The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W3-2023
2nd GEOBENCH Workshop on Evaluation and BENCHmarking of Sensors, Systems and GEOspatial Data
in Photogrammetry and Remote Sensing, 23–24 October 2023, Krakow, Poland

BENCHMARKING THE EXTRACTION OF 3D GEOMETRY FROM UAV IMAGES WITH DEEP LEARNING METHODS

F. Nex[1], N. Zhang[1], F. Remondino[2], E.M. Farella[2], R. Qin[3], C. Zhang[3]

[1] ITC Dep. of Earth Observation Science, University of Twente, Enschede, Netherlands - Email: <f.nex><n.zhang>@utwente.nl

[2] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy - Email: <remondino><elifarella>@fbk.eu

[3] Geospatial Data Analytics Laboratory, The Ohio State University, Columbus, USA - Email: <qin.324><zhang.13596>@osu.edu

Commission I

KEYWORDS: Photogrammetry, UAV, Deep Learning, 3D, NeRF, MVS, monocular, benchmark

ABSTRACT:
3D reconstruction from single and multi-view stereo images is still an open research topic, despite the high number of solutions proposed in the last decades. The surge of deep learning methods has then stimulated the development of new methods using monocular (MDE, Monocular Depth Estimation), stereoscopic and Multi-View Stereo (MVS) 3D reconstruction, showing promising results, often comparable to or even better than traditional methods. The more recent development of NeRF (Neural Radial Fields) has further triggered the interest for this kind of solution. Most of the proposed approaches, however, focus on terrestrial applications (e.g., autonomous driving or small artefacts 3D reconstructions), while airborne and UAV acquisitions are often overlooked. The recent introduction of new datasets, such as UseGeo has, therefore, given the opportunity to assess how state-of-the-art MDE, MVS and NeRF 3D reconstruction algorithms perform using airborne UAV images, allowing their comparison with LiDAR ground truth. This paper aims to present the results achieved by two MDE, two MVS and two NeRF approaches levering deep learning approaches, trained and tested using the UseGeo dataset. This work allows the comparison with a ground truth showing the current state of the art of these solutions and providing useful indications for their future development and improvement.

1. INTRODUCTION

3D reconstruction from images is an enduring research task in the photogrammetric and computer vision communities. Despite the introduction of multiple open-source and commercial solutions for 3D reconstruction, several challenges and limitations still exist: textureless areas, transparencies, or reflective surfaces are just examples of regions where available methods often fail to deliver a correct 3D reconstruction.

In the recent years, deep learning algorithms have demonstrated great potential in several remote sensing tasks, including image-based 3D reconstruction. Nowadays there are several monocular and stereo algorithms leveraging deep learning techniques and achieving comparable results with more conventional methods for depth estimation and 3D reconstruction. However, one of the limitations of such learning-based methods is that they highly rely on large training sets that are often tedious to obtain. Moreover, they are generally applied to close-range scenarios low-resolution images and quantitative evaluations while best practices for daily uses and large-scale scenarios are generally missing.

UAV (Unmanned Aerial Vehicles) are valuable platforms for geospatial data acquisition and have demonstrated their potential in multiple applications and fields (Nex and Remondino, 2014; Candiago et al., 2015; Hassanalian and Abdelkefi, 2017; Nex et al., 2022). These platforms are adopted in a wide range of applications where often 3D reconstruction is one of the main outputs as useful in many cases: most of the inspection, surveying and mapping activities usually need a 3D reconstruction to determine the shape, the extension and the geo-localization of monitored scenes. Ultra-high-resolution UAV images are often an extra challenge to face for achieving accurate 3D reconstructions. Despite the incredible number of sophisticated algorithms developed in the last two decades for image triangulation and dense matching, conventional (hand-crafted) methods often deliver noisy or incomplete point clouds. In that regard, deep learning methods could represent a valid complementary approach to improve and (maybe) overcome traditional methods exploiting the information that can come from one or multiple images. Besides the conventional stereo or multi-view reconstruction (MVS) algorithms (Wang et al., 2021; Stathopoulou and Remondino, 2023), deep learning has also revamped the so-called Monocular Depth Estimation (MDE) algorithms that infer the depth of a scene from a single image (Ming et al., 2021; Masoumian et al., 2022): different approaches using supervised, unsupervised and self-supervised methods have been presented in the last years. At the same time, Neural Radiance Field (NeRF) methods (Mildenhall et al., 2020) have defined a novel way to reconstruct 3D objects by synthesizing novel views of a scene by optimizing a continuous 5D volumetric scene function. Despite the impressive results on relatively small 3D scenes and objects (Remondino et al., 2023), it is unclear if this typology of algorithms will be a valid alternative for wider (i.e., remote sensing) applications.

These methods are then constrained to the use of large training datasets and their performances are still conditioned by transferability limits: as an example, networks trained using terrestrial data deliver generally poor results when tested on other typologies of data such as airborne images.

UAV datasets are not commonly used for training deep learning algorithms. In that regard, the recent ISPRS Scientific Initiative UseGeo (https://usegeo.fbk.eu) has released datasets which represents a good starting point to support the further development of deep learning algorithms considering ultra-high-resolution UAV images.

1.1 Paper aims

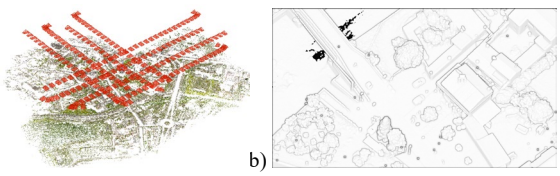
This paper wants to investigate the use of deep learning methods for extracting geometric information from UAV images, evaluating some meaningful state-of-the-art methods and reporting quantitative analyses and lessons learnt for each of them. In particular, the work examines different learning-based approaches for three processes:

- monocular depth estimation (MDE), using single images to predict depths;

- multi-view stereo (MVS), using two or multiple images to reconstruct a 3D scene;
- 3D reconstruction with Neural Radiance Field (NeRF).

