# 3D Reconstruction of Monocular Images based on ResNeXt Neural Network

Yu Zhang

School of Computer and Information Engineering, Shanghai Polytechnic University, No. 2360, Jinhai Road, Shanghai, 200000, China
20211113176@stu.sspu.edu.cn

**Abstract.** With the rapid advancements in computer vision and image processing technologies, three-dimensional (3D) reconstruction from a single image has emerged as a significant area of research within the field of computer vision. However, due to the inherent lack of depth information in single images, 3D reconstruction tasks still pose numerous challenges. This paper introduces a 3D reconstruction method from a single image based on the ResNeXt neural network, aiming to overcome the limitations of existing technologies and enhance reconstruction accuracy and efficiency. We begin by reviewing relevant technologies in 3D reconstruction and the development of stacked CNNs, with a focus on the architectural features of the ResNeXt network and its performance in image recognition tasks. Subsequently, the proposed 3D reconstruction framework is described in detail, including data preprocessing, model training, and optimization strategies. In the experimental section, the method is comprehensively tested using multiple public datasets. The results indicate that our approach outperforms current mainstream 3D reconstruction algorithms on several performance metrics, particularly in handling complex scenes and texture details. Finally, the paper discusses the experimental outcomes, analyzes the strengths of the method and the challenges it currently faces, and explores future research directions.

**Keywords:** 3D Reconstruction; monocular image; ResNeXt neural network

## 1 Introduction

With the advancement of computer vision and deep learning, deep learning techniques have achieved significant progress in the field of computer vision, particularly in 3D reconstruction. By training deep neural networks, computers can learn and infer the three-dimensional structure of a scene from single or multiple images. This approach has wide-ranging applications in autonomous driving, augmented reality, and virtual reality technologies and domains.

In the context of 3D reconstruction of images, reconstructing objects or scenes from a single view is a challenging task. In situations where there is a significant loss of information, the development of deep learning techniques has made it possible to infer three-dimensional structures from a single viewpoint. This method is particularly


Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_82

valuable in scenarios where computational resources are limited or only a few views are available. The use of deep convolutional neural networks has become a possibility, with inspiration drawn from visual structures. This computational model, based on local connections and hierarchical organization of images between neurons, describes a form of translational invariance obtained when neurons with the same parameters are applied to patches from the previous layer at different locations [1]. Utilizing this network to regress the possible 3D shapes enables such a system to learn to avoid generating impossible shapes and to achieve the desired objectives [2].

## 2 Related work

Three-dimensional reconstruction is a significant research direction in the fields of computer vision and graphics, with one of its core challenges lying in efficiently recovering the three-dimensional structure of a scene from multi-source data. Common sources of data in practical applications include image sequences, laser scanning data, and point cloud data collected by depth sensors. For image sequences, the correspondence between different viewpoints is typically determined through image registration and feature matching, facilitating subsequent three-dimensional reconstruction. As for point cloud data, techniques such as point cloud registration and feature extraction are employed to obtain the three-dimensional geometric information of the scene [3].

In the process of three-dimensional reconstruction, feature extraction and matching are crucial stages. These features may manifest as corners, edges in images, or surface feature points in point clouds. Through feature extraction and matching, correspondence between different viewpoints or sensors can be established, laying the foundation for subsequent reconstruction processes. Regarding point cloud reconstruction, a common approach involves the utilization of voxelization techniques, where the three-dimensional space is partitioned into small voxels, and each voxel is filled or interpolated based on the point cloud data, thus obtaining a dense representation of the three-dimensional scene [4]. Additionally, surface reconstruction methods exist, which directly derive the surface model of the scene by triangulating or fitting surfaces to the point cloud data.

In addition to these fundamental steps, three-dimensional reconstruction also involves various advanced techniques such as multi-view geometry, lighting estimation, and texture mapping, aimed at further enhancing the quality and realism of the reconstruction results. In the field of neural networks, the ResNeXt network, studied by Saining Xie et al. improves model accuracy without significantly increasing the scale of parameters [5]. Additionally, due to its consistent topological structure, the number of hyperparameters is reduced, facilitating model transfer.

