

Efficient Visual Place Recognition Through Multimodal Semantic Knowledge Integration

Sitao Zhang^{1*} Hongda Mao² Qingshuang Chen² Yelin Kim²
¹The Pennsylvania State University ²Amazon

Abstract

Visual place recognition is crucial for autonomous navigation and robotic mapping. Current methods struggle with perceptual aliasing and computational inefficiency. We present SemVPR, a novel approach integrating multimodal semantic knowledge into VPR. By leveraging a pre-trained vision-language model as a teacher during the training phase, SemVPR learns local visual and semantic descriptors simultaneously, effectively mitigating perceptual aliasing through semantic-aware aggregation without extra inference cost. The proposed nested descriptor learning strategy generates a series of ultra-compact global descriptors, reduced by approximately $66\times$ compared to state-of-the-art methods, in a coarse-to-fine manner, eliminating the need for offline dimensionality reduction or training multiple models. Extensive experiments across various VPR benchmarks demonstrate that SemVPR consistently outperforms state-of-the-art methods with significantly lower computational costs, rendering its feasibility for latency-sensitive scenarios in real-world applications.

1. Introduction

Visual place recognition (VPR) is a cornerstone of autonomous navigation and robotic mapping systems, enabling machines to identify and recall locations using visual cues from their surroundings [26, 34]. Despite significant research efforts, state-of-the-art VPR methods still struggle with the pervasive problem of perceptual aliasing, where distinct places are incorrectly matched due to similar low-level visual features. This limitation is exacerbated in challenging real-world scenarios involving occlusions, illumination variations, and other appearance changes, leading to failures that can severely compromise the reliability of autonomous systems [29, 46].

The root cause of this issue lies in the reliance on low-level visual features only, which inadvertently learn spurious patterns and fail to capture the true semantic essence

*Work done during an internship at Amazon.

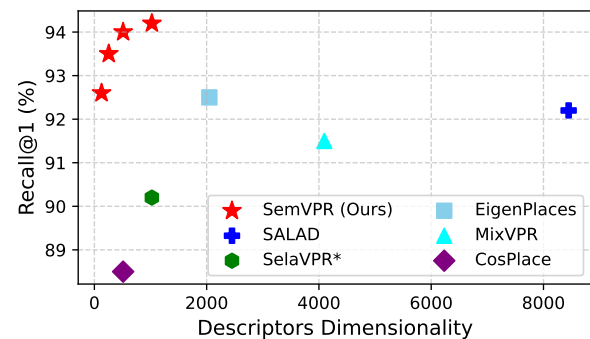


Figure 1. Comparison of Recall@1 and descriptor dimensionality among SOTA methods on Pitts30k. Two-stage methods are indicated with *, showing first-stage results for fair comparison. Our method achieves higher Recall@1 than other methods while maintaining highly compact global features down to 128 dimensions, incurring no additional inference overhead.

of a scene [11]. While recent approaches have attempted to mitigate perceptual aliasing through computationally expensive re-ranking steps [10, 12, 28, 33, 42, 47] or by employing high-dimensional descriptors [1, 2, 12, 18, 28], these solutions are inherently inefficient and may be infeasible for resource-constrained and time-critical applications.

In contrast, humans excel at VPR tasks by leveraging not only visual observations but also a rich understanding of the semantic content of the scene. Humans can effectively disregard transient elements like pedestrians and vehicles, identify reliable visual landmarks, and mitigate the effects of appearance changes by recognizing the underlying semantic similarities between locations. This ability is facilitated by our capacity to reason about abstract concepts, draw upon prior knowledge, and integrate multi-modal information seamlessly [20].

Inspired by this, we propose SemVPR, a novel approach that seamlessly integrates semantic knowledge into the VPR framework, aiming to address the limitations of current methods and deliver an efficient and robust solution for real-world applications. Our key insight is to leverage the complementary strengths of low-level visual patterns and high-

level semantic knowledge by guiding the feature learning and descriptor construction processes with semantic cues derived from state-of-the-art multi-modal models [32].

Specifically, we introduce a pre-trained vision-language model with rich common sense as a teacher during the training phase. While learning local visual descriptors, the model simultaneously learns a high-quality local semantic descriptor under the guidance of the teacher model. This semantic descriptor encodes the scene’s conceptual information, and then contributes to the aggregation of the global descriptor, enabling the model to concentrate on reliable visual landmarks and relevant attributes. In contrast to common unsupervised aggregation methods such as VLAD [18] and GeM [31], the introduction of semantic knowledge guidance allows our model to effectively mitigate perceptual aliasing by focusing on the most relevant and informative aspects of the scene.

Furthermore, we propose a nested descriptor training strategy that allows for learning highly compact global representations. By explicitly imposing constraints at multiple levels of the descriptor construction process, the model is encouraged to aggregate local features from relevant parts only. This yields a series of low-dimensional descriptors that are not only robust to appearance variations but also computationally efficient, making them well-suited for deployment in resource-constrained environments or time-critical applications.

