

MVM-IOD: An Industrial Object-Centric Benchmark Dataset for the Evaluation of 3D Reconstruction Methods

Robert Langendörfer, Markus Hillemann, Markus Ulrich

Machine Vision Metrology, Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany -
(robert.langendoerfer, markus.hillemann, markus.ulrich)@kit.edu

Keywords: industrial dataset, image-based 3D reconstruction, novel view synthesis, camera pose estimation, feed-forward geometry reconstruction

Abstract

3D object reconstruction, and camera pose estimation in industrial applications are challenging tasks, as errors are costly while the computation time is often limited. The complexity of typical industrial objects further complicates these tasks. Most of the existing datasets in this context do not depict realistic industrial scenarios. Therefore, we introduce the Machine Vision Metrology Industrial Object Dataset (MVM-IOD). Images of typical industrial objects are captured systematically, by moving a camera, mounted at the end effector of an industrial robot arm, on a hemisphere around the objects. MVM-IOD contains reference camera poses and reference 3D point clouds, the acquired RGB images of 9 objects and 2 background choices resulting in 18 scenes, which allows evaluation of all image based methods that compute a 3D reconstruction, camera poses, or novel views of a scene. Based on MVM-IOD, we extensively evaluate current SOTA 3D reconstruction and camera pose estimation methods, such as Structure from Motion, Multi-View Stereo, recent feed forward methods (Visual Geometry Grounded Transformer, π^3), and 2D Gaussian Splatting and report our findings as a baseline for future research. The experiments show that capture setups like ours generate out-of distribution images for feed forward methods, leading to suboptimal point clouds and camera poses. However, these out-of distribution images can be shifted closer to the training distribution by applying simple preprocessing steps. Consequently, in certain industrial applications, feed forward methods should be used with caution. The data and code are accessible via our project page: https://pimpilimpo.github.io/projects/Industrial_objects/

1. Introduction

In machine vision applications, such as anomaly detection, object pose estimation, and object grasping, an accurate 3D reconstruction and accurate camera poses are crucial (Steger et al., 2018). The lack of texture on the surface of most industrial objects makes these tasks harder with data captured with cameras (Karami et al., 2021). Furthermore, time is often a limiting factor, as many applications need to be performed in real time for cost efficiency (Yoon et al., 2003). In principle, there are various methods available to solve these tasks. However, to determine which methods work best on objects with certain properties (e.g., geometry, texture, material) a benchmark dataset is needed. Only then, an objective evaluation of current 3D reconstruction and camera pose estimation methods can be achieved. While there are existing datasets containing images and a corresponding 3D point cloud reference (Aanæs et al., 2016; Yan et al., 2023), they are lacking critical aspects that make them unsuitable for the evaluation of image-based 3D reconstruction and camera pose estimation in industrial settings.

In this paper, we introduce MVM-IOD, a novel dataset that can be used in context of several typical computer and machine vision tasks, such as Multi-View Stereo (MVS), Novel View Synthesis (NVS), hand-eye calibration, and camera pose estimation. The dataset contains nine different objects (see Figure 1), which are typical for industrial applications and at the same time challenging for computer and machine vision methods. These challenges include lack of texture, transparent or non-Lambertian surfaces, very fine geometric structures, and different kinds of symmetries. The dataset provides 300 RGB images for each object, captured by an industrial camera mounted to the end effector of a high-precision industrial robot arm.

Further, it provides accurate robot and camera poses for each of the captured images. For the camera poses, the dataset contains two sets: The first one is calculated based on the queried robot poses and the hand-eye pose, which was obtained by hand-eye calibration. The data and the results of the hand-eye calibration are also provided. The second one, which serves as reference, is calculated based on a calibration plate and spatial resection. This also requires the interior orientation of the camera, which was determined by camera calibration and is provided as well. Moreover, the dataset contains an accurate reference for the 3D point cloud of the objects, which has been obtained with a highly-accurate industrial line scanner.

We choose COLMAP as long-term standard method for Structure-from-Motion (SfM) (Schönberger and Frahm, 2016) and MVS (Schönberger et al., 2016) to generate a baseline for evaluating MVM-IOD on the task of pose estimation and 3D reconstruction. The time-critical aspect of most industrial applications motivates us to evaluate recent feed-forward visual geometry reconstruction methods such as Visual Grounded Geometry Transformer (VGGT) (Wang et al., 2025a) and subsequent work such as π^3 (Wang et al., 2025b) as well, as they can solve the camera pose estimation and 3D reconstruction task at inference time within a few seconds. Furthermore, the robustness of these methods is evaluated, which is especially important in industrial applications where errors are costly. To show that the task of NVS can also be evaluated on this dataset, we incorporate 2D Gaussian Splatting (2DGS) (Huang et al., 2024) making a comparison to Gaussian Splatting methods and other NVS algorithms possible, though we focus on its generated 3D reconstruction. We compare the estimated camera poses of VGGT, π^3 and SfM to the reference poses in order to make conclusions about the robustness of these meth-

ods in industrial settings. Furthermore, based on the reference, we evaluate the accuracy, completeness, and Chamfer distance of the estimated point clouds from VGGT, π^3 , COLMAP MVS and 2DGS. To summarize, our contributions are as follows:

1. We introduce MVM-IOD, a novel industrial object dataset, to evaluate typical computer and machine vision tasks in an industrial setting. This addresses the lack of object-centric 3D reconstruction and pose estimation datasets focusing on industrial objects.
2. We evaluate three camera pose estimation methods and four methods for image-based 3D reconstruction on MVM-IOD, to provide a baseline for future research and evaluations.
3. We identify weaknesses of current state-of-the-art feed-forward methods in industrial applications and provide workarounds.

2. Related work

Section 2.1 gives an overview of already existing datasets for image based 3D reconstruction. In Section 2.2, we take a look at different computer vision algorithms that solve tasks such as NVS, 3D reconstruction and camera pose estimation.

