This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Yunxuan Li, Lei Fan, Xiaoying Xing, Jianxiong Zhou and Ying Wu
Northwestern University
{yunxuanli,leifan,xiaoyingxing2026,jianxiongzhou2026}@u.northwestern.edu
yingwu@northwestern.edu

Abstract

Visual localization, the task of determining the position and orientation of a camera, typically involves three core components: offline construction of a keyframe database, efficient online keyframes retrieval, and robust local feature matching. However, significant challenges arise when there are large viewpoint disparities between the query view and the database, such as attempting localization in a corridor previously build from an opposing direction. Intuitively, this issue can be addressed by synthesizing a set of virtual keyframes that cover all viewpoints. However, existing methods for synthesizing novel views to assist localization often fail to ensure geometric accuracy under large viewpoint changes. In this paper, we introduce a confidence-aware geometric prior into 2D Gaussian splatting to ensure the geometric accuracy of the scene. Then we can render novel views through the mesh with clear structures and accurate geometry, even under significant viewpoint changes, enabling the synthesis of a comprehensive set of virtual keyframes. Incorporating this geometry-preserving virtual keyframe database into the localization pipeline significantly enhances the robustness of visual localization.

1. Introduction

Visual localization is essential for numerous applications, including autonomous driving in GPS-denied environments, mobile robotics, augmented reality, and virtual reality. Existing approaches to visual localization can generally be categorized into regression-based methods and structure-based methods. Regression-based methods follow two main strategies: either directly regressing the camera pose from the query image [24, 25, 41, 62] or regressing 3D scene coordinates [7–10, 52] at each pixel. While both camera-pose and pixel-wise regression methods are driven by learned scene representations, they often struggle with challenges associated with scene domain shifts. These domain shifts arise from discrepancies such as changes in appearance due to varying illumination and seasonal differences, as well as substantial viewpoint changes, which are inevitable in both structured and natural environments.

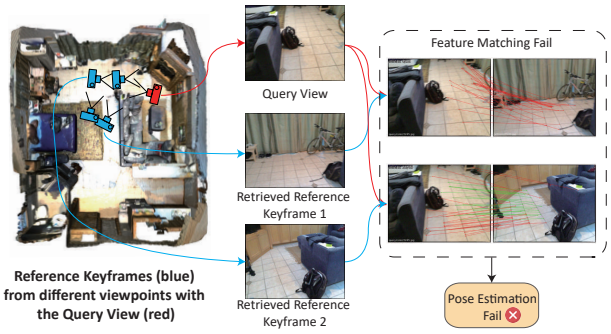


Figure 1. A failure case of the state-of-the-art method in structure-based visual localization toolbox HLoc [2, 47, 48], when the query view undergoes large viewpoint changes relative to the reference keyframes database. The feature matching between the query view and the retrieved reference keyframes contains many false positive matches. As a result, the inlier matches (shown in green) after RANSAC are incorrect, leading to erroneous pose estimation.

In contrast, structure-based methods [30, 47, 49, 56, 57] rely on explicitly constructing a reference keyframe database to represent the environment. Localization is then performed through image retrieval, followed by feature matching and pose estimation. Despite significant progress [37, 42, 51, 58] in addressing challenges associated with appearance changes, a fundamental limitation persists: reference keyframes capture only a partial range of viewpoints, particularly for complex or large-scale environments. A straightforward example of this issue can be explained in corridor environments, where databases are often constructed using images captured from a single direction, with the opposite direction typically neglected. This limited viewpoint diversity often leads to localization failures when

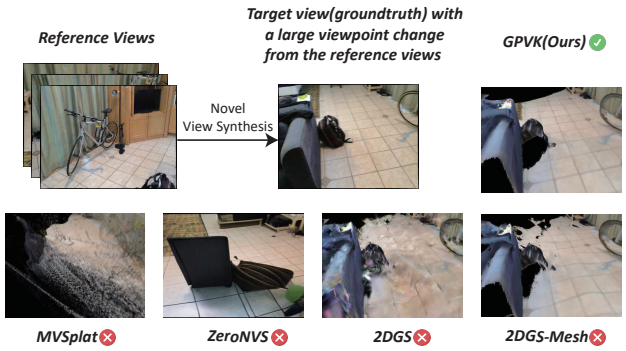


Figure 2. This figure highlights the challenges encountered by existing novel view synthesis methods under large viewpoint changes. The bottom row shows synthesized target views from various methods: from left to right, MVSplat [16], ZeroNVS [46], 2DGS [22], and a method utilizing mesh extracted from 2DGS. These existing methods fail to achieve geometry-preserving results. In contrast, our method, displayed in the upper right, successfully generates geometry-preserving novel views, making it suitable for synthesizing virtual keyframes for localization.

navigating the corridor from the opposite direction, as large viewpoint changes occur.

Before presenting the proposed approach, we first examine the impact of viewpoint changes on the two primary stages of structure-based visual localization: image retrieval and feature matching. In image retrieval [1, 2, 5, 6], handling images with substantial viewpoint changes is typically addressed using global descriptors, which recognize images containing similar objects despite differences in perspective. In the feature matching stage, viewpoint changes introduce additional challenges, leading to missed detections of correct matches and an increase in false positives. Traditional features like ORB [45] and SIFT [31] lack robustness against large viewpoint shifts. Learning-based methods [20, 48, 55], trained under various viewpoint conditions, improve feature matching across diverse perspectives. Nevertheless, substantial viewpoint changes often come with heavy occlusions, limiting the number of available matching points and leaving room for improvement.

