IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3149-3158
Journal homepage: http://iieta.org/journals/ts

Pixel to 3D Voxel Reconstruction Using Octree-Based Network and Deep Learning

Naif Al Mudawi[1], Hamid Ashfaq[2], Abdulwahab Alazeb[1], Nouf Abdullah Almujally[3], Asaad Algarni[4], Khaled Al Nowaiser[5], Ahmad Jalal[2,6*]

[1] Department of Computer Science, College of Computer Science and Information System, Najran University, Najran 55461, Saudi Arabia
[2] Department of Computer Science, Air University, Islamabad 44000, Pakistan
[3] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[4] Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha 91911, Saudi Arabia
[5] Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[6] Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, South Korea

Corresponding Author Email: ahmadjalal@mail.au.edu.pk

https://doi.org/10.18280/ts.420610

Received: 2 April 2024
Revised: 25 September 2024
Accepted: 25 January 2025
Available online: 31 December 2025

Keywords:
2D to 3D reconstruction, 3D reconstruction algorithms, 3D voxel reconstruction, 3D voxelization computer vision for 3D, depth estimation, object detection, single view image processing, and voxel rendering

ABSTRACT

Innovative technology and improvements in intelligent systems have revolutionized the way data is processed, analyzed, and reconstructed in modern times. With the advancements in machine learning and artificial intelligence, it has become possible to reconstruct missing or lost information from a given set of data. It is very challenging to reconstruct data that has been lost during different processes like image acquisition, and dimension reduction. This paper proposed an organized method that has taken a 2D single image of objects and then predicts and reconstructs a voxel-based 3D of that object. A multi-layer encoder-decoder framework has been employed to estimate image depth, which is then combined with the original image to reconstruct the 3D shape of the object. Subsequently, methods such as EfficientNet and octree-based techniques are utilized to generate voxel representations of the 3D structure. For the experiments, three benchmark datasets were utilized alongside state-of-the-art methods for comparison. To evaluate the performance of the proposed model, metrics such as Chamfer Distance (CD), Earth Mover's Distance (EMD), and Intersection over Union (IoU) were employed. We have achieved mean CDs of 0.00387, 0.00317 and 0.00102 on ShapeNetCore, Pascal3D and Pix3D respectively.

1. INTRODUCTION

Creating a 3D interpretation from a single RGB image entails extracting the 3D structural information of a scene from its 2D representation. This process aims to empower robots with the ability to perceive and interact with their surroundings in a way that mimics human spatial understanding. Recent advancements in computer vision have introduced deep learning models capable of estimating 3D depth from 2D images. These models learn the relationship between images and their corresponding depth maps, allowing them to predict the depth and distance of objects accurately.

There are many image-based techniques used in computer vision. For example, Zhang et al. [1] proposed a novel approach for reconstructing a 3D model of a dynamic environment without the use of sensors. The method uses monocular video frames and incorporates a deep learning model for accurate reconstruction. Different multi-view approaches have been used in the past, such as that of Snow [2], which was based on a set of noisy depth measurements; this problem was resolved using graph cuts.

Wen et al. [3] have proposed an approach based on multi-camera views. This method formulates a volumetric fusion problem that integrates information from multiple views into a single consistent 3D model. Shu et al. [4] proposed a generative-model-based approach. The aforementioned methods used single or multiple RGB views of objects.

Wen and Cho [5] proposed a combined deep learning approach that used a recurrent neural network (RNN) with a 3D convolutional neural network (3D-CNN) to reconstruct objects in 3D. The RNN was used to generate a volumetric representation of the object by sequentially processing the 2D images, while the 3D-CNN was used to refine the generated 3D representation and produce a high-quality 3D model.

The proposed method in this research relies on a single RGB image. Later on, Savkin et al. [6] extracted neural and depth features, which contributed to the reconstruction of a 3D model that was further represented in the form of 3D voxels by Liu et al. [7]. For depth estimation, a deep neural network was used, which helped estimate depth maps from RGB

images. Finally, depth information was combined with RGB inputs in the octree-based neural network [8].

The extensive research achievements of this study are as follows:

• 3D visualization of unseen objects from their image data.

• Automatic 3D reconstruction of objects from a single image.

• An improved learning method compared to previous approaches in the field of e-learning and virtual-environment-based training.

• Utilization of EfficientNet as a backbone for feature extraction, known for its balance between high accuracy and low computational cost through compound scaling.

• An octree-based representation is employed to efficiently manage voxel data, reducing memory consumption and accelerating computations by focusing on non-empty regions of the 3D space.

• The combination of EfficientNet and octree-based representation enables the model to handle large-scale 3D data more effectively and at a finer resolution compared with traditional voxel grids.

• Enhanced disease diagnosis using 3D visualization of human organs reconstructed from CT scans.

The structure of this article is organized as follows: the paper begins with the introduction, followed by the related work section. Next, the proposed system is detailed in the materials and methods section, after which the results and performance evaluations are presented, including comparisons with different state-of-the-art methods. Finally, the paper concludes with the conclusion section.

2. RELATED WORK

Many researchers have proposed methods for 3D reconstruction from single-view images. Shu et al. [4] introduced Pix2Vox, a framework capable of handling both single-view and multi-view 3D reconstruction. This approach incorporates an encoder-decoder architecture, a context-aware fusion module, and a refinement stage, achieving higher accuracy and consistency than many existing methods. Additionally, it offers faster inference times and exhibits strong generalization capabilities. The task of recovering 3D representations of objects from single-view or multi-view RGB images using deep neural networks has gained significant attention in recent research. Traditional approaches, such as 3D-R2N2, rely on recurrent neural networks (RNNs) to sequentially integrate feature maps extracted from input images. However, these methods often face challenges such as inconsistent results and memory limitations. Renat and Imangali [9] proposed an end-to-end network for efficient 3D model generation from a single image. This network consists of an encoder, a 2D–3D fusion module, and a decoder, which together produce detailed point clouds from single-object images and retrieve the most similar shapes from the ShapeNetCore dataset. The method demonstrates state-of-the-art performance when compared to volumetric and point-set generation techniques, particularly excelling in capturing intricate details. Additionally, it performs well in environments with complex backgrounds and across diverse viewpoints.

Bae et al. [10] proposed a GAN-based approach for predicting voxel models from a single view. Their method utilizes the alignment of 2D silhouettes and slices within a camera frustum to reconstruct voxel representations of scenes containing multiple object instances. This approach demonstrates excellent performance in reconstructing complex scenes with non-rigid and multi-object configurations.

