Towards Implicit Aggregation: Robust Image Representation for Place Recognition in the Transformer Era

Feng Lu[1,2*] **Tong Jin**[3,4*] **Canming Ye**[1] **Yunpeng Liu**[3†] **Xiangyuan Lan**[2,5†] **Chun Yuan**[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Pengcheng Laboratory [3]Shenyang Institute of Automation, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences [5]Pazhou Laboratory (Huangpu)
lufengrv@gmail.com jintong@sia.cn ycm24@mails.tsinghua.edu.cn
ypliu@sia.cn lanxy@pcl.ac.cn yuanc@sz.tsinghua.edu.cn

Abstract

Visual place recognition (VPR) is typically regarded as a specific image retrieval task, whose core lies in representing images as global descriptors. Over the past decade, dominant VPR methods (*e.g.*, NetVLAD) have followed a paradigm that first extracts the patch features/tokens of the input image using a backbone, and then aggregates these patch features into a global descriptor via an aggregator. This backbone-plus-aggregator paradigm has achieved overwhelming dominance in the CNN era and remains widely used in transformer-based models. In this paper, however, we argue that a dedicated aggregator is not necessary in the transformer era, that is, we can obtain robust global descriptors only with the backbone. Specifically, we introduce some learnable aggregation tokens, which are prepended to the patch tokens before a particular transformer block. All these tokens will be jointly processed and interact globally via the intrinsic self-attention mechanism, implicitly aggregating useful information within the patch tokens to the aggregation tokens. Finally, we only take these aggregation tokens from the last output tokens and concatenate them as the global representation. Although implicit aggregation can provide robust global descriptors in an extremely simple manner, where and how to insert additional tokens, as well as the initialization of tokens, remains an open issue worthy of further exploration. To this end, we also propose the optimal token insertion strategy and token initialization method derived from empirical studies. Experimental results show that our method outperforms state-of-the-art methods on several VPR datasets with higher efficiency and ranks 1st on the MSLS challenge leaderboard. The code is available at `https://github.com/lu-feng/image`.

1 Introduction

Visual place recognition (VPR) involves identifying the coarse geographical location of a query place image by retrieving the most similar images from a geo-tagged database captured at previously visited places [46]. It is a fundamental and essential task in a wide range of computer vision and robotics applications, *e.g.*, augmented reality [53], autonomous driving [21], and SLAM [15]. Thus, it has garnered significant attention and study. Despite recent advances, there still exist some challenges in VPR, including condition variations, viewpoint changes, and perceptual aliasing (images from different places showing high similarity) [46], etc.

*Equal contribution.
†Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Typically, VPR is formulated as an image retrieval problem. For a given query image and a database, all place images are represented using global features, and the nearest neighbor search is conducted in this feature space to get the target place images that best match the query. The global features are usually obtained by employing aggregation methods (*e.g.*, VLAD [35]) to process local features. With the advancement of deep learning, most VPR methods have used a convolutional neural network (CNN) [31] or vision transformer (ViT) [23] as the backbone to extract local (patch) features. Meanwhile, NetVLAD [4] and GeM pooling [56] have become the most popular aggregation methods for aggregating local features to yield global descriptors, which are generally robust against common visual variations. Following this paradigm, some recent studies proposed more aggregation methods (*e.g.*, MixVPR [2], SALAD [34], CricaVPR [49], BoQ [3], and EDTformer [37]), trying to make the global features condition- and viewpoint-invariant, thereby achieving a promising performance.

Although this backbone-plus-aggregator VPR paradigm to obtain global features has become the de-facto standard [10] in the CNN era, it has some potential issues. First, the two-stage process (feature extraction + aggregation) may lead to unnecessary structural complexity and redundancy. Second, the one-shot aggregation of patch features implemented by the aggregator offers no opportunity for correction and refinement. Regarding specific aggregation methods (aggregators), there may exist some particular issues, such as the loss of position information of original patch features in NetVLAD [4]. Designing a perfect aggregator artificially is highly challenging. However, in light of the nature of transformer-based backbones, which are capable of modeling global contextual information and long-range dependencies [24], we argue that it is no longer necessary to design an aggregator separately. Instead, we can leverage the intrinsic self-attention mechanism within the backbone to implicitly aggregate useful information from patch tokens, thereby eliminating the need for an extra aggregator. In fact, previous work BoQ [3] also attempted to utilize self- and cross-attention mechanisms to aggregate useful information, yet it still introduced an extra aggregator that includes encoder blocks and cross-attention layers, as well as a large number of learnable queries. Another study [18] indicated that simply adding some registers, similar to concatenating the class token with patch tokens, can buffer excess global information (so-called "undesirable artifacts") into these registers. Unfortunately, it discarded these registers finally and lacked deeper research on the use of global information on registers.

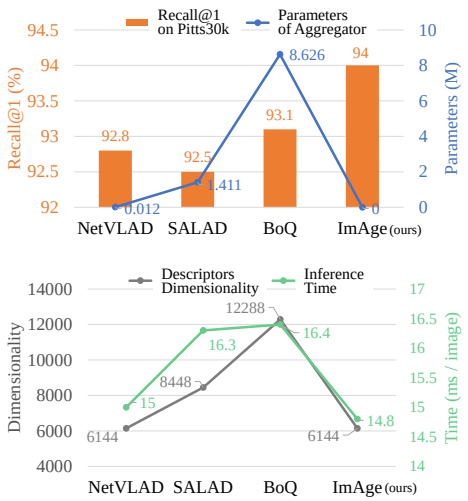

Figure 1: Comparison of three explicit aggregation methods and our ImAge. All methods use DINOv2-base-register as the backbone and are trained on the GSV-Cities dataset. ImAge achieves the best Recall@1 with the smallest descriptor dimension and the lowest inference time. Meanwhile, there is no extra explicit aggregator in our ImAge model.

In view of the issues of the previous explicit aggregation paradigm and the potential implicit aggregation ability of the transformer backbone itself, in this paper, we systematically explore the **Im**plicit **Agg**r**e**gation (abbreviated as ImAge) method, *i.e.*, unify feature extraction and aggregation solely via the backbone for VPR. Specifically, we introduce some learnable aggregation tokens, which are prepended to the patch tokens before a particular transformer encoder block. All these tokens will be jointly processed by the subsequent blocks and interact via the intrinsic self-attention mechanism, thus transmitting useful information within the patch tokens to our aggregation tokens. Finally, we only take aggregation tokens from the output of the last block and concatenate them to serve as the global descriptor, thereby achieving implicit aggregation. The proposed VPR paradigm provides a novel perspective different from the previous paradigm, unifying feature extraction and aggregation into a more cohesive framework. This further enables progressive aggregation in cascaded transformer blocks (rather than one-shot aggregation by a separate aggregator), thus achieving the correction and refinement of global image representations (*i.e.*, our aggregation tokens). Moreover, where and how to add aggregation tokens, as well as the initialization of these tokens, significantly impact performance. To this end, we propose an optimal token insertion strategy and token initialization method to effectively and efficiently yield more robust image representations and thus achieve excellent VPR performance. Our ImAge brings the following **contributions**:

1) We propose an implicit aggregation method to produce robust VPR image representations, which neither modifies the backbone nor needs an extra aggregator. It only adds some aggregation tokens before a specific block of the transformer backbone, leveraging the inherent self-attention mechanism to implicitly aggregate patch features. Our method provides a novel perspective different from the previous paradigm, effectively and efficiently achieving better performance in the transformer era.

2) To further improve the performance and efficiency of our ImAge, we propose: a) an aggregation token insertion strategy that deliberately delays token insertion until a specific transformer block, where patch tokens have acquired sufficient representation capability; b) a token initialization method that uses the L2-normalized cluster centers yielded by the k-means method to initialize added tokens.

3) Extensive experiments show that our ImAge significantly outperforms the latest explicit aggregation methods (*e.g.*, SALAD and BoQ) with the same setup (see Fig. 1). Besides, our method also achieves state-of-the-art (SOTA) results (*e.g.*, ranks 1st on MSLS challenge leaderboard) with high efficiency.