Our work diverges from previous approaches by synthesizing a series of virtual keyframes that comprehensively cover possible viewpoints. However, synthesizing novel views under significant viewpoint changes presents unique challenges. As illustrated in Fig. 2, feed-forward generalizable novel view synthesis methods [13, 14, 16, 63, 69, 70], are effective when there are minor zoom or viewpoint changes, but the quality of generated images degrades significantly with larger viewpoint shifts. Another category of methods leverages generative models to produce high-quality images under extensive viewpoint variations [11, 27, 28, 39, 40, 43, 44, 46, 65, 71]. For example, ZeroNVS [46] generates novel views based on imaginative synthesis under large viewpoint changes, making it less suitable for localization tasks. In contrast, our method incorporates a confidence-aware geometric prior to maintain geometric accuracy during novel view synthesis while avoiding introducing misleading information into the database, enhancing its suitability for localization applications.

Our main contributions are summarized as follows:

- We introduce a confidence-aware geometric prior into the 2D Gaussian splatting process, which enhances geometric accuracy under significant viewpoint changes, making it highly suitable for visual localization tasks.
- We propose a novel visual localization pipeline that generates Geometry-Preserving Virtual Keyframes (GPVKs) to comprehensively cover a wide range of possible viewpoints. Leveraging these virtual keyframes as a reference database significantly increases the robustness of our visual localization system.
- We present a newly collected dataset designed specifically for evaluating localization performance under large viewpoint variations and benchmark our algorithm against multiple baselines on this dataset.

2. Related Works

Visual localization. Visual localization is typically addressed through two main approaches. The first is the regression-based approach, which encodes the scene into a neural network by regressing from an image to an absolute pose [24, 25, 36, 41, 62] or to 3D scene coordinates [7–10, 52]. These methods are generally scene-specific, limiting their adaptability to new environments and resulting in poor generalization under varying viewing conditions.

The second approach is structure-based [23, 29, 30, 30, 47, 49, 56, 57], which constructs a reference database and a 3D point cloud of the scene. Localization of a query image is then achieved by retrieving the most relevant views from the database and establishing correspondences between 2D pixel coordinates and 3D scene coordinates. However, as illustrated in Fig. 1, this method also faces challenges when the query viewpoint significantly deviates from the reference viewpoints.

Some research [59, 66] aims to extract viewpoint-invariant features by warping reference views onto a rectified plane before feature extraction, but this approach is not well suited for non-planar features. In contrast, we address this limitation by synthesizing a set of geometry-preserving virtual keyframes that comprehensively cover all viewpoints. The geometry-preserving nature of our synthesized views makes them suitable for feature matching and effective across various types of feature.

Novel view synthesis. Early attempts [12, 73] at novel view synthesis often relied on Multi-View Stereo (MVS) to generate explicit scene geometries, followed by render-

ing. However, these methods struggled with poor reconstruction in textureless areas and failed to produce photorealistic novel views. The advent of Neural Radiance Fields (NeRF) [32, 35, 67] and 3D Gaussian Splatting [26] marked significant improvements in view synthesis quality. Despite these advancements, techniques that regress from training views can still yield blurry images when confronted with substantial viewpoint changes. To enhance generalizability to novel views, adaptations such as generalizable NeRF [14, 63, 69] and Gaussian Splatting [13, 16, 70] have been developed. These require pretraining on a large number of similar scenes and typically only handle a few input views. Another line of work leverages generative models [11, 27, 28, 39, 40, 43, 44, 65, 71] to produce high-quality images under significant viewpoint changes. However, the synthetic nature of these models makes it challenging to maintain geometric consistency.

Our approach builds upon the use of geometric priors [17, 19, 22, 61] to enhance synthesis quality. While previous methods typically use depth or normal priors to guide training, applying these priors without accounting for their confidence can introduce errors. To address this, we incorporate confidence-aware depth and normal priors to guide the 2D Gaussian Splatting training, achieving improved geometric and feature accuracy in the synthesized images.

View synthesis for visual localization. View synthesis has been explored in visual localization from various angles. In SLAM [54, 68, 72], these methods leverage photometric error between rendered and current views to compute camera poses, primarily for refining coarse pose estimations rather than for re-localization tasks. In contrast, InLoc [56] uses rendered views to validate estimated camera poses without directly contributing to pose computation. Other approaches [33, 36, 38] synthesize multiple images from different camera locations to debias training datasets in pose regression methods. SPP-Net [40] employs GANs to generate novel views, its lack of geometric consistency makes it less effective for feature matching tasks. Some methods [53, 60] combine images with depth sensors or laser scanners to synthesize novel views, although reliance on additional hardware limits their practicality.

Another line of approaches [23, 34] focuses on directly synthesizing feature descriptors for novel views. For example, CROSSFIRE [34] learns to render descriptors using NeRF, but if the initial pose prior has a large viewpoint difference from the query view, the descriptor matching can still suffer. Irschara et al. [23] project feature descriptors from similar viewpoints but different locations on the novel view, limiting effectiveness under large viewpoint changes. In contrast, our method can synthesize geometry-preserving novel views even under substantial viewpoint changes, with or without additional hardware, making it more suitable for visual localization tasks.

