

AANet: Aggregation and Alignment Network with Semi-hard Positive Sample Mining for Hierarchical Place Recognition

Feng Lu, Lijun Zhang, Shuting Dong, Baifan Chen and Chun Yuan

Abstract—Visual place recognition (VPR) is one of the research hotspots in robotics, which uses visual information to locate robots. Recently, the hierarchical two-stage VPR methods have become popular in this field due to the trade-off between accuracy and efficiency. These methods retrieve the top-k candidate images using the global features in the first stage, then re-rank the candidates by matching the local features in the second stage. However, they usually require additional algorithms (e.g. RANSAC) for geometric consistency verification in re-ranking, which is time-consuming. Here we propose a **Dynamically Aligning Local Features (DALF)** algorithm to align the local features under spatial constraints. It is significantly more efficient than the methods that need geometric consistency verification. We present a unified network capable of extracting global features for retrieving candidates via an aggregation module and aligning local features for re-ranking via the DALF alignment module. We call this network **AANet**. Meanwhile, many works use the simplest positive samples in triplet for weakly supervised training, which limits the ability of the network to recognize harder positive pairs. To address this issue, we propose a **Semi-hard Positive Sample Mining (ShPSM)** strategy to select appropriate hard positive images for training more robust VPR networks. Extensive experiments on four benchmark VPR datasets show that the proposed **AANet** can outperform several state-of-the-art methods with less time consumption. The code is released at <https://github.com/Lu-Feng/AANet>.

I. INTRODUCTION

Visual place recognition (VPR) is a fundamental and challenging task for the long-term operation of mobile robots. Its goal is to determine whether the robot has visited the current place and obtain the geographical location by matching the current query image with a set of images of known places. There are two key issues in implementing a robust VPR system: 1) Images taken at the same place may vary greatly over time due to changes in conditions (e.g. illumination and weather) and viewpoints. 2) Images taken at different places may show high similarity (i.e. perceptual aliasing [1]).

VPR is commonly addressed by utilizing image retrieval and matching methods [2], [3]. And the image descriptors to

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SZSTC Grant(JCYJ20190809172201639, WDZC20200820200655001), Shenzhen Key Laboratory (ZDSYS20210623092001004), the Project of Peng Cheng Laboratory (PCL2021A07). (corresponding author: Chun Yuan)

F. Lu, S. Dong and C. Yuan are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen, China. {lf22@mails, dst21@mails, yuanc@sz}.tsinghua.edu.cn.

L. Zhang is with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China. zhanglijun@cigit.ac.cn.

B. Chen is with the School of Automation, Central South University, Changsha 410083, China. chenbaifan@csu.edu.cn.

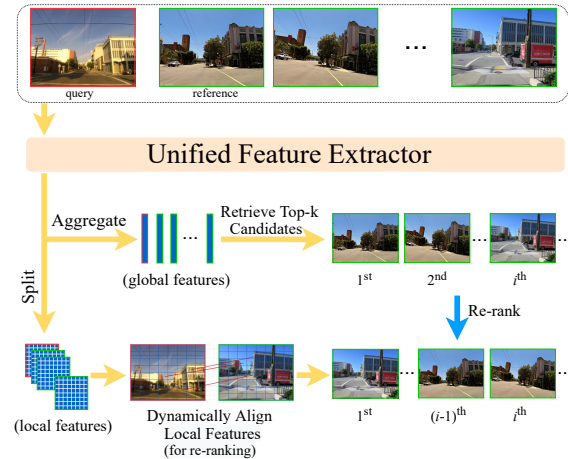


Fig. 1. The hierarchical retrieval process with our network. The unified feature extractor is used to extract feature maps, which can yield global features and local features by aggregating and splitting, respectively. We first use global features for retrieving top-k candidates, then re-rank these candidates by utilizing the proposed DALF method to align local features.

represent places can be roughly divided into two categories: global features and local features. Some global features are also obtained by aggregating local features into a compact feature vector [4]–[6]. Such features are strongly robust to viewpoint change and suitable for fast retrieval of places. However, they ignore the spatial geometrical information of aggregated local features, and thus tend to suffer from perceptual aliasing. Meanwhile, direct matching of local features between image pairs yields better performance, but is time-consuming. A compromise pipeline [3], [7], [8] is to search for top-k candidate images using global features, then re-rank candidates by matching local features. However, these methods usually employ additional algorithms (like RANSAC [9]) for geometric consistency check in the re-ranking stage, which can also entail expensive time overhead.

