2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

CAHIR: Co-Attentive Hierarchical Image Representations for Visual Place Recognition

Guohao Peng[1], Heshan Li[1], Yifeng Huang[1], Jun Zhang[1], Mingxing Wen[1], Singh Rahul[2], and Danwei Wang[1]

***Abstract*— Robust visual place recognition (VPR) against significant appearance changes is crucial for the life-long operation of mobile robots. Focusing on this task, we propose a Co-Attentive Hierarchical Image Representations (CAHIR) framework for VPR, which unifies attention-sharing global and local descriptor generation into one encoding pipeline. The hierarchical descriptors are applied to a coarse-to-fine VPR system with global retrieval and local geometric verification. To explore high-quality local matches between task-relevant visual elements, a cross-attention mutual enhancement layer is introduced to strengthen the information interaction between the local descriptors. Through the proposed selective matching distillation, the mutual enhancement layer can learn from state-of-the-art local matchers in a distillation manner. After weighted cross-matching of the enhanced local descriptors, geometric verification is applied to evaluate the spatial consistency of the compared image pair. Experiments show CAHIR outperforms the existing global and local representations for VPR in terms of performance and efficiency. Quantitatively, it achieves state-of-the-art results on three city-scale benchmark datasets. Qualitatively, CAHIR proves to attach great importance to task-relevant visual elements and excels at finding local correspondences that are discriminative to the VPR task.**

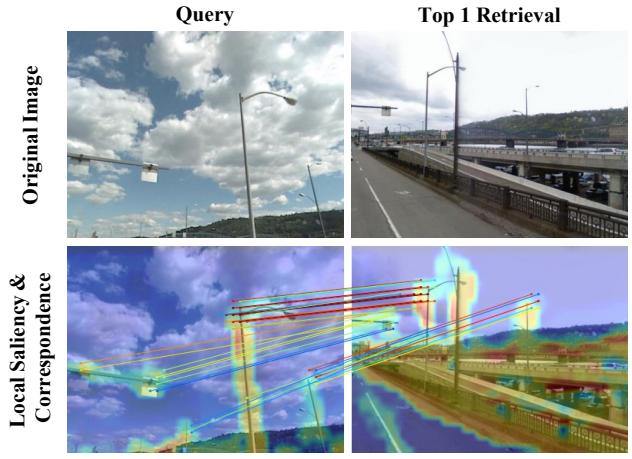

Fig. 1. In this work, a coarse-to-fine VPR framework named CAHIR is proposed. It combines the generation of attention-sharing global and local descriptors into one encoding pipeline. The hierarchical descriptors are applied to two-stage retrieval. As can be seen, CAHIR is able to focus on the static visual elements and discover local correspondences that are discriminative to VPR tasks (as shown in the figures above). By evaluating the geometric consistency of the local matches found, CAHIR can robustly achieve correct retrieval in challenging environments.

I. INTRODUCTION

Visual Place Recognition (VPR) is one of the core capabilities of mobile robots and autonomous systems, serving as the foundation for many practical applications, such as geo-gocalization [1]–[6], topological mapping [7], and robot navigation [8]–[11]. In large-scale environments, VPR is typically solved as an image retrieval task [12]–[16], where the main challenges are significant appearance changes of scenes caused by different illumination, seasons, and weather.

To effectively address the challenges, researchers have attempted to present solutions from a variety of perspectives, which can be roughly divided into two categories. The first type of methods [14], [15], [17]–[19] strives to construct powerful global image descriptors for fast and accurate retrieval. By running a nearest neighbor search on the query image descriptor, candidate reference images with smaller feature space distances to the query image will stand out. However, the global approaches are mainly based on aggregation to obtain compact image descriptors, at the cost of decoupling spatial information and ignoring local details. This may cause confusion in the global retrieval of multiple scenes with similar appearances. As a result, in addition to the global approach, the second type of methods [20]–[23] refocuses on local details. They exploit the spatial consistency of pixels or patches to geometrically verify the candidate reference images obtained by the global retrieval. These methods rely on the cross-matching of hand-crafted [24], [25] or deep-learned [20]–[22], [26] local descriptors, which necessitates the local descriptor extraction independent of the global descriptor encoding pipeline. So far, few studies [21], [23], [27] have attempted to integrate global and local descriptor generation within a single forward pipeline. In addition, the popular local descriptor detectors [24]–[26] and matchers [28], [29] that can be used for geometric verification are not specifically proposed for VPR. They may produce unnecessary dense correspondences on visual cues that are not important to the task. With the above motivations, we make the following contributions in this work.

(1) We propose a coarse-to-fine VPR framework with Co-Attentive Hierarchical Image Representations (CAHIR). In the CAHIR framework, global and local descriptors are extracted concurrently, sharing the same encoding pipeline. By reusing the intermediate features of the global encoding for local descriptor generation, no separate local descriptor extraction pipeline as in other SOTAs [23], [30] is needed. The formulation of the hierarchical descriptors integrates the

[1] Guohao Peng, Heshan Li, Yifeng Huang, Jun Zhang, Mingxing Wen, and Danwei Wang are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

[2] Singh Rahul, is with Continental Automotive Singapore Pte Ltd., 80 Boon Keng Road, 339780, Singapore

979-8-3503-2365-8/23/$31.00 ©2023 IEEE

triple attention from individual, spatial and cluster saliency of local features, which conduces to highlighting static structures while suppressing misleading visual elements in the image representation.