2. THE *USEGEO* DATASET

For the scope of the paper, data from UseGeo's repository are employed (https://github.com/3DOM-FBK/usegeo). UseGeo - UAV-based multi-sensor datasets for geospatial research - was an ISPRS Scientific Initiative which aimed to deliver new and unique datasets for the rigorous assessment of 3D reconstruction algorithms from UAV images (https://usegeo.fbk.eu/). The datasets contains both image and LiDAR data (Figure 1) and aims to support relevant research, contributing with a useful training set for both stereo and monocular 3D reconstruction algorithms. Data have been collected with a RIEGL miniVUX-3UAV scanner and a SONY ILCE-7RM3 camera. The datasets consist of more than 800 images acquired in three different areas and corresponding LiDAR point clouds as ground truth (GT). Each acquisition was performed on average with 80% and 60% forward and side image overlap, respectively. This overlap guarantees a minimum of 8 images on each object point, with a GSD smaller than 2 cm. The available LiDAR data are ground truth for MVS and NeRF algorithms or depth maps (MDE) methods. Different tests have been performed to validate the UseGeo benchmark and guarantee their suitability for the assessment of deep learning algorithms. For more information on UseGeo, please refer to Nex et al. (2023).



a) b)

Figure 1: Camera network (left) and LiDAR ground truth point cloud (right, more than 100 pts/sqm) of an UseGeo dataset used in the experiments of the paper.

3. EXTRACTION OF GEOMETRIC INFORMATION

3.1 Monocular depth estimation with deep learning

Monocular depth estimation (MDE) is an ill-posed process to recover distances between the camera and objects in the 3D scene. Early methods relied on handcrafted features and used complementary cues (Saxena et al., 2008) while recent deep learning approaches employ deep convolutional neural networks (Eigen et al., 2014; Watson et al., 2019; Ranftl et al., 2020; Welponer et al., 2022; Bhat et al., 2023; Zhang et al., 2023). MDE is normally tackled as supervised (Fu et al., 2018; Hu et al., 2019) or self-supervised (Godard et al., 2017; Tosi et al., 2019) problem, incorporating 3D awareness and constraints in order to avoid distortions and artefacts (Yin et al., 2019; Yin et al., 2021). An architecture often adopted for MDE is the encoder-decoder (Fu et al., 2018) with RGB images as input and direct regression of pixel-wise depth maps as output. Depending on the available training data, the scene depth can be estimated as ordinal, i.e., relative (Fu et al., 2018) or Euclidean (Eigen et al., 2014; Yin et al., 2019). Few models were trained and evaluated on UAV and aerial datasets (Hermann et al., 2020; Madhuanand et al., 2021; Chang et al., 2023). MDE could help to complement conventional 3D methods in textureless areas or it could be useful for navigation/visual odometry purposes, obstacle avoidance, etc.

3.2 Learning-based multi-view stereo

Dense image matching and multi-view stereo (MVS) algorithms aim to generate a rich, dense 3D reconstruction of the scene in the form of a dense point cloud or a triangulated mesh (Remondino et al., 2014; Furukawa and Hernandez, 2015; Zhou et al., 2020; Stathopoulou and Remondino, 2023). Starting from camera poses and sparse points, the depth of generally every pixel of the scene has to be calculated. Several methods, either conventional (Bleyer et al., 2011; Rothermel et al., 2012; Schönberger et al., 2018) or, more recently, learning-based (Huang et al., 2018; Yao et al., 2018; Xu and Tao, 2020c; Wang et al., 2021; Liu et al., 2023) have been developed for solving the dense correspondence search problem. Considering the depth estimation with supervised learning methods, the loss function in the training process tries to minimize the discrepancy between the ground truth and the estimated depth along with a regularization smoothness term (Yang et al., 2020; Xu et al., 2021; Wang et al., 2022). In unsupervised and self-supervised learning methods, authors tried to by-pass the requirement of GT depth maps for training: the loss typically aims to minimize the photometric consistency error across the views in an unsupervised way while considering occlusions, photometric and geometric consistency or enforcing cross-view consistency (Zhong et al., 2017; Dai et al., 2019; Huang et al., 2021). Learning-based MVS methods applied to UAV and aerial datasets are still a research frontier in photogrammetry (Liu et al., 2018; Yu et al., 2021; Li et al., 2023a).

3.3 Neural Radiance Field algorithms

A recent and innovative approach for image-based 3D reconstruction is based on Neural Radiance Fields (NeRF) methods (Mildenhall et al., 2020; Barron et al., 2021; Gao et al., 2022; Li et al. 2023b). A NeRF uses implicit representations and combines deep learning methods with physical knowledge from computer graphics to achieve controllable and photo-realistic 3D models of a scene pictured by multi-view images. NeRF is capable of producing novel views of complex scenes by optimizing a continuous scene function from a set of oriented images. NeRF works by training a fully connected network, referred to as a neural radiance field, to replicate the input views of a scene through the use of a rendering loss. From the initial work of Mildenhall et al. (2020), researchers have proposed several modifications and extensions to the original NeRF method in order to improve performance and 3D results (Zhang et al., 2021; Kolodiazhna et al., 2023; Reiser et al., 2023). Beside Instant-NGP (Mueller et al., 2022), several NeRF methods have been included into open frameworks, such as SDFStudio (Yu et al., 2022) and NerfStudio (Tancik et al., 2023). NeRFs provide an alternative solution for 3D reconstruction compared to traditional photogrammetry methods and can produce promising results in situations where photogrammetry may fail to deliver accurate results (Mazzacca et al., 2023; Remondino et al., 2023). The use of NeRF with UAV datasets and large scale scenarios is quite recent and few best practices and models for 3D modeling or navigation purposes are available (Adamkiewicz et al., 2022; Turki et al., 2022; Patel et al., 2023; Turki et al., 2023).

