MULTI-VIEW THREE-DIMENSIONAL RECONSTRUCTION BASED ON A TWO-STAGE MULTI-LEVEL DEPTH NETWORK FOR AGRICULTURE APPLICATIONS

Li Guo, Yinyin Shi, Dinfei Jin, Mingjun Deng and Xu Zhang

School of Automation and Electronic Information, Xiangtan University,
Xiangtan, 411105, China

ABSTRACT

To address the problems appearing in multi-view three-dimensional (3D) reconstruction, such as the improvement of the accuracy and completeness of the 3D reconstructed images, a two-stage multi-level depth network is proposed. In the stage 1 of the proposed network, several convolutional block attention modules (CBAMs) are applied in the lateral connections of the feature pyramid network (FPN). This is targeted to enhance the spatial and channel relativity of the different hierarchical feature maps so as to bring more semantic information. In the stage 2, the obtained multi-scale feature maps in the stage 1 are tackled by a set of cascaded processing procedures, such as adaptive propagation, single-trees transform, and matching cost computation. As a result, a depth map could be generated and then be further refined in the processing. Comparing with other state-of-the-art methods, the subjective and objective experiments based on the DTU dataset show that our method performs better result in completeness meanwhile maintaining a considerable overall metric. The investigation of applying the proposed method for reconstructing agricultural crop images was carried out, which is based on a set of self-collected images. The experiment shows that a suitable human visual perception for the images could be obtained.

KEYWORDS

Multi-view Stereo, Three-dimensional Reconstruction, Deep Learning, Attention Mechanism, Intelligent Agriculture

1. INTRODUCTION

Multi-view stereo (MVS) 3D reconstruction is widely applied in many fields, such as unmanned driving, virtual reality, and intelligent robotics. In agriculture applications, in some cases the crop images need to be reconstructed in 3D visual with high accuracy or other properties for precision operation [1] The employment of traditional MVS methods, for example, the methods of voxel-based, point-cloud based [2], grid based, and depth-map based [3], may obtain a basic image quality and perform good results in accuracy. However, there are still some problems when facing some special conditions, such as sheltering and lighting change, and in some special regions of image, such as weak textured region and non-Lambert surface. These problems include lack of robustness, low completeness, poor human visual perception, etc.[4]. In recent years, the multi-view stereo 3D reconstruction methods based on deep learning have been widely researched [5][6], which are mainly aimed at improving the reconstruction accuracy. However,

David C. Wyld et al. (Eds): ICAITA,CDKP, CBW, SOFT, CMC, Signal, MLT-2023
pp. 129-137, 2023. CS & IT - CSCP 2023 DOI:10.5121/csit.2023.130610

they are still worthy of further improved in the aspect of memory consumption, computing complexity, reconstruction completeness, and reconstruction efficiency [7].

Multi-view stereo 3D reconstruction methods based on deep learning can be usually divided into two research paths including unsupervised [8][9] and supervised learning[10][11]. These methods are usually to extract the feature information of images, and then to calculate the depth maps and to fuse the depth maps so as to reconstruct the 3D point cloud. According to this strategy, many methods and networks have been created. In [12] the network—CasMVSNet adopts FPN to extract multi-scale features.Some methods are developed by introducing modules into depth networks or by improving existing modules. For example, in [13] MVSTR was built by using Transformer architecture to extract the dense features with global context and 3D consistency. In [7] UniMVSNet was constructed by improving the loss functions.Some other MVS improved networks include CVP-MVSNet [14] and MVSNet++ [15]. To improve the 3D reconstruction performance of MVS, some depth networks based on multi-stage deep learning have been proposed, such as MV-GwCNet [9], CasMVSNet [12], UCS-Net [16], PatchmatchNet [4], UniMVSNet [7], and TransMVSNet [17].In [4] the proposed PatchmatchNet network is constructed by introducing the idea of Patchmatch algorithm into the end-to-end deep-learning MVS network. This could bring reduced memory consumption and computing time. However, it is difficult to capture the semantic information in the context by using the limited perceptual-fields CNN and FPN networks. This would result in the local ambiguity in weak textured or non-textured regions so that the feature extraction would be inaccurate. Attention modules have been applied in many depth networks, which could be used to enhance the channel and spatial correlation related to the feature map information. And the feature information could be refined and hierarchal so as to improve the accuracy or other metrics of the feature extraction [18].

