3D Reconstruction in Challenging Sparse View Setup

Nischal Maharjan
Universität des Saarlandes
ETH SSRF Fellowship 2025

Abstract

3D reconstruction is a core task in computer vision, traditionally addressed through Structure-from-Motion (SfM) pipelines that rely on feature matching and bundle adjustment (BA). However, in challenging scenarios with sparse views, these pipelines often fail to provide accurate results. Recent deep learning models, such as the Visual Geometry Grounded Transformer (VGGT), can jointly predict camera parameters, depth maps, point maps, and feature tracks, but their predictions typically lack global alignment. In this project, we explore combining VGGT predictions with BA (VGGT+BA) to improve sparse-view 3D reconstruction. We investigate two complementary directions: improving the inputs to BA and refining the BA block itself. For the inputs, we experiment with alternative tracking modules, including VGGSfM and MASt3R, and study the effects of query point selection and filtering correspondences. For BA, we evaluate iterative re-optimization with filters and different loss formulations with scale adjustments. Our results show that using VGGT priors with optimized BA improves the results compared to standalone baselines.

1. Introduction

3D reconstruction is a fundamental task in robotics, augmented and virtual reality (AR/VR), autonomous driving, and several other domains. A wide range of methods have been proposed to achieve reliable reconstruction, with Structure-from-Motion (SfM) being one of the most widely adopted classical approaches. SfM can provide accurate reconstructions when many viewpoints are available and when reliable correspondences can be established across images. However, its performance degrades in challenging scenarios such as low overlap between views, low-parallax camera trajectories, highly symmetric structures, or textureless scenes where feature point detection and matching become unreliable. In such cases, especially when only sparse views are available, 3D reconstruction becomes significantly more difficult. Recent deep learning–based approaches have attempted to overcome some of these limitations. In particular, the VGGT model has shown promising results in handling challenging settings, especially in texture less regions where classical methods often fail. VGGT predictions tend to preserve local structures, however they often suffer from global misalignment, as illustrated in Figure 1. We can see predicted point cloud(Red) has correct structure but is not perfectly aligned with the ground truth point clouds(green).To address this issue, we try combining VGGT predictions with bundle adjustment (BA). By using VGGT outputs as strong priors for BA, our goal is to achieve more accurate and globally consistent reconstructions, even in sparse and challenging view setups.

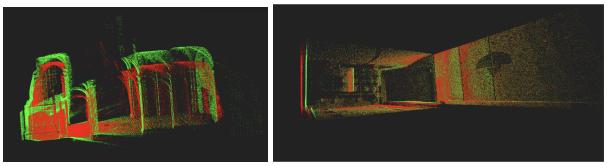

Figure 1. Limitations of VGGT

2. Related Works

Classical 3D Reconstruction Among the traditional approaches to 3D reconstruction, the Structure-from-Motion (SfM) paradigm is the most widely used appraoch. In this pipeline, corner points are detected, and their feature descriptors are computed, and correspondences are established across multiple views. These correspondences are then used to estimate camera poses and triangulate 3D points. Despite the emergence of more modern techniques, such pipelines remain popular due to their robustness, interpretability, and strong theoretical foundations.

Deep Learning Methods

With the advent of deep learning, many elements of SfM pipeline such as keypoint detection and feature matching have been enhanced by deep learning models often achieving state of art performance. VGGSfM [6] is one of the deep learning method which is end to end differentiable. DUSt3R [7] is a method used for Dense and Unconstrained Stereo 3D Reconstruction from a arbitrary set of images. MASt3R-SfM [1] is another method for un-

