatsu Nakajima, Keisuke Tateno, Federico Tombari, and Hideo Saito. 2018. Fast and accurate semantic mapping through geometric-based incremental segmentation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 385–392.
- [21] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9339–9347.
- [22] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. 2021. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13525–13531.
- [23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.
- [24] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.
- [25] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019).
- [26] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. 2019. Multi-view saliency-guided clustering for image cosegmentation. *IEEE Transactions on Image Processing* 28, 9 (2019), 4634–4645.
- [27] Keisuke Tateno, Federico Tombari, and Nassir Navab. 2015. Real-time and scalable incremental segmentation on dense slam. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4465–4472.
- [28] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. 2022. Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12922–12931.
- [29] Han Vanholder. 2016. Efficient inference with tensorrt. In *GPU Technology Conference*, Vol. 1. 2.
- [30] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. 2021. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260* (2021).
- [31] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. 2022. R2L: Distilling Neural Radiance Field to Neural Light Field for Efficient Novel View Synthesis. *arXiv preprint arXiv:2203.17261* (2022).
- [32] Junming Wang, Zekai Sun, Xiuxian Guan, Tianxiang Shen, Zongyuau Zhang, Tianyang Duan, Dong Huang, Shixiong Zhao, and Heming Cui. 2024. AGRNav: Efficient and Energy-Saving Autonomous Navigation for Air-Ground Robots in Occlusion-Prone Environments. *arXiv preprint arXiv:2403.11607* (2024).
- [33] Xiaoyang Wang, Jimin Xiao, Bingfeng Zhang, and Limin Yu. 2022. CARD: Semi-supervised semantic segmentation via class-agnostic relation based denoising. In *Proc. IJCAI*. 1451–1457.
- [34] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12186–12195.
- [35] Zijie Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022. DoF-NeRF: Depth-of-Field Meets Neural Radiance Fields. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1718–1729.
- [36] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5438–5448.
- [37] Rui Xu and Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks* 16, 3 (2005), 645–678.
- [38] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Qing Yang, and Cheng-Lin Liu. 2020. Convolutional prototype network for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2020), 2358–2370.
- [39] Minxiang Ye, Yifei Zhang, Shiqiang Zhu, Anhuan Xie, and Dan Zhang. 2022. Deep Markov Clustering for Panoptic Segmentation. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2380–2384.
- [40] Yuyang Zhao, Zhun Zhong, Nicu Sebe, and Gim Hee Lee. 2022. Novel Class Discovery in Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4340–4349.
- [41] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15838–15847.
- [42] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. 2022. iLabel: Revealing Objects in Neural Fields. *IEEE Robotics and Automation Letters* (2022).
- [43] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847* (2018).