2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL) | 979-8-3503-2644-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVIDL58838.2023.10167367

PT-MVSNet: Overlapping Attention Multi-view Stereo Network with Transformers

Song Zhang[a,b], Lin Li[a,b,c,d,*], Jiangxuan Qiao[a,b]
[a]School of Computer Science, Qinghai Normal University, Xining, China
[b]The State Key Laboratory of Tibetan Intelligent Information Processing and Application,Xining, China
[c]Academy of Plateau Science and Sustainability, Qinghai Normal University, Xining, China
[d]National Qinghai-Tibet Plateau Scientific Data Center Qinghai Sub-center, Xining, China
* Corresponding author: lilin20081@sohu.com

***Abstract*—In this paper, we propose a new multi-view stereo vision model PT-MVSNet based on multi-view stereo (MVS). Multi-view stereo is a successful reconstruction method that uses multiple images to reconstruct a 3D scene. It has been applied in many practical scenes such as architecture, cultural heritage protection, and map making. MVS still faces a lot of challenges, including inaccurate feature matching, excessive image noise, and overly complex computation. To solve the feature-matching inaccuracy problem, we take the Transformer model as the main structure in the feature-matching and add a patch-based overlap attention module (POLA). In this paper, we proposed PT-MVSNet can solve the image feature extraction problem more effectively. To validate the effectiveness of the model, we conducted experiments on the DTU dataset and evaluated its performance by two evaluation metrics. The experiment results show that our method outperforms the latest methods,whose accuracy and completeness reach 0.386 and 0.271 respectively.**

Keywords-feature extraction;Attention mechanism;Multi-view stereo; Deep learning;3D reconstruction

I. INTRODUCTION

Multi-view stereo vision (MVS) is a technique for 3D reconstruction using images from multiple viewpoints, and it is one of the important research directions in the fields of computer vision, computer graphics，and machine learning. The core problem of MVS is to find the correspondence between the pixels in the reference image and the polar lines in the source image for 3D reconstruction by deriving consistent depth values [5]. This technique has a wide range of applications in virtual reality, artificial intelligence, medical image processing, and other fields.

In the current research[7], the research in the field of 3D reconstruction can be divided into two main categories, traditional multi-view geometry-based algorithms, and deep learning-based algorithms. The recently conducted MVS benchmark tests[1,10] show that the deep learning-based methods perform better in terms of 3D reconstruction quality.

Yao et al. introduced an end-to-end MVS architecture called MVSNet [5] that builds cost volumes from various views, employs 3D CNN to learn cost volume regularization to regress the depth values, and implements reconstructed scenes using depth maps. Gipuma [15] proposed a new propagation operation more suitable for GPUs and extended the application of the PatchMatch algorithm in 3D reconstruction from two-view to multi-view. As an open-source 3D reconstruction framework [4], the proposed COLMAP software enables users to perform 3D reconstruction efficiently.

Recent research has demonstrated that the Transformer model performs well in matching tasks, which is explained by the usage of a positional encoding mechanism and a self-attentive mechanism. However, in practice, applying the self-attentive mechanism to the entire feature map may lead to the following problems: first, this approach consumes a large amount of memory because the feature vector is computed for each position. Second, since the self-attentive mechanism is computed based on global contextual information, it may not accurately capture local feature information, which leads to inaccurate matching.

Based on the above issues, we propose a multi-view stereo vision model with a patch-based overlapping attention mechanism called PT-MVSNet.and introduce a Patch-based OverLapping Attention (POLA) module in Transformer for feature extraction to solve the above problem of high memory consumption and local feature extraction is solved. Similar to other MVS methods that use depth maps for reconstruction [2, 9, 14], our network also takes a reference image and multiple source images as input and infers a depth map of the reference image. Experiments show that our method outperforms known methods on the DTU[1] dataset.

II. PT-MVSNET

Our proposed PT-MVSNet is an end-to-end deep-learning neural network. By anticipating the depth maps aligned with the reference image, it seeks to filter and fuse all depth maps to create a dense point cloud representation. In this section, we will mainly introduce the four parts of the model: feature extraction, cost volume aggregation, cost volume regularization, and loss function. The overall architecture of the model is shown in Figure 1. After obtaining the regularized probability volume, a winner-take-all strategy is used to determine the final prediction.