(2) Unlike latest SOTAs [21], [23], [31] which directly match extracted local features, we introduce a locally mutual enhancement layer (ME) to reinforce the local descriptors. It strengthens the information interaction between local descriptors with high correlation in the compared image pair. In order for the mutual enhancement layer to perform optimally, we propose a distillation pipeline with novel selective matching loss, through which the parametric model can be fine-tuned through distillation learning. After cross-matching the enhanced local descriptors, only local correspondences with high task-relevance are preserved for subsequent geometric consistency assessment.

Extensive experiments demonstrate that our CAHIR outperforms the existing global and local representations of VPR on the employed VPR benchmark datasets.

II. Related work

A. Global image representations

The global approaches seek to create a powerful image representation of the entire scene. Traditional methods rely on hand-crafted local features [32], [33] and aggregation-based encoding strategies [13], [34]–[37]. In recent years, more high-performance global representations that embrace the power of deep learning have been proposed. The most typical one is NetVLAD [14], which generalizes VLAD into a differentiable pooling layer. Its multiple variants that integrate contextual re-weighting [15], spatial pyramid enhancement [18], semantic reinforced local weighting [19], or attentional pyramid pooling [38] all show considerable advantages. Cutting in from different angles, SARE [39] and SFRS [40] improve NetVLAD performance by introducing novel training strategies and loss functions. In general, the global approach performs well in terms of retrieval speed. It can be attributed to the compact image representation obtained by feature aggregation. However, this also comes at the cost of decoupling spatial information and ignoring local details. Therefore, researchers have also been studying local representations for spatial consistency checks.

B. Local descriptors and matchers

In a general VPR pipeline, global retrieval is usually followed by geometric verification. Spatial consistency can be used to re-rank a list of globally retrieved candidate images. The most commonly used consistency criterion is the number of correspondences verified by RANSAC [41]. Early works use hand-crafted features [25], [32], [33] for local matching. MagicPoint [42] is a seminal deep learning architecture for finding local feature matches. SuperPoint [26] propose a self-supervised framework to jointly computes pixel-level keypoint locations and descriptors. SuperGlue [28] augments SuperPoint with a graph neural network to emphasize true matches and deemphasize outlier matches. Recently, LoFTR [29] propose a local feature transformer which excels in producing dense matches in low-texture areas. These learning-based local descriptors and matchers can robustly find numerous correspondences between image pairs, which is suitable for VPR tasks. In [23], a strong baseline for VPR is set up by using NetVLAD and SuperGlue for global retrieval and spatial consistency check respectively. However, these local feature extractors and matchers are not dedicated to VPR tasks, and therefore may produce unnecessary dense correspondences on task-irrelevant visual cues.

C. Joint extraction of global and local descriptors

There are serveral works attempting to integrate global and local descriptors into a unified framework. Targeting mobile localization, HF-Net [22] distills NetVLAD and SuperPoint upon a shared MobileNet [43] backbone. DELG [21] is proposed for image retrieval, which can jointly train global and local descriptors within a unified model. While the above two models extract global and local descriptors through separate branches, Patch-NetVLAD [23] derives multi-scale regional VLAD descriptors from global descriptor encoding pipeline. However, its post-processing is a non-learning process. Our CAHIR also unifies global and local descriptor generation in one encoding pipeline. It additionally incorporates a mutual enhancement layer to enhance local descriptors via non-local information interaction. By distillation learning, CAHIR can learn better correspondences from a SOTA local matcher.

III. Preliminaries

Self-attention layer is the core component in the Transformer [44] structure. Through Eq.(1), the input features $F \in \mathbb{R}^{N\times D}$ are first projected to the query, key, and value vectors ($Q \in \mathbb{R}^{N\times S}$, $K \in \mathbb{R}^{N\times S}$, $V \in \mathbb{R}^{N\times L}$) by corresponding projection matrix $W_Q \in \mathbb{R}^{S\times D}$, $W_K \in \mathbb{R}^{S\times D}$, and $W_V \in \mathbb{R}^{L\times D}$.

$$Q = FW_Q^T, \quad K = FW_K^T, \quad V = FW_V^T \tag{1}$$

Then via Eq.(2), an output vector V_i' of the self-attention layer is the weighted sum of the value vectors $\{V_j\}$. The weights are determined by the similarity (typically softmax) scores between the query Q_i and the keys $\{K_j\}$.

$$V_i' = \frac{\sum_j sim(Q_i, K_j)V_j}{\sum_j sim(Q_i, K_j)} = \frac{\sum_j exp(Q_i K_j^T)V_j}{\sum_j exp(Q_i K_j^T)} \tag{2}$$

IV. Proposed Method

A. Overview

As illustrated in Fig.(2), CAHIR is a coarse-to-fine VPR framework, which encompasses three main steps. (1) The global and local image descriptors are first generated by a unified CAHIR extractor (Sec.IV-B). (2) A coarse-level retrieval is then performed by matching the query with the database images using global descriptors. (3) A fine-level geometric verification is finally performed to re-rank the top K retrieved candidates. For better cross-matching and geometric consistency assessment, a locally mutual enhancement layer (Sec.IV-C) is introduced and optimized through a distillation pipeline (Sec.IV-E).

