EfficientVPR: Toward Efficient Visual Place Recognition via Scene-Aware Prompt Tuning and Adaptive Feature Enhancement

Supplementary Material

7. Outline

This supplementary material contains additional experimental details, extended results, and further analysis. Specifically:

- In Sec.8, we provide a cross-scale comparison with SOTA methods under unified dimensions, along with additional qualitative comparisons on challenging cases.
- In Sec.9, we extend the main paper with the following analyses: for SceneVPT, we provide an ablation study on prompt number and additional visualizations; for instance-dependent key local feature enhancement module, we provide an ablation study on OLE.
- In Sec.10, we present more information about testing benchmarks, covering key aspects such as their composition, characteristics, and sources.
- In Sec.11, we offer more details about the comparison methods, including principles and implementation details.
- In Sec.12, we provide a theoretical comparison with FoL to further elucidate the rationale behind our approach.
- In Sec.13, we present more details of other compared prompt tuning methods in Sec.4.4.1.

8. Additional Comprehensive Comparisons with State-Of-The-Art Methods

8.1. Comparisons against SOTA Methods across Different Model Scales with Same Feature Dimensions

While in main paper we have already presented a cross-scale comparison using native feature dimensions, in this section we complement it with an analysis under unified dimensionality. We employ PCA, a standard dimensionality reduction technique in VPR, for this unification. We follow the specific implementation of CricaVPR [33], which also use PCA for dimensionality reduction. Since PCA only introduces additional training and inference time without altering the model parameter count or peak GPU memory usage, we focus our comparison solely on accuracy metrics. Tab.9 presents a cross-model-scale comparison between our method and existing one-stage SOTA methods at a unified feature dimension of 3456. To eliminate potential bias from PCA, we also conduct an additional comparison at 2048 dimensions, where our method also utilized PCA for dimensionality reduction.

A comprehensive analysis of Tab. 9 and 10 leads to the following conclusions:

- Our EfficientVPR achieves superior R@1 using a smaller backbone than the best competitor with a larger model. When the feature dimension is unified to 3456, our method achieves a $+0.7\%$ higher Avg. Acc. R@1 than the sub-optimal method BoQ, despite using a smaller backbone. At the unified dimension of 2048, our method also outperforms the sub-optimal method SALAD by 0.3% in Avg. Acc. R@1, again using a smaller model.
- On the challenging AmsterTime dataset (with significant domain and viewpoint changes), our method outperforms the suboptimal DINOv2-B based SALAD by 4.7% and 2.6% in R@1 at 3456 and 2048 dimensions, respectively. This shows the robustness of our method to both viewpoint and domain changes.
- On the diverse MSLS-val dataset, our method achieves a 0.4% R@1 improvement over the suboptimal DINOv2-B based BoQ at both 3456 and 2048 dimensions. Against BoQ at the same model scale, the gains are even more substantial: $+1.6\%$ R@1 and $+0.8\%$ R@5 at 3456-D; and $+1.8\%$ R@1 and $+1.3\%$ R@5 at 2048-D.
- Our method demonstrates a clear advantage over same-scale competitors on SVOX-night, achieving a 1.2%-1.4% R@1 improvement over the suboptimal method, although it does not exceed larger DINOv2-B-based models. This dataset proves highly sensitive to backbone scale, as evidenced by the drastic performance decrease (3.4%-7.5%) for SALAD and BoQ with a reduced backbone.

8.2. Additional Qualitative Comparison on Challenging Cases

In this subsection, we present further qualitative comparisons on challenging cases, particularly scenes containing non-overlapping regions.

Fig.8 shows challenging cases under significant viewpoint changes across suburban and urban environments. These viewpoint changes result in non-overlapping regions between the query and database images. In the first row, under combined viewpoint and rotation changes, key structures like right-side buildings are absent from the query, while transient objects like trees appear only in the query due to time shift. Other one-stage methods failed by matching trees while overlooking the missing building. In contrast, our method prioritized the persistent structural features to achieve a correct match. In the second row, images are from suburban, where vegetation transitions from being a distractor to a crucial localization cue. While all

