CQVPR: Landmark-aware Contextual Queries for Visual Place Recognition

1st Dongyue Li *Nagoya University*
2nd Jialei Chen *Nagoya University*
3th Ito Seigo *Nagoya University*
4th Hiroshi Murase *Nagoya University*
5rd Daisuke Deguchi *Nagoya University*

***Abstract*—Visual place recognition remains challenging due to significant variations in appearance caused by lighting, viewpoint, and structural similarities across environments. To address this, we propose Contextual Query VPR (CQVPR), a novel method that bridges the gap between pixel-level and segment-level representations. Unlike conventional approaches that either rely on low-level appearance cues or high-level semantic partitions, CQVPR integrates fine-grained visual details with global contextual understanding through a set of learnable queries. These contextual queries capture high-level semantic structures within the scene, which are fused with dense pixel-wise features to form robust descriptors for retrieval. To encourage discriminative query learning, we introduce a query matching loss that promotes similarity among queries from the same location while pushing those from different locations apart. Extensive experiments on several datasets demonstrate that the proposed method outperforms other state-of-the-art methods, especially in challenging scenarios.**

***Index Terms*—visual place recognition, attention mechanism**

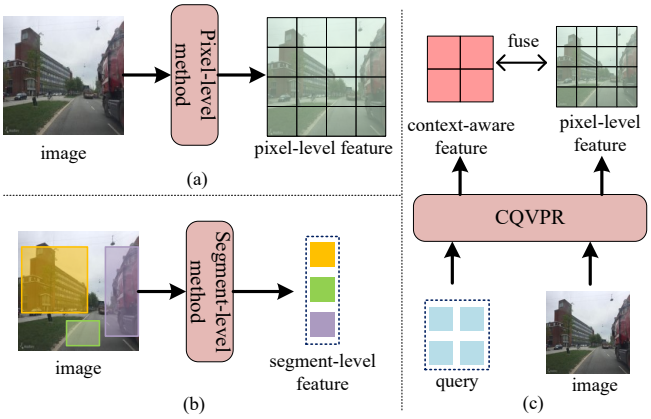

Fig. 1. Conceptual difference among three VPR pipelines. (a) Pixel-level methods. (b) Segment-level methods. (c) The proposed CQVPR.

I. Introduction

Visual Place Recognition (VPR), also referred to as image localization [1] or visual geo-localization [2], aims to estimate the geographic location of a query image by matching it against a database of geo-tagged images. In the context of intelligent transportation systems (ITS), VPR plays a crucial role in enabling robust localization for autonomous vehicles, intelligent navigation, and vehicle-to-infrastructure (V2I) communication. Accurate and efficient VPR is foundational for a wide range of ITS applications, including urban driving, fleet coordination, and infrastructure-aware decision-making. Beyond transportation, VPR has also been widely applied in robotics [3]–[5], augmented reality [6], and pose estimation [7]. Despite its significance, achieving reliable VPR in real-world traffic environments remains challenging due to dynamic lighting, weather variability, viewpoint changes, and perceptual aliasing [8], where visually distinct places may appear similar. Addressing these challenges is essential for deploying scalable and dependable VPR systems in increasingly complex urban mobility scenarios.

With the increasing demand for high-precision localization in fields such as autonomous driving and robotic navigation, the limitations of the Global Positioning System (GPS) have become more evident. GPS suffers from signal blockage in areas under elevated structures, where the GPS receiver cannot capture signals from satellites, leading to a decrease in accuracy. Additionally, in environments like urban canyons, where buildings are closely packed, satellite signals are often reflected, causing multipath effects that further degrade the positioning accuracy. Furthermore, GPS is susceptible to weather disturbances, signal interference, and spoofing attacks, which restrict its reliability in critical scenarios. In contrast, VPR achieves localization by extracting and matching visual features of captured images without relying on satellite signals. VPR demonstrates strong robustness and high precision in complex outdoor environments, making it an effective complement to GPS, especially in situations where satellite-based localization is unavailable or unreliable.

Pixel-level methods [9], [10], which rely solely on visual cues, aim to recognize places based on image appearance. While effective in capturing fine-grained visual information, these methods often lack high-level semantic context, resulting in an overemphasis on visually salient yet semantically uninformative regions. In contrast, segment-level methods typically partition an image into semantically meaningful regions using clustering techniques or semantic segmentation models that delineate object-level components [11]. These segments often correspond to structural elements of landmarks, thereby enhancing place recognition performance. However, segment-level methods generally lack the dense pixel-wise feature representations necessary to preserve spatial consistency and detailed visual cues. To reconcile the strengths of both paradigms, we propose a hybrid method that integrates query-level semantic features with dense pixel-wise representations,

aiming to achieve both semantic robustness and fine-grained visual representations.

