# EVALUATING MONOCULAR DEPTH ESTIMATION METHODS

N. Padkan[1,2], P. Trybala[1], R. Battisti[1], F. Remondino[1], C. Bergeret[1,3]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: <npadkan><ptrybala><rbattisti><remondino>@fbk.eu
[2] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy
[3] ENSG, Paris, France – Email: Clovis.Bergeret@ensg.eu

**Commission II**

**KEY WORDS:** Monocular Depth, Photogrammetry, Deep Learning, 3D, benchmark

## ABSTRACT

Depth estimation from monocular images has become a prominent focus in photogrammetry and computer vision research. Monocular Depth Estimation (MDE), which involves determining depth from a single RGB image, offers numerous advantages, including applications in simultaneous localization and mapping (SLAM), scene comprehension, 3D modeling, robotics, and autonomous driving. Depth information retrieval becomes especially crucial in situations where other sources like stereo images, optical flow, or point clouds are not available. In contrast to traditional stereo or multi-view methods, MDE techniques require fewer computational resources and smaller datasets. This research work presents a comprehensive analysis and evaluation of some state-of-the-art MDE methods, considering their ability to infer depth information in terrestrial images. The evaluation includes quantitative assessments using ground truth data, including 3D analyses and inference time.

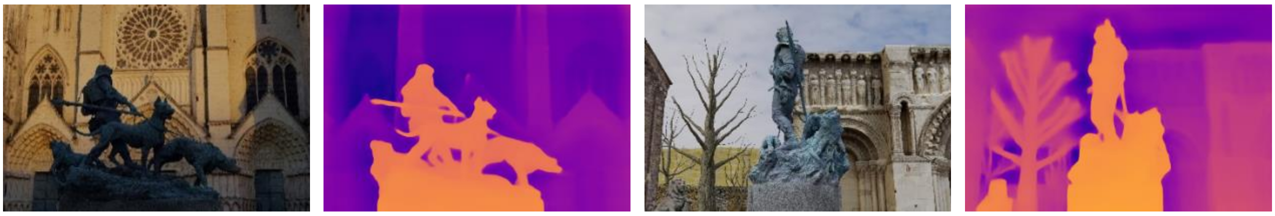

Figure 1: Examples of depth map inferred from single images, commonly known as Monocular Depth Estimation (MDE).

## 1. INTRODUCTION

Estimating depth from a single image, known as Monocular Depth Estimation (MDE, Figure 1), is a fundamental task in the field of computer vision and it is getting also useful in various photogrammetric task. MDE processes hold numerous advantages, being supportive and complementary for simultaneous localization and mapping (SLAM), scene understanding, 3D modelling, robotics, and autonomous driving. The retrieval of depth information becomes especially crucial in scenarios where alternative sources like stereo images, optical flow, or point clouds are unavailable (Bhoi et al, 2019). Traditional depth estimation methods using stereo images or video sequences (Kong and Black, 2015; Cheng and Huang, 2015; Ha et al., 2016; Wei et al., 2022; Hossain and Lin, 2023; Stathopoulou and Remondino, 2023), require extensive computational resources and larger datasets than monocular depth estimation techniques, thus leading to the rise of MDE based on deep learning methods that rely on convolutional neural networks rather than hand-crafted features.

### 1.1 Paper motivations and aims

As MDE has become a key topic of research in the photogrammetric and computer vision communities, a holistic understanding and quantitative evaluation of state-of-the-art methods is necessary. MDE could be a complementary approach to photogrammetric multi-view stereo (MVS) methods or it could support navigation tasks or scene understanding. Therefore, this research work aims to:

- Analyze and test some state-of-the-art methods able to infer depth information from single terrestrial images;
- Perform quantitative evaluations (cloud-to-cloud distances, MAE, RMSE, MTF, etc.) with available ground truth data.

Despite the numerous state-of-the-art depth estimation algorithms, it is important to evaluate their performance across various datasets and environments. Each algorithm is designed with specific assumptions and optimizations, making it challenging to claim that a single algorithm can effectively address all possible applications and scenarios. Therefore, it becomes crucial to test and compare different depth estimation methods to determine their strengths, weaknesses, and suitability for specific applications. By conducting such evaluations, we can gain insights into the applicability and robustness of these algorithms, enabling us to make informed decisions when selecting the most appropriate method for a given task or environment. Other unbiased MDE evaluations are presented in Kock er al. (2020), Dickson et al. (2021), Diab et al. (2022), Marelli et al. (2023), Nex et al. (2023), Theiner et al. (2023).

## 2. MDE METHODS AND EVALUATION

### 2.1 MDE methods

This paper is dedicated to exploring monocular depth estimation (MDE) and thoroughly evaluating specific algorithms that play a crucial role in this area. From various available algorithms like VPD (Zhao et al., 2023), NVDS (Wang et al., 2023), DINOv2 (Oquab et al., 2023), DepthFormer (Li et al., 2023), etc, we have chosen DPT, ZoeDepth, MiDaS, and DenseDepth. The scientific literature and other reputable sources like Paper with Codes have highlighted these algorithms as top performers, driving our selection. By concentrating on these algorithms, our study seeks

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-137-2023 | © Author(s) 2023. CC BY 4.0 License.

to uncover their unique characteristics and contributions, shedding light on their significant roles in the complex realm of monocular depth estimation. It is important to note that in the dynamic landscape of MDE, there is no single algorithm that universally outperforms others across all datasets and scenes. Hence, our study acknowledges the significance of considering and evaluating algorithms based on their application and compatibility with the dataset's images. By concentrating on the chosen algorithms, our study seeks to uncover their unique characteristics and contributions, shedding light on their significant roles in the complex realm of monocular depth estimation.
The selected MDE algorithms are hereafter introduced.

