# ImmersiveDepth: A Hybrid Approach for Monocular Depth Estimation from 360 Images Using Tangent Projection and Multi-Model Integration

Sarshar Dorosti*
Belfast School of Art, Ulster University, Belfast, Northern Ireland, United Kingdom
Email: dorosti-s@ulster.ac.uk

Xiaosong Yang
National Centre for Computer Animation (NCCA), Bournemouth University, Bournemouth, Dorset, United Kingdom
Email: xyang@bournemouth.ac.uk

## Abstract
ImmersiveDepth is a hybrid framework designed to tackle challenges in Monocular Depth Estimation (MDE) from 360-degree images, specifically spherical distortions, occlusions, and texture inconsistencies. By integrating tangent image projection, a combination of convolutional neural networks (CNNs) and transformer models, and a novel multi-scale alignment process, ImmersiveDepth achieves seamless and precise depth predictions. Evaluations on diverse datasets, show an average 37% reduction in RMSE compared to Depth Anything V2 and a 25% accuracy boost in low-light conditions over MiDaS v3.1. ImmersiveDepth thus establishes a robust solution for immersive technologies, autonomous systems, and 3D reconstruction.

*Keywords: Monocular depth estimation, 360-degree images, tangent projection, VR, AR, SfM, MVS*

## Introduction
MDE is crucial for applications such as 3D reconstruction, virtual reality (VR), and robotics [8]. Traditional photogrammetry methods, including Structure from Motion (SfM) and Multi-View Stereo (MVS), often encounter scalability and environmental constraints [15]. While MDE approaches have gained traction for their efficiency, applying them to 360-degree imagery poses significant challenges like spherical distortions, occlusions, and uneven resolution near the poles, leading to suboptimal predictions [12]. Existing solutions range from tangent-based transformations that project 360-degree images onto planar views to minimize distortions [15], to advanced models such as MiDaS [1] and Depth Anything [2], which capture global geometry or refine local details. Nevertheless, integrating global and local features seamlessly particularly in reflective or transparent environments or under conditions involving fog and sharp dark-light contrasts remains a key hurdle.

Capturing a scene in 360 degrees offers complete spatial context, benefitting VR/AR and autonomous navigation, yet distortions near the poles and occlusions persist as major obstacles. To address these issues, we introduce ImmersiveDepth, a hybrid framework that (1) uses tangent image projection to reduce spherical distortions; (2) integrates MiDaS v3.1 (transformer-based) with Depth Anything V2 (CNN-based) for a balance between global consistency and local refinement; and (3) applies multi-scale alignment for seamless, artifact-free depth maps. By leveraging both transformers and CNNs, ImmersiveDepth advances beyond current methods to provide robust, balanced depth estimation for 360-degree imagery.

## Methodology
ImmersiveDepth tackles 360-degree MDE challenges through four stages. Stage 1: Tangent image projection divides spherical images into overlapping planar views using an icosahedron-based projection [15], reducing pole distortions and preserving details for accurate depth estimation.

Stage 2: Depth network integration combines MiDaS v3.1, a transformer-based model with BEiT and Swin backbones for large-scale geometry, and Depth Anything V2, a CNN-based model refining fine details like thin structures and reflective surfaces. MiDaS is trained on ReDWeb, DIML, and MegaDepth datasets, while Depth Anything uses OmniData and synthetic datasets with pseudo-labeling to handle complex environments. Both models, trained for 30 epochs using an AdamW optimizer ($1\times10^{-4}$ initial learning rate, reduced every 10 epochs) and batch size of 8 on an NVIDIA RTX 3090 GPU, are merged through weighted ensembling for balanced depth predictions.

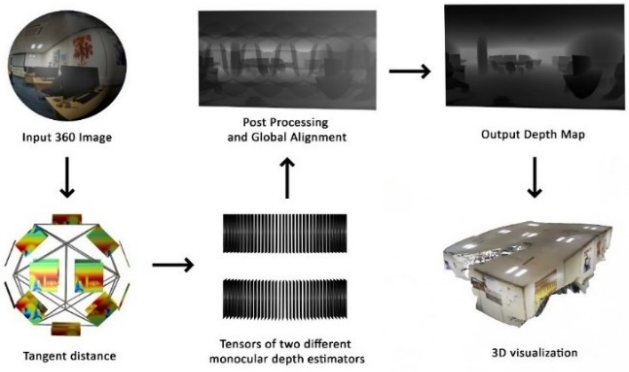


Diagram 1: Conversion of a 360° RGB image into a depth map in the ImmersiveDepth project.