In our previous STA-VPR work [10], we proposed an adaptive Dynamic Time Warping (DTW) method to dynamically align the local features without the expensive geometric consistency verification. It provided a new idea to match the local feature under spatial constraints for boosting the performance of VPR system. Shen et al. proposed BS-DTW to further improve this alignment method in the TCL study [11], and used it to re-rank candidates retrieved by global features. However, both of them align features only in the horizontal direction, and the granularity of the local feature is coarse (we call it the regional feature). For more precise alignment, we propose a Dynamically Aligning Local Features (DALF) algorithm to extend the alignment to both horizontal and vertical directions with high computational efficiency, and

align local features with a finer granularity to achieve better results. The proposed DALF module is integrated into a network together with an aggregation module. This network is named AANet (Aggregation and Alignment Network). The retrieval process with AANet is shown in Fig. 1.

For network training, the images of general large-scale VPR datasets are labeled with noisy GPS coordinates to provide weak supervision. Nevertheless, even geographically close images do not always have enough covisibility region because they may be captured from opposite viewpoints or exist large occlusion. Therefore, GPS tags could only assist in discovering potential positive samples with label noise. To find the true positive sample, most works [2], [12] chose the best matching positive image in feature space (i.e. easiest image among the potential positive samples) for training. However, this limits the ability of the trained network to identify harder positive pairs. To address it, we propose a Semi-hard Positive Sample Mining (ShPSM) strategy. It combines global distance ranking and local distance ranking to select appropriate difficult positive images for training a more robust network with weak supervision.

The main contributions of this paper are:

- 1) We propose a hierarchical VPR architecture AANet, which consists of an aggregation module to extract global features for retrieving candidates, and a DALF alignment module to align local features for re-ranking. The latter can directly align the patch-level local features without time-consuming geometric consistency check.
- 2) We propose an ShPSM strategy that can directly employ the global feature distance and local feature distance to mine semi-hard positive samples for weakly supervised training without additional parameters.
- 3) Extensive experiments show that our method outperforms several state-of-the-art methods. And it is more than two orders of magnitude faster than Patch-NetVLAD [7] that needs to use RANSAC for geometric consistency check.

II. RELATED WORK

One-Stage VPR: Most of the early VPR methods directly retrieved place images without re-ranking. We call these methods one-stage VPR. They typically used global features to represent images, which is obtained by aggregating local features or directly processing the whole image. Some aggregation algorithms, such as Bag of Words (BoW) [13] and Vector of Locally Aggregated Descriptors (VLAD) [4], [5], were used to aggregate local features like SURF [14], [15]. With the great success of deep learning on computer vision tasks, many VPR methods [2], [16]–[25] chose deep features to represent images for better performance. Likewise, some studies [2], [26] have incorporated these aggregation methods into CNN models. Nonetheless, these one-stage methods, which only use aggregated features, tend to suffer from perceptual aliasing due to the neglect of spatial information.

Two-Stage VPR: Recently, the two-stage strategy with re-ranking candidates for VPR (also known as Hierarchical VPR) has received more attention [3], [7], [8], [11], [27]–[29]. These methods typically used compact global features

of place images to retrieve top-k candidates, then re-ranked these candidates by matching the local features. The global features are typically obtained by aggregation methods, such as NetVLAD [2] or Generalized Mean (GeM) pooling [30]. And the re-ranking required geometric consistency verification [3], [7], [8] or homographic transformation estimation [27]. Most related to our proposed DALF method, TCL [11] used a BS-DTW method to dynamically align regional features for re-ranking. However, we align finer-grained local features both horizontally and vertically in a more concise way and achieve better performance.

Positive Samples Mining in Weakly Supervised VPR:

Many CNN architectures for VPR need to be trained on the large-scale place datasets, such as Pittsburgh [31] and MSLS [32]. These datasets usually provide weak supervision of geographic coordinates via noisy GPS tags. To avoid the positive samples used for training being false positives, NetVLAD [2] used the simplest top-1 positive sample for training, which was also adopted by some subsequent works [8], [12]. However, the network trained with the easiest positive samples cannot accurately identify difficult positive pairs with severe changes in viewpoint and condition. To address it, CRN [33] and SFRS [34] mined hard positive samples for training the VPR network. Our proposed ShPSM strategy has a similar motivation. However, CRN [33] required utilizing RANSAC [9] for geometric verification, while SFRS [34] required training the network in generations with self-predicted soft labels. Our method combines global distance and local distance to mine semi-hard positive samples without additional burden.

