2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Pos^{2}VPR: Fast *Pos*ition Consistency Validation with *Pos*itive Sample Mining for Hierarchical Place Recognition

Dehao Zou[1], Xiaolong Qian[1*], Yunzhou Zhang[1], Xinge Zhao[1], Zhuo Wang[1]

***Abstract*— Visual place recognition (VPR) is a challenging issue for robotics and autonomous systems, focusing on utilizing visual information for robot localization. Currently, hierarchical architecture is being employed by growing works, which embraces RANSAC-based geometric verification for re-ranking. However, RANSAC is time-consuming and only employs geometric information while neglecting other potential information that could be useful for re-ranking. Here we propose a fast position consistency via local patch (PCLP) algorithm to take the position of task-relevant patch-descriptor into account. Without training, it only costs little time but performs better than other re-ranking methods that rely on geometric consistency verification. In this paper, we present a unified place recognition framework that incorporates an aggregation module to extract global features for retrieval and a PCLP validation module to filter local patch for re-ranking. Meanwhile, we propose a RANSAC-based tightly coupled learning (R-TCL) strategy to discover the best positive sample for training robust models. Unlike common sample mining methods, we introduce RANSAC into the sample mining process, achieving trade-off between efficiency and accuracy. Due to improved positive sample mining strategy and novel position validation module, our model is named as *Pos*^{2}VPR. Remarkably, *Pos*^{2}VPR outperforms state-of-the-art methods on four major datasets with extremely short running time.**

I. Introduction

Visual Place Recognition (VPR) is an essential problem in mobile robots and computer vision. Its objective is to ascertain if the current place has been visited by a robot and to acquire the geographical location [1]. Currently, there are still two key-challenges that plague VPR systems: 1) Variations in conditions (i.e. illumination and weather) and viewpoints. 2) Perceptual aliasing [2].

VPR is commonly regarded as an image retrieval task [3]. Two common approaches are typically adopted to represent place images in VPR tasks. Global features [4], [5] abstract the entire image into a compact feature vector. These features are treated as appearance-invariant but they often suffer from perceptual aliasing due to the absence of spatial geometric information [1]. Conversely, patch-level descriptors are considered as viewpoint-invariant [6]–[8], which can be applied for spatial matching with techniques like RANSAC [9]. Imperfectly, they are computationally intensive.

*The corresponding author of this paper
[1]Dehao Zou, Xiaolong Qian, Yunzhou Zhang, Xinge Zhao and Zhuo Wang are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. qianxiaolong@ise.neu.edu.cn

This work was supported by National Natural Science Foundation of China (No. 61973066) and Major Science and Technology Projects of Liaoning Province(No. 2021JH1/10400049).

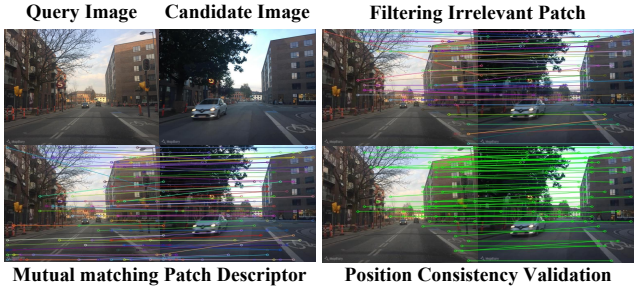

Fig. 1. **Position Consistency via Local Patch for Hierarchical Place Recognition.** PCLP module retain meaningful patch-descriptors and provide strong re-ranking performance while requiring minimal running time.

To pursue complementary advantages of both descriptors, hierarchical pipeline is employed [10]–[12], where it retrieve top-k candidates with global features (Coarse-Retrieval) and re-rank them by matching local features (Fine-Prediction). Nevertheless, the algorithms in re-ranking stage are significantly time-consuming. To address it, we focus on spatial information in coordinates of patch descriptors and propose a novel validation of position consistency via local patch (PCLP) module, as shown in Fig.1.