**DenseDepth – High Quality Monocular Depth Estimation via Transfer Learning** (Alhashim and Wonka, 2019): it employs a convolutional neural network to generate a detailed depth map from a single RGB image via transfer learning. Adhering to a conventional encoder-decoder design, it utilizes feature representations from well-performing pre-trained networks to initialize the encoder. Furthermore, it integrates augmentation and training methodologies that enhance the precision of the outcomes. In Figure 2, an overview of the encoder-decoder network utilized for depth estimation is presented. The encoding process involves transforming the input RGB image into a feature vector via the pre-trained DenseNet-169 (Huang et al., 2017) network, which has been trained on ImageNet (Deng et al., 2009). This resultant vector is then channeled through a sequence of successive up-sampling layers (Lehtinen et al. 2018), facilitating the creation of the final depth map at half the initial resolution. The configuration of these up-sampling layers, coupled with their corresponding skip-connections, constitutes the decoder component. The training of DenseDepth models involved 1 million iterations on the NYU Depth V2 dataset and 300,000 iterations on the KITTI dataset.

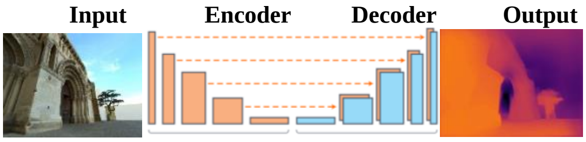

Figure 2: Overview of DenseDepth architecture.

**MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation** (Ranftl et al., 2020): the MiDaS family of models has its roots in a critical study within the realm of relative depth estimation where it is reported the value of dataset integration to attain superior zero-shot performance across diverse datasets. Depth prediction is executed within the realm of disparity space, encompassing inverse depth with considerations for scale and shift. Training employs losses that are invariant to scale and shift to address uncertainties in true depth labels. Existing datasets for depth estimation are amalgamated and enriched with frames and disparity labels extracted from 3D films, thereby creating an extensive meta-dataset. As the evolution of MiDaS releases has unfolded through various iterations, a progressively increasing number of datasets have been integrated over time. The network structure of MiDaS follows a conventional encoder-decoder setup, where the encoder relies on an image-classification network. The original MiDaS v1.0 and v2.0 models utilize the ResNet-based design (He et al., 2016). For the release of MiDaS v3.1, five encoder types are selected based on their perceived potential for downstream tasks. This choice stems from their high quality in depth estimation or their suitability for real-time applications due to low computational requirements. This selection criterion applies equally to the available sizes for encoder types, encompassing both small and large options.

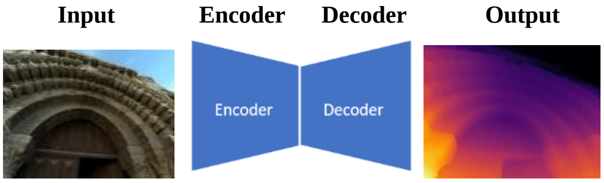

Figure 3: The MiDaS architecture.

**ZoeDepth – Zero-shot Transfer by Combining Relative and Metric Depth** (Bhat et al., 2023): it combines both monocular depth estimation (MDE) and relative depth estimation (RDE) approaches in a two-stage framework (Figure 44). In the first stage, an encoder-decoder structure is trained to estimate relative depths from the input image. This model is trained on a large variety of datasets, which improves its generalization to different scenes and environments. It builds upon the MiDaS (Ranftl et al., 2020) training strategy for relative depth prediction which uses a loss that is invariant to scale and shift. In the second stage, components responsible for estimating metric depth are added as an additional head. This stage helps to refine the depth estimates by incorporating metric depth information, which is the absolute distance between objects in the scene. The ZoeD-M12-NK architecture, employs relative pretraining across 12 distinct datasets, coupled with metric fine-tuning on both indoor and outdoor datasets—namely, NYU Depth v2 (Silberman et al., 2012) and KITTI (Geiger et al., 2013) jointly.

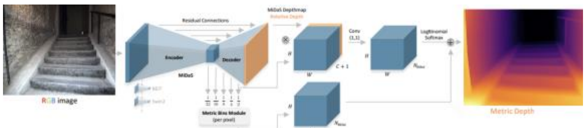

Figure 44: The ZoeDepth architecture (Bhat et al., 2023).

**DPT – Dense Prediction Transformer** (Ranftl et al., 2021): it is an architecture for dense prediction tasks (Figure 5). DPT utilizes an encoder-decoder design, where the encoder incorporates a transformer as the fundamental computational component. The method adopts the Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the foundational architecture, converting its bag-of-words (tokens) representation into image-like feature representations at different resolutions. These representations are then progressively merged using a convolutional decoder to generate the ultimate dense prediction output. DPT-Base, DPT-Large, and DPT-Hybrid are three DPT architectures which are based on ViT architectures (Vit-Base, ViT-Large, and ViT-Hybrid) in terms of reassembling tokens from different layers. ViT-Base employs the patch-based embedding approach and includes 12 transformer layers, while ViT-Large adopts the same embedding technique but incorporates 24 transformer layers and features a broader feature size D (feature dimension of each token). Additionally, ViT-Hybrid utilizes a ResNet50 for calculating the image embedding, and it is followed by the inclusion of 12 transformer layers and is fine-tuned on the KITTI and NYUv2 datasets.

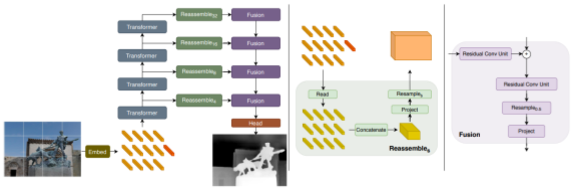

Figure 5: DPT Architecture (Ranftl et al., 2021).

### 2.2 Metrics

For the quantitative evaluation of inferred depth maps, various metrics are commonly employed to assess the disparities between a predicted depth and its corresponding ground truth. These metrics include:

- Root Mean Squared Error (RMSE),
- Mean Absolute Error (MAE),
- Peak Signal-to-Noise Ratio (PSNR) (Johnson, 2006),
- Mean Relative Error (MRE),
- Structural Similarity Index Measure (SSIM) (Brunet et al., 2011)

They all provide effective means to measure the accuracy and quality of the depth estimation process. Valuable insights into the performance of depth prediction algorithms and their alignment with ground truth data can be gained by analyzing these metrics.

