Journal of Image Processing Theory and Applications (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2023.060113
ISSN 2560-6239 Vol. 6 Num. 1

Multi-View Stereo Matching Method Based on Recursive Neural Networks

Jiajia Liu[1,a], Guoliang Jiang[1,b,*]

[1]School of Avionics and Electrical Engineering, Civil Aviation Flight University of China,
Guanghan, Sichuan, 618307, China
[a]cafucjgl@yeah.net, [b]1078079393@qq.com
[*]Corresponding author

Keywords: Cost Volume Aggregation, Recursive Hierarchy, Multi View Stereo, Three-Dimensional Reconstruction

Abstract: This article proposes a recursive layered network reconstruction method with pixel-wise attention cost aggregation to address the problems of textureless areas and poor reconstruction results at scene edges in multi-view stereo matching methods. First, multi-scale features of multiple images are extracted through downsampling and transformed into a cost volume using three-dimensional differentiable homography. Then, a pixel-wise attention aggregation module is added to the cost volume aggregation stage to reweight different pixels and generate a new cost volume. Next, a network with recursive layers is used to regularize the cost volume, replacing the traditional 3D CNN network, and an initial depth map is generated. Finally, the filtered and refined depth maps are merged to generate a three-dimensional dense point cloud. Experimental results show that the proposed network model improves completeness, accuracy, and overall quality by 0.377, 0.363, and 0.370, respectively, compared to other network models, and produces more complete point cloud reconstructions in weak texture areas and scene edges.

1. Introduction

Multi-view stereo (MVS) is a method of reconstructing a real 3D scene based on the principle of multi-view geometry, which utilizes the parallax information between images from multiple viewpoints to infer the depth and 3D shape of the scene, and is weakly interfered by the environment and low-cost compared to the traditional methods that require hardware such as structured-light scanners to acquire the 3D scene. The dense point cloud generated by MVS has rich environmental information and can well reflect the spatial geometric relationship of objects, so it has become a popular method for a wide range of application scenes, such as cultural heritage protection, autonomous driving, virtual reality, etc. [1], so it has become a research hotspot in the field of 3D reconstruction technology. In the traditional multi-view stereo matching method, the SFM (Structure from Motion) algorithm first estimates the camera position and obtains the 3D sparse point cloud of the target, and then performs dense point cloud reconstruction. Traditional methods [2] rely on image changes in obvious regions to design manual features such as SIFT operator and ORB operator for image matching, but for rotating, lighting changes in the image feature extraction becomes difficult, so the traditional MVS method has a narrow range of applications and can not reconstruct texture-poor regions.

In recent years, deep learning has been widely used and achieved great success first in the field of image recognition [3], and then various stereo matching methods based on deep learning have also been proposed. Currently, neural network training is used in all deep learning based 3D reconstruction algorithms, and the global information of the scene is learned through convolutional units on the image features to achieve high accuracy and high integrity reconstruction results [4-6]. YAO et al [7] at the Hong Kong University of Science and Technology proposed an end-to-end network model MVSNet (multi-view stereo network) based on depth maps, which applies the costumers to 3D CNN regularization and thus depth regression, which greatly improves the performance of three-dimensional reconstruction, but it is affected by the number of depth samples, and occupies a high level of computational resources, and then Yao et al [8] improved the network model and proposed RMVSNet (Recurrent multi-view stereo network), which reduces the resource consumption but the effect is not obvious. Chen et al [9] proposed a direct point-based matching cost regularization method PointMVSNet (Point multi-view stereo network). View stereo network). The method adopts a coarse-to-fine depth estimation strategy to directly generate a 3D point cloud based on the initial depth map, but the reconstructed scene represented by the mesh and faceted sheet is not smooth, resulting in the loss of scene edge details. In order to alleviate the problem of high computing resource consumption, there are many methods in this study, such as CasMVSNet(cascaded multi-view stereo network); CVP-MVSNet(Cost volume pyramid Multi-view stereo network); UCS-Net(Uncertainty Sensing Cascade Stereoscopic Network) etc. This study first predicts low-resolution depth maps with large depth intervals, and then successively elevates the coarse sample to a fine strategy to increase the depth range and resolution [10-12]. Although the multi-stage approach reduces the graphics memory consumption, it is not obvious for image feature detail extraction and does not solve the problem of texture scarcity. Yan et al [13] from Peking University proposed D2HC-RMVSNet (Dense Hybrid Recurrent multi-view stereo network) using a recurrent recurrent neural network approach in the depth direction, which introduces different expansion layers to generate multiple scales of background information, and adopts a recurrent recurrent neural network in the regularization process to The regularization process uses a recurrent recurrent neural network to reduce the model memory footprint, and finally replaces the previous fixed viewpoint thresholding method with an overall consistency metric to retain more accurate depth values. This method has excellent results in the end, but the accuracy and completeness of the final generated point cloud model is not obvious. Therefore, the main problems of the current deep learning multi-view stereo are: 1) it is difficult to ensure the accuracy and completeness of the reconstruction effect while reducing the memory consumption; 2) the multi-view matching cost aggregation process rarely considers the pixel-by-pixel visibility problem, which leads to the poor quality of the final reconstruction, especially in the case of texture scarcity and view edges, which will result in a serious loss of details.

