2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Semantic-focused Patch Tokenizer with Multi-branch Mixer for Visual Place Recognition

Zhenyu Xu[1], Ziliang Ren[2], Qieshi Zhang[3,*], Jie Lou[4], Dacheng Tao[5],and Jun Cheng[3]

***Abstract*— Visual Place Recognition (VPR) is critical for navigation and loop closure in autonomous driving tasks, mitigating the impact of shift errors caused by dynamic changes in the environment. Due to the limited ability of backbone networks and extreme environmental changes, current methods fail to capture foundational semantic details that include the distinctive attributes for unique place identification. To address this problem, we propose a new visual token-guided VPR framework that contains a semantic-focused patch tokenizer and a multi-branch Mixer. To mitigate the inference from place-unrelated objects, the semantic-focused patch tokenizer exploits attention-based channel selection and spatial partition, which efficiently captures important semantic information within the channels and preserve spatial relationships among the backbone features. To extract abstract features with spatial structure information, the multi-branch Mixer utilizes a multi-branch structure to aggregate local and global position information, improving the robustness of global representations to environmental changes. Experimental results demonstrate that our method outperforms state-of-the-art methods, achieving 85.3% Recall@1 on the MSLS_val dataset and 59.1% Recall@1 on the Nordland dataset when using ResNet18 as the backbone.**

I. Introduction

Visual Place Recognition (VPR) can find the visited places solely by visual (image) information and update the position information after the mapping process, which is essential in robotics control, visual navigation, and virtual reality tasks. [1]. Similar to image retrieval, a VPR system converts a query image into an image representation, and matches it among reference images in a database to find the best match. Based on the geographic labels associated with the best match, the system determines the current position and realizes accurate re-localization [2]. However, the challenges of this task are appearance changes (illumination, weather, season) and viewpoint changes (camera movement), which change the texture and color of place-related objects [3], [4].

To address these challenges, researchers aim to design distinctive image representations to mitigate the impact of appearance and viewpoint changes [6], [7], [8]. In light of the merits of abstract patterns of Convolutional Neural Networks (CNNs), VPR techniques adopt CNNs as backbones, utilize CNN feature maps as image representations, and make remarkable progress in handling appearance changes [9]. Some designed descriptors based on the regions of interest in the feature maps, integrating high activate feature map by pooling operation [10], [11]. However, interests of origins in the feature map usually focus on place-unrelated objects, which is shown in Fig. 1. In this context, we consider that place-unrelated objects are generic element to the scene, which are unable to identify a particular place. Besides the backbone network, post-processing, and region descriptors, popular VPR methods employed powerful aggregation layers such as NetVLAD [12], Patch-NetVLAD [13], TransVPR [14],

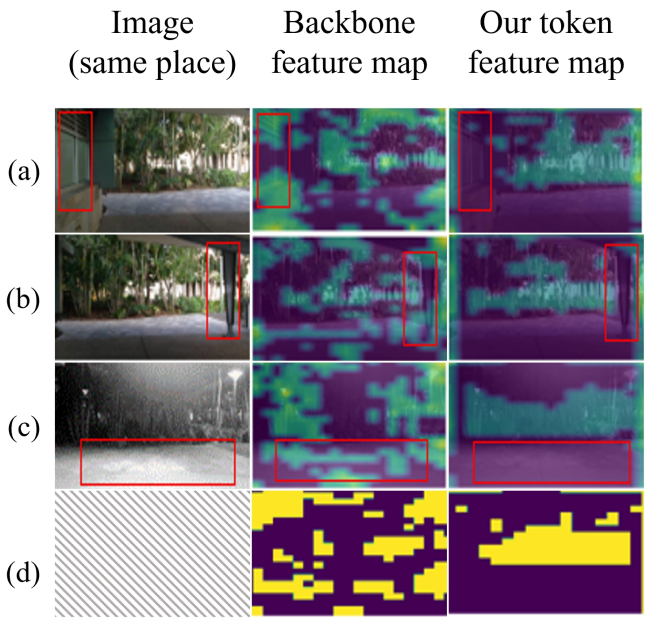


Fig. 1: Visualization of backbone feature and our visual token. (a), (b) and (c) consist of images with different appearance and viewpoint changes in the same place, along with the feature maps from the backbone (ResNet18 [5]) and our token. Red boxes are used to indicate place-unrelated objects. It can be seen that backbone features map respond to the region within red boxes, while our token features map do not react. (d) displays the intersection of the feature maps, showing that our token captures place-related semantic information that remains consistent under illumination and viewpoint changes.