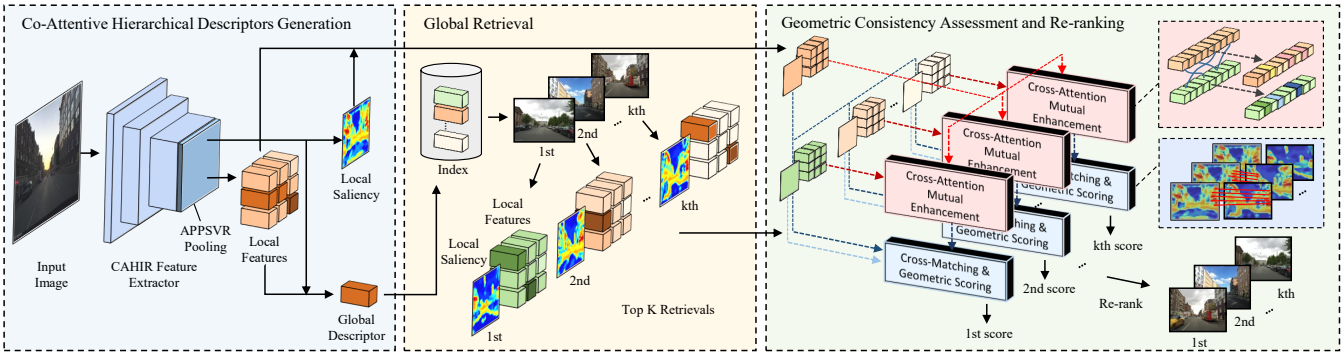

Fig. 2. The illustration of CAHIR inference pipeline. In general, it is a coarse-to-fine VPR framework, which includes three main steps: co-attentive hierarchical descriptor generation, global retrieval, and local geometric consistency verification.

B. Co-Attentive Global and Local Descriptors

We extend our previously proposed global image representation for visual place recognition, APPSVR [38], as a hierarchical descriptor generator. With the entire encoding pipeline preserved for global descriptor generation, we reuse the intermediate local residuals and inferred attention to formulate patch-level local descriptors.

Global image descriptor. Firstly, a VGG-16 [45] cropped at the last convolutional layer is used as a backbone to encode an input image I to feature maps. The spatial activations $\mathbf{x} \in \mathbb{R}^{D\times 1\times 1}$ decomposed from the normalized feature maps $X \in \mathbb{R}^{D\times H\times W}$ are regarded as deep local features. Then soft-assignment [14] measures the probability of each local feature belonging to the k^{th} visual cluster, denoted as $\alpha_k(\mathbf{x}_i)$. Intra-cluster weighting [38] further evaluates the intra-cluster saliency $\beta_k(\mathbf{x}_i)$ of local features, quantifying their significance to cluster-wise feature embedding. As in Eq.(3), a local residual $r_k(\mathbf{x}_i)$ is calculated as the difference between the local feature $\mathbf{x}_i$ and the cluster centroid $\mathbf{c}_k^r$, double weighted by soft-assignment weight and intra-cluster saliency.

$$\mathbf{r}_k(\mathbf{x}_i) = \alpha_k(\mathbf{x}_i)\beta_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_k^r) \tag{3}$$

Next, an overlapping pyramid pooling is applied to generate multi-scale regional features $\mathbf{f}_k$ by summing the local residuals in each spatial pyramid grid. Then visual word vector $\mathbf{V}_k$ is obtained by aggregating the regional features $\mathbf{f}_k$ weighted by their relative spatial saliency μ_k. Finally, the normalized visual word vectors rescaled by their corresponding cluster saliency γ_k are concatenated as the global descriptor. More details can refer to [38]. After PCA whitening and L2-normalization (denoted as $\mathscr{F}_{\text{PCA}}$), a more compact image descriptor with 4096 dimensions is used for global retrieval.

Local patch descriptors. Owing to task-driven attention and representation learning, local residuals are inherently able to characterize the task-relevant local details of the corresponding image patches. Concurrently, the inferred local saliency $\alpha_k\beta_k$ and cluster saliency γ_k can quantify the significance of local patches to the cluster-wise feature representation. With these in mind, we believe that the local residuals $\mathbf{r}_k(\mathbf{x}_i)$ should already contain the discriminative information of the image patches. Thus, we define a local patch descriptor $\mathbf{f}_{patch}(\mathbf{x}_i) \in \mathbb{R}^{(D\cdot K)\times 1\times 1}$ as the concatenation of the normalized cluster-wise local residuals $\tilde{\mathbf{r}}_k(\mathbf{x}_i)$ weighted by their corresponding cluster saliency γ_k. Formally, $\mathbf{f}_{patch}(\mathbf{x}_i)$ is expressed as Eq.(4). For cross-matching efficiency, our local descriptors are reduced to 4096 dimensions by PCA.

$$\mathbf{f}_{patch}(\mathbf{x}_i) = \mathscr{F}_{\text{PCA}_{4096}}([\gamma_1 \cdot \tilde{\mathbf{r}}_1(\mathbf{x}_i), ..., \gamma_K \cdot \tilde{\mathbf{r}}_K(\mathbf{x}_i)]) \tag{4}$$

According to Eq.(3) and Eq.(4), our local descriptors are derived only based on the local residuals and attention weights. They are all intermediate variables for generating the global image descriptor. It enables our hierarchical descriptors to share the same encoding pipeline, and no additional parametric module is required for local feature extraction. During training, the local descriptors are jointly optimized with the global descriptor in a task-driven manner, sharing the same learned attention for VPR.

