Neural Volumetric Representations for Real-Time 3D Scene Reconstruction Using Multi-Modal Learning Algorithm

Pidatala Devendrababu
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, India
pidataladevendrababu@gmail.com

Preeti Jha
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, India
preetijha@klh.edu.in

Abstract: *Deep Learning (DL) is a subfield of Machine Learning (ML) models used in various complex fields. DL algorithms are mostly widely used to reconstruct 3D images collected from multiple online sources. It is a very challenging task for the existing algorithms to reconstruct 2D images into 3D pictures without losing high-quality pixels because of the complex scenes with different lighting situations, dynamic components, and occlusions. This paper presents a novel real-time 3D scene reconstruction using neural volumetric representations combined with a Multi-Modal Learning Algorithm (MMLA). The proposed MMLA focuses on solving issues like volumetric representations of scenes, which are improved by combining numerous modalities such as RGB images, depth sensors, and Inertial Measurement Unit (IMU) data. The MMLA combines the DeepVoxels model and Neural Radiance Fields (NeRF) model, which it calls the Neural Rendering technique, to learn complex patterns in 3D scenes. The pre-trained model EfficientNet accurately obtained the 3D- reconstruction patterns and understood the spatial structures that transfer to the proposed MMLA. The proposed MMLA performance is analyzed using the ShapeNet dataset, which consists of 2D images. Finally, the experimental results show that the proposed MMLA outperforms the superior performance in terms of Mean Squared Error (MSE) of 0.167, Root Mean Squared Error (RMSE) of 0.50, and Mean Absolute Error (MAE) of 1.1. These results may differ from other datasets.*

Keywords: *Deepvoxels, neural rendering technique, NeRF, MMLA.*

Received February 28, 2025; accepted July 29, 2025
https://doi.org/10.34028/iajit/22/6/12

1. Introduction

Real-time 3D scene reconstruction is a significant technique that allows for the development of dynamic, three-dimensional visualizations of circumstances as they are evaluated [28]. The method involves gathering data from the natural environment via various sensors and then processing it to develop a 3D model that can be projected and altered in real time. Generally, 3D construction is implemented in multiple domains and applications that help recreate models like garments [8]; the volumetric garment rendering is parallel and extracted using a neural renderer. Traditional methods were highly reliant on mathematical concepts and multi-view stereo techniques. In this context, reconstructing 2D images into 3D images is tedious for the existing algorithms. The existing algorithms face several issues in the reconstruction of 2D images into 3D images, such as lack of missing depth shapes in the input image, some parts of the image are hidden in the 2D images, which leads to incomplete data, accurate ground truth is required to reconstruct the 3D image which is more difficult, there is a lot of domain gaps identified with the existing dataset such as ShapeNet, and finally, the reconstruction of tiny structures, tedious meshes is challenging for voxel models.

In recent years, advanced Deep Learning (DL) algorithms have been used to develop a better understanding of complex patterns and representations using neural networks [10]. In medical imaging, 3D reconstruction plays a significant role in diagnostics, treatment planning, and surgical decisions [2]. Traditional methods based on effectively reconstructing 2D images into 3D scenes have many drawbacks [36]. In some cases, the single-view reconstruction identified complicated the transformation of images into 2D images, which is the most expensive [11, 26]. On the other hand, detecting objects in occluded regions is also very difficult for existing models.

This paper mainly focused on combining two models, DeepVoxels and Neural Radiance Fields (NeRF). The proposed approach also focused on reconstructing multiple images. Firstly, it recreates the object present in the image. Secondly, it reconstructs the 3D shape present in the input image. The pre-trained model EfficientNet with transfer learning is used to train on ShapeNet by analyzing the 2D to 3D reconstruction of images based on depth maps, orientation of images, and object detection. The proposed approach also focused not only on image reconstruction but also on 3D object reconstruction. These proposed methods predict the camera position, analyze the particular shape from

the input image, and align the 3D shapes. The key contributions of this research work are given as:

- This work aims to reconstruct the 2D images into high-quality 3D scenes using advanced DL algorithms.
- The proposed approach focused on solving contrast variation, obstruction, and dynamic component issues.
- The proposed approach is a novel Multi-Modal Learning Algorithm (MMLA).
- The EfficientNet performs better when training with the ShapeNet dataset and extracting the spatial features.
- MMLA combines DeepVoxels and NeRF based on a neural rendering approach.
- Table 1 explains the following algorithms and its individual functionalities.

Table 1. List of algorithms and its functionalities used in 3D reconstruction.

Method	Functionality
EfficientNet (pre-trained)	The spatial structures extracted and transmit the patterns to MMLA.
MMLA	Merges RGB data, deep sensors, and Inertial Measurement Unit (IMUs) to increase the representation of volumetric scenes.
Neural rendering (DeepVoxels +NeRF)	The 3D scenes represented in a more lightweight and vivid form.
ShapeNet dataset	It helps to increase the learning of complex lighting and structural patterns for high-quality 3D reconstruction.

