International Journal of Natural and Engineering Sciences
E-ISSN: 2146-0086 17(1): 1-3, 2023
Research Article

RGB-Based 3D Reconstruction: An Investigation of SfM and 3D Gaussian Splatting Performances of SIFT, Superpoint, and ALIKED Methods

Mehmet Akış[1,a], Ramazan Tekin[1,b,*]

[1]*Batman University, Faculty of Engineering and Architecture, Department of Computer Engineering, Batman, Turkey*
*Corresponding author: ramazan.tekin@batman.edu.tr

(Received: 7 May 2026 / Accepted: 27 June 2026)

a: ORCID 0009-0002-9998-9682, b: ORCID 0000-0003-4325-6922

Abstract

The three-dimensional (3D) reconstruction of indoor scenes from two-dimensional (2D) RGB images is a fundamental research topic in the field of computer vision. In this study, the effects of classical (SIFT) and learned (SuperPoint, ALIKED) local feature extraction methods, along with learned matching (LightGlue) algorithms, on Structure-from-Motion (SfM) and the final 3D Gaussian Splatting (3DGS) performance in the indoor 3D reconstruction process are comparatively investigated. The findings demonstrate that while learned algorithms provide more consistent matches and stable camera poses, the denser sparse point cloud generated by the SIFT-based classical approach offers higher visual quality (PSNR, SSIM, LPIPS) in the 3DGS stage.

Keywords: 3D gaussian splatting, structure-from-motion, local feature extraction, image matching, indoor reconstruction

Introduction

The three-dimensional (3D) reconstruction of scenes from two-dimensional (2D) images captured by cameras plays a critical role in numerous fields, including autonomous navigation, augmented reality, and the digitization of cultural heritage. Traditionally, this problem is addressed using Structure-from-Motion (SfM) algorithms, which establish geometric relationships between photographs to simultaneously estimate camera poses and the sparse 3D structure of the scene [1]. However, the success of the SfM pipeline is directly dependent on the accurate detection of distinctive local features in the scene and their correct matching across images.

Indoor scenes, in particular, present a challenging test environment for 3D reconstruction due to the presence of low-texture surfaces (flat walls and floors), repetitive architectural patterns, reflective materials, and complex lighting conditions. For many years, the SIFT algorithm, which is robust to scale and rotation changes, has been considered the standard approach for this purpose [2]. However, conventional approaches may fail to find sufficient correspondences in the presence of difficult lighting and textureless regions, resulting in incomplete reconstructions. In recent years, advances in deep learning-based computer vision have introduced powerful alternatives for local feature extraction, such as SuperPoint [3] and ALIKED [4], which surpass the limitations of classical methods.

Meanwhile, Neural Radiance Fields (NeRF) [5] and the more recently proposed 3D Gaussian Splatting (3DGS) method [6] have revolutionized scene representation and novel view synthesis. 3DGS explicitly models the scene using 3D Gaussian ellipsoids and is capable of synthesizing high-quality images in real time. However, the success and convergence speed of 3DGS optimization are strongly dependent on the quality of the initial point cloud and estimated camera poses provided by SfM [6].

The aim of this study is to comprehensively evaluate the effects of early-stage feature extraction and matching choices (SIFT, SuperPoint, ALIKED, and LightGlue) on SfM performance and the final 3DGS scene representation quality in the 3D reconstruction process from RGB images captured in indoor scenes.

Related Work

Indoor 3D reconstruction from RGB images consists of sequential stages including local feature extraction, image matching, camera pose estimation, and scene representation. In this section, we first address the unique challenges of indoor scenes, then trace the evolution of feature extraction and matching methods, and finally examine the dependency of 3DGS on SfM.

Challenges of 3D Reconstruction in Indoor Scenes

Producing 3D reconstructions from RGB images in indoor scenes is a considerably more challenging optimization problem compared to outdoor scenarios, due to textureless surfaces, repetitive patterns, symmetric structures, and restricted fields of view [7, 8]. Finding reliable local correspondences becomes difficult in regions with weak distinctive details, such as walls, ceilings, and furniture surfaces [7]. This limitation reduces inter-image matching quality, leading to errors or model breaks in the camera pose estimation and 3D triangulation stages of the SfM pipeline [1]. Pataki et al. [8] addressed this issue by integrating surface priors into the SfM pipeline, thereby significantly improving model stability in challenging indoor environments.

Local Feature Extraction Methods

Local feature extraction is the stage at which distinctive pixels are mathematically described so that the same physical point can be located across images captured from different viewpoints. The SIFT algorithm served as the standard algorithm for SfM systems for many years, owing to the descriptors it constructs using gradient changes in the image [2]. With the advancement of deep learning, SuperPoint was introduced, performing both interest point detection and descriptor generation in a self-supervised manner within a single convolutional neural network (CNN) [3]. Among more recent approaches, ALIKED aims to adapt local features to shape deformations using deformable transformations, offering high performance with a lighter network architecture [4]. Comprehensive comparative studies have demonstrated that these methods exhibit clear advantages over classical approaches on wide-baseline images [9].