* Qieshi Zhang is corresponding author, qs.zhang@siat.ac.cn
[1]Zhenyu Xu is with CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Zhenyu Xu is also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.
[2]Ziliang Ren is with the School of Computer Science and Technology, Dongguan University of Technology, Dongguan, China.
[3]Qieshi Zhang and Jun Cheng are with CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. They are also with the Chinese University of Hong Kong, Hong Kong, China.
[4]Jie Lou is with China Nuclear Power Operations Co., Ltd., Shenzhen, China.
[5]Dacheng Tao is with Nanyang Technological University, Singapore.

CosPlace [15] and Mixer [16]. These methods effectively aggregated both local and global information, achieving promising recognition results. However, these models come at the cost of multiple layer stacking, and we consider that this kind of stacking structure leads to the loss of fine-grained details during forward propagation, which can compromise recognition performance.

To address the issue of inaccurate feature maps and the coarse-grained features caused by layer stacking, we propose a novel token-guided VPR framework. To find accurate place-related objects in the backbone feature map, we introduce a semantic-focused patch tokenizer. Specifically, it involves an initial channel selection using a channel attention mechanism on the backbone features, followed by spatial partitioning within the selected channels using a self-attention mechanism. After applying our tokenizer, place-related semantic information can be extracted. To obtain fine-grained features and preserve local spatial information, instead of multi-layer stacking, we introduce a multi-branch Mixer, including multi-head attention, coordinate attention and Mixer. In detail, multi-branch Mixer consists of local and global enhancement. In the local enhancement, multiple branches with coordinate attention are utilized to aggregate local spatial information. In the global enhancement, inspired by Mixer [16], two Multi-layer Perceptron (MLP) layers are employed to associate information between positions within a single channel and across multiple channels at the same position. This processing produces diverse feature combinations, enhancing recall performance and generalization ability. Our contributions are as follows:

- We proposed a semantic-focused patch tokenizer that combines channel selection and spatial partitioning, providing precise place-related semantic information for downstream tasks.
- We proposed a multi-branch Mixer that preserves local spatial information from both local and global perspectives, improving the robustness of image representation to appearance and viewpoint changes.
- We conducted experiments on several public datasets and real-world environment for evaluation, and ablation experiments to demonstrate the effectiveness of our method.

II. Related Work

A. Traditional Features

Initially, in the realm of visual feature extraction, conventional handcrafted descriptors such as Scale-Invariant Feature Transform (SIFT) [17], Speeded-Up Robust Features (SURF) [18] and Histogram of Oriented Gradients (HOG) [19] were extensively employed to capture distinctive local features within images. These descriptors offered a level of robustness against diverse viewpoint changes in various computer vision tasks. To integrate these local features into global image representations, methods like the Visual Bag of Words (VBoW) [20], Vector of Locally Aggregated Descriptors (VLAD) [21] and Fisher Vectors (FV) [22] were introduced. These techniques aimed at encoding and capturing the spatial distribution of local features, thus creating a more comprehensive representation for image recognition tasks. These traditional methods were effective at handling viewpoint changes, However, they fail when confronted with extreme changes in appearance.

B. CNN Features

To find effective global image representations, CNN features have been harnessed due to their intrinsic discriminative property [10], [11], [23]. Some researchers focused on training CNNs specifically for place recognition [9], and further employed post-processing techniques like Principal Component Analysis (PCA) dimension reduction [24] and L_2 normalization on the feature maps [25], improving the representation robustness to appearance changes. Some methods use pooling operations with CNN features like Regional Maximum Activation of Convolutions (R-MAC) [26] and Generalized Mean (GeM) [27]. The simplicity and effectiveness of these operations have been evident in practice, and their application has yielded favorable results. However, relying solely on post-processing and pooling operations has limitations in extracting highly informative and discriminative features, due to the loss of spatial information and fine-grained details. This has led researchers to explore more robust feature extraction methods.

C. Advanced CNN Features

Aggregation layer, attention mechanisms and multi-layer stacking architectures have found widespread application in VPR tasks [28], [29], [30], [14], such as NetVLAD [12], Patch-NetVLAD [13], CosPlace [15], and MixVPR [16]. By leveraging the interplay between context, local and global information, they yield impressive recognition outcomes. However, these methods directly utilize the feature maps from the backbone, which can be influenced by place-unrelated objects. Moreover, multi-layer stacking might potentially undermine the fine-grained details.

