CSCPR: Cross-Source-Context Indoor RGB-D Place Recognition

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

***Abstract*— We extend our previous work, PoCo [1], and present a new algorithm, Cross-Source-Context Place Recognition (CSCPR), for RGB-D indoor place recognition that integrates global retrieval and reranking into an end-to-end model and keeps the consistency of using Context-of-Clusters (CoCs) [2] for feature processing. Unlike prior approaches that primarily focus on the RGB domain for place recognition reranking, CSCPR is designed to handle the RGB-D data. We apply the CoCs to handle cross-sourced and cross-scaled RGB-D point clouds and introduce two novel modules for reranking: the Self-Context Cluster (SCC) and the Cross Source Context Cluster (CSCC), which enhance feature representation and match query-database pairs based on local features, respectively. We also release two new datasets, ScanNetIPR and ARKitIPR. Our experiments demonstrate that CSCPR significantly outperforms state-of-the-art models on these datasets by at least 29.27% in Recall@1 on the ScanNet-PR dataset and 43.24% in the new datasets. Website: https://github.com/jingGM/PoCo-CCR.git**

I. Introduction

Place recognition plays an important role in robotics [3], [4], [5], [6], where given query frames the goal is to identify matches from a database that share overlapping fields of view with queries based on image similarities [7], [5], [6]. It is used in various applications, such as augmented reality [8], navigation [9], [6], SLAM [6], [10], etc. However, the place recognition problem is very challenging [5], [6], [4] due to: (1) different sensors, RGB, RGB-D, Lidar, etc., which require modality-specific feature processing; (2) environmental challenges, such as illumination changes, dynamic objects, occlusion, and scale variation; (3) the lack of study around RGB-D indoor place recognition [11], [6], [4], where most of the current approaches only use global retrieval for place recognition [11], [12]. Our approach focuses on the less explored domain of RGB-D indoor place recognition and proposes an end-to-end architecture with a reranking stage to handle the RGB-D place recognition task.

RGB-D indoor place recognition is not well studied: Visual place recognition has been explored for many years, especially in the RGB domain [4], [13]. However, the potential of RGB-D data is overlooked; especially for indoor environments, the depth information can be crucial for place recognition [11], [12], [4]. For RGB-D place recognition, we need to handle different modalities of perceptions (color and geometry), so a good feature extractor is critical for the task. The CoCs [2] method shows comparable performance to the attention mechanism [14], and has been applied [1] to handle different scales of features. Extending from our previous work, PoCo [1], we propose a novel end-to-end architecture, CSCPR, integrating global retrieval and reranking for RGB-D indoor place recognition, and we also keep the consistency of using the CoCs concept for feature processing.

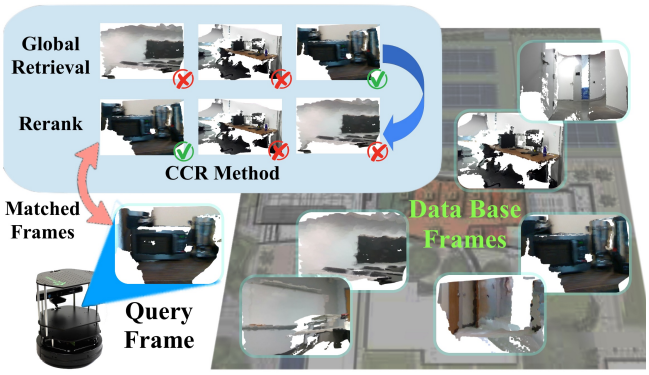

Fig. 1: **Real-world Experiment:** We propose a novel approach, Cross-Source-Context Place Recognition (CSCPR), for RGB-D indoor place recognition. Given a query frame, global retrieval ranks the potentially matched frames, and our novel place-recognition reranking model reranks the candidates to achieve better recognition accuracy.

RGB-D place-recognition reranking is not well studied: Traditional learning-based RGB-D place recognition approaches [11], [15], [16], [6] rely on global retrieval, which extracts global descriptors from each frame and ranks the database frames by the similarities with the query frame. Reranking, as a complement to global retrieval, is meant to improve the accuracy by evaluating the matches of local features between the query and database frames [13], [17], [18]. Therefore, the reranking stage should be fast and accurate in matching local features. The RANSAC-based [19] methods are widely used for reranking by processing the geometric information of query-database frames [17], [18], [20]. However, RANSAC is not a deep-learning-based approach and cannot be empirically trained to calculate the relationships among all local features. Recently, learning-based approaches [13], [21] have been proposed that match empirical local features. However, those approaches are all for RGB-only place recognition. In this work, we generalize the learning-based reranking concept to RGB-D point cloud and propose a novel learning-based method for RGB-D place-recognition reranking.

The scarcity of large-scale clean datasets is a notable gap for RGB-D indoor place recognition: For indoor vision tasks, there are many datasets designed for object classification, segmentation, etc. [22], [23]. However, those datasets are not specifically designed for training RGB-D place recognition tasks, which require positive and negative

[1] University of Maryland, College Park; [2] Amazon, Bellevue, WA, USA

matched frames by overlap and a large scale for training purposes [7]. HPointLoc [24] and ScanNet-PR [11] have recently been published for RGB-D indoor place recognition tasks, but their matched pairs are selected by the distance of camera poses or point cloud distance instead of the overlap of point clouds. This makes the dataset very noisy because of some nearby point clouds with no overlap, and it decreases the efficacy for training learning-based models. To bridge the gap, we introduce two large-scale clean datasets for RGB-D indoor place recognition by choosing frames with overlap.