Reconstructing 3D objects from multiple 2D images has been a common focus in computer vision research. However, Huang et al. [11] addressed a more challenging problem: estimating 3D locations and shapes of multiple objects from just a single 2D image. Unlike prior approaches that either predict a single 3D property or focus exclusively on individual objects, their method employs a comprehensive framework. This includes learning a 3D voxel grid from the input image, utilizing CenterNet-3D for keypoint detection, and applying a coarse-to-fine reconstruction module to achieve efficient and detailed 3D reconstructions. Their approach proved effective for both single- and multi-object scenarios. In another advancement, Yuan et al. [12] presented the Voxel Transformer (VoTr), a novel voxel-based Transformer backbone for 3D object detection using point clouds. By incorporating self-attention mechanisms, VoTr overcomes the limitations of conventional 3D convolutional backbones, such as constrained receptive fields, thereby enabling the modeling of long-range voxel relationships.

Some research studies are based on images captured with sensors such as Light Detection and Ranging (LiDAR) and depth cameras. Kuang et al. [13] proposed a Voxel-CRF model for 3D scene understanding by integrating a voxel-based representation with a conditional random field model to infer semantic labels and object instances in indoor scenes. Shi et al. [14] employed an automated approach for large-scale 3D scene reconstruction of urban areas using LiDAR sensors. They created a meshed representation of a 3.7 km-long area with high detail and no user intervention, and investigated the effects of sensor models on reconstruction quality.

Visualization of 3D objects can be represented in various forms, including volumetric representations using voxels (Liu et al. [7]), mesh representations (Tahir et al. [15]), and point clouds (Ji et al. [16]). Yasir and Ahn [17] investigated different approaches to 3D object shape representation, focusing on surface-based and volumetric methods, as well as viewer-centered versus object-centered reference frames in single-view 3D shape prediction. Their analysis revealed that surface-based techniques perform better than voxel-based representations for novel objects, whereas voxel representations are more effective for familiar objects.

Gbadago et al. [18] introduced a framework known as Hierarchical Surface Prediction (HSP), which utilizes convolutional neural networks (CNNs) to generate high-resolution voxel grids. Their findings indicated that HSP produces more accurate results compared to low-resolution predictions, regardless of the input format.

Traditional methods often use convolutional neural networks (CNNs) such as ResNet, VGG, or other backbone architectures that might not be as parameter-efficient as EfficientNet. Many state-of-the-art approaches rely on regular voxel grids or point-cloud representations. Voxel grids suffer from substantial memory consumption at higher resolutions, while point-cloud methods might lack the explicit structure required for detailed 3D reconstruction. Methods such as 3D-R2N2 or AtlasNet employ RNNs or mesh-based approaches that focus on sequential data processing or direct mesh generation but might not be as efficient in capturing high-resolution details as octree-based approaches.

3. PROPOSED SYSTEM METHODOLOGY

In this section, a discussion about the main idea of our hypothetical methodology has been done for the 3D reconstruction of an object from its RGB image using depth feature predictor and 3D volumetric representation in the form of voxels. Figure 1 represents our method's main flow architecture diagram. The input of this system is in the form of an RGB image and the output is in the form of a voxels 3D model. The first applied deep learning neural network was to estimate the depth of the respective object in the image and also applied background removal preprocessing on the image. Later on, used the depth feature with RGB image to generate an octree-based 3D mesh and then voxelization of that mesh using EfficientNet to reconstruct 3D voxel representation. This system is the alternative to the RGBD input-based system that works using RGB images.

Algorithm 1 outlines a robust method employed in our research for estimating object depth, which plays a crucial role in 3D reconstruction. While depth sensors are traditionally required to capture depth information, this algorithm leverages training on the NYUv2 dataset to predict object depth directly from images. This predictive capability enables accurate estimation of the z-axis, significantly enhancing the 3D reconstruction process by eliminating the dependency on external depth-sensing hardware. The approach demonstrates efficiency and reliability, making it a valuable tool for depth estimation in diverse scenarios.

