ORAP: ORACLE-INSPIRED EARLY TOKEN PRUNING FOR EFFICIENT VISUAL PLACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual place recognition (VPR) has recently achieved remarkable performance with the emergence of powerful foundation models. Nonetheless, their high computational cost remains a major obstacle for efficient deployment. Token pruning offers a promising route to faster inference, yet aggressive early pruning, where the computational benefit is greatest, often causes substantial performance degradation. In this work, we uncover that token rankings induced by the final-block attention provide a highly reliable pruning signal. An oracle study further reveals that these rankings can guide early pruning with near-lossless performance. However, obtaining the oracle rankings requires an additional forward pass and therefore provides no practical acceleration. To bridge this gap, we propose OraP, an oracle-inspired early token pruning framework that anticipates late-emerging token rankings from early representations. Instead of reproducing exact attention values or distributions, OraP preserves only selection-relevant ranking relationships, enabling reliable early pruning in a single forward pass. The proposed OraP can be readily applied to accelerate various representative transformer-based VPR models. Experimental results show that OraP substantially outperforms existing token reduction methods under the same early pruning setting. By removing 50% of the patch tokens within the first transformer block, OraP reduces FLOPs by 48.9% and achieves 2.05× inference throughput with an average absolute R@1 degradation of only 0.8% across twelve benchmark datasets. On an NVIDIA Jetson Xavier NX, OraP achieves a 1.68× speedup in single-query inference, supporting its practicality for edge deployment. Code and models will be released.

1 INTRODUCTION

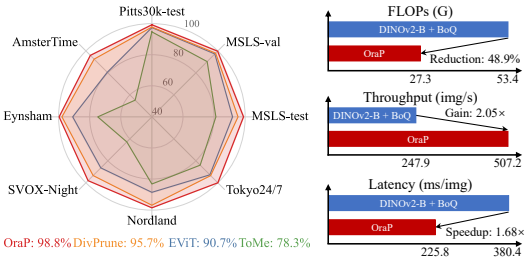
Visual place recognition (VPR) (Lowry et al., 2016) localizes a query image by retrieving its matching places from a geo-tagged reference database, and is widely used in mobile robotics (Xu et al., 2020), autonomous navigation (Bresson et al., 2017), and augmented reality (Liu et al., 2024). Recent VPR methods, exemplified by SALAD (Izquierdo & Civera, 2024a), BoQ (Ali-bey et al., 2024), and SelaVPR++ (Lu et al., 2026), have achieved superior performance by leveraging powerful vision transformer (ViT)-based foundation models (Dosovitskiy et al., 2020; Oquab et al., 2024), exhibiting strong robustness to viewpoint and environmental variations. However, the growing reliance on ViT backbones also introduces substantial computational overhead. To enable real-time place recognition, it is critical to improve inference efficiency while maintaining high retrieval accuracy.

A major computational bottleneck of ViT-based models arises from self-attention, whose complexity grows quadratically with the number of visual tokens. This issue is particularly pronounced in VPR, where higher-resolution images are often used during inference for robust place recognition (Tzachor et al., 2025; Wang et al., 2025). A promising solution is therefore to reduce the token sequence length through pruning or merging (Haurum et al., 2023). Representative methods (Rao et al., 2021; Liang et al., 2022; Bolya et al., 2023) tend to progressively reduce tokens at intermediate or later layers, effectively preserving recognition performance but typically providing only limited acceleration because much of the computation has already been incurred (Jin et al., 2026). More recently, WeiToP (Zeng et al., 2026) explores token pruning immediately after the first transformer block, offering greater computational savings but at the cost of substantial performance degradation. These results reveal a fundamental question: *can we perform reliable token pruning early in the network, where the efficiency gain is greatest, while minimizing the loss in retrieval accuracy?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To answer this question, we first investigate *what* can provide a reliable pruning signal. Prior works (Chen et al., 2017b; Shen et al., 2023; Izquierdo & Civera, 2024a; Wang et al., 2025) suggest that robust VPR models should focus on static discriminative landmarks rather than low-information backgrounds and irrelevant dynamic objects, a pattern also reflected in attention maps of ViT-based VPR models (Lu et al., 2024b; Tzachor et al., 2025; Zhang et al., 2025b; Chen et al., 2026a). Motivated by this alignment between attention and task-relevant semantics, we use the final-block class-to-patch attention scores for token pruning and find that removing 50% of patch tokens causes almost no performance degradation. This observation naturally leads to an idealized *oracle* study, where the token rankings induced by the final-block attention from a full-token forward pass are used to guide early pruning (i.e., within the first transformer block) in a second pass. Remarkably, this oracle-guided early pruning remains nearly lossless, indicating that aggressive early pruning is feasible when guided by reliable supervision. However, obtaining the oracle signal requires an additional pass, eliminating its practical efficiency gains. The key challenge therefore becomes: *how can we obtain such a reliable signal and perform early pruning within a single forward pass?*

In this paper, we propose **Ora**cle-Inspired Early Token **P**runing (**OraP**), a ranking-aware self-distillation pruning framework that anticipates reliable late-emerging token rankings directly from early representations. Instead of matching exact attention values or distributions, OraP introduces a pruning-aware ranking objective that focuses only on the ranking relationships relevant to token selection. This enables reliable token pruning within the first transformer block using a single forward pass, substantially reducing subsequent computational cost. Since the supervision is derived solely from the backbone, OraP does not rely on a specific aggregation design and can be readily applied to various transformer-based VPR models. Compared with existing methods, OraP better preserves recognition performance while significantly improving inference efficiency. As summarized in Fig. 1, OraP retains 98.8% of the full-token performance while reducing FLOPs by 48.9%, achieving 2.05× the inference throughput and a 1.68× single-query speedup on the edge device NVIDIA Jetson Xavier NX, suggesting its suitability for resource-limited deployment. Our main **contributions** are summarized as follows:



(a) Performance Comparison **(b) Efficiency Comparison**

Figure 1: Performance and efficiency of OraP under 50% token pruning. (a) R@1 retention relative to full-token inference across diverse VPR benchmark datasets, with 100% representing the full-token baseline. OraP consistently retains more retrieval performance than representative token reduction methods. (b) OraP retains 98.8% of the full-token retrieval performance while reducing FLOPs by 48.9%, achieving 2.05× throughput and a 1.68× single-query speedup.

1) We reveal through an oracle study that reliable token rankings based on the final-block attention can guide aggressive early pruning with near-lossless recognition performance. This suggests that the key challenge lies in obtaining a reliable early pruning signal rather than early pruning itself.

2) We propose OraP, a ranking-aware self-distillation pruning framework that predicts late-emerging token rankings from early representations with a pruning-aware ranking objective tailored for token selection. Since its supervision is derived solely from the backbone and independent of the aggregator, OraP can be readily applied to accelerate diverse transformer-based VPR models.

3) Extensive experiments across diverse VPR benchmarks demonstrate that OraP substantially outperforms existing token reduction methods while achieving significant inference acceleration, including substantial FLOPs reduction, higher throughput, and practical speedups on edge devices.

2 Related Work

Visual Place Recognition. VPR is commonly formulated as an image retrieval problem, where query and database images are encoded into global descriptors and matched via nearest-neighbor search. Early methods relied on hand-crafted local features (Bay et al., 2008) aggregated by schemes such as Bag of Words (Angeli et al., 2008) and VLAD (Jégou et al., 2010; Arandjelovic & Zisserman, 2013). With the rise of deep learning, CNN-based approaches (Arandjelovic et al., 2016; Radenović et al., 2018; Hausler et al., 2021; Berton et al., 2022a; 2023; Ali-bey et al., 2023) re-

placed hand-crafted features with learned representations and trainable aggregation modules such as NetVLAD (Arandjelovic et al., 2016) and GeM (Radenović et al., 2018). More recently, ViT-based foundation models, particularly DINOv2 (Oquab et al., 2024), have become widely adopted in VPR (Lu et al., 2024a; Izquierdo & Civera, 2024a; Ali-bey et al., 2024; Lu et al., 2024b;c; Qiu et al., 2024; Jin et al., 2025; Tzachor et al., 2025; Wang et al., 2025; Lu et al., 2026; Chen et al., 2026a), substantially improving retrieval performance but also introducing considerable computational cost, especially with the high-resolution inputs commonly used in VPR (Tzachor et al., 2025; Wang et al., 2025). Existing efforts toward efficient VPR mainly compress model parameters through pruning (Grainge et al., 2025a) or quantization (Ferrarini et al., 2022a;b; Grainge et al., 2024; 2025b), or reduce online computation through asymmetric retrieval (Kim et al., 2026). In contrast, our work improves inference efficiency from a token-level perspective by pruning redundant visual tokens early in the network, thus reducing subsequent computation while preserving retrieval performance.

Token Reduction. Token reduction accelerates transformer inference by shortening the token sequence processed by subsequent layers. Existing methods mainly include token pruning (Rao et al., 2021; Liang et al., 2022; Xu et al., 2022b; Yin et al., 2022; Li et al., 2022; Wang et al., 2024; Schlesinger et al., 2026; Su et al., 2026), token merging (Xu et al., 2022a; Bolya et al., 2023; Bonnaerens & Dambre, 2023; Bolya & Hoffman, 2023; Chen et al., 2026b), and hybrid strategies (Kim et al., 2024; Mao et al., 2025), which discard or combine redundant tokens according to criteria such as attention (Liang et al., 2022) or feature similarity (Bolya et al., 2023). More recently, token reduction has been extensively explored in multimodal large language models (Chen et al., 2024; Alvar et al., 2025; Shang et al., 2025; Jiang et al., 2025; Yang et al., 2025; Zhang et al., 2025a; Ye et al., 2025; Baek et al., 2026; Wang et al., 2026; Sun et al., 2026) to reduce the computational burden of visual tokens. However, most existing methods target classification or multimodal generation, while token reduction for retrieval-oriented VPR remains underexplored. A recent benchmark study (Jin et al., 2026) systematically evaluates representative token reduction methods for VPR and reveals a clear performance-efficiency trade-off: later reduction better preserves retrieval performance but offers limited acceleration, whereas early reduction provides greater savings at the cost of larger performance degradation. Closely related to our work, WeiToP (Zeng et al., 2026) transfers token importance derived from its weighted aggregation module to an early predictor through KL divergence distillation. Its supervision, however, is highly tied to a specific aggregation design (i.e., WeiAD) and still incurs noticeable performance degradation. Our work instead starts from an oracle study to investigate how reliable pruning signals emerge across transformer depth, and builds on this insight to develop a ranking-aware self-distillation framework with backbone-derived supervision for reliable early token pruning, thus marking a significant step towards efficient inference.

3 Proposed Method

3.1 Preliminaries

Transformer-based VPR Pipeline. A typical VPR model consists of a visual encoder followed by a feature aggregation stage. The encoder is commonly instantiated with a pre-trained ViT-based foundation model, such as DINOv2. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and patch size p, the patch embedding layer produces $N = HW/p^2$ patch tokens, each with dimension D. After prepending a class token and adding positional embeddings, the input sequence is denoted as $\mathbf{X}_0 \in \mathbb{R}^{(N+1)\times D}$. The sequence is then processed by L transformer blocks, each consisting of a multi-head self-attention (MHSA) sublayer and a feed-forward network (FFN), together with layer normalization (LN) and residual connections. The forward propagation through the l-th block is defined as

$$\mathbf{X}'_l = \mathrm{MHSA}(\mathrm{LN}(\mathbf{X}_{l-1})) + \mathbf{X}_{l-1}, \tag{1}$$

$$\mathbf{X}_l = \mathrm{FFN}(\mathrm{LN}(\mathbf{X}'_l)) + \mathbf{X}'_l, \tag{2}$$

where $\mathbf{X}_l$ denotes the output of the l-th block. After the final block, the output token sequence is fed into an aggregation module to produce a compact global descriptor for retrieval.