In this paper, a two-stage multi-level depth network based on the PatchmatchNet network was proposed, which introduces CBAM modules in the lateral connections of the FPN. The FPN could fuse different hierarchical information in the bottom-up pathway by applying several paralleled attention modules. Furthermore, a multi-level processing was applied. All above processing contribute to multi-scale feature extraction and fusion. The experiments were carried out based on the DTU dataset for validation. The subjective and objective evaluations for the accuracy and the completeness were conducted. Additionally, the computational complexity and the memory consumption were evaluated and contrasted. The network performs good 3D reconstruction completeness without significantly increasing the computational complexity. To investigate the network possible applications in agriculture, the experiment based on selfcollected agricultural crop images had been implemented. This paper is concluded in the end.

2. METHOD

2.1. Network Architecture

The proposed depth network is constructed based on a two-stage multi-level architecture that predicts the depth maps in a coarse-to-fine evolution manner. Its whole network architecture is depicted in Figure 1. In the stage 1, it mainly focuses on the generation of different hierarchical feature maps using the attention-embedded FPN. In the stage 2, it mainly focuses on the depth map prediction by a multi-level processing. In each level, the output depth map is obtained through five critical procedures which are feature extraction, cost volume construction, matching cost computation, cost aggregation, and depth map regression, respectively.

In the stage 1, the inputs are N images which contain 1 reference image and N-1 source images. By embedding convolutional block attention modules (CBAMs) in the differenthierarchical lateral connections of the FPN, the feature maps containing shallow and deep semantic information can be generated. The fusion of different-hierarchical feature maps obtained from the top-down pathway of the FPN and correspondingly from the CBAMs in the lateral connections is carried out. As a result, three different-resolution feature maps are obtained. Their sizes are 1/2, 1/4, and 1/8 scaled in size of the input image. In the stage 2, the different hierarchical feature maps obtained in the stage 1 are further processed by a set of cascaded processing to predict the different-level depth maps. Note that an operation of random local perturbation is introduced in the training to improve the model robustness. And the adaptive propagation is realized resorting to a deformable convolutional network. Moreover, the matching cost for each hypothesis is calculated by using the group correlation to construct the cost volume. The robustness is augmented resorting to the adaptive spatial cost aggregation. The depth regression is realized by leveraging a softmax function. In the output part of stage 2, the depth map generated in the level 1 is fused with the reference image, and then they are refined by a depth residual network.

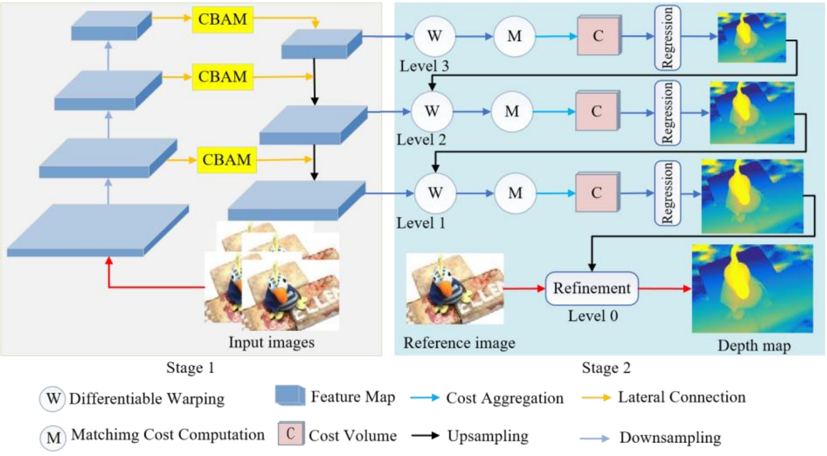

Figure 1. Architecture of the proposed depth network.

