2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

CurricularVPR: Curricular Contrastive Loss for Visual Place Recognition

Dongshuo Zhang[1*] Nanhua Chen[2] Meiqing Wu[1] Siew-Kei Lam[1]

***Abstract*— Visual Place Recognition (VPR) techniques commonly utilize Contrastive Losses (CL) to train models that generate compact and discriminative global descriptors for images. These models often result in poor performance due to one of the following reasons during training: 1) loss functions that focus primarily on easier samples, 2) reliance on time-consuming hard sample mining methods to identify informative supervisory samples, which hinders effective learning from large-scale datasets. To enhance both learning efficiency and effectiveness, we propose a Curricular Contrastive Loss (CCL) and use graded similarity labels as a measure of sample difficulty. Inspired by human learning that begin with easier concepts and progressively tackle more challenging ones, our CCL dynamically emphasizes easier samples during the initial training stages to achieve rapid convergence. The learning gradually focuses on harder samples in later training stages to bolster robustness of the models under challenging conditions. Our proposed method has been extensively evaluated on popular datasets, and the results demonstrate its superior performance compared to the CL and Generalized CL functions.**

I. Introduction

Given a query image, Visual Place Recognition (VPR) aims to retrieve a reference image captured at the same location from a database. VPR plays a pivotal role in many robotics applications. Notably, modern Visual Simultaneous Localization and Mapping (VSLAM) systems rely on VPR for re-localization and loop closure detection. Existing VPR techniques face challenges in recognizing previously visited places that have undergone illumination, weather, and seasonal changes, as well as viewpoint variations.

Many VPR methods approach the task as a metric learning problem, training models with ranking losses like Triplet Ranking Loss (TRL) or Contrastive Loss (CL). The training datasets comprise of urban images from the internet and are assigned binary labels by utilizing GPS tags to determine image pairs that are in close proximity. However, the binary labels can be noisy, leading to stalls in local minima during training [2]. Moreover, TRL and CL necessitates training triplets or training pairs, each consisting of a query image, a positive image (located close to the query), and a negative image (situated far from the query). Due to the noisy binary labels, most methods [3] only select the easiest positive samples (those geographically closest to the query) to formulate positive training pairs, which leads to weak supervision for handling viewpoint changes and occlusions. This contributes to a lack of robustness against challenging viewpoint and environmental changes [4].

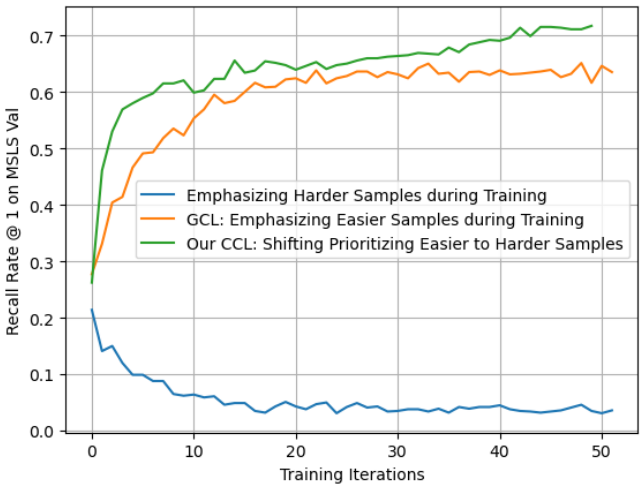


Fig. 1: **Recall Rate @ 1 (R@1) on MSLS Val [1] using Diverse Learning Strategies over Training Iterations.** Our Curricular Contrastive Loss (CCL) (green curve) facilitates faster model convergence and achieves better accuracy compared to the Generalized Contrastive Loss (GCL) [2] (orange curve). Focusing on harder samples from the start can lead to early model divergence (blue curve).

To overcome the challenges arising from noisy binary labels, several hard sample mining strategies have been introduced. Online hard sample mining techniques typically identify both hard positive and negative samples through cosine similarity checks [3], [4]. This enables models to learn from more challenging samples, which exhibit larger viewpoint changes and more severe occlusions, thereby enhancing their robustness. These hard samples are then periodically updated using the latest model. However, the periodic online mining of hard samples lead to extended training times and high consumption of storage space.