In this paper, we leverage a tokenizer to precisely extract CNN feature maps, which aims to alleviate the impact of place-unrelated objects. Additionally, we use a multi-branch structure to enhance spatial and fine-grained details, minimizing spatial information loss during feature abstraction.

III. Method

Our framework is shown in Fig. 2, including a semantic-focused patch tokenizer and multi-branch Mixer. In the semantic-focused patch tokenizer, we assume that not all channels equally contribute to the downstream task. Thus, an Efficient Channel Attention (ECA) is introduced, which captures place-related semantic information by effectively establishing extensive relationships among channels. In the multi-branch Mixer, we enhance tokens from local and global perspectives. Locally, a branch structure denoted as *Branch* is introduced to preserve spatial structure within channels. We then expand this branch, creating a multi-branch architecture that extracts meaningful features L. Globally, two simple

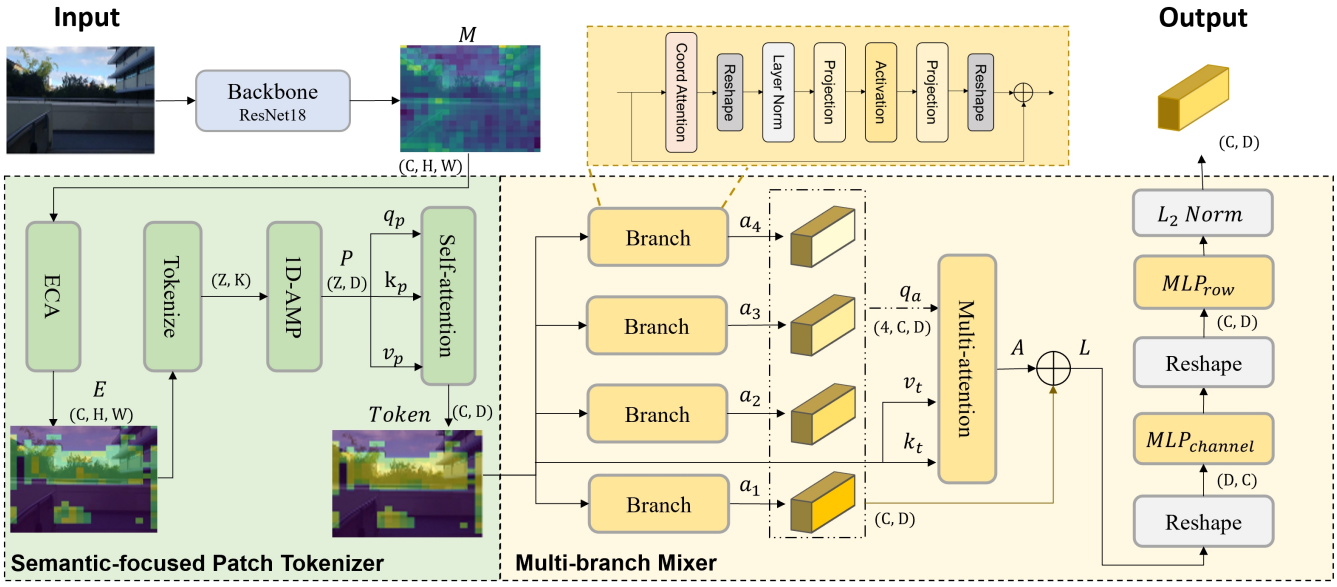


Fig. 2: Our token-guided VPR framework. Given an input image, ResNet18 is employed as the backbone network to extract features. Subsequently, these features are fed into our semantic-focused patch tokenizer, which captures semantic features. Following this, the multi-branch Mixer enhances both local and global semantic information by using multiple branches and multi-head attention structures. Finally, the output global feature is employed as a descriptor for image matching.

MLPs, $P_{channel}$ and P_{row}, are used to capture global position information across multiple channels.

