From NeRF to 3DGS: A Leap in Stereo Dataset Quality?

Magnus K. Gjerde Filip Slezák Joakim B. Haurum
Thomas B. Moeslund
Aalborg University, Denmark
mkgj@create.aau.dk, filip.slezak@agcocorp.com

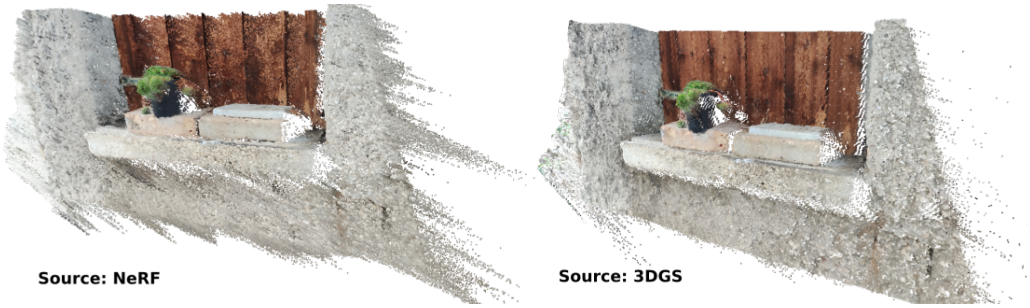


Figure 1. Comparison of reprojected geometry obtained by Tosi *et al*. [19] and our approach with 3D-Gaussian Splatting.

Abstract

Recent advancements in stereo matching, driven by deep learning techniques, have increased the need for datasets containing dense ground truth disparity labels. Yet, the rarity of real-world datasets with these labels presents significant challenges stemming from the difficulties in generating accurate dense disparity maps. Acquisition often involves complex structured light setups, producing a constrained quantity of high-quality samples, or employing laser-based distance sensors, which offer more accessible but sparsely labelled and less accurate data. A promising development in this context is the utilization of Neural Radiance Fields (NeRFs), which leverage a minimal set of RGB images to synthesize stereo images with relatively accurate dense disparity maps. Despite the high quality of synthesized images, NeRF-generated disparity maps exhibit a significant number of outliers, necessitating complex training paradigms for effective use. Our study investigates using 3D Gaussian Splatting (3DGS) over NeRFs to produce stereo training views and dense disparity labels. We demonstrate that 3DGS offers enhanced accuracy in generating disparity labels and propose an efficient strategy for identifying and removing outliers, thereby significantly improving the disparity labels quality.

1. Introduction

Stereo vision is a crucial facet of computer vision. It relies on stereo matching to identify corresponding points in each respective image. Depth information can then be subsequently found using triangulation. Therefore, to estimate valid depth values, correct correspondences among pixels are required. Most early methods used accurately calibrated cameras to capture images, then obtained epipolar rectifications and conducted pixel-based stereo matching in a local neighborhood to complete disparity estimation [6]. With the advancements of deep learning and the steady increase in computing power, disparity estimation using deep learning started to get traction compared to traditional hand-crafted methods [12]. Lately, the advent of deep learning architectures such as RAFT-Stereo [10] and IGEV-stereo [21] has enabled greater accuracy and the ability to find correct correspondences in occlusions or weakly textured areas. Deep learning has since become the mainstream method for generating dense disparity maps.

End-to-end stereo-matching networks require a large amount of labeled data for training [3]. Obtaining real-world ground-truth data is both expensive and time-consuming. The complexity can be observed in the popular Middlebury dataset. They utilized structured light along with a complicated multi-stage calibration procedure. The whole procedure is hardly scalable due to the required equipment and post-processing required [16].