2. Literature Survey

Bernardini *et al*. [7] presented the Ball-Pivoting Algorithm (BPA) that computes the triangle-shaped mesh at a cloud point. The object is retrieved from the surface points using multiple scans. The obtained points form the triangle using the ball specified by the user. The ball plays a significant role in preventing the edges of the triangle from forming. The proposed algorithms are applied to BPA datasets to obtain the scans of complex 3D objects. Wang *et al*. [29] reviewed several 3D reconstruction models that are used for 3D images. Wang *et al*. [29] used several SLAM-based techniques categorized with metrics such as deep network factors, output initialization, datasets, and comparative analysis between various models. Han *et al*. [13] discussed several models based on Machine Learning (ML) and DL to redesign the 3D images. This article provides the literature on multiple algorithms that help reconstruct and convert the images into 3D images. The final results and analysis show the comparison between various DL algorithms demonstrated in this article. Choy *et al*. [9] proposed a novel Recurrent Neural Network (RNN) model that improves the 3D reconstructions. The proposed network helps to learn the mapping from images based on their shapes collected from synthetic data. The network selects one or more images to find the arbitrary viewpoints and outcomes that help reconstruct the object in a 3D occupancy grid. The results show that the proposed approach obtains high performance in 3D reconstruction. The existing models need to solve the issue of finding the accurate texture.

Zhang *et al*. [37] proposed the multi-view stereo network, which helps reconstruct the scene. The proposed Point-based Multi-View Stereo Network with Pyramid Attention (Point-MVSNet) helps generate high-quality 3D images by reconstructing scenes. The performance of the proposed approach is increased by designing the pyramid attention module that increases the 3D reconstruction of the image.

Seitz *et al*. [21] proposed the multi-view stereo algorithms and compared them with various existing algorithms. The proposed approach obtained the multi-view image datasets with a high positive rate. The algorithms are applied to six benchmark datasets and show the evaluation results.

Tatarchenko *et al*. [25] presented the deep Convolutional Neural Network (CNN) model that creates automated 3D outputs effectively utilizing the memory. The proposed approach mainly predicts the structure and tenancy of separate cells. It also makes significant 3D shapes with high-resolution outcomes. The proposed approach obtains the mean value for the algorithms with R2N2-0.560, OGN-0.596, and Dense-0.590. Samavati and Soryani [20] developed a novel approach that reconstructs the 3D image for the given 2D image. The proposed approach removes the key-factor detection and matching. The objects form the input image and redesign the shapes of the objects. The performance is improved using rapid techniques that increase the accuracy of 3D image construction. The performance metrics with mean class Accuracy (mAcc) for ISBNet is 76.1%, and the mean class Intersection over Union (mIoU) for Octree-based Convolutional Neural Network (OCNN) is 85.6%. Tulsiani *et al*. [27] presented the advanced learning CNN model that focused on detecting objects and segmenting input images. It helps to reconstruct the input image into a 3D image. The proposed approach is suitable for denoised photos, which helps increase the silhouette estimations learned from the 2D annotations in datasets. The proposed approach applied to the PASCAL 3D+ dataset that validates the final output.

Laga *et al*. [18] discussed various 2D and 3D vision issues handled by the DL algorithms. The stereo-based prediction is identified by using the VJV, which depicts a wide range of visits, indicating that the data is highly variable. SDT and CWN receive the fewest visits. They are integrating handmade characteristics across various images.

Zheng *et al*. [38] proposed the Parametric Model-Conditioned Implicit Representation (PaMIR), which integrates the dynamic body model with an uncontrolled deep latent functionality. The proposed PaMIR regularises the free-form deep implicit function by leveraging the parametric model's semantic features,

enhancing adaptation capability under challenging positions and various apparel configurations. The training loss is also used to overcome depth ambiguities, resulting in effective surface detail recovery with poor body connection. Finally, the organism reference optimisation approach is applied to improve the precision and uniformity of parametric model estimate using the implicit function. Experimental findings reveal that our method achieves innovative image-based 3D human restoration performance in challenging situations and garment types.

Yu *et al*. [35] presented the new 3D reconstruction model that recreates the 1D RGB images. The proposed approach adopted the pixel-aligned features. The proposed approach's limitations focused on acquiring the voxel-aligned elements from the input image. It also retrieves the fine-tuned aligned features from an accurate cloud point. Finally, the results show that the algorithm applied on ShapeNet dataset and achieved the high accuracy on 1D image. Table 2 explains the performance of various algorithms

Table 2. The performance of several algorithms on reconstruction of 3D images.