3. Method

3.1. Problem Formulation

Our approach takes as input a set of posed reference keyframes and a 3D representation, such as a sparse or dense 3D point cloud. These reference keyframes offer limited viewpoint diversity, which poses challenges for pose estimation under significant viewpoint changes. Our objective is to estimate the 6-DoF pose $(R, t) \in \mathrm{SE}(3)$ for each query image. As illustrated in Fig. 3, our core strategy involves synthesizing a series of geometry-preserving virtual keyframes to comprehensively cover the viewpoint space, enabling more accurate and robust localization. These virtual keyframes are then integrated into a structure-based visual localization framework.

We will introduce our method in the following order. First, we present our novel view synthesis approach, Geometry-Preserving Virtual Keyframes, followed by the selection strategy for virtual keyframes. Then, we briefly introduce the visual localization module. Finally, we propose two metrics for evaluating the quality of synthesized images specifically for localization tasks.

3.2. Geometry-Preserving Virtual Keyframes

Our method builds upon 2D Gaussian Splatting [22], and we begin by briefly introducing the rasterization pipeline. In 2D Gaussian Splatting, a scene is represented using surfel primitives, each defined by a center point $\mathbf{p_i}$, principal tangential vectors $\mathbf{t}_u$ and $\mathbf{t}_v$, and a scaling vector $S = (\mathbf{s}_u, \mathbf{s}_v)$ that specifies the variance of the 2D Gaussian within each surfel. Additional surfel parameters include opacity o and color c, represented via spherical harmonics. The splatting process transforms these 2D Gaussian surfels from world space to camera space. Once transformed, surfels are depth-sorted based on depth d, determined by the intersection of the pixel ray with each surfel. The final pixel color $\hat{C}$ is then computed through alpha composition, integrating the depth-sorted 2D Gaussians:

$$\hat{C} = \sum_{i \in N} c_i \alpha_i T_i, \text{ where } T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{1}$$

where α_i is the blending parameter. For a point $\mathbf{u} = (u, v)$ in uv space, which represents the intersection point of the pixel ray and the i^{th} surfel, its α_i is determined by the product of its opacity o_i and its 2d Gaussian value:

$$\alpha_i = o_i \cdot \exp\left(-\frac{u^2 + v^2}{2}\right) \tag{2}$$

Silimiar to the color rasterization process, depth can be rendered in a the same way:

$$\hat{D} = \sum_{i \in N} d_i \alpha_i T_i \tag{3}$$

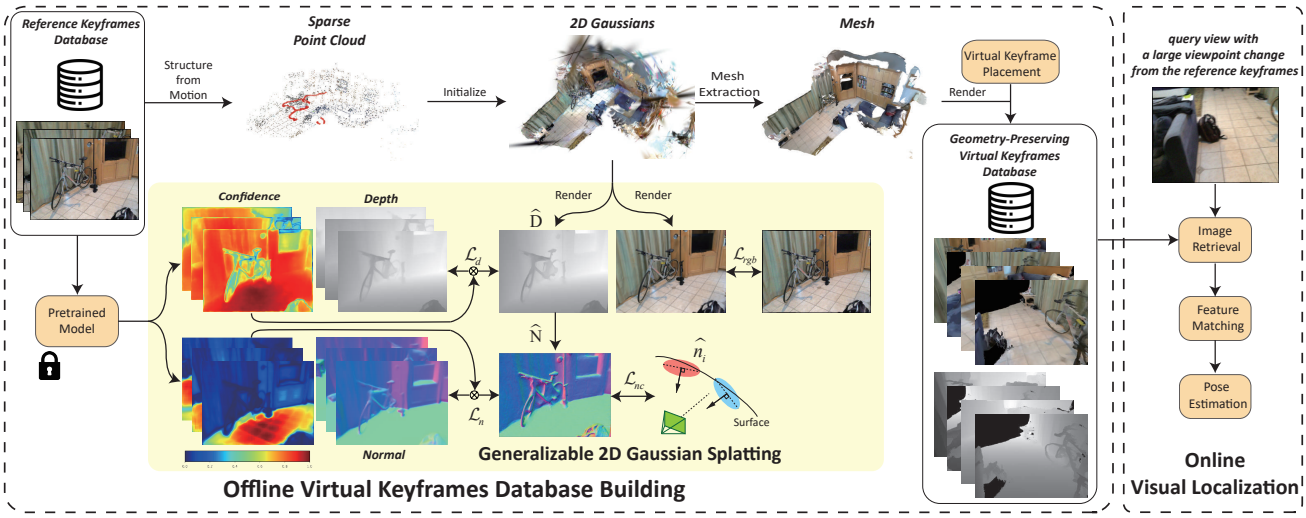


Figure 3. **Geometry-Preserving Virtual Keyframes for Visual Localization (GPVK-VL).** Starting with a database of reference keyframes, we use pretrained models [3, 64] to generate confidence-aware depth and normal maps for each keyframe. For better visualization, we normalize the depth map confidence and the distribution parameter κ of the normal map. Using depth and normal regularization terms along with photometric loss, we train a 2D Gaussian Splatting model from which we can extract a mesh of the scene. This mesh is then used to generate virtual keyframes at selected camera placements. Using this database, we perform visual localization by matching the query view to virtual keyframes with similar viewpoints, enabling accurate feature matching and pose estimation even when the query view has significant perspective changes relative to the reference views.

where d_i represents the depth at the intersection of the pixel ray and the i^{th} surfel.

