Investigating the Potential of SfM, MVS, and Monocular Depth Estimation for Water Surface Reconstruction

Anatol Günthner[1], Markus Brezovsky[2], Frederik Schulte[3], Lukas Winiwarter[3], Gottfried Mandlburger[2], Boris Jutzi[1]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany - (anatol.guenthner, boris.jutzi)@kit.edu
[2] Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria - (markus.brezovsky, gottfried.mandlburger)@geo.tuwien.ac.at
[3] Unit of Geometry and Surveying, Faculty of Engineering Sciences, University of Innsbruck, Innsbruck, Austria - (frederik.schulte, lukas.winiwarter)@uibk.ac.at

Keywords: UAV Photogrammetry, Water Surface Reconstruction, Monocular Depth Estimation, Refractive Neural Radiance Fields, Depth Scaling, Bathymetric Mapping.

Abstract

Reconstructing the water surface in refractive domains such as rivers and lakes is challenging, since light bending at the air-water interface alters the apparent geometry and breaks the straight-ray assumption of conventional image-based 3D reconstruction. Accurate water surface models are therefore a key prerequisite for many refraction-aware applications. This contribution investigates the potential of three passive image-based methods, Structure from Motion (SfM), Multi-View Stereo (MVS), and Monocular Depth Estimation (MDE), to derive a geometrically consistent water surface model from UAV imagery of the Pielach River study site in Austria. The dataset represents a demanding scenario with clear, fast-flowing water and low texture, which causes strong refraction and poor feature stability. Quantitative comparisons against LiDAR-derived reference surfaces show that SfM yields sparse and inconsistent points, MVS reconstructs the riverbed instead of the water surface, and MDE exhibits scale and offset inconsistencies despite explicit calibration using SfM reprojections. Completeness remains below 45 % for all methods with mean vertical deviations in the decimetre-to-metre range. The results indicate that current image-based approaches are insufficient for reliable water-surface reconstruction in such settings, reinforcing the need for an explicitly derived surface model as input to refraction-aware modeling, for example in bathymetric reconstruction and future refractive neural rendering methods, rather than relying on implicitly learned water surfaces.

1. Introduction

Reconstructing three-dimensional geometry in refractive domains such as rivers and lakes is challenging, since light bending at the air-water interface alters the perceived geometry and violates the straight-ray assumption of standard image-based 3D reconstruction. Accurate knowledge of the water surface is therefore a key prerequisite in such settings. Without it, reconstruction algorithms must simultaneously infer both the refractive interface and the submerged geometry, which increases problem complexity and can lead to ambiguous solutions. Approaches that try to recover the water surface implicitly from images alone are typically computationally demanding and often inaccurate, whereas providing an explicit surface prior can simplify reconstruction and improve geometric fidelity. This holds for both classical multi-view pipelines and modern neural scene representations such as Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020), which likewise benefit from an explicit water surface model under refractive conditions.

To address this prerequisite, in this contribution the reconstruction of the water surface from UAV-based imagery captured at the Pielach River in Austria (Mandlburger et al., 2025) is investigated. The dataset provides high-resolution (45 MP) images of a real-world fluvial environment with clear water conditions, enabling the evaluation of image-based reconstruction methods by laser scanning data under challenging refractive conditions. In contrast to structured urban scenes, the river environment poses additional challenges due to its natural, low-texture surroundings, the absence of man-made geometric features, and the exceptionally clear water. These conditions reduce the availability of stable keypoints for feature matching and introduce ambiguities for both multi-view and monocular reconstruction approaches, making the dataset particularly demanding and representative of real-world river environments.

Three principal approaches are compared to derive a water surface model: Structure from Motion (SfM), Multi-View Stereo (MVS), and Monocular Depth Estimation (MDE). SfM provides only sparse geometry and is highly sensitive to water motion, while MVS produces denser point clouds but still fails in deeper or faster-flowing regions. In such clear-water conditions, MVS reconstructions may appear complete at first glance, yet they often represent the riverbed rather than the water surface. MDE, in contrast, delivers per-pixel depth estimates that can capture continuous surfaces, though these depths require metric scaling and alignment to reference data.

The primary objective of this contribution is to evaluate the suitability of these three methods for reconstructing the water surface of a clear, fast-flowing river from UAV imagery, using a LiDAR-derived water surface model (WSM) as geometric reference. The resulting assessment is relevant wherever an explicit WSM is required, for example for refraction-aware bathymetric reconstruction and, in future work, as a prior for refractive neural rendering methods.

2. Related Work

In this contribution, only passive optical methods are considered for reconstructing the geometry of water surfaces. Un-

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-2-2026-585-2026

like active sensing techniques such as laser scanning or radar, which directly measure surface geometry, passive approaches rely solely on image-based information and the inference of three-dimensional structure from visual cues. The following section provides an overview of three representative categories within this domain: SfM, MVS, and MDE.

Structure from Motion techniques estimate camera poses and sparse 3D scene geometry by triangulating feature correspondences across multiple overlapping images (Ullman, 1997). Modern implementations have significantly improved robustness, scalability, and accuracy (Schönberger and Frahm, 2016). SfM assumes static, Lambertian surfaces with consistent appearance across views. In UAV-based river imaging, however, consecutive frames are inherently multi-temporal because the platform moves during acquisition, while the water surface evolves continuously. These temporal variations, combined with specular reflections and refractive effects, violate the brightness and appearance constancy required for stable feature matching.

Building upon SfM, **Multi-View Stereo** methods estimate dense depth by enforcing photometric consistency across multiple calibrated images. Probabilistic formulations allow for view-dependent depth optimization and accurate reconstruction in textured, static regions (Schönberger et al., 2016). Similar to SfM, MVS assumes a rigid scene and consistent radiometry across time. Multi-temporal image acquisition together with reflective and refractive surface properties limits its applicability to flowing water, as both temporal and photometric inconsistencies reduce depth estimation reliability.