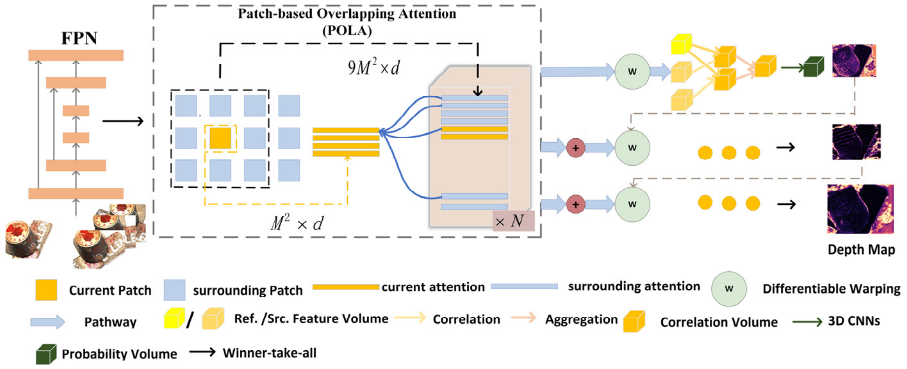

Figure 1. PT-MVSNet architecture

A. *Feature Extraction*

For N input images of size $H \times W$, let I_0 denote the input reference image, and $\{I_j\}_{j=1}^{N-1}$ denotes its neighboring images. First, basic features are extracted using a Feature Pyramid Network (FPN) [6] to deal with target objects at various scales. Next, we apply the Transformer model for feature extraction, which has powerful context awareness and can improve image understanding. Also, to further improve the performance, we have added a POLA module. We first introduce Transformer's attention mechanism, and then explain in detail the working principle and advantages of the POLA module.

1) *Scaled dot-product attention in Transformer*

The three groups of input features are the question Q, the keys K, and the values V [5]. Next, we calculate the dot product of each key vector K and the query vector Q. To make the gradient more stable during backpropagation, the result is divided by a scaling factor $\sqrt{d}$,yielding a collection of attention weights. The attention weight distribution is then calculated using a softmax function, and the attention weights are then applied to the value V. Get the final output. The formula is as in (1):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

2) *Multi-Head attention in Transformer*

Multi-headed attention [8,13] computes attention from the dot product of Q and K. The multi-headed attention mechanism is also applied in our attention module, where we compute the corresponding attention weights for each set of Q and K. Specifically, given a patch vectorized to $S \in R^{m^2 \times d}$ and its surrounding patch vectorized to $T \in R^{9M^2 \times d}$. To obtain the output for the i-th attention mechanism, first do a linear projection of S and T into d_k dimensions and represent the projection results as S_i and T_i. Next, carry out the dot product attention operation on the output. Finally, the h_i connection is merged into H and H is projected into d d dimensions to obtain the final result $U \in R^{m^2 \times d}$, The formula is expressed as follows: (2), (3), (4):

$$h_i = Attention\left(L_i^Q(S),\ L_i^K(T),\ L_i^V(T)\right). \tag{2}$$

$$H \ = Concat([h_1,\ h_2,\ ...,\ h_n]). \tag{3}$$

$$U \ = L^O(H). \tag{4}$$

where n is the number of attention heads, L_i^Q , L_i^K , and L_i^V are linear projection functions, and in the experiment set n=8, $d_k = d/8$.

3) *Patch-based OverLapping Attention*

The POLA module can divide features into $N \times N$ non-overlapping patches and extract features by associating each patch with itself and its neighboring 8 patches. These patches can then be integrated to form a feature map of the entire image, allowing the model to learn a more robust image representation. Patch-based overlapping attention can preserve spatial information better than the global attention mechanism since each patch only covers a small portion of the image.And there is only one intra-block information dissemination during the entire procedure.

The addition of POLA has several benefits. Firstly, it improves the visual perception of the model, we are calculating the attention for each small block separately. This allows the model to better focus on the details and local structures in the image. Secondly, the interpretability of the model can be improved, and by segmenting and processing the features, the model can be made to focus on the important regions in the image.Finally, it is also possible to reduce the number of model parameters, after POLA processing we can share the attention calculation for each chunk, thus avoiding the additional parameters needed to calculate the attention weights for each chunk separately.Figure 2 shows the schematic diagram for n=3.