Normal consistency regularization. The scene is represented using surfels, and the normal for each surfel can be easily estimated through the orthogonal tangential vectors $\mathbf{n} = \mathbf{t}_u \times \mathbf{t}_v$. To enhance geometric accuracy, we incorporate the normal consistency regularization term in 2D Gaussian splatting, as described by Huang et al. [22]. This term aims to align each surfel normal with the surface normal $\hat{N}$ estimated by the gradient of the rendered depth map:

$$\mathcal{L}_{nc} = \frac{1}{|\hat{N}|} \sum_i \omega_i (1 - \hat{\mathbf{n}}_i{}^\top \hat{\mathbf{n}}_i^d) \tag{4}$$

where ω_i denotes the blending weight at the intersection point, $\hat{\mathbf{n}}_i$ represents the normal of the surfel intersected by the i^{th} pixel, $\hat{N}$ is the normal calculated from the depth map gradient, and $\hat{\mathbf{n}}_i^d$ is the normal vector of $\hat{N}$ at pixel i. $|\hat{N}|$ indicates the total number of pixels in $\hat{N}$.

Confidence-aware normal regularization. The normals generated from rendered depth can be noisy, making normal consistency loss insufficient on its own. To address this, we introduce a confidence-aware normal regularization term to further guide training. Specifically, we follow DSINE [3, 4] to produce surface normal estimates along with their associated uncertainty. DSINE models the uncertainty of surface normal estimation as a von Mises-Fisher distribution, with the network outputting both the surface normal estimation mean μ and the parameters κ of the distribution. In our approach, we use the generated surface normals to supervise the normal map $\hat{N}$ obtained from the gradient of the rendered depth. Given μ and κ, the probability density function is formulated as:

$$p_i(\hat{\mathbf{n}}_i^d | \boldsymbol{\mu}_i, \kappa_i) = \frac{\kappa_i \exp(\kappa_i \boldsymbol{\mu}_i^T \hat{\mathbf{n}}_i^d)}{4\pi \sinh \kappa_i} \tag{5}$$

We use the negative log of the probability as our confidence-aware normal loss:

$$-\log p_i = -\log \kappa_i + \log \sinh \kappa_i - \kappa_i \boldsymbol{\mu}_i^\top \hat{\mathbf{n}}_i^d \tag{6}$$

Through derivation and removal of terms unrelated to optimization, we obtain the final form of the confidence-aware normal loss as follows:

$$\mathcal{L}_n = \frac{1}{|\hat{N}|} \sum_i \kappa_i (1 - \boldsymbol{\mu}_i^\top \hat{\mathbf{n}}_i^d) \tag{7}$$

Confidence-aware depth regularization. Following previous works [17, 61], we incorporate a depth prior to further improve geometric accuracy. Unlike previous approaches that use an aligned monocular depth prior, we leverage a state-of-the-art multi-view stereo depth estimation network Dust3r [64] to predict a confidence-aware depth map for training guidance. Since the network provides a confidence value for each pixel, we can integrate it directly into the regularization term. For robustness, We follow DRGS [17] and use the L1 distance to regularize the rendered depth

against the depth prior. As we use an MVS model to estimate depth, the obtained depth is aligned with the scale of the scene. However, slight errors may still occur, resulting in minor differences in scale estimation across frames and introducing inaccuracies in the Gaussian splatting training. We found that the depth prior primarily helps to quickly reconstruct an approximate scene in the early stages of training, providing a good starting point for subsequent training. Gradually reducing the weight of the depth prior in the later stages does not impact the quality of novel view synthesis and can help mitigate errors introduced by the depth prior, so we applied a simple weight decay strategy here:

$$\eta_t = \eta_0 \alpha^t \tag{8}$$

where α is the decay rate and t is the ratio of $\left\lfloor \frac{\text{current step}}{\text{decay step}} \right\rfloor$. We set $\eta_0 = 0.8$ and decay step to 1000 in our experiments. The final form of the confidence-aware depth loss is:

$$\mathcal{L}_d = \frac{\eta_t}{|\hat{D}|} \sum W \|\hat{D} - D\|_1 \tag{9}$$

where $\hat{D}$ is the rendered depth, D is the Dust3r predicted depth, and W is the Dust3r predicted confidence.

Optimization. The final loss term is given by the original photometric loss and normal consistency loss from 2d gaussian splatting [22], together with confidence-aware normal regularization and depth regularization:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_{\text{d}} \mathcal{L}_d + \lambda_{\text{nc}} \mathcal{L}_{nc} + \lambda_{\text{n}} \mathcal{L}_n \tag{10}$$

where λ_{d}, λ_{nc}, and λ_{n} are hyperparameters for each regularization term. We set $\lambda_d = 0.05$, $\lambda_{nc} = 0.05$ and $\lambda_n = 0.005$ in the experiments.

Mesh extraction and rendering. Once a 2D Gaussians model of the scene is trained, we can directly render virtual keyframes using this model. However, when virtual views deviate significantly from the training views, the model lacks sufficient supervision, resulting in noisy rendered images and depth maps, as shown in Fig. 2. To improve the quality of the rendered outputs, a mesh extraction process is necessary. We back-project the rendered image and depth at the reference poses, and utilize truncated signed distance function to fuse them. Using this mesh, we render virtual keyframes with diffuse color and depth. Since our focus is on feature matching, photorealism is not required. Instead, the structural accuracy of the rendered views is crucial.