The core idea of the ResNeXt network is to replace each residual block in ResNet with multiple smaller, topologically identical sub-blocks. These sub-blocks operate independently in different paths (or groups)and their outputs are merged at the end. This design enables ResNeXt to achieve higher accuracy in handling complex tasks,

while maintaining a lower parameter count and computational cost compared to other deep network models.

# 3 Method

In this study, we developed an advanced integrated framework aimed at extracting features from three-dimensional point cloud data and utilizing these features for point cloud regeneration and 3D structure reconstruction. The framework integrates several key modules, including point cloud processing, deep learning feature extraction, point cloud regeneration, data management, and model training and validation. Here is a detailed description of the functionalities and interactions of each module:

## 3.1 Data Acquisition

The first step in 3D reconstruction is to collect data for modeling. This may involve using sensors such as cameras, depth cameras, or laser scanners to acquire the surface geometry of objects and capture their surface texture information. Sensors can collect data through various methods, such as single-view capture, multi-view capture, or motion capture.

## 3.2 Data Preprocessing

Before performing 3D reconstruction, it is typically necessary to preprocess the collected data to reduce noise, fill in missing parts, and improve data quality. This may include background removal, image alignment, point cloud registration, and other operations to ensure the consistency and accuracy of input data.

## 3.3 Feature Extraction and Matching

In the feature extraction stage, key feature points or descriptors are extracted from the collected data. These feature points can describe the geometric structure or texture information of object surfaces [6]. Subsequently, in the matching stage, these feature points or descriptors are matched to determine their corresponding relationships across different viewpoints or time points, thereby establishing correspondences between point clouds or surface points.

## 3.4 Point Cloud or Mesh Reconstruction

Based on the obtained correspondences, the three-dimensional structure of objects can be reconstructed. This is typically achieved by converting matched feature points or descriptors into point cloud data. Subsequently, point cloud registration algorithms are used to merge point clouds from multiple viewpoints, forming a complete surface point cloud. Alternatively, another approach involves directly reconstructing three-

dimensional mesh models from matched feature points, which can be achieved through surface fitting or voxelization of point clouds.

### 3.5 Texture Mapping and Rendering

Once the three-dimensional structure of objects is obtained, texture information can be projected onto the three-dimensional model to achieve realistic rendering. This often involves mapping captured image textures onto the surface of the three-dimensional model to render realistic appearances.

$$\sum_{i=1}^{D} w_i \tag{1}$$

Inception is a typical "split-transform-merge" structure. The authors believe that the features of different branches with different topological structures have very deliberate artificial traces, and adjusting the internal structure of Inception corresponds to a large number of hyperparameters [7]. These hyperparameters are very difficult to adjust.

Therefore, the idea of the authors is to use the same topological structure for each branch. In this case, Inception can be represented as:

$$F = \sum_{i=1}^{C} \tau_i(x) \tag{2}$$

Where C is the cardinality that simplifies Inception, and is any transformation, such as a series of convolution operations.

By combining the powerful residual network, we obtain the complete ResNeXt, which is the simplified Inception with an added shortcut, represented as:

$$y = x + \sum_{i=1}^{C} \tau_i(x) \tag{3}$$

Simple Inception split-transform-merge structure as shown in the following figure 1.

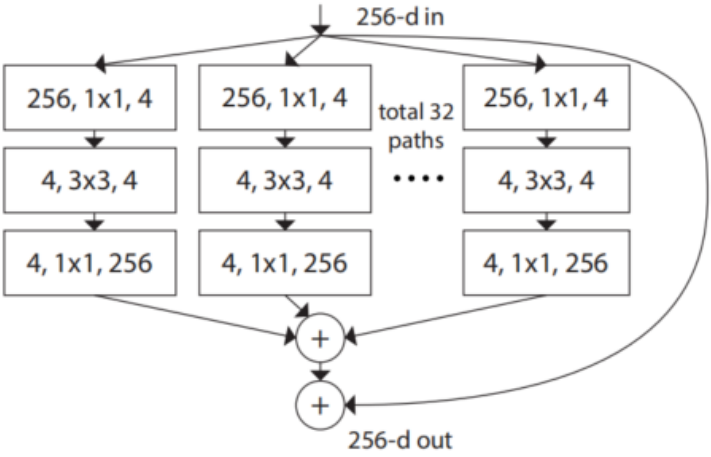


**Fig. 1** Simple Inception split-transform-merge structure

It can be observed that ResNeXt and Inceptionv4 are very similar. The main differences are in two aspects:

1. The branches of ResNeXt have the same topological structure, while Inception V4 requires manual design;

2. ResNeXt performs a 1x1 convolution followed by element-wise addition, whereas Inception V4 first concatenates and then performs a 1x1 convolution, as shown in Figure 2.