Image Matching Approaches

Establishing correct correspondences between extracted local features is of vital importance for camera pose estimation [1, 2]. Classical matching methods compute the Euclidean distance between descriptors and apply strict filtering strategies such as the ratio test [2]. Learned approaches additionally consider contextual relationships between features. SuperGlue, which employs graph neural networks (GNNs), pioneered the field by matching descriptors via attention mechanisms [10]. LightGlue addressed the same problem with an adaptive, more computationally efficient architecture, both accelerating the matching process and improving its accuracy [11]. Sarlin et al. [12] presented a study comprehensively documenting the practical impact of these improvements on camera localization through a back-propagation-based feature learning approach.

HLoc and COLMAP-Based Structure-from-Motion Approaches

The goal of SfM pipelines is to recover camera orientations from 2D correspondences and reconstruct the sparse point cloud of the scene. COLMAP is a leading open-source SfM framework that integrates feature extraction, geometric verification, incremental registration, and bundle adjustment [1]. HLoc (Hierarchical Localization) provides a modular frontend that integrates modern learned feature extraction and matching networks (e.g., SuperPoint, LightGlue) into the SfM pipeline [13]. Routing robust matches obtained via HLoc into COLMAP improves model stability in low-texture indoor environments [1, 13].

3D Gaussian Splatting-Based Scene Representation

Although Neural Radiance Fields (NeRF) models have been widely used for novel view synthesis in recent years, the ray-marching operations in NeRF considerably slow down training and rendering times [5]. 3D Gaussian Splatting (3DGS) represents the scene using explicit 3D Gaussian ellipsoids rather than a volumetric grid [6]. This Gaussian-based representation enables real-time, high-quality rendering through a differentiable rasterizer. However, 3DGS requires initialization at locations where actual surfaces exist in the scene rather than in empty space; consequently, the quality of the sparse point cloud from the SfM stage is a direct

determinant of 3DGS success. This connection is also addressed as a central research question in derivative works aimed at improving initialization density, such as 2DGS [14].

Several recent studies have directly targeted the initialization bottleneck identified in our work. SparseGS [15] proposed a method for training 3DGS from as few as twelve input images by augmenting the sparse SfM point cloud with depth priors and floater pruning strategies, demonstrating that initialization density critically governs reconstruction completeness under sparse-view conditions. DNGaussian [16] introduced depth-regularized Gaussian optimization, coupling monocular depth estimates with the Gaussian training process to compensate for the geometric incompleteness of SfM-derived point clouds, thereby achieving competitive rendering quality with significantly reduced initialization density. More recently, feed-forward approaches such as pixelSplat [17] have explored the elimination of the SfM stage entirely by predicting Gaussian parameters directly from image pairs, representing a fundamentally different paradigm that bypasses the initialization dependency altogether. Collectively, these works underline that the quality and density of the initial point cloud - the central variable examined in our study - is a recognized and actively investigated bottleneck in the 3DGS literature.

Materials and Methods

In this study, three different methods were comparatively evaluated for producing indoor 3D reconstructions from RGB images. The matching outputs from these methods were fed into the Structure-from-Motion stage, and the resulting sparse reconstruction outputs were used as initialization data for 3D Gaussian Splatting. In this way, the effects of the selected approaches on different stages of the reconstruction pipeline were jointly examined.

Dataset

The dataset used in this study consists of 230 RGB images acquired from a closed indoor classroom environment. The images were captured using a Samsung Galaxy S22 smartphone camera with the model number SM-S901E. All images were stored in JPEG format with a resolution of 4000 × 3000 pixels, corresponding to approximately 12 MP. Captures were performed in standard photo mode using a handheld camera from different viewpoints and positions within the scene.

The images represent a challenging indoor scenario containing low-texture wall and ceiling surfaces, repetitive classroom desk arrangements, and variable lighting conditions. The scene was illuminated by a combination of natural window light and existing artificial indoor lighting. During image acquisition, the approximate distance between the camera and the scene varied between 0.5 m and 6 m. The smartphone camera's automatic focus and automatic exposure settings were used throughout the capture process.

Captures were performed using a sequential, overlapping frame strategy, ensuring at least 30% overlap between each image pair. No separate train/test split was applied in this study; therefore, all images were used during the reconstruction and training stages. All methods were evaluated on the same image set under fixed evaluation conditions.

Compared Methods

The first method compared in this study is the classical COLMAP pipeline, in which local feature extraction and matching are performed using SIFT under high-quality settings [1, 2]. In the second method, local features are extracted from images using SuperPoint (maximum 4096 keypoints), and the resulting descriptors are matched using LightGlue [3, 11]. In the third method, local feature extraction is performed with ALIKED (maximum 4096 keypoints), followed by matching using LightGlue [4, 11]. Thus, the SIFT-based classical approach and the learned approaches represented by SuperPoint + LightGlue and ALIKED + LightGlue are evaluated on the same dataset. This comparison aims to investigate the effects of early-stage local feature and matching strategy choices on subsequent stages of the reconstruction pipeline.

