PoCo: Point Context Cluster for RGBD Indoor Place Recognition

Jing Liang[1], Zhuo Deng[2], Zheming Zhou[2], Omid Ghasemalizadeh[2], Dinesh Manocha[1], Min Sun[2], Cheng-Hao Kuo[2], Arnie Sen[2]

***Abstract*— We present a novel end-to-end algorithm (PoCo) for the indoor RGB-D place recognition task, aimed at identifying the most likely match for a given query frame within a reference database. The task presents inherent challenges attributed to the constrained field of view and limited range of perception sensors. We propose a new network architecture, which generalizes the recent Context of Clusters (CoCs) to extract global descriptors directly from the noisy point clouds through end-to-end learning. Moreover, we develop the architecture by integrating both color and geometric modalities into the point features to enhance the global descriptor representation. We conducted evaluations on public datasets ScanNet-PR and ARKit with 807 and 5047 scenarios, respectively. PoCo achieves SOTA performance: on ScanNet-PR, we achieve R@1 of 64.63%, a 5.7% improvement from the best-published result CGis (61.12%); on Arkit, we achieve R@1 of 45.12%, a 13.3% improvement from the best-published result CGis (39.82%). In addition, PoCo shows higher efficiency than CGis in inference time (1.75X-faster), and we demonstrate the effectiveness of PoCo in recognizing places within a real-world laboratory environment. Video: https://youtu.be/D8dObAeMiCw;**

I. Introduction

For mobile robots, localization is a critical capability that enables them to determine their position within a known environment [1]. Visual Place Recognition (VPR) is typically formulated as an image retrieval problem, i.e., retrieving candidate frames from a database by comparing similarities to a given query frame [2], [1]. Consequently, the design of an accurate place recognition system plays an important role in localization tasks by matching current observations with previously visited scenarios from the database [3], [4], [5]. It finds extensive use in applications like navigation [6], [7], and loop closure in SLAM [2] etc. However, the place recognition problem is not trivial. There could be various challenging environments [2], [1], [8], [3], [4] with illumination changes, objects changes, dynamic objects, etc. Furthermore, different sensors could be employed in specific environmental conditions. For example, Lidar [9] and Mono-camera [3], [2] are used for outdoor place recognition, while RGB-D cameras [4], [10] are widely used for indoor scenarios. How to effectively process the perception data from the sensor is also challenging [11], [12], [13]. In this paper, our main focus is on addressing the problem of indoor RGB-D point cloud-based place recognition.

Challenges for Indoor Place Recognition: While indoor and outdoor place recognition encounter similar challenges [8], [2], such as varying illumination conditions, dynamic environments, and changes in view perspective,

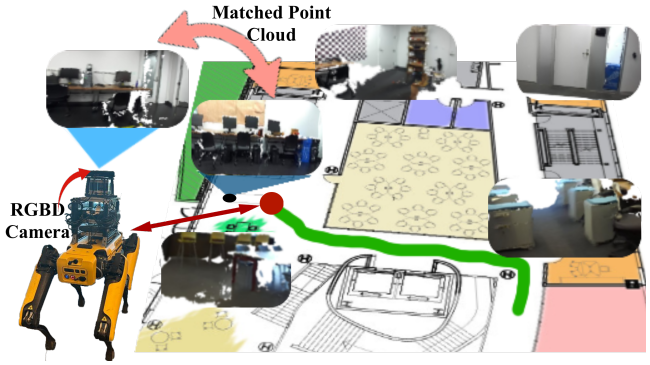

Fig. 1: As shown in the figure, the database contains several frames captured by RGB-D cameras in different rooms. The robot moves along the green trajectory and performs place recognition in real time to locate itself. When the robot moves to the red circle, it observes the blue frame and successfully finds the best-matched frame in the database.

indoor place recognition possesses distinct characteristics. Firstly, indoor environments often exhibit less variability which makes it challenging to distinguish between different indoor spaces that may appear visually similar. Secondly, indoor perceptions often exhibit shorter ranges compared to those of outdoor scenes [4]. Finally, objects in indoor scenes are typically closer to sensors compared to those in outdoor scenes, even a slight camera motion can result in significant visual appearance changes within the field of view. Consequently, approaches tailored for outdoor scenarios may not translate effectively to indoor tasks.

RGB-D Place Recognition is Under Development: A common approach for place recognition involves extracting global descriptors from both query and database frames, followed by ranking retrieved candidate frames based on the computed similarities of these global descriptors [2], [4], [10]. Many approaches have been proposed for RGB place recognition [5], [3], [2], which are mostly for outdoor place recognition because RGB camera captures rich, long-range and dense color information. For indoor scenarios, other than the RGB camera, the RGB-D camera shows more capability of perceiving dense depth images, whereas the geometric information should also be considered for the place recognition task. PointNet-VLAD [9], MinkLoc-3D [14] and Indoor DH3D [15], [16] process point clouds for place recognition, but they are not designed to process color information. To utilize both color and geometric information, Sizikova et al. [17] process RGB and depth images separately by convolutional layers and concatenate them as

[1] University of Maryland, College Park; [2] Amazon, Bellevue, WA, USA;
Code: https://github.com/jingGM/PoCo-CCR

a joint descriptor. To completely fuse the geometric and color information in the global descriptors, CGiS-Net [4] utilizes KP-Conv [18] to process both color and geometric data. In our approach, we present a novel structure to improve the efficacy of processing geometric information and also propose a better feature encoder for RGB-D place recognition tasks.