2.1 Datasets

The DTU dataset (Aanæs et al., 2016) contains a collection of 80 different scenes with 49 or 64 camera positions per scene. The capture setup consists of two cameras and a projector mounted to the end effector of a robot arm, generating images and structured-light reference scans from roughly the same viewing angles. This results in precise reference camera poses and 3D geometry. Most of the objects have a lot of texture, making it unfavorable for industrial applications, where objects often are textureless, reflective, transparent, or induce multiple of these challenges at the same time. The Dataset of Industrial Metal Objects (DIMO) (De Roovere et al., 2022) contains textureless metal objects and uses different cameras mounted to the end effector of an industrial robot to capture the images. While DIMO contains a total of 42.620 scenes, each scene is only captured from 13 camera poses, making it less suited for NVS, as current NVS methods cannot unleash their full potential with such few images (Peng et al., 2024). Furthermore, the objects used in DIMO all have a very primitive geometry such as rectangular cuboids, limiting the relevance of the dataset for industrial applications. A similar setup is used by Durner et al. (2017) in the Top Hat Rail (THR) dataset, which captures images from 200 camera poses per scene. However, there are only a total of four scenes, and the objects are all of a similar, rectangular shape, not being a hard challenge for image-based 3D reconstruction. The T-Less dataset (Hodan et al., 2017) contains scenes with 504 images each of multiple textureless objects placed on a turntable equipped with marker fields for camera pose estimation. Principally, T-Less is suited for evaluating 3D reconstruction and camera pose estimation methods of industrial objects, however, it lacks objects with complex geometry and reflective or transparent surfaces. Hillemann et al. (2024) use the same capturing hardware as we, and a subset of the objects. The main difference is the sampling strategy of the camera poses: Hillemann et al. (2024) sample the poses on a latitude/longitude grid, while we sample the camera poses equidistantly on a hemisphere, which avoids a high sampling density at the pole. Furthermore, we include additional 12 scenes, and evaluate the 3D reconstructions. Langendörfer et

al. (2025) uses the same camera pose sampling strategy and two of the objects used in MVM-IOD for evaluating the 2D and 3D reconstructions of NeRF (Mildenhall et al., 2021) and Gaussian Splatting (Kerbl et al., 2023) methods. In comparison, MVM-IOD contains 14 additional scenes, is evaluated on a broader set of methods, and is made available publicly. The NeRFBK (Yan et al., 2023) dataset is tailored to the evaluation of NVS algorithms. It contains a collection of 21 indoor and outdoor scenes, though only a fraction of them is relevant in an industrial context. The images of the relevant scenes are captured by smartphone cameras or are rendered synthetically, that do not reflect industrial capture scenarios. The NeRF synthetic dataset (Mildenhall et al., 2021) consists of nine synthetic scenes with 600 images each. While not relevant for industrial applications, it can be beneficial to use synthetic scenes because of their perfect ground truth data.

With MVM-IOD, we aim to fill the outlined lack of benchmark datasets in the industrial domain for image based 3D reconstruction and camera pose estimation. MVM-IOD consists exclusively of industry-relevant objects with complex geometry, making it more applicable to industrial scenarios than DTU and THR. Additionally, in contrast to DIMO, it consists of densely sampled images which supports NVS. Furthermore, all scenes are captured with a realistic industrial setup, and do not contain synthetic data in comparison to NeRFBK.

2.2 Methods

SfM and MVS. SfM describes the problem of finding the interior orientation and camera pose along with a sparse point cloud from a set of uncalibrated images (Agarwal et al., 2011). In contrast to current learning-based methods, traditional, rule-based methods divide this process in three stages: keypoint extraction, keypoint matching, and a bundle adjustment. MVS describes the problem of extracting a dense point cloud from a set of images. Typically, the SfM result is used as an input along with the images (Furukawa et al., 2015; Schönberger et al., 2016). COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016) is a widespread implementation of both SfM and MVS and the long-time standard.

NVS. The task of generating novel views of a scene from a set of images with associated interior orientation and camera pose has seen a lot of attention in recent years. Mildenhall et al. (2021) introduced the concept of NeRFs that represent scenes as neural radiance fields, which allow for the generation of accurate novel views, with the drawback of long training times for each single scene. 3D Gaussian Splatting (3DGS, Kerbl et al. (2023)) uses Gaussian primitives with view-dependent color encoding to reconstruct a scene. This approach allows for faster training times per scene at a comparable quality of the generated novel views compared to NeRFs (Hillemann et al., 2024). Subsequent works, such as Guédon and Lepetit (2024), Chen et al. (2024), and Huang et al. (2024) use the basic 3DGS concept, but improve the 3D reconstruction that can be extracted from the Gaussians. Specifically, Huang et al. (2024) do this by using 2D instead of 3D Gaussian primitives, which resolves multi-view geometry inconsistencies.

Feed-forward visual geometry reconstruction. One of the main disadvantages of the mentioned SfM, MVS, and NVS algorithms is that they need to perform a training for each new scene. With the advent of new neural network structures (Vaswani et al., 2017) as well as abundant training data for visual tasks such as Dai et al. (2017); Yao et al. (2020);

Reizenstein et al. (2021), new feed-forward visual geometry reconstruction methods arise. DUST3R (Wang et al., 2024) uses a Visual Transformer architecture to regress pointmaps and camera poses from two images, essentially solving the previously mentioned traditional MVS task in a feed-forward manner by learning a generalizable, scene-agnostic representation. MAST3R (Leroy et al., 2024) builds upon this by introducing local feature heads, making it robust against large viewpoint changes. Recently, VGGT (Wang et al., 2025a) used a transformer with global and local attention to infer pointmaps, depth maps, point tracks, and camera poses from an arbitrary number of input images. By simultaneously estimating these redundant properties, they manage to outperform methods like DUST3R and MAST3R. VGGT can generate a 3D reconstruction directly from the point maps, as well as by projecting the pixels of the depth maps via the camera poses and the interior orientation. π^3 (Wang et al., 2025b) builds upon the idea of VGGT, but it removes the inductive bias of a fixed reference view and adds a local smoothness. The inference time of these methods is in the range of a few seconds, depending on the number of images. Furthermore, they are reporting a competitive quality compared to the results obtained by SfM and MVS. However, almost none of the testing data that was used up to now for evaluating VGGT and π^3 contains industry-relevant scenes or objects, making an evaluation in such critical applications highly relevant.

