2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

PAGTM: Position and Attention-Guided Token Merging for Efficient Visual Place Recognition

Hongchan Cho[1], Youngjo Lee[1], Jinwoo Jang[1], Seunghan Yu[1] and Euntai Kim[1,2,†]

***Abstract*— Recent advances in Vision Transformers (ViTs) have significantly improved the performance of Visual Place Recognition (VPR), but their high computational cost—due to the quadratic complexity of self-attention—limits their practical deployment in real-world scenarios. To address this challenge, we propose PAGTM (Positional- and Attention-Guided Token Merging), a training-free token reduction framework designed specifically for ViT-based VPR models. In VPR, preserving the spatial layout of a scene (e.g. road alignment, building structures) and focusing on semantically meaningful regions are both critical for robust matching under viewpoint and appearance variations. However, existing token reduction methods often overlook these aspects, leading to degraded recognition performance. To address this, PAGTM incorporates two key cues. The first is *positional proximity*, which merges spatially adjacent tokens to maintain the scene's structural layout. The second is *attention-based token protection*, which retains tokens that receive high attention because they represent regions important for distinguishing places, such as signs or distinctive structures. Without requiring any fine-tuning, PAGTM can be directly applied at inference time and consistently outperforms existing token reduction methods such as ToMe and ToFu across multiple ViT-based VPR models and datasets, achieving a better trade-off between computational efficiency and retrieval accuracy.**

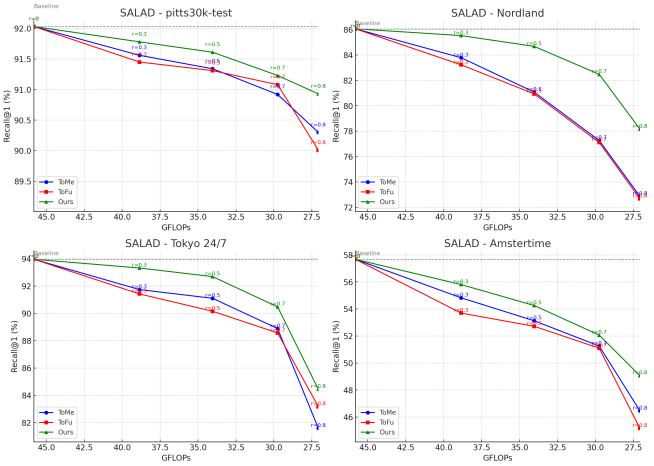

Fig. 1. Comparison of Recall@1 versus GFLOPs for different token reduction strategies on four VPR datasets using the SALAD pipeline. Each curve shows the trade-off between computational cost and recognition performance under varying token reduction rates r (e.g., Blue: ToMe, Red: ToFu, Green: Ours).

I. Introduction

Visual Place Recognition (VPR) [1], [2] is a fundamental task in robotics and autonomous driving that enables an agent to recognize previously visited locations based on visual inputs. A reliable VPR system must handle a wide range of appearance changes, such as variations in lighting, weather, seasons, and viewpoint, making it considerably more challenging than conventional image retrieval. Accurate VPR is essential for long-term navigation, loop closure detection, and topological mapping in real-world deployments.

In earlier VPR research, Convolutional Neural Networks (CNNs) [3], [4], [5], [6], [7], [8], [9], [10] were widely used to extract global or local descriptors from images. Methods such as NetVLAD [5], GeM Pooling [10] and MixVPR [4] achieved competitive performance by aggregating convolutional features for robust retrieval. More recently, Vision Transformers (ViTs) [11], [12], [13] have gained attention as a powerful alternative. By leveraging self-attention, ViTs can capture global contextual information across image regions, which is particularly beneficial for handling large viewpoint and appearance variations in VPR. ViT-based models like AnyLoc [14], SALAD [15], and VLAD-BuFF [16] have achieved state-of-the-art performance across various benchmarks, demonstrating strong generalization to real-world environments.

Despite their impressive performance, ViT-based models are computationally expensive due to the self-attention mechanism, which requires pairwise interactions among all tokens. This results in quadratic complexity with respect to the number of tokens. In VPR, high-resolution images are often needed to capture fine-grained semantic cues such as building facades, intersections, and other stable landmarks, which further increases the token count and, consequently, the memory and latency requirements. These limitations make it challenging to deploy ViT-based VPR models in resource-constrained environments or real-time robotic systems.

To mitigate the computational overhead of ViTs, recent studies have proposed token reduction strategies [17] that aim to decrease the number of tokens while preserving task-relevant information. Notable approaches include Token Merging (ToMe) [18] and Token Fusion (ToFu) [19], which reduce token count by merging or pruning tokens with similar features. While effective in standard vision tasks such as classification, these methods often fall short in VPR scenarios.