Main Results: In this approach, we aim to solve the RGB-D indoor place recognition problem. We extend our previous work PoCo [1] with a reranking stage and integrate the two stages into one end-to-end pipeline for place recognition. We keep the consistency of applying the CoCs concept for feature processing in the reranking stage and propose novel models to process multi-scale and multi-source features for RGB-D place recognition. The following describes our contributions. Collectively, the contributions not only push the boundaries of RGB-D indoor place recognition to an integrated retrieval-reranking learning problem but also equip the community with a data generation pipeline and new datasets for the task.

1) **Novel RGB-D Place-recognition Reranking:** We present a novel *Self-Context Cluster (SCC)* to enhance local features by global information and a novel *Cross Source Context Cluster (CSCC)* to process the local matches between different point clouds (sources) for fast and accurate reranking.
2) **Curated Large-Scale RGB-D Indoor Place Recognition Datasets:** To address the scarcity of suitable large-scale RGB-D place-recognition datasets, we introduce ARKitIPR and ScanNetIPR datasets. We also propose a method to generate datasets by point-cloud overlap. The datasets contain positive and negative frames, keyframes for evaluation, poses, and semantic labels of point clouds.
3) **Superior Performance Across Multiple Datasets:** As shown in Tab. III, our approach outperforms other SOTA RGB-D indoor place recognition approaches by at least 29.27% in ScanNet-PR [11] and 43.75% in ScanNetIR and ARKitIPR. Compared with other place-recognition reranking methods, we improve at least 3.17 points in the two novel datasets. We also demonstrate the effectiveness of our approach in a real-world robot.

II. Related Work

Place Recognition Features and Methods: Normally, place recognition is handled by calculating the similarities/overlap between the database and query frames. Traditional methods rely on RGB image features, such as SIFT, SURF, BoW, etc. [25], [26], [27], to compose descriptors for frame matching. After convolutional networks [28] and attention mechanisms [14] were proved to have better performance in vision tasks, various methods [28], [29], [30] were proposed to encode frame features to global descriptors for place recognition, especially NetVLAD-based methods [15], [16], which improve performance by combining convolutions with VLAD cores for global retrieval. The introduction of Context-of-Clusters (CoCs) [2] offers a comparably effective alternative but with different scales of feature processing, enhancing local features with global context. Our approach uses this concept to process and aggregate features of RGB-D point clouds.

For RGB-D place recognition tasks, NetVLAD [16], ORB [31], etc. use RGB information for place recognition by comparing the global descriptors of images. MinkLoc3D [32] and Point-NetVLAD [15] process point clouds and compare point-cloud global descriptors. CGis-Net [11] applies KP-Conv [33] to encode the geometric information and enhance the features by extracting the semantic information from colors. However, these approaches typically only assess the global descriptors of RGB-D frames. We introduce a novel place-recognition reranking method to improve the accuracy of RGB-D indoor place recognition.

Place-Recognition Reranking: The accuracy of place recognition can be enhanced through a subsequent reranking stage following global retrieval [13], [17], [34]. RANSAC-based geometric verification [18], [17] is a popular choice for reranking candidates. MAGSAC [35] uses a sample-based method NAPSAC to sample the points and apply a quality function to choose the matched pairs. PatchNetVLAD [17] uses RANSAC with scoring methods for reranking after NetVLAD [16]. Although RANSAC focuses on geometric verification, it overlooks non-geometric information. Similarly, registration tasks [36], [37], [38] also find inliers, but those methods still attempt to match points through geometric verification or rendering without empirically learning frame similarities. Matching methods [39], [40] like SuperGlue [40] identify inliers of the matching, but cannot guarantee if the frames are overlapped. Lee et al. [21] proposed CVNet-Rerank based on convolutional neural networks to process image features and compare the similarities among the features for reranking. R2Former [13] used ViTs [41] to extract the local information of images with higher attention values for reranking. However, those methods are all for RGB place recognition. RGB-D point cloud reranking still relies heavily on RANSAC-based geometric verification [18], or registration methods [34]. These approaches cannot use color and geometric information jointly. We propose an end-to-end structure to jointly process RGB and geometric information and we also integrate global retrieval and reranking together for RGB-D indoor place recognition.

III. Method

In this section, we first formulate the problem in Section III-A. Then, in Section III-B we describe the overall architecture of CSCPR, including the Self Context Cluster (SCC) and Cross-Source Context Cluster (CSCC) modules. Finally, the training strategies are described in Section III-C.

A. Problem Definition

As with other place recognition problems [5], [6], [13], [16], the RGB-D indoor place recognition problem is de-

