2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GSLoc: Visual Localization with 3D Gaussian Splatting

Kazii Botashev[1] Vladislav Pyatov[1] Gonzalo Ferrer[1] Stamatios Lefkimmiatis[2]

***Abstract*— We present GSLoc: a new visual localization method that performs dense camera alignment using 3D Gaussian Splatting as a map representation of the scene. GSLoc backpropagates pose gradients over the rendering pipeline to align the rendered and target images, while it adopts a coarse-to-fine strategy by utilizing blurring kernels to mitigate the non-convexity of the problem and improve the convergence. The results show that our approach succeeds at visual localization in challenging conditions of relatively small overlap between initial and target frames inside textureless environments when state-of-the-art neural sparse methods provide inferior results. Using the byproduct of realistic rendering from the 3DGS map representation, we show how to enhance localization results by mixing a set of observed and virtual reference keyframes when solving the image retrieval problem. We evaluate our method both on synthetic and real-world data, discussing its advantages and application potential.**

I. INTRODUCTION

Visual localization, the process of determining the camera pose using a visual representation of a known scene, plays an important role in various applications related to robot navigation, self-driving cars, and augmented/virtual reality [1], [2]. In particular, the main objective of visual localization via camera alignment is, provided an input query image, to determine the 6 degrees of freedom (dof) camera pose (position and orientation) in a 3D environment with a known map representation.

The map representation of a known scene, which is a core part of every localization method, can be of different forms. The most developed and commonly used ones are sparse map representations [3], [4], which rely on a set of 2D-3D feature-landmark correspondences typically estimated using structure-from-motion (SfM) techniques [5]. Despite their effectiveness in various localization scenarios, sparse map representations provide limited scene comprehension, falling short in empty spaces or textureless environments with no distinct features. Dense mapping is an alternative family of representations that aim to utilize information from entire images but may require capturing depth, ensuring continuity of the input frames [6]–[9]. Other methods may operate on dense image descriptors [10], [11], usually extracted with convolutional neural networks (CNN). Methods of this category have proven their efficiency in large-scale scenarios and image retrieval tasks but have limited accuracy and produce only an approximated pose of the query camera.

[1]The authors are with the Skolkovo Institute of Science and Technology (Skoltech), Center for AI Technology. {kazii.botashev, vladislav.pyatov, g.ferrer}@skoltech.ru
[2]Stamatios Lefkimmiatis is with MTS AI, Russia. s.lefkimmiatis@mts.ai

Differentiable mesh-based rendering algorithms have also been employed for the visual localization task, leading to a family of dense map representation methods that can achieve impressive results. However, this comes at the significant cost of requiring a detailed 3D model of the environment [12], [13]. This drawback has been recently mitigated with the introduction of Neural Radiance Field (NeRF) [14] models that can be trained using only a set of posed images. NeRFs can achieve photo-realistic rendering quality by implicitly learning via 2D supervision the 3D scene as a function of a continuous radiance field. While NeRFs were originally introduced to deal with novel-view synthesis, their learned map representation has been recently used in the design of novel pose estimation methods. Started with a simple idea presented in iNeRF [15], it continued with other sophisticated pose estimation approaches [16], [17]. However, despite their initial promising results, such methods still face a limited applicability since they suffer from the same drawbacks of NeRF models, that is extremely long training and rendering times due to the expensive backward mapping ray-casting procedure.

Recently, 3D Gaussian Splatting (3DGS) [18] has been introduced and achieves high-quality real-time novel view synthesis at full HD resolution. This is an alternative learning-based approach that unlike NERF-based methods is based on a forward mapping/rasterization strategy. Specifically, 3DGS represents the 3D scene with a collection of 3D anisotropic Gaussians, which play the role of rendering primitives and whose parameters are directly optimized from a set of available posed images during training. The type of operations required by a 3DGS rasterizer are better suited for GPUs resulting in a very efficient and interactive novel view rendering process.

3DGS introduces a novel and distinctive map representation of the environment, which shows promise for effectively addressing the challenges associated with camera pose estimation and visual localization. The 3DGS strategy is computationally efficient and fully differentiable, facilitating the generation of highly realistic images in arbitrary views. Importantly, it allows for the direct flow of parameter gradients for any given camera pose, enabling real-time dense camera alignment, a capability not offered by other localization methods. Furthermore, it establishes a unique and fully-differentiable rendering-pose relation, enabling the generation of rendering images for any given camera and facilitating gradient-based optimization to refine its pose by minimizing the discrepancy between rendered and query images.

Nevertheless, there are still two challenges associated with

979-8-3503-7769-9/24/$31.00 ©2024 IEEE

this novel approach. The first one is that the accuracy of the initial camera pose used during training can have a significant impact on the success of the method. Secondly, the utilized objective loss, which is based on the photometric difference, is highly non-convex due to the presence of high-frequency details in the images. The non-convex nature of the loss poses difficulties in its optimization, as it can lead to the entrapment of the optimization process to one of the numerous local minima, resulting in suboptimal solutions.