3. The MVM-IOD

In this section, we outline how the MVM-IOD is generated. First, detailed descriptions of the captured objects are given. Subsequently, we give details on the capture setup and the used hardware, and finally we describe how we obtain the reference data for the point clouds and the camera poses.

Objects. The dataset contains industry-relevant objects, as seen in Figure 1, with sizes ranging from 7 to 15 cm diameter. The objects cover various properties, such as transparency, reflectance, lack of texture, and complex geometry, which are challenging for vision-based reconstruction algorithms. The focus of MVM-IOD is to represent typical industrial applications. Each object is captured with two different background choices. The first background is an aluminum table with profile rays, which provides a rich texture for keypoint extraction and matching. The second background is a homogeneous green background, to mimic poor texture in the background, which also often occurs in industrial applications. The background choices are referred to as alu and green, respectively.

Hardware and Setup. All images are acquired by an IDS U3-31J0CP-C-HQ Rev 2.2 RGB industrial camera with a Tamron M23FM08 8 mm lens. The captured images have an image resolution of 2840×2840 px. The camera is mounted on the end effector of a Universal Robots UR3e robot arm with a pose repeatability error of ± 0.03 mm according to ISO 9283. The robot was calibrated by using the approach of Ulrich et al. (2024). It is important to note that the camera is mounted so that all images appear upside-down due to space restrictions. A camera calibration as well as an uncertainty-aware hand-eye calibration (Ulrich and Hillemann, 2021, 2024) is performed with a total of 25 calibration images. We define a pose ${}^b\mathbf{H}_a \in \text{SE}(3)$, indicating a 4×4 rigid transformation from coordinate system a to coordinate system b . The relevant coordinate systems in our setup are base (b), world (w), tool (t) and camera (c).

The images of each scene are captured at the same ${}^w\mathbf{H}_c$ camera poses with the object in the center of the scene. With the



Figure 1. Objects contained in the dataset, Figure adapted from Langendörfer et al. (2025)

result of the hand-eye calibration, we are able to calculate the corresponding ${}^t\mathbf{H}_b$ by matrix multiplication, which is passed to the robot controller to move the robot arm. The ${}^w\mathbf{H}_c$ poses are sampled equidistantly on a sphere with radius of 20 cm via the algorithm described by González (2010) up to a maximum zenith angle of 35° , resulting in a total of 300 ${}^w\mathbf{H}_c$ poses per scene. The radius and the maximum zenith angle are physically limited by the robot workspace. We refer to the ${}^w\mathbf{H}_c$ poses obtained in this way as the robot-derived camera poses, which can be used as a metric alternative to the COLMAP SfM camera pose estimation. We report the Root Mean Square Translation (RMST), the Root Mean Square Rotation (RMSR), and the Reprojection Root Mean Square Error (RRMSE) of the hand-eye calibration, as implemented by Hillemann et al. (2024): $\text{RMST} = 0.11$ mm, $\text{RMSR} = 0.030^\circ$, and $\text{RRMSE} = 0.94$ px. The mean standard deviation is $\sigma_{\text{trans}} = 0.07$ mm for translation and $\sigma_{\text{rot}} = 0.018^\circ$ for rotation, which is roughly two times the robot pose repeatability error.

Reference. The reference point cloud of the objects is captured with a Hexagon AS1 line scanner mounted to an AA85 measurement arm with a system accuracy of 0.047 mm according to ISO 10360-8 D. All reconstructed points in the reference point cloud that do not belong to the scanned object have been filtered out manually. Because the reference point cloud is not represented in the robot coordinate system, a registration is necessary.

The reference for the ${}^w\mathbf{H}_c$ poses is obtained by capturing a scene with a calibration plate in the center of the scene, which

is acquired with the same $300^t H_b$ poses by a calibrated camera and performing spatial resectioning with fixed internal camera parameters. The reference robot poses are still influenced by the pose repeatability error of the robot in each dataset. We report a RRMSE of 0.39 px for the reference poses.

4. Experiments

All experiments are performed on a computer with an NVIDIA® A100 with 40 GB of VRAM. We choose to compare modern feed-forward geometry reconstruction methods to assess their use for real-time industry applications.

4.1 Evaluated Methods

COLMAP. We use COLMAP V3.13.0 (Schönberger and Frahm, 2016; Schönberger et al., 2016) to generate the COLMAP SfM and MVS results. COLMAP SfM takes roughly 5 to 10 minutes, depending on the object and the background choice. COLMAP MVS takes about 3 to 4 hours.

VGGT. We use the pretrained VGGT-1B checkpoint with 1.26 B parameters for training. All images are resized by default to a size of 518×518 px. By default, non-square images are zero-padded. We report an inference time of 17 seconds. VGGT can estimate point clouds via directly via the point maps, which we refer to as VGGT point, or by projecting the estimated depth maps with the estimated camera pose and interior orientation, which we refer to as VGGT depth.

π^3 . All square images are resized to 504×504 px, while non-square images get resized to 588×420 px. We use the standard pretrained model with 0.96 B parameters for inference. We report an inference time of 17 seconds.

