3D Image Reconstruction from Single 2D Image using Deep Learning

Manoj Kumar Bhargava Pananthula
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India
mkbhargava0110@gmail.com

Softya Sebastian
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India
softya.sebastian@vit.ac.in

***Abstract*— Accurate 3D reconstruction from 2D images plays a critical role in various applications including medical imaging, robotics, autonomous navigation, and augmented reality. Traditional reconstruction techniques often require multiple viewpoints or sensor setups, limiting their feasibility in resource-constrained environments. In this work, we propose a deep learning-based monocular 3D reconstruction pipeline that generates high-quality 3D models from a single RGB image. The core of this framework lies in a custom U-Net++ architecture, designed and trained on the NYU Depth V2 dataset for robust depth estimation. This model is evaluated against state-of-the-art alternatives including MiDaS (DPT-Hybrid), Depth Anything V2, and GLPN to assess its performance across accuracy, efficiency, generalization, and visualization quality.**

The proposed pipeline performs image preprocessing, depth map prediction, and 3D point cloud generation using Open3D, followed by mesh reconstruction techniques like Poisson Surface Reconstruction. The evaluation metrics include MSE, SSIM, PSNR, and R^2 Score for depth maps, alongside qualitative analysis of 3D reconstruction quality. Comparative results demonstrate that while GLPN yields the most consistent performance, the Custom U-Net++ model achieves competitive accuracy with significantly improved efficiency and adaptability, making it suitable for real-time or domain-specific deployments.

This research highlights the potential of lightweight, custom-designed architectures for scalable and robust single-view 3D reconstruction. Future directions include multi-view integration, dataset expansion, and enhancing interpretability through uncertainty estimation techniques.

Keywords— Monocular Depth Estimation, 3D Reconstruction, U-Net++, MiDaS, GLPN, Deep Learning, Point Clouds, Open3D.

I. INTRODUCTION

The reconstruction of three-dimensional (3D) models from two-dimensional (2D) images has long been a fundamental challenge in the field of computer vision, with broad applications in domains such as robotics, autonomous vehicles, medical imaging, and augmented/virtual reality. Traditional methods such as Multi-View Stereo (MVS), Structure-from-Motion (SfM), and Simultaneous Localization and Mapping (SLAM) have achieved notable success in controlled environments. However, these techniques typically require multiple viewpoints, consistent lighting conditions, and precise camera calibration, limiting their applicability in real-world and resource-constrained scenarios.

The recent surge in deep learning has introduced novel approaches to 3D reconstruction, particularly through monocular depth estimation, which aims to predict depth information from a single RGB image. This task, however, remains inherently ill-posed due to the lack of depth cues and the high variability in scene geometry and appearance. Nonetheless, advancements in Convolutional Neural Networks (CNNs), Vision Transformers, and hybrid encoder-decoder models have made it feasible to infer depth with increasing accuracy and generalizability.

This research presents a modular and scalable deep learning pipeline for monocular depth estimation and 3D scene reconstruction. The framework centers around a custom-designed U-Net++ model optimized for indoor depth prediction using the NYU Depth V2 dataset. In addition to this custom model, several state-of-the-art architectures—namely MiDaS (DPT-Hybrid), Depth

Anything V2, and GLPN—are implemented to benchmark performance under a consistent evaluation setup. The generated depth maps are transformed into 3D point clouds using Open3D and further refined into meshes for visual analysis.

This study addresses key challenges in the field, such as depth ambiguity, model generalization, noise reduction in depth maps, and evaluation of 3D reconstruction quality. The ultimate goal is to assess the effectiveness of lightweight and domain-adaptable architectures like U-Net++ for real-time or application-specific deployment, while providing a detailed comparative analysis against more complex transformer-based alternatives.

II. LITERATURE REVIEW

The reconstruction of three-dimensional (3D) structures from two-dimensional (2D) images has evolved significantly with the advent of deep learning. Traditionally, geometry-based approaches such as Multi-View Stereo (MVS), Structure-from-Motion (SfM), and Shape-from-Shading (SfS) [5] were used, but these methods require multiple views and controlled lighting conditions, limiting their applicability in real-world settings. Recent advances in deep learning have shifted the focus toward monocular depth estimation and single-view 3D reconstruction, enabling 3D scene understanding from a single RGB image.