Our approach promises to deliver efficient and reliable VPR capabilities, even in challenging scenarios where conventional methods falter, paving the way for more robust and trustworthy autonomous systems. We summarize our major contributions as follows:

- We introduce SemVPR, a VPR framework that integrates multimodal semantic knowledge to address perceptual aliasing. By using a pre-trained vision-language model as a teacher, SemVPR learns visual and semantic descriptors simultaneously at training time, enabling semantic-aware feature aggregation without additional inference costs.
- We introduce a nested descriptor learning strategy that generates a series of ultra-compact global descriptors, reduced by approximately $66\times$ compared to the SOTA method, in a coarse-to-fine manner, which eliminates the need for offline dimension reduction techniques such as PCA or the training of multiple models. To the best of our knowledge, this is the first time such a strategy has been explored for VPR, allowing for ultra-compact global descriptors down to 64 -dim that maintain high performance while significantly reducing computational overhead.
- SemVPR demonstrates state-of-the-art performance across various VPR benchmarks, surpassing existing methods in terms of both accuracy and efficiency. Notably, our approach achieves remarkable results with up to $2,000\times$ memory savings compared to SOTA two-

stage methods and exhibits significantly lower latency. Detailed analysis reveals the robustness of SemVPR, highlighting its ability to maintain high performance in real-world scenarios characterized by severe occlusions, illumination variations, and other challenges.

2. Related Work

Most early works on VPR employed handcrafted features for matching queries against references in the database [26]. With the rise of deep learning, feature extraction using neural networks gradually became the mainstream approach [2, 4, 5, 12]. Recently, researchers began exploring the use of Vision Transformers (ViTs) [9] for VPR, replacing Convolutional Neural Networks (CNNs) [13, 36]. Visual foundation models (VFM) [6, 8, 22, 30, 32] have gained significant traction, demonstrating remarkable generalization capabilities across various downstream tasks. Most recent works have incorporated pre-trained VFMs as backbones. AnyLoc [21] investigated the performance of various VFMs on the VPR task with frozen pre-trained models [6, 8, 30, 32] and classic aggregation strategies [2, 18, 31], where the DINOv2-based [30] model exhibited potential under ultra-high-dimensional descriptors. SelaVPR [28] attempted to fully unleash the capability of pre-trained models for VPR using Parameter Efficient Fine-Tuning (PEFT) [16]. Although it achieved promising performance, it relied on a computationally inefficient re-ranking step based on local feature maps [10, 33]. CricaVPR [27] introduced cross-image correlation-aware representation learning on top of PEFT [16]. However, this method depends on batch construction and exhibits significant performance fluctuations when the test set is shuffled or under different partitions. SALAD [17] proposed an optimal transport aggregation method designed for ViTs, improving upon NetVLAD and achieving excellent results across various benchmarks. Notably, similar to NetVLAD, SALAD’s method relies on the clustering and concatenation of local feature maps. Even with an off-line dimensionality reduction approach like PCA [44], its descriptors remain highly inefficient, being an order of magnitude larger than other methods.

There are a few works that attempt to introduce other forms of external knowledge to enhance scene understanding or train descriptors. StructVPR [35] attempts to use segmentation maps to introduce structural knowledge into the pixel space. However, segmentation maps only contain simple scene layouts without any details and conceptual semantics, which is not helpful for dealing with some situations such as occlusion and perspective changes. Textplace [15] proposes to leverage high-level semantic information with the help of scene texts in the wild, which are naturally robust and under extreme appearance changes and perceptual aliasing. However, this method lacks generalization and cannot be used in scenarios without scene text. We propose

to use a vision-language model [32] to guide the model to learn local semantic descriptors, which can be seen as implicitly learning textual descriptors of the scene. This makes our method not only robust but also highly generalizable, allowing it to be used for recognition in any scenario.

3. Method

3.1. Preliminary

A series of recent works have revealed the powerful representation capability of VFMs in VPR tasks. Inspired by this, we use a pre-trained DINOv2 [30] model as our backbone. DINOv2 uses a ViT minimally adapted as the encoder. Specifically, given an image $x \in \mathbb{R}^{h \times w \times 3}$, the ViT first extracts l non-overlapping patches from the image and projects them into d -dim tokens $\mathbf{x} \in \mathbb{R}^{l \times d}$, where $l = \frac{hw}{p^2}$ be the number of patch tokens extracted from the image, p be the patch size, and d be the dimensionality of the hidden feature space. The input token sequence $\mathbf{x}' \in \mathbb{R}^{l' \times d}$ is formed by prepending a learnable [CLS] token and several optional [REG] tokens to patch tokens [8]. After adding corresponding positional embeddings to preserve the positional information, \mathbf{x}' is fed into a series of transformer blocks to produce the feature representation. A standard transformer block in ViT, including multi-head attention (MHA) [40], a multi-layer perception (MLP) and a Layer Normalization (LN), is formulated as:

$$\mathbf{x}'_i = \text{MHA}(\text{LN}(\mathbf{x}_i)) + \mathbf{x}_{i-1} \quad (1)$$

$$\mathbf{x}_i = \text{MLP}(\text{LN}(\mathbf{x}'_i)) + \mathbf{x}'_i, \quad (2)$$

where \mathbf{x}_i is the output of the i -th block and the input of $(i+1)$ -th block.

3.2. Semantic-Aware Aggregation

Since DINOv2 is trained solely on visual data, the resulting representation captures rich low-level visual information but lacks an understanding of high-level semantic concepts. Conventional approaches that directly apply an aggregation module to DINOv2’s feature map to obtain a global descriptor are suboptimal because the feature map itself does not possess a conceptual understanding of objects or entities within the scene. For instance, the feature map cannot conceptually distinguish between buildings, pedestrians, or other elements, which could lead to shortcut learning where the model learns spurious relationships or patterns due to data bias rather than meaningful associations.

