Advancing Offshore Safety: Monocular Depth Estimation from 360-Degree Images for Enhanced Oil Platform Inspection

Jorge Andres Chamorro Martinez[1], Robinson Garcia[2], Patrick Nigri Happ[1], Daliana Lobo Torres[1], Pedro Pereira Guedes[1], Romeu Fereira de Oliveira[1], Raul Queiroz Feitosa[1], Michelle Facina[2], Maikon Bressani[2], Pedro Soto Vega[3]

[1] Pontifical Catholic University of Rio de Janeiro, Gávea, Rio de Janeiro, Brazil - (cmjorgeandres, patrickhapp, romeu)@tecgraf.com.br
[2] Petrobras, Rio de Janeiro, Brazil
[3] Vision-AD and AutoRob, LabISEN, ISEN Yncréa Ouest, Brest, France

Keywords: Depth Estimation, Offshore Oil Platforms, 360 Imagery, Foundation Models, Depth Anything V2, ZoeDepth, Metric3Dv2, Patchfusion

Abstract

Offshore oil platforms are critical infrastructures that require regular inspection to detect corrosion and maintain structural integrity. While traditional manual inspections are prone to human bias and high operational costs, recent advancements in automated inspection using 360-degree imagery have shown promise. This study presents a comprehensive evaluation of state-of-the-art metric monocular depth estimation methods—Depth Anything V2, ZoeDepth, Metric3Dv2, and Patchfusion—applied to 360-degree images of offshore oil platforms, a novel application in this domain. Metric depth estimation may also benefit downstream tasks such as corrosion and object detection by providing additional spatial context. Our comparative analysis assesses the performance and suitability of these methods in the context of the unique visual characteristics of offshore industrial environments and panoramic imagery. The findings offer valuable insights into the limitations and strengths of current approaches and serve as a basis for future work aimed at improving depth estimates, including domain-specific fine-tuning. This work contributes to ongoing efforts to enhance the efficiency, accuracy, and safety of structural health monitoring in challenging industrial settings. Code is available at `https://github.com/DiMorten/depth_offshore_LAGIRS2025`.

1. Introduction

Corrosion detection plays a crucial role in maintaining the integrity and safety of offshore oil platforms. Traditionally, this process has relied on manual inspections conducted by expert personnel. However, this approach presents several challenges, including high operational costs, potential inconsistencies due to human bias, and risks to the physical safety of inspection workers. Recent advancements have explored automatic corrosion detection methods using 360-degree imagery, as demonstrated by (Garcia et al., 2021). These approaches aim to automate inspections and improve consistency while reducing both costs and safety risks.

Complementary to corrosion detection, object detection techniques have also gained prominence in industrial inspection pipelines. They enable the identification and localization of specific components, such as pipes, valves, and tanks, on which corrosion levels can be assessed more systematically. However, both corrosion and object detection can benefit significantly from geometric context, particularly in complex and cluttered industrial environments.

In this context, understanding metric depth from 360-degree imagery can provide valuable spatial information to support these tasks. Accurate depth maps enable the quantification of object dimensions, separation of foreground from background, and better integration of multi-view data, all of which can lead to more robust and interpretable inspection outcomes. While active depth sensing technologies such as LiDAR offer high accuracy, they are often costly, require specialized equipment, and may not be practical for frequent data collection in operational offshore settings. In contrast, 360-degree imagery can be passively captured by personnel during routine inspections, offering a more accessible and scalable solution.

In this work, we conducted a comparative study aimed at evaluating the suitability of various state-of-the-art monocular depth estimation methods for generating metric depth maps from 360-degree images captured in offshore environments. These environments pose distinct challenges for depth estimation, including corroded surfaces, low-texture regions, non-standard viewpoints, and complex metallic structures that can obscure traditional depth cues.

To the best of our knowledge, this research represents the first comprehensive assessment of metric monocular depth estimation techniques specifically tailored for offshore oil platforms using 360-degree images. The evaluated models include Depth Anything V2, ZoeDepth, Metric3Dv2, and Patchfusion. The contributions of this work are the following:

- A comprehensive comparative assessment of state-of-the-art monocular depth estimation methods applied to 360-degree images of offshore platforms.
- Insights into the challenges and opportunities of adapting these techniques for the unique visual characteristics of offshore industrial environments.
- An assessment of the variability of the depth outcomes for different depth value ranges.