Recently, [2] introduced an offline re-labelling method. This method assigns similarity scores to image pairs based on the extent of their visual overlap, diverging from traditional binary labels. Image pairs exhibiting substantial Field-of-View (FoV) overlap receive higher similarity scores (i.e., easy positive pairs), while those with minimal FoV overlap are assigned lower similarity scores (i.e., hard positive pairs). Image pairs that share no visual content are assigned a

* Corresponding Author.
[1] Dongshuo Zhang, Meiqing Wu and Siew-Kei Lam are with College of Computing and Data Science, Nanyang Technological University, 50 Nanyang Ave, Singapore 639798. (email: DONGSHUO001@e.ntu.edu.sg, {meiqingwu, assklam}@ntu.edu.sg)
[2] Nanhua Chen is with School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China 100081. (email: chennanhua1@foxmail.com)

score of zero (i.e., negative samples). These similarity scores are integrated into a Generalized Contrastive Loss (GCL), explicitly emphasizing learning from image pairs with higher graded similarities. This approach leads to a bias towards easier samples in the learning.

However, both online hard sample mining and offline re-labeling techniques still present drawbacks, impeding effective learning of a robust VPR model across various challenging conditions. Online hard sample mining, involving the model learning from the hardest positive and negative samples from scratch, may encounter convergence issues (see the blue curve in Fig. 1). Conversely, offline graded similarity labels, which focus on easier samples [2], might not offer sufficient informative supervision (see the orange curve in Fig. 1).

To address the above-mentioned limitations, we propose a Curricular Contrastive Loss (CCL) that initially assigns higher learning weights to easier samples, and dynamically increases the training weights on harder samples as the training progresses. The samples difficulty is measured by their graded similarity [2], which reflects their viewpoint changes. We also propose to use an exponential function as the scheduler to regulate the pace and timing of shifting higher weights from easier to harder samples. As depicted by the green curve in Fig. 1, our proposed CCL achieves rapid convergence and enhanced performance. Our main contributions can be summarized as follows:

- We propose a Curricular Contrastive Loss (CCL) that initially assigns higher learning weights to easier samples, and dynamically increases the training weights on harder samples.
- We introduce an appreciation term t to regulate the rate and timing of the transition from focusing on easier to harder samples by adjusting its exponent α.
- We conduct extensive experiments and ablation studies on popular benchmarks under various challenging conditions, and show that our proposed method outperforms current state-of-the-art contrastive loss functions.

II. Related Work

A. Retrieval-based VPR

NetVLAD [3] first treats the VPR as a metric learning problem and adopts Triplet Ranking Loss (TRL) to fine-tune a novel, end-to-end trainable NetVLAD aggregation layer. However, as discussed earlier, a significant limitation of these methods is that they only select the image located closest to the query images as positive samples to formulate positive pairs for training. This leads to insufficient supervision of viewpoint changes and a lack of robustness when confronting challenging conditional changes [4].

Online hard sample mining has been extensively explored to overcome this limitation. For example, SFRS [4] employs self-enhanced fine-grained image-to-image and image-to-region similarities to identify harder samples. These samples exhibit larger viewpoint changes and occlusions relative to the query image, providing more informative supervision. By incorporating these harder samples, SFRS improves both robustness and learning effectiveness. Nonetheless, SFRS requires four phases of increasingly fine-grained training and entails complex calculations for image-to-image and image-to-region similarities, rendering the process time-intensive and potentially impractical for large-scale dataset training. Recent MixVPR [5] incorporates a new Feature-Mixer aggregation layer, along with the Multi-Similarity (MS) loss [6]. The MS miner measures the similarity between negative and positive pairs sharing the same query image, selecting and focusing on the hardest positive samples during training.

Offline re-relabelling method, GCL [2], leverages Field-of-View (FoV) overlap to grade the similarity of image pairs as labels. These graded similarities are subsequently integrated into a Generalized Contrastive Loss (GCL), allowing the learning process to incorporate information from more challenging samples. However, easier samples with high graded similarity receive more emphasis, while harder samples with low graded similarity are less emphasized by GCL during the training process. This result in the model lacking robustness in the face of challenging conditional changes.

Utilizing graded similarity labels as a measure of hardness, our Curricular Contrastive Loss (CCL) is designed to assign higher learning weights to easier samples initially. As the training progresses, the loss function gradually assigns higher weights to harder samples. CCL enables the model to converge quickly in the training, while ensuring robust performance under challenging scenarios.

