Generating Synthetic Stereo Datasets using 3D Gaussian Splatting and Expert Knowledge Transfer

Filip Slezák[1,2] Magnus Kaufmann Gjerde[1] Joakim Bruslund Haurum[1,3] Ivan Nikolov[1]
Morten S. Laursen[2] Thomas B. Moeslund[1,3]
[1]Visual Analysis & Perception Lab, Aalborg University
[2]AGCO A/S, Denmark [3]Pioneer Centre for AI, Denmark

Abstract

In this paper, we introduce a 3D Gaussian Splatting (3DGS)-based pipeline for stereo dataset generation, offering an efficient alternative to Neural Radiance Fields (NeRF)-based methods. To obtain useful geometry estimates, we explore utilizing the reconstructed geometry from the explicit 3D representations as well as depth estimates from the FoundationStereo model in an expert knowledge transfer setup. We find that when fine-tuning stereo models on 3DGS-generated datasets, we demonstrate competitive performance in zero-shot generalization benchmarks. When using the reconstructed geometry directly, we observe that it is often noisy and contains artifacts, which propagate noise to the trained model. In contrast, we find that the disparity estimates from FoundationStereo are cleaner and consequently result in a better performance on the zero-shot generalization benchmarks. Our method highlights the potential for low-cost, high-fidelity dataset creation and fast fine-tuning for deep stereo models. Moreover, we also reveal that while the latest Gaussian Splatting based methods have achieved superior performance on established benchmarks, their robustness falls short in challenging in-the-wild settings warranting further exploration.

1. Introduction

Recovering the 3D structure of a scene captured from images is a widely researched problem that has exploded in popularity through the recent advances in monocular depth estimation as well as novel view synthesis approaches such as Neural Radiance Fields (NeRFs) and Gaussian Splatting.

Initial work in the field required the use of stereo image pairs and, through traditional algorithms like semi-global matching (SGM)[11], achieved remarkable performance using well-designed heuristics. This resulted in accurate disparity maps but still had several shortcomings due to the common challenges of stereo matching. These are, namely, textureless regions and occlusions - both of which make it very challenging to find reliable correspondences, leading to errors or a lack of prediction.

With the introduction of large-scale datasets with accurate disparity labels, such as [19], deep learning-based disparity estimation methods were made possible. However, obtaining per-pixel labels is very challenging, typically addressed by using a simulated environment where accurate ground truth geometry can be directly extracted from the explicit scene representation. While the geometry labels are highly accurate, the realism gap between rendered and real-world images makes these methods perform suboptimal when applied on real-world samples.

Recent advancements include the addition of vision foundation models pre-trained for monocular depth estimation [12, 33] and powerful context networks, such as Dinov2 [22], which have greatly improved the networks' reasoning ability in the ill-posed regions. This motivated the use of monocular depth estimators such as DepthAnythingV2 [33] and Metric3Dv2 [12] to produce high-quality depth estimates of real images. Orthogonally, Tosi *et al*. [27] instead simply collect data via mobile phone recordings and apply the recent advances in the field of Novel View Synthesis in order to obtain realistic images and cheap geometry labels. However, NeRFs, as used by *e.g*. Tosi *et al*., have a considerable shortcoming. The implicit representation of the scene fails to produce accurate, fully dense geometry, requiring a lot of filtering, which makes the disparity maps very sparse. As a consequence, a complicated training procedure requires photometric losses apart from the common L1 disparity loss to compensate for the lack of density. Such a training setup suffers from training instability, and the results are not replicable as outlined in greater detail in our experimentation section and observing the replication efforts of the academic community.

Instead, we investigate the newly introduced 3D Gaussian Splatting [15] based methods, which explicitly model the scene using Gaussians that can be converted to meshes [4, 13, 34]. As these methods offer highly realistic images