Computational Complexity. For a standard transformer block with N patch tokens, embedding dimension D, and an FFN expansion ratio of four, the dominant FLOPs can be approximated as

$$\mathcal{C}(N) \approx 12ND^2 + 2N^2D, \tag{3}$$

where the first term accounts for the token-wise linear projections in MHSA and FFN, while the second corresponds to self-attention operations. Given a pruning ratio $r \in [0, 1)$, we retain $K =$

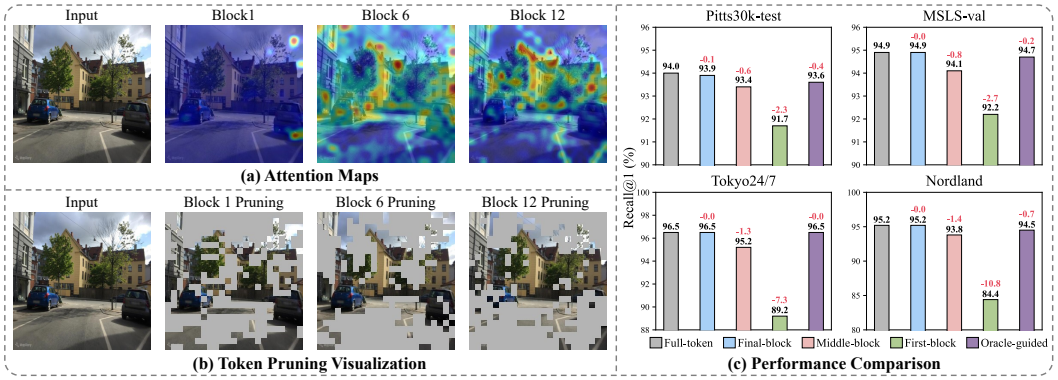


Figure 2: Observation and oracle study. (a) Attention maps across transformer blocks, showing clearer place-discriminative regions at greater depths. (b) Token pruning visualizations using class-to-patch attention scores from different blocks, with pruned tokens denoted by gray patches "■". (c) Performance under 50% pruning. First-, Middle-, and Final-block use attention from blocks 1, 6, and 12, respectively, to perform token pruning within the corresponding block, whereas Oracle-guided uses the final-block attention to guide pruning within block 1. Full-token denotes no pruning.

$\lfloor (1-r)N \rfloor$ patch tokens. The class token and other special tokens, if present, are always retained and are not counted in the pruning ratio. Ignoring the constant number of special tokens, the linear projection computation therefore scales approximately with K/N, while the attention computation scales with $(K/N)^2$. For a fixed pruning ratio r, earlier pruning allows the reduced token sequence to propagate through more subsequent modules, leading to larger end-to-end computational savings (Zeng et al., 2026). This motivates reliable token pruning at an early stage of the backbone.

3.2 Observation and Oracle Study

As discussed in Sec. 1, a well-trained VPR model tends to emphasize static and discriminative regions while suppressing less informative backgrounds and dynamic objects, a behavior also reflected in its attention maps. Thus, we first compare attention maps across different transformer blocks in Fig. 2 (a) and observe that such place-discriminative attention patterns become progressively clearer with the network depth. To examine whether these attention patterns can serve as an effective pruning signal for VPR, we rank patch tokens according to the class-to-patch attention scores. Specifically, for the l-th transformer block, the score of the i-th patch token is defined as

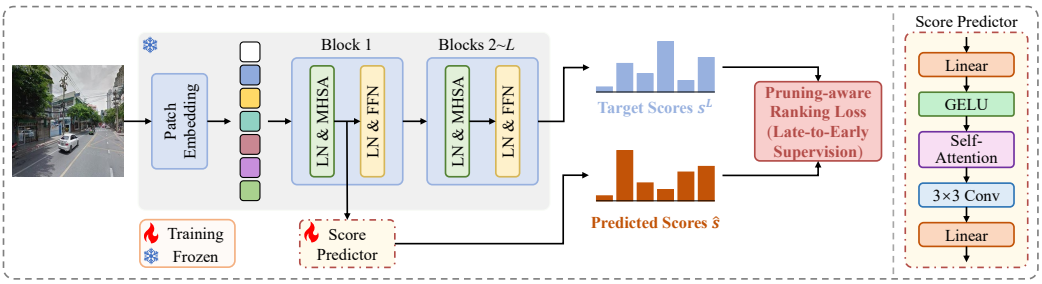
$$s_i^l = \frac{1}{n_h} \sum\nolimits_{h=1}^{n_h} \left[\mathbf{A}_h^l\right]_{\mathrm{cls},i}, \tag{4}$$

where $\mathbf{A}_h^l$ denotes the self-attention matrix of the h-th head and n_h is the number of attention heads. Patch tokens are then ranked by $\mathbf{s}^l$, with the top-K tokens retained. As shown in Fig. 2 (b), tokens retained according to the final-block scores $\mathbf{s}^L$ mainly correspond to place-discriminative structures, whereas discarded tokens largely cover less informative regions. Consistent with this observation, Fig. 2 (c) further shows that pruning 50% of patch tokens using $\mathbf{s}^L$ causes almost no performance degradation, whereas pruning based on attention from earlier blocks leads to noticeably larger drops.

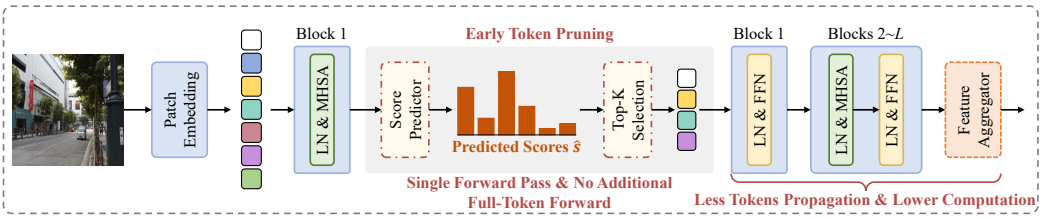
These observations motivate an idealized *oracle* study that decouples the source of the pruning signal from the location where pruning is performed. Specifically, we first conduct a full-token forward pass to obtain the final-block scores $\mathbf{s}^L$, and then use the resulting token rankings to prune patch tokens between MHSA and FFN within the first transformer block in a second pass. As shown in Fig. 2 (c), this oracle-guided strategy remains nearly lossless, with at most 0.7% absolute R@1 degradation across all four benchmarks. This result demonstrates that aggressive early pruning itself is feasible once a reliable token ranking is available. However, obtaining the oracle signal requires an additional full-token forward pass, eliminating the computational benefit of early pruning.

3.3 Learning Late-Emerging Token Rankings via Self-Distillation

To avoid the additional forward pass required by oracle-guided early pruning, we propose a self-distillation framework that learns to predict reliable late-emerging token rankings directly from early



(a) Training: Learn to Predict Late-Emerging Token Rankings via Self-Distillation



(b) Inference: Prediction-Guided Early Token Pruning for Higher Efficiency

Figure 3: Overview of OraP. (a) During training, the final-block class-to-patch attention scores from the frozen full-token backbone supervise an early score predictor via a pruning-aware ranking loss. (b) During inference, the predictor estimates token scores after the first-block MHSA, enabling top-K pruning before the FFN in a single forward pass. Only the retained tokens are propagated through subsequent transformer blocks and the aggregator, substantially reducing computational burden.

representations, as illustrated in Fig. 3. During training, a full-token forward pass produces the final-block attention scores $\mathbf{s}^L = [s_1^L, \ldots, s_N^L]$ defined in Eq. 4, which serve as the distillation target. Meanwhile, after the MHSA sublayer of the first transformer block, a lightweight predictor f_ψ takes the early patch representations $\mathbf{X}_1'$ as input and estimates their future ranking scores

$$\hat{\mathbf{s}} = f_\psi(\mathbf{X}_1') = [\hat{s}_1, \ldots, \hat{s}_N]. \tag{5}$$

Rather than directly relying on the first-block class-to-patch attention scores, whose rankings are still unreliable at this stage, we predict from the richer patch representations $\mathbf{X}_1'$ to recover the rankings induced by $\mathbf{s}^L$. During inference, the predicted scores $\hat{\mathbf{s}}$ are used to retain the top-K patch tokens after the first MHSA sublayer, which are then forwarded to the immediately following FFN and all other subsequent modules. This enables early token pruning within a single forward pass without accessing the final-block attention, thus substantially reducing computational cost.

Since the target rankings emerge after repeated self-attention interactions across L transformer blocks, predicting them from early representations may benefit from contextual information beyond individual token features. We therefore employ a lightweight global-local score predictor instead of relying solely on a token-wise mapping (e.g., an MLP). Specifically, a linear projection first reduces the feature dimension, followed by lightweight self-attention to capture global token dependencies and a convolutional module to model local spatial interactions. The refined features are then mapped to a scalar score for each patch token. This design incorporates both global and local context while introducing only minor overhead relative to the computation saved by early pruning. We next describe how the late-emerging rankings are transferred to the early predictor through a pruning-aware ranking objective designed to align supervision with the subsequent top-K token selection.

3.4 Pruning-Aware Ranking Loss

A straightforward way to transfer the final-block supervision is to regress the attention scores using mean squared error (MSE) (Xia et al., 2022), or match their distributions through KL divergence (Rao et al., 2021; Zeng et al., 2026). However, matching final-block attention values or distributions is not ideal for early prediction. On the one hand, early representations are less semantically mature, making it difficult to reproduce the attention values that emerge after deep feature transformations. On the other hand, such precise prediction is unnecessary for top-K pruning, which mainly depends on the relative ordering of token scores. As long as the retained tokens are ranked above discarded ones, the resulting token selection remains unchanged. We therefore formulate the distillation objective as a ranking problem rather than directly matching attention values or distributions.

A natural ranking objective would preserve the global ordering induced by the final-block scores $\mathbf{s}^L$, as commonly considered in pairwise learning-to-rank formulations (Burges et al., 2005). However, enforcing an identical ranking over all tokens is still stricter than necessary for pruning. Given a pruning ratio r, let $K = \lfloor (1-r)N \rfloor$, and define $\mathcal{T}_r$ as the top-K retained tokens and $\mathcal{D}_r$ as the discarded tokens according to $\mathbf{s}^L$, respectively. Notably, the exact ordering within either subset does not affect the top-K selection. Therefore, we only enforce cross-boundary ranking relations, requiring patch tokens in $\mathcal{T}_r$ to receive higher predicted scores than those in $\mathcal{D}_r$. Specifically, we randomly sample M cross-boundary token pairs $\{(i_m, j_m)\}_{m=1}^{M}$, where $i_m \in \mathcal{T}_r$, $j_m \in \mathcal{D}_r$, and $s_{i_m}^L > s_{j_m}^L$. We then optimize the predictor using the following pruning-aware ranking objective

$$\mathcal{L}_{\text{rank}}^{(r)} = \frac{1}{M} \sum\nolimits_{m=1}^{M} \log\left(1 + \exp(\hat{s}_{j_m} - \hat{s}_{i_m})\right). \tag{6}$$

Minimizing $\mathcal{L}_{\text{rank}}^{(r)}$ encourages $\hat{s}_{i_m} > \hat{s}_{j_m}$ while leaving the within-set ordering unconstrained, thus directly aligning the supervision with the top-K pruning decision.

To avoid specializing the predictor to a single pruning ratio, we use multiple pruning boundaries during training. For each ratio r, cross-boundary pairs are sampled from $\mathcal{T}_r$ and $\mathcal{D}_r$, yielding

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathcal{L}_{\text{rank}}^{(r)}, \tag{7}$$

where $\mathcal{R}$ denotes the set of different training pruning ratios. This multi-boundary supervision improves ranking robustness across different token budgets at no extra inference overhead, enabling the same predictor to support different test-time pruning ratios without retraining or fine-tuning.

4 Experiments

4.1 Experimental Setup

Evaluation Benchmark Datasets. We conduct the experiments on multiple widely used VPR datasets, including Pitts30k-test and Pitts250k-test (Torii et al., 2013), MSLS-val and MSLS-test (Warburg et al., 2020), Tokyo24/7 (Torii et al., 2015), Nordland (Olid et al., 2018), Eynsham (Cummins & Newman, 2010; Berton et al., 2022b), SPED (Chen et al., 2017a), AmsterTime (Yildiz et al., 2022), and SVOX (Berton et al., 2021). Table 1 summarizes their main information. Further details are provided in Appendix A.8.

Table 1: Summary of the evaluation datasets.

Dataset	Description	Number	
		Database	Queries
Pitts30k-test	urban, panorama	10,000	6,816
MSLS-val	urban, suburban	18,871	740
MSLS-test	long-term	38,770	27,092
Tokyo24/7	urban, day/night	75,984	315
Nordland	natural, seasonal	27,592	27,592
Eynsham	rural, panorama	23,935	23,935
Pitts250k-test	urban, panorama	83,952	8,280
SPED-test	various scenes	607	607
AmsterTime	long-term	1,231	1,231
SVOX	cross-domain	17,166	4,356

Evaluation Metrics. We use Recall@N (R@N) as the evaluation metric, i.e., the percentage of query images for which at least one of the top-N candidates falls within a threshold of ground truth. Consistent with prior works (Berton et al., 2022b; Jin et al., 2025; Chen et al., 2026a), we set the threshold to 25 meters and 40° angle for MSLS, 25 meters for Pitts30k-test, Tokyo24/7, Eynsham, Pitts250k-test and SPED, ± 10 frames for Nordland, as well as a special counterpart for AmsterTime.

