

# SUP-VPR: A TRANSFORMER-BASED FRAMEWORK FOR VISUAL PLACE RECOGNITION IN LOW-DATA REGIMES

WhizResearch Nemo

## ABSTRACT

Recently, transformer-based methods have achieved remarkable success in numerous vision tasks, including visual place recognition (VPR). These methods utilize the multi-head attention mechanism to capture complex relationships between images and produce discriminative image embeddings for retrieval. However, the reliance on large-scale annotated data for training presents a major limitation for many real-world applications. To this end, we introduce SUP-VPR, a novel two-stage transformer-based VPR framework designed to operate effectively in low-data regimes (i.e., without the need for large-scale annotated data). In particular, SUP-VPR incorporates supervised real-world data with synthetic data within a unique two-stage retrieval framework. The first stage focuses on intra-city dataset retrieval, while the second stage is dedicated to more challenging inter-city dataset retrieval for city-level localization. Additionally, a cross-attention mechanism is designed to handle inter-city retrieval, and a MixVPR-like fusion module is proposed for effective intra-city feature aggregation. Furthermore, we introduce an adaptive hard negative mining approach for better discrimination against other city databases. Our extensive experiments conducted on seven challenging VPR datasets across diverse scenarios demonstrate the effectiveness of our method for place recognition in low-data regimes.

## 1 INTRODUCTION

Visual place recognition (VPR) is the fundamental objective of determining the location of an image based solely on its visual content, without any auxiliary information (Zhang et al., 2021b; Ali-bey et al., 2022b; Bertol et al., 2022; 2023; Keetha et al., 2023). This task proves particularly useful in augmented reality (AR), where GPS signals are often unavailable or unreliable, and robot navigation, where GPS signals can be noisy or delayed. In particular, VPR systems usually rely on two main components: image descriptors and index-structures. Image descriptors extract a global feature vector for each reference image to create a gallery set, while the index-structures efficiently search for the most similar images in the database (the query set) with the gallery set as the first step for VPR.

Over the past years, deep learning models have dominated the field of landmark retrieval (Radenović et al., 2018; Wang et al., 2022; Zhu et al., 2023), but there are only a few attempts to apply them in the field of VPR due to two main reasons: 1) large-scale annotated datasets are not available for training VPR models as the exact localization of images is not required during annotation; 2) the efficient and compact global descriptors are preferred in VPR applications rather than large-scale and high-dimensional ones.

Recently, benefiting from the powerful transformer architecture (Vaswani et al., 2017), some pioneering works have attempted to study transformer-based VPR systems without re-ranking on some relatively less challenging VPR datasets, such as Brooklyn (Wang et al., 2023b; Zhang et al., 2023a). However, they require more annotations and image pairs during the training phase, and the actual performance is not comparable with those SOTA re-ranking methods as shown in Tab. 1. Moreover, they fail to generalize to more challenging, real-world scenarios. The promising results of these works motivate us to explore the potential of transformer-based VPR systems in more practical scenarios.

In this paper, we propose a novel transformer-based VPR framework, called SUP-VPR, which can perform well in low-data regimes (without the need for large-scale annotated data). The proposed

SUP-VPR follows a two-stage retrieval framework. Fig. ?? shows the detailed framework. The intra-city retrieval stage is responsible for recognizing the specific landmarks within a city, which holds less localization information for city-level retrieval. Hence, we design this stage to be compact, utilizing a well-trained off-the-shelf model (DINOv2 (Oquab et al., 2023)) as the backbone to extract features for 300 Holistic attention blocks as image embeddings. These embeddings are then fused by a MixVPR-like feature-fusion module (Ali-bey et al., 2023) to produce efficient and compact descriptors for intra-city retrieval. Note that supervised real-world images and synthetic images are *jointly* taken into account as the training data for SUP-VPR. Then, as the second stage, the inter-city retrieval stage is responsible for recognizing the target city. Considering the scarce supervision in real-world, we introduce a novel cross-attention mechanism (the part shaded in yellow as shown in Fig. ??) that utilizes 12 attention blocks to focus on inter-city feature modeling. Specifically, these attention blocks are prepended to the transformer blocks to enhance the transformer backbone’s capability in modeling long-range dependencies. Furthermore, a novel adaptive hard negative mining strategy is proposed to make the model more discriminative against other city databases during the training phase. Overall, our main contributions can be summarized as follows:

- We propose a novel two-stage transformer-based VPR framework called SUP-VPR, which can perform well in low-data regimes. To the best of our knowledge, this is the first work that studies the potential of transformer-based VPR under these practical scenarios.
- We propose a novel cross-attention mechanism for inter-city retrieval and a MixVPR-like fusion module for intra-city retrieval. Furthermore, an adaptive hard negative mining strategy is presented to make the model more discriminative against other city databases.
- We conduct extensive experiments on seven popular VPR datasets under both re-ranked and non-re-ranked settings. The promising results show the effectiveness of the proposed SUP-VPR.

