Research on Multi-View Stereo Network Based on Self-Attention Mechanism

Wenkai Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 2562491043@qq.com

Leilei Fan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 2547462712@qq.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yujun@xatu.edu.cn

Zhiyi Hu
Engineering Design Institute
Army Research Laboratory
Beijing, 100042, China
E-mail: 763757335 @qq.com

***Abstract*—As the technologies of virtual reality and augmented reality rapidly advance, the demand for high-quality 3D models has been growing exponentially. However, the Multi-View Stereo Network (MVSNet) for 3D reconstruction has faced issues with the inaccurate extraction of global image information and depth cues. In response to these challenges, this paper presents enhancements to MVSNet. First, the self-attention mechanism is introduced to enhance MVSNet's ability to capture global information in images. Second, a residual structure is added to mitigate the accuracy loss caused by the downsampling and upsampling of feature maps during the regularization process of cost volume, thus ensuring the integrity of information and transmission efficiency. Experimental results indicate that, in comparison with the original MVSNet, the SelfRes-MVSNet reduces the error rate by 1.3% in terms of overall accuracy and completeness, thereby improving the reconstruction effect from 2D images to 3D models.**

Keywords-3D Reconstruction; MVSNet; Depth Map; Multi-View; Deep Learning

I. INTRODUCTION

Multi-view 3D reconstruction technology, as a mainstream research topic in the field of computer vision, has significant research value in areas such as autonomous driving and medical diagnosis. Traditional methods, which rely on manually extracted features, have a more complex reconstruction process and the resulting model accuracy is not high [1]. In 2016, Choy et al. proposed the 3D-R2N2 model, which utilizes the property of Recurrent Neural Networks (RNNs) to accept variable-length sequences, enabling the reconstruction of 3D occupancy grids [2]. However, the model has a large number of training parameters and low accuracy, and may exhibit inconsistency issues when processing different images with variable input. To address this, researchers such as Haozhe Xie proposed new 3D reconstruction models named Pix2Vox [3] and Pix2Vox++ [4]. Both models assign a definite weight value to each voxel in the model, thereby fusing various parts of the object based on the weight values. The models focus on quickly and consistently retrieving the 3D geometry from single or multi-view imagery, improving the inconsistency issues that previous methods might encounter with different image inputs. However, the models' precision is not sufficiently refined for processing single-view images. Therefore, Wang et al. introduced the Pixel2Mesh reconstruction model [10]. The model is designed to generate a 3D mesh model from a single-view RGB image, utilizing Graph Convolutional Neural Networks (GCN) to progressively refine an initial ellipsoid. It is capable of producing a 3D mesh model with rich details and high reconstruction accuracy. Furthermore, in 2018, Yao et al. proposed a

DOI: 10.2478/ijanmc-2025-0021

network model named MVSNet [5-8], which introduced the differentiable homography transformation theory into neural networks and integrated camera model transformation into the neural network architecture. Chen et al. proposed a point cloud-based reconstruction model named Point-MVSNet [9], which directly converts the initially estimated coarse depth map into a 3D point cloud model, and then optimizes based on the point cloud model, which utilizes more efficiently the 3D geometry and 2D texture information. Although Point-MVSNet delivers impressive results, its training process is complex, and it still falls short in integrating global information to enhance model accuracy. In response to the stated issues, we put forward a self-attention mechanism and incorporates a residual structure into the original MVSNet algorithm, enabling the network to concentrate more on the information that is crucial for the current task. We summarize our contributions in three aspects as follows:

- The experimental introduce a self-attention mechanism that, while considering the correlations between different regions in the image, accounts for the modeling capability of long-distance dependencies, which enhances the feature extraction network's proficiency in discerning and incorporating contextual information and learns image features from both local and global perspectives.
- The experimental add the residual module. The improved residual module effectively helps the network suppress less relevant features, focusing on important features, thus enhancing the network's feature representation capability and improving the network's classification performance.
- The experimental results indicate that, in comparison with the original MVSNet algorithm, the improved algorithm has achieved a 0.014mm increase in accuracy, a 0.012mm improvement in completeness, and a 0.013mm reduction.

II. RELATED WORK

In the field of 3D reconstruction from multi-view images, deep learning technology has become the core force driving research progress. According to different technical characteristics, 3D reconstruction methods can be divided into two major categories: methods based on traditional convolutional networks and methods based on attention mechanisms.

