2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Regressing Transformers for Data-efficient Visual Place Recognition

María Leyva-Vallina[1], Nicola Strisciuglio[2], and Nicolai Petkov[1]

***Abstract*— Visual place recognition is a critical task in computer vision, especially for localization and navigation systems. Existing methods often rely on contrastive learning: image descriptors are trained to have small distance for similar images and larger distance for dissimilar ones in a latent space. However, this approach struggles to ensure accurate distance-based image similarity representation, particularly when training with binary pairwise labels, and complex re-ranking strategies are required. This work introduces a fresh perspective by framing place recognition as a regression problem, using camera field-of-view overlap as similarity ground truth for learning. By optimizing image descriptors to align directly with graded similarity labels, this approach enhances ranking capabilities without expensive re-ranking, offering data-efficient training and strong generalization across several benchmark datasets.**

I. Introduction

Visual place recognition (VPR) is the task of identifying a previously visited location by analyzing visual information. It is an important component of visual navigation [1], [2] for autonomous driving and robotics [3]. The task of visual place recognition is formulated as an image retrieval problem. Given a database of map images and a query image, the objective is to retrieve the most similar images to the query image from the database. The query image is recognized based on a distance metric between its descriptor and those of the database images [4]. A key challenge of VPR is the ability to match images from different viewpoints and under varying appearance conditions. Thus, robust image descriptors have to be computed that take into account such issues [5].

Top-performing methods were mostly based on convolutional backbones, trained with image pairs [6] or triplets [7], [8]. In recent works, transformers replaced the convolutional backbones [9], [10], [11]. Training images in benchmark datasets are labelled so that they either depict the same place or not, from different points of view (distance and orientation of the camera) or at different times of the day or year. In this way, a certain level of robustness is expected to be embedded in the descriptors. However, the binary labels may introduce noise in the training, as images of the same place with small visual overlap have the same impact on training as images that share more visual cues. To optimize the training process, and avoid that it stalls in local minima, most methods include hard-pair mining procedures during training [7], [12]. These strategies are time- and memory-wise expensive, but were meant necessary to counteract the negative effect of noisy labels on the training process. The representation power of the learned image descriptors is such that, in many cases and on large datasets (e.g. the Mapillary Street level Sequences [13]), retrieved image lists are post-processed by heavy re-ranking strategies [9], [14], [15], [16] to achieve good results.

A step towards reducing the impact of label noise in training VPR models was done in [17], where graded image similarity labels were computed using a proxy based on approximating the field-of-view overlap of cameras. This was inspired by the frustum overlap included in a camera re-localization pipeline in [18]. We display an example of graded image similarity in Figure 1: a reference image with four image matches taken at different distances. The annotated image similarity ψ decreases gradually as the distance increases, but there is no clear threshold between *similar* and *dissimilar*. In [17], a novel Generalized Contrastive Loss (GLC) was also proposed to embed the graded similarity into the training process. However, the graded ground truth only weighs the descriptor distance in the latent space during training. This influences the gradient updates so that the descriptors of more similar images are pushed more closely to each other in the latent space. It does not ensure that the descriptor distance is an actual metric of how similar two images are.

In this paper, we take a different direction w.r.t. existing approaches, shifting from the contrasting learning approach and recasting training VPR descriptors as a regression task, such that their distance in the latent space is a direct measure of the field-of-view overlap of the cameras in the real world, thus of the image similarity. We revamp a largely disregarded and straightforward regression-based approach, only used in a re-localization pipeline [18], to learn VPR descriptors, and make the following contributions:

- we demonstrate that regression is a powerful solution to train descriptors for VPR (especially for transformers), achieving performance comparable or superior to SoTA methods, without need of pair-mining or re-ranking strategies, while saving time and memory;
- we achieve high data-efficiency, i.e. models trained via regression require only a few thousands training iterations on a small set of image pairs to converge and achieve high retrieval results, and good generalization performance, differently from contrastive approaches that require multiple epochs on the training data.

The direct link between descriptor distance and image similarity imposed by our formulation results in retrieval (ranking) performance higher than or comparable to that of more sophisticated methods that use hard-pair mining for training or re-ranking strategies. The distance in the latent space of descriptors learned by regression is indeed better

[1]María Leyva-Vallina and Nicolai Petkov are with the Bernoulli Institute of the University of Groningen, the Netherlands m.leyva.vallina@rug.nl

[2]Nicola Strisciuglio is with the University of Twente, the Netherlands n.strisciuglio@utwente.nl