## 2 RELATED WORKS

**Visual Place Recognition.** Recent years have seen rapid developments of VPR, with hand-crafted descriptors (Philbin et al., 2007; Lowe, 2004; Torii et al., 2013; Gálvez-López & Tardos, 2012; Arandjelovic & Zisserman, 2013; Jégou et al., 2011; Bay et al., 2006; Rublee et al., 2011), and more recently deep features (Babenko & Lempitsky, 2015; Radenović et al., 2018; Kim et al., 2017; Liu et al., 2019; Ge et al., 2020; Wang et al., 2019; ?; 2022; Zhu et al., 2023) as the global image representations. Supervised by the only GPS coordinates, VPR is a weak-supervised learning task. Much work has been done to improve the generalization ability of VPR models, from hand-crafted descriptors (Torii et al., 2013; Arandjelovic et al., 2016) to deep features (Radenović et al., 2018; Wang et al., 2022; Zhu et al., 2023). The robust VPR feature vectors should encode the most discriminative features within images (Radenović et al., 2018; Wang et al., 2022). To further improve the performance of VPR, recent works have attempted to re-rank the matches after the retrieval phase, which maps as many reference images as possible to the query image to improve the localization accuracy (Barbarani et al., 2023).

Over the past years, the development of VPR has been boosted by the powerful convolutional network, which fuses multi-scale and multi-resolution features to produce robust and compact descriptors (Hausler et al., 2021; Zhang et al., 2021a; Yu et al., 2019; Cao et al., 2020; Masone & Caputo, 2021; Zhang et al., 2023b). The first attention-based VPR method, TransVPR (Wang et al., 2022), focuses on multi-level attention aggregation, which aggregates multi-level features to produce the global feature embedding. MixVPR (Ali-bey et al., 2023) and its derivatives (Hou et al., 2023; Huang et al., 2023) utilize the powerful Mix module from MLFoundations to produce descriptors, which mixes features from different resolutions to produce compact and robust descriptors. GSV cities (Ali-bey et al., 2022b) and GPM (Ali-bey et al., 2022a) explore how a better sampling strategy can improve the performance of VPR models. Recently, Berton *et al.* (Berton et al., 2023) explore viewpoint robustness of VPR models with the large-scale datasets with GPM. However, to the best of our knowledge, there is still no derivative work that explores the potential of attention-based VPR in more practical scenarios, which motivates us to fill in this gap.

**Transformers in Vision.** Recently, transformers (Vaswani et al., 2017) have shown remarkable success in various vision tasks, including image classification (Dosovitskiy et al., 2020; Chen et al., 2021), object detection (Carion et al., 2020), and semantic segmentation (Zheng et al., 2021; Liu et al.,

Table 1: Comparisons to Naive-PR (Wang et al., 2023a), TransVPR (Wang et al., 2022), and ATTR (Wang et al., 2023b). The dimensions of the produced descriptors (e.g., 1024) are indicated in the brackets. The best results are shown in **bold** and the second best results are underlined. The data used for training the models are highlighted in blue color.

Methods	Backbone	Embed. Dim.	Image Pair	Label Pair	Cross-Attn.	Landmark Cls. Classifier
Naive-PR (Wang et al., 2023a)	DINOv2 (Oquab et al., 2023)	1024	3	$\times$	$\times$	$\times$
TransVPR (Wang et al., 2022)				$\checkmark$	$\checkmark$	$\checkmark$
ATTR (Wang et al., 2023b)				$\checkmark$	$\checkmark$	$\checkmark$
ATTR+Naive-PR				$\checkmark$	$\checkmark$	$\checkmark$
SUP-VPR	DINOv2 (Oquab et al., 2023)	1024	3	3	$\checkmark$	$\times$

2021). In contrast to CNNs, transformers treat images as sequences of patches and utilize multi-head attention mechanisms to capture complex relationships between them. Most recently, based on the success of self-supervised pre-training, DINO (?) and its sequel DINOv2 (Oquab et al., 2023) have achieved better results in image recognition. In this work, we use the well-trained DINOv2 as the backbone to extract features for VPR.