B. Classification-based VPR

The recent introduction of the extra-large-scale SF-XL dataset by CosPlace [7] has garnered significant attention in the field. To efficiently train on this dataset, which is 10-100 times larger than previous ones, CosPlace reformulates VPR as a classification problem and split the samples in the dataset into different classes according to their UTM coordinates for training. Its successor, EigenPlaces [8], mines additional images from diverse viewpoints of a location to enhance robustness under viewpoint changes. However, this training paradigm necessitates an extra-large-scale dataset for model convergence, which require more time and larger GPU memory to achieve convergence compared to retrieval-based VPR methods.

C. Curriculum Learning

Curriculum learning [9] is a training strategy that sequences the training based on the difficulty level of samples. In curriculum learning, two components are crucial: 1) the difficulty measure [10], which assesses the difficulty of the samples, and 2) the scheduler [11], which organizes the sequence and timing of presenting these samples during training. In this paper, we employ graded similarity [2] to assess viewpoint changes as sample difficulty and propose a novel appreciation term t as the scheduler. This term regulate the pace and timing of the transition, shifting higher weights from easier to harder samples throughout the training process.

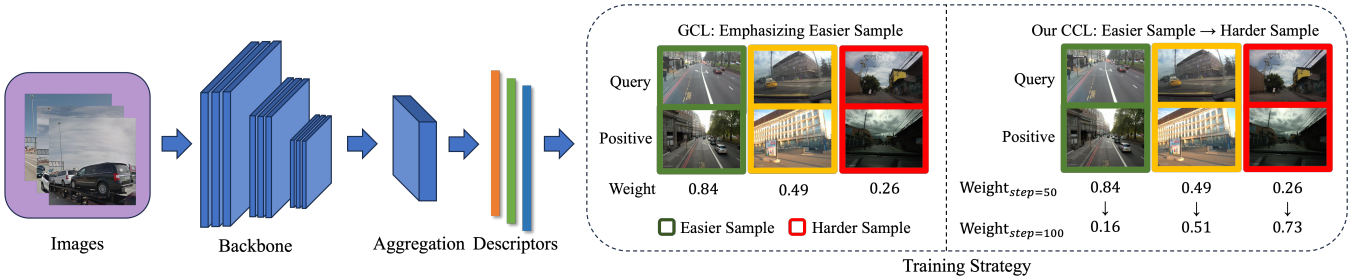


Fig. 2: **Deep Visual Place Recognition (VPR) Pipeline.** During training, GCL [2] focuses on easier samples (outlined in green) with moderate viewpoint changes. In contrast, our Curricular Contrastive Loss (CCL) dynamically shifts the learning emphasis from easier to harder samples (outlined in red) that exhibit larger viewpoint changes as the training progresses.

III. Proposed Method

In this section, we initially discuss prevalent Visual Place Recognition (VPR) pipelines and their objectives during training in Sec. III-A. This is followed by an explanation of how these goals are achieved through current loss functions in Sec. III-B. The existing loss functions struggle to learn from more challenging samples, resulting in sub-optimal robustness. To overcome this limitation, we introduce a novel Curricular Contrastive Loss (CCL) in Sec. III-C, which progressively shifts the learning emphasis from simpler to harder samples during training.

A. Deep VPR Pipeline

As shown in Fig. 2, popular deep VPR pipelines utilize CNN backbones, such as VGG16 [12] and ResNet50 [13], complemented by an aggregation layer. This layer extracts informative features across feature maps, creating a one-dimensional global descriptor for the entire image. Key aggregation layers include NetVLAD [3], GeM [14], and the recent MixVPR [5]. The training objective is to reduce the descriptor distance for images taken within 25 meters of the same place (positive samples), and to increase the distance for images from different locations (negative samples).

B. Revisiting Existing Loss Functions

The primary objective for training VPR models is to separate embeddings for images captured at different locations, while bringing closer together the embeddings for images captured at the same place in the feature space. Suppose there is an image pair (i, j) and their embeddings is (x_i, x_j), the distance $d(x_i, x_j)$ between the two embeddings in the latent space is defined as follows,

$$d(x_i, x_j) = ||f(x_i) - f(x_j)||^2. \tag{1}$$

To achieve the goal, existing methods commonly approach the task as a metric learning problem, utilizing Contrastive Loss (CL) [15]. The definition of CL is outlined as follows,

$$\begin{aligned}\mathcal{L}_{CL}(x_i, x_j) =& \beta_{i,j} \cdot \frac{1}{2} d(x_i, x_j)^2 \\ &+ (1 - \beta_{i,j}) \cdot \frac{1}{2} \max\big(\tau - d(x_i, x_j), 0\big)^2,\end{aligned} \tag{2}$$

where $\beta_{i,j}$ is the binary ground truth label for the image pair (i, j) and τ is the pre-defined margin, which is a hyper-parameter that defines the boundary margin between positive and negative pairs.