We introduce Contextual Query VPR (CQVPR), a novel approach that bridges the gap between pixel-level and segment-level methods. Fig. 1 illustrates the conceptual difference of our method and previous works. Specifically, It leverages a set of learnable queries [12] to encode high-level contextual information about both the landmarks and their surroundings, where each contextual query represents a latent high-level context such as certain objects or structural shapes. Visualization results are present in Fig. 2 and Section IV-D2. The heatmap, which depicts the regions that each query concentrates on, is fused with pixel-level features for producing global and local retrieving descriptors. Additionally, to enable the network to learn more discriminative queries, we designed a loss function that encourages the query embeddings of images from the same place to be as similar as possible, while ensuring that the query embeddings of images from different places are as dissimilar as possible. Experimental results demonstrate the proposed CQVPR can achieve accurate visual place recognition results, even in challenging scenarios involving large scale and viewpoint variations. To summarize, the main contributions of this work are as follows:

(1) Through a learnable Transformer module, the task-related contextual queries are extracted. These inferred queries, containing rich high-level contexts, can be transformed into context-aware features for more effective visual place recognition.

(2) Contextual queries of images from the same place should be similar since they depict the same scene. Therefore, a query matching loss is proposed to maximize the similarity of queries between images from the same place while minimizing it conversely.

(3) Extensive experiments on various well-known benchmark datasets demonstrate that our proposed method outperforms several state-of-the-art baseline methods under different scenarios.

II. Related Works

A. Pixel-level methods

Pixel-level VPR methods are significant approaches in VPR. These methods typically begin by generating pixel-level features, followed by a global retrieval stage. The global retrieval stage aims to efficiently retrieve the top-k candidate images from a large database using global feature representations. The global feature is obtained by pooling pixel-level features across the entire image. For example, CricaVPR [13] generates the pixel-level features through the cross-view interaction and then pools these pixel-level features into a global vector.

In addition to the global retrieval stage, some methods include a re-ranking stage. The re-ranking stage involves refining the global retrieved results by performing local feature matching between the query image and the top-k candidates. Local features are often pixel-level features or their upsampled versions. The re-ranking score between two images is determined by the number of matched points between them.

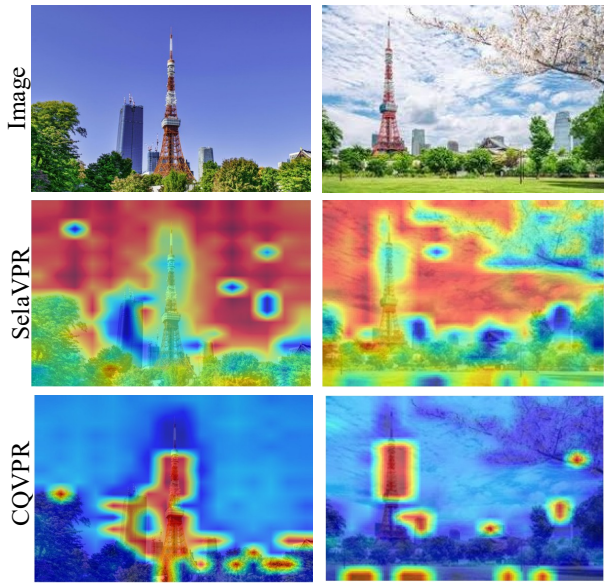

Fig. 2. Comparison of cross attention maps between the previous pixel-level method SelaVPR and our CQVPR. The presented two images are from the same place. Thanks to the introduced high-level contexts, CQVPR focuses on discriminative regions (e.g., buildings and towers). While SelaVPR focuses on less informative regions.

The final re-ranking prediction is derived from this score. For example, AANet [14] proposes an algorithm to align the local features under spatial constraints. R^2Former [15] proposes a unified retrieval and re-ranking framework with only Transformers.

Recently, DINOv2 [16] has achieved impressive performance in the VPR task. SelaVPR [10] achieves the state-of-the-art performance through fine-tuning the DINOv2's features to attend to more distinctive regions. AnyLoc [17] directly adopt the DINOv2 as backbone to establish an universal VPR solution. However, these methods still do not incorporate the explicit high-level contexts, making their descriptors overly rely on visual cues and can not well focus on landmarks during the local matching process.

B. Segment-level methods

Different from pixel-level methods, segment-level methods focus on generating features at a more abstract level, such as object-level, cluster-level and region-level. The global feature of an image can be seen as a special case of segment-level features, where the entire image is treated as a single segment.

In the early stage of VPR, aggregation algorithms like VLAD [18] and Bag of Words (BoW) [19] treat an image as a set of cluster centroids and generate cluster-level features, which are then aggregated for obtaining the final global feature representation. NetVLAD [20] is proposed to make the VLAD algorithm differentiable, allowing it to be seamlessly integrated into any neural network. Patch-NetVLAD [21] tends to assign the NetVLAD pooled feature to each patch of the image. BoQ [22], which is similar to our method, views an image as a set of queries and directly outputs the combination of these