**Transformers for VPR.** Transformers have been applied to VPR in prior works (Wang et al., 2022; 2023b; Zhang et al., 2023a; Wang et al., 2023a; Leyva-Vallina et al., 2023). TransVPR (Wang et al., 2022) is the first attention-based VPR method, which utilizes multi-scale features as a sequence to perform self-attention. This work follows the simple Sum-pooling aggregation for producing global descriptors. However, the compactness of the produced descriptors is not comparable with those produced by CNN-based methods as shown in 1. Besides, the attention-based VPR method (Wang et al., 2023b) leverages the semantic region-level similarity to produce the similarity matrix, which is further used to perform the weighted-sum attention. However, it requires face crops, which should be avoided in most VPR applications. Besides, the VPR performance is significantly influenced by the viewpoint changes and the urban scenes. To alleviate these issues, Eigenplaces (Berton et al., 2023) proposes a viewpoint-aware minibatch sampling strategy along with a novel soft and hard minibatch triplet loss for viewpoint-robust VPR. However, all these works require more labeled data for training purposes, which is often costly and not available in reality. Note that some works (Berton et al., 2021; 2023; Milford & Wyeth, 2008; Yhdiz et al., 2022) use more annotated data for training, but they use the data for different tasks (e.g., training an auxiliary classifier for training). In this work, we propose a novel two-stage transformer-based VPR framework, called SUP-VPR, which can perform well in low-data regimes. The compact and robust descriptors produced by our SUP-VPR are *competitive* with those produced by CNN-based methods.

### 3 APPROACH

This section details the proposed two-stage VPR framework, which aims to determine the location of a query image. Fig. ?? shows an overview of our proposed SUP-VPR. The proposed SUP-VPR tackles the VPR task from two stages. In the first stage, the intra-city retrieval stage is responsible for recognizing the specific landmarks within a city, which holds less localization information for city-level retrieval. Hence, we design this stage to be compact, utilizing a well-trained off-the-shelf model (DINOv2 (Oquab et al., 2023)) as the backbone to extract features for images. Furthermore, these image embeddings are fed to a MixVPR-like feature fusion module to produce compact and robust descriptors for retrieval. In the second stage, the inter-city retrieval stage is responsible for recognizing the target city. Considering the scarce supervision in real-world, we introduce a novel cross-attention mechanism (the part shaded in yellow as shown in Fig. ??) that utilizes 12 attention blocks to focus on inter-city feature modeling. Specifically, these attention blocks are prepended to the transformer blocks to enhance the transformer backbone’s capability in modeling long-range dependencies.

### 3.1 INTRA-CITY RETRIEVAL STAGE

The first retrieval stage is similar to re-ranked VPR methods, which focus on intra-city place recognition. This stage concentrates on identifying the same landmarks within a city. Considering the unique visual characteristics of each city, we utilize the off-the-shelf self-supervised model DINOv2 (Oquab et al., 2023) as the backbone to extract image embeddings, which can enhance the learning of discriminative features for images. Furthermore, we propose a novel feature fusion module based on the architecture from MixVPR (Ali-bey et al., 2023) to produce robust global feature embeddings for this stage.

**Backbone.** Recently, pre-trained vision models (especially Transformers) have received a lot of attention due to their outstanding performance in various vision tasks. Among these models, DINOv2 (Oquab et al., 2023) is a prominent one, which relies solely on the self-supervised pre-training and has even outperformed some fully supervised models. In this paper, we utilize this off-the-shelf model to extract image embeddings as a sequence of patches for intra-city retrieval.

Specifically, for a query image  $I_q \in \mathbb{R}^{H \times W \times 3}$  with its metadata (e.g., GPS coordinates and view directions), we adopt DINOv2 to extract a sequence of image patches  $\hat{\mathbf{x}}_q = \{\hat{\mathbf{x}}_q^b \in \mathbb{R}^{64 \times 64 \times 3}\}_{b=1}^B$ , where  $B = HW/P^2$  is the number of patches with the patch size  $P \times P$  (e.g.,  $16 \times 16$ ). Note that DINOv2 uses  $64 \times 64$  patches to perform self-supervised pre-training, we here downsample the feature map from  $384 \times 384$  to  $64 \times 64$  to enhance the runthrough under limited computation resources. These image patches,  $\hat{\mathbf{x}}_q^b$ , are further mapped into latent embeddings  $\tilde{\mathbf{x}}_q^b = \phi_b(\hat{\mathbf{x}}_q^b)$  with the linear layer  $\phi_b(\cdot) : \mathbb{R}^{64 \times 64 \times 3} \rightarrow \mathbb{R}^d$ . We then obtain the image embeddings by performing mean-pooling on the sequence of latent embeddings as  $\mathbf{x}_q \in \mathbb{R}^d$ . For the reference images in the database, we use the same operation to extract their image embeddings  $\mathbf{x}_r \in \mathbb{R}^d$ .