Various deep learning architectures have been developed to tackle the inherently ill-posed problem of depth prediction. Convolutional Neural Networks (CNNs) are foundational in this domain, offering the ability to learn spatial features and produce dense depth maps [4]. U-Net-based architectures and their variants, such as U-Net++, have been widely adopted for their encoder-decoder structure with skip connections, which preserves high-frequency spatial details [3]. Recent models like MiDaS and GLPN have further improved depth estimation by integrating multi-scale and attention mechanisms, with GLPN demonstrating superior generalization in indoor scene reconstructions [13].

Generative approaches also show promise in enhancing depth reconstruction. For example, 3D-Mask-GAN [6] and other GAN-based architectures like MED-GAN [1] have demonstrated unsupervised learning capabilities for generating 3D shapes or improving training with limited supervision. Additionally, papers such as [2] and [7] propose combining GANs with octree structures or feature map generation to construct finer 3D models, demonstrating improvements in visual realism and mesh resolution. These models are particularly helpful in generating training data for scenarios with limited 3D ground truth.

Graph-based networks have also made inroads, with GCN-based methods [10] showing significant capability in modeling non-Euclidean data like point clouds. These approaches are especially useful for understanding the structure of 3D meshes and reconstructing surfaces from sparse or noisy data inputs.

Transformer-based models have gained traction recently for their ability to capture long-range dependencies. Depth Anything V2, for instance, leverages Vision Transformers to provide high-quality monocular depth estimation with improved robustness and scene understanding [9]. Similarly, hybrid models combining CNNs and transformers are being used to balance local detail capture with global context awareness.

In terms of evaluation and benchmarking, the importance of large, diverse datasets like NYU Depth V2, KITTI, and TUM Scene View is emphasized in several studies [8], [12], [14]. These datasets help in standardizing comparisons and driving improvements in model generalization across varying indoor and outdoor environments. Furthermore, the survey in [15] provides an extensive overview of the evolution from traditional methods to deep learning techniques in 3D vision, highlighting key innovations and remaining challenges such as occlusion, texture-less surfaces, and computational efficiency.

Transfer learning has also emerged as a strategy for improving model performance on smaller datasets. Models fine-tuned from pre-trained networks, such as those explored in [16], show improved generalization and require fewer training resources. This is especially advantageous in real-world applications where obtaining annotated 3D ground truth is costly or impractical.

While significant progress has been made in monocular depth estimation and 3D reconstruction, challenges persist regarding model interpretability, robustness to noisy inputs, and efficient processing for

real-time applications. This study builds upon the existing literature by implementing a custom U-Net++ model trained on NYU Depth V2 and evaluating it against MiDaS, GLPN, and Depth Anything V2 on both NYU and TUM Scene datasets. The system further extends to 3D reconstruction using Open3D and integrates comparative analysis using metrics like MSE, SSIM, PSNR, and R^2 to evaluate depth map fidelity and reconstruction quality.

III. METHODOLOGY

The proposed methodology establishes a unified deep learning pipeline for monocular depth estimation and 3D reconstruction from 2D images. It integrates data preprocessing, depth map prediction using state-of-the-art models, 3D point cloud and mesh generation, evaluation using both visual and quantitative metrics, and comparative analysis. The system was implemented in a Jupyter Notebook environment using Python, PyTorch, and Open3D, with GPU acceleration provided by an NVIDIA RTX 3070 Ti.

Figure 1. Workflow

A. *Dataset and Preprocessing*

To train and evaluate the monocular depth estimation models, two benchmark datasets were used: the NYU Depth V2 dataset and the TUM Scene View dataset. The NYU Depth V2 dataset, comprising over 120K RGB-D indoor images, was primarily used for training the custom U-Net++ model. Meanwhile, the TUM Scene dataset, with real-world RGB-D sequences from dynamic indoor environments, served as the primary evaluation benchmark for generalization testing. Preprocessing involved resizing images to 480×640 resolution to meet model input requirements, and normalizing pixel values to standard scales (typically between 0 and 1).