In the conventional training setup for the VPR tasks, the binary ground truth label $\beta_{i,j}$ is indicative. Specifically, $\beta_{i,j} = 1$ signifies that the pair of images is captured at the same place within a 25-meter range in geometric space. Conversely, $\beta_{i,j} = 0$ denotes that the two images are not captured at the same place within 25 meters. However, binary labels often result in stalling in local minima during training.

To address this limitation, recent Generalized Contrastive Loss (GCL) [2] was introduced to change the binary ground truth $\beta_{i,j}$ to a continuous label $\gamma_{i,j}$ in the range $[0, 1]$ based on the Field-of-View (FoV) overlap of the image pair. This approach allows VPR models to learn from samples with larger viewpoint changes, thereby enhancing their robustness in challenging scenarios. It is important to note that these continuous labels $\gamma_{i,j}$ also serve as learning weights during the backpropagation process, as depicted in Eqn. 4. In the GCL, labels for image pairs with larger visual overlaps (i.e., easier samples) are close to 1, leading to higher learning weights during training. Conversely, labels for image pairs with smaller visual overlaps (i.e., harder samples) are close to 0, resulting in lower learning weights during training.

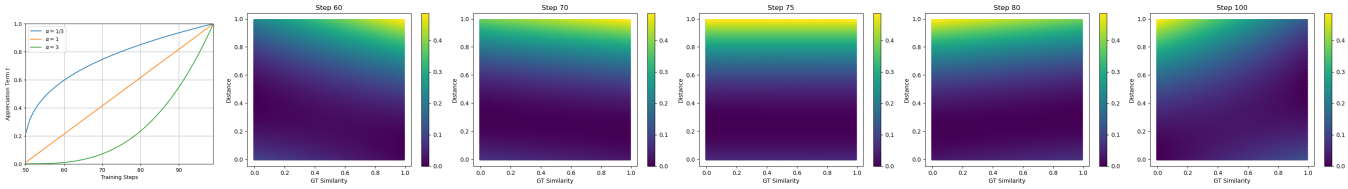
$$\begin{aligned}\mathcal{L}_{GCL}(x_i, x_j) =& \gamma_{i,j} \cdot \frac{1}{2} d(x_i, x_j)^2 \\ &+ (1 - \gamma_{i,j}) \cdot \frac{1}{2} \max\big(\tau - d(x_i, x_j), 0\big)^2,\end{aligned} \tag{3}$$

$$\frac{\partial \mathcal{L}_{GCL}}{\partial d(x_i, x_j)} = \begin{cases} d(x_i, x_j) + \tau(\gamma_{i,j} - 1), & \text{if } d(x_i, x_j) < \tau \\ d(x_i, x_j) \cdot \gamma_{i,j}. & \text{if } d(x_i, x_j) \geq \tau \end{cases} \tag{4}$$

As such, while GCL, with its utilization of continuous labels, exhibits better robustness compared to CL, which employ binary labels, we found that it places more emphasis on easier samples during training. However, this emphasis on easier samples can result in the model being less adept at handling more challenging cases.

C. Proposed Curricular Contrastive Loss (CCL)

Taking inspiration from how humans typically learn new knowledge, starting with easier concepts before gradually delving into more challenging ones, our proposed CCL



(a) Appreciation term (b) $\alpha = 1, t = 0.2$ (c) $\alpha = 1, t = 0.4$ (d) $\alpha = 1, t = 0.5$ (e) $\alpha = 1, t = 0.6$ (f) $\alpha = 1, t = 1$

Fig. 3: **Visualization of Appreciation Term t *w.r.t.* Different Exponents α and Our Curricular Contrastive Loss (CCL).** The closer the color is to yellow, the higher the loss and weights during backpropagation. Initially, when t is close to 0, the model focuses on learning easier samples. It then progressively shifts its attention to more challenging samples as t reaches 1. The parameter α controls the pace and timing of this learning emphasis from easier to harder samples.

adopts a similar strategy to tackle the aforementioned issues. The formula for CCL is defined as follows,

$$\begin{aligned}\mathcal{L}_{CCL}(x_i, x_j) =& \delta_{i,j} \cdot \frac{1}{2} d(x_i, x_j)^2 \\ &+ (1 - \delta_{i,j}) \cdot \frac{1}{2} \max\big(\tau - d(x_i, x_j), 0\big)^2,\end{aligned} \tag{5}$$

where $\delta_{i,j}$ gradually changes throughout the training process and is defined as follows,

$$\delta_{i,j} = \begin{cases} \gamma_{i,j} & \text{if } s_{\text{cur}} < \frac{s_{\text{total}}}{2} \\ t + (1 - 2t) \cdot \gamma_{i,j}, & \text{if } s_{\text{cur}} \geq \frac{s_{\text{total}}}{2} \end{cases} \tag{6}$$

where s_{cur} represents the current training step number and s_{total} denotes the total number of training steps.