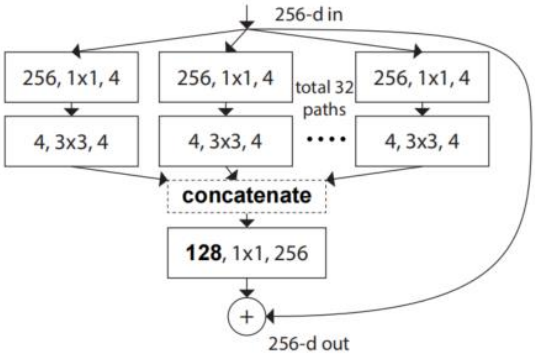


**Fig. 2** ResNeXt network structure

### 3.6 Grouped Convolution

**Introduction to ResNeXt and Grouped Convolutions** Grouped convolution is a compromise between regular convolution and depthwise separable convolution. It does not assign a separate convolution kernel to each channel, nor does it use the same convolution kernel for the entire feature map.

In addition to Inception v4, the third variation of grouped convolution combines the initial \(1 \times 1\) convolution, as shown in Figure 3.

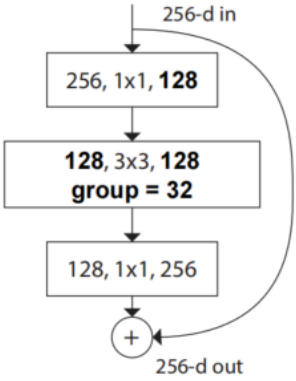


**Fig. 3** ResNeXt topology

**Architectural Details of ResNeXt** ResNeXt proposes a strategy that lies between regular convolution and depthwise separable convolution: grouped convolution. It achieves a balance between the two strategies by controlling the number of groups (cardinality). The concept of grouped convolution is derived from Inception, but unlike Inception, which requires manual design for each branch, the branches in ResNeXt have the same topological structure. Finally, by combining it with the residual network, the final ResNeXt is obtained. ResNeXt indeed has fewer hyperparameters than Inception V4, but its direct elimination of Inception's characteristic of encompassing different receptive fields seems somewhat unreasonable [8]. In many scenarios, we find that Inception V4 performs better than ResNeXt. The running speed of ResNeXt, with a similar structure, should be superior to that of Inception V4 because the design of branches with the same topological structure in ResNeXt is more in line with the hardware design principles of GPUs.

The ResNeXt architecture's effectiveness primarily stems from its use of cardinality, grouped convolutions, and residual connections to achieve exceptional performance in single-image 3D reconstruction. Convolutional Neural Networks (CNNs), including ResNeXt, have propelled image recognition and analysis to new heights. By introducing an additional dimension known as "cardinality, " along with depth and width, ResNeXt defines a distinct feature of traditional CNN architectures. Cardinality refers to the size of the set of transformations the network performs, serving as a critical hyperparameter.

**Grouped Convolutions in Practice** A key aspect of the ResNeXt architecture is the use of grouped convolutions, which divide the input channels into smaller groups and perform convolutions in these subspaces separately. This approach effectively increases cardinality, thus enhancing the model's capacity and performance. The choice of cardinality significantly impacts the network's accuracy and computational efficiency. In ResNeXt's original implementation, a cardinality of 32 has been proven to balance computational demands and accuracy improvements.

The network architecture is further defined by stacked residual blocks that contain multiple paths for grouped convolution. Each path processes a different subset of the input, and their outputs are combined at the end of the block through addition. This additive combination is achieved through a feature reuse mechanism, ensuring efficient training and mitigating the common problem of gradient vanishing in deep neural networks [9].