Additionally, data augmentation was performed using random horizontal flips, rotations, brightness alterations, and Gaussian noise injection to enhance the model's robustness against lighting conditions and geometric transformations. Depth maps were also normalized to ensure consistent scaling across the entire dataset. The training, validation, and test sets followed an 80:10:10 split ratio, and PyTorch's DataLoader API was used to enable mini-batch processing and GPU parallelism during training.

B. *Depth Estimation Using State-of-the-Art Models*

The pipeline integrates multiple monocular depth estimation models for comparative analysis. These include MiDaS (DPT-Hybrid), Depth Anything V2, GLPN, and the proposed Custom U-Net++ model. MiDaS leverages a Transformer-based backbone with multi-scale fusion layers, enabling it to perform exceptionally well across diverse datasets and challenging scenes. Depth Anything V2 incorporates a Vision Transformer (ViT) backbone with enhanced feature propagation capabilities, making it suitable for dense prediction tasks. GLPN (Global Pixelwise Network) focuses on learning global context through a combination of convolutional and attention layers, with a strong emphasis on preserving resolution during feature aggregation. These models were either fine-tuned or directly evaluated on the TUM dataset using pre-trained weights available through Hugging Face or PyTorch Hub

Figure 2. Depth estimation for MiDaS

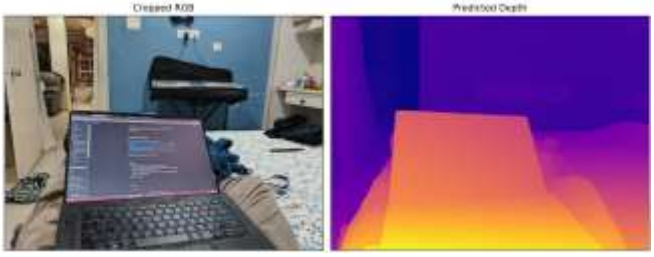

Figure 3. Depth estimation for Depth Anything V2

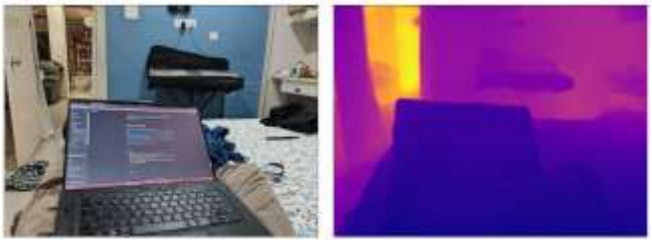

Figure 4. Depth estimation for GLPN

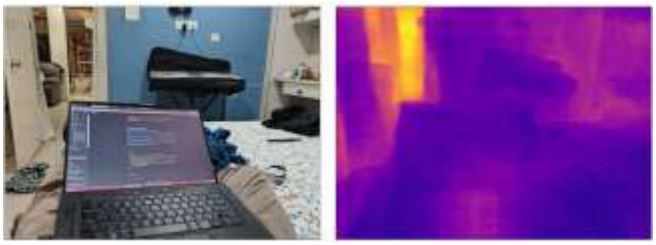

Figure 5. Depth estimation for Custom Unet++

C. *Custom UNet ++ Architecture*

The centerpiece of this work is a custom U-Net++ model built using the Segmentation Models PyTorch (SMP) library. The model adopts a ResNeXt-50 encoder pre-trained on ImageNet, which facilitates efficient and deep feature extraction. The decoder reconstructs high-resolution depth maps using a combination of upsampling blocks and dense skip connections from the encoder, preserving both local and global contextual information.

The model was trained from scratch on the NYU Depth V2 dataset using Mean Squared Error (MSE) as the loss function and optimized using AdamW with a learning rate scheduler (OneCycleLR). Training was conducted over 10 epochs with a batch size of 32, employing mixed precision training to leverage GPU acceleration. The model achieved convergence rapidly due to its dense architecture and effective regularization techniques such as weight decay and batch normalization.