In contrast, humans do not rely solely on visual observations for place recognition but also leverage common sense knowledge to comprehend the abstract information within a scene, facilitating spatial reasoning by identifying reliable visual landmarks and focusing on relevant attributes. Therefore, we propose introducing a pre-trained vision-language model [37] as a teacher during the training phase. This

teacher model has already acquired concept-level knowledge through multi-modal pre-training on both visual and textual data. By distilling the semantic knowledge from the teacher model into our backbone network [14], we can guide the model to focus on reliable visual landmarks and their relevant attributes when aggregating low-level local visual descriptors. Crucially, once the training process is complete, the teacher model is no longer required during inference, minimizing computational overhead and ensuring efficient performance.

Specifically, we introduce a separate branch in the k -th block of the backbone to learn semantic concepts. This is formulated as

$$\mathbf{s} = \text{Conv}(\sigma_s(\text{Conv}(\mathbf{x}_{k-1}))), \quad (3)$$

where Conv is 1D-convolution operator, σ_s is a non-linear activation function, and \mathbf{s} is the local semantic descriptors with the same shape as \mathbf{x}_k . We use a fine-tuned CLIP model as the teacher model to provide semantic-level guidance. The details of this teacher model will be discussed in a later section.

Given the image x , we obtain the target semantic feature map by extracting from the teacher model \mathcal{F}_t

$$\mathbf{s}_t = \mathcal{F}_t(x). \quad (4)$$

We discard the [CLS] token from the teacher’s outputs such that the target semantic feature map \mathbf{s}_t has the same shape with \mathbf{s} . The feature-level distillation loss is formulated as:

$$\mathcal{L}_d = \frac{1}{p^2} \sum_i^p \sum_j^p \left(1 - \frac{\mathbf{s}^{ij} \cdot \mathbf{s}_t^{ij}}{\|\mathbf{s}^{ij}\| \cdot \|\mathbf{s}_t^{ij}\|} \right), \quad (5)$$

where p is the patch size and $\mathbf{s}^{ij}, \mathbf{s}_t^{ij}$ are the features of the corresponding patches.

In contrast to most works that introduce extra layers for global descriptor aggregation, including popular GeM [31] and NetVlad [2], we stick to the ViT design and use the [CLS] token and attention mechanism for aggregation. Specifically, let \mathbf{x}_{n-1} be the output of the penultimate block as the local visual descriptors, the semantic-aware aggregation is formulated as

$$\mathbf{x}_n^* = \sigma_a(\mathbf{W}_a \mathbf{s} + \mathbf{b}_a) \cdot \mathbf{x}_{n-1} \quad (6)$$

$$\mathbf{x}'_n = \text{MHA}(\text{LN}(\mathbf{x}_n^*)) + \mathbf{x}_n^* \quad (7)$$

$$\mathbf{x}_n = \text{MLP}(\text{LN}(\mathbf{x}'_n)) + \mathbf{x}'_n, \quad (8)$$

where $\mathbf{W}_a, \mathbf{b}_a$ are learnable parameters and σ_a is an activation function. The full-size global descriptor is defined as the [CLS] token of the output sequence

$$\mathbf{g} = \mathbf{x}_n[0]. \quad (9)$$

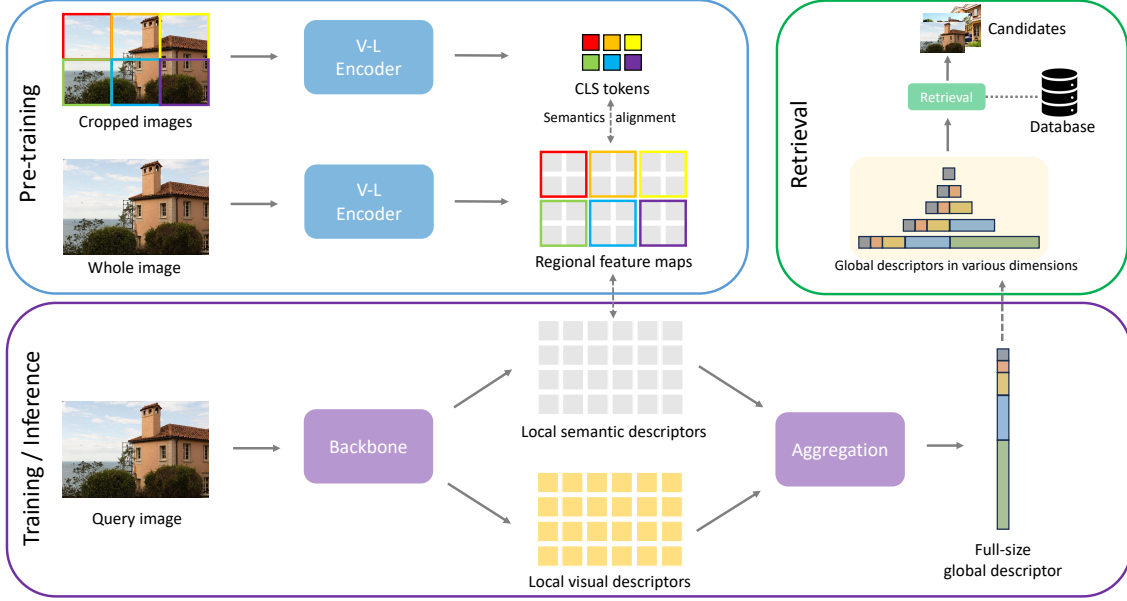


Figure 2. **Overview of the proposed SemVPR.** During pre-training, Local Semantic Alignment (LSA) is applied to a pre-trained vision-language encoder to enhance the quality of its local feature maps. For the training phase, the backbone model simultaneously learns local semantic descriptors and local visual descriptors, guided by the vision-language teacher model, and performs Semantic-Aware Aggregation (SAA) to create global descriptors. Nested Descriptor Learning (NDL) strategy is deployed, producing a set of descriptors with varying capabilities from the same model output, which can be directly utilized for retrieval tasks.