C. Locally Mutual Enhancement

Taking into account the representation deviation between a query image and its positive reference image caused by appearance changes, we introduce a mutual enhancement layer to obtain better cross-matching. It strengthens the correlation between the local descriptors of the compared image pairs through mutual information interaction.

Mutual enhancement layer. Via the self-attention layer in Eq.(2), an output vector V_i' aggregates multiple highly correlated value vectors $\{V_j\}$ retrieved by the similarity function. The core idea of mutual retrieval and fusion is in line with the essence of local information interaction. Therefore, a self-attention layer can be inherently used to strengthen the cross-matching ability of local descriptors. Following [29], we extend Eq.(2) as a cross-attention layer to mutually enhance the local descriptors $\{\tilde{F}_A, \tilde{F}_B\}$ of the compared image pair.

Specifically, as in Eq.(5), the queries Q_A are projected from the first set of local descriptors $\tilde{F}_A$, while the keys K_B and values V_B are projected from $\tilde{F}_B$. Then the encoded vectors V_A' are generated by Eq.(6), which is formally the weighted average of the values V_B. Intuitively, an encoded $V_A'(i)$ integrates information from all value vectors whose keys have a high correlation with the query $Q_A(i)$. To combine relevant information into the local representation, the encoded vectors V_A' are first L_2-normalized (L_2) and

concatenated with the original local descriptors $\tilde{F}_A$. Then a multilayer perception (MLP) network projects the concatenated vectors into the residual features, which are normalized and added on $\tilde{F}_A$ to formulate the final descriptors $\tilde{F}_A^O$.

$$Q_A = \tilde{F}_A W_Q^T, \quad K_B = \tilde{F}_B W_K^T, \quad V_B = \tilde{F}_B W_V^T \tag{5}$$

$$V_A' = \text{Softmax}\left(\frac{Q_A K_B^T}{\sqrt{D}}\right) V_B \tag{6}$$

$$\tilde{F}_A^O = \text{L}_2(\tilde{F}_A + \text{L}_2(\text{MLP}(\text{Concat}(\tilde{F}_A, \text{L}_2(V_A'))))) \tag{7}$$

Eq.(5)$\sim$(7) can be unified into a one-way cross-attention enhancement layer $\mathscr{F}_{\text{CE}}$: $\tilde{F}_A^O = \mathscr{F}_{\text{CE}}(\tilde{F}_A, \tilde{F}_B)$, where $\tilde{F}_A$ encodes the interactive information from $\tilde{F}_B$. To realize the mutual enhancement between the local descriptors of a query image and its compared reference image, the two sets of descriptors $\{\tilde{F}_q, \tilde{F}_r\}$ are fed into the cross-attention enhancement layer $\mathscr{F}_{\text{CE}}$ and mutually enhanced in sequence. Formally, the mutual enhancement is expressed as Eq.(8).

$$\tilde{F}_q^O = \mathscr{F}_{\text{CE}}(\tilde{F}_q, \tilde{F}_r), \quad \tilde{F}_r^O = \mathscr{F}_{\text{CE}}(\tilde{F}_r, \tilde{F}_q^O) \tag{8}$$

D. Cross-Matching and Geometric Consistency

Relative local saliency. According to Eq.(3)$\sim$(4), the contribution of a local feature $\mathbf{x}_i$ to the cluster-wise local representation $\gamma_k \cdot \tilde{\mathbf{r}}_k(\mathbf{x}_i)$ is determined by its intra-cluster saliency $\alpha_k(\mathbf{x}_i)\beta_k(\mathbf{x}_i)$ and the cluster saliency γ_k. Therefore, the overall saliency η of a local descriptor $\mathbf{f}_i$ is defined as the accumulation of its cluster-wise saliency weights:

$$\eta(\mathbf{f}_i) = \sum_{k=1}^{K} \alpha_k(\mathbf{x}_i)\beta_k(\mathbf{x}_i)\gamma_k \tag{9}$$

To amplify the difference in the significance of local descriptors to geometric consistency assessment, we introduce a contrast enhancement strategy based on standardization. As in Eq.(10), the relative local saliency of each descriptor $\tilde{\eta}(\mathbf{f}_i)$ is calculated as the standardized local saliency $\eta(\mathbf{f}_i)$ across all local descriptors $F \in \mathbb{R}^{HW \times 4096}$. It can be seen from Fig.(I) that the relative local saliency can well quantify the varying significance of visual elements to the task.

$$\tilde{\eta}(\mathbf{f}_i) = (\eta(\mathbf{f}_i) - \min_{\mathbf{f}_j \in F} \eta(\mathbf{f}_j)) / (\max_{\mathbf{f}_j \in F} \eta(\mathbf{f}_j) - \min_{\mathbf{f}_j \in F} \eta(\mathbf{f}_j)) \tag{10}$$