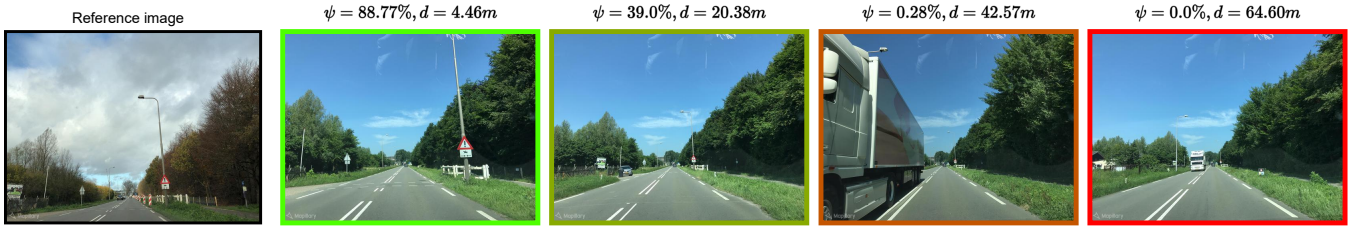


Fig. 1. A reference image (leftmost) and four match images, taken at different distances. The larger the distance w.r.t the reference image, the lower the annotated similarity ground truth ψ, and the smaller the amount of shared visual features.

representative of image similarity than that of descriptors learned via contrastive methods;

II. Previous work

Visual place recognition. Methods for visual place recognition are usually designed to solve an image retrieval task: given a query image, similar images are to be retrieved from a database (map). Traditional VPR methods build on extracting local features and combining them into holistic image descriptors, such as Bag-of-Words [19], [20], Fisher vectors [21] and VLAD [22]. Also, information about image sequences was used to improve performance [1].

Current state-of-the-art approaches rely on descriptors learned via deep learning and convolutional networks (CNNs) [23], [24]. Initially, pre-trained networks were shown to extract effective descriptors [25], [8], [26]. However, training them using image pairs or triplets (depicting the same place or not) in an end-to-end fashion was demonstrated to boost the performance results largely [6], [27]. The combination of a CNN backbone with a trainable VLAD layer, in NetVLAD, contributed to achieving state-of-the-art performance on several benchmarks [7]. Recently, Transformers have been deployed as backbones for VPR, to learn more powerful and effective image descriptors. TransVPR [9] and CosPlace [10] trained a transformer backbone respectively using a triplet learning architecture and a classification-based approach. R2Former [11] combined learning transformer-based descriptors and re-ranking end-to-end. These methods achieved high performance and generalization, attributable to the attention mechanism intrinsic to the transformer architecture and to the large datasets used for training (e.g. MSLS and SF-XL).

Metric Learning for VPR. Image descriptors for VPR are learned so that their distance is small in a latent space for images of the same place, and large for images of different places [6], [28]. This approach is referred to as metric learning. Image pairs or triplets, labelled as depicting the same place or not, are used to train neural networks to compute descriptors in a metric space. Most popular and top-performing approaches are based on triplet network learning, e.g. NetVLAD [7] (with an architecture composed of a VGG backbone and a trainable VLAD layer). These methods use image triplets with a reference image (anchor), a positive image (depicting the same place of the anchor), and a negative image (depicting a different place). A triplet loss is then optimized to reduce the distance between the anchor and positive image descriptor, and maximize that between the anchor and negative image descriptor. Several improvements of NetVLAD were proposed, such as the fusion of patch- and global-level VLAD descriptors in PatchNetVLAD [15], or training with a stochastic attraction-repulsion triplet loss function in NetVLAD-SARE [12]. PointNetVLAD [29] combined 3D information from PointNet [30] into NetVLAD. Loss functions with weighting schemes for image pairs based on GPS distance were also explored to train NetVLAD [31].

These methods are trained using hard-pair mining to compose training batches, which requires costly computations so that the noise in data and labels is counteracted. However, the learned descriptors show some limited performance when used on larger diverse datasets. Thus, two-stage approaches are deployed, with the top-k retrieved candidates being re-ranked using geometric verification, based on keypoint matching, either from convolutional features [15] or attention maps [9]. Our work differs from existing metric learning approaches for VPR as we do not deploy contrastive learning, and demonstrate that a regression-based learning approach allows training (larger) models efficiently, resulting in descriptors with higher representation capabilities, and with no need for hard mining or re-ranking.

Limitations of contrastive learning for VPR. The common assumption to train VPR methods is that two images either depict the same place or not, in a binary way. Ground truth labels in benchmark datasets (e.g. MSLS [13] or Pittsburgh30k [32]) are available for image pairs, indicating whether they show the same place or not. This approach does not take into account that images of the same place may share more or less visual cues, depending for example on perspective changes in the camera position. Hence, image pairs depicting the same place with large or small visual cue share are weighted in the same manner, introducing noise during training [17]. Computation-heavy hard-pair mining is used to select hard image pairs and triplets to avoid training stalling [7]. These methods are data-hungry, due to the necessity of overcoming the label noise problems.