This work focuses on utilizing the 3D Gaussian Splatting rendering technique as a map representation for visual localization tasks and aims to overcome the existing challenges described above. Our study includes an investigation of the viability of the 3DGS method as a map representation, a comprehensive convergence analysis for various camera initialization scenarios, an exploration of convergence limitations arising from the highly non-convex nature of the problem, and the proposal of a coarse-to-fine optimization strategy to mitigate such limitations. The main contributions of this work are summarized as follows:

- We analytically derive the gradients of the 3DGS renderings with respect to to camera poses and implement a 3DGS-based visual localization pipeline.
- We propose a coarse-to-fine optimization strategy where we apply gradually fading Gaussian blur on the query and rendered images that allows us to overcome the problem of suboptimal convergence for high-frequency image details.
- We propose an effective way of improving the localization results by enhancing camera initialization obtained via image retrieval by extending its image base with rendered camera frames.
- We evaluate our approach on indoor synthetic and real scenes, provide a comprehensive quantitative analysis of camera pose convergence based on various initial camera pose priors and parameterizations, and compare it with a sparse feature-based localization baseline.

II. RELATED WORK

A. Visual Localization methods

Classic sparse feature-based localization focuses on detecting and matching a sparse set of distinctive features or keypoints in the camera images [3], [4]. The initial approach to feature matching involved manual design of keypoint detection algorithms that can identify visually salient image details: points, edges, and corners [19]. However, the recent progress in the field, which has been driven by the introduction of dedicated neural networks for feature extraction [20], has led to a revision of the feature extraction and matching stages. Indeed, these network architectures have demonstrated exceptional performance, achieving precise and robust feature matching results.

Consequently, the map representation in sparse feature-based localization can be constructed using network-extracted keypoints alongside their related 3D positions or descriptors. At the localization stage, the query image is processed to extract keypoints that are afterwards matched against the map to estimate the 6-DoF camera pose. Sparse feature-based methods are computationally efficient and have demonstrated robustness in a variety of applications. However, they cannot be used for tasks that require scene understanding while they also disregard useful volumetric context and, thus, can fail in featureless or empty environments.

On the other hand, dense visual localization methods aim to utilize an entire image dense map representation by matching visual information across the entire images using dense descriptors, such as pixel-level descriptors or dense feature maps. This type of representations encodes information about the appearance, texture, or semantic context of the scene [11].

B. Rendering-based Pose Estimation

The introduction of Neural Radiance Fields (NeRF) [14] and the large array of follow-up work [15]–[17] has brought a new paradigm to novel view synthesis and subsequently to camera pose estimation. In particular, NeRF represents the scene as a continuous 3D volume and learns the radiance field properties, resulting in a more accurate and realistic map representation of static scenes with intricate geometry and complex lighting. While NeRF-based camera pose estimation methods have certain advantages and can achieve competitive results, they also face a limited applicability due to the principal drawbacks of the NeRF model itself, including long model training and inference time due to the computationally expensive utilized backward mapping/ray-casting procedure.

Meanwhile, recent studies indicate that 3DGS-based camera pose estimation methods effectively circumvent these drawbacks and hold significant promise for future applications [21], [22].

III. RENDERING WITH 3D GAUSSIAN SPLATTING

For a given set of $\mathcal{N}$ RGB images $\{I_k\}_{k=1}^{\mathcal{N}}$ and corresponding camera poses $T_{w\,k}^{c} \in SE(3)$, 3DGS can learn a 3D scene representation that enables photo-realistic novel view rendering for an arbitrary camera pose. This is achieved by modeling the scene using a collection of $\mathcal{M}$ 3D Gaussians, which are defined in a world coordinate frame w:

$$\mathbf{G} = \{G_i^w : (\boldsymbol{\mu}_i^w, \boldsymbol{\Sigma}_i^w, \boldsymbol{\sigma}_i, \boldsymbol{c}_i)\}_{i=1}^{\mathcal{M}}. \tag{1}$$

These Gaussians serve as rendering primitives and are fully described by their centers $\boldsymbol{\mu}_i^w \in \mathbb{R}^3$, covariance $\boldsymbol{\Sigma}_i^w \in \mathbb{R}^{3\times3}$, opacity $\boldsymbol{\sigma}_i \in \mathbb{R}$ and view-dependent color $\boldsymbol{c}_i \in \mathbb{R}^3$.

Having the 3D Gaussian scene representation $\mathbf{G}$ at hand, rendering the image for a novel view, which is determined by a camera pose defined using a world w to camera c rigid body transformation $T_w^c = \{\mathbf{R}_w^c \in SO(3), \mathbf{t}_w^c \in \mathbb{R}^3\} \in SE(3)$, proceeds by perspectively projecting the Gaussians G_i^w to the image plane $\mathcal{I}$. To do so, we first express the Gaussians w.r.t the camera frame, which leads to:

$$\boldsymbol{\mu}_i^c = T_w^c \boldsymbol{\mu}_i^w; \quad \boldsymbol{\Sigma}_i^c = \mathbf{R}_w^c \boldsymbol{\Sigma}_i^w \mathbf{R}_w^{c\,\intercal} \tag{2}$$

Then, all the Gaussians are perspectively projected to the image plane using an affine approximation of the projective non-linear transformation π that involves its Jacobian $\mathbf{J}$. This

approximation leads to a mapping of the initial 3D Gaussians to 2D Gaussians whose centers and covariances are expressed as:

$$\boldsymbol{\mu}_i^{\mathcal{I}} = \pi(\boldsymbol{\mu}_i^c); \quad \boldsymbol{\Sigma}_i^{\mathcal{I}} = \mathbf{J}\boldsymbol{\Sigma}_i^c\mathbf{J}^\intercal = \mathbf{J}\mathbf{R}_w^c\boldsymbol{\Sigma}_i^w\mathbf{R}_w^{c\intercal}\mathbf{J}^\intercal \quad (3)$$