**MixVPR-like Feature Fusion.** To capture multi-scale and multi-resolution information in the image embeddings, we further adopt a MixVPR-like feature fusion module (Ali-bey et al., 2023) to produce robust and compact global feature embeddings for intra-city retrieval. In particular, the MixVPR-like module performs element-wise addition on image embeddings at different scales to produce the global feature embeddings. Mathematically, it can be formulated as follows:

$$f_{\text{MixVPR}}(\mathbf{x}_q, \mathbf{x}_r) = \sum_{i=1}^K \gamma_i \cdot [\Phi_i(\mathbf{x}_q) \oplus \Phi_i(\mathbf{x}_r)], \quad (1)$$

where  $\Phi_i(\cdot)$  is the  $i_{th}$  layer norm operation.  $\gamma_i$  is the weight for the specific layer, which is learned automatically during the training phase. Furthermore,  $\oplus$  represents the operation of element-wise addition, which is utilized to aggregate multi-scale and multi-resolution features. Note that the original feature fusion module from MixVPR utilizes a specific mixing operation to perform feature mixing for the aggregated global feature embeddings. However, considering the computation resources and the specific task of intra-city retrieval, we remove this operation to achieve higher inference efficiency and better performance. In the following, we refer to this modified feature fusion module as the MixVPR-like fusion module.

With the proposed MixVPR-like fusion module, we can fuse the image embeddings  $\mathbf{x}_q$  and  $\mathbf{x}_r$  to produce the global feature embeddings  $f_{\text{MixVPR}}(\mathbf{x}_q, \mathbf{x}_r) \in \mathbb{R}^d$ . These global feature embeddings are utilized to perform inner-product similarity to obtain the similarity scores between the query image and the reference images. During the training phase, we further perform contrastive learning on these global feature embeddings to enhance the discriminative capability of the feature fusion module. Note that we perform random data augmentation (e.g., color distortion, Gaussian blur) during the training phase to enhance the robustness of the model.

### 3.2 INTER-CITY RETRIEVAL STAGE

The second retrieval stage focuses on recognizing the city of the query image. We propose a novel cross-attention mechanism that utilizes 12 attention blocks to focus on inter-city feature modeling. Specifically, these attention blocks are prepended to the transformer blocks to enhance the transformer backbone’s capability in modeling long-range dependencies. Furthermore, a novel adaptive hard negative mining strategy is proposed to make the model more discriminative against other city databases.

**Cross-attention for inter-city retrieval.** Considering the limited supervision for city recognition, we propose a novel cross-attention mechanism that utilizes specific attention blocks to focus on inter-city feature modeling. Specifically, we prepend these attention blocks to the transformer blocks to enhance the transformer backbone’s capability to model long-range dependencies. The architecture of these attention blocks is similar to that of the multi-head attention blocks, which consists of three weight matrices, namely query weights  $\mathbf{W}'_Q \in \mathbb{R}^{d \times d}$ , key weights  $\mathbf{W}'_K \in \mathbb{R}^{d \times d}$ , and value weights  $\mathbf{W}'_V \in \mathbb{R}^{d \times d}$ . With these three weight matrices, for a query image with its image embeddings  $\mathbf{x}_q$ , its database counterpart  $\mathbf{x}_a$  and the image embeddings of other cities  $\mathbf{x}_o \in \mathbb{R}^{d' \times d}$ , where  $d'$  is the number of other cities, we can produce the inter-city image embeddings as follows:

$$\begin{aligned} \mathbf{y}_q &= \mathbf{x}_q + \mathbf{h}_q, \\ \mathbf{h}_q &= \text{Softmax}\left(\frac{\mathbf{W}'_Q \cdot \mathbf{x}_q}{\sqrt{d}} \cdot \frac{\mathbf{W}'_K \cdot [\mathbf{x}_a, \mathbf{x}_o]}{\sqrt{d'}}\right) \cdot \mathbf{W}'_V \cdot [\mathbf{x}_a, \mathbf{x}_o], \end{aligned} \quad (2)$$

where  $\mathbf{y}_q \in \mathbb{R}^d$  serves as the inter-city image embeddings for city recognition. In particular,  $\mathbf{h}_q$  is the output of the proposed attention block, which captures the cross-attention between the query image and its database counterpart as well as the other cities. It is a summation of the similarity results with the value matrix  $\mathbf{W}'_V$  applied to the image embeddings of both the database counterpart  $\mathbf{x}_a$  and other cities  $\mathbf{x}_o$ . Furthermore, before performing the softmax operation, we divide the query matrix  $\mathbf{W}'_Q \cdot \mathbf{x}_q$  and the key matrix  $\mathbf{W}'_K \cdot [\mathbf{x}_a, \mathbf{x}_o]$  by a scaling factor  $\sqrt{d}$ . This scaling factor controls the magnitude of the softmax operation to make the training more stable and effective. Furthermore, we utilize the  $\mathbf{x}_a$  to perform contrastive learning for training to ensure the feasibility of the training. Note that we do not perform any data augmentation or other operations on the inter-city retrieval stage to ensure the consistency between the training phase and the inference phase.