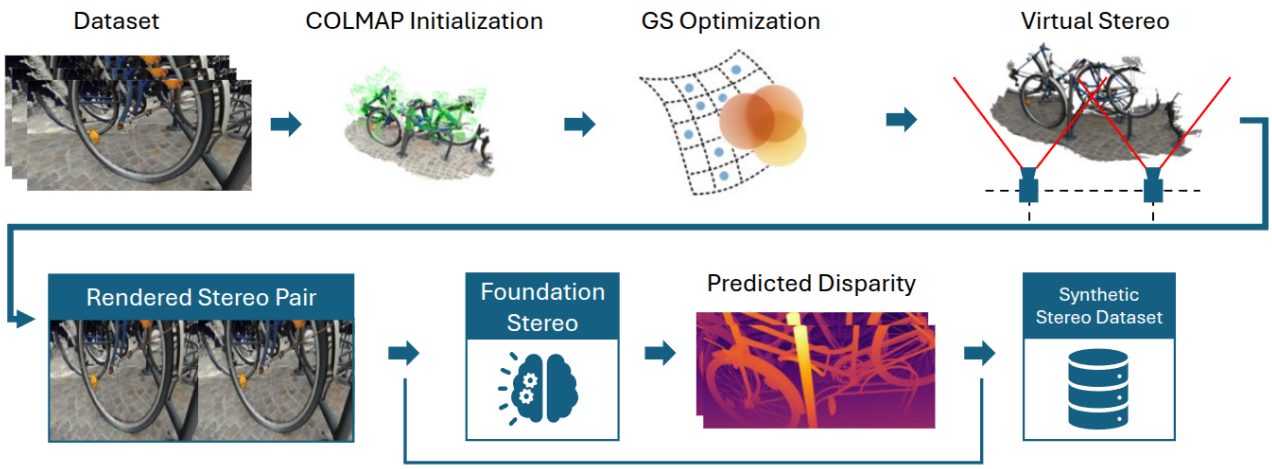


Figure 1. **Expert Knowledge Transfer using FoundationStereo.** Overview of the proposed expert knowledge transfer setup where stereo pairs rendered with Gaussian Splatting are supplied to FoundationStereo [30], which produces high-quality pseudo depth estimates. The process starts with the 3D Gaussian Splatting (3DGS) pipeline [15], where COLMAP is used to initialize the optimization of 3DGS. Stereo image pairs are rendered after the scene has been fitted with 3DGS. As the underlying geometry of 3DGS is poorly reconstructed, the stereo pairs are instead sent to FoundationStereo, producing ground-truth disparity images. The generated synthetic stereo dataset is then used to train a lightweight stereo network, RAFT-Stereo [18] (not depicted in this figure).

and seemingly accurate surface reconstruction, we hypothesize that they might be the missing link to the easy acquisition of stereo-matching datasets. Furthermore, our choice of 3DGS based methods is further motivated by their excellent rendering speed compared to NeRF based methods of comparable reconstruction quality. Concurrently, we are inspired by the recent development within monocular and stereo vision foundation models, and investigate whether using depth estimates from a large stereo vision foundation model [30] outperforms the performance obtained with the reconstructed geometry, as shown in Figure 1.

Using a fine-tuning training setup, we evaluate the effect of the different 3D Gaussian Splatting-based synthetic datasets, and compare directly with NeRF-Stereo. We do so by performing zero-shot evaluation on a set of previously unseen datasets commonly used within the stereo depth estimation field.

Our contributions are as follows:

- We qualitatively and empirically find that the rendered disparity from 3DGS and the reconstructed mesh of PGSR are insufficient for fine-tuning a RAFT-Stereo network.
- We propose a knowledge transfer setup where predicted depth estimates from FoundationStereo are used as ground truth for 3DGS stereo renderings.
- We find that fine-tuning a RAFT-Stereo network using our proposed expert knowledge transfer setup consistently outperforms other methods, including the state-of-the-art NeRF-Stereo method.

2. Related Works

2.1. Disparity Estimation Algorithms

Large-scale disparity map estimation became possible with the development of SGM algorithms [11], which optimized the calculated disparity along preset directional paths, instead of purely global optimization, which can be costly, or local methods that tended to leave artifacts around edges of objects. With the rise of deep learning models, architectures like PSM-Net [3] and models built on top of it [14], using features extracted by ResNet and a pyramidal structure which utilizes features from different scales. Later on, the work presented in RAFT-Stereo [26] and later on in RAFT-3D [18] models the disparity and full 3D scene flow between consecutive frames and iteratively optimizes them using GRU layers. Taking the iterative optimization idea and expanding on it through the use of Geometry Encoding Volume is the IGEV-Stereo [32]. It incorporates more local details through the iterative optimization of the disparity maps to capture smaller 3D surface features. One problem that persists in both RAFT and IGEV models is the ambiguity in smooth and featureless regions, which translates into incorrect disparity values propagated through the iterative optimization. The Selective-Stereo model [28], which introduces the Selective Recurrent Unit (SRU), instead of the normally used GRU ones, together with a Contextual Spatial Attention (CSA) module, tries to mitigate this problem. It aggregates disparity information at different levels of detail and frequencies to capture hidden disparity information, especially in edge regions and at featureless areas. As trans-

formers have proven to provide high-quality and robust results in many other computer vision fields, their use in disparity estimation has also proven extremely useful. Both the transformer self-attention and relative pixel distance encoding of STTR [16] and the positional embeddings and pretraining of CroCoV2 [29] boost their accuracy compared to CNN-based models and can handle large discrepancies in the estimated disparity.

