GPVK-VL: Geometry-Preserving Virtual Keyframes for Visual Localization under Large Viewpoint Changes

Supplementary Material

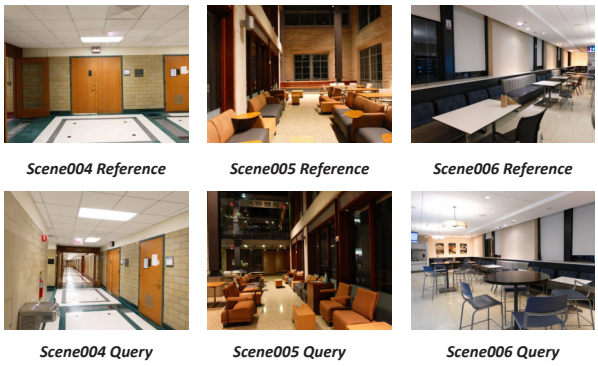

Figure 6. These are examples for scene004 to scene006 from our proposed dataset. The top row shows frames from the reference keyframes database, and the bottom row shows query frames.

A. Investigation Dataset

Our dataset. We specifically collected 6 indoor sequences for testing our problem. These sequences consist of environments such as rooms, hallways, and corridors. When constructing the ground truth poses for this dataset, we utilized the Hloc [20, 48] tool, sampling as many viewpoints as possible within the environment. Frames with registration errors were manually removed. After filtering out frames with registration errors, we manually selected reference images. Frames covisible with the references but with viewpoint changes under 45° were removed, ensuring query images have viewpoint differences between 45° and 180°.

Scannet dataset [18]. We selected a subset of 12 sequences from the ScanNet dataset for our Virtual Keyframe Synthesis Evaluation. The subset was curated by manually selecting initial query views, with the following criteria for choosing reference keyframes: (1) Each reference keyframe's viewing frustum must overlap with at least one query keyframe above a threshold o. (2) if a reference keyframe's overlap with the i^{th} query keyframe satisfies this condition, its rotation and translation must also exceed specified threshold values relative to that query keyframe. To evaluate performance under varying levels of viewpoint changes, different sequences were assigned different thresholds. All experiments were conducted at an image resolution of (512, 384) for this dataset.

B. Virtual Keyframe Synthesis Evaluation

In this section, we analyze the performance differences of various methods on the ScanNet dataset using the GC-FMR metric. As shown in the fig. 7, generalizable feed-forward novel view synthesis methods, such as MVSplat [16] and ZeroNVS [46], are not suitable for synthesizing virtual keyframes for localization purposes. Methods like 2DGS [22] and DRGS [17] fail to maintain sufficient geometric fidelity in the synthesized virtual keyframes, as reflected in the GC-FMR results. Our approach, by introducing confidence-aware geometry regularization, ensures that the extracted mesh retains structural integrity and exhibits minimal deviation from the ground truth in terms of keypoint positions, making it more suitable for feature matching and pose estimation.

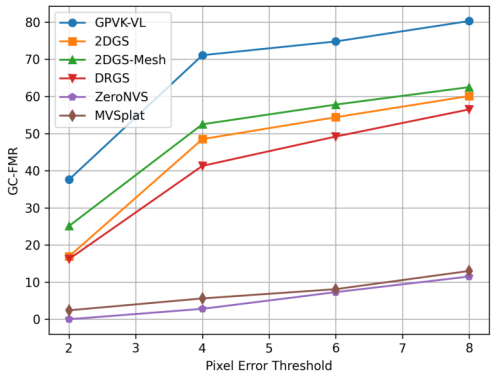

Figure 7. Comparison of Different Methods on GC-FMR metric on the ScanNet Dataset