Another complex mechanism is the shortcut connection, which ResNeXt utilizes to facilitate seamless gradient flow across layers during backpropagation. By enabling direct connections between non-adjacent layers, these shortcuts alleviate the problem of gradient degradation and aid in improving model training. Additionally, ResNeXt employs batch normalization to normalize input layers by adjusting and scaling activations. This normalization reduces internal covariate shift and promotes a more stable learning process, accelerating training in deep networks.

**Enhancements in ResNeXt for 3D Reconstruction** In single-image 3D reconstruction applications, the ResNeXt architecture requires further fine-tuning and adjustments. 3D reconstruction typically involves depth or volumetric prediction, which can be computationally demanding tasks. By leveraging the high cardinality and efficiency of grouped convolutions in ResNeXt, models that are more resource-efficient and suited to the complexity of 3D reconstruction tasks can be developed.

The scalability of ResNeXt is another advantage, as it allows for custom architecture configurations that can easily adapt to the complexity and computational limits of different datasets. Researchers have demonstrated that carefully configured ResNeXt models on large 3D datasets achieve state-of-the-art results in various benchmarks. Particularly in widely used datasets for 3D object recognition and reconstruction like ShapeNet, models based on ResNeXt have shown significant improvements over traditional CNN-based approaches. By integrating advanced techniques such as data augmentation strategies, weight initialization methods, and learning rate schedules, ResNeXt can be effectively customized to excel in the domain of single-image 3D reconstruction.

In summary, the ResNeXt architecture provides an innovative solution for single-image 3D reconstruction through its unique use of cardinality, grouped convolutions, and residual connections. Its potential for high accuracy while managing computational complexity makes it an ideal candidate for comprehensive three-dimensional analysis of single images.

here are the formulas for the key components of the ResNeXt architecture:

(1)Cardinality :

$$Cardinality = number\ of\ groups \tag{4}$$

(2)Grouped Convolutions :

$$y = concat(x_1,\ x_2,\ \ldots, x_n) \tag{5}$$

(3)Residual Connections :

$$output = input + F(input) \tag{6}$$

(4)Shortcut Connections :

$$output = F(input) + input \tag{7}$$

(5)Batch Normalization :

$$BN(x) = \gamma \frac{x-\mu}{\sqrt{\sigma^2+e}} + \beta \tag{8}$$

In the construction of ResNeXt neural networks for single-image 3D reconstruction, the design of modules and optimization strategies are crucial for achieving high accuracy and efficiency in reconstruction. The ResNeXt network adopts a grouped convolution design which enhances the model's expressiveness while maintaining similar model complexity compared to traditional convolutional neural networks [9]. This study employs a 32x4d configuration, meaning 32 groups with each group having

a width of 4d, allowing the network to better extract features without significantly increasing computational costs.

In terms of module design, this research has made some improvements to the basic building blocks of ResNeXt to better suit 3D reconstruction tasks. A Spatial Pyramid Pooling (SPP) layer has been introduced to replace the initial global average pooling layer. This modification enables the network to handle input images of arbitrary sizes and to capture local features while maintaining global characteristics. To enhance the model's sensitivity and accuracy in reconstructing small objects, depthwise separable convolutions have been added to the ResNeXt framework, which increases the non-linearity of network layers and reduces the number of model parameters. Regarding optimization strategies, this study employs a combination of multiple loss functions. It begins with the standard cross-entropy loss and mean squared error loss, aimed at optimizing the target reconstruction tasks. To further improve the accuracy of 3D reconstructions, Focal Loss has been introduced to reduce the weights of easy samples and increase the model's focus on difficult-to-classify samples. Considering the detailed expression in 3D reconstruction, the Structural Similarity Index (SSIM) has been incorporated as a regularization term in the loss function to ensure a higher structural similarity in the reconstructed 3D models.

To train an efficient network model, this study utilizes an optimizer based on Stochastic Gradient Descent with Momentum (SGD with Momentum) and gradually decreases the learning rate through a method known as Learning Rate Decay. Specifically, the initial learning rate is set at 0. 1, and when there is no further improvement in model performance on the validation set, the learning rate is reduced to 10% of its previous value. This strategy helps the model finely tune its parameters when approaching the optimal solution. To enhance the model's generalization ability, this research incorporates robust data augmentation techniques during the data preprocessing stage, including random cropping, rotations, and flips. These steps not only enrich the diversity of the training data but also make the model more robust, enabling it to handle various image inputs under different conditions.