Better Feature Extraction Can Be Used: Because place recognition (retrieval) problem requires encoding a global descriptor for each frame, feature extraction plays an important role in the process. Convolution [18], [9], [5] and Tranformers [3], [19] are used for feature processing in place recognition tasks. Vision transformers (ViTs) [20] show better performance in the feature extraction for different vision tasks [21]. CoCs [12] is recently introduced to efficiently extract features from RGB images with less computational cost but with comparable performance in vision tasks as ViTs. CoCs leverage learned centers with larger receptive fields in an image to cluster and aggregate local pixels. Then it dispatches the center features to local features for further enhancement. However, CoCs is only designed for 2D images, and cannot be directly applied for point cloud inference. Inspired by the novel aggregation and dispatch mechanism where semantic-related features will be dynamically grouped and learned together, generalize this idea to jointly extract appearance and geometric features from RGB-D data for place recognition.

Main Results: We propose a novel architecture, as shown in Fig. 1, to handle both color appearance and geometric features from the RGB-D point cloud. We generalize CoCs to process point clouds and make it capable of extracting both color appearance and geometric features jointly from RGB-D data. Our approach is trained and evaluated in the ScanNet-PR [4] and ARKit [22] datasets with 807 and 5047 indoor scenarios, respectively. These two datasets exhibit various illuminations, layouts, sizes, and color features across these scenarios. In particular, the ARKit dataset is collected by a mobile device, resulting in lower depth resolution and sparser RGB-D point clouds compared to ScanNet-PR. Our contributions include:

1) We propose a novel end-to-end architecture to generalize CoCs concept from operating on 2D image domain to point clouds, where local point features are learned by interacting with all higher-level center points in a novel aggregation-and-dispatch approach.
2) We develop the architecture to jointly process different modalities, color and geometric information, to improve the performance in indoor RGB-D place recognition task. Especially, we explicitly encode geometric information into point features to enhance the representation of global descriptors.
3) Our method consistently outperforms SOTA baseline models by a significant margin on multiple datasets. We observe a relative improvement of 5.7% and 13% in Recall@1 over baselines on challenging large-scale datasets, ScanNet-PR [4] and ARKit [22].

II. Related Works

Visual Place Recognition: Place recognition is often formulated as an image retrieval problem, where the candidate frames are ranked by calculating similarities between their global descriptors and the descriptor of a query frame [4], [3]. Traditional place recognition methods extract classical RGB image features, SIFT, SURF, etc [23], [24], [25] and use the features to compose descriptors, e.g. Bag-of-Words [26], to match the frames. Then learning-based approaches generate more representative and emperical features for visual place recognition and improved the performance [5], [2]. Especially, CNN-based methods [9], [27], [28] become predominant methods for visual place recognition tasks. By extending feature-engineered VLAD into a data-driven learning approach, NetVLAD-based approaches [5] have resulted in higher accuracy in terms of matching than the traditional methods. However, extracting 3D spatial geometric features from static images is still challenging. On the other hand, numerous works directly consume point clouds as their inputs. [9] employs PointNet [11] structure and upgrade [5] to directly extract features from a 3d point cloud. To improve the limitation of [9] on capturing local geometric structures, [14] proposed to compute more discriminative 3D descriptors based on a sparse voxelized point cloud representation and sparse 3D convolutions. [15], [16] designed a Siamese network that jointly learns 3D local feature detection and description directly from raw 3D points. While point clouds offer numerous advantages such as encoded 3D spatial information and viewpoint invariance, it is arguable that integrating 3D features with 2D color features into the descriptors could potentially enhance performance further as this integration exploits the complementary nature of 3D and 2D features, leveraging the strengths of both modalities for more comprehensive scene representation.

Indoor RGB-D Place Recognition: This research field is under-explored as only a few works are available for indoor RGB-D place recognition. For example, [10] employs Patch-NetVLAD [29] for image retrieval and the estimate 6 DoF camera poses based on refined feature matches with synthetic RGB-D images. The system combines various components from existing techniques and follows a piece-wise optimization approach, whereas ours adopts an end-to-end learning approach. The most relevant to our work is CGiS-Net [4], which aggregates both color, semantic, and geometric information together using real RGB-D images, and its geometry feature is extracted by KPConv [18]. Again, CGis-Net has two distinct training stages for semantic encoder/decoder and feature encoder respectively, whereas ours is end-to-end and the network architecture is significantly different from theirs.

Feature Processing: Traditional methods [30], [24] use manual-crated features for place recognition. After that CNN-based methods demonstrated better performance in different vision tasks [9], [27], [28] and also place recognition. NetVLAD-based approaches [5] show good per-

formance for RGB place recognition, and KPConv-based approaches [31], [4] show promising results in point cloud place recognition. Vision transformers (ViTs) [20] is the next generation of the feature extractor and outperform CNN-based approaches in multiple vision tasks [32], [33], [3]. Recently, a new light-weighted feature extraction method Context of Clusters (CoCs) [12], is proposed and demonstrates state-of-the-art performance in different vision tasks but is more computationally effective than ViTs. However, CoCs is only designed for RGB features. In our work, we generalize it to RGB-D point cloud feature extraction.

III. Our Approach

In this section, we first formulate the problem in Section III-A, then describe the architecture of our PoCo methon. Finally, we discuss the training strategies in Section III-C.