Finally, the image intensity $\hat{\mathbf{C}}$ is computed via depth-ordered α-blending of the projected Gaussians as follows:

$$\hat{\mathbf{C}} = \sum_{i\in\mathcal{M}} \mathbf{c}_i(\mathbf{d}_i)\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \quad (4)$$

where the density α_i is computed as a multiplication of the covariance $\boldsymbol{\Sigma}_i^{\mathcal{I}}$ and opacity $\boldsymbol{\sigma}_i$, $\mathbf{c}_i(\mathbf{d}_i)$ is the view-dependent color of the 3D Gaussians defined with spherical harmonics and computed based on the view-direction vector $\mathbf{d}_i = (\boldsymbol{\mu}_i^w - \mathbf{t}_c^w)/\|(\boldsymbol{\mu}_i^w - \mathbf{t}_c^w)\|$.

Starting with some sparse SfM [5] point cloud initialization and following the above described fully-differentiable rendering procedure, 3DGS gradually optimizes the 3D Gaussian parameters with gradient descent by minimizing the weighted combination of L_1 and D-SSIM losses between the rendered image $\hat{I}_k(T_{wk}^c, \mathbf{G})$ and the ground-truth posed image I_k. As a result, the 3D scene is learned with a high fidelity representation allowing for photo-realistic novel view rendering.

IV. METHOD

The learned 3DGS scene model $\mathbf{G}$ serves as a novel and distinctive map representation of the 3D environment and is potentially highly suitable for being utilized in the visual localization task. Specifically, for a query image $\tilde{I}$ the corresponding pose $\tilde{T}_w^c = \tilde{T} \in SE(3)$ can be found by minimizing the discrepancy between the rendered $\hat{\tilde{I}}$ and query images:

$$\tilde{T} = \underset{T\in SE(3)}{\arg\min}\, \mathcal{L}_1(\hat{\tilde{I}}(T, \mathbf{G}), \tilde{I}). \quad (5)$$

Ostensibly, the solution of this task seems to be straightforward thanks to the photo-realistic real-time rendering capabilities of 3DGS. However, there are several aspects that require specific attention and which we address next.

A. Camera Pose Gradients

To achieve real-time rendering performance, 3DGS utilizes GPU capabilities and its official implementation of the rasterization step is based in CUDA. This precludes out-of-the-box automatic differentiation and instead requires the derivation of the explicit form of gradients for all the parameters to be optimized. To enable camera pose optimization, it is also required to express analytically all the gradients related to the poses parameters.

Since in the original work of 3DGS the authors did not optimize the camera poses we need to derive all the gradients related to the pose parameters that we wish to estimate. In this work, we implement the camera pose optimization on the Riemannian manifold and use Lie algebra to derive the camera pose Jacobians for the terms in Eq. (3) and Eq. (4) via the chain rule as follows:

$$\frac{\partial \boldsymbol{\mu}_i^{\mathcal{I}}}{\partial T_w^c} = \frac{\partial \boldsymbol{\mu}_i^{\mathcal{I}}}{\partial \boldsymbol{\mu}_i^c}\frac{\partial \boldsymbol{\mu}_i^c}{\partial T_w^c} \quad (6)$$

$$\frac{\partial \boldsymbol{\Sigma}_i^{\mathcal{I}}}{\partial T_w^c} = \frac{\partial \boldsymbol{\Sigma}_i^{\mathcal{I}}}{\partial \mathbf{J}}\frac{\partial \mathbf{J}}{\partial \boldsymbol{\mu}_i^c}\frac{\partial \boldsymbol{\mu}_i^c}{\partial T_w^c} + \frac{\partial \boldsymbol{\Sigma}_i^{\mathcal{I}}}{\partial \mathbf{R}_w^c}\frac{\partial \mathbf{R}_w^c}{\partial T_w^c} \quad (7)$$

$$\frac{\partial \mathbf{c}_i}{\partial T_w^c} = \frac{\partial \mathbf{c}_i}{\partial \mathbf{d}_i}\frac{\partial \mathbf{d}_i}{\partial \mathbf{t}_c^w}\frac{\partial \mathbf{t}_c^w}{\partial T_w^c}. \quad (8)$$

We compute the derivatives on the manifold following the general approach detailed in [23]. Due to space limitations we omit their detailed derivations and provide only their final forms:

$$\frac{\partial \boldsymbol{\mu}_i^c}{\partial T_w^c} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\mu}_i^{c\wedge} \end{bmatrix}; \quad \frac{\partial \mathbf{t}_c^w}{\partial T_w^c} = \begin{bmatrix} \mathbf{R}_w^{c\intercal} & \mathbf{0} \end{bmatrix} \quad (9)$$

$$\frac{\partial \mathbf{R}_w^c}{\partial T_w^c} = \begin{bmatrix} \mathbf{0} & -\mathbf{r}_{c1}^\wedge \\ \mathbf{0} & -\mathbf{r}_{c2}^\wedge \\ \mathbf{0} & -\mathbf{r}_{c3}^\wedge \end{bmatrix} \quad (10)$$

where $^\wedge$ denotes the skew symmetric matrix constructed from the corresponding input vector and $\mathbf{r}_{cj}$ denotes the j-th column of the rotation matrix $\mathbf{R}_w^c$.