Naively reducing tokens without considering their spatial or semantic importance often leads to performance degradation in VPR. As illustrated in Figure 2, VPR

†is that the corresponding author
[1]All the authors with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea, {hongchan98, lzozo95, 0126jang, ysh, etkim}@yonsei.ac.kr
[2]Korea Institute of Science and Technology (KIST), Seoul, South Korea

models frequently retrieve candidates with globally similar layouts—such as road geometry or surrounding structures—even when the match is incorrect. However, the presence of distinctive landmarks, such as building facades, or utility poles, is often the key to identifying true matches. This highlights the importance of token reduction strategies that preserve both the structural composition of the scene and semantically informative regions, which are essential for accurate place recognition.

To this end, we propose PAGTM (Positional- and Attention-Guided Token Merging), a token reduction strategy specifically tailored for VPR. Rather than relying solely on feature similarity, PAGTM introduces two additional key cues: (1) 2D positional proximity and (2) attention-based token protection.

First, 2D positional proximity is leveraged to preserve the spatial structure of the scene, which is essential for robust location matching in VPR. PAGTM encourages merging between tokens that are spatially adjacent in the image plane, under the assumption that such tokens are more likely to belong to the same semantic region (e.g., road surface, sidewalk, or building facade). This strategy helps maintain local geometric consistency and prevents the unintended blending of unrelated regions. Since VPR often relies on the global arrangement of structural elements, preserving these spatial relationships during token reduction is key to retaining matchable representations across different viewpoints.

Second, we leverage attention-based token protection to retain semantically important features essential for place recognition. Specifically, attention scores from a ViT backbone fine-tuned for VPR are used to identify and preserve tokens that consistently receive high attention across layers. As visualized in Figure 4, these high-attention regions correspond to landmarks such as intersections, or building facades. By protecting these tokens from merging, our method ensures that discriminative visual cues are maintained even under aggressive reduction.

Also, a key advantage of PAGTM lies in its practicality: it is entirely training-free and can be directly applied to existing ViT-based VPR models without any architectural changes or fine-tuning. Operating solely at inference time, it requires only the attention maps from a pre-trained backbone. This plug-and-play design makes PAGTM well-suited for real-world robotics and autonomous systems, where computational budgets are tight and retraining is often infeasible.

We evaluate PAGTM on five standard VPR datasets, Pitts30k-Test, Nordland, Tokyo 24/7, AmsterTime, and Eynsham under three ViT-based VPR models (SALAD, Clique-Mining and VLAD-BuFF). Experimental results show that PAGTM consistently outperforms existing token reduction methods in both retrieval performance and inference speed, demonstrating a superior trade-off between accuracy and efficiency.

Our main contributions are summarized as follows:

- We propose a token merging strategy that combines feature similarity, positional proximity, and attention-based protection for Visual Place Recognition (VPR) task.
- We design a dynamic protection mechanism that uses attention maps to preserve high-importance tokens and key visual cues.
- Our merging method can be seamlessly applied at inference time, enabling plug-and-play integration into various existing Transformer-based VPR pipelines.
- Experiments on standard benchmarks demonstrate a strong trade-off between efficiency and accuracy, surpassing existing token reduction methods.

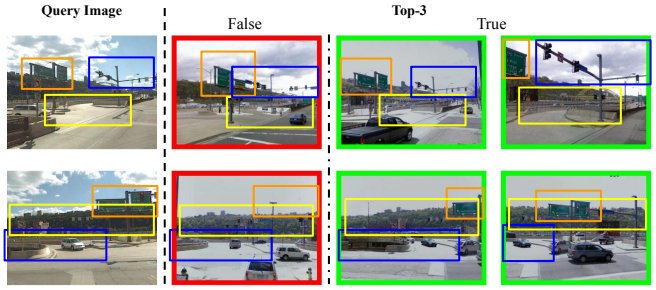

Fig. 2. Retrieval results for two query images using the SALAD baseline model. Boxes of the same color indicate structurally similar regions, while orange boxes highlight landmarks that help distinguish true positives from false matches.

II. Related Work

1) Vision Transformers: Vision Transformers (ViTs) [11], [12], [20], [21], [22] have achieved strong performance across various vision tasks by modeling pairwise relationships between image patches via self-attention. Unlike CNNs, ViTs capture global context without relying on inductive biases such as locality or translation invariance, and scale well with data and model size. However, their self-attention mechanism introduces a computational bottleneck [23], [24], [25], with quadratic time and memory complexity in the number of tokens. This limitation becomes prominent for high-resolution inputs, resulting in increased latency and GPU memory usage, and motivates efficient token reduction strategies.

2) ViT-based Models for Visual Place Recognition: Due to their strong representational power, ViTs have been widely adopted in visual place recognition (VPR) [26], [27], [28], where modeling global spatial context is critical. Works like AnyLoc [14] and SALAD [15] apply ViT-based backbones to enhance robustness against appearance and viewpoint variations. More recent methods such as Clique-Mining [29] and VLAD-BuFF [16] improve retrieval performance by combining contrastive learning with fine-tuned transformer features. These approaches leverage data augmentation and local descriptor mining to extract place-aware representations. Nonetheless, they still suffer from the inefficiency of full self-attention, limiting their scalability in real-world applications.

