TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation

Ruotong Wang*   Yanqing Shen*   Weiliang Zuo   Sanping Zhou   Nanning Zheng†

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

{wrt072@stu., qing1159364090@stu., weiliang.zuo@, spzhou@, nnzheng@mail.}xjtu.edu.cn

Abstract

Visual place recognition is a challenging task for applications such as autonomous driving navigation and mobile robot localization. Distracting elements presenting in complex scenes often lead to deviations in the perception of visual place. To address this problem, it is crucial to integrate information from only task-relevant regions into image representations. In this paper, we introduce a novel holistic place recognition model, TransVPR, based on vision Transformers. It benefits from the desirable property of the self-attention operation in Transformers which can naturally aggregate task-relevant features. Attentions from multiple levels of the Transformer, which focus on different regions of interest, are further combined to generate a global image representation. In addition, the output tokens from Transformer layers filtered by the fused attention mask are considered as key-patch descriptors, which are used to perform spatial matching to re-rank the candidates retrieved by the global image features. The whole model allows end-to-end training with a single objective and image-level supervision. TransVPR achieves state-of-the-art performance on several real-world benchmarks while maintaining low computational time and storage requirements.

1. Introduction

Visual Place Recognition (VPR) is an essential and challenging problem in autonomous driving and robot localization systems, which is usually defined as an image retrieval problem [27]. Given a query image, the algorithm has to determine whether it is taken from a place already seen and identify the corresponding images from a database. There are two types of image representations commonly used in VPR tasks. Global image features [2, 8, 18, 21, 37, 38, 54] abstract the whole image into a compact feature vector without geometrical information. Patch-level descriptors

---

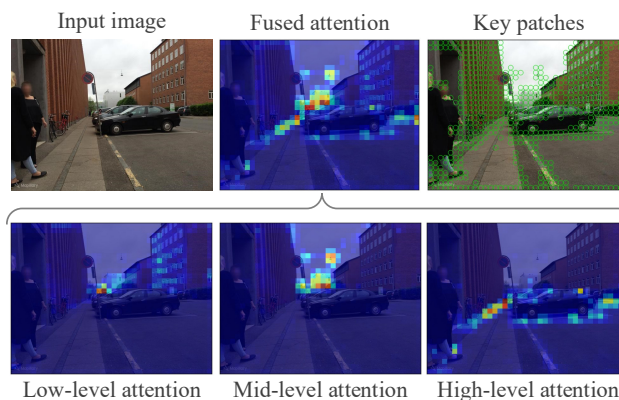*Equal contribution.

†Corresponding author.

Figure 1. **Visualization of multi-level attentions from TransVPR.** Low-level attention maps mainly focus on small objects and textural areas on the surface of buildings. Mid-level attentions focus on objects in the air, such as street lamps and tree canopies, while high-level attentions tend to outline the contours of the ground and the lane lines. All these attention masks are combined to generate global image representations as well as key-patch descriptors.

[12, 14, 23, 26, 32, 60] describe particular patches or key-points in an image and can be used to perform spatial matching between image pairs using cross-matching algorithms (*e.g.* RANSAC [16]). To achieve a good trade-off between accuracy and efficiency, a commonly used two-stage strategy is to retrieve candidates with global features and then re-rank them using patch-level descriptor matching [40,51]. Several recent researches [6, 19, 45, 52] have attempted to design a holistic system to extract both types of features. Most recently, Patch-NetVLAD [19] has used an integral feature space to derive patch descriptors from the global image feature and has achieved state-of-the-art performance in several benchmarks. However, an important factor that may reduce the robustness of Patch-NetVLAD is that its extracted features unselectively encode the information from all regions of an image.

It needs to be emphasized that the ability to identify task-relevant regions in an image is critical to VPR systems.

This is because distracting elements and dynamic objects in a scene (*e.g.* the sky, the ground, untextured walls, cars, pedestrians, *etc.*) are not helpful to recognize a place and seriously harm the VPR performance [27]. In order to detect keypoints or regions of interest, several CNN-based methods have been proposed [9, 12, 23, 24, 57, 59, 60].

Recently, the Transformer [56] architecture has obtained competitive results in multiple computer vision tasks [7, 13]. Unlike CNNs, the self-attention operation in vision Transformers can dynamically aggregates global contextual information and implicitly select task-relevant information. To benefit from this property of vision Transformers and improve the robustness of place recognition, this work brings the following **contributions**: Firstly, we propose a Transformer-based novel place recognition model, TransVPR, which can adaptively extract robust image representations from distinctive regions in an image. Secondly, inspired by previous studies on CNNs which combine multi-level feature maps to enrich image representations [9, 59, 61], we fuse multi-level attentions, which focus on different semantically meaningful regions (see Fig. 1), to generate global image representations. The effectiveness of this procedure is demonstrated by qualitative and quantitative experiments. Finally, the output tokens of Transformer layers filtered by the fused attention mask are further employed as patch-level descriptors to perform geometrical verification. All components in TransVPR are tightly coupled so that the whole model allows end-to-end optimization with a single training objective and only image-level supervision. Experimental results show that the proposed TransVPR achieves superior performance on VPR benchmark datasets with low computational time and memory requirements. It outperforms the state-of-the-art VPR approaches [2, 6, 17, 19, 41] by significant margins (5.8% absolute increase on Recall@1 compared with the best baseline method, DELG [6]).

2. Related Work

We review previous works on image description techniques, especially related to place recognition.