In [17], a first approach to use graded instead of binary similarity labels for image pairs was explored in combination with a Generalized Contrastive Loss (GCL) function. A relevant outcome was that hard-pair mining deemed not necessary to ensure training convergence. Larger backbones were trained efficiently, while achieving higher results than

existing methods. However, the contrastive learning paradigm still introduces an artificial binarization of the problem by dividing the loss function into a positive and a negative term.

In this work, we further exploit the concept of graded image similarity [17] and train VPR models by a regression task. In this way, the descriptor distance in the latent space has a direct relation with the similarity degree of images. The outcome is a more robust distance-based ranking of retrieved images based on similarity to the query image.

III. Methodology

A. Architecture, optimization and training batches

We use a siamese architecture to train the encoder of a (hybrid) visual transformer [33] or a convolutional backbone with a GeM pool layer [6]. We optimize a Mean Squared Error loss function, which is formulated as:

$$\mathcal{L}(x_i, x_j, \psi_{i,j}) = \|d(\theta(x_i), \theta(x_j)) - (1 - \psi_{i,j})\|_2^2 ,$$

where x_i and x_j are the input images, which have ground truth similarity $\psi_{i,j}$, with $\theta(x_i)$ and $\theta(x_j)$ their respective descriptors computed with a model $\theta(\cdot)$. We denote $d(\cdot, \cdot)$ as a distance function in the descriptor space. We consider the Euclidean distance, which led to better results than the Cosine distance. We L2-normalize the descriptors.

We do not perform hard mining to select samples to compose the training batches. We rely on the graded similarity ground truth available with the images to compose the training batches so that 50% of the pairs have $\psi \in (0.5, 1]$, 25% with $\psi \in (0, 0.5]$ and the remaining 25% has $\psi = 0$ [17]. This ensures that a batch has pairs with approximately uniformly distributed ground truth similarity in the interval $[0, 1]$.

B. Image search and retrieval

In visual place recognition, a set of query images Y taken from unknown positions are localized in an environment by comparing them with similar map images retrieved from a set X, for which the camera pose is known. We compute descriptors of the map images by $\theta(x),\ \forall x \in X$, and of the query images $\theta(y),\ \forall y \in Y$ using the encoder models that we train. For a given query image descriptor $\theta(y)$, image retrieval is performed by an exhaustive nearest neighbour search among the descriptors of the map set $\theta(x)\ \ \forall x \in X$. We retrieve a set of map images, which are ranked according to the closest distance of their descriptor in the latent space with respect to the query image descriptor. We point out that we do not address the pose estimation and camera localization tasks. Our focus is on training encoder models for visual place recognition to compute effective image descriptors and enable image retrieval to provide high-quality ranking results. The improved ranked retrieved images have the potential of boosting the performance of camera localization pipelines.

C. Training data

We train on the Mapillary Street Level Sequences (MSLS) dataset, a large-scale place recognition dataset that contains images taken in 30 cities across six continents [13]. It includes challenging variations of camera viewpoint, season, time and illumination. The training set contains over 500k query images and 900k map images from 22 cities. The validation set consists of 19k map images and 11k query images from two cities, and the test set has 39k map images and 27k query images from six different cities. For training, we use the FoV overlap introduced in [17] as ground truth image similarity.

D. Evaluation data

We evaluate the trained models on the following datasets.
MSLS. We use the MSLS validation set and the MSLS test set. For the latter, since the ground truth is not publicly available, we submit the predictions of our methods to the official MSLS evaluation server. Following the protocol in [13], two images are referenced as similar for evaluation if they are taken by cameras located within $25m$, and with less than $40°$ of viewpoint variation.
Pittsburgh. It contains images recorded via Google Street View in the city of Pittsburgh, Pennsylvania, over a span of several years [20]. We use the test set of Pittsburgh30k, which is a widely used benchmark [7]. It contains 10k map images and 7k query images, and we use it to evaluate the generalization and out-of-distribution performance of our models trained on the MSLS dataset.
Tokyo 24/7. The dataset consists of 315 query images and 76k map images taken in the city of Tokyo, Japan. The dataset images show large variations of illumination, as they are taken during day and night [32]. It is a commonly used dataset for benchmarking, explicitly designed to test robustness to changes of illumination.

E. Implementation details

We trained our models for one epoch (520k image pairs) using a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate equal to 0.1. For the training of models with the Generalized Contrastive Loss, we decrease the learning rate by a factor of 10^{-1} every 250k iterations. For the optimization with the Mean Squared Error loss, we keep the learning rate constant at 0.1 for the whole training. All our experiments were carried out using PyTorch. The models and training code will be publicly available.