In the first half of the training steps ($s_{\text{cur}} < \frac{s_{\text{total}}}{2}$), $\delta_{i,j} = \gamma_{i,j}$, emphasizing easier samples for rapid convergence. In the remaining half of the training steps ($s_{\text{cur}} \geq \frac{s_{\text{total}}}{2}$), we shift our focus from easier to harder samples by the appreciation term t, which is an exponential function, defined as follows,

$$t = \left(\frac{2 s_{\text{cur}}}{s_{\text{total}}} - 1 \right)^{\alpha}, \tag{7}$$

where α is an exponent that regulates the rate of the transition from prioritizing easier to harder samples.

As depicted in Fig. 3a, the appreciation term t initiates at 0 and progressively reaches 1 towards the end of training. At the beginning of the last half of the training steps ($s_{\text{cur}} = \frac{s_{\text{total}}}{2}$), our curricular loss mirrors that of the GCL, prioritizing easier samples, as shown in Eqn. 4. When training approaches completion ($s_{\text{cur}} = s_{\text{total}}$), our curricular loss exhibits an opposite characteristic compared to the GCL. Specifically, in Eqn. 8, smaller values of $\gamma_{i,j}$ (i.e., harder training pairs) in the derivative *w.r.t.* $d(x_i, x_j)$ receive larger weights during backpropagation.

$$\frac{\partial \mathcal{L}_{CCL}(x_i, x_j)}{\partial d(x_i, x_j)} = \begin{cases} d(x_i, x_j) - \tau \cdot \gamma_{i,j}, & \text{if } d(x_i, x_j) < \tau \\ d(x_i, x_j) \cdot (1 - \gamma_{i,j}), & \text{if } d(x_i, x_j) \geq \tau \end{cases} \tag{8}$$

If α falls within the range $[0, 1]$, the training emphasis shifts swiftly from easier samples to harder samples in the initial stages of the second half of the training, gradually stabilizing at 1. The smaller the value of α, the faster the rate of shifting at the beginning of the second half of the training process. For $\alpha = 1$, the shifting rate changes linearly. When $\alpha > 1$, the shifting rate changes slowly and steadily at the beginning, and becomes steeper in the last 20 steps. The larger the value of α, the faster the rate of shifting towards the end of the training process.

Fig. 3b-3f illustrate how α affects the loss during training. On the x-axis of each graph, we have the Field-of-View (FoV) overlap similarity, where a higher value (right of the x-axis) signifies less viewpoint changes and thus easier training samples, while a lower value (left of the x-axis) indicates harder samples. In these visualizations, the closer the color is to yellow, the higher the loss and weights during backpropagation. We can observe a shift in focus from easier samples on the right to harder samples on the left in each row, with varying shifting rates from prioritizing easier to harder samples for different α values. The detailed ablation study of α can be found in Sec. IV-C.2.

IV. Experiments and Results

A. Implementation Details

1) Pipeline and Baselines: Our training frameworks are based on the publicly available GCL framework [2]. Our proposed Curricular Contrastive Loss (CCL) is applied to train this setup. To comprehensively evaluate our method, we incorporate various SOTA aggregation layers such as NetVLAD [3], GeM [14], and MixVPR [5] into our CCL-enhanced pipeline. We benchmark our approach against other established loss functions, including Contrastive Loss (CL), Generalized Contrastive Loss (GCL) [2], with all models trained on the MSLS dataset [1] using the same setup for a consistent comparison.

2) Graded Similarity Label γ: It is calculated by measuring the overlap of the FoV of cameras associated with images. For MSLS dataset [1], FoV overlap is computed as the intersection-over-union (IoU) of camera FoVs. The graded similarity labels for MSLS are provided by GCL [2].