In contrast to these multi-view methods, **Monocular Depth Estimation** predicts a per-pixel depth map directly from a single RGB image using convolutional or transformer-based models (Eigen et al., 2014; Liu et al., 2015; Laina et al., 2016; Godard et al., 2017). Since MDE does not rely on cross-view consistency, each UAV frame can be processed independently, which is advantageous in dynamic fluvial environments where the water surface changes between acquisitions. Depth-Anything-V2 (Yang et al., 2024) provides dense depth predictions normalized to a relative value range (e.g., 0 to 1), requiring metric scaling for quantitative use. A related feed-forward approach, MapAnything (Keetha et al., 2025), performs metric 3D reconstruction for multiple tasks including SfM, MVS, and MDE; to the best of our knowledge, it has not yet been applied to refractive or clear-water environments.

Further Approaches to Reconstruct Water Surfaces. Some methods estimate dynamic water surfaces from stereo imagery by analyzing refractive distortions of a known pattern placed beneath the water or by explicitly seeding the surface to introduce measurable texture in field conditions (Chandler et al., 2008). However, these approaches are limited to controlled laboratory environments and rely on predefined calibration targets on or below the surface, which makes them unsuitable for in-situ measurements in natural waters without reference patterns (Morris, 2004). Learning-based methods have also been proposed to correct refractive distortions or infer surface geometry from synthetic or laboratory data, which typically requires a known background pattern and stable imaging conditions (Thapa et al., 2020). Other studies attempted to jointly reconstruct both the dynamic water surface and the underwater scene using multi-view optimization frameworks, demonstrating the feasibility of coupled refractive geometry estimation under controlled setups (Qian et al., 2018). More recent neural rendering approaches extended this concept using differentiable refractive radiance fields to recover transparent surfaces and the underlying scene geometry, yet these methods have so far only been validated in laboratory environments with static cameras and shallow water volumes (Zhan et al., 2023). A first real-world application of refractive NeRFs to UAV-based imagery demonstrated that the network can implicitly learn the water surface when provided with binary water masks defining refractive regions, although this process remains computationally demanding and would likely benefit from an explicitly known surface geometry (Günthner et al., 2025). Single-camera approaches have further shown that refractive cues alone can be exploited to recover the water surface and the scene below in relatively calm, clear-water settings such as fountains, provided that illumination and optical conditions are favorable (Xiong and Heidrich, 2021). In inland waters, image-based monitoring has mainly focused on observable surface properties such as turbidity or flow velocity derived from RGB or multispectral imagery. These approaches depend on visible optical contrast within the water column, which is typically provided by suspended matter or surface tracers, and therefore become unreliable in clear, fast-flowing rivers where specular reflections and refraction dominate the imagery (Manfreda et al., 2024).

Photogrammetric Bathymetry and Refractive Modelling. Previous studies have shown that refraction at the air-water interface causes systematic depth biases and instability in standard SfM-MVS pipelines under clear-water conditions, confirming earlier findings from classical multimedia photogrammetry that already analysed light bending and bundle adjustment in the presence of refractive interfaces (Höhle, 1971; Kotowski, 1988; Mandlburger, 2019; Maas et al., 2025; Mulsow et al., 2024). Dense image matching can recover the riverbed in favourable situations, but accuracy degrades rapidly in the presence of waves, specularities, or low texture. Simulation-based analyses further demonstrate strong errors when planar surface assumptions are used and highlight the limited determinability of refractive bundle adjustment. Additional work has proposed image- or object-space strategies to suppress wave-induced artefacts, yet these approaches primarily stabilise underwater geometry rather than recovering the water surface itself. Overall, the literature indicates that explicit surface modelling or multi-temporal compensation is required for reliable through-water photogrammetry. Beyond these through-water photogrammetric analyses, several refractive SfM frameworks explicitly incorporate planar or otherwise parametrized air-water interfaces into multi-view geometry, demonstrating that accurate reconstruction is feasible when the surface can be described by a simple model and calibration data are available (Jordt-Sedlazeck and Koch, 2013; Elnashef and Filin, 2022).

3. Methodology

The following section describes the workflow for reconstructing the water surface from multi-view and monocular imagery. It comprises sparse geometry estimation using SfM, dense reconstruction via MVS, and depth prediction through MDE, followed by metric calibration and quantitative evaluation.

3.1 Structure from Motion

SfM techniques estimate camera poses and sparse three-dimensional scene geometry by identifying and triangulating corresponding features across multiple overlapping images. The resulting sparse point cloud represents the main structural elements of the observed area and provides the geometric basis for subsequent processing steps. Georeferencing is achieved by integrating the GNSS information embedded in the UAV image metadata, which enables the reconstruction to be scaled and aligned within a consistent metric coordinate framework.

3.2 Multi-View Stereo

The dense reconstruction is subsequently obtained through MVS, which estimates per-pixel depth information for each calibrated image by identifying photometrically consistent correspondences across multiple overlapping views. The individual depth maps are then fused into a dense surface representation, resulting in a high-resolution point cloud that captures fine-scale geometric detail. The MVS reconstruction inherits the metric scale and orientation from the preceding SfM stage.

3.3 Offset and Scale Calibration for Monocular Depth Estimation

As third reconstruction method, the monocular depth estimator Depth-Anything-V2 (Yang et al., 2024) is used to predict dense monocular depth maps for each image in the dataset. To obtain metric monocular depth maps, the predictions from Depth-Anything-V2 are aligned with a globally georeferenced point cloud. The reference, e.g. the sparse SfM point cloud is reprojected into each image plane to create a sparse depth map at specific pixel locations. Since Depth-Anything-V2 predicts inverse depths, the reprojected SfM depths are likewise inverted for consistency. This enables a direct comparison of inverse depths at identical image coordinates. The pixel positions obtained through reprojection of the reference define the sampling mask. Inverse depths from both the monocular prediction and the reference are extracted at these positions to estimate a global scale and offset. The scaling methodology for each individual image follows the depth-regularized 3D Gaussian Splatting approach (Kerbl et al., 2023).