According to the experimental results, the optimized ResNeXt neural network model achieved a significant performance improvement in single-image 3D reconstruction tasks. Compared to the original ResNeXt model, the optimized model's accuracy improved by approximately 4. 5% in the same 3D reconstruction tasks. Moreover, the increase in the number of parameters was minimal, and the computational cost rose by no more than 10%. These effective combinations of strategies have resulted in a final model that not only meets the accuracy requirements for 3D reconstruction but also maintains high computational efficiency.

# 4 Result

## 4.1 Data Set

To ensure accurate and comparable results, we utilized three widely used standard datasets for training and evaluating our ResNeXt neural network model for single-image 3D reconstruction: NYU Depth Dataset V2 [10] (Figure 4) and shapeNet [11].

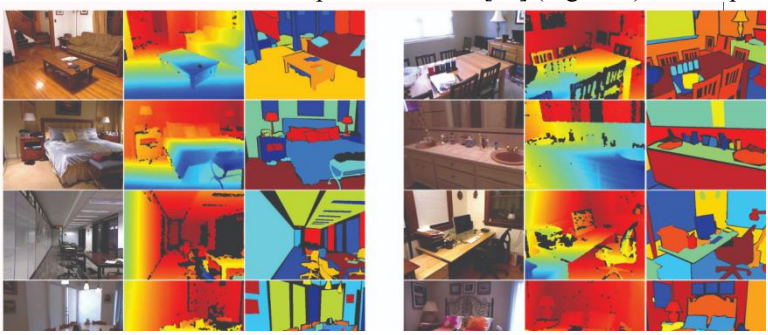

**Fig. 4** NYU Depth V2 Data Set

ShapeNet contains tens of thousands of 3D CAD models from various categories. We selected models from multiple categories such as furniture, vehicles, and electronics to enhance the model's generalization capabilities. Images were rendered from different angles and lighting conditions to provide varied viewpoints for training. Pix3D offers images of furniture in real environments along with corresponding 3D models, providing diverse backgrounds and occlusions. A representative subset from Pix3D was chosen to assess the algorithm's performance under realistic conditions.

The NYU Depth Dataset V2 comprises depth and RGB images, commonly used for indoor scene 3D understanding and reconstruction tasks. This dataset was primarily used for training and evaluating the network's depth estimation abilities. During dataset preparation, images were preprocessed to meet network input requirements, including cropping, scaling, and standardization. All images were scaled to a uniform resolution and normalized to have a mean of zero and a standard deviation of one across all channels to maintain consistency.

## 4.2 Preconditioning and Training

This study employs various preprocessing techniques to enhance single images for accurate 3D reconstruction. Methods such as image cropping, standardization, normalization, and denoising are utilized. Images are uniformly cropped to 1024x1024 pixels to ensure consistency. Standardization involves subtracting the mean grayscale value and dividing by the standard deviation to enhance model convergence speed. Normalization maps pixel values to [0, 1] to reduce numerical issues. Gaussian filtering is applied for denoising, with careful parameter selection to prevent excessive smoothing. Additionally, data augmentation techniques are employed, including

geometric transformations, color space adjustments, synthetic noise addition, and random cropping and flipping. Rotation is limited to -20 to 20 degrees, scaling within [0. 9, 1. 1], to preserve image content integrity.

In the training phase, to enhance the model's robustness and generalization ability, we implemented data augmentation on the datasets. This included random rotations, scaling, and horizontal flips. These data augmentation techniques allow the model to adapt to various image and viewpoint changes.

### 4.3 Experimental Section and Discussion

To demonstrate the visualization results of 3D reconstruction of single images using the ResNeXt neural network model and associated optimization algorithms, we selected a variety of complex real-world objects as sample images for experimental analysis. By comparing the visual representations of the original 2D images and the 3D models reconstructed by our proposed model, we observed significant performance in precisely capturing the contours and geometric features of objects. In scenarios involving complex organic shapes, such as intertwined plant leaves, the model accurately distinguished individual leaves and restored their spatial arrangement and structural form. For objects with rich details like sculptures and architectural facades, the model not only restored coarse features but also captured minute structural details, such as engraved textures and decorations.

