ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Uncertainty Quality of VGGT: An Analysis on the DTU Benchmark Dataset

Markus Hillemann[1], Robert Langendörfer[1], Steven Landgraf[1], Markus Ulrich[1]

[1]Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany -
(markus.hillemann@, steven.landgraf@, robert.langendoerfer@, markus.ulrich@)kit.edu

Keywords: 3D Reconstruction, 3D Foundation Models, Feed Forward, Multi-View Stereo, Uncertainty Estimation

Abstract

Visual Geometry Grounded Transformer (VGGT) has already attracted a great deal of attention in a short period of time, not least due to the Best Paper Award at CVPR-2025. Similar to DUSt3R and MASt3R, VGGT aims to bring about a paradigm shift by replacing established methods like bundle adjustment and feature matching with a simple, unified, feed-forward neural network that predicts camera poses, depth maps, and dense 3D structure directly from multiple images of a scene in a few seconds. A key aspect is its ability to process an arbitrary number of views consistently in a single forward pass without any post-processing or iterative optimization. For photogrammetry, this opens new possibilities for real-time, scalable, and accessible 3D reconstruction. In this context, not only high reconstruction accuracy but also high-quality uncertainty estimates are crucial, as they foster trust and enable robust quality assurance. This paper therefore investigates the quality of VGGT’s uncertainty predictions. The analysis identifies an effective confidence threshold for filtering VGGT’s raw output and demonstrates that enhancing uncertainty quality holds strong potential for improving the accuracy of its 3D reconstructions.

1. Introduction

Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025) has received substantial attention for pushing feed-forward, learning-based 3D reconstruction towards a unified paradigm. Similar to DUSt3R (Wang et al., 2024b) and MASt3R (Leroy et al., 2024), VGGT aims to replace classical, multi-stage SfM+MVS pipelines with a single forward pass that jointly estimates camera poses, depth maps, and dense 3D structure from multiple images. A key innovation is its ability to process an arbitrary number of views consistently without explicit post-optimization, which is particularly appealing for photogrammetric use cases requiring scalability and fast processing. In this context, not only geometric accuracy but also reliable uncertainty estimates are essential for trustworthy operation and rigorous quality assurance (Landgraf et al., 2025b). Moreover, uncertainty predictions can be used to improve 3D reconstruction accuracy, e.g., simply by filtering the redundant output of VGGT as depicted in Fig. 1.

Despite VGGT’s impressive progress in unifying multi-view reconstruction (Wang et al., 2025), the reliability of its uncertainty estimates remains largely unexplored. As noted by Zhang et al. (2025b), “another critical and underexplored area is uncertainty quantification. For safety-critical applications such as robotics, principled methods for estimating the reliability of the reconstructed geometry are essential”. Existing works have primarily focused on improving geometric accuracy and generalization, while overlooking the potential of leveraging model-intrinsic uncertainties as practical indicators of reliability or failure. In particular, it remains unclear whether VGGT’s uncertainty already encodes meaningful information about reconstruction quality. This motivates our study, which systematically investigates how useful VGGT’s native uncertainty estimates are out of the box without introducing additional uncertainty quantification mechanisms.

We deliberately evaluate the feed-forward variant of VGGT, i.e., the raw predictions without post-hoc bundle adjustment, because only then are the reported uncertainties directly interpretable and not altered by subsequent optimization.

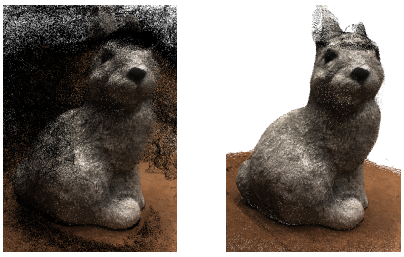

Figure 1. Qualitative effect of filtering the raw output of VGGT with a confidence threshold. Left: raw VGGT point maps, right: filtered with a confidence threshold of 2.0.

Our contributions are threefold:

- We present the first systematic analysis of VGGT’s uncertainty quality. We perform the evaluation on the DTU benchmark and consider both reconstruction branches, via point maps and via depth maps.
- We show that a causally meaningful confidence threshold of 2.0 effectively filters the raw outputs and improves the accuracy–completeness trade-off across scenes.
- We provide a comprehensive evaluation of the uncertainties with established metrics like PAvPU (Mukhoti and Gal, 2018) and AUSE (Ilg et al., 2018), revealing potential for improvement. Enhancing uncertainty quality can also contribute to improved reconstruction accuracy.

In summary, we investigate whether VGGT’s native aleatoric uncertainties already encode actionable signals for photogrammetric quality assessment, and how such signals can be leveraged in practice. The remainder of the paper reviews related work (Section 2), introduces the basics of VGGT and its uncertainty formulation (Section 3), details the evaluation setup (Section 4), and presents results on the DTU dataset (Aanæs et al., 2016), including the quality of VGGTs 3D reconstruction as well as its uncertainty quality (Section 5).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-665-2026

2. Related work

Traditional Structure from Motion and Multi-view Stereo. Reconstructing 3D scenes from a set of 2D images has been one of the most fundamental tasks in photogrammetry (Zhou et al., 2024). Thereby, Structure from Motion (SfM) recovers the camera orientation, i.e., interior and exterior orientation, and a sparse 3D point cloud by detecting and matching local features across multiple views and solving a bundle-adjustment optimization (Hartley and Zisserman, 2003; Schönberger and Frahm, 2016). Multi-view Stereo (MVS) takes this process one step further by densely reconstructing the geometry of a scene from multiple overlapping images (Seitz et al., 2006; Stathopoulou and Remondino, 2023). Even though traditional SfM + MVS pipelines, e.g., COLMAP, (Schönberger and Frahm, 2016; Schönberger et al., 2016), can achieve a very high accuracy and completeness, they typically require images with generous overlap as well as clearly textured areas, and require lengthy processing times (Liu et al., 2025; Wu et al., 2025).