Patch-Level Descriptors. In early VPR systems [1, 11, 25, 30, 31, 46], traditional methods such as SIFT [26], SURF [4] and ORB [39] have been widely used to represent a small patch centered around a detected keypoint. However, these handcrafted features cannot handle severe appearance changes. More recently, CNN-based methods have achieved superior performances [5, 19, 23, 32, 52]. In order to extract sparse patch descriptors, some approaches have proposed to firstly detect keypoints based on local structures and then describe them with a separate CNN [34, 43, 48, 60, 63], while others have used shared network to preform both detection and description [12, 14]. In addition to these general methods, several attempts have also been devoted to learn task specific patch-level features for place recognition [6, 32, 62]. Besides, Patch-NetVLAD [19] provided an alternative solution using a global descriptor technique, NetVLAD [2], to extract descriptors from predefined image patches.

In most previous studies, patch-level descriptors refer to as *local descriptors*, which encode the content in local patches around keypoints. In contrast, our Transformer-based patch-level descriptors are not local, since each output token from a Transformer layer has global perception field. In this way, the patch descriptors are able to capture more semantically meaningful structures with long-range dependency.

Global Image Representation. Global image features are usually obtained by aggregating local descriptors. Some traditional techniques, such as Bag of Words (BoW) [10, 49], Fisher Kernel [21, 35, 36], and Vector of Locally Aggregated Descriptors (VLAD) [3, 20], have been used to assign visual words to images. Likewise, in deep learning context, some works [2, 29] have incorporated these clustering methods into CNN architectures, while other studies [18, 22, 37, 53] have focused on pooling from CNN feature maps. Recently, unified networks have been developed to jointly extract global features and patch-level descriptors [6, 40, 47, 51]. Previous CNN-based approaches which extract features from high-level convolutional layers require deep networks with down sampling layers to integrate enough contextual information. As a first attempt, El-Nouby *et al.* [15] introduced vision Transformers to image retrieval tasks by using the [class] token [56] from the final layer as a global feature. Differently, we aggregate multi-level attentions to generate global features and explicitly learn corresponding attention maps which can be further employed to detect key-patches.

Attentions for Place Recognition. In order to adaptively identify task-relevant regions in a complex scene image, attention mechanism has been adopted in several VPR approaches. Among them, the learned attention maps can be considered as patch descriptor filters [32, 59, 62] or weight maps which modulate the CNN feature maps to generate global features [9, 24]. The attention module in CNN-based methods has usually been implemented as a shallow CNN which is trained separately [32] or jointly [6, 9, 59, 62] with the backbone network. In our work, a new formulation is proposed, where the attention module is as simple as a linear layer which decodes the attention information from Transformer tokens.

3. Methodology

TransVPR jointly extracts both patch-level and global image representations by aggregating multi-level attentions in vision Transformers. Given an input image, its raw patch descriptors are firstly extracted by a shallow CNN and em-
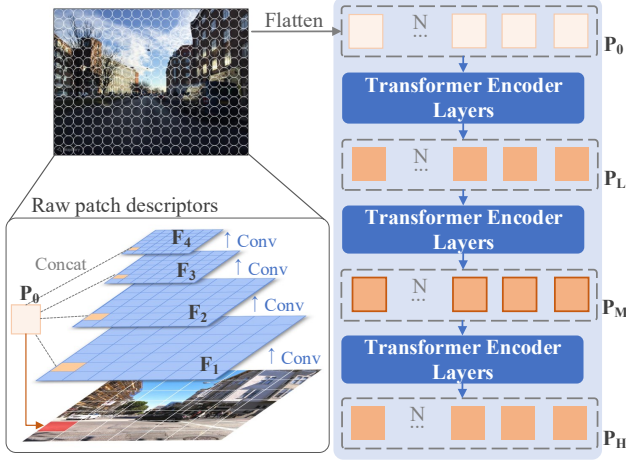
Figure 2. **Patch descriptor extraction.** For an input image, pyramid feature maps are generated by a CNN and each reshaped into a sequence of flattened 2D patches. Raw patch-level descriptors are obtained by concatenating patch embeddings at the same position from each feature map. They are then sent into a Transformer encoder to integrate global contextual information. Output patch tokens from the low-level, mid-level, and high-level Transformer layers are selected for subsequent processing.

bedded as input tokens of a vision Transformer. Attentions from shallow, middle, and deep layers of the Transformer are merged to generate global image features and detect task-relevant patches. Fig. 2 and Fig. 3 illustrate the whole feature extraction pipeline.

3.1. Patch Descriptor Extraction

A four-layer CNN is applied on an input image to extract raw patch-level features, as illustrated in Fig. 2. Given an image or a feature map $\mathbf{F_{i-1}} \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$, the output of a convolutional layer is:

$$\mathbf{F_i} = \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv}(\mathbf{F_{i-1}})))), \quad (1)$$

where $\mathbf{F_i} \in \mathbb{R}^{\frac{H_{i-1}}{2} \times \frac{W_{i-1}}{2} \times C_i}$. In practice, 3×3 convolutional kernel is used, and the number of output channels are set to 64, 128, 256, and 512 respectively. In this way, we obtain a feature pyramid $\{\mathbf{F_1}, \mathbf{F_2}, \mathbf{F_3}, \mathbf{F_4}\}$, where the size of feature map is reduced by half in order.

