Supplementary Material for CaseVPR: Correlation-Aware Sequential Embedding for Sequence-to-Frame Visual Place Recognition

I. Effects of Query Sequence Length

As mentioned in the main text, the sequence length S adopted in the coarse-level sequence retrieval is set as 5 following previous works [1]–[3]. Fine-tuning S does not make a big difference in the retrieval performance, since the nearest consecutive frames already contain sufficient spatio-temporal information to be correlated and encoded as sequence descriptors for identifying potential starting frames. In fine-level sequence matching, the query sequence length qds has a large impact on the frame correspondence precision by determining how much historical data is incorporated.

Therefore, in this section, we explore the effect of different query sequence lengths by gradually increasing qds starting from S. As illustrated in Fig. 1, the figures on various datasets reveal a general trend that increasing qds enhances the sequence matching performance. Specifically, results on Oxford1_v, NYL, and Sm2Sp show improvements in precision with increasing qds, peaking at $qds = 15$. Results on Sp2F, Oxford2, and W2F show performance peaks at a certain value and then decreases, suggesting varying optimal qds values for different datasets. Based on these observations, we selected $qds = 5$ in our experiments in the main text to balance computational efficiency and accuracy.

II. More Qualitative Results

Using the same example presented in Fig. 4 of the main text, Fig. 2 supplements the top-5 retrieved sequences obtained by sequence retrieval and sequence matching in our CaseVPR pipeline. Specifically, the top-5 candidate reference sequences from the coarse-level sequence retrieval (3) are visualized as colored columns horizontally aligned with the difference matrix (4), with each column representing the range of reference frames covered by a single candidate sequence (6). The range of each candidate column corresponds to a restricted candidate region (5) on the difference matrix for mining a starting point, upon which fine-grained sequence matching is applied to mine frame-level correspondences (7). Among the sequences refined by sequence matching, the top-ranked one is selected as the final retrieval (2) of our CaseVPR pipeline. Note that although the candidate sequences obtained through coarse-level sequence retrieval (6) roughly match the query (1) in appearance, their corresponding frame pairs may not be strictly aligned. Hopefully, fine-grained sequence matching can further mine precise frame-to-frame correspondence, ensuring accurate intra-sequence alignment.

Fig. 3 presents two additional qualitative examples comparing CaseVPR with three state-of-the-art single-frame benchmark methods on the Oxford-RobotCar dataset. As illustrated in Fig. 3.(7), the single-frame benchmark methods erroneously retrieve false positives due to perceptual aliasing, where distinct locations appear visually indistinguishable because of repeated building facades or misleading objects (e.g., cars of the same color). Instead of relying on a single image, CaseVPR mitigates perceptual aliasing by correlating consistent VPR-related visual cues in consecutive frames into feature embedding, resulting in discriminative hierarchical descriptors for sequence retrieval and frame-to-frame correspondence mining. In addition to robust descriptors, the adaptive search strategy in the sequence matching stage of CaseVPR can flexibly adjust step sizes between consecutive frames in the retrieved sequence, which ensures consistent matching of consecutive frame pairs. As a result, CaseVPR is able to correctly retrieve challenging queries when other benchmark methods fail.

Another example is presented in Fig. 3.(b), where the single-frame baseline MixVPR fails to retrieve the correct place due to the occlusion by a moving van (see the 7th column in (b)). Occlusion may cause critical visual elements (the traffic light in this example) to not be visible in a single frame, making the scene appear more similar to an incorrect location. Such partial visibility is common in real-world scenarios, where dynamic objects occasionally block essential cues crucial for the task. CaseVPR mitigates this limitation by integrating spatial-temporal cues from multiple frames. It correlates the scenes in consecutive frames so that reliable landmarks that are beneficial to the task can be identified and embedded into features even if they are temporarily obscured.

Overall, the qualitative examples in Fig. 3 demonstrate that CaseVPR is capable of addressing the limitations of single-frame approaches under varying and dynamic conditions. By integrating temporal information and environmental continuity, our CaseVPR can achieve robust and accurate place recognition, even under challenging circumstances with occlusions, dynamic objects, or perceptual aliasing.

References

[1] R. Mereu *et al.*, "Learning sequential descriptors for sequence-based visual place recognition," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10383–10390, 2022.

[2] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Jist: Joint image and sequence training for sequential visual place recognition," *IEEE Robot. Automat. Lett.*, pp. 1–8, 2023.

[3] J. Zhao *et al.*, "Learning sequence descriptor based on spatio-temporal attention for visual place recognition," *IEEE Robot. Automat. Lett.*, vol. 9, p. 2351–2358, Mar. 2024.