Model Settings. Unless otherwise specified, we use DINOv2-B with the BoQ aggregator, which is trained on the merged training set introduced by SelaVPR++ (Lu et al., 2026), following prior work (Jin et al., 2026). All input images are resized to 322×322 for inference, yielding 529 patch tokens. Detailed training settings for the VPR model and score predictor are provided in Appendix A.1.

4.2 Main Results

Comparison Methods. We compare OraP with representative token reduction methods covering diverse criteria, including random pruning, learning-based DynamicViT (Rao et al., 2021), attention-based EViT (Liang et al., 2022), similarity-based ToMe (Bolya et al., 2023) and ToFu (Kim et al., 2024), diversity-based DivPrune (Alvar et al., 2025), graph-based G-Prune (Jiang et al., 2025), and

Table 2: Comparison to representative token reduction methods on six benchmarks. The best and second-best results among token reduction methods are shown in **red** and **blue** bold, respectively.

Method	Pitts30k-test		MSLS-val		MSLS-test		Tokyo24/7		Nordland		Eynsham	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Full-token Reference, All 529 Patch Tokens (100%)												
Full-token Inference	94.0	97.3	94.9	97.7	76.5	86.2	96.5	98.7	95.2	98.4	92.5	95.9
Retain 264 Patch Tokens (↓ 50%)												
Random	90.1	96.4	91.4	96.3	69.9	81.7	92.6	96.5	85.9	94.4	86.7	93.4
DynamicViT NIPS' 2021	91.0	96.0	**93.9**	**97.4**	**73.9**	**85.1**	**93.0**	96.8	91.7	96.8	**91.1**	**95.2**
EViT ICLR' 2022	91.6	96.7	92.3	97.0	70.3	82.0	89.5	94.3	84.3	93.3	84.1	91.4
ToMe ICLR' 2023	89.0	95.6	85.7	94.7	62.1	76.2	81.0	91.1	79.2	91.3	69.2	82.9
ToFu WACV' 2024	91.5	96.6	92.0	96.9	71.6	82.9	**93.0**	**97.5**	87.4	95.1	87.2	93.7
DivPrune CVPR' 2025	91.9	96.7	93.0	**97.4**	73.2	84.5	90.2	96.5	**92.2**	**97.2**	90.5	95.1
G-Prune AAAI' 2025	87.7	95.1	87.3	93.5	62.4	74.9	88.9	94.0	32.6	45.7	86.9	93.4
AgilePruner ICLR' 2026	**92.3**	**96.8**	93.4	96.9	73.2	84.1	92.4	96.5	88.5	95.5	90.3	94.9
OraP (Ours)	**93.3**	**97.0**	**94.7**	**97.7**	**75.7**	**85.6**	**96.5**	**98.4**	**93.8**	**97.8**	**92.2**	**95.7**

Table 3: Comparison to selected strong token reduction baselines on supplementary VPR datasets.

Method	Pitts250k-test		SPED-test		AmsterTime		SVOX-Night		SVOX-Rain		SVOX-Sun	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Full-token Reference, All 529 Patch Tokens (100%)												
Full-token Inference	96.2	99.0	90.6	95.9	64.3	84.7	97.3	99.6	98.4	99.7	97.5	99.1
Retain 264 Patch Tokens (↓ 50%)												
DynamicViT NIPS' 2021	93.3	97.9	**89.6**	**95.4**	**60.1**	**81.9**	**92.2**	**98.3**	97.2	99.1	**96.7**	**98.8**
EViT ICLR' 2022	93.7	98.3	87.5	94.2	51.8	73.8	84.1	93.9	92.1	97.0	91.0	96.4
ToFu WACV' 2024	93.7	98.3	89.1	94.7	55.2	78.2	86.3	95.7	96.8	99.3	95.0	98.4
DivPrune CVPR' 2025	94.5	**98.5**	89.1	95.1	59.5	79.9	90.9	98.1	**97.9**	**99.5**	96.4	98.7
AgilePruner ICLR' 2026	**94.7**	**98.8**	89.3	95.2	57.4	78.6	91.3	**98.3**	**97.9**	99.4	96.3	98.7
OraP (Ours)	**95.5**	**98.8**	**90.3**	**95.6**	**61.7**	**82.5**	**95.4**	**98.7**	**98.1**	**99.7**	**97.5**	**99.1**

the attention-diversity hybrid AgilePruner (Baek et al., 2026). Detailed descriptions are provided in Appendix A.9. For a consistent comparison, all methods perform token reduction at the same location, i.e., between the MHSA and FFN sublayers of the first transformer block, with a fixed reduction ratio of 50%. Results under additional reduction ratios are provided in Appendix A.2.

Performance Analysis. As shown in Table 2 and Table 3, OraP consistently achieves the best R@1 and R@5 among the evaluated token reduction methods on all twelve VPR datasets. The advantages are particularly pronounced on several challenging benchmarks. Compared with the best competing method in R@1 on each dataset, OraP achieves absolute improvements of 1.8% on MSLS-test, 3.5% on Tokyo24/7, 1.6% on Nordland, 1.6% on AmsterTime, and 3.2% on SVOX-Night. Averaged across all twelve datasets, the absolute R@1 improvement over the best competing method on each dataset is 1.4%. Notably, OraP substantially outperforms EViT, although both methods rely on attention-derived signals for token selection. EViT directly uses the attention scores available at the pruning location, where the attention-derived token rankings are less reliable, as suggested by our oracle study. In contrast, OraP predicts late-emerging token rankings from early representations and uses the resulting scores to guide early pruning. This distinction leads to substantial performance gaps, with OraP achieving absolute R@1 improvements of 7.0% on Tokyo24/7, 9.5% on Nordland, and 11.3% on SVOX-Night over EViT. More importantly, despite removing 50% of the patch tokens within the first transformer block, OraP remains remarkably close to full-token inference, with an average absolute R@1 degradation of only 0.6% on the six main benchmarks and 0.8% across all twelve datasets. OraP even matches the full-token R@1 on Tokyo24/7 and SVOX-Sun. These results further demonstrate that OraP can effectively anticipate reliable late-emerging token rankings from early representations, enabling early pruning with only minor degradation in retrieval performance.

Qualitative Analysis. Fig. 4 provides a qualitative comparison of different token reduction methods under diverse challenging conditions. Competing methods such as EViT and DivPrune often discard place-discriminative structures or retain fragmented regions, leading to incorrect retrievals, including visually similar but geographically wrong matches beyond the predefined threshold, as shown in the first and second examples. Under viewpoint changes and occlusions, OraP better preserves stable cues such as building facades and railings, while competing methods retain less coherent visual evidence. Similar behavior appears under severe lighting and seasonal changes, where OraP preserves visual cues consistent between the query and its ground-truth reference despite substantial appearance variations. Consequently, OraP yields more reliable retrievals across these challenging cases. Additional qualitative examples and analyses of failure cases are provided in Appendix A.7.

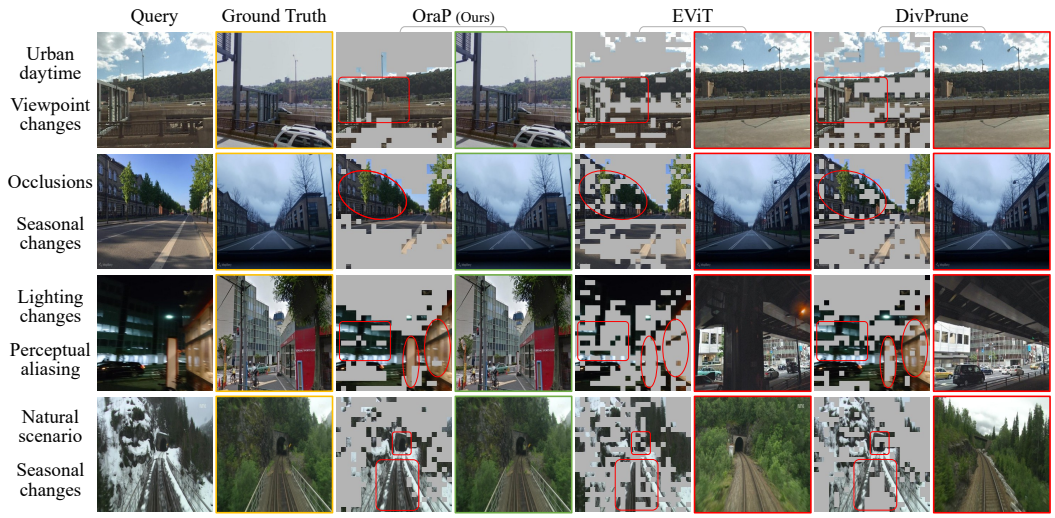


Figure 4: Qualitative results. For each case, the first two columns show the query and its ground truth. For each method, the left image visualizes the retained patches after pruning, and the right shows the top-1 retrieved result. The highlighted regions indicate place-discriminative structures that are better preserved by OraP, enabling correct retrieval under various challenging conditions.

Table 4: Inference efficiency comparison. Throughput is measured on an NVIDIA RTX 4090 with the batch size set to 32, while single-query latency is evaluated on an NVIDIA Jetson Xavier NX.

Model	Reduction	Params		FLOPs		Throughput		Latency	
		(M)	rel.	(G)	rel.	(img/s)	rel.	(ms/img)	rel.
DINOv2-B + BoQ	Full-token	92.9	100%	53.4	100%	247.9	1.0×	380.4	1.0×
	Random	92.9	100%	26.7	50.0%	516.8	2.08×	221.9	1.71×
	OraP	93.9	101.1%	27.3	51.1%	507.2	2.05×	225.8	1.68×

Table 5: Generalization of OraP across diverse VPR models with different aggregation paradigms. All models use the DINOv2-B backbone. We compare full-token inference with OraP under a 50% reduction ratio, with pruning performed between the MHSA and FFN of the first transformer block.

VPR Models	FLOPs	Throughput	Pitts30k-test		MSLS-val		MSLS-test		Tokyo24/7		Nordland	
	(G)	(img/s)	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD	50.9	261.2	93.8	97.1	94.7	97.4	76.3	86.0	97.1	98.7	97.0	98.9
NetVLAD+OraP	**26.0**	**534.3**	93.2	97.1	94.2	97.4	75.5	85.6	96.5	98.4	95.3	98.4
SALAD	51.4	253.5	93.6	97.0	94.7	97.6	76.1	86.0	97.5	98.1	96.5	98.7
SALAD+OraP	**26.2**	**513.6**	93.0	96.8	94.5	97.4	75.1	85.3	96.5	97.8	95.0	98.1
ImAge	51.2	262.1	93.9	97.2	94.6	97.4	76.9	86.7	97.1	98.7	97.3	99.1
ImAge+OraP	**26.2**	**536.7**	93.6	97.2	94.3	97.3	76.1	86.2	95.9	98.1	95.2	98.4

Efficiency Analysis. We evaluate the inference efficiency of OraP in terms of parameters, FLOPs, throughput, and single-query latency. As shown in Table 4, OraP introduces only 1.0M additional parameters, corresponding to a 1.1% increase over the original model. Nonetheless, OraP reduces the computational cost from 53.4G to 27.3G FLOPs, achieving a 48.9% reduction. This translates into practical acceleration: 1) on an NVIDIA RTX 4090, OraP increases the throughput from 247.9 to 507.2 img/s, reaching 2.05× the baseline throughput; 2) on an NVIDIA Jetson Xavier NX (i.e., edge device), it reduces single-query latency from 380.4 ms to 225.8 ms, corresponding to a 1.68× speedup. Compared with random pruning that introduces no extra module, OraP incurs only marginal additional overhead, indicating that our lightweight predictor preserves most of the efficiency gains of early pruning. Additional training efficiency analysis is provided in Appendix A.4.

4.3 Accelerating Diverse VPR Models with OraP

Beyond the default "DINOv2-B + BoQ" setting, an important advantage of OraP is that its supervision is derived from the visual backbone rather than a specific aggregation design. We further apply OraP to three other representative VPR models with different aggregation paradigms: NetVLAD (Arandjelovic et al., 2016), SALAD (Izquierdo & Civera, 2024a), and the implicit aggregation method ImAge (Lu et al., 2025). As shown in Table 5, OraP consistently preserves strong retrieval

Table 6: Effect of pruning location on retrieval performance and inference efficiency.