A. Approaches Relying on Traditional Convolutional Networks

In 2021, Liu et al. introduced the Swin Transformer [11], a novel architecture that merges the strengths of CNNs in processing large - scale images and Transformers in capturing long - range dependencies. By leveraging local attention mechanisms and shifted window techniques, the Swin Transformer can effectively manage both global and local image features. In 2022, Dong and Yao et al. put forward the PatchMVSNet [12], an unsupervised multi - view stereo approach. It is tailor - made for reconstructing the 3D structure of weakly textured surfaces. By implementing a patch matching and depth estimation framework, the network is able to resolve the problem of matching ambiguity that often occurs under weak lighting and texture - deficient conditions. Different from other methods, PatchMVSNet focuses on enhancing the accuracy of 3D reconstruction in such challenging scenarios, providing a new solution for the field of multi - view stereo reconstruction. In 2023, Li Chenghuan et al. introduced the innovative R3D - SWIN [13] model. This model broke new ground by being the first to apply the Shifted Window Attention mechanism in voxel - based 3D reconstruction tasks. R3D - SWIN combines a Transformer Encoder and a CNN Decoder, aiming at single - view reconstruction problems. By establishing connections across different windows, it elevates the model's capacity to acquire information at multiple scales, enabling a more comprehensive and in - depth understanding of the data.

B. Methods Based on Attention Mechanisms

In 2020, Dosovitskiy et al. proposed the Vision Transformer [14] model, an innovative approach that applies the self-attention mechanism to image recognition. The Vision Transformer enables the model to pick up on long - range links between various image regions when processing image sequences. In 2021, Zehao Yu et al. introduced LA-Net [14]. It is a multi - view stereo matching network that capitalizes on attention mechanisms. LA-Net utilizes a long-range attention network to selectively aggregate reference features to each location, capturing long-range interdependencies across the entire scene. In 2023, Yu et al. presented a new method called ACR (Attention Collaboration-based Regressor) [16] for reconstructing hands in arbitrary scenes from monocular RGB images. ACR utilizes central and partial basis attention mechanisms to reduce interdependencies between hands and better handle hand interactions through cross-hand priors. In 2024, Xuanhao Yan et al. developed a deep learning network model called SA-Pmnet [20], which combines self-attention mechanisms and multi-scale feature extraction modules specifically for 3D reconstruction of forest scenes. SA-Pmnet enhances the capacity to seize image particulars through self-attention mechanisms, significantly improving 3D reconstruction rates and breast diameter extraction accuracy in complex forest environments.

III. MVSNet

The fundamental architecture of MVSNet [6] comprises the following steps: feature extraction, differentiable homography transformation, cost volume construction, cost volume regularization, and depth map estimation and optimization. The following sections detail the cost volume construction and cost volume regularization modules.

A. Cost Volume Construction

Cost volume construction primarily involves merging multiple feature volumes obtained from the feature extraction module. To achieve this, Yao et al. proposed a construction method based on the variance cost metric M in the MVSNet algorithm [6]. This method can handle an arbitrary number of 2D images, and its core principle is to calculate the variance of image feature volumes of the same object from different viewpoints. The result of the variance cost calculation is used to construct the cost volume, where each voxel represents the consistency of all input image feature maps at a specific hypothetical depth. The smaller the variance, the higher the consistency between the feature maps at that depth. The calculation method of the feature volume variance is shown in Equation (1).

$$\begin{cases} V = \frac{W}{4} \cdot \frac{H}{4} \cdot D \cdot F \\ C = M\left(V_1, \ldots, V_2\right) = \frac{\sum_{i=1}^{N} (V_i - \bar{V})^2}{N} \end{cases} \quad (1)$$

In this equation, V represents the size of the feature volume; W is the width of the input image; H is the height of the input image; D is the number of depth hypotheses; F is the number of channels in the feature map. $\bar{V}$ is the average volume of all feature volumes, C is the cost volume, and M is the variance cost metric of the image feature volumes from different viewpoints.

B. Cost volume regularization

The network structure for cost volume regularization is depicted in Figure 1 Cost volume regularization primarily involves the refinement of the cost volume and the generation of probabilities P at all depth hypothesis values d. Subsequently, the most suitable depth position is selected for each pixel based on the P values. The MVSNet algorithm employs a multi-scale 3D CNN model for cost volume regularization, utilizing an "encoder-decoder" architecture. As 3D convolutional kernels include an additional depth dimension compared to 2D kernels, 3D convolution inherently results in a substantial increase in computational parameters and requires more GPU resources. To minimize resource consumption as much as possible, the network structure reduces the number of channels from 32 to 8 in the initial 3D convolutional layer, includes 2 convolutional layers within each module, and the final convolutional layer outputs a single-channel volume. A softmax operation is then performed to normalize the probabilities of the cost volume along the depth direction, yielding the probability distribution of pixels at each hypothesized depth position. However, due to the excessive use of downsampling operations throughout the network structure and the simple addition fusion used during the decoding stage for information integration, the MVSNet algorithm fails to fully leverage the structure, resulting in inaccurate probability values for the pixel information generated.