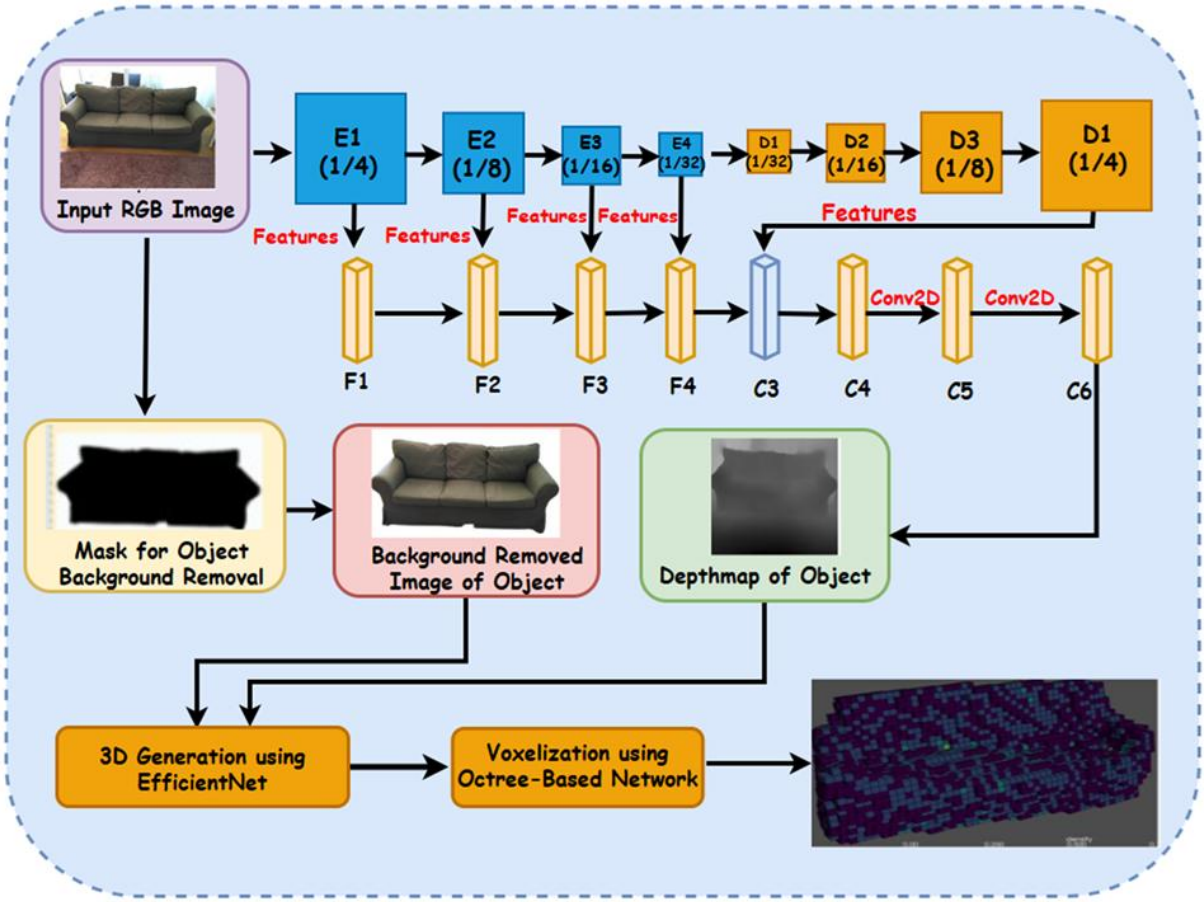


Figure 1. The main system architecture of our proposed system

3.1 Preprocessing of the data

For the generation of the 3D model using deep learning model. We need to preprocess RGB data is usually required to segment the object from the image. Moreover, object segmentation of images also dramatically reduced the computational cost. Hence, generally, it's easy for a model to map the 3D of the object. For background removal. Deng et al. [19] described research on image content-based indexing and retrieval for digital image libraries. Existing techniques also used global image features, leading to background features being mistaken as object features. The proposed approach analyzed background regions using colour clusters and removed them from the indexing process to avoid interfering with the retrieval of meaningful image content. The goal is to improve the accuracy of image retrieval based on colour features. We used a similar technique to remove the background of the image using colour histogram fuzzy clustering. Approaches using regular voxel grids often suffer from the curse of dimensionality, requiring significant computational resources to process high-resolution data.

Point-based methods, while efficient in representing sparse data, can struggle to maintain accurate reconstruction details due to their lack of inherent spatial organization.

Mesh-based approaches directly predict 3D surfaces but are often more computationally expensive and require post-processing steps to ensure smoothness and connectivity of the mesh.

Figure 2 shows the original image which was employed as input for the 3D reconstruction. Figure 3 shows the image with its background noise removed, alongside its corresponding original image. The octree structure inherently adapts to the density of the object, refining the representation only in areas where details are required. This leads to efficient memory use

and faster processing times. Leveraging EfficientNet's compound scaling strategy ensures that the features are extracted with optimized accuracy and efficiency, which is not always the focus of traditional backbone networks.

(a) (b) (c)

Figure 2. Original images of objects (a) aeroplane, (b) bottle and (c) sofa

(a) (b) (c)

Figure 3. Original images without background (a) aeroplane, (b) bottle and (c) sofa

3.2 Feature extraction

In this subsection, a deep learning algorithm helped in the extraction of depth prediction using a deep learning method based on generative modelling.

3.2.1 Depth prediction

In order to obtain the depth information from the image, the approach proposed by Khan et al. [20] and Alzahrani et al. [21] has been adopted in this work. This research presented two enhancements for single-picture depth estimation, first a convolutional neural network (CNN) for efficiently fusing information at different scales and secondly used the use of three loss terms to assess errors in depth, gradients, and surface normal. The results show that these enhancements have improved accuracy, particularly when reconstructing small objects and object borders with finer precision.

Algorithm 1: Depth Prediction Algorithm

Input: RGB Image
Output: Depth Image
//NYU v2 Dataset for training
Model: SENet-154
No_of_features = 2048
Block_Channel = [256, 512, 1024, 2048]
Transformation:
//Scale and normalize to reduce computational cost
Scale (Image, 320, 240)
Normalize (Image, Self.Mean, Self.SD): -SD: Standard Deviation
Block 1:
C1=Conv2d(Input)
/*
//DownSampling/EnCoding Block:
E1=DownSampling(1/4, C1)
E2=DownSampling(1/8, E1)
E3=DownSampling(1/16, E2)
E4=DownSampling(1/32, E3)

Block 2:
C2=Conv2d(E4)
//UpSampling/DeCoding Block:
D1=UpScale(1/2, C2)
D2=UpScale(1/2, D1)
D3=UpScale(1/2, D2)
D4=UpScale(1/2, D3)
Block 3:
//Concatenation Features:
F1=Upscale(1/16,E1)
F2=Upscale(1/16,E2)
F3=Upscale(1/16,E3)
F4=Upscale(1/16,E4)
Concatenate:
//Concatenation and Convolution:
F=F1+F2+F3+f4
C3=Conv2d(D4)
FN=F+C3
C4=Conv2d(FN)
C5=Conv2d(C4)
C6=Conv2d(C5)
Output: Depth Imagev

$$Conv_{2d} = w * F(x,y) = (\sum_{i}^{H}\sum_{j}^{W} w(\delta x, \delta y).F(x+\delta x, y+\delta y) \quad (1)$$

where, $Conv_{2d}$ is convolutional function that extracts features using weight w on image dimension x, y. Figure 4 shows the depth results.

(a) (b) (c)

Figure 4. Depth results on (a) aeroplane, (b) bottle and (c) sofa

3.2.2 3D shape features extraction using EfficientNet

To predict 3D shape points from RGB images and their depth, our proposed system utilizes a CNN model, specifically EfficientNet. Reconstructing 3D shapes from a single RGB image poses significant challenges due to potential variations in object configurations and the inherent ambiguity of depth information. Traditional methods for accurate 3D reconstruction from monocular images often depend on extensive 3D annotations during training, which are costly and labor-intensive to produce. To overcome this limitation, a self-supervised 3D reconstruction network, S2HAND, has been developed to estimate pose, shape, texture, and camera viewpoint without the need for labeled data. In this work, 2D key points detected in the input image are used to extract geometric information, and the consistency between 2D and 3D representations is leveraged to train a precise 3D reconstruction model. Additionally, a novel set of loss functions was introduced to enhance the neural network's outputs. This approach demonstrates the feasibility of training accurate 3D reconstruction models without manual annotations and has proven valuable for extracting 3D shape features, which are integral to generating 3D voxels. A typical efficient model consists of different layers of convolutional layers and fully connected layers, pooling layer and shortcut layers. The shortcut connections are what make this neural network model unique, as they allow the network to learn residual functions that can be added to the input. This helps to alleviate the vanishing gradients problem and allows EfficientNet models to be trained effectively even when they have hundreds of layers. Figure 5 shows the detailed architecture of the EfficientNet model used in this suggested system. EfficientNet, due to its scaled architecture, provides a more computationally efficient solution compared to older architectures, maintaining accuracy even with fewer parameters.

