ART & TECHNOLOGY PUBLISHING PTE. LTD.

Computer System Networking and Telecommunications

2023 Volume 2, Issue 1

ISSN: 2630-4635

3D Point Cloud Reconstruction Technique from 2D Image Using Efficient Feature Map Extraction Network

Jeong-Yoon Kim, Seung-Ho Lee*

Department of Electronic Engineering, Hanbat National University, Daejeon, Republic of Korea

***Corresponding author:** Seung-Ho Lee, shlee@cad.hanbat.ac.kr

Abstract

In this paper, we proposed a 3D point cloud reconstruction technique from 2D images using an efficient feature map extraction network. The uniqueness of the method proposed in this paper is as follows. First, we used a new feature map extraction network that is about 27% more efficient than existing techniques in terms of memory. The proposed network did not downsize the input image until mid-way into the deep-learning network, so important information required for 3D point cloud reconstruction was preserved. The problem of increasing memory caused by using full-sized images is mitigated by efficiently configuring the deep learning network to be shallow. Second, by preserving the high-resolution features of the 2D image, the accuracy was further improved compared to the conventional technique. Feature maps extracted from the full-size image contained more detailed information than the existing method, thereby improving the accuracy of 3D point cloud reconstruction. Third, we used a divergence loss that did not require shooting information. Requiring information about not only 2D images but also the shooting angles in training can make dataset collection challenging. In this paper, the accuracy of the reconstruction of the 3D point cloud was increased by increasing the diversity of information through randomness without additional shooting information. In order to objectively evaluate the performance of the proposed method, experiments were conducted using the ShapeNet dataset following the same procedures as previous studies. The proposed method yielded a Chamfer distance (CD) of 5.87, an Earth mover's distance (EMD) of 5.81, and FLOPs of 2.9G. Lower CD and EMD values indicate greater accuracy in 3D point cloud reconstruction, while a lower FLOPs value indicates reduced memory requirements for deep learning networks. Consequently, the CD, EMD, and FLOPs performance evaluation results of the proposed method demonstrated approximately 27% improvement in memory efficiency and approximately 6.3% improvement in accuracy compared to other methods, validating its objective performance.

Keywords

Point cloud
Feature map
Reconstruction
Reparameterization trick
Latent vector
Deep learning

1. Introduction

Recently, the market size of 3D applications, including metaverses, has been growing rapidly, and the demand for 3D models has been increasing rapidly. In order to create a realistic space in the virtual world, it is becoming more and more important to represent 3D model data in as much detail as possible and as efficiently as possible. Conventional methods of acquiring 3D model data include LiDAR and image processing algorithms based on multiple images taken from different angles. However, there were many difficulties in reconstructing invisible parts with only partial information. Later, as artificial intelligence emerged and began to be applied to various fields, a 3D model reconstruction method using 2D images based on deep learning emerged. Choy *et al.* [1] demonstrated that the accuracy of deep learning-based 3D model reconstruction using 2D images can exceed the existing methods by reconstructing 2D images in a 3D voxel method in the form of a three-dimensional grid using RNN (Recurrent Neural Network). Fan *et al.* [2] presented a training method for a deep learning network that reconstructs 2D images into 3D point clouds. A 3D point cloud is a decimal coordinate representation of each point that forms the outline of an object, allowing for a more detailed representation with less memory than the 3D voxel method, which represents shapes as 0s or 1s on a 3D grid. Later, Mandikal *et al.* [3] proposed a latent matching technique that matches the latent vector of a Variational AutoEncoder (VAE) [4] with the feature map of a 2D image, which has the same input and output and is highly accurate, to improve accuracy. However, this technique has the disadvantage of slow learning speed and large memory consumption during the learning process because it requires additional training of the VAE in addition to the deep learning network for 3D reconstruction. In addition, divergence loss, which prevents the value of the latent vector from diverging, requires additional acquisition information such as *x*-axis rotation and *y*-axis rotation of 2D images for learning. Later, Bin *et al.* [5] generated latent vectors directly using feature maps of 2D images without using VAEs, achieving better accuracy and learning speed and reducing the amount of memory required for learning. However, it still has the disadvantage of requiring additional acquisition information.