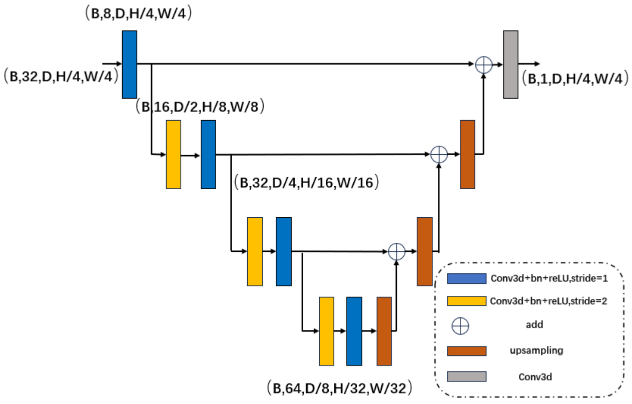


Figure 1. Cost Volume Regularization Network Architecture

Addressing the issues in the feature extraction module and cost volume regularization module of the MVSNet algorithm, We propose SelfRes-MVSNet, which is an improvement on MVSNet. Firstly, to mitigate the lack of detailed feature information during extraction, this paper introduces a self-attention mechanism that jointly learns image features from both local and global perspectives. Secondly, to tackle the issue of lost regularization information in the cost volume, a residual structure is introduced to ensure the integrity of the input information.

IV. SELFRES-MVSNET

A. *Self-Attention Mechanism Module*

CNNs face inherent challenges when dealing with global image associations, limiting their capacity to fully extract global image information. In the MVSNet framework, the feature extraction module usually relies on convolutional operations. It uses local receptive fields and shared weights to establish relationships within local regions, but this approach has limitations in capturing global context.

However, this approach does not account for the associations between different areas in the image, resulting in a relatively limited modeling of long-range dependencies. To overcome this limitation, we introduce a self-attention mechanism that enables the feature extraction network to consider dependencies between different regions of the image from a global perspective, thereby enhancing the network's capability to capture contextual information.

The essence of the self-attention mechanism is to analyze the significance of current information for the task at hand and to allocate attention based on its importance. This is primarily achieved through the computation of attention weight coefficients, which reflect the degree of importance of the information. The network architecture is illustrated in Figure 2.

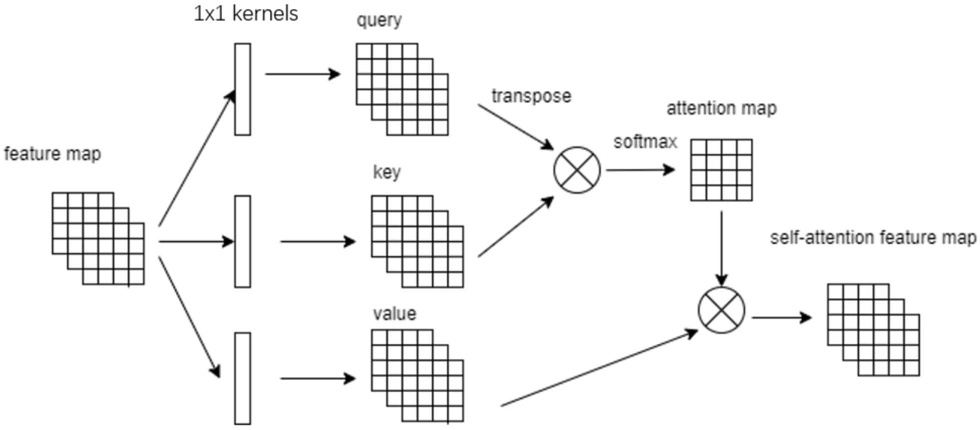


Figure 2. Self-attention Architecture

Initially, the self-attention mechanism is defined by three 1×1 convolutions, which correspond to three learned weight matrices W^q、W^k and W^v. By conducting three separate 1×1 convolutions on the feature map, we obtain the query, key, and value, as indicated in Equation (2).

$$Q_{\mathrm{ij}} = W^q x_{ij}, K_{\mathrm{ij}} = W^k x_{ij}, V_{\mathrm{ij}} = W^v x_{ij} \qquad (2)$$

In the formula, x is the feature map vector, while W^q、W^k and W^v are the weight matrices.