4. DATA PROCESSING AND TESTED METHODS

The extraction of geometric information following Section 2 methods have been tested to show their performances with UAV images from the UseGeo dataset. In the following some relevant results are briefly summarized.

4.1 Monocular depth estimation methods

For the assessment of MDE on UAV images, two self-supervised algorithms have been used:

- Madhuanand et al. (2021): it adopts a self-supervised architecture for video sequences with two 2D CNN encoders and a 3D CNN decoder for extracting information from consecutive temporal frames. A contrastive loss term is introduced for improving the quality of image generation.
- Zhang et al., 2023: it refers to Lite-Mono, a self-supervised MDE approach initially developed using the KITTI benchmark. The architecture is lightweight and it takes advantage of hybrid CNN and Transformer to extract multi-scale image features. Specifically, Lite-Mono uses consecutive dilation convolutions (CDC) to expand receptive fields and learn enhanced local features, and it uses the Local-Global Feature Interaction (LGFI) module to model long-range global contexts. The network has four variants with different parameters, originally designed for real-time depth estimation on edge devices.

The method of Madhuanand et al. (2021) was trained using two subsets of images from the Hessigheim3D (Kölle, et al., 2021) dataset and Zeche Zollern (Nex et al., 2015) datasets together with a subsample of UseGeo data (Nex et al., 2023). The original image size of these two other datasets was modified to have a similar format as the UseGeo depth maps (i.e., 1898×1320 px). A total of 1036 images were used for training, 136 images for validation and 88 for testing. The testing was performed using only images from the UseGeo dataset.

The method presented in Zhang et al. (2023) was trained using the only UseGeo dataseet 1. To train and evaluate the largest model Lite-Mono-8m (see Zhang et al., 2023) was used. The UseGeo dataset was split into a training set of 728 images and a test set of 100 images. The model was trained with a batch size of 14. All the input images were resized to 768×448 pixels, and data augmentations such as horizontal flips, brightness adjustment (±0.2), saturation adjustment (±0.2), contrast adjustment (±0.2), and hue jitter (±0.1) were applied with a 50% chance. The initial learning rates to train the depth network and the pose network were set to $5e^{-5}$ and $1e^{-4}$, respectively.

Method	Abs Rel	Sq Rel	RMSE	$\delta_{1.25}$	$\delta_{1.15}$	$\delta_{1.05}$
#1	0.049	0.377	5.967	0.999	0.968	0.579
#2	0.076	0.834	9.215	0.989	0.878	0.355

Table 1. Quantitative results with the two MDE methods.

Both methods were scaled (with an average scale factor) to determine their residuals with respect to the available ground truth. Table 1 reports quantitative results on the testing images whereas some predicted depths are shown in Figure 2. Method #1 was trained for 500 epochs while method #2 for 1500 epochs (method X2), taking approximately 12 hours on an NVIDIA A40 GPU.

