2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

GIDP: Learning a Good Initialization and Inducing Descriptor Post-enhancing for Large-scale Place Recognition

Zhaoxin Fan[1], Zhenbo Song[2], Hongyan Liu[3], Jun He[1]

Abstract— Large-scale place recognition is a fundamental but challenging task, which plays an increasingly important role in autonomous driving and robotics. Existing methods have achieved acceptable good performance, however, most of them are concentrating on designing elaborate global descriptor learning network structures. The importance of feature generalization and descriptor post-enhancing has long been neglected. In this work, we propose a novel method named GIDP to learn a Good Initialization and Inducing Descriptor Pose-enhancing for Large-scale Place Recognition. In particular, an unsupervised momentum contrast point cloud pretraining module and a reranking-based descriptor post-enhancing module are proposed respectively in GIDP. The former aims at learning a good initialization for the point cloud encoding network before training the place recognition model, while the later aims at post-enhancing the predicted global descriptor through reranking at inference time. Extensive experiments on both indoor and outdoor datasets demonstrate that our method can achieve state-of-the-art performance using simple and general point cloud encoding backbones.

I. INTRODUCTION

Large-scale place recognition is a challenging but important task, which plays an increasingly important role in autonomous driving [1], [2], [3] and robotics [4], [5], [6]. The place recognition result is always used to provide the agent with accurate localization information. For instance, suppose there is a robot waiter in a hotel who is asked to transport luggage to the guest's room. To locate the robot itself, place recognition is always required, especially in an indoor environment where GPS signal is not available.

For large-scale place recognition task, a HD map of the zone where the robot serves is always pre-built. Then, the HD-map is cut into thousands of different submaps. The localization information of each submap in the HD map is known information. When the robot walks in the zone, it compares the current captured scene with submaps in the database and finds the most similar one. To this end, the location of current scene is equivalent to the known location of the most similar submap. A straight forward way to achieve the above place recognition task is to use images to represent scenes [7], [8], [9]. However, images are sensitive to environmental changes such as weather change and season change [10]. Compared to image, point cloud captured by LiDAR is much more robust towards variants caused by the above environmental changes. Therefore, in this paper, we study the more robust point cloud based large-scale place recognition.

[1]Renmin University of China, {fanzhaoxin, hejun}@ruc.edu.cn
[2]Nanjing University of Science and Technology, {songzb}@njust.edu.cn
[3]Tsinghua University, {hyliu}@tsinghua.edu.cn

Fig. 1: The different between our work and previous methods. (a) Training and inference pipeline of previous methods. (b) Training and inference pipeline of our GIDP.

PointNetVLAD [11] is the first deep learning based work for large-scale place recognition. It uses PointNet [12] and NetVLAD [7] to learn point cloud descriptors. Then, following works [13], [14], [15], [16], [17], [18], [10] try to improve the performance through the perspective of using graph networks [14], [17], [16], designing attention mechanism [13], [15], and using sparse convolution [18], [10]. Though all of them achieve remarkable performance, the above mentioned methods all focus on designing novel network architectures for better global point cloud descriptor learning. The equally important network weights initialization and descriptor post-processing have long been neglected as shown in Fig. 1 (a).

To tackle the above issue, we propose a novel method named GIDP to learn a Good Initialization and Inducing Descriptor Pose-enhancing for Large-scale Place Recognition as shown Fig. 1 (b). The intuition behind GIDP is that we support the opinion that learn a pretraining model tailored for point cloud-based large-scale recognition is very important. Moreover, we are also motivated by the fact that the learned global descriptors can be further enhanced before similarity calculation. To this end, GIDP consists of two main modules. The first one is an unsupervised momentum contrast point cloud pretraining module, which aims at learning a good initialization for the descriptor encoding network. Benefited from the power of our pretraining module, a simple point cloud encoding network such as PointNet [12] and DGCNN [19] can achieve promising result. Result in a more light-weight network architecture and more efficient inference than existing methods. The second module is a re-ranking based descriptor post-enhancing module. In this module, a descriptor is interacted with other descriptors in an simple yet effective manner and then re-ranked at inference time. In this way, the descriptor adopts more semantic information to describe the scene and finally result in performance improvement. To the best of knowledge, we are the first to design pretraining models and post-processing modules for point cloud-

based large-scale place recognition.