A. Problem Formulation

Following a similar paradigm to VPR tasks [5], [3], we define the indoor RGB-D place recognition task as a point cloud retrieval problem. In subsequent sections, we will refer to each frame as a colorized point cloud. Denote a query frame $\mathbf{Q} \in \mathcal{Q}$ and a candidate frame $\mathbf{D} \in \mathcal{D}$, where $\mathcal{Q}$ and $\mathcal{D}$ represent query set and database respectively. In general, the model $\mathcal{M}$ transforms each frame into a global descriptor via representation learning, $\mathcal{M} : frame \rightarrow \mathbf{v} \in \mathbb{R}^n$. The goal is to retrieve likely matched candidate frames from the database based on descriptor similarities between query and database. Therefore, the model is expected to learn a representation such that positive frame pairs are close together in the embedding space, while negative frame pairs are far apart. To be more specific, $s(\mathcal{M}(\mathbf{Q}), \mathcal{M}(\mathbf{D}_p)) \gg s(\mathcal{M}(\mathbf{Q}), \mathcal{M}(\mathbf{D}_n))$ where $s(\cdot)$ is the similarity function, and $\mathbf{D}_p$ and $\mathbf{D}_n$ are positive and negative frames.

B. Architecture of Our PoCo Model

The architecture of our PoCo model is shown in Fig. 2. The input frames possess identical tensor shapes as $N \times 9$, where N is the number of points in a frame and each point is asociated with a 9-D feature vector, i.e., color $\{r, g, b\}$, position $\{x, y, z\}$ and normal $\{n_x, n_y, n_z\}$. As the Context of Clusters (CoCs) [12] is originally designed for 2D images, PoCo generalizes it to process 3D point cloud. In the high-level, the basic layer of the network is consisting of one Point Reducer block and one Context Cluster block. The Point Reducer downsamples point cloud and aggregate point features, while the Context Cluster enhances the point features further in an aggregation and dispatch way. To facilitate jointly learning of color and geometric features, each point feature vector is designed to have two parts: feature part and geometry part. The feature part carries the learned embedding, while geometry part stores fixed content, i.e., position and normal. The motivation to include the normal is that helps to encode points relative positional relationships for better model generalizability.

Reducer Blocks: Each reducer block is used to (1) downsample points population from the last layer; (2) estimate relative geometric relationships and aggregate embedding features for reduced points. For points downsampling, we choose the Farthest Point Sampling (FPS) strategy [34] for its excellent computational efficiency, with the complexity of $O(nk)$, while preserving the structure of the point cloud. Then for each point $p \in \mathbf{P}_r^{l+1}$, we apply KNN [35] to choose the K nearest points from $\mathbf{P}_r^l$ as shown in Fig. 3, where $\mathbf{P}_r^l$ and $\mathbf{P}_r^{l+1}$ represent input and output point sets in the reducer block. The feature aggregation from $\mathbf{P}_r^l$ to $\mathbf{P}_r^{l+1}$ is based on PointConvFormer [13] as shown in Eq. 1, where $\mathcal{N}(p)$ represents p's neighbor set. $f_{1,2,3,4}(\cdot)$ are learnable Linear Layers, and $\psi(\cdot)$ is a multi-head attention module calculating the similarity score between $f(p_k)$ and $f(p)$.

$$\mathbf{f}_p = \sum_{p_k \in \mathcal{N}(p)} f_4(g(p_k, p)) \psi(f_1(p_k), f_2(p)) f_3(p)^\top. \quad (1)$$

The vector $g(p_k, p)$ encodes coordinate-independent geometric information as shown in Eq.2, where $\mathbf{r}_k = p - p_k$ is a translation vector from p_k to p, $\mathbf{n}$ and $\mathbf{n}_k$ are normals of p and p_k. Thus, $g(p_k, p)$ only encodes the relative geometric information between the two points [36].

$$\hat{\mathbf{r}}_k = \frac{\mathbf{r}_k}{\|\mathbf{r}_k\|}; \quad \mathbf{v} = \frac{\mathbf{n}_k - (\mathbf{n}_k \cdot \hat{\mathbf{r}}_k)\hat{\mathbf{r}}_k}{\|\mathbf{n}_k - (\mathbf{n}_k \cdot \hat{\mathbf{r}}_k)\hat{\mathbf{r}}_k\|}; \quad \mathbf{w} = \frac{\hat{\mathbf{r}}_k \times \mathbf{v}}{\|\hat{\mathbf{r}}_k \times \mathbf{v}\|};$$

$$\begin{aligned} g(p_k, p) = [&\mathbf{n} \cdot \mathbf{n}_k, \frac{\mathbf{r}_k \cdot \mathbf{n}_k}{\|\mathbf{r}_k\|}, \frac{\mathbf{r}_k \cdot \mathbf{n}}{\|\mathbf{r}_k\|}, \mathbf{n} \cdot \mathbf{v}, \mathbf{n} \cdot \mathbf{w}, \mathbf{r}_k \cdot \mathbf{n}_k, \\ &\mathbf{r}_k \cdot (\mathbf{n} \times \mathbf{n}_k), \|\mathbf{r}_k\|], \end{aligned} \quad (2)$$

Context Cluster Blocks: Points in the block are dynamically clustered into groups/centers based on their learned affinities in the embedding space during the aggregation stage. As semantically correlated point features are more likely to be grouped together, each center point can effectively learn richer features from its associated member points. After that, aggragated center features are adaptively dispatched to each member point based on the similarity. In this way, member points implicitly communicate with each other via their center point and learn to optimize features jointly. We provide an overview of context cluster in Fig.4 and will delve into further details in the following sections.