Next, a dot product operation is performed between the query matrix and the key matrix. The result is then divided by a constant $\sqrt{d_k}$, where d_k is the feature dimension. A softmax operation is applied to normalize the weight values obtained from the previous step, so that the sum of all weights is 1. The resulting values represent the importance of the relevant information for the task. Finally, the normalized weights are used to perform a weighted sum operation with the corresponding values to acquire the ultimate attention. The calculation method is shown in Equation (3).

$$Attention\left(Q,K,V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

By incorporating the self-attention mechanism into the feature extraction network, the issue of relying solely on convolutional layers to expand the receptive field is addressed. This not only enables the network to focus on learning more important information but also allows it to rapidly capture long-range dependencies, thus selecting information that is more closely related to the current task. In this paper the self-attention mechanism is introduced after the last convolutional layer. Through the union of convolutional operation and the self-attention mechanism, the network can extract features from both local and global regions, thereby enhancing its ability to capture contextual information from the data. The enhanced feature extraction structure is shown in Figure 3.

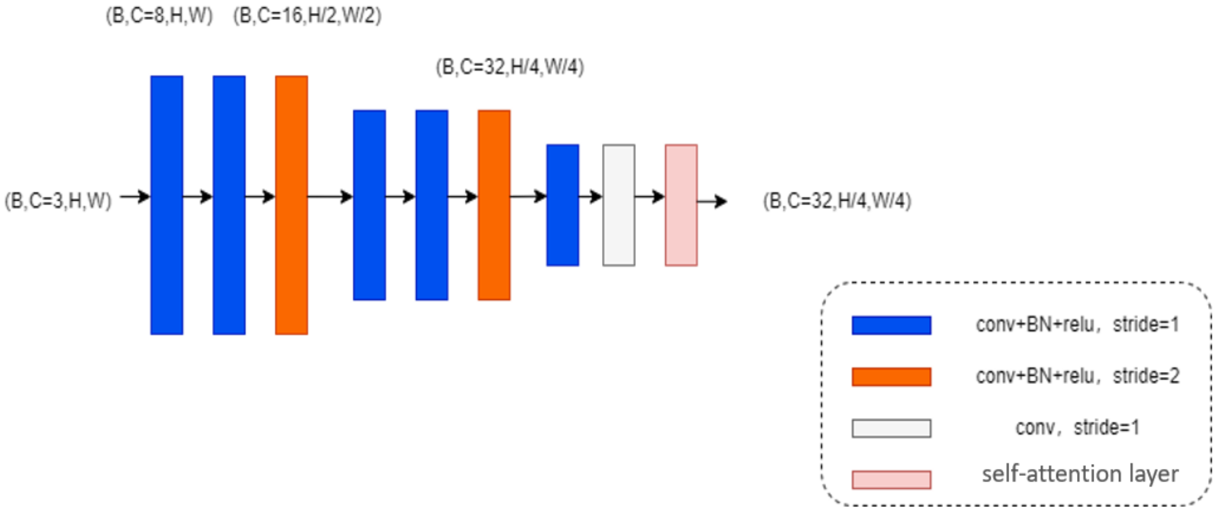


Figure 3. The Enhanced Feature Extraction Block

B. ResNet Block

In Convolutional Neural Networks (CNNs), the convolution operation is particularly adept at extracting features from specific parts of an image with a robust capability. However, it has inherent limitations in dealing with the internal connectivity of the entire global image, which in turn restricts the neural network's capacity to extract global contextual information from the image. In MVSNet, the feature extraction module conventionally relies solely on convolutional operations to extract local image features, primarily modeling local regions through local receptive fields and shared weights, without considering the interconnections between different areas within the image. This leads to a relatively limited modeling of long-range dependencies. To address this issue, MVSNet

Incorporates a residual structure, as illustrated in Figure 4 the residual module enhances the integration of information by adding the input data to the output results obtained after passing through the convolutional layers, thereby enriching the data information contained in the feature maps. However, in deep networks, this structure can result in a significant amount of redundant information.