A. Semantic-focused Patch Tokenizer

Given the backbone feature map $M \in R^{C\times W\times H}$, C represents the number of feature channels, and W and H denote the width and height of the feature map, respectively. The enhanced feature map $E \in R^{C\times H\times W}$ is obtained by applying the ECA to the M.

$$E = ECA(M). \tag{1}$$

Then, E is transformed into a patch representation from (C,W,H) to (Z,K), where Z and K are defined as $Z = C\times P_l^2$ and $K = \frac{W\cdot H}{P_l^2}$, respectively. P_l is the length of the patches. To enhance computational efficiency, a 1-Dimensional Adaptive Max Pooling (1D-AMP) layer is employed, effectively reducing the feature dimension. This results in a segmented patch representation denoted as $P \in R^{Z\times D}$, described by the following equation:

$$P = 1D-AMP(Tokenize(E)), \tag{2}$$

where the D is the scale factor of 1D-AMP.

Next, patch representation P is mapped to queries ($q_p \in R^{Z\times D}$), keys ($k_p \in R^{Z\times D}$), and values ($v_p \in R^{Z\times D}$), respectively. Then, The self-attention matrix $S(q_p,k_p) \in R^{D\times D}$ is computed between different regions within the token features as follows:

$$S(q_p,k_p) = SoftMax\left(\frac{q_p k_p^T}{\sqrt{d}}\right). \tag{3}$$

Here, $SoftMax$ denotes the softmax function that normalizes the similarity scores between queries and keys. The scaling factor $\sqrt{d}$ is used to control the magnitude of the dot product and stabilize the learning process.

Finally, $Token \in R^{Z\times D}$ is created by computing the dot product between the values v_p and $S(q_p,k_p)$, followed by element-wise addition of these enhanced values to the original token features v_p.

$$Token = S(q_p,k_p)\cdot v_p + v_p. \tag{4}$$

B. Multi-branch Mixer

1) Local Enhancement: First, a coordinate attention layer [35] is applied to the $Token$ to generate an attention feature map $C_a \in R^{Z\times K}$, which capture local spatial information. Then, C_a is processed by a two-layer MLP MLP_{cor} to produce abstract features. For brevity, we omit details of feature shape transformations. Finally, the $Branch(\cdot)$ is formulated by element-wise addition of the MLP output and the $Token$.

$$C_a = Coord_{att}(Token), \tag{5}$$

$$MLP_{cor}(C_a) = W_2(ReLU(W_1\cdot Norm(C_a))), \tag{6}$$

$$Branch(Token) = MLP_{cor}(C_a) + Token. \tag{7}$$

To improve the diversity of features, a multi-branch architecture is adopted, followed by a multi-head attention module MA to identify relevant attention features. For each branch indexed as j, the respective feature is computed as $a_j = Branch_j(Token)$. Then, $[a_1,a_2,\ldots,a_j]$, $Token$, and $Token$ are mapped as q_a, k_t, v_t respectively, and the multi-head attention $MA(\cdot)$ is applied to obtain the output A:

$$A = MA(q_a,k_t,v_t). \tag{8}$$

TABLE I: Recall performance comparison with existing methods on the public benchmarks.

Method	Backbone	MSLS_val [31]			Nordland [32]			Pitts_30k [12]		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ConAP [33]	ResNet18	75.8	83.2	86.2	30.0	45.7	52.9	89.3	94.8	96.3
Gem [27]	ResNet18	65.3	78.1	81.2	11.6	20.8	25.8	75.1	88.5	91.8
CosPlace [15]	ResNet18	78.0	86.2	89.2	32.4	49.5	56.8	87.7	93.8	95.4
MixVPR [16]	ResNet18	83.6	88.9	91.3	47.8	66.1	73.0	89.8	94.5	96.2
Patch-NetVLAD† [13]	VGG16	79.5	86.2	87.7	44.9	50.2	52.2	88.7	94.5	95.9
SuperGlue* [30]	Superpoint [34]	78.0	82.0	84.3	29.1	33.4	35.0	87.2	94.8	96.4
DELG* [29]	ResNet50	83.2	90.0	91.0	51.3	66.8	69.8	**90.0**	95.0	**96.7**
NetVLAD† [12]	VGG16	60.8	74.3	80.0	10.4	16.3	19.7	83.5	91.3	94.0
Ours	**ResNet18**	**85.3**	**90.5**	**92.7**	**59.1**	**75.2**	**80.8**	**90.0**	**95.8**	96.2

Note: Results of models labeled with '†' were evaluated using pre-trained models in local testing. Results of models marked with '*' are evaluated from the proposed paper. The remaining results were obtained through local training and testing.