2DGS. For 2DGS input, we use the COLMAP SfM result to undistort the input images, as well as for the camera poses and the initial point cloud. We use the default training parameters and train for 30000 iterations, which roughly takes 3 hours.

The tested methods use different amounts of VRAM. VGGT uses too many resources to process all 300 images of each dataset at once. Therefore, we subsample each dataset randomly to $n = 100$ images, but keep the same random subsampling for all datasets and experiments. For COLMAP and 2DGS, all images are processed in full resolution.

4.2 Evaluation

First, we show that VGGT and π^3 neither generate accurate camera poses nor accurate 3D reconstruction with the original MVM-IOD images and how to improve the results. Then, we evaluate both the accuracy of the poses and the accuracy and completeness of the generated point cloud of the different methods by comparing the results with the reference. A comparison of the internal camera parameters is not performed, as VGGT and π^3 do not estimate the distortion parameters. However, methods that estimate the distortion parameters can also be assessed with MVM-IOD.

Image preprocessing for VGGT and π^3 . One initial insight is that the images of the dataset are difficult to process with feed-forward geometry transformers such as VGGT and π^3 , and the 3D reconstruction as well as the pose estimation is

visibly off most of the time with the original images, which can be seen, for example, in Figure 2 for object 6 with background alu. This makes a registration and subsequent comparison to the reference point cloud nearly impossible because the reconstructed object is broken into multiple instances with different orientations. The most likely reason for this is that the image contents and the image format are out of domain for these models, which were not trained on similar data. It is shown empirically that by rotating the images by 180° and changing the aspect ratio from 1 : 1 to 4 : 3, VGGT and π^3 are able to drastically improve the pose estimation and 3D reconstruction accuracy of the scenes in most cases (Table 1). These preprocessing steps are motivated by the fact that the camera is mounted upside down to the end-effector, obviously making the scenes appear unusual for the trained models. Also, most of the training images used for VGGT and π^3 are non-square. Therefore, the MVM-IOD images are adapted to be more in-distribution also in this regard. Applying only one of the two preprocessing steps did not achieve satisfying improvements.

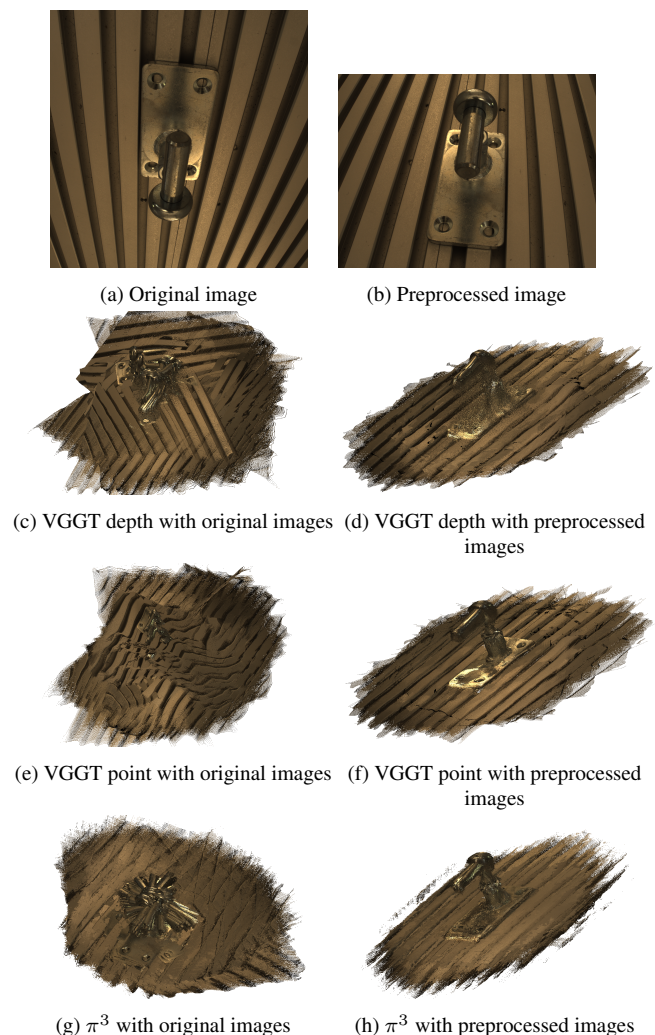


Figure 2. a) and b) show images of object 6 with background alu before (original) and after preprocessing by rotating the image by 180° and changing the aspect ratio to 4 : 3. c), e), and g) show the VGGT point, VGGT depth and π^3 point clouds when using the original images. d), f), and h) show the same point clouds of VGGT and π^3 when using the preprocessed images.

Camera Pose Evaluation. For the evaluation of the camera poses, the EVO tool (Grupp, 2017) is used. Each set of estimated camera poses is compared to the corresponding reference poses obtained as described in Section 3. As COLMAP SfM, VGGT, and π^3 do not output metric camera poses or point clouds, an alignment and scaling to the reference poses is necessary to enable a comparison. This is achieved by applying the Umeyama algorithm (Umeyama, 1991). Then, the Absolute Pose Error $\mathbf{H}_{APE,i}$ is calculated as

$$\mathbf{H}_{APE,i} = \mathbf{H}_{est,i}^{-1} \mathbf{H}_{ref,i} \in SE(3) . \quad (1)$$

Where $\mathbf{H}_{est,i}$ is the estimated camera pose and $\mathbf{H}_{ref,i}$ the reference pose. The $\mathbf{H}_{APE,i}$ between a set of N reference and estimated poses is used to calculate the Mean Translation Error (MTE, Equation (2)) and the Mean Rotation Error (MRE, Equation (3)):

$$MTE = \frac{1}{N} \sum_{i=1}^N \|\text{trans}(\mathbf{H}_{APE,i})\| \quad (2)$$

$$MRE = \frac{1}{N} \sum_{i=1}^N |\text{angle}(\log_{SO(3)}(\text{rot}(\mathbf{H}_{APE,i})))| \quad (3)$$