3.3. Virtual Keyframes Placement

Our application focuses on localization in indoor or urban environments, allowing us to restrict the placement of the virtual camera at the eye level for robots or humans. We first perform gravity alignment of the point cloud. Then, by projecting the 3D point cloud onto the floor plane to create a 2D occupancy map, we transform the problem into view selection within the 2D map. Specifically, we are allocated a budget of K virtual keyframes location. We use θ_y as the yaw angular resolution, and we start with an initial spatial resolution s_i to divide the shorter side of the 2D occupancy map's bounding box, obtaining a fixed distance. This distance is then used to sample the map. After initial sampling, we eliminate samples that are closer than a threshold τ_d to walls or obstacles in the occupancy map, and then remove samples where the co-visible area ratio with the reference database falls below the threshold τ_r. If the remaining number of virtual keyframes k_i is less than K, we increase the spatial resolution to $s_{i+1} = s_i + \sigma_s$ and repeat the process until k_i is greater than K.

For some scenarios, having only horizontal viewpoints for virtual keyframes is insufficient, and upward and downward perspectives are also needed. Therefore, for each virtual keyframe location, we selected three pitch angles, $[0, +\theta_p, -\theta_p]$, to generate the virtual keyframes.

3.4. Visual Localization

After completing the placement of virtual keyframes, we synthesize the color and depth maps for each virtual keyframe location, forming a virtual keyframe database. We then perform localization using a structure-based localization pipeline, which involves image retrieval, feature matching, and pose estimation, as shown in the 'Online Visual Localization' part in Fig 3. Here, we employ the state-of-the-art methods from the visual localization toolbox HLoc, specifically using NetVLAD [2] for image retrieval, SuperGlue [48] for feature matching, and finally, 3D-2D Perspective-n-Point RANSAC for pose estimation.

3.5. Evaluation Metrics for Virtual Keyframes

Feature descriptor similarity. We observed that existing evaluation metrics for novel view synthesis are not well-suited to our task, as they primarily assess the realism, restoration quality, and signal-to-noise ratio of synthesized images. Our focus, however, is on whether the features of the synthesized images can be effectively matched and if these matches can be accurately used for localization. To address this, we adopt a similar idea from InLoc [56] and employ a metric based on the difference in feature descriptors. Specifically, we first extract feature points from the ground truth query view and obtain the corresponding feature descriptors, f_q. For the rendered query view, we calculate feature descriptors f_r at the same pixel locations and compute the mean descriptor distance between them, which we refer to as Feature Descriptor Similarity (FDS).

Geometrically-correct feature matching ratio. We propose another metric called the Geometrically-Correct Feature Matching Ratio (GC-FMR), which provides a more intuitive measure of the quality of novel view synthesis results for localization. First, we extract feature point from the



Figure 4. These are examples from our proposed dataset. The top row shows frames from the reference keyframes database, and the bottom row shows query frames that have view overlap with them but exhibit significant viewpoint changes. Most of the dataset is composed of examples like these.

ground truth query view. Then, we conduct feature matching between the ground truth query view and the rendered query view. Finally, we calculate the ratio of correct feature matches to the total number of feature points in the ground truth query view under different error thresholds.

4. Experiments

Datasets. We evaluate our approach on four datasets. We collected a dataset for visual localization evaluation under large viewpoint changes, comprising 6 indoor scenes, with examples shown in Fig. 4. 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. The dataset was then divided into query frames and reference frames, following the principle of maximizing the viewpoint differences between the query frames and the reference keyframes in the database. ScanNet [18] is an indoor dataset that contains various indoor scenarios. For ScanNet, we selected a subset by manually choosing initial query views. The criteria for selecting reference keyframes were as follows: (1) The viewing frustum of each reference keyframe must overlap with that of at least one query keyframe above a threshold o. (2) if the overlap of a reference keyframe with the query keyframe i^{th} satisfies this condition, its rotation and translation must also exceed the specified threshold values relative to the query keyframe. This approach enables us to simulate our target problem, such as observing the same scene from different perspectives. 7Scenes [21] is a standard indoor localization benchmark comprising 7 scenes, while Cambridge Landmarks [25] is a standard outdoor localization benchmark that contains buildings observed from multiple viewpoints. We used the 7Scenes and Cambridge datasets to evaluate the performance of our method on standard localization benchmarks.

Baselines. For the virtual keyframe synthesis component, we compared our method with the traditional MVS method Colmap [50], the latest generalizable view synthesis method MVSplat [16], geometry regularized view synthesis methods DRGS [17] and 2DGS [22]. DRGS applies Gaussian splatting regularized by an aligned monocular depth, while 2DGS uses normal consistency loss to regularize surfel normals. We also included ZeroNVS [46], a generative view synthesis model, as a baseline. For MVSplat, we selected the top 10 co-visible images from the reference database as context views, while for ZeroNVS, which takes one context view, so we used the top co-visible image.

In our localization experiments, we compared our approach with the visual localization toolbox HLoc [47, 48], as well as DFNET [15], LENS [33], CROSSFIRE [34], and InLoc [56]. DFNET and LENS leverage NeRF-rendered views to debias pose regression, CROSSFIRE employs rendered descriptors for pose estimation, and InLoc uses rendered views at estimated poses for verification and reranking. We enhance InLoc by incorporating synthesized views for pose verification within HLoc.

Metrics. For evaluating virtual keyframe synthesis in the ScanNet dataset, we used PSNR, SSIM, LPIPS, as well as our proposed FDS and GC-FMR to assess the performance of various view synthesis methods. Specifically, FDS used mean L2 distance of SuperPoint descriptors, while GC-FMR utilized SuperGlue for feature matching.