Using the equations (6)-(10) it is possible to propagate all the necessary gradients for the camera pose optimization task. We iteratively solve the optimization problem of (5) for 2000 steps or until convergence using the first-order Adam [24] optimizer with exponentially decaying learning rate. More details are provided in the appendix.

Manifold-derived pose Jacobians have a minimal 6 DoF representation and lead to a better convergence compared to alternative parametrizations. In particular, according to our ablation study, the manifold optimization achieves better results and shows clear advantages compared to the common alternative parametrization utilizing quaternions for rotation and 3D vectors for translation.

B. Impact of Initial Camera Pose Proximity

Fig. 1: Visual explanation of 3D Intersection over Union (IoU) metric used for camera frames proximity estimation. Computed with voxels of the scene, this metric naturally describes both proximity of the camera poses and the visual similarity of their image frames. Here for the visualized frames the 3D IoU is equal to 0.15.

Among the most important factors that affect the final result of visual localization is the proper initialization of

the camera poses. Finding an initial camera pose that exhibits a sufficiently large overlap with the query camera frame is crucial and can be one of the main factors of success or failure. This problem, which is also known as the Image Retrieval task, is a separate long-standing computer vision problem that requires special attention on its own. While, a 3DGS-based solution of this task is an intriguing possible research direction, here we focus on solving the exact visual localization task. As a result, finding the best possible existing Image Retrieval algorithm for the camera pose initialization for GSLoc is out of scope of this work.

Instead, we seek to perform a comprehensive analysis of our method. We aim to estimate the dependency between obtaining the correct result with GSLoc and the proximity of the initial camera frame to the target one. In other words, we want to answer the following questions: 1) How close/far our initial guess of the camera pose need to be so that our method leads to the correct solution? and 2) What are the chances of this convergence?

To answer these questions, we propose to measure the camera frames proximity using the 3D Intersection over Union (IoU) metric that is computed using voxels of the scene. As visualized in Fig. 2, this metric naturally describes both proximity of the camera poses and the visual similarity of their corresponding image frames and enable us to quantitatively assess their impact on the final result. For our particular task the 3D IoU is more informative and intuitive compared to simple rotation and translation distances.

C. Extending image retrieval database with renderings

Besides the 3D IoU criterion, we also evaluate the results of the visual localization for initial poses corresponding to the closest dense image descriptors. Following a widely adopted approach, we extract global image descriptors using NetVLAD [11] and for each query image we find the most similar ones in the pool of images used for 3DGS training. Next, we solve the visual localization task by initializing the camera pose with these closest matches.

This is a very common approach widely adopted as a first step for sparse feature based visual localization. Its results directly depend on the number and diversity of images used as a map for comparison. However, 3DGS-based map representation allows us to overcome this limitation by extending the original set of images used both for 3DGS training and image retrieval step with any number of posed photorealistic scene renderings produced with the optimized 3DGS scene. Starting with a limited set of images, 3DGS allows us to extend our image base used for Image Retrieval by creating arbitrary novel-view renderings. This increases the probability of obtaining a good initial pose and as a result increases our chances of converging to the correct solution. Further, we show the effectiveness of this proposed technique both for synthetic and real scenes in evaluation.

D. Coarse-to-fine Rendering Scheduling

The highly non-convex nature of the photometric $\mathcal{L}_1$ loss w.r.t the 6 DoF space of camera poses in $SE(3)$ poses a significant challenge related to its optimization. This non-convexity is caused by high-frequency image details and can lead to the entrapment of the optimization process to a bad local minima, which in turn can lead to a sub-optimal solution.

The visual representation of one such case is depicted in Fig. 2(a)-(b). Iteratively minimizing the objective function in (5) in a standard way leads to the convergence of the first-order method to a sub-optimal solution. Indeed, it is clearly visible that during standard optimization using Adam [24] does not manage to escape the local minima caused by the sub-optimal overlap between the intermediate rendering and the target query image. This results to an unsuccessful image alignment (highlighted with yellow).

To overcome this problem we propose a simple yet effective coarse-to-fine strategy of applying a progressively decaying Gaussian blur both on the rendered and target images. Specifically, we convolve both the target query image and the intermediate rendering with a 2D Gaussian kernel $\mathcal{N}_{2d}(\delta_j) \in \mathbb{R}^{L \times L}$ of fixed size L while gradually decreasing its covariance δ_j. This results in a modified objective function $\mathcal{L}_1(\mathcal{N}_{2d} * \hat{\tilde{I}}(T, \mathbf{G}), \mathcal{N}_{2d} * \tilde{I})$ with smoothed gradients and a stabilized camera pose estimation as depicted in Fig. 2(c)-(d). Smoothing the image gradients allows us to avoid being trapped in local minima and converge to the correct camera pose. We have also found it to be effective the strategy of running several passes of coarse-to-fine optimization, restarting each new pass with the result from the previous one. Based on the above, we have concluded that the highest efficiency is achieved by the following two-step GSLoc algorithm: 1) In the first step a standard camera pose optimization takes place. 2) If the first step does not recover the correct pose (the photometric loss between the rendered image and the query image exceeds a user-defined threshold), then we restart the entire process and apply the described coarse-to-fine optimization strategy.