As a first step, the inverse depths of the reference point cloud $\mathbf{D}_{\text{ref}}^{-1}$ are median-centered:

$$\mathbf{s}_{\text{ref}} = \mathbf{D}_{\text{ref}}^{-1} - \text{median}(\mathbf{D}_{\text{ref}}^{-1}) \tag{1}$$

The monocular inverse depths $\mathbf{D}_{\text{mono}}^{-1}$ are centered analogously:

$$\mathbf{s}_{\text{mono}} = \mathbf{D}_{\text{mono}}^{-1} - \text{median}(\mathbf{D}_{\text{mono}}^{-1}) \tag{2}$$

A scale ratio is computed for each depth pair. The global scale factor is then defined as the median of all ratios:

$$\text{scale} = \text{median}\left(\frac{\mathbf{s}_{\text{ref}}}{\mathbf{s}_{\text{mono}}}\right) \tag{3}$$

The offset is derived by aligning the medians of both distributions after applying the scale factor to the monocular inverse depths:

$$\text{offset} = \text{median}(\mathbf{D}_{\text{ref}}^{-1}) - \text{median}(\mathbf{D}_{\text{mono}}^{-1}) \cdot \text{scale} \tag{4}$$

The estimated global scale and offset are then applied to the entire inverse monocular depth map:

$$\mathbf{D}_{\text{mono,scaled}}^{-1} = \text{scale} \cdot \mathbf{D}_{\text{mono}}^{-1} + \text{offset} \tag{5}$$

Finally, the calibrated inverse depths are converted to absolute depth values:

$$\mathbf{D}_{\text{mono,scaled}} = \frac{1}{\mathbf{D}_{\text{mono,scaled}}^{-1}} \tag{6}$$

The availability of a sufficient range of depth differences in the reference data is crucial for robust estimation of offset and scale.

3.4 Evaluation

The reconstructed water surfaces are evaluated against a LiDAR-derived water surface model (WSM) that serves as a geometric reference. The evaluation follows the general principles of established MVS benchmarks (e.g. Seitz et al., 2006; Hermann et al., 2024), using distance-based metrics to quantify geometric accuracy and spatial consistency. In contrast to these bidirectional completeness-accuracy protocols, the present analysis computes one-way cloud-to-mesh (C2M) distances from the reconstructed surface toward the LiDAR-derived WSM. The WSM has a spatial resolution of 0.5×0.5 m and is interpolated onto a 5 cm grid, which is subsequently meshed (max. edge length 10 cm) for distance computation. The mean and standard deviation of the resulting distances describe the average deviation and its spatial variability, respectively. Following Hermann et al. (2024), points of the reconstructed surface located within a distance of 20 cm from the WSM are considered geometrically consistent and indicate completeness.

In addition to the original point clouds, mesh-based variants of the SfM and MVS reconstructions are generated to analyse the effect of surface interpolation. Owing to the sparse and irregular SfM point distribution, small maximum triangle edge lengths (1 m, 5 m, 10 m, 20 m) still produce holes in the water-covered area; only a value of 50 m yields a topologically closed surface. This maximum edge length of 50 m is therefore adopted for meshing both the SfM and the denser MVS point cloud to obtain meshes of comparable topology. The resulting meshes are smoothed using Laplacian smoothing (20 iterations, smoothing factor 0.2) and then uniformly resampled with an average spacing of 5 cm to obtain dense mesh-derived point clouds for the C2M evaluation.

4. Study Area and Data Processing

Imagery and LiDAR data used in this contribution originate from an openly available benchmark dataset (Mandlburger et al., 2025) acquired along the Pielach River in Lower Austria. It includes both nadir and oblique UAV imagery as well as airborne laser bathymetry from October 2024. The measurements were conducted shortly after a major flood, providing exceptionally clear-water conditions that enable seamless reconstruction of the shallow fluvial area with optical methods. In addition to the water surface model, the dataset also includes a LiDAR-derived digital terrain model (DTM) of the riverbed, generated from the same airborne bathymetric laser data, and both products are provided as gridded elevation models (GeoTIFF) in a common ETRS89/UTM 33N reference frame. Accuracy metrics of the topo-bathymetric LiDAR campaign and the derived digital water surface model are reported in Mandlburger et al. (2025), including local comparisons between image-based water surface tie points and the LiDAR-derived WSM with millimetre-level mean offsets and a standard deviation on the order of 5 cm. In this study, the WSM was additionally validated by intersecting it with the LiDAR-derived DTM, deriving the shoreline, and visually confirming its consistency with the UAV orthophoto.

The UAV image block is processed to derive an SfM-based camera orientation followed by dense MVS reconstruction using the open-source software COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016). Monocular depth maps are generated with Depth-Anything-V2 (Yang et al.,

2024). Metric scaling of the monocular outputs follows the depth-regularized 3D Gaussian Splatting procedure (Kerbl et al., 2023), implemented through a modified version of the public `make_depth_scale.py` script[1]. The processed images were georeferenced using black-and-white checkerboard ground control targets measured in the same terrestrial reference network as the LiDAR trajectories, ensuring consistent alignment of image-based reconstructions, the WSM, and the DTM within the ETRS89/UTM 33N system.

The UAV images were captured from approximately 80 m altitude using a DJI M350 RTK platform equipped with a 45 MP RGB camera, resulting in a ground sample distance of approximately 1 cm. For the monocular depth evaluation, 23 nadir images are selected that together cover the study reach. Binary masks are applied to restrict all analyses to the water-covered region. Table 1 lists the approximate number of reconstructed 3D points for each dataset variant within the investigated area, both before and after voxel-based downsampling with a 5 cm grid. For reference, the interpolated water surface model (WSM) used for validation contains about 600 000 points within the water-covered area.