Authors	Proposed approach	Dataset	Performance
Xie *et al*. [32]	Pix2vox++	ShapeNet, Pix3D, and Things3D	IoU-0.670, 0.436, 0.430
Choy *et al*. [9]	3D-R2N2	PASCAL 3D and ShapeNet	IoU-0.634
Banani *et al*. [5]	Novel approach	ShapeNet	IoU-0.82
Gwak *et al*. [12]	GAN	ShapeNet	0.62
Bautista *et al*. [6]	Inductive biases encoded	ShapeNet	IoU-0.749
Rezende *et al*. [19]	Deep Generative Model (DGM)	ShapeNet	IoU-0.751
Tatarchenko *et al*. [24]	Feed-forward network	ShapeNet dataset	Average error for nomral image-0.0057, depth-0.0207
Worrall *et al*. [30]	Encoder-decoder networks	Basel face dataset	Test error-2.14
Isola *et al*. [15]	Conditional Adversarial Networks (CAN)	Cityscapes dataset	Class IOU-0.29

3. Methodology

In the methodology section, the methods and algorithms are briefly described, along with the mathematical models. The subsections explain each algorithm clearly, each accompanied by a diagram. Firstly, DeepVoxels for the 3D Reconstruction Method is explained in section 3.1. Section 3.2 introduced NeRF; section 3.3 presented the volumetric scene function; section 3.4 described the volume rendering equation; section 3.6 proposed the neural network architecture; and section 3.7 provided the dataset description.

3.1. DeepVoxels for 3D Reconstruction Method

DeepVoxels is a 3D reconstruction technique that transforms 2D images into deep 3D photos. The proposed DeepVoxels aims to develop a 3D model that can be analyzed and operated from multiple perspectives [23, 34]. Figure 1 describes the components used and elaborates on the flow of DeepVoxels. This technique is most potent in computer vision, Augmented Reality (AR), and Virtual Reality (VR) applications. Representing 3D space in voxel grids A voxel grid representation is popular in various tasks for representing an object or scene in full 3D, with each voxel value (volumetric pixel) corresponding to a region in 3D space and holding information about the properties within 3D space. In this context, a 3D-based pixel grid is used in 2D images [14, 22]. The information, such as color and occupancy, is stored in a voxel grid, and then a 3D scene is reconstructed from this voxel grid. It represents the view-dependent appearance of a 3D scene without explicitly modeling its geometry. A Vertical 3D grid is trained on different constant aspects using the basic 3D scene structure.

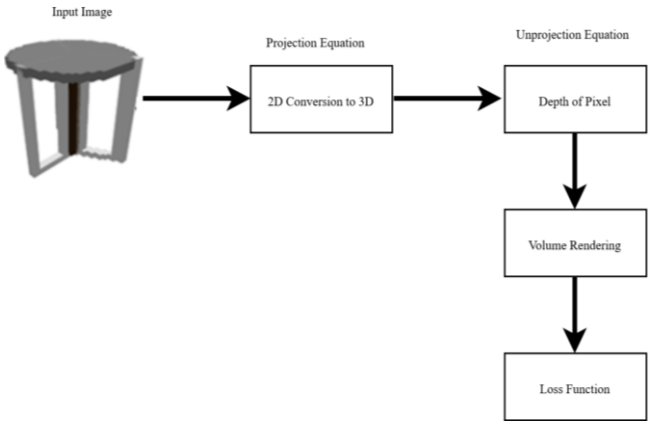

Figure 1. Process of deepvoxels.

It is supervised, does not require a 3D reconstruction of the scene, and employs a 2D re-rendering loss to enforce context and multi-view topology in a logical manner.

- **Projection equation**: to project a 3D point onto a 2D image plane, we use:

$$X = [A, B, C, 1]^T \quad (1)$$

$$x = KRX \quad (2)$$

Where $x=[u, v, w]^T$ the projected point in homogeneous coordinates, K is is the intrinsic matrix of the camera, and R is the extrinsic matrix (rotation and translation).

- **Unprojection equation**: to unproject a 2D pixel $x=[u,v]^T$ into 3D space, given depth

$$A = dK^{-1}x \quad (3)$$

- **Volume rendering**: the color of a pixel in the novel view can be computed using volume rendering:

$$C(r) = \int_{t_n}^{t_f} \tau(t)c(t)exp\,(-\int_{t_n}^{t} \tau(s)ds)dt \quad (4)$$

Here, the $C(r)$ represents the colour along a ray r is, $\tau(t)$ is the density along the ray, and $c(t)$is the color at position t.

- **Loss function**: the reconstruction loss used to train the network is typically based on the difference between the predicted image and the ground truth image:

$$\mathcal{L} = \sum_{a,b} ||I_{pred}(a,b) - I_{gt}(a,b)||^2 \quad (5)$$

Where I_{pred}- predicted image and I_{gt} is the ground truth image.

3.2. Neural Radiance Fields (NeRF)

NeRF is an innovative 3D scene representation and reconstruction approach that has gained significant attention in recent years. Figure 2 describes the step-by-step process for 3D reconstruction [1]. NeRF utilizes DL techniques to render highly realistic images from sparse sets of 2D photographs. Unlike traditional 3D reconstruction methods, NeRF incorporates scene dimensions and appeal into a neural network, allowing for the formation of fresh views of a scene with high accuracy [31]. It models the scene as a continuous volumetric field, where each point in 3D space emits light (radiance) in different directions. In this context, the radiance field is represented by neural networks that select the pairs of 3D axes and view the directions and outcomes of the ray samples with RGB color and density values [3, 4, 17, 33]. The 2D images are generated by using radiance fields that integrate the long rays using volumetric rendering techniques. This procedure simulates light traveling through the scene and combining the color and opacity to create pixel values. The difference between the rendered images and input photographs is reduced, and the network parameters are optimized to learn the implicit representation of the scene. This neural network encodes the scene's geometry and appearance and allows high-quality rendering from any viewpoint.