In this case, we utilize different branches as queries to generate appropriate weights for v_t, producing more diverse token combinations.

To ensure the spatial information, the aggregated output A is applied with residual structure, resulting in the local representation L:

$$L = A + Branch_1(x). \tag{9}$$

This enhancement improves the discriminative attributes of local features by leveraging the complementary information gathered from multiple branches and attention mechanisms.

2) Global Enhancement: The L is transposed to L^T, then projected along the channel and row dimensions using learnable operations $MLP_{channel}$ and MLP_{row}, respectively. The resulting tensor undergoes L_2 normalization along the row dimension, generating the final output $Output$.

$$L_c = MLP_{channel}(L^T), \tag{10}$$

$$Output = L_2Norm(MLP_{row}(L_c^T)). \tag{11}$$

IV. Experiments

A. Implementation

Our method and comparisons were implemented by using PyTorch [36]. The input images were cropped to 320 $\times$ 320 and augmented with RandAugment [37] to expand the training data. For the neural network backbone, we used a ResNet18 [5] pre-trained on ImageNet [38]. To leverage the pre-trained weights while adapting to our target domain, we froze the first and second layers during fine-tuning. For the tokenizer, we grid-searched suitable token sizes and scale factor of the adaptive max pooling layer, which are set to 4 and 256, respectively. As for the multi-branch Mixer, we used 4 branches in all experiments, with related ablation studies on the branch number presented later.

During the training phase, the batch size is set to 80, and each batch is included a minimum of 4 images per place (class). We choose the Stochastic Gradient Descent (SGD) [39] optimizer with a learning rate of 0.05, a weight decay of 0.001, and a momentum of 0.9. Our loss function was the multi-similarity loss [40], which is widely used for visual place recognition tasks. Our computational resources comprised an NVIDIA Tesla K80 GPU. The training process was carried out for up to 80 epochs to ensure convergence.

B. Dataset and Evaluation Metrics

Our method was trained on the GSV-Cities dataset [33] and evaluated on the Pitts30k-test [12], Mapillary Street-Level Sequences (MSLS) [31], and Nordland datasets [41], [32]. These datasets include challenging places with diverse variations in appearance and viewpoint, including different times of day, seasonal changes, and varying camera angles. The Pitts30k-test dataset [12] comprises 8k query-reference pairs collected from Google Street View. The MSLS dataset [31] offers a comprehensive evaluation platform encompassing a wide range of geographic, viewpoint, and seasonal variations. The Nordland dataset presents an demanding benchmark, featuring scenes that span from snowy winters to sunny summers, exhibiting extreme appearance changes.

For evaluation, we adopt the Recall@N (R@N) metric, a widely-used measure for comparing VPR techniques [42], [33], [16]. In this context, a successful retrieval is achieved when a query image is correctly matched to its corresponding image within the top-k retrieved reference images.

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

Recall performance comparisons with existing methods are shown in Table I. Various methods employ different network backbone architectures, including ResNet18, ResNet50, and others. Across the MSLS_val [31], Nordland [41], [32], and Pitts_30k [12] datasets, our method achieves the highest top 1 recall. Notably, our framework relies solely on ResNet18, a relatively lightweight backbone, yet it competes favorably with or even outperforms state-of-the-art place recognition methods that use more complex backbones like Patch-NetVLAD (VGG16) and DELG (ResNet50). Furthermore, when evaluated on the Nordland dataset, our methods exhibit a noteworthy enhancement in recognition accuracy, surpassing other methods by nearly 10%. These results show the effectiveness and generalization capability of our methods.

D. Ablation Studies

1) Branch: The Ablation results of different branch configurations on parameters, latency and recall performance on MSLS_val dataset is shown in Table II. The R@1 accuracy exhibits noticeable trends with varying branch counts. As the number of branches increases, there's a shift in the trade-off between model complexity, latency, and recall performance.

TABLE II: Ablation studies of different branches B.