Therefore, in this paper, we proposed a technique for 3D point cloud reconstruction from 2D images using an efficient feature map extraction network. The uniqueness of the proposed method is as follows. First, it uses a novel feature map extraction network, which is about 27% more efficient than the existing methods in terms of memory. Second, by preserving the high-resolution features of 2D images, the accuracy can be improved by about 6.3% over the existing methods. Third, it uses divergence loss, which does not require information acquisition.

2. The bottom line

2.1. Overview

The overview of the 3D point cloud reconstruction technique from 2D images using an efficient feature map extraction network proposed in this paper is shown in **Figure 1**.

2.2. Efficient feature map extraction feature map extraction with high-resolution feature preservation using residual networks

2.2.1. Problems with Residual Network (ResNet)

ResNet [6] is a network designed to address the problem of reduced accuracy in existing deep networks for classifying objects in 2D images due to gradient vanishing. ResNet solved the problem of initial feature being lost in deep networks by effectively transferring initial features deep into the network. This helps prevent gradient vanishing and significantly improves accuracy. Most of the deep learning network structures created since then have applied ResNet's initial feature transfer method. To solve the problem of excessive memory usage due to its depth, ResNet applies a convolutional layer with a mask size of 7 × 7 and a stride of 2 in the first layer to reduce the image size by