III. PROPOSED METHOD

A. Architecture Overview and Place Description

As shown in Fig. 2, the architecture of AANet consists of a global (aggregation) branch and a local (alignment) branch. They share the unified backbone to extract the $W \times H \times C$ -dimensional (width by height by channel) feature map. However, they are responsible for computing the global distance to search top-k candidate places and the local distance to re-rank these candidates, respectively. The global branch aggregates the dense feature map into a compact vector (i.e. global features) by an aggregation module. Then, the global distance is calculated using the L2 distance between the global features of two place images. The local branch splits the $W \times H \times C$ -dimensional dense feature map into $W \times H$ separate C -dimensional patch features (i.e. local features). And we propose the DALF algorithm to quickly align the local features between images and obtain the local distance.

In the training stage, the ShPSM strategy is proposed to select a semi-hard *positive* sample corresponding to the query image based on the global distance and local distance. Consistent with most previous works [2], [11], we directly use the global distance to select hard *negative* samples. The obtained triplets are used to compute the global loss and local loss to jointly optimize the network.

The details of the DALF algorithm and the ShPSM strategy will be introduced in the next two sections. Considering

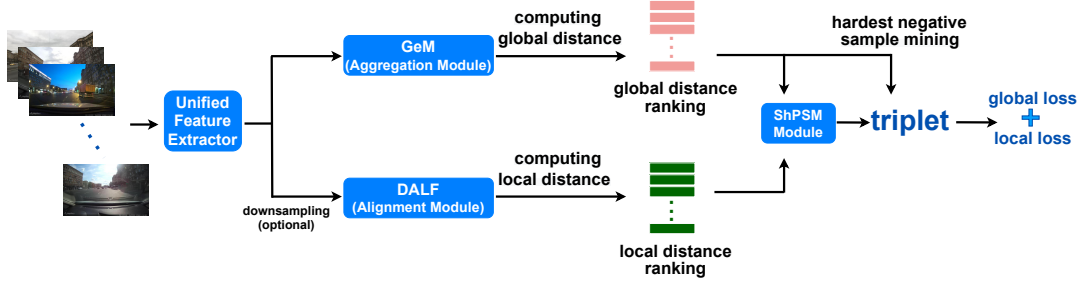


Fig. 2. The network architecture of the proposed AANet. The top is the global branch, and the bottom is the local branch.

that Vision Transformer can capture long-distance feature dependencies, we use the Compact Convolutional Transformer (CCT) [35] as backbones, whose output size is $M \times C$ -dimensional (576×384 -D, M is the number of patch tokens). Then the output is reshaped as a $W \times H \times C$ -dimensional ($24 \times 24 \times 384$ -D) feature map to restore spatial position. In the global branch, we use GeM [30] to aggregate the feature map to get the 384-D global feature. In the local branch, we downsample the feature map through 3×3 max pooling to get an $8 \times 8 \times 384$ tensor, which can be regarded as 8×8 separate 384-D local features. The L2 normalization is used for both global features and local features.

B. Dynamically Aligning Local Features Algorithm

In STA-VPR [10], we used adaptive DTW to align regional features to improve the robustness of the deep features against viewpoint changes. However, this work only aligned regional features in the horizontal direction. To extend the alignment to the vertical direction, we propose a DALF algorithm to calculate the local feature distance between images, which is able to align local features in both horizontal and vertical directions with high computational efficiency. The following first describes how to align regional features in the horizontal direction, then introduces how to get the final local feature alignments through the horizontal and vertical regional feature alignments.

Given the $8 \times 8 \times 384$ -D feature map of an image, we can obtain 8 regional features by splitting it vertically. By flattening each regional feature, each image can be represented as a feature sequence of N ($N=8$) elements (keeping the order from left to right). Each element is a 3072-D vector.

For an image pair r and q (i.e. reference and query), the corresponding feature sequence can be denoted as RX ($rx_1, \dots, rx_i, \dots, rx_N$) and QX ($qx_1, \dots, qx_j, \dots, qx_N$), respectively. By calculating the L2 distance between all pairwise regional features in these two sequences, an $N \times N$ distance matrix \mathbf{D} is established. Formally, The (i, j) -element $d_{i,j}$ in matrix \mathbf{D} is calculated as

$$d_{i,j} = \|rx_i - qx_j\| \quad i, j \in \{1, 2, \dots, N\}. \quad (1)$$

The DTW [36], [37] algorithm aims to find an optimal warping path in the distance matrix \mathbf{D} that shortens the total distance. And the alignment between sequence RX and QX is expressed by the points in this warping path P :

$$P = \{p_1, p_2, \dots, p_k, \dots, p_K\} \quad (2)$$

where $N \leq K < 2N - 1$.

The path should meet three conditions:

- 1) **Boundary:** $p_1 = (1, 1)$ and $p_K = (N, N)$.
- 2) **Continuity:** Given $p_k = (c, d)$ and $p_{k-1} = (c', d')$, then $c - c' \leq 1$ and $d - d' \leq 1$.
- 3) **Monotonicity:** Given $p_k = (c, d)$ and $p_{k-1} = (c', d')$, then $c - c' \geq 0$ and $d - d' \geq 0$.