3.3. Local Semantics Alignment

The original CLIP encoder was trained on image-text pairs, where the objective was to establish a holistic correspondence between the entire image and its textual description, without imposing explicit constraints on local regions within the image [32]. Consequently, the quality of CLIP’s local feature map, which captures region-specific representations, is known to be suboptimal, thereby not ideal for our purpose. To overcome this challenge, we finetune the CLIP model focusing on refining its local feature map. Let \mathcal{F}_c be the visual encoder of the original CLIP model, we initialized \mathcal{F}_t with the same weights from \mathcal{F}_c . Given an image, we divide it into a grid of $H \times W$ local regions, denoted as $R = \{R^{ij} | 0 \leq i \leq H, 0 \leq j \leq W\}$. The goal is to align the region representations within the local feature map from \mathcal{F}_t with the image-level representations of the corresponding image crops from \mathcal{F}_c . This is formulated as a self-alignment pre-training process with loss:

$$\mathcal{L}_{pre} = \frac{1}{HW} \sum_i^H \sum_j^W \left(1 - \frac{\mathbf{t}^{ij} \cdot \mathbf{c}^{ij}}{\|\mathbf{t}^{ij}\| \cdot \|\mathbf{c}^{ij}\|} \right), \quad (10)$$

where \mathbf{t}^{ij} is the average pooling of the patch features within region R^{ij} from model \mathcal{F}_t , and \mathbf{c}^{ij} is the [CLS] token of the output from model \mathcal{F}_c with input image crop R^{ij} . This process can be seen as implicitly aligning the regional features with the textual description of the image crop, such

that the local feature map could be equipped with rich semantic-level information.

3.4. Nested Descriptor Learning

Obtaining high-performance and efficient descriptors has long been a crucial challenge in VPR tasks. Current methods suffer from the drawback of producing descriptors with excessively high dimensions, often reaching tens of thousands, resulting in inefficiency. This stems from their natures of clustering and concatenation operations on local features [2, 12, 17], which can be regarded as simulating the local matching of two-stage methods using retrieval.

We introduce a nested descriptor learning strategy to address this issue. With the help of concept-level semantic knowledge, our approach encourages the model to extract information from the most relevant parts of the image rather than exhausting all of them, thereby learning a series of extremely compact descriptors. In particular, this set of descriptors with varying capabilities is derived from the same global vector output by a single model, eliminating the dependence on offline dimensionality reduction methods such as PCA [44], or the necessity of training multiple models of different sizes.

Specifically, we follow the map partitioning of EigenPlaces [5] and adopt classification as a proxy training task using the Large Margin Cosine Loss [41], formulated as:

Dataset Name	Pitts30k [38]	MSLS-val [43]	Tokyo 24/7 [39]	SPED [7]	SF-XL test-v1 [4]	SF-XL test-v2 [4]	SF-XL test-occlusion [3]	SF-XL test-night [3]
# queries	6.8k	740	315	607	1000	598	76	266
# database	10k	19k	76k	607	2.8M	2.8M	2.8M	2.8M
Orientation	panorama	frontal-view	multi-view	frontal-view	panorama	panorama	panorama	panorama
Scenery	urban	urban, suburban	urban	country	mostly urban	mostly urban	mostly urban	mostly urban
Domain Shift	none	day/night	day/night	season	viewpoint, day/night	viewpoint	severe occlusion	extreme day/night

Table 1. Summary of the evaluation benchmarks.

$$\mathcal{L}_{lmc}(x; W, m, s) = \frac{1}{N} \sum_i -\log \frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)} + \sum_{i \neq j} e^{s \cos(\theta_j)}}, \quad (11)$$

subject to

$$\begin{aligned} \cos(\theta_j) &= \tilde{W}_j^T \tilde{x}_i \\ \tilde{W} &= \frac{W}{\|W\|}, \quad \tilde{x} = \frac{x}{\|x\|}, \end{aligned} \quad (12)$$

where N is the number of training samples, x_i is the i -th feature vector corresponding to the ground-truth class y_i , W is a learnable weight matrix, and s, m are hyperparameters for scaling and margin respectively.

We optimize the global descriptor in a coarse-to-fine manner by directly imposing constraints in multi-granularity [23]. The CosFace loss is applied on multiple nested sub-vectors of the full-size global descriptor \mathbf{g} and aggregated together. This is formulated as:

$$\mathcal{L}_c = \sum_{k \in \mathcal{K}} \beta_k \cdot \mathcal{L}_{lmc}(\mathbf{g}_{1:k}; \mathbf{W}^{(k)}, m^{(k)}, s^{(k)}). \quad (13)$$

Here \mathcal{K} is a collection of nested dimensions and $\mathbf{g}_{1:k}$ represents a k -dimensional global descriptor consisting of the first k elements of the full-size global descriptor \mathbf{g} . For the optimization of each k -dimensional global descriptor, a separate linear projector, parameterized by $\mathbf{W}^{(k)}$, is used, as well as the corresponding scaling factor $s^{(k)}$ and margin $m^{(k)}$. Specifically, we would have

$$\begin{cases} m^{(k_i)} \leq m^{(k_j)} & \text{if } k_i \leq k_j, \\ s^{(k_i)} \geq s^{(k_j)} & \text{if } k_i \leq k_j. \end{cases} \quad (14)$$

The first constraint sets a larger decision margin for higher-dimensional descriptors in classification tasks, forcing them to capture more nuanced patterns, thereby learning representations in a coarse-to-fine manner. The second constraint ensures the stability of metric learning. All classification losses are aggregated with corresponding weight β_k .