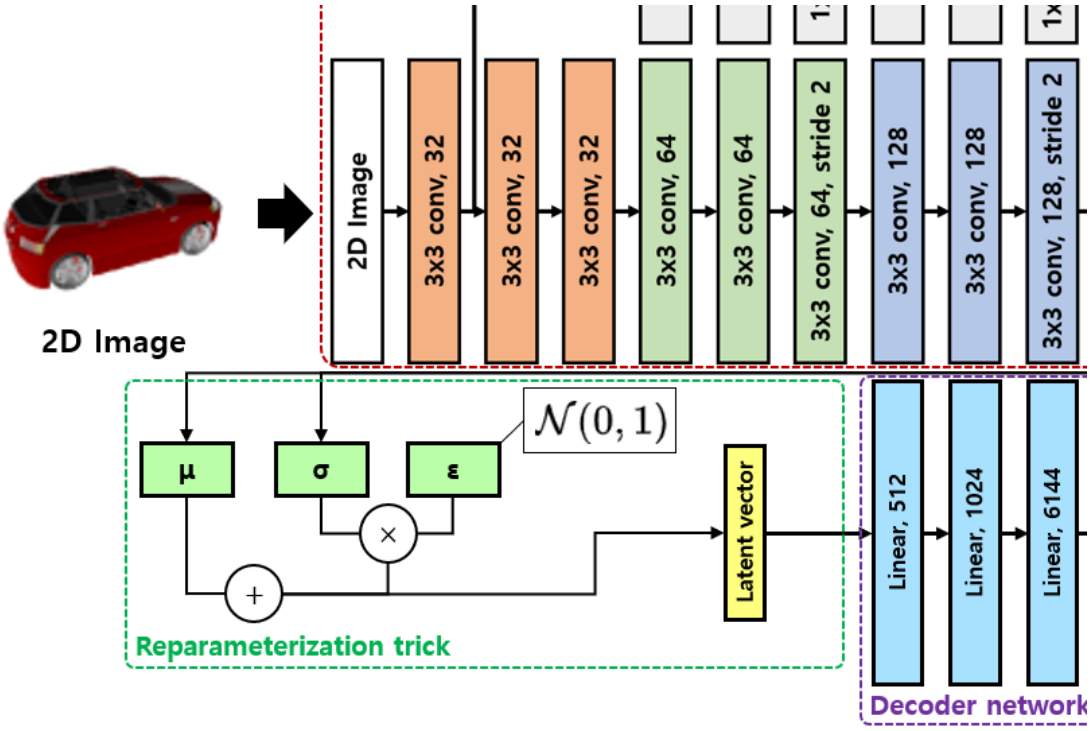

Figure 1. An overview of the proposed method

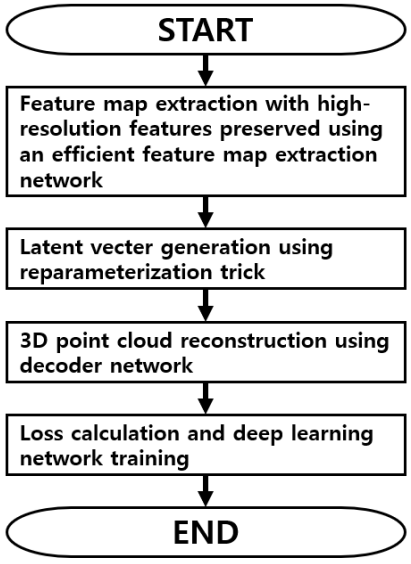

Figure 2. The learning process of the proposed method in this paper

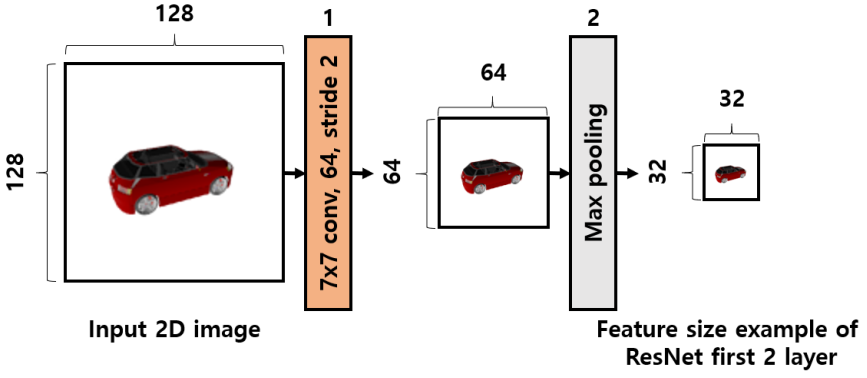

Figure 3. Problems in the process of reducing the size of the input 2D image passing through ResNet 1st and 2nd layers.

half, and then performs max pooling with a mask size of 2 × 2 in the next layer to further reduce the size of the image by half. Therefore, a large amount of features from the original image are lost in order to create and maintain deep layers. **Figure 3** illustrates the problem of reducing the size of the input 2D image as it passes through the first and second layers of ResNet.

2.2.2. Efficient feature map extraction network proposed in this paper

In this paper, we proposed a network to solve the feature loss problem of ResNet and perform efficient feature map extraction in terms of memory. The proposed network does not downsize the input image until mid-way into the deep-learning network, so important information required for 3D point cloud reconstruction is preserved, hence improving accuracy. The problem of increased memory due to image size is solved by reducing the number of channels and making the depth of the deep learning network shallow. **Figure 4** shows an efficient feature map extraction network with no reduction in the size of the input 2D image.

ResNet uses multiple bottleneck layers as a way to convey the initial features of a 2D image in depth. However, multiple layers of bottleneck layers can also cause a decrease in accuracy due to gradient vanishing. Therefore, in this paper, unlike ResNet, we directly passed the initial features to the last layer of the feature map network without passing through multiple layers, which minimized the loss of initial features and further improved the accuracy. **Figure 5** shows the difference between the 34th layer of ResNet and the initial feature