Branch (B)	parameter (M)	Latency (ms)	MSLS_val [31] R@1	R@5	R@10
1	0.701	6	80.6	90.2	91.7
2	1.132	9.2	82.5	89.3	90.8
4	**1.994**	**15.5**	**85.2**	**90.5**	**92.7**
8	3.717	27	80.4	88.2	90.4

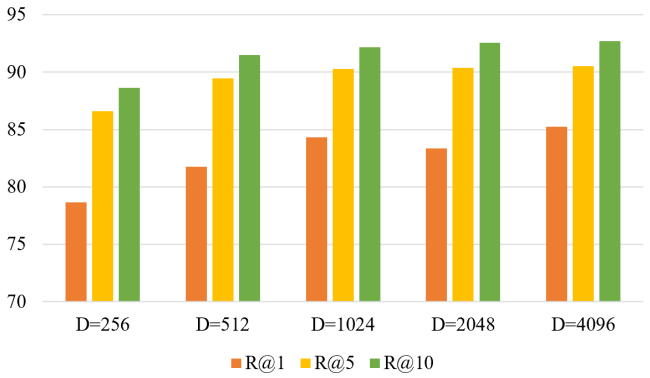


Fig. 3: Recall performance of different dimension in our method.

TABLE III: Ablation studies of different token size P_l.

Token size (P_l)	parameter (M)	Latency (ms)	Pitts_30k [12] R@1	R@5	R@10
1	1.994	15.5	90.0	95.1	96.2
2	0.476	13	89.1	94.8	95.9
4	**0.322**	**12.5**	**89.2**	**94.8**	**95.9**

Specifically, the model with 4 branches achieves a balance, yielding an R@1 accuracy of 85.2% while maintaining acceptable latency. These results highlight the significance of selecting an optimal branch configuration to strike a practical equilibrium between accuracy and efficiency.

2) Descriptor Dimension: The recall performance of our method with different dimensions on the MSLS_val dataset is presented in Fig. 3. A general trend of improved performance is observed as the dimension increases. Specifically, higher dimensions, such as D=2048 and D=4096, result in higher recall rates compared to lower dimensions like D=256. This observation shows that enhancing recognition accuracy is achieved by increasing the dimension of the feature representations.

3) Token Size: The parameter size, latency, and recall rate on the pitts_30k dataset for different Token sizes are presented in Table III. As Token size increases, there is a significant reduction in parameter size (from 1.994M to 0.322M) and latency (from 15.5 ms to 12.5 ms). Remarkably, the R@1 accuracy remains competitive across token sizes, ranging from 89.2% to 90.0% on the pitts_30k dataset. This outcome verifies the efficacy of our token-guided method in preserving the positions of place-related objects, maintaining performance in coarse granularity, and simultaneously reducing memory storage and computational resources.

4) Components: To validate the effectiveness of both our tokenizer and multi-branch Mixer, ablation experiments

TABLE IV: Ablation studies of different components on MSLS dataset.

Component Backbone (ResNet18)	Tokenizer	Multi-branch Mixer	MSLS_val R@1	R@5	R@10
✓			39.4	47.8	50.2
✓	✓		40.9	48.2	52.9
✓		✓	83.2	89.4	92.0
✓	✓	✓	**85.2**	**90.5**	**92.7**

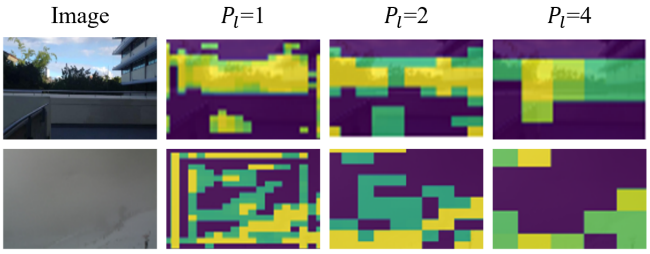


Fig. 4: Visualization of Token feature map with varying token size. We divide the feature values into three intervals, with yellow representing the highest response, followed by green, and finally purple. Encountering significant visual variations, our token can capture semantic information that remains invariant across scenes.

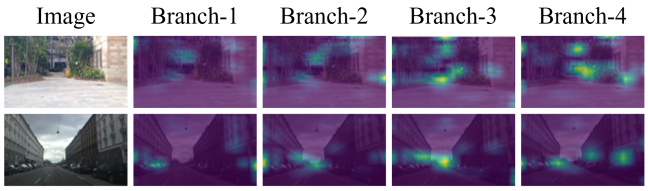
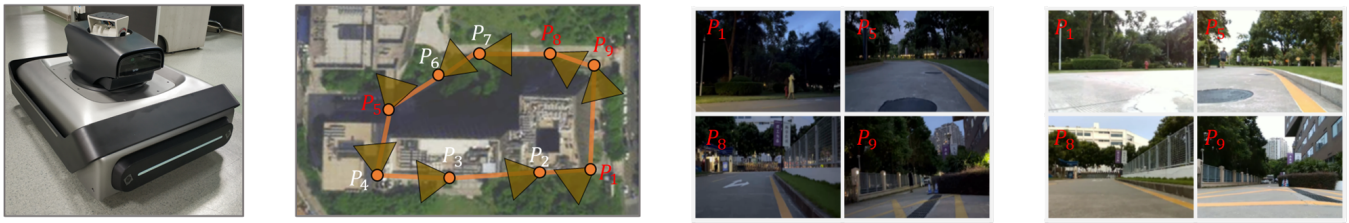