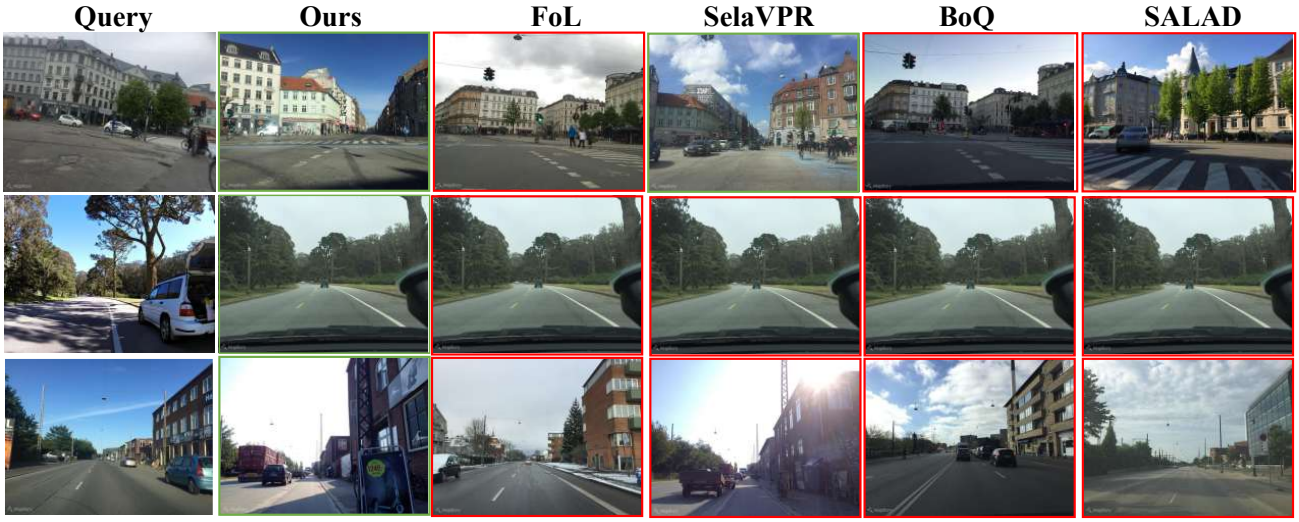

Figure 8. **Qualitative comparison on challenging cases under various viewpoint changes**. Green boxes: correct matches. Red boxes: errors. Viewpoint changes with rotation (1th row), viewpoint changes coupled with dynamic occlusions in suburban(2nd row), viewpoint changes coupled with dynamic occlusions in urban (3rd row). Under significant viewpoint shifts and temporal occlusions, competing methods consistently return images with transient elements (e.g., trees) or non-discriminative structures (incorrect matches). Unaffected by these factors, our method consistently retrieving the correct reference image from the same location.

Method	Back -bone	Pitts250k -test		MSLS -val		Amster Time		Eynsham		SVOX Night		SVOX Overcast		SVOX Snow		Avg. Acc.	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
SALAD	B	95.0	98.5	91.5	96.4	55.5	78.2	91.6	95.0	95.5	**99.1**	**98.4**	99.3	98.5	99.5	89.4	95.1
BoQ	B	94.5	98.0	92.4	95.8	55.2	74.2	91.4	94.9	**96.6**	98.8	**98.4**	**99.7**	99.0	**99.7**	89.6	94.4
SALAD	S	94.0	98.2	89.6	95.4	54.1	75.2	90.6	94.4	88.7	96.2	97.7	98.6	98.4	99.1	87.6	93.9
BoQ	S	94.6	98.3	91.2	95.8	52.2	73.5	90.9	94.5	93.2	97.6	97.9	99.2	99.0	99.6	88.4	94.1
EfficientVPR (Ours)	S	**95.4**	**98.8**	**92.8**	**96.6**	**60.2**	**78.4**	**92.0**	**95.3**	94.4	97.9	**98.4**	99.4	**99.1**	**99.7**	**90.3**	**95.2**

Table 9. **Comparison with SOTA methods across different model scales with** 3456 **feature dimensions on seven benchmarks.** B: DINOv2-base, S: DINOv2-small. The best results are highlighted in **bold** and the second best are underlined.

comparative methods fail by retrieving images from adjacent locations, our approach consistently identifies the correct same-site matches. In the third row, our method yields reference image with multiple temporary interfering objects from exactly the same location. The other two one-stage methods returned reference images of other locations with similar distractors (i.e., chimney, trees, cars) but different main buildings, and the two-stage methods likewise failed to handle this dynamic occlusion scenario. These examples indicate that our proposed one-stage method EfficientVPR can better avoid being affected by non-overlapping regions, and thus more robust to diverse viewpoint changes and dynamic occlusions across various scenes.