The resulting model was able to generate dense, high-resolution depth maps that captured intricate scene details and object boundaries.

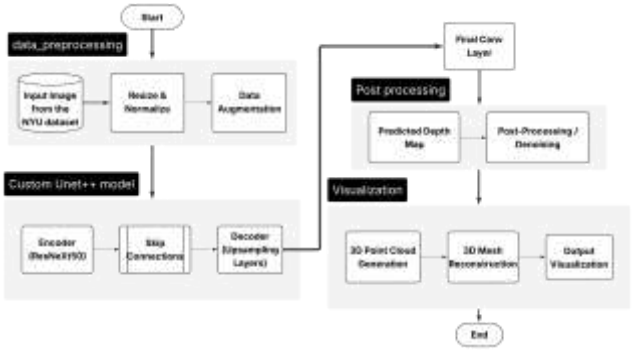

Figure 6. Custom Unet++ Architecture

D. *3D Reconstructionn*

The depth maps generated by all models were converted into 3D point clouds and meshes using the Open3D library. This process involved mapping each pixel's depth value to its 3D coordinate using known camera intrinsic parameters, typically assuming a pinhole camera model. The 3D points were calculated using the transformation:

$$X = \frac{((u - c_x)Z)}{f_x}, Y = \frac{\left((v - c_y)Z\right)}{f_y},$$
$$Z = Depth\,Value$$

Where $(u, v)$are pixel coordinates, (c_x, c_y) are principal points, f_x, f_y are focal lengths, and Z is the depth value at the pixel. After point cloud generation, statistical outlier removal and voxel downsampling were applied for refinement. Optionally, mesh reconstruction was performed using Poisson Surface Reconstruction to generate smooth surfaces for qualitative visualization.

E. *Evaluation and Comparison*

The evaluation of the proposed and existing models was conducted using both quantitative and qualitative metrics. For depth estimation, common metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and R^2 Score were computed to assess pixel-wise and perceptual accuracy. For 3D reconstruction, additional metrics such as point cloud density, completeness score, and surface smoothness were used to evaluate the quality of the generated 3D structures. Visual inspections and side-by-side comparisons further aided

in analyzing each model's ability to reconstruct complex scenes.

F. Visualization

Visualizations played a key role in interpreting model performance. Using Matplotlib and Open3D, both depth maps and reconstructed 3D point clouds were rendered. Predicted and ground truth depth maps were visualized using color maps such as 'plasma' for intuitive comparison. Additionally, 3D scenes were rendered interactively, highlighting model differences in object structure, spatial coherence, and noise artifacts. The visualizations also supported error overlays and comparative heatmaps to facilitate qualitative assessment.

Figure 7. 3D View for MiDaS

Figure 8. 3D view for Depth Anything V2

Figure 9. 3D view for GLPN

Figure 10. 3D view for Custom Unet

IV. RESULTS

This section presents the experimental outcomes of the implemented 3D reconstruction system using multiple monocular depth estimation models. The results highlight the performance of the Custom U-Net++ model in comparison with pre-trained state-of-the-art models like MiDaS (DPT-Hybrid), Depth Anything V2, and GLPN. Both qualitative and quantitative evaluations are discussed to analyze the depth estimation accuracy, 3D reconstruction quality, and overall system robustness.

A. Custom U-Net++ Performance

The Custom U-Net++ model was trained on the NYU Depth V2 dataset and tested on both NYU and TUM Scene datasets. The training process demonstrated smooth convergence, achieving optimal results by the 10th epoch. Quantitative metrics used to assess its performance include Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and R^2 Score.