IV. Experiments and results

A. Results

Comparison with state-of-the-art. We compared the results of our models with existing approaches, considering both methods that perform only retrieval (i.e. NetVLAD [7], TransVPR w/o re-ranking [9], GCL [17]) and methods that apply re-ranking strategies to improve the list of retrieved images (i.e. PatchNetVLAD [15], SP-SuperGlue [14], DELG [16] and TransVPR with attention-based re-ranking [9]). We compute the performance in terms of recall rate at k (R@k), typically used in VPR. All considered models are trained on the MSLS dataset [13]. We test their generalization abilities on the Pittsburgh30k [7] and the Tokyo24/7 [7] datasets. Although we do not do re-ranking, our results compare in many cases favourably or on par with re-ranking approaches. This highlights the quality and effectiveness of the descriptors

TABLE I

COMPARISON TO STATE-OF-THE-ART METHODS TRAINED ON THE MSLS DATASET. THE MARK ⋆ DENOTES METHODS THAT PERFORM RE-RANKING. TL INDICATES THE USE OF A TRIPLET LOSS. WE UNDERLINE THE BEST RESULTS BY OUR METHODS AND SHOW THE BEST RESULTS OVERALL IN BOLD.

			MSLS-Val			MSLS-Test			Pitts30k			Tokyo24/7		
Encoder	**PCA$_w$**	**Dim**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
NetVLAD (TL)	N	32768	44.6	61.1	66.4	28.8	44.0	50.7	40.4	64.5	74.2	11.4	24.1	31.4
NetVLAD (TL)	Y	4096	70.1	80.8	84.9	45.1	58.8	63.7	68.6	84.7	88.9	34.0	47.6	57.1
TransVPR (TL) [9]	-	-	70.8	85.1	86.9	48	67.1	73.6	73.8	88.1	91.9	-	-	-
ResNeXt-GeM-GCL [17]	N	2048	75.5	86.1	88.5	56.0	70.8	75.1	64.0	81.2	86.6	37.8	53.6	62.9
ResNeXt-GeM-GCL [17]	Y	1024	80.9	90.7	92.6	62.3	76.2	81.1	79.2	90.4	93.2	58.1	74.3	78.1
ViT-GCL	N	768	71.2	84.9	88.4	50.0	67.9	73.9	70.8	88.0	91.9	40.3	60.0	68.9
ViT-GCL	Y	768	80.0	91.1	92.4	57.9	74.0	78.8	83.6	93.8	95.8	72.1	83.2	87.0
ViT-R50-GCL	N	768	69.7	81.6	85.3	46.8	62.7	69.5	70.9	86.7	91.2	42.9	62.9	70.5
ViT-R50-GCL	Y	768	78.6	88.1	90.4	55.9	71.1	77.2	82.6	92.8	95.4	76.2	86.0	87.3
SP-SuperGlue⋆ [14]	-	-	78.1	81.9	84.3	50.6	56.9	58.3	87.2	94.8	96.4	88.2	90.2	90.2
DELG⋆ [16]	-	-	83.2	90.0	91.1	52.2	61.9	65.4	89.9	**95.4**	**96.7**	**95.9**	**96.8**	**97.1**
Patch NetVLAD⋆ [15]	Y	4096	79.5	86.2	87.7	48.1	57.6	60.5	88.7	94.5	95.9	86.0	88.6	90.5
TransVPR⋆ [9]	-	-	86.8	91.2	92.4	63.9	74	77.5	89	94.9	96.2	-	-	-
R2Former [11]	-	-	**89.7**	**95**	**96.2**	**73.0**	**85.9**	**88.8**	**91.1**	95.2	96.3	88.6	91.4	91.7
NetVLAD-MSE (ours)	N	32768	72.3	82.7	85.5	51.5	64.7	70.7	74.7	87.2	90.9	20.7	41.0	50.8
NetVLAD-MSE (ours)	Y	4096	71.4	82.7	85.8	51.3	66.0	71.3	53.5	75.2	82.9	44.8	60.3	66.7
ViT-MSE (ours)	N	768	80.4	90.5	93.2	58.7	73.7	79.6	82.1	92.7	95.1	55.9	73.0	76.5
ViT-MSE (ours)	Y	768	82.4	90.5	92.3	60.5	73.8	78.4	85.1	94.2	95.9	71.4	85.4	89.8
ViT-R50-MSE (ours)	N	768	82.6	91.1	**93.6**	61.8	77.2	80.7	83.9	93.0	95.1	59.4	75.9	82.5
ViT-R50-MSE (ours)	Y	768	84.3	91.1	93.6	64.4	77.3	81.2	86.4	94.7	96.0	78.1	88.9	91.8

that we learn by regressing image similarity in the form of field-of-view overlap. We summarize the results in Table I.