Feed-forward Learning-based 3D Reconstruction. Recent learning-based approaches have begun to replace the traditional SfM + MVS pipelines with end-to-end networks. While earlier work used learning for individual parts, for example, SuperPoint (DeTone et al., 2018) and SuperGlue (Sarlin et al., 2020) for keypoint detection and feature matching, they still relied on the iterative and often fragile nature of the SfM stage (Zhang et al., 2025b). However, after the seminal work of DUSt3R (Wang et al., 2024b), a new paradigm, which embeds the entire workflow into a single feed-forward model, was created. DUSt3R and its successors (e.g., MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025)) employ transformer-based architectures (Vaswani et al., 2017) to jointly regress camera poses and dense 3D geometry from an arbitrary set of images in a single forward pass. After this groundbreaking paradigm shift in 3D reconstruction, there has been an explosion of follow-up work (Tang et al., 2025; Yang et al., 2025; Zhang et al., 2025a,b).

Uncertainty Quantification in Deep Vision Models. Neural networks often produce some notion of confidence or uncertainty. In Bayesian modeling, uncertainty is usually decomposed into aleatoric (data) and epistemic (model) uncertainty (Der Kiureghian and Ditlevsen, 2009). Aleatoric uncertainty captures inherent noise in the observations (e.g., sensor noise, low-texture regions) and cannot be reduced by more data, whereas epistemic uncertainty reflects the model's lack of knowledge and can be mitigated with more data (Kendall and Gal, 2017; Gawlikowski et al., 2023). This distinction is critical since Wang et al. (2025) state that VGGT outputs aleatoric uncertainty maps only, and does not account for epistemic uncertainty. We will discuss VGGT's uncertainty modeling in detail in Section 3. Unlike epistemic uncertainty, which typically requires multiple forward passes (e.g., via Monte Carlo dropout), aleatoric uncertainty can be learned directly from the data in a single forward pass. While these can be incorporated by various techniques like Bayesian Neural Networks (Gal and Ghahramani, 2016; Neal, 2012) or Ensembling (Lakshminarayanan et al., 2017; Ganaie et al., 2022), they usually introduce high computational costs, lack rigorous theoretical analysis, or require careful design choices and modifications to the training process (He et al., 2023).

3. Basics of VGGT and its Uncertainty Estimation

VGGT (Wang et al., 2025) processes an arbitrary number of input images from a scene using a transformer backbone. It is designed as a simple, pure learning-based model, i.e., without any rule-based assumptions like camera models or epipolar constraints. The architecture alternates between frame-wise and global self-attention layers, allowing the model to capture both local and global geometric relationships across views. Each image is tokenized using a pretrained DINO encoder (Oquab et al., 2024), and special tokens are introduced to predict camera intrinsics and extrinsics. The outputs of VGGT are:

- exterior orientation and field of view for each image,
- depth maps per image and corresponding confidence maps,
- point maps in a global coordinate frame, i.e., one predicted 3D coordinate per pixel, and corresponding confidence maps,
- and dense feature maps for point tracking.

The model is trained to predict all these quantities jointly, by minimizing a weighted sum of four loss components, i.e., one loss component for each output type. This formulation is mathematically redundant (e.g., point maps can be derived from depth maps and camera parameters). This redundancy, however, improves performance through multi-task learning (Wang et al., 2025). In this paper, we focus on the 3D reconstruction, and hence on the depth maps and point maps as well as their corresponding confidence maps.

Depth Map and Point Map Loss of VGGT. The depth loss

$$\begin{aligned}\mathcal{L}_{\text{depth}} = &\left\| \boldsymbol{\Sigma}_D \odot (\hat{\boldsymbol{D}} - \boldsymbol{D}) \right\| \\ &+ \left\| \boldsymbol{\Sigma}_D \odot (\nabla \hat{\boldsymbol{D}} - \nabla \boldsymbol{D}) \right\| - \alpha \log \boldsymbol{\Sigma}_D\end{aligned} \tag{1}$$

is adapted from DUSt3R (Wang et al., 2024b) and includes both residual and gradient terms, weighted by the predicted uncertainty map $\boldsymbol{\Sigma}_D \in \mathbb{R}^{H \times W}$, where H and W are the image height and width. In Equation (1), $\hat{\boldsymbol{D}}$ is the predicted depth map, $\boldsymbol{D}$ the ground truth depth map, ∇ the spatial gradient operator, $\odot$ the element-wise product, and α a regularization weight.

The point map loss

$$\begin{aligned}\mathcal{L}_{\text{pmap}} = &\left\| \boldsymbol{\Sigma}_{\boldsymbol{P}} \odot (\hat{\boldsymbol{P}} - \boldsymbol{P}) \right\| \\ &+ \left\| \boldsymbol{\Sigma}_{\boldsymbol{P}} \odot (\nabla \hat{\boldsymbol{P}} - \nabla \boldsymbol{P}) \right\| - \alpha \log \boldsymbol{\Sigma}_{\boldsymbol{P}}\end{aligned} \tag{2}$$

is defined analogously, where $\hat{\boldsymbol{P}}$ is the predicted 3D point map, $\boldsymbol{P}$ the ground truth point map, and $\boldsymbol{\Sigma}_{\boldsymbol{P}} \in \mathbb{R}^{H \times W}$ is the predicted uncertainty map of the point maps.