Here, $\text{trans}(\cdot)$ and $\text{rot}(\cdot)$ are the translation and rotation parts of the $\mathbf{H}_{APE,i}$ and $\text{angle}(\log_{SO(3)}(\cdot))$ is the Rodrigues angle in degrees. Moreover, the RRMSE (Section 3) of the COLMAP SfM and robot-derived camera poses is reported. Of course, the RRMSE of the robot-derived camera poses is constant over all objects in the dataset at 1.05 px, while the mean RRMSE over all COLMAP SfM results is 1.17 px with $\sigma_{RRMSE} = 0.096$ px. RRMSE values for VGGT and π^3 are not calculated, because both methods reconstruct a point for each pixel and hence RRMSE would be zero. While it would be possible to perform a bundle adjustment after inference to refine the results, this approach would not evaluate the performance of the actual methods and would compromise their real-time capability. The resulting MTE and MRE are reported in Table 1 for all methods and all scenes in the dataset. For VGGT and π^3 , we report the results for the original images as well as for the preprocessed images as described in Section 4.2. We observe that the COLMAP SfM and robot-derived MTE are in the same order of magnitude, though for some scenes with green background, COLMAP SfM completely fails likely due to missing texture. The robot-derived MRE is one order of magnitude lower than the COLMAP SfM MRE. In contrast to Wang et al. (2025b,a) our results show that both COLMAP SfM and robot-derived have significantly lower MRE and MTE values than VGGT and π^3 in all cases. For VGGT, preprocessing the images results in significantly lower MRE and MTE values in 14 cases while MRE and MTE stay roughly the same for objects 2, 3, 4, and 7 with the alu background. Roughly the same applies to π^3 , though the MRE and MTE values increase for object 7 with background alu when applying the preprocessing to the images. Generally, it can be observed that VGGT and π^3 seemingly struggle with the alu background, as after preprocessing only scenes with background alu have large MRE and MTE values. At the same time, even without the preprocessed images, somewhat low MRE and MTE values are achieved for objects 1, 5/6 (only π^3) and 8 with green background.

3D Reconstruction. We report the mean of accuracy and the mean of completeness, as well as the overall metric of the point clouds. The accuracy (ACC_i) for each point \mathbf{x}_i in the estimated

point cloud \mathbf{P}_{est} is calculated as seen in Equation (4), while the completeness ($COMP_j$) of each point \mathbf{x}_j in the reference point cloud \mathbf{P}_{ref} is calculated as in Equation (5).

$$ACC_i = \|\mathbf{x}_i - \text{NN}(\mathbf{x}_i, \mathbf{P}_{ref})\|_2 \quad (4)$$

$$COMP_j = \|\mathbf{x}_j - \text{NN}(\mathbf{x}_j, \mathbf{P}_{est})\|_2 \quad (5)$$

Where $\text{NN}(\mathbf{x}, \mathbf{P})$ is the nearest neighbor of a point \mathbf{x} in a point cloud \mathbf{P} . Finally, the overall metric is the mean of accuracy and completeness, also referred to as the Chamfer distance (Barrow et al., 1977). All point cloud operations are done in CloudCompare (CloudCompare Development Team, 2025).

The reference and estimated point clouds are registered with an Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) following an automatic coarse registration. In cases of poor initial reconstruction, a manual coarse alignment is occasionally applied to ensure a robust registration. We verify the registration success visually and only report results in case of a registration success. Complete registration failures are usually related to a poor estimated point cloud. Furthermore, the estimated point clouds are cropped automatically, so that no points from the background falsify the metrics. We provide masks for cropping regions of the estimated point cloud, that are not reconstructed in the reference point cloud, as otherwise the accuracy would be falsified. Both VGGT and π^3 output confidences for the estimated points. For π^3 , we apply the default confidence threshold for cropping points with low confidence. For VGGT, the default confidence cropping threshold of 5 removed too many points, so we empirically chose a threshold of 2.

While the number of points in the VGGT and π^3 point clouds are determined by the total number of images and their size, the COLMAP MVS and 2DGS reconstruction typically consist of considerably fewer points and are not necessarily sampled evenly. Therefore, a spatial subsampling similar to the one used in Aanæs et al. (2016), but with a different subsampling distance threshold, is applied to \mathbf{P}_{est} and \mathbf{P}_{ref} . The subsampling distance threshold is set to the scanner accuracy of 0.047 mm (see Section 3). This assures, that regions with high point density do not bias the results. Table 2 shows the mean accuracy, mean completeness, and overall metric for all objects. As in Aanæs et al. (2016), we disregard observations beyond a threshold of 20 mm, which we consider outliers. As we have shown in the camera pose evaluation of Section 4.2, the preprocessing steps of Section 4.2 lead to better camera poses. Figure 2 exemplarily shows that this applies to the point clouds as well. Therefore, for VGGT and π^3 we only report the results for the preprocessed images, as in most cases even the registration is failing with the original images. For both VGGT and π^3 , the object and background combinations that resulted in bad camera poses also tend to lead to either failed registrations or high accuracy and completeness metrics (compare Tables 1 and 2), which makes sense as they are estimated simultaneously. VGGT point and COLMAP MVS achieve mostly submillimeter accuracy with both background choices, given that the registration worked. π^3 , 2DGS, and VGGT depth perform slightly worse. The worst completeness results are reported for 2DGS, which is mainly caused by erroneous depth estimation on textureless metallic surfaces. This issue can be seen in Figure 3. The bad completeness in combination with the long training of 2DGS appears to be not worth the effort, except if the generated novel views are of interest. π^3 along with COLMAP MVS generally tend to achieve the best completeness values. The main reason for VGGT point falling behind in this regard, is that the

Table 1. Mean Translation Error (MTE) and Mean Rotation Error (MRE) for all dataset objects with both background choices (green/alu). For VGGT and π^3 , we use the preprocessed images as described in Section 4.2, while we use the original images for COLMAP SfM and robot-derived. Still, we report the metrics of VGGT and π^3 using the original images in parentheses and mark them in red if they are worse and green otherwise. Lowest metric for each object and background combination is print bold.