The final training objective is formulated as the weighted sum of distillation loss and classification loss:

$$\mathcal{L} = \alpha_d \mathcal{L}_d + \alpha_c \mathcal{L}_c \quad (15)$$

4. Experiments

4.1. Datasets

To rigorously evaluate the effectiveness of the proposed method, we conducted comprehensive experiments on a variety of VPR benchmarks, as summarized in Table 1.

4.2. Implementation Details

Training strategy. For the pre-training, we initialize the \mathcal{F}_t model with weights from EVA-CLIP ViT-L [37] and fine-tune it on the COCO dataset [24] using \mathcal{L}_{pre} loss. We train for 5 epochs using the AdamW [25] optimizer with a learning rate of 1e-5 and a weight decay of 0.1 [45]. For the training, we initialize our backbone with the weights from DINOv2 ViT-L [30] and fine-tune the last 4 blocks of the model on the training set of SF-XL [4]. We follow the map partitioning design of EigenPlaces [5] to split the map into mesh with 15×15 -meters and use non-adjacent cells as classes. We train the model for 20 epochs with 5000 iterations each using the AdamW optimizer [25]. The learning rates of the backbone, semantic branch, and projection layer in CosFace loss [41] are set to 1e-5, 1e-4, and 1e-3, respectively. The nesting dimension \mathcal{K} is set to $\{1024, 512, 256, 128, 64\}$, and we use $m^{(1024)} = 0.4$ and decay by 0.5 for the remaining dimensions. We use $s^{(k)} = 100$ and $\beta^{(k)} = 1$. For the final loss, we have $\alpha_c = 1$ and $\alpha_d = 5$.

Evaluation protocol. We evaluate the recognition performance using Recall@N (R@N), following the standard evaluation procedure [4, 5, 17, 28]. We also report detailed computational costs, including descriptor dimensions, latency, and memory footprint [19].

4.3. Comparisons with State-of-the-Art Methods

Comparisons with single-stage baselines. We compare our SemVPR with a single-stage baseline [2] and several single-stage SOTA methods [1, 4, 5, 17]. These methods use only one global descriptor for retrieval, which is the same as our setting. In addition, we also compare with a ViT-based SOTA method proposed recently [28]. However, this method includes a time-consuming re-ranking stage relying on local feature maps. For a fair comparison, we report its first-stage retrieval results here. As

Method	Venue	Desc. Dim	Pitts30k [38]			Tokyo24/7 [39]			MSLS-val [43]			SPED [7]		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [2]	CVPR'16	32768	81.9	91.2	93.7	60.6	68.9	74.6	53.1	66.5	71.1	78.7	88.3	91.2
CosPlace [4]	CVPR'22	512	88.5	94.5	95.2	82.8	90.0	92.7	79.5	87.2	89.4	67.2	80.4	87.8
MixVPR [1]	WACV'23	4096	91.5	95.5	96.3	85.1	91.7	94.3	88.0	92.7	94.6	85.2	92.1	94.6
EigenPlaces [5]	ICCV'23	2048	92.5	96.8	97.6	93.0	96.2	97.5	89.1	93.8	95.0	69.9	82.9	88.7
SelaVPR* [28]	ICLR'24	<u>1024</u>	90.2	96.1	97.1	81.9	95.9	96.5	87.7	<u>95.8</u>	96.6	86.6	93.1	95.2
SALAD [17]	CVPR'24	8448	<u>92.2</u>	<u>96.3</u>	<u>97.6</u>	<u>94.6</u>	<u>96.9</u>	<u>97.8</u>	91.2	96.3	96.6	<u>90.9</u>	<u>95.7</u>	<u>96.4</u>
SemVPR (ours)	-	<u>1024</u>	94.2	97.7	98.5	96.8	98.1	98.7	<u>89.7</u>	94.7	<u>95.9</u>	91.4	95.9	97.2
SemVPR (ours)	-	512	94.0	97.6	98.3	97.1	98.1	98.4	<u>89.3</u>	94.6	<u>95.4</u>	91.1	96.2	97.0
SemVPR (ours)	-	256	93.5	97.0	97.9	95.9	97.8	98.7	88.9	94.2	<u>95.5</u>	89.0	95.6	96.7
SemVPR (ours)	-	128	92.6	96.7	97.4	92.1	95.2	96.8	86.9	92.8	93.9	88.6	94.2	95.6
SemVPR (ours)	-	64	90.1	95.8	97.0	82.9	90.5	93.9	82.2	90.8	92.4	80.4	89.6	94.2

Table 2. Comparison against retrieval-only baselines. Two-stage methods are marked with *, and report results from the first-stage for fair comparison. The best results are **bolded**, and the second bests are underlined.