Furthermore, Fig.9 extends the scenario from Fig.8 though domain shift, increasing the difficulty. Under domain shift, the non-overlapping regions become more misleading. In the first row, our method successfully retrieves the correct image containing the main building with non-overlapping area, while all other methods fail. The two-stage methods find images from nearby but not same locations that contained non-overlapping regions, while the one-stage methods retrieve images of a dominant but entirely different building. In the second row, under extreme illumination variations, our method still retrieves the correct match with non-overlapping regions. In the third row, under the dual influence of domain and viewpoint shift, only our method succeeds in matching the correct location despite occlusions. All other methods fail, returning incorrect images of a different building that merely share some similar features like windows. In the fourth row, with the absence of color, vegetation features in natural scenes become ambiguous. It causes SelaVPR, BoQ, and SALAD to retrieve completely incorrect locations with dissimilar distant vegetation. FoL finds a visually similar yet merely a nearby one. Only our method achieves the correct match. Only our method secures the correct match. The ability to effectively

Method	Back -bone	Pitts250k -test		MSLS -val		Amster Time		Eynsham		SVOX Night		SVOX Overcast		SVOX Snow		Avg. Acc.	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
SALAD	B	**94.8**	98.4	91.4	**96.4**	55.9	77.2	91.5	95.1	95.0	98.2	**98.2**	99.1	98.5	99.4	89.3	94.8
BoQ	B	94.3	97.8	91.8	95.5	51.1	69.6	91.2	94.8	**95.7**	**98.4**	97.8	99.1	**98.7**	**99.7**	88.7	93.6
SALAD	S	93.9	98.2	89.5	95.0	53.5	74.5	90.6	94.4	87.5	96.0	97.6	98.6	98.3	99.2	87.3	93.7
BoQ	S	93.0	97.8	90.4	95.1	48.8	69.0	90.7	94.4	92.0	96.5	97.5	98.6	98.6	99.2	87.3	92.9
EfficientVPR (Ours)	S	**94.8**	**98.5**	**92.2**	**96.4**	**58.5**	**77.6**	**91.7**	**95.2**	93.4	97.8	**98.2**	**99.2**	**98.7**	**99.7**	**89.6**	**94.9**

Table 10. **Comparison with SOTA methods across different model scales with** 2048 **feature dimensions on seven benchmarks.** B: DINOv2-base, S: DINOv2-small. The best results are highlighted in **bold** and the second best are underlined.