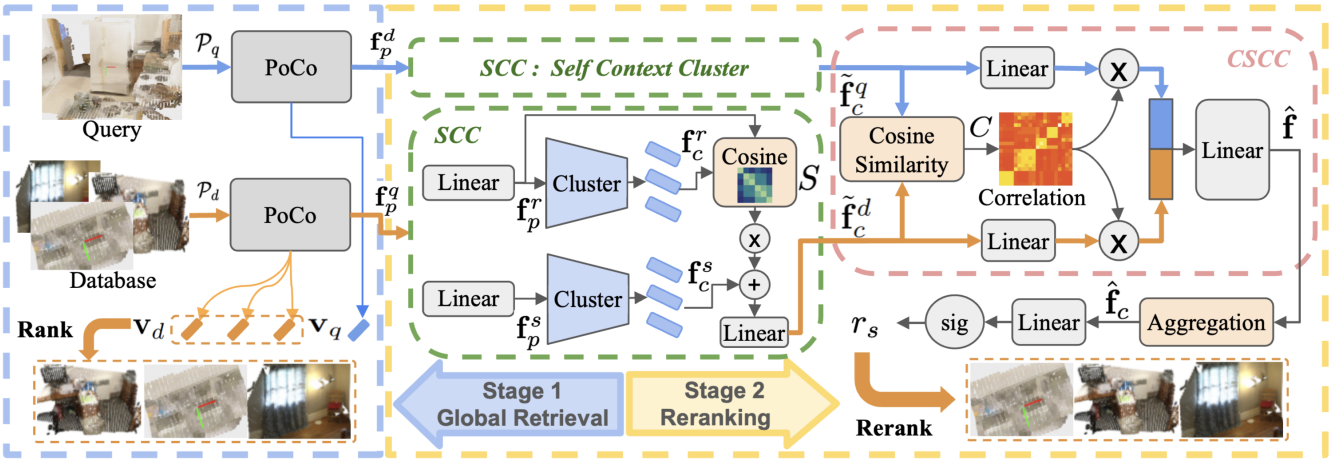

Fig. 2: **Architecture of CSCPR:** The blue box indicates Global Retrieval, and the yellow box represents Reranking. We use PoCo [1] for global retrieval. After the database frames are ranked by comparing the global descriptors, $\mathbf{v}_q$ and $\mathbf{v}_d$, the reranking stage calculates the similarity of local features between query and database frames and reranks the database frames. The reranking model composes two Self Context Clusters (SCC) to process multi-scale features of each frame and a Cross Source Context Cluster (CSCC) to calculate the similarity of the local features between these two sorts of frames.

fined as a retrieval problem. It's separated into two consecutive tasks, Global Retrieval [17], [15], [16], [11] and Reranking [34], [13]. In the paper, we use "frame" as an RGB-D colorized point cloud. As shown in Fig. 2, CSCPR is an end-to-end structure to process the color and geometric information of the frames jointly. The Global Retrieval (blue box) is from our previous work, PoCo [1], with the same input format. It calculates the global descriptors, $\mathbf{v}_q$ of a query frame $\mathcal{P}_q \in \mathcal{Q}$ and $\mathbf{v}_d$ of each of the database frames $\mathcal{P}_d \in \mathcal{D}$. Then we compare the similarity between $\mathbf{v}_q$ and all $\mathbf{v}_d$ to rank the database frames. However, the Global Retrieval model only matches the global descriptors, but some of the local information could be missed. The Reranking stage (yellow box) reranks database frames based on the similarities of local features, $\mathbf{f}_p^q$ and $\mathbf{f}_p^d$, between $\mathcal{P}_q$ and $\mathcal{P}_d$. The learning objective is to ensure the similarity between the $(\mathbf{Q}, \mathbf{D}_p)$ is much bigger than the similarity between $(\mathbf{Q}, \mathbf{D}_n)$, where $\mathbf{D}_p \in \mathcal{D}$ and $\mathbf{D}_n \in \mathcal{D}$ correspond to the positive and negative matches to $\mathbf{Q}$.

B. Architecture of CSCPR

Considering that reranking is the second stage and compares local features of frames, the structure should: 1. **effectively process the relation** of the local features from different frames; 2. be **fast** and light in parameters and computation. To address these criteria, as shown in Fig. 2, we 1. propose a novel Self Context Clusters (SCC) to process multi-scale features and a novel Cross Source Context Cluster (CSCC) model to process the relation between two frames. 2. design a simple structure (yellow box) with only three small but powerful modules (two SCC and one CSCC).

Self Context Cluster (SCC) performs two functions: 1. **Small memory usage:** It downsamples the point features to a smaller number of center features. 2. **Multi-scale feature processing:** Because indoor RGB-D data have large variations w.r.t. perspectives, scales, and similar structures for different rooms, SCC learns features from different scales: local fine-scale feature $\mathbf{f}_p \in \mathcal{R}^{N \times D_p}$ with denser points and coarse-scale feature $\mathbf{f}_c \in \mathcal{R}^{M \times D_c}$ with sparser center points, where $N > M$ are the numbers of points and centers. To make the features more representative, we enhance each of the center features with the global information (all point features) of the frame.

As shown in the green box of Fig. 2, inspired by the attention mechanism [14], we transform the point features by projecting them into two distinct branches in SCC, reference point feature $\mathbf{f}_p^r = \|l_r(\mathbf{f}_p)\|_g$ and source point feature $\mathbf{f}_p^s = \|l_s(\mathbf{f}_p)\|_g$, where l_r and l_s are different linear layers to separate the point features into reference and source features. $\|\cdot\|_g$ represents Group Norm. The geometric information of the points is used to downsample the points to a smaller number of centers by the Farthest Downsampling method, and we find the K-nearest points for each center and average neighbors' features to a center feature $\mathbf{f}_c$. Thus, we have reference and source center features as $\mathbf{f}_c^r = \underset{k}{\text{mean}}(\mathbf{f}_p^r)$ and $\mathbf{f}_c^s = \underset{k}{\text{mean}}(\mathbf{f}_p^s)$.