Leveraging assistance of large-scale VPR datasets with noisy GPS label, most existing VPR methods are trained in a weakly supervised manner [3], [6], [13]. However, proximity in GPS cannot promise the sufficient co-visible regions in sample images. Hence, we propose RANSAC-based tightly coupled learning (R-TCL) strategy which make improvements by replacing the local feature distance with image similarity to find the best positive samples. This modification leads to a faster and more effective sample mining strategy. With a novel ***pos***ition validation and effective ***pos***itive sample mining strategy, we call our model as *Pos*^{2}VPR.

In summary, our work makes the following contributions:

- We propose a hierarchical VPR architecture *Pos*^{2}VPR, which consists of an aggregation module to extract global features for retrieving candidates, and a training-free PCLP validation module to filter patch descriptor for re-ranking, preserving higher matching accuracy with an extremely short processing time.
- We propose a RANSAC-based tightly coupled learning (R-TCL) strategy for triplet networks to mine best positive sample. It guides network to learn features that are more favorable for place recognition, thus improving performance without additional burden.
- A series of comparison experiments have been conducted to validate the effectiveness of every proposed

module in *Pos*^{2}VPR. Results demonstrate that our approach outperforms state-of-the-art methods, with latency being less than half of AANet [14].

II. Related Work

Global Image Retrieval. In the early of VPR, most methods employed global features to represent images and retrieve place images [3], [15]. The current predominant methods are based on deep learning, many VPR methods [5], [16]–[18] have adopted deep features to achieve enhancing performance. In recent study by Berton et al. [13], they introduced a VPR benchmark and implemented various global-retrieval-based methods within a unified framework.

Hierarchical Retrieval Pipeline. Recent SOTA methods [6], [14], [19]–[21] introduced re-ranking module into VPR that typically requires geometric consistency verification [6], [11]. Patch-NetVLAD [6] adopted NetVLAD [3] for global retrieval and applied RANSAC [9] based geometric verification on multi-scale patch descriptors. AANet [14] designed DALF algorithm to align local features for re-ranking. Unlike them, to tackle significant time consumption with re-ranking, we propose a strong, concise and efficient position consistency verification algorithm that leverages the positional information among local patch features.

Positive Sample Strategy. Larg-scale place datasets embracing Pittsburgh [22] and MSLS [23] only provide weak supervision approach with noisy GPS labels. To eliminate false positives in training tuple, NetVLAD [3] mined the simplest top-1 positive sample in triplet for training. This strategy was also adopted by some subsequent works [6], [11], [24]. But the top-1 sample doesn't imply a enough co-vision region with the query image, which can be detrimental to the training of model. To address it, CRN [25] and SFRS [26] mined hard positive samples for training satge. While TCL [19] combined global and local descriptors for mining the best positive sample by re-ranking mining candidate with local distance. In contrast, we propose R-TCL strategy and achieve balance between efficiency and performance.

III. Proposed Method

A. Overview and Place Representation

As shown in training stage of Fig.2, here we introduce R-TCL strategy and PCLP module for training and testing stages, respectively. Given a query and reference images, we first extract their corresponding $W \times H \times C$ dimensional patch-level features. Then, for selecting mining candidate images, the local features are aggregate into global features to calculate L2 distance between the query and reference images. As indicated by the red dashed box in Fig.2, the mining candidates are utilized in the positive sample mining process, where our R-TCL strategy is proposed to mine the best positive sample for triplets. After acquiring triplet training tuple, DALF module [14] in local branch provides local loss that is combined with global loss to jointly optimize model. In right part of Fig.2, the testing stage of our retrieval pipeline is depicted. We propose PCLP module by focusing on positional correlation between patch-feature pairs during the re-ranking stage. The top-K candidates of the query are coarsely retrieved using global features, and then reordered with PCLP module to get final retrieval result against the query. Detailed information about the R-TCL strategy and PCLP module will be presented in the following sections.