Experimental Setup

The experiments were conducted in two different computational environments. The feature extraction, feature matching, and SfM reconstruction stages prior to 3D Gaussian Splatting were performed on a local system

equipped with an 11th Gen Intel(R) Core(TM) i5-11400H processor with 6 cores and a 2.70 GHz base clock, 16 GB of 3200 MHz RAM, and a 465.76 GB SSD storage device (MSI M450 500GB). The primary GPU used in these stages was an NVIDIA GeForce RTX 3050 Laptop GPU with 4 GB of dedicated VRAM. The local experiments were conducted on the Windows 11 Pro 64-bit operating system.

Due to the higher GPU memory requirement of the 3D Gaussian Splatting training stage, this stage was carried out in the Google Colab environment using an NVIDIA Tesla T4 GPU with 16 GB of GPU memory. Thus, while the preprocessing stages, including feature extraction, feature matching, and SfM reconstruction, were completed on the local hardware, the Gaussian-based scene representation was trained in the Colab environment.

The software environment consisted of Python 3.10, PyTorch 2.5, CUDA 12.1 in the local environment, CUDA 13.0 in the Colab environment, and COLMAP 4.1.0. The HLoc framework was used for SuperPoint feature extraction and LightGlue matching. ALIKED feature extraction and LightGlue matching were performed through the graphical user interface of COLMAP 4.1.0. The 3D Gaussian Splatting implementation was based on the official repository by Kerbl et al. [6]. All methods were evaluated under the same two-stage computational setup to ensure a fair and reproducible comparison. In this setup, the preprocessing stages were consistently performed on the local system, while the 3D Gaussian Splatting training stage was consistently carried out in the Google Colab environment.

Local Feature Extraction and Image Matching Stage

In this stage, local features are first extracted from each image to enable the establishment of correspondences between images, candidate image pairs are then identified, and point matches are generated between these pairs. In the classical method, these operations are carried out through COLMAP's own pipeline, whereas in the learned methods, feature and match outputs are prepared through HLoc-based preprocessing [1, 13]. In this way, the feature and match data to be used in the subsequent Structure-from-Motion stage are obtained for each method.

Structure-from-Motion (SfM) Stage

Based on the obtained correspondences, the position and orientation matrices (R, t) of the cameras and the 3D world coordinate points (X) are computed. In the SfM stage, reprojection error is minimized. Formally, given the 2D observation x_{ij} of a 3D point X_j in the i-th camera, the objective function is as follows:

$$\min_{R,t,X} \sum_{i,j} \rho\left(\left\|x_{ij} - \pi\left(R_i X_j + t_i\right)\right\|^2\right)$$

Equation 1

Here, π denotes the camera projection function, and ρ denotes a robust loss function against outliers. This nonlinear optimization (Bundle Adjustment) is solved via COLMAP, yielding the sparse point cloud that serves as the primary input for 3DGS [1].

3D Gaussian Splatting Stage

Each 3D point from SfM is initialized as a Gaussian ellipsoid in the 3DGS framework. A 3D Gaussian distribution is expressed by the following equation:

$$G(x) = \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Equation 2

Here, μ denotes the center (mean) position of the Gaussian and Σ denotes the covariance matrix. To ensure the differentiability of the optimization, the covariance matrix is decomposed into a rotation (R) and a scaling (S)

matrix and parameterized as $\Sigma = RSS^{T}R^{T}$ [6]. During training, the position (μ), covariance (Σ), opacity (α), and Spherical Harmonics (SH) coefficients of each Gaussian element, which capture view-dependent color, are optimized via a pixel-based photometric rendering loss computed against the training images. The cameras and sparse points obtained from the SfM stage constitute the input layer of this process [6].

Evaluation Metrics

In this study, evaluation is conducted at three levels. First, at the local feature and matching level, the number of keypoints, the number of matches, and the correspondence structure obtained after geometric verification are examined. Second, at the Structure-from-Motion stage, the number of registered images, the sparse point cloud, and model completeness are assessed [1]. Third, the scene representation obtained at the 3D Gaussian Splatting stage is compared. 3DGS outputs are evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) metrics, which are adopted as standard quality measures in the 3DGS literature [6]. Higher PSNR and SSIM values, along with lower LPIPS values, indicate a more successful scene representation.

Results and Discussion

In this section, the findings obtained from the methods compared in the study are evaluated in terms of local feature extraction, image matching, Structure-from-Motion, and 3D Gaussian Splatting stages. The performance of the methods at the feature and matching level is first examined, followed by a discussion of the impact of these differences on sparse reconstruction and the final scene representation.

Local Feature Extraction Results

The keypoint statistics produced by the methods are summarized in Table 1. The SIFT-based classical approach produced an average of 10,569 keypoints per image, forming a considerably denser yet more variable (higher standard deviation) feature set compared to the two learned methods. ALIKED and SuperPoint remained at averages of approximately 2,980 and 2,649 keypoints, respectively, yielding more constrained but more uniformly distributed results. ALIKED achieves higher average and median keypoint counts compared to SuperPoint; however, its higher standard deviation indicates greater variability in the number of keypoints produced per image (Figure 1).