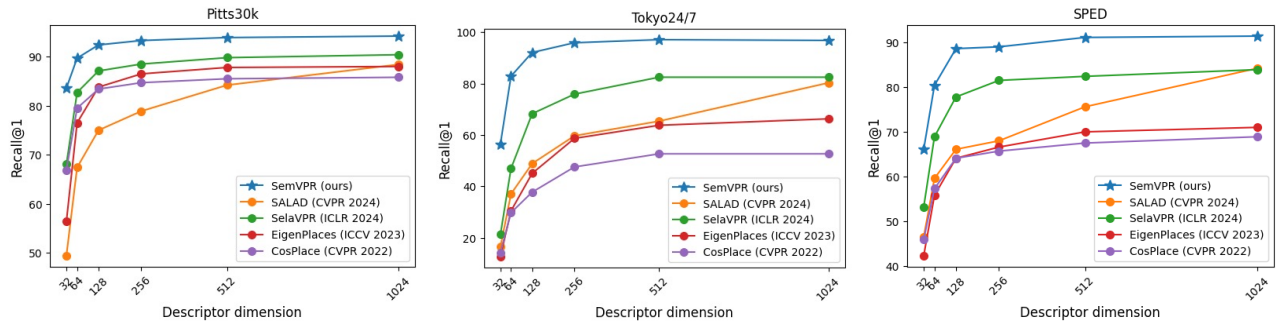


Figure 3. Comparison of performance against single-stage methods under same descriptor dimensions.

shown in Table 2, our SemVPR achieves SOTA or comparable performance across various benchmarks. Despite using a much more compact global descriptor for retrieval, SemVPR significantly outperforms other single-stage methods on Pitts30k and Tokyo 24/7. It’s worth noting that our method remains SOTA even with an ultra-compact descriptor in 128 dimensions on Pitts30k and 256 dimensions on Tokyo 24/7, which is $66\times$ and $33\times$ more efficient compared to the most recent SOTA method SALAD [17] with a 8448-dimensional descriptor. The global descriptor of SALAD is constructed by concatenating clustered local feature maps and global features. Using such a concatenated ultra-high-dimensional descriptor for retrieval can actually be regarded as a simulation of the re-ranking based on local features. Meanwhile, the descriptor of SemVPR is aggregated under semantic guidance, making it capture the most relevant semantic information while discarding useless local information, thereby achieving good performance while maintaining compactness. Compared to CosPlace [4] and SelaVPR [28] which have relatively compact descriptors, SemVPR achieves a substantial improvement in performance in the same dimensionality, rendering the superiority of the proposed method. Furthermore, SemVPR can still achieve competitive results against SOTAs on all

benchmarks while using an ultra-compact descriptor as low as 64 dimensions. For a more fair comparison, we further evaluate the performance of these methods by truncating the descriptors to the same dimension. Figure 3 intuitively demonstrates the superiority of SemVPR.

To the best of our knowledge, our work is the first to explore such low-dimensional descriptors and achieve competitive performance, rendering the efficiency and robustness of SemVPR. It is worth noting that unlike previous methods that rely on training multiple models to obtain descriptors of different sizes or use offline dimensionality reduction methods such as PCA, all SemVPR descriptors come from the output of a single model, i.e., the 128-dimensional descriptor is simply a truncation of the 1024-dimensional descriptor. This makes SemVPR seamlessly adaptable to different scenarios without any overhead.

Comparisons with baselines with re-ranking. We compare our SemVPR with the best two-stage SOTA methods. All these methods have a re-ranking stage that relies heavily on local feature maps, which leads to huge latency and memory overhead. Patch-NetVLAD [12] and TransVPR [42] deploy RANSAC-based geometric verification on patch descriptors [10]. R2Former [47] introduces an

Method	Venue	Desc. Dim		Memory (GB)	Latency (ms)		Pitts30k [38]			Tokyo24/7 [39]		
		Global	Local		Retrieval	Re-ranking	R@1	R@5	R@10	R@1	R@5	R@10
Patch-NetVLAD [12]	CVPR'21	4096	2826 × 4096	908.30	0.35	12642.55	88.7	94.5	95.9	86.0	88.6	90.5
TransVPR [42]	CVPR'22	256	1200 × 256	22.72	0.18	4642.17	89.0	94.9	96.2	79.0	82.2	85.1
R2Former [47]	CVPR'23	256	500 × 131	4.70	0.18	249.73	91.1	95.2	96.3	88.6	91.4	91.7
SelaVPR [28]	ICLR'24	1024	256 × 128	2.47	0.25	197.81	92.8	96.8	97.7	94.0	96.8	97.7
SemVPR (ours)	-	1024	0	0.76	0.25	0	94.2	97.7	98.5	96.8	98.1	98.7
SemVPR (ours)	-	256	0	0.54	0.18	0	93.5	97.0	97.9	95.9	97.8	98.7
SemVPR (ours)	-	128	0	0.48	0.14	0	92.6	96.7	97.4	92.1	95.2	96.8

Table 3. Comparison against baselines with local re-ranking. The best results are **bolded**. Latency refers to the average retrieval time and re-ranking time for a single query image. Memory refers to the peak memory footprint during the retrieval process.

additional transformer module for correlation-based reranking while SelaVPR [28] performs local matching based on mutual nearest neighbor. As shown in Table 3, despite using global descriptors solely, SemVPR surpasses all two-stage methods by a large margin on both Pitts30k and Tokyo 24/7. We also report the latency and memory footprint of each method. Note that the computational cost of two-stage methods is usually dominated by the re-ranking stage, as it requires storing huge local feature maps and performing a time-consuming matching algorithm. However, SemVPR does not rely on re-ranking, making its computational overhead several orders of magnitude smaller. This result renders the efficiency of SemVPR, with up to **2,000×** memory savings and **100,000×** latency reduction compared with two-stage baselines, making SemVPR adaptable to a wider range of application scenarios, such as edge devices with limited computing power and latency-sensitive navigation scenarios.