The current center features only contain local information about the frame, and we need to enhance it with the frame's global information, which is introduced by all the point features. Thus, we enhance the center features by calculating similarities between each center and all point features and aggregating the most globally similar point features to the center. The global similarity is calculated as Eq. 1:

$$S = \text{sig}(\alpha \cos(\mathbf{f}_c^r, \mathbf{f}_p^r) + \beta), \tag{1}$$

where α and β are trainable parameters. sig and cos are Sigmoid and Cosine Similarity functions. The similarity matrix is $S \in \mathcal{R}^{M \times N}$. We threshold the similarity matrix to $\hat{S}$ by only using the most similar center for each point and setting other values as 0. To learn the sparse but representative multi-scale feature $\tilde{\mathbf{f}}_c$ for reranking, according to the similarity matrix S, we aggregate the point features, $\mathbf{f}_p^r$, of the reference branch to center features $\mathbf{f}_c^s$ of the source

branch and normalize them by the similarities:

$$\tilde{\mathbf{f}}_{c,i} = l_c \left(\frac{1}{1 + \sum_{j=1}^{N} \hat{S}_{i,j}} \left(\mathbf{f}^s_{c,i} + \sum_{j=1}^{N} \hat{S}_{i,j} * \mathbf{f}^r_{p,j} \right) \right) \quad (2)$$

i, j are the indices of centers and points. $l_c(\cdot)$ indicates linear layers. As demonstrated in Table II, the learned "fine-coarse" scale features $\tilde{\mathbf{f}}_c$ are more effective and efficient than Attention [14], which only processes single-scale features.

Cross Source Context Cluster (CSCC) captures cross-source correlation between the query and database frames. As in the red box of Fig. 2, the query and database frames have sparse multi-modality and multiple-scale center features $\tilde{\mathbf{f}}^q_c$ and $\tilde{\mathbf{f}}^d_c$, output from SCC. To calculate the correlation of the two frames, we use the same calculation method as Eq. 1 but the points are from the same scale, $C = \text{sig}(\alpha \cos(\tilde{\mathbf{f}}^q_c, \tilde{\mathbf{f}}^d_c) + \beta)$, where $C \in \mathcal{R}^{M_q \times M_d}$ is the correlation matrix and M_q and M_d are the center numbers of query and the database frames. We choose the top $K = 500$ center pairs according to the correlation matrix, and other correlation values are masked out as 0. To fuse the multi-sourced feature together for reranking, given the masked correlation, we use it as weights to concatenate the query and retrieved candidate features together, where each concatenated feature is $\hat{\mathbf{f}}_{i,j} = l_c(C_{i,j}[l_q(\tilde{\mathbf{f}}^q_{c,i}), l_d(\tilde{\mathbf{f}}^d_{c,j})])$, where l_c, l_q, l_d indicating different linear layers. Finally, the aggregation function in Fig. 2 aggregates the most correlated 500 features to one reranking score, r_s:

$$\hat{\mathbf{f}}_c = \frac{1}{1 + \mathbf{c}_n} \sum_{j=1}^{M_d} \left(\frac{1}{1 + \mathbf{c}_m} \sum_{i=1}^{M_q} \hat{\mathbf{f}}_{i,j} \right) \quad (3)$$

where $\mathbf{c}_m = \sum_{i=1}^{M_q} C_{i,j}$ is the aggregation of the correlation matrix along the dimension of query features and $\mathbf{c}_n = \text{mean}_i(\sum_{j=1}^{M_d} C_{i,j})$ is the aggregation of the correlation matrix along the dimension of database features and average in the dimension of query features. Then we have the reranking score of the database frames and the query frame to rerank database frames: $r_s = \text{sig}(l_f(\hat{\mathbf{f}}_c))$, where l_f indicates linear layers. Our overall design is simple but effective in capturing relationships between two RGB-D clouds (see Table II).

C. Training

Training the end-to-end place recognition, CSCPR, requires multiple loss functions for Global Retrieval and Reranking. A cosine annealing scheduler [42] and Adam optimizer [43] are used to schedule the learning rates and the learning rate changes from 10^{-4} to 10^{-7}. Given ranked candidates from global retrieval, in the reranking stage we need to refine the rank by candidates' local features, so distinguishing between query and hard negatives is crucial in this stage. Therefore, we apply hard negative mining [44] in training and jointly train the two stages together.

For Global Retrieval, we use the same loss functions as [1] and we denote it as $\mathcal{L}_g$. For Reranking, we use the entropy loss function:

$$\mathcal{L}_r = -(y * \log(r_s) + (1 - y) * \log(1 - r_s)) \ . \quad (4)$$

where y is the label of the matching frames. $y = 1$ when the database frame matches the query frame, and $y = 0$ when they are not matched. The total loss function is $\mathcal{L} = \beta_g \mathcal{L}_g + \beta_r \mathcal{L}_r$, where β_g and β_r are hyperparameters, and we choose 1.0 in the training.

IV. Dataset Generation

Algorithm 1 Overlap is the major metric in dataset generation. The database frames are chosen by the overlap of frames from the original datasets and the positive and negative matches are chosen by overlaps with query frames.