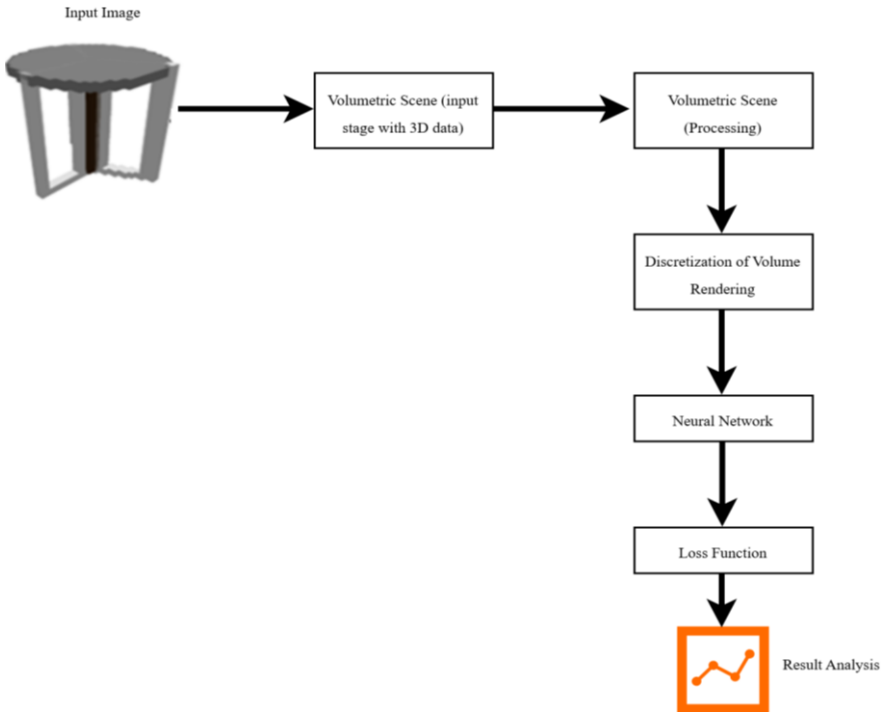

Figure 2. Process steps for NeRF.

3.3. Volumetric Scene Function

The input scene is initializes continuously with the function F0 which arrange the 3D position a=(a, b, c) and a 2D direction d=(θ, ∅) and outcomes the colour c=(r, g, b) and σ represents the density of volume:

$$F_\theta: (x, d) \rightarrow (c, \sigma) \quad (6)$$

3.4. Volume Rendering Equation

Initialize the ray obtained from camera $r(t)=o+td$ where o represents the origin of rays and d represents the direction of ray, the color $C(r)$ of the pixel is measured by using volume rendering:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt \quad (7)$$

Where $T(t)$ the cumulative transmission with the ray from t_n to t:

$$T(t) = exp\,(-\int_{t_n}^{t} \sigma(r(s))ds) \quad (8)$$

In practice, the continuous integral is approximated using quadrature, specifically a discrete sum over sampled points:

$$C(r) \approx \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))c_i \quad (9)$$

Where: δ_i is the distance between adjacent sample points.

T_i is the transmittance from the origin to the i^{th} sample.

3.5. Neural Network Architecture

The neural network F_θ is typically a Multi-Layer Perceptron (MLP) that takes as input the position x and direction d. Positional encoding is often applied to the inputs to enable the network to learn high-frequency functions:

$$\gamma(a) = (sin(2^0\pi x), cos(2^0\pi x), \ldots., sin(2^{L-1}\pi x), cos(2^{L-1}\pi x)) \quad (10)$$

3.6. Loss Function

The model undergoes training with a loss function that reduces the variance within the projected colour $C(r)$ and the ground truth colour C_{gt} for each ray.

$$\mathcal{L} = \sum_{r\in R} ||C(r) - C_{gt}(r)||_2^2 \quad (11)$$

To summarize, NeRF uses a neural network to model a 3D scene by learning a volumetric representation from a set of 2D images. The network predicts the color and density at any 3D point and viewing direction, which are then used to render images from novel viewpoints via volume rendering. This involves:

- Defining a neural network that predicts colour and density.
- The volume rendering is used to measure the colour rays transferring into the scene.
- Training the network using the discrepancy between rendered and ground truth images.

This approach enables the creation of detailed and realistic 3D reconstructions from 2D images.

3.7. Dataset Description

It is a well-annotated database for 3D shapes of general objects. In ShapeNet, these are object classes like tables or chairs that serve as important mini tasks to many different computer vision or ML problems, such as 3D shape recognition, reconstruction, or segmentation. In all, there are 13255 images for experimentation [4]. Of these, 7000 images were used for training and 6255 for testing. Figure 3 shows the sample images of ShapreNet dataset.