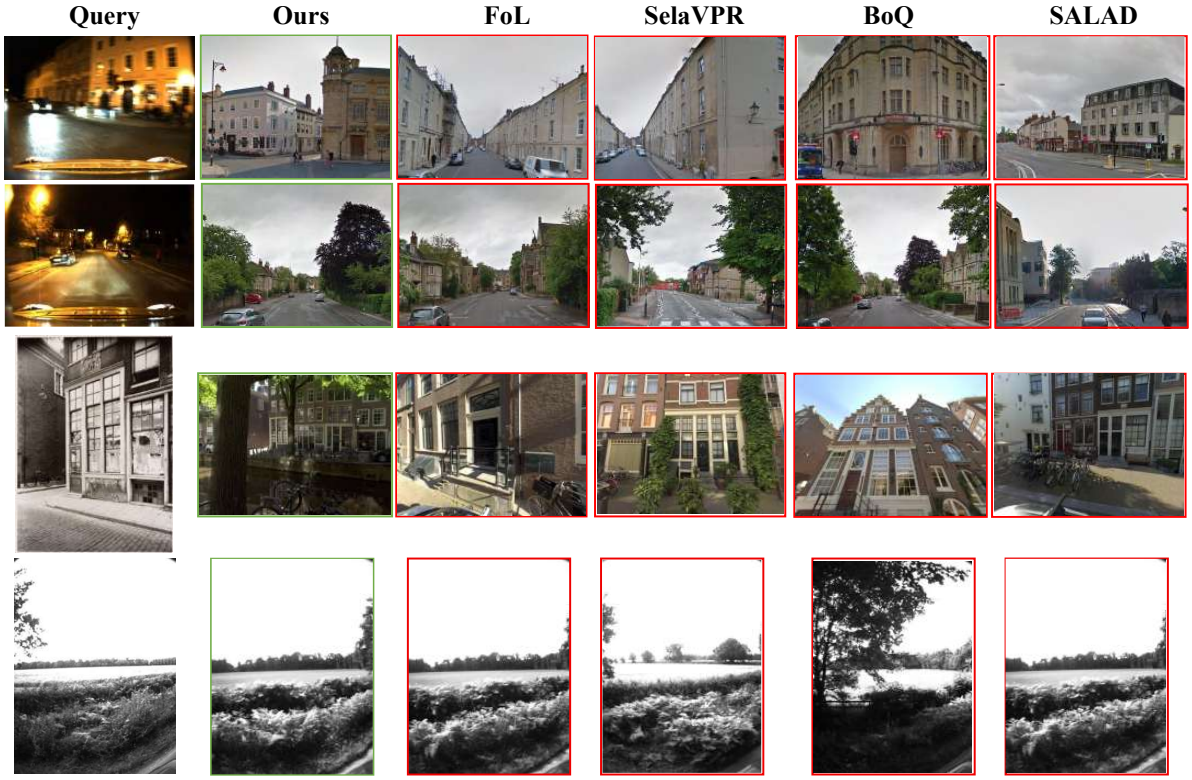

Figure 9. **Qualitative comparison on challenging cases under domain shift**. Green boxes: correct matches. Red boxes: errors. Drastic viewpoint changes (1th row), viewpoint changes coupled with time shift (2nd row), viewpoint changes coupled with dynamic occlusion in urban (3rd row), viewpoint changes in natural (4th row). In the first and third row, our method successfully retrieves the correct image containing the main building with non-overlapping area, while all other methods fail. In the second and fourth row, under conditions of extreme illumination change and color loss, our method correctly handles the viewpoint variations and returns accurate matches, while all other methods fail.

handle such a wide range of challenging factors, showcases the versatility and effectiveness of our proposed method in real-world applications where environmental conditions can vary drastically.

9. Additional Ablation Studies

9.1. Ablation on Prompt Number

Tab.11 shows the impact of the number of prompts on model performance. It can be seen that increasing the number of prompts up to 32 leads to varying degrees of performance improvement across various datasets. Compared to using only 8 prompts, fine-tuning the backbone with 16 prompts delivers a significant performance boost, achieving a 0.7% gain in R@1 on both Pitts250k and AmsterTime. When the prompt number is further increased to 32, it yielded an additional R@1 gain of 0.3%, 0.7%, and 0.3% on Pitts250k, AmsterTime, and SVOX-Overcast, respectively. However, further increasing the number of prompts

to 64 does not yield additional performance gains but instead leads to a slight degradation in R@1 on both Pitts250k and AmsterTime. This suggests that excessive prompts can adversely affect the fine-tuning of the backbone. Therefore, we opt to use 32 prompts for backbone fine-tuning.

Num.	#Params.	Pitts250k -test		Amster Time		SVOX Overcast	
		R@1	R@5	R@1	R@5	R@1	R@5
8	0.04	94.0	98.2	53.8	75.8	97.5	**99.3**
16	0.07	94.7	**98.5**	54.5	75.5	97.6	99.2
32	0.15	**95.0**	98.4	**55.2**	**77.0**	**97.9**	99.2
64	0.29	94.9	**98.5**	54.6	75.7	**97.9**	99.0

Table 11. **Ablation on the number of prompts**. We adjust the number of prompts used in SceneVPT and record the corresponding fine-tuning parameter count along with R@1 and R@5 over Pitts250k-test, AmsterTime and SVOX-overcast. To ensure consistency with the ablation studies in the main paper, all test images are resized to 266×266.

9.2. Additional Visualization and Analysis of Fine-tuning Methods

This subsection presents further visualizations and analysis of the fine-tuning methods, building upon the main paper. Additionally, it provides an enlarged version of Fig. 6 from the main paper to facilitate a more detailed examination of the visual results.

The heatmaps are generated using the same methodology as in the main paper. Specifically, they are mean attention maps obtained from the final layer of the backbone network by averaging across all attention heads. Tab.10 shows additional visualization cases. Specifically, in the first query, street lamps and signs shift from common distractions to the core discriminative features. While both VPT-shallow and VPT-deep reduce focus on the sky compared to the frozen backbone, they still over-attend to clouds and fail to capture critical elements like distant buildings, lamps, and road sign. The adapter method, though reducing attention on left-side clouds, over-focuses on the right and incompletely recognizes the sign, again missing the distant architecture. Consequently, these methods produced incorrect matches. In contrast, our method effectively identifies and prioritizes the discriminative features while suppressing attention on irrelevant sky regions, thus achieving a correct match. For the second query, fine-tuning with VPT-shallow draws significantly more attention to the main building than the frozen backbone, yet remains overly distracted by the clouds, indicating insufficient adaptation. VPT-deep, while markedly reducing focus on the sky compared to VPT-shallow, still attends to it more than our SceneVPT. Furthermore, it only partially captures the building's structure, emphasizing the mid-lower section while neglecting upper details like the glass facade, leading to an incorrect match.