Stage 3: Multi-scale alignment seamlessly integrates depth maps using disparity alignment and Poisson blending [12], while a bilateral filter refines boundaries for consistency.

Stage 4: Post-processing and fine-tuning apply mean-std normalization for depth scale unification, Gaussian filtering to reduce noise, and CLAHE-based contrast enhancement for low-light clarity. Pseudo-labeled 360-degree images adapt the model for reflective or transparent surfaces.

ImmersiveDepth is evaluated on KITTI [4] (dynamic outdoor scenes), NYU Depth V2 [5] (360-degree indoor environments), and Matterport3D [6] (cluttered indoor spaces). Datasets, split into training (70%), validation (15%), and testing (15%), are resized to 512×512 pixels and augmented with random rotations, flips, and color jitter, ensuring robust and accurate depth estimation across varied scenarios.

## Result and Discussion
To evaluate ImmersiveDepth, we use RMSE, AbsRel, δ-thresholds ($\delta_1$, $\delta_2$, $\delta_3$), squared relative error (sq_rel), and Log10 error. These metrics confirm significant advancements over MiDaS v3.1 and Depth Anything V2 across various scenarios.

Table 1. ImmersiveDepth outperforms Depth Anything V2 with higher d1-d3 precision and lower errors, demonstrating superior depth estimation fidelity.

| Metric | ImmersiveDepth | DepthAnything v2 |
|---|---|---|
| d1↑ | 0.1237 | 0.0733 |
| d2↑ | 0.285 | 0.1758 |
| d3↑ | 0.4523 | 0.2946 |
| abs_rel↓ | 1.0761 | 1.8249 |
| sq_rel↓ | 0.2373 | 0.6056 |
| rmse↓ | 0.1872 | 0.287 |
| rmse_log↓ | 0.8282 | 1.0679 |

ImmersiveDepth achieves a 37% reduction in RMSE (0.187 vs. 0.287) compared to Depth Anything V2 and a $\delta_1$ accuracy of 78.2%, outperforming MiDaS v3.1 in all tested settings. With CLAHE-based contrast enhancement, it yields a 25% accuracy improvement in low-light conditions. Reflective surfaces are handled effectively, with an 18% gain in robustness against specular highlights.

In Matterport3D indoor scenes, ImmersiveDepth eliminates banding artifacts and preserves intricate edge details, such as furniture boundaries. In KITTI outdoor datasets, it maintains consistent depth predictions across objects at varying distances, offering realistic spatial representations for VR applications.

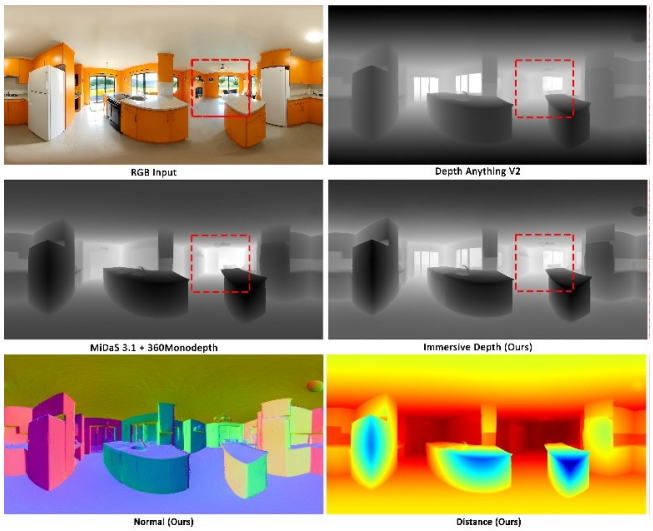


Figure 1: Comparison of GT, Depth Anything V2, ImmersiveDepth (ours), with histogram and bar chart evaluation factors for assessing depth estimation performance on transparent and reflective surfaces.