We conduct extensive experiments on several widely used indoor and outdoor datasets. Results show that our method achieves new state-of-the-art performance using simple backbones. Our contributions can be summarized as follows:

- We propose GIDP, a novel point cloud based large-scale place recognition method, which achieve state-of-the-art performance using simple backbones.
- We propose an unsupervised momentum contrast point cloud pretraining module and a re-ranking based descriptor post-enhancing module to improve the place recognition performance.
- We conduct extensive experiments on both indoor and outdoor datasets to show the effectiveness of our method and the proposed modules.

II. Related Work

In our work, in this section, we introduce the related work. Since we aim at improving the performance of point cloud-based large-scale place recognition methods, we first introduce the state-of-the-art place recognition models. Then, we present recently popular unsupervised training methods. Finally, we introduce related re-ranking methods.

A. Large-scale place recognition

Large-scale place recognition can be divided into image-based methods [7], [8], [9] and point cloud based methods [13], [14], [15], [16], [17], [18], [10]. Image based methods are proven to be sensitive to environmental change, hence attract less attention in recent year. In contrast, point cloud-based methods are very robust benefited from the inherent properties of LiDAR. PointNetVLAD [11] is the first deep learning method for point cloud large-scale place recognition simply uses PointNet [12] and NetVLAD [7]. Then, DAGC [14] and SR-Net [17] utilizes dynamic graph network and static graph network respectively to for this task, while SOE-Net [15] and PCAN [13] propose different attention mechanism for this task. LPD-Net [16] uses hand-craft features to further improve performance. Recently, sparse convolution [20] becomes popular, therefore, methods like [18], [10] are proposed to use sparse convolution to learn point cloud descriptors. Though promising, all the above proposed methods are trying to design better network structures for performance improvement. In contrast, we propose to learn a good initialization and good post-enhancing to improve place recognition performance.

B. Unsupervised pretraining

Unsupervised pretraining has become very popular in recent years due to its effectiveness. In the natural language processing (NLP) field, BeRT [21] is a phenomenal work that utilizes unsupervised pretrain to train a transformer model, which now is the most widely used NLP backbone. In computer vision field, MOCO V1/V2 [22], [23] propose to use contrastive learning for pretraining and use data augmentation to generate positive and negative samples. SimCLR [24] researches the combination of different data augmentation methods and proposes a projection head in its pretraining framework. DINO [25] introduces an unsupervised pretraining framework and finds that the attention map can well describe the foreground objects. The above mentioned methods are all image-level methods, while DenseCL [26] presents a dense contrastive training pipeline to densely pretrain the network in pixel level, which is more suitable for dense tasks such as segmentation. In point cloud processing, there are also some works [27], [28] are proposed using contrastive learning. In our work, we also use contrastive learning for unsupervised point cloud encoding network pretraining. Our method is tailored for large-scale place recognition.

C. Re-ranking methods

In the current large-scale place recognition framework, the task is essentially a retrieval problem. In retrieval, re-ranking of learned descriptors is a hot topic. [29] proposes K-reciprocal feature, which encodes the K- reciprocal features into a singe vector for re-ranking. [30] proposes a CNN semantic re-ranking system that greatly improves the retrieval performance. [31] introduces a method to meta-learn the re-ranking updates for image retrieval. In UED [32], the re-ranking step is involved by running a diffusion process on the underlying data manifolds. In the filed of NLP, Rocketqav2 [33] proposes to adaptively improve the retriever and the re-ranker according to each other's relevance information. Though the above proposed methods demonstrate good performance, they are designed for images or texts. In our work, we propose a simple yet effective method for re-ranking descriptors learned from point clouds for the large-scale place recognition task.