Subject to the continuity and monotonicity conditions, if the warping path has passed through the point (i, j) , then the next point must be one of the following three cases: $(i+1, j)$, $(i, j+1)$, and $(i+1, j+1)$.

To find the optimal warping path, DTW acquires a cumulative distance matrix \mathbf{S} by dynamic programming. Its (i, j) -element $s_{i,j}$ is the cumulative distance of the optimal path from $(1,1)$ to (i, j) , which is computed as

$$s_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ d_{i,j} + s_{i,j-1} & i = 1, j \in \{2, \dots, N\} \\ d_{i,j} + s_{i-1,j} & i \in \{2, \dots, N\}, j = 1 \\ d_{i,j} + \min\{s_{i-1,j-1}, s_{i-1,j}, s_{i,j-1}\} & i, j \in \{2, \dots, N\}. \end{cases} \quad (3)$$

In STA-VPR, adaptive DTW was proposed to avoid incorrect alignment. It needs to add an adaptive parameter in Eq. 3, which is calculated by estimating the degree of viewpoint change. Here, we propose a normalized DTW algorithm to align regional features without computing this parameter. It calculates the cumulative distance as follows:

$$s_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ d_{i,j} + s_{i,j-1} & i = 1, j \in \{2, \dots, N\} \\ d_{i,j} + s_{i-1,j} & i \in \{2, \dots, N\}, j = 1 \\ d_{i,j} + \arg \min_{s_{i',j'}} \left(\frac{s_{i',j'}}{k_{i',j'}} \right) & i, j \in \{2, \dots, N\} \end{cases} \quad (4)$$

where $(i', j') \in \{(i-1, j-1), (i-1, j), (i, j-1)\}$ and $k_{i',j'}$ is the length (i.e. the number of the points passed) of the warping path from $(1,1)$ to (i', j') . That is, the normalized DTW uses the normalized cumulative distance instead of the cumulative distance when it selects the previous point in the recursive formula. This can also solve the misalignment problem caused by the vanilla DTW mentioned in STA-VPR [10].

After calculating the complete cumulative distance matrix \mathbf{S} , the alignment between sequence RX and QX is indicated by the warping path from the start point $(1, 1)$ to the end point (N, N) . We use X_align to record the alignment between sequences RX and QX . For example, if rx_1 is aligned with qx_1, qx_2 and qx_3 , we will get $X_align[1] = \{1, 2, 3\}$.

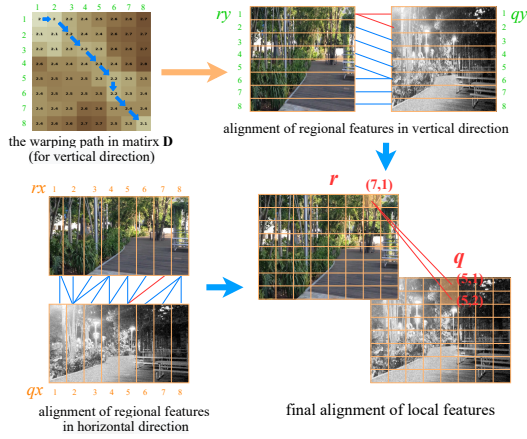


Fig. 3. An instance of computing the alignment between local features using the DALF algorithm. The regional feature rx_7 is aligned with qx_5 , and ry_1 is aligned with qy_1 as well as qy_2 , so local feature $r_{7,1}$ should align with $q_{5,1}$ and $q_{5,2}$.

Similarly, we can horizontally split the $8 \times 8 \times 384$ -D feature map to get the sequence RY ($ry_1, \dots, ry_i, \dots, ry_N$) and QY ($qy_1, \dots, qy_j, \dots, qy_N$), so as to align the regional features from the vertical direction. The Y_align is used to record the vertical alignment between two images.

Finally, if the regional feature rx_i is aligned with $qx_{i'}$, and ry_j is aligned with $qy_{j'}$, then we consider the local feature $r_{i,j}$ should align with $q_{i',j'}$ (see Fig. 3). The local distance of an image pair is represented by averaging the L2 distances between all aligned local feature pairs. That is

$$d_L(r, q) = \frac{\sum \|r_{i,j} - q_{i',j'}\|}{A} \quad i, j \in \{1, \dots, N\} \quad (5)$$

where $i' \in X_align[i]$, $j' \in Y_align[j]$, and A is the number of all aligned local feature pairs.