In response to these challenges, exploring novel stereo data generation and enhancement techniques is essential. Tosi *et al*. [19]. used NeRFs to generate a stereo-matching dataset using only 100 images per scene captured by a handheld camera, offering a potential low-cost solution. However, the reconstructed disparity exhibited large errors and a complicated training protocol using trinocular photometric loss was required to achieve good zero-shot generalization. Consequently, the training compute required has been substantially increased when compared to methods only supervised with ground truth disparity. Recently, a technique called 3D Gaussian Splatting [8] offered the possibility to generate images of higher fidelity than NeRFs. While the input to both methods is just images, 3DGS also constructs an explicit 3D representation, offering a greater potential for downstream tasks such as scene editing, and scene relighting [4, 20].

Our work addresses the feasibility of using 3DGS for stereo-dataset generation. The primary focus of our evaluation is on the quality of the disparity map in an attempt to push the frontier of low-acquisition cost realistic stereo datasets while removing the need for complicated training protocol due to low accuracy disparity labels. We summarize the contents of this work as follows:

- A comprehensive evaluation of disparity accuracy generated using NeRFs and 3DGS
- Quantitatively showing that 3DGS has potential to generate higher quality stereo matching datasets than NeRF.

2. Related work

Deep learning for stereo matching Deep learning techniques has attracted great interest from the research community. Laga *et al*. [9], identified more than 150 published papers in this area between 2014 and 2019, and since then the trend of using deep learning techniques has only continued. Today, most published papers rely on iteratively disparity refinement architectures such as RAFT-Stereo and IGEV-Stereo [10, 21], however, hybrid systems are also an option. Aleotti *et al*. [1] published a neural disparity refinement technique with a switchable disparity estimator as the backbone. They tested with both handcrafted AD-CENSUS [22], SGM [6], and the learned C-CNN [11] stereo matchers. Their approach achieved good zero-shot generalization by refining disparity maps obtained by SGM. [1] However, regardless of hybrid or end-to-end deep learning systems, these models are dependent on ground truth training data.

Self-supervised learning A method to train deep models without the use of ground-truth depth data is self-supervised stereo. Originally used in optical flow estimation, it has been proposed as a possible solution in the absence of sufficient ground truth. [7] A common approach is to use traditional image features to generate sparse stereo matches with high confidence which is subsequently used to aid a deep stereo estimator. [9] Another approach uses image reconstruction as the supervisory signal. Here, the input is a set of images, and by hallucinating depth for an image and projecting it into nearby views, the model is trained by minimizing the image reconstruction error. [5]. However, according to Tosi *et al*. [19] self-supervised methods provide good results in single domains, but often lack generalization to other domains.

NeRF-Supervised Deep Stereo Leveraging NeRF to supervise deep stereo networks marks a significant leap forward as presented by Tosi *et al*. [19]. This approach enables the training of stereo matching models without expensive ground-truth depth data, relying instead on synthetic views generated by NeRF from a real-world scene that was captured using a single handheld camera. They proposed a training procedure from rendered stereo triplets and showed that stereo networks were capable of predicting sharp and detailed disparity maps using only this procedure. However, it is observed that a lot of the points need to be filtered due to inaccurate disparities. Figure 1 shows an example of using such an unfiltered disparity map.

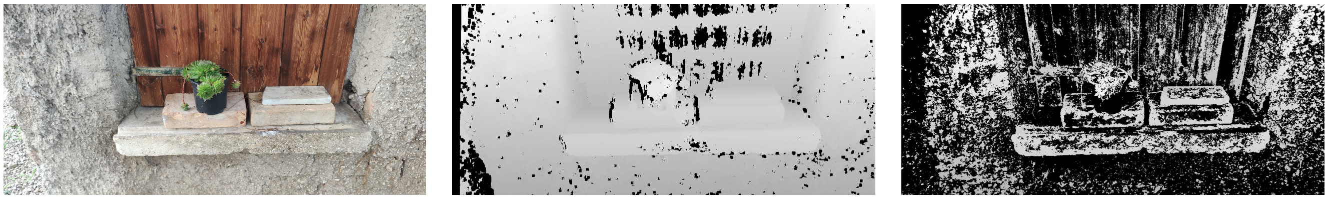
3D Gaussian Splatting Unlike the widely adopted Neural Radiance Fields, 3DGS adopts an explicit 3D Gaussian point representation, forming our method's basis for generating synthetic disparity labels. A 3D Gaussian point is defined as:

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

Where μ and Σ denote the spatial mean and covariance matrix respectively [24]. The Gaussians are also associated with a learned opacity o and a view-dependent color $\mathbf{c}$. In the rendering process the Gaussians are projected from 3D to 2D onto the image plane and then rasterized into an image. The spatial position of the Gaussians are asserted through ordinary projection and the 2D covariance matrices are approximated as $\Sigma' = \mathbf{J}\mathbf{W}\mathbf{\Sigma}\mathbf{W}^\mathbf{T}\mathbf{J}^\mathbf{T}$, where $\mathbf{W}$ and $\mathbf{J}$ denote the viewing transformation and the jacobian of an affine approximation of the perspective projection transformation. [24] From Kerbl *et al*. [8], A neural point-based approach computes the color C of a pixel as:

$$C = \sum_{i\in N} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \qquad (2)$$

For each point in N, Where c_i is the learned color of each point and α_i is the evaluated 2D Gaussian with the 2D screen projected covariance Σ' and opacity o. In implementation, the points in the set N must be ordered from back to front. Once the color is saturated above a certain threshold the pixel is colored and presented. Given the explicit 3D representation of the scene, optimized Gaussian spatial



(a) Original input view from the dataset provided by Tosi *et al.* [19]

(b) 3DGS filtered by right-to-left consistency check with SSIM

(c) NeRF filtered with Ambient Occlusion map according to Tosi *et al.* [19]

Figure 2. Comparison of disparity images of the same viewpoint for NeRF and 3DGS. In this sample 3DGS provides a more dense disparity map than NeRF. Additionally, the accuracy compared with RAFT-oracle is also higher for 3DGS, showing strong potential for dense disparity supervision.

means can be used to render a depth image based of the Gaussian's distance to the camera z_{camera}. Furthermore, disparity image can be generated by using equation 3 below:

$$d = \frac{fb}{z_{camera}} \tag{3}$$

Where b is a virtual baseline and f is the focal length for the camera used to capture the input views.

3. Evaluation methodology

To isolate variables for a fair comparison, we are directly replicating the data generation procedure from Tosi *et al.* [19]. The only change was utilizing 3DGS [8] instead of Instant-NGP [13] for stereo image and disparity map generation. Otherwise, the dataset is identical across all parameters.

3.1. Comparison metrics

To establish the accuracy of our generated disparity maps, we utilize the concept of *Oracle* as a reference following [19]. As the synthetized disparity maps contain rendering errors, we further explore various filtering methods that can potentially identify outliers which will be tested against ambient occlusion (AO) [13] filtering strategy used in [19].

Oracle stereo network The disparity labels extracted by 3DGS pipeline do not have an associated ground truth, prohibiting quantitative evaluation. To rectify this, we use the idea of *Oracle* stereo network following [19]. Therefore, the disparity map obtained by *Oracle* will be assumed to be the ground truth and used for evaluation in subsequent experiments.

Left-right consistency check Provided that the disparity map is accurate, it can be used to warp the right image to the left one. Incorrect disparity values will result in faulty alignment, which can be easily identified on a per-pixel level using Structural Similarity Index Measure (SSIM) [14, 15]. Once every pixel is assigned a similarity score, a subset can be chosen by selecting a similarity threshold, trading off density for accuracy. One downside is that such a check is not occlusion-aware, typically removing all the occluded points from the disparity map.