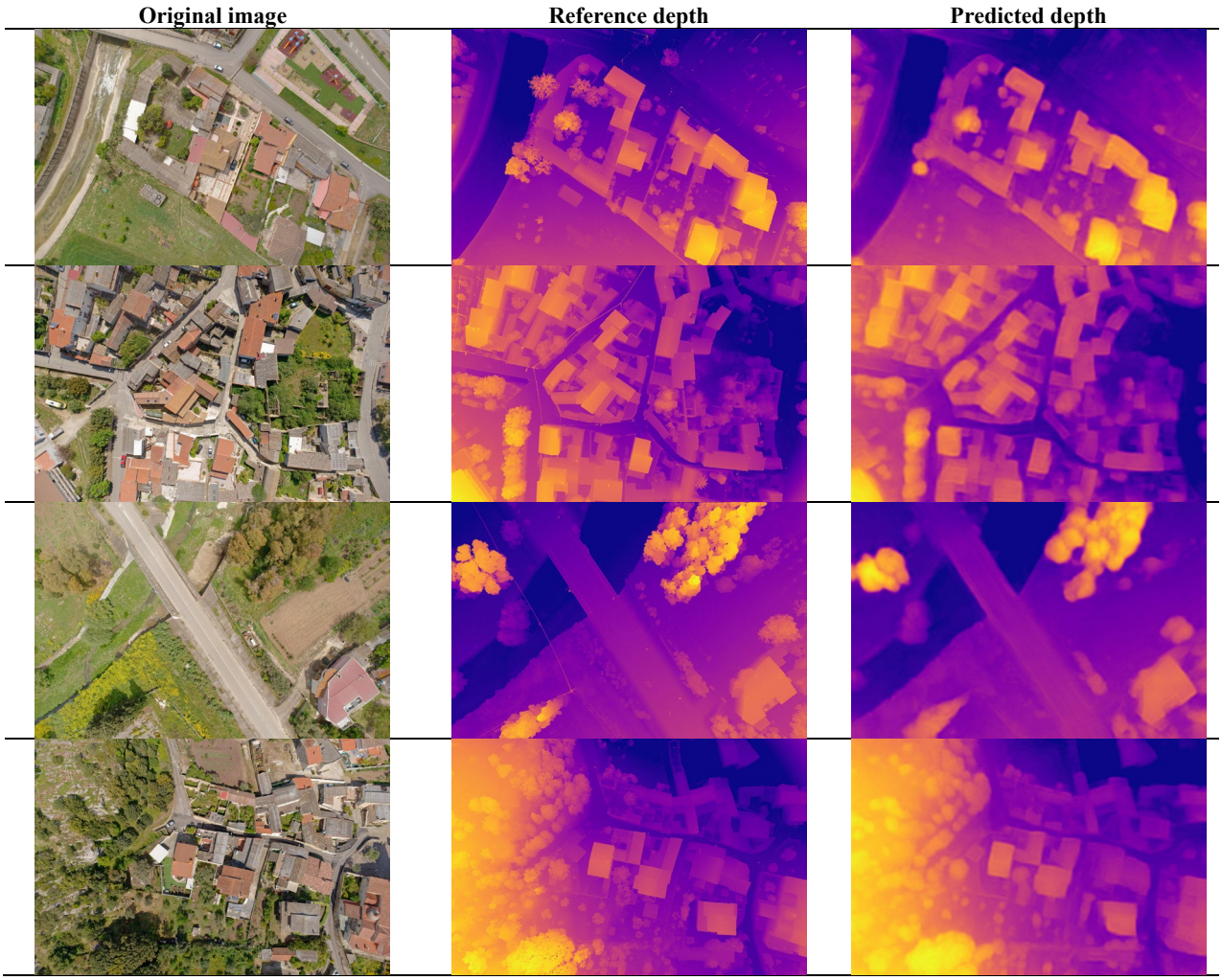


Figure 2: Reference and predicted depth maps achieved with MDE (Zhang et al., 2023) on same UAV images.

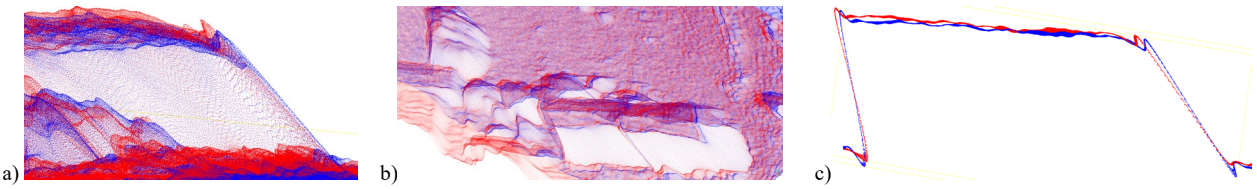


Figure 3. Overlap of different contiguous point clouds derived from the inferred depth maps (a,b) and a cross-section (c) showing misalignments and errors nearby man-made objects.

From the visual results, it can be observed that the predicted depth distribution is close to the ground-truth, although the sharpness of inferred depth maps can still be improved. The residuals are generally low and the reported metrics show some promising results for future adoption of MDE in UAV projects.
UseGeo provides also camera poses, hence depth maps were converted into point clouds and overlapped to evaluate their consistency (Figure 3). Although the results are generally good in the flat rural areas, some misalignments are visible in correspondence of man-made objects, mainly due to noisy 3D reconstructions and erroneous shapes in correspondence of man-made objects' borders due to the different perspectives of the images used for each depth estimation.

4.2 Multi-view stereo reconstruction

For the assessment of learning-based MVS methods on UAV images, the following methods have been tested:

- MVSFormer (Cao et al., 2023): it is a MVS learning architecture based on Vision Transformers which can be generalized to various input resolutions with efficient multi-scale training strengthened by gradient accumulation.
- UniMVS (Peng et al., 2022): it is a unified approach to exploit the advantages of regression and classification in depth estimation for MVS tasks. It constrains the cost volume like classification methods but also realizes the sub-pixel depth prediction like regression methods.