Section 2 reviews related work. Section 3 outlines the evaluated models. Section 4 details the dataset and experimental protocol.

Section 5 presents the results and discussion. Finally, Section 6 presents the conclusions.

2. Related works

Early approaches to monocular depth estimation relied on hand-crafted features and geometric principles. Structure from Motion (SfM) reconstructed 3D structure by analyzing motion across multiple 2D frames (Hartley and Zisserman, 2003, Snavely et al., 2006). However, SfM required multiple images from a moving camera, struggled with static scenes, and produced depth maps only up to an arbitrary scale. Regarding metric depth estimation, its accuracy tended to be insufficient and inconsistent across the overlapping views (Pataki et al., 2025). Other classical techniques included shape from shading, which inferred depth from lighting and shadows (Horn, 1989), but relied heavily on assumptions about lighting and surface reflectance. Depth from defocus estimated depth by analyzing blur across images with varying focus (Favaro and Soatto, 2005), requiring multiple captures and being limited by optical constraints. Probabilistic models, such as Markov Random Fields, used spatial relationships to infer depth (Szeliski and Golland, 1999), but were dependent on hand-crafted features and often computationally expensive.

Fully convolutional networks (FCNs) enabled end-to-end monocular depth estimation by predicting dense depth maps from a single image using large labeled datasets (Eigen et al., 2014, Laina et al., 2016, Li et al., 2015, Fu et al., 2018). FCNs eliminated the need for hand-crafted features and geometric assumptions, efficiently capturing multi-scale spatial information. However, their reliance on local receptive fields limited their ability to model global scene structure.

To overcome this, self-attention transformers were introduced, offering improved performance by capturing long-range dependencies across the image (Ranftl et al., 2021, Bhat et al., 2021,Chang et al., 2021,Bhat et al., 2022). Unlike FCNs, transformers weigh relationships between all pixels, enhancing the understanding of large-scale geometry.

Depth Prediction Transformer (DPT) (Ranftl et al., 2021) introduced a novel architecture for dense prediction tasks, including monocular depth estimation, based on Vision Transformers (ViT). Unlike convolutional approaches that downsample spatial resolution early in the network, DPT preserves high-resolution representations by dividing the input image into non-overlapping patches and embedding them into tokens, which are processed globally using self-attention mechanisms. The encoder captures long-range dependencies and global context, while the decoder reassembles these features into spatial maps using multi-scale feature fusion.

AdaBins introduces a transformer-based model for monocular depth estimation, using self-attention to adaptively learn depth binning, improving depth accuracy by better handling global context (Bhat et al., 2021). An improvement of this technique called LocalBins uses all layers of the decoder to predict depth distributions instead of using only the end of the decoder (Bhat et al., 2022).

Depth Anything (DA) is a foundation model for monocular depth estimation that integrates large-scale supervised and self-supervised training to support both relative and metric depth prediction (Yang et al., 2024a). In this study, we evaluate its enhanced version, Depth Anything V2 (DAv2) (Yang et al., 2024b), alongside other state-of-the-art methods. DA employs a teacher–student distillation framework and incorporates semantic consistency losses to improve generalization across diverse indoor and outdoor environments. Our study differs from previous work by assessing DAv2 on 360-degree imagery of offshore platforms, a setting not considered in the original publications.

3. Evaluated Models

In this section, we briefly describe the four state-of-the-art monocular depth estimation methods evaluated in this study: Depth Anything V2 (DAv2), ZoeDepth, Metric3Dv2, and Patchfusion.

3.1 Depth Anything V2 (DAv2)

DAv2 (Yang et al., 2024b) is a foundation model for monocular depth estimation. It improves upon the original DA (Yang et al., 2024a) by relying exclusively on synthetic data for labeled supervision. This approach eliminates the depth noise typically found in real-world datasets, which can arise from factors such as reflections, occlusions, and sensor inaccuracies. High-fidelity synthetic labels provide more stable and accurate supervision during training.