Aggregation: For the point set $\mathbf{P}_r^{l+1}$, we downsample it to generate center points $\mathbf{P}_c^{l+1}$, where the cardinality $|\mathbf{P}_r^{l+1}| = N_r > |\mathbf{P}_c^{l+1}| = N_c$. N_c is chosen by FPS, less than half of the N_r. Each initialized center feature $\mathbf{f}_c'$ is calculated as the mean vector of its K spatially nearest point features $\{\mathbf{f}_p\}$ from $\mathbf{P}_r^{l+1}$. As is shown in Fig. 4,in the aggregation stage, we firstly calculate the cosine similarity scores between all the centers and all the points: $s_{c,p} = \sigma(\alpha \cdot cosine(\mathbf{f}_p, \mathbf{f}_c') + \beta)$, where $\sigma(\cdot)$ is the sigmoid activation function and α and β are trainable parameters. Then point features are aggregated into center points according to previous computed similarities. The center feature $\mathbf{f}_c'$ is incorporated as an anchor feature for numerical stability as

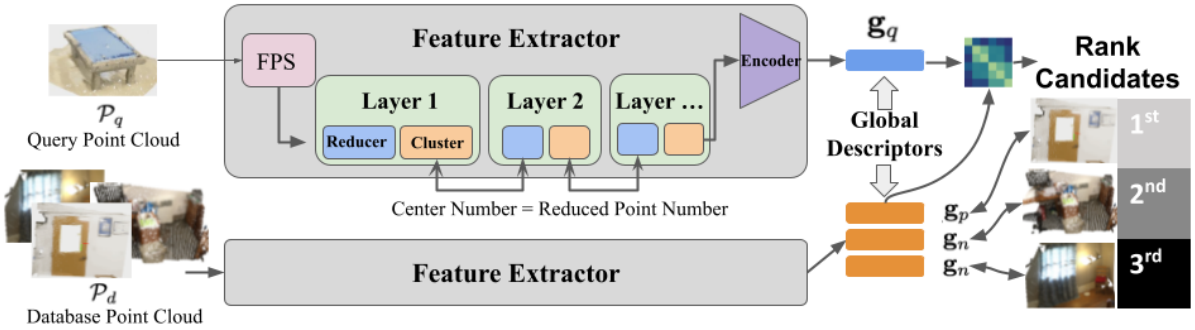

Fig. 2: **PoCo Architecture.** The input contains query and database frames and the model generates a global descriptor for each frame, and the descriptors are used to rank database frames by the similarities to the query frame. Note that the extractor consists of Farthest Point Sampling (FPS), an encoder, and layers of Reducers and Cluster blocks (see Fig. 3, 4).

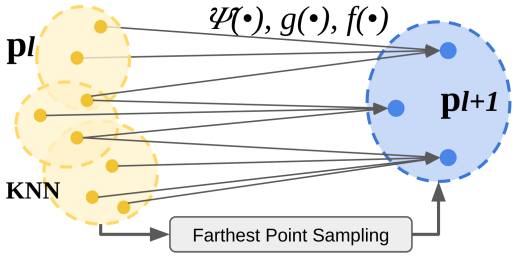

Fig. 3: **Reducer Block.** The blue circles are points in different levels. In Reducer Block, we use the Farthest Point Sampling method to downsample points $\mathbf{P}^l$ to $\mathbf{P}^{l+1}$ and aggregate the geometric and feature information of the K-nearest neighbors to the downsampled $\mathbf{P}^{l+1}$ by Eq. 1.

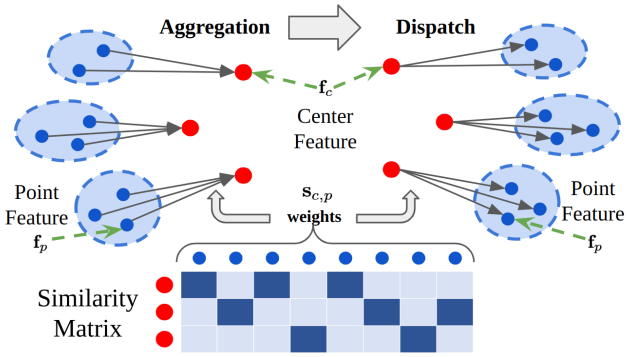

Fig. 4: **Cluster Block.** Red points are centers downsampled from the blue points. The blue points are grouped by the similarity of w.r.t. the red centers. Then the features of the blue points are aggregated to the centers by Eq. 3 and then the center features are dispatched to the points by Eq. 4.

well as further emphasize the locality in Eq. 3

$$\mathbf{f}_c = \frac{1}{\mathcal{C}} \left(\mathbf{f}'_c + \sum_{p=1}^{N_r} \mathbf{s}_{c,p} \cdot \mathbf{f}_p \right), \tag{3}$$

$\mathcal{C} = 1 + \sum_{p=1}^{N} \mathbf{s}_{c,p}$, is the normalization factor.

Dispatch: The previous learned similarities are used to dispatch the aggregated center features to their member cluster points via Eq.4 and $h(\cdot)$ indicates Linear Layers:

$$\mathbf{f}_p = \mathbf{f}_p + h(s_{c,p} \cdot \mathbf{f}_c)). \tag{4}$$

Global Descriptor Extraction: The Encoder shown in Fig. 2 aggregates all the point features from the last Context Cluster block into a global descriptor, which is used to measure the similarity between the query frame and candidate frames of database. We adopt cosine similarity as the metric function. The encoder is implemented as a variant of the Reducer Block, which is configured to have a single point as the output. Its input point features are all normalized and aggregated by the geometric and feature weights as described in Equation 1. In the experiments, the dimension of the global descriptor is set to 256.

C. Training

During training, models are optimized to distinguish positive examples from negative examples. We utilize the widely used Triplet Loss [37] for this purpose. In addition, we also aim to provide guidance to estimate the similarity within certain boundaries for stable training. Therefore, Circle Loss is also employed in the training loss [38].