B. RANSAC-based Tightly Coupled Learning Strategy

When training with VPR dataset [22], [23], each query image I_q is accompanied with a set of potential positive samples $\{p_i^q\}$ and definite negative samples$\{n_j^q\}$, where potential positives refer that there is at least one co-visible image with the query. While the purpose of sample mining is to identify the best positive sample among them. To tackle this problem, we propose the RANSAC-based TCL strategy, which prioritizes the co-visible region during the positive sample mining process. This strategy aims to improve model robustness without compromising training speed.

As illustrated in Fig.3, we first calculate global feature distances between the query and potential positive samples, and then rank them accordingly. Note that, unlike TCL [19], which utilizes $[CLS]$ token as global feature. We employ GEM [27] to aggregate local patch-level features in our strategy. Here, we consider that top-ranked potential positive samples in ranking list already exhibit sufficient co-visible regions with the query. In second stage, we apply the powerful RANSAC algorithm [9] to compare similarity between the query and potential positive images. Image similarity is defined as the number of inliers when estimating homography based on matched patches with RANSAC algorithm. To address time-consuming problem of RANSAC, as shown in Fig.3, we mitigate computational burden by selecting only top five images (i.e. mining candidate) in ranking list. We then perform RANSAC verification on the candidate and re-rank them in accordance with similarity. After that, top-1 in mining candidate is regarded as the best positive sample p_*^q. We follow the common practice of previous works [3], [19] to identify the negative samples. In this way, training tuple (q, p_*^q, n_j^q) embracing the best positive sample p_*^q is obtained.

To jointly optimize the model with triplets (q, p_*^q, n_j^q), we apply triplet ranking loss in [3] for the global and local branch, where the losses are given by:

$$L_{global} = \sum_j \max\left\{\left(D_g\left(q, p_*^q\right) + m - D_g\left(q, n_j^q\right)\right), 0\right\} \quad (1)$$

$$L_{local} = \sum_j \max\left\{\left(D_l\left(q, p_*^q\right) + m - D_l\left(q, n_j^q\right)\right), 0\right\} \quad (2)$$

Here, D_g denotes global distance and D_l is local distance captured by DALF module. Then, total loss L_T is integrated from global loss and local loss with weight λ, that is:

$$L_T = L_{global} + \lambda L_{local} \quad (3)$$

C. Validation of Position Consistency via Local Patch

To re-rank retrieval candidates in re-ranking stage, DALF module [14] was proposed to align local features. However, in practice, we have found that performing brute-force matching followed by verification on local features