Figure 3. ShapeNet dataset table and chairs images.

4. Performance Metrics

The performance of proposed 3D image reconstruction is mainly based on various scenarios that analyze the model's strength. These algorithms are most widely implemented using the Python language, 16 GB RAM, 1 TB hard drive, and an Intel I7 processor to handle the overhead issues in the system. The 2D input image contains several occlusions that appear at the time of input processing. The following parameters show the quality of the final output image.

4.1. Accuracy Metrics

The parameters compared with reconstructed model with a ground truth.

- **Mean Squared Error (MSE):** this is the parameter that shows the difference between similar points in predicted (reconstructed) and actual label.

$$MSE = \frac{1}{x \cdot y}\sum_{a=1}^{x}\sum_{b=1}^{y}[I(a,b) - K(a,b)]^2 \quad (12)$$

- **Root Mean Squared Error (RMSE)**: it provides the error magnitude.

$$RMSE = \sqrt{\frac{1}{a}\sum_{a=1}^{x}(y_a - \hat{y}_a)^2} \quad (13)$$

- **Mean Absolute Error (MAE)**: the average of the quantitative variations among each point.

$$MAE = \frac{1}{a}\sum_{a=1}^{x}(y_a - \hat{y}_a) \quad (14)$$

5. Ablation Study

Table 3 compares various DL algorithms by their 3D reconstruction performance on the sofa. Three typical error measures, MSE, RMSE, and MAE, are employed to measure the reconstruction quality; they are used as quality metrics to evaluate the performance; the lower, the better. The Convolutional Network (CN) and Encoder-Decoder Networks (EDN) obtain higher error rates than the other models; this represents low reconstruction ability. +AMask+SFB performs better and has smaller error values, implying its 3D output is more detailed. Notably, the MMLA model outperforms the other methods in terms of three metrics, MSE (0.167), RMSE (0.50), and MAE (1.1), indicating that the MMLA model has better performance in generating accurate and good-quality 3D sofa reconstructions. It demonstrates that MMLA can reduce structural errors and improve visual realism in 3D modeling. Finally, the Figure 4 shows the visualization graph for comparison of existing algorithms with proposed approach.

Table 3. Performance of algorithms based on 3D reconstruction quality (conversion of sofa).

	MSE	RMSE	MAE
CN [9]	0.431	0.101	3.3
EDN [5]	0.347	0.991	2.3
+AMask + SFB [16]	0.301	0.78	1.9
MMLA	0.167	0.50	1.1

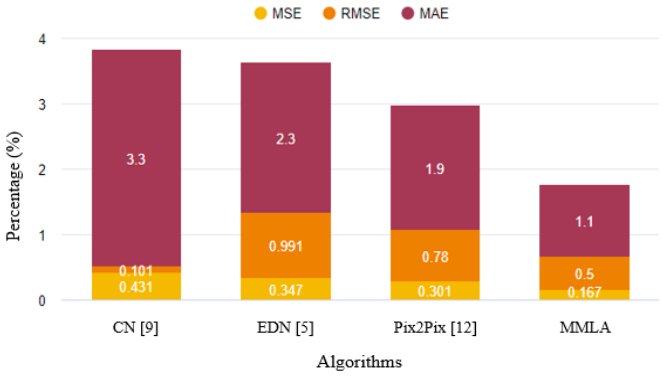

Figure 4. Comparison between algorithms based on reconstruction quality (conversion of sofa).

Table 4 compares algorithms regarding the quality of 3D reconstruction, which evaluates the following performances. The CN obtains the low performance in terms of (MSE: 0.443, RMSE: 0.99, MAE: 3.6) and reconstructs the low-quality images of the models. The error values are reduced, and the EDN's performance is better, recommending a better ability to learn spatial features. The +AMask+SFB [16] is a GAN-style reconstruction model that obtains better performances in terms of MAE (1.6) that reflects high visual features. Finally, the MMLA outperforms high performance, representing the lowest errors (MSE: 0.143, RMSE: 0.51, MAE: 0.99), representing high learning accuracy and high-quality 3D reconstructions. Finally, the Figure 5 shows the visualization graph for comparison of existing algorithms with proposed approach.

Table 4. Performance of algorithms based on 3D reconstruction quality (conversion of table).

	MSE	RMSE	MAE
Convolutional Network [9]	0.443	0.99	3.6
Encoder-Decoder Networks [5]	0.352	0.88	2.4
+AMask + SFB [16]	0.312	0.77	1.6
MMLA	0.143	0.51	0.99

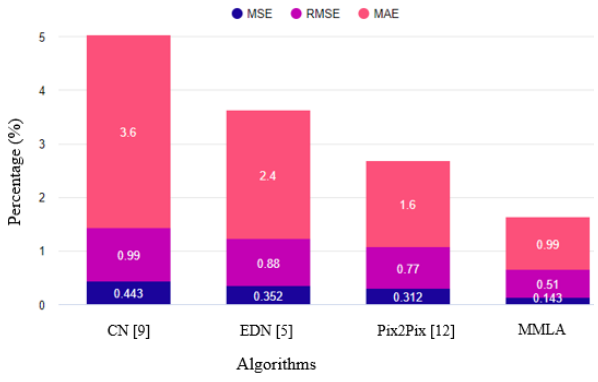

Figure 5. Comparison between algorithms based on reconstruction quality (conversion of table).