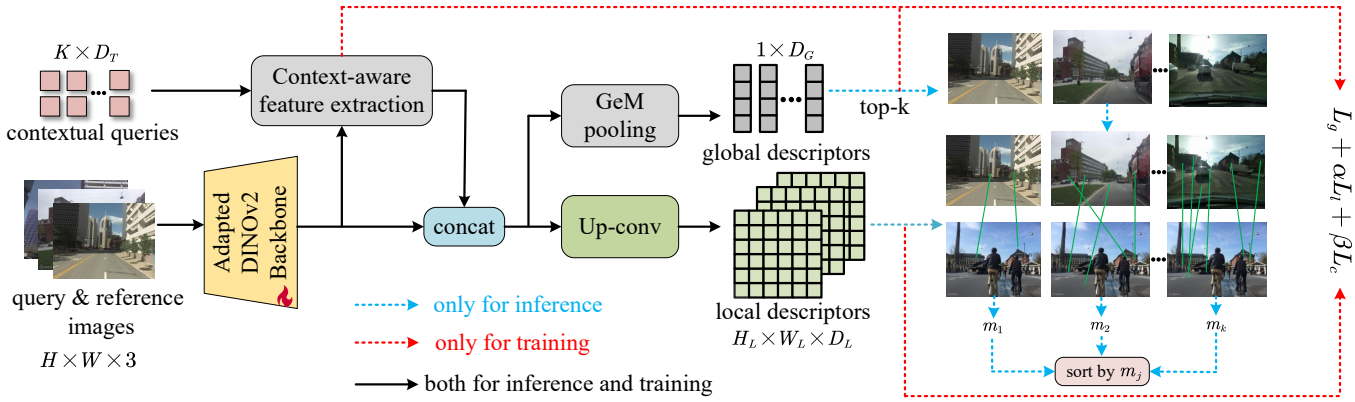

Fig. 3. The overview of the proposed CQVPR. Learnable contextual queries are first randomly initialized and then transformed into context-aware features. The context-aware features are then fused with pixel-level features from the backbone.

queries as the global feature. Although BoQ also introduces learned queries, it can not generate pixel-wise features and therefore can not do local matching. Specifically, the queries in BoQ are processed into global features. Due to the discrete nature of these queries, generating local features is infeasible. In contrast, our proposed CQVPR can generate both global and local features, providing a more comprehensive and versatile representation.

Recently, some methods have leveraged the explicit semantic segmentation model to generate object-level features. For example, [23] adopts the semantic segmentation backbone to generate features at each semantic class. Similar to PatchNetVLAD, SegVLAD [11] leverages the recent SAM [24] model to assign the NetVLAD pooled features to each semantic object and directly do matching at the object level. Despite the impressive performance achieved by these segment-level methods, their features lack spatial, appearance, and contextual information, which are crucial for distinguishing between different landmarks.

III. Methodology

Fig. 3 provides a detailed overview of the Contextual Query VPR (CQVPR) pipeline. Humans recognize places not just by landmarks' appearance and semantics, but also by their surrounding context, like nearby objects, trees, and streets. Building on this perspective, CQVPR is proposed to bridge the gap between pixel-level and segment-level approaches by integrating visual features with contextual information through a novel mechanism. CQVPR achieves this by leveraging a set of learnable queries to encode high-level latent contexts within an image. Each query captures specific contextual features, such as object shapes or structural elements, providing a broader understanding of the scene beyond just the landmarks.

A. Context-aware feature extraction

Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the adapted DINOv2 [10] is employed to obtain the pixel-level feature $\mathbf{F_V}$, which is then combined with context-aware features for obtaining the global and local descriptors.

We hypothesize that each image can be viewed as K distinct contextual queries. The learnable contextual embedding $\mathbf{T} \in \mathbb{R}^{K \times D_T}$ representing K contextual queries is first randomly initialized, and then updated through a cross attention layer,

$$\mathbf{T} = CA(\mathbf{T}, \text{conv}(\mathbf{F_V})), \tag{1}$$

where $CA(\cdot)$ is the cross attention layer between queries $\mathbf{T}$, keys $\mathbf{F_V}$, and values $\mathbf{F_V}$. $\mathbf{F_V}$ is the pixel-level feature mentioned above and conv$(\cdot)$ stands for the convolution layer. Since D_C, the number of channels of $\mathbf{F_V}$, is very large, the convolution layer here is employed to reduce it to D_T for efficiency. The heatmap $\mathbf{H}$ is generated through computing the similarity between $\mathbf{F_V}$ and $\mathbf{T}$.

$$\mathbf{H} = \langle \mathbf{T}, \text{conv}(\mathbf{F_V}) \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product and conv$(\cdot)$ stands for the same convolution layer in Eq 1. The heatmap $\mathbf{H}$ is then transformed to be the context-aware feature $\mathbf{F_T}$.

$$\mathbf{F_T} = \text{MLP}(\text{norm}(\mathbf{H})), \tag{3}$$

where norm$(\cdot)$ means the normalization along the channel dimension and MLP$(\cdot)$ denotes the multi-layer perceptron layer.

