3D Gaussian Splatting in the Dark

Joonkyu Min[1], Junyoung Lee[2] and Yongchan Jo[2]

[1]Department of ECE, Seoul National University, Seoul, Korea
[2]Department of CSE, Seoul National University, Seoul, Korea
{timothy0609, juncong, jych1109}@snu.ac.kr

Abstract

In this paper, we present a method to render realistic novel views from dark and blurry scenes by gaussian splatting. 3D scene representation with Gaussian splats achieves remarkable results. However, existing pipelines easily fail in noisy and blurry images taken in low light. The existing pipelines require point cloud and camera pose initialization from COLMAP, which easily fails in dark and blurry scenes. Also, the noise and blur of the frames create floating Gaussians that degrade the rendering quality. To address this issue, we propose efficiently initializing the camera poses and point cloud using the depth map optimization method. We additionally take advantage of depth map-based initialization by introducing a depth regularization loss in training Gaussian splatting. We verify the proposed method on the LOL-blur dataset and a few custom datasets of dark and blurry scenes. Our approach demonstrates robust reconstruction and geometrical consistency from low-quality images compared to the previous methods.

1. Introduction

Reconstructing 3D space from videos or images has been a hot topic in the field of computer vision for a long time. Recently, Neural Radiance Fields (NeRF) [9] have shown that using synthetic neural networks, realistic novel view synthesis can be achieved given sufficient computing power and sophisticated images. However, due to the NeRF's disadvantage of slow rendering speed, there has been extensive research on alternatives to NeRF. 3D Gaussian Splatting (3DGS) [7] has appeared with remarkable reproduction quality and fast rendering speed, thus real-time rendering can be achieved. 3DGS requires initialization through an SfM (Structure from Motion) model before performing optimization. Traditional SfM models, such as COLMAP [13], work by identifying common 3D keypoints across multiple images and performing feature matching based on these points. However, this approach has limitations in performing on low-quality input images, such as dark scenes containing noise and motion blur. We aim to address these limitations and build a robust 3D scene reconstruction from images taken in low-light conditions.

To handle the major challenges of dark scenes, we decided to utilize various different method of initialization and optimization of the gaussian splatting process. Instead of using feature extraction, we match correspondence by dense point tracking method based on optical flow estimation, with an assumption of continuous frame inputs. We compute the camera pose and point cloud by using a deep-learning-based SfM model. Also, using optimization method from a blur aware method, BAD-Gaussians[17], we could extract sharp Gaussian splatting rendering from blurry images. We furthermore introduce depth map regularization loss, to take advantage of dense matching initialization method, as well as handle the noisy inputs.

2. Related Work

Novel view synthesis Structure from motion (SfM) [16] and Multi-view stereo (MVS) [15] are used frequently to reconstruct the 3D scene with multiple images. COLMAP [13] stands as the state-of-art SfM methods to find the camera pose and sparse 3D keypoints considering the epipolar constraint [5] of images with various views. More recently, neural-network-based 3D reconstruction methods have emerged to take advantage of highly developed deep-learning strategies. Among them, Neural radiance fields (NeRF) [9] is one of the neural-network-based methods which reconstruct the 3D scene with remarkable quality. However, due to the structure of NeRF, slow rendering speed has prompted many researchers to make efforts to achieve real-time rendering. Among them, 3D Gaussian Splatting (3DGS) [7] presented with the fast and high quality with the alpha-blending rasterization. It uses Gaussian

attenuated spherical harmonic splats as the primary primitives. These splats, defined by their position, orientation, and opacity, represent various parts of a scene with high detail.