#### 2.2.1 Edge sharpness analysis with MTF, ESF and LSF

The Modulation Transfer Function (MTF) serves as a crucial specification in the design of imaging systems to assess system quality (Kohm, 2004). MTF is essentially the normalized magnitude of the Fourier Transform of the point spread function of the imaging system. Alternatively, it characterizes how sinusoidal wave patterns are weakened based on their spatial frequency. Therefore, the MTF quantifies the level of sharpness present in the resulting image after reconstruction (Kohm, 2004). The MTF is a measure derived from the contrast of the image in relation to the contrast of the object (Winston et al., 2005) and is obtained by normalizing the absolute value of the Fourier Transform of the Point Spread Function (PSF) (Williams and Becklund, 2002). One of the approaches for determining MTF is utilizing edges. The edge spread function (ESF) characterizes how the system responds to a high-contrast edge. The Fourier Transform of the LSF's (Line Spread Function) normalized magnitude yields a one-dimensional cross-section within the two-dimensional MTF structure. Alternative techniques exist to directly compute the system's MTF from the ESF, thereby obviating the necessity for differentiation (Tatian, 1965). A prerequisite for determining MTF from edges is the presence of an accurate ESF representation. The slanted edge algorithm capitalizes on the edge's phase shift across the sampling grid to generate an enhanced ESF, termed "super-resolved". In-house scripts are used to compute MTF, ESF and LSF on specific edges in the inferred depth maps and compare them to the GT (Section 4).

## 3. DATASETS

The considered datasets focus mainly on terrestrial architectural outdoor and indoor scenarios, with large depth changes in the scene. Two publicly available datasets were considered and used for evaluating the MDE methods presented in Section 2:

1) ENRICH[1] (Marelli et al., 2023) is a versatile dataset designed for evaluating photogrammetric and computer vision algorithms, including MDE. In contrast to other datasets, ENRICH provides higher resolution images with varying lighting conditions, camera orientations, scales, and fields of view. It comprises three distinct sub-datasets, namely ENRICH-*Aerial*, ENRICH-*Square*, and ENRICH-*Statue*, each showcasing unique characteristics for algorithm testing.
2) ArchDepth[2] (Welponer et al., 2022) constitutes an extensive collection of synthetic RGB images paired with accurate metric depth maps. Comprising approximately 24,000 images, this dataset showcases photorealistic outdoor views featuring primarily historical buildings and squares.

Moreover, images collected with our in-house stereo-vision SLAM-based system GuPho (Torresani et al., 2021; Menna et al., 2022) were also used to apply MDE methods and perform metric evaluations of derived point clouds.

## 4. EXPERIMENTS

The processing and analyses were performed using an Intel E5-1650 @3.2GHz, 32 GB RAM and NVIDIA GeForce GTX 1050 Ti GPU. In the following, comparative analyses of the described algorithms (Section 2) are given.

### 4.1 ENRICH-*Statue* images

ENRICH-*Statue* comprises four sub-sets of images. For our evaluation, we selected the dataset labeled "camera 1" consisting of 50 images captured in partly cloudy weather conditions. Figure 6 present samples of estimated depth maps obtained using ZoeDepth, MiDaS, and DPT algorithms. From a visual standpoint, it is evident that DPT demonstrates superior performance in estimating the depth of nearby objects compared to ZoeDepth while it exhibits a higher likelihood of producing inaccurate depth estimates for far away objects. Table 1 and 2 report the evaluation of the MDE methods using various metrics (Section 2.2), including inference time, on all images of the ENRICH-*Statue* dataset. Following Marelli et al. (2023), we provide evaluation results using a depth threshold suitable as well as non-capped predictions. Such depth cap is used to exclude far away areas belonging, e.g., to the sky. Our assessment shows that ZoeDepth demonstrates the fastest inference on the ENRICH-*Statue* dataset, surpassing the second fastest algorithm (MiDaS) by approximately twice the speed. In terms of metrics, the DPT algorithms (DPT-Large and DPT-Hybrid) outperform both ZoeDepth and MiDaS in most of the metrics, although ZoeDepth exhibits superior performance in terms of SSIM.

The predicted depth maps are also converted into point clouds using the Open3D library (Zhou et al., 2018). Figure 7 shows the resulting shaded point clouds from the viewpoints close to the original camera perspective. Two distinctive cases are presented: the former image, partially containing the sky (upper row) and the latter one, representing the same object from another viewpoint, without any distant object in the frame (bottom row). The results of the ZoeDepth model, although subjectively the best in terms of predicting the relative depth between objects in the foreground and the background, are heavily blurred at the object edges, which is visible as an outline around the objects due to point shading. Those errors heavily influence the metrics for images containing the sky, as ZoeDepth predicts there high and inconsistent values, which also "blend in" the objects due to edge smoothness. The complex geometry of the statue placed in the foreground also resulted in a very noisy 3D reconstruction.

Compared to ZoeDepth, the results of DPT-Hybrid are better in terms of consistency: the noise level at the statue, the façade in the background and the sky is lower and the sharpness of the edges slightly improves. However, the error of predicting the relations between distances from the camera to the foreground and the background are high. In the case of the picture without the sky, a small part of the point cloud is missing due to incorrectly predicting very small depth values. A similar, but more severe case happens for the same area for MiDaS output.

The MiDaS model tackled well the task of predicting the depth of the façades located in the image backgrounds. The network output for mostly flat areas such as walls, as well as for the sky, is coherent, which together with the sharp edges between the

[1] https://github.com/davidemarelli/ENRICH

[2] https://3dom.fbk.eu/projects/on-going/archdepth

foreground and the background, would be helpful for filtering out the depth values belonging to the sky. However, MiDaS created highly noisy point clouds of the statue, which deteriorated its final scores and resulted in highly erroneous points in the middle bottom areas of both analyzed depth maps.