2 Related Work

Visual Place Recognition: Early research on VPR primarily focused on aggregating the hand-crafted descriptors [7] to global descriptors using some classical aggregation algorithms, such as Bags of Words [60] and VLAD [35, 45, 64, 5, 41]. In light of the remarkable achievements of deep learning across numerous computer vision tasks, contemporary VPR approaches [63, 4, 38, 17, 54, 26, 27, 69, 70, 28, 11, 44, 22] have increasingly utilized diverse deep features for better performance. Besides, traditional aggregation algorithms are gradually replaced by trainable aggregation layers, *e.g.*, NetVLAD [4] and GeM pooling [56]. Although some methods [30, 9, 59, 47, 50] employ local feature matching for re-ranking after initial global feature retrieval to boost performance, the backbone-plus-aggregator paradigm has been the de-facto standard [10] in VPR over the past decade. Some recent research [8, 1, 2, 34, 3, 49] has proposed several alternative approaches following this paradigm. For instance, CricaVPR [49] leveraged a cross-image encoder to produce cross-image correlation-aware global representations. SALAD [34] redefined the soft assignment in NetVLAD as an optimal transport problem and used the Sinkhorn algorithm to solve it. BoQ [3] employed distinct learnable queries to probe the input features through cross-attention, facilitating better information aggregation. These methods achieved excellent results using the ViT-based foundation model DINOv2 as the backbone. Unlike these methods that meticulously design auxiliary aggregators to yield global features, our ImAge method presents a novel paradigm that only introduces some additional tokens to the transformer backbone to conduct implicit aggregation via the inherent self-attention mechanism in transformers, thus achieving a simpler architecture and more powerful performance.

Additional Tokens in Transformers: Popularized by BERT [20], integrating special tokens into the token sequence in transformers has been a promising design choice for various purposes. We group such extra tokens into 3 categories based on their functional roles. **1)** *Output-oriented tokens* are learnable anchors that collect information from patch tokens, whose output values are then transmitted as task-specific outputs, *e.g.*, the class tokens used in BERT [20] and ViT [23] for classification, as well as detection tokens in YOLOS [25] for object detection. **2)** *Prompt tokens* act as trainable continuous vectors that replace traditional discrete text prompts, efficiently guiding pretrained transformer models to adapt to specific tasks by adjusting the model input, without modifying the parameters of models [43, 42, 36], which has become an essential branch of parameter-efficient fine-tuning methods [32]. **3)** *Memory tokens* act as registers that hold intermediate states during sequential processing steps, tracing their roots to neural memory architectures [14, 13]. This approach gains critical support from the DINOv2-register work [18], which observed that vision transformers improperly re-purpose background patch tokens as implicit memories when the standard class token lacks the capacity to accommodate global semantics. To address this, they prepend multiple memory tokens called registers to input tokens, which provide extra storage for buffering of global context, thus eliminating artifacts. Inspired by this work, we introduce the concept of **aggregation tokens** to effectively absorb global context from patch tokens. However, register tokens are discarded from the final output after temporary use, contrasting with our aggregation tokens that directly form the output descriptor for VPR (*i.e.*, our method falls into the "output-oriented tokens" category). Among VPR methods, BoQ [3] also advocated the introduction of a bag of output-oriented tokens named queries for aggregation (but in the aggregator rather than backbone). While effective, BoQ uses extra encoder blocks and cross-attention layers as the aggregator. In contrast, our method directly employs the inherent self-attention mechanism of the backbone, offering unique advantages.

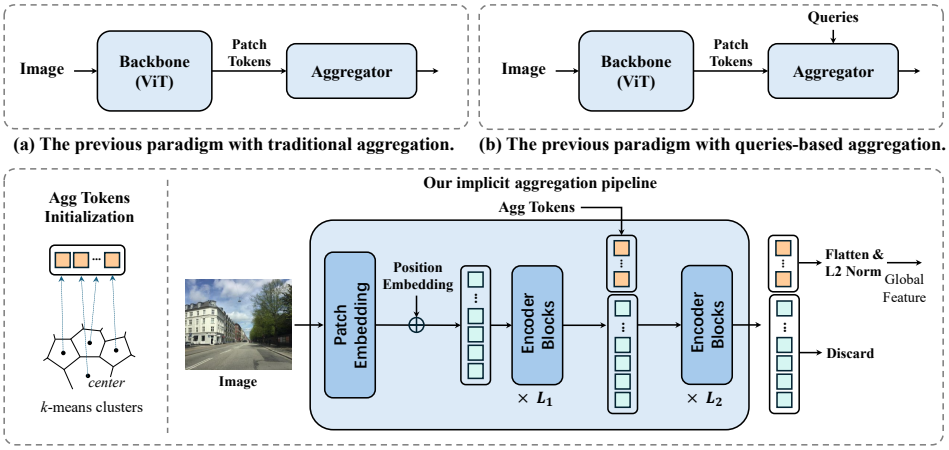

Figure 2: Illustration of the previous paradigm and our ImAge paradigm. (a) The backbone-plus-aggregator paradigm with the traditional aggregator. (b) The backbone-plus-aggregator paradigm with a queries-based aggregator that introduces some queries to learn global information from the patch tokens. (c) Our ImAge only prepends a set of aggregation tokens to the patch tokens before a specific block in transformer backbone, making them interact globally via self-attention to achieve implicit aggregation. Notably, these aggregation tokens are simply initialized by the k-means algorithm.

3 Methodology

This section begins with a review of the ViT [23] and the self-attention mechanism in it, which serves as the foundation for our ImAge method. Following that, we first present the pipeline of our method. Then, we introduce the insertion strategy of our aggregation tokens and their initialization method.

3.1 Preliminary

ViT [23] and its variants have rapidly emerged as the preferred backbones for a variety of computer vision tasks [71, 74, 73, 72, 50], owing to their exceptional capacity for modeling global relationships [58]. Given an input image of size $H \times W$, ViT partitions it into $N = HW/P^2$ non-overlapping patches. Each patch is then flattened and linearly projected to create a D-dimensional token x_p^i. A learnable class token $x_{\mathsf{CLS}} \in \mathbb{R}^D$ is prepended to this sequence, and positional embeddings are added to encode spatial information, forming the initial input token sequence $z_0 = [x_{\mathsf{CLS}}, x_p^1, \ldots, x_p^N] \in \mathbb{R}^{(N+1)\times D}$. This sequence is iteratively processed through L transformer encoder blocks. Each block comprises three core components: layer normalization (LN), multi-head self-attention (MHSA), and multi-layer perceptron (MLP). The l-th block updates the input z_{l-1} to z_l via

$$\begin{aligned} z'_l &= \mathrm{MHSA}\big(\mathrm{LN}(z_{l-1})\big) + z_{l-1}, \\ z_l &= \mathrm{MLP}\big(\mathrm{LN}(z'_l)\big) + z'_l. \end{aligned} \tag{1}$$

Within the MHSA module, the input sequence undergoes parallel linear transformations to generate h independent sets of queries Q, keys K, and values V, each parameterized by learnable projection matrices. For each attention head, the scaled dot-product attention

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\big(QK^\top/\sqrt{d}\big)\, V, \quad d = D/h, \tag{2}$$

computes context-aware similarity scores and dynamically aggregates information across all tokens. This mechanism facilitates rich cross-token interactions, where each token selectively assimilates features from others based on pairwise affinities. The outputs of all heads are concatenated to integrate multi-subspace representations and then linearly projected again, synthesizing position-wise updated embeddings z'_l that encode global contextual relationships. These properties of ViT indicate its potential to aggregate patch tokens by introducing additional tokens, which we will introduce below.

3.2 Implicit Aggregation via the Transformer Backbone

After extracting the patch features/tokens via the backbone, there are primarily two ways in previous works to obtain robust global descriptors. One is to directly aggregate these patch tokens with a common aggregator (*e.g.*, NetVLAD [4] and SALAD [34]), as in Fig. 2 (a). The other uses the queries-based aggregator to learn global information from the patch tokens (*e.g.*, BoQ [3] and EDTformer [37]), as in Fig. 2 (b). However, our ImAge will essentially eliminate the use of aggregators.

An overview of our ImAge is presented in Fig. 2 (c). Unlike existing VPR methods, ImAge removes the explicit aggregator and uses only the backbone network to achieve implicit feature aggregation. In this work, we utilize the vision transformer as the backbone, making the first L_1 encoder blocks process the patch tokens as usual. After these encoder blocks, a set of M learnable aggregation (agg) tokens, formulated as $a \in \mathbb{R}^{M \times D}$, is introduced and prepended to the other tokens z, getting a new sequence $[a, z]$. Then, these combined tokens will be uniformly processed by the subsequent L_2 encoder blocks and perform global interactions via the internal self-attention mechanism. Specifically, $[a, z]$ is first linearly transformed to produce the query $Q = [Q_a, Q_z]$, key $K = [K_a, K_z]$, and value $V = [V_a, V_z]$. Next, the interactions are computed according to Eq. 2 as follows:

$$Attn(Q, K, V) = [Q_a, Q_z][K_a, K_z]^\top [V_a, V_z] = [\underbrace{Q_a K_a^\top V_a}_{\text{Agg-Agg}} + \underbrace{Q_a K_z^\top V_z}_{\text{Agg-Patch}}, Q_z K_a^\top V_a + Q_z K_z^\top V_z], \tag{3}$$

where we omit the Softmax and scaling operations for simplicity. Based on Eq. 3, it is evident that the self-attention layers within the backbone enable us to achieve two key objectives: 1) Agg tokens can focus on their own features by agg-agg attention, thereby enhancing their intrinsic representation capabilities; 2) More importantly, agg tokens can fully learn and capture the global contextual information within the patch tokens by agg-patch attention, thus achieving robust implicit aggregation. Finally, we take the agg tokens from the output of the last encoder block, which are flattened into a vector and L2-normalized to form the final global image representation. It is worth noting that in the previous backbone-plus-aggregator paradigm, the global image representation is formed after one-shot aggregation of patch features implemented by the aggregator and is immediately output (without opportunity for refinement). Our method, however, adds agg tokens before a specific block of the transformer backbone. These agg tokens serve as global representations, and they are subsequently corrected and refined in subsequent blocks (synchronously with the refinement of patch tokens), rather than being aggregated/yielded all at once. This is an advantage over the previous paradigm.