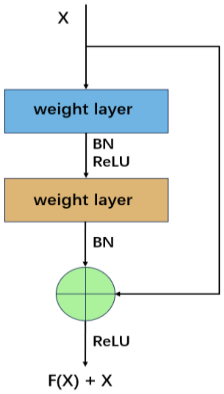


Figure 1.

Figure 4. Residual Block

In Convolutional Neural Networks (CNNs), the convolution operation excels at extracting features from specific parts of an image with a strong capability. However, it falls somewhat short when dealing with the internal relationships within the entire global image, thereby limiting the network's capacity to capture global contextual information from the image. In MVSNet, the feature extraction module relies predominantly on convolutional operations to capture local image features, primarily by using local receptive fields and shared weights to model relationships within local regions, without taking into account the connections between different areas of the image. This results in a relatively constrained modeling of long-range dependencies. To effectively assist the network in suppressing weakly correlated features, focusing on important features, and thereby enhancing the network's feature representation capabilities and improving its classification performance, this paper introduces a residual module into MVSNet and improves upon the residual module, with the improved residual module shown in Figure 5.

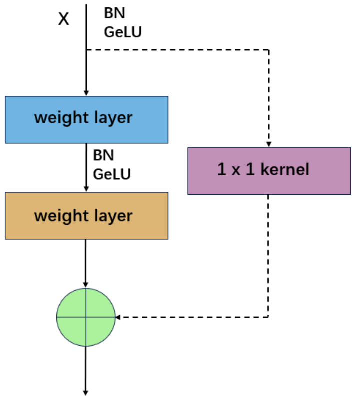


Figure 5. Pre-activated Residual Block

Firstly, since the shape of the input features may not match the shape of the output features, a 1×1 convolutional layer is added to the shortcut connection to ensure consistency in shape between the input and output features.

Secondly, the enhanced ResNet block in this paper adopts a pre-activation approach, where the data undergoes a batch normalization (BN) layer and an activation function layer before the convolutional operation. In the original ResNet block, the data first undergoes the convolutional operation and then the BN and activation operations. Although the data is standardized during this process, after merging with the values from the shortcut path, the combined data may no longer be standardized when input to the next convolutional layer. Adopting the pre-activation approach ensures that the data remains standardized when input to the convolutional layer, which helps maintain the same distribution of network inputs, reduces network training time, and mitigates gradient vanishing.

Finally, we replace the ReLU activation function with the GeLU activation function. The highlight of the GeLU function lies in its concept of random regularization; its output is probabilistic and its derivative is continuous, which makes gradient propagation smoother. GeLU introduces characteristics similar to the Sigmoid function in its nonlinear transformation, allowing the output of the GeLU function to span a broader range, thus helping to accelerate the model's convergence rate. The formula for the GeLU function is expressed as $\mathrm{GeLU}(x) = x * \Phi\ (x)$, where $\Phi\ (x)$ is the cumulative distribution function of the Gaussian distribution. In this function, if the value of x decreases, the value of $\Phi\,(x)$ also decreases accordingly, which means the probability that the input is "dropped out" increases. Therefore, the GeLU activation function is stochastically dependent on the input value, making the entire activation process more flexible compared to the ReLU activation function. The improved cost volume regularization network structure is shown in Figure 6, where only the enhanced residual network structure is applied to the encoding stage of the original network structure. This improvement aims to further enhance the network's ability to capture global and local features, as well as to strengthen the network's stability and convergence speed.