Then, *patch embedding* is applied on each feature map following the process proposed in [13]. A feature map $\mathbf{F_i} \in \mathbb{R}^{H_i \times W_i \times C_i}$ is reshaped into a sequence of flattened 2D patches $\mathbf{F_i'} \in \mathbb{R}^{N \times (R_i^2 \cdot C_i)}$, where (R_i, R_i) is the resolution of a feature map patch and $N = H_i W_i / R_i^2$ is the number of patches. To maintain a fixed number of patches on each feature map, R_i is set to $R_{i-1}/2$. The flattened patches are mapped to $D/4$ dimensions, where D is the latent embedding dimension of the subsequent Transformer blocks.

Next, patch embeddings at the same position of different feature maps are concatenated as a raw patch-level local descriptor. We denote the group of raw patch descriptors as $\mathbf{P_0} \in \mathbb{R}^{N \times D}$. The location of a patch descriptor is approximated to the center coordinate of the corresponding image patch.

Finally, in order to integrate global contextual information, raw patch descriptors are then sent into a Transformer encoder as input tokens. We follow the standard implementation of Transformer encoder, which consists of a stack of Multi-headed Self-Attention (MSA) and Multi-Layer Perceptron (MLP) modules [13,56]. In pre-training, a learnable `[class]` token is added in the front of the token sequence to obtain an image representation for classification. Since the spatial positional information can be implicitly encoded in raw patch descriptors by the CNN architecture, *positional embedding* is removed from Transformer blocks so that the model can be flexible to different input sizes.

3.2. Multi-Level Attentions

Although the Transformer has global perception field from the lowest layers, it is observed that its mean attention distance increases with depth [13]. In other words, there are some differences in the scales of structures captured by different Transformer layers. To integrate information across multiple levels, three groups of output patch tokens from the low-level, mid-level, and high-level layers of Transformer are selected, denoted as $\{\mathbf{P_L}, \mathbf{P_M}, \mathbf{P_H}\}$. A group of multi-level patch tokens $\mathbf{P}$ are firstly composed by concatenating these three groups of tokens along the channel.

$$\mathbf{P} = \text{Concat}([\mathbf{P_L}, \mathbf{P_M}, \mathbf{P_H}]) \in \mathbb{R}^{N \times 3D}. \quad (2)$$

For each level, an attention mask is estimated individually over all spatial positions, indicating the contribution of the information encoded in each specific patch token to recognize a place. Note that while computing these attention masks, the concatenated patch tokens $\mathbf{P}$ are used (see Fig. 3). Formally:

$$\mathbf{a_i} = \text{softmax}(\mathbf{P}\mathbf{W_i^a}) \in \mathbb{R}^{N \times 1}, \quad (3)$$

where $i \in \{L, M, H\}$ and $\mathbf{W_i^a} \in \mathbb{R}^{3D \times 1}$ maps a concatenated patch token to a scalar. Then, a multi-level attention map $\mathbf{A} \in \mathbb{R}^{N \times 1}$ is generated by merging the three attention maps.

$$\mathbf{A} = \text{MinMaxNorm}(\sum_i \text{MinMaxNorm}(\mathbf{a_i})). \quad (4)$$

3.3. Final Image Representations

Key-Patch Descriptors. In theory, output tokens from any Transformer layer can be used as patch-level descriptors to perform geometrical verification. In practice, we choose the mid-level patch tokens ($\mathbf{P_M}$) which give the most stable result in experiments. Patches where attention score $\mathbf{A}$
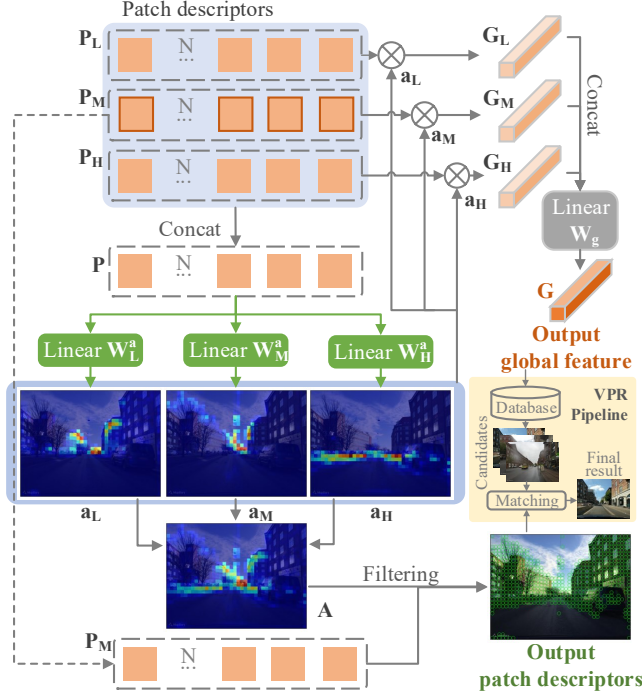
Figure 3. **Multi-level attention aggregation.** Given patch tokens from three different layers of Transformer, three attention maps are generated by applying a linear projection on their concatenation. The multi-level global feature is generated by combining single level global features, which are computed by summarizing the patch tokens weighted by the corresponding attention maps. Multi-level attention maps are further fused and used to select task-relevant patch descriptors. In VPR pipeline, global representations are used to retrieve candidates by nearest neighbour searching, while patch descriptors are used to perform geometrical verification to re-rank these candidates.

greater than a threshold τ are defined as key-patches, and their corresponding descriptors are used in final geometrical verification stage.