Table 1. *Local Feature Extraction Statistics.*

Method	No. of Images	Avg. Keypoints	Median Keypoints	Min	Max	Std. Dev.	Images with <1000 Keypoints
SIFT + COLMAP	230	10569.25	10331.50	4513	16572	2121.06	0
ALIKED + LightGlue	230	2980.40	2995.00	894	4094	787.91	1
SuperPoint + LightGlue	230	2648.98	2687.00	822	4096	668.45	2

Figure 1. *Comparison of local feature distributions obtained with a) SIFT, b) ALIKED, and c) SuperPoint methods on sample images from the same scene.*

Image Matching and Geometric Verification Results

The fact that feature density does not directly translate into matching success is clearly evident in the Zero-Match and Zero-Inlier statistics in Table 2. Despite its high feature count, SIFT failed to find any geometrically valid correspondences (zero-inlier) in 20,416 out of 26,335 potential image pairs. The approaches using deep learning-based LightGlue, however, produced more balanced inlier counts despite matching over far fewer features. In particular, the ALIKED + LightGlue approach achieved the lowest Zero-Match (3,980) and Zero-Inlier (12,233) counts as well as the highest average inlier count (75.03), making it the method that produced the most consistent geometric match structure between image pairs (Figure 2).

Table 2. *Matching and Geometric Verification Results.*

Method	Total Pairs	Avg. Raw Matches	Median Raw Matches	Zero-Match Pairs	Avg. Inliers	Median Inliers	Zero-Inlier Pairs
SIFT + COLMAP	26335	32.29	0.00	18115	67.14	0.00	20416
ALIKED + LightGlue	26335	106.67	34.00	3980	75.03	16.00	12233
SuperPoint + LightGlue	26335	89.52	18.00	10981	59.98	0.00	15738

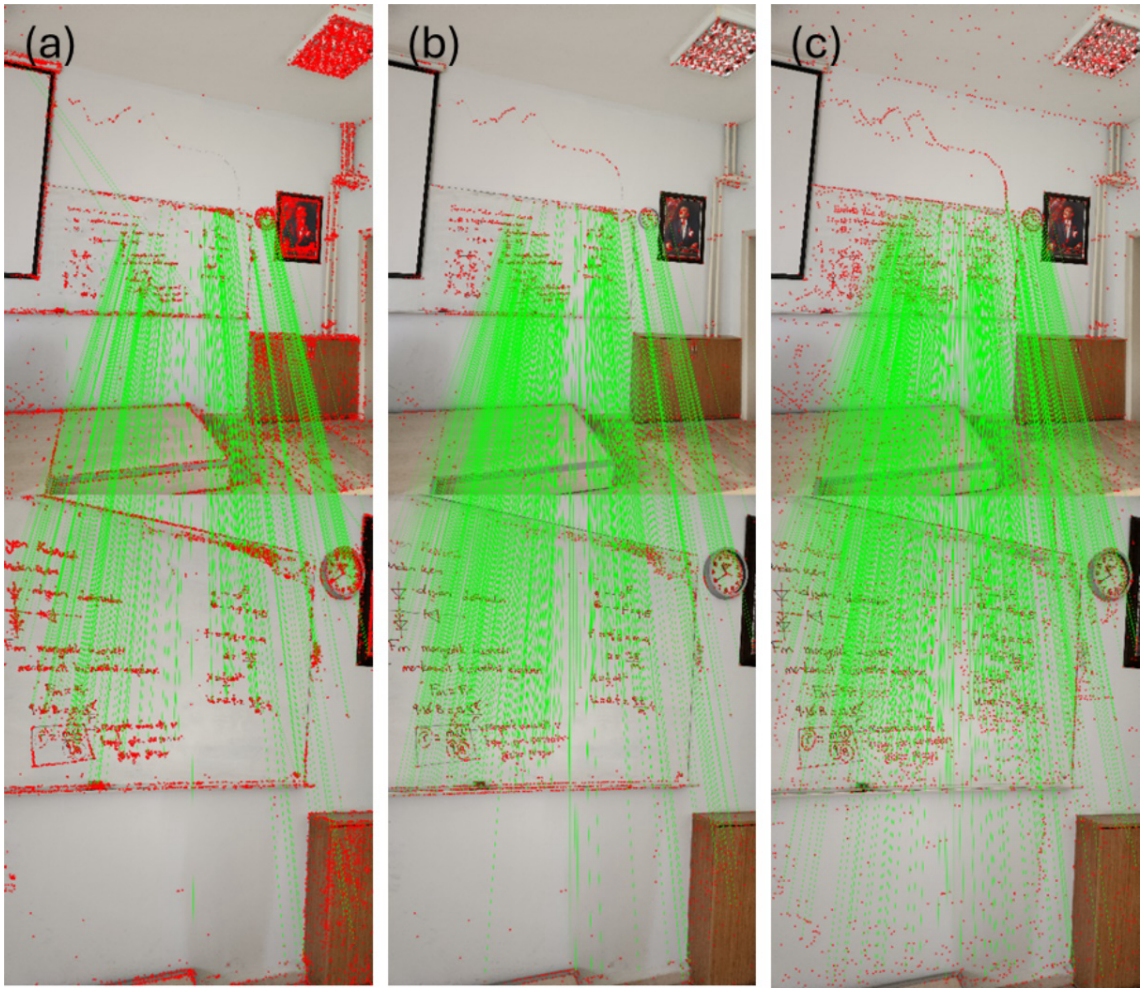

Figure 2. *Comparison of matching results obtained with a) SIFT, b) ALIKED + LightGlue, and c) SuperPoint + LightGlue on the same image pairs.*