In these loss formulations, the residuals are weighted by the predicted uncertainty $\boldsymbol{\Sigma}_{\boldsymbol{D}}$ or $\boldsymbol{\Sigma}_{\boldsymbol{P}}$, meaning the model learns to downweight regions with large errors. The gradient terms enforce local smoothness and consistency in the spatial structure. The log penalty on the uncertainty prevents the model from trivially inflating $\boldsymbol{\Sigma}_{\boldsymbol{D}}$ or $\boldsymbol{\Sigma}_{\boldsymbol{P}}$ to reduce the residuals. This formulation is not probabilistic in the strict sense (i.e., not derived from a likelihood), but rather a heuristic uncertainty-aware loss that balances accuracy and confidence.

This mechanism allows VGGT to express spatially varying reliability in its predictions, which is particularly relevant for photogrammetric applications where geometric accuracy and error quantification play an important role. These uncertainties are not post-hoc estimates but are intrinsic outputs of the model,

making them potentially suitable for downstream tasks such as uncertainty-aware outlier rejection, multi-view fusion, or meshing, as well as error propagation analysis.

Confidence and Uncertainty in VGGT. In the supplementary material of VGGT, Wang et al. (2025) describe that VGGT implements per-pixel *aleatoric uncertainty* maps Σ_D and Σ_P. The term *uncertainty* usually means that large values express high uncertainty and vice versa. In contrast, the official implementation actually predicts per-pixel *confidence* maps, where larger values represent higher confidence (lower uncertainty). Concretely, the weights $\boldsymbol{\Sigma} \in \{\boldsymbol{\Sigma_D}, \boldsymbol{\Sigma_P}\}$ which are passed through an `expp1` activation, yield confidence values

$$\boldsymbol{C} \;=\; \mathrm{expp1}(\boldsymbol{\Sigma}) \;=\; e^{\boldsymbol{\Sigma}} + 1\,. \tag{3}$$

Because of this activation, the values of the confidence maps are always larger than 1. However, this means that the confidence maps can contain values that represent negative weights $\boldsymbol{\Sigma}$, which are not meaningful in this context. Conversely, meaningful weights have a value larger than 2 in the confidence maps. We investigate this observation empirically in our experiments.

For analyses that conceptually operate on uncertainty, we convert confidence into a monotonic, scale-free uncertainty proxy $\boldsymbol{U}$, with

$$\boldsymbol{U} \;=\; -\log(\boldsymbol{C} - 1)\,. \tag{4}$$

We filter raw predictions in the confidence space (e.g., keep predictions with $\boldsymbol{C} > 2.0$), but compute uncertainty metrics (PAvPU, pAC, pUI, sparsification, AUSE) in the uncertainty space using $\boldsymbol{U}$.

4. Evaluation Methodology

This section describes how we applied VGGT, provides relevant details about the DTU dataset and evaluation, and specifies the metrics used to evaluate point cloud and uncertainty quality.

VGGT. We consistently evaluate the variant referred to as *Feed-Forward* in Wang et al. (2025), i.e., the raw outputs of VGGT without parameter adjustments, fine-tuning, or post-processing steps. This means in particular that we do not perform bundle adjustment in the post-processing. The reason for this is that the uncertainties can only be meaningfully evaluated for the feed-forward variant of VGGT, since they are not propagated by the integrated bundle adjustment with VGGSfM (Wang et al., 2024a). We use the pretrained VGGT-1B checkpoint with 1.26 B parameters. By default, VGGT resizes all images to a size of 518×518 pixels and uses centered zero padding, i.e., padded pixels at the edges of the images are black. The points in the point maps refer to a coordinate system that is defined by the first image, so that these points can be evaluated directly as a point cloud. To compute point clouds from the depth maps, the estimated exterior orientation and field of view is used. In the following, reconstruction results obtained from the point map branch are referred to as *VGGT-p*, while results from the depth map branch are denoted as *VGGT-d*.

DTU Dataset. The DTU dataset (Aanæs et al., 2016) is an established benchmark for evaluating MVS and 3D reconstruction algorithms. It was specifically designed to provide high-quality ground truth geometry under controlled conditions, making it particularly relevant for photogrammetric evaluation. It consists of 80 indoor scenes captured using a robotic arm equipped with a high-resolution camera. Each scene is captured from 49 or 64 viewpoints arranged on a hemispherical grid. For each scene, the dataset provides:

- High-resolution RGB images,
- Interior and exterior orientation,
- Structured-light-based reference geometry (ground truth),
- Surface normals and visibility masks.

The exterior orientation has been determined with a calibration board and resection in space with high accuracy. The scenes are illuminated under seven multiple lighting conditions (L1 - L7) to enable testing the robustness against photometric variation. Typically, light settings L3 or L7 are used for evaluations. We select L3 because it results in less overexposure in the images. The reference geometry was acquired using a structured light scanner with an average precision of approximately 0.14 mm for the surface points, which corresponds to approximately 0.6 pixels, allowing for precise evaluation of reconstructed point clouds and depth maps. The scenes include both Lambertian and non-Lambertian surfaces, as well as varying levels of geometric complexity and occlusion.