3) Token Reduction in Transformers: To address the computational cost of Vision Transformers, various token reduction techniques have been proposed. Token Merging

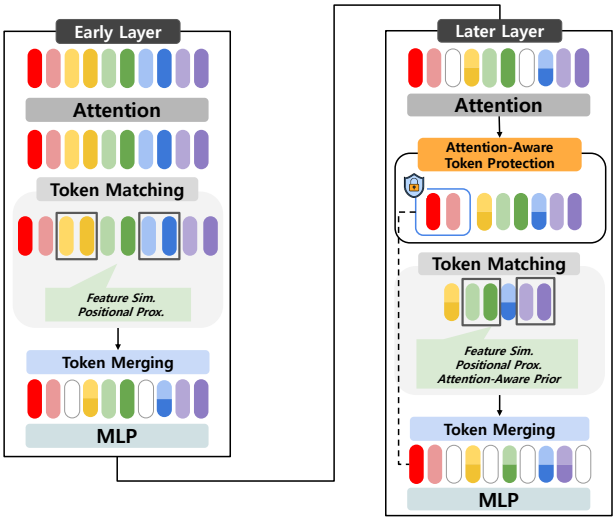

Fig. 3. PAGTM framework: Early layers use feature and positional cues for merging, while later layers add attention-aware token protection and penalties to preserve semantic tokens.

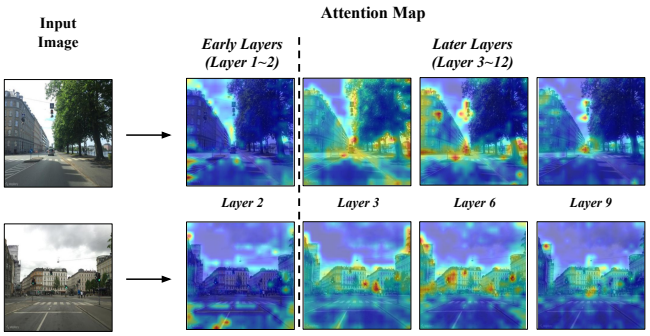

Fig. 4. Visualization of attention maps from a DINOv2 backbone fine-tuned for VPR. Early layers (1–2) show broad and weak attention, while later layers (3–12) focus more clearly on meaningful regions such as buildings and traffic lights.

(ToMe) [18] reduces redundancy by merging tokens with similar feature embeddings. Token Fusion (ToFu) [19], in contrast, adopts a hybrid strategy that combines aspects of merging and pruning: it retains a subset of important tokens while aggregating the rest into a single fused token.

While effective in classification and segmentation, these methods often overlook spatial and semantic distinctiveness essential for visual place recognition. Relying solely on feature similarity can lead to merging spatially distant or semantically critical tokens, reducing scene discriminability. Furthermore, many of other approaches [30], [31], [32] require model modification or re-training, limiting their applicability in training-free scenarios. In contrast, our method (PAGTM) combines feature similarity, positional proximity, and attention-based priority for context-aware, architecture-agnostic token reduction.

III. Method

We propose PAGTM (Positional- and Attention-Guided Token Merging), a training-free framework for efficient token reduction in ViT-based Visual Place Recognition (VPR). Unlike previous methods that rely solely on feature similarity, PAGTM introduces two key innovations: (1) leveraging attention-based token importance to dynamically protect critical tokens, and (2) incorporating positional proximity into the merging process to preserve the geometric structure of scenes. PAGTM is applied progressively across all Transformer layers and designed to be seamlessly integrated into standard ViT backbones.

A. Overall Pipeline

Given a ViT model with L layers and an initial token sequence $\mathcal{T} = \{T_1, \ldots, T_N\}$, PAGTM progressively reduces the number of tokens across layers such that the final layer operates on a compact yet informative set of tokens. At each layer, the token reduction process consists of the following steps:

1) Estimate token importance using multi-head attention maps.
2) Protect high-importance tokens from merging.
3) For the remaining tokens, compute merging priority using feature similarity, positional proximity, and attention-aware penalty.
4) Select top-scoring pairs and merge them using weighted feature averaging.

The total number of merged tokens across all layers corresponds to a predefined reduction ratio r, evenly distributed over L layers. The detailed process is illustrated in Figure 3 and in Algorithm 1.

B. Token Importance Estimation

In Visual Place Recognition, not all tokens contribute equally. Certain regions—such as street signs, architectural edges, or intersections—carry higher semantic salience and are often critical for accurate retrieval. Interestingly, we observe that ViT models fine-tuned for VPR consistently assign higher attention to such regions, reflecting their utility in distinguishing places.