B. Place Recognition Performance and Training Efficiency

We compared our methods against the latest published SOTA approaches, including two classification-based techniques, CosPlace [7] and EigenPlaces [8], along with five retrieval-based methods [3], [4], [18], [2], [5]. The detailed results are presented in Table I. Our method outperforms other retrieval-based VPR methods on the Pitts30k and Tokyo 24/7 test sets, achieving the best performance. Remarkably, compared to our baseline method, GCL [2] that

TABLE I: **Comparison of Place Recognition Performance with SOTA Methods on Popular Datasets. Bold** text signifies the highest accuracy across all VPR methods, while underline denotes the highest accuracy among models trained with the same setup as GCL [2].

Method	Venue	Backbone	Training Dataset	Loss	Pitts250k [16]			Pitts30k [16]			Tokyo 24/7 [17]			MSLS Val [1]		
					R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CosPlace †	CVPR 22'	ResNet50	SF-XL	CosFace	92.3	97.4	98.4	90.9	95.7	96.7	87.3	94.0	95.6	87.4	94.1	94.9
EigenPlaces †	ICCV 23'	ResNet50	SF-XL	CosFace	94.1	97.9	98.7	**92.5**	**96.8**	**97.6**	**93.0**	**96.2**	**97.5**	**89.1**	**93.8**	**95.0**
SFRS ‡	ECCV 20'	VGG16	Pittsburgh	CL	90.7	96.4	97.6	-	-	-	85.4	91.1	93.3	-	-	-
MixVPR ‡	WACV 23'	ResNet50	GSV-Cities	MS	**94.6**	**98.3**	**99.0**	-	-	-	-	-	-	88.0	92.7	94.6
StructVPR Global ‡	CVPR 23'	MobileNetv2	Pittsburgh	TRL	-	-	-	85.1	92.3	94.3	-	-	-	83.0	91.0	92.6
NetVLAD §	CVPR 16'	VGG16	MSLS	TRL	-	-	-	68.6	84.7	88.9	34.7	47.6	57.1	70.1	80.8	84.9
ResNet50-GeM-GCL §	CVPR 23'	ResNet50	MSLS	GCL	-	-	-	79.9	90.0	92.8	58.7	71.1	76.8	74.6	84.7	88.1
ResNet152-GeM-GCL §	CVPR 23'	ResNet152	MSLS	GCL	-	-	-	80.7	91.5	93.9	69.5	81.0	85.1	79.5	88.1	90.1
ResNeXt-GeM-GCL §	CVPR 23'	ResNeXt	MSLS	GCL	-	-	-	79.2	90.4	93.2	58.1	74.3	78.1	80.9	90.7	92.6
NetVLAD-GCL ¶		VGG16	MSLS	GCL	54.1	74.2	82.0	53.6	75.6	82.2	28.7	41.6	55.0	60.4	75.8	80.1
MixVPR-GCL ¶		ResNet50	MSLS	GCL	71.3	85.7	90.6	72.5	84.9	90.7	70.3	79.2	84.6	73.9	84.2	88.1
NetVLAD-CCL (Ours) ¶		VGG16	MSLS	**CCL**	71.5	89.3	91.2	69.1	84.0	89.8	49.3	57.1	72.6	64.0	76.3	79.8
MixVPR-CCL (Ours) ¶		ResNet50	MSLS	**CCL**	78.9	90.1	93.1	76.7	89.2	94.6	75.3	88.7	90.5	75.4	85.7	88.8
ResNet50-GeM-CCL (Ours) ¶		ResNet50	MSLS	**CCL**	84.1	93.0	95.0	83.1	92.0	94.4	86.9	91.3	93.6	79.3	87.7	90.5
ResNet152-GeM-CCL (Ours) ¶		ResNet152	MSLS	**CCL**	86.5	94.6	96.5	85.6	93.6	95.5	87.2	92.9	95.0	82.7	89.7	91.1
ResNeXt-GeM-CCL (Ours) ¶		ResNeXt	MSLS	**CCL**	86.6	94.4	96.1	85.8	93.4	95.4	87.8	93.6	95.1	85.3	92.6	93.9

† Results are cited from EigenPlaces [8].
‡ Results are cited from their original papers [4], [5], [18].
§ Results are cited from GCL [2], presenting the highest accuracy reported.
¶ All methods were trained and evaluated employing the same setup as GCL [2], with results reported **using PCA whitening**.

TABLE II: **Ablation Study on Backbone. Bold** text signifies the highest accuracy among methods within each group, characterized by the same backbone, aggregation layer, and descriptor dimension.