Version	Pitts250k -test		Amster Time		SVOX Overcast	
	R@1	R@5	R@1	R@5	R@1	R@5
concat	94.6	98.1	53.2	75.6	97.1	98.7
nonlinear orthogonal fusion mechanism	**95.0**	**98.4**	**55.2**	**77.0**	**97.9**	**99.2**

Table 12. **Ablation on OLE.** "nonlinear orthogonal fusion mechanism" refers to our proposed OLE, i.e. the variant that contains nonlinear orthogonal fusion mechanism. "concat" means using direct feature concatenation in place of nonlinear orthogonal fusion mechanism. All test images are resized to 266×266.

The adapter-based method heavily focuses on the clouds and critically ignores the left side of the building, which results in a retrieval with a structurally dissimilar main structure.

9.3. Ablation on OLE

We have verified the OLE module in Sec.3.2.2. In this subsection, we further explore its nonlinear orthogonal fusion mechanism to analyze its specific contribution to the performance. Specifically, we replace the nonlinear orthogonal fusion mechanism with a direct concatenation of features to investigate its impact. In Tab.12, when employing direct feature concatenation, the model exhibits performance degradation across all datasets. This is particularly noticeable on AmsterTime, where R@1 and R@5 drop by 2.0% and 1.4%, respectively. It indicates that the simple aggregation of directional information compromises the discriminative power of the features.

10. Datasets Details

GSV-Cities [5] provides $570,000$ training images collected from $67,000$ locations, covering viewpoint variations and seasonal changes, but excluding other condition variations such as illumination. Therefore, using only GSV-Cities for training allows evaluating the model's generalization ability on datasets containing other types of condition variations.

Pitts250k [41] features urban street-view scenarios, with its primary challenge lying in substantial viewpoint variations. It is sourced from Google Street View, covering the complex urban environment of Pittsburgh, USA. Its test set includes $84,000$ database images and $8,200$ query images. Although this dataset does not include condition variations, it contains significant viewpoint changes, making it well-suited for evaluating a model's robustness to viewpoint variations.

MSLS [47] encompasses urban, suburban, and natural scenes captured by vehicle-mounted cameras, featuring challenging variations including temporal, viewpoint, seasonal, illumination changes and dynamic occlusions. This dataset consists entirely of vehicle-mounted camera images,