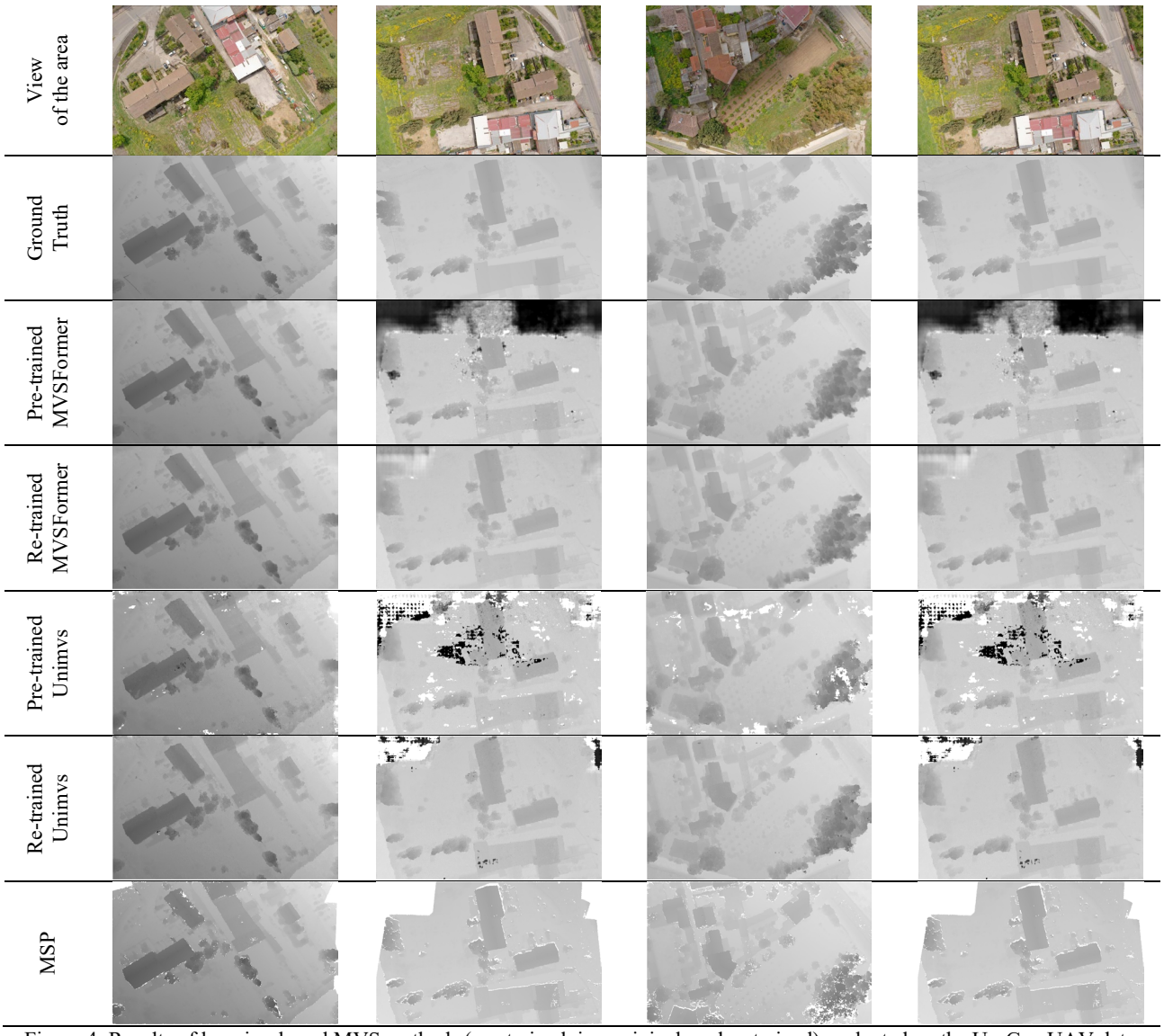


Figure 4: Results of learning-based MVS methods (pre-trained, i.e., original, and re-trained) evaluated on the UseGeo UAV data.

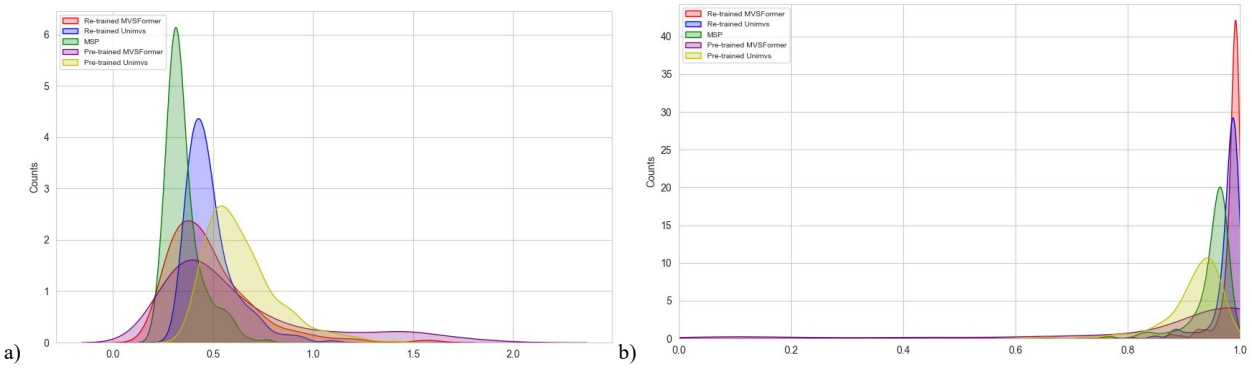


Figure 5: Graphical summary of MAE (a) and completeness (b) for the UseGeo dataset and the different MVS methods.

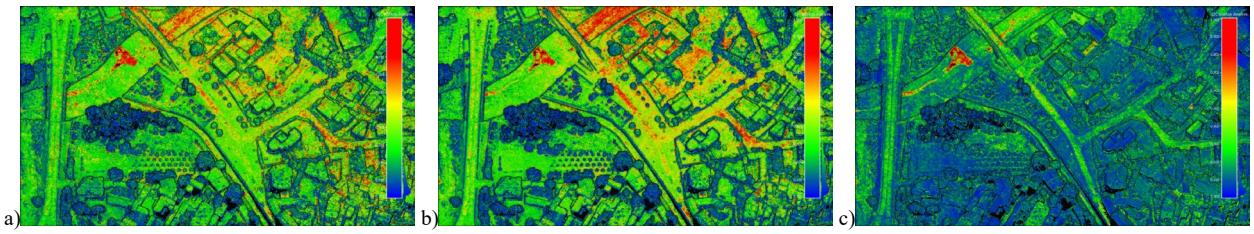


Figure 6: Cloud-to-cloud colour-coded difference between LiDAR GT and Re-trained MVSFormer (a), Re-trained Unimvs (b) and MSP (c). See Table 3 for metrics.

Both methods were first applied "as is" to dataset 1 of UseGeo (224 images) to check their replicability to an UAV context ("pre-trained" rows in Figure 4). Then both methods were re-trained using dataset 2 of UseGeo (327 images, 10 epochs) and, again, evaluated on dataset 1 ("re-trained" rows in Figure 4). A conventional SGM-based encoded in the MSP method[1] was also tested on dataset 1. A summary of accuracy (Mean Absolute Error – MAE) and completeness on all 224 images of UseGeo dataset 1 with respect to the available GT is presented in Table 2 and Figure 5. It is clear how conventional MVS are still outperforming in terms of accuracy, as also reported in other analyses (Mazzacca et al., 2023; Remondino et al., 2023). On the other hand, learning-based approaches are able to reach much higher completeness in the 3D scene.