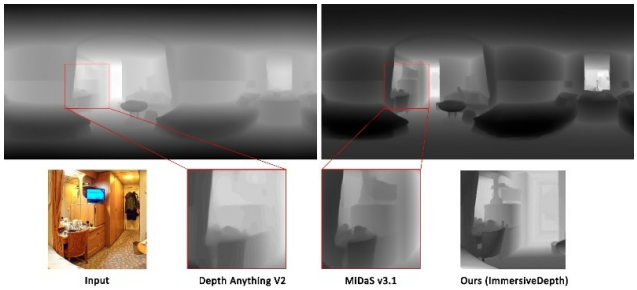


Figure 2: Comparison of depth maps from Depth Anything V2, MiDaS v3.1, and our method, combined via weighted averaging to enhance depth prediction accuracy.

Figures 1 and 2 showcase its superior performance compared to Depth Anything V2 and MiDaS v3.1, providing smoother transitions and sharper boundaries. Optimized tangent image projection and normalization enable scalable, high-resolution depth map generation for VR or AR.

## Conclusion

This study introduces ImmersiveDepth, a hybrid framework for MDE from 360-degree images, addressing key challenges such as spherical distortions, occlusions, and texture inconsistencies. By integrating tangent image projection, CNNs, transformer models, and multi-scale alignment techniques, ImmersiveDepth achieves seamless and precise depth predictions, establishing a benchmark for immersive technologies and real-world applications. It demonstrates a 37% reduction in RMSE and 25% improved accuracy in low-light conditions compared to MiDaS v3.1 and Depth Anything V2, excelling in complex scenarios like reflective surfaces and occlusions. Its robust validation across diverse datasets underscores its adaptability for VR or AR. Future work aims to optimize computational efficiency and simplify the framework for real-time, scalable applications.

## References

[1] R. Birkl, D. Wofk, and M. Müller, "MiDaS v3.1 -- A Model Zoo for Robust Monocular Relative Depth Estimation," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.14460.

[2] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," 2024, *arXiv*. doi: 10.48550/ARXIV.2401.10891.

[3] L. Yang *et al.*, "Depth Anything V2," 2024, *arXiv*. doi: 10.48550/ARXIV.2406.09414.

[4] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 10766–10776. doi: 10.1109/ICCV48922.2021.01061.

[5] A. Chang *et al.*, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *2017 International Conference on 3D Vision (3DV)*, Qingdao: IEEE, Oct. 2017, pp. 667–676. doi: 10.1109/3DV.2017.00081.

[6] R. Baskar, P. M. Krishnammal, A. Aeron, S. S. Ali, T. T. Leonid, and M. R. Arun, "3D Image Reconstruction and Processing for Augmented and Virtual Reality Applications: A Computer Generated Environment," in *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, Greater Noida, India: IEEE, Nov. 2023, pp. 866–870. doi: 10.1109/ICCSAI59793.2023.10421067.

[7] Y. A. Shleibik, "3D RECONSTRUCTION OF 2D IMAGES USING DEEP LEANING," 2023, doi: 10.13140/RG.2.2.33309.69607.

[8] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, "Monocular Depth Estimation Using Deep Learning: A Review," *Sensors*, vol. 22, no. 14, p. 5353, Jul. 2022, doi: 10.3390/s22145353.

[9] S. Tang, F. Zhang, J. Chen, P. Wang, and Y. Furukawa, "MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.01097.

[10] S. Guttikonda and J. Rambach, "Single Frame Semantic Segmentation Using Multi-Modal Spherical Images," 2023, *arXiv*. doi: 10.48550/ARXIV.2308.09369.

[11] D. Seichter, B. Stephan, S. B. Fischedick, S. Müller, L. Rabes, and H.-M. Gross, "PanopticNDT: Efficient and Robust Panoptic Mapping," 2023, *arXiv*. doi: 10.48550/ARXIV.2309.13635.

[12] M. Rey-Area, M. Yuan, and C. Richardt, "360MonoDepth: High-Resolution 360° Monocular Depth Estimation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 3752–3762. doi: 10.1109/CVPR52688.2022.00374.

[13] Z. Cheng and G. Ji, "Evaluating Panoramic 3D Estimation in Indoor Lighting Analysis," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.14836.

[14] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth," 2023, *arXiv*. doi: 10.48550/ARXIV.2302.12288.

[15] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, "Tangent Images for Mitigating Spherical Distortion," 2019, *arXiv*. doi: 10.48550/ARXIV.1912.09390.