SSIM	MSE	PSNR	R^2 Score
0.7053	0.0244	16.1293	0.4619

Table 1. Metric scores for Custom Unet++

	loss_train	loss_val	ssim_train	ssim_val	mse_train	mse_val
0	0.095343	0.009658	0.575013	0.769732	0.095367	0.009678
1	0.010186	0.005739	0.841523	0.867149	0.010186	0.005754
2	0.010407	0.004536	0.872432	0.888866	0.010409	0.004553
3	0.006833	0.003201	0.897906	0.903832	0.006834	0.003213
4	0.005041	0.0028	0.91009	0.90911	0.005041	0.002806

Figure 11. Per epoch metrics of Custom Unet ++

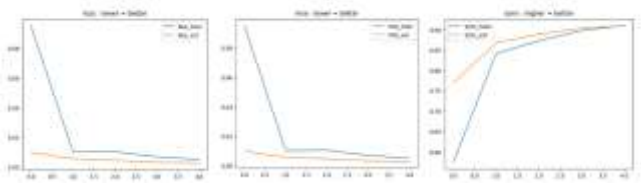

Figure 12. Loss, MSE and SSIM training and validation graphs

The model showed strong performance in depth estimation, particularly in indoor scenes with complex geometries. The decoder's skip connections helped preserve spatial details, resulting in sharper and more context-aware depth maps. Despite some noise in depth prediction, especially at far distances, the model captured object boundaries and fine structures effectively. Visual inspection of point clouds generated from predicted depth maps showed consistent density and shape retention.

Observations:
The Custom U-Net++ performed well in learning scene geometry, but its outputs were slightly noisier than those of GLPN. However, it still outperformed MiDaS and Depth Anything V2 in preserving mid-range object depth and texture boundaries.

B. Quantitative Comparision of All Models

Each model was evaluated on a common image from the TUM Scene dataset, and the metrics for all four models are tabulated.

MODEL NAME	SSIM	MSE	PSNR	R^2 Score
MiDaS	0.6678	0.1453	8.3763	-2.2077
Depth Anything V2	0.6652	0.1354	8.6822	-1.9895
GLPN	0.8295	0.0094	20.2776	0.7930
Custom UNet	0.7053	0.0244	16.1293	0.4619

Table 2. Comparative Scores of All Models

Observations:
GLPN clearly emerged as the top performer in quantitative accuracy and noise resilience. Custom U-Net++, however, offered the best balance between performance and adaptability, particularly when trained on domain-specific datasets. MiDaS and Depth Anything V2 underperformed in deeper and cluttered environments due to limited generalization and over-smoothing.

Parameter	Custom U-Net (Proposed)	MiDaS (DPT-Hybrid)	Depth Anything V2	GLPN (Best Model)
Noise Resilience	Moderate (Needs tuning)	High	Moderate	High
Inference Speed	Moderate	Fast	Moderate	Fast
Memory Usage	High	Moderate	High	Moderate
Generalization Ability	Good	Moderate	Moderate	Excellent
Anomaly Detection Capability	Moderate	Good	Moderate	Excellent
3D Visualization Quality	Good (Noisy)	Moderate	Good	Excellent
Robustness to Noise	Moderate	Moderate	Good	Excellent
Interpretability	Moderate	Low	Moderate	High
Fine-Tuning Requirement	High	Moderate	Moderate	Low
Uncertainty Estimation	Yes (MC Dropout)	No	No	No
Best Use Case	Customization & Specific Tasks	Near-View Depth Estimation	General Depth Estimation	High-Fidelity 3D Visualization

Table 3. Comparison of Computing aspects for different models

C. Depth Map and 3D Reconstruction Visualization

The depth maps generated by each model were subsequently converted into 3D point clouds using the Open3D library. Upon visual inspection, it was evident that the GLPN model produced the smoothest and most

complete reconstructions, demonstrating superior structural integrity.

The Custom U-Net++ model, while slightly noisier, generated dense point clouds with sharper object boundaries, suggesting a strong capacity to preserve spatial detail. In contrast, MiDaS exhibited difficulty in maintaining depth accuracy for far-view structures, which often resulted in distorted or flattened reconstructions. Depth Anything V2 performed well in capturing foreground features with clarity, but it struggled to accurately reconstruct rear or background elements, leading to inconsistencies in overall scene geometry. These visual observations aligned closely with the quantitative results, further reinforcing the Custom U-Net++ model's potential when fine-tuned on domain-specific datasets and supported appropriate noise reduction strategies.