Reconstruction method	Original points	Downsampled (5 cm)
SfM	470 000	430 000
MVS	32 000 000	5 300 000
MDE (1 image)	45 500 000	2 500 000
MDE (23 nadir images)	1 046 500 000	22 100 000
WSM (5 cm water mask)	–	600 000

Table 1. Approximate number of reconstructed 3D points within the investigated area for the different reconstruction methods. Listed are the total point counts of the original reconstructions before any spatial filtering, and the number of points remaining after voxel-based downsampling with a 5 cm grid size. The last row indicates the number of points of the interpolated water surface model (WSM) at 5 cm resolution within the water-covered area. Values are rounded for readability.

5. Results

In this section, we present both qualitative and quantitative evaluations of the reconstructed water surfaces. The qualitative assessment analyses spatial error patterns and visual differences between the various reconstruction methods. The quantitative analysis subsequently focuses on geometric accuracy and completeness metrics derived from LiDAR reference data, including histograms of cloud-to-mesh (C2M) distances and statistical summaries.

5.1 Qualitative Assessment

The qualitative evaluation begins with a visual overview of the reconstructed 3D points for the investigated part of the river in Figure 1. The sparse SfM reconstruction (Figure 1a) clearly demonstrates that neither the water surface nor the riverbed could be reconstructed, highlighting the failure of conventional feature matching in low-texture and refractive areas. Only a few isolated points appear on the water surface, typically caused by coincidental feature matches on transient reflections or small floating structures, and these do not represent the true refractive interface. Consequently, alternative reconstruction methods are required to recover geometry in these regions. In contrast, the dense MVS reconstruction (Figure 1b) appears more complete, with substantially higher point density and improved coverage. However, the water surface still contains large gaps or noisy point clusters, reflecting the breakdown of photometric consistency in the presence of refraction, transparency, and fast-changing surface patterns. Only deeper areas with strong flow dynamics and surface reflections remain incomplete. The MDE results (Figure 1c) show the potential to further close these remaining gaps, suggesting that this approach can complement or extend classical multi-view reconstruction pipelines.

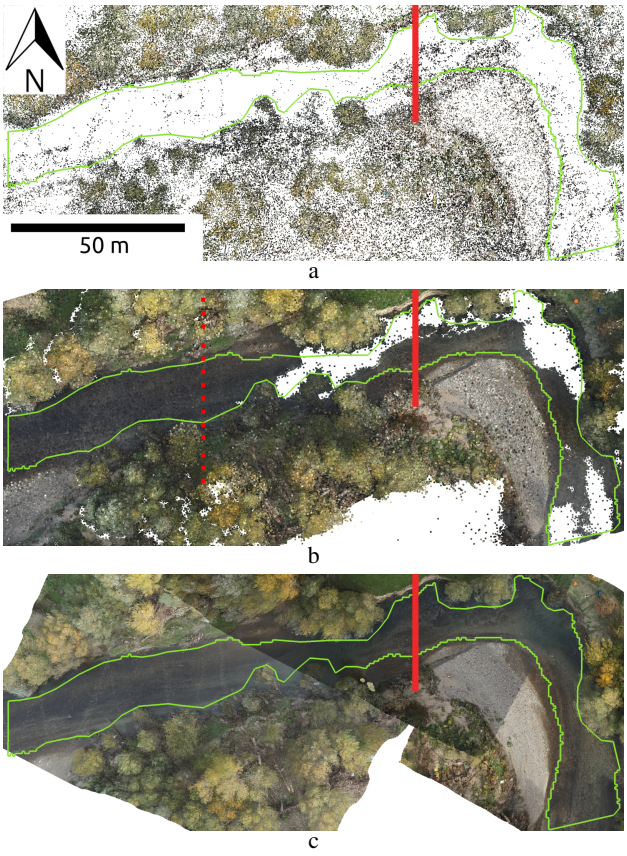

Figure 1. Top view of the investigated area of the Pielach River showing the distribution of reconstructed 3D points. a) Sparse SfM, b) Dense MVS, c) Three coregistered MDE reconstructions. The red solid bar in the East marks the position of the extracted cross section visualized in Figure 3. Similary, the dotted bar in b) marks the extracted cross section shown in Figure 2.

To examine these findings in more detail, cross-sectional profiles are analyzed in the following. A representative cross section extracted from a shallower part of the river, where the MVS reconstruction appears visually complete, is shown in Figure 2. Its location is marked as dotted line in Figure 1b. It becomes evident that the reconstructed geometry corresponds to the riverbed rather than the water surface. Even in regions that seem fully reconstructed from above, the geometry in fact represents the riverbed. This results from the exceptional water clarity, which causes the algorithm to match features on the bed rather than on the refractive interface. The cross section represents a part of the river with a depth of up to 80 cm and illustrates the fundamental limitation of standard multi-view stereo techniques in optically transparent waters, where the water surface cannot be reconstructed without an explicit refractive model or an alternative approach.

[1] https://github.com/graphdeco-inria/gaussian-splatting/blob/main/utils/make_depth_scale.py

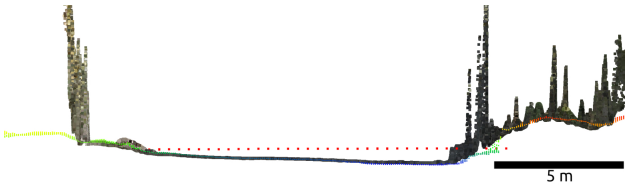

Figure 2. Cross section of the river of a part where the MVS reconstruction appears fully reconstructed. It is evident that the riverbed has been reconstructed. Here, the river is up to 80 cm deep. The location of the cross section is marked as a dotted line in Figure 1b.