3.2.3 Octree-based network for 3D voxel reconstruction

To generate the 3D in the form of voxels in this research octree-based network has been used. Octree-based networks [22] are a type of deep neural network that makes use of octrees, a data structure for efficiently storing 3D volumetric data. Octrees are used to represent 3D space as a hierarchy of cubic cells, with each cell being subdivided into eight child cells until a certain depth is reached. The speciality of octree-based networks lies in their ability to efficiently process 3D data, such as point clouds or voxel grids while maintaining high spatial resolution. By using octrees, these networks selectively processed only the parts of the input that contained relevant information, which significantly reduced the computational cost and memory requirements of the network.

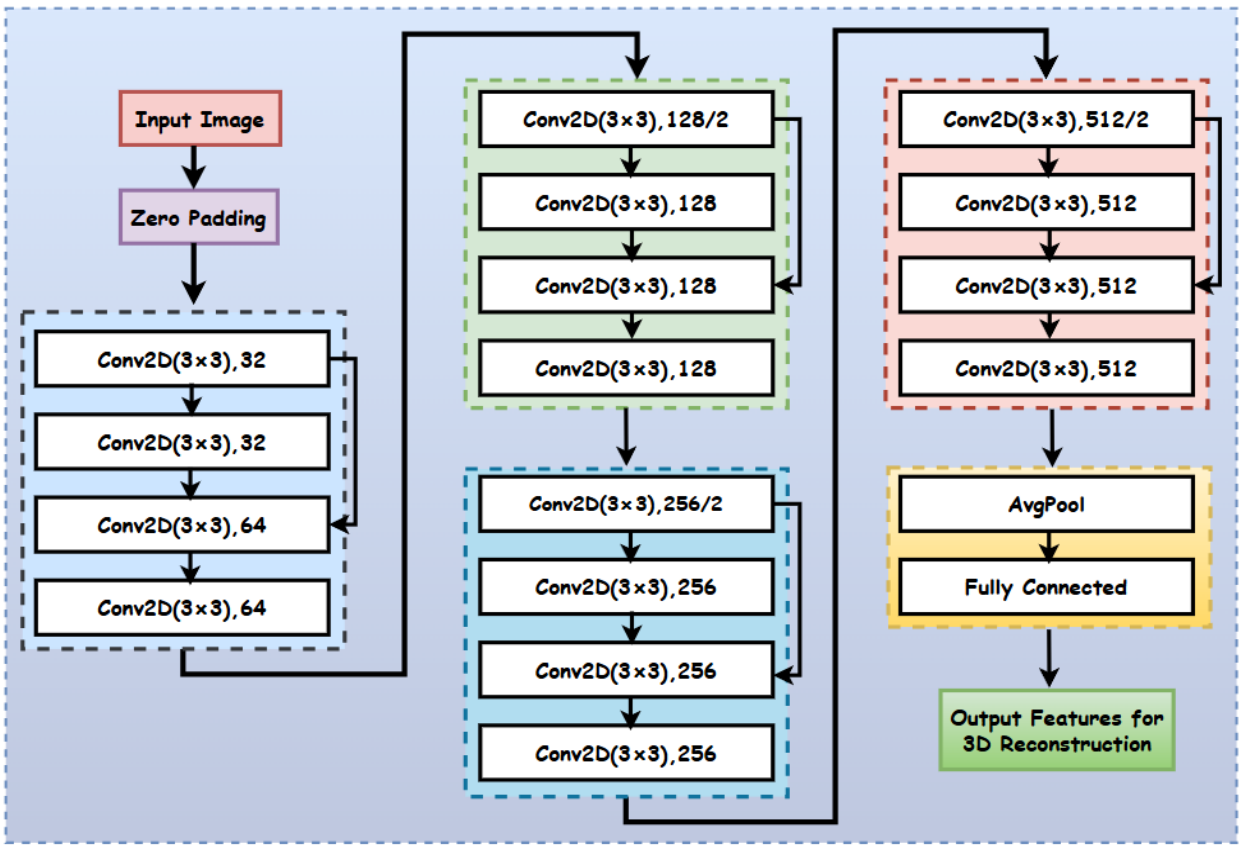


Figure 5. Detailed architecture of EfficientNet for 3D features extraction

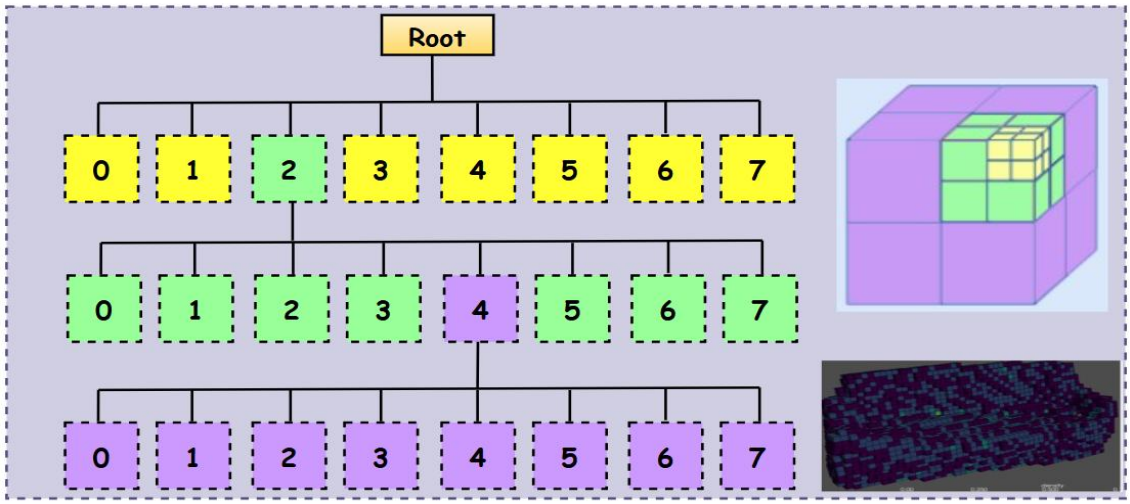
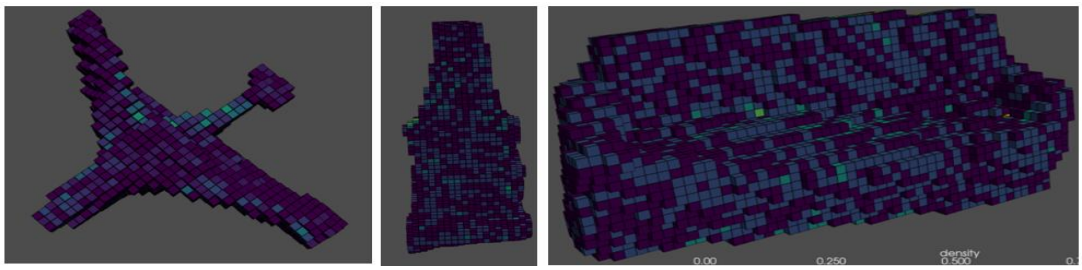