Obviously, our ImAge fundamentally diverges from the practices of prompt tuning (aim to fine-tune models) [42] and register tokens (aim to remove artifacts) [18], which discard the newly added tokens finally. Besides, our method also differs from the class token. Our agg tokens have better scalability, along with different insertion strategies and initialization methods, which will be described below.

3.3 The Insertion Strategy of Aggregation Tokens

Our implicit aggregation method provides a robust image representation for VPR in an extremely simple manner. It requires neither explicit aggregators nor any modifications to the original backbone. However, where and how to add our agg tokens remains an open issue worthy of further exploration. For instance, previous works such as prompt tuning and DINOv2-register prepend additional tokens to the patch tokens (and class token) before the first transformer block, as

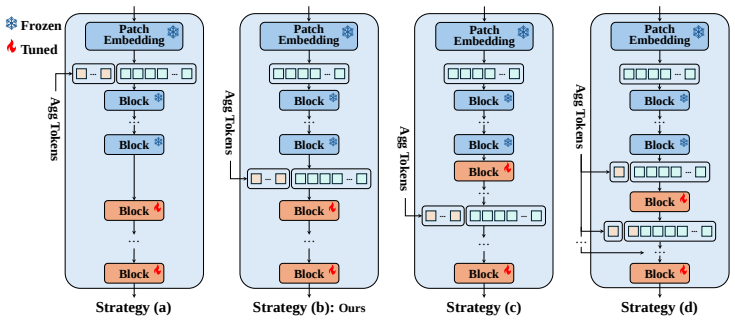

Figure 3: Illustration of 4 insertion strategies for agg tokens. (a) Agg tokens are added before all transformer blocks. (b) Agg tokens are added at the junction between frozen and trainable blocks (our strategy). (c) Agg tokens are added at a deeper tunable block. (d) Agg tokens are added incrementally across multiple blocks rather than all at once.

shown in Fig. 3 (a). Our objective differs from these works, and we no longer follow this way regarding the specific placement of agg tokens. More specifically, there are two reasons for this: 1) Our goal is to aggregate patch tokens with meaningful representations. Since early transformer blocks produce relatively weak features, adding agg tokens at the beginning is unnecessary and potentially detrimental to model performance. 2) In the field of VPR, the common practice for model training is to fine-tune only the last few blocks (layers) of the pre-trained model on the VPR dataset, while freezing the previous blocks. If agg tokens are added at the beginning, although most of the shallow and middle blocks are frozen, the added agg tokens need to be tuned. According to the chain rule of back-propagation [52], the gradients of the parameters in these frozen blocks still need to be calculated, leading to significant GPU memory and computational overhead.

In light of the above considerations, our strategy is to prepend agg tokens only when the patch tokens have acquired sufficient representational capability. A more specific criterion is to add the agg tokens at the junction between frozen and trainable transformer blocks, as illustrated in Fig. 3 (b). For example, in the case of the DINOv2 backbone, most previous VPR methods [34, 51] only fine-tune the last four blocks. Accordingly, we prepend the agg tokens to the patch tokens before the fourth-to-last block. Since the preceding blocks are frozen, it indicates that the features output here are general enough. The subsequent blocks are then trained on the VPR dataset to produce features more suitable for the VPR task, so our agg tokens can also learn better task-specific global representations. Additionally, we consider two alternative strategies. One is to add agg tokens before a deeper trainable block, as shown in Fig. 3 (c). The other is to add agg tokens progressively instead of all at once, as shown in Fig. 3 (d). However, both ways reduce the opportunities for the refinement and correction of image representations, which leads to suboptimal performance. Based on these objective factors, we finally propose the aforementioned strategy (b) for the insertion of agg tokens.

3.4 The Initialization of Aggregation Tokens

The agg tokens are learnable parameters, and their initialization can significantly impact the model performance. Prior to training on VPR datasets, the model is typically pre-trained on large-scale datasets. As a result, the patch tokens output by the specific block of such a model already have good representational capabilities. If our agg token is inappropriately initialized and prepended to patch tokens, it will instead cause damage to the representation of patch tokens in the subsequent processing of the MHSA layer. So, proper initialization of the agg token is essential.

Fortunately, a similar issue has been discussed in NetVLAD [6]. This method determines k cluster centers (and the parameters of the assignment layer) through training. The residual statistics from patch features to cluster centers are used as the global representation. At the beginning, it also requires initializing k cluster centers and the soft-assignment layer through the unsupervised k-means algorithm to achieve good performance. Although our ImAge method uses the self-attention mechanism to perform implicit aggregation, its essence can be regarded as each added agg token representing a unique category (but not necessarily corresponding to an object category in human semantics, such as building or vegetation) that is helpful to VPR, similar to each cluster in NetVLAD. Therefore, we can learn from NetVLAD, using the k-means algorithm to perform unsupervised clustering for the initialization of agg tokens. Besides, NetVLAD uses L2-normalized cluster centers to initialize parameters (weight $\mathbf{w}$) in the assignment layer. Through our empirical research, the L2-normalized centers can reduce the impact of extreme cases and are more suitable for initializing agg tokens than the original centers, *i.e.*, it is our final method.

4 Experiments

4.1 Datasets and Performance Evaluation

Datasets. We conduct the experiments on several VPR benchmark datasets. These datasets exhibit various challenges, including viewpoint changes, condition variations, and the perceptual aliasing issue. Table 1 provides a summary of the main evaluation datasets. **MSLS** [68] is a particularly challenging dataset, in which images are taken from urban, suburban, and natural scenes, covering diverse visual changes. **Pitts30k** [65] extracted from Google Street View, mainly presents severe variations in viewpoint. **Tokyo24/7** [64] shows dramatic condition (light) changes. **Nordland** [62] is gathered across four seasons with a fixed perspective from the front of a train. Moreover, we also use

the Baidu Mall [61], SPED [16], Pitts250k [65], St. Lucia [29], and SVOX [12] datasets for a few supplementary experiments.

Performance Evaluation. We follow the previous work [8, 10] using the Recall@N (R@N) as the evaluation metric for recognition performance. R@N is the proportion of queries for which at least one of the top-N predicted images is within a threshold of ground truth. We set the threshold to 10 frames for Nordland and 25 meters for others, as in this benchmark [10].

Table 1: Summary of the main evaluation datasets.

Dataset	Description	Number	
		Database	Queries
Pitts30k	urban, panorama	10,000	6,816
MSLS-val	urban, suburban	18,871	740
MSLS-challenge	long-term	38,770	27,092
Tokyo24/7	urban, day/night	75,984	315
Nordland	natural, seasonal	27,592	27,592

4.2 Implementation Details

The experiments are conducted on the NVIDIA RTX A6000 GPU using PyTorch. We use DINOv2-base-register as the backbone and only fine-tune the last four transformer blocks with the previous layers frozen. The token dimension in backbone is 768, and the number of our aggregation tokens is 8, thus outputting 6144-dim global descriptors. The image resolution is 224×224 for training and 322×322 for inference, as in SALAD [34]. We employ the multi-similarity loss [67] for training, with hyperparameters set following the GSV-Cities work [1]. The model is trained using the Adam optimizer with an initial learning rate of 0.00005, halved every 3 epochs. Each training batch contains 120 places, with 4 images per place (*i.e.*, 480 images). Besides, we set the maximum epochs to 20.