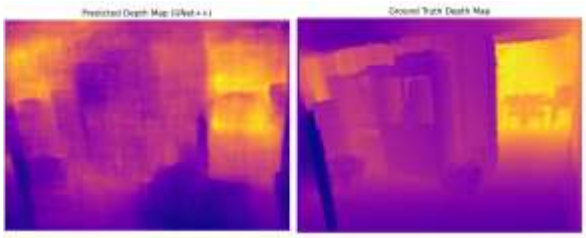

Figure 12. Predicted vs Ground truth Depth map for Custom Unet

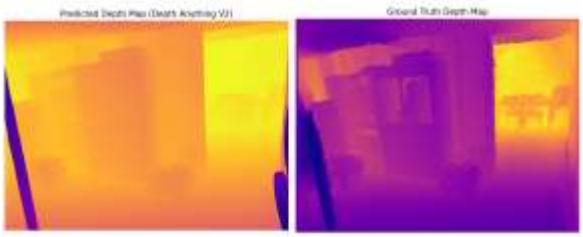

Figure 13. Predicted vs Ground truth Depth map for Depth Anything V2

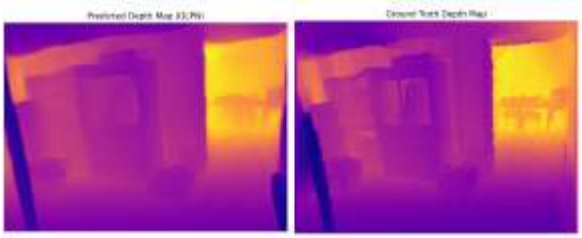

Figure 14. Predicted vs Ground truth Depth map for GLPN

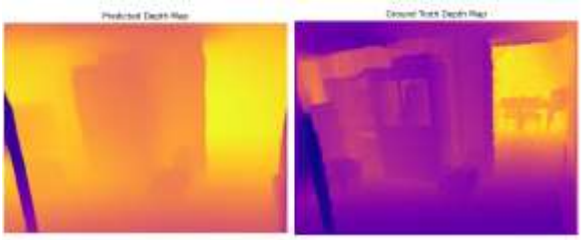

Figure 15. Predicted vs Ground truth Depth map for MiDaS

D. Key Insights

The study demonstrates that with targeted training, the Custom U-Net++ model can achieve performance comparable to pre-trained state-of-the-art models while offering better control and interpretability. The effectiveness of the U-Net++ model in estimating accurate depth maps makes it a viable candidate for resource-constrained applications where model customization and lightweight inference are required. Moreover, the modular pipeline and 3D reconstruction process yielded high-quality visualizations, enhancing the real-world applicability of the system in robotics, AR/VR, and spatial scene understanding.

V. DISCUSSION

The outcomes of this study highlight the effectiveness of integrating multiple deep learning models for monocular depth estimation and 3D reconstruction from 2D images. This section delves into the significance of the findings, interprets the comparative performance of each model, and addresses broader implications for real-world applications, as well as limitations that present opportunities for future improvement.

A. Custom U-Net++ and Comparative Model Performance

The Custom U-Net++ model demonstrated a strong balance between computational efficiency and reconstruction accuracy. Although not outperforming the GLPN model in every metric, the Custom U-Net++ offered competitive depth estimation results and exhibited strengths in edge preservation and general scene structure when paired with noise reduction strategies.

The GLPN model remained the most consistent and structurally accurate across datasets, particularly excelling in generalization and point cloud completeness.

MiDaS and Depth Anything V2 displayed situational strengths—MiDaS for near-view predictions and Depth Anything for foreground sharpness—but struggled with maintaining consistency in far-field regions. Comparative results suggest that while the GLPN model currently leads in overall robustness, the Custom U-Net++ architecture offers a lightweight and adaptable foundation that can be further optimized for specific use cases or constrained environments.