Fig. 5: Visualization of Multi-branch Mixer attention feature map. It can be observed that each branch extracts distinct scene cues while avoiding low-texture regions like the sky and the road, reducing noise interference.

are shown in Table IV. The backbone exhibits a baseline performance, while the tokenizer demonstrates a modest enhancement. However, the important advancement arises from the combination between of tokenizer and multi-branch Mixer, yielding substantial recall improvements: R@1 increases from 39.4% to 85.2%, R@5 rises from 47.8% to 90.5%, and R@10 climbs from 50.2% to 92.7%. This powerful combination demonstrates the crucial role played by these components in enhancing recognition performance on the MSLS_val dataset.

E. Visualization

To validate the semantic attention of *Token* across different places, the *Token* feature maps are visualized in Fig. 4. With the increase in token size, an evident improvement in feature map granularity is observed, while the focus on place-related objects remains unchanged. This enhancement leads to a reduction in the impact of noise interference. Furthermore, superior proficiency in capturing advanced semantic attributes of objects is demonstrated by larger Token



(a) Our mobile robot (b) Trajectory in our campus (c) Query examples (d) Reference examples

Fig. 6: Deploying VPR method on mobile robot validated within the our campus. (a) Our mobile robot, equipped with a monocular camera at its top, captures images. (b) The trajectory (orange line), key frames (dots), and corresponding camera angles (triangle regions) recorded in the robot's loop closure. Each place is denoted as P_i. (c) and (d) are images taken at the same place but at different appearance changes (illumination and weather) and viewpoint changes, serving as query and reference images, respectively.

sizes, thereby contributing to the enhancement of image recognition accuracy and localization precision. Notably, in sparse scenes like snowy landscapes, P_l=4 effectively determines object positions.

To validate the multiple branches to capture meaningful spatial cues, the feature maps obtained from the multi-branch Mixer are shown in Fig. 5. The visualization illustrates that, within the context of the same place, position relationships are effectively preserved by each branch. Notably, a remarkable consistency is observed in the high-response regions of the feature maps across different branches, indicating the robustness of our method in capturing fundamental place attributes. Furthermore, distinct points are captured by each branch, thereby showing the capability to extract diverse and informative features. This observation is further substantiated by the varying degrees of feature map responses, thereby providing empirical evidence for the efficacy of our method.

F. Deployment

In real-world environments, challenges like lens blur, dynamic object occlusion and fluctuating lighting conditions serve as rigorous tests to validate the effectiveness of VPR methods. Therefore, we deployed VPR methods on a robotic platform to conduct loop closure detection within our campus, as shown in Fig. 6. To ensure images spatial consistency captured within a short time frame, we introduce a parameter denoted as T. This parameter refers to a threshold that allows for a certain frame tolerance, since consecutive frames within a short time span correspond to the same place. Assuming a query image Q_i, where i denotes the ith frame, VPR methods successfully retrieve frames from the reference images within the range of $[i-T, i+T]$.

The recall performance of our VPR method with varying T values is summarized in Table V. Compared with other methods, our method exhibits a notable 5% improvement in both R@5 and R@10, highlighting its advantage in longer time intervals between queries. While our method lags slightly behind MixVPR [16] in R@1, it stands out with the lowest latency of 12.5 ms/img, as opposed to 13-14 ms/img for other methods, showing its superior speed. Furthermore, these marginal differences become more pronounced during loop closure time. Here, loop closure time is defined as the time of feature extraction and matching processes. Notably, our method exhibits the fastest performance at 2975 ms, whereas alternative methods range from 3000 to 3300 ms. The significantly lower latency and loop closure time show the enhanced efficiency of our method.