Method	Background	Object 1	Object 2	Object 3	Object 4	Object 5	Object 6	Object 7	Object 8	Object 9	Mean
<i>Mean Translation Error (MTE) [mm] ↓</i>											
Robot-derived	green/alu	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
π^3	green	5.34 (4.65)	3.88 (40.88)	5.75 (22.47)	4.94 (62.98)	4.56 (5.62)	3.49 (5.12)	14.10 (65.42)	1.54 (2.46)	3.04 (10.39)	5.18 (24.44)
VGGT	green	2.35 (6.90)	4.61 (70.96)	7.82 (66.20)	3.07 (57.68)	2.36 (37.75)	2.10 (44.01)	2.47 (76.31)	2.65 (4.47)	2.06 (11.38)	3.28 (41.74)
COLMAP SfM	green	-	-	-	-	-	0.09	-	0.15	0.11	0.12
π^3	alu	6.42 (60.62)	79.90 (63.86)	79.70 (58.25)	78.17 (68.81)	4.22 (52.66)	4.39 (74.38)	72.11 (39.34)	2.33 (33.56)	2.86 (69.15)	36.70 (57.84)
VGGT	alu	4.73 (78.19)	80.21 (79.95)	79.00 (76.06)	79.11 (77.38)	3.78 (64.74)	4.25 (78.05)	78.57 (79.55)	3.20 (50.29)	5.40 (67.35)	37.58 (72.40)
COLMAP SfM	alu	0.27	0.07	0.17	0.22	0.24	0.14	0.26	0.08	0.24	0.19
<i>Mean Rotation Error (MRE) [°] ↓</i>											
Robot-derived	green/alu	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
π^3	green	2.98 (3.27)	2.39 (29.27)	1.15 (18.94)	2.17 (43.06)	4.22 (3.03)	1.16 (2.29)	4.94 (61.31)	1.32 (1.16)	1.21 (6.03)	2.39 (18.71)
VGGT	green	2.14 (3.44)	3.64 (170.10)	6.35 (164.35)	1.12 (40.49)	1.93 (20.75)	1.71 (15.12)	2.33 (82.21)	1.78 (4.01)	1.18 (4.47)	2.46 (56.10)
COLMAP SfM	green	-	-	-	-	-	0.72	-	0.73	0.72	0.72
π^3	alu	6.68 (66.28)	150.87 (159.00)	98.44 (54.01)	85.36 (92.43)	1.82 (162.58)	2.40 (138.00)	79.44 (24.28)	0.90 (23.67)	2.29 (76.91)	47.58 (88.57)
VGGT	alu	2.06 (99.49)	122.42 (121.30)	124.30 (95.12)	136.69 (97.48)	2.04 (55.48)	1.36 (101.39)	135.00 (113.64)	1.34 (25.54)	1.91 (143.70)	58.57 (94.79)
COLMAP SfM	alu	0.70	0.72	0.71	0.69	0.69	0.71	0.71	0.72	0.70	0.71

Table 2. Mean accuracy, mean completeness and overall in mm for the estimated point clouds of VGGT, π^3 , COLMAP MVS and 2DGS when compared with their respective reference point cloud. For VGGT and π^3 we use the point clouds generated with the preprocessed images as input. For VGGT, we report the results for the pointcloud generated via the pointmap branch (point) as well as the pointcloud generated via the depthmap branch (depth). Lowest metric for each object and background combination is print bold.

Method	Background	Object 1	Object 2	Object 3	Object 4	Object 5	Object 6	Object 7	Object 8	Object 9	Mean
<i>Accuracy [mm] ↓</i>											
VGGT depth	green	3.74	1.25	2.25	1.21	1.29	0.70	1.00	0.70	0.67	1.45
VGGT point	green	3.43	0.63	0.70	1.01	0.81	0.40	0.65	0.43	0.29	0.93
π^3	green	-	1.45	2.39	1.79	1.45	1.18	-	0.57	0.66	1.36
COLMAP MVS	green	-	-	-	-	-	0.29	-	0.40	0.29	0.33
2DGS	green	-	-	-	-	-	1.41	-	3.86	4.01	3.09
VGGT depth	alu	7.90	-	-	1.77	1.34	0.84	1.44	0.87	1.27	2.20
VGGT point	alu	6.71	-	-	0.57	0.99	0.53	0.60	0.58	0.50	1.50
π^3	alu	-	-	-	-	1.43	1.14	-	0.68	0.86	1.02
COLMAP MVS	alu	9.47	0.70	1.88	0.63	0.80	0.27	0.66	0.37	0.25	1.67
2DGS	alu	6.80	1.16	2.55	1.86	0.78	0.97	1.98	0.55	0.53	1.91
<i>Completeness [mm] ↓</i>											
VGGT depth	green	3.52	0.48	1.54	0.69	0.20	0.11	0.13	0.12	0.21	0.89
VGGT point	green	5.29	3.63	2.78	1.61	0.30	0.21	0.33	0.15	0.70	1.78
π^3	green	-	0.45	1.11	0.30	0.56	0.11	-	0.10	0.11	0.39
COLMAP MVS	green	-	-	-	-	-	0.11	-	0.15	0.52	0.26
2DGS	green	-	-	-	-	-	1.43	-	1.02	1.54	1.33
VGGT depth	alu	7.48	-	-	0.27	0.16	0.12	0.12	0.12	0.14	1.20
VGGT point	alu	7.02	-	-	0.18	0.18	0.32	0.38	0.17	0.27	1.22
π^3	alu	-	-	-	-	0.20	0.12	-	0.09	0.18	0.15
COLMAP MVS	alu	3.73	0.96	2.66	0.43	0.89	0.11	0.25	0.14	0.16	1.04
2DGS	alu	6.03	1.31	3.02	1.57	0.66	1.73	0.85	0.50	1.00	1.85
<i>Overall (Chamfer distance) [mm] ↓</i>											
VGGT depth	green	3.63	0.87	1.89	0.95	0.75	0.41	0.56	0.41	0.44	1.17
VGGT point	green	4.36	2.13	1.74	1.31	0.55	0.31	0.49	0.29	0.49	1.35
π^3	green	-	0.95	1.75	1.04	1.01	0.65	-	0.34	0.39	0.87
COLMAP MVS	green	-	-	-	-	-	0.20	-	0.28	0.41	0.30
2DGS	green	-	-	-	-	-	1.42	-	2.44	2.77	2.21
VGGT depth	alu	7.69	-	-	1.02	0.75	0.48	0.78	0.50	0.71	1.70
VGGT point	alu	6.86	-	-	0.37	0.59	0.42	0.49	0.38	0.39	1.36
π^3	alu	-	-	-	-	0.82	0.63	-	0.39	0.52	0.59
COLMAP MVS	alu	6.60	0.83	2.27	0.53	0.84	0.19	0.46	0.25	0.20	1.35
2DGS	alu	6.42	1.24	2.79	1.72	0.72	1.35	1.42	0.53	0.76	1.88