	MAE [m]	Completeness [%]
MSP	0.35	94
Pre-trained MVSFormer	0.617	88
Re-trained MVSFormer	0.48	98
Pre-trained Unimvs	0.638	91
Re-trained Unimvs	0.485	97

Table 2: MAE and completeness for conventional and learning-based MVS methods on UseGeo Dataset 1. See Figure 5 for visuals.

From each MVS methods, point clouds were also derived and compared to the LiDAR data: Cloud-to-Cloud results are shown in Figure 6 whereas metrics in Table 3. It can be noticed how a conventional MVS method is still outperforming learning-based approaches.

	Mean [m]	Stand. Dev. [m]
MSP	0.0845	0.0805
Re-trained MVSFormer	0.1316	0.1099
Re-trained Unimvs	0.1682	0.1305

Table 3: Cloud-to-cloud statistics for learning-based and conventional MVS methods on UseGeo Dataset 1. See Figure 6 for visuals.

4.3 Neural Radiance Field (NeRF)

For the evaluation of NeRF methods on UAV images, a method called MCT-Nerf (Xu et al., 2023) built upon Mip-NeRF (Barron et al., 2021) was used and trained with the 224 images from UseGeo's dataset 1. To enable the use of the large-format images from UseGeo (7953 by 5279 pixels), MCT-Nerf uses a tile-based approach which divides the scene into small areas (e.g., 50 x 50 m^2) based on the ground plane and tiles the images into small patches (e.g., 800 by 800 pixels) with re-adjusted principal points. After the training, depth maps for each view are derived using standard approaches (Kangle et al., 2022) that take a weighted average of depths based on the accumulated radiance intensity (Figure 7a-b-c). These depth maps are finally combined and point clouds are derived. In addition, meshes are generated (Figure 7d) by extracting the triangulated surface from the fused depth map (Izadi et al., 2011) using marching cube algorithms (Newman and Hong, 2006). The derived depth maps are compared to the available GT data to derive metrics: Mean Absolute Error is 1.72 m whereas completeness reached 88%. A cloud-to-Cloud (C2C) comparison with the available GT data revealed a mean of 0.175 m and a standard deviation of 0.253 m.

5. CONCLUSIONS

This paper has shown the performances of different 3D reconstruction algorithms, based on deep learning methods, on UAV images. Most of the existing contributions focus on the use of these algorithms in terrestrial or close-range applications. In contrast, this paper has tested MDE, MVS and NeRF on the UseGeo UAV dataset with the aim of assessing their performances on high-resolution drone/UAV data.
MDE approaches deliver results that have still a lower quality and smoothed edges compared to ground truth data: the point clouds are usually noisier than the conventional photogrammetric ones and show deformations that prevent their correct alignment when combined using external orientation parameters of individual images.

[1] https://u.osu.edu/qin.324/msp/

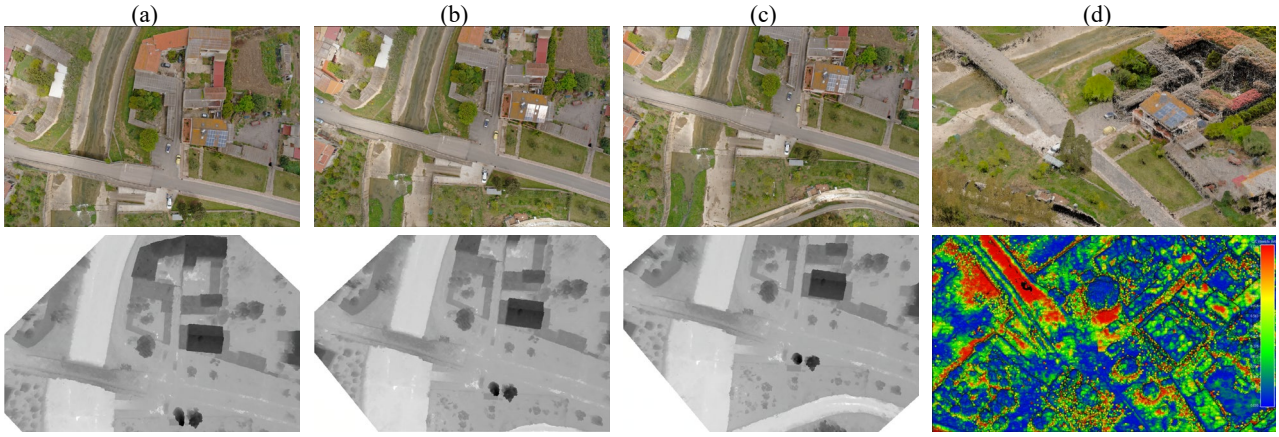


Figure 7: Results of the NeRF-based depth maps on the UseGeo UAV dataset 1 (a,b,c), point cloud and C2C results (d).

when combined using external orientation parameters of individual images. The training used on the images also deeply influences the results: larger training datasets might allow better results, although generalization and domain adaptation could be compromise. In addition, self-supervised approaches suffer from the lack of a ground-truth scale factor, increasing the discrepancies in the merge of different point clouds.

Learning-based MVS methods needed some re-training to be tailored to UAV scenarios. More complete results were produced although accuracy is below conventional methods, hence complementarity is the actual key word. Cloud-to-Cloud analyses also highlighted larger errors with respect to a traditional photogrammetric method.