Structure-from-Motion Results

As shown in Figure 3, the Structure-from-Motion results reveal notable differences between the methods in terms of sparse reconstruction density, registration performance, and geometric quality. As seen in Table 3, the SIFT-based classical COLMAP pipeline produced the highest number of 3D points (93,022) and achieved the lowest reprojection error (1.247). The ALIKED + LightGlue method produced a denser sparse reconstruction among the learned methods, with a 99.1% registration rate and 55,491 3D points. Its average track length of 4.05 was the highest among all methods, indicating a robust structure for inter-image point tracking. The SuperPoint + LightGlue method yielded results close to ALIKED with an average track length of 3.91, but remained more limited in geometric consistency due to its lower 3D point count (39,870) and higher reprojection error (1.437).

Table 3. *SfM Sparse Reconstruction Statistics.*

Method	Reg. Rate	No. of 3D Points	Total Observations	Avg. Track Length	Med. Track Length	Avg. Reproj. Error	Med. Reproj. Error	Avg. Triangulated Pts./Image	Med. Triangulated Pts./Image
SIFT + COLMAP	99.6%	93022	314048	3.38	2	1.247	1.246	1371.39	1242
ALIKED + LightGlue	99.1%	55491	224809	4.05	3	1.383	1.437	985.99	953
SuperPoint + LightGlue	98.7%	39870	156041	3.91	3	1.437	1.512	687.41	642

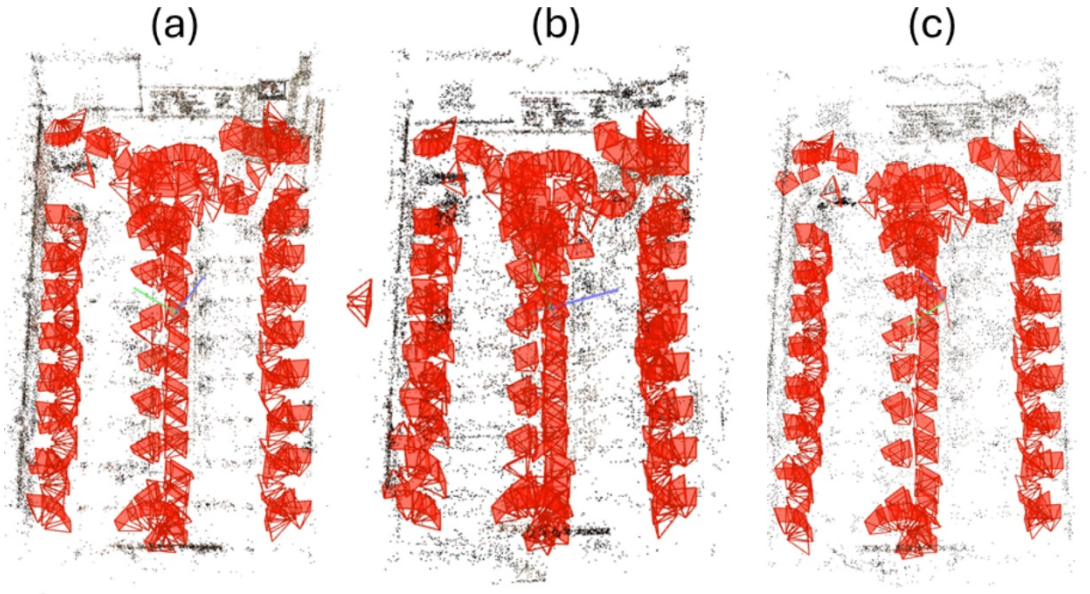

Figure 3. *Comparison of Structure-from-Motion results. a) SIFT, b) ALIKED + LightGlue, and c) SuperPoint + LightGlue.*

3D Gaussian Splatting Results

The 3D Gaussian Splatting results reveal the impact of the compared methods on the final stage of scene representation (Figure 4). As seen in Table 4, the SIFT-based classical COLMAP pipeline achieved the best results across all metrics (PSNR=27.11, SSIM=0.8714, LPIPS=0.3617). Despite their lower keypoint and sparse point counts, ALIKED + LightGlue and SuperPoint + LightGlue produced results close to the SIFT-based classical COLMAP pipeline. The ALIKED + LightGlue method delivered the best performance among learned methods with PSNR=26.74, SSIM=0.8686, and LPIPS=0.3658, while SuperPoint + LightGlue exhibited slightly lower but comparable performance with PSNR=26.51, SSIM=0.8621, and LPIPS=0.3714.