confidences of the point branch appear to be lower than those of the depth branch, resulting in more points cropped. Regarding the overall metric, VGGT depth/point, π^3 , and COLMAP MVS are not significantly different in the big picture, though exceptions exist (e.g., for object 6 alu, COLMAP MVS is significantly better). Taking the processing times of VGGT and π^3 into account, the results are remarkable.

4.3 Ablation Study: NeRF Synthetic

We verify our findings that VGGT as well as π^3 are sensitive to the rotation of input images. We use a subset of $n = 100$

images of the NeRF Synthetic dataset (Mildenhall et al., 2021) and report the MRE (Equation (3)) and MTE (Equation (2)) for the original images and for the images rotated by 180° in Table 3. The MTE and MRE values cannot be compared directly to the MTE and MRE values of MVM-IOD or to the other scenes within the NeRF synthetic dataset, as each scene has a different extent. However, if VGGT and π^3 were invariant regarding the rotation of the input images, one would expect similar results for the MRE and MTE, independent of the rotation. While this is the case for the scenes materials and ficus, where the MRE and MTE values do not change significantly, the MRE and MTE

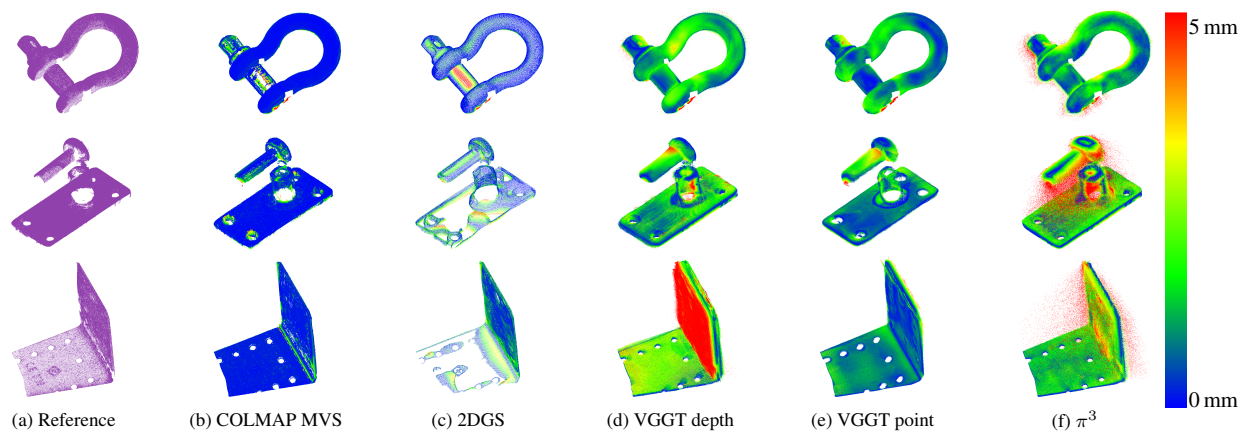


Figure 3. Pointwise accuracy of the registered point clouds. A saturation threshold of 5 mm is used for visualization. From top to bottom we show objects 8,6 and 9, all with alu background.

Table 3. Mean Rotation Error (MRE) and Mean Translation Error (MTE) for VGGT and π^3 on the NeRF Synthetic dataset. All experiments are done with a subsampling of $n = 100$ images. The results for the original images, objects naturally facing up, are shown in black. In parentheses, we report the results for the rotated images, objects facing down and mark them in green if the rotation improved the result and in red if the opposite is the case. Lowest metric for each object is print bold.

Method	chair	drums	ficus	hotdog	lego	materials	mic	ship	Mean
<i>Mean Rotation Error (MRE) [°] ↓</i>									
VGGT	0.67 (0.63)	0.23 (1.01)	3.01 (2.94)	0.41 (0.68)	0.40 (1.05)	1.67 (1.87)	0.86 (1.30)	0.42 (2.45)	0.96 (1.49)
π^3	0.57 (1.22)	0.44 (1.37)	2.41 (3.99)	0.60 (1.76)	0.47 (1.23)	0.72 (1.02)	0.59 (1.16)	0.50 (6.83)	0.79 (2.32)
<i>Mean Translation Error (MTE) [mm] ↓</i>									
VGGT	33.17 (44.40)	13.42 (59.97)	199.83 (188.28)	21.34 (41.64)	16.73 (62.29)	103.04 (101.75)	44.35 (64.24)	25.11 (210.16)	56.62 (96.59)
π^3	48.78 (78.61)	26.55 (82.51)	122.83 (164.30)	76.39 (97.93)	25.10 (87.25)	71.27 (83.51)	41.95 (81.91)	42.90 (152.63)	56.97 (103.58)

do worsen significantly for the other scenes. In most cases, the π^3 camera poses are significantly more affected by the rotation than the VGGT camera poses.