Figure 6. Octree based network for 3D visualization in voxels



(a) (b) (c)

Figure 7. Visualization of 3D voxel models

In addition, octree-based networks incorporated specialized layers and operations that have been tailored to 3D data, such as octree convolution and max pooling, which further enhanced their performance on tasks such as 3D object recognition and Segmentation. The octree-based approach is highly scalable, making it suitable for high-resolution reconstructions with lower memory requirements. This allows it to handle larger and more complex datasets like ShapeNetCore and Pix3D without a significant increase in computational cost [23]. Figure 6 shows the octree-based network conceptual model that processed the information to generate 3D voxels.

Figure 7 shows a 3D representation of reconstructed voxel models using the intended system. The 3D voxel model has been segmented into different colour ranges from yellow to green and in the dark blue intensity voxels the inner layer starts from yellow to the external layer till dark blue.

4. EXPERIMENTAL SETUP AND RESULTS

This section shows the experimental setup, dataset details, testing and validation performed on our proposed architecture. For Experimentation 3 Benchmark datasets have been used. A description of these datasets is given below and then Different types of evaluation have been performed.

4.1 Datasets description

4.1.1 ShapeNet dataset

The experimentation for 3D voxel reconstruction has been done using the ShapeNetCore dataset [7]. ShapeNetCore is a large repository of 3D CAD models that are annotated with semantic information like consistent alignments, parts, and sizes. It is organized by the WordNet taxonomy and contains over 3 million models with more than 220,000 models classified into 3,135 categories. ShapeNetCore provides a web-based interface for data visualization and serves as a benchmark for computer graphics and vision research. Our proposed system achieves high accuracy and detail in reconstruction due to the octree's ability to focus on relevant areas, effectively handling the variety of object geometries. The SOTA methods Perform well in capturing general object shapes but often struggle with fine details due to voxel resolution limits or inefficient use of mesh structures.

4.1.2 Pix3D sun dataset

The Pix3D sun dataset consists of 395 3D models spanning nine item categories, with each model paired with real-world photographs taken in diverse settings. This dataset includes 10,069 image-shape pairs, all of which are annotated with precise 3D data, enabling accurate pixel-level alignment between object shapes and their silhouettes in the images. Our proposed system utilizes the octree structure to enhance reconstruction quality, excelling in capturing fine details, particularly in complex indoor scenes and furniture models. While state-of-the-art methods based on mesh and point-based approaches achieve good results, they often struggle to preserve intricate details or require post-processing steps to mitigate noise in the data.4.1.3 PASCAL3D+Silberman Dataset.

PASCAL3D+Silberman, is a novel and difficult dataset for 3D object detection and pose estimation. PASCAL3D+adds 3D annotations to the PASCAL VOC 2012's 12 rigid

categories. Additionally, new photographs from ImageNet are added to each category. PASCAL3D+images are substantially more variable than previous 3D datasets, with more than 3,000 object occurrences per category on average. Our proposed system excels in reconstructing real-world objects with occlusions and complex surfaces thanks to the efficient feature extraction of EfficientNet and the octree's adaptability. The other methods typically have issues with occluded or partially visible objects, leading to lower reconstruction accuracy when compared to the octree-based approach.

In this article, the ShapeNetCore dataset has been used for the training purposes of the proposed system. For testing and evaluation purposes Pix3D and PasCAL3D+datasets have been used.

4.2 Results

Experiment I: Loss functions CD and earth movers distance (EMD)

To evaluate the reconstructed 3D models, two loss functions were applied for point-to-point comparisons with the ground truth data. The first metric, CD, was computed using three benchmark datasets, with the corresponding results presented in Table 1.

The second metric, Earth Mover's Distance (EMD), was employed to measure the dissimilarity between the reconstructed models and the ground truth. EMD was calculated using the Sinkhorn and Wasserstein distances, with the results illustrated in Figure 8. The mathematical equations used for calculating CD and EMD are provided.

$$CD(X,Y) = \frac{1}{|X|} * \left(\sum min\left(||x-y||^2\right)\right) + \frac{1}{|Y|} * \left(\sum min\left(||x-y||^2\right)\right) \quad (2)$$

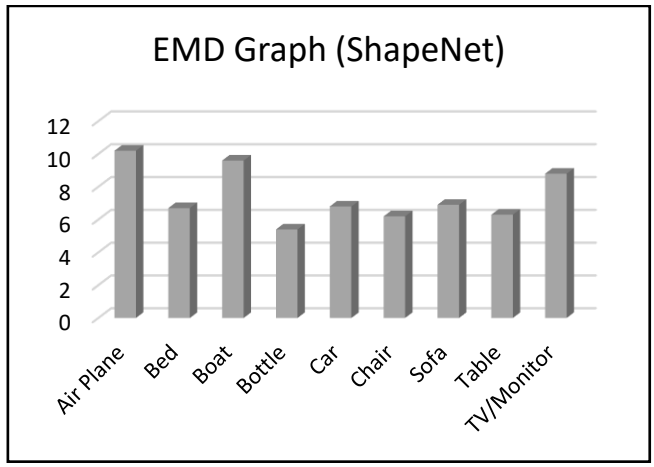
In this equation, X irepresents the predicted point set of the reconstructed 3D model, while denotes the ground truth point set. The cardinalities $|X|$ and $|Y|$ correspond to the number of points in X and Y respectively. The term $||x-y||^2$ refers to the squared Euclidean distance between a point x in X and a point y in Y. This equation calculates the minimum distance between the points in the predicted set and those in the ground truth set.