Global Image Features. As illustrated in Fig. 3, global feature in a single level $\mathbf{G_i}$ is computed by aggregating patch tokens $\mathbf{P_i}$ which are weighted by the corresponding attention map $\mathbf{a_i}$:

$$\mathbf{G_i} = \mathbf{a_i}^T \mathbf{P_i} \in \mathbb{R}^D. \quad (5)$$

Multi-level global image feature $\mathbf{G}^* \in \mathbb{R}^{3D}$ is defined as a concatenation of $\mathbf{G_L}$, $\mathbf{G_M}$ and $\mathbf{G_H}$, and then final global representation $\mathbf{G}$ is obtained by post-processing $\mathbf{G}^*$:

$$\mathbf{G} = \text{L2Norm}(\text{L2Norm}(\mathbf{G}^*)\mathbf{W_g}), \quad (6)$$

where learnable matrix $\mathbf{W_g} \in \mathbb{R}^{3D \times D}$ is used for dimensionality reduction.

3.4. Training Strategy

At first, the feature extraction backbone (CNN and Transformer) is pre-trained jointly on Places365 [64], an

image classification dataset containing 1.8 million images from 365 scene categories. The [class] token from the last Transformer layer is followed by a fully connected layer for classification.

Then, the pre-trained model is transferred to image retrieval task by removing the classification layer and adding the attention and the dimensionality reduction modules. The commonly used triplet margin loss [44] is adopted to be the training objective, defined as:

$$L(\mathbf{G^q}, \mathbf{G^p}, \mathbf{G^n}) = \max(d(\mathbf{G^q}, \mathbf{G^p}) - d(\mathbf{G^q}, \mathbf{G^n}) + m, 0), \quad (7)$$

where $\mathbf{G^q}$, $\mathbf{G^p}$ and $\mathbf{G^n}$ are global features of query, positive and negative samples. The margin m is a constant hyper-parameter. Parameters in attention and dimensionality reduction modules (W_i^a and W_g) are initialized by training for a few epochs on a large scale VPR dataset, Mapillary Street Level Sequences (MSLS) [58] training set, with frozen backbone parameters.

After initialization, the whole TransVPR can be further finetuned in an end-to-end fashion on VPR datasets.

4. Experiments

In this section, we evaluate the proposed TransVPR model on several benchmark datasets compared with some state-of-the-art VPR methods. We give the details of experimental settings, datasets, evaluation metrics, and compared methods in the following.

4.1. Implementation Details

Model Settings. TransVPR is implemented in PyTorch framework. The base TransVPR model contains six transformer encoder layers for feature aggregation. The latent embedding dimension D of the Transformer is 256. Without loss of generality, output tokens from the second, forth and sixth transformer layers are selected as $\mathbf{P_L}$, $\mathbf{P_M}$ and $\mathbf{P_H}$. The total parameter size of TransVPR is 19.86MB. The key-patch filtering threshold τ is set to 0.02 in practice. The patch size on the original image is set to 16×16. The dimensionality of output patch-level and global features are all set to 256.

In geometrical verification, given an image pair, their key-patch descriptors are matched in a brute-force manner. Cross checking is performed to ensure that matched descriptors are mutual nearest neighbors. The image similarity is defined as the number of inliers when estimating the homography based on the matched patches with RANSAC algorithm. Maximum allowed reprojection error of inliers is set to 1.5 times of the patch size.

Training. We finetuned the pre-trained TransVPR model on MSLS training set and pittsburgh 30k (Pitts30k) [55] training set. The former aims to deal with evaluation datasets containing diverse scenes (MSLS and Nordland

Dataset	Environment			Variation				
	Urban	Suburban	Natural	Viewpoint	Day/night	Weather	Seasonal	Dynamic
MSLS [58]	✓	✓	✓	+	+	+	+	+
Nordland [33,50]		✓	✓	−	−	−	+	−
Pitts30k [55]	✓			+	−	−	−	+
RobotCar-S2 [28,42]	✓			+	+	+	+	+

Table 1. **Summary of datasets used for evaluation.** $+$ indicates that the dataset contains the particular environmental variation, and $-$ is the opposite.

[33, 50] datasets) while the latter is particularly for urban scenes (Pitts30k and Robotcar Seasons V2 [28, 42] datasets). In MSLS training set, both GPS coordinates and compass angles are provided, so the positive sample is selected as the image with the most similar field of view to the query. For Pitts30k dataset where the angle labels are not given, the weakly supervised positive mining strategy proposed in [2] is adopted. Hyper-parameters and further details for training are presented in Supplementary Material.

4.2. Datasets

We evaluated our model on several public benchmark datasets: MSLS [58], Nordland [33, 50], Pitts30k [55] and RobotCar Seasons v2 (RobotCar-S2) [28, 42]. All of these datasets contain some challenging environmental variations. Tab. 1 summarizes the qualitative nature of them. More details of dataset usage are given in Supplementary Material. All images are resized to 640×480 while evaluation.

4.3. Metrics

For MSLS, Nordland and Pitts30k datasets, we use Recall@N metric which computes the percentage of query images that are correctly localized. A query image is considered as correctly localized if at least one of the top N ranked reference images is within a threshold distance from the ground truth location of the query. Default threshold definitions are used for all datasets [33, 55, 58].

For RobotCar-S2 dataset, we follow [19] and directly use the pose of the best matched reference image as the estimated pose of the query without calculating explicit 6-DOF poses. Recall@1 scores under three default error tolerances are used as evaluation metrics.