4.3 Comparisons with State-of-the-Art Methods

This section shows the experimental comparison of our ImAge with SOTA methods, including 11 single-stage VPR methods: NetVLAD [4], SFRS [28], CosPlace [8], MixVPR [2], EigenPlaces [11], CricaVPR [49], SALAD [34], SALAD-CM [33], BoQ [3], SuperVLAD [51] and EDTformer [37], as well as 2 two-stage VPR methods (TransVPR [66] and SelaVPR [50]) that leverage local features for re-ranking. The latest studies, CricaVPR, SALAD, SALAD-CM, BoQ, SelaVPR, SuperVLAD, and EDTformer, all use the foundation model DINOv2 as the backbone to extract deep features and achieve SOTA results. Our method mainly adopts DINOv2-base-register in experiments. Additionally, Cosplace and EigenPlaces construct an extra large-scale dataset (SF-XL) for training. CircaVPR, SALAD, BoQ, and EDTformer are trained on the GSV-Cities dataset, while SALAD-CM combines GSV-Cities and MSLS-train for training. Our work further merges Pitts30k-train, MSLS-train, SF-XL, and GSV-Cities for training, following the process in SelaVPR++ [48]. Table 2 presents the comprehensive quantitative results. Moreover, to enable a fairer comparison among three leading aggregation methods (NetVLAD, SALAD, and BoQ) and our ImAge, we conduct a consistent comparison using the same setup (backbone, training data, image resolution), as shown in Table 3. The experiments using other transformer backbones (ViT and CLIP) are shown in Appendix D.

For the comprehensive comparison in Table 2: Compared to existing SOTA methods (*e.g.*, SALAD-CM, BoQ, and EDTformer), our ImAge removes the explicit aggregator and only uses the backbone to obtain robust global descriptors, thus achieving a promising performance. On Pitts30k, a benchmark known for its extreme viewpoint variations, EDTformer and BoQ achieve 93.4% and 93.7% R@1, respectively. In comparison, our ImAge achieves a notable 94.1% R@1, attaining a new level. This indicates that global descriptors produced by our ImAge are highly robust to viewpoint changes. SALAD-CM significantly outperforms other methods on the MSLS dataset, which presents greater challenges due to diverse visual changes and perceptual aliasing. Nevertheless, our ImAge method further advances recognition performance, achieving 94.5% R@1 on MSLS-val and 93.8% R@5 on MSLS-challenge (ranks 1st on the official leaderboard). On Tokyo24/7, which is characterized by severe illumination changes, our ImAge also achieves the best performance with 97.1% R@1. In addition to its competitive performance on urban and suburban datasets, our ImAge still performs well on natural image datasets suffering from seasonal variations. Specifically, ImAge achieves an almost perfect R@5 (*i.e.*, $>99.0\%$) on Nordland. Overall, compared with other SOTA methods, our ImAge delivers substantial performance improvements across diverse scenarios. More importantly, our method no longer relies on a dedicated aggregator to obtain such robust global features.

For the fairer comparison in Table 3: In this comparison, we use the same training dataset (GSV-Cities), backbone (DINOv2-base-register), and input image resolution (224×224 in training and

Table 2: Comprehensive comparison to existing SOTA VPR methods on multiple benchmark datasets. All methods follow the settings of their respective original works, so there are differences in the backbone, training set, image resolution, etc. The best results are highlighted in **bold** and the second are underlined. † CricaVPR and SuperVLAD use a cross-image encoder to correlate multiple images from the same place to achieve better performance on Pitts30k. They are not included in the comparison with others (on all datasets).

Method	Dim	Pitts30k			MSLS-val			MSLS-challenge			Tokyo24/7			Nordland		
		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
NetVLAD [4]	32768	81.9	91.2	93.7	53.1	66.5	71.1	35.1	47.4	51.7	60.6	68.9	74.6	6.4	10.1	12.5
SFRS [28]	4096	89.4	94.7	95.9	69.2	80.3	83.1	41.6	52.0	56.3	81.0	88.3	92.4	16.1	23.9	28.4
TransVPR [66]	/	89.0	94.9	96.2	86.8	91.2	92.4	63.9	74.0	77.5	79.0	82.2	85.1	63.5	68.5	70.2
CosPlace [8]	512	88.4	94.5	95.7	82.8	89.7	92.0	61.4	72.0	76.6	81.9	90.2	92.7	58.5	73.7	79.4
MixVPR [2]	4096	91.5	95.5	96.3	88.0	92.7	94.6	64.0	75.9	80.6	85.1	91.7	94.3	76.2	86.9	90.3
EigenPlaces [11]	2048	92.5	96.8	97.6	89.1	93.8	95.0	67.4	77.1	81.7	93.0	96.2	97.5	71.2	83.8	88.1
SelaVPR [50]	/	92.8	96.8	97.7	90.8	96.4	97.2	73.5	87.5	90.6	94.0	96.8	97.5	87.3	93.8	95.6
CricaVPR† [49]	4096	94.9†	97.3†	98.2†	90.0	95.4	96.4	69.0	82.1	85.7	93.0	97.5	98.1	90.7	96.3	97.6
SuperVLAD† [51]	3072	95.0†	97.4†	98.2†	92.2	96.6	97.4	75.3	86.8	89.9	95.2	97.8	98.1	91.0	96.4	97.7
SALAD [34]	8448	92.5	96.4	97.5	92.2	96.4	97.0	75.0	88.8	91.3	94.6	97.5	97.8	89.7	95.5	97.0
SALAD-CM [33]	8448	92.7	96.8	97.9	94.2	97.2	97.4	82.7	91.2	92.7	96.8	97.5	97.8	96.0	98.5	99.2
BoQ [3]	12288	93.7	97.1	97.9	93.8	96.8	97.0	79.0	90.3	92.0	96.5	97.8	**98.4**	90.6	96.0	97.5
EDTformer [37]	4096	93.4	97.0	97.9	92.0	96.6	97.2	78.4	89.8	91.9	**97.1**	**98.1**	**98.4**	88.3	95.3	97.0
ImAge (Ours)	6144	**94.1**	**97.3**	**98.1**	**94.5**	**97.3**	**98.0**	**84.5**	**93.8**	**95.4**	**97.1**	**98.1**	**98.4**	**97.7**	**99.3**	**99.6**

Table 3: Consistent comparison to SOTA VPR aggregation algorithms. *All methods consistently use the same backbone (DINOv2-base-register), training dataset (GSV-Cities), and image resolution.

Method	Dim	Param. in Aggre.	Inference Time (ms)	Pitts30k			MSLS-val			Tokyo24/7			Nordland		
				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*	6144	0.012 M	15.0	92.8	96.6	97.8	91.8	96.5	96.6	95.6	**98.1**	98.7	90.5	96.5	97.8
SALAD*	8448	1.411 M	16.3	92.5	96.6	97.5	92.6	96.6	97.0	95.6	97.5	**99.0**	86.5	93.6	95.7
BoQ*	12288	8.626 M	16.4	93.1	**97.2**	**98.0**	92.8	96.5	97.0	95.2	97.7	98.2	87.0	94.0	95.9
ImAge*	6144	0 M	14.8	**94.0**	**97.2**	**98.0**	**93.0**	**97.0**	**97.2**	**96.2**	**98.1**	98.4	**93.2**	**97.6**	**98.6**

Table 4: Consistent comparison to SOTA VPR aggregation algorithms on supplementary datasets. *All methods consistently use the same backbone (DINOv2-base-register), training dataset (GSV-Cities), and image resolution.

Method	Dim	Baidu Mall		SPED		Pitts250k		St. Lucia		SVOX-Night		SVOX-Sun	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD*	6144	69.8	82.5	91.1	94.9	95.6	98.5	**99.9**	99.9	97.0	98.9	97.7	99.2
SALAD*	8448	67.3	81.2	90.3	94.6	95.4	98.8	**99.9**	**100**	96.1	99.0	97.2	99.4
BoQ*	12288	65.6	79.2	90.3	**96.0**	95.6	98.9	**99.9**	**100**	97.4	**99.5**	97.4	99.3
ImAge*	6144	**70.6**	**83.8**	**91.6**	95.6	**96.5**	**99.1**	**99.9**	**100**	**97.6**	99.4	**98.0**	**99.5**

322×322 in inference) for all methods. It is worth mentioning that Fig. 1 has shown some of the results of Table 3. In summary, our ImAge achieves the best overall performance on all datasets with the smallest descriptor dimension, the fastest inference speed, and the fewest model parameters. Note that even considering the additional parameters brought by our agg tokens, it is only 0.006M, *i.e.*, half of NetVLAD (0.07% of BoQ). This further supports our statement that an elaborately designed aggregator is not indispensable in the transformer era for robust global descriptors.

Besides, we also conduct the consistent comparison experiments on some supplementary datasets, including Baidu Mall [61], SPED [16], Pitts250k [65], St. Lucia [29], and SVOX [12], and the results are shown in Table 4. Compared to the three SOTA explicit aggregation methods, our ImAge achieves the best R@1 performance on all supplementary datasets. In particular, on Baidu Mall, which is the only indoor dataset and exhibits a distinct visual distribution from the other outdoor datasets, our method achieves the best performance, outperforming NetVLAD, SALAD, and BoQ with 0.8%, 3.3%, and 5.0% absolute R@1 improvements, respectively. This demonstrates that the global descriptors produced by our ImAge method through implicit aggregation are not only highly robust against common visual changes but also exhibit superior generalization ability.