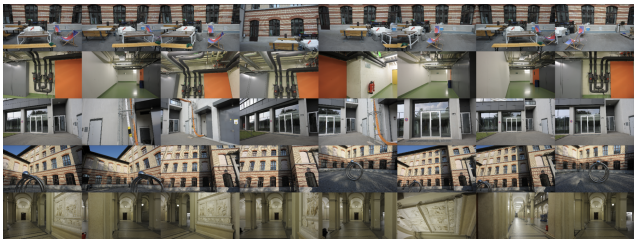

Figure 2. ETH3D Dataset

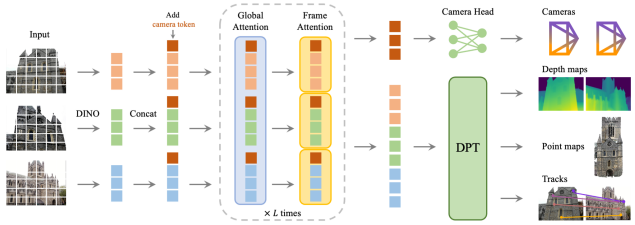

Figure 3. VGGT architecture

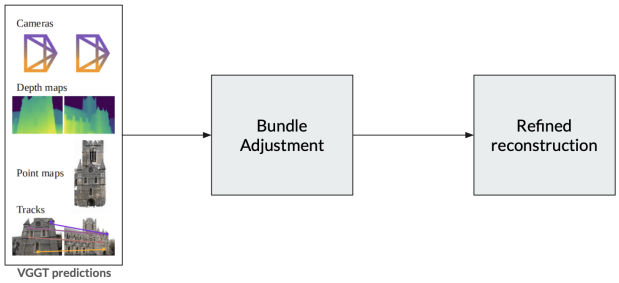

Figure 4. Overview of VGGT+BA

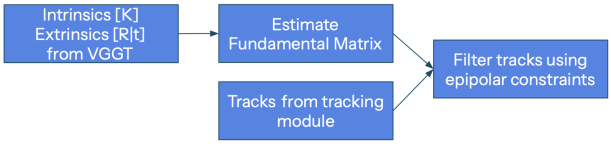

Figure 5. Filtering tracks

constrained SfM based upon MASt3R [2] which is tailored version of DUSt3R trained for with additional objective of dense matching. however these methods still require set of images and fails when the view points are limited. In this work we try to achieve similar performance with limited sparse number of viewpoints.

3. Methodology

3.1. Dataset

For this project we have used ETH3D dataset for comparison and evaluation of different methods. The dataset is comprised of images, depth and camera poses. We sample 8 images of a particular scene that has significant change in viewpoints but also has overlapping regions as shown in figure 2. Using only 8 sparse view points we aim to reconsrtuct the scene from them.

3.2. VGGT+BA

VGGT [5] is a transformer based model that predicts camera parameters, depth maps, point maps, and feature tracks directly. The outputs of VGGT looks promising. Our proposed method is to use the VGGT prediction as priors for bundle adjustment optimization (VGGT+BA). Figure 3 shows the architecture of VGGT model. It predicts camera parameters, depth maps, point maps, and tracks. We used VGGT+BA as our baseline as shown in figure 4 and tried to improve the performance over that benchmark.

3.3. Improving track predictions

The quality of bundle adjustment (BA) strongly depends on the accuracy of the input tracks. To investigate this, we replaced the default VGGT track predictions with alternative tracking modules. Specifically, we experimented with the VGGSfM [6] tracking module and MASt3R [2] tracks. Since MASt3R was specifically trained to find the dense matches it has a high potential to find more accurate correspondences leading to better input tracks. These alternative tracking strategies were integrated into our pipeline to study their effect on reconstruction performance.

3.4. Filtering Matches

Next way of improving tracks were to apply some filters to output of tracking module. Since the trackers we used has their own limitation we were passing tracks without any refinement. Hence we also experimented with refining the matches using epipolar constraint. Given matches $\{x1, x2\}$ from two view points they should satisfy the epipolar constraint $x_2^T F x_1 = 0$ where F is the fundamental matrix across two views. We have the camera parameters estimated by VGGT model which are used to estimate the fundamental matrix F and used it to filter the matches as shown in figure 5. However since there is uncertainty in the prediction of VGGT model we still can't be sure the filter is more accurate hence we also have relaxed the filtering threshold.

3.5. Reapplying BA(ReBA)

As per [4] since BA is severely affected by outliers, a second step of BA can significantly improve the results. So we also experimented by filtering certain points on basis of reprojection error and triangulation angle and reapplying the BA.

3.6. BA Optimization

Beside the inputs to the BA we can also change the optimization parameters. We specifically experimented with loss functions for bundle adjuster. We replaced trivial L2 loss with Cauchy and Soft_L1 loss.

4. Experiments and Results

VGGSfM vs MASt3R VGGSfM tracks had already performed better than the VGGT track predictions. Therefore here we experimented with MASt3R matches for predicting tracks and compared it with the VGGSfM tracking module. However the issue with MASt3R is that it estimates pairwise matches. We first used tracks with **tracklen=2** for all pairs using dense matches. But the performance was not good for track length of 2. We also experimented by estimating sparse query points using superpoint for query image and finding the matches across remaining 7 images in order to increase the tracklen.