In order to solve the above problems, this paper proposes a hierarchical recursive multi-view stereo network with pixel-by-pixel attention aggregation module based approach. The hierarchical recurrent network consumes less memory than the classical hierarchical network and can reconstruct higher resolution images; meanwhile, the pixel-by-pixel attentional aggregation module is added to assign higher weights in the matched views to overcome the difficulties of texture scarcity and less edge information in complex fields.

2. Network Models

The design of the network model in this paper follows the rules of camera geometry, draws on the experience of previous MVS methods, and uses a typical learning-based MVS flow model. The network goes through several major modules including: feature extraction, costum construction and aggregation, costum regularization, and depth map estimation, and the network architecture is shown in Fig. 1. The input image is divided into 1 reference image and N-1 source images, and the feature maps of all images share weights using an encoder, and then 3D costomes are constructed by the differentiable singular response transform, and after regularizing the costomes to obtain the probability bodies, the probability bodies can be used to generate predicted depth maps as shown in Fig. 1, and finally all depth maps are filtered and fused to obtain a dense point cloud.

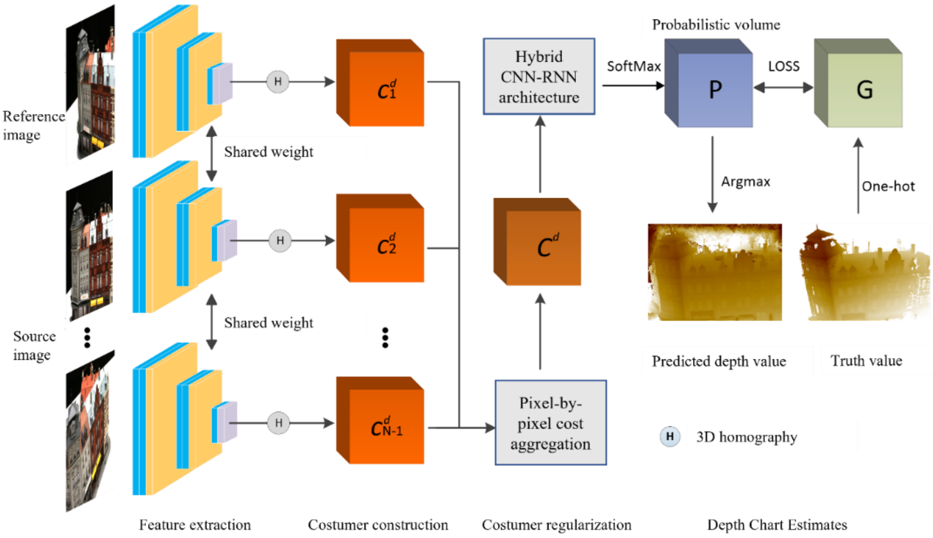

Figure 1: Overall network structure

2.1 Feature Extraction

First, the network performed feature extraction to extract N feature maps from N images. Then, a 2D CNN network was utilized to obtain the depth feature information of the images at each spatial scale. Then, using eight 2D convolutional layers, the features were downsampled in the third and sixth layers, resulting in three different scale features. Thanks to the three different scale features, different levels of feature information can be extracted. Convolutional neural network is adopted to extract features from the image and the local features in the image are represented by the convolutional kernel. Each layer in the feature extraction comes with BN regularization as well as ReLU activation, which is used to improve the fitting ability of the model. Like the traditional matching task, by using the same network, weight sharing of image features can be achieved to enhance the learning. After feature extraction the network outputs N 32-bit channel feature maps that are quadruple reduced in length and width, after downsampling all the pixel information of the image has been encoded into the extracted feature maps, so that no contextually important information is lost when stereo matching is performed, which significantly improves the quality of the reconstruction as compared to stereo matching using the original image.