Figure 4: Qualitative results. In these four challenging scenarios (involving dynamic objects, severe viewpoint variations, condition changes, etc.), our proposed ImAge method consistently retrieves the correct results from the database, while other methods all return the wrong images.

Fig. 4 presents qualitative retrieval results, where the proposed ImAge consistently demonstrates high robustness in various extreme scenes. For example, the first three cases exhibit severe lighting changes, viewpoint variants, and seasonal transitions, respectively. Other methods often retrieve visually similar but actually incorrect results due to perceptual aliasing. However, our ImAge effectively addresses these challenges in VPR and successfully returns the right results.

4.4 Ablation Studies

In this section, we conduct a series of ablation studies on our ImAge. We uniformly use the DINOv2-base-register backbone and train models on GSV-Cities with the batch size set to 120 (as the experiment in Table 3). Unless stated otherwise, we only fine-tune the last four transformer blocks.

Effect of tokens insertion strategy. In Section 3.3, we discussed several strategies for adding agg tokens and proposed the optimal strategy. To validate its effectiveness, we conduct an ablation to compare different strategies. To be fair, we consistently add 8 agg tokens. Strategy (a) and (â) both add agg tokens before the first transformer block. The only difference is that all transformer blocks in (â) are trainable. Strategy (b) is our optimal strategy. Strategy (c) introduces agg tokens before the penultimate block. Strategy (d) progressively adds 2 agg tokens before each of the last four blocks. Results are presented in Table 5. Among them, (a) performs the worst, because the early frozen transformer blocks produce weak and less informative features for VPR, harming the agg tokens to effectively capture meaningful global information. The issue is mitigated in (â), which further confirms our hypothesis (*i.e.*, adding agg tokens before the first trainable blocks). However, (â) trains all blocks, which incurs a lot of computational overhead and damages the excellent transferability of foundation models, thus failing to get optimal results. When fine-tuning only the last four transformer blocks, our proposed strategy (b) consistently outperforms all alternatives on all datasets by a large margin. This is because the last four tunable blocks can produce more suitable features for the VPR task, so our agg tokens can fully learn task-specific global representations. Although (c) and (d) also show relatively competitive performance, the late or gradual addition of agg tokens provides fewer opportunities to interact with patch features, thus limiting their ability to learn better representations.

Effect of aggregation tokens initialization. To validate the effectiveness of our proposed initialization methods for agg tokens, we conduct an ablation study using four initialization strategies: zero initialization (*i.e.*, no initialization), normal distribution initialization (commonly used for the class token or register tokens initialization [18]), vanilla cluster centers (yielded by k-means) initialization, and L2-normalized cluster centers initialization (*i.e.*, ours). We consistently use 8 agg tokens and prepend them to the patch tokens before the fourth-to-last transformer block. The experimental results are presented in Table 6. Zero initialization produces uniform representations at the beginning, limiting (even harming) interaction between agg tokens and patch features, and hindering global context modeling. In contrast, normal initialization provides a better inductive bias during early training and introduces slight randomness into the agg tokens, which helps break symmetry to get better performance. However, both initialization methods lack any visual prior, forcing the agg tokens

Table 5: Comparison of different insertion strategies for agg tokens. The **strategy (b)** is ours.

Method	Pitts30k		MSLS-val		Nordland	
	R@1	R@5	R@1	R@5	R@1	R@5
Strategy (a)	88.5	94.1	83.6	90.9	40.4	56.2
Strategy (â)	92.6	96.9	92.0	96.6	89.0	95.6
Strategy (b)	**94.0**	**97.2**	**93.0**	**97.0**	**93.2**	**97.6**
Strategy (c)	93.2	97.1	92.2	96.5	88.1	95.0
Strategy (d)	93.3	97.1	92.4	96.6	90.3	96.4

Table 6: Comparison of different initializations for agg tokens. The **centers-L2N** is ours.

Method	Pitts30k		MSLS-val		Nordland	
	R@1	R@5	R@1	R@5	R@1	R@5
zero	92.1	96.6	89.6	95.1	68.9	82.7
normal_distrib	92.9	96.9	92.0	96.8	88.6	95.3
centers	93.5	96.9	92.6	96.9	91.7	97.0
centers-L2N	**94.0**	**97.2**	**93.0**	**97.0**	**93.2**	**97.6**

to learn the patterns relevant to VPR from scratch, which constrains their final performance. Initializing agg tokens with cluster centers can be viewed as injecting a data-driven prior. These centers, obtained via unsupervised clustering of descriptors from randomly sampled training images, tend to capture common visual patterns. Such initialization can facilitate agg tokens to learn meaningful global information and diminish useless elements. Moreover, L2-normalized cluster centers offer more robust initializations for agg tokens by mitigating the influence of outliers, thereby achieving the optimal performance on all datasets.

Effect of the number of aggregation tokens. In this subsection, we investigate the impact of the number of added agg tokens (and use the class token, *i.e.*, cls, as baseline). The agg tokens are all added before the fourth-to-last block, and the results are in Table 7. Even with a single agg token, ImAge demonstrates a clear advantage over the class token with the same dimensionality, notably achieving an 11.3% absolute R@1 improvement on Nordland. This proves the differences and advantages of our method compared with directly using the class token, as well as the excellent performance of our method with low-dimensional descriptors. Furthermore, performance consistently improves as the number of agg tokens increases, with the best results obtained using 8 agg tokens. This is because a moderate increase of agg tokens enables more sufficient interaction and finer aggregation from the patch tokens. However, when the number of agg tokens becomes excessively large (*e.g.*, 64), a noticeable decline is observed. This may be attributed to the global nature of self-attention, where an excessive number of agg tokens can interfere with the contextual information of patch tokens, thereby indirectly degrading their own representational capability. Thus, adding 8 agg tokens is a promising choice overall.

Table 7: Comparison with the ImAge ablated versions with different numbers of aggregation tokens.

Number	Pitts30k		MSLS-val		Nordland	
	R@1	R@5	R@1	R@5	R@1	R@5
cls	91.8	96.5	89.1	95.3	63.5	79.0
1	92.2	96.6	90.7	95.4	74.8	87.0
4	93.4	97.0	92.3	96.6	89.6	96.1
8	**94.0**	**97.2**	**93.0**	**97.0**	**93.2**	**97.6**
16	93.7	**97.2**	92.8	96.9	92.2	97.2
32	93.1	96.9	92.6	96.8	90.3	96.2
64	92.8	96.8	92.2	96.5	85.4	93.0

5 Conclusions

In this paper, we presented ImAge, an innovative paradigm that explores implicit aggregation with a transformer to produce robust global image representation for VPR. Our method only adds some aggregation tokens and leverages the inherent self-attention of the transformer backbone to implicitly aggregate patch features. It overcomes the limitations of the previous backbone-plus-aggregator paradigm in an extremely simple manner, which neither modifies the original backbone nor requires an extra aggregator. Moreover, we propose an aggregation token insertion strategy and a token initialization method for our ImAge method to further improve the performance and efficiency. Experimental results show that ImAge obviously outperforms the latest explicit aggregation methods with higher efficiency under the same setup and achieves SOTA results on common VPR datasets.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), National Natural Science Foundation of China (62402252, 62536003), and Guangdong High-Level Talent Programme (2024TQ08X283).

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022.

[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023.

[3] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17794–17803, June 2024.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[5] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[6] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, 2018.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[8] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022.

[9] Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *IEEE/CVF International Conference on Computer Vision*, pages 12169–12178, 2021.

[10] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022.

[11] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023.

[12] Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2918–2927, 2021.

[13] Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. 2022.

[14] Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.

[15] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[16] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE international conference on robotics and automation*, pages 3223–3230. IEEE, 2017.

[17] Zetao Chen, Fabiola Maffra, Inkyu Sa, and Margarita Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16. IEEE, 2017.

[18] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[21] Anh-Dzung Doan, Yasir Latif, Tat-Jun Chin, Yu Liu, Thanh-Toan Do, and Ian Reid. Scalable place recognition under appearance change for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9319–9328, 2019.

[22] Shuting Dong, Mingzhi Chen, Feng Lu, Hao Yu, Guanghao Li, Zhe Wu, Ming Tang, and Chun Yuan. Vpr-cloak: A first look at privacy cloak against visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7197–7208, 2025.

[23] 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. In *International Conference on Learning Representations*, 2020.

[24] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.

[25] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.

[26] Sourav Garg, Adam Jacobson, Swagat Kumar, and Michael Milford. Improving condition- and environment-invariant place recognition with semantic place categorization. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6863–6870. IEEE, 2017.

[27] Sourav Garg, Niko Suenderhauf, and Michael Milford. Don't look back: Robustifying place categorization for viewpoint- and condition-invariant place recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3645–3652, 2018.

[28] 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*, pages 369–386. Springer, 2020.