Comparisons on large-scale benchmarks. We evaluate SemVPR on SF-XL [3, 4] and compare against several SOTAs [4, 5, 17, 28]. We intentionally compare only with single-stage methods, partly for fairness and partly because the computational overhead of re-ranking-based methods is unacceptable on such a large dataset. As shown in Table 4, SemVPR achieves the best on all the test sets, including challenging ones with severe occlusions and extreme day/night changes. It is worth noting that although the recall rates of SALAD [17] seem promising, its high-dimensional descriptor hinders its practicality in large-scale applications, as its huge latency and memory overhead of nearly 200GB make it infeasible for usage in real-time applications or deployment on common devices. Instead, SemVPR achieves competitive results using descriptors as low as 256 dimensions. Note that even with a large-scale database at the city level and without extra engineering optimization, SemVPR can seamlessly run on a device with 8GB memory with decent latency and performance, while all baselines fail to do so [19]. This result demonstrates the superiority of SemVPR in real-world applications.

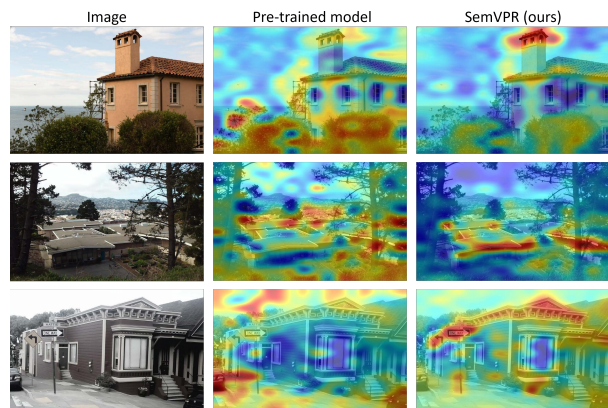


Figure 4. Example queries and attention map visualizations of the pre-trained backbone model (DINOv2) and SemVPR. The regions attended by the pre-trained model have no relevance to place recognition while SemVPR focuses on the most discriminative parts such as the attic (1st example), the shaped building (2nd example), and the unique roof and traffic signs (3rd example).

4.4. Ablation Studies

We perform a series of ablation experiments to validate the effectiveness of the proposed SemVPR both quantitatively and qualitatively. We first evaluate the performance of pre-trained models as the borderlines using their [CLS] token. As shown in Table 5, DINOv2 achieves decent results because of its powerful representation capability while the CLIP performs poorly as its [CLS] token contains merely high-level semantics but lacks low-level visual features essential for VPR tasks. Fine-tuning DINOv2 on SF-XL training set bridges the gap between pre-training and downstream VPR tasks, significantly improving the performance of the model. Equipped with the semantic-aware aggregation, the 1024-dimensional full-size descriptor has reached the level of SOTAs. Local semantic alignment plays an important role in this because it makes the local feature map of the teacher model \mathcal{F}_t have rich semantic information, thereby guiding the backbone to learn useful local semantic descriptors. We note that the nested descriptor learning strategy does not improve the performance of the full-size

Method	Venue	Desc. Dim	Latency (ms)	Memory (GB)	SF-XL test-v1 [4]			SF-XL test-v2 [4]			SF-XL Occlusion [3]			SF-XL Night [3]		
					R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CosPlace [4]	CVPR'22	2048	53.25	45.28	76.4	83.6	85.3	88.8	94.2	95.9	26.3	38.2	46.1	23.8	29.0	31.5
EigenPlaces [5]	ICCV'23	2048	53.25	45.28	84.1	89.1	90.7	90.7	95.7	96.7	32.9	48.6	52.6	23.6	30.7	34.5
SelaVPR* [28]	ICLR'24	<u>1024</u>	<u>32.47</u>	<u>22.77</u>	74.9	80.7	82.1	89.3	95.7	96.2	35.5	47.4	55.3	38.4	50.9	55.4
SALAD [17]	CVPR'24	8448	196.67	186.4	<u>88.6</u>	<u>93.5</u>	<u>94.4</u>	<u>94.8</u>	<u>97.3</u>	<u>98.3</u>	<u>50.7</u>	62.8	<u>66.2</u>	<u>46.6</u>	<u>58.9</u>	<u>62.2</u>
SemVPR (ours)	-	<u>1024</u>	<u>32.47</u>	<u>22.77</u>	93.8	96.9	97.7	94.9	98.0	98.5	56.6	<u>61.8</u>	67.1	56.2	71.7	75.8
SemVPR (ours)	-	512	21.15	11.82	93.0	96.2	97.2	<u>94.0</u>	98.3	98.5	<u>50.0</u>	<u>57.9</u>	<u>63.2</u>	52.8	67.8	73.0
SemVPR (ours)	-	256	15.70	6.35	89.4	93.8	94.9	<u>92.6</u>	97.7	<u>98.2</u>	<u>39.5</u>	<u>52.6</u>	<u>57.9</u>	48.1	60.3	65.9
SemVPR (ours)	-	128	11.39	3.61	81.8	86.7	88.8	88.0	95.5	<u>96.8</u>	31.6	35.5	42.1	33.5	44.4	50.6

Table 4. Comparison on large-scale datasets and their challenging variants. Two-stage methods are marked with *, and report results from the first-stage for fair comparison. The best results are **bolded**, and the second bests are underlined. Latency refers to the average retrieval time for a single query image. Memory refers to the peak memory footprint during the retrieval process.