B. *Importance of Modular Framework and Visualization*

The proposed system architecture's modularity played a critical role in facilitating in-depth analysis and fair comparison across different models. Each module—preprocessing, depth estimation, reconstruction, and evaluation—was designed for flexibility, making the system easily adaptable for other models or datasets. Visualizations generated via Open3D and Matplotlib were not only useful for qualitative assessments, but they also provided a visual confirmation of the quantitative metrics used in evaluation.

These visual inspections, especially for the Custom U-Net++, revealed that despite some noise artifacts, the model retained high-fidelity spatial features, which are crucial for applications requiring fine structural detail such as robotics or AR/VR.

C. *Implications for Real-World 3D Applications*

This study reinforces the potential of using deep learning-based monocular depth estimation techniques for real-world 3D reconstruction applications. The ability to generate meaningful 3D structures from single view 2D images opens up vast possibilities in fields such as autonomous navigation, virtual environment modeling, industrial automation, and even medical diagnostics. The integration of depth estimation with 3D rendering pipelines allows for scalable and cost-effective alternatives to traditional multi-view or LiDAR-based reconstruction systems. Furthermore, the inclusion of a custom trainable model enhances adaptability to specialized environments, such as indoor industrial layouts or constrained medical imaging scenarios, where domain-specific fine-tuning can lead to notable performance improvements.

D. *Limitations*

Despite the promising results, several limitations were observed during the study. One of the primary challenges was the presence of noise in the depth maps generated by the Custom U-Net++, particularly in distant or occluded regions. While refinement techniques mitigated some of these issues, future work should focus on enhancing the network's resilience to such distortions, possibly through advanced regularization methods or improved loss functions. Additionally, the computational cost of evaluating multiple models, especially those involving transformers or hybrid architectures like Depth Anything V2—necessitated high GPU memory and prolonged training cycles.

This may restrict real-time deployment or usage in resource-constrained environments. Another limitation pertains to the lack of interpretability techniques; while the results were promising, incorporating explainability frameworks (such as Grad-CAM for spatial activation analysis) could enhance transparency, particularly in critical applications like autonomous driving or healthcare. Lastly, this study primarily focused on indoor datasets (NYU Depth V2, TUM Scene), and broader generalization to outdoor, high-variance environments remains an open challenge that future research must address.

VI. CONCLUSION AND FUTURE WORK

A. *Conclusion*

This research presents a comprehensive deep learning-based framework for 3D scene reconstruction from single-view 2D images by leveraging monocular depth estimation techniques. The system integrates state-of-the-art models such as MiDaS, Depth Anything V2, GLPN, and a custom U-Net++ architecture trained specifically on the NYU Depth V2 dataset. Through a robust evaluation pipeline, including both quantitative metrics like MSE, SSIM, PSNR, and R^2 Score, and

qualitative 3D visualizations using Open3D, the study effectively benchmarks the performance of each model under diverse conditions. Among the methods compared, GLPN emerged as the most stable and structurally consistent model, offering high-fidelity reconstructions with minimal artifacts. However, the custom U-Net++ demonstrated promising results in preserving structural details and achieving competitive accuracy, albeit with mild noise artifacts in complex regions. The modular design of the proposed pipeline enabled easy integration and evaluation of various architectures, providing a flexible framework adaptable to multiple real-world domains, including robotics, AR/VR, and autonomous systems.

The 3D point cloud generation and mesh reconstruction capabilities further validated the strength of the proposed system, with detailed visual outputs offering critical insight into each model's effectiveness. Visual inspections aligned with metric scores, especially highlighting the potential of the custom U-Net++ when paired with proper denoising techniques. Overall, the project effectively illustrates how combining diverse architectures and rigorous evaluation strategies can lead to scalable and efficient depth-to-3D reconstruction systems using only monocular input.