Figure 4 illustrates this pattern: attention maps from different layers of a DINOv2 backbone show strong activations near meaningful structures such as building corners, traffic lights, and road junctions. This observation motivates our design choice to use attention scores as a proxy for token importance.

For each token T_i, we compute its importance score by averaging its incoming attention across all heads:

$$I_i = \frac{1}{H} \sum_{h=1}^{H} A_i^{(h)},$$

where $A_i^{(h)}$ denotes the attention weight received by token i from head h, and H is the total number of heads.

C. Dynamic Token Protection

We introduce a dynamic token protection strategy to prevent performance degradation under high token reduction. Instead of protecting a fixed number of tokens, we scale the number based on the global reduction ratio $r \in [0, 1]$:

$$k_\ell = (1 - r) \cdot k_{\max}, \tag{1}$$

where $k_{\max}$ is the maximum number of tokens allowed to be protected (typically 20% of the input tokens).

Tokens with the highest attention scores are selected and excluded from the merging pool, ensuring that critical regions are preserved intact. This adaptive strategy avoids suboptimal merging caused by an overly small candidate pool and maintains a balance between efficiency and content preservation.

D. Token Merging Priority

The remaining unprotected tokens are split into two disjoint sets: source tokens $\mathcal{T}_{src}$ and retained tokens $\mathcal{T}_{ret}$, forming a bipartite structure. For each candidate pair $(T_i, T_j) \in \mathcal{T}_{src} \times \mathcal{T}_{ret}$, we compute a merging priority score based on three factors:

a) Feature Similarity: We use the cosine similarity of key embeddings from the self-attention module:

$$S_{feat}(i, j) = K_i^\top K_j, \tag{2}$$

where $K_i, K_j \in \mathbb{R}^d$ are L2-normalized key vectors of tokens i and j. This formulation follows the common practice used in prior token merging methods [18], [19], and serves as the base similarity metric in our framework.

b) Positional Proximity: To preserve the structural layout of scenes, we incorporate 2D positional information into the merging process. Tokens that are spatially close in the image plane are more likely to belong to the same semantic region (e.g., road, sidewalk), and are thus better candidates for merging.

We define the positional proximity score as:

$$S_{pos}(i, j) = 1 - \frac{\|P_i - P_j\|_2 - d_{\min}}{d_{\max} - d_{\min} + \epsilon}, \tag{3}$$

where $P_i, P_j \in \mathbb{R}^2$ are token positions, and $d_{\min}, d_{\max}$ denote the minimum and maximum pairwise distances in the current layer. This normalization ensures scale-invariance and encourages merges between spatially adjacent tokens while preserving geometric layout.

c) Attention-Aware Priority: Although token protection preserves the most critical regions, it is inherently limited in scope to avoid excessively reducing the merging candidate pool. As a result, some semantically important tokens may remain unprotected.

To further prevent the merging of such informative tokens, we introduce an attention-aware penalty that downweights the merging priority of token pairs with high attention scores. Given attention-based importance scores I_i and I_j, the penalty is defined as:

$$S_{att}(i, j) = 1 - \frac{(I_i + I_j)/2 - I_{\min}}{I_{\max} - I_{\min} + \epsilon}, \tag{4}$$

where $I_{\min}$ and $I_{\max}$ are the minimum and maximum token importance values in the current layer.

While token protection preserves the most critical tokens, the attention-aware penalty further discourages merging of unprotected yet salient ones. Together, these mechanisms adaptively preserve important regions and support strong retrieval performance under high reduction.

d) Final Score: The final merging score is computed as a weighted sum of the three factors:

$$S(i, j) = \lambda_f S_{feat}(i, j) + \lambda_p S_{pos}(i, j) + \lambda_a S_{att}(i, j), \tag{5}$$

where $\lambda_f + \lambda_p + \lambda_a = 1$. Unless stated otherwise, we set $\lambda_p = \lambda_a = 0.2$ and $\lambda_f = 0.6$.

Algorithm 1 PAGTM: Positional- and Attention-Guided Token Merging

Input: Initial token set $\mathcal{T} = \{T_i\}_{i=1}^N$, target reduction rate r, total layers L, attention maps A
Output: $\mathcal{T}_{\text{reduced}}$