The model adopts a semi-supervised teacher-student framework: a larger teacher trained on synthetic data generates pseudo-depth labels for unlabeled real images, which are then used to train a student model. Strong data augmentations (e.g., color jitter, blur, spatial distortions) are applied during student training to improve generalization. Additionally, a feature alignment loss based on DINOv2 (Oquab et al., 2023) semantic features promotes depth consistency across semantically similar regions.

DAv2 supports both relative and metric depth estimation. It is initially trained on relative depth and then fine-tuned on synthetic datasets with metric annotations (Hypersim (Roberts et al., 2021) for indoor scenes and Virtual KITTI (Cabon et al., 2020) for outdoor) following the ZoeDepth (Bhat et al., 2023) protocol to produce outputs in real-world units.

3.2 ZoeDepth

ZoeDepth (Bhat et al., 2023) is a monocular depth estimation framework designed to output depth in metric units, with strong generalization across indoor and outdoor environments. It follows the MiDaS protocol (Ranftl et al., 2020), consisting of two stages: (1) pre-training an encoder-decoder architecture for relative depth on 12 heterogeneous datasets, and (2) fine-tuning the entire network for metric depth on two standardized datasets: NYUv2 for indoor scenes and KITTI for outdoor scenes. This two-step training approach enables the model to leverage the diversity of relative data while achieving scale-aware metric predictions. The backbone of ZoeDepth can be selected from a variety of options including BEiT, Swin2, and DPT, with DPT (Ranftl et al., 2021) being reported as a good trade-off between FPS and accuracy.

ZoeDepth introduces a metric bins module inspired by AdaBins (Bhat et al., 2021), framing depth estimation as a classification task over a set of discrete depth bins. The depth range is divided into N_{bins} intervals (e.g., 64), and the network outputs

a softmax distribution over these bins for each pixel. The final depth is computed as a weighted sum of the bin centers:

$$\text{depth}(x, y) = \sum_{i=1}^{N_{\text{bins}}} p_i(x, y) \cdot c_i(x, y)$$

where p_i is the predicted probability for bin i and c_i is the dynamically generated center value for that bin. To generate adaptive bin centers per pixel, ZoeDepth leverages a hierarchy of MLPs operating on multi-scale features. The coarsest feature map (1/32 resolution) predicts the base bin centers, while progressively finer features produce *attractor points* that perturb these centers, enhancing local adaptation.

The probabilities p_i are computed using a LogBinomial Softmax function, which encourages smoother transitions between neighboring bins compared to traditional softmax, reflecting the continuous nature of depth. This representation is particularly effective for capturing non-linear depth behaviors.

3.3 Metric3Dv2

Metric3Dv2 (Yin et al., 2023) is a state-of-the-art zero-shot monocular depth estimation method that predicts metric-scale depth across diverse image domains and camera models. Traditional approaches often rely on fixed camera intrinsics or adopt affine-invariant strategies (e.g., MiDaS), which generalize well but discard metric scale. Metric3Dv2 resolves this trade-off using a Canonical Camera Space Transformation Module (CSTM), which standardizes input data during training by projecting all samples into a shared canonical camera space. After prediction, a de-canonical transformation recovers the original metric scale using the image's focal length.

The model is trained on over 8 million images from 11 (real-world and synthetic) datasets and 10,000+ cameras, allowing it to learn to decouple scene geometry from camera intrinsics and avoid overfitting to dataset-specific scale cues. This large-scale, mixed-camera training enables strong generalization in zero-shot settings.

At inference, Metric3Dv2 requires the focal length to output depth in real-world units. While this is typically available in curated datasets or SLAM applications, a default value can be used for web or in-the-wild images, with some loss in absolute accuracy.

The architecture employs a ViT encoder and a combination of loss functions, including scale-invariant log loss and patch-based local normalization, to enforce both metric consistency and fine-grained spatial coherence. These design choices enable Metric3Dv2 to produce accurate and generalizable depth predictions directly in metric units.