B. Global and local descriptors

After the context-aware feature $\mathbf{F_T}$ and pixel-level feature $\mathbf{F_V}$ are obtained, the global descriptor $\mathbf{G}$ can be generated as follows

$$\mathbf{G} = \text{L2}(\text{GeM}([\mathbf{F_V}, \mathbf{F_T}])), \tag{4}$$

where L2$(\cdot)$ denotes the L2 normalization, and GeM$(\cdot)$ represents the GeM pooling [25]. $[\cdot, \cdot]$ indicates the concatenation along the channel dimension. Based on global descriptors, a similarity search is performed in the global feature space across reference images, retrieving the top-k most similar candidate images to the query image. To obtain the final predictions, the local descriptors are leveraged for re-ranking

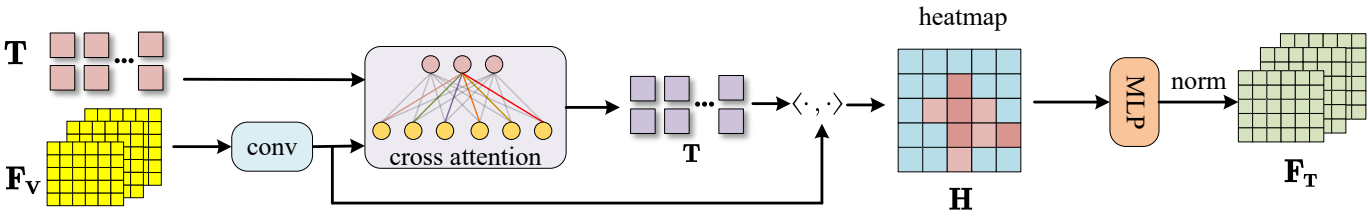

Fig. 4. The illustration of the context-aware feature extraction module.

these candidates. For each image, the local descriptor $\mathbf{L}$ can be obtained through up-sampling the fused $\mathbf{F_V}$ and $\mathbf{F_T}$

$$\mathbf{L} = \text{L2}(\text{up-conv}([\mathbf{F_V}, \mathbf{F_T}])), \tag{5}$$

where the up-conv$(\cdot)$ is the up-convolution layer and the L2$(\cdot)$ denotes the L2 normalization. Local descriptors are leveraged to do local feature matching [26] and the number of matches is treated to be the score for re-ranking the top-k candidates.

C. Loss functions

To optimize the model for generating global descriptors, a global loss L_g based on the triplet loss [20] is proposed to weakly supervise the overall network,

$$L_g = \sum_j l(\|\mathbf{G}_q - \mathbf{G}_p\| + m - \|\mathbf{G}_q - \mathbf{G}_{n,j}\|), \tag{6}$$

where $l(x) = \max(x, 0)$ is the hinge loss, m is the margin. $\mathbf{G}_q$, $\mathbf{G}_p$, and $\mathbf{G}_{n,j}$ denote the global descriptors of the query, positive, and negative images, respectively, which are computed through Eq. 6.

For local matching, a mutual matching loss L_l [10] is leveraged for optimizing the network to produce local descriptors that are easier to be matched. Additionally, to better supervise the extraction of contextual queries, a contextual query matching loss L_c is introduced. When two images are from the same location, the similarity between their learned contextual embeddings is enlarged, whereas for images from different locations, the similarity is reduced.

$$\begin{aligned}
L_c &= \sum_k l(s_{n,k} - s_p), \\
s_p &= \frac{1}{|\mathbf{M}_t|} \sum_{(i,j)\in\mathbf{M}_t} \mathbf{T}_q^T(i)\mathbf{T}_p(j), \\
s_{n,k} &= \frac{1}{|\mathbf{M}'_t|} \sum_{(i',j')\in\mathbf{M}'_t} \mathbf{T}_q^T(i')\mathbf{T}_{n,k}(j'), \\
\mathbf{M}_t &= \{(i,j) \mid \forall\,(i,j) \in \text{MNN}\left(\mathbf{T}_q^T\mathbf{T}_p\right)\}, \\
\mathbf{M}'_t &= \{(i',j') \mid \forall\,(i',j') \in \text{MNN}\left(\mathbf{T}_q^T\mathbf{T}_{n,k}\right)\},
\end{aligned} \tag{7}$$

where $\mathbf{T}_q$, $\mathbf{T}_p$, and $\mathbf{T}_{n,k}$ represent the learned contextual embeddings of the query, positive, and negative images, respectively. MNN$(\cdot)$ denotes the mutual nearest neighbor criteria [26]. $l(x)$ is the hinge loss. Finally, the overall loss L can be obtained as

$$L = L_g + \alpha L_l + \beta L_c, \tag{8}$$

where α, β are the hyperparameters used to weight L_l and L_c.

IV. Experiments

A. Datasets and Metric

We evaluate the proposed CQVPR on several benchmark datasets that are widely used in the VPR task, including Tokyo 24/7, MSLS, Pitts30k, Pitts250k, SPED, AmsterTime and SVOX. These datasets are selected to cover diverse environments and challenging conditions such as illumination changes, viewpoint and seasonal variations.

In the experiments, the performance is measured by using Recall@N (R@N), which indicates the percentage of queries for which at least one of the N retrieved database images falls within a specified distance threshold of the ground truth. Following previous literature, a threshold of distance is usually set to 25 meters, except for AmsterTime, where the distance threshold is set to be 10 meters.

B. Implementation details

Given a 224×224 input image, the backbone would generate a 1024-dimensional feature $\mathbf{F_V}$, which has the spatial resolution of 14×14 pixels. The number of queries, K, is set to 10, and the channel dimension of the contextual embeddings, D_T, is set to 256. Before calculating the heatmap $\mathbf{H}$, a 1×1 convolution is leveraged to map $\mathbf{F_V}$ into a 256-dimensional feature for efficiency. After $\mathbf{H}$ is obtained, the MLP layer would expand the channel dimension of $\mathbf{H}$ to D_T, namely, 256. The 256-dimensional context-aware feature $\mathbf{F_T}$ and 1024-dimensional pixel-level feature $\mathbf{F_V}$ are concatenated along the channel dimension to establish the 1280-dimensional feature.