5. Conclusion

We present MVM-IOD, a dataset for the evaluation of camera pose estimation and 3D reconstruction on industrial objects. We show, that MVM-IOD is highly relevant, as novel feed-forward geometry transformers struggle with these type of input images more than with non-industrial data. While the tested 3D reconstruction methods do not generate accurate 3D reconstructions for some objects in the MVM-IOD, we see this as a challenge for future methods, as our motivation is to provide an extremely challenging dataset. We also empirically show that VGGT and π^3 have some sort of scene gravity awareness, as the results worsen when the scene contents seem upside down. This means, that in industrial scenarios, the operator should be aware that current feed-forward geometry transformers can fail, when the input images are lying too far out of distribution. This is especially relevant, as the feed-forward methods will always produce an output unlike other tested methods like COLMAP, which simply fails in adverse conditions. The images can be shifted closer to the distribution by applying pre-processing steps to the images. Or, these findings can be used to train or finetune new feed-forward geometry transformers with upside down images and images of varying aspect ratios.

References

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., Dahl, A. B., 2016. Large-Scale Data for Multiple-View Stereopsis. *IJCV*, 1–16.

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., Szeliski, R., 2011. Building rome in a day. *Communications of the ACM*, 54(10), 105–112.

Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., Wolf, H. C., 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report.

Besl, P. J., McKay, N. D., 1992. Method for registration of 3-d shapes. *Sensor fusion IV: control paradigms and data structures*, 1611, Spie, 586–606.

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*.

CloudCompare Development Team, 2025. Cloudcompare software, version 2.13.2. Open Source Project. <https://www.cloudcompare.org/> (26 Feb 2025).

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CVPR*, 5828–5839.

De Roovere, P., Moonen, S., Michiels, N., Wyffels, F., 2022. Dataset of industrial metal objects. *arXiv preprint arXiv:2208.04052*.

Durner, M., Kriegel, S., Riedel, S., Brucker, M., Márton, Z.-C., Bálint-Benczédi, F., Triebel, R., 2017. Experience-based optimization of robotic perception. *2017 18th ICAR*, IEEE, 32–39.

Furukawa, Y., Hernández, C. et al., 2015. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2), 1–148.

- González, Á., 2010. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Mathematical geosciences*, 42(1), 49–64.
- Grupp, M., 2017. evo: Python package for the evaluation of odometry and slam. <https://github.com/MichaelGrupp/evo>.
- Guédon, A., Lepetit, V., 2024. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 5354–5363.
- Hillemann, M., Langendörfer, R., Heiken, M., Mehlretter, M., Schenk, A., Weinmann, M., Hinz, S., Heipke, C., Ulrich, M., 2024. Novel View Synthesis with Neural Radiance Fields for Industrial Robot Applications. *ISPRS Archives*, 48, 137–144.
- Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X., 2017. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. *2017 IEEE WACV*, IEEE, 880–888.
- Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S., 2024. 2d gaussian splatting for geometrically accurate radiance fields. *ACM SIGGRAPH 2024 conference papers*, 1–11.
- Karami, A., Menna, F., Remondino, F., 2021. Investigating 3D reconstruction of non-collaborative surfaces through photogrammetry and photometric stereo. *ISPRS Archives*, 43, 519–526.
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.
- Langendörfer, R., Hillemann, M., Ulrich, M., 2025. Industry-relevant benchmark dataset for the evaluation of 3D reconstruction methods and evaluation of different NeRF methods. *45. Jahrestagung der DGPF in Muttens/Basel 2025. Hrsg.: TP Kersten*, 99.
- Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with mast3r. *ECCV*, Springer, 71–91.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.
- Peng, R., Xu, W., Tang, L., Liao, L., Jiao, J., Wang, R., 2024. Structure consistent gaussian splatting with matching prior for few-shot novel view synthesis. *Thirty-Eighth Conference on NeurIPS*.
- Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D., 2021. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. *ICCV*, 10901–10911.
- Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *CVPR*.
- Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *ECCV*, 501–518.
- Steger, C., Ulrich, M., Wiedemann, C., 2018. *Machine vision algorithms and applications*. John Wiley & Sons.
- Ulrich, M., Hillemann, M., 2021. Generic hand–eye calibration of uncertain robots. *IEEE ICRA*, 11060–11066.
- Ulrich, M., Hillemann, M., 2024. Uncertainty-Aware Hand–Eye Calibration. *IEEE Transactions on Robotics*, 40, 573–591.
- Ulrich, M., Steger, C., Butsch, F., Liebe, M., 2024. Vision-Guided Robot Calibration using Photogrammetric Methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, Volume 218(Part A), 645–662.
- Umeyama, S., 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04), 376–380.
- 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., 2025a. Vggt: Visual geometry grounded transformer. *CVPR*, 5294–5306.
- Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. *CVPR*, 20697–20709.
- Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2025b. π^3 : Scalable Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*.
- Yan, Z., Mazzacca, G., Rigon, S., Farella, E. M., Trybala, P., Remondino, F. et al., 2023. NeRFBK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *ISPRS Archives*, 48(1), 219–226.
- Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *CVPR*, 1790–1799.
- Yoon, Y., DeSouza, G., Kak, A., 2003. Real-time tracking and pose estimation for industrial objects using geometric features. *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, 3, 3473–3478 vol.3.