The circle loss, $\mathcal{L}_c$, is defined as Equation 6. $m = 0.2$ is a margin to reinforce the optimization and $\delta_p = 1 - m$ and $\delta_n = m$. $cl(\cdot)$ is a clamp function, which clamps the value to 0 if it is negative. $\gamma = 1$ is a hyperparameter. s_p and s_n are cosine similarities of the query-positive frame pair and query-negative frame pair, respectively. The circle loss guides the similarities between the query and positive cases to 1 and negative cases to 0.

$$a_p = cl(s_p - 1 - m, 0), a_n = cl(s_n + m, 0) \tag{5}$$

$$\mathcal{L}_c = softplus(\log \sum \exp(\gamma a_n (s_n - \delta_n)) - \log \sum \exp(\gamma a_p (s_p - \delta_p))) \tag{6}$$

Triplet loss, $\mathcal{L}_t$, is defined as Equation 7. Triplet loss is to maximize the difference between the distances of the query-positive frame pair and query-negative frame pair. The distance between the query and database descriptors is calculated by the L2 distance function, and the circle loss uese similarity is cosine similarity. Those two metrics can be converted by $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2 = 2 - 2\ cos(\mathbf{x}, \mathbf{y})$ for two vectors $\{\mathbf{x}, \mathbf{y}\}$. The margin is set as $m = 0.2$.

$$\mathcal{L}_t = \max(d(\mathbf{g}_p, \mathbf{g}_q) - d(\mathbf{g}_n, \mathbf{g}_q) + m, 0) \tag{7}$$

In the final loss function, we combine the two losses $\mathcal{L} = \alpha\mathcal{L}_c + \beta\mathcal{L}_t$. In experiments, we find the best weights for these two terms as $\alpha = 10$ and $\beta = 0.1$ via hyper-parameters tuning. By assigning a higher weight to the circle loss, the model can rapidly converge, facilitating the differentiation between positive and negative examples. Subsequently, circle loss aids in refining the estimation process, resulting in more accurate similarity values that closely align with the boundaries.

IV. Experiments

In this section, we firstly introduce two large-scale datasets with various challenging scenarios for training and evaluation. Then we describe the settings of the implementation. For evaluation, we demonstrate our approach outperforms other state-of-the-art indoor RGB-D based or point-cloud based approaches by both quantitative and qualitative results. Then we conduct ablation studies to evaluate contribution and impact of individual components on the indoor RGBD place recognition task.

A. Dataset and Training Settings

Our approach is trained and evaluated on two indoor RGB-D datasets, ScanNet-PR [4] and ARKit [22]. They are popular large scale datasets suitable for data-driven learning approaches and encompass a variety of object and scene categories presenting diverse illumination conditions. ScanNet-PR is derived from the ScanNet dataset [39] captured by a commodity RGB-D sensor combined with a iPad RGB camera and the ARKit dataset is captured by an iPad camera together with a dense Lidar scanner. The datasets contain numerous scenarios, each representing a room with multiple frames. Each frame contains a point cloud with both color ($\{r, g, b\}$) and geometric ($\{x, y, z\}$) information for each point. However, the number of points varies for different frames. Following previous work dataset split settings, we split ScanNet-PR's 807 scenarios into training (565), validation (142), and testing (100) datasets. For the ARKit dataset, we employ a similar processing strategy as that used for ScanNet-PR [4]. The ARKit dataset has 5047 scenarios and we split it into training (3958), validation (1089), and testing (100) datasets.

Our PoCo model is implemented by Pytorch and trained on 8 Tesla-V100 GPUs. The input points are constrained to 2000 points through voxelization. Besides $\{r, g, b\}$ and $\{x, y, z\}$, we also use the normal vector $\{n_x, n_y, n_z\}$ of each point to provide additional relative geometric information as in Eq. 2, which is estimated by the plane within a 0.2-meter radius of the point. For the training strategy, we use cosine annealing scheduler [40] with Adam optimizer [41]. The learning rate is from 10^{-4} to 10^{-7}.

B. Implementations of Comparison and Ablation Study

To ensure a fair comparison, we use the SOTA evaluation strategy of CGiS-Net [4], which is the current best approach in the ScanNet-PR dataset. In each scenario of the dataset, the database frames are chosen by a distance threshold of 3 meters, and other frames are query frames. For each query frame, we use the Recall metric to evaluate the place recognition performance: The similarities between the query frame and the database frames from **all testing scenarios in the dataset** are calculated and the database frames are ranked by the similarity. The Recall@k represents the percentage of matched frames in the top k candidates with the query frame. Besides Recall@k, we include other metrics such as running time (inference time), FLOPs, and model size as well. All those metrics are reported on the same PC equipped with one NVIDIA GeForce RTX 3060 GPU and an Intel Xeon(R) W-2255 CPU.

To evaluate the performance of our approach, PoCo, we compare with different SOTA indoor RGB-D and point cloud-based place recognition approaches, CGiS-Net [4], MinkLoc-3D [14], NetVLAD [5], PointNet-VLAD [9], Indoor DH3D [16]. All these approaches are trained and tested in the two datasets, separately.

For the ablation study, to demonstrate the capability of our model in handling both color and geometric information, we designed three types of experiments: 1. Vanilla RGB CoCs, which uses the CoC-Medium [12] as the backbone and applied our Encoder block of Fig. 2. This is to test the capability of CoCs in RGB place recognition, and also shows how much effect color information takes in place recognition task; 2. PoCo w/o Color, which only processes geometric information of the point cloud, where the input feature changes from $\{r, g, b, x, y, z\}$ to $\{x, y, z\}$. This ablation study tests the effectiveness of our PoCo model using geometric information in place recognition tasks; 3. PoCo w/o Eq. 2, which removes the relative geometric information and uses absolute positions of the points instead. This model is to test how much effect the relative geometric function has on the place recognition performance.