Table 5 explains a comparison of various algorithms utilized for the 3D reconstruction of the chair images and shows that the CN achieves the most significant error values (MSE: 0.398, RMSE: 0.98, MAE: 3.23), which leads to low accuracy of reconstruction. The EDN performs moderately better, getting lower errors (MSE: 0.292, RMSE: 0.87, MAE: 2.23), while the better performance of +AMask+SFB [16] (MSE: 0.241, RMSE: 0.74, MAE: 1.57). Nevertheless, the performances of the proposed MMLA model were best in all indices, and the error rate was far lower (MSE: 0.131, RMSE: 0.52, and MAE: 0.987). These results confirm that MMLA can more accurately capture the structural and spatial details of the 3D chair pattern models with better quality than existing methods. Finally, the Figure 6 shows the visualization graph for comparison of existing algorithms with proposed approach.

Table 5. Performance of algorithms based on 3D reconstruction quality (conversion of chair).

	MSE	RMSE	MAE
CN [9]	0.398	0.98	3.23
EDNs [5]	0.292	0.87	2.23
+AMask + SFB [16]	0.241	0.74	1.57
MMLA	0.131	0.52	0.987

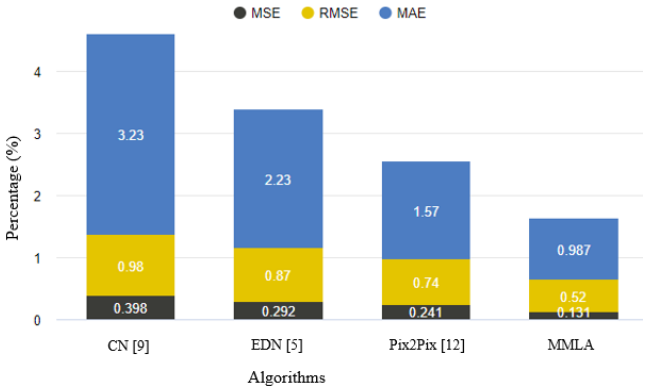

Figure 6. Comparison between algorithms based on reconstruction quality (conversion of chair).

6. Conclusions

Neural representations of volumes (e.g., DeepVoxels, neural rendering technique, and NeRF) have revolutionized the real-time 3D scene representation mechanisms through unconventional approaches. These approaches use DL to learn how to scan and capture realistic 3D scenes from sparse or dense multimodal sets of observed data, like images or videos. In this context, the NeRF is one of the effective 3D reconstruction models focused on geometry and appearance at a refined stage. It is one of the practical features for applications that require accurate rendering patterns. Multidimensional data like images and depth maps significantly increase the strength and perfectness of 3D reconstruction. The proposed approach alleviates the data sparsity and enhances the reconstruction quality under challenging scenarios. It allows the rendering and reconstruction of 3D scenes in real-time or near real-time, which makes these methods also useful for interactive work, including VR, AR, and gaming. DeepVoxels and neural rendering techniques provide scalable solutions for complex scenes and varied object shapes. The performance of MMLA is measured using an MSE of 0.167, RMSE of 0.50, and MAE of 1.1 for

the sofa-type images. For the table and chair type images, the performance is MSE of 0.143 and 0.131, RMSE of 0.51 and 0.52, and MAE of 0.99 and 0.987. Furthermore, these techniques demonstrate good generalization performance over multiple scenes and datasets.

References

[1] Abate D., Themistocleous K., and Hadjimitsis D., "The Application of Neural Radiance Fields (NeRF) in Generating Digital Surface Models from UAV Imagery," *in Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, Athens, pp. 10228-10231, 2024. https://ieeexplore.ieee.org/document/10641392

[2] Ahmad B., Floor P., Farup I., and Hovde O., "3D Reconstruction of Gastrointestinal Regions Using Single-View Methods," *IEEE Access*, vol. 11, pp. 61103-61117, 2023. https://ieeexplore.ieee.org/document/10154004

[3] Ahmed M., Alazeb A., Al Mudawi N., Sadiq T., and et al., "Perception of Natural Scenes: Objects Detection and Segmentations Using Saliency Map with AlexNet," *The International Arab Journal of Information Technology*, vol. 22, no. 3, pp. 461-475, 2025. https://doi.org/10.34028/iajit/22/3/4

[4] Anciukevicius T., Xu Z., Fisher M., Henderson P., and et al., "RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, pp. 12608-12618, 2023. DOI: 10.1109/CVPR52729.2023.01213

[5] Banani M., Corso J., and Fouhey D., "Novel Object Viewpoint Estimation through Reconstruction Alignment," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 3110-3119, 2020. DOI: 10.1109/CVPR42600.2020.00318