Method	Loss	Dim.	PCA_w	Pitts30k [16]			Pitts250k [16]			SPEDTEST [19]			Tokyo 24/7 [17]			MSLS Val [1]		
				R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VGG16-GeM	CL	512	No	51.25	71.93	79.69	44.43	63.06	70.07	47.61	62.77	68.70	28.25	42.54	50.79	41.47	55.13	60.89
VGG16-GeM	GCL	512	No	61.55	79.94	86.00	53.31	72.42	79.20	**53.71**	68.20	74.14	**36.19**	**56.51**	**66.35**	58.25	71.74	76.23
VGG16-GeM	**CCL (Ours)**	512	No	**63.09**	**81.51**	**86.97**	**55.30**	**74.14**	**80.40**	52.72	**69.52**	**74.46**	33.33	51.43	60.63	**66.35**	**78.92**	**82.16**
ResNet50-GeM	CL	2048	No	61.53	80.08	86.90	54.82	74.15	80.75	**58.81**	**76.61**	**80.89**	32.38	51.11	58.41	44.61	59.55	65.53
ResNet50-GeM	GCL	2048	No	72.33	87.25	91.39	68.26	84.54	89.19	56.34	69.69	75.29	40.00	58.10	68.89	58.49	71.77	76.36
ResNet50-GeM	**CCL (Ours)**	2048	No	**74.78**	**88.32**	**92.02**	**69.86**	**84.86**	**88.80**	58.65	72.65	78.91	**41.90**	**63.49**	**74.60**	**67.97**	**80.81**	**84.05**
ResNet152-GeM	CL	2048	No	66.42	83.83	89.44	60.71	78.91	85.01	53.71	71.99	77.92	36.19	57.46	66.35	51.19	66.19	71.61
ResNet152-GeM	GCL	2048	No	72.67	87.90	91.51	67.95	84.92	89.82	47.61	63.76	70.18	42.54	58.41	**68.25**	62.41	75.69	79.64
ResNet152-GeM	**CCL (Ours)**	2048	No	**75.21**	**89.47**	**92.65**	**70.71**	**86.82**	**90.80**	**59.47**	**73.81**	**79.41**	**43.17**	**60.32**	**68.25**	**73.65**	**84.59**	**88.11**
ResNext-GeM	CL	2048	No	55.94	77.48	84.95	50.18	70.69	78.89	47.61	65.07	72.32	33.65	53.02	58.10	53.99	68.87	73.69
ResNext-GeM	GCL	2048	No	64.03	81.32	86.72	58.54	76.17	81.71	56.01	72.65	79.74	39.68	58.41	65.40	68.77	80.95	84.72
ResNext-GeM	**CCL (Ours)**	2048	No	**76.47**	**88.97**	**92.40**	**73.44**	**86.80**	**91.03**	**67.55**	**78.91**	**81.88**	**43.81**	**63.49**	**75.24**	**81.62**	**89.73**	**91.49**

* All methods were trained and evaluated using the same setup as GCL [2], with all results reported **w/o the use of PCA whitening**.

shares the same setup as ours including the backbone, aggregation layer, and training on the MSLS dataset, our method shows improvements of 5.1%, 18.3%, and 4.4% on the Pitts30k, Tokyo 24/7, and MSLS Val datasets, respectively. More detailed comparisons between our CCL and both CL and GCL are reported in Sec. IV-C.1.

TABLE III: **Comparison of Training Efficiency and Resource Consumption with SOTA Methods.**

Method	Backbone	Aggregation	Batch Size	Training Time t $(min/iter)$	GPU Memory (MB)
CosPlace	ResNet50	GeM	32	34	7175
EigenPlaces	ResNet50	GeM	32	14	3743
MixVPR	ResNet50	MixVPR	32	11.67	21661
GCL	ResNet50	GeM	32	**2.17**	**1885**
CCL (Ours)	ResNet50	GeM	32	**2.17**	**1885**

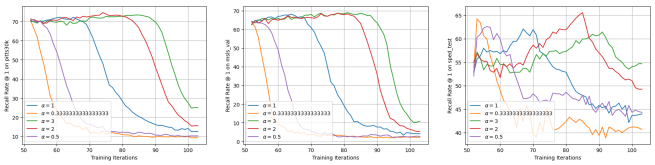
* All methods tested on the same NVIDIA GeForce RTX 4090 platform.