A more comprehensive view of the reconstruction behavior is provided by additional cross sections extracted from a deeper part of the river (Figure 3). These profiles visualize the spatial distribution of points reconstructed by SfM, MVS, and MDE at one common location, marked by a red bar in Figure 1, where even the dense MVS reconstruction fails to recover the riverbed, resulting in large gaps and missing geometry. For readability, Figure 3c shows a single representative MDE point cloud only, although multiple MDE reconstructions were generated; these exhibit similar coverage but inconsistent scale and offset, causing vertical misalignment with respect to the WSM. This is also evident in the histograms in Figure 5, where only a limited fraction of MDE distances falls within 20 cm of the reference surface, while the majority of points are widely scattered around, mainly above it. In this area, the river reaches a maximum depth of approximately 1.1 m.

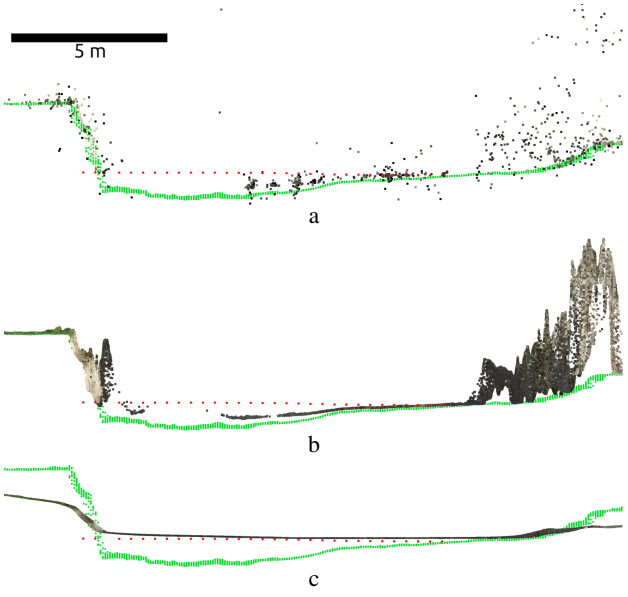

Figure 3. River cross section showing the digital terrain model (DTM) in green, the WSM as dotted red line, and RGB point clouds. The location of this cross section is marked as red solid bar in Figure 1. Here, the river is up to 1.1 m deep. Subfigures: WSM and DTM with a) Sparse SfM, b) Dense MVS, c) one selected MDE point cloud. Multiple MDE reconstructions were produced but are not overlaid for readability, since they exhibit similar coverage yet inconsistent scale and offset that lead to vertical misalignment with respect to the WSM.

Beyond the cross-sectional analysis, the completeness distribution maps in Figure 4 provide an additional perspective on the reconstruction performance. Each of the sub-figures represents one reconstruction variant, including two SfM-based results (raw and meshed) and two MVS-based results (raw and meshed) as well as the MDE-based result derived from the merged and subsampled point cloud of 23 monocular depth estimations. For each point of this aggregated cloud, the distance to the LiDAR-derived WSM was computed, and only those within 20 cm are visualized. These blue areas represent regions considered as reconstructed or complete, while white areas mark zones without a valid match within this distance threshold. The corresponding complete set of C2M distances, including those beyond the 20 cm range, is shown in the histograms of Figure 5.

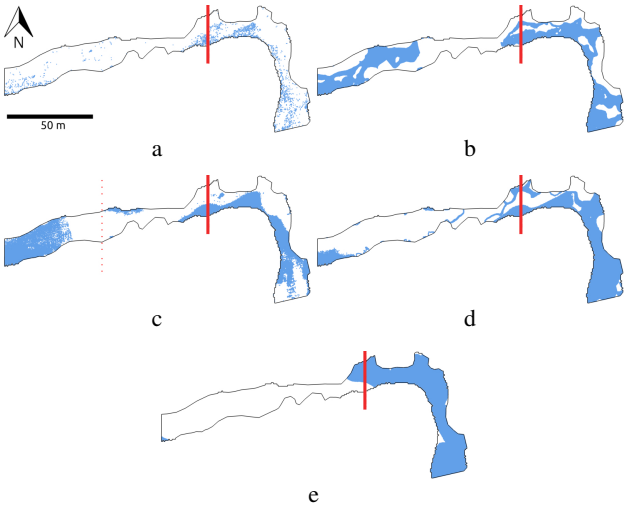

Figure 4. Spatial distribution of C2M distances smaller than 20 cm between the reconstructed surfaces and the WSM. a) Sparse SfM cloud, b) SfM mesh, c) dense MVS cloud, d) MVS mesh, e) MDE-based reconstruction (merged from 23 frames). Histograms of the full distance distributions are provided in Figure 5. Cross sections from Figure 2 and 3 in red.

Across all visualizations, the completeness remains low, as large portions of the water surface are not reconstructed. Overall, the qualitative evaluation confirms that none of the conventional image-based methods reliably reconstruct the water surface under clear and fast-flowing conditions, motivating the subsequent quantitative analysis of C2M distance distributions and completeness statistics.

5.2 Quantitative Assessment

Following the qualitative assessment, the quantitative analysis provides a numerical evaluation of reconstruction accuracy and completeness. While Figure 4 illustrates the spatial distribution of points within 20 cm from the LiDAR-derived WSM, the histograms in Figure 5 summarize these results statistically. Each histogram represents one reconstruction variant and shows the distribution of C2M distances between the reconstructed surface and the WSM. The red portion indicates the share of distances smaller than 20 cm, corresponding to the regions previously marked in blue in Figure 4. In addition, the mean (μ) and standard deviation (σ) of the distance distribution are reported, providing a quantitative measure of geometric deviation.

To facilitate direct comparison across reconstruction methods, Table 2 compiles the corresponding statistical metrics. Although histograms depict the overall distribution of deviations, the table summarizes the mean distance, standard deviation, and the proportion of points in each reconstruction that are within 20 cm of the WSM, allowing a direct evaluation of relative performance and geometric consistency.