2.2 Construction of Costumers

After completing the feature extraction, the costal body is obtained by making depth assumptions on the feature maps of N-1 source images and 1 reference image. A planar scanning algorithm is used to obtain the depth of each feature map, due to the different viewpoints of each image, it is necessary to convert one viewpoint to another, the mapping relationship between the reference feature map and the source feature map can be described as a 3D differentiable monoresponsive transform as in Equation (1)

$$H_i^{(d)} = dK_i T_i T_{ref}^{-1} K_{ref}^{-1} \tag{1}$$

Through the microsingle reactivity transform, the viewpoint of the reference image is converted to the stereo space corresponding to the viewpoint of the source feature map, and after N-1 feature map mapping a feature costum is formed C, where K and T denote the camera internal and external parameters, respectively. The cost body is calculated as follows

$$c_i^{(d)} = (f_{srci}^{(d)} - f_{ref})^2 \tag{2}$$

f_{src} denotes the i th source image feature extraction and f_{ref} denotes the reference image feature extraction.

After constructing the volume of the costumers for each view, the next step is to aggregate all the image costumers into one costumers for regularization. It is common practice to perform variance computation on N-1 costomes, with the underlying principle that all views should be equally important. However, in practice this does not make sense because of the presence of varying shooting angles, problems such as occlusion or different lighting conditions causing non-Lambertian surfaces, and the presence of weakly textured regions in some views. Faced with differences between pixels in different views, different weights need to be weighted, and averaging does not work for different pixel weights.

$$C^{(d)} = \frac{\sum_{i=1}^{N-1} (c_i^{(d)} - \overline{c_i^{(d)}})^2}{N-1} \tag{3}$$

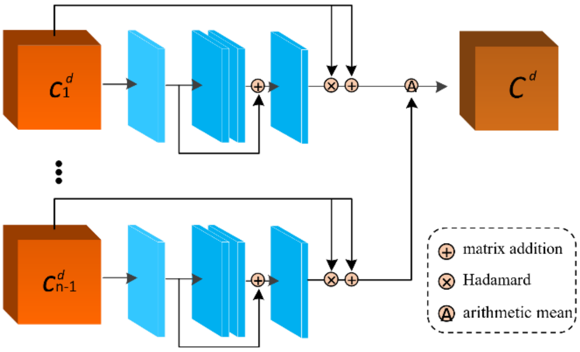

Figure 2: Pixel-by-pixel attention aggregation module

Therefore, in this paper, a pixel-by-pixel attention aggregation module, as shown in Fig. 2, is used to deal with costumers with different viewpoints, which is defined as

$$C^{(d)} = \frac{1}{N-1} \sum_{i=1}^{N-1} [1 + \omega(c_i^{(d)})] \otimes c_i^{(d)} \tag{4}$$

For the input H × W × 32 costumers, the number of 32-bit channels is divided into 4, 4, 4, 1, and after reweighting by the H × W × 1 attention map, all costumers are summed and divided by $N-1$. In Eq. (4) $\otimes$ denotes Hadamard multiplication and ω is a pixel-by-pixel attention map generated adaptively based on the per view cost body in such a way that pixels that may be confusingly matched are suppressed and pixels with critical contextual information are given greater weight and $1+\omega$ prevents over-smoothing better than ω.