2.2. Stereo Datasets and Training Paradigm

The advent of deep learning-based disparity estimation methods can be traced back to the origin of large-scale datasets that contain accurate disparity labels, such as the seminal work by Mayer et al. [19]. However, obtaining per-pixel labels is very challenging, typically addressed by using a simulated environment where accurate ground truth geometry can be directly extracted from the synthetic rendering pipeline. However, while the geometry labels are highly accurate, the rendered images have a realism gap compared to images captured in the real world. While realistic images can be easily captured by building a real stereo pair and capturing the real world, extracting accurate geometry becomes the primary challenge. Accurate dense geometry can be captured by either using structured light projectors [24] or using an expensive Laser Scanner [25]. Both of these methods require significant post-processing and static scenes, but they achieve a dense geometry reconstruction. An additional approach is to use a multi-beam LiDAR, which can also support dynamic scenes, such as [7] [8] and [9]. Still, these require post-processing to address the typical scanning distortion, accurate calibration, and offer only sparse depth maps due to the current LiDAR sensor limitations. Furthermore, neither synthetic nor real stereo dataset collection methodology offers flexibility to easily capture custom scenes to improve disparity estimation performance on specific tasks. Both methodologies are prohibitively expensive, as custom synthetic scenes require experts to create custom digital assets, while the real dataset requires a sensor suite with state-of-the-art depth sensors. Recently, the advancements in novel view synthesis (NVS) have opened up the possibility to use freely captured images to generate training data for stereo depth estimation [10, 23, 27] and optical flow [17]. However, due to the limitations of NeRFs [21], it is not possible to extract accurate geometry, and the training procedure requires a complicated loss function to achieve good performance, deviating from the simple L1 loss used in the state-of-the-art works, which use existing stereo datasets.

2.3. Novel View Synthesis and Mesh Extraction

Given a freely captured set of RGB images, novel view synthesis encodes the scene in a representation which can be queried to obtain novel views. Pioneering work NeRFs [21] has revolutionized the field, achieving unprecedented realism. Subsequent works have improved on the original idea [1], offering improved visuals, but at the cost of rendering time. While some works have also explored improving the rendering speeds [2], it came at the cost of visual quality. The trade-off between rendering speed and visual quality has been addressed in the work by Kerbl et al. [15], offering superior visual fidelity while achieving real-time rendering speeds. Furthermore, unlike NeRFs, which embed the scene implicitly in the weights of the neural network, 3D Gaussian Splatting presents an explicit geometric representation of the scene. Despite this, the original formulation is not suitable for accurate geometry extraction, presenting surface artifacts. The most recent works tackle this issue by structuring the scene composition in a manner that allows them to extract accurate meshes. For example, 2DGS [13] employs a direct depth loss to the Gaussian primitives, and collapses one of the splat axes, representing the scene as a composition of 2D discs which are aligned with the surfaces, offering a better starting point for geometry extraction. Gaussian opacity fields [34] aim to remove the extraction of the meshes as a post-processing step and extract accurate surfaces directly from the Gaussian splat cloud. However, all of these methods use only the information in the images, which can be problematic for textureless areas where multiple 3D geometries can render to the same 2D images. To correct this problem, researchers are looking to include pretrained networks, which can be a powerful regularizer for the ill-posed textureless regions of the image. Gaussian Surfels [5] uses a monocular normal estimation based on [6].

3. Methodology

In this work, we investigate whether the recent advancements in 3D Gaussian Splatting can be used to synthesize novel views, reconstruct accurate surface geometries from images, and apply the results to the downstream task of stereo-matching with the RAFT-Stereo network [18].

3.1. Stereo Dataset Rendering with 3DGS

When rendering the disparity image from a 3DGS scene representation, the naive approach is to use the explicit representation of the gaussians and calculate the disparity based on the 3D location of the splat following a hard threshold on the opacity. Kerbl *et al.* [15] proposed a neural point-based approach to compute the color C of a pixel as:

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

Where c_i is the learned color of each point and α_i is the screen projected opacity of the splat. The naive depth rendering approach replaces the learned color of each point