ScanNetPR	R@1 ↑	R@2 ↑	R@3 ↑
SIFT [23] + BoW [26]	16.16	21.17	24.38
NetVLAD [5]	21.77	33.81	41.49
PointNet-VLAD [9]	22.43	30.81	36.58
Indoor DH3D [16]	16.10	21.92	25.30
MinkLoc-3D [14]	10.13	16.63	20.80
CGiS-Net [4]	61.12	70.23	75.06
CGiS-Net w/o color [4]	39.62	50.92	56.14
CGiS-Net w/o geometry [4]	40.07	51.28	58.96
Vanilla RGB CoCs [12]	41.82	57.65	66.82
PoCo w/o color	44.34	54.27	59.78
PoCo w/o Equation 2	62.23	73.81	79.63
PoCo	**64.63**	**75.02**	**80.09**

TABLE I: The table shows our PoCo method outperforms other state-of-the-art methods by at least 5.7% improvement in Recall@1 value in the ScanNetPR dataset.

C. Results

In Table I and Table II, we present the performance of different SOTA approaches compared with our PoCo method and also ablation studies in two datasets. From Table I, the learning-based methods show better accuracy in Recall@1 than SIFT+BoW. In the ScanNetPR dataset (Table I), our PoCo method outperforms the other approaches by

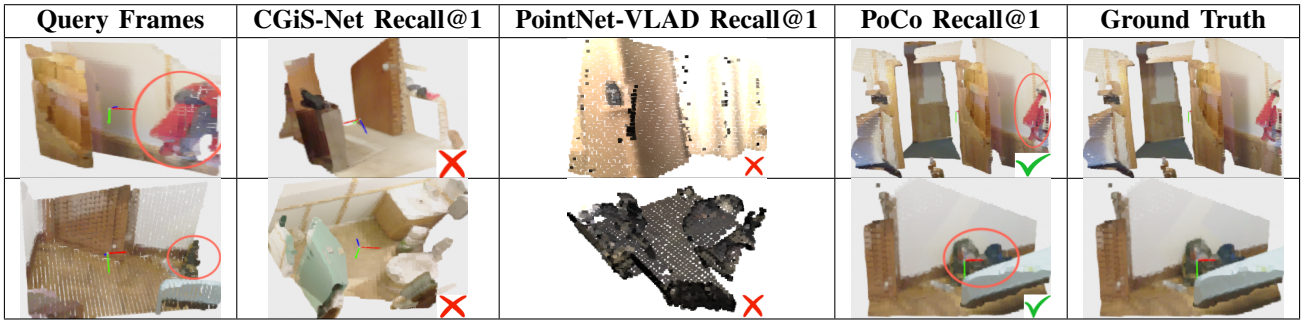

Fig. 5: **Qualitative Comparison in the ScanNetPR dataset:** Our PoCo method outperforms other approaches in the challenging scenarios in the ScanNetPR dataset, with small overlap areas, marked by red circles, between query and positive frames.

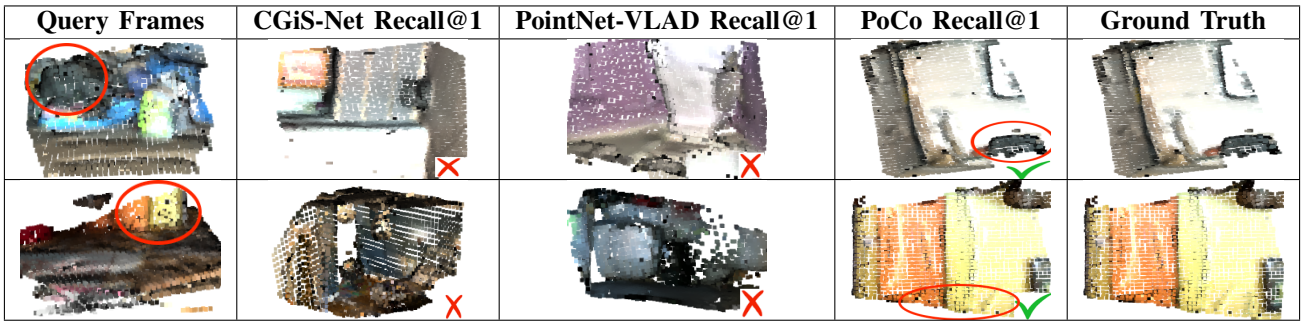

Fig. 6: **Qualitative Comparison in the ARKit dataset:** Our PoCo method can detect small overlap areas, circled in red, but CGiS-Net and PointNet-VLAD fails.