2.3 Costumer Regularization

The costum regularization is to generate a probabilistic volume P from the costums obtained above for generating the depth map. In classical networks, the cost body regularization phase uses the U-Net [14] network based on the 3D CNN "encoding-decoding" architecture, due to the excessive computation of 3D CNNs this method causes huge memory and graphics memory consumption, which restricts the improvement of the image resolution, and also reduces the improvement of the accuracy. So in this paper we use recurrent recurrent neural network to adjust the costumers and use hybrid approach of RNN and CNN to avoid loss of accurate pixels. Thus, the depth map can be estimated over a very large depth range using recurrent networks and the memory consumption can be effectively reduced.

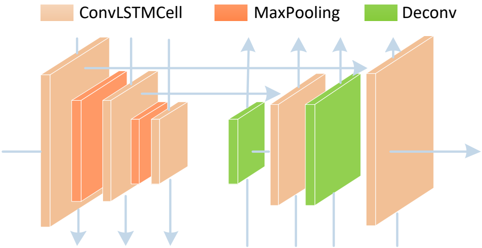

Figure 3: CNN-RNN hybrid network architecture

In this paper, costum regularization is done by a recursive approach using a hybrid CNN-RNN network structure with LSTM module [15]. In the regularized network, feature transfer occurs in both horizontal and vertical directions. In the vertical direction, each 3D costumed body is regularized by a CNN with an encoder-decoder architecture. In the horizontal direction, there are five parallel RNNs transferring the results of the previous recursive convolution to the latter through a recursive convolutional layer (ConvLSTMCell), as shown in Fig. 3. Assuming that the cost body at depth d is processed by the jth convolutional layer denoted as $C_j^{(d)}$, the depth $d-1$ of the output of this layer is C_j^{d-1} and the storage is kept in state $m_j^{(d)}$, ConvLSTMCells operates as follows.

First $C_j^{(d)}$ and C_j^{d-1} are connected together by convolutional layer processing, and the LSTM portion of each ConvLSTMCell is divided into 4 variables as

$$\begin{cases} i = \sigma(w) \\ f = \sigma(x) \\ o = \sigma(y) \\ g = \tanh(z) \end{cases} \tag{5}$$

All signals in the convolutional processing space are two-dimensional and the final output is

$$m_j^d = m_j^{(d-1)} \otimes f + i \otimes g \tag{6}$$

$$C_j^{(d)} = o \otimes \tanh(m_j^{(d)}) \tag{7}$$

2.4 Depth Map Estimation and Optimization

After obtaining the costum C, external conditions such as ambient lighting, lens resolution, camera angle, etc. will negatively affect the image sampling quality, and direct depth prediction will affect the prediction accuracy due to the presence of noise, so preprocessing is needed to remove the noise effect. Since the 2D image features have been transformed into spatial 3D model features with depth, ordinary convolutional networks are not capable of feature processing of depth information, so it is necessary to convert the regularized cost body C into the probability body P by applying the SoftMax classification operation [16], so as to generate the predicted depth maps in the following.

The traversal of the pixel points is performed next, and the final depth estimate will be determined based on the probability distribution of the depth sampling values corresponding to each pixel point. When the probability body is known, the simplest method is to directly use the maximum value in the depth probability map to estimate the depth maps of all pixels in the reference image according to the "winner-takes-all" principle [17]. The network uses the Argmax operation [18] for depth value regression to estimate the depth values of the pixels as follows

$$d_E = \sum_{d=d\min}^{d\max} d \times P(d) \tag{8}$$

In Eq. (8), $P(d)$ is the probabilistic predicted value of the pixel at depth $d\max$ and $d\min$ are the maximum and minimum values of depth sampling, respectively. Compared with directly using the depth value with the largest probability, the evaluation method of Eq. (8) better considers the result predicted by the pixel at all depths, and obtains a smoother and more continuous depth map. At the same time, this summation is derivable, so an end-to-end network model can be constructed. To facilitate the subsequent optimization operation, the depth of the depth image is quantized to take a value between 0 and 1 during the summation and converted back to the normal depth at the end of the optimization.