The loss function for the proposed network model is the addition of all the depth estimation loss and the rendered ground-truth loss [4]. The reason for embedding CBAM in the lateral connections of the FPN is mainly because that in usual FPN it is difficult to distinguish the foreground and background in different-resolution feature maps. The CBAMs could enhance the correlation of channel and spatial information of the feature maps and mines the interdependencies of the different-scale features so as to extract finer features.

2.2. Convolutional Block Attention Module (CBAM)

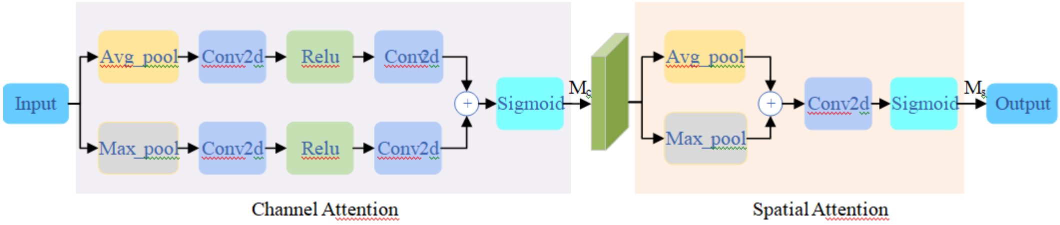

Figure 2. Structure of the convolutional block attention module (CBAM).

The structure of the applied convolutional block attention module (CBAM) is illustrated in Figure 2 **Error! Reference source not found.**. The channel and spatial attention block are sequentially connected. Given that the input feature map is expressed as $\mathrm{F} \in R^{H \times W \times C}$. The channel attention block could produce a one-dimensional channel attention feature map expressed as $\mathrm{M}_C \in R^{1 \times 1 \times C}$. The spatial attention block could produce a two-dimensional spatial attention feature map expressed as $\mathrm{M}_\mathrm{S} \in R^{H \times W \times 1}$ **Error! Reference source not found.**. The whole processing procedure of the CBAM module can be expressed as equations (1) and (2).

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_\mathrm{S}\left(F'\right) \otimes F' \tag{2}$$

where $\otimes$ denotes element-wise multiplication. F'' is the final refined output.

3. EXPERIMENTS

3.1. Experimental Settings

The experiments were implemented mainly based on the DTU dataset [19] and a set of self-collected agricultural crop images to investigate and evaluate the proposed network. The DTU dataset was divided into a training set, a validation set, and an evaluation set.The network was built based on the Pytorch computing architecture. The network was trained on the DTU training set and evaluated on the DTU evaluation set without any fine-tuning. And then the network was tested based on the self-collected agricultural crop images for agricultural research.The selection and preprocessing for the input images are consistent with the method presented in [4].For the training on the DTU dataset, the number of input images is set to N=5 and the resolution of the input image is set to 640×512. The iteration number in the phase 3, 2, and 1 of the Patchmatch algorithm is set to 2, 2, and 1, respectively. The model is trained with 9 epochs using the optimizer Adam. The initial learning rate is set to 0.01.The batch size is set to 2. The graphic processing unit (GPU) used for the training is NVIDIA RTX3080. For the evaluating on the DTU dataset, the resolution of the input image is set to 1600×1200.When the depth maps undergo filtering, the threshold values for the photometric consistency filtering and the geometric consistency filtering are set to 1 and 0.8, respectively.

3.2. Experiments and Analysis based on the DTU Dataset

The proposed method was evaluated in terms of three assessment metrics, namely accuracy, completeness, and overall.The quantitative parameters of fifteen methods of multi-view stereo vision 3D reconstruction were calculated and contrasted. The results are listed and contrasted in Table 1. The bold indicate the best value. It can be seen that the method—Gipuma performs the best performance in accuracy. TransMVSNet performs the best value in overall. Our method significantly outperforms other methods in completeness.

Table 1. Quantitative metrics contrast of different methods based on the DTU evaluation set.