We demonstrate that a hybrid transformer architecture (ViT-R50 [33]), trained for a single epoch by optimizing a straightforward regression loss function can outperform methods with more complex algorithms (e.g. re-ranking of the retrieved results) and training strategies (e.g. hard-pair mining for training), such as NetVLAD [7] and TransVPR [9]. We indeed achieved higher results on the MSLS, Pittsburgh30k and Tokyo24/7 datasets than methods that perform retrieval only based on descriptor distance ranking. We achieve in some cases higher results than methods that also apply re-ranking (R@5 3.3% higher than TransVPR on the MSLS test set), or otherwise comparable and very competitive results (e.g. R@5 0.1% lower than TransVPR on MSLS val, or 0.7% lower than DELG on Pittsburgh30k). R2Former [11] leads the results on MSLS. It is, however, trained with a complex joint optimization of retrieval and re-ranking, thus requiring substantial more complexity than our method.

The other method that relies on graded similarity ground truth, namely GCL, achieves lower results than our models trained using regression. This highlights the fact that our regression-based approach further exploits the information provided by the graded ground truth, and result in descriptors that are more suited for image retrieval and place recognition.

Data-efficiency. We study the data-efficiency of the proposed method, namely the ability to exploit few samples to train highly-performing methods in low-data regimes. We show that the proposed regression-based training requires only a few thousand image pairs (thus training iterations) to learn very effective and robust descriptors. In Fig. 2, we report the results (R@5) achieved at intermediate steps while training on the MSLS dataset. We evaluate the data-efficiency of a hybrid transformer trained using the MSE loss and compare the results to those of the same backbone trained using the GCL and the binary Contrastive loss (CL). We test a model snapshot every 10k iterations (i.e. after seeing 10k training pairs) on the MSLS validation set. We also test the generalization of the model snapshots to Pittsburgh30k and Tokyo24/7.

The regression-based training is able to very quickly learn robust and generalizable image descriptors for image retrieval and VPR. It uses only a few thousand image pairs (no need to complete a single epoch) to achieve very high results, superior to those of the GCL and CL training. The training regime of the MSE loss stabilizes very quickly to retrieval performance that is much higher than other methods. This has implications on several aspects, e.g. saving unnecessary computations and energy needed to train models with complex hard-pair mining for a longer time, quickly obtaining very powerful descriptors that do not require expensive re-ranking algorithms and allowing for hyperparameter optimization or model search that would not be feasible with existing models. It is worth pointing out that GCL is the only other method that uses data efficiently and can exploit graded image similarity ground truth. State-of-the-art methods require several training epochs and days to converge [17].

Quality of ranked retrieval results. We measured the quality of the retrieval ranking of our models using three different metrics. We compute the Recall@5 (R@5), which considers the place of a query image as correctly recognized if there is at least one true positive among the top-5 retrieved database images. This is a standard metric in visual place recognition and image retrieval but does not take into account the position of those true positives in the ranking. We compute the top-5 Mean Reciprocal Ranking (MRR@5) to measure the position in which the first positive hit appears in the top-5 retrieved images. It is equal to 1 if the first retrieved image is a true positive, and f.i. is equal to $4/5$ if the first positive hit is in

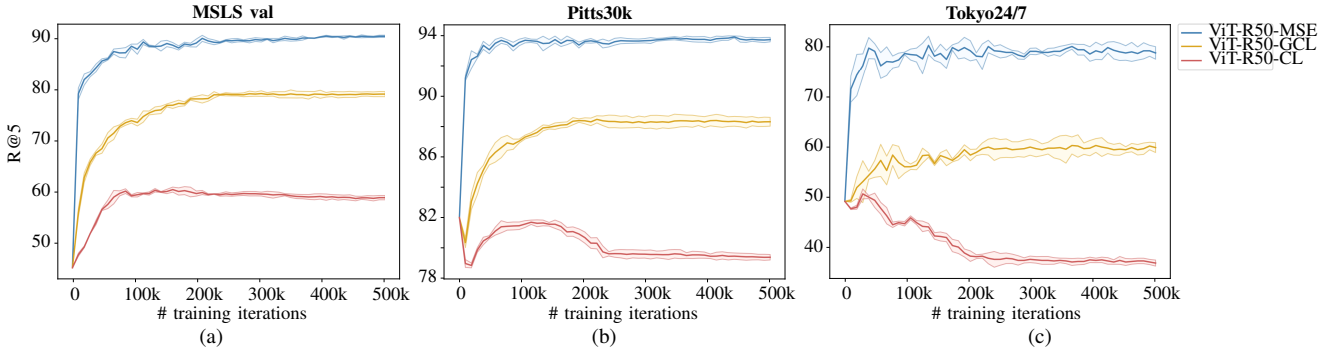


Fig. 2. Retrieval performance every 10k training iterations on the (a) MSLS-val, (b) Pitts30k, and (c) Tokyo24/7 for the same ViT-R50 encoder trained with CL, GCL, and MSE loss functions. We ran each experiment three times and report the average, minimum and maximum R@5.