Pruning Location	FLOPs (G)	Throughput (img/s)	Pitts30k-test		MSLS-test		Tokyo24/7		Nordland	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
After Patch Embedding	26.3	530.3	92.3	96.8	74.4	84.9	94.9	97.8	91.9	97.1
Within Block 1 (Ours)	27.3	507.2	93.3	97.0	75.7	85.6	96.5	98.4	93.8	97.8
After Block 1	28.5	490.1	93.3	97.0	75.7	85.6	96.5	98.4	93.9	97.9
After Block 2	30.7	450.5	93.4	97.1	75.8	85.7	96.5	98.4	94.0	98.0
After Block 3	32.9	416.7	93.4	97.2	75.9	85.8	96.5	98.7	94.1	98.1

performance while substantially improving inference efficiency across all three models. Specifically, it reduces FLOPs by approximately 49% and achieves over 2× the full-token throughput for each model. Despite pruning 50% of the patch tokens within the first transformer block, the average absolute R@1 degradation across five benchmark datasets and three models is only 0.9%. These results demonstrate that OraP generalizes across different aggregation paradigms and provides an effective token-level acceleration mechanism for diverse transformer-based VPR models.

4.4 Ablation Studies

We conduct ablation studies on several key design choices of OraP: the pruning location, distillation objective, and score predictor. Unless otherwise specified, all experiments use the "DINOv2-B + BoQ" model with a 50% pruning ratio. Ablation on the predictor is provided in Appendix A.3.

Effect of Pruning Location. We first investigate the pruning location by moving the score predictor and pruning operation to different depths of the backbone. As shown in Table 6, pruning immediately after patch embedding achieves the highest efficiency but leads to noticeably larger retrieval performance degradation, with R@1 dropping to 92.3%, 74.4%, 94.9%, and 91.9% on Pitts30k-test, MSLS-test, Tokyo24/7, and Nordland, respectively. Moving pruning after the first MHSA significantly improves performance with only a small efficiency cost, requiring 27.3G FLOPs while maintaining 507.2 img/s throughput. Further postponing pruning to the end of block 1 or later blocks yields only marginal gains. For example, moving pruning from within block 1 to after block 3 improves R@1 by at most 0.3%, while increasing FLOPs from 27.3G to 32.9G and reducing throughput from 507.2 to 416.7 img/s. These results indicate that the representations after the first MHSA already contain sufficient information for the proposed OraP to estimate effective token rankings, making the position between the first MHSA and FFN a favorable performance-efficiency trade-off.

Table 7: Effect of distillation objectives on performance.

Objective	Pitts30k-test		MSLS-test		Tokyo24/7		Nordland	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
MSE	91.9	96.6	74.4	84.6	94.3	97.5	92.3	97.3
KL	92.3	96.8	74.9	85.1	94.9	97.8	92.8	97.5
Full Ranking	92.9	96.9	75.3	85.4	95.9	98.1	93.3	97.6
Ours	**93.3**	**97.0**	**75.7**	**85.6**	**96.5**	**98.4**	**93.8**	**97.8**

Effect of Distillation Objective. As shown in Table 7, ranking-based distillation objectives consistently outperform MSE and KL divergence, suggesting that relative token ordering is more relevant to pruning than reproducing attention values or distributions. Full-ranking supervision already yields strong performance, but it imposes additional within-set ordering constraints that are harder to satisfy from early representations. By focusing only on cross-boundary ordering, our pruning-aware ranking objective relaxes these unnecessary constraints and improves R@1 by 0.4-0.6% over full ranking across the four benchmarks. Thus, aligning the objective with the top-K pruning decision is more suitable than preserving unnecessary score or ranking information when predicting from semantically immature early representations.

5 Conclusion

In this work, we investigate early token pruning for efficient visual place recognition and reveal that reliable token rankings induced by the final-block attention can guide aggressive early pruning with minimal performance degradation. Building on this finding, we propose OraP, a ranking-aware self-distillation framework that predicts these late-emerging token rankings from early representations and enables pruning within the first transformer block in a single forward pass. Experiments show that OraP outperforms existing token reduction methods under the same setting while substantially improving inference efficiency, and generalizes well across diverse transformer-based VPR models. We believe that the proposed OraP provides an effective and practical way to accelerate advanced VPR models through reliable early token pruning for real-world large-scale applications.

AI USE STATEMENT

Generative AI tools were primarily used for language polishing, improving the clarity and organization of the manuscript, and assisting with figure and table design. All research ideas, methodological decisions, implementations, experiments, analyses, and conclusions were independently reviewed and verified by the authors. The authors take full responsibility for the content of this work, including all AI-assisted text and artifacts.

REPRODUCIBILITY STATEMENT

We make efforts to ensure the reproducibility of our work by providing detailed descriptions of the proposed method, model architecture, training objectives, and inference procedure in the main paper. Implementation details, hyperparameter settings, and training configurations are provided in the appendix. All datasets used in our experiments are publicly available and can be obtained from the official sources provided in the corresponding references. We also provide detailed descriptions of the datasets and data processing procedures in the appendix to facilitate consistent evaluation. The source code and configuration files will be released upon publication to support reproducibility.

REFERENCES

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2998–3007, 2023.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803, 2024.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401, 2025.

Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, 2008.

Relja Arandjelovic and Andrew Zisserman. All about vlad. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1578–1585, 2013.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.

Changwoo Baek, Jouwon Song, Sohyeon Kim, and Kyeongbo Kong. Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models. In *International Conference on Learning Representations*, 2026.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4878–4888, 2022a.

Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5396–5407, 2022b.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11080–11090, 2023.

Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2918–2927, 2021.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4599–4603, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023.

Maxim Bonnaerens and Joni Dambre. Learned thresholds token merging and pruning for vision transformers. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Guillaume Bresson, Zayed Alsayed, Li Yu, and Sébastien Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220, 2017.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Shunpeng Chen, Changwei Wang, Rongtao Xu, Peixingtian, Yukun Song, Jinzhou Lin, Wenhao Xu, Jingyi Zhang, Li Guo, and Shibiao Xu. SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yutian Chen, Yuheng Qiu, Ruogu Li, Jay Patrikar, and Sebastian Scherer. Co-me: Confidence guided token merging for visual geometric transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14590–14599, 2026b.

Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *IEEE international conference on robotics and automation (ICRA)*, pp. 3223–3230. IEEE, 2017a.

Zetao Chen, Fabiola Maffra, Inkyu Sa, and Margarita Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9–16, 2017b.

Mark Cummins and Paul Newman. Highly scalable appearance-only slam–fab-map 2.0. 2010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Bruno Ferrarini, Michael Milford, Klaus D. McDonald-Maier, and Shoaib Ehsan. Highly-efficient binary neural networks for visual place recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5493–5500, 2022a.

Bruno Ferrarini, Michael J. Milford, Klaus D. McDonald-Maier, and Shoaib Ehsan. Binary neural networks for memory-efficient and effective visual place recognition in changing environments. *IEEE Transactions on Robotics*, 38(4):2617–2631, 2022b.

Oliver Grainge, Michael Milford, Indu Bodala, Sarvapali D. Ramchurn, and Shoaib Ehsan. Design space exploration of low-bit quantized neural networks for visual place recognition. *IEEE Robotics and Automation Letters*, 9(6):5070–5077, 2024.

Oliver Grainge, Michael Milford, Indu Bodala, Sarvapali D. Ramchurn, and Shoaib Ehsan. Structured pruning for efficient visual place recognition. *IEEE Robotics and Automation Letters*, 10 (2):2024–2031, 2025a.

Oliver Grainge, Michael J. Milford, Indu Bodala, Sarvapali D. Ramchurn, and Shoaib Ehsan. Tetravpr: A ternary transformer approach for compact visual place recognition. *IEEE Robotics and Automation Letters*, 10(8):8396–8403, 2025b.

Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 773–783, 2023.

Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14152, 2021.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17658–17668, 2024a.

Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *European Conference on Computer Vision*, pp. 240–257. Springer, 2024b.

Hervé Jégou et al. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311, 2010.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4075–4083, 2025.

Tong Jin, Feng Lu, Shuyu Hu, Chun Yuan, and Yunpeng Liu. Edtformer: An efficient decoder transformer for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8835–8848, 2025.

Tong Jin, Yunpeng Liu, Shuyu Hu, Qinghua Zhang, Ruize Han, Song Wang, and Feng Lu. Are all tokens necessary for visual place recognition? an empirical study of token reduction for efficient inference, 2026.

Jihwan Kim, Yuseok Cho, and Sung-Eui Yoon. Towards test-time efficient visual place recognition via asymmetric query processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5673–5681, 2026.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1372–1381, 2024.

Ling Li, David Thorsley, and Joseph Hassoun. Sait: Sparse vision transformers through adaptive token pruning. *arXiv preprint arXiv:2210.05832*, 2022.

Youwei Liang, GE Chongjian, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Evit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Haomin Liu, Linsheng Zhao, Zhen Peng, Weijian Xie, Mingxuan Jiang, Hongbin Zha, Hujun Bao, and Guofeng Zhang. A low-cost and scalable framework to build large-scale localization benchmark for augmented reality. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2274–2288, 2024.

Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1): 1–19, 2016.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun 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 (CVPR)*, pp. 16772–16782, 2024a.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *International Conference on Learning Representations*, 2024b.

Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5789–5816, 2024c.

Feng Lu, Tong Jin, Canming Ye, Xiangyuan Lan, Yunpeng Liu, and Chun Yuan. Towards implicit aggregation: Robust image representation for place recognition in the transformer era. In *Advances in Neural Information Processing Systems*, 2025.

Feng Lu, Tong Jin, Xiangyuan Lan, Lijun Zhang, Yunpeng Liu, Yaowei Wang, and Chun Yuan. Selavpr++: Towards seamless adaptation of foundation models for efficient place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2731–2748, 2026.

Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

Junzhu Mao, Yang Shen, Jinyang Guo, Yazhou Yao, Xiansheng Hua, and Hengtao Shen. Prune and merge: Efficient token compression for vision transformer with spatial information preserved. *IEEE Transactions on Multimedia*, 27:4670–4683, 2025.

Daniel Olid, José M. Fácil, and Javier Civera. Single-view place recognition under seasonal changes, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Qibo Qiu, Shun Zhang, Haiming Gao, Honghui Yang, Haochao Ying, Wenxiao Wang, and Xiaofei He. Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing. In *Advances in Neural Information Processing Systems*, pp. 120928–120950, 2024.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

Oded Schlesinger, Amirhossein Farzam, J Matias Di Martino, and Guillermo Sapiro. Spot: Sparsification with attention dynamics via token relevance in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2335–2345, 2026.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Yanqing Shen, Sanping Zhou, Jingwen Fu, Ruotong Wang, Shitao Chen, and Nanning Zheng. Structvpr: Distill structural knowledge with weighting samples for visual place recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11217–11226, 2023.

Wei-Yuan Su, Ruijie Zhang, and Zheng Zhang. Rényi entropy: A new token pruning metric for vision transformers. *arXiv preprint arXiv:2603.27900*, 2026.

Zixun Sun, Yubo Dong, Hehe Fan, and Yi Yang. Hi-lo prune: Look at what you'll lose before pruning with hierarchical token selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31941–31951, 2026.

Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 883–890, 2013.

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1808–1817, 2015.

Issar Tzachor, Boaz Lerner, Matan Levy, Michael Green, Tal Berkovitz Shalev, Gavriel Habib, Dvir Samuel, Noam Korngut Zailer, Or Shimshi, Nir Darshan, and Rami Ben-Ari. EffoVPR: Effective foundation model utilization for visual place recognition. In *International Conference on Learning Representations*, 2025.

Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, et al. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7536–7544, 2025.

Hongjie Wang, Bhishma Dedhia, and Niraj 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 (CVPR)*, pp. 16070–16079, 2024.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5022–5030, 2019.

Ziyang Wang, Mengwei Li, Hao Yin, Wenhao Liu, and Zilei Wang. Posprune: Visual token pruning with positional bias correction for efficient large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10485–10493, 2026.

Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2626–2635, 2020.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1513–1528, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18113–18123, 2022a.

Ming Xu, Niko Sünderhauf, and Michael Milford. Probabilistic visual place recognition for hierarchical localization. *IEEE Robotics and Automation Letters*, 6(2):311–318, 2020.

Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2964–2972, 2022b.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2025.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22128–22136, 2025.

Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2749–2755, 2022.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10809–10818, 2022.

Zichao Zeng, June Moh Goo, Junwei Zheng, Weijia Fan, Jiaming Zhang, Rainer Stiefelhagen, and Jan Boehm. Faster or stronger: Towards flexible visual place recognition via weighted aggregation and token pruning, 2026.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867, 2025a.

Sitao Zhang, Hongda Mao, Qingshuang Chen, and Yelin Kim. Efficient visual place recognition through multimodal semantic knowledge integration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5601–5610, 2025b.