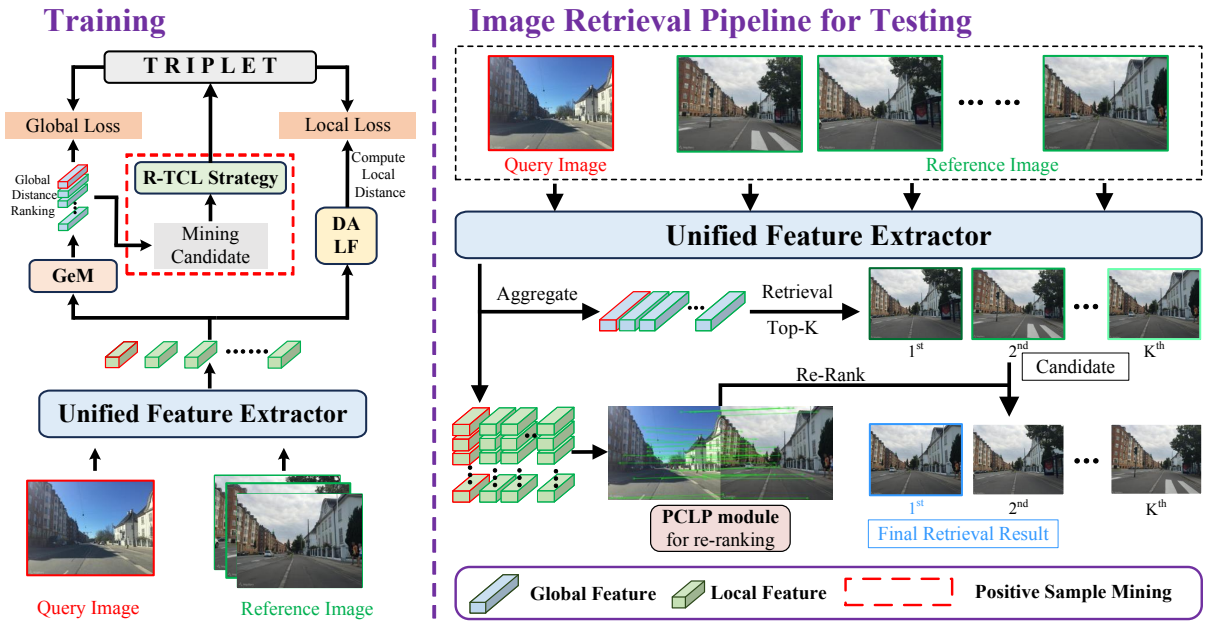

Fig. 2. **The overview of *Pos*^{2}VPR.** The left part is training stage where the R-TCL strategy is proposed to achieve positive sample mining. The right part illustrates testing stage and the PCLP module is designed to re-rank the coarse retrieval candidates.

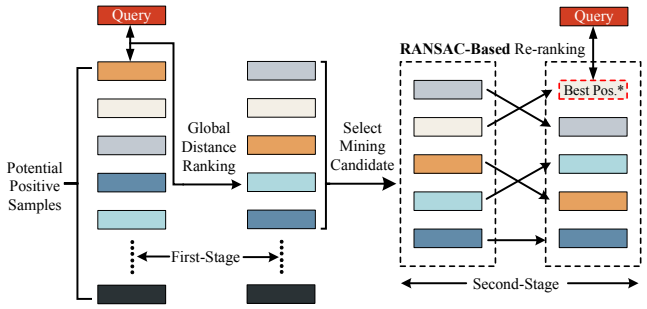

Fig. 3. **The depiction of the R-TCL strategy.** It contains two stages, and the sample in red dashed line is the best positive sample p_*^q.

yields higher accuracy. But traditional geometric verification method like RANSAC [9] is highly time-consuming. In order to chase a balance between accuracy and efficiency, we observe that positional information of each patch-level feature within the image also contains crucial information. They reflect whether the content between image pairs is distributed in the same spatial pattern. For example, if the feature in top-left patch of query image is similar to that in bottom-right patch of candidate image, but due to a significant disparity in their positions, it is evident that these features are mismatched. Based on this observation, we propose a training-free and fast PCLP module, which includes three parts: mutual matching, mask filtering and position verification (visualized in Fig.1). The following provides a detailed description.

The backbone is employed to obtain feature maps $\{f_c^i\}_{i=1}^N$ and $\{f_q^j\}_{j=1}^N$corresponding the candidates I_c and the query image I_q, where N represents number of patches. Then, cosine similarity is adopted to measure the similarity $SIM \in \mathbb{R}^{N\times N}$ between local feature pairs which can be replaced by the inner product as L2-normalized. Formally:

$$SIM(i,j) = \left(f_c^i\right)^T \left(f_q^j\right), \quad i,j \in N \tag{4}$$

After calculating the similarity map, we obtain patch-level descriptor pairs from mutual nearest neighbor matching set $\mathbb{MNN}$ which is defined as:

$$\mathbb{MNN} = \left\{(\mathrm{a},\mathrm{b}) : a = MS_c(\mathrm{i},\mathrm{b}), b = MS_q(\mathrm{a},\mathrm{j})\right\} \tag{5}$$

where $MS_c = \arg\max_{\mathrm{i}} SIM$ and $MS_q = \arg\max_{\mathrm{j}} SIM$.It presents that if the local feature f_q^b is the most similar to f_c^a in I_c, and the local feature f_c^a is the most similar to f_q^b within I_q. Then we consider the patch pair (a,b) to be part of the nearest neighbor matching set.