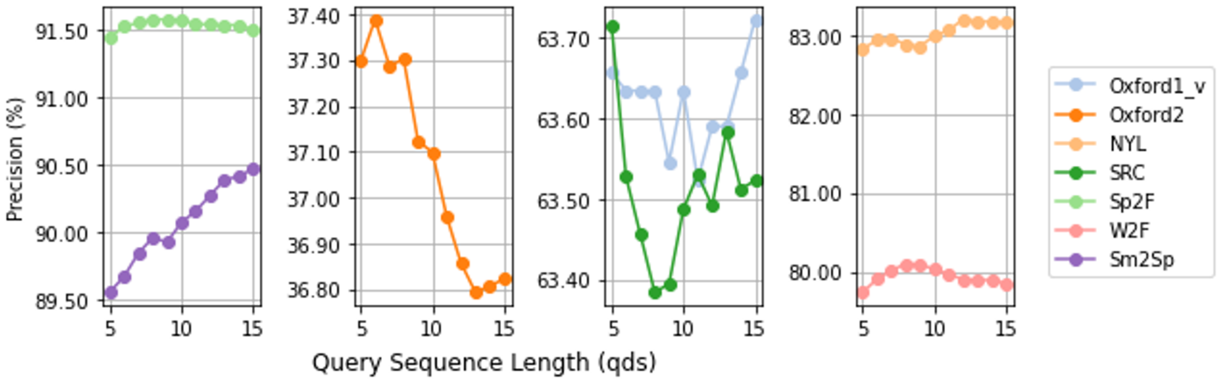

Fig. 1: Ablation study: effect of adjusting the query sequence length qds on the sequence matching performance of our CaseVPR pipeline. The figure is split into 4 subplots for better visualization.

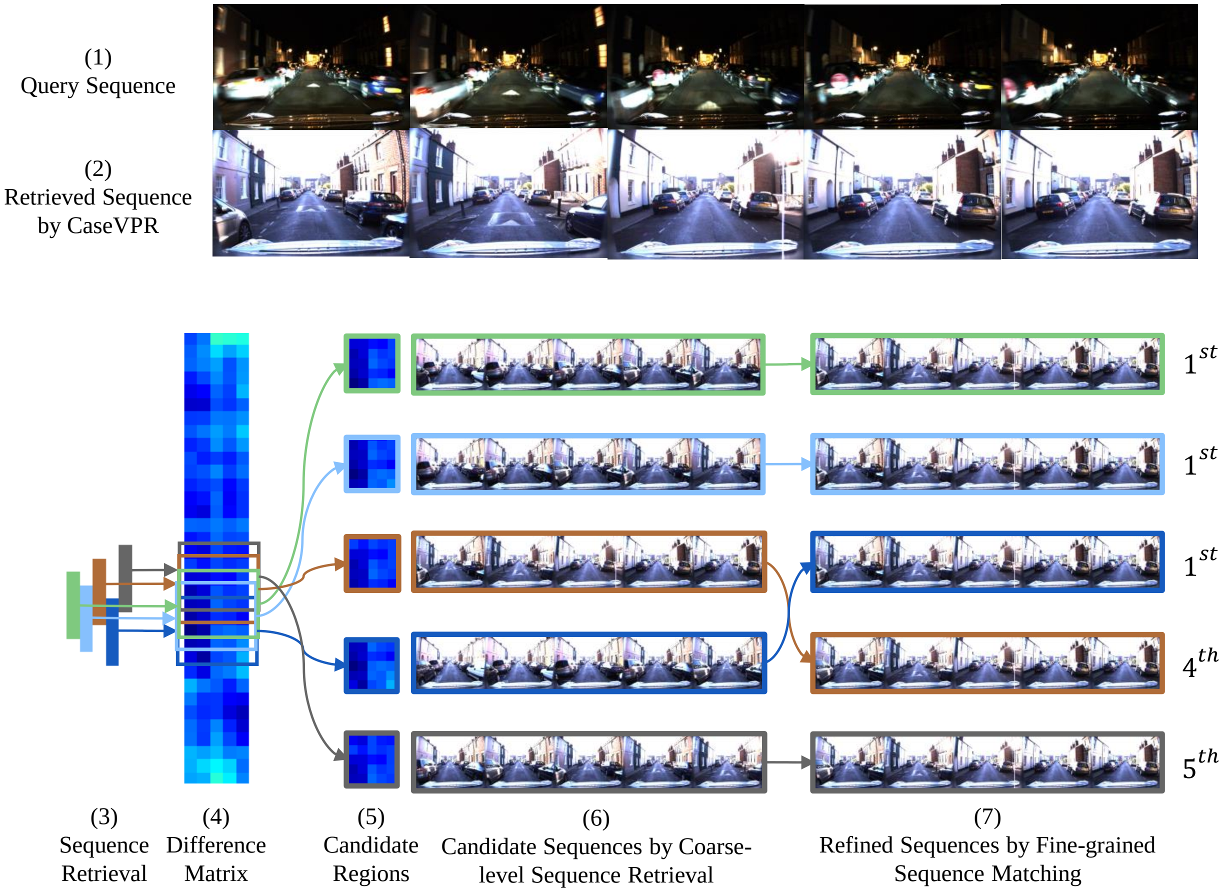

Fig. 2: Visualization of the top-5 retrieved sequences obtained by coarse-level sequence retrieval and fine-grained sequence matching in our CaseVPR pipeline. The darker the blue in the difference matrix, the smaller the difference between the query and reference frames.