A Appendix

A.1 Implementation Details

VPR Model Preparation. Unless otherwise specified, we use DINOv2-B as the backbone and BoQ as the aggregator. Following (Jin et al., 2026), the VPR model is trained on the merged training set introduced by SelaVPR++ (Lu et al., 2026) using the multi-similarity loss (Wang et al., 2019). During training, images are resized to 224×224, and the model is optimized with AdamW using an initial learning rate of 5×10^{-5}, which is halved every three epochs. Each training batch contains 120 places with 4 images per place (i.e., 480 images in total). During inference, the input resolution is set to 322×322, yielding 529 patch tokens for DINOv2 with a patch size of 14.

Score Predictor Training. After obtaining the well-trained VPR model, we freeze all its parameters and optimize only the lightweight score predictor. The predictor is trained on GSV-Cities (Ali-bey et al., 2022) with an input resolution of 322×322. Each training batch contains 60 places with two images sampled per place, yielding 120 images in total. Place labels are not used for optimization, as the supervision is derived solely from the frozen backbone of the VPR model. The predictor is inserted between the MHSA and FFN sublayers of the first transformer block and is supervised by the rankings derived from the final-block class-to-patch attention. To improve robustness across different token budgets, we use multiple pruning ratios $\mathcal{R} = \{0.3, 0.4, 0.5, 0.6, 0.7\}$ during training. For each ratio $r \in \mathcal{R}$, 10,000 cross-boundary patch pairs are randomly sampled from the corresponding retained and discarded sets, where $M = 10,000$ in Eq. 6. This results in a total of 50,000 sampled pairs for the multi-boundary ranking objective. We train the predictor using AdamW with an initial learning rate of 1×10^{-4}, decayed by a factor of 0.5 every three epochs. The training is terminated when the performance on MSLS-val does not improve for 3 epochs. Finally, the actual number of effective epochs is 5, and the total training time is 0.47 hours.

Hardware and Efficiency Evaluation. All training experiments are conducted using PyTorch on a server equipped with four NVIDIA GeForce RTX 4090 GPUs. For efficiency evaluation on the high-performance server, throughput is measured on a single RTX 4090 with a batch size of 32. To further assess practical deployment efficiency on resource-constrained edge hardware, we measure single-query inference latency on an NVIDIA Jetson Xavier NX Developer Kit with 8 GB unified memory running JetPack 5.1.2. A photograph of the device used in our experiments is shown in Fig. 5. The device provides a substantially more constrained deployment environment than the RTX 4090 server, complementing the throughput evaluation with an edge-device latency assessment.



Figure 5: The NVIDIA Jetson Xavier NX used for single-query latency in our experiments.

A.2 Additional Results with Different Pruning Ratios

The main paper reports comparisons under a 50% token pruning ratio. In this section, we further evaluate all methods under pruning ratios of 30%, 40%, 60%, and 70% to examine their robustness across different token budgets. All experiments use the same VPR model (i.e., DINOv2-B + BoQ) with an input resolution of 322×322. For OraP, the same score predictor trained with multi-boundary supervision is directly used across all pruning ratios without ratio-specific retraining. The results are presented in Table 8, and we summarize four key findings as follows.

1) Consistent superiority across different pruning ratios. OraP maintains consistently strong retrieval performance across all evaluated token budgets. At each pruning ratio from 30% to 70%, OraP achieves the best or tied-best R@1 on all six benchmark datasets. At moderate pruning ratios of 30% and 40%, OraP remains particularly close to full-token inference, while consistently outperforming or matching other competing token reduction methods. These results indicate that the learned rankings are not specialized to the 50% setting emphasized in the main paper, but remain effective over a broad range of retained token budgets. To achieve a better balance between retrieval performance and inference efficiency, we mainly adopt a 50% pruning ratio in the main experiments.

Table 8: Retrieval performance under different token pruning ratios across six VPR benchmark datasets. All methods perform token reduction between the MHSA and FFN sublayers of the first transformer block. Full-token inference is included as a reference. The best and second-best results among token reduction methods are highlighted in **red** and **blue** bold, respectively.

Method	Pitts30k-test		MSLS-val		MSLS-test		Tokyo24/7		Nordland		Eynsham	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Full-token Reference, All 529 Patch Tokens (100%)												
Full-token Inference	94.0	97.3	94.9	97.7	76.5	86.2	96.5	98.7	95.2	98.4	92.5	95.9
Retain 370 Patch Tokens (↓ 30%)												
Random	92.2	96.9	93.6	97.4	74.5	85.0	95.9	97.8	91.5	96.7	91.7	95.6
DynamicViT NIPS' 2021	93.4	**97.1**	94.3	**97.6**	**76.2**	**85.9**	**96.2**	**98.4**	**94.8**	**98.3**	92.3	**95.9**
EViT ICLR' 2022	93.1	**97.2**	93.4	**97.6**	74.7	85.0	94.6	97.5	91.9	97.0	90.5	95.1
ToMe ICLR' 2023	**93.5**	**97.2**	94.2	97.4	75.6	85.4	**96.2**	**98.4**	**94.8**	**98.2**	92.0	95.6
ToFu WACV' 2024	**93.7**	**97.1**	**94.6**	**97.7**	76.0	85.8	**96.5**	**98.4**	94.7	**98.2**	92.2	**95.8**
DivPrune CVPR' 2025	**93.7**	**97.2**	93.9	**97.7**	76.0	**85.9**	95.2	97.5	**94.9**	**98.3**	**92.4**	**95.8**
G-Prune AAAI' 2025	92.4	96.8	90.4	96.2	71.7	82.7	95.6	98.1	71.5	84.0	90.9	95.4
AgilePruner ICLR' 2026	**93.5**	**97.1**	94.1	97.3	75.2	85.0	**96.5**	**98.4**	91.7	96.9	92.0	95.7
OraP (Ours)	**93.7**	**97.1**	**94.9**	**97.7**	**76.3**	**86.1**	**96.5**	**98.7**	**94.9**	**98.3**	**92.5**	**95.9**
Retain 317 Patch Tokens (↓ 40%)												
Random	91.7	96.8	93.0	97.3	73.1	84.4	95.2	97.5	90.2	96.3	90.6	95.3
DynamicViT NIPS' 2021	92.6	96.6	**93.9**	97.4	**75.5**	**85.7**	**96.2**	**98.1**	**94.2**	**98.1**	91.9	95.5
EViT ICLR' 2022	92.4	96.9	93.1	97.4	73.2	84.1	93.7	95.6	89.1	95.8	88.5	94.1
ToMe ICLR' 2023	93.0	**97.1**	92.3	97.3	73.4	84.2	92.7	97.1	93.2	97.7	90.1	94.6
ToFu WACV' 2024	**93.3**	**97.1**	93.8	**97.6**	75.1	85.3	**96.2**	**98.1**	93.2	97.5	91.7	95.5
DivPrune CVPR' 2025	**93.5**	96.9	93.8	**97.6**	75.3	85.5	94.3	96.8	94.0	98.0	**92.1**	**95.7**
G-Prune AAAI' 2025	90.4	96.1	89.3	95.4	67.9	79.7	94.0	97.5	51.9	66.5	89.7	94.8
AgilePruner ICLR' 2026	**93.3**	**97.0**	93.2	97.0	74.5	84.7	94.6	97.1	90.3	96.4	91.6	95.5
OraP (Ours)	**93.5**	**97.1**	**94.7**	**97.7**	**76.0**	**85.8**	**96.5**	**98.4**	**94.5**	**98.2**	**92.4**	**95.8**
Retain 211 Patch Tokens (↓ 60%)												
Random	88.5	95.9	89.4	96.4	66.9	80.3	92.1	95.9	76.6	89.0	82.3	91.2
DynamicViT NIPS' 2021	88.7	95.1	**93.6**	**97.2**	**72.6**	**83.9**	**92.4**	**96.5**	**90.9**	**96.5**	**90.1**	**94.7**
EViT ICLR' 2022	89.1	96.1	89.9	96.9	65.3	77.8	78.7	88.3	75.0	87.5	75.6	85.1
ToMe ICLR' 2023	85.8	94.7	83.0	93.9	57.8	72.3	64.1	79.4	74.1	88.4	61.6	76.4
ToFu WACV' 2024	90.8	96.4	92.2	96.8	70.2	81.9	90.2	95.2	85.2	94.0	84.9	92.1
DivPrune CVPR' 2025	91.2	96.2	93.0	**97.2**	72.4	83.2	88.3	93.7	89.9	96.1	89.7	94.6
G-Prune AAAI' 2025	82.9	93.2	83.1	89.7	55.1	67.8	80.6	89.5	18.5	28.2	80.8	89.6
AgilePruner ICLR' 2026	**91.3**	**96.5**	91.9	97.0	69.9	81.4	87.0	92.7	76.5	88.5	87.4	93.2
OraP (Ours)	**92.5**	**96.7**	**93.8**	**97.6**	**75.1**	**85.2**	**95.6**	**97.8**	**91.7**	**97.2**	**91.9**	**95.5**
Retain 158 Patch Tokens (↓ 70%)												
Random	84.3	93.8	86.0	94.2	58.4	72.9	76.8	88.6	52.4	70.9	59.3	73.7
DynamicViT NIPS' 2021	85.2	93.7	**92.2**	**96.8**	**70.0**	**82.3**	**86.7**	**94.0**	**84.8**	**94.2**	**89.1**	**94.0**
EViT ICLR' 2022	82.8	93.3	82.4	92.4	54.9	68.8	61.9	72.4	57.4	73.7	62.1	74.7
ToMe ICLR' 2023	76.0	90.0	75.1	87.3	44.3	60.2	43.5	63.5	49.8	69.5	34.0	51.0
ToFu WACV' 2024	84.5	93.4	85.9	94.3	59.6	73.2	70.2	80.0	63.8	79.6	63.5	76.2
DivPrune CVPR' 2025	88.1	**95.2**	92.0	96.6	68.0	80.0	74.0	88.3	83.6	93.0	86.8	92.2
G-Prune AAAI' 2025	71.9	86.8	70.5	81.8	42.8	56.1	61.9	78.7	9.4	15.3	67.2	79.8
AgilePruner ICLR' 2026	**88.6**	**95.2**	89.5	96.5	63.5	76.2	76.5	87.3	61.7	77.9	81.4	89.4
OraP (Ours)	**91.3**	**96.3**	**93.5**	**97.3**	**73.6**	**84.3**	**93.7**	**96.5**	**87.1**	**95.0**	**90.7**	**95.0**

2) Graceful degradation with increasingly aggressive pruning. As expected, retrieval performance gradually decreases as more patch tokens are removed. Nevertheless, OraP exhibits a relatively smooth degradation relative to the full-token reference. Averaged across the six benchmarks, the absolute R@1 degradation is only 0.1% and 0.3% at pruning ratios of 30% and 40%, respectively, and remains limited to 1.5% when 60% of the patch tokens are removed. Even under the highly aggressive 70% pruning setting, where only 158 of the original 529 patch tokens are retained, the average absolute R@1 degradation is 3.3%. This trend suggests that the token rankings predicted by OraP remain informative even under substantially reduced token budgets. Importantly, the larger performance degradation under aggressive pruning is accompanied by greater computational savings and inference acceleration. These settings therefore provide useful operating points for latency- or resource-constrained applications where efficiency is prioritized over peak retrieval performance.

3) Larger advantages at high pruning ratios. The advantage of OraP becomes increasingly evident as the pruning ratio grows. With fewer tokens retained, more visual information is inevitably

Table 9: Effect of the score predictor architecture. We consistently use the "DINOv2-B + BoQ" model with a 50% token pruning ratio. Starting from a token-wise MLP predictor, we progressively introduce global self-attention and local convolution.

Predictor	Self-Attn	Conv	Pitts30k-test		MSLS-test		Tokyo24/7		Nordland		AmsterTime	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
MLP	✗	✗	91.5	96.2	74.5	85.1	94.3	97.5	92.5	97.4	59.1	81.7
+ Global	✓	✗	92.3	96.5	75.3	85.5	95.6	97.8	93.2	97.6	60.6	82.0
+ Local	✗	✓	92.5	96.9	75.1	85.4	96.2	97.5	93.1	97.6	60.9	82.3
Global-Local (Ours)	✓	✓	**93.3**	**97.0**	**75.7**	**85.6**	**96.5**	**98.4**	**93.8**	**97.8**	**61.7**	**82.5**

discarded, placing greater demands on the quality of token selection within the limited token budget. At a 70% pruning ratio, OraP outperforms the strongest evaluated competing method on each dataset by an average absolute R@1 of 3.1%. The gains are particularly pronounced on challenging benchmark datasets, reaching 3.6% on MSLS-test and 7.0% on Tokyo24/7. On Nordland, despite retaining only 30% of the patch tokens, OraP still achieves 87.1% R@1, compared with 84.8% for the strongest competing method. These results further highlight the importance of reliable token rankings at high pruning ratios, where effectively allocating the limited token budget to informative visual content becomes critical for preserving retrieval performance.