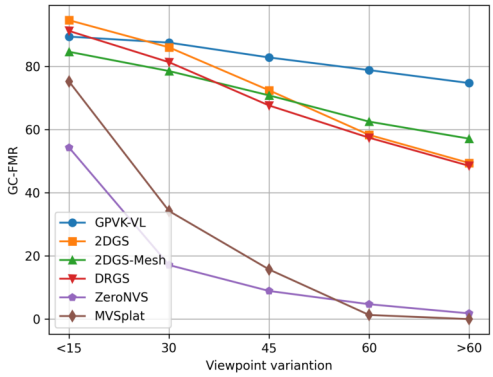

Figure 8. Comparison of Different Methods under Viewpoint Variations

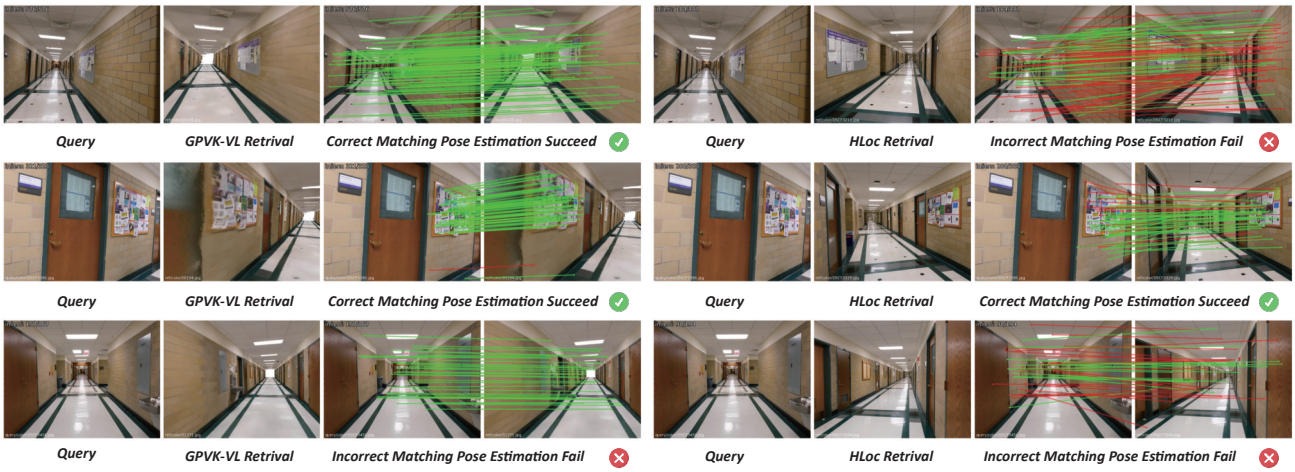

Figure 9. Qualitative comparison of GPVK-VL and HLoc on visual localization pipeline. The left half, from left to right, shows the query view, the frame with the most inliers retrieved from the virtual keyframe database using GPVK-VL, and the feature matching results between the query view and the retrieved keyframe. The right half presents the corresponding results of our main baseline method, HLoc. Green lines indicate inlier matches selected by the Perspective-n-Points RANSAC localization algorithm, while red represents outliers.

C. Viewpoint Variantions

We conducted tests on ScanNet with varying viewpoint differences between query and reference frames. We used the top retrieval's relative angle of the query frame as a measure of its viewpoint difference from the reference database. As shown in the Fig. 8, most methods perform well when the viewpoint difference is small. However, as the viewpoint variation increases, the synthesis quality of most methods deteriorates rapidly. In contrast, our method maintains high synthesis quality, making it suitable for localization under large viewpoint variations.

D. Visual Localization Qualitative Result

When the difference in viewpoint between the query view and the reference keyframe database is large, the HLoc method, despite NetVLAD successfully retrieving relevant keyframes, struggles with feature matching. This issue is particularly severe in indoor scenes with many repetitive patterns, leading to a high number of false positive matches and consequently to localization failures, as shown in the first row of Fig. 9.

In contrast, our method synthesizes and stores query-view-similar perspectives in the virtual keyframe database. As a result, the retrieved image has a much closer viewpoint, and due to the geometry-preserving nature of our approach, the difficulty of feature matching is significantly reduced, enabling successful localization of the query view. Even in cases where HLoc achieves successful localization, the large viewpoint difference often results in highly sparse correct matches, negatively impacting localization accuracy, as illustrated in the second row of Fig. 9. In comparison, our method generates highly dense and accurate feature matches, leading to precise localization.

However, if our method fails to produce high-quality virtual keyframes at query-similar viewpoints, it degenerates into a behavior similar to HLoc, retrieving the most relevant yet significantly viewpoint-different virtual keyframe. This results in false positive matches and eventual localization failure, as demonstrated in the third row of Fig. 9.