$M \leftarrow \lfloor N \cdot r \rfloor$ // total number of tokens to remove
$m_\ell \leftarrow \lfloor M/L \rfloor$ for each layer ℓ
for layer $\ell = 1$ to L **do**
  Extract token embeddings and attention maps from $\mathcal{M}$ for current layer
  Compute attention-based importance $I_i \leftarrow \frac{1}{H} \sum_{h=1}^{H} A_i^{(h)}$
  Protect top-k_ℓ tokens with highest I_i
  Perform bipartite matching: $(T_i, T_j) \in \mathcal{T}_{src} \times \mathcal{T}_{ret}$
  for each pair $(T_i, T_j) \in \mathcal{T}_{src} \times \mathcal{T}_{ret}$ **do**
    Compute feature similarity $S_{feat}(i, j)$
    Compute positional proximity $S_{pos}(i, j)$
    if $\ell \leq 2$ **then**
      $S(i, j) \leftarrow \lambda_f S_{feat} + \lambda_p S_{pos}$
    else
      Compute attention-aware penalty $S_{att}(i, j)$
      $S(i, j) \leftarrow \lambda_f S_{feat} + \lambda_p S_{pos} + \lambda_a S_{att}$
    end if
  end for
  Select top-m_ℓ pairs with highest $S(i, j)$ scores
  for each selected pair (i, j) **do**
    Merge: $T_j \leftarrow \frac{T_i + T_j}{2}$, remove T_i from $\mathcal{T}$
    Update token set: $\mathcal{T}_{\text{reduced}} \leftarrow \mathcal{T} \setminus \{T_i\}$
  end for
end for
return $\mathcal{T}_{\text{reduced}}$

E. Token Reduction via Merging

To reduce the number of tokens progressively, we allocate the global reduction ratio r uniformly across the L Transformer layers. The total number of tokens to remove is $M = \lfloor N \cdot r \rfloor$, where N is the initial token count. At each

layer ℓ, we merge $m_\ell = \left\lfloor \frac{M}{L} \right\rfloor$ token pairs with the highest merging priority scores.

We select the top m_ℓ scoring token pairs (T_i, T_j), where $T_i \in \mathcal{T}_{src}$ and $T_j \in \mathcal{T}_{ret}$, and update the retained tokens by averaging features:

$$\tilde{T}_j = \frac{T_i + T_j}{2}. \tag{6}$$

The spatial position of T_j is preserved, and T_i is discarded. This operation yields a compact token set passed to the next layer.

F. Layer-Wise Reduction Strategy

Since early-layer attention maps are diffuse and do not yet capture meaningful semantics, we defer attention-based token protection and merging until layer 3, where attention becomes more focused on distinctive regions (Figure 4).

Accordingly, we set $\lambda_a = 0$ and exclude token protection in the early layers ($\ell \leq 2$). From layer 3 onward ($\ell \geq 3$), both attention-based protection and the full merging score are applied. This design follows the structure shown in Figure 3 and Algorithm 1.

G. Architectural Integration

We insert PAGTM between the self-attention and MLP blocks of each ViT layer, following prior works such as ToMe [18]. This placement allows PAGTM to utilize key embeddings and attention maps produced by the attention module, enabling informed token reduction before the computationally expensive feed-forward stage. PAGTM requires no retraining and is compatible with any ViT-based backbone without architectural modification.

IV. Experiments

A. Experimental Setup

We evaluate PAGTM on five standard VPR benchmarks: Pitts30k-Test, AmsterTime, Tokyo 24/7, Nordland, and Eynsham, covering diverse environments and challenging conditions. For evaluation, we use the official pretrained VPR models released by SALAD [15], Clique-Mining [29], and VLAD-BuFF [16], which are built on the DINOv2-ViT-B backbone [13]. These checkpoints follow the original training protocols of the respective methods, including partial fine-tuning of the backbone or training of the aggregation module where applicable. PAGTM is applied strictly at inference time as a token reduction module and does not modify or retrain any model parameters.

The effectiveness is measured using Recall@N, the standard metric in visual place recognition. Ground truth is defined as within 25 meters for Pitts30k, Tokyo 24/7, and Eynsham; ±10 frames for Nordland; and designated counterparts for AmsterTime. All images are resized to 322×322 with a batch size of 32, and experiments are conducted on an NVIDIA RTX A5000 GPU. Unless noted otherwise, the token reduction ratio is set to $r = 0.5$, meaning that 50% of tokens are removed during inference.

B. Main Results

Table I shows the retrieval performance of PAGTM compared to two state-of-the-art token reduction methods, ToMe and ToFu, under a token reduction ratio of $r = 0.5$. We evaluate on five representative VPR datasets—Pitts30k-Test, AmsterTime, Nordland, Eynsham, and Tokyo 24/7—using three widely adopted aggregation heads: SALAD, Clique-Mining, and VLAD-BuFF. All methods use the same ViT backbone and pretrained weights to ensure a fair comparison.

PAGTM consistently achieves higher Recall@1 and Recall@5 across all combinations of datasets and aggregation frameworks. For instance, on AmsterTime with SALAD, PAGTM improves Recall@1 by +1.87%p over ToMe and +2.48%p over ToFu. On Nordland with Clique-Mining, the gain reaches +2.95%p, highlighting its robustness in long-term appearance changes. Even with VLAD-BuFF—which aggregates residuals across cluster centers and is therefore relatively robust to token count changes—PAGTM yields consistent improvements across all datasets.