4) Strong candidate retrieval under limited token budgets. Interestingly, OraP maintains highly competitive R@5 even under aggressive pruning. At pruning ratios of 60% and 70%, the average R@5 across the six benchmark datasets remains 95.0% and 94.1%, respectively. In particular, at 70% pruning, OraP still achieves R@5 above 95.0% on five of the six datasets. This suggests that more aggressive token pruning can be attractive in scenarios where coarse retrieval or candidate generation is sufficient, such as systems followed by a subsequent re-ranking stage. In such cases, a larger fraction of tokens can be removed to further reduce computation while still preserving a high probability of retrieving the correct place within the top-ranked candidates.

A.3 Ablation on Score Predictor

We further investigate the design of the score predictor to understand the contribution of global and local contextual modeling. As shown in Table 9, a simple token-wise MLP already provides a reasonable baseline, indicating that early patch representations themselves contain useful information for predicting the late-emerging rankings. Introducing either global self-attention or local convolution consistently improves retrieval performance, suggesting that contextual relationships among tokens provide complementary cues beyond individual token features. Specifically, the global-only and local-only variants improve the average R@1 across the five benchmark datasets by 1.0% and 1.2% over the MLP baseline, respectively. Combining both components yields the strongest performance on all five benchmarks, improving the average R@1 by 1.8% over the MLP baseline. By jointly modeling global and local interactions, our predictor exploits complementary contextual information to more effectively recover late-emerging token rankings from early representations.

A.4 Additional Training Efficiency Analysis

Although OraP requires training an additional score predictor, its training cost remains substantially lower than that of DynamicViT (Rao et al., 2021), which similarly relies on learned prediction modules. As shown in Table 10, the two methods introduce a comparable number of trainable parameters (1.05M vs. 0.96M), yet OraP requires only 5.1 GB of peak GPU memory, compared with 33.6 GB for DynamicViT. Notably, the VPR model is frozen for both methods, and only the prediction modules are optimized. The substantial difference therefore arises from their training pipelines. In OraP, the early representations and final-block supervision are extracted from the frozen backbone, while the ranking loss is applied directly to the score predictor. Consequently, gradient propagation is confined to the predictor. In contrast, the token decisions of DynamicViT affect subsequent transformer blocks, requiring gradients to propagate through the downstream frozen modules to optimize its prediction modules and thereby retaining a substantially larger training graph. Moreover, OraP requires only 5.6 minutes per epoch, corresponding to 20.1% of DynamicViT's 27.8 minutes. The complete training process takes 0.47 hours, compared with 1.85 hours for DynamicViT. These results suggest that OraP introduces only modest training overhead despite learning an additional score predictor.

Table 10: Training efficiency comparison between DynamicViT and OraP. Both methods use the same frozen VPR Models (i.e., DINOv2-B + BoQ) and train only their prediction modules. Due to the hardware memory limit, GPU memory is measured using 60 images per batch.

Method	Trainable Params (M)	GPU Memory (GB)	Time/Epoch (min)	Total Time (h)
DynamicViT	**0.96**	33.6	27.8	1.85
OraP (Ours)	1.05	**5.1**	**5.6**	**0.47**

Table 11: Comparison between token pruning and input resolution reduction under a matched computational budget. Both settings require 24.8G FLOPs. Full-token inference with a resolution of 224×224 processes 256 patch tokens throughout the whole network, whereas OraP retains 238 of 529 patch tokens from 322×322 inputs after 55% pruning between the MHSA and FFN sublayers of the first transformer block. Retrieval performance is reported as R@1 (%).

Setting	Throughput	Pitts30k-test	MSLS-val	MSLS-test	Tokyo24/7	Nordland	AmsterTime
$\text{Full}_{224\times224}$	530.5	92.9	94.1	74.7	95.6	91.0	58.7
$\text{OraP}_{322\times322}$	**569.7**	**93.0**	**94.2**	**75.3**	**95.9**	**93.0**	**61.1**

A.5 Token Pruning vs. Input Resolution Reduction

Reducing the input resolution provides another straightforward way to decrease the computational cost of transformer-based VPR models. We therefore compare OraP at 322×322 resolution with full-token inference at 224×224 under a matched computational budget. As shown in Table 11, OraP retains 238 of the original 529 patch tokens after 55% pruning and requires 24.8G FLOPs, exactly matching the computational cost of full-token inference at 224×224 with 256 patch tokens.

Under this matched computational budget, OraP effectively removes redundant tokens from the higher-resolution input and achieves better performance than low-resolution full-token inference. Specifically, OraP achieves higher R@1 on all six benchmarks, with an average absolute improvement of 0.9% over the 224×224 full-token baseline. The advantage becomes more pronounced on challenging benchmarks, reaching 2.0% on Nordland and 2.4% on AmsterTime. Moreover, OraP achieves a higher throughput of 569.7 img/s, compared with 530.5 img/s for low-resolution full-token inference. Beyond these performance and efficiency gains, the comparison highlights a fundamental distinction between token pruning and input resolution reduction. Reducing the input resolution directly decreases the number of patch tokens and thus lowers computation, but at the cost of globally reduced spatial detail before feature extraction. In contrast, OraP preserves the richer high-resolution input and selectively removes redundant tokens after early contextualization. This allows OraP to better preserve the benefits of high-resolution inference while achieving comparable or even higher efficiency than low-resolution inference. More broadly, token pruning provides a complementary acceleration mechanism without requiring the input resolution itself to be reduced.

A.6 Generalization Across Other Transformer Backbones

The main experiments use DINOv2-B as the backbone to validate the effectiveness of OraP. DINOv2 is a ViT-based vision foundation model pre-trained on the large-scale LVD-142M dataset and has been widely adopted in recent transformer-based VPR systems (Izquierdo & Civera, 2024a; Ali-bey et al., 2024; Lu et al., 2024a; Izquierdo & Civera, 2024b; Jin et al., 2025; Lu et al., 2024c; 2025; Chen et al., 2026a). Nevertheless, OraP is not limited to DINOv2-B. We first extend the evaluation to DINOv2-S and DINOv2-L to investigate its generalization across different model scales within the same family. We further evaluate OraP on the CLIP model (Radford et al., 2021) and a vanilla ViT (Dosovitskiy et al., 2020) to assess its applicability beyond the DINOv2 family.

Generalization across DINOv2 series. As shown in Table 12, OraP consistently achieves the highest R@1 among all evaluated token reduction methods on all six benchmark datasets for both backbone scales. Compared with the strongest competing method on each dataset, OraP provides an average absolute R@1 improvement of 1.4% on DINOv2-S and 1.3% on DINOv2-L. These results show that the proposed ranking-prediction strategy is not specific to the DINOv2-B backbone and remains effective across different model scales within the DINOv2 family.

Table 12: Generalization of OraP across DINOv2 series. All methods are evaluated with 322×322 input, yielding 529 patch tokens, and use a 50% token pruning ratio, retaining 264 patch tokens. The best and second-best results are highlighted in **red** and **blue** bold, respectively.

Backbone	Method	Pitts30k-test		MSLS-val		MSLS-test		Tokyo24/7		Nordland		Eynsham	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
DINOv2-S	Full-token Inference	93.3	96.4	94.2	96.8	73.6	84.3	94.9	97.1	91.8	96.8	91.9	95.4
	DynamicViT NIPS' 2021	88.8	94.7	**92.0**	**96.6**	**70.6**	**82.5**	**91.1**	**94.9**	**87.2**	**94.7**	**90.8**	**94.7**
	EViT ICLR' 2022	88.4	95.0	86.2	93.6	61.8	75.3	76.8	88.9	53.6	70.6	73.6	83.6
	ToFu WACV' 2024	90.0	95.1	91.0	**96.4**	66.9	79.2	87.0	94.3	69.9	83.9	80.2	88.7
	DivPrune CVPR' 2025	90.4	95.3	91.5	**96.4**	70.3	81.8	82.5	91.7	85.4	93.7	90.3	94.4
	AgilePruner ICLR' 2026	**91.1**	**95.8**	91.4	**96.4**	68.9	80.8	85.7	92.1	75.6	88.2	88.5	93.6
	OraP (Ours)	**92.2**	**96.0**	**93.9**	**96.6**	**72.0**	**83.6**	**92.7**	**95.9**	**89.3**	**95.0**	**91.3**	**95.1**
DINOv2-L	Full-token Inference	94.7	97.5	95.5	98.0	78.3	87.4	97.8	98.7	97.5	99.2	92.7	96.1
	DynamicViT NIPS' 2021	91.8	96.6	94.2	97.6	74.2	84.5	**94.9**	**97.5**	**94.2**	**98.1**	**91.5**	**95.5**
	EViT ICLR' 2022	91.9	96.6	93.5	97.6	73.7	84.7	90.2	95.9	89.7	95.9	90.1	94.7
	ToFu WACV' 2024	90.4	95.9	92.3	97.4	70.2	82.0	91.4	95.9	77.8	89.4	78.3	88.3
	DivPrune CVPR' 2025	92.3	**96.8**	**94.6**	97.4	**75.5**	**85.6**	90.8	96.8	92.5	97.3	91.3	95.3
	AgilePruner ICLR' 2026	**92.5**	**96.8**	94.2	**97.7**	74.3	84.9	93.3	**97.5**	89.2	96.3	89.0	94.3
	OraP (Ours)	**93.7**	**97.1**	**95.0**	**97.8**	**77.0**	**86.6**	**97.1**	**98.4**	**95.5**	**98.6**	**92.5**	**95.8**

Table 13: Generalization of OraP on CLIP-ViT-B/16 and ViT-B/16. All methods use 224×224 input images, yielding 196 patch tokens, with a 50% token pruning ratio that retains 98 patch tokens.

Backbone	Method	Pitts30k-test		MSLS-val		MSLS-test		Tokyo24/7		Nordland		Eynsham	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP-ViT-B/16	Full-token Inference	91.5	96.1	94.1	97.3	72.1	81.8	90.2	94.9	85.7	93.3	90.8	94.7
	DynamicViT NIPS' 2021	86.6	93.8	**92.4**	**96.6**	**68.9**	**80.0**	74.6	87.0	**72.8**	84.7	**89.2**	**93.9**
	EViT ICLR' 2022	88.8	95.0	90.3	96.2	66.3	78.1	68.9	82.2	71.6	**84.9**	87.6	93.0
	ToFu WACV' 2024	88.6	94.4	88.8	95.1	65.0	77.2	**79.0**	**89.2**	58.4	75.6	86.4	93.0
	DivPrune CVPR' 2025	89.9	95.2	91.8	96.4	67.0	78.6	71.7	86.7	70.8	84.1	87.1	92.7
	AgilePruner ICLR' 2026	**90.1**	**95.3**	89.9	95.8	66.2	78.2	78.1	88.9	65.7	81.0	87.4	93.0
	OraP (Ours)	**90.3**	**95.5**	**93.4**	**96.9**	**71.0**	**81.0**	**81.9**	**90.2**	**80.5**	**88.9**	**89.7**	**94.0**
ViT-B/16	Full-token Inference	91.3	96.1	93.2	96.9	71.6	81.5	87.3	93.0	83.6	91.0	90.8	94.5
	DynamicViT NIPS' 2021	85.1	92.9	**90.1**	**96.1**	**66.8**	**78.3**	73.3	80.6	**70.5**	**83.1**	**89.0**	**93.2**
	EViT ICLR' 2022	82.5	92.8	70.4	83.5	36.3	50.6	61.9	77.1	7.3	13.2	47.4	63.5
	ToFu WACV' 2024	88.0	**95.1**	87.8	95.1	62.3	74.4	75.2	86.3	54.4	71.0	86.0	92.2
	DivPrune CVPR' 2025	**89.5**	**95.1**	89.2	95.4	65.9	77.4	76.2	**87.3**	64.3	78.8	87.1	92.7
	AgilePruner ICLR' 2026	88.9	95.0	86.6	94.5	60.9	73.2	**76.5**	**87.3**	47.0	65.0	84.0	91.5
	OraP (Ours)	**90.1**	**95.3**	**91.5**	**96.5**	**68.7**	**79.5**	**77.8**	**88.3**	**76.1**	**85.9**	**89.5**	**93.7**

Interestingly, the degradation from full-token inference varies across the DINOv2 series. The average R@1 degradation increases from 0.6% on DINOv2-B to 1.4% on DINOv2-S and 1.0% on DINOv2-L, likely for different reasons. DINOv2-S has the same 12-block depth as DINOv2-B but lower model capacity, which may offer less representational redundancy to compensate for information removed by pruning. In contrast, DINOv2-L contains 24 transformer blocks. Since OraP always prunes within the first block, the pruning location is relatively earlier with respect to the overall network depth, and the predictor must anticipate the final-block rankings that emerge after substantially more feature transformations and token interactions. In addition, discarded tokens are excluded from a longer sequence of subsequent blocks, potentially amplifying the impact of early information removal. Overall, these observations suggest that the robustness of early token pruning depends not only on model scale, but also on representational redundancy and the relative pruning depth. Nevertheless, OraP still provides substantial computational savings and improved efficiency while keeping retrieval degradation limited across different DINOv2 scales.