The depth estimation task is different from the regression task due to the fact that the cost body regularization transforms the matching cost computation into a pixel-by-pixel probability distribution for depth prediction. So the cross-entropy loss function is used to measure the difference between the probability body P and the one-hot coding body G of the true depth map, defined as

$$L = \sum_{\mathrm{p}\in\{\mathrm{pv}\}} \sum_{d=d_0}^{d_{D-1}} -G^{(d)}(\mathrm{p}) \log[P^{(d)}(\mathrm{p})] \tag{9}$$

In Eq. (9), pv denotes the valid set of pixels, $G^{(d)}(\mathrm{p})$ denotes the probability of the true depth map at pixel p, and $P^{(d)}(\mathrm{p})$ denotes the predicted probability of the depth map at pixel p.

3. Experiments and Results

In this paper, we use the DTU dataset for training and testing. The DTU dataset is a classical large-scale dataset that is widely used for multi-view 3D reconstruction. The dataset was collected under well-controlled laboratory conditions with fixed camera trajectories. It contains 49 views covering 128 scans under 7 different lighting conditions. These views are divided into 79 training sets, 18 validation sets, and 22 evaluation sets, and each image has a resolution of 1600 × 1200. in total, there are 27,097 training samples for each image. In addition, the dataset provides the truth values of the dense point cloud, which facilitates the evaluation of the method.

3.1 Network Training

In this paper, the PyTorch framework is used to implement the network model, and the training data comes from the DTU training set, which contains a total of 3791 images. During the training process, the image size is set to 160×128, the number of input images is N=7, and the depth direction is uniformly sampled, so the number of depth layers is D=192.The network is trained using an end-to-end approach using the Adam optimizer, with an initial learning rate of 0.001 for a total of 16 epochs, and the learning rate will be adjusted after each epoch with a 0.9 decay rate. The training process is performed on a server with a Tesla P40 GPU with 24G of video memory and the batch size is set to 2.

3.2 Network Test

The method in this paper has high storage efficiency and can handle higher resolution images and finer depth planar scans. In the testing phase the number of input images is set N=7 and depth layer assumption D=512 to obtain finer depth maps. The height and width of the input image must be a multiple of 8 due to the network parameters. Finally, the input image with 800×600 resolution is used for DTU evaluation. Before testing on other scenarios or datasets, the corresponding structure of the network model needs to be fine-tuned to improve various scenario adaptations. If self-built datasets are used, the parameters of the estimated depth range and the camera need to be requested to adapt the network thus preserving the content information near the image boundaries. The experimental tests were conducted on a GPU GTX 1080 Ti and a CPU of 6-core Intel(R) Xeon(R) CPU E5-2650 v4 2.20GHz with 30G of memory.

3.3 Point Cloud Test Results

This was accomplished using the official evaluation protocol for the DTU dataset [19], under the point cloud model. The reconstruction accuracy (Acuracy.Acc) and completeness (Completeness, Comp), as well as the combination of both, i.e., Overall (OA), were evaluated.

Accuracy (Acc): represents the distance within the visual mask from the MVS reconstruction point to the nearest point of the structured light scanning model (the mask is calculated based on the effective measurement area of the structured light model).

Completeness (Comp): defined as the distance of each point in the structured light scan model to the nearest point of the MVS reconstructed model.

Overall (OA): The average of Accuracy and Completeness, calculated by the following formula

$$Acc = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \| x - y \|^2 \tag{10}$$

$$Comp = \frac{1}{|S_2|} \sum_{x \in S_2} \min_{y \in S_1} \| x - y \|^2 \tag{11}$$

$$OA = \frac{Acc + Comp}{2} \tag{12}$$

S_1 in Eq. (10) denotes the set of all points in the 3D point cloud after reconstruction by the algorithm, and S_2 in Eq. (11) denotes the set of all points in the standard point cloud generated by structured light scanning. Lower values of the above three evaluation indexes indicate better reconstruction quality. Finally, Num represents the number of points in the point cloud model, and in general, the larger the value of points in the same model, the better the reconstruction effect.