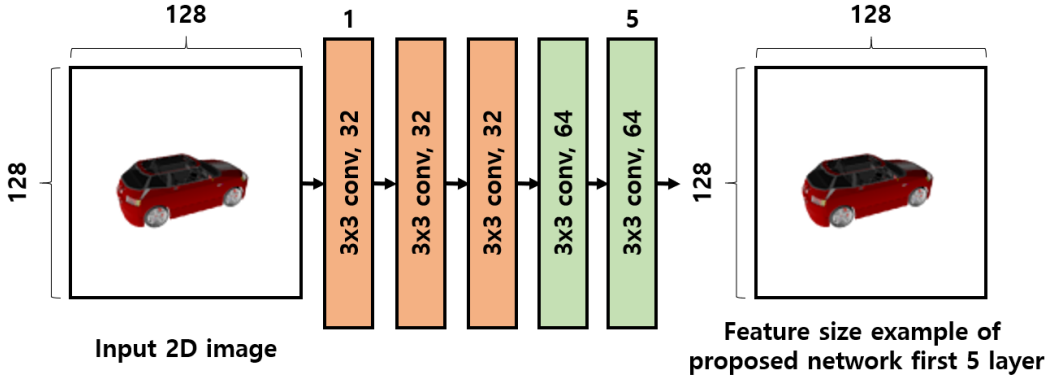

Figure 4. Efficient feature map extraction network without reducing the size of the input 2D image

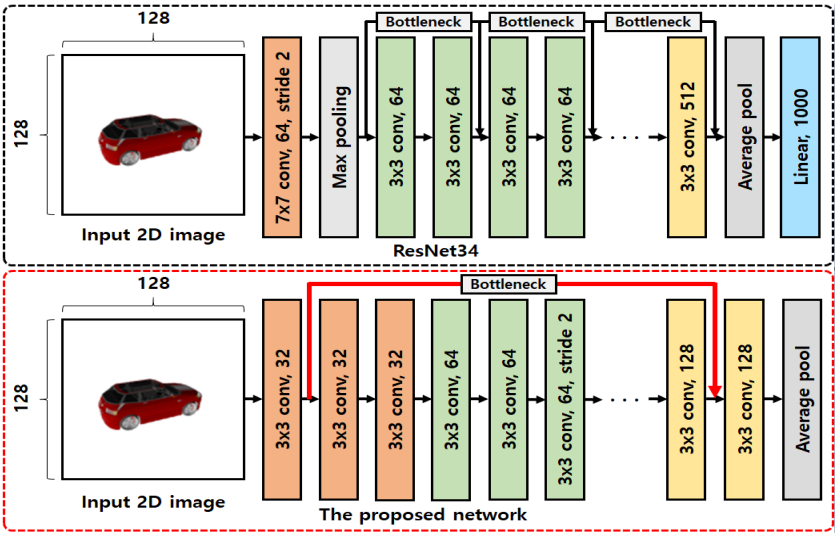

Figure 5. Difference between ResNet 34 layer and the initial feature delivery method of the proposed network.

passing method of the proposed network.

2.3. Generating a latent vector using the reparameterization trick

Reparameterization trick is a technique mainly used in Variational AutoEncoder (VAE) to generate a latent vector in the form of a normal distribution using a feature map extracted from the input data. Data constrained in the form of a normal distribution has less randomness, and it is convenient to estimate some values using a deep learning network. In this paper, we extracted a feature map of size (1, 1000) from a 2D image and used it as input data for a linear layer to generate a mean μ and standard deviation σ of size (1, 100). Then, σ was multiplied by ε, a random value of size (1, 100) that followed a normal distribution, to limit the randomness of the standard deviation. The latent vector was then generated by adding μ and σ together. **Figure 6** shows the process of generating a latent vector using a feature map.

2.4. Reconstructing a 3D point cloud using a decoder network

The latent vector generated from the feature map of the 2D image by the reparameterization trick was reconstructed into a 3D point cloud using a Decoder network with three linear layers. The LeakyReLU activation function was applied in multiple layers to