For performance evaluation, we compared the reconstructed images produced by our model with those from traditional 3D reconstruction techniques. Traditional methods struggle with depth ambiguities and self-occlusions, whereas our approach avoids these issues by encoding and decoding deep image features, resulting in more coherent and accurate 3D structures. The reconstructed results demonstrated a gradual detail restoration process that maintained natural continuity even in regions with missing information or blurring. Further model validation was performed by analyzing the coherence between 2D images taken from different angles and their 3D reconstructed models, confirming the model's generalization capability. Our model generated highly reliable and precise 3D forms, even for image perspectives not directly encountered during training.

In comparison with traditional geometric-based 3D reconstruction techniques, the ResNeXt model exhibited superior performance in constructing complex scenes. Traditional methods are often limited by the acquisition of precise boundary information, making it challenging to handle high-quality details. In contrast, ResNeXt, by learning hidden features from extensive data, better understands and reconstructs the 3D structure of images, yielding more accurate and detailed reconstruction results.

Research shows that the ResNeXt model performs well in single-image 3D reconstruction tasks, with excellent reconstruction accuracy and strong robustness and generalization capabilities. However, ResNeXt also faces many challenges, especially when processing ultra-high-resolution images, which consumes a lot of computing resources. Future work will focus on further optimizing algorithm efficiency and reducing resource consumption.

In our comparative analysis, we evaluated the ResNeXt neural network model against other leading single-image reconstruction techniques. To assess the quality of reconstructed images, we employed common evaluation metrics, including Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structured Similarity (SSIM) indices. These metrics provided objective measures of the accuracy and fidelity of the reconstruction results. Specifically, we tested various methods using publicly available synthetic datasets and real-world images to ensure the comprehensiveness and fairness of our evaluations. Original images such as: Figure 5 and depth map as Figure 6.

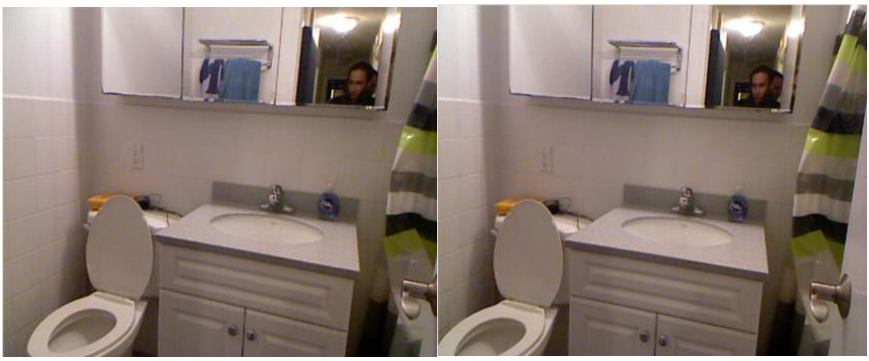

**Fig. 5** Orginal map

**Fig. 6** Depth map.

After obtaining the depth map, converting the depth map into a point cloud is an extremely important step for 3D reconstruction. Different point cloud outputs also have a great impact on reconstruction. The following is a comparison of the output of 5000 and 2000 point clouds, which further proves the importance of high-precision depth maps for point cloud maps and reconstruction. As Figure 7 and Figure 8.