TABLE II

COMPARISON OF RANKED RETRIEVAL RESULTS OF BACKBONES TRAINED WITH GCL [17] AND MSE (OURS) LOSS.

Dataset	Metric	ViT-R50-GCL	ViT-R50-MSE
MSLS-val	R@5	81.6	**91.1**
	MRR-5	0.4399	**0.5433**
	KLDiv	14×10^{-4}	1×10^{-4}
Pitts30k	R@5	86.7	**93.0**
	MRR@5	0.5024	**0.6351**
	KLDiv	30.6×10^{-3}	5.7×10^{-3}
Tokyo24/7	R@5	62.9	**75.9**
	MRR@5	0.3226	**0.4505**
	KLDiv	37×10^{-4}	8×10^{-4}

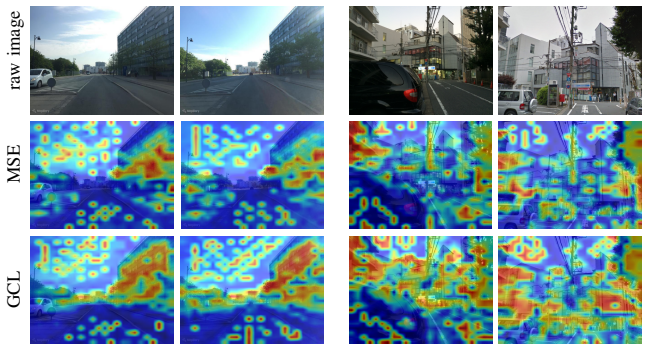


Fig. 3. Example attention maps on the last layer of ViT-R50-MSe and ViT-R50-GCL models for pairs of similar images from the MSLS validation dataset (columns 1-2), and the Tokyo24/7 dataset (columns 3-4).

the second position, until the score equals to 0 if there is no positive retrieved among the 5 first results. When training a model with a ViT-R50 using a regression loss, not only the recall is higher, but correct images are placed earlier in the ranked list than with the counterpart model trained using a contrastive approach with graded similarity labels (see Table II). This also applies in out-of-distribution tests on the Pitts30k and Tokyo24/7 datasets, demonstrating generalization of ranking capabilities. We compare with the GCL as it is the only method trained using graded similarity at large-scale.

Furthermore, we compute the Kullback-Leibler divergence between the distribution of query-map pairwise Euclidean distance and the annotated similarity ground truth of image pairs. For MSLS, the ground truth is the FoV overlap [17]. For Pitts30k and Tokyo24/7 datasets, we consider the binary ground truth publicly available. The Kullback-Leibler divergence measures how different two distributions are. As such, a lower value indicates a higher similarity between distributions. The KL of the MSE-trained model is at least one order of magnitude smaller than that of its GCL counterpart (Table II). The MSE-trained descriptors are thus better suited for VPR, as their distance in the latent space is a more robust measure of image visual similarity and ranking.

Attention maps. We show example attention maps of our regression-trained transformer compared to those of its counterpart trained with GCL in Fig. 3. Although both models tend to have attention maps with focus on relevant parts of the images, the model trained with MSE reacts less to non-permanent features, such as the cars in the left images (both MSLS and Tokyo24/7). This supports the evidence that regression-trained models perform better than contrastive-based ones, as they focus on structural shared clues in images.

B. Ablation experiments

Encoder. We trained several encoders on the MSLS dataset [13] by optimizing the MSE loss, and studied the impact of transformers and convolutional backbones. Following [7], [17] we chose the following encoders: NetVLAD with a VGG16 backbone, and fully-convolutional backbones VGG16 and ResNeXt with GeM pooling [6]. We also considered two transformers, namely a Vision Transformer (ViT) and a Hybrid Vision Transformer (ViT-R50) with knowledge distilled from ResNet50, without a pooling layer nor a projection head. All encoders (and the VGG16 encoder of NetVLAD) were pre-trained on ImageNet.