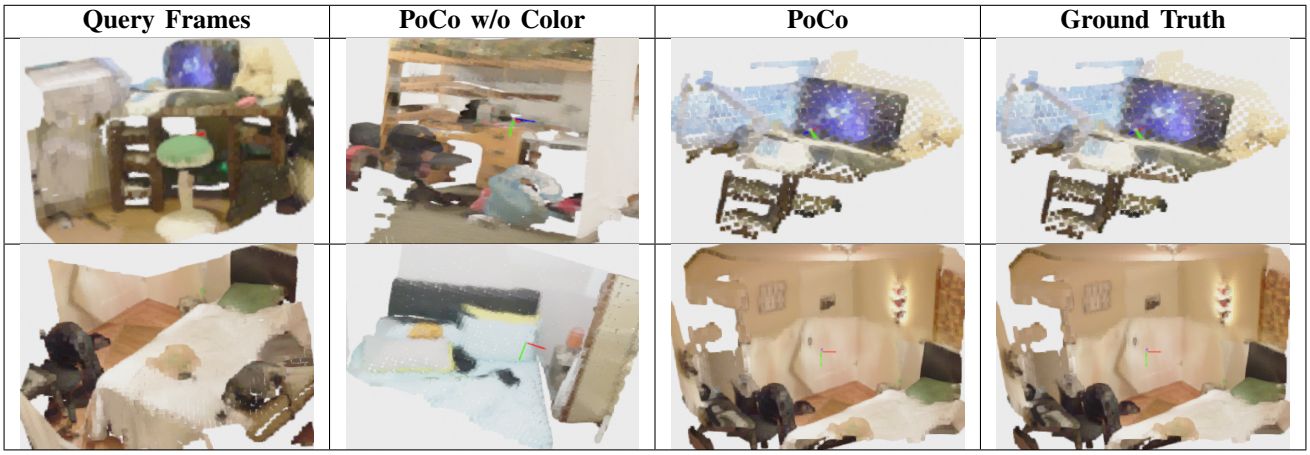

Fig. 7: **Benefits of Color Information.** The second and third columns show recall@1 frame from models w/o and with color information, respectively. After we remove the color information, the PoCo w/o Color model selects candidate frames only based on geometric information, where the desk (1st row) and the bed (2nd row) in matched candidate frames are structurally similar to the query frames, but their colors are very different.

ARKit	R@1 ↑	R@2 ↑	R@3 ↑
PointNet-VLAD [9]	11.04	16.57	20.57
MinkLoc-3D [14]	8.14	10.95	13.79
CGiS-Net [4]	39.82	49.01	56.02
Vanilla RGB CoCs [12]	17.19	24.74	30.06
PoCo w/o color	21.58	30.00	35.66
PoCo w/o Equation 2	41.41	51.86	58.11
PoCo	**45.12**	**57.10**	**62.14**

TABLE II: The table shows our PoCo model outperforms other approaches by at least 4 points in Recall@1 in the ARKit dataset.

Methods	R@1↑	Running Time (s)	FLOPs (Mb)	Model Size (Mb)
PointNet-VLAD	22.43	0.01	1319.67	75.47
MinkLoc-3D	10.13	0.02	-	10.17
CGiS-Net	61.12	0.07	738.60	26.18
PoCo	64.63	0.04	285.63	29.94

TABLE III: Our approach is 1.75X faster than CGiS-Net and also has better recall values. Our computational cost is the smallest compared with other approaches.

at least 5.7% in Recall@1 and In the ARKit dataset (Table II) we have at least 4 points improvement. Considering PointNet-VLAD and MinkLoc-3D only process geometric

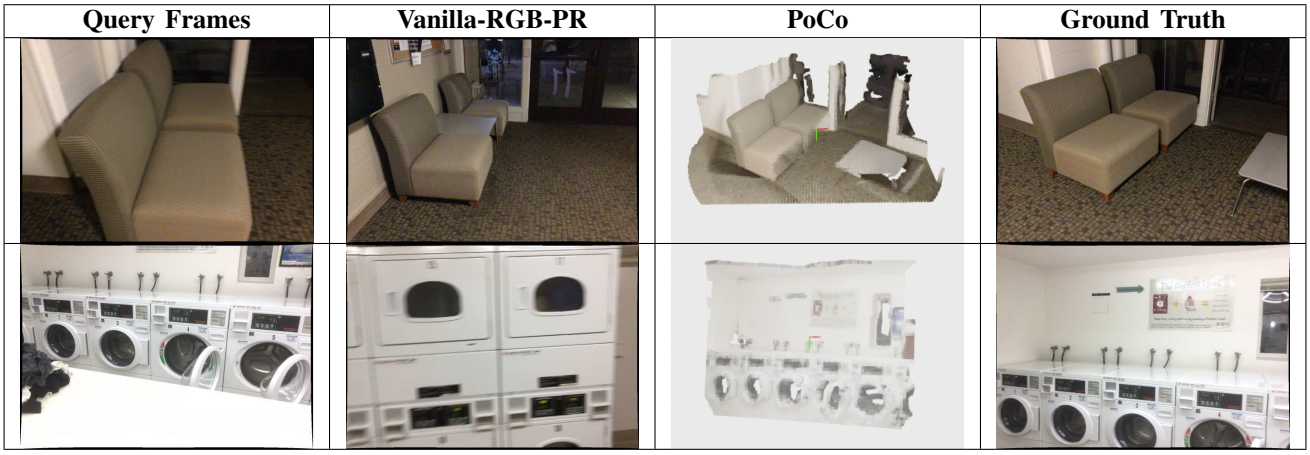

Fig. 8: **Benefits of Geometric Information.** The Vanilla-RGB-PR can choose similar objects (e.g., chairs or washing machines). But it cannot capture spatial relationships of objects well. For example, in the second row, the query image has lined machines but the matched image has stacked machines.

information, in Table I, we compare PoCo w/o color with these two methods and CGiS-Net w/o color. Our pure geometric model still outperforms other approaches by at least 11.91% in Recall@1. In the ARKit dataset, our PoCo w/o color also outperforms other pure-geometric methods. Those results indicate our model can effectively encode geometric information for place recognition and support our innovation in generalizing CoCs to point cloud place recognition. As shown in Table I, compared with other RGB-based approaches NetVLAD [5] and SIFT+BoW, our RGB model (Vanilla RGB CoCs) demonstrates better accuracy in Recall@1, which means our model is capable of effectively encoding the RGB information for place recognition. From Fig. 5 and Fig. 6, we observe our model is much better than baselines in those challenging cases, where the overlapping (red-circled) areas between query and positive frames are very small. From these two figures, we can also observe that PointNet-VLAD cannot recognize different colors for matching, whereas, in Fig. 5, it chooses a scenario with a completely different color as the best match. Besides the accuracy, as shown in Table III, we also compared the computational cost of different approaches. Compared with CGiS-Net, our PoCo model achieves both higher accuracy in Recall and less computational cost. The PointNet-VLAD and MinkLoc-3D have less inference time, but their accuracy is much worse than our model.