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-2-2026-585-2026 | © Author(s) 2026. CC BY 4.0 License.

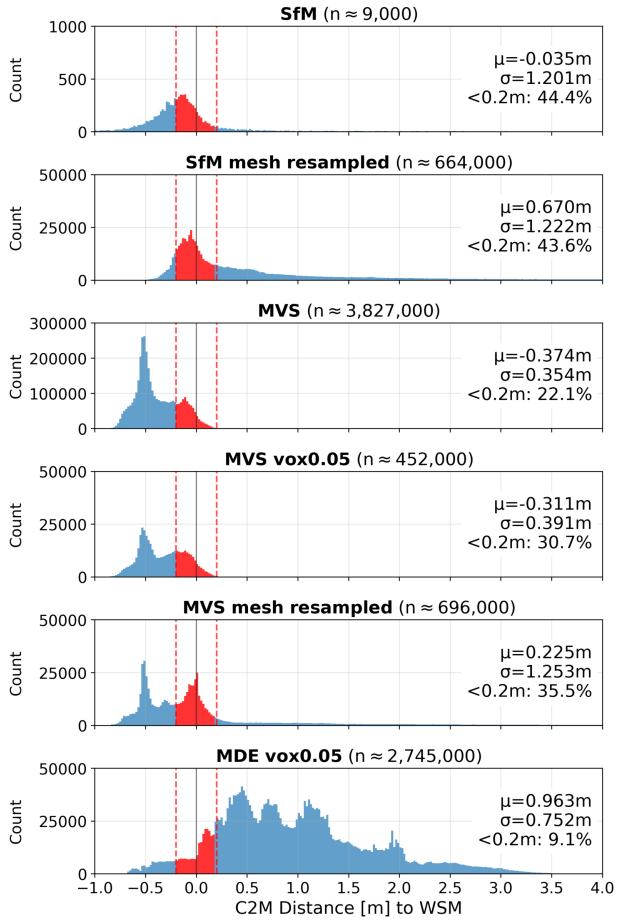

Figure 5. Histograms of C2M distances between reconstructed point clouds and the WSM mesh. The evaluation includes the original point clouds from SfM and MVS as well as their corresponding meshes that were smoothed and resampled to a dense point representation with an average spacing of approximately 5 cm. These derived point clouds were assessed by computing C2M distances to the WSM mesh. Additionally, a resampled point cloud from MDE is included for comparison. Each histogram uses a uniform bin width of 2 cm, with individually adjusted y-axis scales for visual clarity. The proportion of points within 20 cm of the WSM mesh is highlighted in red. Mean (μ), standard deviation (σ), and the percentage of points within this threshold are reported for each point cloud.

In the MDE-based reconstruction, the numerical results provide further evidence of the scale and offset inconsistencies already observed qualitatively. The large mean deviation and high standard deviation in Table 2 reflect a systematic misalignment with the WSM. This behaviour is consistent with the structural mismatch between the noisy SfM reference used for calibration and the strongly smoothed monocular depth predictions, which reduces local depth variation and limits the accuracy of the estimated global scale and offset. As a consequence, only a small fraction of MDE points falls within 20 cm of the WSM, despite the dense spatial coverage of the underlying depth maps.

6. Discussion

The following discussion interprets the quantitative and qualitative findings in the context of image-based reconstruction under refractive conditions, highlighting methodological limitations, error sources, and implications for future refraction-aware reconstruction methods.

Reconstruction method	Mean [m]	Std. [m]	<0.2 m [%]
SfM	**-0.035**	1.201	44.4
SfM mesh resampled	0.670	1.222	43.6
MVS	-0.374	**0.354**	22.1
MVS vox0.05	-0.311	0.391	30.7
MVS mesh resampled	0.225	1.253	35.5
MDE vox0.05	0.963	0.752	9.1

Table 2. Statistical summary of C2M distances between the reconstructed point clouds and the WSM. Mean (μ), standard deviation (σ), and the percentage of points within 20 cm of the WSM are reported for each reconstruction method.

Data Challenges. The Pielach dataset is a demanding scenario as clear water, strong flow dynamics, and low surface texture reduce the number of stable keypoints and compromise photometric consistency. Under these refractive conditions, purely image-based pipelines struggle independent of implementation details.

Histograms and Maps Interpretation. The histograms of the signed C2M distances (Figure 5) together with the maps of distances <20 cm (Figure 4) indicate low overall coverage of the water surface. The <20 cm column in Table 2 must be read strictly as the fraction of *points in the respective reconstruction* whose nearest neighbor to the WSM is within 20 cm. It does not represent the fraction of the WSM that is reconstructed, nor does it by itself indicate geometric correctness or accuracy. Interpretation should therefore rely on this metric together with the histograms and cross-sections.

SfM. The very small mean C2M distance in Table 2 is not meaningful due to the extremely high spread and sparse, irregular support. Isolated points may fall close to the WSM, but the surface is not reconstructed. In addition to these limitations, the multi-temporal nature of the UAV image sequence further violates the assumption of a static scene. Small surface motions between views break epipolar consistency, which further reduces the number of stable matches and contributes to the sparse and irregular SfM output.

MVS (Raw, Voxel, Mesh). The raw MVS reconstruction exhibits less spread than SfM reconstruction but the histogram is predominantly negative, consistent with the cross-sections (Figures 2, 3b): reconstructed 3D-points align with the riverbed rather than the water surface in clear shallow areas. While this behaviour is expected and even desirable in classical photobathymetry, where bottom topography is the primary target, it fundamentally limits the suitability of standard MVS as a method of explicit water surface reconstruction. The multi-temporal acquisition further contributes to this behavior, since even minor surface motion disrupts photometric consistency across views and shifts the depth optimization toward the static riverbed.