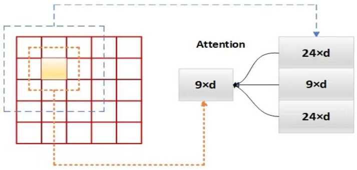

Figure 2. Local attention diagram

B. *Cost Volume*

The next work is mainly to construct a 3D cost volume based on the previously obtained feature map and camera parameters. This cost body mainly includes three parts: homography matrix transformation, Cost Metric, and Cost Volume Regularization.

1) Differentiable Homography

In the feature extraction stage, we obtain the feature maps of the reference image as well as the sourc e image. we need to convert the viewpoint of the source image to the viewpoint of the reference image by homography matrix transformation [8,13].In the equation, we use R to denote the rotation between the two viewpoints and t to denote the translation. K_0 denotes the pixel matrix in the reference camera and K_1 denotes the pixel matrix in the source camera. Under the depth assumption d, the reference view pixel H is transformed with the source view pixel H_1, The conversion relationship is as in (5):

$$H_1 = K\left[R\left(K_0^{-1}Hd\right)+t\right] \tag{5}$$

2) Cost Metric

Next, the cost body is obtained by the above single-response transformation. Referring to TransMVSNet [8], we first discretize the known depth space into D depth values and obtain the pairwise feature correlation at position p as follows(6):

$$c_i^{(d)}(\boldsymbol{p}) =< L_0(\boldsymbol{P}), \hat{L}_i^{(d)}(\boldsymbol{p}) > \tag{6}$$

$\hat{L}_i^{(d)}$ denotes the value of the mapping corresponding to the i-th feature at depth d. Through this operation, we reduce the number of channels to 1. The purpose is to construct the cost body of $H'\times W'\times D'\times 1$ for Cost Volume Regularization. where H' is the height. W' is the width, D' is the depth assumption, and 1 indicates the number of views. To aggregate cost volumes, we use the following mechanism: cost volume pixels have the same features in the depth direction, Therefore, we use pixel-level weight maps to perform the operation. The volume associated with the token is defined as follows(7):

$$C^{(d)}(\boldsymbol{p}) = \sum_{i=1}^{N-1} \max_d \left\{c_i^{(d)}(\boldsymbol{p})\right\} \cdot c_i^{(d)}(\boldsymbol{p}). \tag{7}$$

3) Cost Volume Regularization

The covariance regularization is an effective method to reduce the noise and error of the volume and improve the reconstruction quality. Inspired by the regularization method of the MVSNet network [5], we also apply to regularize in the regularization step. Finally, the softmax function is applied to perform probability normalization.

4) LOSS Function

Unlike MVSNet[5],we use depth estimation as a classification task and therefore use a focal loss function for computation [16]. The focal loss for each depth estimation stage is expressed as Equation (8):

$$L = \sum_{p\in\{p_v\}} -(1-p^{(\tilde{d})}(\boldsymbol{p}))^{\alpha} \log\left(p^{(\tilde{d})}(\boldsymbol{p})\right) \tag{8}$$

where $p^{(d)}(\boldsymbol{p})$ denotes the predicted value of the depth hypothesis d at pixel p, p_v denotes the amount by which the current depth value is closest to the true depth value, and q denotes the union with valid pixel values. Compared with the traditional cross-entropy loss function, the focal loss function can make the network focus more on the pixels that are difficult to classify, thus improving the accuracy of depth estimation.

III. EXPERIMENT AND RESULT ANALYSIS

A. *Experimental dataset and environment*

To validate the valid values of our model, we will train on the DTU [1]dataset, which [1] is a large-scale MVS dataset that was acquired with a camera under strict laboratory conditions. It contains 128 scans and 49 views at 7 different illumination intensities, with a total of 27097 training samples. The image resolution was 1600×1200 .we divided the dataset into 79 training scans, 18 validation scans, and 22 evaluation scans. For the DTU [1]dataset training, we used the same training approach as other learning-based methods [3,5,11,12].

All experiments in this paper were conducted under Windows and the Pytorch deep learning framework. The computer hardware configuration is a Tesla PG503 graphics card and 64 GB RAM.

B. *Implementation Details*

In the training phase, we set the number of input images to N = 5 and the image resolution to 512×640. For coarse to fine regularization, the depth is assumed to be sampled in the range of 425 mm to 935 mm. The number of planar sweep depths assumed for each stage was 48, 32, and 8. The corresponding depth intervals decayed from the coarsest stage to the finest stage by 0.25 and 0.5, respectively. the model was trained for 16 rounds.