4.4. Compared Methods

We compared TransVPR against several state-of-the-art algorithms, including two VPR methods based on nearest-neighbor searching using global image representations: **NetVLAD** [2] and **SFRS** [17], and two models which extract both global and patch descriptors for two-stage
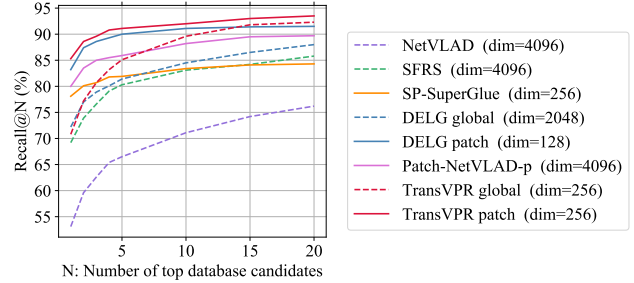

Figure 4. **Recall@N curves on MSLS val dataset.** Retrieval results using only global representation are depicted in dotted line, while results after re-ranking are depicted in solid line. TransVPR achieves the best performance in both global retrieval stage and re-ranking stage.

pipeline (*i.e.*, retrieval and re-ranking): **Patch-NetVLAD** [19] and **DELG** [6]. For Patch-NetVLAD, we tested both its speed-focused and performance-focused configurations, denoted as Patch-NetVLAD-s and Patch-NetVLAD-p respectively. In addition, we also compared against a strong hybrid baseline, **SP-SuperGlue**, which re-ranks NetVLAD retrieved candidates by using SuperGlue [41] matcher to match SuperPoint [12] patch-level descriptors. For all two stage methods, top-100 images retrieved by global features are further re-ranked by geometrical verification results. More installation details of the compared methods are explained in the Supplementary Material.

5. Results and Discussion

5.1. Quantitative Results

The quantitative results of TransVPR compared with other approaches are shown in Tab. 2. Our TransVPR convincingly outperforms all compared methods on MSLS validation, MSLS challenge and Nordland datasets, with an absolute increase on Recall@1 of 3.6%, 11.7% and 7.5% respectively in compared with the best baseline, DELG. It also achieves competitive results on Pitts30k and Robotcar-S2 datasets. Note that when training TransVPR on Pitts30k dataset, we only used the weakly supervised learning strategy in [2], and we can expect further boost of TransVPR performance using the fine-grained supervision proposed by [17]. Taking the average of all datasets, our full model surpasses the global feature retrieval based methods by a large margin, and outperforms the two-stage approaches, SP-SuperGlue, DELG and Patch-NetVLAD, with absolute gains of 10.8%, 5.9% and 7.2% on Recall@1 score.

The Recall@N curves of all methods including global retrieval results and re-ranking results are plotted in Fig. 4. TransVPR also achieves the best result in global retrieval stage. Note that among all compared methods, TransVPR is the only method which selectively integrates task-relevant information when generating global representations.

Method	MSLS val			MSLS challenge			Nordland test			Pitts30k test			Robotcar-S2 test		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	.25m/2°	.5m/5°	5.0m/10°
NetVLAD [2]	53.1	66.5	71.1	35.1	47.4	51.7	7.7	13.7	17.7	81.9	91.2	93.7	5.6	20.7	71.8
SFRS [17]	69.2	80.3	83.1	41.5	52	56.3	18.8	32.8	39.8	89.4	94.7	95.9	8.0	27.3	80.4
TransVPR (w/o re-ranking)	70.8	85.1	89.6	48.0	67.1	73.6	15.9	38.6	49.4	73.8	88.1	91.9	2.9	11.4	58.6
SP-SuperGlue [12,41]	78.1	81.9	84.3	50.6	56.9	58.3	29.1	33.5	34.3	87.2	94.8	96.4	9.5	**35.4**	85.4
DELG [6]	83.2	90.0	91.1	52.2	61.9	65.4	51.3	66.8	69.8	**89.9**	**95.4**	**96.7**	2.2	8.4	76.8
Patch-NetVLAD-s [19]	77.8	84.3	86.5	48.1	59.4	62.3	34.9	49.8	53.3	87.5	94.5	96.0	2.7	8.9	33.9
Patch-NetVLAD-p [19]	79.5	86.2	87.7	48.1	57.6	60.5	46.4	58.0	60.4	88.7	94.5	95.9	9.6	35.3	**90.9**
TransVPR	**86.8**	**91.2**	**92.4**	**63.9**	**74.0**	**77.5**	**58.8**	**75.0**	**78.7**	89.0	94.9	96.2	**9.8**	34.7	80.0

Table 2. Comparison to state-of-the-art methods on benchmark datasets.

Figure 5. **Comparison of retrieval results on MSLS validation dataset.** In these challenging examples, TransVPR successfully retrieves the matching database image, while all other methods produce false results.

Method	Extraction latency (ms)	Matching latency (s)	Memory (MB)
NetVLAD [2]	17	–	–
SFRS [17]	203	–	–
SP-SuperGlue [12,41]	166	7.83	1.93
DELG [6]	197	36.04	0.37
Patch-NetVLAD-s [19]	63	1.73	1.82
Patch-NetVLAD-p [19]	1336	7.65	44.14
TransVPR (ours)	45	3.19	1.17

Table 3. Feature extraction time, descriptor matching time, and memory footprint of all methods. Latency is measured on an NVIDIA GeForce RTX 2080 Ti GPU. For global retrieval methods, matching latency and memory requirements are negligible.