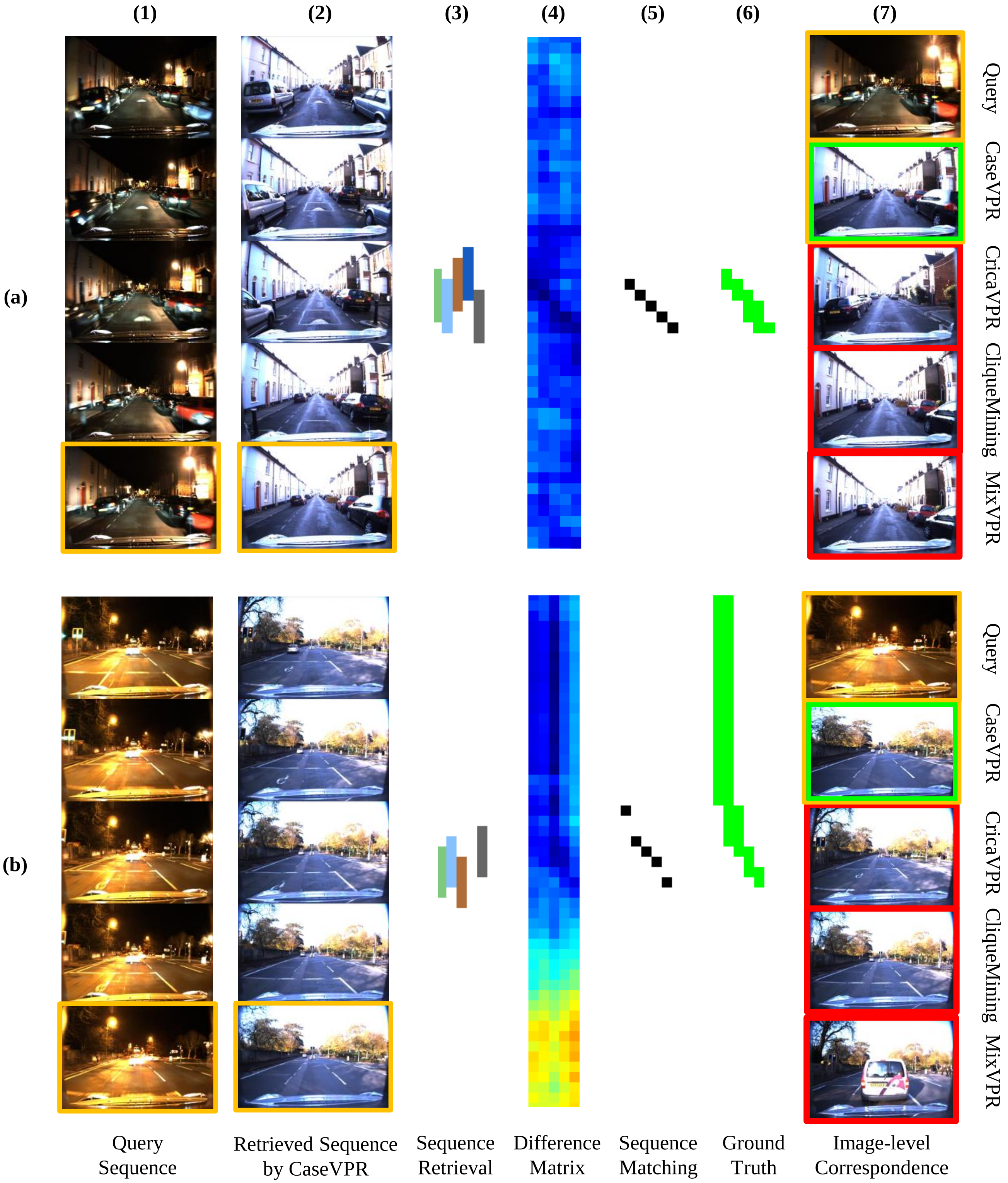

Fig. 3: Two more challenging examples (1) correctly retrieved by CaseVPR (2). With sequence retrieval (3) restricting the candidate regions (4) for fast starting point mining, subsequent sequence matching (5) enables CaseVPR to flexibly find the correct image-level correspondence (7) while other baseline methods fail due to perceptual aliasing.