Generalization beyond the DINOv2 family. We further evaluate OraP on CLIP-ViT-B/16 and vanilla ViT-B/16 to examine its generalization to other transformer backbones. As shown in Table 13, OraP consistently achieves the highest R@1 among all evaluated token reduction methods on all six benchmarks for both backbones. Compared with the strongest competing method on each dataset, OraP provides an average absolute R@1 improvement of 2.4% on CLIP-ViT-B/16 and 1.9% on ViT-B/16. The advantage is particularly evident on Nordland, where OraP surpasses the strongest competing method by 7.7% and 5.6% R@1, respectively. These results demonstrate that the effectiveness of predicting late-emerging token rankings extends beyond the DINOv2 family.

Compared with the DINOv2-B backbone, however, the degradation from full-token inference becomes more noticeable, increasing from 0.6% to 2.9% on CLIP-ViT-B/16 and 4.0% on ViT-B/16. This difference may stem from both the reduced token redundancy and the representation robustness of the underlying backbones. Under the 224×224 input setting, CLIP-ViT and ViT contain only 196 patch tokens and retain 98 after 50% reduction, substantially fewer than the 264 retained tokens in our DINOv2 experiments at 322×322. Removing half of this smaller token set therefore introduces a stronger information bottleneck and leaves less redundancy for the remaining tokens to compensate for discarded visual cues. Meanwhile, compared with DINOv2-B, these backbones provide different levels of visual representation robustness, which may make them less tolerant to early information removal. The degradation is particularly evident on Tokyo24/7 and Nordland, where retrieval requires robustness to severe viewpoint, illumination, or seasonal variations and often depends on complementary structural cues distributed across the scene. Pruning within the first transformer block may remove part of these cues before sufficient interaction occurs in later blocks. Overall, these results suggest that the tolerance to early token pruning depends jointly on backbone representation robustness and available token redundancy. Nevertheless, OraP remains consistently more robust than competing token reduction methods in these challenging settings.

A.7 Additional Qualitative Results and Failure Cases Analysis

We provide additional qualitative comparisons across diverse VPR scenarios as a supplement to Fig. 4 to further examine how different token reduction methods preserve place-relevant visual information. Unless otherwise specified, each example consists of two rows. The upper row shows the query image followed by the retained-token visualizations of different methods under 50% token pruning, while the lower row shows the corresponding top-1 retrieval results. Green and red borders indicate correct and incorrect retrievals, respectively, and highlighted regions denote representative place-discriminative structures discussed in the corresponding captions. These visualizations provide an intuitive understanding of how token selection affects retrieval performance.

Additional Qualitative Results. As shown in Fig. 6, Fig. 7, Fig. 8, and Fig. 9, across the Pitts30k, MSLS, Nordland, and AmsterTime datasets, OraP generally preserves more coherent and place-discriminative visual structures than competing token reduction methods. The retained regions often correspond to stable architectural elements, scene layouts, or environmental structures that remain informative despite large viewpoint, seasonal, illumination, or cross-domain appearance changes. In contrast, competing methods more frequently produce fragmented token selections or discard parts of these informative regions, which often coincide with incorrect retrievals. *Notably, the examples from Nordland also show that successful retrieval does not require different methods to retain exactly the same token subset. Multiple selections can lead to the same correct match as long as sufficient place-discriminative information is preserved.* The examples from AmsterTime further demonstrate that reliable token selection remains important under severe modality and long-term appearance changes, where historical grayscale queries must be matched against contemporary RGB references.

Failure Case Analysis. We further examine some corner cases in which all evaluated token pruning methods fail, as presented in Fig. 10. These examples reveal several common challenges of token pruning for VPR. First, scenes dominated by visually similar architectural structures contain relatively few obviously redundant regions, making localization highly dependent on subtle facade details or small distinctive landmarks. Second, some scenes lack a single dominant landmark and instead encode place identity through multiple weak and spatially distributed cues, whose collective information can be disrupted by aggressive token removal. Natural environments pose a related challenge, as retrieval may depend more heavily on global terrain, route geometry, and overall spatial layout than on isolated local regions. Finally, severe domain shifts, such as historical grayscale imagery versus contemporary RGB observations, can weaken the correspondence of otherwise discriminative regions and make reliable token selection more difficult. Overall, these cases indicate that token pruning remains particularly challenging when place-discriminative information is subtle, distributed across the scene, or strongly affected by appearance and domain changes. *In such extremely challenging scenarios, a more conservative pruning ratio (e.g., retaining 70% of patch tokens) can provide additional information capacity to preserve weak but complementary cues. Moreover, exploring effective adaptive pruning strategies that dynamically adjust the token budget according to scene complexity may further improve robustness in these challenging conditions.*

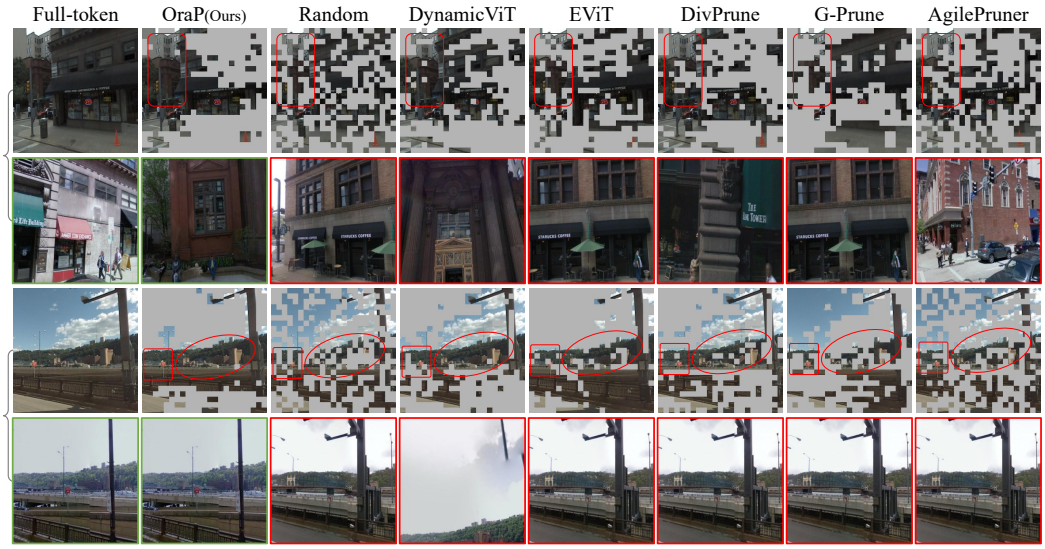


Figure 6: Additional qualitative results on Pitts30k-test. In the first example, OraP preserves a coherent vertical facade and entrance structure that remains discriminative across viewpoint changes, whereas competing methods retain more fragmented or incomplete regions, resulting in incorrect retrievals. In the second example, OraP better preserves the distant building structures and rail-side scene layout shared across the query and reference views, while competing methods discard or fragment these cues under large viewpoint variation. Overall, these cases illustrate the advantage of preserving spatially coherent structural cues for robust retrieval under viewpoint changes.

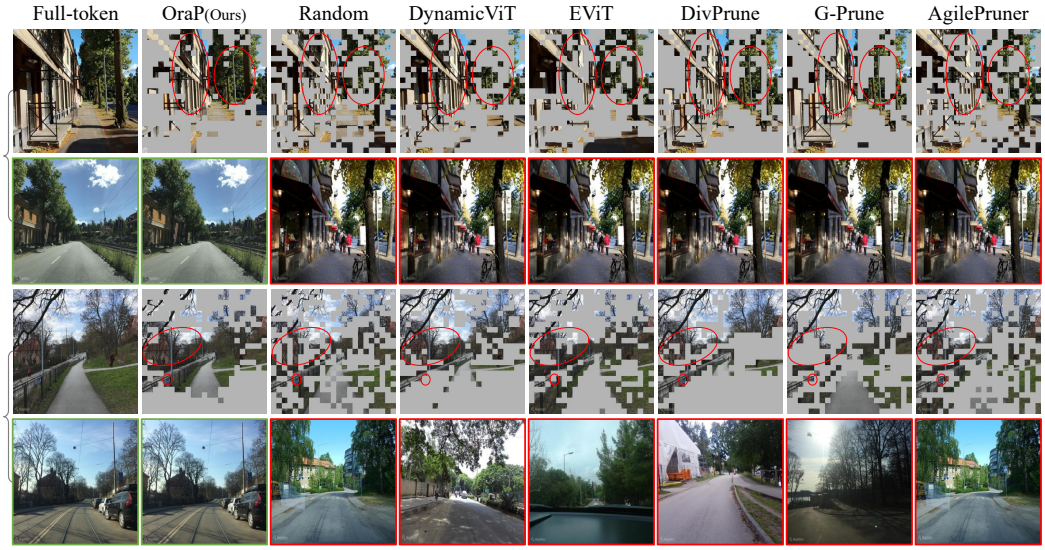


Figure 7: Additional qualitative results on the MSLS dataset. In the first example, OraP preserves both the distinctive facade structures along the left side of the street and the vegetation-related cues on the opposite side, whereas competing methods tend to retain these regions less completely or introduce more fragmented selections. This more coherent preservation is accompanied by correct retrieval despite substantial viewpoint and appearance variations. In the second example, the discriminative structures occupy only a relatively small portion of the query image. OraP retains the distant building-related structures and several localized scene cues more completely, while competing methods frequently discard or fragment them, resulting in incorrect retrievals. These examples highlight the ability of OraP to preserve complementary and spatially localized cues that remain informative under challenging urban appearance changes.

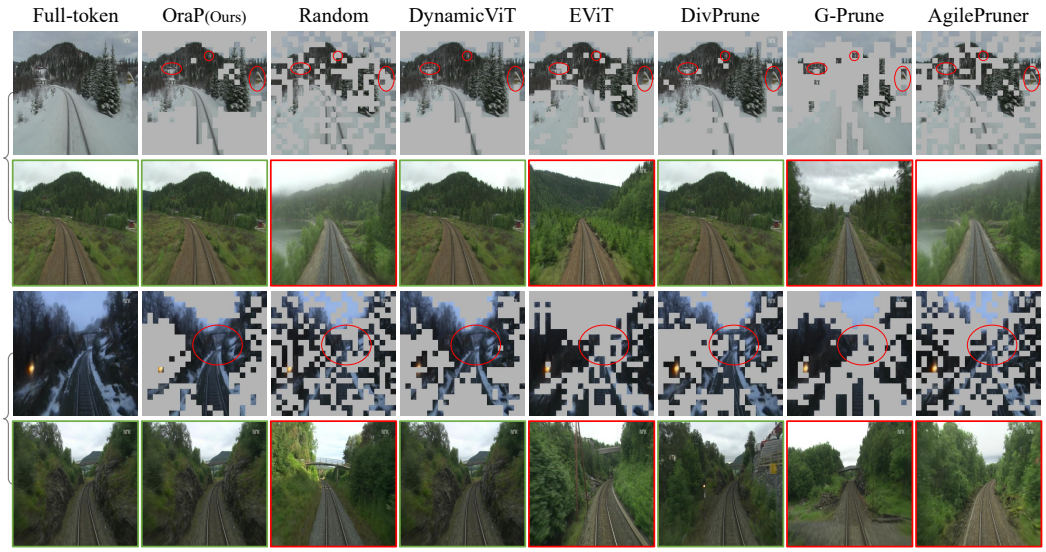


Figure 8: Additional qualitative results on the Nordland dataset. In the first example, OraP, DynamicViT, and DivPrune all produce correct retrievals, although the retained token subsets are not identical. These methods preserve sufficient discriminative cues around the railway, mountain contours, and surrounding vegetation, while other methods discard or fragment more of these regions and lead to incorrect matches. A similar pattern appears in the second example: OraP, DynamicViT, and DivPrune retain different subsets of patches, yet each preserves enough stable structural and environmental cues to support the correct retrieval under severe seasonal and illumination changes. Overall, these cases suggest that successful VPR does not require an identical token subset, but rather sufficient preservation of place-discriminative information.