C. Semi-hard Positive Sample Mining Strategy

In the VPR training dataset, we can get a set of potential positive samples $\{p_i^q\}$ and a set of definite negative samples $\{n_j^q\}$ for each query image q according to GPS tags. In the NetVLAD study [2], the network was optimized using a triplet ranking loss. For a training tuple $(q, \{p_i^q\}, \{n_j^q\})$, this loss is defined as

$$L_\theta = \sum_j l(d_G(q, p_{i^*}^q) + m - d_G(q, n_j^q)) \quad (6)$$

where m is a constant parameter giving the margin. l is the hinge loss: $l(x) = \max(x, 0)$. d_G is the global distance. n_j^q is the hard negative sample selected based on global distance.

To ensure a sufficient covisibility region between the positive sample and the query image, the positive sample $p_{i^*}^q$ in NetVLAD is the image with the smallest global distance from the query image in the set $\{p_i^q\}$. Obviously, this strategy ignored the potential of hard but informative positive images. However, it is difficult to mine the hard positive sample from the set $\{p_i^q\}$ while ensuring that it is not a false positive sample. In this paper, we propose the ShPSM strategy to mine a semi-hard positive sample p_{sh}^q for training.

As shown in Fig. 4, we calculate the global distance and local distance between the query image and all potential

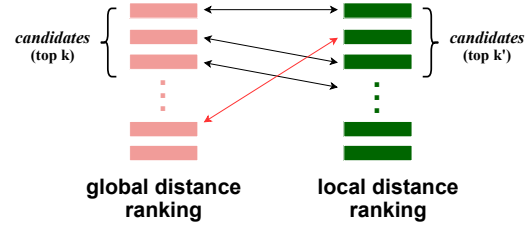


Fig. 4. The schematic of the ShPSM strategy. The sample corresponding to the red arrow is the semi-hard positive sample p_{sh}^q .

TABLE I
SUMMARY OF THE DATASETS.

Dataset	Description	Number		Variation	
		Database	Queries	Cond.	View.
MSLS-train	long-term	900k	500k	✓	✓
MSLS-val	urban, suburban	19k	11k	✓	✓
Pitts30k-train	urban	10k	7,416	✓	✓
Pitts30k-test	panorama	10k	6,816	✓	✓
GP	urban, campus	200	200	✓	✓
Nordland	seasonal	3,450	3,450	✓	×

positive samples to get the global distance ranking and local distance ranking. Meanwhile, we consider the potential positive samples corresponding to the top k distances in the ranking list as *candidates*. That means it already has enough covisibility region with the query sample and can be used as a training sample. Then, we evaluate the ranking difference between the two ranking lists for each *candidate*. The one with the largest difference is selected as p_{sh}^q . This implies that the sample p_{sh}^q ranks high in one ranking list, but relatively low in another. That is, it is an easy sample when calculated with one distance, but a hard sample used another distance. We call such a sample the semi-hard positive sample (which is different from the semi-hard positive sample in some other works). Formally, Given a potential positive sample p_i^q , its global distance with the query image q in the global distance ranking is denoted as $G_{rank}(d_G(q, p_i^q))$. Similarly, the local distance in the local distance ranking is $L_{rank}(d_L(q, p_i^q))$. Then the semi-hard positive sample is calculated as follows:

$$p_{sh}^q = \arg \max_{p_i^q} |G_{rank}(d_G(q, p_i^q)) - L_{rank}(d_L(q, p_i^q))|$$

$$s.t. \quad G_{rank}(d_G(q, p_i^q)) \leq k \quad or \quad L_{rank}(d_L(q, p_i^q)) \leq k'. \quad (7)$$

We obtain the final loss function that integrates the global loss term L_g and local loss term L_l by weight λ as follows:

$$L = L_g + \lambda L_l \quad (8)$$

where

$$L_g = \sum_j l(d_G(q, p_{sh}^q) + m - d_G(q, n_j^q)) \quad (9)$$

and

$$L_l = \sum_j l(d_L(q, p_{sh}^q) + m - d_L(q, n_j^q)). \quad (10)$$

IV. EXPERIMENT

A. Datasets and Performance Evaluation

Our experiments are conducted on four VPR benchmark datasets. The key information of these datasets is summarized

in Table I. Here we briefly describe the usage of these datasets. **Mapillary Street Level Sequences (MSLS)** [32] contains over 1.6 million images captured in both urban and suburban environments over seven years. Since its test set labels have not yet been publicly available, we test the model on its validation set as in [11], [38]. **Pittsburgh (Pitts30k)** [31] includes 30k reference images and 24k query images, and has been geographically divided into train/validation/test sets. We evaluate models on its test set. **Gardens Point (GP)** [17] contains images of two traverses during the day and night. We use the night-right (reference) and the day-left (query) images as test set. **Nordland** [39] consists of train and test set. We use the summer (reference) and winter (query) images from the down-sampled (224x224) test set in the following experiments.