To obtain the global descriptor, the GeM pooling [25] is leveraged, which is a general pooling mechanism. For the local descriptor, two 3×3 up-convolutions with a stride of 2 and padding of 1 are employed to up-sample the 1280-dimensional feature, resulting in a 128-dimensional local descriptor with a spatial resolution of 61×61 pixels. In the re-ranking process, the top-100 candidates are re-ranked to obtain final results.

CQVPR is trained using the Adam optimizer, configured with a learning rate of 10^{-5} and a batch size of 4. Training is terminated when the Recall@5 (R@5) on the validation set fails to improve for three consecutive epochs. The training procedure defines positive images as reference images located within 10 meters of the query image, while definite negatives are those positioned beyond 25 meters. The margin parameter m in Eq. 6 is set to 0.1, and α, β in Eq. 8 are both set to 1.

TABLE I
COMPARISON ON PITTS30K, TOKYO24/7 AND MSLS-VAL DATASETS. THE BESTS RESULTS ARE HIGHLIGHTED IN BOLD. THE REPORTED PERFORMANCE OF BOQ IS DIRECTLY EXTRACTED FROM ITS ORIGINAL PAPER.

Method	**Pitts30k**			**Tokyo24/7**			**MSLS-val**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [20]	81.9	91.2	93.7	60.6	68.9	74.6	53.1	66.5	71.1
SFRS [27]	89.4	94.7	95.9	81.0	88.3	92.4	69.2	80.3	83.1
Patch-NetVLAD [21]	88.7	94.5	95.9	86.0	88.6	90.5	79.5	86.2	87.7
CosPlace [28]	88.4	94.5	95.7	81.9	90.2	92.7	82.8	89.7	92.0
TransVPR [9]	89.0	94.9	96.2	79.0	82.2	85.1	86.8	91.2	92.4
StructVPR [29]	90.3	96.0	97.3	-	-	-	88.4	94.3	95.0
GCL [30]	80.7	91.5	93.9	69.5	81.0	85.1	79.5	88.1	90.1
MixVPR [31]	91.5	95.5	96.3	85.1	91.7	94.3	88.0	92.7	94.6
EigenPlaces [32]	92.5	96.8	97.6	93.0	96.2	97.5	89.1	93.8	95.0
R^2Former [15]	91.1	95.2	96.3	88.6	91.4	91.7	89.7	95.0	96.2
BoQ [22]	92.4	96.5	97.4	-	-	-	91.2	95.3	96.1
SelaVPR [10]	92.8	96.8	97.7	**94.0**	**96.8**	97.5	90.8	**96.4**	**97.2**
CQVPR (Ours)	**93.3**	**96.9**	**98.1**	**94.0**	**96.8**	**98.1**	**91.5**	**96.4**	97.0

TABLE II
COMPARISON ON PITTS250K AND SPED DATASETS. THE BESTS RESULTS ARE HIGHLIGHTED IN BOLD. THE REPORTED BOQ'S PERFORMANCE IS DIRECTLY EXTRACTED FROM ITS ORIGINAL PAPER.

Method	**Pitts250k**		**SPED**	
	R@1	R@5	R@1	R@5
NetVLAD [20]	90.5	96.2	78.7	88.3
GeM [25]	87.0	94.4	66.7	83.4
Conv-AP [33]	92.9	97.7	79.2	88.6
CosPlace [28]	92.1	97.5	80.1	89.6
MixVPR [31]	94.6	98.3	85.2	92.1
EigenPlaces [32]	94.1	98.0	69.9	82.9
BoQ [22]	95.0	98.5	86.5	93.4
SelaVPR [10]	95.7	**98.8**	89.0	94.6
CQVPR (Ours)	**96.0**	98.7	**89.1**	**95.1**

TABLE III
COMPARISON (R@1) ON MORE CHALLENGING DATASETS.

Method	**AmsterTime**	**SVOX-NIGHT**	**SVOX-SUN**
SFRS [27]	29.7	28.6	54.8
CosPlace [28]	38.7	44.8	67.3
MixVPR [31]	40.2	64.4	84.8
EigenPlaces [32]	48.9	58.9	86.4
SelaVPR [10]	54.6	88.8	90.9
CQVPR (Ours)	**55.8**	**90.3**	**94.1**

CQVPR is first trained on MSLS and then subsequently trained on Pitts30k. For evaluation on MSLS-val, performance of the model only trained on MSLS is reported. For other datasets, the performance of the fully trained model is reported.

C. Comparison with State-of-the-Art Methods

As shown in Table I, CQVPR achieves the highest R@1 of 93.3 on Pitts30k, surpassing all other methods, including the recent SOTA method SelaVPR (ICLR 2024). On MSLS-val, which includes some suburban or natural scene images and is therefore prone to perceptual aliasing, our CQVPR still can achieve the best R@1, demonstrating its ability to produce more discriminative global and local feature descriptors to differentiate similar images from different places. CQVPR also excels in the Pitts250k and SPED datasets, as shown in Table II. It achieves the highest R@1 on Pitts250k, which demonstrates its ability of being employed in the large-scale datasets.