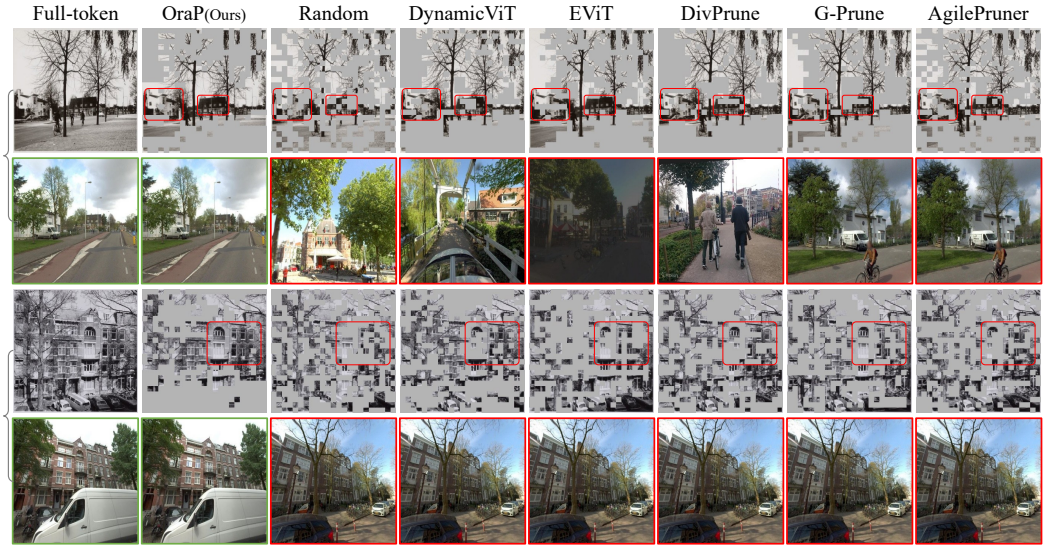


Figure 9: Additional qualitative results on the AmsterTime dataset. The query images are historical grayscale photographs, while the reference database consists of contemporary RGB images, making retrieval more challenging due to the large cross-domain appearance gap. In the first example, OraP preserves the most place-discriminative structures and successfully retrieves the correct location, whereas the competing methods are more easily distracted by visually similar but irrelevant regions. The second example is particularly challenging because strong perceptual aliasing is present. OraP retains the key discriminative building region and still produces the correct retrieval, while the competing methods tend to preserve less informative or misleading regions and therefore retrieve incorrect matches. These examples further show that OraP can preserve robust place-discriminative cues even under severe modality and appearance changes.



Figure 10: Failure cases. Cases 1 and 2 are dominated by visually similar architectural structures, with few obviously redundant regions. Therefore, successful localization tends to depend on subtle facade details and small distinctive landmarks that may be removed in error during pruning. Case 3 contains no dominant landmark, and place identity is instead distributed across multiple weak cues such as road geometry, roadside structures, and vegetation layout, making token removal more likely to disrupt their collective spatial evidence. Case 4 illustrates a natural scene in which retrieval relies heavily on global terrain and railway layout rather than isolated local landmarks, while substantial seasonal appearance changes further increase the difficulty. Case 5 exhibits a severe cross-domain gap between a historical grayscale query and contemporary RGB references, which weakens the correspondence of discriminative architectural regions and increases the difficulty of preserving the most informative tokens. Overall, these cases suggest that token pruning becomes particularly challenging when discriminative information is subtle, spatially distributed, or strongly affected by domain changes, rather than concentrated in a small number of stable visual regions.

A.8 More Details about Datasets

Pitts30k-test (Torii et al., 2013) is collected from Google Street View panoramas with geographic annotations. The dataset contains images sampled from various locations in Pittsburgh, where the same place can be observed from different viewpoints and under diverse appearance conditions. Due to substantial viewpoint changes and environmental variations, Pitts30k-test has become a widely used benchmark for evaluating VPR methods in urban scenarios. Moreover, Pitts30k-test is a carefully selected subset of Pitts250k-test, providing a more challenging evaluation setting with fewer reference images and larger retrieval difficulties.

MSLS (Mapillary Street-Level Sequences) (Warburg et al., 2020) is a large-scale VPR dataset containing over 1.6 million street-level images collected from 30 cities across six continents. It covers diverse environments, including urban, suburban, and natural scenes, and provides geographic coordinates and camera orientation information for each image. The dataset exhibits substantial appearance variations caused by viewpoint changes, illumination differences, weather conditions, seasonal changes, and dynamic objects. MSLS is originally divided into three subsets: training, public validation (MSLS-val), and withheld test (MSLS-test/challenge). Originally, MSLS-test/challenge was evaluated through an official online evaluation server, where participants submitted retrieval results without access to the test ground truth. Since the original evaluation server (Codalab) is no longer available and the test ground truth has been publicly released, we directly evaluate all query images using the publicly available ground truth to compute retrieval metrics. Therefore, our MSLS-test results may differ from previous reports obtained under the original online evaluation protocol.

Tokyo24/7 (Torii et al., 2015) is a challenging urban VPR benchmark collected in Tokyo, containing 75,984 database images and 315 query images. The query set is selected from 1,125 candidate images captured at 125 different locations, where each place is observed from multiple viewpoints and at different times of the day. The dataset is specifically designed to evaluate robustness against severe appearance variations, including viewpoint changes and significant illumination differences, particularly day-to-night transitions.

Nordland (Olid et al., 2018) consists of image sequences captured from a fixed camera mounted at the front of a train across four seasons. Unlike typical VPR datasets that involve viewpoint changes, Nordland focuses on extreme appearance variations caused by seasonal and illumination changes while maintaining consistent viewpoints. The dataset mainly contains suburban and natural scenes, and provides frame-level correspondences as ground-truth information. Following previous works (Lu et al., 2024a; Ali-bey et al., 2024; Lu et al., 2026), we uniformly sample images at 1 FPS and use winter sequences as queries and summer sequences as database images for evaluation.

Eynsham (Cummins & Newman, 2010) contains grayscale image sequences captured by cameras mounted on a vehicle traversing the same route twice around Oxford, covering both urban and rural environments. We use images from the first traversal as the database and those from the second traversal as queries. The original equirectangular panoramas are further divided into five perspective crops, which are used as individual images for VPR evaluation.

Pitts250k-test (Torii et al., 2013) is the test split of the Pittsburgh 250k benchmark, containing a large collection of geo-referenced street-view images acquired from Google Street View in Pittsburgh and its surrounding areas. Compared with Pitts30k-test, it provides a substantially larger database with broader spatial coverage, enabling the evaluation of VPR methods under large-scale urban retrieval scenarios. The dataset includes diverse urban environments and appearance variations, making it a challenging benchmark for assessing the robustness of VPR systems.

SPED (Chen et al., 2017a) is a long-term VPR dataset collected from fixed surveillance cameras monitoring different locations over extended periods. The dataset contains image sequences captured under diverse environmental conditions, exhibiting substantial appearance variations caused by changes in illumination, weather, and seasons. Each place is associated with hundreds of images collected at different times of the day and year, providing a challenging benchmark for evaluating VPR methods under long-term appearance changes.

AmsterTime (Yildiz et al., 2022) contains over one thousand query-reference image pairs collected from Amsterdam. Each pair consists of a historical grayscale image as the query and a contemporary RGB image from the same location as the reference, where the correspondence is manually verified by human experts. The dataset introduces extreme long-term temporal changes and sub-

stantial domain gaps, including variations in viewpoint and visual modality, making it a challenging benchmark for evaluating VPR methods under historical-to-modern place recognition scenarios.

SVOX (Berton et al., 2021) is a cross-domain VPR benchmark designed to evaluate place recognition under substantial appearance and domain variations. Its database consists of large-scale Google Street View imagery covering the city of Oxford, while the query images are obtained from the Oxford RobotCar dataset (Maddern et al., 2017). The query set is further divided according to different conditions, including variations in weather and illumination. In our experiments, we report results on the three most challenging subsets, namely SVOX-Night, SVOX-Rain, and SVOX-Sun.

A.9 Compared Methods Details

Random Pruning directly retains K patch tokens from the original N tokens with equal sampling probability, without estimating their relevance or redundancy. As a criterion-free baseline, it provides a reference for assessing the benefit brought by more deliberate token selection strategies.

DynamicViT (Rao et al., 2021) introduces lightweight prediction modules at multiple transformer blocks to estimate token importance and progressively sparsify less informative tokens. During training, binary token decisions are sampled with Gumbel-Softmax, while an attention-masking mechanism suppresses interactions involving discarded tokens to enable differentiable optimization. The original full-token backbone further serves as a teacher to regularize the pruned model through token-level feature distillation and prediction-level KL divergence. During inference, tokens with low predicted keeping probabilities are directly removed, reducing the computation of subsequent transformer blocks.

EViT (Liang et al., 2022) evaluates patch-token significance using the class-to-patch attention scores. Tokens receiving larger attention weights from the class token are preserved, while the remaining tokens are aggregated into a single token through attention-weighted fusion. The retained tokens together with the fused token are then passed to the subsequent transformer blocks.

ToMe (Bolya et al., 2023) reduces the token sequence through merging instead of discarding tokens. It employs bipartite soft matching to identify highly similar token pairs and combines each matched pair into a single representative token. By progressively consolidating redundant tokens, ToMe shortens the sequence while retaining information from the merged tokens for subsequent transformer layers.

ToFu (Kim et al., 2024) combines token pruning and merging within a unified reduction framework. It identifies similar token pairs through bipartite soft matching and applies different reduction operations according to network depth. Specifically, matched tokens are pruned in earlier layers, whereas average merging or MLERP is employed in later layers to form representative tokens.

DivPrune (Alvar et al., 2025) performs token selection by explicitly promoting feature diversity among the retained tokens. It formulates the selection process as a maximum-diversity problem, favoring tokens that are mutually dissimilar in the feature space. By avoiding selecting highly redundant tokens, DivPrune aims to preserve a more diverse subset of visual information.

G-Prune (Jiang et al., 2025) formulates token selection from a graph perspective, treating visual tokens as nodes and their pairwise feature similarities as edges. Token importance is estimated through information propagation over the resulting graph, allowing each token to incorporate relationships with other tokens. The top-K tokens according to the propagated importance scores are then retained for subsequent computation.

AgilePruner (Baek et al., 2026) combines attention- and diversity-aware token selection in an image-adaptive manner. Motivated by the observation that attention-based pruning is more effective for simple images whereas diversity becomes more important for visually complex images, it measures the feature diversity of each input using effective rank. Tokens are first ranked by their attention scores, after which highly similar candidates are progressively removed according to an adaptive similarity threshold determined by the input effective rank. In this way, the method relies more on attention-based rankings for visually simple inputs, while placing greater emphasis on diversity-aware selection for visually complex ones.

A.10 Limitations and Future Work

In this paper, the proposed OraP provides an effective and practical approach to early token pruning, achieving substantial computational savings while maintaining excellent retrieval performance across diverse VPR benchmarks and architectures. Despite these advantages, several limitations remain and motivate our future research. **1) Intrinsic gap of oracle-guided early pruning.** Although the oracle study shows that late-emerging token rankings can enable nearly lossless early pruning in VPR models built upon foundation models such as DINOv2, a small performance gap from full-token inference remains. This indicates that even if OraP perfectly recovers the oracle rankings, fully lossless early pruning cannot be guaranteed. This gap arises because the ranking induced by the final-block attention is derived from token representations refined through repeated self-attention interactions (i.e., across multiple transformer blocks) over the complete token sequence, whereas early pruning alters subsequent information propagation and removes the potential contribution of discarded tokens to later representations. *Future work may therefore explore mechanisms that compensate for the information carried by discarded tokens.* Moreover, our generalization experiments show that this gap becomes larger on vanilla ViT and CLIP backbones, indicating that the effectiveness of early pruning also depends on backbone robustness to token removal. *However, this also suggests that, as visual foundation models continue to advance, such robustness may further increase, potentially allowing OraP to operate with smaller token budgets while preserving higher retrieval performance and achieving greater efficiency gains.* **2) Compatibility with downstream aggregation modules.** OraP performs structural token pruning and therefore requires downstream modules to support variable-length token sequences. While this property is compatible with a wide range of dominant aggregation mechanisms such as GeM, NetVLAD, SALAD, BoQ, and ImAge, some VPR models explicitly depend on a fixed number of input tokens. For example, the token-mixing layers in MixVPR (Ali-bey et al., 2023) operate along a predefined length of token sequence, making direct integration with OraP infeasible without modifying or retraining the aggregation module. *Extending token pruning to such fixed-token architectures, potentially through adaptive aggregation or token restoration mechanisms, is another promising direction for future work.*