3D reconstruction using low-quality images 3D reconstruction in low-quality environments is an even more challenging problem. Because it is not only difficult to find 3D keypoints, but also challenging to accurately calibrate the camera poses. RawNeRF [10] enhances NeRF by using linear raw images for training, enabling high dynamic range (HDR) view synthesis and robust scene reconstruction from extremely noisy images captured in low-light conditions. However, this approach is slow and not stable against noise. There are also several existing works on reconstructing 3D gaussians from blurry frames [4] [12] [17]. BAD-Gaussians [17] is a novel approach that uses explicit Gaussian representation to handle severe motion-blurred images and inaccurate camera poses, achieving high-quality scene reconstruction and real-time rendering, outperforming previous deblur neural rendering methods. Furthermore, this method splits a single low-quality image and its camera pose into multiple sharp virtual images and a virtual camera pose trajectory. So, it does not require highly accurate SfM points initialization.

3. Method

3.1. Structure from Motion Initialization

We first process our video frames on CoTracker [6] to compute the patch correspondences across the frames, rather than extract image features such as SIFT. [8] CoTracker [6] is one of the state-of-the-art point-tracking methods, based on a transformer model that tracks dense points jointly across a video sequence, considering their correlation.
Using the robust correspondence results as the input, we replace the original SfM method by using a neural network based SfM method. Our method follows one of the existing works, FlowMap [1], which operates SfM based on depth map optimization by finetuning a pre-trained depth estimator network and obtain SfM outputs as well as the optimized depth maps of each frame. In this method, solving camera intrinsics are done by selecting from the set of candidates $\mathbf{K}_k$ which are obtained with a pinhole camera estimation and discretized set of focal lengths. For each candidate, the equation 2 was used to compute a corresponding set of poses, then the equation 3 was used to compute the camera-induced flow loss $\mathcal{L}_k$. The resulting camera intrinsics $\mathbf{K}$ are obtained with a softmin-weighted sum of the candidates, with the assumption that $\mathbf{K}$ was shared across frames:

$$\mathbf{K} = \sum_k w_k \mathbf{K}_k \qquad w_k = \frac{\exp(-\mathcal{L}_k)}{\sum_l \exp(-\mathcal{L}_l)} \tag{1}$$

Following this methodology, the depth maps $\mathbf{D}_i$ and $\mathbf{D}_j$ are back-projected with camera intrinsic $\mathbf{K}_i$ and $\mathbf{K}_j$ to generate point clouds $\mathbf{X}_i$ and $\mathbf{X}_j$. The optical flow between frames i and j to matched points in $\mathbf{X}_i$ and $\mathbf{X}_j$ yields $\mathbf{X}_i^{\leftrightarrow}$ and $\mathbf{X}_j^{\leftrightarrow}$, which are two filtered point clouds of corresponding patches. To make the process of solving for the best-aligned relative pose differentiable and closed-form, depth map alignment should be cast as an orthogonal Procrustes problem [2]. The formulation aims to find the rigid transformation that minimizes the total distance between matched points:

$$\mathbf{P}_{ij} = \underset{\mathbf{P} \in SE(3)}{\arg\min} \|\mathcal{W}^{1/2}(\mathbf{X}_j^{\leftrightarrow} - \mathbf{P}\mathbf{X}_i^{\leftrightarrow})\|_2^2, \tag{2}$$

where the diagonal matrix $\mathcal{W}$ contains correspondence weights. This problem is solved in closed form via singular value decomposition. [2, 14] With the known correspondence $\mathbf{u}_{ij}$ and obtained poses, re-projecting the transformed 3D points from the resulting point with the relative pose P_{ij} yields $\hat{\mathbf{u}}_{ij}$. The overall camera-induced flow loss is the following:

$$\mathcal{L} = \|\hat{\mathbf{u}}_{ij} - \mathbf{u}_{ij}\| \tag{3}$$

3.2. Depth Regularization

For the depth regularization, we added the depth map regularization term, which is an L1 loss between the rendered depth map and the pseudo ground truth depth map. The loss function thus can be expressed as:

$$\mathcal{L} = (1-\lambda_{\text{SSIM}})\mathcal{L}_{\text{rgb}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}, \tag{4}$$