5.2. Qualitative Results

Fig. 5 illustrates some retrieval and matching results of hard examples with challenging conditions. In these cases, TransVPR produces correct matches while all other methods fail. For example, observing the first and fourth rows, where there are severe viewpoint changes or occlusions caused by dynamic objects, TransVPR can successfully perform matching based on distinctive regions and avoid distracting areas. However, other methods show a tendency to retrieve images with similar global layout as the query.

To further have an intuitive interpretation of the semantic cues captured by multi-level attentions, some visualization examples of learned attention maps are presented in Fig. 6. This confirms attention maps from different levels tend to focus on areas with different semantic information. For example, a_L mainly focuses on the small objects and textural areas on the surface of buildings. a_M focuses on objects in the air, such as street lamps and tree canopies, while a_H outlines the contours of the ground and the lane lines. All these attention maps avoid distracting areas such as the sky, the ground, dynamic objects, and untextured walls, which may change over time or have no effect on recognizing a scene. Note that we did not add any semantic constraints during the training process. These semantic information can be learned automatically by the attention mechanism in TransVPR under only image-level supervision.
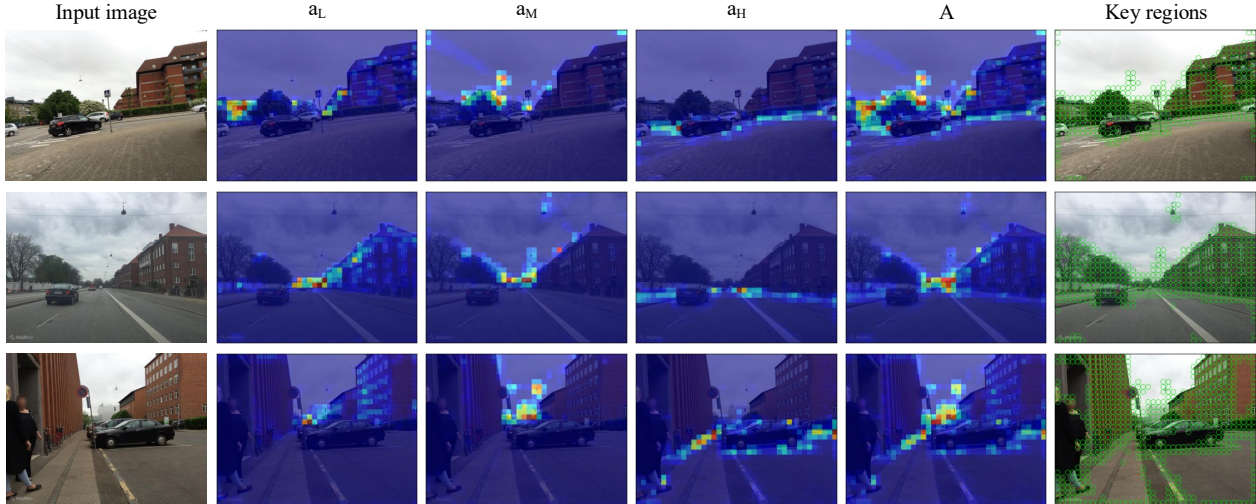
Figure 6. **Some visualisations on multi-level attentions.** From left to right: Input images, attention maps from each level, final attention maps, output key-patches. All images used here are from MSLS validation set and unseen during the network training stage. Different semantic cues are captured by attention maps from different levels. The network only pays attention to distinctive regions in the image and filters out confusing information.

5.3. Latency and Memory

In real-world VPR systems, latency and scalability are important factors that need to be considered. Tab. 3 shows the computational time and memory requirements for all compared techniques to process a single query image. TransVPR is 4.4 times and 29.7 times faster than DELG and Patch-NetVLAD-p in terms of feature encoding, when 11.3 times and 2.4 times faster than them in terms of spatial matching.

The memory footprint of TransVPR is 1.17 MB per image, the same order as SP-SuperGlue and Patch-NetVLAD-s. Considering patch features account for the main part of storage, using sparse and low-dimensional patch features can reduce the memory cost substantially. Patch-NetVLAD-p has an extremely large memory footprint due to its multi-scale features and high dimensionality (dim = 4096), while TransVPR requires less memory with relatively low-dimensional patch features (dim = 256).

5.4. Ablations and Analysis

We conduct several ablation experiments to further validate the design of TransVPR.

Choice of Patch Descriptor Sets. In Fig. 7, we show the performance of TransVPR when using different patch descriptor sets. Patch descriptors outputted from Transformer layers significantly outperform raw patch descriptors, demonstrating that the global contextual information encoded in Transformer tokens can improve the patch representation. The performances of patch tokens are similar no matter which Transformer layer they are from. Because despite there is no patch-level supervision while training,
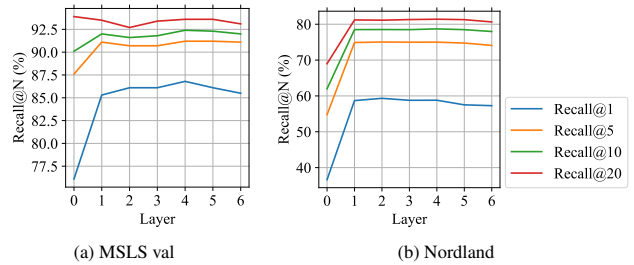

(a) MSLS val (b) Nordland

Figure 7. **Ablations of local descriptor set selection.** Recall performance of TransVPR with local descriptors from varying Transformer layer. Patch descriptors from any Transformer layer have similar performances and significantly outperform raw patch descriptors (*i.e.*, layer 0). Besides, a slight degradation of performance is observed at the last layer.

the patch tokens retain some locality which ensures the accuracy of spatial matching. It might because the residual connections in Transformers give the output tokens the ability to retain original information. However, there is a slight decay of performance when using patch tokens from the last layer which may contain relatively more contextual information.