Finally, an edge sharpness analysis was also used to assess the achieved details of estimated depths by ZoeDepth, DPT, and MiDaS. The aim is to assess the edge quality within the estimated depth and determine whether it effectively conveys the intricate details of objects in the scene. A specific area from the ground truth of an image within the ENRICH-Statue dataset and from the corresponding inferred depth maps was chosen (Figure 8a) to compute the ESF, LSF and MTF curves for ZoeDepth, DPT, and MiDaS methods (Figure 8b-d). As an indicative metric, the length of a 10% – 90% edge transition interval, i.e., edge width in pixels, was computed from ESF plots. Results for the ground truth depth image are also included for reference as almost perfect target scores. All graphs indicate clearly that the worst sharpness was achieved by ZoeDepth output. Its edge width reached 16 px, compared to the GT baseline of just 0.3 px. Both DPT and MiDaS models achieved similar edge widths (7.1 px and 6.3 px, respectively). Even if they clearly outperform ZoeDepth in this aspect, their curves are still far from the ground truth, especially in the case of the MTF.

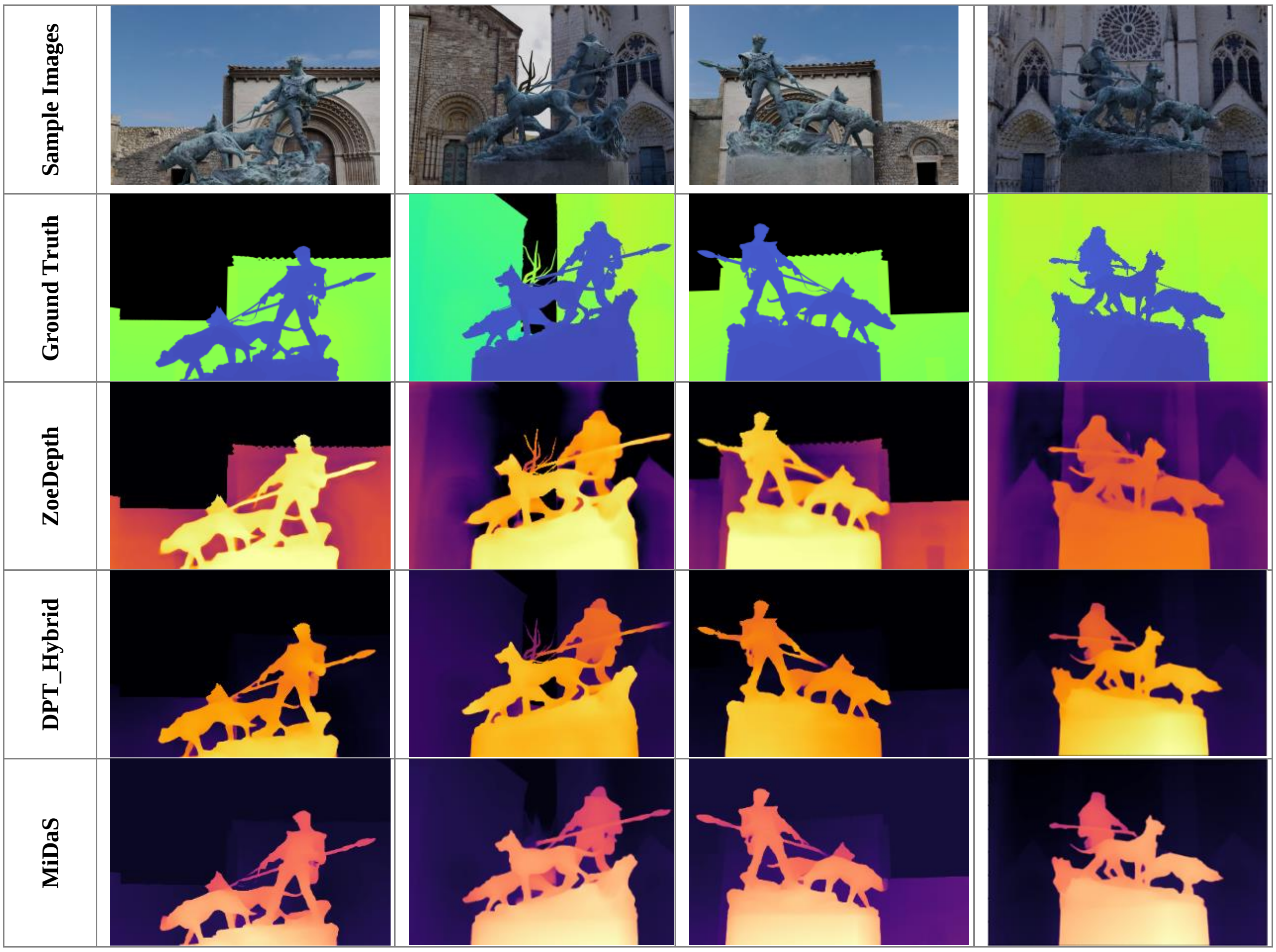

Figure 6: Examples of estimated depth maps using ZoeDepth, DPT-Hybrid, and MiDaS on the ENRICH-*Statue* images.

| | **Inference Time (s)** | **MAE (m) ↓** | **MRE (m) ↓** | **RMSE (m) ↓** | **PSNR ↑** | **SSIM ↓** |
|---|---|---|---|---|---|---|
| **ZoeDepth** | 1.996 | 22.647 | 0.532 | 35.056 | 25.646 | 0.616 |
| **DPT_Hybrid** | 6.126 | 7.555 | 0.368 | 20.250 | 33.375 | 0.790 |
| **DPT_Large** | 9.299 | 7.253 | 0.372 | 20.491 | 33.104 | 0.802 |
| **MiDaS** | 4.490 | 20.133 | 1.728 | 53.456 | 23.125 | 0.699 |

Table 1: Inference time and different metrics computed on the entire ENRICH-*Statue* dataset (no depth limit).

| | **Inference Time (s)** | **MAE (m) ↓** | **MRE (m) ↓** | **RMSE (m) ↓** | **PSNR ↑** | **SSIM ↓** |
|---|---|---|---|---|---|---|
| **ZoeDepth** | 1.996 | 20.591 | 0.530 | 24.598 | 27.839 | 0.619 |
| **DPT_Hybrid** | 6.126 | 5.559 | 0.365 | 6.797 | 39.0622 | 0.795 |
| **DPT_Large** | 9.299 | 5.440 | 0.369 | 6.843 | 38.977 | 0.806 |
| **MiDaS** | 4.490 | 13.939 | 1.736 | 18.796 | 30.158 | 0.707 |

Table 2: Inference time and different metrics computed on the entire ENRICH-*Statue* dataset (depth limited to 70 m).