[6] Bautista M., Talbott W., Zhai S., Srivastava N., and Susskind J., "On the Generalization of Learning-based 3D Reconstruction," *arXiv Preprint*, vol. arXiv:2006.15427v1, pp. 1-10, 2020. https://arxiv.org/abs/2006.15427

[7] Bernardini F., Mittleman J., Rushmeier H., Silva C., and Taubin G., "The Ball-Pivoting Algorithm for Surface Reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, no. 4, pp. 349-359, 1999. DOI: 10.1109/2945.817351

[8] Chen Y., Xie R., Yang S., Dai L., and et al., "Single-View 3D Garment Reconstruction Using Neural Volumetric Rendering," *IEEE Access*, vol. 12, pp. 49682-49693, 2024. DOI: 10.1109/ACCESS.2024.3380059

[9] Choy C., Xu D., Gwak J., Chen K., and Savarese S., "3D-R2N2: A Unified Approach for Single and Multi-View 3D Object Reconstruction," *in Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, pp. 628-644, 2016. https://link.springer.com/chapter/10.1007/978-3-319-46484-8_38

[10] Farshian A., Gotz M., Cavallaro G., Debus C., and et al., "Deep-Learning-based 3-D Surface Reconstruction-a Survey," *Proceedings of the IEEE*, vol. 111, no. 11, pp. 1464-1501, 2023. DOI: 10.1109/JPROC.2023.3321433

[11] Gotz M., Cavallaro G., Geraud T., Book M., and Riedel M., "Parallel Computation of Component Trees on Distributed Memory Machines," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 11, pp. 2582-2598, 2018. DOI: 10.1109/TPDS.2018.2829724

[12] Gwak J., Choy C., Chandraker M., Garg A., and Savarese S., "Weakly Supervised 3D Reconstruction with Adversarial Constraint," *in Proceedings of the International Conference on 3D Vision*, Qingdao, pp. 263-272, 2017. https://ieeexplore.ieee.org/document/8374579

[13] Han X., Laga H., and Bennamoun M., "Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1578-1604, 2021. DOI: 10.1109/TPAMI.2019.2954885

[14] Huang Y., Huang S., Hsu H., and Wang Y., "Interpreting Latent Representation in Neural Radiance Fields for Manipulating Object Semantics," *in Proceedings of the IEEE International Conference on Image Processing*, Kuala Lumpur, pp. 470-474, 2023. https://ieeexplore.ieee.org/document/10222650

[15] Isola P., Zhu J., Zhou T., and Efros A., "Image-to-Image Translation with Conditional Adversarial Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 5967-5976, 2017. https://ieeexplore.ieee.org/document/8100115

[16] Jo S., Lee D., and Rhee C., "Occlusion-Aware Amodal Depth Estimation for Enhancing 3D Reconstruction from a Single Image," *IEEE Access*, vol. 12, pp. 106524-106536, 2024. https://doi.org/10.1109/access.2024.3436570

[17] Ko K., Kim S., and Lee M., "Zero-Shot 3D Scene Representation with Invertible Generative Neural Radiance Fields," *IEEE Access*, vol. 13, pp. 68561-68576, 2025. https://ieeexplore.ieee.org/document/10967257

[18] Laga H., Jospin L., Boussaid F., and Bennamoun M., "A Survey on Deep Learning Techniques for Stereo-based Depth Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1738-1764, 2022. https://ieeexplore.ieee.org/document/9233988

[19] Rezende D., Ali Eslami S., Mohamed S., Battaglia

P., and et al., "Unsupervised Learning of 3D Structure from Images," *in Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, pp. 5004-5011, 2016. https://dl.acm.org/doi/10.5555/3157382.3157656

[20] Samavati T. and Soryani M., "Deep Learning-based 3D Reconstruction: A Survey," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 9175-9219, 2023. https://doi.org/10.1007/s10462-023-10399-2

[21] Seitz S., Curless B., Diebel J., Scharstein D., and Szeliski R., "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, pp. 519-528, 2006. https://ieeexplore.ieee.org/document/1640800

[22] Shan Y., Liang C., and Xu M., "3D Reconstruction and Estimation from Single-View 2D Image by Deep Learning-A Survey," *in Proceedings of the IEEE Conference on Artificial* Intelligence, Singapore, pp. 1-7, 2024. DOI: 10.1109/CAI59869.2024.00010

[23] Sitzmann V., Thies J., Heide F., Niebner M., and et al., "DeepVoxels: Learning Persistent 3D Feature Embeddings," *arXiv Preprint*, vol. arXiv:1812.01024v2, pp. 1-10, 2018. https://arxiv.org/abs/1812.01024

[24] Tatarchenko M., Dosovitskiy A., and Brox T., "Multi-View 3D Models from Single Images with a Convolutional Network," *arXiv Preprint*, vol. arXiv:1511.06702v2, pp. 1-20, 2016. https://arxiv.org/abs/1511.06702