3.4 Patchfusion

Most foundational models face spatial resolution limitations due to computational constraints, making them unsuitable for high-resolution inputs from modern cameras (e.g., 4K). Patchfusion (Li et al., 2024) introduces a tile-based framework to address this issue, using a custom loss function that improves the fusion of high-level and patch-level features by enforcing consistency across overlapping training patches. PatchFusion is compatible with various depth estimation models (e.g., ZoeDepth, DA). It trains three networks: (1) a global coarse estimator that takes the downsampled full image and outputs D_c; (2) a local fine-grained estimator that processes cropped patches and outputs D_f; and (3) a fusion network that takes as input the cropped RGB patch, the corresponding region from the coarse depth map $roi(D_c)$, and the fine-grained depth D_f. During inference, the image is divided into non-overlapping patches, with optional additional patches generated by shifting patch locations by half the patch size or sampling random positions.

4. Experiments

4.1 Dataset

The dataset used in this study comprises a total of 500 photospheres, which are 360-degree panoramic images capturing the entire surrounding scene from a single viewpoint. Both the RGB photospheres and the corresponding depth maps were generated from LiDAR point clouds captured with the Leica BLK 360 sensor, using the pipeline of (Zang et al., 2022), which converts raw point clouds into equirectangular panoramic images by projecting 3D points onto the spherical image plane and encoding per-pixel depth in meters.

These images were captured aboard two offshore oil platforms, with 250 photospheres collected per platform. Each photosphere was subsequently converted into a cubemap representation, producing six perspective images per photosphere. This transformation resulted in 1,500 images per platform, for a total of 3,000 images across the dataset. The image and reference resolution is 1344×1344 pixels, corresponding to the cubemap representation. The two platforms evaluated in this work are referred to as Platform A (PA) and Platform B (PB).

As shown in Figure 1, the majority of reference depth values are concentrated within the 0 to 10-meter range, accounting for 95.5% at site PA and 99.5% at site PB.

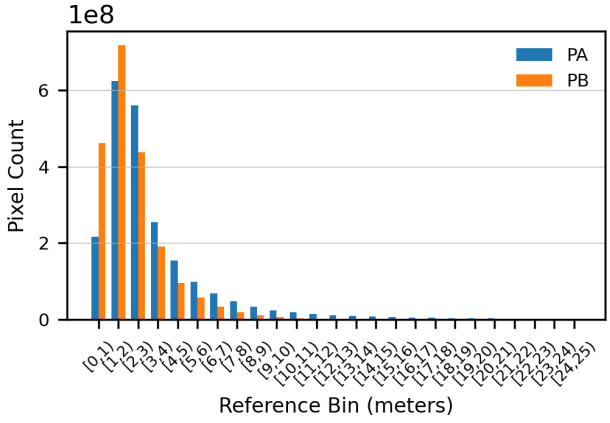

Figure 1. Sample distribution by reference value bins, for PA and PB sites.

4.2 Experimental Protocol

We evaluated 4 foundational monocular depth estimation models on the dataset described in Section 4.1. Each model was tested independently, using images from Platforms PA and PB, to assess their generalization performance across different deployment scenarios. Given that the target application involves both indoor and outdoor environments, we evaluated models

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-X-3-W4-2025-71-2026 | © Author(s) 2026. CC BY 4.0 License.

pre-trained on datasets representative of each setting. Specifically, for DAv2, we assessed versions pre-trained on indoor (Hypersim) and outdoor (vKitti) datasets. Similarly, for ZoeDepth, we evaluated three pre-trained variants: one trained on indoor data (NYUv2), one on outdoor data (Kitti), and one on a combination of both (NYUv2+Kitti).

For the DAv2 model, the predicted depth values (originally normalized in the range $[0, 1]$) were rescaled by multiplying them with a scalar factor max_depth. We empirically set $max\_depth = 10m$ for the indoor model and $max\_depth = 17.5m$ for the outdoor model. These values were chosen based on a grid search over plausible depth ranges, selecting the ones that yielded the best overall performance. For the Patchfusion model, each image is initially divided into a 4×4 grid, yielding 16 non-overlapping patches. Additional patches were generated by shifting the original patch positions horizontally and vertically by half the patch size. Further, randomly positioned patches were added, resulting in a total of 128 patches per inference. This configuration is referred to by the authors as `r128`. The Patchfusion strategy was tested with the ZoeDepth and Depth Anything (DA) base models.