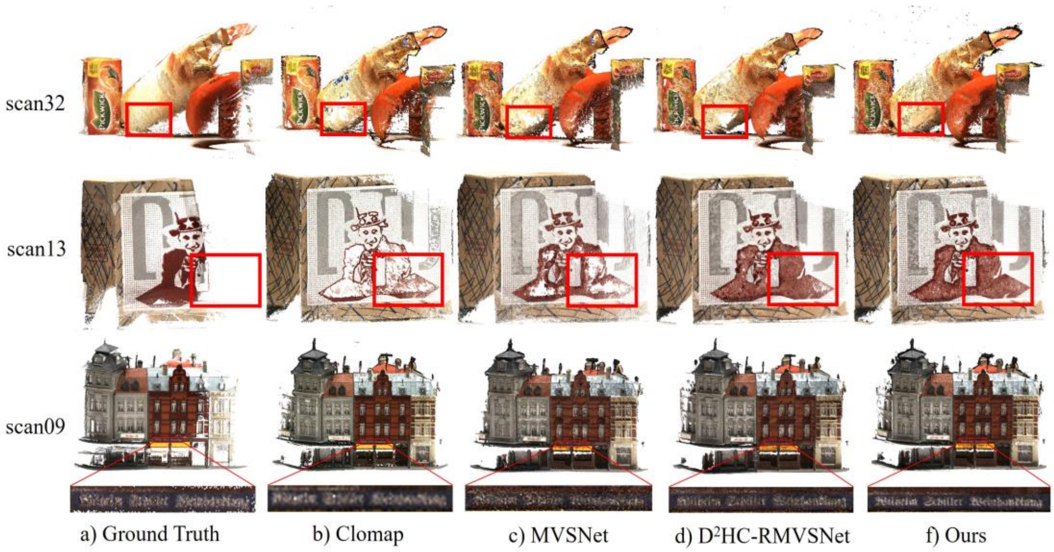

Figure 4: Comparison effect of densely reconstructed point cloud of DTU dataset

To validate the reconstruction results of the proposed method, a comparison is made with the classical traditional method Clomap and the deep learning based methods D2HC-RMVSNet, RMVSNet and MVSNnet, as shown in Figure 4. Where the rectangular box is the detail comparison area. In terms of reconstruction details, the method in this paper is more complete than the other methods after reconstructing the texture-deprived regions, and the scene edges are more sufficiently detailed. In the final scan9, the scene edge details are fuller, and the rectangular boxed area also clearly presents the text, which shows that the network effectively improves the reconstruction effect. At the same time, the characteristics of these methods can be observed through experiments, of which a) is a structured light scanner to generate real point cloud data, and the results produced by this kind of hardware devices can truly reflect the scene information. However, due to environmental factors, it is difficult to reconstruct a non-diffuse reflective scene and the reconstruction area is small. b) The traditional depth-based image method cannot predict the depth information of the texture-deprived region, which leads to the phenomenon of obvious voids in the weak texture region. c) The deep learning-based method can predict some information of the texture-deprived region through learning, but the traditional 3D CNN network cannot predict the context information, and there is still a reconstructed scene. d) The deep learning-based method cannot estimate the texture information of the texture-deprived region through learning, but it cannot use the traditional 3D CNN network to reconstruct the scene. Contextual information, and there is still the phenomenon of reconstructed scene voids. d) The method uses a recurrent network to regularize the costumers, and is able to reconstruct a complete dense point cloud, but it does not consider the pixel visibility problem, and there is still the problem of indistinctness and difficulty in reconstructing the weak texture.

From the comparison effect of DTU dataset, it can be seen that the enhancement of this paper's method comparing with the traditional method, especially in the texture-deprived region and the edge region of the scene, the reconstruction of dense point cloud model is more complete. Then comparing with other network models, this paper's method can obtain more information at the edge of its reconstructed scene, which makes the edge details richer. The method in this paper effectively improves the above phenomenon, thanks to the pixel-by-pixel attention aggregation module and the recursive network's processing of texture-deprived regions and scene edge details. In addition, this paper also calculates the quantitative results of the above objective metrics, accuracy, completeness and wholeness, as shown in Table 1.