For a comprehensive localization evaluation on our dataset, we define success as achieving localization errors below a certain threshold and report the success rate accordingly. Our dataset emphasizes robustness under large viewpoint changes, where many methods fail in more than half of the queries, rendering the median errors infinite. Thus, we use the success rate to assess the robustness. This success rate highlights the focus of our method on improving localization robustness under significant viewpoint changes. Additionally, we report the median translation and rotation errors for our method on the 7Scenes and Cambridge datasets, allowing for accuracy comparisons with baseline methods on standard visual localization benchmarks.

Implementation. We utilize the MVS model Dust3r [64] to generate a confidence-aware depth prior for each reference keyframe, leveraging 10 covisible frames. For normal estimation, we employ the monocular model DSINE [3] to produce a confidence-aware normal map for each keyframe. These confidence-aware depth and normal maps are then used to guide the 2D Gaussian Splatting training. To extract a mesh, we use a voxel size of 0.01 during TSDF fusion. For virtual keyframe generation, we allocate a budget K of 10 times the number of reference keyframes for generating virtual keyframes, selecting them according to the method

	Scene001	Scene002	Scene003	Scene004	Scene005	Scene006
DFNET	10.8/12.9/12.9	24.5/27.8/29.5	40.2/43.2/44.7	11.1/14.8/15.5	23.0/25.6/26.4	19.1/21.9/21.9
HLoc	25.7/29.7/29.7	85.2/86.8/86.8	**86.5/89.5**/91.0	31.8/33.3/34.0	59.8/62.3/62.3	54.7/57.5/57.5
InLoc	34.6/35.6/36.6	85.2/85.2/86.8	**86.5/89.5/92.5**	37.0/39.3/39.3	65.8/68.3/68.3	57.5/61.6/61.6
Colmap+VK	36.6/39.6/39.6	78.6/78.6/80.3	74.6/80.6/82.1	44.4/46.7/48.1	53.8/58.1/58.9	65.7/69.8/69.8
MVSplat+VK	11.9/14.8/17.8	32.7/40.9/40.9	34.3/38.8/43.2	11.1/14.0/17.0	22.2/27.3/29.0	27.3/32.8/34.2
DRGS-Mesh+VK	22.7/24.7/24.7	65.5/70.4/73.7	71.6/76.1/80.1	29.6/36.3/40.0	55.5/58.9/59.8	52.0/57.5/60.2
2DGS-Mesh+VK	27.7/29.7/29.7	62.3/70.4/72.1	65.6/71.6/74.6	31.8/34.8/37.0	46.1/48.7/50.4	47.9/53.4/54.7
GPVK-VL (Ours)	**59.4/64.4/64.4**	**89.5/91.2/91.2**	83.5/88.0/**92.5**	**57.7/62.2/63.7**	**76.9/80.3/82.0**	**82.1/86.3/86.3**

Table 1. Comparison of various visual localization pipelines on our proposed dataset. Success rates (%) are shown at different orientation and distance thresholds, with the best-performing method highlighted. The thresholds are set at $(2^\circ, 5\text{cm})$, $(4^\circ, 8\text{cm})$, and $(6^\circ, 10\text{cm})$.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FDS ↓
MVSplat	13.28	0.34	0.70	2.23
ZeroNVS	11.51	0.23	0.68	2.51
DRGS	**16.53**	**0.69**	0.48	1.81
Colmap	13.36	0.42	0.51	1.72
2DGS	16.35	0.66	0.51	1.92
2DGS-Mesh	15.78	0.61	0.49	1.60
GPVK-VL (Ours)	16.32	0.62	**0.45**	**1.20**

Table 2. Comparisons of different virtual keyframe synthesis methods on ScanNet.

described in Section 3.3, with $s_0 = 1$, $\sigma_s = 1$, $\tau_d = 0.2m$, $\tau_r = 0.7$, $\theta_y = 15^\circ$, and $\theta_p = 30^\circ$. In the visual localization process, we use NetVLAD for image retrieval and SuperGlue for feature matching, followed by Perspective-n-Points with RANSAC for 3D-2D pose estimation.

4.1. Virtual Keyframe Synthesis Evaluation

We evaluated the quality of Virtual Keyframe Synthesis using the ScanNet dataset. Reference keyframes were used to render images at query views' locations, and these rendered query views were compared with ground truth. The tests were conducted using traditional image quality metrics as well as feature description similarity. As shown in Table 2 , our method does not show significant advantages under traditional evaluation metrics; however, it distinctly outperforms other approaches in Feature Descriptor Similarity. We also evaluated our method against other novel view synthesis approaches on the ScanNet dataset using the Geometrically-Correct Feature Matching Ratio, due to space limitations, we discuss this part in the Appendix B.

This shows that our method effectively maintains geometric accuracy even under large viewpoint changes. The features extracted from the rendered images are closer to the ground truth, with a higher number of points matching the ground truth features. This further confirms that our view synthesis approach is well suited for visual localization tasks. Some qualitative results on both ScanNet and our proposed data sets are shown in Fig. 5; as can be seen, our method achieves the highest accuracy in rendering textures and geometric structures.

	FDS ↓	GC-FMR
2DGS	1.60	52.5/57.8/62.5
+normal	1.52	53.3/56.6/59.2
+c-normal	1.43	60.8/66.7/72.1
+depth	1.71	44.9/46.3/51.8
+c-depth	1.38	63.5/69.4/74.0
GPVK-VL (Ours)	1.20	71.1/74.8/80.3

Table 3. Ablation study on the use of confidence-aware depth and normal priors. Our method incorporates both c-normal and c-depth for regularization. The thresholds for GC-FMR are 4/6/8 pixels. All methods in this table use Mesh for rendering.