Model performance was assessed using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Additionally, thresholded accuracy metrics δ_1, δ_2, and δ_3 indicate the percentage of predictions within a given factor of the ground truth (GT). Specifically, $\delta_1 = 1.25$ reflects predictions within ±25% of GT, $\delta_2 = 1.25^2$ within ±56%, and $\delta_3 = 1.25^3$ within ±95%. Thus, δ_1 is the most rigorous, followed by δ_2 and then δ_3.

In addition to these metrics, we include a binned comparison of predicted depth values. Specifically, we group the predictions into bins and compute the corresponding mean reference depth within each bin. Assuming a normal distribution of reference depths within each predicted bin, we estimate the confidence intervals by showing the $\pm 1 * \sigma$ and $\pm 2 * \sigma$ bounds, with σ being the per-bin standard deviation. These intervals correspond to 68.3% and 99.7% confidence levels, respectively, providing insights into the dispersion and reliability of the predictions.

The experiments were carried out in a RTX 4090 GPU with 24 GiB VRAM.

5. Results and Discussions

Table 1 summarizes the evaluation metrics for all methods across the two study platforms (PA and PB). The DAv2 model trained on indoor data consistently outperformed its outdoor-trained counterpart. Likewise, ZoeDepth achieved superior performance when trained exclusively on indoor datasets, compared to its outdoor or hybrid (indoor + outdoor) configurations across both platforms. This suggests that, although the 360-degree images were captured in outdoor environments, the visual characteristics of the scenes (such as enclosed metallic corridors, dense pipework, and shaded regions) may resemble indoor settings, where objects are closer and scene geometry is more constrained. Additionally, the effective depth range in the reference data may be limited by scanner capabilities, as reflected in the distribution shown in Figure 1. Among the Patchfusion variants, the Depth Anything-based version consistently outperformed the ZoeDepth-based alternative across all metrics.

In terms of regression-based metrics—namely MAE, RMSE, and R^2—the best-performing model across both platforms was DAv2 (Indoor). Likewise, DAv2 achieved the best thresholded accuracies with the exception of δ_3 in PA, where it obtained the second-best result. Metric3Dv2 ranked second on platform PA, although its performance on platform PB was notably lower compared to other methods. The strong performance of DAv2 may be attributed to its pre-training on a synthetic depth dataset, which provides clean reference signals and avoids common errors found in real-world data. Similarly, Metric3Dv2 was also trained on synthetic datasets, potentially explaining its competitive performance on PA. In addition, Metric3Dv2 achieved the highest R^2, δ_2, and δ_3 scores in PA and the second highest thresholded accuracies in PB, reinforcing its effectiveness in capturing the overall depth structure. Despite this, Metric3Dv2 did not attain the best error or correlation scores in PB, showing an R^2 of -0.42. This discrepancy between poor regression metrics and strong threshold accuracy suggests the presence of outliers, and also indicates R^2's sensitivity to challenging data points in the PB ground truth. Threshold accuracy, being based on relative ratios, is more robust to such outliers compared to absolute error metrics.

The regression and thresholded metrics were lower than those reported on traditional datasets such as NYU. For instance, the authors reported $\delta_1 = 0.98$ for the NYU dataset using DAv2, whereas the best performing model (DAv2 indoor) achieved $\delta_1 = 0.45$ and $\delta_1 = 0.47$ for PA and PB, respectively. This difference may be attributed to the domain shift between general-purpose training sets and the specific context of offshore platforms, and it may be narrowed in future works by fine-tuning to the specific domain.

Table 1 also presents inference times for all methods. Despite the advantages of Metric3Dv2, the model did not fit in the assessed GPU, and thus its inference time was prohibitively larger than other methods (70.4s per sample). The fastest approaches were ZoeDepth and DAv2 with 1.5s and 3.5s per sample. As expected, Patchfusion had a larger inference time compared to its base models due to its patch-wise multi-inference strategy.