III. Methodology

A. Problem Formulation

Given a HD-Map of a zone, we firstly cut it into a database of submaps: $\mathcal{D} = \{M_1, M_2, \cdots, M_n\}$, where M_i is the $i-th$ submap in database $\mathcal{D}$. Note the localization information of M_i is known information. Then, assume that a robot is walk in the zone. The robot is asked to capture a scene $S \in R^{N\times 3}$ represented as point cloud at each walking step. Our goal is to find the most similar submap M_s of S. Then, the location of S is equal to the location of M_s. To do so, we learn a deep learning model $\mathcal{F}$ to encode S and M_i to a vector v_s and v_i. Then, a KNN algorithm ϕ can be used to find M_s. The whole process can be defined as:

$$M_s = \phi(\mathcal{F}(S), \mathcal{F}(M_1, M_2, \cdots, M_n)) \tag{1}$$

Fig. 2: Pipeline of our method. We use a simple backbone along with a GeM pooling module to consist of the point cloud encoding network. To train the network, an unsupervised momentum contrast point cloud pretraining module is proposed to first pretrain the network. Then, we supervisely train the network using contrastive learning. Finally, we introduce a re-ranking based descriptor post-enhancing module to improve the powerful of predicted global descriptors.

B. Overview

As mentioned before, training a point cloud encoding network and post-processing the predicted global point cloud descriptors are the most important factors of obtaining an excellent large-scale place recognition model. Therefore, in this work, we mainly research the two factors.

Fig. 2 illustrates the pipeline of our work GIDP. As shown in the figure, our work can be divided into three stages. The first stage is an unsupervised momentum contrast point cloud pretraining module. This module takes a batch of point clouds as input and uses contrastive learning to unsupervisely pretrain the point cloud encoding network. Then, in the second stage, we use the weights pretrained in the last stage to initialize the network and construct a supervised contrastive learning framework to train the network. In this stage, a GeM pooling [] is used to aggregate global descriptor. After training, we can use the trained network to predict point cloud descriptor for place recognition. However, we find the output descriptor is not powerful enough for large-scale place recognition. Therefore, at inference time, we use the third stage, the re-ranked based post-enhancing module to enhance the representational ability of the descriptor. Finally, a K-Nearest-Neighbor (KNN) algorithm can be used to find the most similar submap for each scene using these enhanced (or called re-ranked) descriptors. Next, we introduce the detail of the unsupervised momentum contrast point cloud pretraining module, the re-ranking based post enhancing module and the loss function used to train the network.

C. Unsupervised momentum contrast point cloud pretraining module

Though existing large-scale place recognition methods have achieved promising performance, there is still much room for their performance improvement. The main reason is that existing methods training the point cloud encoding network from scratch, which is easy to cause over-fitting. The consequence is that the abundant scene semantic and scene geometry hidden in the point cloud can not be fully parsed, and so the finally predicted descriptors will be less effective. A possible solution is to pretrain the network before training for the particular place recognition task. Hence, the network would learn a good initialization where the feature is more powerful and more general. In this section, we introduce such a pretraining module named unsupervised momentum contrast point cloud pretraining module to tackle the issue.

Constructing training triplets: Utilizing contrastive learning for pretraining is proven to be very effective. For every anchor point cloud P_a, at least a positive anchor P_{pos} and a negative anchor P_{neg} are required. However, since we aim at pretraining the network unsupervisely, there is no ground-truth information we can use except the point cloud itself to construct the {anchor, positive sample, negative sample} triplet. To this end, we propose to construct the triplet by data augmentation as introduced in [24]. Specifically, for P_a, we apply random data augmentation to generate its positive samples. Then, any other point clouds in the dataset can be its negative samples. Since we design the pretraining module for the large-scale place recognition task, the data augmentation should be carefully designed to maintain the scene semantic and geometry. Therefore, the specific data augmentation we choose includes: random jitter, random points removal, random block removal, and random shear.