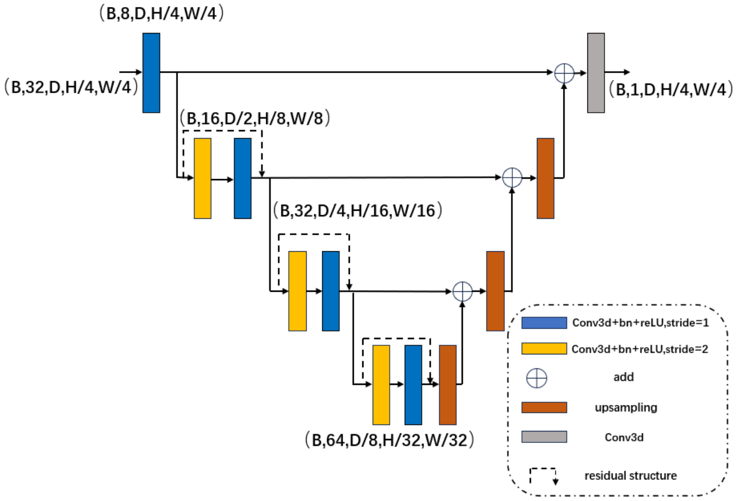


Figure 6. Improved Cost Volume Regularization Network

V. EXPERIMENTS

A. Dataset

In this research, the experiments were carried out using the open - source DTU dataset from the website http://roboimagedata.compute.dtu.dk. This dataset contains 124 different scenes, consisting of 42,532 RGB images in total. Each object is captured in 49 or 64 images taken by cameras positioned at various locations under 7 different environmental lighting conditions, with each image having a resolution of 1600×1200. We allocated images from 79 scenes to the training set, images from 18 scenes to the validation set, and images from 22 scenes to the test set. Figure 7 illustrates the 3D point cloud models of some objects in the DTU-dataset.

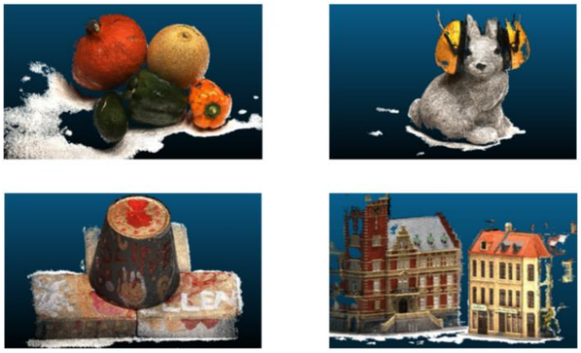


Figure 7. Partial Scenes in DTU Dataset

The experiments in this paper were conducted on an Ubuntu 18.04 operating system. The hardware setup included an Intel Xeon Gold 5118 processor, an Nvidia TITAN Xp GPU, and 32GB of RAM. The software environment comprised Python version 3.6.8 and the PyTorch framework.

We utilize three metrics to quantify the discrepancy between the reconstructed model and the actual model, which are: reconstruction accuracy (Acc), completeness (Comp), and the average of the two, namely overall performance (OA). The lower the values of these three evaluation indicators, the closer the 3D reconstruction effect is to the actual 3D model of the object. The specific calculation process is shown in Equation (4).

$$\begin{cases} Acc = \frac{1}{|S_1|}\sum_{x\in S_1}\min_{y\in S_2}||x-y||^2 \\ Comp = \frac{1}{|S_2|}\sum_{x\in S_2}\min_{y\in S_1}||x-y||^2 \\ OA = \frac{Acc+Comp}{2} \end{cases} \tag{4}$$

Here, S_1 is the aggregate of all points in the 3D point cloud model generated by the 3D reconstruction algorithm, and S_2 is represents the aggregate of all points in the actual 3D point cloud model from the dataset.

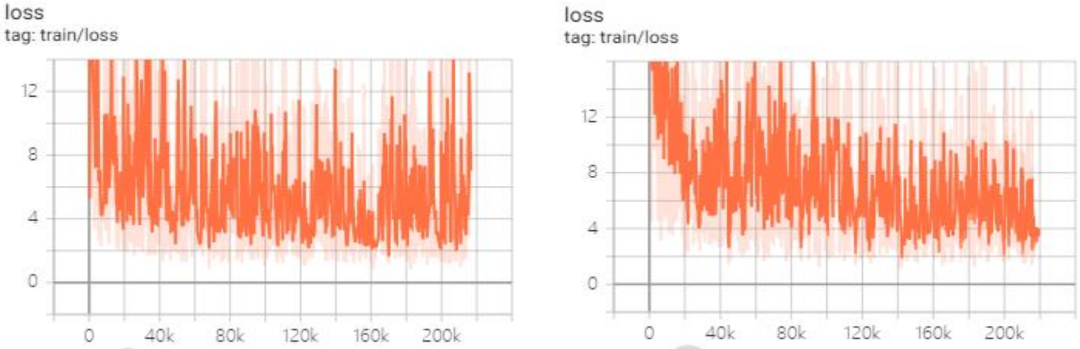
B. Results

The primary objective of this experiment is to verify the reconstruction efficacy of the SelfRes-MVSNet algorithm on the DTU dataset. It is assumed that the number of input images is N=3, and the parameter settings used during the training process are as shown in Table 1.

TABLE I. EXPERIMENTAL PARAMETERS

parameter	value
batch size	1
learning rate	0.001
epoch	16
Adam-β_1	0.9
Adam-β_2	0.999

The original MVSNet and the SelfRes-MVSNet were both trained on the DTU dataset. After the completion of the training process, the TensorBoard tool was utilized to visualize the Loss data generated during training. As shown in Figure 8.