About Ablation studies, as shown in Table I and Table II, geometric information and color information are both important to place recognition and they have very similar effects on the Recall values. Geometric information matters more than color information, and by only using each of them, in Table I PoCo w/o color has Recall@1 2.52 points higher than the RGB model. In Table II, it has a larger difference of 4.39 points, which means that our design can effectively process geometric information for place recognition. Comparing Poco with PoCo w/o Equation 2, we observe that relative geometric information makes the method more generalized to different point clouds and achieves 2.4 and 3.71 improvement in Recall@1 in ScanNetPR and ARKit datasets, respectively. Fig. 7 provides the qualitative difference between PoCo and RGB models. RGB model can only detect similar objects, but cannot tell complex positional relations among the object. Compared with it, our PoCo method can both recognize the objects and their relative positions in the scenarios. From Fig. 8, PoCo w/o Color is capable of recognizing very similar structures, such as the table and the chair in the first row. But because it doesn't have color information, it falsely matched the frames with very obvious color differences. However, PoCo is capable of utilizing color information to find the best match. From the ablation study, we demonstrate the efficacy of our innovation in jointly processing different modalities, color and geometric information, and achieves higher accuracy in indoor RGB-D place recognition tasks.

V. Conclusion, Limitations and Future Work

We are the first to propose an indoor RGB-D Point-Cloud-based place recognition model based on the concept of CoCs. We explicitly encode geometric information during feature extraction with transformer-based models. PoCo outperforms previous indoor RGB-D place recognition methods significantly in recall@K. However, our model still has its limitations, where sometimes our model relies much on geometric information which hurts the performance, and in some extreme cases, the overlapping areas between the ground truth and query frame are small and do not have many features. To solve the issues, we could use local features to help re-ranking candidate frames to improve accuracy. Therefore, in the future, a second stage of re-ranking candidates will be explored.

Acknowledgment: This work was supported in part by ARO Grants W911NF2310046, W911NF2310352 and U.S. Army Cooperative Agreement W911NF2120076

References

[1] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence

Organization, 8 2021, pp. 4416–4425, survey Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/603

[2] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *ieee transactions on robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[3] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 370–19 380.

[4] Y. Ming, X. Yang, G. Zhang, and A. Calway, "Cgis-net: Aggregating colour, geometry and implicit semantic features for indoor place recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 6991–6997.

[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[6] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 4158–4163.

[7] P. Mirowski, M. Grimes, M. Malinowski, K. M. Hermann, K. Anderson, D. Teplyashin, K. Simonyan, A. Zisserman, R. Hadsell *et al.*, "Learning to navigate in cities without a map," *Advances in neural information processing systems*, vol. 31, 2018.

[8] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[9] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4470–4479.

[10] D. Yudin, Y. Solomentsev, R. Musaev, A. Staroverov, and A. I. Panov, "Hpointloc: Point-based indoor place recognition using synthetic rgb-d images," in *International Conference on Neural Information Processing*. Springer, 2022, pp. 471–484.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[12] X. Ma, Y. Zhou, H. Wang, C. Qin, B. Sun, C. Liu, and Y. Fu, "Image as set of points," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=awnvqZja69

[13] W. Wu, L. Fuxin, and Q. Shan, "Pointconvformer: Revenge of the point-based convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 802–21 813.

[14] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1790–1799.

[15] J. Du, R. Wang, and D. Cremers, "Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 744–762.

[16] X. Yang, Y. Ming, Z. Cui, and A. Calway, "Fd-slam: 3-d reconstruction using features and dense matching," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8040–8046.

[17] E. Sizikova, V. Singh, B. Georgescu, M. Halber, K. Ma, and T. Chen, *Enhancing Place Recognition Using Joint Intensity - Depth Analysis and Synthetic Data*, 11 2016, pp. 901–908.

[18] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and cnns in digital pathology," 2022.

[22] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman, "ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. [Online]. Available: https://openreview.net/forum?id=tjZjv_qh_CE

[23] L. David, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[24] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International journal of robotics research*, vol. 27, no. 6, pp. 647–665, 2008.

[25] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[26] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 591–606, 2008.

[27] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2136–2145.

[28] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.

[29] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[31] P. Shi, Y. Zhang, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *ArXiv*, vol. abs/2306.10561, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259204004

[32] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[33] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5396–5407.

[34] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi, "The farthest point strategy for progressive image sampling," *IEEE Transactions on Image Processing*, vol. 6, no. 9, pp. 1305–1315, 1997.

[35] O. Kramer and O. Kramer, "K-nearest neighbors," *Dimensionality reduction with unsupervised nearest neighbors*, pp. 13–23, 2013.

[36] X. Li, W. Wu, X. Z. Fern, and L. Fuxin, "The devils in the point clouds: Studying the robustness of point cloud convolutions," *arXiv preprint arXiv:2101.07832*, 2021.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[38] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6398–6407.

[39] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[40] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.