To further reduce memory usage and speed up the matching process, a relevance mask is estimated to selectively filter features that are irrelevant for place recognition, defined as:

$$RV(m) = \mathrm{MinMaxNorm}((\mathbb{F}(m))), m \in N \tag{6}$$

Then we set threshold t_m, if the value of $RV(m)$ is below the threshold t_m, the corresponding m^{th} patch is regarded as a meaningless region and filtered out, as illustrated in gray patch of Fig.4.

Then we propose a novel validation module to verify the matched features. For P matched patch pairs (A,B), we can obtain their positional coordinates $(a_p,\ b_p)$ and $(a_p^{'},\ b_p^{'})$ in images I_c and I_q, respectively. By calculating the Euclidean distance between these two patches in the spatial domain, we can infer whether they are located in close proximity to each other.

$$D_{qc}(p) = \sqrt{\left(a_p - a_p'\right)^2 - \left(b_p - b_p'\right)^2} \tag{7}$$

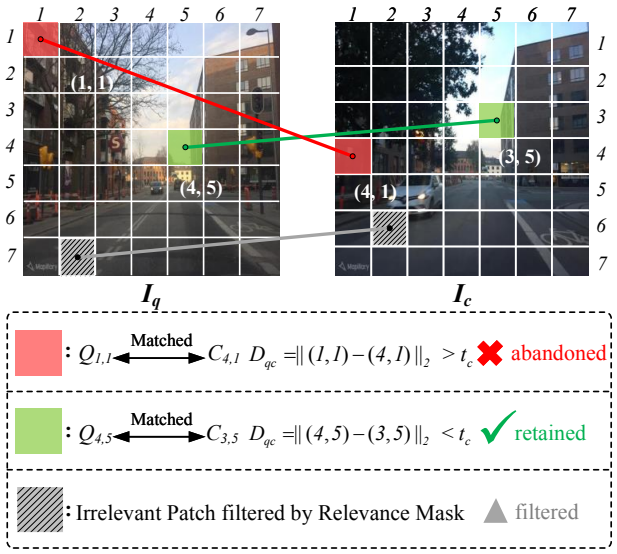

Fig. 4. **Illustration of position consistency validation.**Green patches pair is retained for close proximity, while red pair is abandoned. The gray patch is filtered out as irrelevant patch although their coordinates are close.

As shown in the red patch pairs in Fig.4, if distance between coordinates of the matched patches exceeds threshold, it indicates that although the two patches exhibit high similarity, the distribution of matched features shows disparity. This suggests that it is a wrong match and thus abandoned. By summing the number of patch-pairs whose distance between coordinates is below the threshold t_c, we obtain a validation score. This validation score will be ultimately used as basis for re-ranking the candidate images. That is:

$$Score = \sum_{p=1}^{P} (D_{qc}(p) < t_c), \quad p \in P \tag{8}$$

IV. Experiment

A. Implementation Details

Pos^{2}VPR is implemented in PyTorch framework. During both training and testing, the resolution of all input images is resized to 384×384. As for default backbone, the CCT-14 architecture is employed and initialized with off-the-shelf pre-trained weights on ImageNet-1k. Patch size is set to 16×16. Additionally, we set the batch size to 4, and the model is trained using the Adam optimizer with a learning rate of 0.00001. The number of candidates for re-ranking is set 32. And the key-patch filtering threshold t_m is set to 0.2 in practice and the distance threshold t_c in PCLP module is assigned to 192, which is half the length of input image. In triplet loss function, we assign weight λ=1 and margin m=0.1.

B. Datasets and Evaluation Metrics

To evaluate the generalization ability of *Pos*^{2}VPR, we conduct experiments on several benchmark datasets, including **Mapillary Street Level Sequences (MSLS)** [23], **Pittsburgh (Pitts30k)** [22], **St_Lucia** [28] and follow public split of validation/test sets. Table I provides the summary of datasets and we present the usage of each dataset, which facilitates evaluation of the results.