Contrastive learning: Taking a batch of point cloud as input, suppose the output of the point cloud encoding network is $\{v_a, v_{p,1}, v_{p_2}, \cdots, v_{p,m_p}, v_{n,1}, v_{n,2}, \cdots, v_{p,m_n}\}$, where v_a is the descriptor of an anchor point cloud, $v_{p,i}$ is the descriptor of the positive sample and $v_{n,i}$ is the descriptor of the negative sample. We follow [24] to use a projection head to project v_i to u_i,i.e., $\{u_a, u_{p,1}, u_{p21}, \cdots, u_{p,m_p}, u_{n,1}, u_{n,2}, \cdots, u_{p,m_n}\}$. The projection head is implemented as a MLP layer. Then, for v_a, we choose one positive sample $u_{p,+}$ and K negative samples to calculate the InfoNCE [34] loss:

$$L_{pretrain} = -log\frac{exp(u_a \cdot u_{p,+})}{\sum_1^K exp(u_a \cdot u_{n,k})} \tag{2}$$

Momentum update: Following MOCO [22], we record all descriptors as a queue to calculate the InfoNCE loss, which can make the dictionary large, therefore the back-propagation will not be limited by the information in a mini-batch. However, it also makes it intractable to update the point cloud encoding network during back-propagation. To this end, we also use the momentum update to address this issue. Specifically, we use two point cloud encoding networks during training, one is the anchor encoder which takes the anchor point cloud as input, and the another one is the momentum encoder which takes the positive samples and negative samples as input. The two encoders share the same network structure. Suppose the parameters of the anchor encoder is θ_a and the parameters of the momentum encoder is

θ_{pn}. We update θ_a by back-propagation while update θ_{pn} by:

$$\theta_{pn} \leftarrow m\theta_{pn} + (1-m)\theta_a \tag{3}$$

Without bells and whistles, through the above process, the point cloud encoding network can be comprehensively pretrained. Hence, performance of the downstream large-scale place recognition task can be greatly improved, evidenced by experimental results.

D. Re-ranking based descriptor post-enhancing module

After the unsupervised pretraining stage and the supervised encoding network training stage, the point cloud encoding model owns the ability of predicting discriminative global point cloud descriptors. However, the information hidden in the training dataset is still fully utilized. That is to say that the descriptor can still be enhanced using the training dataset. Motivated by this, in this section, we introduce a re-ranking based descriptor post-enhancing module, which can utilize the training set to further improve the power of descriptors. Related descriptors exploration: For a scene presented by point cloud P_a, suppose its descriptor is $v_a \in R^C$, we argue that other descriptors $v_{o,i} \in \mathcal{D}$ in the training set share some general semantic and geometric features with it. Some of these features can be used to enhance v_a. To utilize such kind of information, we should first find the K most relevant descriptors of v_a. In our work, we use a KNN algorithm in the latent high-dimensional feature space to find them:

$$\{v_{o,1}, v_{o,2}, \cdots, v_{o,k}\} = KNN(v_a, \mathcal{D}) \tag{4}$$

Inverse distance based enhancing: After finding $\{v_{o,1}, v_{o,2}, \cdots, v_{o,k}\}$, we need to use them to enhance v_a. Note since we use contrastive learning to train the encoding network, theoretically, similar descriptors would locate close in the feature space while dissimilar ones are far away from each other. Therefore, distance in the feature space is an important factor to reflect the descriptor relationship. To this end, we propose an inverse distance based enhancing method to utilize the distance relationship of descriptors to enhance v_a. The enhancing process can be formulated as:

$$\hat{v_a} = \lambda v_a + (1-\lambda)\sum_{1}^{K}(w_k \cdot v_{o,k}) \tag{5}$$

where $w_k = \frac{exp(-|v_a - v_{o,k}|)}{\sum_1^K exp(-|v_a - v_{o,i}|)}$, $\hat{v_a}$ is the enhanced global descriptor, and λ is a balance term.

Inductive vs Transductive: Since the training set is pre-collected, it is very straight forward to re-use it at inference time for enhancing the descriptor in an on-line manner. This is in essence an inductive setting. We also note that in some cases, at inference time, we will collect many scenes at different location and then retrieve the the location of each scene in an off-line manner. In such case, other scene collected at inference time can also be additionally used to enhance the descriptor. This inference time enhancing is called a Transductive setting. In our work, we experiment with both settings. Both of them demonstrate good performance.