Methods	Acc.(mm)	Comp.(mm)	Overall(mm)
Furu[2]	0.613	0.941	0.777
Gipuma[3]	**0.283**	0.873	0.578
SurfaceNet[6]	0.450	1.040	0.745
MVSNet[5]	0.396	0.527	0.462
R-MVSNet[10]	0.383	0.452	0.417
UCS-Net[16]	0.338	0.349	0.344
Point-MVSNet[11]	0.342	0.411	0.376
CVP-MVSNet[14][14]	0.296	0.406	0.351
CasMVSNet[12]	0.325	0.385	0.355
MVSNet++[15]	0.407	0.345	0.376
UniMVSNet[7]	0.352	0.278	0.315
TransMVSNet[17]	0.321	0.289	**0.305**
M^{3}VSNet[8]	0.881	1.073	0.977
MV-GwCNet[9]	0.383	0.415	0.399
PatchmatchNet[4]	0.427	0.277	0.352
Ours	0.451	**0.273**	0.362

The subjective evaluating results are depicted in Figures 3 and 4. In Figure 3, the depth maps and the point cloud maps for the typical PatchmatchNet network and our proposed network, the ground-truth (GT) depth map, and the input image are provided. It can be seen that the depth map obtained by our method has less noise and looks smoother. This is primarily due to the introduction of the attention mechanism. In Figure 4, the images processed by our network, R-MVSNet, and PatchmatchNet are provided and contrasted. Meanwhile the standard GT is given. For the first row of images (Scan9), it can be seen that our proposed method performs better human visual perception and obvious features in the region framed by the red line. For the second row of images (Scan11), the dashboard scale is clearer for our method as seen from the amplifying region. For the third row of images (Scan33), both the dog's earlobe profile and the detail features in the middle region are more abundant for our method.

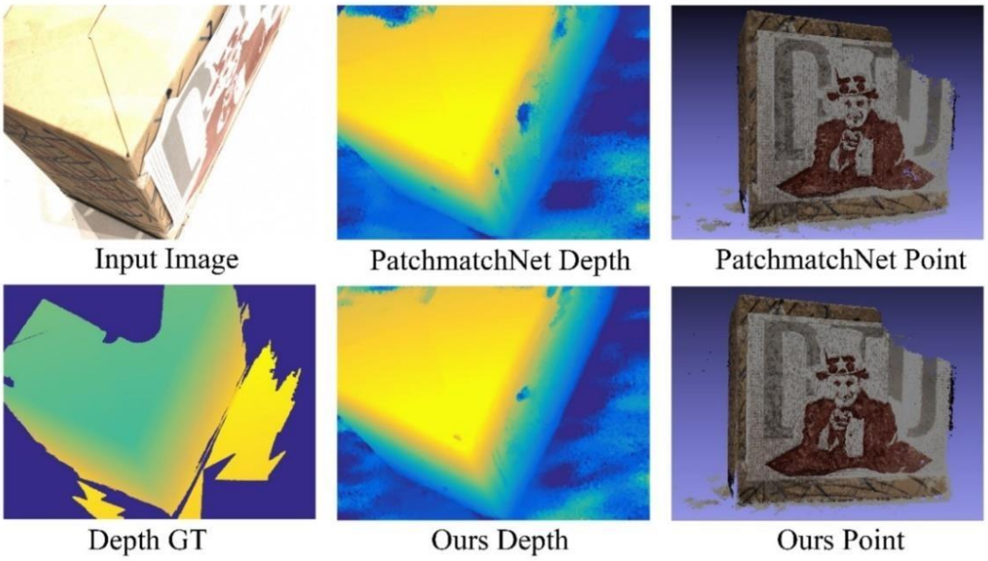

Figure 3. Depth maps and 3D point cloud reconstruction of the image—Scan 13 in the DTU dataset.