TABLE I
Summary of Dataset for Testing

Dataset	Description			Variation		
	Urban	Suburban	Natural	Illumination	Viewpoint	Dynamic
MSLS	✓	✓	✓	++	++	++
Pitts30k	✓	-	-	-	++	+
St_Lucia	✓	-	-	+	+	+

We address $Recall@N (N = 1, 5, 10)$ as evaluation metrics. Following previous work, we set the tolerance for correct localization as 25 meters for MSLS, Pitts30k and St_Lucia. Furthermore, the latency and feature dimensions are also presented in Table II.

C. Comparison with State-of-the-Art

We compare *Pos*^{2}VPR with several SOTA methods, embracing two global image representation methods: NetVLAD [3] and GCL [29], and five two-stage models: SP-SuperGlue [20], [30], Patch-NetVLAD [6], TCL [19], ETR [21] and AANet [14]. Among them, NetVLAD is trained on Pittsburgh250k dataset, while GCL (ResNet152-GeM-PCA), TCL (DeiT-S, TCL-R100), and AANet following optimal configurations in their original works. For Patch-NetVLAD, we test its both speed-focused and performance-focused configurations. While ETR-D version which incorporates an excellent re-ranking module is selected in our experiment. The quantitative results of *Pos*^{2}VPR compared with other approaches are shown in Table II.

It can be observed that our method without re-ranking, denoted as *Pos*^{2}VPR (w/o re-ranking), convincingly outperforms the compared methods on all datasets even if our feature dimension is more compact than that of NetVLAD and GCL. When comparing R@1 with the NetVLAD, our approach achieved absolute improvements of 19.9% on MSLS_val, 25.5% on MSLS_challenge, 0.8% on Pitts30k dataset, and 47.5% on St Lucia.

Compared to the two-stage pipeline, *Pos*^{2}VPR also achieves the best results on MSLS validation, MSLS challenge and St_Lucia datasets, which demonstrates effectiveness of the R-TCL strategy and PCLP module. On the Pitts30k dataset, R@1 of *Pos*^{2}VPR obtains second-best performance with a 1.7% lower than TCL model. We believe the cause is that there is a large rotational variation in images of Pitts30k. Compared to the BS-DTW in TCL for aligning feature, our PCLP module focuses more on the spatial positions of similar patch pairs. However, due to the rotational variations in dataset, the distances between similar patch pairs are more likely to exceed the threshold and be filtered out by PCLP module. Nevertheless, benefit from PCLP, *Pos*^{2}VPR achieves a 4.4% higher R@1 value than AANet on MSLS_val dataset and a 12.2% higher R@1 value than ETR on MSLS_challenge dataset. The MSLS dataset has rich perceptual aliasing and illumination variations, where *Pos*^{2}VPR demonstrates impressive results compared to other latest methods. This directly proves that our model is capable

TABLE II

QUANTITATIVE RESULTS OF *Pos*^{2}VPR WITH STATE-OF-THE-ART METHODS ON MAJOR VPR DATASET

Method	Feature Dim	Latency (s)	MSLS_val			MSLS_Challenge			Pitts30k_test			St_Lucia		
			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 [3]	32768	0.033	60.8	74.3	79.5	31.5	42.1	46.2	81.9	91.2	93.7	49.0	68.1	76.2
GCL [29]	2048	0.087	78.3	85.0	86.9	-	-	-	75.6	91.5	92.6	69.4	75.1	80.6
***Pos*^{2}VPR(w/o re-ranking)**	384	**0.016**	**80.7**	**90.1**	**92.4**	**57**	**75.2**	**80**	**82.7**	**92.0**	**94.3**	**96.5**	**99.2**	**99.7**
SP-SuperGlue [20], [30]	256	7.340	78.1	81.9	84.3	50.6	56.9	58.3	87.2	94.8	96.4	86.5	92.1	93.4
Patch-NetVLAD-s [6]	512	0.97	77.8	84.3	86.5	48.1	59.4	62.3	87.5	94.5	96	90.2	93.6	95
Patch-NetVLAD-p [6]	4096	17.630	79.5	86.2	87.7	48.1	57.6	60.5	88.7	94.5	95.9	93.9	95.5	96.2
TCL [19]	384	-	78.7	82.5	85.3	-	-	-	**90.5**	**95.9**	**97.5**	-	-	-
ETR [21]	1024	-	79.3	88.0	89.6	50.6	62.1	65.8	84.2	91.6	93.8	98.7	99.0	99.2
AANet [14]	384	0.066	80.1	88.9	91.0	-	-	-	88.0	94.2	95.8	98.2	99.6	99.7
***Pos*^{2}VPR(ours)**	384	**0.028**	**84.5**	**92.2**	**93.5**	**62.8**	**76.8**	**81.8**	88.8	94.5	96.0	**99.2**	**99.8**	**99.9**