V. EVALUATION

We assess the performance of our proposed method by performing extensive experiments on 5 synthetic scenes from the Replica [25] dataset. Our motivation for using this data in our evaluation stems from several reasons that we discuss next. The first reason is that the use of synthetic data ensures that the 3DGS map representation $\mathbf{G}$ is adequately learned. This can be achieved by exploiting the available accurate ground truth poses and depth information during the 3DGS training. In turn, this allows us to neglect any possible negative effects of the incorrect map representation on the visual localization results and validate the efficiency of the proposed method without worrying about data-related inaccuracies. Another reason is that we have access to the ground truth poses which allows us to accurately compute the localization errors and perform a precise evaluation of the proposed method. Finally, in order to conduct a comprehensive study of the effect of pose initialization based on the camera proximity according to the 3D voxel-based IoU, we need access to the detailed 3D voxel model of the scene, which we can accurately extract

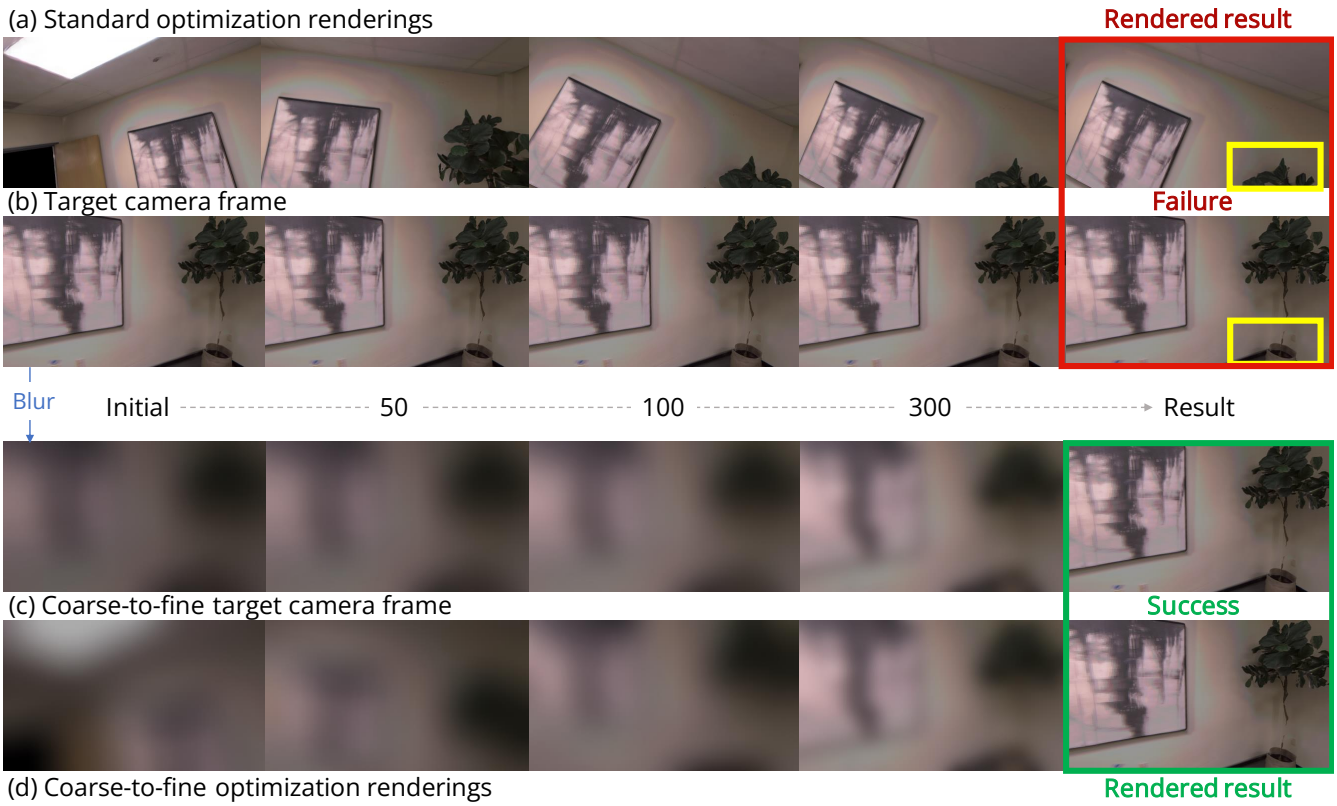


Fig. 2: Visualization of the camera pose alignment process induced by iterative optimization of photometric loss between intermediate renderings and target images for standard (a)-(b) and coarse-to-fine (c)-(d) strategies. Standard optimization decsribed with (a)-(b) leads to convergence to a sub-optimal solution: it does not manage to escape the local minima caused by the sub-optimal overlap between the intermediate rendering and the target query image (highlighted with yellow) resulting to an unsuccessful image alignment. On the contrary, smoothing the image gradients with our coarse-to-fine approach (c)-(d) allows us to avoid being trapped in local minima and converge to the correct camera pose.

from such synthetic data. Nevertheless, we also evaluate our method on 2 real scenes from the Deep Blending dataset [26] and show the coherence of the obtained real-data results with the synthetic ones.