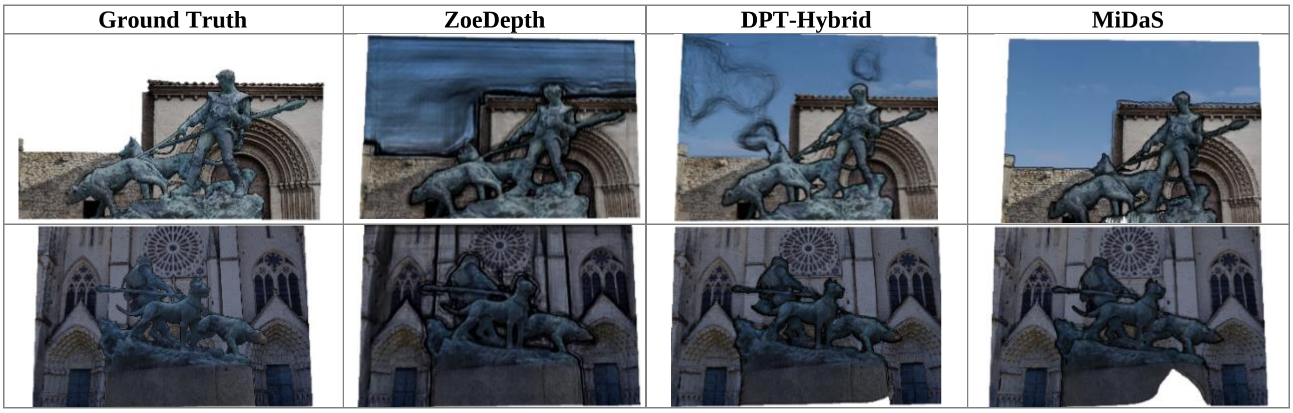

Figure 7: Visual comparison of the generated point clouds from the considered MDE methods with respect to the available GT for the ENRICH-*Statue* dataset.

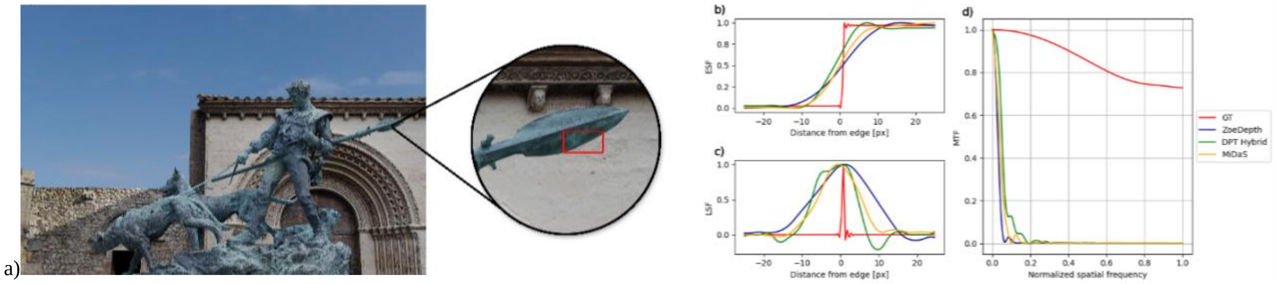

Figure 8: Edge sharpness analysis for an ENRICH-*Statue* image: the region of interest marked in red (a) and the results of ESF (b), LSF (c) and MTF (d) methods on depth maps estimated with ZoeDepth, DPT and MiDaS.

### 4.2 ArchDepth images

Within the repository, the Piazza subset includes 9,000 pairs of RGB images and depth maps, organized based on cardinal directions and rendering cameras. The Piazza subset comprises four smaller datasets, named *CityE, CityN, CityS,* and *CityW*. We specifically utilized the *CityW* dataset (Folder cam6 - 640x480 px, pinhole camera, 36mm sensor size, 20mm focal length), which contains 800 pairs of RGB images and depth maps. ZoeDepth, DPT, MiDaS, and DenseDepth algorithms are evaluated and Figure 9 and Table 3 display visuals and metrics. Once again, it is notable ZoeDepth's remarkable inference speed. Despite the DPT algorithm's superior performance in MAE, MRE, RMSE, and PSNR, it lags behind in terms of depth estimation speed per image, taking approximately 3x more time. In the terms of SSIM, DenseDepth outperforms the other three algorithms, while in terms of MAE, PSNR, and MRE, it exhibits superior performance compared to MiDaS and ZoeDepth.

Figure 10 reports the edge sharpness analysis on an image (and predicted depth maps) from the ArchDepth dataset, which is of much lower resolution (640 x 480 px). The resulting edge widths for ZoeDepth, DenseDepth, DPT_Hybrid and MiDaS reached 5.2 px, 3.4 px, 1.2 px and 1.0 px, respectively, with the ground truth baseline of 0.1 px. Both DenseDepth and ZoeDepth failed to create sharp depth edges. Similarly, to the analysis presented for the ENRICH dataset (Figure 8), MiDaS and DPT achieved reasonable and comparable results, but still severely less sharp compared to the ground truth.