DTU Evaluation. The DTU benchmark provides a standardized evaluation protocol. The evaluation metrics are

$$\text{Accuracy} = \frac{1}{|\boldsymbol{\mathcal{P}_r}|} \sum_{\boldsymbol{p}_r \in \boldsymbol{\mathcal{P}_r}} \min_{\boldsymbol{p_g} \in \boldsymbol{\mathcal{P}_g}} \|\boldsymbol{p_r} - \boldsymbol{p_g}\|_2, \tag{5}$$

i.e., the mean Euclidean distance from the points $\boldsymbol{p_r}$ in the reconstructed point cloud $\boldsymbol{\mathcal{P}_r}$ to the closest points $\boldsymbol{p_g}$ in the ground truth point cloud $\boldsymbol{\mathcal{P}_g}$,

$$\text{Completeness} = \frac{1}{|\boldsymbol{\mathcal{P}_g}|} \sum_{\boldsymbol{p_g} \in \boldsymbol{\mathcal{P}_g}} \min_{\boldsymbol{p_r} \in \boldsymbol{\mathcal{P}_r}} \|\boldsymbol{p_g} - \boldsymbol{p_r}\|_2, \tag{6}$$

i.e., the mean Euclidean distance from ground truth points to the closest reconstructed point, and the overall score, i.e., the average of Accuracy and Completeness, which is also known as Chamfer Distance. Note that for all three metrics, lower values indicate better performance (↓). For the terms ‘Accuracy’ and ‘Completeness’, this differs from the usual interpretation. If the metrics from Equations (5) and (6) are explicitly meant, we use the capitalized terms throughout the paper.

To compute reasonable distances, the reconstructed point clouds need to be registered to the ground truth. Thus, we estimate a coarse 3D similarity transformation based on the bounding boxes which is refined with the Iterative Closest Point algorithm by minimizing the Accuracy. The correctness of the registration is verified manually for all point clouds.

In the DTU evaluation, only points within a predefined evaluation mask are considered, ensuring that only visible and relevant regions are evaluated. Additionally, the point clouds are downsampled to a voxel size of 0.2 mm, which corresponds to a conservative estimate of the accuracy of the ground truth (Aanæs et al., 2016). Following Yariv et al. (2020) and a series of subsequent works like from Huang et al. (2024) or Chen et al. (2024), we evaluate on a fixed subset of 15 scenes.

Uncertainty Evaluation. To assess the uncertainty quality of VGGT, we employ the Patch Accuracy vs. Patch Uncertainty (PAvPU) metric (Mukhoti and Gal, 2018)

$$\text{PAvPU} = \frac{(AC + IU)}{(AC + AU + IC + IU)}\;, \tag{7}$$

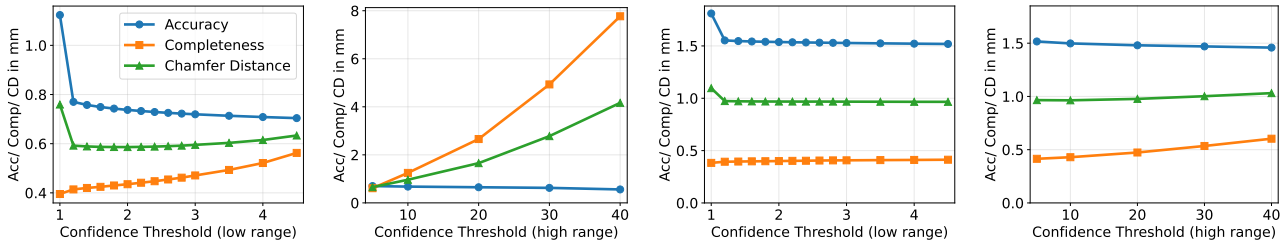

Figure 2. Accuracy, Completeness and Chamfer Distance (in mm) ↓ of VGGT-p (left) and VGGT-d (right) on the DTU dataset depending on the selected confidence threshold for filtering the raw output.

where AC, AU, IC, and IU denote the number of accurate–certain, accurate–uncertain, inaccurate–certain, and inaccurate–uncertain predictions, respectively. PAvPU is based on two intuitive desiderata: (1) if a model is certain (confident) about its prediction, it should be accurate, and (2) if a model is uncertain, it may or may not be accurate (Mukhoti and Gal, 2018). To determine whether a point is accurate or inaccurate and certain or uncertain, two thresholds τ_a and τ_u are necessary. We consider a point as accurate, when its Accuracy is smaller than an intuitive threshold of $\tau_a = 1\,\text{mm}$. Following Landgraf et al. (2025a), we further consider a prediction as certain, when its uncertainty value is smaller than the median of the uncertainties ($< \tau_u = 0.5$), and vice versa. Mukhoti and Gal (2018) additionally propose two intuitive probability metrics, that we also report: $\text{pAC} = \frac{AC}{AC+IC}$, i.e., the probability that the model is accurate given that the uncertainty is below τ_u and $\text{pUI} = \frac{IU}{IU+IC}$, i.e., the probability that the uncertainty of the model exceeds τ_u given that the prediction is inaccurate. Consequently, higher values for PAvPU, pAC, and pUI generally represent a higher uncertainty quality.

We further evaluate the uncertainty quality using Sparsification Curves and the Area Under the Sparsification Error (AUSE) (Ilg et al., 2018). The former visualize how well uncertainties correlate with actual prediction errors. They are constructed by sorting the predictions according to their uncertainty. Then, the most uncertain samples are gradually removed, while the remaining subset is used to compute the corresponding error metric (e.g., Chamfer Distance). If the uncertainty estimates perfectly reflect prediction errors, the error metric should decrease rapidly as uncertain samples are removed. The AUSE quantifies the deviation between the actual sparsification curve and an ideal curve, which represents the best possible ranking, which is obtained using ground-truth Accuracy instead of predicted uncertainty. It is calculated as the area between both curves, with lower AUSE values indicating higher uncertainty quality.