NeRF methods are promising although geometric results are still not competing with conventional photogrammetric 3D reconstruction approaches where much higher and detailed surface models can be achieved. An approach to handle high resolution images is proposed but the need of many overlapping images with very short baselines is somehow limiting NeRF applications in aerial cases.

Evaluations of more methods are necessary. Nevertheless, performed tests have certainly shown that 3D reconstruction from UAV images can benefit from deep learning methods, ideally as complementary to conventional method and promising results are on the horizon.

ACKNOWLEDGEMENTS

The realization of the UseGeo dataset was supported by ISPRS. The work was partly supported by the project "AI@TN" funded by the Autonomous Province of Trento, Italy. Authors are also thankful to Chaoyi Zhou, Debao Huang and Ningli Xu (OSU) for their support within the paper experiments.

REFERENCES

Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., Schwager, M., 2022. Vision-Only Robot Navigation in a Neural Radiance World. *IEEE Robotics and Automation Letters*, Vol. 7(2), pp. 4606-4613.

Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R. and Srinivasan, P.P., 2021. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. Proc. *IEEE CVPR.*

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv*:2302.12288.

Bleyer, M., Rhemann, C., and Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows. *Proc. BMVC*, Vol. 11, pages 1-11.

Candiago, S., Remondino, F., De Giglio, M., Dubbini, M., Gattelli, M., 2015. Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. *Remote Sensing*, Vol. 7(4), pp. 4026-4047.

Cao, C., Ren, X., Fu, Y., 2023. MVSFormer: Multi-View Stereo by learning robust image features and temperature-based *Depth. Transactions of Machine Learning Research*.

Chang, R., Yu, K., Yang, Y., 2023. Self-supervised monocular depth estimation using global and local mixed multi-scale feature enhancement network for low-altitude UAV remote sensing. *Remote Sensing,* 15, 3275.

Dai, Y., Zhu, Z., Rao, Z., and Li, B., 2019. MVS2: Deep unsupervised multi-view stereo with multi-view symmetry. Proc. IEEE *Int. 3DV Conference*, pp. 1-8.

Eigen, D., Puhrsch, C., and Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Fu, H., Gong, M., Wang, C., Batmanghelich, K. and Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. Proc. *IEEE CVPR*, pp. 2002-2011.

Furukawa, Y. and Hernández, C., 2015. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, Vol. 9(1-2), pp. 1-148.

Gao, K., Gao, Y., He, H., Lu, D., Xu, L. and Li, J., 2022. NeRF: Neural radiance field in 3D vision, a comprehensive review. *arXiv preprint arXiv*:2210.00379.

Hassanalian, M., Abdelkefi, A., 2017. Classifications, applications, and design challenges of drones: A review. *Progress in Aerospace Sciences*, Vol. 91, 99-131.

Hermann, M., Ruf, B., Weinmann, M., Hinz, S., 2020. Self-supervised learning for monocular depth estimation from aerial

imagery. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* V-2–2020, pp. 357-364.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, Vol. 30(2), pp. 328-341.

Huang, B., Yi, H., Huang, C., He, Y., Liu, J., and Liu, X., 2021. M3VSnet: unsupervised multi-metric multi-view stereo network. *Proc. IEEE Int. Conf. on Image Processing*, pp. 3163-3167.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., and Huang, J.-B., 2018. DeepMVS: Learning multi-view stereopsis. Proc. *IEEE CVPR*, pp. 2821-2830.

Kangle, D., Liu, A., Zhu, J.Y., Ramanan, D., 2022. Depth-supervised nerf: Fewer views and faster training for free. Proc. *IEEE CVPR*, pp. 12882-12891.

Kolodiazhna, O., Savin, V., Uss, M. and Kussul, N., 2023. 3D Scene reconstruction with Neural Radiance Fields (NeRF) considering dynamic illumination conditions. Proc. *Int. Conference on Applied Innovation in IT*, Vol. 11(1), pp. 233-238.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. Proc. 24th *ACM UIST*, pp. 559-568.

Li, J., Huang, X., Feng, Y., Ji, Z., Zhang, S., Wen, D., 2023a. A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multiview stereo reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61.

Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y. and Lin, C.H., 2023b. Neuralangelo: high-fidelity neural surface reconstruction. Proc. *IEEE CVPR*, pp. 8456-8465.

Liu, J., Ji, S., Zhang, C., Qin, Z., 2018. Evaluation of deep learning based stereo matching methods: From ground to aerial images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 42.

Liu, J., Gao, J., Ji, S., Zeng, C., Zhang, S., Gong, J., 2023. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images. *ISPRS J. of Photogrammetry and Remote Sensing*, Vol.204, 42-60.

Madhuanand, L, Nex, F, Yang, M.Y., 2021. Self-supervised monocular depth estimation from oblique UAV videos. *ISPRS J. of Photogrammetry and Remote Sensing*, Vol. 176, pp.1-14.

Masoumian, A., Rashwan, H.A., Cristiano, J., Asif, M.S., Puig, D., 2022. Monocular depth estimation using deep learning: A review. *Sensors*, 22, 5353

Mazzacca, G., Karami, A., Rigon, S., Farella, E. M., Trybala, P., Remondino, F., 2023. NeRF for heritage 3D reconstruction. *Int. Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, Vol. XLVIII-2/W4-2023.

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*, Vol. 65(1), pp. 99-106.

Ming, Y., Meng, X., Fan, C., Yu, H., 2021: Deep learning for monocular depth estimation: A review. *Neurocomputing*, Vol. 438, pp. 14-33.

Mueller, T., Evans, A., Schied, C. and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, Vol. 41(4), pp. 1-15.