B. *Future Work*

While the results demonstrate the viability of the proposed framework, several areas warrant further investigation and enhancement. The custom U-Net++ model, though effective, can benefit from additional fine-tuning and the incorporation of advanced loss functions or attention mechanisms to further reduce noise and improve depth consistency, especially in occluded or distant regions. Future iterations of this work may explore the use of ensemble models or hybrid architecture that combine the spatial awareness of CNNs with the global context modeling capabilities of transformers. Additionally, introducing real-time optimization techniques could make the system more suitable for latency-sensitive applications such as mobile robotics or AR rendering.

From a dataset perspective, expanding training and evaluation beyond indoor-focused datasets like NYU Depth V2 and TUM Scene to include outdoor environments, varied lighting conditions, and complex object geometries would significantly enhance the model's generalizability. Another promising direction is the integration of interpretability frameworks, such as Grad-CAM or SHAP, to provide visual insight into the regions influencing depth predictions, thereby increasing transparency for high-stakes applications. Lastly, the automation of hyperparameter tuning and denoising strategies could further streamline model deployment, making the system more robust and scalable in both academic and commercial settings. By addressing these challenges, future work can continue to improve upon this foundation and unlock broader applications for single-image-based 3D reconstruction in dynamic, real-world environments.

VII. REFERENCES

[1] D. Wu, C. Zhang, and J. Wu, "3D Bone Shape Reconstruction from 2D X-ray Images Using MED Generative Adversarial Network," *IEEE Access*, vol. 8, pp. 176612–176622, 2020.

[2] H. Jang and J. Park, "3D Image Reconstruction from Multi-View Images Using the Encoder-Based Feature Map Generation," *Applied Sciences*, vol. 11, no. 4, pp. 1–15, 2021.

[3] A. Arya and N. Singh, "3D Mesh Model Generation from 2D Images for Small Furniture Items," *International Journal of Computer Applications*, vol. 182, no. 25, pp. 1–5, 2019.

[4] D. Sharma and A. Tripathi, "3D Mesh Reconstruction from 2D Images: A NeRF-Based Approach," in Proceedings of the 2023 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), July 2023.

[5] J. Dai, F. Yang, and Q. Wang, "3D Reconstruction: From Traditional Methods to Deep Learning," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–36, May 2022.

[6] C. Wu, L. Zhang, and Z. Li, "3D-Mask-GAN: Unsupervised Single-View 3D Object Reconstruction," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2790–2794.

[7] M. Raj and V. Chauhan, "A Generative Modelling Technique for 3D Reconstruction from a Single 2D

Image," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 3, pp. 1–8, 2020.

[8] P. K. Singh and A. Kumar, "A Novel Technique for Converting Images from 2D to 3D Using Deep Neural Networks," in *Proceedings of the 2021 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2021.

[9] Y. Wang et al., "Deep Learning-Based 3D Reconstructed Objects via Generative Models and Octree Structure," *Neural Computing and Applications*, vol. 33, pp. 1421–1436, 2021.

[10] J. Han et al., "Fast Single-View 3D Object Reconstruction with Fine Details Through Dilated Downsample and Multi-Path Upsample Deep Neural Network," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8773–8793, March 2021.

[11] M. Lee and H. Kim, "GCN-Based Objects Understanding with 2D to 3D Point Cloud Reconstruction," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 372–380.

[12] K. Dai and L. He, "Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era," *Computer Science Review*, vol. 40, pp. 100410, 2021.

[13] W. Liu et al., "Monocular Depth Estimation Based on Deep Learning: A Survey," *Multimedia Tools and Applications*, vol. 81, pp. 18933–18966, 2022.

[14] M. Gupta and A. Rathore, "A Comprehensive Survey on Monocular Depth Estimation Techniques Using Deep Learning," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, pp. 1–15, 2023.

[15] K. Singh et al., "Recent Developments in Deep Learning for 3D Reconstruction from Single Images," *Journal of Imaging*, vol. 9, no. 2, pp. 1–25, Feb. 2023.

[16] Y. Zhang, M. Kumar, and R. S. Rana, "Transfer Learning-Based Approach for 3D Reconstruction from a Single 2D Image," *International Journal of Computational Vision and Robotics*, vol. 14, no. 1/2, pp. 75–86, 2023.