（a）MVSNet

（b）SelfRes-MVSNet

Figure 8. Loss Curve

From the loss curve chart in Figure 8, it can be observed that the fluctuation range of the SelfRes-MVSNet's loss curve is smaller compared to the original MVSNet, and the convergence value of the Loss has also decreased. This indicates that the 3D model reconstructed by the SelfRes-MVSNet is closer to the actual point cloud model. To better demonstrate the effect of 3D reconstruction, we have visualized the depth maps generated by the algorithms. As shown in the true depth map of Figure 9, it is a comparative result of the depth maps produced by the original MVSNet and the algorithm presented in this paper.

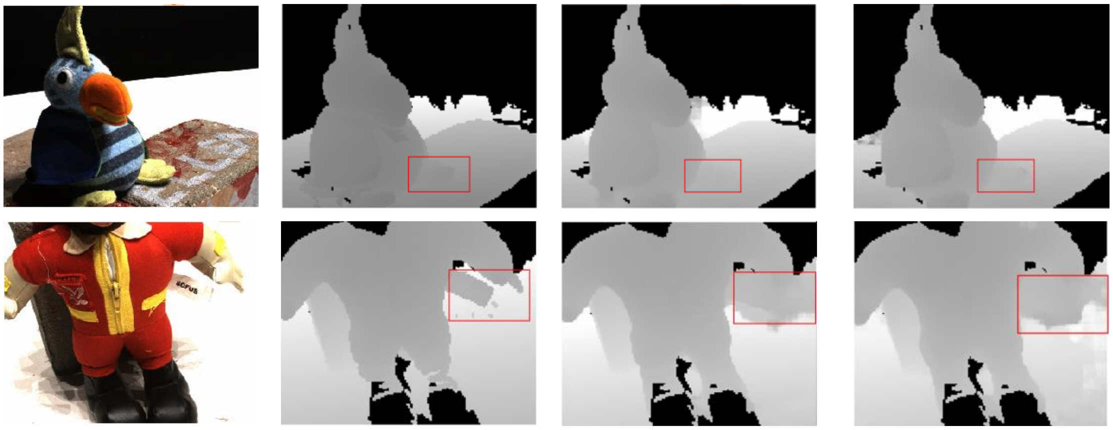

（a）Original Image （b）Real Depth Map （c）MVSNet （d）SelfRes-MVSNet

Figure 9. Depth Map Results

In Figure 9, in comparison of the feet and hands parts, (d) is closer to (b), thus (d) better demonstrates a more complete two-dimensional form compared to (c).

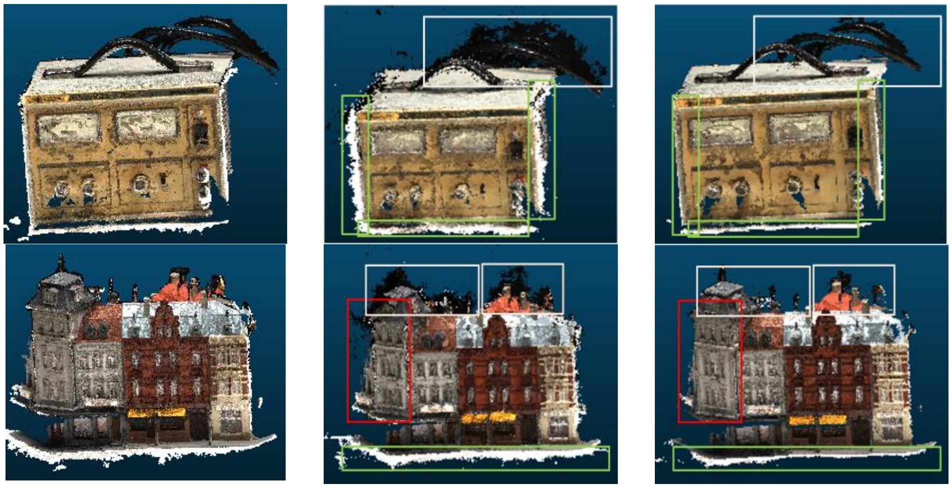

（a）Real Point Cloud （b）MVSNet （c）SelfRes-MVSNet

Figure 10. Comparison of the generated point cloud models

The SelfRes-MVSNet algorithm was quantitatively compared with other 3D reconstruction algorithms in terms of modeling quality, and the results are shown in Table 2. From Table 2, it can be concluded that compared to traditional 3D reconstruction methods, 3D reconstruction algorithms based on deep learning generally have lower errors than conventional approaches. The experimental results also show that, compared to the original MVSNet our selfRes MVSNet has reduced the error in accuracy by 0.014mm("mm" refers to the average error per pixel or per unit length (millimeter) of measurement.), the error in completeness by 0.012mm, and an overall error reduction of 0.013mm.