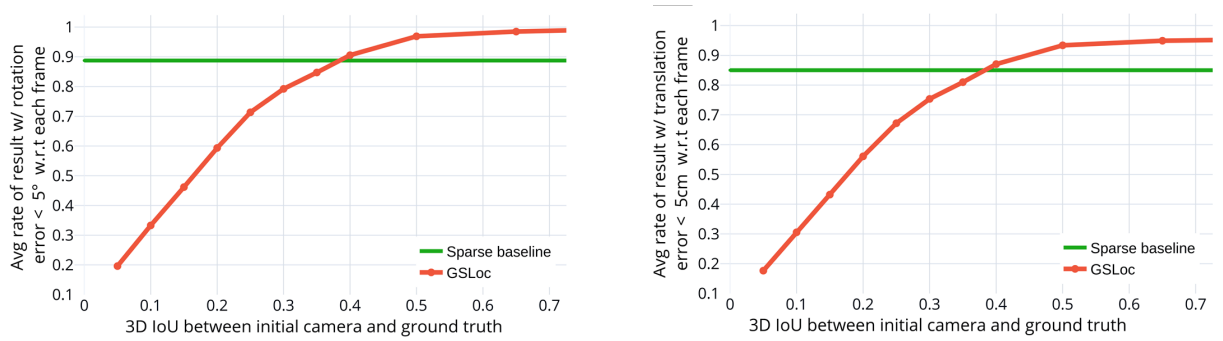
A. Synthetic Data and Initial Camera Analysis

1) Setting: We perform our study using 5 scenes from the Replica [25] dataset. For each scene we assume that the camera intrinsics are known and utilize a base trajectory of approximately 200 frames that describe in detail the environment. These frames are accompanied by a depth estimated point cloud and camera poses.

Next, for each scene we capture a diverse set of 32 test query frames that we then use for the visual localization evaluation. For each of these query frames we randomly generate 16 different pose initializations related to each one of 10 different 3D IoU levels lying in the range between 0.05 and 0.65. Overall, we utilize 5 scenes - 32 query frames - 10 IoU levels - 16 different initalization poses. In total this amounts to approximately 25k different localization tasks, which allow us to thoroughly validate the performance of GSLoc, its strengths and possible limitations.

In order to not significantly deviate from a realistic setup, we process the base images using SfM to estimate the camera poses and the 3D points, with hloc serving as the SfM reconstruction method. Specifically, we first extract local feature descriptors with SuperPoint [27] and global image descriptors with NetVLAD [11]. Based on the similarity of NetVLAD decriptors we then estimate the top 5 neighbors for each image. After that we use SuperGlue to perform feature matching of each image with it's top 5 neighbors. The rest of the reconstruction is performed with COLMAP's [5] incremental mapping.

This SfM reconstruction is a priori not perfect due to the inherent flaws of such methods. Therefore, in order to minimize the impact of these map reconstruction errors on visual localization, we replace all the poses and 3D points in the SfM reconstruction with the ground truth ones. One can interpret this choice as if we had a flawless SfM reconstruction. With this strategy, we manage to avoid overly refined experiments setup, keeping the evaluation clean but still realistic. After this, we use this clean SfM reconstruction both for learning the 3DGS scene map representation and as a database for the sparse matching based visual localization baseline. The localization baseline that we use for comparison consists of the following steps. For the query image I_q we perform feature extraction with SuperPoint [27]. Then we extract its NetVLAD [11] global descriptor, find the top 5 neighbors from the database and perform feature matching



(a) Rotation results

(b) Translation results

Fig. 3: Quantitative results of GSLoc on synthetic scenes from Replica [25] dataset compared with sparse feature-matching baseline. Provided results show the dependency between obtaining the correct pose with GSLoc and the proximity of the initial camera frame to the target one. With the increase of the frames' proximity, GSLoc first reaches and then surpasses the baseline. We report the results separately for rotation (a) and translation (b) pose components.

with SuperGlue. This gives us 2D-3D correspondences that we use in PnP RANSAC to estimate the absolute pose. The final pose of the query image is obtained after the non-linear refinement with the Levenberg-Marquardt algorithm.

2) Results: For all the 5 scenes, we have in total 160 test query frames to localize. Following [15], we identify the localization as successful if the resulting pose error is less than 5 degrees for rotation and 5cm for translation. While for the baseline the result is binary: either success or failure, for GSLoc method we separately evaluate each of 10 IoU proximity levels trying to localize each frame with 16 different initializations. Hence, for each IoU level the result of localizing each of the 160 test frames is not binary but the 0-1 ratio describing the percentage of successful results for 16 different initialization per query frame. By averaging this ratios w.r.t to all query frames, we evaluate the efficiency of our method and present the results along with the baseline comparison in Fig. 3.

These results indicate that with an increased proximity of the initial camera pose the 3DGS based visual localization improves its results getting close and outperforming the sparse feature based localization after reaching the threshold of ~ 0.4 3D IoU. Although, it may seem that the baseline results are not perfect, it has in fact managed to solve most of the localization tasks, failing only for extreme featureless images. While such cases are almost impossible to solve with sparse methods, GSLoc handles them well due to its dense alignment nature. Based on the above, we conclude that the 3DGS based map representation used by GSLoc is suitable for solving visual localization tasks and can lead to competitive results. We also show that as any other visual localization methods, GSLoc requires a certain level of proximity of the initial camera pose used for optimization. We elaborate on this aspect of the problem in the next section.