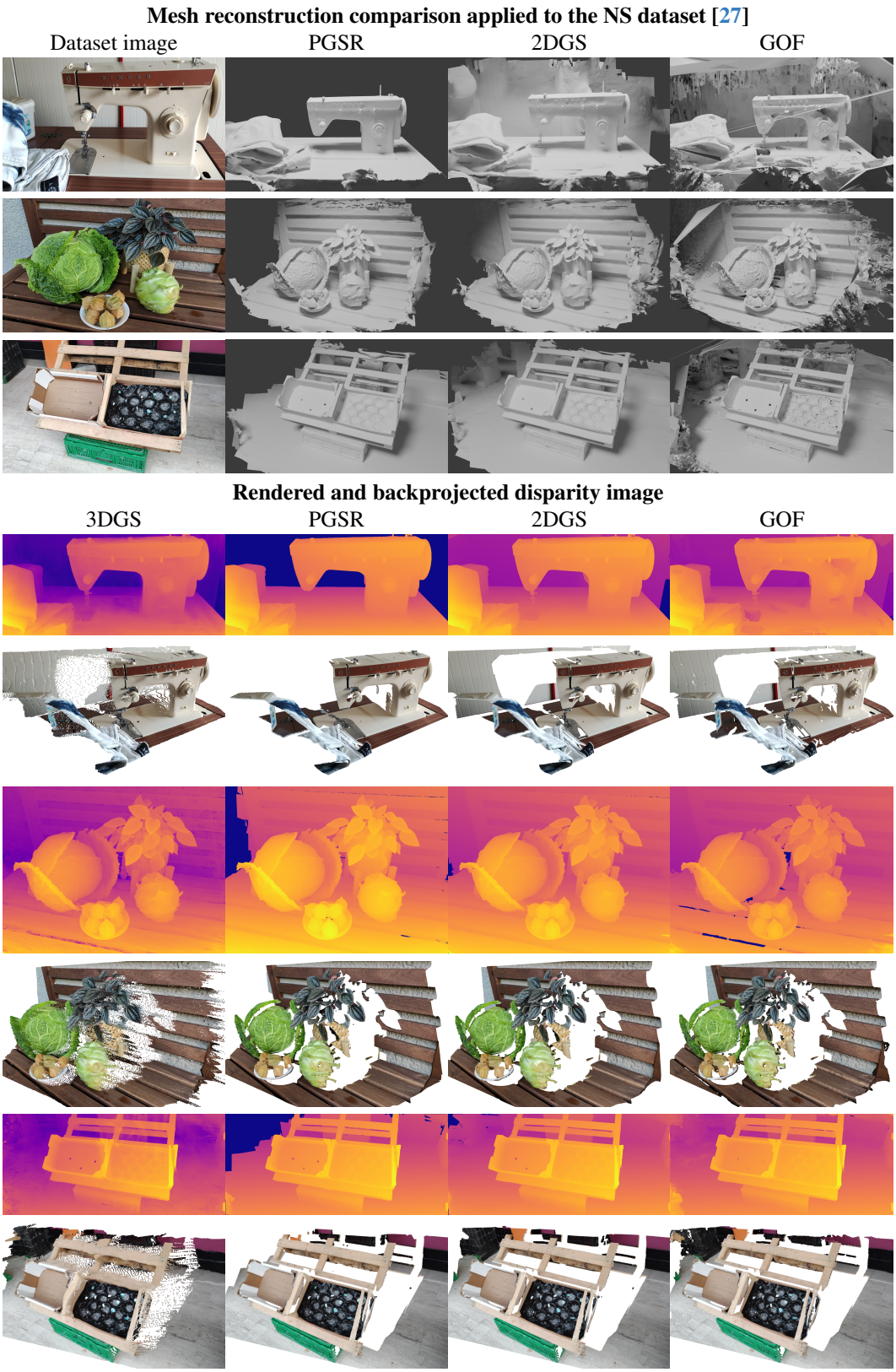
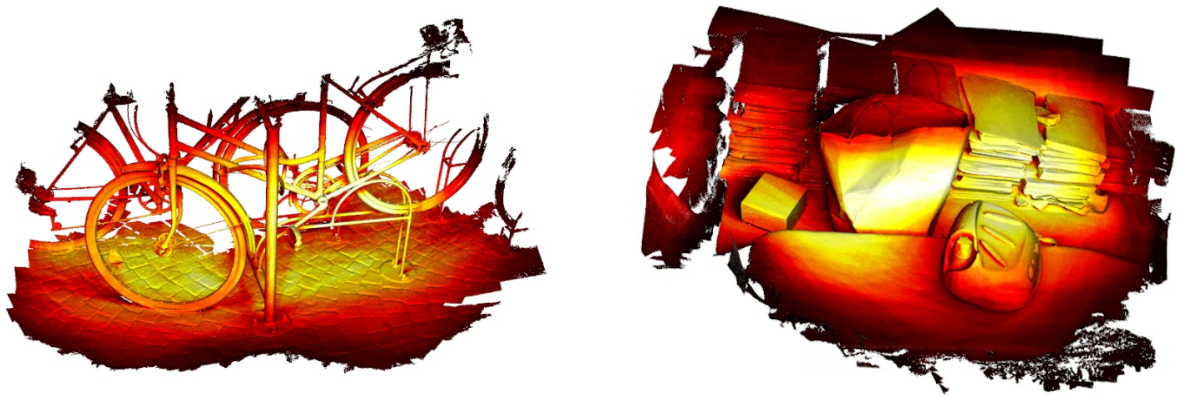


Figure 2. **Reconstructed Geometry with Gaussian Splatting.** Qualitative examples of mesh reconstructions, renderings, and backprojected disparity images produced using the 3DGS, PGSR, 2DGS, and GOF methods. Note that, for illustrative clarity, spurious faces occluding the camera view in the GOF method have been manually removed.



(a) High Observability

(b) Low Observability

Figure 3. **Observability filtering:** Examples of the observability heatmaps where brighter colors represent vertices that have been seen by more cameras. Figure 3a shows a high observability scene where an object is seen from many positions, and Figure 3b shows a low-observability scene. Notice how there are many more dark areas in Figure 3b, indicating they are poorly observed.

with the z-depth from the viewing transform. Since the new color represents a depth value, it can be converted to disparity d in the rendered stereo pair with a virtual baseline b and focal length f as:

$$d = \frac{fb}{z_{depth}} \tag{2}$$

This approach generates 'comet tails' smooth transitions at edges which should be sharp, similar to the rendered images used in the NeRF-Stereo work [27], showing little improvement over different Novel View Synthesis methods. Therefore, in order to improve the quality of the depth estimates for the stereo pairs, recent advances in mesh generation are considered [4, 13, 34].