TABLE II. COMPARISON OF DIFFERENT 3D RECONSTRUCTION METHODS

Methods	（Acc）/mm	（Comp）/mm	（OA）/mm
Colmap	0.400	0.664	0.532
Gipuma	0.283	0.873	0.578
Camp	0.835	0.554	0.695
MVSNet	0.456	0.574	0.515
PointMVSNet	0.435	0.475	0.455
SelfResMVSNet	0.442	0.562	0.502

To observe differences between MVSNet & SelfRes - MVSNet more intuitively, the point - cloud models are shown in Figure 10.

From Figure 10, it can be observed that the point cloud model generated in (c) is clearer and has a higher degree of completeness compared to the model generated in (b). In Figure 10 (b), the original MVSNet is evidently affected by noise, appearing quite rough overall, with many dark shadows around, blurred edge areas, and distinct jagged shapes. In Figure 10 (c), the model generated by the SelfRes-MVSNet is overall clear, with smoother edge areas, and the gap on the upper left side of the house model is also clearly visible. This fully illustrates that the improved algorithm has a stronger capability in handling detailed information compared to the original MVSNet.

VI. CONCLUSIONS

This paper proposes an SelfRes-MVSNet based on deep learning, which effectively addresses the issues of missing detail feature information and loss of cost volume regularization information in the original MVSNet for 3D reconstruction by optimizing the feature extraction module and refining the cost volume regularization strategy.The experimental results indicate that, compared to the original MVSNet algorithm, the improved MVSNet algorithm has reduced the error in accuracy by 0.014m, decreased the error in completeness by 0.012mm, and also achieved an overall performance error reduction of 0.013mm. In summary, the MVSNet outperforms the original MVSNet in terms of accuracy, completeness, and overall performance.

REFERENCES

[1] Xie Qiqi. Multi-view 3D reconstruction based on MVSNet.Qinghai Normal University, 2024.

[2] Shi Shuaijie. Research on 3D reconstruction technology for monocular vision based on voxels and point clouds. Harbin Institute of Technology, 2021.

[3] Xie, Haozhe et al. "Pix2vox: Context-Aware 3d Reconstruction from Single and Multi-View Images", arXiv: Computer Vision and Pattern Recognition abs/1901.11153.1 (2019): 2690-2698.

[4] Xie H, Yao H, Zhang S, et al. Pix2Vox++: Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images. International Journal of Computer Vision, 2020, 128(12): 2919-2935.

[5] Feng Yajuan. Research on MVS 3D Reconstruction Algorithm Based on Deep Learning. Shanxi University, 2023.

[6] Yao, Luo et al. "Mvsnet: Depth Inference For Unstructured Multi-View Stereo",European Conference on Computer Vision 11212. (2018): 785-801.

[7] Wang Siqi, Zhang Jiaqiang, Li Liyuan, Li Xiaoyan, Chen Fansheng Application of MVSNet in 3D reconstruction of spatial targets. China Laser: 1-18 [2022-12-24].

[8] Yu Jingwei Research on Multi perspective Deep Estimation Methods Based on Deep Learning. Shenyang University of Technology, 2022.

[9] Rui C, Songfang H, Jing X, Hao S, et al. Point-Based Multi-View Stereo Network[C], IEEE International Conference on Computer Vision, 2019, 2019(1): 1538-1547.

[10] N. Wang, Y. Zhang, Z. Li, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 52-67.

[11] Liu, Ze et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 9992-10002.

[12] Dong, Hao-Chen and Jian Yao. “PatchMVSNet: Patch-wise Unsupervised Multi-View Stereo for Weakly-Textured Surface Reconstruction.”ArXiv abs/2203.02156 (2022)

[13] Li, Chenhuan et al. “R3D-SWIN: Use Shifted Window Attention for Single-View 3D Reconstruction.” ArXiv abs/2312.02725 (2023)

[14] Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ArXiv abs/2010.11929 (2020)

[15] Zhang, Xudong et al. “Long-range Attention Network for Multi-View Stereo.” 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (2021): 3781-3790.

[16] Yu, Zheng-Lun et al. “ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction.” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023): 12955-12964.