### 4.3 GuPho images

Further experiments are conducted using indoor images collected with the FBK GuPho system (Torresani et al., 2021; Menna et al., 2022) and the ZoeDepth method (Figure 11). Inferred depth maps are converted into point clouds using the Open3D library (Zhou et al., 2018) and, given the flat indoor surfaces, best fitting analyses are performed for floor, ceiling, and wall areas. Metrics are reported in Table 4, highlighting few cm errors from predicted and ground truth geometries.

| | **Inference Time (s)** | **MAE (m) ↓** | **MRE (m) ↓** | **RMSE (m) ↓** | **PSNR ↑** | **SSIM ↓** |
|---|---|---|---|---|---|---|
| **ZoeDepth** | 1.421 | 5.710 | 1.339 | 6.154 | 40.185 | 0.551 |
| **DPT_Hybrid** | 4.106 | 2.735 | 0.630 | 4.358 | 44.016 | 0.640 |
| **DenseDepth** | 1.846 | 3.157 | 0.710 | 5.050 | 42.844 | 0.320 |
| **MiDaS** | 1.473 | 3.244 | 0.798 | 4.821 | 42.747 | 0.446 |

Table 3: Different metrics computed on the entire *CityW* section of the ArchDepth dataset.

| | **Case 1 - RMSE [mm]** | **Case 2 - RMSE [mm]** | **Case 3 - RMSE [mm]** |
|---|---|---|---|
| **Floor** | 9 | 22 | 9 |
| **Ceiling** | 19 | 25 | 25 |
| **Right Wall** | 9 | 67 | 36 |
| **Left Wall** | 21 | 49 | 49 |
| **Mean** | 14 | 41 | 30 |

Table 4: Metrics for the ZoeDepth MDE on the GuPho images (Figure 11): the RMSE (σ) was calculated individually for the floor, ceiling, right wall, and left wall. Additionally, the average (Mean) of RMSE was also determined as the part of the analysis.

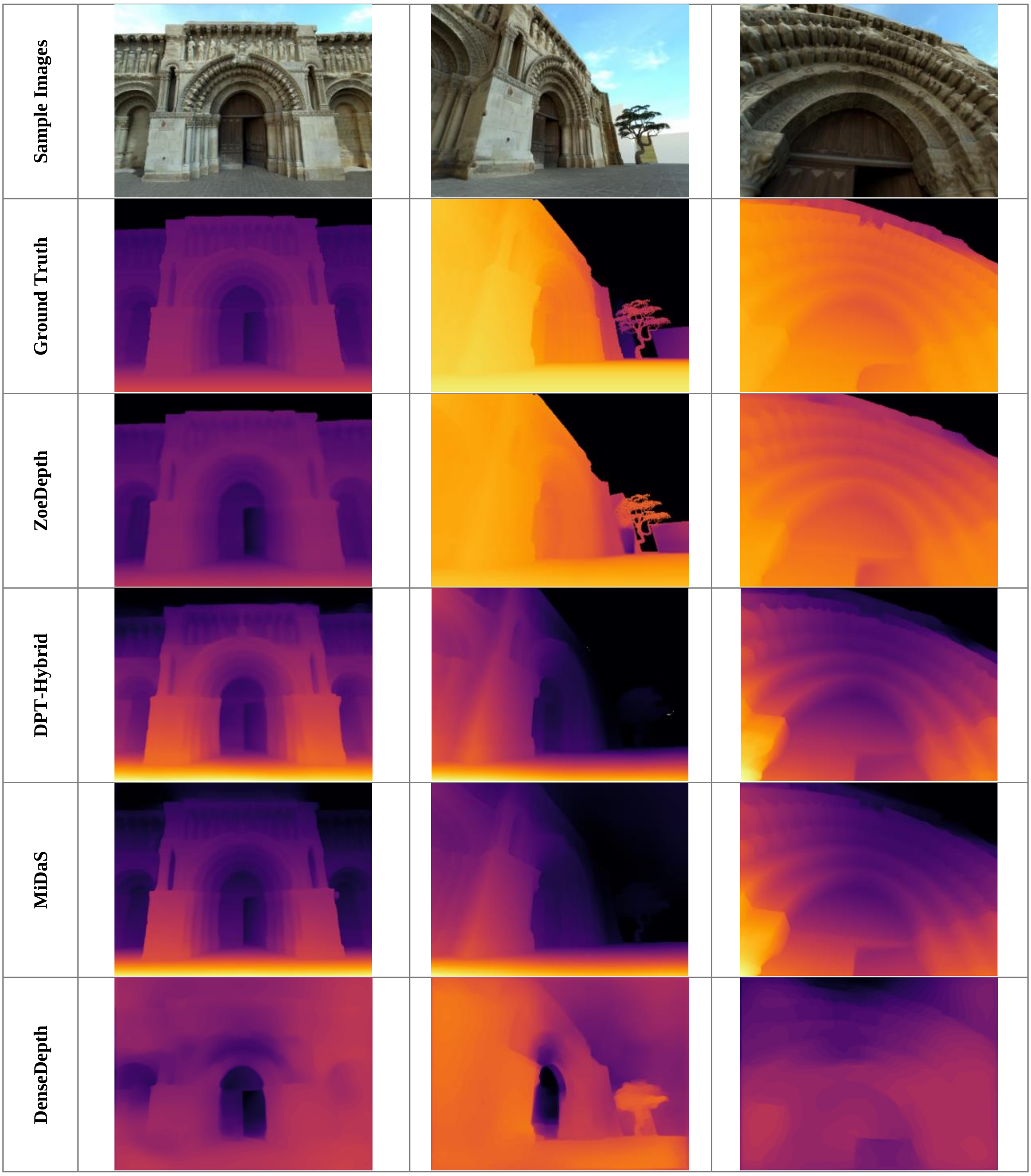

Figure 9: Examples of estimated depth on single images of the ArchDepth dataset using ZoeDepth, DPT, MiDaS, and DenseDepth.