We evaluate the recognition performance mainly based on Recall@N (R@N). It is the percentage of queries for which at least one of the top-N retrieved images is taken within a certain threshold of the query. The threshold is 25 meters for MSLS and Pitts30k, and ± 2 frames for GP and Nordland, following standard procedure [2], [11], [38].

B. Implementation Details

We use the CCT-14 model pre-trained on ImageNet [40] as the backbone. The 9th and subsequent transformer encoder layers are removed, and the layers before the 3rd encoder layer are frozen. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. The input image is resized to 384x384 resolution. Adam optimizer is used to train our models with the learning rate = 0.00001 and the batch size = 4. All test models are trained on MSLS, and the test model for Pitts30k is finetuned with random horizontal flipping and random resized cropping on Pitts30k-train. The model training process stops when R@1+R@5 on the validation set does not improve for 3 epochs. (But we only use 7k queries on MSLS to train the model for the GP dataset.) An epoch is defined as passing 40k queries for MSLS and 5k queries for pitts30k.

During training, the potential positive samples are the database images within a 10 meters radius from the query, while the definite negatives are those further than 25 meters. The number of hard negative samples in triplets is 2, which are selected from 1000 random negatives. We empirically set the weight $\lambda = 1$, the margin $m = 0.1$. The k and k' in Eq. 7 equal 30% of the number of all potential positives for MSLS, and equal 1 and 2 for Pitts30k. And we only re-rank the top-20 candidates to get the final results.

C. Ablation Study

To evaluate the effectiveness of the proposed DALF algorithm, the ShPSM strategy, as well as the joint optimization combining global loss L_g and local loss L_l , we perform a series of ablation experiments:

- **GeM(U)**: Untrained vanilla GeM.
- **GeM(V)**: Vanilla GeM trained with global loss.
- **GeM(J)**: Use joint loss ($L_g + L_l$) of AANet for training. But only use global features for direct retrieval.

TABLE II
RESULTS (R@1) OF DIFFERENT ABLATED VERSIONS.

Method	$L_g + L_l$	ShPSM	DALF	MSLS	Pitts30k	Nordland
GeM(U)	×	×	×	20.2	51.6	2.7
GeM(V)	×	×	×	72.4	82.5	40.7
GeM(J)	✓	×	×	75.6	83.7	52.3
GeM(J+S)	✓	✓	×	76.5	84.8	54.8
AANet(-S)	✓	×	✓	79.7	87.2	70.4
AANet	✓	✓	✓	80.1	88.0	72.9

- **GeM(J+S)**: Use joint loss ($L_g + L_l$) of AANet for training and the ShPSM strategy to mine hard positive samples. But only use global features for direct retrieval.
- **AANet(-S)**: Use AANet for Two-stage retrieval. But do not use ShPSM to mine hard positive samples.
- **AANet**: Our complete pipeline.

In this subsection, we mainly use R@1 for performance evaluation. The R@1 results are shown in Table II. We also provide the PR curves on the Nordland as shown in Fig 5. The results show that training with global loss significantly improves performance compared to off-the-shelf GeM(U). And the performance is further improved after using joint loss $L_g + L_l$ for optimization. This indicates that the combination of global loss yielded by the global (aggregation) branch and local loss yielded by the local (alignment) branch to optimize the unified backbone is successful. AANet(-S) always performs better than GeM(J), and AANet also always outperforms GeM(S), illustrating that the re-ranking method can improve performance with or without the ShPSM strategy. Likewise, GeM(J+S) always outperforms GeM(J), while AANet also always behaves better than AANet(-S). It shows that the performance can be improved by the ShPSM strategy on these datasets, with or without re-ranking. The above results indicate that the proposed DALF algorithm and ShPSM strategy work well in our AANet.

In addition, the improvements of our method on these datasets are different. The most significant improvement is achieved on the Nordland dataset, where AANet achieves more than 70% absolute improvement over GeM(U) and more than 30% absolute improvement over GeM(V). This is due to the extreme appearance changes caused by seasonal changing in the Nordland dataset. The network trained with hard positive samples can better cope with the appearance changes than that trained with easiest positive samples. Besides, we use patch-level local features to compute the local distance through aligning. In addition to solving the viewpoint change problem, this could also provide more detailed information (especially compared with keypoint-level local features), which is critical under severe appearance changes. Although the ShPSM strategy brings less improvement than using the DALF algorithm for re-ranking on these datasets, its process is concise and has no additional burden. It is still worthwhile to use the ShPSM strategy for training.