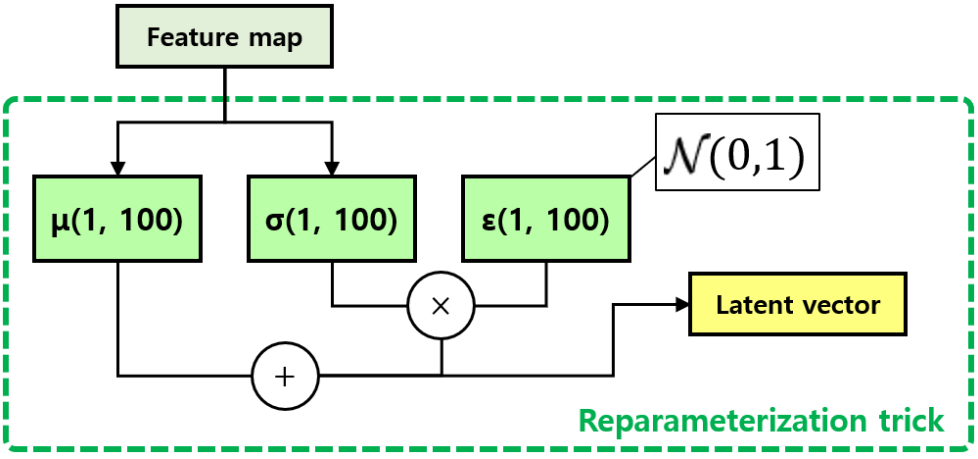

Figure 6. Latent vector generation process using feature map

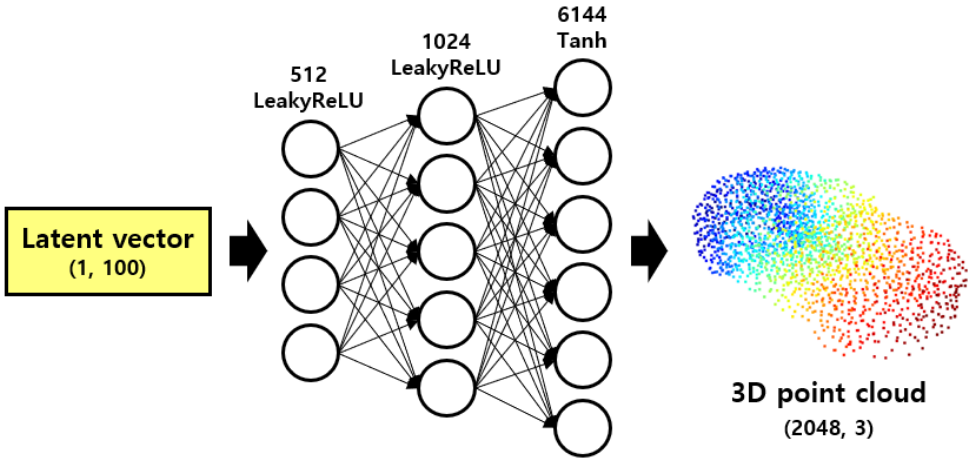

Figure 7. Decoder network structure used in this paper.

make the data nonlinear to improve the estimation performance of the decoder network and improve the reconstruction accuracy of the 3D point cloud. The latent vector of size (1, 100) was reconstructed into a 3D point cloud with three-dimensional coordinate information for 2048 points of size (1, 512), (1, 1024), (1, 6144), and finally (2048, 3). **Figure 7** shows the structure of the decoder network used in this paper.

2.5. Computing loss and training a deep learning network

In this paper, a composite loss function that combines several loss functions is used to train a deep learning network. The divergence loss proposed in this paper prevents the denominator from becoming zero and diverging to infinity when generating the latent vector. In addition, the Chamfer distance (CD) loss and Earth mover's distance (EMD) loss improve the accuracy of the reconstructed 3D point cloud by making the distance between points in different 3D coordinate systems closer. The loss function used to train the deep learning network in this paper is shown in Equation (1), where L_{DIV} represents the divergence loss.

$$Loss = L_{DIV} + CD_{Loss} + EMD_{Loss} \quad (1)$$

2.5.1. Divergence loss

In this paper, we use KullbackLeibler (KL) divergence as the divergence loss. KL divergence constrains μ and σ so that the latent vector generated by it follows the form of the standard deviation. On the other hand, the divergence loss 2D image such as Equation (2) used by Mandikal *et al.* and Bin *et al.* requires the rotation angle information of the acquired camera and uses it to ensure that the latent vectors generated from the feature maps of 2D images with similar acquisition angles follow a similar shape. Where η is the weight that determines the learning reflection degree of divergence loss, Φ_I is the shooting angle of the 2D image, and δ is the penalty angle that determines the reflection degree of angle difference.