$$EMD(X,Y) = min_{S_{gt} \to S_{pr}} \sum ||x - \emptyset(x)|| \quad (3)$$

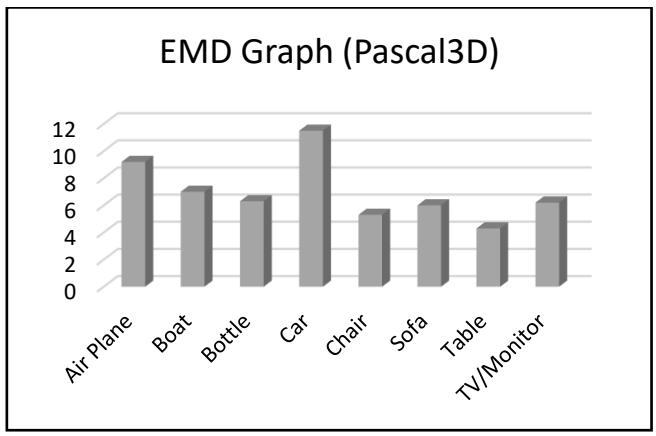
where, $x \in S_{pr}$ and $\emptyset(x) \in S_{pr}.\emptyset(x)$ is the closest point with the ground truth as shown in Table 1.

Table 1. CD on ShapeNetCore, Pascal 3D and Pix3D

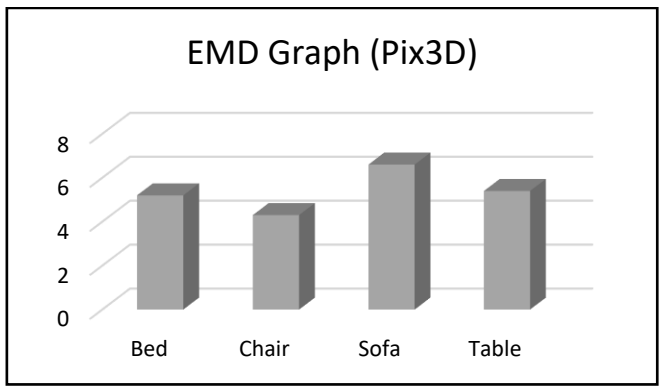
Objects	ShapeNetCore	Pascal3D	Pix3D
Air Plane	0.00070	0.00090	--
Bed	0.00240	--	0.00080
Boat	0.00480	0.00150	--
Bottle	0.00610	0.01050	--
Car	0.00130	0.00390	--
Chair	0. 00090	0.00080	0.00100
Sofa	0. 00180	0.00400	0.00140
Table	0. 00070	0.00100	0.00090
TV/Monitor	0.00790	0.00280	--
Mean	0.00387	0.00317	0.00102



(a)



(b)



(c)

Figure 8. Graphical representation of EMD values on (a) ShapeNetCore, (b) Pascal3D and (c) Pix3D

Experiment II: Evaluation matrix: accuracy, precision, recall and F1-Score

When using machine learning methods and artificial intelligence methods precision, recall, F1-Score and accuracy are commonly used matrices for evaluation purposes. Precision is the fraction of the 3D reconstructions that are correct among all the reconstructed 3D models. In other words, it measures how many of the reconstructed models are relevant to the ground truth. In 3D reconstruction, precision is typically used to evaluate the accuracy of shape and pose estimation. Recall is the fraction of the ground truth models that are correctly reconstructed among all the ground truth models. It measures how many of the relevant models were correctly identified by the algorithm. In 3D reconstruction, recall is

often used to evaluate the completeness of the reconstructed models [24]. The F1-Score is a performance metric that integrates precision and recall into a single value by calculating their harmonic mean. Precision reflects the proportion of accurately detected objects out of all detected objects, whereas recall represents the proportion of correctly identified objects relative to the total number of ground truth objects. The F1-Score ranges from 0 to 1, with higher values signifying improved performance as shown in Tables 2 and 3. Accuracy is the overall correctness of the reconstructed models. It measures how many of the reconstructed models are both relevant and correctly identified. In 3D reconstruction, accuracy is a composite metric that considers both precision and recall. The following equations are used to calculate precision and recall using max distance.

$$D(a,b) = \sqrt{\sum_{i}^{Gt_n} (a_i - b_i)^2} \tag{4}$$

$$Max\_Distance = Max(D_1(gt,pt), D_2(pt,gt)) \tag{5}$$

$$Precision = D(gt, D_1)/Max\_Distance \tag{6}$$

$$Recall = D(pt, D_2)/Max\_Distance \tag{7}$$

where, $D(a,b)$ is distance between two different point cloud models. Max Distance is the maximum distance computed between the gt ground truth and the pt predicted point cloud. Precision and Recall are calculated using distances and Max_Distance results (See Table 4). To calculate the F1-Score the harmonic mean of precision and recall is used in the following equation:

$$F1Score = 2 * Recall * \frac{Precision}{Recall + Precision} \tag{8}$$

where, $F1Score$ is the precision and recall combined matric to measure the accuracy of the model? To calculate accuracy, we used different loss functions to calculate overall loss and then this overall loss has been used to compute accuracy.

$$L_{Edge}(P_C, Target_{Length}) = \sum (i,j)\ ||P_i - P_j|| \tag{9}$$

$$L_{Normal(P_C, Target_{Length})} = \sum (i) \left(\left| P_i - P_j \right| \right)^2 \tag{10}$$

$$L_{Laplacian}(P_C, Target_{Length}) = \sum (i)\ (||\Delta^2 P_i||)^2 \tag{11}$$

where, P_C is predicted point cloud and $Target_{Length}$. The number of points in the point cloud in this article is set to 1024. P_i is the point in the predicted point set and P_j is the closest point in the ground truth point set. Δ^2 is the Laplacian operator that calculates the 2nd derivative of the position of the point (Table 5).

$$Overall\ loss = \lambda_E * L_{Edge} + \lambda_N * L_{Normal} + \lambda_L * L_{Laplacian} \tag{12}$$

where, $Overall\ loss$ is the loss function calculated using the combined effect of different weighted loss functions. λ_E, λ_N and λ_L is the weighted value used at the end to control the importance of edge, normal and Laplacian loss functions respectively in the overall loss function.

$$Accuracy(X,Y) = 1 - overall\ loss \tag{13}$$

where, $Accuracy$ is the overall evaluation of the proposed system. Overall loss has been computed using a combination of different types of loss functions with specific weights.