Figure 2 groups the predicted values into 1-meter bins and shows the mean reference value per bin on the vertical axis for PA. Each line represents the best variant of each method in terms of MAE: the indoor version for DAv2 and ZoeDepth, and the DA version for Patchfusion. The plot shows that DAv2 and ZoeDepth rarely predicted beyond 10 meters, while Metric3Dv2 and Patchfusion reached values over 40 meters. This upper limit may not have significantly affected performance metrics, as most reference values are also below 10 meters (see Figure 1). Shaded areas represent 1×Std (56% confidence) and 2×Std (96% confidence). The figure highlights the high variability of predictions, with most methods showing deviations of around 40% of the mean at 56% confidence and around 80% at 96%.

Figure 3 presents the corresponding predicted vs. reference plot for test site PB. In this case, the variability was lower compared to test site PA, with deviations of around 20% of the mean at 56% confidence. However, Metric3Dv2 and Patchfusion produced higher errors in general for all bins, and especially for larger reference values. These errors indicate the need for a fine-tuning of these models to the specific domain of the assessed data set.

Figure 4 presents qualitative results for the best-performing variant of each model. DAv2 and ZoeDepth failed to capture the

Table 1. Evaluation metrics for different models on test offshore platforms PA and PB. Best results are in **bold** and second bests are underlined.

Method	Environment	Test platform PA						Test platform PB						Time↓ (s)
		MAE↓ (m)	RMSE↓ (m)	R^2↑	δ_1↑	δ_2↑	δ_3↑	MAE↓ (m)	RMSE↓ (m)	R^2↑	δ_1↑	δ_2↑	δ_3↑	
DAv2	Indoor	**1.2**	3.0	0.39	**0.45**	**0.73**	0.88	**0.6**	**1.0**	**0.63**	**0.47**	**0.77**	**0.91**	3.5
DAv2	Outdoor	1.3	3.1	0.34	0.43	**0.73**	0.87	0.8	1.2	0.51	0.38	0.66	0.84	3.5
ZoeDepth	Indoor	1.5	3.5	0.17	0.39	0.67	0.82	0.9	1.3	0.38	0.32	0.59	0.78	**1.5**
ZoeDepth	Outdoor	2.0	4.0	-0.06	0.21	0.41	0.61	0.9	1.5	0.24	0.28	0.54	0.75	**1.5**
ZoeDepth	In+Out	2.0	4.0	-0.06	0.36	0.61	0.76	1.5	3.0	-2.10	0.27	0.52	0.70	**1.5**
Metric3Dv2	In+Out	1.3	**2.8**	**0.47**	0.41	**0.73**	**0.90**	0.9	2.0	-0.42	0.39	0.70	0.88	70.4
Patchfusion(DA)	Indoor	1.7	3.0	0.39	0.31	0.61	0.83	1.1	1.7	-0.02	0.28	0.54	0.77	32.0
Patchfusion(ZoeDepth)	Indoor	2.0	3.5	0.15	0.28	0.54	0.75	1.2	2.0	-0.44	0.28	0.54	0.75	28.0

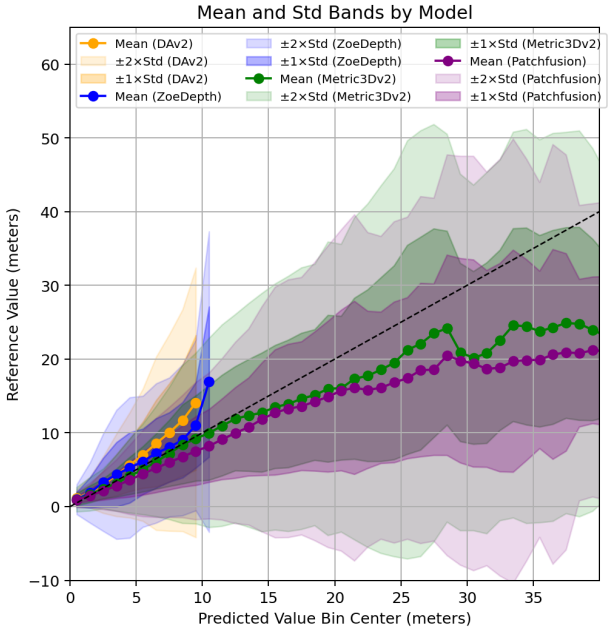

Figure 2. Overlaid histogram for DAv2, ZoeDepth, Metric3Dv2 and Patchfusion in PA test site.