Correspondence mining. The enhanced local descriptors $\tilde{F}_q^O \in \mathbb{R}^{HW \times 4096}$ and $\tilde{F}_r^O \in \mathbb{R}^{HW \times 4096}$ are two sets of unit vectors, so the cross-matching similarities between local descriptors can be intuitively obtained by matrix multiplication of the descriptors, weighted by the binary term $\tilde{\eta}(\tilde{F}^O) \geq t$ indicating if their local saliency exceed the threshold t. To mine potential correspondences from the cross-matching, we follow the latest SOTAs [23], [29], [46] to use dual-softmax ($\mathscr{F}_{\text{DS}}$) and mutual nearest neighbor search ($\mathscr{F}_{\text{MNN}}$) strategies. By defining a confidence matrix $\mathscr{M}_c$ as Eq.(11), a set of mutual matches $\mathscr{P}_{matches}$ can be filtered based on the criteria in Eq.(12). The constant ξ in Eq.(11) is chosen to be a large positive number, so that each element in $\mathscr{M}_c$ tends to be binary after dual-softmax.

$$\mathscr{M}_c = \mathscr{F}_{\text{DS}}(\xi((\tilde{\eta}(\tilde{F}_q^O) \geq t)\tilde{F}_q^O)((\tilde{\eta}(\tilde{F}_r^O) \geq t)\tilde{F}_r^O)^T) \tag{11}$$

$$\mathscr{P}_{matches} = \{(i,j) | \forall (i,j) \in \mathscr{F}_{\text{MNN}}(\mathscr{M}_c)\} \tag{12}$$

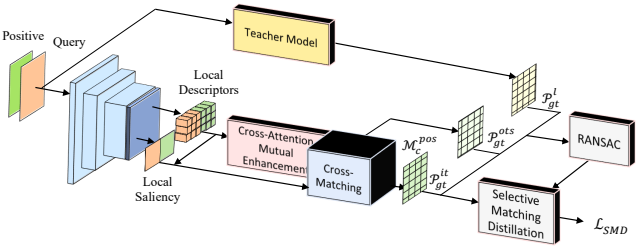

Fig. 3. The diagram of local correspondence distillation pipeline.

After obtaining the mutual matches $\mathscr{P}_{matches}$, the geometric consistency of the compared image pair can be evaluated by the number of inlier matches verified by RANSAC.

E. Selective Matching Distillation

In order for a positive reference I_r^p to have better geometric consistency with the query I_q than any negative reference I_r^n, there should be more potential correspondences between a truly matched image pair. To this end, we propose a distillation pipeline as in Fig.(3), through which the model can learn from SOTA local matchers.

The latest SOTAs [26], [28], [29] for feature matching excel at finding dense correspondences, although they may focus on areas that are not discriminative for VPR. (*e.g.*, repetitive building structures). To enable CAHIR to selectively learn more potential matches, we introduce the pre-trained LoFTR [29] as a teacher model for distillation. Specifically, the three batches of RANSAC-verified matches predicted by LoFTR ($\mathscr{P}_{gt}^{l}$), off-the-shelf CAHIR ($\mathscr{P}_{gt}^{ots}$), and CAHIR in training ($\mathscr{P}_{gt}^{it}$) are first fused in the order of priority through Eq.(13)$\sim$Eq.(14).

$$\begin{aligned}\mathscr{F}_{\text{fuse}}(\mathscr{P}_1, \mathscr{P}_2) = \mathscr{P}_1 \cup \{(i,j) | (i,j) \in \mathscr{P}_2 \ \ and \\ (i,\cdot) \notin \mathscr{P}_1 \ \ and \ \ (\cdot,j) \notin \mathscr{P}_1\}\end{aligned} \tag{13}$$

$$\mathscr{P}_{gt}^{+} = \mathscr{F}_{\text{fuse}}(\mathscr{F}_{\text{fuse}}(\mathscr{P}_{gt}^{l}, \mathscr{P}_{gt}^{ots}), \mathscr{P}_{gt}^{it}) \tag{14}$$

$$\mathscr{P}_{gt}^{-} = \{(i,k),(l,j) | (i,j) \in \mathscr{P}_{gt}^{+}, \ k \neq j, \ l \neq i\} \tag{15}$$

We treat $\mathscr{P}_{gt}^{+}$ as the positive matches for distillation supervision. The corresponding negative matches $\mathscr{P}_{gt}^{-}$ can be obtained by Eq.(15). Considering that not all true matches can be found by LoFTR and retained by RANSAC filtering, to avoid CAHIR being restricted by $\mathscr{P}_{gt}^{+}$, we propose a selective matching distillation loss $\mathscr{L}_{SMD}$ as in Eq.(16). In $\mathscr{L}_{\text{SMD}}$, the predicted matches that are not contained in $\mathscr{P}_{gt}^{+}$ and $\mathscr{P}_{gt}^{-}$ will not be penalized by the distillation. This allows the model to explore more potential matches with transformation patterns similar to the ground-truth $\mathscr{P}_{gt}^{+}$.

$$\begin{aligned}\mathscr{L}_{\text{SMD}} = -\frac{1}{|\mathscr{P}_{gt}^{+}|}\Big(\sum_{(i,j)\in\mathscr{P}_{gt}^{+}} \log \mathscr{M}_{conf}(i,j) \\ + \sum_{(k,l)\in\mathscr{P}_{gt}^{-}} \log(1 - \mathscr{M}_{conf}(k,l))\Big)\end{aligned} \tag{16}$$

TABLE I

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS. **GV** IS AN ABBREVIATION FOR GEOMETRIC VERIFICATION, WHICH INDICATES WHETHER A METHOD INCLUDES ADDITIONAL GEOMETRIC CONSISTENCY ASSESSMENT TO RE-RANK THE TOP K=20 RETRIEVED CANDIDATES.