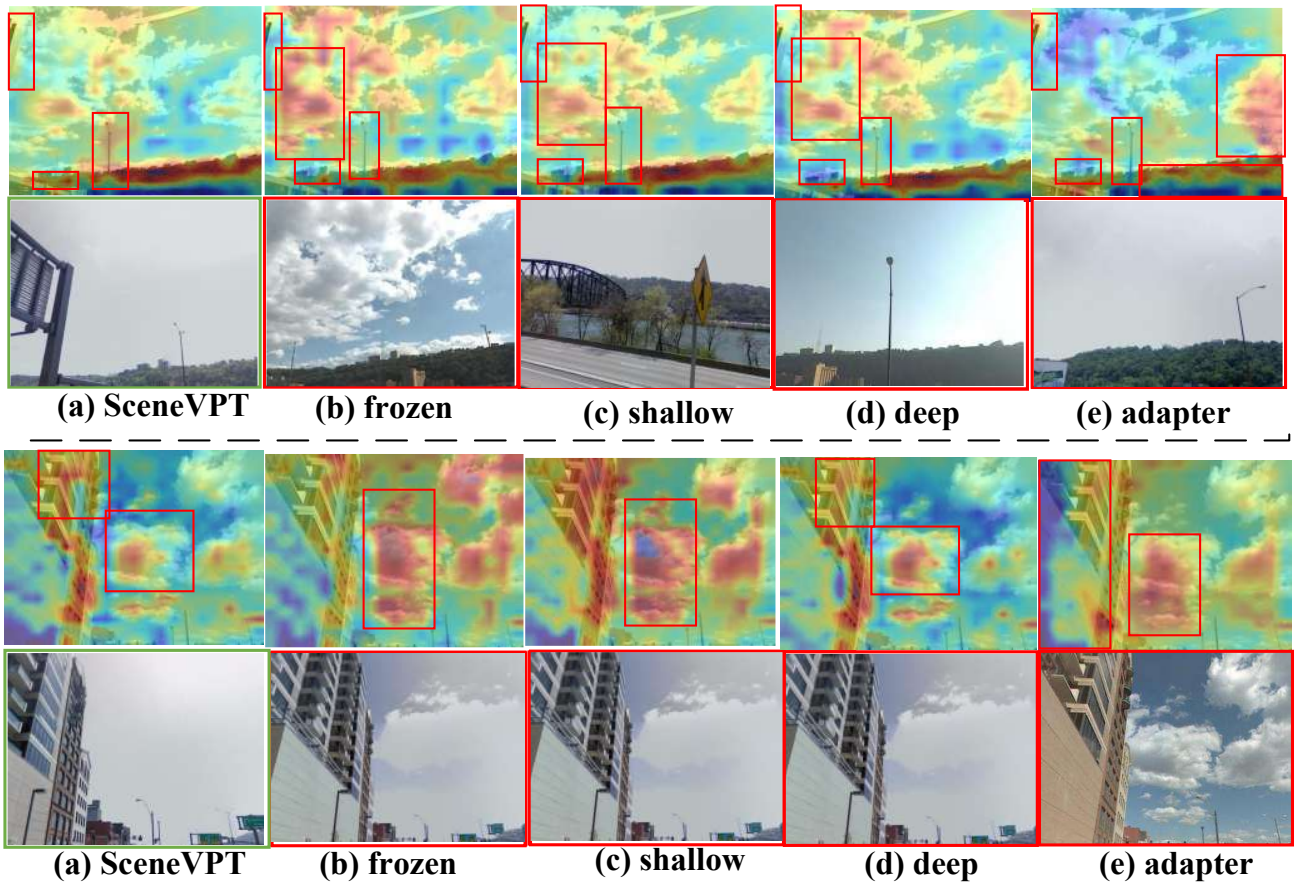

Figure 10. **Additional visualization of fine-tuning methods**. Consistent with the body of the text, the first and third row visualizes features extracted by backbones fine-tuned with different methods. The second and fourth row shows their retrieval results (green boxes: correct matches; red boxes: errors). In both cases, while the sky and clouds occupy most of the area, they lack distinctiveness. The core discriminative features lie in the smaller objects, such as street lamps, signs, and building details. Fine-tuned by SceneVPT, the backbone effectively learns to focus on image-specific discriminative regions, such as the signs and street lamps in the first query and the architectural details in the second, while suppressing interference from irrelevant areas like the clouds. In contrast, with other approaches the backbone is prone to either missing the core discriminative features or paying excessive attention to non-essential parts of the image.

capturing changes across over 30 countries and regions over a 7-year period. Its validation set, MSLS-val, includes 740 query images and $18,871$ reference images.

AmsterTime [52] is an urban dataset with severe domain shift, pairing historical grayscale queries with modern color references. It documents Amsterdam's historical and contemporary urban transitions, where query images consist of archival grayscale photographs while reference samples are modern color images. Additionally, it incorporates seasonal variations, illumination changes, and significant viewpoint shifts, making it one of the most challenging VPR benchmarks that closely approximates real-world conditions.

Eynsham [14] is a suburban-scene dataset where all images are grayscale, preventing models from leveraging color information for VPR. Additionally, the rural environment contains numerous low-texture scenes with similar appearances, further increasing the challenge for appearance-based matching. The test set of Eynsham consists of $24,000$ reference images and $24,000$ query images.

SVOX [11] is specifically designed for urban scenarios with diverse weather variations, consisting of multiple specialized subsets. Its SVOX-night subset uses nighttime scenes as query images paired with daytime scenes as reference images. The full SVOX test set comprises $17,200$ database images and $14,300$ query images.

11. Details of Comparative Methods

NetVLAD [7] As a cornerstone VPR method, NetVLAD differentiably reparameterizes traditional VLAD aggregation, with CNN-learned feature assignment weights substantially improving discriminative power of global descriptors. In our experiments, we select the most widely adopted VGG-16 variant with 4096-D features to better match our proposed method.