MDE (23 maps, merged, 5 cm voxel). Despite dense coverage, the C2M statistics reveal large offsets and scale inconsistencies across images (mean 0.963 m, σ 0.752 m, <20 cm share 9.1%). The current *offset-and-scale calibration*, estimated per image from SfM reprojections, is applied; however, the residuals indicate that this approach is not yet sufficient. A more robust estimation strategy, incorporating both per-image and potentially spatially varying scale refinement, appears necessary to achieve consistent metric alignment. The low completeness and large spread in the MDE evaluation can be explained by the struc-

https://doi.org/10.5194/isprs-annals-XI-2-2026-585-2026

tural mismatch between the reference and the monocular depth maps: the SfM reprojections used for calibration are noisy and exhibit high-frequency variations, whereas the monocular predictions are strongly smoothed, particularly in vegetated areas. This reduces local depth variation and limits the accuracy of the estimated global offset and scale.

MapAnything. This method (Keetha et al., 2025) is designed for unified metric three-dimensional reconstruction from single images and multi-view inputs and therefore represents a promising candidate for bridging the gap between monocular and multi-view depth inference. As mentioned in Section 2, it aims to predict geometry directly in a consistent metric space, which could, in principle, mitigate the scale ambiguity inherent in monocular approaches. This capability motivates its inclusion in the comparison. The reconstruction shown in Figure 6 appears visually plausible in top view, yet the cross section reveals substantial vertical inconsistencies. The deviations are similar to those observed for the calibrated monocular depth maps derived from Depth-Anything-V2 predictions aligned with SfM reference data. Offset and scale remain inconsistent across the scene, leading to poor alignment with the WSM. This indicates that the internal metric calibration of MapAnything is not yet sufficient for reliable water surface reconstruction in refractive real-world environments.

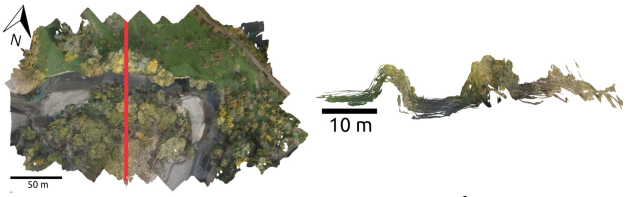

Figure 6. Top and side views of the reconstruction generated with MapAnything using 210 overlapping nadir images. a) Top view showing the reconstructed scene, with the red vertical bar indicating the position of the cross section shown in b). b) Side view along the marked cross section, illustrating the vertical inconsistencies and poor alignment of the estimated geometry.

DTM Reference Validation. To further substantiate the interpretation that MVS predominantly reconstructs the riverbed rather than the water surface, an additional comparison against a LiDAR-derived digital terrain model (DTM) was performed, but without refraction correction of the MVS point cloud since we actually aim for the water surface. While the WSM evaluation focuses on deviations relative to the water surface, the DTM provides an independent reference for the submerged topography. The histograms in Figure 7 show the signed C2M distances between the reconstructed point clouds and the DTM representing the riverbed, with distances within 20 cm highlighted in red. The corresponding spatial distributions in Figure 8 reveal that many MVS points fall close to the DTM in shallow areas, confirming that the algorithm locks onto features on the riverbed instead of the water surface. SfM exhibits a similar tendency, although with substantially lower point density. This complementary evaluation supports the conclusion that, under clear-water conditions, standard multi-view reconstruction aligns with the submerged topography rather than the refractive surface, underscoring the need for an explicit refractive model or alternative recovery strategy.

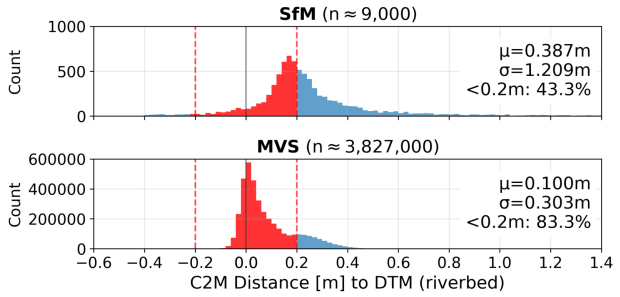

Figure 7. Histograms of signed C2M distances between the reconstructed point clouds and the LiDAR-derived DTM representing the riverbed. a) SfM reconstruction, b) MVS reconstruction. The mean (μ), standard deviation (σ), and completeness (within ± 0.2 m) are reported. The proportion of distances within this interval to the DTM are shown in red. This red fraction corresponds to the spatially distributed completeness visualized in Figure 8. The bin width is 2 cm. Due to the substantially lower number of points in the SfM reconstruction compared to MVS, the y-axes differ in scale.

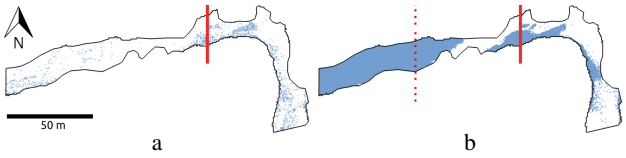

Figure 8. Spatial distribution of C2M distances smaller than 20 cm between the reconstructed point clouds and the LiDAR-derived digital terrain model (DTM) representing the riverbed. a) SfM reconstruction, b) MVS reconstruction. Cross sections from Figure 2 and 3 in red.

7. Conclusion and Outlook

In this contribution, it is shown that none of the investigated image-based methods, that is SfM, MVS and MDE, can reliably reconstruct the refractive water surface from UAV imagery of a clear, fast-flowing river. This systematic comparison demonstrates that SfM fails to produce consistent points, MVS reproduces the riverbed instead of the surface, and MDE exhibits scale and offset inconsistencies. These results demonstrate that current image-based pipelines are insufficient for metric water-surface reconstruction under refractive conditions. Providing an explicitly derived, geometrically accurate water surface model remains essential for robust refraction-aware modeling and accurate bathymetric reconstruction and is expected to benefit future refractive neural rendering methods as well.