Method	GV	Pitts30k-test			Pitts250k-test			Tokyo24/7			Mapillary-val		
		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 [14]	×	85.2	92.8	95.0	86.5	93.8	95.5	67.9	78.4	81.3	62.7	75.3	79.7
APPSVR [38]	×	87.4	94.3	95.8	88.8	95.6	96.8	77.1	85.7	89.5	63.7	77.6	81.9
SuperGlue [28]	√	88.2	94.9	95.9	90.7	95.6	96.7	82.5	83.2	83.8	78.4	82.4	83.2
LoFTR [29]	√	88.4	94.1	95.3	91.1	95.6	96.4	81.6	83.8	84.1	77.6	81.6	83.2
PatchNetVLAD [23]	√	89.8	95.1	**96.5**	91.6	95.9	96.5	79.4	83.5	83.5	74.5	80.1	82.2
CAHIR	√	**90.1**	**95.4**	96.0	**92.3**	**96.9**	**97.5**	**90.5**	**92.1**	**92.4**	**81.9**	**88.2**	**90.3**

V. EXPERIMENTS

A. Datasets and Evaluation Metric

Four benchmark datasets for city-scale VPR are used to evaluate our method: Pittsh250k [47], Pitts30k [14], Tokyo24/7 [13], and Mapillary [48]. Following the latest SOTAs [18], [23], [38], we use Pitts30k as the training set for evaluation on Pitts30k, Pitts250k, and Tokyo24/7. Models trained on Mapillary are tested on the validation set, Mapillary-val. We use the recommended configuration of datasets for benchmarking, where the performance of models is evaluated by the *Recall@N* metric.

B. Implementation Details

We implement CAHIR in Pytorch framework. The VGG16 backbone is initialized with MatConvNet pretrained weights. Since our hierarchical image descriptors are generated through the shared encoding pipeline, we first train the CAHIR feature extractor using the same protocol as in [38]. After training and freezing the feature extractor, other parametric modules in CAHIR are then fine-tuned through the distillation pipeline in Sec.IV-E. A Stochastic Gradient Descent optimizer is used to minimize the loss function Eq.(16) for 30 epochs, where the learning rate 0.001 is reduced by a factor of 10 every 15 epochs.

C. Comparison with State-of-The-Arts

Among global descriptors for VPR, we compare against **NetVLAD** [14] and **APPSVR** [38]. As for local representations, we adapt **SuperGlue** [28] and **LoFTR** [29] to the VPR task. Equipped with NetVLAD for global retrieval, they are used to re-rank the retrieved candidates through RANSAC scoring. We also compare with **PatchNetVLAD** [23], which fuses multi-scale locally global descriptors and achieves the SOTA performance on benchmark datasets.

Table I compares the performance of CAHIR to other benchmark models. Provided with the top $K = 20$ globally retrieved candidates, geometric verification (GV) re-ranks these candidates according to either the number of verified matches or the customized geometric consistency score. It can be seen that the methods with additional geometric verification (SuperGlue, LoFTR, PatchNetVLAD, and CAHIR) unsurprisingly outperform the global descriptors (NetVLAD, APPSVR). Besides, the local feature matchers SuperGlue and LoFTR set up strong baselines when adapting to the VPR task, especially on Tokyo24/7 and Mapillary where the appearance of scenes changes drastically. PatchNetVLAD eliminates the computational overhead of performing local feature extraction separately, while still obtaining remarkable results on par with SuperGlue. Our CAHIR steadily surpasses all other baselines, which demonstrates the comprehensive advantages of the proposed components. Compared with our base global descriptor APPSVR, an increase in *Recall@1* of 2.7%, 3.5%, 13.4%, and 18.2% can be observed on Pitts30k, Pitts250k, Tokyo24/7, and Mapillary.

TABLE II

ABLATION STUDY. 'GV' DENOTES GEOMETRIC VERIFIATION. 'ME' DENOTES MUTUAL ENHANCEMENT BEFORE CROSS-MATCHING.

Method	Components		Pitts30k-test			Tokyo24/7		
	GV	ME	r@1	r@5	r@10	r@1	r@5	r@10
NetVLAD	×	×	85.2	92.8	94.9	67.9	78.4	81.3
CAHIR-G	×	×	88.2	94.1	95.7	79.4	89.2	91.4
CAHIR-OTS	√	×	89.4	95.1	**96.2**	90.4	91.7	92.4
CAHIR	√	√	**90.1**	**95.4**	96.0	**90.5**	**92.1**	**92.4**

TABLE III

EVALUATE THE DISTILLED LOCAL MATCHING. N_m REPRESENTS THE AVERAGE NUMBER OF VALID MATCHES FOUND FOR EACH IMAGE PAIR.