[29] Arren J Glover, William P Maddern, Michael J Milford, and Gordon F Wyeth. Fab-map + ratslam: Appearance-based slam for multiple times of day. In *2010 IEEE international conference on robotics and automation*, pages 3507–3512. IEEE, 2010.

[30] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021.

[31] 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*, pages 770–778, 2016.

[32] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[33] Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *European Conference on Computer Vision*, pages 240–257, 2024.

[34] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[35] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[36] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

[37] Tong Jin, Feng Lu, Shuyu Hu, Chun Yuan, and Yunpeng Liu. Edtformer: An efficient decoder transformer for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025.

[38] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2145, 2017.

[39] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *arXiv preprint arXiv:2308.00688*, 2023.

[40] Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[41] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Predicting good features for image geo-localization using per-bundle vlad. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1170–1178, 2015.

[42] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[43] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.

[44] Bingxi Liu, Yujie Fu, Feng Lu, Jinqiang Cui, Yihong Wu, and Hong Zhang. Npr: Nocturnal place recognition using nighttime translation in large-scale training procedures. *IEEE Journal of Selected Topics in Signal Processing*, 18(3):368–379, 2024.

[45] Stephanie Lowry and Henrik Andreasson. Lightweight, viewpoint-invariant visual place recognition in changing environments. *IEEE Robotics and Automation Letters*, 3(2):957–964, 2018.

[46] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *ieee transactions on robotics*, 32(1):1–19, 2015.

[47] Feng Lu, Shuting Dong, Lijun Zhang, Bingxi Liu, Xiangyuan Lan, Dongmei Jiang, and Chun Yuan. Deep homography estimation for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10341–10349, 2024.

[48] Feng Lu, Tong Jin, Xiangyuan Lan, Lijun Zhang, Yunpeng Liu, Yaowei Wang, and Chun Yuan. Selavpr++: Towards seamless adaptation of foundation models for efficient place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[49] Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[50] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[51] Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. In *Advances in Neural Information Processing Systems*, volume 37, pages 5789–5816, 2024.

[52] Otniel-Bogdan Mercea, Alexey Gritsenko, Cordelia Schmid, and Anurag Arnab. Time- memory- and parameter-efficient visual adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5536–5545, 2024.

[53] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 268–283. Springer, 2014.

[54] Tayyab Naseer, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2614–2620. IEEE, 2017.

[55] 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.

[56] 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.

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[58] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[59] Yanqing Shen, Sanping Zhou, Jingwen Fu, Ruotong Wang, Shitao Chen, and Nanning 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*, pages 11217–11226, 2023.

[60] Sivic and Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*, pages 1470–1477. IEEE, 2003.

[61] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7436–7444, 2017.

[62] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. page 2013, 2013.

[63] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4297–4304. IEEE, 2015.

[64] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015.

[65] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013.

[66] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022.

[67] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew 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*, pages 5022–5030, 2019.

[68] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2626–2635, 2020.

[69] Zhe Xin, Yinghao Cai, Tao Lu, Xiaoxia Xing, Shaojun Cai, Jixiang Zhang, Yiping Yang, and Yanqing Wang. Localizing discriminative visual landmarks for place recognition. In *2019 International conference on robotics and automation (ICRA)*, pages 5979–5985. IEEE, 2019.

[70] Peng Yin, Lingyun Xu, Xueqian Li, Chen Yin, Yingli Li, Rangaprasad Arun Srivatsan, Lu Li, Jianmin Ji, and Yuqing He. A multi-domain feature learning method for visual place recognition. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 319–324. IEEE, 2019.

[71] WU Yue, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025.

[72] Lijun Zhang, Kangkang Zhou, Feng Lu, Zhenghao Li, Xiaohu Shao, Xiang-Dong Zhou, and Yu Shi. Esmformer: Error-aware self-supervised transformer for multi-view 3d human pose estimation. *Pattern Recognition*, 158:110955, 2025.

[73] Lijun Zhang, Kangkang Zhou, Feng Lu, Xiang-Dong Zhou, and Yu Shi. Deep semantic graph transformer for multi-view 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7205–7214, 2024.

[74] Quan Zhang, Xiaoyu Liu, Wei Li, Hanting Chen, Junchao Liu, Jie Hu, Zhiwei Xiong, Chun Yuan, and Yunhe Wang. Distilling semantic priors from sam to efficient image restoration models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25409–25419, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction 1 of our paper clearly outline our contributions, which include the development of the Implicit Aggregation and its advantages over explicit aggregators like NetVLAD[4] and SALAD[34], providing a clear clarification of the scope and contributions of our work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our paper includes a dedicated "Limitations and Future Work" section B where we comprehensively discuss the limitations of our approach.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our paper only involves a small amount of theory, which has been described in the methodology section 3 and relevant references are also provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper provides detailed formulations and descriptions of our proposed algorithms called ImAge in the methodology section 3. We also provide implementation details in the experiments section 4.2. In addition, more details about datasets I and compared methods J are included in the appendix. What's more, open access to our code and model checkpoints will be provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The full codes and model checkpoints for reproducing our methods will be publicly available upon paper publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our experimental setup is detailed in sub-section 4.2, including key hyperparameters. Additional information on the compared methods can be found in apendix J. More importantly, detailed ablation studies 4.4 are performed to show how the design strategies of our methods were chosen.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not commonly used in the VPR research.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We presented our compute resources in sub-section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, it is.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discussed both the potential positive and negative societal impacts of the work in appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our model achieved state-of-the-art performance in the task of VPR, and while there is a small possibility of unintended and malicious use, our project is not at high risk since we do not release new datasets. We will include a reminder of the risks in the `README.md` for the upcoming release of our open-source project.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In the paper, we have accurately cited the original sources of the datasets in sections 4.1 and I. We respect the licenses of the referenced code and data and will properly acknowledge them in the project's `README.md`.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Codes and model checkpoints for reproducing our methods will be publicly available upon paper publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper neither involves crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: In this research, the development of the core methods 3 is entirely based on our own original thinking without any involvement of LLMs as important, original, or non-standard components. LLMs were only utilized for paper polishing purposes, which falls under the category of writing, editing, or formatting.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Broader Impacts

Visual place recognition plays an important role in several applications, including autonomous driving, augmented reality, and robot localization. Our work proposes an implicit aggregation method to produce robust image representation for VPR with the transformer-based backbone and shows SOTA performance. While our exploration on VPR remains fundamental and application-agnostic, the potential utilization of VPR technology for intrusive surveillance and social media monitoring raises some privacy issues. It is crucial to prevent the misuse of VPR research for detrimental purposes.

B Limitations and Future Work

While our study provides a novel insight (*i.e.*, implicit aggregation without an explicit aggregator) into achieving robust global image representation for VPR, we acknowledge three limitations of our work: **1)** Although our ImAge method demonstrates universality for the transformer-based models, compared to the foundation models pre-trained on the massive dataset (*e.g.*, DINOv2 pre-trained on the LVD-142M dataset [55]), the performance improvement is less pronounced when using backbones without pre-training on sufficiently massive data (*e.g.*, the ViT pre-trained only on ImageNet [19]). This will be shown in Appendix D. However, this also suggests that with the advancement of increasingly powerful foundation models, the superiority of our approach compared to existing VPR methods may become more prominent. **2)** The proposed method may not be a good choice when we want to keep the backbone frozen (like AnyLoc [39]) or when the backbone is extremely expensive to fine-tune. However, it is also worth noting that we only fine-tune the last few blocks of the transformer backbone in most cases, which is relatively cheap. **3)** Although this work focuses on the VPR task, we believe that the proposed ImAge is broadly applicable to a wide range of image (or other modalities) retrieval tasks. The potential of our ImAge method for more applications in the machine learning community needs to be further explored through more experiments in future work.

C More Details about the Relations & Differences to Other Methods

Our ImAge design draws inspiration from prior works, particularly DINOv2-register [18] and prompt tuning [42]. Both approaches introduce additional tokens to the transformer backbone before the first encoder block. However, their objectives and usage differ fundamentally from ours. Prompt Tuning aims to adapt a frozen model to downstream tasks by learning a set of prompt tokens in a parameter-efficient manner, while these tokens are typically excluded from the final representation. DINOv2-register introduces additional register tokens to mitigate artifacts in the feature maps. Although the study shows that these registers may capture certain global information, they are ultimately discarded, and only the patch and class tokens are used for downstream tasks. In contrast, our ImAge method introduces the aggregation (agg) tokens before a particular transformer block, and utilizes agg tokens from the output of the last block as the final global image representation, which provides a reverse perspective compared to these approaches. In addition, while the class token within the transformer backbone is sometimes used as a global representation, our ImAge differs in several key aspects and demonstrates superior scalability. First, the class token is typically introduced at the beginning of the transformer and starts to learn from shallow features, which may limit its flexibility and make it difficult to fully capture task-specific complex semantics. Second, the class token is a single fixed embedding, which inherently restricts its representational capacity. Although it may suffice for relatively simple classification tasks, it often proves inadequate for more complex scenarios requiring richer and more flexible representations. In contrast, our ImAge introduces agg tokens with customizable positions and quantities, allowing them to fully learn task-specific global features. Moreover, we also design a tokens initialization method based on the k-means algorithm, which is significantly different from other works. Additionally, among existing VPR methods, BoQ [3] also offers some valuable insights for our work. However, it elaborately designs an explicit aggregator consisting of additional encoder blocks and cross-attention layers, which aggregates global information from patch tokens into a set of extra learnable queries. In contrast, we focus on the transformer backbone itself and make use of the inherent self-attention mechanism. Our study reveals a new insight: the aggregation function, previously implemented through an exquisitely designed aggregator, already appears naturally in the transformer backbone. We demonstrate that, by adding just some additional tokens, we can fully develop this implicit and progressive aggregation behavior.