The following directions may support further progress in refractive surface reconstruction: i) Improvement of offset and scale calibration for MDE, including robust per-image estimation on SfM reprojection masks and systematic outlier rejection; ii) Exploration of joint alignment of multiple depth maps through bundle-like optimization on inverse depth with separate offset and scale per frame; iii) Integration of water and land masks during calibration to suppress vegetated outliers; iv) Development and validation of the calibration strategy in controlled synthetic refractive scenes before re-deployment in the field, for example using recent physically based simulation datasets of fluvial environments (Schulte et al., 2025), which enable systematic testing of refraction handling and geometric consistency.

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-2-2026-585-2026 | © Author(s) 2026. CC BY 4.0 License.

Acknowledgments

The research presented was carried out within the transnational WEAVE project BathyNeRF funded by the Austrian Science Fund (FWF) [10.55776/PIN1353223] and the German Research Foundation (DFG) [538522540].

References

Chandler, J. H., Wackrow, R., Sun, X., Shiono, K., Rameshwaran, P., 2008. Measuring a dynamic and flooding river surface by close range digital photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII, 211–216.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*.

Elnashef, B., Filin, S., 2022. Target-Free Calibration of Flat Refractive Imaging Systems Using Two-View Geometry. *Optics and Lasers in Engineering*, 150, 106856.

Godard, C., Mac Aodha, O., Brostow, G. J., 2017. Unsupervised monocular depth estimation with left-right consistency. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 270–279.

Günthner, A., Brezovsky, M., Schulte, F., Winiwarter, L., Mandlburger, G., Jutzi, B., 2025. Exploring the Potential of Refractive NeRFs for Photogrammetric Bathymetry - First Application to UAV-based Data from the Pielach River. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2-W10-2025, 107–114.

Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E., Remondino, F., Jutzi, B., Ruf, B., 2024. Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 13.

Höhle, J., 1971. Zur Theorie und Praxis der Unterwasser-Photogrammetrie. *Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften, Series C*, 163.

Jordt-Sedlazeck, A., Koch, R., 2013. Refractive Structure-from-Motion on Underwater Images. *Proceedings of the IEEE International Conference on Computer Vision*, 57–64.

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S. R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P., 2025. MapAnything: Universal Feed-Forward Metric 3D Reconstruction. *arXiv*.

Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4), 139:1–139:14.

Kotowski, R., 1988. Phototriangulation in multi-media photogrammetry. *International Archives of Photogrammetry and Remote Sensing*, 27(B5), 324–334.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. *2016 Fourth international conference on 3D vision (3DV)*, IEEE, 239–248.

Liu, F., Shen, C., Lin, G., 2015. Deep convolutional neural fields for depth estimation from a single image. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5162–5170.

Maas, H.-G., Sardemann, H., Mulsow, C., Gueguen, L.-A., Mandlburger, G., 2025. New Approaches in Photo-Bathymetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W10-2025, 185–190.

Mandlburger, G., 2019. Through-Water Dense Image Matching for Shallow Water Bathymetry. *Photogrammetric Engineering & Remote Sensing*, 85, 445–455.

Mandlburger, G., Rhomberg-Kauert, J., Gueguen, L.-A., Mulsow, C., Brezovsky, M., Dammert, L., Haines, J., Glas, S., Himmelsbach, T., Schulte, F., Amon, P., Winiwarter, L., Jutzi, B., Maas, H.-G., 2025. Mapping shallow inland running waters with UAV-borne photo and laser bathymetry : The Pielach River showcase. *Journal of Applied Hydrography*, 130(3), 42–54.

Manfreda, S., Miglino, D., Saddi, K. C., Jomaa, S., Eltner, A., Perks, M., Peña-Haro, S., Bogaard, T., van Emmerik, T. H., Mariani, S., Maddock, I., Tauro, F., Grimaldi, S., Zeng, Y., Gonçalves, G., Strelnikova, D., Bussettini, M., Marchetti, G., Lastoria, B., Su, Z., Rode, M., 2024. Advancing river monitoring using image-based techniques: challenges and opportunities. *Hydrological Sciences Journal*, 69(6), 657–677.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1), 99–106.

Morris, N. J. W., 2004. Image-based Water Surface Reconstruction with Refractive Stereo. Master’s thesis, University of Toronto.

Mulsow, C., Sardemann, H., Gueguen, L.-A., Mandelburger, G., Maas, H.-G., 2024. Concepts for compensation of wave effects when measuring through water surfaces in photogrammetric applications. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2-2024, 289–295.

Qian, Y., Zheng, Y., Gong, M., Yang, Y.-H., 2018. Simultaneous 3D Reconstruction for Water Surface and Underwater Scene. *Proceedings of the European Conference on Computer Vision (ECCV)*, 754–770.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-From-Motion Revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. *European conference on computer vision*, 501–518.

Schulte, F., Brezovsky, M., Günthner, A., Jutzi, B., Mandlburger, G., Winiwarter, L., 2025. Simulation and validation of underwater scenes for two-media optical 3D reconstruction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2-W10-2025, 271–278.

Seitz, S., 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, 519–528.

Thapa, S., Li, N., Ye, J., 2020. Dynamic Fluid Surface Reconstruction Using Deep Neural Network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21–30.

Ullman, S., 1997. The Interpretation of Structure from Motion. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 203(1153).

Xiong, J., Heidrich, W., 2021. In-the-Wild Single Camera 3D Reconstruction Through Moving Water Surfaces. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 12538–12547.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth Anything V2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

Zhan, Y., Nobuhara, S., Nishino, K., Zheng, Y., 2023. NeRFrac: Neural Radiance Fields through Refractive Surface. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 18356–18366.