Method	$d = 64$		$d = 256$		$d = 1024$	
	R@1	R@5	R@1	R@5	R@1	R@5
Frozen CLIP	35.0	61.6	46.3	71.9	49.6	74.2
Frozen DINOv2	67.6	87.1	76.7	91.1	78.0	92.1
SemVPR w/o SAA	77.4	89.2	86.3	94.0	91.7	96.2
SemVPR w/o LSA	80.3	91.2	87.8	94.5	93.5	97.3
SemVPR w/o NDL	82.6	92.4	91.4	96.3	94.4	97.9
SemVPR	90.1	95.8	93.5	97.0	94.2	97.7

Table 5. Ablation on components on Pitts30k. LSA: Local Semantics Alignment. SAA: Semantic-Aware Aggregation. NDL: Nested Descriptor Learning.

descriptor. Instead, the performance of low-dimensional nested descriptors has been significantly enhanced (+9%) at the cost of only a small sacrifice in the performance of full-sized descriptors (-0.3%), which is within expectation.

To further analyze the effect of the semantic-aware aggregation, we visualize the attention maps of [CLS] token in SemVPR by averaging the channel dimension, and compare them with the ones generated from the pre-trained backbone model. As shown in Figure 4, the attention of the pre-trained model tends to be scattered and lacks focus in most scenarios, sometimes biased towards the foreground. However, we can clearly see that the SemVPR pays strong attention to the buildings and moderate attention to the vegetation, as they are reliable visual landmarks for VPR tasks. Moreover, we notice that the attention of SemVPR is highly concentrated: it always focuses on the most discriminative parts of the scene, such as the attic (in the first example), the shaped building (in the second example), and the unique roof and traffic signs (in the third example).

We also noticed that the pre-trained model has a strong tendency to over-focus on some common objects in daily life such as vehicles. This may be a bias introduced by visual-only pre-training tasks [30]. As shown in Figure 5, our model is able to correct this bias with the help of semantic guidance and stays focused on reliable landmarks, which greatly improves the robustness of descriptors.

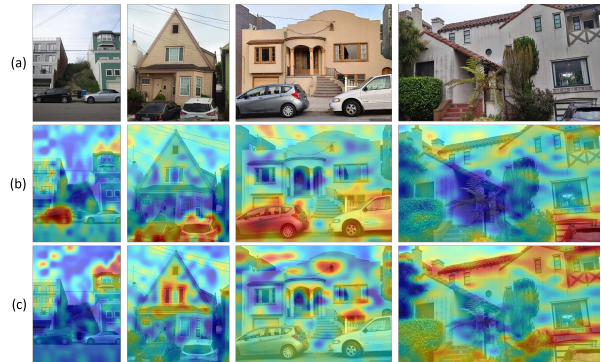


Figure 5. Example queries (a) and attention map visualizations of the pre-trained DINOv2 model (b) and SemVPR (c). The pre-trained model tends to focus on common objects such as vehicles while SemVPR corrects this bias and extracts features from reliable landmarks such as buildings.

5. Conclusions

We introduce SemVPR, an efficient and robust visual place recognition framework. By integrating a VLM during training, SemVPR learns both local visual and semantic descriptors. The semantic-aware aggregation approach mitigates perceptual aliasing and enhances robustness. Our nested descriptor learning strategy further enables a single model to produce extremely compact descriptors without additional overhead, eliminating the need for multiple models or offline dimensionality reduction. SemVPR outperforms previous state-of-the-art methods, including two-stage re-ranking approaches, across various benchmarks while maintaining lower latency and memory requirements. Its performance on large-scale and challenging datasets further demonstrates its efficiency and robustness. SemVPR offers a promising solution for latency-sensitive and computationally constrained scenarios.

References

- [1] 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. 1, 5, 6
- [2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Paszka, 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. 1, 2, 3, 4, 5, 6
- [3] Giovanni Barbarani, Mohamad Mostafa, Hajali Bayramov, Gabriele Trivigno, Gabriele Berton, Carlo Masone, and Barbara Caputo. Are local features all you need for cross-domain visual place recognition? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2023. 5, 7, 8
- [4] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 2, 5, 6, 7, 8
- [5] 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. 2, 4, 5, 6, 7, 8
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [7] Zetao Chen, Lingqiao Liu, Inkyu Sa, Zongyuan Ge, and Margarita Chli. Learning context flexible attention model for long-term visual place recognition. *IEEE Robotics and Automation Letters*, 3(4):4015–4022, 2018. 5, 6
- [8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 2, 3
- [9] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 6
- [11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1
- [12] 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. 1, 2, 4, 6, 7
- [13] 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. 2
- [14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [15] Ziyang Hong, Yvan Petillot, David Lane, Yishu Miao, and Sen Wang. Textplace: Visual place recognition and topological localization through reading scene texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2861–2870, 2019. 2
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [17] 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*, pages 17658–17668, 2024. 2, 4, 5, 6, 7, 8
- [18] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International journal of computer vision*, 87:316–336, 2010. 1, 2
- [19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5, 7
- [20] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. 1
- [21] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 2
- [22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2
- [23] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022. 5
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [26] 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. 1, 2

- [27] 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*, pages 16772–16782, 2024. 2
- [28] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024. 1, 2, 5, 6, 7, 8
- [29] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C de Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2020. 1
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, 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. 2, 3, 5, 8
- [31] 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. 2, 3
- [32] 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. 2, 3, 4
- [33] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2
- [34] Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial. *arXiv preprint arXiv:2303.03281*, 2023. 1
- [35] 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. 2
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2
- [37] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 5
- [38] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 5, 6, 7
- [39] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2): 257–271, 2018. 5, 6, 7
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 4, 5
- [42] 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. 1, 6, 7
- [43] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5, 6
- [44] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 2, 4
- [45] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 5
- [46] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020. 1
- [47] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 1, 6, 7