where $\mathcal{L}_{depth} = ||D_{render}, D_{GT}||_1$. Following similar methodology from previous works of sparse view NeRF and 3DGS method [11] [3], the depth map rendering is done by alpha blending of depth as follows:

$$D = \sum_{i \in N} d_i \alpha_i T_i, \tag{5}$$

where D is the rendered depth and $d_i = (R_i p_i + T_i)_z$ is the depth of each splat from the camera with $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$, $R_i \in \mathbb{R}^{3\text{x}3}$ and $t_i \in \mathbb{R}^3$ are the camera pose, p_i is the 3d point, and α is the opacity learned and multiplied by the covariance of 2D Gaussian. The depth of Gaussians is not a differentiable variable, so only alphas are affected by the loss. Theoretically, the alpha of Gaussians that create noisy spots in the depth map should decrease because of the regularization term.

3.3. 3DGS robust to motion blur

Our method follows BAD-Gaussians [17] in the gaussian splatting optimization, represent the process of creating an

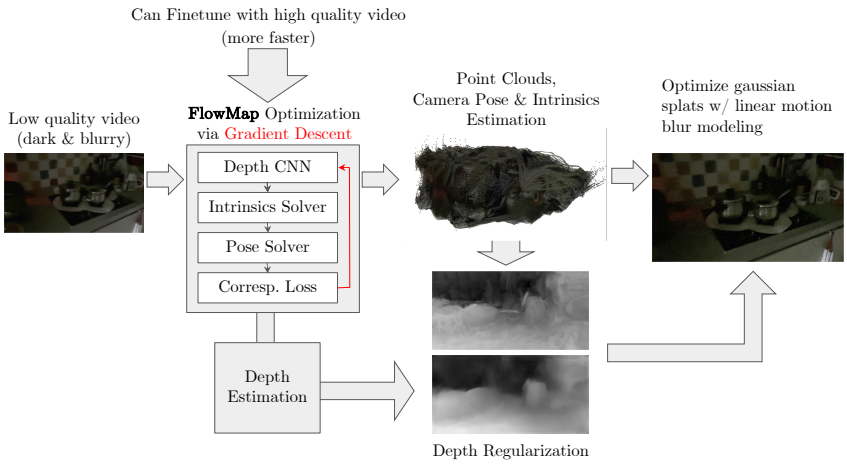
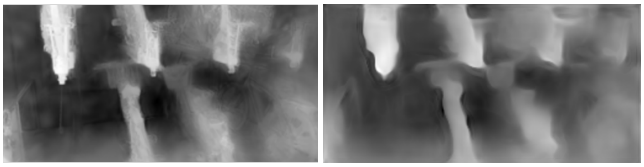


Figure 1. **Overview.** Processing Low-quality video frames processed through CoTracker [6] yields correspondence of patches. Training from the pre-trained neural network by optimizing the correspondence loss outputs depth maps which are geometrically more reliable than using other monocular depth map estimation methods. The floating Gaussians can be further handled by the usage of depth regularization terms during the optimization process.



(a) rasterized depth map (b) ground truth depth map

Figure 2. **Comparison** of the rendered depth map and the pseudo ground truth depth map that is optimized at the initialization process.

image with the integration through the flow of virtual latent sharp images. The integration is approximated by averaging n discrete samples of virtual images $\mathbf{C}_t(\mathbf{u})$, as denoted:

$$\mathbf{B}(\mathbf{u}) \approx \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{C}_i(\mathbf{u}). \tag{6}$$