B. Enhanced Camera Initialization with Image Retrival on Extended Image Base

The camera initialization for visual localization problems is typically obtained by solving the Image Retrieval task. A common way to do this is by finding the closest image descriptors in an image database. For instance, such approach corresponds to the preliminary step used in our sparse baseline. In this section we investigate whether such resulting poses are suitable enough for being used as initialization within GSLoc.

1) Setting: We follow the same setup as the one used in the previous experiment. The only major difference here is that instead of investigating different 3D IoU levels for initial camera poses, we simply follow our sparse baseline and initialize the pose with those that have 5 closest NetVlad [11] descriptors among images in our base trajectory. Here we switch on the binary result classification, assuming the localization of the query to be successful if at least one out of five camera initialization lead GSLoc to the correct final solution.

Further, we utilize the learned 3DGS scenes to extend our base trajectory image databases used for the image retrieval task. We carefully capture an additional set of camera frames extending the existing database and increasing its diversity and scene comprehension.

2) Results: We report the results for the experimental setup described above in Fig. 4. We describe the percentage of the successfully localized frames for each individual room of Replica [25] dataset. Detailing the previously reported results, the baseline method initialized with closest descriptors manages to achieve the correct result for most of the cases, failing only for frames that are dominated by empty space as in the "office 2" scene.

In contrast, GSLoc initialized with the original base map dense matches, exhibits a slightly worse performance, on average having less successfully localized results. What is interesting is that extending the image base with the renderings actually helps to improve its performance and on average increases the GSLoc success rate by 10% bringing it closer to the baseline. Investigating the reasons for this improvement, we figured out that the median 3D IoU for the camera initializations estimated with dense matching on the original database is ~ 0.3 and increases to ~ 0.4 with the database rendering extension. This is an indirect confirmation

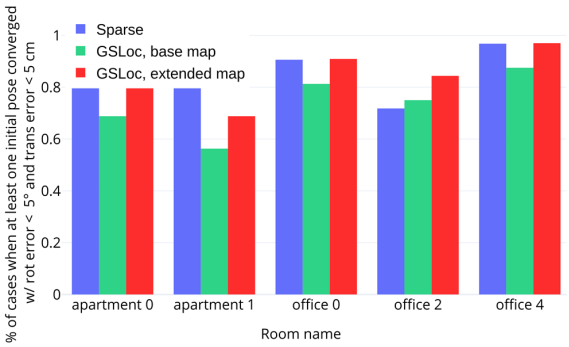


Fig. 4: Quantitive results on the synthetic scenes from Replica [25] dataset. Enhancing the GSLoc camera initializations obtained by the image retrieval with the rendering-extended imagebase leads to consistent success rate improvement proving the efficiency of the proposed method.

of the results previously reported in Fig. 3. This experiment proves the efficiency of the proposed base map extension technique and reveals a potential of utilizing 3DGS-based methods for successfully solving the image retrieval task.

Concluding the discussion of the results obtained on synthetic data, we report the average numerical pose estimation errors of successfully localized frames for all previously reported experiments in Tab. I. We show that the successful localization with GSLoc leads to comparable or even smaller pose errors compared with the sparse baseline method. Besides that, we also show that successfully localized frames also achieve higher visual metrics, which directly relate to the quality of the pose estimation and therefore might be used as a criterion for deciding if localization has been successful.

Method	Rotation error, deg		Translation error, cm		Mean PSNR, dB
	Mean	Median	Mean	Median	
Sparse baseline	0.098	0.058	0.498	0.295	-
GSLoc, init with 5 desc.	0.091	0.054	**0.487**	0.286	**35.52**
GSLoc, avg for all IoUs	**0.078**	**0.035**	0.534	**0.264**	35.36

TABLE I: Average pose estimation errors for successfully localized frames. GSLoc leads to comparable or even smaller pose errors compared with the sparse baseline method.

C. Real Data Results

In the last experiment, we show that the results obtained for the synthetic data remain valid for the real-world datasets. To do so, we conducted similar experiments with camera initialization with NetVLAD descriptors and image base map rendering extension for 2 real indoor scenes from the Deep Blending dataset [26]. Each scene is represented with a comprehensive set of arbitrary posed photos.