Finally, we present correlation curves. These represent the mean uncertainty across all points that lie within an Accuracy interval. The intuition behind the correlation curves is that there should tend to be a linear relationship between accuracy and uncertainty, such that the worse the accuracy, the higher the uncertainty should be. If this linear relationship exists and the uncertainty is calibrated, a metric statement about accuracy can be derived from the uncertainty prediction, which is a prerequisite for subsequent statistical evaluations such as hypothesis testing. Deviations from this intuition are represented by accurate but uncertain points, which are acceptable and should therefore have no negative impact on an uncertainty quality metric. Consequently, for points with an Accuracy better than τ_a, this requirement does not apply strictly.

5. Results of VGGT on the DTU-Dataset

This section presents the evaluation results of VGGT on the DTU dataset, with a particular focus on the uncertainty quality. Since uncertainty quality is best interpreted relative to the underlying prediction error, we first report the reconstruction performance using standard DTU evaluation metrics, as outlined in Section 4. Lastly, we analyze the correlation between VGGT's prediction errors and its estimated uncertainties.

5.1 Preliminary study on confidence filtering

Filtering the raw output from VGGT based on confidence values has a significant impact on point cloud quality (see Fig. 1). Therefore, we first conduct a preliminary study to determine the best confidence threshold.

Fig. 2 presents the results of the DTU evaluation in terms of Accuracy, Completeness, and Chamfer Distance averaged over all scenes, depending on the selected confidence threshold for filtering. For both VGGT-p and VVGT-d, the analysis shows that the higher the confidence threshold is set, the lower the Accuracy and the higher the Completeness. However, Completeness increases faster than Accuracy decreases. Very similar Chamfer Distances are achieved after filtering with confidence thresholds ranging from 1.2 to 5. However, if all points are included in the evaluation (threshold 1.0), the result is significantly worse. This clearly indicates that data points with the lowest confidences should be excluded from further analysis. For VGGT-d (Fig. 2, right), filtering based on the confidence threshold has only a very minor effect. This is due to poor confidence predictions, which we will discuss again in Section 5.3. For VGGT-p (Fig. 2, left), the optimal confidence threshold across the scenes is always between 1.2 and 3.5 with a mean value of 1.9 and a standard deviation of 0.6. Consequently, empirical evidence also shows that the causally meaningful (cf. Section 3) threshold of 2.0 is a good choice. Therefore, further investigations are carried out based on the outputs filtered with a threshold of 2.0.

5.2 Performance of VGGT on the DTU-Dataset

Table 1 provides an overview of a number of commonly used, current methods for 3D reconstruction from multiple images. Some of these methods require known external orientations as input, which is usually calculated by SfM using COLMAP. The values are not exactly comparable, as different subsets of scenes were used for evaluation in some cases, and the runtimes were determined using different hardware. However, the table is intended to serve as a rough overview only. VGGT achieves mean Chamfer Distances of less than one millimeter for both branches, even though it was not trained on the DTU data and requires only a few seconds of computation time on an A100

Table 1. Quantitative results of mean Chamfer Distance (CD) (in mm) ↓ across multiple scenes of the DTU dataset. The table is intended to serve as a rough overview. The values are not exactly comparable, as different subsets of scenes were used for evaluation in some cases, and the runtimes were determined using different hardware. We were unable to find runtimes for the DTU dataset for MASt3R, COLMAP, and DUSt3R in the literature, so we report a rough estimate based on the processing of 49 images. The results for the two VGGT variants are based on our evaluation with an A100 and measure the inference time of the model.

Method	**CD**	**Time**
Known exterior orientation		
Neuralangelo (Li et al., 2023)	0.61	> 128h
2DGS (Huang et al., 2024)	0.80	∼20min
PGSR (Chen et al., 2024)	0.52	∼30min
MASt3R (Leroy et al., 2024)	0.37	3-10min
Unknown exterior orientation		
COLMAP (Schönberger et al., 2016)	0.53	1-2h
DUSt3R (Wang et al., 2024b)	1.74	3-10min
VGGT-p (Wang et al., 2025)	0.59	∼ 5sec
VGGT-d (Wang et al., 2025)	0.97	∼ 5sec

GPU. Most other methods are considerably slower. Despite the longer runtimes, however, these methods hardly achieve a higher point cloud quality. One exception is MAST3R, which is fast and has a lower Chamfer Distance than both VGGT variants. However, ground truth poses were used to achieve this result (Wang et al., 2025). For comparison, Wang et al. (2025) also report results for the DTU dataset. They achieve a lower Chamfer Distance of 0.38 mm and report a runtime of 1.8 seconds with a more powerful graphics card than ours. We suspect that Wang et al. (2025) perform bundle adjustment in post-processing, thereby achieving a lower Chamfer Distance. However, this and other details of their experiment are not explicitly mentioned in the paper, so the experiments cannot be reproduced exactly.

Table 2 shows the results of VGGT on the DTU dataset in terms of Accuracy (Acc), Completeness (Comp), and Chamfer Distance (CD). VGGT-p, i.e., the point map branch, is significantly more accurate than VGGT-d on the DTU dataset. In contrast, VGGT-d is more complete on average. The quality of the results is very similar for the different scenes; no pattern can be discerned with this regard.

Fig. 3 shows the ground truth together with qualitative results of VGGT-p and VGGT-d for scenes 24, 69, and 122. The point clouds are colored based on the predicted color, the pointwise Accuracy (cf. Equation (5) without averaging over the ground truth points), the uncertainty (cf. Equation (4)), and the classification of each point into accurate–certain (AC), accurate–uncertain (AU), inaccurate–certain (IC), and inaccurate–uncertain (IU). In this section, we will focus on a discussion of the first two columns of the figure. The uncertainties are discussed in Section 5.3. The predicted point clouds have high visual quality. The black points in scene 24 show minor reconstruction artifacts, whereby the reconstruction with VGGT-p contains more of these artifacts than VGGT-d. These points are artifacts that are caused by the zero padding that is applied by VGGT during cropping. They could therefore be easily filtered out. The coloring with pointwise Accuracy clearly illustrates the higher Accuracy of VGGT-p. Furthermore, Accuracy is good on flat surfaces but tends to be poorer for points directly at edges of the 3D geometry. The Accuracy is slightly worse for the windows in scene 24. Note that these are not real glass windows, as the building shown is a miniature model.