We also observe that CosPlace, EigenPlaces and MixVPR outperform our method. This is attributed to the difference in training datasets used. In particular, CosPlace and EigenPlaces utilize the extra-large-scale dataset SF-XL, which comprises of 41.2 million panorama images, for model training. As a result, these classification-based methods require longer training time and higher GPU memory consumption to reach model convergence, as detailed in Table III. In contrast, our approach employs the MSLS training set, which contains only 1.6 million frontal-view images, approximately 25 times smaller than SF-XL. This difference leads to more efficient training and lower GPU usage in our method. We achieve training times that are 15.6, 6.5, and 5.4 times faster, along with 3.8, 2, and 11.5 times lesser GPU memory consumption compared to CosPlace, EigenPlaces, and MixVPR, respectively. We also trained MixVPR on the same MSLS dataset as our method. Its performance significantly degraded, with an average decrease of 31.8% for R@1. As such, our method achieves the best trade-off among all the methods in terms of training efficiency and place recognition performance.

C. Ablation Studies

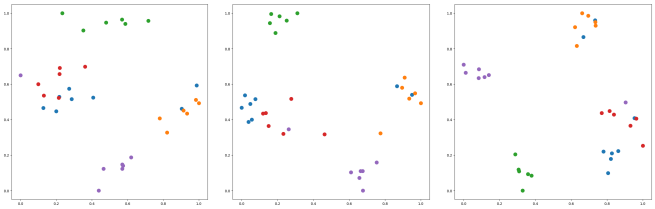
1) Study on Backbone: We integrate our proposed CCL into various backbones to compare its performance with both CL and GCL [2]. Detailed results are presented in Table II. From the table, it is observable that our CCL significantly outperforms both CL and GCL across different backbones on various datasets. Therefore, it is evident that adopting CCL leads to improved performance.

2) Study on Exponent α: We selected five values for the exponent α: $\frac{1}{3}$, $\frac{1}{2}$, 1, 2, and 3, respectively, and trained the model using our CCL with each of these values. Since the initial half of the steps maintains the same training strategy as GCL [2], we have plotted the evaluation metrics only for the latter half of the steps. As illustrated in Fig. 4, the model achieved the best performance when trained using $\alpha = 2$.



(a) R@1 on the Pitts30k Dataset (b) R@1 on the MSLS Val Dataset (c) R@1 on the SPEDTEST Dataset

Fig. 4: **Study on Exponent α.** The model achieved the best performance when trained using $\alpha = 2$.



(a) CL (b) GCL (c) Our CCL

Fig. 5: **T-SNE [21] Visualization for Images from 5 Places.** Data points of the same color belong to the same class. Closer distances between data points indicate proximity in the feature space. The model trained with our CCL reduces intra-class distances, signifying that the samples within the same class are closer in the feature space, compared to CL and GCL [2].

D. Qualitative Visualization

We randomly selected 5 places from GSV-Cities [20] and visualized latent distance of descriptors of images from these 5 places using T-SNE visualization tool [21]. The inter-class distance of the descriptors trained using our CCL is larger compared to those trained using CL and GCL. Meanwhile, the intra-class distance of the descriptors trained with our CCL is smaller than those trained using CL and GCL. As shown in Fig. 5c, we can easily distinguish 5 clusters, which is not evident in Fig. 5a and 5b. The intra-class distances for all classes are short, although there are a few harder samples present. This indicates that our CCL enables the model to learn more discriminative descriptors.

V. Conclusions

We proposed a novel Curricular Contrastive Loss (CCL) for VPR tasks. By leveraging graded similarity as a difficulty measure and introducing an appreciation term t as the scheduler, our CCL progressively integrates more challenging samples. Our study demonstrated that harder samples are invaluable for their informative supervisory capacity. Furthermore, as the VPR research field enters the "big data" era, the efficiency, adaptability, and performance improvements offered by CCL become a promising approach for developing more robust and accurate VPR systems.

VI. Acknowledgement

This work was supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1, under Grant RG78/21; the computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2626–2635.

[2] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 487–23 496.

[3] 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.

[4] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 369–386.

[5] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2998–3007.

[6] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5022–5030.

[7] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[8] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 080–11 090.

[9] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[10] S. Basu and J. Christensen, "Teaching classification boundaries to humans," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 109–115.

[11] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[15] F. Radenovic, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *European conference on computer vision*. Springer, 2016, pp. 3–20.

[16] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.

[17] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1808–1817.

[18] Y. Shen, S. Zhou, J. Fu, R. Wang, S. Chen, and N. Zheng, "Structvpr: Distill structural knowledge with weighting samples for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 11 217–11 226.

[19] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4015–4022, 2018.

[20] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, pp. 194–203, 2022.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.