Table 2. Precision, Recall, F1-Score and Accuracy on 3D using ShapeNetCore

Objects	Precision	Recall	F1-Score	Accuracy (%)
Air Plane	0.7224	0.8342	0.8311	91.2078
Bed	0.9772	0.9561	0.9662	94.2341
Boat	0.6921	0.8352	0.8231	92.1291
Bottle	0.7941	0.9332	0.8571	95.2978
Car	0.8972	0.8162	0.7432	91.2822
Chair	0.7674	0.9361	0.8432	96.0736
Sofa	0.8547	0.8313	0.8421	93.7224
Table	0.9921	0.8135	0.8939	95.2635
TV/Monitor	0.7772	0.8164	0.7867	89.9726
Mean	0.8440	0.8673	0.8444	93.4969

Table 3. Precision, Recall, F1-Score and Accuracy on 3D using Pascal3D

Objects	Precision	Recall	F1-Score	Accuracy (%)
Air Plane	0.7224	0.8342	0.8311	91.2078
Bed	0.6921	0.8352	0.8231	92.1291
Boat	0.7941	0.9332	0.8571	95.2978
Bottle	0.8972	0.8162	0.7432	91.2822
Car	0.7674	0.9361	0.8432	96.0736
Chair	0.8547	0.8313	0.8421	93.7224
Sofa	0.9921	0.8135	0.8939	95.2635
Table	0.7772	0.8164	0.7867	89.9726
TV/Monitor	0.8250	0.8546	0.827	93.3916
Mean	0.7224	0.8342	0.8311	91.2078

Table 4. Precision, Recall, F1-Score and Accuracy on 3D using Pix3D

Objects	Precision	Recall	F1-Score	Accuracy (%)
Bed	0.9538	0.9621	0.9575	96.2619
Chair	0.722	0.832	0.7731	93.7112
Sofa	0.8738	0.882	0.878	94.0075
Table	0.9101	0.8016	0.8518	95.0922
Mean	0.8166	0.8738	0.8409	94.3708

Table 5. IoU on ShapeNetCore, Pascal3D and Pix3D

Objects	ShapeNetCore	Pascal3D	Pix3D
Air Plane	6.02	5.78	--
Bed	5.57	--	5.40
Boat	3.22	5.88	--
Bottle	2.11	3.41	--
Car	7.80	7.84	--
Chair	8.72	8.42	9.29
Sofa	5.42	6.04	5.98
Table	5.55	5.97	6.20
TV/Monitor	5.87	5.64	--
Mean	5.58	6.13	6.72

Experiment III: Evaluation object overlapping: Intersection over Union (IoU).

This system has been also evaluated using a comparison of predicted and ground truth overlapping using Kato, Intersection over Union (IoU) calculation. This method is best

for the evaluation of a volumetric 3D reconstruction system (See Table 6). The higher the value of IoU the better the reconstruction will be.

$$IoU = \frac{Obj_{gt} \cap Obj_{pr}}{Obj_{gt} \cup Obj_{pr}} = \frac{\sum\{I(Obj_{pr} > \epsilon) * I(Obj_{gt})\}}{\sum\{I(I(Obj_{pr} > \epsilon) + I(Obj_{gt}))\}} \quad (14)$$

where, I is the indicator function that shows the ith voxel in the volumetric 3D shape. ϵ is the threshold where the value has been computed.

Experiment IV: Comparison with the State-of-the-art systems

Comparing the performance of 3D reconstruction methods with the state of the art (SOTA) is an important task in evaluating the effectiveness of these methods. The state of the art refers to the best-known method or model that achieves the highest performance on a given task. In 3D reconstruction, the state of the art can be determined by comparing the performance of different methods on standard benchmark datasets ShapeNetCore, Pascal3D and Pix3D for object reconstruction. To compare the performance various metrics have been compared like CD, EMD and IoU.

Table 6. Comparison of IoU with different state-of-the-art methods

Objects	ShapeNetCore (Ours)	Pascal3D (Ours)	Pix3D (Ours)	DRC [18]	CSD M
Air Plane	**6.02**	5.78	--	5.70	5.00
Bed	5.57	--	5.40	--	--
Boat	3.22	5.88	--	--	9.94
Bottle	2.11	3.41	--	---	--
Car	7.80	7.84	--	7.60	5.18
Chair	8.72	8.42	9.29	4.70	5.20
Sofa	5.42	6.04	5.98	--	6.58
Table	5.55	5.97	6.20	--	--
TV/Monitor	5.87	5.64	--	--	9.64
Mean	5.58	6.13	6.72	6.00	6.92

5. CONCLUSIONS

This research article has utilized RGB-based data to reconstruct voxel-based 3D shapes. The system can be used in various world real-world applications like E-learning, E-commerce, medical diagnostics, scene understanding and 3D game development. Initially, RGB image has been given as input to the model for preprocessing, background removal and object detection. Next, depth features have been extracted to estimate the 3D shape of the object. Later on, the points have been estimated in 3D space. Two neural networks have been used for voxelization and visualization of objects: The first one is the EfficientNet which is a convolutional neural network that uses RGB and depth as input and in return provides the predicted 3D tensor of objects. The second 3D Tensor has been used in the octree-based network, each node has a further 8 nodes and the tree expanded till the required number of points in the 3D shape. The three benchmark datasets; ShapeNetCore, Pascal3D and Pix3D have been used for the experimentation of this proposed system. These datasets are based on world real-world objects. We achieved mean CDs of 0.00387, 0.00317 and 0.00102 on ShapeNetCore, Pascal3D and Pix3D respectively. Also, IoU has been used to measure the performance of this system. We achieved mean IoU of 5.58, 6.13 and 6.72 on the above datasets. We compare the proposed methods with different state-of-the-art methods, showing our system has much better results. For future directives, we will find various datasets related to the human face and human pose. Also mapping 3D of medical images dataset will be done like X-ray scans and CT-scan.

3D voxel reconstruction has significant real-world applications in areas such as medical imaging, geospatial mapping, and industrial design. In healthcare, voxel-based models are used for precise organ and tumour visualization in MRI and CT scans. In geospatial studies, they enable detailed 3D terrain mapping. Case studies in archaeology and urban planning showcase their utility in reconstructing ancient sites and simulating city landscapes for infrastructure development.

ACKNOWLEDGMENT

This research is supported and funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The research team thanks the Deanship of Graduate Studies and Scientific Research at Najran University for supporting the research project through the Nama'a program, with the project code NU/GP/SERC/13/18-3.