5.3 Quality of the Predicted Uncertainties

Table 3 shows the quality of the predicted uncertainties of VGGT-p and VGGT-d in terms of PAvPU, pAC, pUI, and AUSE. VGGT-p achieves a better uncertainty quality than VGGT-d for all metrics. In particular, the probability that a point is accurate when it is classified as certain by VGGT (pAC) is significantly lower for VGGT-d. VGGT-d thus exhibits the typical phenomenon of overconfidence (Guo et al., 2017). However, the quality of the uncertainty prediction of VGGT-p is also suboptimal. For example, the probability that an inaccurate point will be rated as uncertain by the network is only 61.2%. Therefore, inaccurate points cannot simply be removed based on filtering uncertain points.

As mentioned, Fig. 3 also visualizes the uncertainty predictions and quality. As a reminder, the most important aspects for meaningful uncertainties are that (1) points with low uncertainty should be accurate and (2) inaccurate points should be uncertain. The right column of Fig. 3 can be interpreted as follows: Many green and orange points, but few red points indicate high uncertainty quality. Many green and blue points, but few orange points indicate good accuracy. Red points (inaccurate but certain) are the most problematic, as overconfidence can lead to critical misjudgments. The figure shows that the estimated uncertainties are suboptimal. For example, VGGT-d often provides a low uncertainty for points on the ground, even though these are often inaccurate. But even with VGGT-p, there are inaccurate points that are considered certain, e.g., in scene 24 on the roof on the left or in scene 69 on the hat of the left snowman. In general, VGGT-d tends to report lower uncertainty estimates compared to VGGT-p, despite being less accurate. This underscores the above statement about the overconfidence of VGGT-d. We suspect that the predominant source of this overconfidence arises from VGGT providing uncertainty estimates only for depth maps, but not for the interior and exterior orientation. However, interior and exterior orientation are required to compute the point cloud from the depth maps. Interestingly, VGGT is always very uncertain about the windows in scene 24, even though the accuracy there is only slightly below average. We suspect that the reason for this lies in contextual knowledge of VGGT: During training, the model learned that windows are generally difficult to predict. However, the building model of scene 24 does not contain real glass windows, and the accuracy is actually better than the uncertainty would suggest.

Fig. 4 shows the sparsification curves for VGGT-p and VGGT-d for three scenes. When points are filtered out based on uncertainty, accuracy improves only slowly for both VGGT-p and VGGT-d. This clearly illustrates that the correlation between uncertainty and accuracy is suboptimal. Consequently, there is room for improvement in terms of uncertainty quality, which should be addressed in future research.

Finally, Fig. 5 visualizes the correlation curves. Ideally, the curves should rise linearly (with an arbitrary slope, since the uncertainties are unscaled). This is the case for points with a good Accuracy (<2 mm). However, the linear relationship does not apply to points with poorer Accuracy for many scenes. There are not many points with very poor Accuracy (>10 mm), which is why the curves in this area are very noisy.

Table 2. Accuracy (Acc), Completeness (Comp), and Chamfer Distance (CD) (in mm) per scene on the DTU dataset. VGGT-p denotes the point map branch and VGGT-d the depth map branch of VGGT.

Method	Metric	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
VGGT-p	acc ↓	0.86	0.95	0.97	0.68	0.82	0.56	0.85	0.76	0.75	0.69	0.58	0.63	0.71	0.61	0.65	**0.74**
	comp ↓	0.44	0.60	0.58	0.36	0.43	0.45	0.51	0.40	0.55	0.36	0.32	0.31	0.51	0.36	0.33	0.44
	CD ↓	0.65	0.77	0.78	0.52	0.62	0.50	0.68	0.58	0.65	0.53	0.45	0.47	0.61	0.49	0.49	**0.59**
VGGT-d	Acc ↓	1.24	1.55	1.85	1.55	2.03	1.61	1.38	1.71	1.71	1.50	1.28	1.39	1.11	1.38	1.65	1.53
	Comp ↓	0.30	0.42	0.42	0.33	0.75	0.39	0.31	0.62	0.49	0.37	0.30	0.36	0.25	0.34	0.35	**0.40**
	CD ↓	0.83	0.99	1.14	0.94	1.39	1.00	0.84	1.17	1.10	0.93	0.79	0.88	0.68	0.86	1.00	0.97