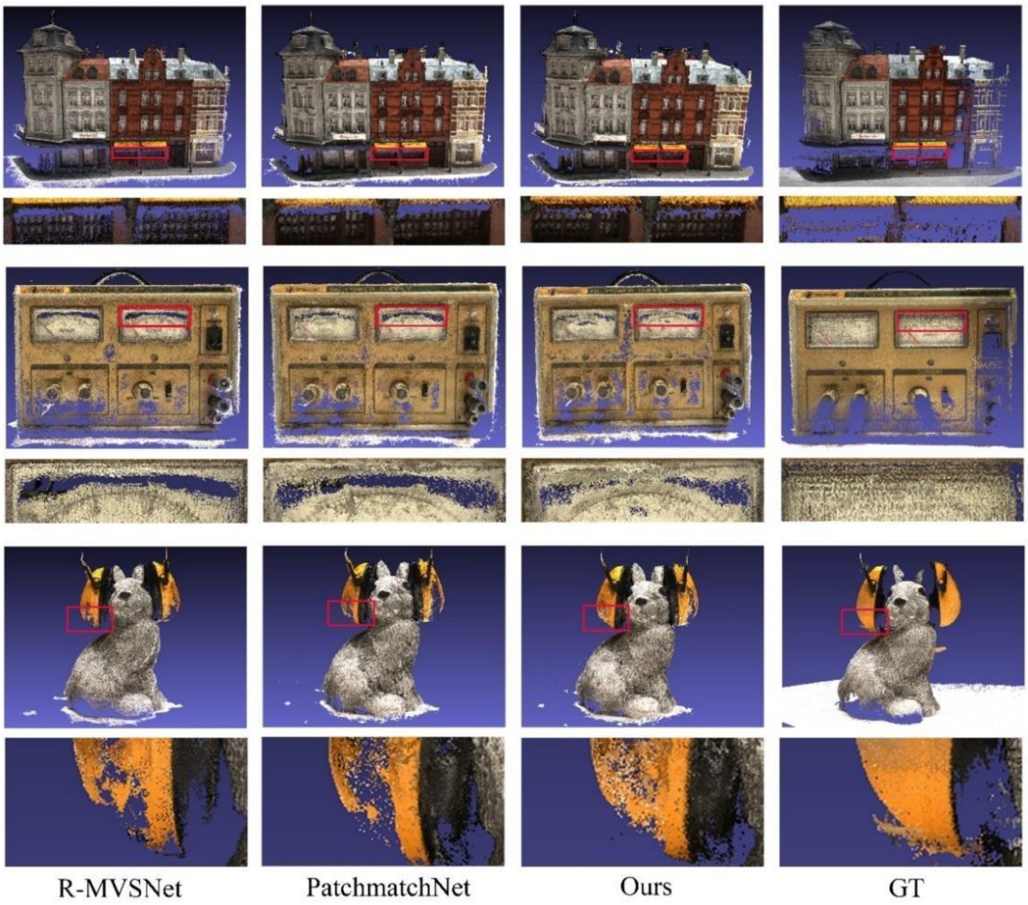

Figure 4. 3D reconstruction of the images—Scan9 (the first row), Scan11 (the second row), and Scan33 (the third row) in the DTU dataset by using different methods.

3.3. Computational Complexity and Memory Consumption

To investigate the computational complexity and the memory consumption for the proposed method, the experiments were carried out on the DTU evaluation set. The resolution of the input images is set to 864×1152, considering the memory limitation. The results are listed in Table 2. It can be seen that the testing time for our method can be reduced by 74%, 64%, and 51% compared to the methods of TransMVSNet, CasMVSNet, and MVSNet, respectively. In comparison to PatchmatchNet, the memory consumption of our method is almost equal, and there is a slight reduction in testing time.

Table 2. Testing time and memory consumption of different methods based on the DTU evaluation set.