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

We evaluate the recognition performance by comparing AANet against several state-of-the-art VPR methods, including two VPR methods using global features for direct

TABLE III
COMPARISON TO STATE-OF-THE-ART METHODS ON FOUR BENCHMARK DATASETS.

Method	Dim	Time (s)	MSLS val			Pitts30k			Nordland			Gardens Point		
			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]	32768	0.025	60.8	74.3	79.5	81.9	91.2	93.7	7.7	13.7	17.7	38.0	74.5	83.5
GCL [41]	2048	0.061	78.3	85.0	86.9	75.6	91.5	92.6	27.2	45.4	53.7	42.0	79.5	88.0
AANet (w/o re-ranking)	384	0.014	76.5	87.4	90.3	84.8	93.4	95.5	54.8	77.9	85.7	54.0	89.0	93.5
SP-SuperGlue [42], [43]	256	-	78.1	81.9	84.3	87.2	94.8	96.4	29.1	33.5	34.3	-	-	-
Patch-NetVLAD-s [7]	512	0.90	77.0	84.2	86.9	87.1	94.0	95.6	34.9	49.8	53.3	71.5	93.5	96.5
Patch-NetVLAD-p [7]	4096	16.51	79.5	86.2	87.7	88.7	94.5	95.9	46.4	58.0	60.4	78.0	94.0	97.0
TCL [11]	384	-	78.7	82.5	85.3	90.5	95.9	97.5	43.5	51.7	54.3	80.5	95.0	97.0
AANet (ours)	384	0.053	80.1	88.9	91.0	88.0	94.2	95.8	72.9	87.6	90.6	77.0	97.0	99.5

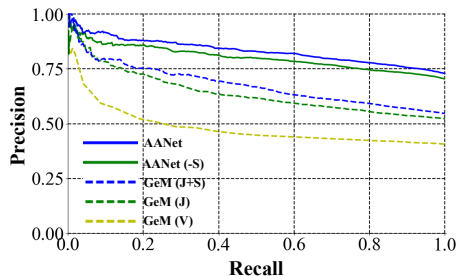


Fig. 5. Precision-recall curves on the Nordland dataset.

retrieval: NetVLAD [2] and GCL [41], and three two-stage VPR methods: SP-SuperGlue [42], [43], Patch-NetVLAD [7] and TCL [11]. The NetVLAD was trained on the Pitts250k. GCL (ResNet152-GeM-PCA) and TCL (DeiT-S, TCL-R100) follow the optimal testing configuration of the original work. For Patch-NetVLAD, our approach is compared with its two different versions, where Patch-NetVLAD-s is speed-focused and Patch-NetVLAD-p is performance-focused. SP-SuperGlue was first used for VPR in [7]. It retrieved the candidates by NetVLAD, then used the SuperGlue [43] matcher to match the SuperPoint [42] local descriptors for re-ranking. The quantitative results are shown in Table III.

Our method is denoted as AANet (w/o re-ranking) when it does not use re-ranking. Although we use 384-dimensional global features, which are more compact than those of NetVLAD and GCL, it performs better on all four datasets (except R@1 on MSLS) in comparison with these two methods. When compared with GCL, it has achieved 2.4%, 1.9%, 32.5%, and 9.5% absolute improvements for R@5 on the MSLS, Pitts30k, Nordland, and GP datasets, respectively.

AANet also achieves competitive results in comparison with two-stage methods that require re-ranking. It outperforms other methods on MSLS, Nordland, and GP (except R@1). The R@1 of AANet on the Pitts30k dataset is better compared to SP-SuperGlue and Patch-NetVLAD-s, but worse than Patch-NetVLAD-p, which used RANSAC for spatial verification. We think this is because the place images of Pitts30k have a large rotational variation. The DALF algorithm in AANet mainly considers horizontal and vertical shifts. It is not good at addressing rotational changes. Similar to this, Rapid Spatial Scoring used in Patch-NetVLAD-s also considers horizontal and vertical movements. But our method outperforms it on all datasets. Benefiting from our re-ranking method, more than absolute improvements of 18% for R@1 are achieved on both Nordland and GP datasets compared to

the case without re-ranking.