The generated meshes can be efficiently rendered into a depth map and converted to disparity using Eq. 2. As meshing-based methods are designed with accurate surface reconstruction in mind, artifacts such as comet tails should be reduced. To decide on the viewing transform for the dataset, we re-use the COLMAP poses along with the camera intrinsics. We use the same baselines as Tosi *et al*. [27] to achieve the same disparity distribution in the dataset.

3.2. Qualitative Assessment of Gaussian Splatting Methods

As mentioned, since the introduction of the 3D Gaussian Splatting method, there has been a rapid development of newer models with a focus of improving the underlying geometry. While it is possible to generate visually pleasing and convincing images with poor underlying geometry, training on noisy ground truth structural information will result in models of poor quality which produce blurred outputs. Therefore, we qualitatively evaluate three of the most recent and state-of-the-art Gaussian Splatting methods from which we can extract meshes explicitly. Namely we consider PGSR, 2DGS and GOF [4, 13, 34]. Using each method, we reconstruct and render the dataset used in the NeRF-Stereo paper [27], and qualitatively investigate the reconstruction. For every method, we have used base settings for the rendering pipeline. Mesh rendering for PGSR and 2DGS is based on TSDF fusion and produces numerous floating artifacts, and a post-processing step is applied to keep only the largest cluster. However, while this improves the overall mesh quality, it also results in overfiltering, removing the finer correct details.

While each method can produce visually pleasing images, we find that they suffer from noisy and insufficient geometry reconstructions. In the qualitative comparison in Figure 2, we observe that 2DGS provides a balanced trade-off between finer details and geometric accuracy, whereas GOF contains significant artifacts and holes in the textureless areas. In contrast, PGSR results in the cleanest geometry with less artifacts, even though the back wall is excluded making the reconstruction very object centric. From Figure 2, it is trivial to see the noisy results of plain 3DGS depth, with a significant amount of comet tails present. Based on these results, we decided to solely use the mesh reconstruction obtained from PGSR, as it produces the cleanest meshes.

3.3. Mesh Filtering

Filtering the 3D structure directly can help with removing artifacts caused by the presence of insufficient information. Even though PGSR employs a filtering strategy masking all the surfaces where the camera rays intersect them at an an-

gle greater than 80 degrees, the rasterized input depth maps contain artifacts due to under-reconstructed areas, causing incorrect surface estimation during the TSDF process.

A prerequisite for accurate estimation of 3D structure from 2D views is the presence of parallax, meaning that each 3D point should be captured by cameras that are sufficiently displaced in relative pose. This principle is well aligned with our observation of the rendered meshes, where the quality of the 3D reconstruction was proportional to how well observed a given surface was given the camera system. Similarly, we have observed that the reconstructed visual quality was also proportional to how many training cameras have been observing a given surface during the novel view rendering. As such, we have turned these observations into a method which could be used to further filter existing meshes, but also provide valuable information about visual fidelity of each camera, which is especially critical considering that the proposed pipeline by Nerf-Stereo [27] leverages the poses of training cameras.

Given the rendered mesh and the known poses of all the training cameras relative to it, we calculate how many cameras have unoccluded view of every single vertex. A visual example can be seen in figure 3. Once the meshes are enriched by the observability information, depth maps are rendered for every single training camera. Besides obtaining per-pixel depth, the observability is also propagated as additional attribute to every pixel. To rank the camera poses based on observability, we sum the observability across all the projected pixels. In summary, the observability based filtering can not only help in removing low-quality geometry from rendered meshes, but also ranks the training camera poses based on how well observed area they observe, providing an informed way to select fewer rendering poses, rather than choosing all as in [27].

3.4. Transferring Expert Knowledge from FoundationStereo

Motivated by our findings that the Novel View Synthesis methods produce poor geometry for in-the-wild datasets, as demonstrated in both the NeRF-Stereo paper and our explorations of the Gaussian Splatting meshing methods, we consider an orthogonal direction. Instead of relying on the computed geometry, we investigate the effectiveness of estimating depth using large-scale foundation models. Specifically, we use the FoundationStereo [30] model to produce a pseudo disparity estimates for all rendered stereo pairs, which we use to train a RAFT-Stereo network, see Figure 1. This can be seen as a way of performing knowledge distillation, as we are effectively training a smaller-scale network on predictions of a large-scale network, which would be prohibitively slow in real-time applications. It is also worth noting that FoundationStereo was trained purely on synthetic data, eliminating any concerns of data leakage from the test datasets. This approach is similar, but distinct, from the prior GS2Mesh work [31] which used depth estimates to produce detailed geometry of scenes rendered with Gaussian Splatting.