where $\mathbf{B}(\mathbf{u}) \in \mathbb{R}^{H \times W \times 3}$ is the captured motion-blurred image with $\mathbf{u} \in \mathbb{R}^2$ as the pixel location in the image, and $\mathbf{C}_t(\mathbf{u}) \in \mathbb{R}^{H \times W \times 3}$ is the latent sharp image captured at time t. The extent of motion blur depends on the camera movement during the exposure time so slower movement results in motion-blurred images, especially in low-light scenarios with longer exposure times. In addition, the blurred image $\mathbf{B}(\mathbf{u})$ is differentiable with respect to each virtual sharp image $\mathbf{C}_i(\mathbf{u})$. For the equation (6), a virtual sharp image $\mathbf{C}_t(\mathbf{u})$ can be rendered from a specified camera pose $\mathbf{T}_i$ in the 3D-GS framework. To model the poses of each images, BAD-Gaussians used a camera motion trajectory with the linear interpolation between the start pose $\mathbf{T}_{start} \in \mathbf{SE}(3)$ and the end pose $\mathbf{T}_{end} \in \mathbf{SE}(3)$, therefore the virtual camera pose at time t can be expressed as:

$$\mathbf{T}_t = \mathbf{T}_{\text{start}} \cdot \exp\left(\frac{t}{\tau} \log(\mathbf{T}_{\text{start}}^{-1} \cdot \mathbf{T}_{\text{end}})\right), \tag{7}$$

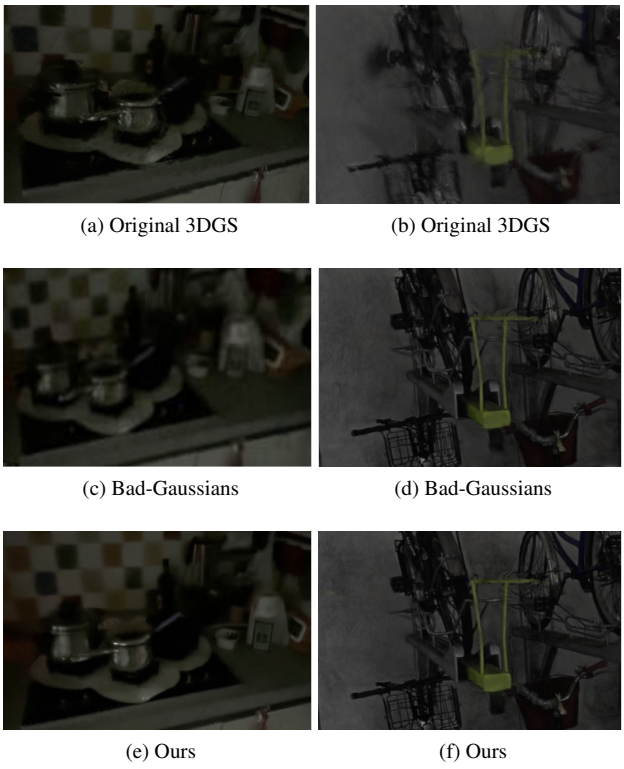
where τ represents the exposure time. The objective is to estimate both $\mathbf{T}_{start}$ and $\mathbf{T}_{end}$ for each frame with the learnable parameters of Gaussians $\mathbf{G}_\theta$.

4. Experiment Results

From Figure 3, we can observe that the original 3DGS shows noisy and aliased scene view. Our method renders overall better results than previous methods. Figure 4 shows the side viewpoint rendering quality of the scene. We found some results that depth regularization shows better geometry, which can be interpreted as depth map giving some information of the relative 3D positions of Gaussians.

5. Limitations and Future Work

Despite the promising results, our current work has several limitations that need to be addressed in future research. Firstly, our method has not been sufficiently tested across multiple datasets. This limits the generalizability of our findings, as the robustness and performance of our approach under varying conditions and scenarios remain unverified. Secondly, our study lacks a comprehensive quantitative evaluation of metrics. Without detailed quantitative analysis, it is challenging to fully assess the improvements and the effectiveness of our proposed method compared to existing approaches. To further enhance our method, future work could consider the following aspects:

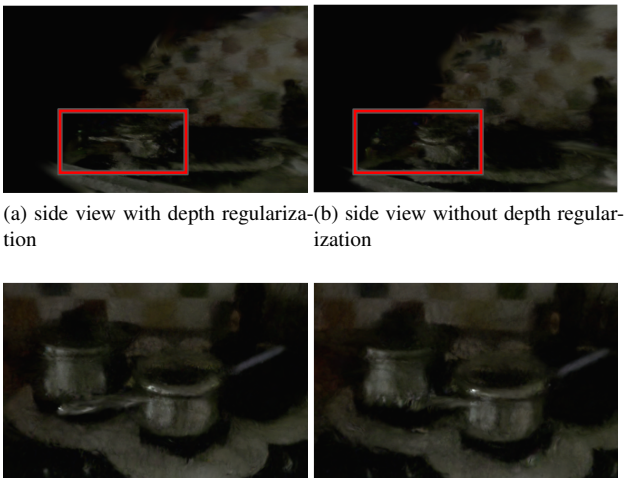


(a) Original 3DGS (b) Original 3DGS

(c) Bad-Gaussians (d) Bad-Gaussians

(e) Ours (f) Ours

Figure 3. **Experiment results.** The left figures show the compared results of a custom dataset of a dark kitchen scene, and the right figures show the result of the bicycle scene of LOL-blur dataset [18].



(a) side view with depth regularization (b) side view without depth regularization

(c) front view with depth regularization (d) front view without depth regularization

Figure 4. **Rendered side views** show that depth map based regularization leads to better geometrical consistency, such as the handle of the pot shown in the figure.

Rendering Brighter Images While Maintaining Color Quality Improving the rendering process to produce brighter images without compromising color fidelity could significantly enhance the visibility and quality of features, particularly in dark scenes.

Reducing Noise During the Correspondence Search Process Implementing advanced noise reduction techniques during the feature correspondence search could lead to more accurate and reliable feature matching, thus improving the overall performance of the SfM pipeline. Addressing these limitations and exploring these future directions will be crucial for refining our approach and validating its efficacy across diverse and challenging visual conditions.

6. Conclusion

In this work, we addressed the significant challenges of Structure-from-Motion (SfM) in dark and blurry scenes. Traditional SfM methods, such as COLMAP [13], often fail in these conditions due to the lack of reliable feature correspondences. Our approach leverages Cotracker to identify more robust correspondences across video frames, significantly enhancing performance in challenging environments. By utilizing a pre-trained network, we achieved a slight but consistent improvement in the accuracy of camera pose estimation, maintaining the quality of SfM-style output parameters. The primary issue of noise and blur, often manifesting as floating Gaussians in dark scenes, was mitigated through our implementation of a depth regularization method. This method produced geometry-corrected depth maps, effectively reducing the presence of noisy Gaussians. Furthermore, we integrated a set of virtual sharp images corresponding to each blurred frame, which allowed us to account for camera motion during exposure. This was modeled with a continuous trajectory in SE(3) space, ensuring accurate pose estimation and robust 3D reconstruction. Our method demonstrates significant improvements over traditional approaches, making it a valuable contribution to the field of computer vision, particularly for applications requiring reliable SfM in adverse visual conditions.

References

[1] Ayush Tewari Cameron Smith, David Charatan and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. 2024. 2

[2] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, 2020. 2

[3] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *arXiv preprint arXiv:2311.13398*, 2023. 2

[4] François Darmon, Lorenzo Porzi, Samuel Rota-Bulò, and Peter Kontschieder. Robust gaussian splatting, 2024. 2

[5] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 1

[6] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. 2023. 2, 3

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

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

[9] 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 *ECCV*, 2020. 1

[10] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *CVPR*, 2022. 2

[11] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[12] Jeongtaek Oh, Jaeyoung Chung, Dongwoo Lee, and Kyoung Mu Lee. Deblurgs: Gaussian splatting for camera motion blur, 2024. 2

[13] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 4

[14] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow, 2023. 2

[15] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 1992. 1

[16] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1979. 1

[17] Lingzhe Zhao, Peng Wang, and Peidong Liu. BAD-Gaussians: Bundle Adjusted Deblur Gaussian Splatting. 2024. 1, 2

[18] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022. 4