Here we show more visualization results to further understand the proposed attention block. Fig. ?? shows the attention map of the proposed attention block and the down-scaled self-attention block from ATTR (Wang et al., 2023b). We observe that the attention maps of the down-scaled self-attention block mostly focus on the foreground regions, which cannot accurately reflect the correspondence between the query image and the database counterpart, as well as the correspondence between the query image and the other cities. In contrast, the proposed attention block can focus on the most discriminative regions and has a better feasibility to model the correspondence between images from different cities.

**Adaptive Hard Negative Mining.** Over the past years, a lot of works (Radenović et al., 2018; Wang et al., 2022; Zhu et al., 2026) have explored how an appropriate sampling strategy can improve the performance of VPR. In this paper, we propose to sample hard negatives for the second retrieval stage to make the model more discriminative against other city databases. Specifically, we minimize the distance of the global feature embeddings  $\mathbf{f}_q$  of the query image and the global feature embeddings  $\mathbf{f}_a$  of its database counterpart, while maximizing the distance of the global feature embeddings  $\mathbf{f}_q$  of the query image and the global feature embeddings  $\mathbf{f}_o$  of the image embeddings of other cities. This can be formulated as:

$$\mathcal{L} = -\log \frac{\exp(\langle \mathbf{f}_q, \mathbf{f}_a \rangle / \tau)}{\exp(\langle \mathbf{f}_q, \mathbf{f}_a \rangle / \tau) + \sum_{i=1}^{d'} \exp(\langle \mathbf{f}_q, \mathbf{f}_{o_i} \rangle / \tau)}, \quad (3)$$

where  $\langle \cdot, \cdot \rangle$  is the operation of inner-product similarity, and  $\tau$  is the temperature factor. The utilization of hard negative samples can guide the second retrieval to focus more on the discriminative regions of the city database images, which can improve the generalization ability of the model to other city databases.

Note that the hard negative mining strategy should be different according to the specific task. In this paper, we propose to adaptively mine hard samples based on the similarity distance of the image embeddings. Specifically, we calculate the average similarity of the image embeddings within a city, and sample the images from other cities that are closest to the average similarity of the specific city as hard samples. With this adaptive strategy, the model can concentrate on the most challenging samples, which can enhance the generalization ability of the model to other city databases. Furthermore, we only mine hard samples from the images of other cities to enhance the cross-city generalization ability.

Table 2: The impact of using image pairs of different classes during the training phase.

Image Pair	Brooklyn		Oxford II		London		Pittsburgh		Boston		Svoboda	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Different	88.54	98.80	88.29	96.50	77.06	91.34	85.98	96.36	71.17	87.65	63.17	77.98
Same	89.06	99.05	87.17	95.73	75.46	90.61	85.40	96.11	72.46	88.85	62.93	77.61
None	88.54	98.84	87.30	95.74	75.93	91.12	85.62	96.28	68.68	86.96	62.64	77.76

Table 3: The impact of utilizing the proposed cross-attention mechanism on the performance of VPR.

Cross-Attention	Brooklyn		Oxford II		London		Pittsburgh		Boston		Svoboda	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Yes	88.54	98.80	88.29	96.50	77.06	91.34	85.98	96.36	71.17	87.65	63.17	77.98
No	88.00	98.42	86.60	95.59	75.63	90.56	85.18	95.98	69.08	86.94	62.31	76.99

### 3.3 OVERALL LOSS FUNCTION

To train SUP-VPR, we perform contrastive learning on the global feature embeddings of the two retrieval stages, which aligns with the weakly-supervised learning paradigm of the VPR task. For the intra-city retrieval stage, we utilize the similarity of the global feature embeddings of the query image and the reference images from the same city as the ground truth. For the inter-city retrieval stage, we utilize the global feature embeddings of the database counterpart as the ground truth. According to Eq.3, we only mine hard samples from the images of other cities to enhance the cross-city generalization ability. Consequently, the overall loss function can be formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{intra} + \lambda \mathcal{L}_{inter}, \quad (4)$$

where  $\mathcal{L}_{intra}$  and  $\mathcal{L}_{inter}$  are the loss functions for the intra-city retrieval stage and the inter-city retrieval stage, respectively.  $\lambda$  is the weight factor for the loss function of the inter-city retrieval stage, which is set to 1e5 in our paper.

## 4 EXPERIMENTS

### 4.1 DATASETS AND METRICS

We conduct our experiments on seven challenging VPR datasets, including Brooklyn (Arandjelovic et al., 2016), Oxford II, London, Pittsburgh, Svoboda (Sünderhauf et al., 2013), Boston (Babenko & Lempitsky, 2015), and Mapillary Street-Level Sequences (MSLS) (Warburg et al., 2020).

Following the original VPR datasets and relevant papers (Warburg et al., 2020), we use the training set composed of 64096 images from Brooklyn, Oxford II, London, and Pittsburgh, to train the proposed SUP-VPR. Furthermore, the testing phase involves the remaining images and images from Boston, Svoboda, and MSSL. Specifically, following the original VPR datasets and relevant papers (Warburg et al., 2020), we use 70% of the images for training and 30% for testing. We perform five runs and report the averaged results.