Through above process, we can post-enhance all point cloud embeddings at inference time. In this way, the final retrieval stage is equivalent to adopting a re-ranking process. Through this re-ranking, the distribution of descriptors in the feature space would be more distinctive, hence the final place recognition results can be significantly increased.

E. Loss

To supervisely train or finetune the point cloud encoding network after the unsupervised pretraining, we adopt the following triplet loss to train our model on view of its superior performance in [18]:

$$L(v_a, v_p, f_n) = max\{d(v_a, v_p) - d(v_a, v_n) + m, 0\} \tag{6}$$

where v_a is the descriptor of the query scan, v_p and v_n are descriptors of positive sample and negative sample respectively, and m is a margin. $d(x, y)$ means the Euclidean distance between x and y. To build informative triplets, we use batch-hard negative mining following [18].

IV. Experiments

In this section, we first introduce the datasets we use. Then, we present the implementation details. Next, the comparison results are discussed. Finally, we present the ablation study.

A. Dataset

For fair comparison, we conduct experiments on the benchmark datasets proposed by [11], which is the most widely datasets for point cloud-based large-scale place recognition. Its benchmark consists of four datasets: one outdoor dataset named Oxford generated from Oxford RobotCar [35] and three in-house datasets: university sector (U.S.), residential area (R.A.) and business district (B.D.). The four datasets contain 21711, 400, 320, 200 submaps for training and 3030, 80, 75, 200 submaps for testing for Oxford., U.S., R.A. and B.D. respectively. Each point cloud contains 4096 points. During training, point clouds are regarded as correct matches if they are at maximum 10m apart and wrong matches if they are at least 50m apart. In testing, the retrieved point cloud is regarded as a correct match if the distance is within 25m between the retrieved point cloud and the query scan. We use average recall at top 1% and average recall at top 1 as main metrics as previous methods [11], [10], for a fair comparison.

B. Implementation Details

We implement our method using PyTorch. We train two different version of models using PointNet and DGCNN as backbones respectively. At pretraining state, the batch size is 64. The learning rate is 0.03. The model is pretrained for 100 epochs using Adam optimizer. The

Method	Average recall at top-1% (%)				Average recall at top-1 (%)			
	Oxford	U.S.	R.A.	B.D.	Oxford	U.S.	R.A.	B.D.
PointNetVLAD	80.3	72.6	60.3	65.3	-	-	-	-
PCAN	83.8	79.1	71.2	66.8	-	-	-	-
DAGC	87.5	83.5	75.7	71.2	-	-	-	-
SOE-Net	96.4	93.2	91.5	88.5	-	-	-	-
SR-Net	94.6	94.3	89.2	83.5	86.8	86.8	80.2	77.3
LPD-Net	94.9	96.0	90.5	89.1	86.3	87.0	83.1	82.3
Minkloc3D	97.9	95.0	91.2	88.5	93.0	86.7	80.4	81.5
SVT-Net	97.8	96.5	92.7	90.7	93.7	90.1	84.3	85.5
GIDP+PointNet+Inductive	86.8	75.8	72.1	68.2	73.6	61.1	57.8	58.6
GIDP+PointNet+Transductive	88.7	77.3	73.3	66.5	77.1	64.9	60.8	57.1
GIDP+DGCNN+Inductive	98.0	98.1	94.8	91.5	92.6	91.7	87.8	86.0
GIDP+DGCNN+Transductive	98.6	98.8	95.5	91.1	94.5	93.5	90.5	85.3

TABLE I: Comparison with state-of-the-art under the baseline setting.