In Table I and II, besides the improvement of the CQVPR, it is also worth highlighting the metrics saturation observed in all the above five datasets. Therefore, the proposed CQVPR is additionally evaluated on more challenging datasets Amster-Time, SVOX-NIGHT and SVOX-SUN. The proposed CQVPR achieves improvements of +1.2%, +1.5%, and +3.2% in R@1 across these three datasets compared to the second-best method, demonstrating its superior performance in challenging scenarios with severe modality changes and illumination variations. This phenomenon proves that the introduced high-level contexts can improve the robustness of feature descriptors.

D. Visualization results

1) Qualitative results of local matching: In this section, we present the qualitative local matching results of our CQVPR method compared to SelaVPR, as shown in Fig. 5. Homography verification is employed here for clear visualization.

For two images from the same place, the more matches, the better. Conversely, for two images from different places, fewer matches are preferred since all matches in this case are incorrect. As illustrated in Fig. 5, our method extracts a

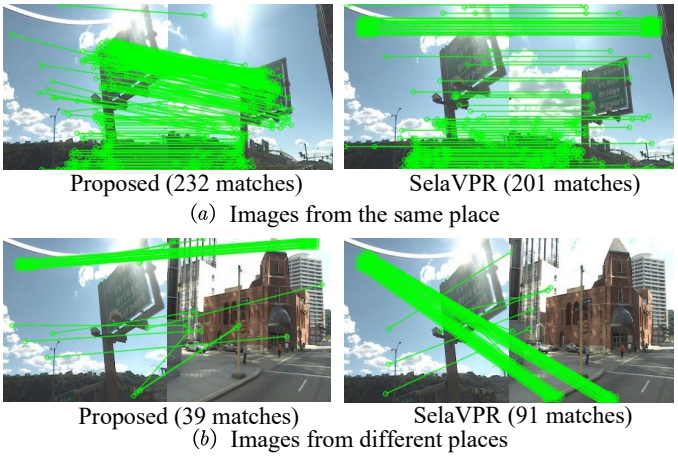

Fig. 5. Comparison of local matching between CQVPR and SelaVPR. (a) presents the local matching between images from the same place. The more matches means the better performance. (b) presents local matching between images from different places. The fewer matches the better.

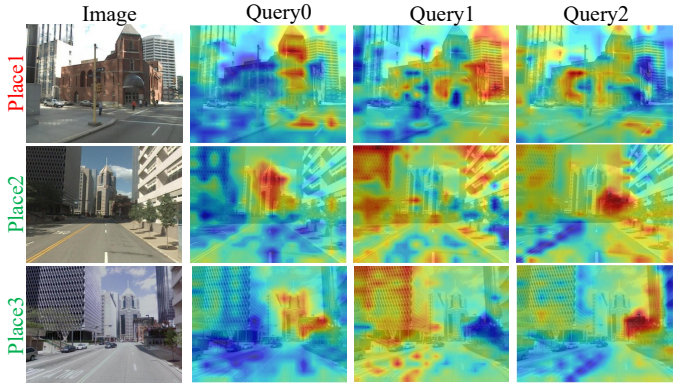

Fig. 6. Visualization of the heatmap. For clarity, the number of queries is set to 3. The place 2 and place 3 are from the same place while place 1 and place 2 are from different places.

higher number of correspondences between images captured at the same location. Notably, it establishes a greater proportion of correct matches, particularly by accurately identifying point correspondences along prominent landmarks, such as the highway sign in this case. In contrast, when processing images from different locations, CQVPR effectively reduces the number of matches, demonstrating its robustness and discriminativeness.

2) Visualization of the heatmaps: Fig. 6 illustrates the visualization of the heatmap $\mathbf{H}$, highlighting the latent semantic regions that each query individually focuses on. It can be found that each query not only corresponds to the landmark alone, but takes the surrounding environment into account. This is different from the segment-level methods. Queries of images from the same place focus on similar regions, as both describe the same scene. Conversely, heatmaps of images from different places show no overlap, as these images correspond to entirely different places.

TABLE IV
ABLATION ON THE EFFECTIVENESS OF EACH CONTRIBUTION.

DINOv2	fine-tuned DINOv2	context-aware features	query matching loss	Pitts30k R@1	R@5	R@10	R@20
✓				87.8	93.8	96.3	97.6
	✓			92.8	96.8	97.7	98.4
		✓	✓	81.2	92.0	93.9	95.3
	✓	✓		93.0	96.8	97.8	98.7
	✓	✓	✓	**93.3**	**96.9**	**98.1**	**98.9**

TABLE V
ABLATION ON THE CHOICE OF CONTEXT-AWARE FEATURES.

context-aware feature	Pitts30k R@1	R@5	R@10	R@20
$\mathbf{F}_{\mathbf{T}}^{*}$	92.0	96.7	97.7	98.6
$\mathbf{F}_{\mathbf{T}}$	**93.3**	**96.9**	**98.1**	**98.9**

E. Abaltion Study

In this section, a series of ablation experiments are conducted to better understand CQVPR. Experiments are conducted under the same training and evaluation protocol as in section IV-C.