Evaluating these three metrics together indicates that the differences between the methods are limited rather than representing a significant quality degradation.

Table 4. *3D Gaussian Splatting Image Quality Metrics.*

Method	PSNR	SSIM	LPIPS
SIFT-based classical COLMAP pipeline	27.11	0.8714	0.3617
ALIKED + LightGlue	26.74	0.8686	0.3658
SuperPoint + LightGlue	26.51	0.8621	0.3714

To contextualize these results within the broader 3DGS literature, it is instructive to compare the obtained metric values against those reported in related studies. The original 3DGS paper by Kerbl et al. [6] reports PSNR values ranging from approximately 23 dB to 30 dB across standard benchmark scenes, including Tanks and Temples, Deep Blending, and Mip-NeRF 360 datasets, with indoor and bounded scenes generally yielding values at the higher end of this range. The PSNR values obtained in the present study - 27.11, 26.74, and 26.51 for the SIFT-based, ALIKED + LightGlue, and SuperPoint + LightGlue pipelines, respectively - fall within this range and indicate a competitive level of rendering quality for a single-scene indoor evaluation conducted with a handheld capture setup. Similarly, the SSIM values reported here (0.8714, 0.8686, and 0.8621) are consistent with those observed in comparable indoor 3DGS evaluations in the literature. It should be noted, however, that direct numerical comparison with published benchmark results is inherently constrained by differences in scene content, image count, camera configuration, and 3DGS training settings; the published values are therefore presented here as contextual reference points rather than strict baselines.

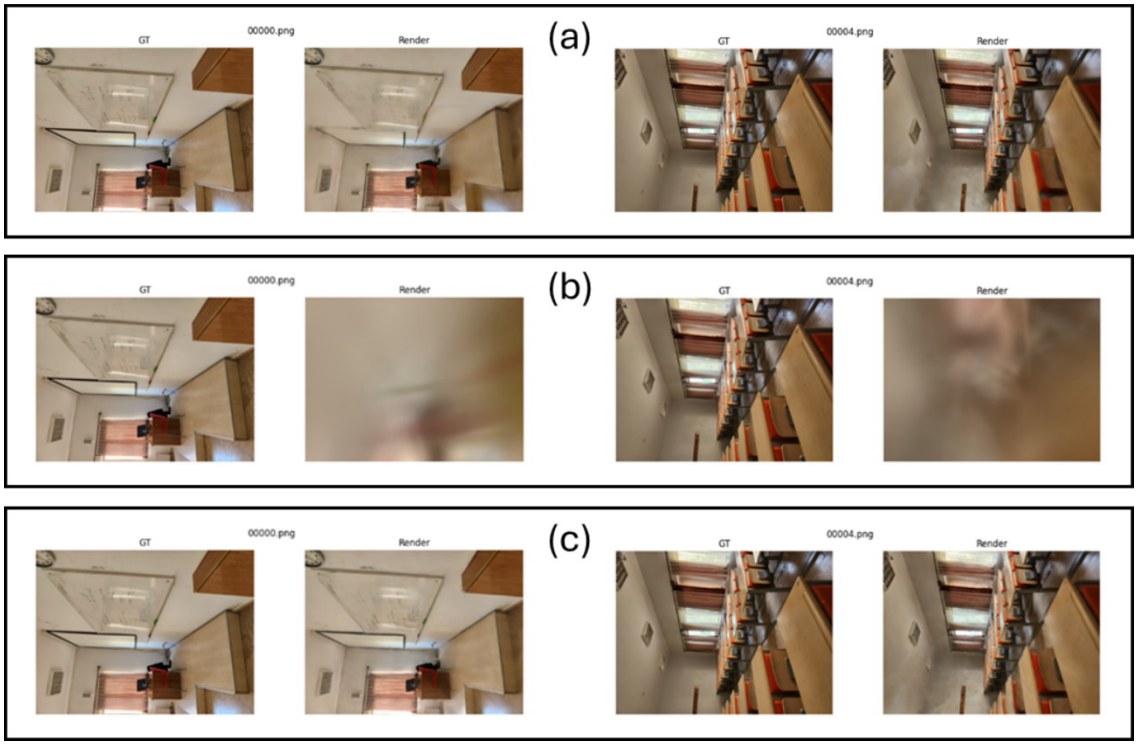

Figure 4. *Comparison of 3DGS rendering results. Rendering outputs obtained with a) SIFT-based classical COLMAP, b) ALIKED + LightGlue, and c) SuperPoint + LightGlue methods.*

Computational Performance

Table 5 presents a comprehensive comparison of the computational performance of the three evaluated pipelines, measured on the experimental hardware described in Section 3.1. All times were recorded end-to-end under identical conditions, with the exception of the ALIKED + LightGlue matching stage, which is discussed separately below.