Method	Average recall at top-1% (%)				Average recall at top-1 (%)			
	Oxford	U.S.	R.A.	B.D.	Oxford	U.S.	R.A.	B.D.
PointNetVLAD	80.1	90.1	93.1	86.5	63.3	86.1	82.7	80.1
PCAN	86.4	94.1	92.3	87.0	70.7	83.7	82.3	80.3
DAGC	87.8	94.3	93.4	88.5	71.4	86.3	82.8	81.3
SOE-Net	96.4	97.7	95.9	92.6	89.3	91.8	90.2	89.0
SR-Net	95.3	98.5	93.6	90.8	88.5	93.5	86.8	85.9
LPD-Net	98.2	98.2	94.4	91.6	93.0	90.5	97.4	85.9
Minkloc3D	98.5	99.7	99.3	96.7	94.8	97.2	96.7	94.0
SVT-Net	98.4	99.9	99.5	97.2	94.7	97.0	95.2	94.4
GIDP+DGCNN+Inductive	98.0	99.7	98.7	96.2	93.3	96.4	94.9	93.1
GIDP+DGCNN+Transductive	98.6	99.8	99.2	95.8	95.1	97.3	97.8	92.5

TABLE II: Comparison with state-of-the-art under the refined setting.

feature dimension of both u_i and v_i is 256. At encoding network supervised training stage, we train our model under two setting: the baseline setting which only uses training set of Oxford to train the model, and the refined setting which additionally add training set of U.S. and R.A. In the baseline setting, the initial batch size is 32 and the initial learning rate is 10^{-3}. The model is trained for 40 epochs and the learning rate is decayed by 10 at the end of the 30th epoch. The refined model is trained with an initial batch size of 16 and an initial learning rate of 10^{-3}. The model is trained for 80 epochs and the learning rate is decayed by 10 at the end of the 60th epoch. At the final re-ranking based post-enhancing stage, λ is set to 0.2. The dimension of the final global point cloud descriptor is 256. The number of neighbors K is 5. All experiments are conducted on a single A6000 GPU.

C. Results

In this section, we conduct experiments on Oxford., U.S., R.A. and B.D. and compare our method with the state-of-the-art methods include PointNetVLAD [11], PCAN [13], DAGC [14], SR-Net [17], LPD-Net [16], SOE-Net [15], Minkloc3D [18] and SVT-Ne [10]. PointNet and DGCNN are used are backbones respectively. We show results on both baseline setting and refined setting. We also show results on both Inductive setting and Transductive setting.

Results on the baseline setting: In Table I, we show the results on the baseline setting. We can find from the table that 1) Even using the simplest baseline PointNet, our method performs very well. PointNetVLAD also use PointNet as their backbone network, while our method outperforms it for a large margin, showing the superiority of our method. Our method using PointNet even outperform DAGC that uses DGCNN as the backbone, greatly demonstrating the effectiveness of our method. 2) When using DGCNN as the backbone, our method outperforms all previous methods and achieves new stae-of-the-art. Note DGCNN is very simple and light-weight backbone compared backbone used by other methods. 3) The performance of our method under the Transductive setting outperforms that of the Inductive setting. That is because under the Transductive setting, other point clouds collected at the inference time are used for descriptor post-enhancing, bringing more semantic information and geometry information to the query descriptor. 4) Stronger backbones can bring better performance. It can been seen that the model using DGCNN performs significantly better than the model using PointNet. That is because DGCNN is stronger in learning local features, therefore it will benefit more from our pretraining stage and post-enhancing stage. In a word, our GIDP is superior than all existing methods though only simple backbones are used. We contribute the superiority to the proposed unsupervised momentum contrast point cloud pretraining module and the re-ranking based descriptor post-enhancing module.

Results on the refined setting: In Table II, we show the refined results of our method and other methods. In this experiment, we only show the result of using DGCNN as backbone. It can be found that after adding the training set of indoor datasets U.S. and R.A., the

Variants	Avg recall at top 1% Oxford	U.S.	R.A.	B.D.
Random Init+PoinNet	83.0	69.5	72.5	62.5
Random Init+PoinNet+Inductive	87.3	70.9	74.7	64.0
Random Init +PoinNet+Transductive	85.1	69.1	72.5	63.2
GIDP+PoinNet+Inductive	86.8	75.8	72.1	68.2
GIDP+PoinNet+Transductive	88.7	77.3	73.3	68.5