1) Ablation on the effectiveness of each contribution: In this section, an ablation study is conducted on Pitts30k to verify the effectiveness of each of our contributions, as shown in Table IV. The third row of Table IV presents results of only using context-aware features to do VPR. The single pixel-level or context-aware feature can not achieve the optimal performance. These results demonstrate the complementary nature of our contributions.

2) Ablation on the choice of context-aware features: Besides processing the heatmap $\mathbf{H}$ to be the context-aware feature $\mathbf{F}_{\mathbf{T}}$, in this section, we try another way to obtain the context-aware feature

$$\mathbf{F}_{\mathbf{T}}^{*} = \text{softmax}(\mathbf{H}) \cdot \mathbf{T}, \tag{9}$$

where $\mathbf{F}_{\mathbf{T}}^{*}$ directly aggregates each query's embedding with a weighted summation. As shown in Table V, employing $\mathbf{F}_{\mathbf{T}}^{*}$ results in worse performance. We attribute this to the fact that directly aggregating the embedding of each query through the weighted summation may reduce the robustness of the features.

3) Ablation on the number of queries: In this section, we analyze the effect of the number of queries. As shown in Table VI, setting the number of queries to 10 achieves the best performance. The reason for this phenomenon may be that the number of queries should be consistent with the number of landmarks. Since existing VPR datasets are mainly about urban scenes, an excessive or insufficient number of queries could degrade the performance.

4) Efficiency analysis: In this section, we analyze the efficiency of each component for feature extraction in CQVPR. Since the local matching process is only leveraged during inference and follows a standard procedure, it is not included in the comparison. Notably, the GeM pooling operation is excluded from the analysis as its parameters and runtime

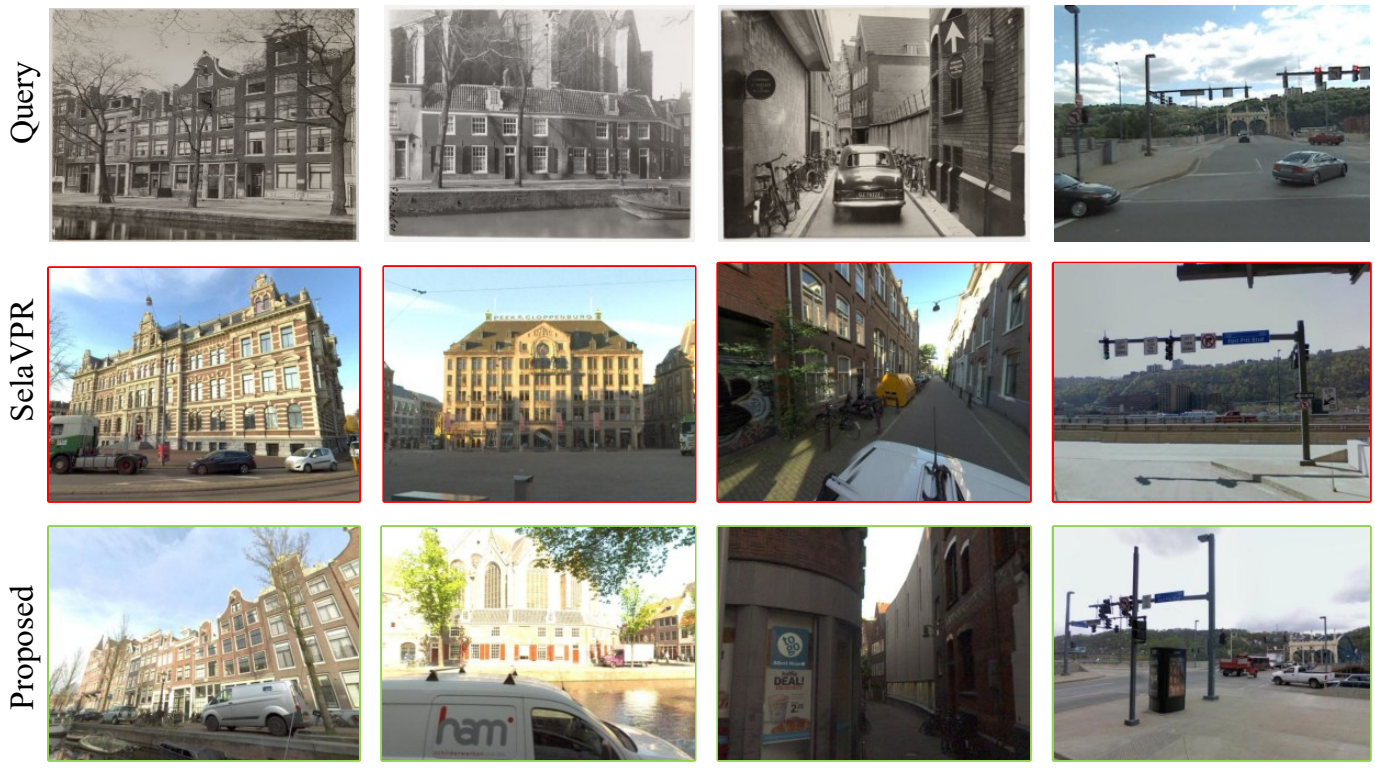

Fig. 7. Qualitative results of SelaVPR and the proposed CQVPR are presented. The Red line denotes the false prediction while the green line stands for the right.