TABLE III

ABLATIONS ON DEGENERATE CONFIGURATIONS

Configurations	R-TCL	PCLP	Joint Loss	MSLS_val	Pitts30k	St_Lucia
OLS-GR-j	×	×	✓	75.3	81.9	96.3
OLS-RR-j	×	✓	✓	81.5	86.1	98.6
RLS-GR-j	✓	×	✓	80.7	82.7	96.5
RLS-RR-g	✓	✓	×	81.4	86.9	95.5
***Pos*^{2}VPR**	✓	✓	✓	**84.2**	**88.8**	**99.2**

of handling such challenges.

D. Ablation Study

We evaluate the impact of the proposed R-TCL strategy, the PCLP algorithm, and the local loss in training by conducting ablation studies between *Pos*^{2}VPR and four degenerate configurations:

- OLS-GR-j: Original triplet Learning Strategy & Global Retrieval & joint loss.
- OLS-RR-j: Original triplet Learning Strategy & Re-Ranking & joint loss.
- RLS-GR-j: RANSAC-based tightly coupled Learning Strategy & Global Retrieval & joint loss.
- RLS-RR-g: R-TCL Strategy & Re-Ranking & global loss (without local loss).
- *Pos*^{2}VPR: R-TCL Strategy & Re-Ranking & joint loss.

Here, we only use R@1 to evaluate the performance of each degenerate model. According to results in Table III, adopting R-TCL strategy leads to a significant improvement compared to OLS-GR-j. Similarly, *Pos*^{2}VPR outperforms OLS-RR-j, illustrating that R-TCL strategy can improve performance, regardless of whether re-ranking is performed. Likewise, since OLS-RR-j outperforms OLS-GR-j while *Pos*^{2}VPR is better than RLS-GR-j. There are obvious improvements after introducing PCLP module with or without the R-TCL strategy. Above results demonstrate effectiveness of R-TCL strategy and PCLP module in improving performance for VPR.

Due to non-differentiability of PCLP module, we introduce DALF module into training to provide local loss. Here, we also investigate necessity of local loss. From last two rows of Table III, *Pos*^{2}VPR outperforms RLS-RR-g when training with joint loss. It also indicates the success of introducing local loss for joint optimization.

Furthermore, the improvement on the MSLS dataset is higher compared to that on the Pitts30k dataset, with *Pos*^{2}VPR achieving an absolute improvement of over 8% on msls_val compared to OLS-GR-j. We attribute this phenomenon to the presence of more prominent appearance variations in the MSLS dataset. While networks trained with R-TCL are better equipped to handle such appearance variations than those trained with original learning strategy. And MSLS dataset provides more detailed information that benefits spatial position verification.

E. Latency Performance Analysis

To analyze efficiency, we evaluate runtime for processing a single query on Pitts30k test dataset which includes both the feature extraction time and the matching time. NetVLAD and GCL are used to compare the retrieval time (first-stage) against *Pos*^{2}VPR. As shown in Table II, our *Pos*^{2}VPR exhibits significantly lower runtime compared to NetVLAD and GCL, which can be attributed to usage of lower feature dimensions. Instead of 32768-*D* in NetVLAD and 2048-*D* in GCL, our model uses 384-*D* feature and results in a notable advantage in terms of computational efficiency.