Following (Wang et al., 2022), we use Recall at N ( $R@N$ ) with  $N \in \{1, 5\}$  and Top-5 Recall at 1m ( $R1m$ ) as the evaluation metrics. For  $R@N$ , the dirac delta function  $\delta\{\cdot\}$  is used to return 1 when the top-N matches (e.g.,  $N=1$ ) include the correct answer, and 0 otherwise.  $R@N$  is the ratio of the sum of these top-N match results to the query image number within the tested set. For  $R1m$ , it is the ratio of the sum of the correctness of the first match within the tested set.

### 4.2 EXPERIMENTAL SETTINGS

**Training Settings.** We use PyTorch (Paszke et al., 2019) with the Torchmeta (Abadi et al., 2016) package to implement our SUP-VPR with an NVIDIA V100 GPU. We use the Adam optimizer (?) with a weight decay of 1e-4, a learning rate of 1e-5, and a batch size of 16 to train our SUP-VPR. For the training phase of the intra-city retrieval stage, we train the proposed SUP-VPR for three epochs. For the training phase of the inter-city retrieval stage, we only train the proposed model for one epoch. Note that we use the MixVPR-like fusion module for one epoch before we use it

Table 4: The impact of utilizing the proposed hard negative mining strategy on the performance of VPR.

Hard Negative Mining	Brooklyn		Oxford II		London		Pittsburgh		Boston		Svoboda	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Yes	88.54	98.80	88.29	96.50	77.06	91.34	85.98	96.36	71.17	87.65	63.17	77.98
No	87.89	98.51	86.93	95.99	76.57	91.10	85.84	96.34	69.48	86.99	62.48	77.41

Table 5: Comparison with state-of-the-art methods. The methods are divided into two groups: those using pre-trained models (e.g., ResNet101 and DINOv2) trained on large-scale datasets, and those using only data from VPR datasets. The best results are shown in **bold**, and the second-best results are underlined. Methods using pre-trained models are displayed in **magenta**.

Methods	Venue	Brooklyn		Oxford II		London		Pittsburgh		Boston		Svoboda		MSLS	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
SURF+NM (Arandjelovic et al., 2016)	CVPR'06	59.6	87.7	79.0	95.8	64.8	86.9	70.1	91.8	49.3	73.3	51.6	70.6	-	-
SURF+VLAD (Arandjelovic et al., 2016)	CVPR'06	63.0	91.1	79.0	95.8	64.8	86.9	70.1	91.8	49.3	73.3	51.6	70.6	-	-
BRIEF+NM (Torii et al., 2013)	CVPR'08	62.1	88.0	79.0	95.8	64.8	86.9	70.1	91.8	49.3	73.3	51.6	70.6	-	-
GIST+VLAD (Torii et al., 2013)	CVPR'11	63.0	91.1	79.0	95.8	64.8	86.9	70.1	91.8	49.3	73.3	51.6	70.6	-	-
DIS+VLAD (Torii et al., 2013)	CVPR'13	63.0	91.1	79.0	95.8	64.8	86.9	70.1	91.8	49.3	73.3	51.6	70.6	-	-
NetVLAD (Arandjelovic et al., 2016)	CVPR'16	85.0	97.2	85.9	97.0	74.5	91.8	82.9	96.3	65.8	86.1	59.9	76.9	-	-
SPIN+VLAD (DeTone et al., 2018)	CVPR'18	-	-	89.3	97.5	75.3	92.3	83.3	96.6	69.2	87.6	60.5	77.2	-	-
Eigenplaces (Berton et al., 2023)	ICCV'23	-	-	91.0	98.0	79.0	93.0	86.0	96.5	75.0	90.0	64.0	79.0	-	-
TransVPR (Wang et al., 2022)	CVPR'22	96.0	99.4	91.0	97.8	79.2	92.1	83.6	96.3	67.6	87.0	61.4	78.9	-	-
MixVPR (Ali-bey et al., 2023)	WACV'23	<b>96.6</b>	<b>99.4</b>	<b>91.3</b>	<b>97.8</b>	<b>80.2</b>	<b>92.5</b>	<b>84.5</b>	<b>96.6</b>	<b>70.1</b>	<b>87.3</b>	<b>62.3</b>	<b>79.2</b>	-	-
AnyLoc (Keeha et al., 2023)	NeurIPS'23	-	-	85.1	-	69.3	-	76.7	-	56.0	-	57.6	-	-	-
ATTR (Wang et al., 2023b)	WACV'23	<b>96.0</b>	<b>99.2</b>	<b>92.1</b>	<b>97.9</b>	<b>80.3</b>	<b>93.1</b>	<b>84.3</b>	<b>96.4</b>	<b>70.0</b>	<b>87.8</b>	<b>62.8</b>	<b>79.5</b>	-	-
SUP-VPR (Ours)	-	<b>96.3</b>	<b>99.3</b>	91.6	<b>98.0</b>	80.1	92.4	<b>84.7</b>	<b>96.4</b>	<b>70.3</b>	<b>88.0</b>	<b>63.6</b>	<b>80.1</b>	<b>65.2</b>	<b>88.7</b>