TABLE III: Results of ablation study

variants	Avg recall at top 1% Oxford	U.S.	R.A.	B.D.
without Jitter	98.0	98.1	93.0	91.3
without Randompoints Removal	97.5	97.2	93.1	89.1
without Randomblock Removel	97.8	97.0	94.1	90.3
without RandomShear	97.6	97.2	92.7	88.8
GIDP+DGCNN+Iransductive	98.6	98.8	95.5	91.1

TABLE IV: Impact of data augmentation in unsupervised pretraining.

performance of all methods are improved. Under the Transductive setting, our method performs comparable with the current state-of-the-art model SVT-Net, though we use a much simpler backbone while SVT-Net use sparse Transformers. Under the Transductive setting, benefited from the additional inference time captured point clouds, our method wins 3 out 4 datasets at the more strict average recall at top 1 metric. This demonstrates the superiority of our re-ranking based post-enhancing module. However, we also find that the difference between our method and SVT-Net is tiny. This is partly because the current dataset is too small so the performance become saturated. Therefore, in the future work, a new more large-scale dataset should be built to benchmark novel methods.

D. Ablation study

In this section, we study the impact of our proposed models and other key designs. All experiments are conducted at the baseline training setting.

Effectiveness of unsupervised momentum contrast point cloud pretraining module: The unsupervised momentum contrast point cloud pretraining module is one of the key contributions of our work. This module plays a role of learning a good initialization for the point cloud encoding network. In Table III, we show the result of without using this module. It can be found that without this module, the performance of the model is decreased significantly. This greatly demonstrate the importance of introducing the pretraining stage for large-scale place recognition.

In the unsupervised momentum contrast point cloud pretraining module, we use random data augmentation to generate positive samples. In Table IV, we show the results of removing one kind of data augmentation each time. It can be found from the table that all four kinds of random data augmentation plays a significant role of improving performance. It is because these data augmentation methods can simulate different situations of capturing the same scene with different geometries using LiDAR, so the model can learn more general features through pretraig.

In the pretraining stage, as introduced before, we adopt a projection head after the point cloud encoding network. In Table V, we investigate the impact of the projection head. It can be seen that without the projection head. The performance decreases a lot. This demonstrates that the effectiveness of adding the projection head, which increases the generalization ability of the pretrained weights.

Effectiveness of the re-ranking based descriptor post-enhancing module: In the re-ranking based descriptor post-enhancing module, each descriptor is interact with its neighbors in the feature space to enhance its representational ability. In the top 3 rows of Table III, we show the impact of this module. To avoid the influence of the unsupervised momentum contrast point cloud pretraining module, we randomly initialize point cloud encoding network during training. PointNet is used as the backbone. It can be seen from the table that without this module, the model performs relatively poor. After adding the module, the performance of the model increases for a large margin. Then, if we use the Transductive setting, the result is further increased. These experiments greatly demonstrate the correctness of designing the re-ranking based descriptor post-enhancing module.

Impact of different backbones: In Table VI, we show the performance of our method with different backbones. It can be seen that stronger backbones can bring better results. For the DGCNN backbone, we test the results of indexing different number of neighbors when learning local features. It can be found that with the increasing of number of neighbors, the performance increases. However, when it is larger than 16, the improvement is tiny. Therefore, we choose 16 as our default setting.

variants	Avg recall at top 1% Oxford	U.S.	R.A.	B.D.
without Projection Head	97.6	96.2	94.6	89.5
GIDP+DGCNN+Iransductive	98.6	98.8	95.5	91.1

TABLE V: Impact of projection head in unsupervised pretraining.

variants	Avg recall at top 1% Oxford	U.S.	R.A.	B.D.
GIDP+PoinNet+Inductive	86.8	75.8	72.1	68.2
GIDP+DGCNN8+Inductive	97.6	96.2	92.8	87.8
GIDP+DGCNN16+Inductive	97.6	96.2	94.6	89.5
GIDP+DGCNN24+Inductive	97.4	97.0	93.8	90.3

TABLE VI: Impact of different backbones

V. CONCLUSIONS AND LIMITATION

In this paper, we propose a novel method named GIDP to learn a good initialization and inducing descriptor post-enhancing for point cloud based large-scale place recognition. In GIDP, an unsupervised momentum contrast point cloud pretraining module and a re-ranking based descriptor post-enhancing module are proposed to

achieve our goal. We have conducted extensive experiments on both indoor and outdoor datasets. Results show that our method can achieve state-of-the-art performance using simple backbones.