Table 8: Results of NetVLAD and our ImAge using CLIP (base version, only vision encoder) and ViT (base version) as backbone. All models are trained on the GSV-Cities dataset with the batch size equal to 120. The learning rate is 0.00006 for the CLIP-based model and 0.0003 for the ViT-based model. For ViT, the last two blocks are directly truncated and all other blocks are trainable, as in [10, 51]. For CLIP, we only train the last 6 blocks with the previous layers frozen. All methods produce 768*8-dimensional descriptors, *i.e.*, 8 clusters for NetVLAD and 8 aggregation tokens for ImAge, the same as in the main paper.

Method	Pitts30k			MSLS-val			Nordland		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-NetVLAD	90.6	**95.7**	**97.2**	87.2	**94.1**	94.6	60.6	**74.6**	**80.2**
CLIP-ImAge (Ours)	**91.2**	95.6	96.9	**88.2**	**94.1**	**95.5**	**61.0**	**74.6**	80.1
ViT-NetVLAD	90.1	95.3	96.4	82.4	90.7	93.0	52.1	67.6	74.1
ViT-ImAge (Ours)	**90.3**	**96.1**	**97.3**	**86.2**	**92.2**	**93.8**	**53.3**	**69.3**	**75.6**

Table 9: Results of CricaVPR and the CricaVPR boosted by ImAge (*i.e.*, CricaVPR+ImAge).

Method	Pitts30k			MSLS-val			Nordland		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CricaVPR	**94.9**	97.3	**98.2**	90.0	95.4	96.4	90.7	96.3	97.6
CricaVPR+ImAge	**94.9**	**97.5**	**98.2**	**92.0**	**97.2**	**97.3**	**94.1**	**97.9**	**98.7**

D Comparison to NetVLAD Using Other Transformer Backbones

In the main paper, we conduct experiments using the DINOv2 backbone to validate the effectiveness of our method. Notably, DINOv2 is a foundation model based on the ViT architecture and pre-trained on the large-scale curated LVD-142M dataset. However, our method is also applicable to other transformer models. To this end, we conduct additional experiments using the CLIP [57] model and a ViT model pre-trained only on ImageNet. The results are shown in Table 8. We observe that our proposed method, ImAge, consistently achieves higher R@1 performance across all datasets compared to the explicit aggregation method NetVLAD. However, the performance gains of ImAge are less pronounced (except for ViT-ImAge on MSLS-val) compared to using DINOv2-base-register as the backbone. This observation aligns with the prior study [40], which suggests that foundation models pre-trained on large-scale datasets (significantly larger than ImageNet) are more capable of utilizing additional tokens to capture global information. Moreover, using CLIP as the backbone yields significantly less improvement than DINOv2. Although CLIP is a widely used foundation model (*i.e.*, a vision-language model), its pre-training data and objectives differ considerably from those of the VPR task, making it not a promising choice. This is consistent with the prior work AnyLoc [39], which suggests that CLIP performs significantly worse than DINOv2 in outdoor VPR scenarios.

E Improving Other VPR Methods with ImAge

Since our ImAge is a general image representation method for VPR, it can not only be implemented based on different transformer backbones, but also can be combined with some other VPR methods to improve their performance. This section uses the CricaVPR [49] method as an example to conduct experiments, and the results are shown in Table 9. It can be seen that our method significantly improves the performance.

F The GPU Memory Usage and Computational Efficiency in Training

Our method does not add agg tokens before the first block of the transformer backbone, which can significantly reduce GPU memory usage and computational burden. Here, we not only compare our method with adding tokens before the first block, but also use NetVLAD and SALAD as baselines. The results are shown in Table 10. Our method has significant advantages over adding tokens before the first block in terms of GPU memory usage and training time, and also outperforms NetVLAD and SALAD.

Table 10: The comparison of training GPU memory usage and training time.

Method	Training GPU Memory (GB)	Training Time/Epoch (min)
NetVLAD	17.54	9.93
SALAD	21.81	9.98
Adding tokens before 1st block	34.00	15.12
ImAge (Ours)	**16.73**	**9.87**

G The Attention Visualization of Aggregation Tokens

Here we provide the visualization of attention weights of our agg tokens to other patch tokens, as in Fig. 5. This vividly demonstrates that our agg tokens can effectively focus on objects beneficial for VPR (*e.g.*, buildings and vegetation) while ignoring irrelevant or even detrimental elements (*e.g.*, sky and moving vehicles). Additionally, we can observe that: 1) Our method maintains consistent attention on key objects under significant illumination and seasonal changes, indicating high robustness. 2) The attention on critical objects is sparse rather than uniform, suggesting that typically only the most distinctive features need to be considered for VPR. Even for buildings, there is no need to focus on (aggregate) their full area. 3) Some agg tokens focus on both buildings and vegetation, and there are also multiple tokens that focus on buildings. Therefore, there is not a one-to-one correspondence between agg tokens and human-defined object categories.

H Additional Qualitative Results and Failure Cases

In this section, we provide additional qualitative results (*i.e.*, visual examples) as a supplement for Fig. 4 in the main paper. As shown in Fig. 6, our ImAge method demonstrates exceptional robustness in retrieving correct database images across various challenging scenarios, including seasonal/viewpoint/lighting variations and occlusions. In contrast to other methods that fail to distinguish critical landmarks or are misled by superficial similarities, the proposed ImAge accurately captures key features (*e.g.*, building textures, positional relationships) to identify right matches.

Moreover, Fig. 7 illustrates some representative failure cases. While our method achieves relatively close retrievals (within 50 meters) in ambiguous natural scenes without distinct landmarks, it occasionally exceeds the predefined threshold (*i.e.*, 25 meters) due to geographic proximity but insufficient visual discriminability. The fourth example, which is the most challenging, involves nighttime images with over-exposure and motion blur, where all methods (including ours) even fail to meet the 50-meter criterion, highlighting persistent challenges in low-quality visual conditions. These results underscore both the advancements of our approach and the remaining difficulties in VPR, which may require increasing the geographical density of image collection for the database to solve. Additionally, for the last two samples, SelaVPR based on local feature re-ranking obtains the correct results, while other methods (including ours) all fail. This points to a possible way to further enhance the robustness of our approach in the future.

I More Details about Datasets

The testing datasets used in our experiments, including Pitts30k, Pitts250k, Tokyo24/7, Nordland, SPED, St. Lucia, and SVOX, are organized following the Visual Geo-localization (VG) benchmark [10]. Notably, we use the official version MSLS dataset as in previous work [68, 2, 49, 3, 34]. This version of MSLS-val only consists of 740 query images, which is different from the version in the VG benchmark [10]. In addition, there are also several versions of the Nordland dataset in the VPR task. In our experiments, we use the version in the VG benchmark [10], which employs the summer images as the database and the winter images as queries, each containing 27592 images. Baidu Mall [61] is a well-known indoor dataset for image-based localization. All images are collected at a shopping mall that is over 5000 square meters with many challenging elements, such as transparent windows, reflective materials, repetitive structures, dynamic pedestrians, etc.

Moreover, in the comprehensive comparison (*i.e.*, Table 2 in main paper) with other SOTA methods, we merge Pitts30k-train, MSLS-train, SF-XL, and GSV-Cities for training, following the approach in SelaVPR++ [48]. Specifically, we process datasets other than GSV-Cities to divide places into a finite number of categories, thus facilitating fully supervised training with the multi-similarity loss [67].