Require: N consecutive frames,
Require: $F_l \leftarrow \{\mathbf{P}_0, p_0\}$
Require: database = $\{F_l, \}$
  while $n < N$ **do**
    $F_n \leftarrow \{\mathbf{P}_n, p_n\}$
    if FrameOverlap(F_l, F_n) $< T_c$ **then**
      database.add(F_n)
      $F_l = F_n$
    end if
  end while

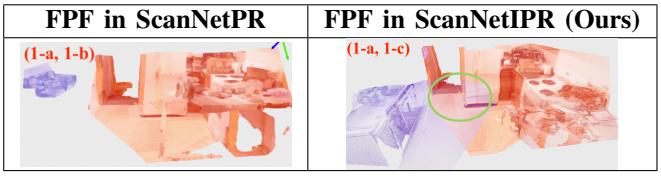

Fig. 3: **Furthest Positive Frame (FPF) of ScanNetPR vs. ScanNetIPR (ours):** FPF depicts the least overlapping matched frame to the query in both datasets. For the same query frame (red), ScanNetPR using center distance to determine matched frames leads to erroneous matching with no overlapped areas. In ScanNetIPR, the overlap (green) is the only criterion for matching; thus it is more accurate for training and evaluating place recognition task.

There are not many RGB-D datasets for indoor place recognition. ScanNetPR [11] tries to bridge the gap, but because the matching frames are chosen by the distance between frames' centers, there is some noise in the dataset. As shown in Fig. 3, there are cases in which point clouds are very close, but they have no overlap. Without overlapping areas, the model cannot tell whether the two point clouds are in the same place, and it is also difficult to apply hard negative mining with the ScanNet-PR directly, limiting the final performance of the models. To solve this issue and better evaluate the RGB-D indoor place recognition, we propose a dataset generation method and two large-scaled (in terms of the number of scenarios and frames) datasets, ScanNetIPR and ARKitIPR, based on ScanNet-V2 [22] and ARKit [23]. The generated ARKitIPR dataset has 5037 scenarios (4196 for training, 741 for evaluation, and 100 for testing) and 377625 frames of point clouds in total. ScanNetIPR has 1605 scenarios (1193 for training, 312 for evaluation, and 100 for testing) and 53201 frames of point clouds in total. Both datasets contain 1. positive and negative matching frames, 2. the pose of each frame, 3. database

keyframes for the testing datasets and for recall calculation, and 4. point clouds that each contains positional x, y, z, normal n_x, n_y, n_z, color r, g, b information and semantic labels. Because overlap of frames is the major reference for place recognition, we use overlap to determine positively and negatively matched cases for the dataset generation, as shown in Alg. 1.

First, we generate database frames according to the coverage of frames from original datasets, ARKit and ScanNet-V2. In contrast to the odometry tasks, place recognition does not require very close frames to estimate the transformation matrices because those frames do not provide much different information for place retrieval and add burden in training. Therefore, the database frames are chosen by some overlap threshold along the trajectory of the camera motion, where if the current frame overlaps too much with the last frame it will be dropped, as shown in Alg. 1. The overlap in this step is calculated by the symmetric intersection of the union (IoU) of the voxelized frames. The FrameOverlap() in Alg. 1 is as in Eq. 6, where $\mathbf{P}_n$ and $\mathbf{P}_l$ are points chosen by frustums in the scenario. $\mathbf{v}_n$ and $\mathbf{v}_l$ are the voxels of the n_{th} frame and the last selected frame, respectively. $F_n = p_n \mathbf{P}_n$ represents the point cloud transformed in the global frame. The threshold $T_c = 0.5$ is used in the data generation task.

$$[\mathbf{v}_n, \mathbf{v}_l] = \text{voxelize}([p_n \mathbf{P}_n,\ p_l \mathbf{P}_l]), \tag{5}$$

$$\text{FrameOverlap}(\mathbf{P}_n, \mathbf{P}_l) = \frac{|v_n \bigcap v_l|}{|v_n \bigcup v_l|}. \tag{6}$$

Second, we choose the positive frames in the same scenario and negative cases in all scenarios. Since this step treats each frame as a query frame to calculate the overlap with other frames in the scenario, the overlap is mostly w.r.t. the query frame. Therefore, an asymmetric overlap metric can be used:

$$\text{Overlap}(\mathbf{P}_q, \mathbf{P}_d) = \frac{|v_q \bigcap v_o|}{|v_q|}, \tag{7}$$

where $\mathbf{P}_q$ and $\mathbf{P}_d$ are query and database frames. $\mathbf{v}_q$ and $\mathbf{v}_d$ are the voxels of these two frames with the same voxelization as Eq. 6. Then the positive frames are selected by the threshold Overlap $> T_p$. Negative frames can be selected by Overlap $<= T_n$. Normally, the negative threshold $T_n = 0$.

Third, we extract the key frames of each scenario, where we build a graph for each scenario using the positively matched frames and choose the dominating nodes as the keyframes of the scenario as [7].

V. Experiment

Our experimental design aims to evaluate the effectiveness and efficiency of our CSCPR against SOTA methods in RGB-D place recognition. To ensure a fair comparison, we conducted training and evaluation across multiple datasets, including ScanNet-PR, which uses a 3m threshold to select the database frames for place recognition, and newly proposed datasets, ScanNetIPR and ARKitIPR, which use overlap to select frames. For each dataset, we perform both training and testing for all models. The models are trained on 8 Tesla-V100 GPUs, and the input point clouds of CSCPR are constrained to 3000 points by voxelization downsampling. Evaluations are conducted in the device with an NVIDIA RTX A5000 GPU and an Intel Xeon(R) W-2255 CPU. The primary metric used for evaluation is the Recall@1-3, which is the percentage of cases where at least one within top-k candidates is positive. Our experiments are structured into three distinct evaluations:

E1: Place-Recognition Reranking Evaluation This experiment is to evaluate the performance of our innovations, SCC and CSCC, on reranking. To make a fair comparison, we compare all the place-recognition reranking approaches with the same global retrieval stage. We quantitatively and qualitatively demonstrate CSCPR's superior reranking performance in terms of accuracy and processing speed, which are our design criteria, over various SOTA approaches.