Though our method achieves good performance, there still exist limitations. For example, currently, we only investigate pretrain point-based backbones, how to pretrain other type of backbones, such as sparse voxel-based backbones, is not studied. In the future, we plan to investigate more novel pretrain methods and make the training pipeline more efficient.

VI. Acknowledgement

This work was supported in part by National Key Research and Development Program of China under Grant No. 2020YFB2104101 and National Natural Science Foundation of China (NSFC) under Grant Nos. 62172421, 71771131, and 62072459. We acknowledge the support from the National Demonstration Center for Experimental Education of Information Technology and Management (Renmin University of China).

References

[1] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," arXiv preprint arXiv:2203.10638, 2022.

[2] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 2583–2589.

[3] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," arXiv preprint arXiv:2207.02202, 2022.

[4] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," IEEE Access, vol. 8, pp. 39 830–39 846, 2020.

[5] D. Belanche, L. V. Casaló, C. Flavián, and J. Schepers, "Service robot implementation: a theoretical framework and research agenda," The Service Industries Journal, vol. 40, no. 3-4, pp. 203–225, 2020.

[6] M.-H. Huang and R. T. Rust, "Engaged to a robot? the role of ai in service," Journal of Service Research, vol. 24, no. 1, pp. 30–41, 2021.

[7] 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.

[8] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," IEEE transactions on neural networks and learning systems, vol. 31, no. 2, pp. 661–674, 2019.

[9] 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.

[10] Z. Fan, Z. Song, H. Liu, Z. Lu, J. He, and X. Du, "Svt-net: Super light-weight sparse voxel transformer for large scale place recognition." AAAI, 2022.

[11] 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.

[12] 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.

[13] W. Zhang and C. Xiao, "Pcan: 3d attention map learning using contextual information for point cloud based retrieval," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 436–12 445.

[14] Q. Sun, H. Liu, J. He, Z. Fan, and X. Du, "Dagc: Employing dual attention and graph convolution for point cloud based place recognition," in Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 224–232.

[15] Y. Xia, Y. Xu, S. Li, R. Wang, J. Du, D. Cremers, and U. Stilla, "Soe-net: A self-attention and orientation encoding network for point cloud based place recognition," in Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2021, pp. 11 348–11 357.

[16] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu, "Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2831–2840.

[17] Z. Fan, H. Liu, J. He, Q. Sun, and X. Du, "Srnet: A 3d scene recognition network using static graph and dense semantic fusion," in Computer Graphics Forum, vol. 39, no. 7. Wiley Online Library, 2020, pp. 301–311.

[18] 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.

[19] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

[20] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[23] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297, 2020.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[25] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[26] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3024–3033.

[27] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra, "Self-supervised pretraining of 3d features on any point-cloud," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10 252–10 263.

[28] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding," in European conference on computer vision. Springer, 2020, pp. 574–591.

[29] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1318–1327.

[30] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by cnn semantic re-ranking," IEEE transactions on cybernetics, vol. 50, no. 7, pp. 3330–3342, 2019.

[31] X. Shen, Y. Xiao, S. X. Hu, O. Sbai, and M. Aubry, "Re-ranking for image retrieval and transductive few-shot classifi-

cation," Advances in Neural Information Processing Systems, vol. 34, pp. 25 932–25 943, 2021.
[32] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 740–749.
[33] R. Ren, Y. Qu, J. Liu, W. X. Zhao, Q. She, H. Wu, H. Wang, and J.-R. Wen, "Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking," arXiv preprint arXiv:2110.07367, 2021.
[34] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.
[35] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," The International Journal of Robotics Research, vol. 36, no. 1, pp. 3–15, 2017.