4.2. Visual Localization Evaluation

We conducted an initial evaluation of our method on our proposed dataset. As shown in Table 1, our approach significantly outperformed all other methods in terms of localization robustness. The results for DFNet highlight the limitations of pose regression methods in scenarios with large viewpoint changes, as these methods lack training on viewpoints near the query view with high-quality image supervision. InLoc does not use rendered views in the retrieval and pose estimation stages, which limits its localization success rate. For Colmap, MVSplat, DRGS-Mesh, and 2DGS-Mesh, we incorporated the same virtual keyframe placement strategy and localization pipeline to assess their performance in localization tasks. Our method demonstrated particularly strong improvements in long corridor environments, such as Scene001 and Scene004. These scenes pose significant challenges for feature matching due to large viewpoint changes and repetitive patterns. By synthesizing views closely aligned with the query view, our method substantially enhances the robustness of the localization.

We also tested our method on the 7Scenes and Cambridge datasets to evaluate its performance on standard localization benchmarks, as shown in Table 4. The results demonstrate that our method remains competitive on standard localization datasets and enhances the localization accuracy of structure-based methods on certain sequences.

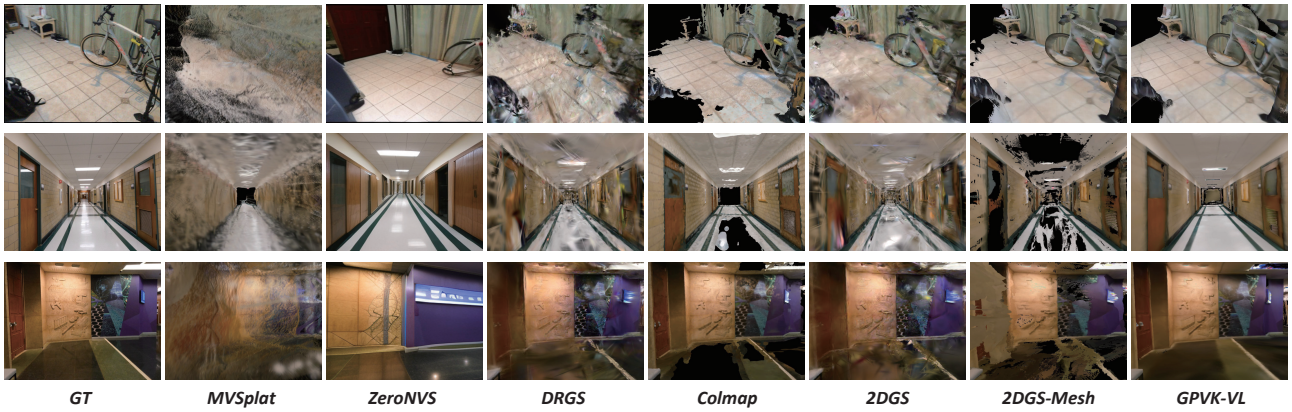


Figure 5. Qualitative comparison of different methods on ScanNet and our dataset. MVSplat struggles with view synthesis under large viewpoint changes. ZeroNVS, as a generative model, cannot guarantee geometric accuracy. DRGS uses aligned monocular depth for supervision, but inaccuracies in depth negatively impact the quality of the synthesized images. COLMAP MVS struggles with missing regions in texture-less areas, while 2DGS lacks geometric supervision, leading to structural inaccuracies or distortions. In contrast, our method preserves both structural and geometric accuracy, making it more suitable for localization.

Dataset		LENS	DFNet	CROSSFIRE	HLoc	GPVK-VL (Ours)
Cambridge	Kings College	0.33m / 0.5°	0.73m / 2.4°	0.47m / 0.7°	**0.12m / 0.2°**	0.25m / 0.3°
	Old Hospital	0.44m / 0.9°	2.00m / 3.0°	0.43m / 0.7°	**0.14m / 0.3°**	0.21m / 0.4°
	Shop	0.27m / 1.6°	0.67m / 2.2°	0.20m / 1.2°	**0.04m / 0.2°**	0.06m / 0.2°
	Church	0.53m / 1.6°	1.37m / 4.0°	0.39m / 1.4°	**0.07m / 0.2°**	0.08m / 0.4°
7Scenes	Chess	0.03m / 1.3°	0.05m / 1.9°	**0.01m / 0.4°**	0.02m / 0.8°	0.01m / 0.7°
	Fire	0.10m / 3.7°	0.17m / 6.5°	0.05m / 1.9°	0.02m / 0.9°	**0.02m / 0.8°**
	Heads	0.07m / 5.8°	0.06m / 3.6	0.03m / 2.3°	**0.01m / 0.7°**	**0.01m / 0.7°**
	Office	0.07m / 1.9°	0.08m / 2.5°	0.05m / 1.6°	0.03m / 0.9°	**0.03m / 0.8°**
	Pumpkin	0.08m / 2.2°	0.10m / 2.8°	**0.03m / 0.8°**	0.05m / 1.3°	0.03m / 0.9°
	Kitchen	0.09m / 2.2°	0.22m / 5.5°	**0.02m / 0.8°**	0.04m / 1.4°	0.03m / 1.0°

Table 4. Comparison of different localization pipelines on 7Scenes and Cambridge, the metric is median translation and rotation error in (m/°). We did not conduct tests on the 'Stairs' sequence due to significant motion along the z-axis in that scenario break our assumption.