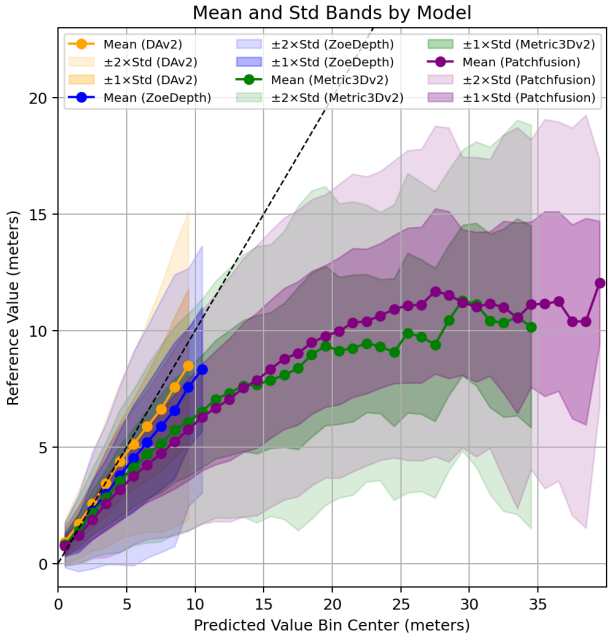

Figure 3. Overlaid histogram for DAv2, ZoeDepth, Metric3Dv2 and Patchfusion in PB test site.

farthest depth regions, underestimating long-range distances. In contrast, Metric3Dv2 and Patchfusion tended to overestimate these distant regions, often predicting values exceeding those in the reference data. Despite these differences, all models produced accurate estimates at short- and mid-range depths (i.e., up to 10 meters), with mean absolute errors remaining below 1.4 meters across all methods.

6. Conclusions

This work presented a comparative study of state-of-the-art monocular depth estimation methods applied to 360-degree imagery of offshore oil platforms, a challenging setting characterized by low-texture surfaces, non-standard viewpoints, and complex metallic structures. The objective was to evaluate the suitability of these methods for producing accurate metric depth maps in such environments.

Based on standard regression metrics (MAE, RMSE, and R^2), the DAv2 Indoor model demonstrated the most balanced performance, with MAE values of 1.2 meters for PA and 0.6 meters for PB, followed closely by ZoeDepth Indoor. The DAv2 Indoor model also achieved the best or second-best thresholded accuracy scores, further indicating its suitability for depth estimation in offshore platforms. While Metric3Dv2 achieved competitive thresholded accuracies and also delivered the lowest RMSE and highest R^2 in PA, its inference time was significantly longer due to high memory requirements, making it impractical on the evaluated GPU. As expected, Patchfusion also resulted in increased inference times. However, it did not consistently improve prediction quality. This finding suggests that patch-based approaches may be unnecessary for this dataset's spatial resolution, where models like DAv2 and Metric3Dv2 can process full-resolution inputs without downsampling.

The inference times, ranging from 1.5s to 70.4s per sample, underscore a trade-off between accuracy and practical deployment. For real-time or continuous monitoring scenarios, faster models like DAv2 and ZoeDepth would be more suitable, despite potential accuracy compromises, whereas Metric3Dv2's current inference speed makes it less practical for such scenarios without significant hardware upgrades or model optimization.

Across all models, prediction variability remained high, and prediction errors were higher for greater depth ranges. However, since approximately 97.5% of ground truth values were concentrated below 10 meters, these larger errors had limited overall impact on the evaluated metrics. This suggests that current models may be adequate for near-field depth estimation in offshore environments but may require adjustments for more distant structures.

As future work, we aim to fine-tune the assessed models on domain-specific data from offshore oil platforms, with the goal of reducing prediction variability and improving accuracy across the entire depth range. Additionally, we will explore the

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-X-3-W4-2025-71-2026 | © Author(s) 2026. CC BY 4.0 License.