[25] Tatarchenko M., Dosovitskiy A., and Brox T., "Octree Generating Networks: Efficient Convolutional Architectures for High-Resolution 3D Outputs," *in Proceedings of the IEEE International Conference on Computer Vision*, Venice, pp. 2107-2115, 2017. https://ieeexplore.ieee.org/document/8237492

[26] Tian Y., Zhang H., Liu Y., and Wang L., "Recovering 3D Human Mesh from Monocular Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, 15406-15425, 2023. DOI: 10.1109/TPAMI.2023.3298850

[27] Tulsiani S., Kar A., Carreira J., and Malik J., "Learning Category-Specific Deformable 3D Models for Object Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 719-731, 2017. https://ieeexplore.ieee.org/document/7482798

[28] Vinodkumar P., Karabulut D., Avots E., Ozcinar C., and Anbarjafari G., "Deep Learning for 3D Reconstruction, Augmentation, and Registration: A Review Paper," *Entropy*, vol. 26, no. 3, pp. 1-44, 2024. https://doi.org/10.3390/e26030235

[29] Wang C., Reza M., Vats V., Ju Y., and et al., "Deep Learning-based 3D Reconstruction from Multiple Images: A Survey," *Neurocomputing*, vol. 579, pp. 128018, 2024. https://doi.org/10.1016/j.neucom.2024.128018

[30] Worrall D., Garbin S., Turmukhambetov D., and Brostow G., "Interpretable Transformations with Encoder-Decoder Networks," *in Proceedings of the IEEE International Conference on Computer Vision*, Venice, pp. 5737-5746, 2017. https://ieeexplore.ieee.org/document/8237873

[31] Wu D., Li Y., Yang R., Li S., and et al., "Neural Radiance Field Reconstruction Technique Under Layer Training Strategy," *in Proceedings of the International Conference on HVDC*, Urumqi, pp. 747-750, 2024. https://ieeexplore.ieee.org/document/10723007

[32] Xie H., Yao H., Zhang S., Zhou S., and Sun W., "Pix2Vox++: Multiscale Context-Aware 3D Object Reconstruction from Single and Multiple Images," *International Journal of Computer Vision*, vol. 128, pp. 2919-2935, 2020. https://link.springer.com/article/10.1007/s11263-020-01347-6

[33] Yang J., Zhang G., Li Y., and Yang L., "VST3D-Net: Video-based Spatio-Temporal Network for 3D Shape Reconstruction from a Video," *in Proceedings of the International Conference on 3D Immersion*, Brussels, pp. 1-7, 2020. https://ieeexplore.ieee.org/document/9376350

[34] Yang L., Yang C., Xie R., Liu J., and et al., "3D Reconstruction from Traditional Methods to Deep Learning," *in Proceedings of the IEEE 10th International Conference on Cyber Security and Cloud Computing, and 9th International Conference on Edge Computing and Scalable Cloud*, Xiangtan, pp. 387-392, 2023. https://ieeexplore.ieee.org/document/10195547

[35] Yu X., Tang J., Qin Y., Li C., and et al., "PVSeRF: Joint Pixel-, Voxel- and Surface-Aligned Radiance Field for Single-Image Novel View Synthesis," *in Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa, pp. 1572-1583, 2022. https://doi.org/10.1145/3503161.3547893

[36] Zhang J., Dong Y., Kuang M., Liu B., and et al., "The Art of Defense: Letting Networks Fool the Attacker," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3267-3276, 2023. https://ieeexplore.ieee.org/document/10130393

[37] Zhang K., Liu M., Zhang J., and Dong Z., "PA-MVSNet: Sparse-to-Dense Multi-View Stereo with Pyramid Attention," *IEEE Access*, vol. 9, pp. 27908-27915, 2021. https://ieeexplore.ieee.org/document/9352763

[38] Zheng Z., Yu T., Liu Y., and Dai Q., "PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human

Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3170-3184, 2022. https://ieeexplore.ieee.org/document/9321139

Pidatala Devendrababu has completed his B.Tech from Anurag Engineering College, Kodad, and M.Tech from MITS College, Kodad. He is currently pursuing his Ph.D. at KL University. With over 12 years of experience as an Assistant Professor, he has served in various engineering colleges under the Jawaharlal Nehru Technological University, Hyderabad (JNTUH). His primary areas of interest include Artificial Intelligence (AI) and Machine Learning (ML), Natural Language Processing (NLP), and Deep Learning. His academic and professional journey reflects a deep commitment to teaching and research in the field of intelligent computing. He can be contacted at email: pidataladevendrababu@gmail.com.

Preeti Jha is an Associate Professor in the Department of Computer Science and Engineering at KLH Hyderabad. She earned her Ph.D. and completed post-doctoral research under the guidance of Dr. Aruna Tiwari at IIT Indore. Her research focuses on Fuzzy Clustering, Data Mining, Machine Learning, and Large-Scale Genomic Data Analysis. She is actively involved in academic research and contributes to advancements in soft computing techniques. Dr. Jha is also a member of the Soft Computing Research Society, reflecting her dedication to interdisciplinary research and collaboration within the field of intelligent data analysis and computational methodologies. She can be contacted at email: preetijha@klh.edu.in.