Statistical outlier removal Outliers in 3D pointclouds can be identified by computing for each point P_i its average Euclidean distance $\hat{D}_i = \frac{1}{k}\sum_{j=1}^{k} D_{ij}$ to its k nearest neighbors. A point P_i is flagged as an outlier if its $\hat{D}_i$ exceeds a threshold defined by the global mean μ plus a multiplier α times the standard deviation σ of all average distances $\hat{D}$ in the point cloud, as in: $\hat{D}_i > \mu + \alpha \cdot \sigma$. The 3D point clouds can be generated by re-projecting the estimated disparity map using the camera's intrinsic parameters.[23]

4. Experimental Results

Tosi *et al.* provides 270 scenes captured with handheld cameras. Each scene contains 100 bounded views of a static scene such as the one presented in Figure 2a. Furthermore, each scene is provided with image poses, estimated camera intrinsics, and a sparse point cloud generated with COLMAP [17, 18]. The 3DGS models are trained with the official implementation of Kerbl *et al.* [8]. No changes have been made to their implementation and we are using the 30k iteration models. Out of the 270 scenes, only 200 scenes are sufficiently reconstructed with 3DGS and are used for evaluation. An example of an unusable scene is depicted in Figure 4, in the worst cases just a few instances of large Gaussians were covering most of the rendered image. The rendered views are generated using the SIBR interactive viewer which has been repurposed for 3DGS by Kerbl *et al.* [2, 8], and further modified by us to generate disparity images. We select RAFT-Stereo [10] as the *Oracle* network with the officially supplied sceneflow checkpoint.

Ambient occlusion filtering The application of ambient occlusion filtering within the NeRF framework leads to a significant reduction in the number of disparity points, averaging a 44% decrease with a standard deviation of 18.4%, Almost half of the dataset is discarded due to inaccurate disparities. This suggests that ambient occlusion filtering is highly effective at eliminating poor-quality disparity. However, the extensive removal of points also indicates a general

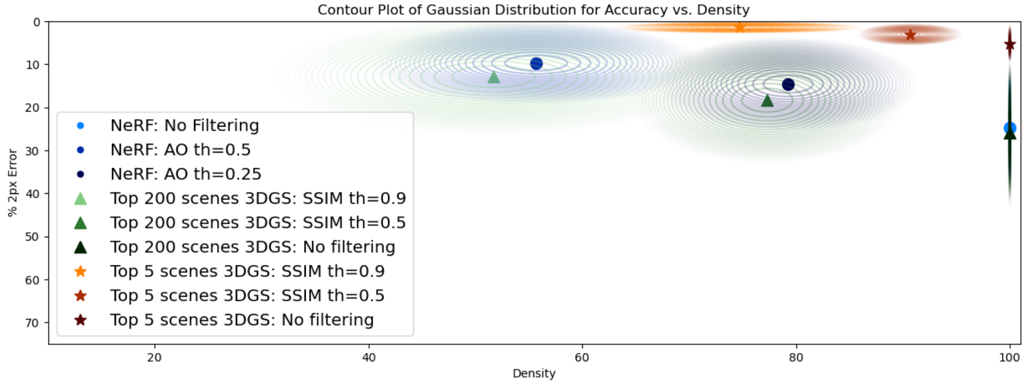
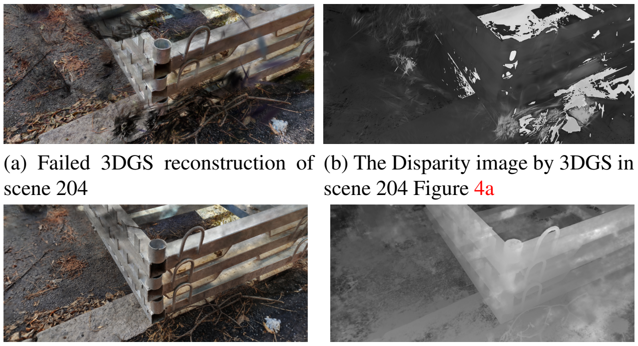


Figure 3. We assess the quality of the data with respect to the RAFT-Oracle procedure. Each data point is calculated by randomly selecting 9 views from each of the reconstructed scenes. The 2px error is presented along the vertical axis and density along the horizontal axis. Notice the y-axis is flipped. A density of 100% means that no points in the disparity image has been filtered, and no error means that the generated Disparity is identical to RAFT-Oracle. The NeRF shows on average a lower error than 3DGS across all the scenes, but for the top 5 well-reconstructed scenes the 3DGS as scored by the RAFT-Oracle, 3DGS shows an impressive density-to-error ratio.