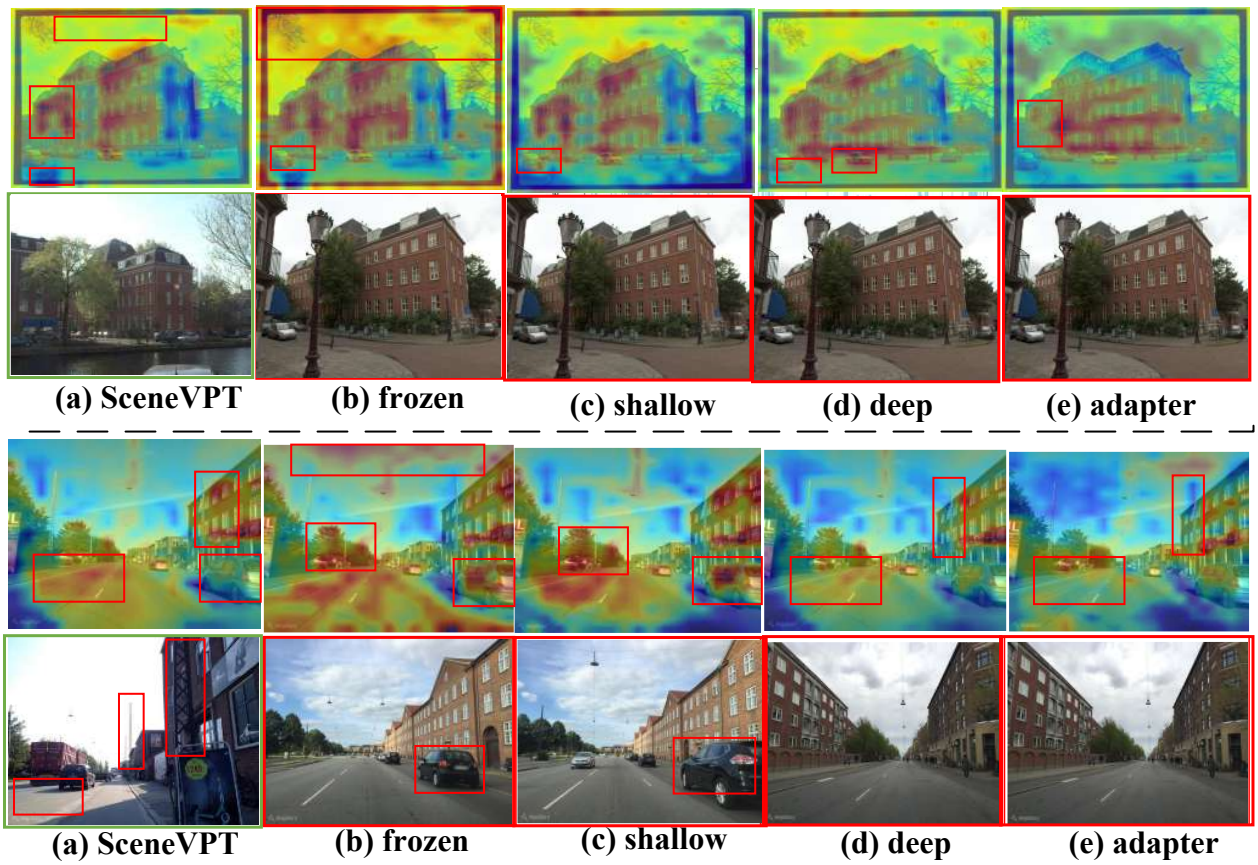

Figure 11. **An enlarged version of Fig.6**.

SFRS [18] SFRS introduces a hard positive samples mining method. It employs an iterative self-supervision mechanism to progressively optimize similarity labels throughout the training process. In the comparison experiments, we choose its official version, which uses VGG-16 as the backbone and VLAD as the feature aggregator.

CosPlace [9] CosPlace employs a classification task as a proxy to train an inference model for extracting discriminative descriptors used in retrieval. It is trained on the self-constructed large-scale dataset SF-XL [9], which encompasses not only the condition variations covered by GSV-Cities but also includes changes in lighting, weather conditions, and camera equipment. For the comparative experiments, we select its optimal configuration, which uses the ResNet-50 as the backbone with 2048-D features.

MixVPR [2] MixVPR incorporates a multi-scale feature aggregator. It alleviates the issue encountered in previous place recognition methods where fixed-scale feature encoding via global coding struggles to adapt to the feature representation of objects of varying sizes within a scene. Consistent with its official implementation, we adopt the optimal configuration, which use ResNet-50 as the backbone with 4096-D features.

EigenPlaces [10] Similar to CosPlace, EigenPlaces also utilizes a classification task as an intermediary to train a model for inference. Unlike CosPlace, which solves the coherent representation learning problem by dividing images into multi-class cells based on geographical direction, EigenPlaces adopts eigenvalue-based techniques to provides samples from different viewpoints for model training. We adopt its optimal configuration, which uses ResNet-50 as the backbone with 4096-D features and adopt SF-XL as training set.

CricaVPR [33] CircaVPR introduces a cross-image encoder that employs Transformer-based attention mechanisms to establish inter-image relationships within a batch, effectively capturing cross-image variations. For backbone fine-tuning, it adopts an adapter-based approach. In our experiments, we follow the official implementation using DINOv2-base as the backbone and apply PCA to reduce extracted features to 4096 dimensions. Notably, since the proposed cross-image encoder leverages the sequential query image characteristics from test datasets while modeling inter-image relationships, we additionally provide test results of a single query image to eliminate this bias.