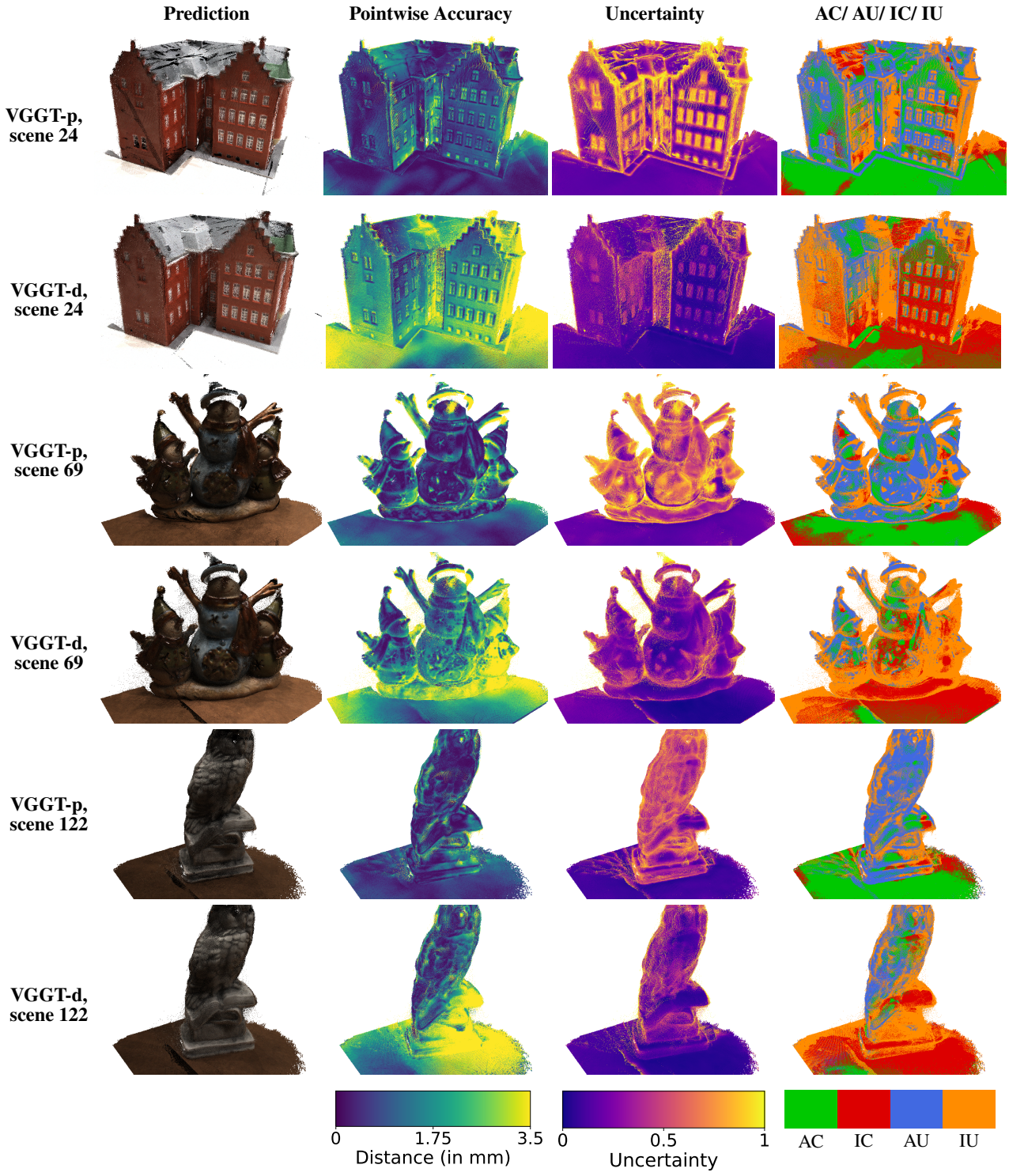

Figure 3. Qualitative results of VVGT for scenes 24, 69, and 122 of the DTU dataset.

Table 3. Uncertainty quality of VGGT-p (top) and VGGT-d (bottom).

Method	Metric	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
VGGT-p	PAvPU (in %) ↑	50.4	54.6	54.6	55.8	56.3	53.9	58.4	50.9	55.7	52.8	54.0	55.0	57.1	58.2	57.0	**55.0**
	pAC (in %) ↑	67.2	73.0	64.2	83.5	80.7	89.4	78.4	75.9	81.1	80.8	88.1	87.8	82.7	90.0	86.0	**80.6**
	pUI (in %) ↑	50.7	57.3	55.7	63.0	62.4	63.4	64.0	51.7	61.6	56.4	62.6	64.6	64.6	72.5	66.7	**61.2**
	AUSE ↓	0.47	0.46	0.46	0.28	0.36	0.26	0.34	0.45	0.34	0.39	0.26	0.26	0.31	0.24	0.24	**0.34**
VGGT-d	PAvPU (in %) ↑	52.7	53.2	52.3	54.3	51.8	52.8	56.4	52.6	54.5	53.2	54.4	59.7	60.0	53.7	50.3	54.1
	pAC (in %) ↑	52.4	49.9	36.9	47.2	35.3	44.5	53.5	46.9	44.6	50.4	53.9	56.8	66.4	49.9	38.9	48.5
	pUI (in %) ↑	52.6	53.0	51.8	53.7	51.3	52.4	56.0	52.4	53.7	53.0	54.4	59.2	61.4	53.5	50.2	53.9
	AUSE ↓	0.69	0.76	0.79	0.74	0.90	0.79	0.60	0.86	0.77	0.72	0.62	0.55	0.44	0.62	0.89	0.72

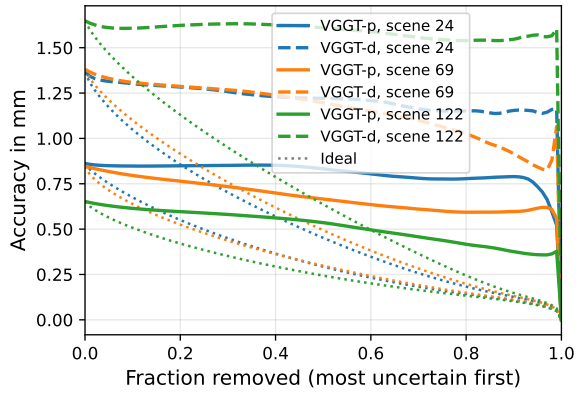

Figure 4. Sparsification curves of VGGT-p and VGGT-d for scenes 24, 69, and 122.