	Metrics(Support =109)	VGGSfM	MASt3R
Intrinsics	fovx_error(deg) ↓	0.98	**0.96**
	fovy_error(deg) ↓	1.60	**1.00**
Extrinsics	auc@01(%) ↑	**74.90**	71.13
	auc@03(%) ↑	**83.38**	80.23
	auc@05(%) ↑	**86.78**	84.26
	auc@10(%) ↑	**90.49**	88.96
	auc@20(%) ↑	**93.42**	92.36
	auc@30(%) ↑	**94.77**	94.03

Table 1. Camera metric comparison for VGGSfM vs MASt3R tracking Module

	Metrics(Support =109)	VGGSfM	MASt3R
Error	rmse_mean(cm) ↓	899.36	**434.32**
	rmse_median(cm) ↓	**6.69**	10.56
AUC	auc@02cm(%) ↑	**20.67**	16.59
	auc@04cm(%) ↑	**32.28**	27.97
	auc@06cm(%) ↑	**40.23**	35.97
	auc@08cm(%) ↑	**46.20**	42.10
	auc@10cm(%) ↑	**50.92**	47.00

Table 2. 3D metric comparison for VGGSfM vs MASt3R tracking Module

However we found that even though MASt3R was trained for task of finding matches tracking accuracy was better for VGGSfM than the MASt3R as shown in table 3. We also notice that using MASt3R matches we have better intrinsic paramters prediction for camera but the extrinsic parameters are better for VGGSfM as shown in table 1. Regarding the 3D reconstruction accuracy the VGGSfM tracks are better than the MASt3R tracks as indicated in table. 2

	Metrics(Support =109)	VGGSfM	MASt3R
Error	tracking_error/mean ↓	**2.13**	4.07
	tracking_error/median ↓	**0.90**	2.14
Statistics	mean_track_length ↑	3.79	3.85
	median_track_length ↑	3.89	3.87
	max_track_length ↑	7.07	7.11
	full_track_percentage ↑	6.85	5.43

Table 3. Tracking Statistics and error comparison for VGGSfM vs MASt3R tracking Module

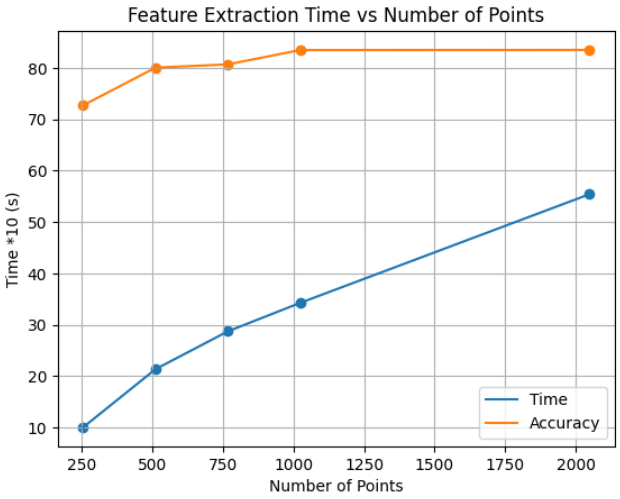

Figure 6. Query points effect on camera accuracy and time complexity

Effect of number of query points Since we are using the superpoint for the estimation of the query points in the query image and finding matches in remaining ones, we also tried to see the effect of specifying certain number of query points. Increasing the query point to maximum limit has only slight improvement in the metrices but the time complexity was very high due to large number of points to be optimized. If we are to just optimize the camera paramters few points should have been sufficient. Decreasing the query points it should be able to achieve similar performance in camera parameters estimation without significant decrease in accuracy but within very less time complexity. We can see in figure 6 that the time taken for the bundle adjustment decreases with decrease in number of query points with very insignificant drop in the accuracy.

Epipolar constraint Filter Using Epipolar constraint filter did increase the tracking accuracy as shown in table 4 however the final reconstruction perfromance was not im-

proved. Using epipolar constraint filter reduced the tracklen across the images and that might explain the underperformance in reconstruction metrics eventhough we have the better tracking error.

	Metrics(Support =84)	without filter	with filter
Error	mean_error ↓	2.11	**1.70**
	median_error ↓	0.89	**0.82**
Statistics	mean_tracklen ↑	3.99	2.77
	median_tracklen ↑	4.12	2.68
	max_tracklen ↑	7.29	6.52
	full_track% ↑	7.95	3.12

Table 4. Tracking Statistics and error comparison for with and without epipolar filter

Reapplying BA Reapplying bundle adjustment with some filter on the resulting reconstruction from first pass BA has slight improvement in the performance but not that significant indicating we might need some better filtering before reapplying the second pass of bundle adjustment.

Loss Functions Among trivial(L2 loss), Soft_L1 loss and robust(Cauchy loss) we find that Cauchy Loss has better improvement in the reconstruction metrics as shown in tables 5 and 6. However we see there are some outltiers based on mean rmse error in 3d reconstruction points.