Notably, PAGTM achieves these performance gains while reducing token count by 50% at inference. On the challenging Tokyo 24/7 dataset with Clique-Mining, it even surpasses the full-token baseline (94.92% vs. 93.97% Recall@1), showing that PAGTM not only preserves essential visual cues but can enhance recognition by filtering out redundant tokens. These results underscore its ability to maintain or improve accuracy under significant computational compression.

We also evaluate PAGTM under other token reduction levels ($r = 0.3, 0.7, 0.8$), and observe consistent performance advantages over ToMe and ToFu, as detailed in figure 1. These results confirm that PAGTM remains robust across varying compression ratios and can serve as a practical solution for computation-constrained visual place recognition systems.

C. Ablation Study

To assess the contribution of each component in our token reduction strategy, we conduct an ablation study on the Pitts30k-Test and AmsterTime datasets using a fixed token reduction ratio of $r = 0.5$, as shown in Table II. We progressively activate the four key criteria used in our merging framework: feature similarity, spatial proximity, attention-aware priority, and attention-based protection.

Using only feature similarity yields a strong baseline. Adding spatial proximity improves performance, particularly on AmsterTime where spatial context is critical. Incorporating attention-aware priority provides further gains across all aggregators by helping preserve semantically important tokens. Finally, enabling attention-based protection from Layer 3 onward—based on key token attention scores—delivers the best performance across all metrics and datasets. These results demonstrate that each component plays a crucial role in improving retrieval accuracy under token reduction.

TABLE I

RETRIEVAL PERFORMANCE (RECALL@1 / RECALL@5) OF EACH METHOD ON FIVE DATASETS UNDER REDUCTION RATIO $r = 0.5$, USING THREE DIFFERENT AGGREGATORS: SALAD, CLIQUE-MINING, AND VLAD-BUFF. THE BEST SCORES ARE HIGHLIGHTED IN BOLD.

SALAD [15]										
Methods	Pitts30k-Test		AmsterTime		Nordland		Eynsham		Tokyo 24/7	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Baseline	92.03	96.41	57.68	77.25	86.06	93.99	91.35	95.02	94.92	96.83
ToMe [18]	91.34	96.11	53.13	75.14	81.08	91.34	90.88	94.85	91.34	96.11
ToFu [19]	91.31	95.98	52.72	74.25	80.94	91.43	90.86	94.86	90.16	**96.19**
PAGTM (Ours)	**91.61**	**96.27**	**55.00**	**75.87**	**84.68**	**93.27**	**91.23**	**95.07**	**92.70**	95.56
Clique-Mining [29]										
Methods	Pitts30k-Test		AmsterTime		Nordland		Eynsham		Tokyo 24/7	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Baseline	91.75	96.49	54.75	74.82	94.14	97.89	91.64	95.07	93.97	97.46
ToMe [18]	90.73	96.13	51.58	71.57	91.34	96.76	91.06	94.83	94.60	96.19
ToFu [19]	90.76	96.13	51.75	71.81	91.37	96.87	91.02	94.81	93.33	96.19
PAGTM (Ours)	**91.21**	**96.18**	**53.53**	**72.46**	**92.84**	**97.35**	**91.35**	**95.04**	**94.92**	**96.83**
VLAD-BuFF [16]										
Methods	Pitts30k-Test		AmsterTime		Nordland		Eynsham		Tokyo 24/7	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Baseline	92.02	96.11	56.70	78.88	83.61	92.80	91.80	95.39	96.83	98.41
ToMe [18]	91.43	95.99	53.05	75.95	76.03	88.67	91.27	95.27	94.60	96.83
ToFu [19]	91.53	96.07	54.18	76.36	75.78	88.55	91.33	95.14	94.60	**98.10**
PAGTM (Ours)	**91.55**	**96.07**	**55.32**	**76.93**	**80.15**	**90.68**	**91.74**	**95.34**	**95.87**	97.46

TABLE II

ABLATION RESULTS OF OUR TOKEN REDUCTION STRATEGY AT REDUCTION RATIO $r = 0.5$ ON THE PITTS30K-TEST AND AMSTERTIME DATASETS, EVALUATED USING SALAD, CLIQUE-MINING, AND VLAD-BUFF. RETRIEVAL PERFORMANCE IS MEASURED BY RECALL@1 AND RECALL@5.