SuperPoint + LightGlue achieved the shortest total pipeline time of approximately 3 hours 19 minutes, and the highest rendering speed at ≈0.46 FPS (≈2190 ms/frame). SIFT + COLMAP completed the full pipeline in approximately 3 hours 39 minutes, with a rendering speed of ≈0.30 FPS (≈3320 ms/frame). The ALIKED + LightGlue pipeline incurred the longest total runtime of approximately 5 hours 34 minutes, largely attributable to the matching stage.

Notably, the ALIKED + LightGlue matching stage could not be executed on the NVIDIA GeForce RTX 3050 Laptop GPU (4 GB VRAM) due to an ONNX Runtime out-of-memory error. Since reducing image resolution or keypoint count to fit within the available VRAM would alter the experimental configuration and compromise the fairness of cross-method comparisons, these modifications were not applied. Consequently, the ALIKED + LightGlue matching time (2 h 19 min) was measured on a CPU-based system equipped with an 11th Gen Intel Core i5-11400H processor, and this is explicitly noted in Table 5. This hardware limitation is itself a practically relevant finding, as it indicates that the ALIKED + LightGlue pipeline, in its standard configuration, requires GPU resources beyond those available in the present study for the matching stage.

Across all methods, 3DGS training time constituted the dominant computational cost, ranging from approximately 2 hours 29 minutes (SuperPoint + LightGlue) to 2 hours 39 minutes (ALIKED + LightGlue), reflecting the intensive optimization process inherent to Gaussian splatting regardless of the upstream feature pipeline. These results suggest that SuperPoint + LightGlue offers the most computationally efficient end-to-end pipeline among the evaluated methods while maintaining competitive reconstruction quality.

Table 5. Computational performance comparison of the three evaluated pipelines.

Method	Feature Extraction	Matching	SfM	3DGS Training	Rendering	ms/frame	FPS	Total
SIFT + COLMAP	19 min 51 s	8 min 52 s	21 min 14 s	2 h 37 min	12 min 40 s	≈3320	≈0.30	3 h 39 min 54 s
ALIKED + LightGlue	10 min 36 s	2 h 19 min (CPU†)	14 min 33 s	2 h 39 min	11 min 18 s	≈2980	≈0.34	5 h 34 min 27 s
SuperPoint + LightGlue	27 s	31 min 29 s	9 min 50 s	2 h 29 min	8 min 17 s	≈2190	≈0.46	3 h 19 min 03 s

† The ALIKED + LightGlue matching stage was executed on a CPU (11th Gen Intel Core i5-11400H) due to an ONNX Runtime out-of-memory error encountered on the NVIDIA GeForce RTX 3050 Laptop GPU (4 GB VRAM).

Conclusion

In this study, the effects of different feature extraction and matching methods on the final 3D Gaussian Splatting (3DGS) quality were investigated. It is well established in the literature that learned feature matching networks [11] find considerably more robust correspondences than classical approaches [2] in challenging scenes. Our findings confirm that ALIKED and SuperPoint-based methods provide an extremely strong SfM foundation in terms of camera registration and matching stability.

However, the technological superiority of deep learning-based algorithms alone is not sufficient to maximize 3DGS quality. The 3D Gaussian Splatting algorithm requires a set of initialization points with correct geometric placement at the start of optimization [6]. Although the classical SIFT approach has many unmatched points, it generates a volumetrically massive (93,000+) initialization point set, allowing the 3DGS optimization to capture high-frequency fine details (texture) in the scene far more successfully. The ALIKED + LightGlue and SuperPoint + LightGlue methods produced 55,491 and 39,870 sparse 3D points, respectively, and despite their lower initialization density, delivered PSNR results close to the SIFT-based approach in the 3DGS stage.

The observed limited quality difference is attributed not to algorithmic shortcomings of the learned models, but rather to the maximum keypoint constraints (hyperparameter limits) applied during the feature extraction stage. Ultimately, learned AI approaches hold a stability advantage for resolving camera orientations in indoor environments. However, if the goal is to produce a high-fidelity photorealistic scene using 3D Gaussian Splatting, the density of the initial point cloud remains a critical parameter. Hybrid approaches that

simultaneously optimize point cloud density and matching quality represent a promising research direction for unifying these two objectives.

Several concrete directions emerge from these findings for future research. First, integrating Multi-View Stereo (MVS) densification between the SfM and 3DGS stages could bridge the initialization density gap between classical and learned pipelines, enabling learned matching methods to benefit from denser point clouds without sacrificing pose estimation stability. Second, adaptive keypoint limit strategies — in which the number of extracted keypoints is dynamically adjusted based on local scene complexity rather than fixed at a global maximum — could simultaneously improve matching consistency and initialization density for SuperPoint and ALIKED. Third, depth-prior-augmented Gaussian initialization approaches, such as those explored in DNGaussian [16], represent a direct architectural solution to the sparse initialization problem identified in this study and warrant comparative evaluation against the pipelines examined here. Fourth, method-specific tuning of 3DGS hyperparameters — particularly the Gaussian densification threshold and the number of training iterations — may partially compensate for initialization density differences between methods and should be systematically investigated. Finally, extending the evaluation framework to diverse indoor scene types, including corridors, large open halls, and scenes with predominantly specular surfaces, will be essential for assessing the generalizability of the observed trade-offs.