to fuse image features as it is more stable to train. During the training phase of the MixVPR-like fusion module, we adopt a step-wise decay learning rate scheduler with a decay step of 500 and a decay rate of 0.95. During the training phase for SUP-VPR, we extend the training phase to seven epochs. For the evaluation phase, we utilize the image embeddings produced by the MixVPR-like fusion module to perform intra-city retrieval, while we utilize the image embeddings produced by the inter-city retrieval stage to perform city-level retrieval. For the evaluation metrics, we utilize Recall at N (R@N) with  $N \in \{1, 5\}$  and Top-5 Recall at 1m (R1m) as the evaluation metrics.

**Hyper-parameters.** We use a temperature factor  $\tau$  of 0.1 for Eq. 3. For the adaptive hard negative mining strategy, we set the number of negative samples to 5, which is sampled from the top-5 closest images from other cities according to the similarity with the database counterpart. For the feature sequence extracted by the DINOv2 backbone, we use the grid size of  $16 \times 16$  as the default setting and extract  $256 \times 256$  image embeddings. For the MixVPR-like fusion module, we use the number of layers  $K$  as 4, which is consistent with that used in MixVPR (Ali-bey et al., 2023). We use PyTorch’s built-in data augmentation (e.g., random horizontal flip, random crop, etc.).

#### 4.3 IMPACT OF IMAGE PAIRING

As shown in Tab. 2, we investigate the impact of using image pairs of different classes during the training phase. Specifically, we randomly sample the image pairs with different classes from different cities, while the image pairs with the same class are randomly sampled from the same city. The results show that using image pairs of different classes can significantly improve the performance of VPR. Especially for city-level retrieval, utilizing image pairs with different classes can improve the performance by over 10% for Svoboda and Boston. The main idea is to make the model more discriminative against other city databases.

Here we show more visualization results to further understand the impact of image pairs with different classes. Fig. ?? shows the attention maps of the image pairs with the same class and the image pairs with different classes. We observe that the attention maps of the image pairs with the same class mainly focus on the global context regions of images, which cannot effectively represent the discriminative regions. In contrast, the attention maps of the image pairs with the different classes can better represent the most discriminative regions. Since the data is more discriminative against other cities during training, the model can perform better for city-level retrieval.

#### 4.4 IMPACT OF CROSS-ATTENTION

As shown in Tab. 3, we evaluate the impact of utilizing the proposed cross-attention mechanism on the performance of VPR. The results show that utilizing the proposed cross-attention mechanism can significantly improve the performance of inter-city retrieval. Especially for the challenging Boston dataset and the large-scale MSSL dataset, we can observe more significant improvements. The main idea is to enhance the capability of the model to model long-range dependencies.

Here we show more visualization results to further understand the impact of the proposed cross-attention mechanism. Fig. ?? shows the attention map of the proposed attention block and the down-scaled self-attention block from ATTR (Wang et al., 2023b). We observe that the attention maps of the down-scaled self-attention block mainly focus on the foreground regions, which cannot accurately represent the correspondence between the query image and the database counterpart, as well as the correspondence between the query image and the other cities. In contrast, the attention maps of the proposed attention block can focus on the most discriminative regions.

#### 4.5 IMPACT OF HARD NEGATIVE MINING

As shown in Tab. 4, we evaluate the impact of utilizing the proposed hard negative mining strategy on the performance of VPR. The results show that utilizing the proposed hard negative mining strategy can significantly improve the performance of VPR. Especially for city-level retrieval, we can observe more significant improvements. The hard negative mining strategy makes the model more discriminative against other city databases.

#### 4.6 COMPARISON WITH STATE-OF-THE-ART METHODS

As shown in Tab. 5, we compare the proposed SUP-VPR method with other state-of-the-art methods on popular VPR benchmarks. Note that we divide the methods into two groups: using the pre-trained model (e.g., ResNet101 (He et al., 2016) and DINOv2 (Oquab et al., 2023)) trained on large-scale datasets, and only using the data from the VPR datasets. The results show that our proposed SUP-VPR performs favorably compared to the recently proposed transformer-based methods, which are usually computationally heavy. Furthermore, our proposed SUP-VPR shows better favorably for those large-scale datasets. The main idea is to utilize the compact and robust descriptors produced by SUP-VPR to enhance the VPR performance.