Method Configuration				**Pitts30k-Test**						**AmsterTime**					
Feat.	**Pos.**	**Attn.**	**Attn.**	SALAD		Clique-Mining		VLAD-BuFF		SALAD		Clique-Mining		VLAD-BuFF	
Sim.	**Prox.**	**Prior.**	**Prot.**	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
✓				90.45	96.05	90.49	95.91	90.68	95.75	49.63	71.97	47.52	70.02	49.39	71.41
✓	✓			91.29	96.20	90.52	96.17	90.92	95.88	51.42	74.49	50.45	71.32	53.05	76.28
✓	✓	✓		91.53	96.14	90.86	96.17	91.34	96.07	54.67	75.22	51.42	71.41	54.75	76.04
✓	✓	✓	✓	**91.61**	**96.27**	**91.21**	**96.18**	**91.55**	**97.07**	**55.00**	**75.87**	**53.53**	**72.46**	**55.32**	**76.93**

D. Qualitative Results

We visualize and compare token merging behaviors between ToMe and our method (PAGTM) to qualitatively assess the effectiveness of our design components.

Figure 5 presents results at a reduction ratio of $r = 0.5$, illustrating the impact of our attention-aware token protection strategy. Red boxes highlight semantically important regions. In ToMe, these regions are often merged with background or adjacent textures, leading to a loss of fine-grained cues. In contrast, PAGTM dynamically protects high-attention tokens and discourages their merging using attention-aware penalties. As shown in the closed-up views, by explicitly protecting informative tokens, PAGTM preserves fine-grained semantic structure that would otherwise be lost—ultimately leading to improved retrieval performance.

Figure 6 shows token grouping results at a higher reduc-

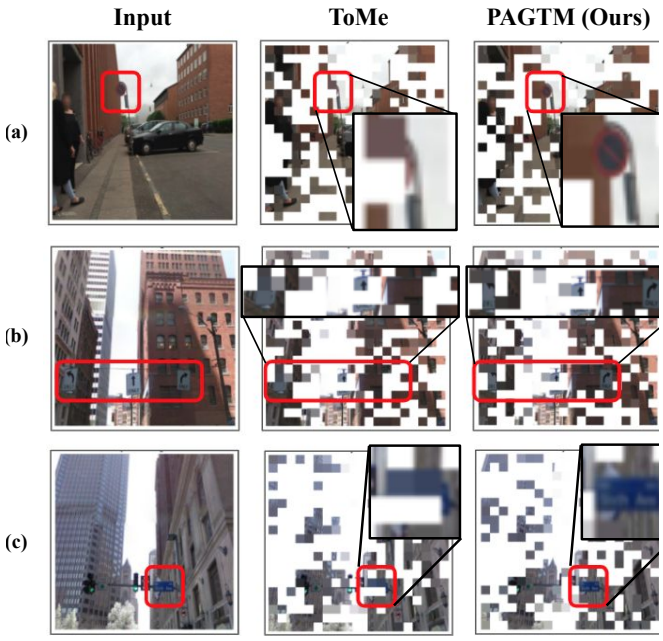

Fig. 5. Visualization of retained tokens after token reduction. Each row compares the original input with the reduced token maps from ToMe and our method (PAGTM).

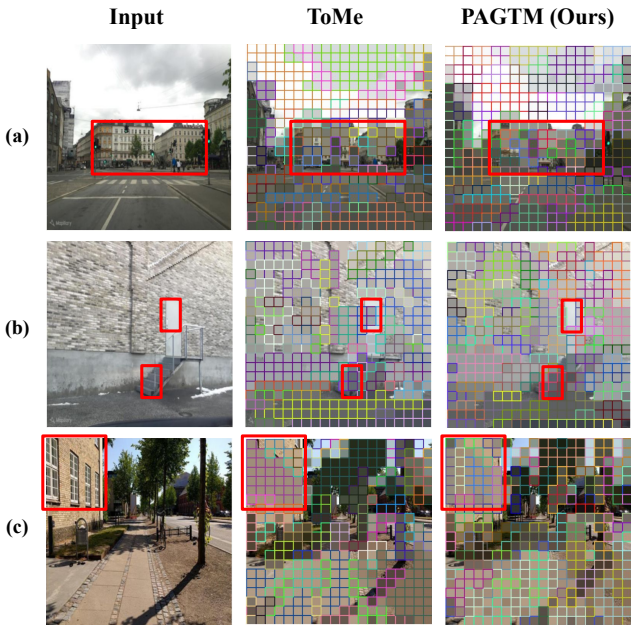

Fig. 6. Visualization of token merging results at a reduction ratio of $r = 0.7$. Each colored bounding box indicates a group of tokens merged into a single token. Red boxes indicate regions where PAGTM better preserves semantic or structural coherence compared to ToMe.

tion ratio of $r = 0.7$, focusing on the effect of our positional proximity term. ToMe, which merges primarily based on feature similarity, often disregards spatial boundaries. In (a), two visually similar buildings are merged into a single region, whereas PAGTM maintains distinct groups for each building. In (b), a metal door and cement wall are grouped together in ToMe due to similar textures, but separated in PAGTM. In (c), ToMe merges a window and an exterior wall, whereas PAGTM preserves structural separation. These results demonstrate that positional proximity improves geometric consistency under aggressive token reduction.