4. Experimental Design

Disparity estimation models pretrained on synthetic datasets possess impressive reasoning skills about spatial geometry, but still fall short due to the sim2real gap in the image space. However, only a small amount of real data is required to improve the matching accuracy on real data, as demonstrated by RAFT-Stereo. We follow the exact finetuning procedure and augmentations outlined in their work, designed for finetuning SceneFlow checkpoint on smaller scale, but high quality real data. We further consider this training setup appropriate, as the scenes from Nerf-Stereo dataset looks visually similar to Middlebury style datasets. While RAFT-Stereo was fine-tuned on a Middlebury 2014 dataset [24], we generate the training dataset following the rendering procedure of Nerf-Stereo [27] outlined in Section 3.1.

As the extracted poses from the camera systems using COLMAP are used as anchors to generate the stereo pairs, their observations vary in quality. Our experiments show that using the observability score per camera is a great predictor of stereo image quality, ensuring that the very left camera image observes the best reconstructed area of the scene. Given the five cameras with maximized observability, we generate three baselines following Nerf-Stereo, ensuring disparity distribution is similar to the test and validation datasets. We also show results when training using the 3DGS rendered images and FoundationStereo pseudo depth estimates. For all the experiments, regardless of the method, we select the same scenes and camera locations to ensure a fair comparison.

Lastly, we compare the performance to the RAFT-Stereo model trained on the Sceneflow dataset as well as NeRF-Stereo. It is worth highlighting that NeRF-Stereo was trained from scratch with a trinocular and photometric loss, depending on filtering pixels with poor geometry estimates based on their transparency[1]. We also report NeRF-Stereo results obtained from trying to reproduce the training process using an unofficial code reproduction[2] by the authors of the RAFT-Stereo paper, as the original stereo-network training code is not available.

We follow the NeRF-Stereo evaluation protocol and report results on 200 stereo images from KITTI 2015 [20], 15 stereo pairs from Middlebury v3 training set (Midd-T) [24] and 27 pairs from ETH3D [25]. The best checkpoint

[1] Incorrectly called Ambient Occlusion (AO) filtering in the NeRF-Stereo paper, and in this work referred to as α filtering.

[2] https://github.com/husheng12345/Unofficial-NeRF-Supervised-Deep-Stereo

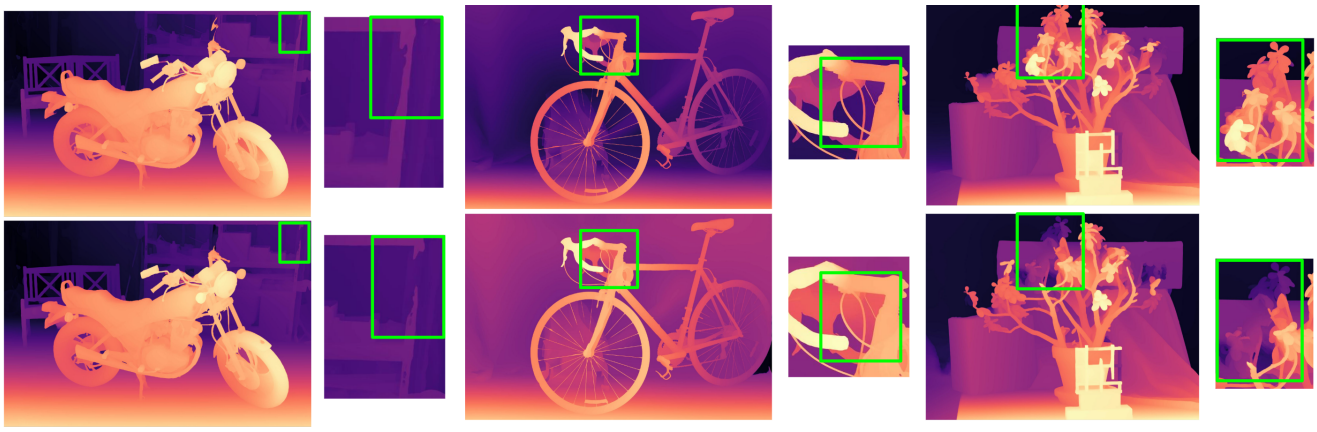


Figure 4. **Qualitative Comparison of RAFT-Stereo and 3DGS + FS.** We make a visual comparison between RAFT-Stereo SceneFlow checkpoint, and our finetuned version on dataset generated using our proposed pipeline leveraging 3DGS for stereo images and FoundationStereo for ground truth pseudolabels. Our proposed 3DGS + FS method is shown in the top row, and predictions from the baseline RAFT-Stereo method are shown in the bottom row. Improvements can be observed in thin structures and complex depth discontinuities

per method is chosen using a suite of validation datasets, specifically 194 stereo images from KITTI 2012 [8], 13 additional images from the training set of Middlebury v3 (Midd-A) at Full, Half, and Quarter resolutions (F, H, Q), and the Middlebury 2021 (Midd-21) dataset [24]. Performance is measured by determining the ratio of pixels with a disparity error larger than τ pixels. τ is dataset dependent and set to 1 for ETH3D, 2 for Middlebury, and 3 for KITTI, following common practice in the stereo matching field.