E. Runtime Analysis

We analyze the computational efficiency of the VPR methods by the running time (feature encoding time + feature matching time) of each query on the Pitts30k test dataset. We compare the retrieval time of the first stage of our method with two direct retrieval approaches (NetVLAD and GCL) and the overall time consumption with Patch-NetVLAD (a two-stage method). The results are also shown in Table III. Since our method uses 384-D global features in the first stage, whose size is much smaller than that used in NetVLAD and GCL, it has a significant advantage in computational efficiency (it is also due to the different backbones). Moreover, our method is more than two orders of magnitude faster than Patch-NetVLAD-p, which uses RANSAC for geometrical consistency verification. Although a fast spatial verification method (Rapid Spatial Scoring) is proposed in Patch-NetVLAD-s, it is still more time-consuming than our method. It is worth mentioning that to extend the horizontal regional features alignment in STA-VPR [10] and TCL [11] to both horizontal and vertical directions, a naive solution is to first align the local features in all pairs of regional features in the vertical direction using the DTW-based algorithm (requiring $N \times N$ times alignments), and then align regional features in the horizontal direction. In this way, a total of $N \times N + 1$ times DTW-based alignments are performed. However, our proposed DALF method only needs to perform the DTW-based alignment process twice, which greatly improves the alignment efficiency.

V. CONCLUSION

Here we proposed the hierarchical VPR architecture AANet that utilizes the aggregation module to extract global features for retrieving candidate places and the DALF alignment module to align local features under spatial constraints for re-ranking. Compared to the mainstream re-ranking methods that require geometric consistency check, our method is significantly more efficient. To overcome the limitation that many methods use the easiest positive samples during training and have difficulty in identifying harder positive pairs during inference, we proposed the ShPSM strategy to mine semi-hard positive images for training a more robust network. The experimental results illustrated that our method outperformed the state-of-the-art methods on several benchmark datasets while showing high computational efficiency.

REFERENCES

- [1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.
- [2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.
- [3] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *European Conference on Computer Vision*. Springer, 2020, pp. 726–743.
- [4] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.
- [5] S. Lowry and H. Andreasson, "Lightweight, viewpoint-invariant visual place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 957–964, 2018.
- [6] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2020.
- [7] S. Hausler, S. Garg, M. Xu, M. Milford, and T. 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*, 2021, pp. 14 141–14 152.
- [8] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 648–13 657.
- [9] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [10] F. Lu, B. Chen, X.-D. Zhou, and D. Song, "Sta-vpr: Spatio-temporal alignment for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4297–4304, 2021.
- [11] Y. Shen, R. Wang, W. Zuo, and N. Zheng, "Tcl: Tightly coupled learning strategy for weakly supervised hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2684–2691, 2022.
- [12] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2570–2579.
- [13] A. Angeli, D. Filliat, S. Doncieux, and J. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.
- [14] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [15] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.
- [16] N. Sünderhauf, S. Shirazi, A. Jacobson, E. Pepperell, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems XI*, 2015.
- [17] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4297–4304.
- [18] S. Garg, A. Jacobson, S. Kumar, and M. Milford, "Improving condition-and environment-invariant place recognition with semantic place categorization," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 6863–6870.
- [19] S. Garg, N. Sünderhauf, and M. Milford, "Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3645–3652.
- [20] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 9–16.
- [21] L. G. Camara, C. Gbert, and L. Peuil, "Highly robust visual place recognition through spatial matching of cnn features," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3748–3755.
- [22] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, and Y. Wang, "Localizing discriminative visual landmarks for place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5979–5985.
- [23] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, and M. Milford, "Deep learning features at scale for visual place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3223–3230.
- [24] P. Yin, L. Xu, X. Li, C. Yin, Y. Li, R. A. Srivatsan, L. Li, J. Ji, and Y. He, "A multi-domain feature learning method for visual place recognition," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 319–324.
- [25] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2614–2620.
- [26] Y. Hou, H. Zhang, and S. Zhou, "Bocnf: efficient image matching with bag of convnet features for scalable and robust visual place recognition," *Autonomous Robots*, vol. 42, no. 6, pp. 1169–1185, 2018.
- [27] G. Berton, C. Masone, V. Paolicielli, and B. Caputo, "Viewpoint invariant dense matching for visual geolocalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 169–12 178.
- [28] S. Hausler and M. Milford, "Hierarchical multi-process fusion for visual place recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3327–3333.
- [29] S. Garg and M. Milford, "Seqnet: Learning descriptors for sequence-based hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4305–4312, 2021.
- [30] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.
- [31] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.
- [32] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2626–2635.
- [33] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2136–2145.
- [34] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *European conference on computer vision*. Springer, 2020, pp. 369–386.
- [35] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.
- [36] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [37] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '00, 2000, pp. 285–289.
- [38] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5396–5407.
- [39] D. Olid, J. M. Fàcil, and J. Civera, "Single-view place recognition under seasonal changes," *arXiv preprint arXiv:1808.06516*, 2018.
- [40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.
- [41] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Generalized contrastive optimization of siamese networks for place recognition," *arXiv preprint arXiv:2103.06638*, 2021.
- [42] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings*

of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 224–236.

- [43] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superglue: Learning feature matching with graph neural networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.