$$(\sigma - \eta e^{-\frac{(\phi_i - \pi)^2}{\delta^2}})^2 \quad (2)$$

Requiring information about not only 2D images but also the shooting angles in training can make dataset collection challenging. Therefore, in this paper, we calculated the divergence loss using only μ and σ which form the latent vector without using the shooting angle. Unlike a previous paper [7] that limited the randomness of the latent vector, we used randomness to increase the diversity of information to improve the accuracy of 3D point cloud reconstruction. Equation (3) shows the divergence loss used to train the deep learning network in this paper.

$$L_{DIV} = -0.05\,(1 + \sigma - \mu^2 - e^{\sigma}) \quad (3)$$

2.5.2. CD loss

CD loss is a function that finds the closest points between two different 3D coordinate arrays. The CD loss function calculates the points closest to each other using a bidirectional calculation. However, when there are points that share the closest point, three or more points may merge into one, resulting in the loss of a point that forms the outline of an object. To solve this problem, deep learning-based 3D point cloud reconstruction was used in conjunction with EMD loss, which preserved the overall shape. Equation (4) shows the formula for CD, and **Figure 8** shows an example of the calculation of CD in both directions.

$$CD(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \| x - y \|_2^2 + \sum_{y \in S_2} \min_{x \in S_1} \| x - y \|_2^2 \quad (4)$$

2.5.3. EMD loss

EMD loss is similar to CD loss in that it is a function that finds the closest point between two different 3D coordinate arrays. EMD loss calculates distances through unidirectional calculations. Therefore, the EMD loss has a significantly lower likelihood of

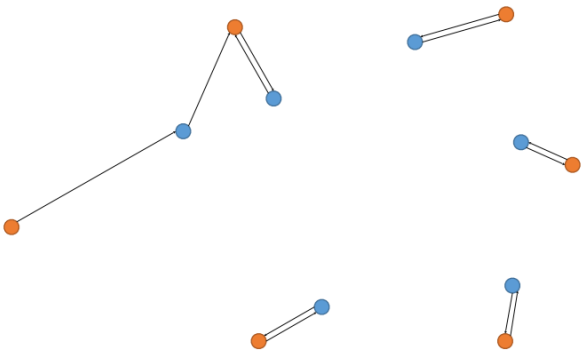

Figure 8. Example of CD calculation

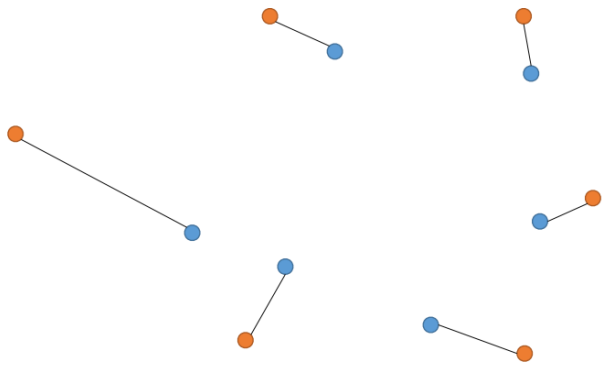

Figure 9. Example of EMD calculation.

multiple points merging into one compared to the CD loss. Equation (5) shows the formula for EMD loss and **Figure 9** shows an example of a unidirectional calculation using EMD loss.

$$EMD(S_1, S_2) = \sum_{\Phi: S_1 \to S_2} \min_{x \in S_1} \| x - \Phi(x) \|_2 \quad (5)$$

2.6. Performance experiments

2.6.1. Experimental environment

The OS and hardware used in the experiments consist of an Intel i7-10700k 3.8GHz CPU, 16GB RAM, and NVIDIA GeForce RTX 2770 (VRAM 8GB) GPU based on Windows 10 64bit OS. The deep learning libraries used were Pytorch 1.9.1, CUDA 11.1, and cuDNN 8.2.1. In this paper, 13 categories of the ShapeNet dataset [8] were used in the experiments to create the same experimental environment as other studies. **Table 1** shows the categories of the ShapeNet dataset used in our experiments, and **Figure 10** shows an example of the ShapeNet dataset used in our experiments. for objects built by researchers at Princeton, Stanford, etc.