Method	Dataset	Off-the-shelf		Distilled	
		N_m	r@1	N_m	r@1
CAHIR	Pitts30k-test	122	89.4	178	90.2

D. More Results and Discussions

Ablation Studies. To validate the integrated components in our proposed method, we compare CAHIR variants that gradually enables each functional module. CAHIR-G retains only the global branch of CAHIR, disabling candidate re-ranking based on geometric verification. CAHIR-OTS performs geometric verification by cross-matching off-the-shelf local descriptors without mutual enhancement. CAHIR is our proposed model with all components enabled. As shown in Table II, applying each component incrementally results in steady performance improvements. CAHIR-OTS convincingly outperforms CAHIR-G, which demonstrates the benefits of exploiting the spatial relationships of visual elements that are ignored in the global descriptor generation. Further enabling the mutual enhancement layer brings another improvement on all three datasets, proving that the local information interaction can enhance the cross-matching ability of their local descriptors.

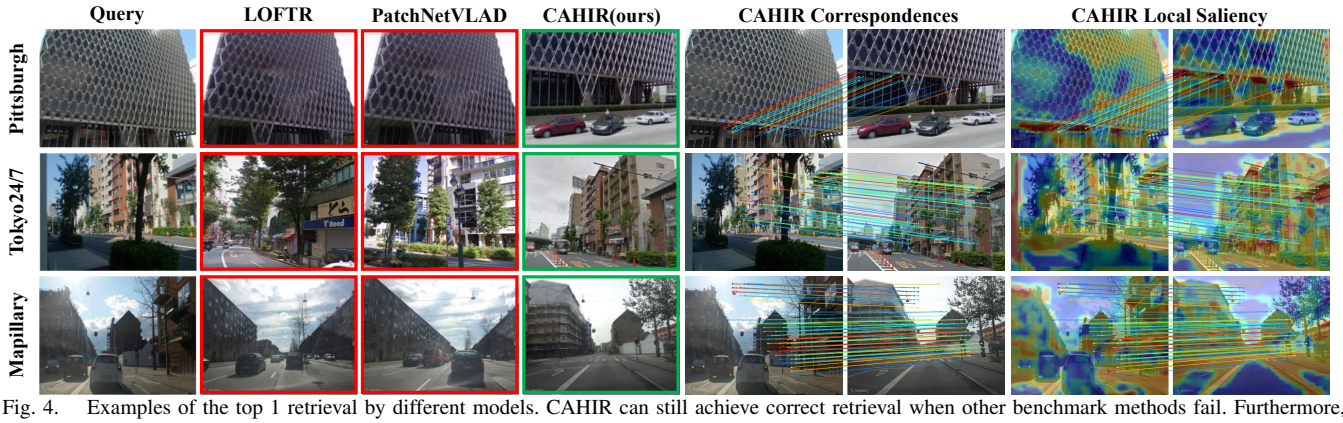

Fig. 4. Examples of the top 1 retrieval by different models. CAHIR can still achieve correct retrieval when other benchmark methods fail. Furthermore, the correspondences on saliency heat map shows that CAHIR prioritizes finding high-quality correspondences between task-relevant visual elements. For example, in the first row, matching between repetitive patterns on buildings is suppressed by CAHIR.

TABLE IV

COMPUTATIONAL AND MEMORY COST FOR PROCESSING AN IMAGE PAIR.

Method	Extraction	Matching	Memory
SuperGlue [28]	87.48 *ms*	29.33 *ms*	2.0 *MB*
LoFTR [29]	168.52 *ms*	78.43 *ms*	9.8 *MB*
PatchNetVLAD [23]	516.66 *ms*	37.19 *ms*	37.0 *MB*
CAHIR	17.31 *ms*	27.23 *ms*	3.8 *MB*

Distilled local descriptor matching. The mutual enhancement layer coupled with distillation learning is dedicated to finding local matches suitable for the VPR task. Rather than exploring as many matches as possible, it puts emphasis on finding the high-quality correspondences between task-relevant visual elements. From this perspective, it is destined that CAHIR cannot predict dense matching like its teacher model LoFTR. Nonetheless, Table.III shows that the trained CAHIR is able to find more valid local matches for each image pair than the off-the-shelf local descriptors. The performance improvement brought also verifies the rationality of the mutual enhancement of local descriptors and the distilled descriptor matching.

Inference latency and memory footprint. Table IV presents a comparison between local approaches in terms of computational time and memory footprint for processing an image pair. For full local descriptor extraction directly from images, CAHIR is 5.1 times faster than the second best SuperGlue. The extraction latency of CAHIR can be further reduced to negligible if reusing the intermediate features cached in the global encoding pipeline, while other models require separate local feature extraction or post-processing. CAHIR also has the best matching latency, which is 2.9 times faster than the teacher model LoFTR. In terms of memory footprint, CAHIR is slightly inferior to the best SuperGlue due to the different local descriptor dimensions. Overall, CAHIR shows strong competitiveness in model efficiency.

E. Qualitative Results.

Fig.(4) depicts a few challenging queries sampled from the three employed datasets, and the top retrieved image by different methods. It can be noted that in the case of failure of other benchmark models, CAHIR can still achieve correct retrieval. In addition, the visualized local saliency indicates that CAHIR prioritizes long-term static objects that are more stable and discriminative for VPR. The local matches found by CAHIR shows that our method places great emphasis on the correspondences between task-relevant visual elements. Fig.(5) illustrates two typical cases of the selective matching distillation, where CAHIR is able to selectively learn from the teacher model and further improve on this basis.