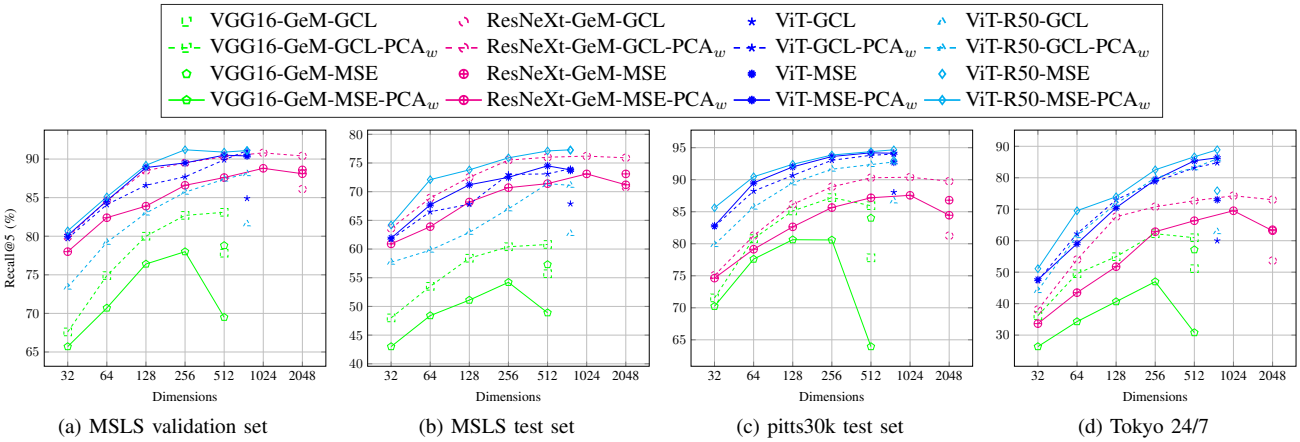
In Table III, we report results on the MSLS validation and test sets and the generalization performance on the Pitts30k and Tokyo24/7 benchmarks. In all cases, the transformers achieve higher results, with the hybrid ViT-R50 models achieving the highest performance in all considered datasets. However, the ablation results support that the MSE loss can be applied to any architecture for VPR successfully.

PCA and whitening. We study the effect of PCA and whitening on our descriptors [7], [6]. We observed that

TABLE III

ABLATION STUDY OF ENCODERS AND PCA. ALL MODELS ARE TRAINED ON THE MSLS DATASET WITH AN MSE LOSS.

Encoder	PCA_w	Dim	MSLS-val			MSLS-test			Pitts30k			Tokyo24/7		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD	N	32768	72.3	82.7	85.5	51.5	64.7	70.7	74.7	87.2	90.9	20.7	41.0	50.8
	Y	4096	71.4	82.7	85.8	51.3	66.0	71.3	53.5	75.2	82.9	44.8	60.3	66.7
VGG16-GeM	N	512	69.3	81.1	84.2	43.9	57.3	62.6	66.1	84.0	89.2	33.3	57.1	64.8
	Y	256	65.4	78	81.6	38.6	54.2	60.8	62.3	80.6	86.4	28.3	47.0	52.4
ResNeXT-GeM	N	2048	78.8	88.6	90.7	58.5	73.1	78.6	71.9	86.8	91.3	41.6	63.5	71.4
	Y	1024	81.1	90.8	92.4	59.7	73.1	77.4	74.3	87.6	91.5	51.8	69.7	78.4
ViT	N	768	80.4	90.5	93.2	58.7	73.7	79.6	82.1	92.7	95.1	55.9	73.0	76.5
	Y	768	82.4	90.5	92.3	60.5	73.8	78.4	85.1	94.2	95.9	71.4	85.4	89.8
ViT-R50	N	768	82.6	**91.1**	**93.6**	61.8	77.2	80.7	83.9	93.0	95.1	59.4	75.9	82.5
	Y	768	**84.3**	**91.1**	**93.6**	**64.4**	**77.3**	**81.2**	**86.4**	**94.7**	**96.0**	**78.1**	**88.9**	**91.8**



(a) MSLS validation set (b) MSLS test set (c) pitts30k test set (d) Tokyo 24/7

Fig. 4. Results obtained on the MSLS validation, MSLS test, Pittsburgh 30k and Tokyo 24/7 datasets by MSE-trained models with and without PCA whitening. Reducing the dimensionality of the descriptors and applying the whitening transform contribute to an increase of the retrieval performance.