Additionally, we compared the total latency against SP-SuperGlue, Patch-NetVLAD, and AANet. According to Table II, *Pos*^{2}VPR is several orders of magnitude faster than Patch-NetVLAD-p and SP-SuperGlue and is also 30 times faster than the speed-focused version of Patch-NetVLAD. Moreover, the runtime of our model is less than half of AANet's that is considered as fast model. The latency indicates that PCLP module presents exceptional computational efficiency. The reason is that PCLP module only utilizes the coordinates of similar patch pairs for verification. Thanks to the powerful parallel computing capability of GPUs, the computation time required for this multiplication operation is significantly reduced. In contrast with AANet, which uses DALF to align local features, our approach is simpler but more efficient, resulting in lower latency. Hence, it is more suitable for real-world scenarios.

V. Conclusion

We present *Pos*^{2}VPR, a novel hierarchical VPR architecture that leverages an aggregation module to extract global features for candidate retrieval and incorporates the PCLP module to filter similar patch pairs based on their coordinates for re-ranking. The proposed PCLP re-ranking module explores the rich information contained in spatial relationship between task-relevant patch pairs and is applicable to other transformer backbones without training. Additionally, we introduce the R-TCL strategy to mine the best positive samples for training a more robust model. By retaining a small set of potential positive image samples, we achieve a trade-off between accuracy and efficiency using the RANSAC algorithm, addressing the limitations of training network solely rely on simple triplets. Remarkably, in comparison to the state-of-the-art methods, experimental results on multiple benchmark datasets demonstrate that our *Pos*^{2}VPR achieves higher accuracy and efficiency. The results also prove that *Pos*^{2}VPR possesses powerful performance to tackle the challenges for VPR in real-world scenarios.

References

[1] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, vol. 113, p. 107760, 2021. 1

[2] 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, 2015. 1

[3] 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. 1, 2, 4, 5

[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*. IEEE, 2010, pp. 3304–3311. 1

[5] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. 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*, 2024, pp. 16 772–16 782. 1, 2

[6] 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. 1, 2, 4, 5

[7] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3456–3465. 1

[8] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101. 1

[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. 1, 2, 3

[10] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," *arXiv preprint arXiv:2402.14505*, 2024. 1

[11] 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. 1, 2

[12] F. Lu, S. Dong, L. Zhang, B. Liu, X. Lan, D. Jiang, and C. Yuan, "Deep homography estimation for visual place recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 10 341–10 349. 1

[13] 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. 1, 2

[14] F. Lu, L. Zhang, S. Dong, B. Chen, and C. Yuan, "Aanet: Aggregation and alignment network with semi-hard positive sample mining for hierarchical place recognition," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 771–11 778. 2, 4, 5

[15] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4297–4304. 2

[16] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Robotics: Science and Systems XI*, pp. 1–10, 2015. 2

[17] S. Garg, A. Jacobson, S. Kumar, and M. Milford, "Improving condition-and environment-invariant place recognition with semantic place categorization," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6863–6870. 2

[18] S. Garg, N. Suenderhauf, 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)*. IEEE, 2018, pp. 3645–3652. 2

[19] 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. 2, 4, 5

[20] 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. 2, 4, 5

[21] H. Zhang, X. Chen, H. Jing, Y. Zheng, Y. Wu, and C. Jin, "Etr: An efficient transformer for re-ranking in visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5665–5674. 2, 4, 5

[22] 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. 2, 4

[23] 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. 2, 4

[24] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00266 2

[25] 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. 2

[26] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 369–386. 2

[27] 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. 2

[28] M. J. Milford and G. F. Wyeth, "Mapping a suburb with a single camera using a biologically inspired slam system," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1038–1053, 2008. 4

[29] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Generalized contrastive optimization of siamese networks for place recognition," *arXiv preprint arXiv:2103.06638*, 2021. 4, 5

[30] 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. 4, 5