Multi-Level Attentions for Key-Patch Detection. To verify the effectiveness of key-patch detection using the fused multi-level attention mask $\mathbf{A}$, we evaluated the TransVPR performance using patch descriptors ($\mathbf{P_M}$) without filtering and with filtering by each individual attention mask or by their combinations. The results are shown in Tab. 5. While using individual attention masks, the low-level attention mask $\mathbf{a_L}$ achieves the best performance, and the performance of $\mathbf{a_H}$ is dramatically poor. It indicates

	Method	MSLS val			Nordland test			Pitts30k test			Robotcar-S2 test		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	.25m/2°	.5m/5°	5.0m/10°
Global retrieval	sL-sATT	69.2	84.6	88.9	13.1	33.9	45.5	68.6	85.2	90.3	2.2	10.6	55.6
	mL-sATT	70.7	84.3	88.0	13.7	34.7	45.6	71.2	86.5	90.8	2.5	11.0	54.9
	mL-mATT-plain	**71.5**	**85.7**	**89.9**	13.2	32.7	43.0	71.2	87.0	91.3	**3.3**	**12.0**	56.2
	mL-mATT-standard	70.8	85.1	89.6	**15.9**	**38.6**	**49.4**	**73.8**	**88.1**	**91.9**	2.9	11.4	**58.6**
Re-ranking	sL-sATT	87.4	92.7	93.2	54.0	70.1	73.7	87.4	94.0	95.3	9.4	32.3	78.1
	mL-sATT	**87.7**	**91.5**	**93.0**	55.7	70.8	74.2	88.6	94.7	96.0	9.4	33.0	77.3
	mL-mATT-plain	84.7	89.6	91.5	54.1	68.4	71.7	88.1	94.3	95.5	9.5	34.2	78.3
	mL-mATT-standard	86.8	91.2	92.4	**58.8**	**75.0**	**78.7**	**89.0**	**94.9**	**96.2**	**9.8**	**34.7**	**80.0**

Table 4. **Ablations on multi-level attention aggregation strategy.** The proposed TransVPR configuration (mL-mATT-standard) achieves the best results.

Attention mask	MSLS val			Nordland test		
	R@1	R@5	R@10	R@1	R@5	R@10
None	81.2	87.6	90.1	58.8	73.9	77.7
a_L	86.4	91.1	92.2	55.1	72.9	76.9
a_M	84.1	90.4	91.8	47.8	69.1	74.3
a_H	61.2	77.3	82.8	23.5	40.7	48.6
a_L & a_M	**86.9**	90.9	91.9	57.4	74.3	77.7
A	86.8	**91.2**	**92.4**	**58.8**	**75.0**	**78.7**

Table 5. Performance of TransVPR when using different combinations of attention masks learned from multiple Transformer levels to select key-patch descriptors. The fused multi-level attention mask **A** performs best.

that building surfaces and fixed objects contribute the most to place recognition. Attention mask combining a_L and a_M achieves better result than using any single attention mask, but its performance is still inferior than **A**. Besides, matching using all patch descriptors without filtering leads to weak performance compared with **A**.

Multi-Level Attention Aggregation Strategy. We study how the proposed multi-level attention aggregation strategy affects the model performance by comparing the standard TransVPR with three degenerate configurations:

- *Multi-level & multiple attention maps & plain connection* (mL-mATT-plain). We remove the concatenation operation of multi-level patch tokens before computing attention maps. The three attention maps are computed only using patch tokens from the same level.
- *Multi-level & single attention map* (mL-sATT). Instead of estimating three attention maps separately, a single attention map **A** is computed based on concatenated patch tokens **P**. The global feature is expressed by the summation of **P** weighted by **A**.
- *Single level & single attention map* (sL-sATT). Only patch tokens from the last Transformer layer are used to estimate both the attention map and the global feature.

Detail architectures of these configurations are illustrated in Supplementary Material. Tab. 4 presents the eval-

uation results. There is in general a performance degradation from the standard TransVPR (mL-mATT-standard) to mL-sATT and then to sL-sATT, and mL-mATT-standard largely outperforms mL-mATT-plain on all datasets after re-ranking. Besides, the standard TransVPR has a better generalization ability on datasets with data distribution very different from the training set. These results demonstrate the effectiveness of combining multi-level information across Transformer layers and estimating separate attention maps for each level. In addition, results of all configurations are significantly improved by re-ranking using key-patch descriptors, especially on Nordland dataset which suffers severe perceptual aliasing and thus relies more on fine-grained spatial matching.

6. Conclusion

In this work, we have designed a novel vision Transformer-based place recognition model, TransVPR, which jointly extracts distinctive global and patch-level image features by aggregating multi-level attentions. All components of TransVPR are integrated in a single and lightweight network, enabling end-to-end optimization with image-level supervision. TransVPR outperforms some state-of-the-art VPR techniques on several benchmark datasets and achieves superior trade-off between accuracy and efficiency. Ablation results further verify the effectiveness of the design of our model.

This demonstration of TransVPR is limited in VPR tasks, where the major limitation is that the camera localization would be not precise enough when reference images are sparse. Therefore, one research topic of future work is to estimate camera poses in a regression framework by exploiting the TransVPR descriptors.