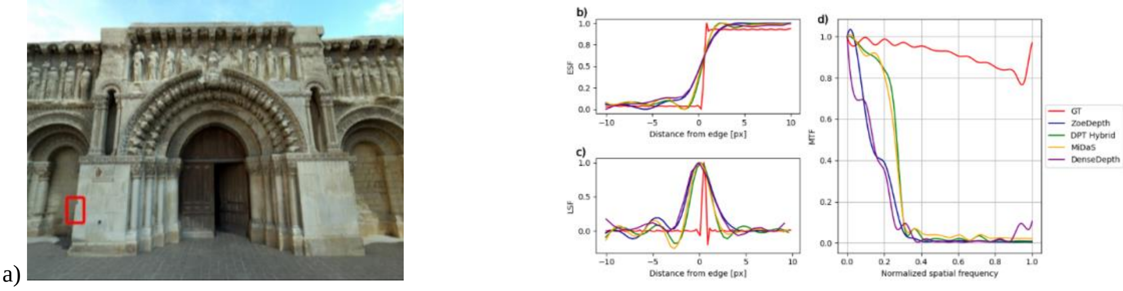

Figure 10: Edge sharpness analysis for an ArchDepth building facade image: the region of interest marked in red (a) and the results of ESF (b), LSF (c) and MTF (d) methods on depth maps estimated with ZoeDepth, DenseDepth, DPT and MiDaS.

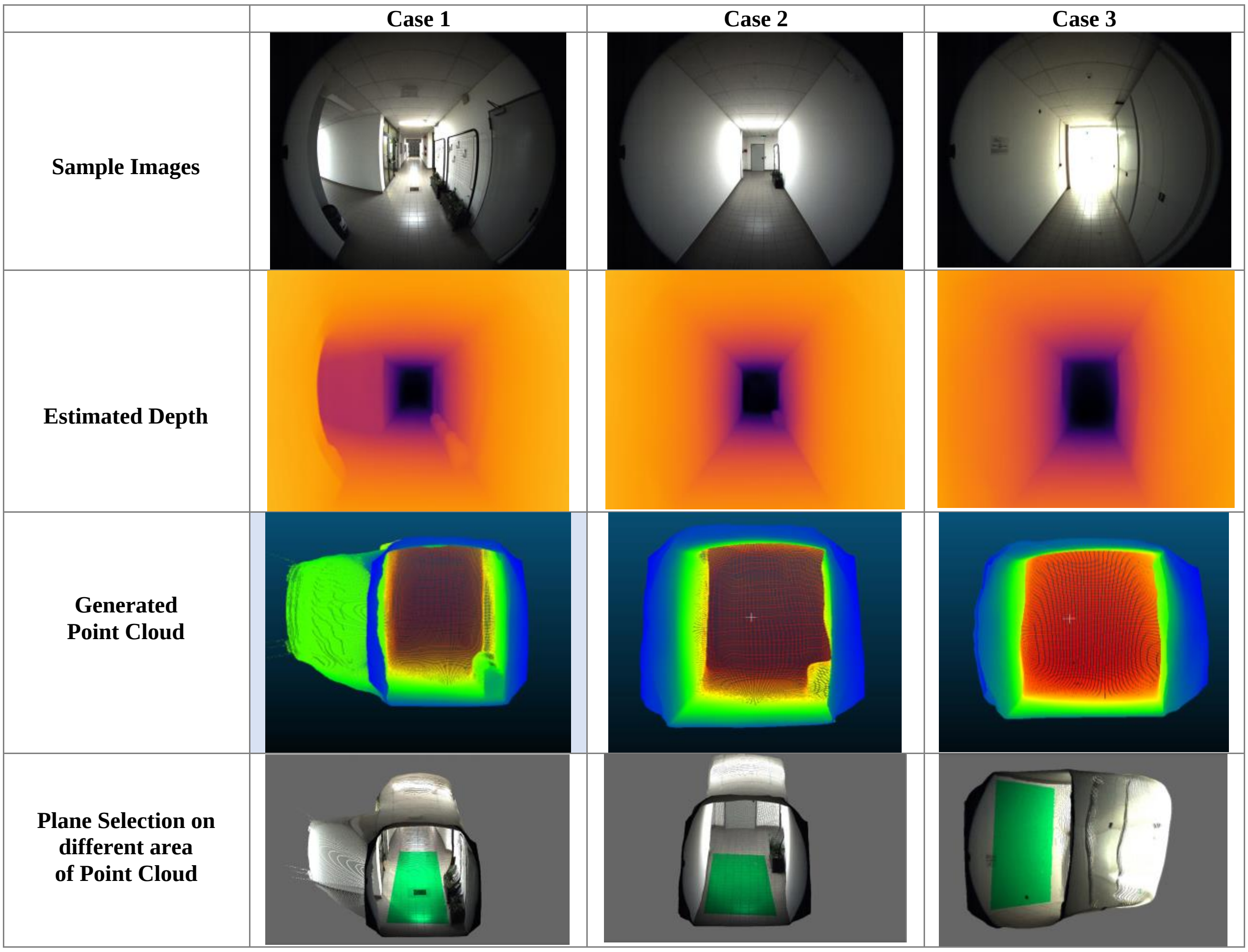

Figure 11: Visual impressions and comparisons (Table 4) for the recovered depths and point clouds derived from ZoeDepth on the GuPho dataset.

## 5. CONCLUSIONS

The crucial role of MDE methods and performances on some datasets were presented. MDE is of interest in various fields such as scene comprehension, 3D modeling, robotics, and autonomous driving. MDE is becoming more and more popular due to their reliability and deep learning performances. This study has provided an analysis and evaluation of some MDE methods using different metrics such as inference time, MAE, MRE, RMSE, PSNR, and SSIM. According to our evaluating, defining an algorithm as a winner is impossible. Determining a singular optimal algorithm exhibiting superior performance across all datasets proves to be a challenging task. This challenge arises due to the nuanced influence of application, training and varying scenarios on the algorithm process. When considering the aspect of time, ZoeDepth demonstrated the quickest inference time compared to the other algorithms. While when we consider different metrics, DPT and MiDaS had better performances.