2.6.2. Experimental results

As a performance evaluation method, we compared the floating-point operations per second (FLOPs), which represent the memory required for the computation of CD, EMD, and network. To objectively evaluate the performance of the proposed method, we calculated the average results for 13 categories under the same conditions as similar studies. **Table 2** shows the comparison results of our proposed method with the methods in other papers, and **Figure 11** shows the results of 3D intra-cloud reconstruction from 2D

Table 1. 13 categories of Shape Net dataset used in the experiment

Category	Content
Object	Bench, chair, lamp, speaker, gun, table, cabinet, monitor, long chair, cell phone
Vehicle	Ship, airplane, car

Table 2. Comparison of the ShapeNet dataset reconstruction accuracy of the proposed method and the methods of other papers

Method	CD ↓	EMD ↓	FLOPs ↓
PSGN[2]	9.11	12.29	-
3D-1mnet	8.91	9.70	4G
3D-reconstnet	6.26	6.20	4G
Proposed method	5.87	5.81	2.9G

Figure 10. Example of the ShapeNet dataset used in the experiment

images of the ShapeNet dataset. From the results in **Table 2**, the CD value was 5.87, the EMD value was 5.81, and the FLOPs value was 2.9G for the method proposed in this paper. The lower the CD and EMD values, the more accurate the reconstructed 3D point cloud, which means that it is closer to the original. Also, the lower the FLOPs value, the less memory is required for the deep-learning network. Therefore, it can be concluded that the CD, EMD, and FLOPs performance of the proposed method was better than the methods published in other papers.

In the existing method, the memory optimization method through initial image reduction would result in the loss of important features. In contrast, the technique proposed in this paper preserved important features by maintaining the resolution and reducing the number of channels. In addition, by preserving important features, the accuracy was improved even with an efficient network structure, achieving lower FLOPs than the existing method. We found that the latent vector generated by the reparameterization trick becomes too stereotyped when constrained by KL divergence loss. In addition, the reflection ratio of KL divergence is reduced and the latent vector can contain more diverse information. Therefore, the 3D point cloud reconstruction using this method was more accurate. **Figure 11** shows the results of 3D point cloud reconstruction from 2D images of the ShapeNet dataset.