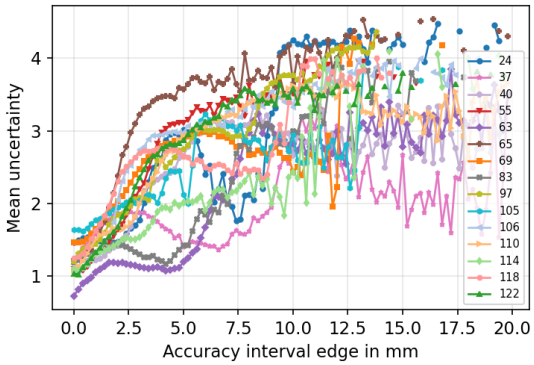

Figure 5. Correlation Curves of VGGT-d for all DTU scenes. The lines are interrupted when there are no points in an Accuracy interval.

6. Conclusion

We analyze the uncertainty quality of the depth maps and point maps predicted by VGGT on the DTU benchmark dataset. Our evaluations show that a confidence threshold of 2.0 is a good choice for filtering the raw output of VGGT. Since we have both a causal justification and empirical evidence from the evaluation on the DTU dataset, we believe that this threshold can be generally recommended as a starting point for potential dataset-specific fine tuning, since it does not necessarily lead to the best results for every single scene. Depending on specific application requirements, a different trade-off between accuracy and completeness may be more appropriate, which also affects the choice of the confidence threshold (higher for a focus on accuracy, lower for a focus on completeness). Moreover, we find that the point map branch is more accurate than the depth map branch for 3D reconstruction, when the feed-forward results are evaluated (i.e., no bundle adjustment in post-processing). The quality of the predicted uncertainties is also significantly better for the point map branch, while the depth map branch suffers from overconfidence in particular.

The analysis shows that improving the uncertainty prediction of VGGT has high potential to improve the accuracy of the 3D reconstruction. In principle, the output of VGGT is redundant, as it outputs a 3D point for each pixel. If the uncertainties were of higher quality, they would have more potential to increase the 3D reconstruction accuracy, e.g., by filtering uncertain (and therefore inaccurate) points, or by utilizing them for multi-view fusion. Potentially, higher uncertainty quality could be achieved by taking into account the epistemic uncertainty component. Finally, metric uncertainty quantification per predicted 3D point is currently not yet possible, but would be advantageous for many photogrammetric and metrology tasks and safety-critical applications.

References

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., Dahl, A. B., 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120, 153–168.

Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G., 2024. PGSR: Planar-based Gaussian Splatting for Efficient and High-fidelity Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2), 105–112.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, PMLR, 1050–1059.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., Suganthan, P. N., 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115, 105151.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R. et al., 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1), 1513–1589.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On calibration of modern neural networks. *International conference on machine learning*, PMLR, 1321–1330.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.

He, W., Jiang, Z., Xiao, T., Xu, Z., Li, Y., 2023. A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*.

Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S., 2024. 2d gaussian splatting for geometrically accurate radiance fields. *ACM SIGGRAPH 2024*, 1–11.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T., 2018. Uncertainty estimates and multi-hypotheses networks for optical flow. *Proceedings of the European Conference on Computer Vision (ECCV)*, 652–667.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M., 2025a. A comparative study on multi-task uncertainty quantification in semantic segmentation and monocular depth estimation. *tm-Technisches Messen*, 92(7-8), 298–310.

Landgraf, S., Hillemann, M., Ulrich, M., 2025b. Rethinking Semi-supervised Segmentation Beyond Accuracy: Reliability and Robustness. *arXiv preprint arXiv:2506.05917*.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, Springer, 71–91.

Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., Lin, C.-H., 2023. Neuralangelo: High-fidelity neural surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8465.

Liu, S., Yang, M., Xing, T., Yang, R., 2025. A Survey of 3D Reconstruction: The Evolution from Multi-View Geometry to NeRF and 3DGS. *Sensors*, 25(18), 5748.

Mukhoti, J., Gal, Y., 2018. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*.

Neal, R. M., 2012. *Bayesian Learning for Neural Networks*. 118, Springer Science & Business Media.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 1–31.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European conference on computer vision*, Springer, 501–518.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 1, IEEE, 519–528.

Stathopoulou, E. K., Remondino, F., 2023. A survey on conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record*, 38(183), 374–407.

Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z., 2025. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5283–5293.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. Vggt: Visual geometry grounded transformer. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, J., Karaev, N., Rupprecht, C., Novotny, D., 2024a. Vggsfm: Visual geometry grounded deep structure from motion. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21686–21697.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024b. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wu, X., Landgraf, S., Ulrich, M., Qin, R., 2025. An Evaluation of DUSt3R/MASt3R/VGGT 3D Reconstruction on Photogrammetric Aerial Blocks. *arXiv preprint arXiv:2507.14798*.

Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21924–21935.

Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y., 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2492–2502.

Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.-X., Liang, H., Xu, Z., Su, H. et al., 2025a. Advances in feed-forward 3d reconstruction and view synthesis: A survey. *arXiv preprint arXiv:2507.14501*.

Zhang, W., Wu, Y., Li, S., Ma, W., Ma, X., Li, Q., Wang, Q., 2025b. Review of Feed-forward 3D Reconstruction: From DUSt3R to VGGT. *arXiv preprint arXiv:2507.08448*.

Zhou, L., Wu, G., Zuo, Y., Chen, X., Hu, H., 2024. A comprehensive review of vision-based 3d reconstruction methods. *Sensors*, 24(7), 2314.