(a) Failed 3DGS reconstruction of scene 204

(b) The Disparity image by 3DGS in scene 204 Figure 4a

(c) Nerf reconstructed image of scene 204 in Tosi *et al.* [19]

(d) NeRF Rendered disparity image of the scene rendered Figure 4c [19]

Figure 4. Comparison of disparity images of the same viewpoint for NeRF and 3DGS. The results are from the same scene which has been used for evaluation in section 4

lack of confidence in the accuracy of the disparity map produced. This trend continues for left-right consistency check of all the 3DGS scenes; however, high confidence is seen in the well-reconstructed scenes.

Left-Right consistency check Figure 3 shows that a left-right consistency check filtering positively correlates with increased accuracy. Furthermore, a more aggressive threshold can be used to compensate density for accuracy.

Statistical outlier removal Given that the reconstructed pointclouds have exhibited comet tail artefacts as can be observed in figure 1, we have evaluated SOR filter as a potential candidate method to identify outliers. However, our experiments have shown that SOR removed more good points than outliers, reducing the density and accuracy of the disparity map regardless of the selected threshold. As such, we deem it unsuitable.

Discussion In general, the data quality generated by 3DGS shows comparable performance to NeRF, however some of the scenes reconstructed with 3DGS contain large amounts of Noise. A qualitative sample is presented in Figure 4. This observation suggests that not every scene effectively modeled by NeRF guarantees successful reconstruction by 3DGS. Given that the data collected by Tosi *et al.* [19] is used to assess the quality of synthetic data generation by NeRF, it implies that these scenes have already met the criteria for successful modeling by NeRF. When examining the quality for the top 5 reconstructed scenes depicted in Figure 3, it raises the question of whether it's possible to reconstruct all scenes in the dataset at the same quality level using 3DGS and, if achievable, the effectiveness of training a 3DGS-supervised deep stereo network with such data.

5. Conclusion

Our work proposes an alternative method to generate training data for stereo matching by leveraging 3D Gaussian Splatting (3DGS) models, addressing the domain gap between synthetic and real-world data. Initial findings show that compared to NeRF, our approach with 3DGS shows promising results in generating dense and accurate ground truth disparity maps. Furthermore, we outline a simple filtering strategy which can be used to trade density for accuracy. However, further investigation into how training data rendered using 3DGS could improve the performance of a stereo-matching network is required.

This research was partially funded by Innovation Fund Denmark, grant number 3129-00060B

References

[1] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, pages 207–217. IEEE, 2021. 2

[2] Sebastien Bonopera, Peter Hedman, Jerome Esnault, Siddhant Prakash, Simon Rodriguez, Theo Thonat, Mehdi Benadel, Gaurav Chaurasia, Julien Philip, and George Drettakis. sibr: A system for image based rendering, 2020. 3

[3] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13022–13032, 2022. 1

[4] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv:2311.16043*, 2023. 2

[5] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. 2019. 2

[6] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 1, 2

[7] Rico Jonschkowski, Austin Stone, Jonathan T. Barron, A. Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *European Conference on Computer Vision*, 2020. 2

[8] 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. 2, 3

[9] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1738–1764, 2020. 2

[10] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1, 2, 3

[11] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 2

[12] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2015. 1

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

[14] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 3

[15] Diana Sadykova and Alex Pappachen James. Quality assessment metrics for edge detection and edge-aware filtering: A tutorial review. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2366–2369. IEEE, 2017. 3

[16] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, 2014. 2

[17] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[18] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[19] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 855–866, 2023. 1, 2, 3, 4

[20] Cyrus Vachha and Ayaan Haque. Instruct-gs2gs: Editing 3d gaussian splats with instructions, 2024. 2

[21] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2

[22] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3*, pages 151–158. Springer, 1994. 2

[23] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 3

[24] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 371–378, 2001. 2