5. Experimental Results

We report our results in Table 1. We find that finetuning on the 3DGS and PGSR based datasets is comparable to the performance of the reproduced NeRF-Stereo training paradigm from random initialization on their full dataset. However, none of them (including the reproduced NeRF-Stereo) are comparable to the RAFT-Stereo baseline, which was solely trained on SceneFlow [19]. This is a clear indication that the reconstructed geometry in the Gaussian Splatting-based methods is of too poor quality, as discussed in Section 3.2. We also find that the observation filtering used with PGSR does not consistently improve performance. Instead, performance is only improved on Midd-T at quarter resolution as well as on the ETH3D dataset.

Interestingly, we observe that when fine-tuning using the NeRF-Stereo renderings, the performance becomes significantly better and even beats the RAFT-Stereo baseline on the Midd-T dataset. We attribute this to the α-filtering used by NeRF-Stereo, which aggressively filters out pixels with poor underlying geometry. Lastly, we find that our expert knowledge transfer setup (3DGS + FS) outperforms all other methods on the Midd-T and ETH3D datasets, and is close to matching the RAFT-Stereo performance on the KITTI-15 dataset. We visually examine the predicted disparity maps from the baseline RAFT-Stereo and the proposed 3DGS+FS, see Figure 4. We find that our proposed methods are better at recovering very thin structures (such as the brake cables on the bicycle) as well as complex layered scenes (such as the flowers on the table). It is worth noting that the original reported results from NeRF-Stereo do outperform 3DGS+FS on KITTI-15 and parts of Midd-T. However, as the results are not reproducible, we argue that they are not fully representative.

6. Dataset and Training Instabilities

Throughout the proposed experiments, we encountered several instances of instability when constructing the synthetic datasets and training the stereo networks.

Firstly, we encountered several problems with the NeRF-Stereo dataset. As the data is collected with handheld cameras being moved around, there is a considerable amount of motion blur and poor quality images, which we found the Gaussian Splatting methods are quite sensitive towards. We also found that the provided COLMAP poses result in poor reconstruction with Gaussian Splatting-based methods, as the principal point was refined and did not lie in the center of the image. This is a known problem[3] and was rectified by recomputing COLMAP without refining the principal points. While it is practically possible to convert images with non-centred principal points by simple cropping and adjusting the camera paramters, we have chosen to rerun the scenes with the provided *convert.py* script common in gaussian splatting repositories, as that would align with the core idea of generating custom training datasets only from a

[3] https://github.com/graphdeco-inria/gaussian-splatting/issues/144#issuecomment-1786762565

Method	Filtering	Fine-Tuned	KITTI-15 (>3px) All	KITTI-15 (>3px) Noc	Midd-T F (>2px) All	Midd-T F (>2px) Noc	Midd-T H (>2px) All	Midd-T H (>2px) Noc	Midd-T Q (>2px) All	Midd-T Q (>2px) Noc	ETH3D (>1px) All	ETH3D (>1px) Noc
RAFT-Stereo	-	✗	5.46	5.27	15.72	12.00	11.23	8.67	10.52	7.42	2.61	2.29
NeRF-Stereo	α	✗	**5.41**	**5.23**	16.45	12.08	9.67	**6.42**	**8.05**	**4.82**	2.94	2.23
NeRF-Stereo*	α	✗	6.65	6.22	20.10	15.43	12.82	8.89	11.18	7.68	4.37	3.91
NeRF-Stereo	α	✓	5.63	5.41	15.89	11.72	11.03	7.84	9.71	6.54	3.48	3.00
3DGS	-	✓	5.77	5.59	22.74	19.97	15.53	11.96	10.41	7.11	4.65	4.28
PGSR	-	✓	6.26	6.09	16.64	12.81	11.23	8.26	10.32	7.52	3.19	3.04
PGSR	obs.	✓	6.51	6.34	17.14	13.34	11.75	8.69	9.83	7.26	3.08	2.89
3DGS + FS	-	✓	5.52	5.31	**14.78**	**11.25**	**9.62**	6.80	9.00	6.18	**2.35**	**2.14**

Table 1. **Zero-Shot Generalization Benchmark.** We compare the zero-shot performance of various synthetic dataset generation methods. All results are obtained from the RAFT-Stereo model. We report both the original NeRF-Stereo results as well as the reproduced results (denoted by *), with the model in both cases trained from scratch. The Fine-tuned column indicates whether the model was initialized with the SceneFlow checkpoint. The filtering column indicates which, if any, filtering methods are used on the data. Obs. indicates observability filtering was used to select scenes, whereas α indicates the pixel transparency was thresholded (denoted AO filtering in the NeRF-Stereo paper). 3DGS + FS denotes 3DGS with FoundationStereo [30] depth estimates. We highlight the **best** and second-best results.

set of freely captured images. However, this in turn resulted in COLMAP failing to estimate poses for several scenes, even when attempted multiple times.