**MSLS.** We observe that our proposed SUP-VPR can achieve better R@1 than MixVPR (Ali-bey et al., 2023) by 0.8%, and the second best is ATTR (Wang et al., 2023b) with -2.8%. For R@5, our method is the best with a 1.3% margin compared to ATTR. For R@1m, our method is the best with a 3.5% margin compared to ATTR. The promising results show the effectiveness of our method on those large-scale datasets. The main idea is to enhance the generalization ability of the model to other city databases.

### 5 CONCLUSIONS

In this paper, we propose SUP-VPR, a novel two-stage transformer-based framework designed for VPR in low-data regimes. To achieve this, we propose a MixVPR-like fusion module for intra-city dataset retrieval and a cross-attention mechanism for inter-city dataset retrieval. Furthermore, an adaptive hard negative mining strategy is proposed to make the model more discriminative against other city databases. Our extensive experiments on seven challenging VPR datasets under both re-ranked and non-re-ranked settings demonstrate the effectiveness of our method for place recognition in low-data regimes. We believe that this work reveals the potential of transformer-based VPR under practical and realistic scenarios, which opens up new avenues for future research.

### REFERENCES

Marín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

- Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Global proxy-based hard mining for visual place recognition. In *British Machine Vision Conference (BMVC)*, 2022a.
- Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. GSV-CITIES: Toward Appropriate Supervised Visual Place Recognition. *Neurocomputing*, 513:194–203, 2022b.
- Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. MixVPR: Feature mixing for visual place recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2998–3007, 2023.
- Relja Arandjelovic and Andrew Zisserman. All about VLAD. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1578–1585, 2013.
- Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.
- Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1269–1277, 2015.
- Giovanni Barbarani, Mohamad Mostafa, Hajali Bayramov, Gabriele Trivigno, Gabriele Berton, Carlo Masone, and Barbara Caputo. Are local features all you need for cross-domain visual place recognition? In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154–6164, 2023.
- Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pp. 404–417, 2006.
- Gabriele Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2918–2927, January 2021.
- Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4878–4888, 2022.
- Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11080–11090, 2023.
- Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision (ECCV)*, pp. 726–743, 2020.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.
- Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 357–366, 2021.
- Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 224–236, 2018.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2020.
- Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

- Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European Conference on Computer Vision (ECCV)*, pp. 369–386, 2020.
- Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14152, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Qingfeng Hou, Jun Lu, Haitao Guo, Xiangyun Liu, Zhihui Gong, Kun Zhu, and Yifan Ping. Feature relation guided cross-view image based geo-localization. *Remote Sensing*, 15(20):5029, 2023.
- Gaoshuang Huang, Yang Zhou, Xiaofei Hu, Chenglong Zhang, Luying Zhao, Wenjian Gan, and Mingbo Hou. Dino-mix: Enhancing visual place recognition with foundational vision model and feature mixing. *arXiv preprint arXiv:2311.00230*, 2023.
- Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2011.
- Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatayallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *arXiv preprint arXiv:2308.00688*, 2023.
- Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3251–3260, 2017.
- María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23487–23496, 2023.
- Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic attraction-repulsion embedding for large scale image localization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2570–2579, 2019.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.
- David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.
- Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9: 19516–19547, 2021.
- Michael Milford and G. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24:1038–1053, 2008.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 8026–8037, 2019.
- James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

- Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to sift or surf. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2564–2571, 2011.
- Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In *Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013.
- Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 883–890, 2013.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13648–13657, 2022.
- Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022–5030, 2019.
- Yuwei Wang, Yuanying Qiu, Peitao Cheng, and Junyu Zhang. Hybrid CNN-transformer features for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1109–1122, 2023a.
- Yuwei Wang, Yuanying Qiu, Peitao Cheng, and Junyu Zhang. Transformer-based descriptors with fine-grained region supervisions for visual place recognition. *Knowledge-Based Systems*, 280: 110993, 2023b.
- Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2626–2635, 2020.
- Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. *arXiv preprint arXiv:2203.16291*, 2022.
- Jun Yu, Chaoyang Zhu, Jian Zhang, Qingming Huang, and Dacheng Tao. Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):661–674, 2019.
- Hao Zhang, Xin Chen, Heming Jing, Yingbin Zheng, Yuan Wu, and Cheng Jin. ETR: An efficient transformer for re-ranking in visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5665–5674, January 2023a.
- Jian Zhang, Yunyin Cao, and Qun Wu. Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recognition*, 116:107952, 2021a.
- Qieshi Zhang, Zhenyu Xu, Yuhang Kang, Fusheng Hao, Ziliang Ren, and Jun Cheng. Distilled representation using patch-based local-to-global similarity strategy for visual place recognition. *Knowledge-Based Systems*, 280:111015, 2023b.
- Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 113:107760, 2021b.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6881–6890, 2021.

Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2Former: Unified retrieval and reranking transformer for place recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19370–19380, 2023.

Generated by WhizResearch