Comparisons with Other Reranking Approaches: ***a.*** *classical RANSAC-based geometric verification*, including RGB-based MAGSAC [47] to fit the homography of images [17] and RGB-D-based RANSAC + Kabsch [45]; ***b.*** *learning-based place recognition reranking approaches.* Because of the lack of learning-based place recognition reranking works in the RGB-D domain, we adapt the SOTA method, R2Former [13], from the RGB domain to RGB-D point space by extending the pixel-position encoding method to $\{x, y, z\}$ positions. ***c.*** *matching methods*, including RGB-based SuperGlue [40] and RGB-D-based URR [36] to calculate the matching pairs and use the number of matched pairs to determine if the frames overlap. ***d.*** *registration methods* [49], [46], [36], including TEASER++ [46], URR [36], and PointMBF [49], which are mostly used for localization, where given a pair of matched frames, the corresponding features are extracted from overlapped areas and used to calculate transformation matrices.

As shown in Tab I, we observe RGB-D-based approaches outperform corresponding RGB-based approaches. Our approach outperforms all other approaches by at least 3.19 and 3.17 at Recall@1 in ARKitIPR and ScanNetIPR, respectively. The learning-based approaches outperform RANSAC-based approaches. We also observe our approach has less inference time than other approaches by at least 20%, and this satisfies our design criterion; fast and effective. As shown in Fig 4, there are two scenarios. We compare our approach, CSCPR, with the closest place-recognition reranking R2Former and TEASER++. R2Former, based on attention models, does not perform well in scenarios with very different scales but with similar geometric features, as shown in the first two rows of Fig. 4. TEASER++ cannot match the color information of frames.

Ablation Study: As shown in Tab. II, we compare our approach with three different modified versions to highlight the benefits of our design. Attention [14] composes a sequence of self-attention and cross-attention models to substitute for our SCC and CSCC. The CoCs concept allows for multi-scale feature processing by relating large-scale environmental representations to small-scale object details,

Approaches	Data Type	Inference Time (ms)	*ARKitIPR*			*ScanNetIPR*		
			R@1 ↑	R@2 ↑	R@3 ↑	R@1 ↑	R@2 ↑	R@3 ↑
PoCo [1]	RGB-D	-	45.12	57.10	62.14	58.10	70.33	75.95
Kabsch [45]	RGB-D	93	53.00	59.37	65.59	66.07	74.89	80.99
TEASER [46]	RGB-D	107	56.03	62.35	66.75	69.40	80.94	82.82
R2Former [13]	RGB-D	5	71.94	76.95	79.24	80.05	82.09	85.02
MAGSAC [47]	RGB	18	20.60	33.03	38.49	41.00	47.25	52.30
SuperGlue [40]	RGB	40	49.19	55.50	59.44	69.84	75.83	78.72
SelaVPR [48]	RGB	11.4	56.32	63.81	67.56	67.31	76.47	80.60
URR [36]	RGB-D	90	31.86	40.03	46.57	46.50	56.10	62.28
PointMBF [49]	RGB-D	300	50.12	57.30	63.47	61.27	70.92	77.40
Ours	RGB-D	4	**75.13**	**80.24**	**82.33**	**83.22**	**87.76**	**89.30**

TABLE I: **Quantitative Reranking Results:** PoCo [1] is the global retrieval baseline. Our place-recognition reranking method outperforms other reranking approaches by at least **3.19** points in Recall@1.

Approaches	Data Type	Inference Time (ms)	*ARKitIPR*			*ScanNetIPR*		
			R@1 ↑	R@2 ↑	R@3 ↑	R@1 ↑	R@2 ↑	R@3 ↑
Attention [14]	RGB-D	11	64.14	67.65	68.81	74.95	82.94	86.22
Ours/SCC	RGB-D	3	65.56	68.39	69.11	76.52	80.03	81.03
Ours/Correlation	RGB-D	4	45.56	60.18	66.06	47.33	65.99	74.95
Ours	RGB-D	4	**75.13**	**80.24**	**82.33**	**83.22**	**87.76**	**89.30**

TABLE II: **Reranking Ablation Study:** Ours, Ours/SCC, and Ours/Correlation represent our complete reranking, without SCC and without the correlation matrix, respectively. Attention represents the method with a sequence of self-cross attention blocks [14]. The table shows effectiveness of our components and the outperformance w.r.t. an attention-based alternative by at least 11% in ARKitIPR and ScanNetIPR.