Newman, T.S., Hong, Y., 2006. A survey of the marching cubes algorithm. *Computers & Graphics*, Vol. 30(5), pp. 854-879.

Nex F., Remondino, F., 2014. UAV for 3D mapping applications: a review. *Applied Geomatics*, Vol.6(1), pp. 1-15.

Nex, F, Armenakis, C, Cramer, M, Cucci, DA, Gerke, M, Honkavaara, E, Kukko, A, Persello, C, Skaloud, J., 2022. UAV in the advent of the twenties: where we stand and what is next. *ISPRS J. of Photogrammetry and Remote Sensing*, Vol. 184, pp. 215-242.

Patel., D., Pham, P., Bera, A., 2023. DroNeRF: Real-time multi-agent drone pose optimization for computing Neural Radiance Fields. *arXiv preprint arXiv*: 2303.04322v1.

Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R., 2002. Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation. Proc. *IEEE CVPR.*

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, Vol. 44(3), pp. 1623-1637.

Reiser, C., Szeliski, R., Verbin, D., Srinivasan, P.P., Mildenhall, B., Geiger, A., Barron, J.T. and Hedman, P., 2023. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *arXiv preprint arXiv:2302.12249.*

Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NeRF-based 3D reconstruction. *Remote Sensing*, 15, 3585.

Rothermel, M., Wenzel, K., Fritsch, D., and Haala, N., 2012. Sure: Photogrammetric surface reconstruction from imagery. *Proc. LC3D Workshop*, Berlin, Vol. 8.

Saxena, A., Sun, M., and Ng, A. Y., 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE TPAMI*, 31(5), pp. 824-840.

Schönberger, J. L., Sinha, S. N., and Pollefeys, M., 2018. Learning to fuse proposals from multiple scanline optimizations in semi-global matching. *Proc. ECCV*, pp. 739-755.

Stathopoulou, E.K., Remondino, F., 2023. A survey of conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record*, 10.1111/phor.12456.

Tancik, M., Weber, E., Ng, E.,Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A., 2023. Nerfstudio: A modular framework for Neural Radiance Field development. *Proc. ACM SIGGRAPH*.

Tosi, F., Aleotti, F., Poggi, M., and Mattoccia, S., 2019. Learning monocular depth estimation infusing traditional stereo knowledge. Proc. *CVPR*, pp. 9799-9809.

Turki, H., Ramanan, D., Satyanarayanan, M., 2022. Mega-NeRF: scalable construction of large-scale NeRFs for virtual fly-throughs. Proc. *CVPR*.

Turki, H., Zhang, J., Ferroni, F., Ramanan, D., 2023. SUDS: Scalable Urban Dynamic Scenes. Proc. *CVPR*.

Wang, X., Wang, C., Liu, B., Zhou, X., Zhang, L., Zheng, J. and Bai, X., 2021. Multi-view stereo in the Deep Learning Era: A comprehensive review. *Displays*, 70, p. 102102.

Wang, X., Zhu, Z., Huang, G., Qin, F., Ye, Y., He, Y., Chi, X. and Wang, X., 2022. MVSTER: Epipolar transformer for efficient multi-view stereo. *Proc. ECCV*.

Watson, J., Firman, M., Brostow, G. J., and Turmukhambetov, D., 2019. Self-supervised monocular depth hints. Proc. *CVPR*, pp. 2162-2171.

Welponer, M., Stathopoulou, E.-K., and Remondino, F., 2022. Monocular depth prediction in photogrammetric applications. *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, pp. 469-476.

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D., 2022. BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering. *Proc. ECCV*.

Xu, Q., Tao, W., 2020. PVSnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv*:2007.07714.

Xu, H., Zhou, Z., Qiao, Y., Kang, W., Wu, Q., 2021. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. *Proc. Conf. on Artificial Intelligence*, Vol. 2.

Xu, N., Qin, R., Huang, D., Remondino, F., 2023. Enabling Neural Radiance Fields (NeRF) for Large-scale Aerial Images – A Multi-tiling Approaching and the Geometry Assessment of NeRF. *arXiv:2310.00530.*

Yang, J., Mao, W., Alvarez, J. M., and Liu, M., 2020. Cost volume pyramid based depth inference for multi-view stereo. Proc. *IEEE CVPR*, pp. 4877-4886.

Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L., 2018. MVSnet: Depth inference for unstructured multi-view stereo. Proc. *ECCV*, pp. 767-783.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C., 2021. Learning to recover 3D scene shape from a single image. Proc. *IEEE CVPR*, pp. 204-213.

Yin, W., Liu, Y., Shen, C., and Yan, Y., 2019. Enforcing geometric constraints of virtual normal for depth prediction. Proc. *IEEE CVPR*, pp. 5684-5693.

Yu, D., Ji, S., Liu, J., Wei, S., 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol., 171, pp. 155-170.

Yu, Z., Chen, A., Antic, B., Peng, S. P., Bhattacharyya, A., Niemeyer, M., Tang, S., Sattler, T., Geiger, A., 2022. SDFStudio: A Unified Framework for Surface Reconstruction. Retrieved from https://github.com/autonomousvision/sdfstudio

Zhang, N., Nex, F., Vosselman, G., Kerle, G., 2023. Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation. Proc. *IEEE CVPR*, pp. 18537-18546.

Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T. and Barron, J.T., 2021. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, Vol. 40(6), pp.1-18.

Zhong, Y., Dai, Y., and Li, H., 2017. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv*:1709.00930.

Zhou, K., Meng, X., and Cheng, B., 2020. Review of stereo matching algorithms based on deep learning. *Computational intelligence and neuroscience*.