4.3. Ablation Study

We conducted an ablation study to evaluate the impact of our confidence-aware depth and normal priors. As shown in Table 3, using standard mono depth or normal priors directly for supervision can introduce inaccuracies due to noise or misalignment in the data, which can disrupt the geometry of the generated images. This misalignment negatively impacts metrics such as FDS and GC-FMR, resulting in poorer localization performance. However, incorporating confidence-aware normal and MVS depth priors for supervision mitigates these issues by weighting the contributions of each prior based on confidence. This enables the generation of geometry-preserving virtual keyframes, which are crucial for accurate and robust localization.

5. Conclusion

In this paper, we introduce GPVK-VL, a novel visual localization pipeline. Our approach uses confidence-aware depth and normal priors to train a 2D Gaussian splatting model, from which a mesh is extracted. This mesh enables the rendering of geometry-preserving virtual keyframes that comprehensively cover the viewpoint space. GPVK-VL shows strong performance in scenarios with significant viewpoint variations, such as long corridor scenes.

Future work includes: 1) Scalability for large maps: Mesh extraction limits our approach for large-scale maps. Dividing the scene into blocks and rendering each independently could address this. 2) Extending beyond 2D planes: Currently, virtual keyframe selection is restricted to a 2D plane, which limits applications to mobile robots. Extending to 3D is feasible but requires additional storage, which can be reduced by storing only global descriptors and poses, with images and depth maps rendered on-the-fly.

6. Acknowledgement

This work was supported in part by National Science Foundation grant IIS-2007613.

References

[1] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 2

[2] 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*, pages 5297–5307, 2016. 1, 2, 5

[3] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024. 4, 6

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021. 4

[5] 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*, pages 4878–4888, 2022. 2

[6] 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*, pages 11080–11090, 2023. 2

[7] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021. 1, 2

[8] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017.

[9] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Luigi Di Stefano, and Philip HS Torr. On-the-fly adaptation of regression forests for online camera relocalisation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4457–4466, 2017.

[10] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip Torr, and Stuart Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. In *2019 International Conference on 3D Vision (3DV)*, pages 564–573. IEEE, 2019. 1, 2

[11] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 2, 3

[12] SC Chan, Heung-Yeung Shum, and King-To Ng. Image-based rendering and synthesis. *IEEE Signal Processing Magazine*, 24(6):22–33, 2007. 2

[13] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2, 3

[14] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2, 3

[15] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 6

[16] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 2, 3, 6, 1

[17] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. 3, 4, 6, 1

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 1

[19] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 6, 1

[21] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179. IEEE, 2013. 6

[22] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4, 5, 6, 1

[23] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2599–2606. IEEE, 2009. 2, 3

[24] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983, 2017. 1, 2

[25] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof cam-

era relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1, 2, 6

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 3

[27] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021. 2, 3

[28] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 2, 3

[29] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *European conference on computer vision*, pages 15–29. Springer, 2012. 2

[30] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 1, 2

[31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020. 3

[33] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 3, 6

[34] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 252–262, 2023. 3, 6

[35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[36] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530. IEEE, 2017. 2, 3

[37] Tayyab Naseer, Wolfram Burgard, and Cyrill Stachniss. Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302, 2018. 1

[38] Tony Ng, Adrian Lopez-Rodriguez, Vassileios Balntas, and Krystian Mikolajczyk. Ood-pose: Camera pose regression from out-of-distribution synthetic views. In *2022 International Conference on 3D Vision (3DV)*, pages 722–732. IEEE, 2022. 3

[39] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[40] Pulak Purkait, Cheng Zhao, and Christopher Zach. Synthetic view generation for absolute pose regression and image synthesis. In *BMVC*, page 69, 2018. 2, 3

[41] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, 2018. 1, 2

[42] Ananth Ranganathan, Shohei Matsumoto, and David Ilstrup. Towards illumination invariance for visual localization. In *2013 IEEE international conference on robotics and automation*, pages 3791–3798. IEEE, 2013. 1

[43] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022. 2, 3

[44] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 2, 3

[45] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[46] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 2, 6, 1

[47] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1, 2, 6

[48] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 5, 6

[49] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 1, 2

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[51] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6896–6906, 2018. 1

[52] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 1, 2

[53] Dominik Sibbing, Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Sift-realistic rendering. In *2013 International Conference on 3D Vision-3DV 2013*, pages 56–63. IEEE, 2013. 3

[54] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 3

[55] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[56] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1, 2, 3, 5, 6

[57] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018. 1, 2

[58] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2074–2088, 2020. 1

[59] Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, Fredrik Kahl, and Gabriel J Brostow. Single-image depth prediction makes feature matching easier. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 473–492. Springer, 2020. 2

[60] 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*, pages 1808–1817, 2015. 3

[61] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. *arXiv preprint arXiv:2403.17822*, 2024. 3, 4

[62] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE international conference on computer vision*, pages 627–637, 2017. 1, 2

[63] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2, 3

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 4, 6

[65] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2, 3

[66] Changchang Wu, Brian Clipp, Xiaowei Li, Jan-Michael Frahm, and Marc Pollefeys. 3d model matching with viewpoint-invariant patches (vip). In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[67] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3

[68] Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 499–507. IEEE, 2022. 3

[69] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 3

[70] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024. 2, 3

[71] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 2, 3

[72] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 3

[73] Sveta Zinger, Luat Do, and Phn De With. Free-viewpoint depth image based rendering. *Journal of visual communication and image representation*, 21(5-6):533–541, 2010. 2