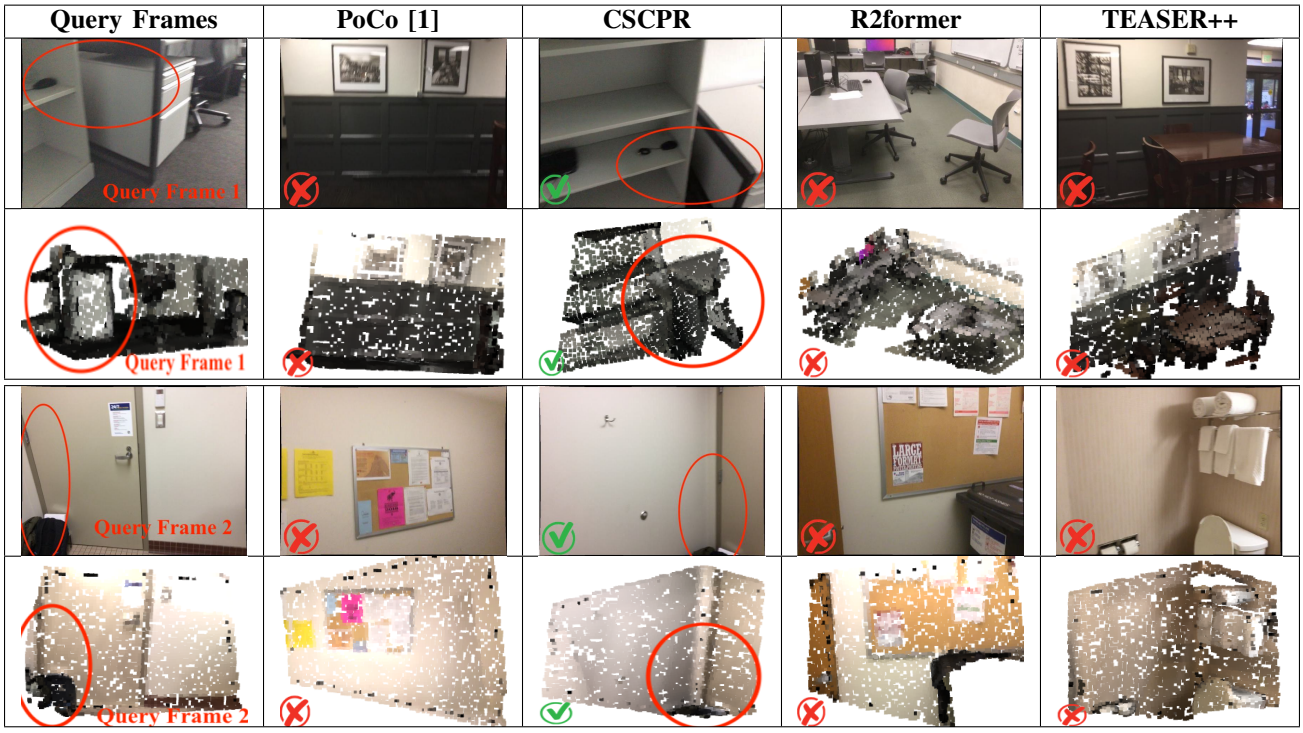

Fig. 4: **Qualitative Comparisons:** 1st and 3rd rows show RGB images corresponding to point clouds in 2nd and 4th rows. The red circles mark the overlapping areas between query frames (1st column) and later Recall@1 frames from different approaches. R2Former performs closest to our approach, but it does not perform well in different scaled frames. Our overall algorithm (CSCPR) balances the geometric and RGB information well and achieves the best performance, even for the scenarios that have very small overlapping areas.

unlike attention mechanisms, which operate at a single scale without multi-scale interactions. Therefore, our design based on the CoCs concept is faster and outperforms attention-based alternatives by at least 11% on ARKitIPR and ScanNetIPR in Recall@1. The attention model costs 12.5 Mb and 4.42 GMACs for a single batch, but our CSCPR costs 10 Mb and 3.01 GMACs, which is more computationally efficient. For the performance of single components, SCC improves the performance by at least 9% on the two datasets by enhancing the local features with global information. We also observe the correlation matrix is critical in reranking and improves by at least 65% on two datasets, as discussed in Section III-B, which provides the relationship between two frames. These results demonstrate the efficacy of our

innovative components (SCC and CSCC) in RGB-D place recognition reranking.

E2: End-to-End Solution Evaluation As mentioned in Section I, RGB-D indoor place recognition reranking is not well explored, and we also did not find any integrated approaches with two stages for RGB-D place recognition. To make a fair comparison, we compare SOTA RGB-D approaches with global retrieval with CSCPR in ScanNet-PR [11], as shown in Tab. III. Our approach outperforms other approaches by at least 29.27% in Recall@1. This experiment also evaluates the performance of new RGB-D point cloud-based datasets, ScanNetIPR and ARKitIPR, as shown in Tab. IV. We observe at least 43.24% improvement over other approaches in Recall@1 in the datasets.

E3: Proposed Dataset Analysis This experiment is to validate the performance of CSCPR but also to contribute valuable datasets to the community, facilitating further advancements in RGB-D indoor place recognition. Fig. 3 shows the benefits of our datasets with less noise in the datasets. From the comparison of Tab. III and IV, we observe approaches in our ScanNetIPR dataset have lower recall, meaning our ScanNetIPR dataset is more difficult than ScanNet-PR. The reason is ScanNetPR has more positive frames, 19.94, on average for each frame in each scenario than ScanNetIPR, which has 15.98 on average. From Tab. IV, we show our new generated datasets are valid for RGB-D indoor place recognition with performance comparable to the results in Tab. III.