	Metrics	L2_loss	Soft_L1	Cauchy
Intrinsics	fovx_error(deg) ↓	1.16	1.07	**0.99**
	fovy_error(deg) ↓	1.64	1.41	**1.22**
Extrinsics	auc@01(%) ↑	72.44	76.63	**78.60**
	auc@03(%) ↑	81.36	84.09	**85.44**
	auc@05(%) ↑	85.04	86.98	**88.10**
	auc@10(%) ↑	89.11	90.27	**90.95**
	auc@20(%) ↑	92.31	93.04	**93.40**
	auc@30(%) ↑	93.83	94.42	**94.67**

Table 5. Camera metric comparison for loss functions

	Metrics	L2_loss	Soft_L1	Cauchy
Error	rmse_mean(cm) ↓	**847.48**	2121.21	3621.12
	rmse_median(cm) ↓	7.55	8.87	**6.31**
AUC	auc@02cm(%) ↑	20.12	21.80	**22.96**
	auc@04cm(%) ↑	31.68	33.40	**34.74**
	auc@06cm(%) ↑	39.56	41.30	**42.63**
	auc@08cm(%) ↑	45.48	47.24	**48.53**
	auc@10cm(%) ↑	50.16	51.92	**53.19**

Table 6. 3D metric comparison for different loss functions

Also the change of scale show prominent effect on the performance. As the scale is decreased the performance increases and vice versa as shown in tables 7 and 8. So cauchy loss with scale of 0.05 has best performance in comparison to other setting of loss and hyperparameters.

Metrics	Cauchy Loss scales						
	0.05	0.1	0.2	0.5	1	2	3
fovx_err ↓	**0.66**	0.66	0.69	0.85	0.99	1.07	1.10
fovy_err ↓	**0.81**	0.81	0.93	1.12	1.22	1.40	1.46
auc@01 ↑	**84.54**	84.42	83.65	81.40	78.60	76.44	75.36
auc@03 ↑	**88.98**	88.42	88.39	87.34	85.44	84.01	83.29
auc@05 ↑	**90.76**	90.23	90.28	89.54	88.10	86.93	86.45
auc@10 ↑	**92.92**	92.43	92.47	92.01	90.95	90.24	89.99
auc@20 ↑	**94.80**	94.35	94.33	94.06	93.40	92.98	92.89
auc@30 ↑	**95.87**	95.39	95.35	95.13	94.67	94.39	94.32

Table 7. Effect of different scale for Cauchy Loss function on Camera metrics

Metrics	Cauchy Loss scales						
	0.05	0.1	0.2	0.5	1	2	3
median ↓	**5.49**	5.65	5.59	5.82	6.31	7.05	9.11
auc@02 ↑	24.93	24.93	**25.07**	24.24	22.96	22.04	21.38
auc@04 ↑	**37.13**	37.13	37.03	36.05	34.74	33.70	32.93
auc@06 ↑	**45.17**	45.16	45.00	43.98	42.63	41.61	40.79
auc@08 ↑	**51.08**	51.06	50.92	49.87	48.53	47.53	46.72
auc@10 ↑	**55.68**	55.64	55.53	54.48	53.19	52.18	51.40

Table 8. Effect of different scale for Cauchy Loss function on 3D metrics

5. Future Enhancements

There is lot of ways we can improve on the current state. From all the experiments conducted it seems having increased track length across all views has most prominent effect on the final outcome. Also there could be redundant tracks so completing and merging tracks could improve the final results. Use of method mentioned in Pixel-Perfect SfM [3] in order to refine the keypoint for better tracks can be another direction which can be explored.

6. Conclusion

In conclusion, our experiments demonstrate that using VGGT predictions as priors for Bundle Adjustment (VGGT+BA) yields better reconstruction results than relying on VGGT predictions alone, particularly in terms of

achieving improved global alignment. Furthermore, we observe that VGGSfM tracks outperform those generated by MASt3R, as the latter is based on pairwise matching. Our analysis also suggests that track length plays a more significant role in the final reconstruction quality than tracking accuracy itself. Thus, enhancing track length across images could further improve performance. Finally, careful optimization of bundle adjustment parameters can also lead to better results. Among the loss functions experimented, the Cauchy loss with a lower scale provides more robust results compared to the trivial loss and the Soft L1 loss.

References

[1] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion, 2024. 1

[2] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 2

[3] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement, 2021. 4

[4] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[5] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025. 2

[6] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual geometry grounded deep structure from motion, 2023. 1, 2

[7] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2024. 1