We form a test query set by taking each 8^{th} image of the set. We use the rest of the frames as the image base firstly for the SfM reconstruction and 3DGS training and secondly for the initial pose estimation with closest descriptors. Following the same experimental design, we again extend the image base with 3DGS renderings and showcase that the effectiveness of this technique remains valid also for the real scenes. We report our results in Fig. 5. We observe that the results achieved for the real scenes are consistent with those of the synthetic scenes. This is a strong

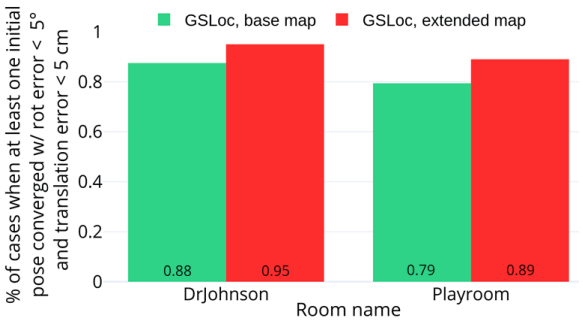


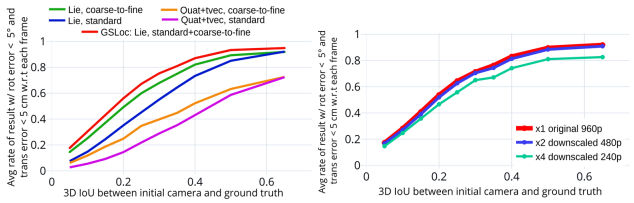
Fig. 5: Quantitive results on the real scenes from Deep Blending [26] dataset. Enhancing the GSLoc camera initializations obtained by the image retrieval with the rendering-extended imagebase leads up to 10 % success rate improvement matching the observations obtained with synthetic data.

indication of the suitability of GSLoc for real-world visual localization.

VI. ABLATION STUDY

We perform a ablation of our GSLoc method by re-running the experiment described in section V.A utilizing different optimization strategies and pose parametrizations. We summarize the results in Fig. 6(a) showing the advantage of the proposed two step standard-coarse-to-fine GSLoc on manifold optimization over other methods.

Furthermore, we estimate how the rendering resolution affects the GSLoc localization results, describing the details in Fig. 6(b). While we run all our experiments on an image resolution of 960x540 pixels we see that the decrease of image size results in almost identical results for 2x downscaling and starts suffering localization degradation only at a 4x image downscaling.



(a) Method ablation (b) Resolution ablation

Fig. 6: Ablation study on optimization strategy and pose parametrization (a) and rendering resolution (b). Proposed two step standard+coarse-to-fine GSLoc optimization on manifold outperforms other methods and allows running on 2x downscaled images without results degradation.

VII. LIMITATIONS AND FUTURE WORK

While GSLoc shows promising results for visual localization, there are a few issues and limitations that we did not address in this work. An important one is the running-time efficiency of the method, which has not been optimized. Both, the multi-step sequential optimization as well as the use of a first order gradient descent method can lead to an increased execution time. In the future, we plan to improve the time efficiency of GSLoc and mitigate the current limitations by utilizing second order optimization algorithms.

VIII. CONCLUSIONS

We have presented GSLoc - a novel visual localization technique based on 3D Gaussian Splatting environment map representation. We have demonstrated both on synthetic and real data that our method is capable of performing accurate camera pose estimation. We have confirmed it via a comprehensive convergence analysis of various camera initializations and parametrizations. We have thoroughly explored the convergence limitations due to non-convexity of the photometric loss and proposed a coarse-to-fine strategy to mitigate this issue. Finally, we have proposed an effective way to improve the localization results by enhancing the GSLoc camera initialization, which is obtained by image retrieval with a refined image base that is extended with 3DGS-rendered camera frames.

APPENDIX

IX. IMPLEMENTATION AND RUNTIME DETAILS

We modify the original CUDA-based implementation [18] of the differentiable renderer enabling camera pose-related gradients. We solve optimization using Adam [24] optimizer for 2000 steps or until convergence when the loss change is smaller than 10^{-5} for 3 consecutive iterations. On average, for cases that achieve a successful outcome the number of necessary iterations may vary between 100-300 iterations and take around 5-15 seconds to converge on a modern GPU. The optimization learning rate starts with 10^{-2} and exponentially decays with iterations to 10^{-5}. The Gaussian blur is applied for the first 1000 iterations, its kernel covariance δ_j decays linearly from 10^{-1} to 10^{-4}, its kernel size L is 200 pixels for standard resolution experiments and is decreased for resolution ablation according to the image downscale factors. For the coarse-to-fine optimization strategy we decide that the localization successfully converged and should not be restarted if the rendered image PSNR has reached 25 dBs.

ACKNOWLEDGEMENTS

The work was supported by the Analytical center under the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021

REFERENCES

[1] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, vol. 1, 2015, p. 1.

[2] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger *et al.*, "Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4695–4702.

[3] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 48–55.

[4] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.

[5] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[6] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[8] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.

[9] J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel, "Direct sparse mapping," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1363–1370, 2020.

[10] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[11] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges, "Category level object pose estimation via neural analysis-by-synthesis," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 139–156.

[13] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 710–10 719.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[15] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "inerf: Inverting neural radiance fields for pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.

[16] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-nerf: Monte carlo localization using neural radiance fields," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4018–4025.

[17] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6229–6238.

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[19] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[20] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[21] H. Matsuki, R. Murai, P. H. J. Kelly, and A. J. Davison, "Gaussian Splatting SLAM," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[22] Y. Sun, X. Wang, Y. Zhang, J. Zhang, C. Jiang, Y. Guo, and F. Wang, "icomma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching," *arXiv preprint arXiv:2312.09031*, 2023.

[23] K. Botashev and G. Ferrer, "Analytical jacobian approximation for direct optimization of a trajectory of interpolated poses on se (3)," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 9198–9204.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[25] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[26] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 6, pp. 1–15, 2018.

[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.