SALAD [25] SALAD reformulate NetVLAD as an opti-

mal transport problem. It pays attention to both the feature-to-cluster relationship and the cluster-to-feature relationship, and selectively discards features deemed uninformative through a "dustbin" mechanism. In our experiments, in addition to the original version, we also provide the DINOv2-small version of the model according to the official optimal configuration.

BoQ [4] To bridge the performance gap between CNN based models and ViT based models, Ali-Bey et al. introduces a novel Transformer-based aggregation technique called BoQ, which adopts the concept of learnable object queries from DETR [13] in object detection. It employs multiple sets of learnable queries to aggregate multi-scale image features, which are then concatenated to form the final global representation, effectively capturing the universal yet distinctive attributes of locations. In our experiments, in addition to the officially released DINOv2-base version, we also reproduce their DINOv2-small based variant in strict accordance with the configurations in their paper, thereby providing a more comprehensive comparison.

R2Former [55] R2Former is a two-stage method based on ViT-small. It uses the last two layers of the Transformer as the rearrangement module, comprehensively considering feature correlation, attention values, and xy coordinates. During the model training stage, R2Former adopted a multi-stage training method, training the global retrieval and reranking modules respectively and then fine-tuning them together. In the experiment, we adopted the optimal parameter settings in the original paper.

SelaVPR [34] SelaVPR is a two-stage method based on DINOv2-large. It introduces a mutual nearest neighbor (MNN) local feature loss to train the feature extractor, eliminating the computationally intensive spatial verification step in reranking. As for the fine-tuning of backbone, it adopts an adapter based fine-tuning method. Consistent with the original paper, we employed the officially released model trained on Pitts30k for evaluation on Pitt250k and Tokyo24/7, while using their MSLS-trained model version for testing on the remaining datasets.

FoL [42] FoL is a recent cutting-edge two-stage method also based on DINOv2-large. It builds upon SALAD. The novel part is that it models local regions with two losses (SAL and CEL) and introduces weakly supervised learning with pseudo-correspondences. A detailed discussion of this method will be presented in the following section. In our experiment, we not only adopt its original version, but also provide a DINOv2-small based version for fair comparison. All experiments adhere to the official settings.

12. Relations to Other Methods

The two-stage method FoL mines reliable local regions during training, yet it still relies on generic VPR features during testing. Therefore, its global features remain static and cannot adapt to the unique challenges presented by each test sample. In our work, the sample-specific information originates from prompts. Due to our self-adaptive prompt selection mechanism, prompts are dynamically adjusted for each input during testing, allowing our model to extract features that are both sample- and task-specific during testing.

Additionally, during training, although FoL also takes into account the utilization of discriminative local regions, its utilization method is vulnerable to environmental changes in VPR. This is because, in FoL, the determination of the sample-specific discriminative region essentially relies on calculating the similarity between the CLS token and each image patch. However, in extreme lighting changes (such as night to day change), the patch token degrade. Thus, it affects the selection of local regions. In contrast, our method employs CLS token to select prompts, making prompts both sample- and task-specific. These informative prompts are then used to enhance the sample-specific features. This fundamental difference makes our method inherently more robust to environmental changes. This is also reflected in our performance on SVOX-night and MSLS-val.

13. Details of Other Compared Prompt Tuning Methods

In this section, we further introduce IAPT [56] and LoR-VP [27], the other methods (besides VPT) chosen for comparison in Sec.4.4.1.

IAPT is an efficient, representative open-source method for input-adaptive prompt tuning. Like VPT and our SceneVPT, it also belongs to the prefix prompt-based paradigm. It generates context-adaptive soft prompts through layer-wise prompt generators, a common design in existing instance-aware prompt tuning methods. For our experiments, we strictly adhere to the official token count and generator architecture, and re-optimize its parameters for our task to ensure optimal performance.

To compare the performance of other common prompt tuning methods, we select LoR-VP, a recent open-source approach based on non-prefix prompts. It leverages low-rank matrix multiplication to generate visual prompts. Unlike prefix-based methods, it directly adds the low-rank-derived prompt to the uniformly resized image, enabling full-patch interaction while maintaining parameter and training efficiency. In our experiments, we strictly follow the original paper to integrate its open-source code with our model. We then re-optimized both the model parameters and training hyperparameters for VPR task to maximize performance.