Table 1: Quantitative results of DTU dataset

	Acc	Comp	Overall	Num
PMVS[20]	0.613	0.941	0.777	117320
Clomap[2]	0.400	0.664	0.532	1310014
MVSNet[7]	0.396	0.527	0.462	3682198
RMVSNet[8]	0.385	0.459	0.422	5343617
D2HC-RMVSNet[13]	0.395	0.378	0.386	13373332
Ours	**0.377**	**0.363**	**0.370**	13996472

In Table 1, it can be seen that comparing the various methods, although the accuracy of this paper's method is not significantly improved compared to other networks, the quality of completeness and integrity is better than that of the above methods, and has a significant advantage. Compared with other deep learning based methods, this paper's method improves accuracy by 3.7%, completeness by 23.9%, and integrity by 14.4% on average, which proves the effectiveness of this paper's method.

3.4 Ablation Experiment

In order to validate the performance of the network proposed in this paper, ablation experiments are performed to describe the effect of adding certain parts of the network on the results in order to better understand the network behavior. The addition and deletion of the pixel-by-pixel attention aggregation module proposed in this paper is studied on Baseline to qualitatively analyze the effectiveness and costuming of the proposed method, and the results are shown in Table 2. The following ablation experiments are performed on the DTU dataset using the same parameters comparing different network architectures. Baseline generalizes 2D CNN for feature extraction and takes the same hybrid CNN-RNN network for costum regularization without any additional modules.

Table 2: Quantitative results of ablation experiment

Model	Acc	Comp	Overall	Mem(GB)
Baseline	0.408	0.374	0.391	2.42
Ours	0.377	0.363	0.370	2.63
MVSNet	0.396	0.527	0.462	15.4
R-MVSNet	0.385	0.459	0.422	6.7

As shown in Table 2, the Mem values in the table are the average memory consumption for estimating the depth maps of the above scenarios, it can be seen that the advantage of integrating the costal regularized CNN-RNN hybrid network is significantly lower than that of MVSNet and R-MVSNet in terms of memory consumption. and with the addition of the pixel-by-pixel attention aggregation module, the accuracy improves by 0.031 after increasing the memory consumption by only 0.21 GB, and the Overall decreased from 0.391 to 0.370.Both this paper's method and Baseline use recurrent networks, and from Figure 5, it can be seen that the memory consumption is smaller in specific scenarios, which proves that the hybrid CNN-RNN network structure of LSTM reduces the memory consumption. This also indicates the advantage of this paper's method in terms of memory consumption for denser reconstruction work on higher resolution images.

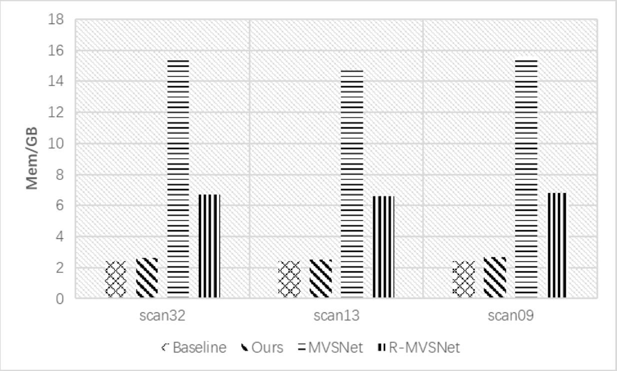

Figure 5: Comparison of memory consumption in each scenario

4. Conclusion

In this paper, we propose a pixel-by-pixel attentional aggregation approach for multiview stereo matching, which an end-to-end is deep learning architecture based on depth maps. Adopting the streaming mode of the classical MVSNet architecture and making improvements in the costum aggregation process, the use of pixel-by-pixel attention aggregation is able to process more details in the scene and effectively improves the reconstruction performance at the edges of the scene and in weakly textured regions. A recurrent network with recursive layering is used instead of the traditional 3D CNN, which significantly reduces the memory consumption and is more efficient. And experiments are conducted on the dataset DTU to compare the quantitative effect with the existing methods, and the final experiments show that the method in this paper significantly improves the reconstruction effect compared with the methods with excellent reconstruction effect in the current stage, which confirms the effectiveness of the method.