REFERENCES

[1] Zhang, Y., Liu, Z., Liu, T.P., Peng, B., Li, X. (2019). RealPoint3D: An efficient generation network for 3D object reconstruction from a single image. IEEE Access, 7: 57539-57549. https://doi.org/10.1109/ACCESS.2019.2914150.

[2] Poux, F., Billen, R. (2019). Voxel-based 3D point cloud semantic segmentation: Unsupervised geometric and relationship featuring vs deep learning methods. ISPRS International Journal of Geo-Information, 8(5): 213. https://doi.org/10.3390/ijgi8050213

[3] Wen, Y.F., Wang, Z.T., Li, Z.Y., Wei, D.X., Sun, Y. (2024). Efficient 3D view synthesis from single-image utilizing diffusion priors. In International Symposium on Neural Networks, pp. 93-102. https://doi.org/10.1007/978-981-97-4399-5_9

[4] Shu, D.W., Park, S.W., Kwon, J. (2019). 3D point cloud generative adversarial network based on tree structured graph convolutions. In 2019 IEEE/CVF International

Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 3858-3867. https://doi.org/10.1109/ICCV.2019.00396

[5] Wen, M.Y., Cho, K. (2024). Depth prior-guided 3D voxel feature fusion for 3D semantic estimation from monocular videos. Mathematics, 12(13): 2114. https://doi.org/10.3390/math12132114

[6] Savkin. A., Ellouze. R., Navab, N., Tombari. F. (2021). Unsupervised traffic scene generation with synthetic 3D scene graphs. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, pp. 1229-1235. https://doi.org/10.1109/IROS51168.2021.9636318

[7] Liu, K., Liu, J., Liu, S.D. (2024). Enhanced semi-supervised medical image classification based on dynamic sample reweighting and pseudo-label guided contrastive learning (DSRPGC). Mathematics, 12(22); 3572. https://doi.org/10.3390/math12223572

[8] Hu, C., Dong, Y.H., Peng, S.B., Wu, Y.H. (2025). Open-world semi-supervised learning for FMRI analysis to diagnose psychiatric disease. Information, 16(3): 171. https://doi.org/10.3390/info16030171

[9] Renat, A., Imangali, K. (2024). Learning Latent Representations for 3D voxel grid generation using variational autoencoders. In 2024 IEEE AITU: Digital Generation, Astana, Kazakhstan, pp. 169-173. https://doi.org/10.1109/IEEECONF61558.2024.10585546

[10] Bae, S.J., Kim, S., Lee, H., Lee, J., Lim, S.C., Bang, G., Kang, J.W. (2020). 2D reprojection of plenoptic 3D voxel data using Gaussian intensity spreading. In 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), pp. 1395-1397. https://doi.org/10.1109/ICTC49870.2020.9289557

[11] Huang, Y.S., Zou, S.L., Liu, X.W., Xu, K. (2025). Part-aware shape generation with latent 3D diffusion of neural voxel fields. In IEEE Transactions on Visualization and Computer Graphics, https://doi.org/10.1109/TVCG.2025.3562871

[12] Yuan, G.Z.Q., Fu, Q.C., Mi, Z.X., Luo, Y.M., Tao, W. B. (2022). Ssrnet: Scalable 3d surface reconstruction network. IEEE Transactions on Visualization and Computer Graphics, 29(12): 4906-4919. https://doi.org/10.1109/TVCG.2022.3193406

[13] Kuang, H., Wang, B., An, J.P., Zhang, M., Zhang, Z.H. (2020). Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from lidar point clouds. Sensors, 20(3): 704. https://doi.org/10.3390/s20030704

[14] Shi, C., Lv, Z., Yang, X.H., Xu, P.F., Bibi, I. (2020). Hierarchical multi-view semi-supervised learning for very high-resolution remote sensing image classification. Remote Sensing, 12(6): 1012. https://doi.org/10.3390/rs12061012

[15] Tahir, R., Sargano, A.B., Habib, Z. (2021). Voxel-based 3D object reconstruction from single 2D image using variational autoencoders. Mathematics, 9(18): 2288. https://doi.org/10.3390/math9182288

[16] Ji. Y., Qian, J.F., Yang, R.M., Ji, T.S., Liao, H.H. (2024). A 3D tile generation method based on adaptive meshing and octree segmentation. In 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), Wuhan, China, pp. 652-657. https://doi.org/10.1109/ISCEIC63613.2024.10810165

[17] Yasir, S.M., Ahn, H. (2024). Deep learning-based 3D instance and semantic segmentation: A review. Journal on Artificial Intelligence, 4(2): 99-114. https://doi.org/10.32604/jai.2022.031235

[18] Gbadago, D.Q., Moon, J., Kim, M., Hwang, S. (2021). A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and genetic algorithm: A case of butadiene synthesis. Chemical Engineering Journal, 409: 128163. https://doi.org/10.1016/j.cej.2020.128163

[19] Deng, J.J., Shi, S.S., Li, P.W., Zhou, W.G., Zhang, Y.Y., Li, H.Q. (2021). Voxel R-CNN: Towards high performance voxel-based 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(2): 1201-1209. https://doi.org/10.1609/aaai.v35i2.16207

[20] Khan, M.S., Lingam, M.S., Mankar, V.H. (2016). Volumetric feature extraction of 3D images defined over hexagonal prism lattice. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, pp. 790-796. https://doi.org/10.1109/NGCT.2016.7877518

[21] Alzahrani, M., Usman, M., Jarraya, S.K., Anwar, S., Helmy, T. (2024). Deep models for multi-view 3D object recognition: A review. Artificial Intelligence Review, 57(12): 323. https://doi.org/10.1007/s10462-024-10941-w

[22] Naseer, A., Jalal, A. (2024). Holistic scene recognition through U-Net semantic segmentation and CNN. In 2024 19th International Conference on Emerging Technologies (ICET), Topi, Pakistan, pp. 1-6. https://doi.org/10.1109/ICET63392.2024.10935069

[23] Abro, I.A., Jalal, A. (2024). Multi-modal sensors fusion for fall detection and action recognition in indoor environment. In 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, pp. 1-6. https://doi.org/10.1109/ETECTE63967.2024.10823705

[24] Abbas, Y., Jalal, A. (2024). Drone-based video surveillance using Yolov6 and neuro fuzzy classifier. In 2024 19th International Conference on Emerging Technologies (ICET), Topi, Pakistan, pp. 1-6. https://doi.org/10.1109/ICET63392.2024.10935116