Acknowledgement

References

[1] Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Trans. Robot.*, 24(5):1027–1037, 2008. 2

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016. 1, 2, 5, 6

[3] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *CVPR*, pages 1578–1585, 2013. 2

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006. 2

[5] Luis G Camara and Libor Přeučil. Visual place recognition by spatial matching of high-level cnn features. *Robot. Autom. Syst.*, 133:103625, 2020. 2

[6] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, pages 726–743, 2020. 1, 2, 5, 6

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[8] 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 *IEEE Int. Conf. Robot. Autom.*, pages 3223–3230, 2017. 1

[9] Zetao Chen, Lingqiao Liu, Inkyu Sa, Zongyuan Ge, and Margarita Chli. Learning context flexible attention model for long-term visual place recognition. *Robot. Autom. lett.*, 3(4):4015–4022, 2018. 2

[10] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *ECCV*, volume 1, pages 1–2, 2004. 2

[11] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *Int. J. Rob. Res.*, 27(6):647–665, 2008. 2

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, pages 224–236, 2018. 1, 2, 5, 6

[13] 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 *ICLR*, 2020. 2, 3

[14] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, pages 8092–8101, 2019. 1, 2

[15] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021. 2

[16] 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

[17] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, pages 369–386, 2020. 2, 5, 6

[18] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 124(2):237–254, 2017. 1, 2

[19] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, pages 14141–14152, 2021. 1, 2, 5, 6

[20] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010. 2

[21] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE TPAMI*, 34(9):1704–1716, 2011. 1, 2

[22] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *ECCV*, pages 685–701, 2016. 2

[23] Ahmad Khaliq, Shoaib Ehsan, Zetao Chen, Michael Milford, and Klaus McDonald-Maier. A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes. *IEEE Trans. Robot.*, 36(2):561–569, 2019. 1, 2

[24] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *CVPR*, pages 3251–3260, 2017. 2

[25] Jana Košecká, Fayin Li, and Xialong Yang. Global localization and relative positioning based on scale-invariant keypoints. *Robot. Autom. Syst.*, 52(1):27–38, 2005. 2

[26] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999. 1, 2

[27] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Trans. Robot.*, 32(1):1–19, 2015. 1, 2

[28] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *Int. J. Rob. Res.*, 36(1):3–15, 2017. 5

[29] Eva Mohedano, Kevin McGuinness, Noel E O'Connor, Amaia Salvador, Ferran Marques, and Xavier Giró-i Nieto. Bags of local convolutional features for scalable instance search. In *ACM Int. Conf. Multimedia Retr*, pages 327–331, 2016. 2

[30] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robot.*, 33(5):1255–1262, 2017. 2

[31] Paul Newman and Kin Ho. Slam-loop closing with visually salient features. In *IEEE Int. Conf. Robot. Autom.*, pages 635–642, 2005. 2

[32] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, pages 3456–3465, 2017. 1, 2

[33] Daniel Olid, José M. Fácil, and Javier Civera. Single-view place recognition under seasonal changes. In *PPNIV Workshop at IROS 2018*, 2018. 5

[34] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. In *NeurIPS*, 2018. 2

[35] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8, 2007. 2

[36] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, pages 3384–3391, 2010. 2

[37] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 41(7):1655–1668, 2018. 1, 2

[38] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, pages 5107–5116, 2019. 1

[39] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pages 2564–2571, 2011. 2

[40] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, pages 12716–12725, 2019. 1, 2

[41] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 2, 5, 6

[42] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018. 5

[43] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *CVPR*, pages 1822–1830, 2017. 2

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4

[45] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *CVPR*, pages 2556–2565, 2019. 1

[46] Stephen Se, David Lowe, and Jim Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int. J. Rob. Res.*, 21(8):735–758, 2002. 2

[47] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *CVPR*, pages 11651–11660, 2019. 2

[48] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, pages 118–126, 2015. 2

[49] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, volume 3, pages 1470–1470, 2003. 2

[50] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *IEEE Int. Conf. Robot. Autom. Worksh.*, 2013. 5

[51] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, pages 7199–7209, 2018. 1, 2

[52] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *CVPR*, pages 5109–5118, 2019. 1, 2

[53] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. pages 1–12, 2016. 2

[54] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, pages 1808–1817, 2015. 1

[55] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *CVPR*, pages 883–890, 2013. 4, 5

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3

[57] Yannick Verdie, Kwang Yi, Pascal Fua, and Vincent Lepetit. Tilde: A temporally invariant learned detector. In *CVPR*, pages 5279–5288, 2015. 2

[58] 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 *CVPR*, pages 2626–2635, 2020. 4, 5

[59] Liqi Yan, Yiming Cui, Yingjie Chen, and Dongfang Liu. Hierarchical attention fusion for geo-localization. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 2220–2224, 2021. 2

[60] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, pages 467–483, 2016. 1, 2

[61] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *CVPR*, pages 4709–4717, 2017. 2

[62] Fangming Yuan, Peer Neubert, Stefan Schubert, and Peter Protzel. Softmp: Attentive feature pooling for joint local feature detection and description for place recognition in changing environments. In *IEEE Int. Conf. Robot. Autom.*, pages 5847–5853, 2021. 2

[63] Linguang Zhang and Szymon Rusinkiewicz. Learning to detect features in texture images. In *CVPR*, pages 6325–6333, 2018. 2

[64] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 4