Secondly, we found that the scene reconstruction with 3D Gaussian Splatting based methods can be prohibitively expensive, with some scenes failing to be reconstructed due to insufficient VRAM even when using an A100 with 40GB VRAM. On average, 10% of scenes or meshes were not reconstructed with 3DGS, GOF, 2DGS, or PGSR.

7. Discussion

Despite the excellent performance of the latest 3DGS-based methods on popular 3D reconstruction benchmarks from multi-view images, our study has revealed that they are still insufficient for in-the-wild scenarios. Although it was impossible to evaluate the accuracy of the 3D reconstructions due to the absence of associated ground truth data, assessing accuracy by using a proxy downstream task can be equally or even more valuable for judging the robustness of such methods. As such, the poor performance when training on the data rendered from meshes has been unexpected and suggests that the current benchmark performance is not representative of in-the-wild settings, which can be regarded as a valuable result and inspire future works. On the other hand, the FoundationStereo performance has shown to be representative even in a zero-shot setup, suggesting that the methodology from GS2Mesh [31] can be further improved. Moreover, while it could be argued that 3DGS-based methods could be replaced with NeRFs, it should be considered that 3DGS-based methods offer superior rendering speeds at comparable visual fidelity, which is essential when rendering large-scale datasets.

8. Conclusion

In this paper, we have investigated the feasibility of using 3D Gaussian Splatting-based methods to generate synthetic datasets to train a stereo network. Through qualitative assessment of state-of-the-art meshing-based methods, we find that the reconstructed geometry is consistently of insufficient quality with artifacts, noise, and holes present. This is consequently reflected in poor performance when fine-tuning a RAFT-Stereo network, even when applying filtering to only use the most well-observed parts of the meshes. We further show that by utilizing an expert knowledge transfer setup, where depth estimates from the FoundationStereo model are used as pseudo ground truth, we achieve much better performance. We attribute the success of this methodology to the fact that Gaussian Splatting methods capture high-fidelity visual images even if they are unable to capture underlying geometry due to the limitations of traditional algorithms. On the other hand, superior zero-shot performance of FoundationStereo can infer accurate geometry even from very limited visual signal. We consider our results to be a promising avenue for future work on generating synthetic stereo datasets and transferring expert knowledge into small and lightweight stereo networks, while offering unprecedented flexibility of capturing custom datasets from just freely captured images.

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

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. 3

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19640–19648, 2023. 3

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2

[4] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. 2024. 1, 5

[5] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *International Conference on Computer Graphics and Interactive Techniques*, 2024. 3

[6] Ainaz Eftekhar, Alexander Sax, Roman Bachmann, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10766–10776, 2021. 3

[7] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6:4947–4954, 2021. 3

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 3, 7

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013. 3

[10] Magnus Kaufmann Gjerde, Filip Slezák, Joakim Bruslund Haurum, and Thomas B. Moeslund. From neRF to 3DGS: A leap in stereo dataset quality? In *Synthetic Data for Computer Vision Workshop @ CVPR 2024*, 2024. 3

[11] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 807–814. IEEE, 2005. 1, 2

[12] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[13] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *ArXiv*, abs/2403.17888, 2024. 1, 3, 5

[14] Zedong Huang, Jinan Gu, Jing Li, and Xuefei Yu. A stereo matching algorithm based on the improved psmnet. *Plos one*, 16(8):e0251657, 2021. 2

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 1, 2, 3

[16] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 3

[17] Han Ling, Quansen Sun, Yinghui Sun, Xian Xu, and Xingfeng Li. Adfactory: An effective framework for generalizing optical flow with nerf. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20591–20600, 2024. 3

[18] 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. 2, 3

[19] 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, 3, 7

[20] M. Menze, C. Heipke, and A. Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5:427–434, 2015. 6

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf. *Communications of the ACM*, 65:99 – 106, 2020. 3

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[23] Sadra Safadoust, Fabio Tosi, Fatma Güney, and Matteo Poggi. Self-evolving depth-supervised 3d gaussian splatting from rendered stereo pairs. In *British Machine Vision Conference (BMVC)*, 2024. 3

[24] 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. 3, 6, 7

[25] Thomas Schöps, Johannes L. Schönberger, S. Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. *2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 2538–2547, 2017. 3, 6

[26] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8375–8384, 2021. 2

[27] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 855–866, 2023. 1, 3, 4, 5, 6

[28] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 2

[29] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3

[30] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *arXiv*, 2025. 2, 6, 8

[31] Yaniv Wolf, Amit Bracha, and Ron Kimmel. GS2Mesh: Surface reconstruction from Gaussian splatting via novel stereo views. In *European Conference on Computer Vision (ECCV)*, 2024. 6, 8

[32] 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. 2

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1

[34] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *ArXiv*, abs/2404.10772, 2024. 1, 3, 5