Figure 5: The visualization of the attention weights of our agg tokens to patch tokens. The first column (a) represents the input images. The middle 2-5 columns (b) separately display the attention weights of a single agg token to all patch tokens (reshaped to restore spatial position), meaning each image shows the attention of only one agg token. The last column (c) shows the merged attention of all 8 agg tokens. The first five examples (*i.e.*, five rows) show five different places, with buildings, vegetation, and dynamic interference. While different agg tokens attend to distinct regions (or objects) in the images, they consistently focus on stable and discriminative areas (*e.g.*, buildings and vegetation), while largely ignoring variable elements (*e.g.*, cars). The sixth and seventh examples show two images taken at the same place in different seasons. Our agg tokens can consistently focus on buildings (and some discriminative regions where the terrain and railroad tracks change). The last two examples demonstrate that agg tokens can consistently focus on buildings and landmarks even after undergoing severe lighting changes.

J More Details about Compared Methods

In the main paper, we compare our method with several other approaches and briefly introduce them. Here, we provide more details about them (for the results in Table 2).

NetVLAD [4] and **SFRS** [28] both consist of a VGG16 backbone and a NetVLAD aggregator, and use Pitts30k as the training dataset. The latter employs self-supervised image-to-region similarities to mine hard positive samples for training a more robust model. In our experiments, we use their PyTorch implementations for comparison.

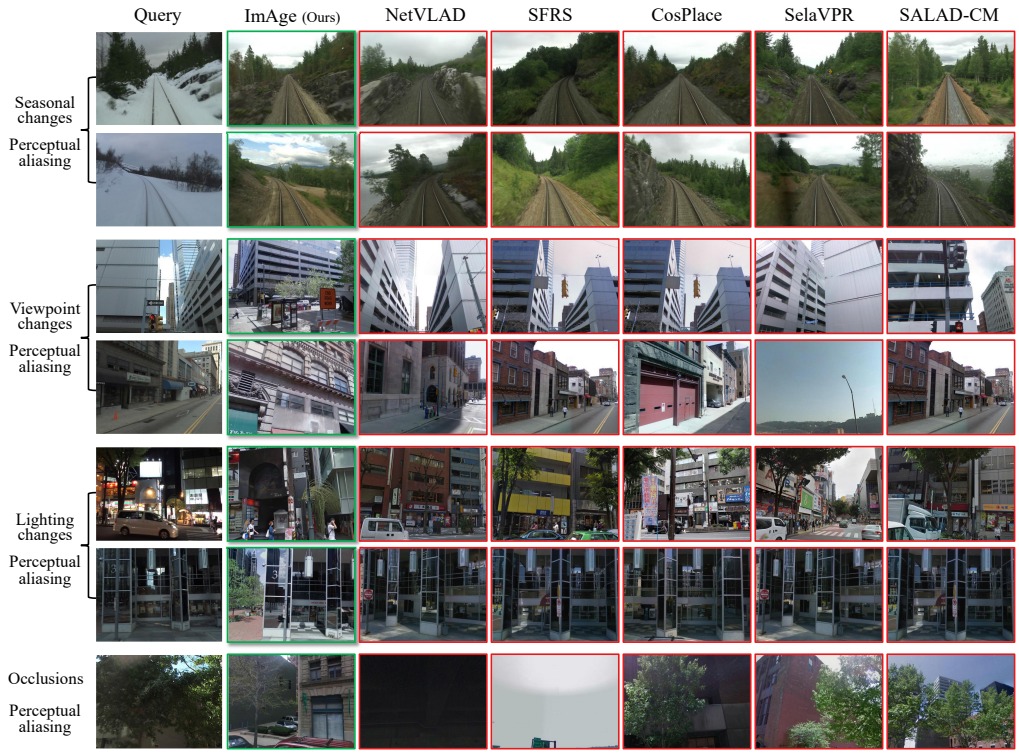

Figure 6: Qualitative results. In four challenging groups of examples (covering seasonal changes, viewpoint variations, lighting changes, occlusions, etc.), our ImAge successfully retrieves the correct database images, while other methods fail. In the first group, most other methods return incorrect images with landscapes absent from the query image (*e.g.*, lakes, cliffs, and hills) or railway tracks of contradicting shapes. The examples in the second group exhibit significant viewpoint variations, where our ImAge consistently gets the right results and demonstrates high robustness. In contrast, other methods return images that appear similar in viewpoint but are actually wrong. Still, they cannot distinguish the critical difference of the landmarks (*e.g.*, the texture of the buildings and their positional relationship). As for the third group, the dim nature of the query image likely interferes with the judgment of the other approaches, resulting in low-luminosity images with different buildings. Dynamic objects like cars in the first example query of this group are also misleading. Nevertheless, our method successfully caught the key features (*e.g.*, the texture of buildings). The final group shows a complex query with severe occlusions by a colossal tree. It is so difficult that all these methods except ours have crashed, returning perceptually similar but wrong images that are also extensively covered (by darkness, brightness, and trees). In summary, our ImAge method demonstrates unparalleled capacity to recognize the truly identical place against various perceptual variations.

CosPlace [8] and **EigenPlaces** [11] both frame VPR training as a classification task and use the SF-XL dataset to train their models. For Cosplace, we use the official model based on the VGG16 backbone (with the 512-dim output feature) for testing. For EigenPlaces, we utilize its official implementation and the best configuration based on the ResNet50 backbone to output 2048-dim global descriptors for comparison.

MixVPR [2] aggregates the deep features using the multi-layer perceptrons and trains the model with multi-similarity loss [67] on the GSV-Cities [1] dataset. We apply its best-performing configuration (ResNet50 with 4096-dim output features) for comparison.

CricaVPR [49], **SuperVLAD** [51], **SALAD** [34], **BoQ** [3], and **EDTformer** [37] all use the foundation model DINOv2-base [55] as the backbone to extract deep features, and train their models on GSV-Cities with the multi-similarity loss. In the comparison experiments, we consistently use their official implementations and the best configurations.

Figure 7: Failure cases. In the first three examples, our method retrieves database images that are geographically close to the query images. However, the distance (radius) between these retrieved images and the query images exceeds the predefined threshold (*i.e.*, 25 meters), although it remains below 50 meters. These cases (partially correct) are labeled in yellow. For the first two cases, ImAge tolerably retrieves results with distances of 33.11 meters and 27.32 meters, while other methods even fail to find an image within 50 meters. In the third example, the images are captured in natural scenes without discriminative landmarks. Nonetheless, ImAge can effectively exclude incorrect answers involving ponds and rivers, while some other methods fail to do so. The distance for this retrieval is a fair 35.70 meters, compared to other methods ranging from 161.90 meters to 4592.90 meters. In the fourth challenging example, all methods, including ours, fail to get an answer within 50 meters. This challenge arises from the complex lighting conditions at night, where over-exposure in bright areas, such as lights, affects the overall texture and the visibility of landmark details. For the last two challenging cases (involving large changes in viewpoint), all methods (including ours) that rely solely on global features for retrieval fail. Notably, SelaVPR, which is based on local features re-ranking, yields the right results. This provides a potential direction for further improving the accuracy of our method. In short, some challenges for current VPR methods remain, despite our method moving a step forward from others.

SALAD-CM [33] is an improvement of SALAD. This work analyzes the Geographic Distance Sensitivity of VPR embeddings and proposes a novel mining strategy to address it. Moreover, SALAD-CM first trains the model using both the GSV-Cities and MSLS datasets for better performance. In the comparison experiments, we follow its official implementation.

The rest **TransVPR** [66] and **SelaVPR** [50] are two-stage VPR methods. These works provide two models: one trained for testing on urban datasets (*e.g.*, Pitts30k and Tokyo24/7), and another trained for testing on datasets that may contain suburban and natural scenes (*e.g.*, MSLS and Nordland). We follow the usage in their original paper for comparison experiments.

K The Snapshot of MSLS Leaderboard

Fig. 8 is the snapshot of the MSLS place recognition challenge [68] leaderboard at the time of submission, and our ImAge method ranks 1st.

Results

#	User	Entries	Date of Last Entry	recall@5 ▲
1	ImAge4VPR	1	05/11/25	0.94 (1)
2	SelaVPRplusplus	3	01/31/25	0.94 (1)
3	anonymous456	9	03/02/25	0.94 (2)
4	amaralibey	1	07/07/24	0.90 (3)
5	mapillary_challenge	11	04/17/24	0.90 (3)
6	SKyxuan	16	07/04/24	0.90 (4)
7	anonymous123	9	07/08/24	0.90 (5)
8	ningzuotao	16	12/20/23	0.89 (6)
9	magnus	1	06/05/24	0.89 (6)
10	razor	1	06/05/24	0.89 (7)
11	izquierdo	25	11/15/23	0.89 (8)
12	anonymous02	3	09/17/23	0.89 (9)
13	uno	30	06/12/24	0.89 (9)
14	qixi	6	12/19/23	0.89 (9)
15	Pleiades	1	03/27/25	0.88 (10)

Figure 8: The snapshot of MSLS place recognition challenge leaderboard. Our ImAge method (named "ImAge4VPR" for double-blind policy) ranks 1st at the time of submission.