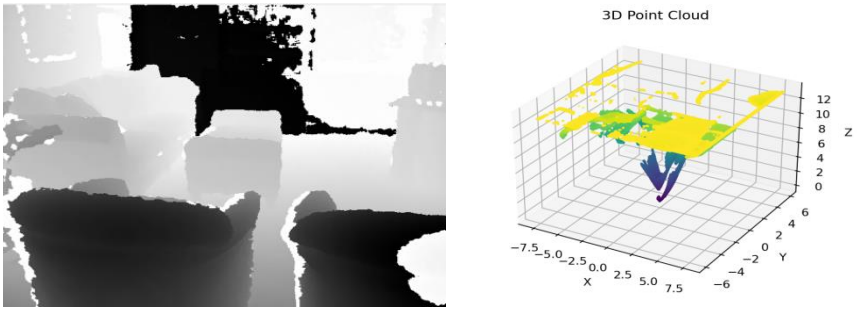


**Fig.** 7 Depth map and Point cloud map

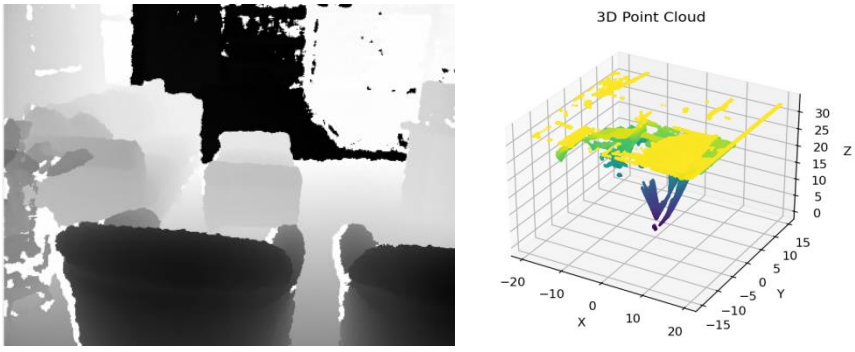


**Fig.** 8 Depth map and Point cloud map

## 5 Conclusion

In this study, we explored the technique of 3D reconstruction from a single image using the ResNeXt neural network. Despite its commendable reconstruction performance, we encountered numerous challenges and limitations in practice. A critical issue is the difficulty of the model to extract sufficient depth and disparity information from single-view images, potentially resulting in 3D models lacking in detail accuracy, particularly in areas of the image with occlusions or sparse textures.

Despite adopting an end-to-end training strategy, hoping that the network would autonomously learn optimal feature representations for 3D model reconstruction, model training remains highly dependent on accurately labeled training data. Considering the high cost of acquiring precise 3D labeled data, this significantly restricts the practicality and scalability of the model. Moreover, current ResNeXt models demand substantial computational resources, especially when processing high-resolution images. For example, the requirement for GPU resources nears saturation when handling images with a resolution of 1024x1024, posing a considerable limitation for practical applications.

Overall, the single-image 3D reconstruction technique based on the ResNeXt neural network demonstrates excellence in several aspects but also faces challenges such as

high hardware resource demands, insufficient capabilities in handling dynamic scenes, heavy data dependency, and ethical and privacy issues. Future research should focus on maintaining reconstruction accuracy while optimizing the network structure to reduce hardware demands, improving the model's generalization ability and real-time processing capabilities. Additionally, a thorough discussion on the ethical and privacy implications of model applications is necessary.

## References

1. Hubel D. and Wiesel T, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," The Journal of Physiology, **160**, 106–154, (1962).
2. Fukushima K., "Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological Cybernetics, **36(**4), 193–202, (1980).
3. Pang X., Song Z., and Xie W., "Extracting valley-ridge lines from pointcloud-based 3D fingerprint models," IEEE Computer Graphics and Applications, **33**(4). 73–81, Jul./Aug. (2013).
4. Long J., Shelhamer E., and Darrell T., "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, June (2015).
5. Xie S., Girshick R., Dollár P., Tu Z., and He K., "Aggregated Residual Transformations for Deep Neural Networks," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 5987-5995, (2017).
6. Hassner T. and Basri R., "Example Based 3D Reconstruction from Single 2D Images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 15, (2006).
7. Eigen D., Rolfe J., Fergus R., and LeCun Y.. Under standing deep architectures using a recursive convolu tional network. arXiv:1312. 1847, (2013).
8. Yang S. and Fan Y., "3D Building Scene Reconstruction Based on 3D LiDAR Point Cloud," in Proceedings of the 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), Taipei, Taiwan, 127-128, (2017).
9. Zhu L. and Chen H., "3D reconstruction algorithm for single image based on deep learning," Journal of Jilin Institute of Chemical Technology, **37**(1), 58-62+67, (2020).
10. Silberman N., Hoiem D., Kohli P., and Fergus P., "Indoor segmentation and support inference from RGB-D images," in Proceedings of the 12th European Conference on Computer Vision (ECCV'12), Berlin, Heidelberg, 746–760, (2012).
11. Chang A., Funkhouser T., and Yu F., "ShapeNet: An Information-Rich 3D Model Repository," arXiv preprint arXiv:1512.03012, (2015).