although the whitening does not lead to significantly better performance on the MSLS dataset, it can help the generalization capabilities of the MSE-trained descriptors, especially in the case of the Tokyo 24/7 datasets (see Table III). We report the results achieved by applying whitening and PCA to decrease the size of the descriptors in Figure 4. For convolutional architectures, PCA whitening leads to a boost in the retrieval performance on the MSLS dataset, with less improvement in the out-of-distribution test benchmarks Pitts30k and Tokyo 24/7. For the MSE-trained models that use VGG16 as backbone (VGG16-GeM and NetVLAD), the PCA and whitening impact negatively on the results, especially in generalization tests. The whitening tends to improve the generalization performance of transformer backbones trained with the MSE loss to the Pittsburgh30 and Tokyo24/7 datasets, while not having an impact on in-distribution tests on the MSLS validation and test sets. This is attributable to the fact that MSE-trained models tend to learn features with lower covariance (see Figure 5). PCA of the transformer descriptors contributes to maintaining good performance even when down-scaling the feature space size to 128 dimensions.

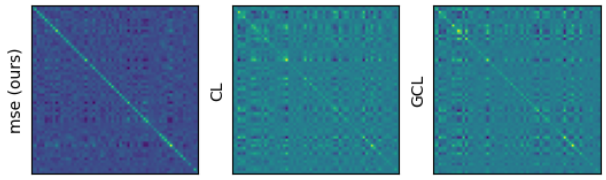


Fig. 5. Example of the covariance matrices of features (from the MSLS validation set) learned with MSE and contrastive losses.

V. CONCLUSIONS

We shifted from the contrastive learning paradigm for visual place recognition and showed that this task can be treated as a regression problem at large-scale with high results. We train Vision Transformers to learn global image descriptors whose distance in a latent space is a direct measure of similarity of the images. This straightforward training scheme results in descriptors that have exceptional ranking capabilities. Furthermore, they do not require complex and time-consuming re-ranking to curate the retrieval results. The training process is data-efficient as it requires a few iterations on a limited view of the data, namely a few thousand training pairs. Our training scheme dispenses VPR pipelines from the need for large datasets to learn effective descriptors, and can foster training high-performing models in energy-saving settings. We achieve better performance than other VPR methods and good generalization to out-of-distribution test sets. We achieved results higher than or comparable with methods that use re-ranking strategies or that are trained using triplet loss, hard-pair mining and that are trained for several epochs.

REFERENCES

[1] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012, pp. 1643–1649.

[2] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[3] D. Doan, Y. Latif, T. Chin, Y. Liu, T. Do, and I. Reid, "Scalable place recognition under appearance change for autonomous driving," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9318–9327.

[4] N. Pion, M. Humenberger, G. Csurka, Y. Cabon, and T. Sattler, "Benchmarking image retrieval for visual localization," in *International Conference on 3D Vision*, 2020.

[5] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2136–2174, 2021.

[6] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *TPAMI*, vol. 41, no. 7, pp. 1655–1668, 2018.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.

[8] M. Lopez-Antequera, M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Place and object recognition by cnn-based cosfire filters," *IEEE Access*, vol. 7, pp. 66 157–66 166, 2019.

[9] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zhen, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *CVPR*, 2022.

[10] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4868–4878.

[11] 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.

[12] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *CVPR*, 2019, pp. 2570–2579.

[13] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *CVPR*, 2020.

[14] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *CVPR*, 2020. [Online]. Available: https://arxiv.org/abs/1911.11763

[15] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *CVPR*, 2021, pp. 14 141–14 152.

[16] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 726–743.

[17] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," *CVPR*, 2023.

[18] V. Balntas, S. Li, and V. Prisacariu, "Relocnet: Continuous metric learning relocalisation using neural nets," in *ECCV*, 2018, pp. 751–767.

[19] D. Galvez-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.

[20] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *TPAMI*, 2015.

[21] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*. IEEE, 2010, pp. 3384–3391.

[22] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *TPAMI*, vol. 34, no. 9, pp. 1704–1716, 2011.

[23] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, p. 107760, 2020.

[24] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, pp. 1–1, 2021.

[25] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *ACRA*, 2014.

[26] M. Leyva-Vallina, N. Strisciuglio, M. López-Antequera, R. Tylecek, M. Blaich, and N. Petkov, "Tb-places: A data set for visual place recognition in garden environments," *IEEE Access*, 2019.

[27] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Place recognition in gardens by learning visual representations: data set and benchmark analysis," in *CAIP*. Springer, 2019, pp. 324–335.

[28] A. Gordo, J. Almazán, 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] M. Angelina Uy and G. Hee Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *CVPR*, 2018, pp. 4470–4479.

[30] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.

[31] J. Thoma, D. P. Paudel, and L. Van Gool, "Soft contrastive learning for visual localization," *NeurIPS*, 2020.

[32] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2013.

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