TABLE V: Recall performance of the VPR methods with varying T in real-world environment.

Method	Latency ms/img	Loop closure ms	T = 5			T = 7		
			R@1	R@5	R@10	R@1	R@5	R@10
ConAP [33]	13.9	3308.0	32.07	52.32	76.79	42.62	63.71	84.39
Gem [27]	13.1	3118.8	34.18	65.82	81.43	46.41	73.42	87.76
CosPlace [15]	13.0	3094.0	30.8	57.81	80.59	41.35	64.98	87.34
MixVPR [16]	14.2	3379.6	**37.55**	57.81	80.17	**49.37**	67.93	86.92
Ours	**12.5**	**2975.0**	35.02	**62.03**	**83.12**	48.1	**70.89**	**90.3**

V. Conclusion

In this paper, we proposed a new visual token-based method to capture unique place identification for VPR. First, we designed a semantic-focused patch tokenizer, which efficiently captures place-related semantic information. Then, we introduced a multi-branch Mixer to improve the robustness of the global representation to environmental changes. Based on the experiments, our method demonstrates the desired recall performance with smaller backbones. Moreover, it maintains performance while reducing parameters, showing the effectiveness and efficiency of the proposed method. The visualization and real-world deployment further show the effectiveness of the visual token and multi-branch architecture, enhancing their capability to capture unique place identification of real-world environments.

Acknowledgment

We acknowledge the supports from the National Natural Science Foundation of China (Nos. U21A20487, 62376261), Shenzhen Technology Project (Nos. JCYJ20220818101206014), Natural Science Foundation of Guangdong Province (Nos. 2022A1515140119 and 2023A1515011307), CAS Key Technology Talent Program Shenzhen Engineering Laboratory for 3D Content Generating Technologies (No. [2017]476).

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.

[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.

[3] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4416–4425, 2021.

[4] Q. Zhang, Z. Xu, Y. Kang, F. Hao, Z. Ren, and J. Cheng, "Distilled representation using patch-based local-to-global similarity strategy for visual place recognition," *Knowledge-Based Systems (KBS)*, vol. 280, p. 111015, 2023.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[6] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, 2017.

[7] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Robotics: Science and Systems (RSS)*, pp. 1–10, 2018.

[8] F. Maffra, Z. Chen, and M. Chli, "Tolerant place recognition combining 2D and 3D information for UAV navigation," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2542–2549, 2018.

[9] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304, 2015.

[10] 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.

[11] S. Garg, N. Suenderhauf, and M. Milford, "Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3645–3652, 2018.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.

[13] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14152, 2021.

[14] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "TransVPR: Transformer-based place recognition with multi-level attention aggregation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13648–13657, 2022.

[15] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4878–4888, 2022.

[16] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "MixVPR: Feature mixing for visual place recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2998–3007, 2023.

[17] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 1150–1157, 1999.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, 2005.

[20] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *International Workshop on Multimedia Information Retrieval (MIR)*, pp. 197–206, 2007.

[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311, 2010.

[22] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision (IJCA)*, vol. 105, pp. 222–245, 2013.

[23] Q. Zhang, Z. Xu, Z. Yang, Z. Ren, S. Yuan, and J. Cheng, "Enhancing visual place recognition using discrete cosine transform and difference-based descriptors," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 1–1, 2024.

[24] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[25] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2238–2245, 2015.

[26] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *International Conference on Learning Representations (ICLR)*, pp. 1–12, 2016.

[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 (TPAMI)*, vol. 41, no. 7, pp. 1655–1668, 2018.

[28] H. Wu, M. Wang, W. Zhou, Y. Hu, and H. Li, "Learning token-based representation for image retrieval," in *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2703–2711, 2022.

[29] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *European Conference on Computer Vision (ECCV)*, p. 726–743, 2020.

[30] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020.

[31] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2626–2635, 2020.

[32] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," *10th Planning, Perception and Navigation for Intelligent Vehicles (PPNIV) Workshop at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–6, 2018.

[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.

[34] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 224–236, 2018.

[35] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, 2021.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *International Conference on Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[37] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 702–703, 2020.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2017.

[39] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[40] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022–5030, 2019.

[41] S. Skrede, "Nordlandsbanen: minute by minute, season by season," 2013. https://bit.ly/2QVBOym.

[42] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision (IJCV)*, vol. 129, no. 7, pp. 2136–2174, 2021.