ScanNet-PR	Data Type	R@1 ↑	R@2 ↑	R@3 ↑
SIFT [25] + BoW [50]	RGB	16.16	21.17	24.38
NetVLAD [16]	RGB	21.77	33.81	41.49
PointNetVLAD [15]	Point Cloud	27.10	32.10	37.01
MinkLoc3D [32]	Point Cloud	15.21	19.25	22.79
Indoor DH3D [12]	RGB-D	16.10	21.92	25.30
CGiS-Net [11]	RGB-D	61.12	70.23	75.06
PoCo [1]	RGB-D	64.63	75.02	80.09
AEGIS-NET [51]	RGB-D	65.09	74.26	79.06
CGiS-Net w/o color [11]	Point Cloud	39.62	50.92	56.14
PoCo [1] w/o color	Point Cloud	44.34	54.27	59.78
CSCPR w/o color	Point Cloud	60.58	73.59	78.29
CSCPR	RGB-D	**84.14**	**89.82**	**91.25**

TABLE III: **Quantitative Results:** Compared with other SOTA methods, our approach CSCPR with reranking improved the Recall@1 performance by at least **29.27**%.

VI. Conclusion, Limitations, and Future Work

We explored RGB-D place recognition with an integrated pipeline with both global retrieval and reranking. By developing a fast and effective reranking model, we close the gap between RGB-D place recognition reranking and learning-based algorithms. We generalized the CoCs concept to noisy colorized point cloud feature processing and demonstrated better performance in place recognition tasks. We handle the scarcity of RGB-D place recognition datasets and propose a data generation pipeline for the community to explore more datasets. We introduce two large-scale RGB-D datasets for training and testing purposes. We push forward the boundary of RGB-D indoor place recognition accuracy by demonstrating that our design outperforms other SOTA approaches by at least 29.27% in Recall@1 in the ScanNet-PR dataset. For limitations, if the overlapping areas do not have many features, our method may not work well. As part of future work, we would like to apply semantic pretraining to the model to improve its understanding of the environment.

	Approaches	Data Type	R@1 ↑	R@2 ↑	R@3 ↑
ScanNetIPR	PointNetVLAD	Point Cloud	22.43	30.81	36.58
	MinkLoc3D	Point Cloud	10.13	16.63	20.80
	CGiS-Net	RGB-D	57.89	69.95	75.51
	AEGIS-NET [51]	RGB-D	58.00	69.12	74.79
	PoCo [1]	RGB-D	58.10	70.33	75.95
	CSCPR	RGB-D	**83.22**	**87.76**	**89.30**
ARKitIPR	PointNetVLAD	Point Cloud	11.04	16.57	20.57
	MinkLoc3D	Point Cloud	8.14	10.95	13.79
	CGiS-Net	RGB-D	39.80	49.30	55.64
	AEGIS-NET [51]	RGB-D	45.71	57.71	63.34
	PoCo [1]	RGB-D	45.12	57.10	62.14
	CSCPR	RGB-D	**75.13**	**80.24**	**82.33**

TABLE IV: **Comparisons in New ScanNetIPR and ARKitIPR:** The ScanNetIPR is more difficult than ScanNetPR, where methods have relatively smaller Recall@1. Our method still outperforms other approaches for both RGB-D and pure point-cloud place recognition in both ScanNetIPR and ARKitIPR by at least 43.24% in Recall@1.

References

[1] J. Liang, Z. Deng, Z. Zhou, O. Ghasemalizadeh, D. Manocha, M. Sun, C.-H. Kuo, and S. Arnie, "Poco: Point context cluster for rgbd indoor place recognition," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[2] 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

[3] 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

[4] 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.

[5] 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.

[6] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[7] A. Kornilova, I. Moskalenko, T. Pushkin, F. Tojiboev, R. Tariverdizadeh, and G. Ferrer, "Dominating set database selection for visual place recognition," *arXiv preprint arXiv:2303.05123*, 2023.

[8] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 268–283.

[9] 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.

[10] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[11] 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.

[12] 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.

[13] 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.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] 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.

[16] 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.

[17] 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.

[18] K. Vidanapathirana, P. Moghadam, S. Sridharan, and C. Fookes, "Spectral geometric verification: Re-ranking point cloud retrieval for metric localization," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2494–2501, 2023.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[20] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," in *proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 105–12 115.

[21] S. Lee, H. Seong, S. Lee, and E. Kim, "Correlation verification for image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5374–5384.

[22] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[23] 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

[24] 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.

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

[26] 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.

[27] 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.

[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] 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.

[30] 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.

[31] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[32] 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.

[33] 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.

[34] W. Zhang, H. Zhou, Z. Dong, Q. Yan, and C. Xiao, "Rank-pointretrieval: Reranking point cloud retrieval via a visually consistent registration evaluation," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[35] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[36] B. et al., "Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering," 2021.

[37] G. Mei, H. Tang, X. Huang, W. Wang, J. Liu, J. Zhang, L. Van Gool, and Q. Wu, "Unsupervised deep probabilistic approach for partial point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 611–13 620.

[38] A. Hatem, Y. Qian, and Y. Wang, "Point-tta: Test-time adaptation for point cloud registration using multitask meta-auxiliary learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 494–16 504.

[39] M. Li, Z. Qin, Z. Gao, R. Yi, C. Zhu, Y. Guo, and K. Xu, "2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 128–14 138.

[40] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[41] 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.

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

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

[44] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

[45] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, no. 5, pp. 698–700, 1987.

[46] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[47] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1304–1312.

[48] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," *arXiv preprint arXiv:2402.14505*, 2024.

[49] M. Yuan, K. Fu, Z. Li, Y. Meng, and M. Wang, "Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 17 648–17 659. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01622

[50] 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.

[51] Y. Ming, J. Ma, X. Yang, W. Dai, Y. Peng, and W. Kong, "Aegis-net: Attention-guided multi-level feature aggregation for indoor place recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4030–4034.