Methods	Testing Time (s)	GPU memory (GB)
MVSNet[5]	0.342	7.7
CasMVSNet[12]	0.468	4.7
UniMVSNet[7]	0.330	6.2
TransMVSNet[17]	0.662	4.4
PatchmatchNet[4]	0.169	6.8
Ours	0.167	6.8

3.4. Experiments and Analysis based on a Set of Self-Collected Agricultural Images

To investigate the proposed network applied in agriculture, the network model was trained on the DTU dataset without any fine-tuning and then it was tested on a set of self-collected agricultural crop images. The images of three types agricultural crops are collected, which are of pakchoi, garlic, and romaine lettuce plants, respectively. They were obtained in the laboratory with 49 views for each crop under the same lighting conditions and different perspectives. The obtained original images are shown in the left column of Figure 5. The device used for taking photographs for these crops is Canon EOS 80D. The resolution of the acquired images is 6000×4000. To obtain the camera parameters, the method of COLMAP [20] was applied for the sparse reconstruction of the crop images, which would calculate the intrinsic matrix and the extrinsic matrix of the camera and the sparse point cloud information. The original images and the output information from the COLMAP are inputted into the proposed network to reconstruct the dense point cloud. The experimental results are illustrated in Figure 6. It can be seen that the point clouds reconstructed by the proposed network has fewer false matches in comparison to PatchmatchNet. And the reconstructed point clouds are more complete and perform better human visual perception.

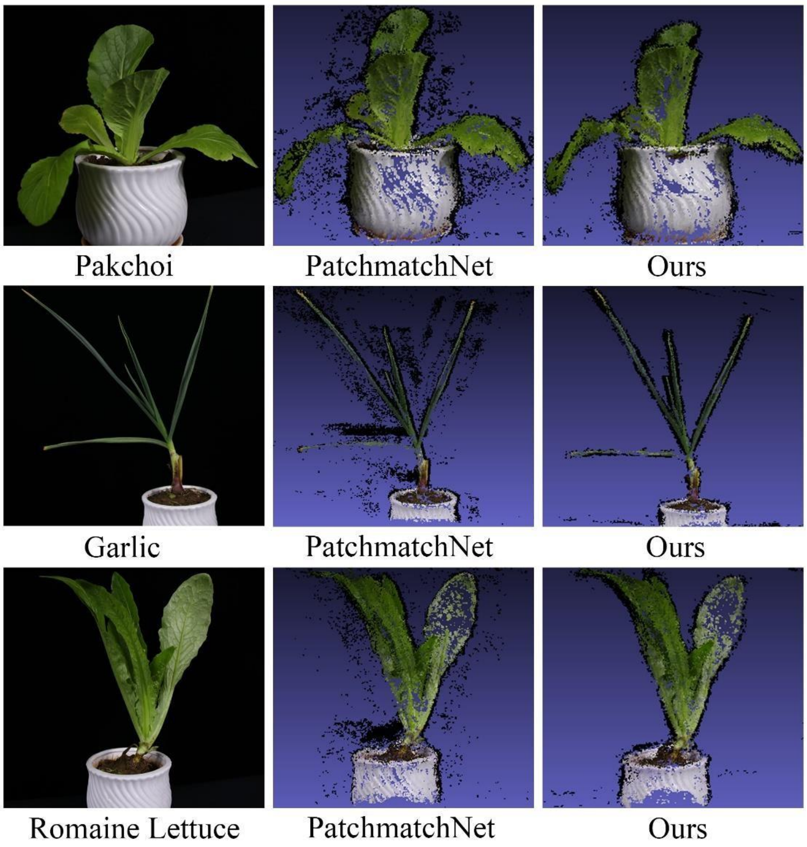

Figure 5. 3D point cloud reconstruction of PatchmatchNet and our network based on the selfcollected agricultural crop images.

4. Conclusion

In this paper, a two-stage multi-level 3D reconstruction depth network using an attention mechanism is presented. The network is constructed based on the PatchmatchNet network. Through the employment of the CBAMs in the stage 1, the spatial and channel correlation in different hierarchical feature map in the FPN could be enhanced. And more abundant texture information could be generated by the following multi-level processing in the stage 2. The objective experiments show that our method performs better 3D reconstruction effect in completeness and lower computational complexity. The subjective experiments show that our method performs good human visual perception. Moreover, the proposed network could be a candidate for agricultural crop image processing.