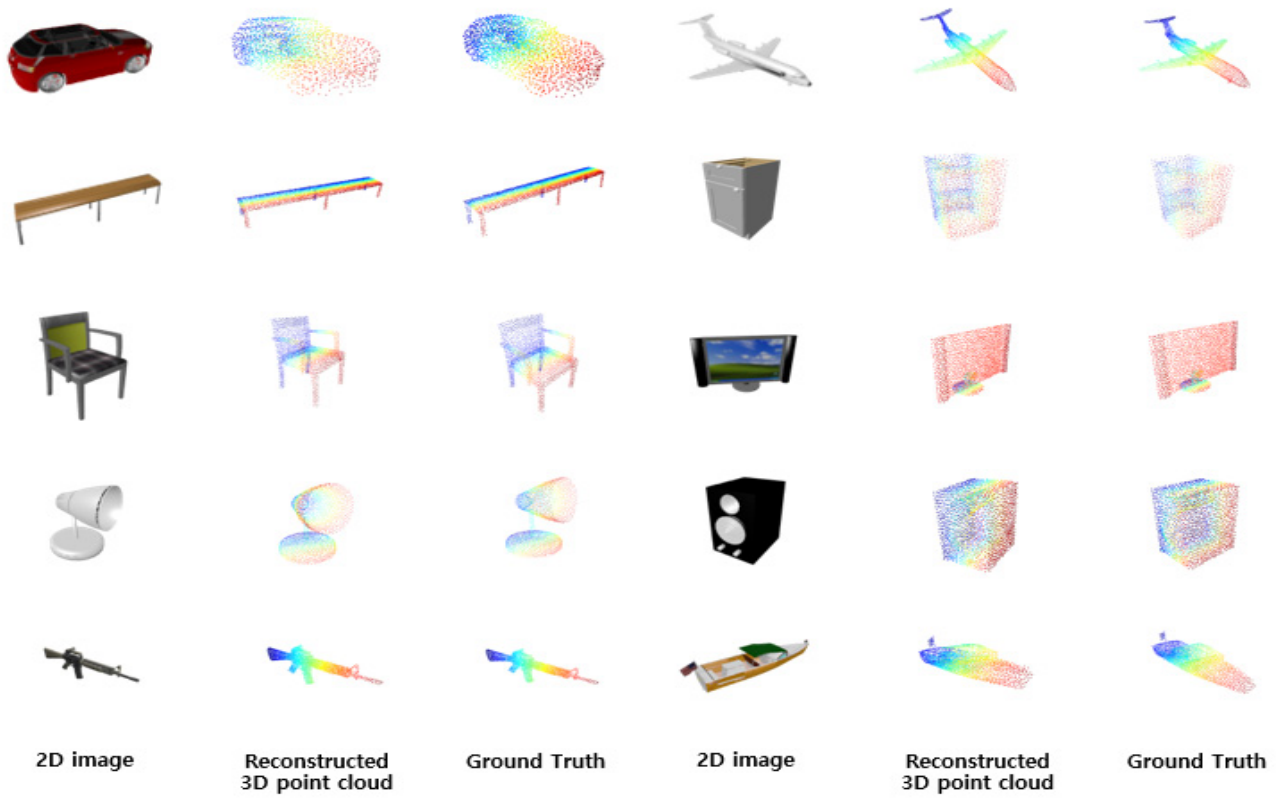

Figure 11. 3D point cloud reconstruction results from a 2D image of the ShapeNet dataset

3. Conclusion

In this paper, we proposed a technique for 3D point cloud reconstruction from 2D images using an efficient feature map extraction network. This method preserved the high-resolution features of 2D images, which improved the accuracy by about 6.3% over the existing method. Second, a new feature map extraction network was used, which was about 27% more efficient than the existing method in terms of memory. Third, we used a divergence loss that does not require image information. To objectively evaluate the performance of the proposed method, we used the ShapeNet dataset and experimented with the same method as similar studies, and found that the CD value of the proposed method was 5.87, the EMD value was 5.81, and the FLOPs value was 2.9G. The lower the CD and EMD values, the more accurate the reconstructed 3D point cloud. Besides, the lower the FLOPs value, the less memory is required for the deep learning network.

For future research, it is necessary to study a new divergence loss derivation that can contain more information in the latent vector for 3D point cloud reconstruction than the generalized KL divergence.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1F1A1066371).

Disclosure statement

The authors declare no conflict of interest

References

[1] Choy CB, Xu D, Gwak J, 2016, 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. Proceedings of the European Conference on Computer Vision, 628–644.

[2] Mandikal P, Navaneet KL, Agarwal M, 2018, 3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image. Proceedings of the British Machine Vision Conference (BMVC). https://doi.org/10.48550/arXiv.1807.07796

[3] Fan H, Hao S, Guibas L, 2016, A Point Set Generation Network for 3D Object Reconstruction from a Single Image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 605–613. https://doi.org/10.48550/arXiv.1612.00603

[4] Kingma DP, Welling M, 2014, Auto-Encoding Variational Bayes. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1312.6114

[5] Li B, Zhang Y, Zhao B, et al., 2020, A Single-View 3D-Object Point Cloud Reconstruction Network, IEEE Access, 8: 83782–83790. https://doi.org/10.1109/ACCESS.2020.2992554

[6] He K, Zhang X, Ren S, et al., 2016, Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.48550/arXiv.1512.03385

[7] Higgins I, Matthey L, Pal A, et al., 2016, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Proceedings of the International Conference on Learning Representation.

[8] Chang AX, Funkhouser T, Guibas L, et al., 2015, Shapenet: An Information-Rich 3D Model Repository. https://doi.org/10.48550/arXiv.1512.03012