Limitations

The primary limitation of this study is that the evaluation was conducted on a single indoor scene. This choice was deliberate: by fixing the scene, all extraneous variables were held constant, enabling a controlled and fair comparison of the three reconstruction pipelines. However, it also means that the generalizability of the findings to other indoor environments cannot be assumed without further validation. The consistency of the methods needs to be tested across multi-scene scenarios encompassing different lighting conditions, spatial scales, and camera motion profiles. Future work should prioritize evaluation across a diverse set of indoor scenes - including corridors, office rooms, laboratories, and large open halls - to assess whether the observed trade-off between SIFT's initialization density advantage and the matching stability of learned methods holds under varying structural complexity and illumination conditions. Additionally, the hyperparameters in the 3DGS stage (training iterations, Gaussian densification threshold) were held constant. It should be noted that results may differ if these parameters are individually optimized for each method.

Furthermore, the maximum keypoint limit was set to 4,096 for both SuperPoint and ALIKED, following the common configuration of the HLoc framework. Whether higher keypoint limits would yield denser SfM point clouds and, consequently, improved 3DGS rendering quality for learned methods remains an open question; a dedicated ablation study systematically varying this parameter across different limit values (e.g., 1,024, 2,048, 4,096, and 8,192) would provide valuable insight into the sensitivity of the full reconstruction pipeline to this design choice.

Finally, all experiments were conducted as single runs; no repeated-run analysis was performed to quantify the variability of the reported metrics. While the SIFT-based SfM stage is largely deterministic and produces consistent outputs across runs, the 3DGS optimization involves stochastic components that may introduce measurable run-to-run variation in PSNR, SSIM, and LPIPS values. Reporting mean and standard deviation across multiple independent runs would provide a more statistically robust characterization of the results, and is identified as a methodological improvement for future studies.

Moreover, the present study does not incorporate hybrid SfM–3DGS approaches or advanced Gaussian initialization strategies - such as depth-prior augmentation or sparse-view reconstruction methods - into the comparison. As the related work demonstrates, these emerging techniques directly address the initialization density bottleneck identified in our findings, and benchmarking the compared pipelines against such methods represents a natural and important extension of this work.

References

[1] Schonberger JL, Frahm J-M. (2016). Structure-from-motion revisited. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp 4104–4113.

[2] Lowe DG. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of*

Computer Vision. 60(2): 91–110.

[3] DeTone D, Malisiewicz T, Rabinovich A. (2018). SuperPoint: self-supervised interest point detection and description. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp 224–236.

[4] Zhao X, Wu X, Chen W, Chen PCY, Xu Q, Li Z. (2023). ALIKED: a lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*. 72: 1–16.

[5] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. (2021). NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*. 65(1): 99–106.

[6] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*. 42(4): 131–139.

[7] Taira H, Okutomi M, Sattler T, Cimpoi M, Pollefeys M, Sivic J, Pajdla T, Torii A. (2018). InLoc: indoor visual localization with dense matching and view synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp 7199–7209.

[8] Pataki Z, Sarlin P-E, Schönberger JL, Pollefeys M. (2025). MP-SfM: monocular surface priors for robust structure-from-motion. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp 21891–21901.

[9] Jin Y, Mishkin D, Mishchuk A, Matas J, Fua P, Yi KM, Trulls E. (2021). Image matching across wide baselines: from paper to practice. *International Journal of Computer Vision*. 129(2): 517–547.

[10] Sarlin P-E, DeTone D, Malisiewicz T, Rabinovich A. (2020). SuperGlue: learning feature matching with graph neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 4938–4947.

[11] Lindenberger P, Sarlin P-E, Pollefeys M. (2023). LightGlue: local feature matching at light speed. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp 17627–17638.

[12] Sarlin P-E, Unagar A, Larsson M, Germain H, Toft C, Larsson V, Pollefeys M, Lepetit V, Hammarstrand L, Kahl F, Sattler T. (2021). Back to the feature: learning robust camera localization from pixels to pose. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 3247–3257.

[13] Sarlin P-E, Cadena C, Siegwart R, Dymczyk M. (2019). From coarse to fine: robust hierarchical localization at large scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 12716–12725.

[14] Huang B, Yu Z, Chen A, Geiger A, Gao S. (2024). 2D Gaussian splatting for geometrically accurate radiance fields. In: *ACM SIGGRAPH 2024 Conference Papers*. pp 1–11.

[15] Xiong H, Muttukrishnan S, Kerr J, Kanazawa A. (2024). SparseGS: real-time 360° sparse view synthesis using Gaussian splatting. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[16] Li J, Zhang J, Bai X, Zheng J, Ning X, Zhou J, Gu L. (2024). DNGaussian: optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 20775–20785.

[17] Charatan D, Li S, Tagliasacchi A, Sitzmann V. (2024). 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*. pp 19457–19467.