C. *Experimental Analysis*

1) Experimental Performance

We evaluate the proposed PT-MVSNet model using official evaluation metrics on the evaluation set of the DTU [1]dataset. The evaluation phase was set to N = 5 and the input resolution was 864×1152. The quantitative comparisons are shown in Table 1. Accuracy and Completeness are the dataset's two official measures; the lower their values, the better. where bolded indicates the best result and underlined indicates the second best result.Overall measures the model's overall performance is calculated as the average of accuracy and completeness.As can be shown, our model performs better than other methods.

Table 1: Quantitative results on the DTU [1]dataset

Methods	Acc.(mm)	Comp.(mm)	Overall(mm)
Gipuma[15]	**0.283**	0.873	0.578
COLMAP[4]	0.400	0.664	0.532
R-MVSNet[3]	0.385	0.459	0.422
AA-RMVSNet[11]	0.376	0.339	0.357
Vis-MVSNet[12]	0.369	0.361	0.365
CasMVSNet[9]	0.325	0.385	0.355
UCS-Net[14]	0.338	0.349	0.344
PatchmatchNet [2]	0.427	0.277	0.352
MVSNet[5]	0.396	0.527	0.462
Ours	0.386	**0.271**	**0.329**

2) Reconstruction results

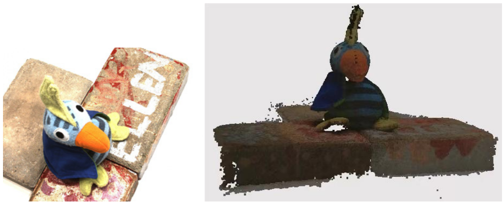

Figure 3. Reconstruction results on the DTU [1]dataset evaluation set

Figure 3 shows the results of our model reconstruction on the DTU[1] dataset evaluation set. The selected dataset is the scan4 dataset in DTU[1]. The left side is one of the original images given by the dataset, and the right side is the result of our reconstruction. It can be seen that our reconstructed effect is more realistic and clear.

IV. Conclusions

In this study, we propose PT-MVSNet, a novel MVS-based end-to-end neural network. PT-MVSNet integrates the overlapping attention mechanism with the Transformer to extract local information during the 3D reconstruction feature extraction. This novel feature extraction method is capable of increasing the algorithm's accuracy and efficiency. And our loss model also employs a focal loss function to significantly increase the precision of depth estimation.

Acknowledgment

Research and Application of Key Technology for 3D Modeling of Ecological Environment Based on Semantic-Driven Augmented Reality Technology (21104), State Key Laboratory of Mapping and Remote Sensing Information Engineering, Wuhan University

Research on Image Semantic 3D Reconstruction Based on Deep Learning (2022-SKL-103), an independent project of the State Key Laboratory of Tibetan Intelligent Information Processing and Application jointly established by the province and the Ministry

Research on the Cooperative Guarantee Mechanism of Qinghai-Tibetan Scientific Research Integrated Service, Science, and Technology Program of the Ministry of Science and Technology of the People's Republic of China

Qinghai Province Science and Technology Assistance to Qinghai Cooperation Special Project (2022-QY-203).

References

[1] Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., & Dahl, A. B. (2016). Large-scale data for multiple-view stereopsis. International Journal of Computer Vision, 120, 153-168.

[2] Wang, F., Galliani, S., Vogel, C., Speciale, P., & Pollefeys, M. (2021). Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14194-14203).

[3] Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent mvsnet for high-resolution multi-view stereo depth inference. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5525-5534).

[4] Schönberger, J. L., Zheng, E., Frahm, J. M., & Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14 (pp. 501-518). Springer International Publishing.

[5] Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European conference on computer vision (ECCV) (pp. 767-783).

[6] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[7] Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12179-12188).

[8] Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., & Liu, X. (2022). Transmvsnet: Global context-aware multi-view stereo network with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8585-8594).

[9] Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2495-2504).

[10] Knapitsch, A., Park, J., Zhou, Q. Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG), 36(4), 1-13.

[11] Wei, Z., Zhu, Q., Min, C., Chen, Y., & Wang, G. (2021). Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6187-6196).

[12] Zhang, J., Yao, Y., Li, S., Luo, Z., & Fang, T. (2020). Visibility-aware multi-view stereo network. arXiv preprint arXiv:2008.07928.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[14] Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L. E., Ramamoorthi, R., & Su, H. (2020). Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2524-2534).

[15] Galliani, S., Lasinger, K., & Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision (pp. 873-881).

[16] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).