Although this paper's method is excellent under the laboratory indoor dataset, the network needs to improve the generalization ability considering the practical operation afterwards. In the future, reconstruction using large outdoor datasets while ensuring the existing accuracy is a key direction for research.

References

[1] Wang Siqi, Zhang Jiaqiang, Li Liyuan, et al. Application of MVSNet in Three-dimensional Reconstruction of Spatial Objects. Chinese Laser, 2022, (23): 176-185.

[2] Schonberger J L, Frahm J M. Structure-from-motion revisited. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4104-4113.

[3] Zhou Min, Zhang Junran, Li Nanxin. Single Image 3D Reconstruction Model Based on Axial Spatial Attention and Intermediate Fusion Representation. . Semiconductor Optoelectronics, 2023, 44(01): 122-127.

[4] Huang P H, Matzen K, Kopf J, et al. Deepmvs: Learning multi-view stereopsis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2821-2830.

[5] Xie Qiqi, Xin Yuelan, Zeng Xi. Multi-view 3D Reconstruction Based on Attention Mechanism. . Laser Journal, 2023, (1): 136-142.

[6] Liu Huijie, Bai Zhengyao, Cheng Wei, et al. Multi-view Stereo Reconstruction with Fusion Attention Mechanism and Multi-layer U-Net. Journal of Image and Graphics, 2022, (2): 475-485.

[7] Yao Y, Luo Z, Li S, et al. Mvsnet: Depth inference for unstructured multi-view stereo. Proceedings of the European conference on computer vision (ECCV). 2018: 767-783.

[8] Yao Y, Luo Z, Li S, et al. Recurrent mvsnet for high-resolution multi-view stereo depth inference. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5525-5534

[9] Chen R, Han S, Xu J, et al. Point-based multi-view stereo network. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1538-1547.

[10] Gu X, Fan Z, Zhu S, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2495-2504.

[11] Yang J, Mao W, Alvarez J M, et al. Cost volume pyramid based depth inference for multi-view stereo. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4877-4886.

[12] Cheng S, Xu Z, Zhu S, et al. Deep stereo using adaptive thin volume representation with uncertainty awareness. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2524-2534

[13] Yan J, Wei Z, Yi H, et al. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. European conference on computer vision. Springer, Cham, 2020: 674-689.

[14] Jiao L, Huo L, Hu C, et al. Refined Une-t: Unet-based refinement network for cloud and shadow precise segmentation. Remote Sensing, 2020, 12(12): 2001.

[15] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: lstm cells and network architectures. Neural computation, 2019, 31(7): 1235-1270.

[16] Wan Lei, Tong Xin, Sheng Mingwei, et al. Overview of Deep Learning Image Classification Methods with Softmax Classifier. . Navigation and Control, 2019, 18(6): 1-9.

[17] Li W, Hu C. Multi-focus Image Fusion and Depth Map Estimation Based on Iterative Region Splitting Techniques. Journal of Imaging, 2019, 5(9): 73.

[18] Liu, X., Li, Y., & Wang, Q. "Multi-View Hierarchical Bidirectional Recurrent Neural Network for Depth Video Sequence Based Action Recognition", International Journal of Pattern Recognition and Artificial Intelligence, (2018), 32(10): 1850033.

[19] Aanæs H, Jensen R R, Vogiatzis G, et al. Large-scale data for multiple-view stereopsis. International Journal of Computer Vision, 2016, 120: 153-168.

[20] Wang A, An N, Zhao Y, et al. 3D Reconstruction of remote sensing image using region growing combining with cmvs-pmvs. International Journal of Multimedia and Ubiquitous Engineering, 2016, 11(8): 29-36.