Together, these qualitative examples highlight that PAGTM preserves both semantic saliency and spatial structure more effectively than prior methods, even under substantial token compression.

E. Efficiency Analysis

We evaluate the efficiency of PAGTM under a standardized inference setting using a ViT-B backbone and input resolution of 322×322. The reported GFLOPs and throughput (Imgs/s) measure the feature extraction stage, including the ViT encoder and the token reduction operation. Since all aggregation frameworks produce descriptors with fixed dimensionality, the aggregation and retrieval stages remain identical across methods and therefore do not affect the reported speedup.

Table III reports GFLOPs and inference throughput across three aggregators and reduction rates $r = \{0, 0.3, 0.5, 0.7\}$. As r increases, PAGTM significantly reduces computation and boosts throughput—achieving over 34% lower GFLOPs and 50% higher throughput at $r = 0.7$ compared to the full-token baseline. These improvements come without additional FLOPs over ToMe and ToFu, while consistently delivering better retrieval performance under the same computational budget. This efficiency–accuracy trade-off, achieved without architectural changes, makes PAGTM practical for deployment on resource-limited platforms such as mobile robots.

TABLE III
COMPUTATION COST AND THROUGHPUT WITH DIFFERENT AGGREGATORS. GFLOPS AND IMAGES/SEC INCLUDE RELATIVE IMPROVEMENTS IN PARENTHESES.

Aggregator	Reduction Rate	GFlops	Imgs/s
SALAD & Clique-Mining	$r = 0$	45.82 (–%)	97.72 (–%)
	$r = 0.3$	38.86 (-15.2%)	108.30 (+10.8%)
	$r = 0.5$	34.04 (-25.7%)	124.78 (+27.7%)
	$r = 0.7$	29.75 (-35.0%)	143.56 (+46.9%)
VLAD-BuFF	$r = 0$	45.35 (–%)	95.41 (–%)
	$r = 0.3$	38.53 (-15.0%)	110.93 (+16.3%)
	$r = 0.5$	33.80 (-25.5%)	127.42 (+33.6%)
	$r = 0.7$	29.60 (-34.7%)	146.00 (+53.1%)

V. CONCLUSION

We introduced PAGTM, a training-free token merging framework that leverages attention-based importance and positional proximity to reduce tokens efficiently in ViT-based place recognition. Our method preserves semantically and spatially important information by dynamically protecting key tokens and guiding merges in a structure-aware manner, without requiring any model retraining or architectural changes.

Experiments across multiple datasets and aggregation methods show that PAGTM consistently outperforms existing approaches under the same FLOPs budget, achieving better accuracy with significantly reduced computation. Its compatibility with standard ViTs and improved inference-time efficiency make PAGTM a practical and scalable solution for resource-constrained visual localization systems.

VI. Future Work

Recent methods such as CricaVPR [26] and SelaVPR [28] reshape patch tokens into feature maps before aggregation, which makes the direct use of token reduction less straightforward. A potential future direction is to design a mechanism that restores reduced tokens prior to this reshaping step, so that efficient reduction strategies like PAGTM can also be integrated into these frameworks.

Acknowledgment

This work was supported by Korea Evaluation Institute Of Industrial Technology (KEIT) grant funded by the Korea government(MOTIE) (No.20023455, Development of Cooperate Mapping, Environment Recognition and Autonomous Driving Technology for Multi Mobile Robots Operating in Large-scale Indoor Workspace)

References

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

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

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

[4] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007.

[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

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

[7] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 080–11 090.

[8] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 726–743.

[9] 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 *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3223–3230.

[10] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[13] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[14] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, 2023.

[15] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 17 658–17 668.

[16] A. Khaliq, M. Xu, S. Hausler, M. Milford, and S. Garg, "Vlad-buff: burst-aware fast feature aggregation for visual place recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 447–466.

[17] H. Wang, B. Dedhia, and N. K. Jha, "Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 070–16 079.

[18] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," *arXiv preprint arXiv:2210.09461*, 2022.

[19] M. Kim, S. Gao, Y.-C. Hsu, Y. Shen, and H. Jin, "Token fusion: Bridging the gap between token pruning and token merging," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1383–1392.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[22] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *ICCV*, 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.

[25] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," *arXiv preprint arXiv:1911.03584*, 2019.

[26] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 772–16 782.

[27] G. Huang, Y. Zhou, X. Hu, C. Zhang, L. Zhao, and W. Gan, "Dino-mix enhancing visual place recognition with foundational vision model and feature mixing," *Scientific Reports*, vol. 14, no. 1, p. 22100, 2024.

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

[29] S. Izquierdo and J. Civera, "Close, but not there: Boosting geographic distance sensitivity in visual place recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 240–257.

[30] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13 937–13 949, 2021.

[31] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 420–14 430.

[32] X. Chen, Z. Liu, H. Tang, L. Yi, H. Zhao, and S. Han, "Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2061–2070.