TABLE VI
ABLATION ON THE NUMBER OF QUERIES.

The number of queries	Pitts30k R@1	R@5	R@10	R@20
5	92.3	96.6	97.8	98.8
10	**93.3**	**96.9**	**98.1**	**98.9**
20	92.6	96.5	97.8	98.6

TABLE VII
EFFICIENCY ANALYSIS OF EACH COMPONENT FOR FEATURE EXTRACTION IN CQVPR.

Module	Params (M)	Runtime (ms)
fine-tuned DINOv2	354.77	27.0
context-aware feature	0.656	0.9
up-convolution	3.24	1.0

are negligible. The results in Table VII demonstrate that the fine-tuned DINOv2 module constitutes the majority of the computational cost. Notably, when referred to Table IV, the context-aware feature extraction module not only achieves high accuracy independently but also is highly efficient compared to the fine-tuned DINOv2.

V. CONCLUSION

In this work, we propose the Contextual Query VPR (CQVPR), which integrates contextual information with detailed pixel-level features. By introducing learnable contextual queries, our method effectively captures high-level contextual information about landmarks and their surrounding environments. Furthermore, we propose a query matching loss to supervise the context extraction process, ensuring robust and accurate context modeling. Extensive experiments conducted on multiple datasets demonstrate CQVPR's superior performance compared to SOTA methods.

ACKNOWLEDGEMENTS

This work is supported by JSP KAKENHI Grant Number 23K28164.

REFERENCES

[1] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2570–2579.
[2] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5386–5397.
[3] M. Xu, N. Snderhauf, and M. Milford, "Probabilistic visual place recognition for hierarchical localization," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 311–318, 2021.
[4] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4015–4022, 2018.
[5] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, 2019.
[6] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt, "Scalable 6-dof localization on mobile devices," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 268–283.

[7] N. Pion, M. Humenberger, G. Csurka, Y. Cabon, and T. Sattler, "Benchmarking image retrieval for visual localization," in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 483–494.

[8] S. Lowry, N. Sünderhauf, P. Newman, J. Leonard, D. Cox, P. Corke, and M. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[9] R. Wang, Y. Shen, W. Zuo, and et al., "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13638–13647.

[10] F. Lu, L. Zhang, X. Lan, and et al., "Towards seamless adaptation of pre-trained models for visual place recognition," in *Proc. Int. Conf. Learn. Represent.*, 2024.

[11] K. Garg, S. Puligilla, S. Kolathaya, and et al, "Revisit anything: Visual place recognition via image segment retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2024.

[12] K. Giang, S. Song, and S. Jo, "Topicfm: Robust and interpretable topic-assisted feature matching," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, pp. 2447–2455.

[13] F. Lu, X. Lan, L. Zhang, and et al., "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[14] F. Lu, L. Zhang, S. Dong, and et al., "Aanet: Aggregation and alignment network with semi-hard positive sample mining for hierarchical place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 11771–11778.

[15] S. Zhu, L. Yang, C. Chen, and et al., "R2former: Unified retrieval and reranking transformer for place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19370–19380.

[16] M. Oquab, T. Darcet, T. Moutakanni, and et. al, "Dinov2: Learning robust visual features without supervision," in *Transactions on Machine Learning Research*, 2023.

[17] N. Keetha, A. Mishra, J. Karhade, and et. al, "Anyloc: Towards universal visual place recognition," *IEEE Rob. Autom. Lett.*, pp. 1286–1293, 2024.

[18] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[19] A.en 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.

[20] R. Arandjelovic, P. Gronat, A. Torii, and et al., "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[21] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14136–14147.

[22] A. Ali-bey, B. Chaib-draa, and P. Giguère, "BoQ: A place is worth a bag of learnable queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17794–17803.

[23] S. Garg, N. Suenderhauf, and M. Milford, "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," *Robotics: Science and Systems XIV*, 2018.

[24] A. Kirillov, E. Mintun, N. Ravi, and et al., "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[25] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2019.

[26] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-Free Local Feature Matching with Transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8918–8927.

[27] Y. Ge, H. Wang, F. Zhu, and et al., "Self-supervising fine-grained region similarities for large-scale image localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[28] Gabriele B., Carlo M., and Barbara C., "Rethinking visual geo-localization for large-scale applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4878–4888.

[29] Ya. Shen, S. Zhou, J. Fu, and et al., "Structvpr: Distill structural knowledge with weighting samples for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11217–11226.

[30] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23487–23496.

[31] A. Ali-Bey, B. Chaib-Draa, and P. Giguére, "Mixvpr: Feature mixing for visual place recognition," in *Proc. Winter Conf. on Appli. of Comput. Vis.*, 2023, pp. 2997–3006.

[32] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 11046–11056.

[33] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, pp. 194–203, 2022.