In our view, the majority of existing MDE algorithms suffer from a limitation in training diversity, specifically in terms of various areas and image types. For instance, when the training dataset lacks a sufficient number of images featuring the sky, it can lead to challenges in accurately estimating depth in those particular regions.

In future work, our focus will be on the integration of MDE into the GuPho V-SLAM system to achieve real-time depth prediction. By combining the capabilities of V-SLAM with MDE, we aim to enhance the robustness and completeness of 3D reconstruction e.g., in textureless environments. Additionally, our research will explore alternative methods for MDE, considering various deep learning architectures and traditional computer vision approaches.

## ACKNOWLEDGMENTS

This study was partially carried out within the Interconnected Nord-Est Innovation Ecosystem (iNEST) and received funding from the European Union Next-GenerationEU (Piano Nazionale Di Ripresa e Resilienza (PNRR) – Missione 4, Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS00000043). The work is also partly supported by the project “AI@TN” funded by the Autonomous Province of Trento, Italy.

## REFERENCES

Alhashim, I. and Wonka, P., 2018. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv*:1812.11941.

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

Brunet, D., Vrscay, E.R. and Wang, Z., 2011. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, *21*(4), pp.1488-1499.

Bhoi, A., 2019. Monocular depth estimation: A survey. *arXiv preprint arXiv:*1901.09402.

Cheng, F.H., Huang, K.Y., 2015. Real-time stereo matching for depth estimation using GPU. *Proc. IEEE UMEDIA*.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. Proc. *IEEE CVPR*, pp. 248-255.

Diab, A., Sabry, M., El Mougy, A., 2022. Comparing Monocular Camera Depth Estimation Models for Real-time Applications. Proc. *ICAART*, pp. 673-680.

Dickson, A., Knott, A., Zollman, S., 2021. Benchmarking Monocular Depth Estimation Models for VR Content Creation from a User Perspective. Proc. 36th *IVCNZ.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv*:2010.11929.

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11), pp.1231-1237.

Ha, H., Im, S., Park, J., Jeon, H.-G, Kweon, I.S., 2016. High Quality Depth from Uncalibrated Small Motion Clip. *Proc. IEEE CVPR.*

Hossain, S. and Lin, X., 2023. Efficient Stereo Depth Estimation for Pseudo-LiDAR: A Self-Supervised Approach Based on Multi-Input ResNet Encoder. *Sensors*, *23*(3), p.1650.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. Proc. *IEEE CVPR*, pp. 770-778.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. *IEEE CVPR*, pp. 4700-4708).

Johnson, D.H., 2006. Signal-to-noise ratio. *Scholarpedia*, *1*(12), p.2088.

Koch, T., Liebel, L., Körner, M., Fraundorfer, F., 2020. Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. *Computer Vision and Image Understanding*, Vol. 191, 102877.

Kong, N. and Black, M.J., 2015. Intrinsic depth: Improving depth transfer with intrinsic images. *Proc. ICCV*.

Kohm, K., 2004. Modulation transfer function measurement method and results for the Orbview-3 high resolution imaging satellite. *Int. Arch. Photogramm. Remote Sens.,* Vol. 35, pp. 12-23.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M. and Aila, T., 2018. Noise2Noise: Learning image restoration without clean data. *arXiv preprint arXiv*:1803.04189.

Li, Z., Chen, Z., Liu, X. and Jiang, J., 2023. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, pp.1-18.

Marelli, D., Morelli, L., Farella, E.M., Bianco, S., Ciocca, G., Remondino, F., 2023. ENRICH: Multi-purposE dataset for benchmarking, In Computer vision and pHotogrammetry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198, pp.84-98.

Menna, F., Torresani, A., Battisti, R., Nocerino, E., Remondino, F., 2022. A modular and low-cost portable VSLAM system for real-time 3D mapping: from indoor and outdoor spaces to underwater environments. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-2/W1-2022, 153–162.

Nex, F., Zhang, N., Remondino, F., Farella, E.M., Qin, R., Zhang, C., 2023. Benchmarking the extraction of 3D geometry from UAV images with deep learning methods. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W3-2023.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. and Assran, M., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv*:2304.07193.

Padkan, N., Battisti, R., Menna, F., Remondino, F., 2023. Deep learning to support 3D mapping capabilities of a portable VSLAM-based system. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W1-2023, 363–370.

Ranftl, R., Bochkovskiy, A. and Koltun, V., 2021. Vision transformers for dense prediction. *Proc. ICCV*, pp. 12179-12188.

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*, 44(3), pp.1623-1637.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBd images. *Proc. ECCV*, pp. 746-760.

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

Torresani, A., Menna, F., Battisti, R., Remondino, F., 2021. A V-SLAM Guided and Portable System for Photogrammetric Applications. *Remote Sensing*, Vol.13(12), 2351.

Tatian, B., 1965. Method for obtaining the transfer function from the edge response function. *JOSA*, 55(8), pp.1014-1019.

Theiner, J., Nommensen, Rhotert, J., Springstein, M., Müller-Budack, E., Ewerth, R., 2023. Analyzing Results of Depth Estimation Models With Monocular Criteria. Proc. *CVPR*.

Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., Xian, K. and Lin, G., 2023. Neural Video Depth Stabilizer. *arXiv preprint arXiv*:2307.08695.

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

Wei, K., Kuno, Y., Arai, M. and Amano, H., 2022. Rt-libsgm: An implementation of a real-time stereo matching system on FPGA. In *International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies*, pp. 1-9.

Winston, R., Miñano, J.C. and Benitez, P.G., 2005. *Non-imaging optics*. Elsevier.

Williams, C.S. and Becklund, O.A., 2002. Introduction to the optical transfer function, Vol. 112. *SPIE Press.*

Zhou, Q.Y., Park, J. and Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:*1801.09847.

Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J. and Lu, J., 2023. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153.*