Acknowledgements

This research was supported in part by the Starting Research Fund Project of Xiangtan University under Grant 19QDZ16.

References

[1] Sampaio G. S., Silva L. A., & Marengoni M., (2021) "3D reconstruction of non-rigid plants and sensor data fusion for agriculture phenotyping", *Sensors,* Vol. 21, No.12, pp 4115.

[2] Furukawa Y & Ponce J,(2010) "Accurate, dense, and robust multiview stereopsis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 8, pp 1362-1376.

[3] Galliani S. Lasinger K., & Schindler K. (2015) "Massively parallel multiview stereopsis by surface normal diffusion". *InProceedings of the IEEE International Conference on Computer Vision,* Santiago, Chile. pp 873-881.

[4] Wang F., Galliani S., Vogel C., *et al*, (2021) "PatchmatchNet: Learned multi-view patchmatch stereo". *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Virtual, Online, United States, pp 14189-14198.

[5] Yao Y., Luo Z., Li S., *et al*, (2018) "MVSNET: Depth inference for unstructured multi-view stereo". *15th European Conference on Computer Vision*, Munich, Germany, Berlin, pp.785-801.

[6] Ji M., Gall J., Zheng H., *et al*, (2017) "SurfaceNet: An end-to-end 3d neural network for multiview stereopsis". *In Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp 2326-2334.

[7] Peng R., Wang R., Wang Z., *et al*, (2022) "Rethinking depth estimation for multi-view stereo: A unified representation". *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, United states, pp 8635-8644.

[8] Huang B., Yi H., Huang C., *et al*, (2021) "M3VSNET: Unsupervised multi-metric multi-view stereo network". *In Proceedings - International Conference on Image Processing*, Anchorage, AK, United states, pp 3163-3167.

[9] Qi S., Sang X., Yan B., *et al*, (2022) "Unsupervised multi-view stereo network based on multi-stage depth estimation", *Image and Vision Computing*, Vol. 122, No.104449.

[10] Yao Y., Luo Z., Li S., *et al*, (2019) "Recurrent MVSNet for high-resolution multi-view stereo depth inference" *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, United States, pp 5520-5529.

[11] Chen R., Han S., Xu J., *et al*, (2019) "Point-based multi-view stereo network" *In Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, Republic of, pp 1538-1547.

[12] Gu X., Fan Z., Zhu S., *et al*, (2020) "Cascade cost volume for high-resolution multi-view stereo and stereo matching" *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Virtual, Online, United states, pp 2492-2501.

[13] Zhu J., Peng B., Li W., *et al*, (2021) "Multi-view stereo with transformer". *ArXiv*.

[14] Yang J., Mao W., Alvarez J. M., *et al*, (2022) "Cost volume pyramid based depth inference for multiview stereo", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No.9, pp 4748-4760.

[15] Chen P. –H., Yang H. –C., Chen K. –W., *et al*, (2020) "MVSNet++: Learning depth-based attention pyramid features for multi-view stereo", *IEEE Transactions on Image Processing*, Vol. 29, pp 72617273.

[16] Cheng S., Xu Z., Zhu S., et al, (2020) "Deep stereo using adaptive thin volume representation with uncertainty awareness". *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Virtual, Online, United states, pp 2521-2531.

[17] Ding Y., Yuan W., Zhu Q., *et al*, (2022) "Transmvsnet: Global context-aware multi-view stereo network with transformers". *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, United states, pp 8585-8594.

[18] Woo S., Park J., Lee J. Y., *et al*, (2018) "CBAM: Convolutional block attention module". *15th European Conference on Computer Vision*, Munich, Germany, pp 3-19.

[19] Aanas H., Jensen R. R. , Vogiatzis G., *et al*, (2016) "Large-scale data for multiple-view stereopsis" , *International Journal of Computer Vision*, Vol. 120, No. 2, pp 153–168.

[20] Schonberger J. L. & Frahm J. M., (2016) "Structure-from-motion revisited". *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Virtual, Online, United states, pp 8949-8958.