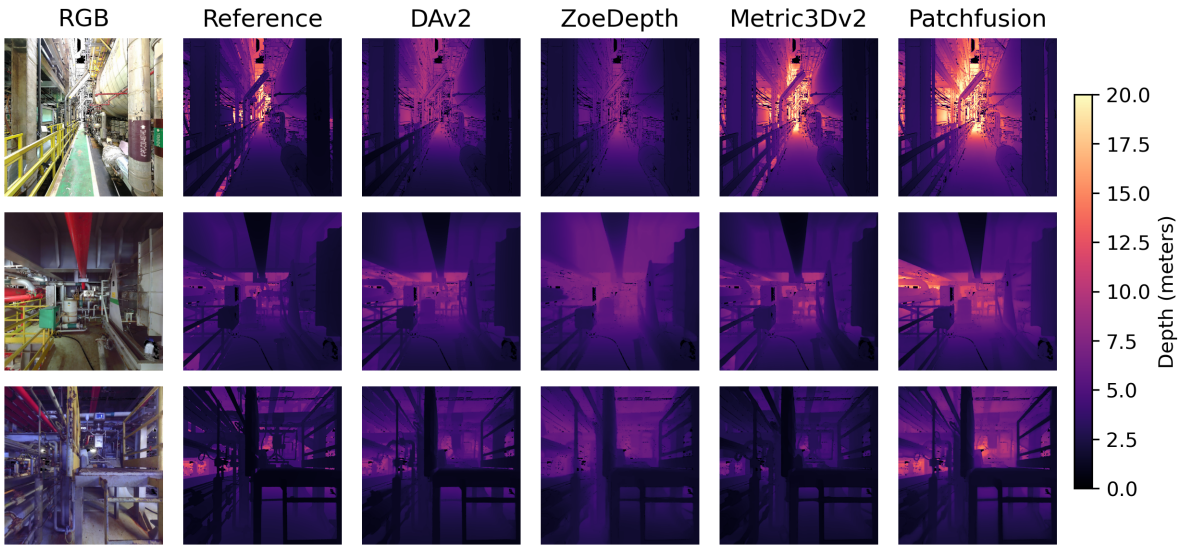

Figure 4. Qualitative results of depth estimation. The first row shows a sample from PA, and the second and third rows from PB.

integration of estimated depth maps into broader automated inspection pipelines to support more robust structural health monitoring. Finally, as a future work, we will involve field experts to evaluate the practical utility of the estimated depth, assessing whether the observed errors are compatible with real inspection requirements and the intended use within offshore platforms.

References

Bhat, S. F., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4009–4018.

Bhat, S. F., Alhashim, I., Wonka, P., 2022. Localbins: Improving depth estimation by learning local distributions. *European Conference on Computer Vision*, Springer, 480–496.

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*.

Cabon, Y., Murray, N., Humenberger, M., 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.

Chang, W., Zhang, Y., Xiong, Z., 2021. Transformer-based monocular depth estimation with attention supervision. *BMVC*, 6, 7.

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*, 27.

Favaro, P., Soatto, S., 2005. A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), 406–417.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2002–2011.

Garcia, R., Happ, P., Feitosa, R., 2021. Large scale semantic segmentation of virtual environments to facilitate corrosion management. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 465–470.

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

Horn, B. K., 1989. Obtaining shape from shading information. *Shape from shading*, 123–171.

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.

Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M., 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1119–1127.

Li, Z., Bhat, S. F., Wonka, P., 2024. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10016–10025.

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

Pataki, Z., Sarlin, P.-E., Schönberger, J. L., Pollefeys, M., 2025. Mp-sfm: Monocular surface priors for robust structure-from-motion. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21891–21901.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3), 1623–1637.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., Susskind, J. M., 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *Proceedings of the IEEE/CVF international conference on computer vision*, 10912–10922.

Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3d. *ACM siggraph 2006 papers*, 835–846.

Szeliski, R., Golland, P., 1999. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32(1), 45–61.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024b. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C., 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9043–9053.

Zang, A., Carvalho, F., Teixeira, M. A., Raposo, A. B., Dos Reis, L. P., 2022. Transforming point cloud data into full panoramic maps for real-time vr applications. *Proceedings of the 24th Symposium on Virtual and Augmented Reality*, 125–131.