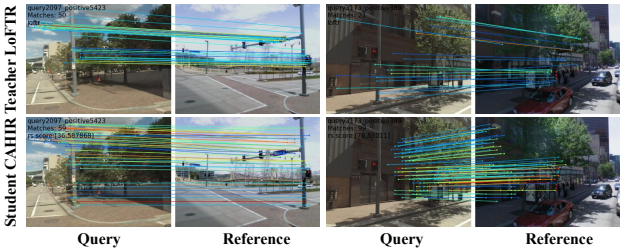

Fig. 5. Selective matching distillation allows CAHIR to explore more potential correspondences (left) and selectively ignore bad supervision from the teacher model (right).

VI. CONCLUSIONS

In this paper, we propose a coarse-to-fine VPR framework named CAHIR. It unifies global and local descriptor generation into one encoding pipeline. The hierarchical descriptors can be used for global retrieval and geometric verification progressively. To obtain high-quality local matching for geometric verification, a locally mutual enhancement layer is introduced to strengthen the information interaction between the local descriptors to be compared. By introducing a distillation pipeline with novel selective matching loss, the parametric model can further learn from the SOTA local matcher. Extensive experiments demonstrate that CAHIR outperforms the SOTA global and local image representations for VPR on benchmark datasets.

ACKNOWLEDGEMENT

This study is supported under the RIE2020 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

REFERENCES

[1] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 920–929, 2017.

[2] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *BMVC*, 2012.

[3] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1043–1050, 2012.

[4] Z. Wu, Y. Yue, M. Wen, J. Zhang, J. Yi, and D. Wang, "Infrastructure-free hierarchical mobile robot global localization in repetitive environments," *IEEE Trans. Instrum. Meas. (TIM)*, vol. 70, pp. 1–12, 2021.

[5] Z. Wu, Y. Yue, M. Wen, J. Zhang, G. Peng, and D. Wang, "MSTSL: Multi-sensor based two-step localization in geometrically symmetric environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2021, pp. 5245–5251.

[6] Y. He, J. Wang, D. Su, K. Nakadai, J. Wu, S. Huang, Y. Li, and H. Kong, "Observability analysis of graph slam-based joint calibration of multiple microphone arrays and sound source localization," in *2023 IEEE/SICE International Symposium on System Integration (SII)*, 2023, pp. 1–8.

[7] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *CVPR*, 2011.

[8] E. Chalmers, E. B. Contreras, B. Robertson, A. Luczak, and A. J. Gruber, "Learning to predict consequences as a method of knowledge transfer in reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 2259–2270, 2018.

[9] C. McManus, W. Churchill, W. P. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 901–906, 2014.

[10] R. Mur-Artal, J. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, pp. 1147–1163, 2015.

[11] S. Eiffert, N. D. Wallace, H. Kong, N. Pirmarzdashti, and S. Sukkarieh, "Resource and response aware path planning for long-term autonomy of ground robots in agriculture," *ArXiv*, vol. abs/2105.10690, 2021.

[12] R. Arandjelovic and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *ACCV*, 2014.

[13] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[14] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[15] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3251–3260, 2017.

[16] A. Khaliq, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes," *ArXiv*, vol. abs/1811.03032, 2018.

[17] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *ACM Multimedia*, 2018.

[18] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019.

[19] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[20] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3476–3485, 2016.

[21] B. Cao, A. F. de Araújo, and J. Sim, "Unifying deep local and global features for image search," in *ECCV*, 2020.

[22] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019.

[23] 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.

[24] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf," *2011 International Conference on Computer Vision*, pp. 2564–2571, 2011.

[26] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 337–33 712, 2018.

[27] M. Teichmann, A. F. de Araújo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5104–5113, 2019.

[28] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946, 2020.

[29] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021.

[30] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *ECCV*, 2014.

[31] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," *ArXiv*, vol. abs/1812.03506, 2018.

[32] M. C. Dorst, "Distinctive image features from scale-invariant keypoints," 2011.

[33] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.

[34] G. Schindler, M. A. Brown, and R. Szeliski, "City-scale location recognition," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.

[35] T. Sattler, M. Havlena, F. Radenović, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2102–2110, 2015.

[36] A. Babenko and V. S. Lempitsky, "Aggregating local deep features for image retrieval," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1269–1277, 2015.

[37] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3384–3391, 2010.

[38] G. Peng, J. Zhang, H. Li, and D. Wang, "Attentional pyramid pooling of salient visual residuals for place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 885–894.

[39] 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)*, 2019, pp. 2570–2579.

[40] 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*.

[41] M. A. Fischler and R. C. Bolles, "A paradigm for model fitting with applications to image analysis and automated cartography," 1987.

[42] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Toward geometric deep slam," *arXiv preprint arXiv:1707.07410*, 2017.

[43] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[44] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *ICML*, 2020.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[46] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," in *NeurIPS*, 2018.

[47] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition

with repetitive structures," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, 2013.

[48] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2623–2632, 2020.