Global Contextual Complementary Network for Multi-View Stereo

Yongrong Cao[1]
yongrongc3@gmail.com
Suping Wu[1]*
pswuu@nxu.edu.cn
Xing Zheng[1]
zxing269@gmail.com
Bin Wang[1]
mrbin2021@163.com
Pan Li[1]
18235448104@163.com
Zhixiang Yuan[1]
yzxnxu@163.com
Lei Lin[1]
2409937088@qq.com
Yuxin Peng[1]
pyx_9999@163.com

[1]School of Information Engineering
Ningxia University
Yinchuan, China

Abstract

Multi-View Stereo (MVS) has always been a challenging problem. Existing reconstruction methods mostly rely on convolutional neural networks, which limits the ability of the network to capture the global context of images, resulting in a lack of a certain complete representation of the final depth map. In this paper, we propose a Global Context Complementary Network (GCCN), which aims to enhance the complete representation of depth maps with a global context complementary learning strategy. Specifically, for the feature maps, we first exploit the advantages of convolution neural network (CNN) and self-attention to extract 2D local features and long-term dependence information, respectively. Thus, GCCN achieves maximizing the preservation of the complementary information. Furthermore, in the 3D cost volume regression stage, in order to obtain richer 3D depth information, we design a Contextual-feature Complementary Learning Module (CCLM), which utilizes global feature interaction in the cost volume to achieve complementary learning of cost volumes at different scales. We conduct experiments on the DTU benchmark dataset and the Tanks and Temples dataset. The results show that our approach achieves significant performance compared to state-of-the-art methods.

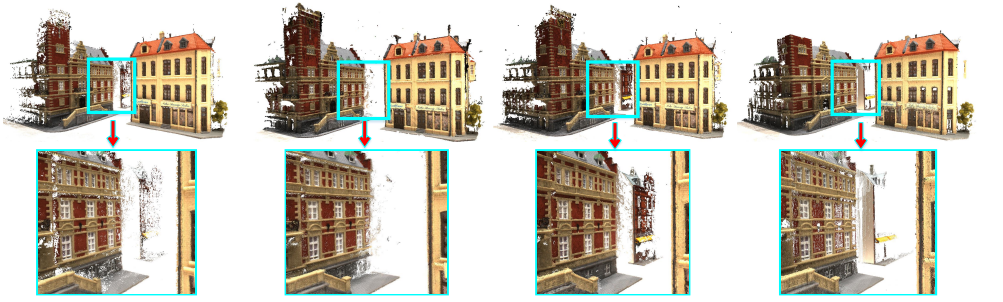

Figure 1: Comparison Between the State-of-the-art Learning-based Multi-view Stereo Approach [23, 25] and Ours. Best View on Screen. From left to right: Reconstructed Point Clouds of CVP-MVSNet [25], AA-RMVSNet [23], CVP-MVSNet+Ours, and The Ground Truth Point Clouds.

1 Introduction

Multi-View Stereo (MVS) reconstruction is still a hot topic over the past decade. MVS can be regarded as an extensive process on the basis of structure-from-motion(SfM) [4, 5]. SFM extracts and matches the feature points from multi-view images and then reconstructs the sparse point clouds [19, 21]. Multi-view Stereo (MVS) aims to restore the realistic dense 3D scenes through images taken from different viewpoints and calibrated cameras [18]. It's a core problem in the computer vision field and has wide applications in autonomous driving, augmented reality, robotics, etc [1, 6].

One of the key advantages of deep learning-based MVS is cost volume regularization, where most networks employ multi-scale 3D CNNs [3, 11, 12, 26] to regularize 3D cost volumes. Nevertheless, 3D convolution operations bring great memory consumption and time computational complexity. Therefore, these MVS algorithms are difficult to apply to high-resolution scenes. To solve the above-mentioned problems, Yao et al. proposed a scalable multi-view stereo vision framework based on recurrent neural networks (R-MVSNet) [27]. Zehao Yu et al. proposed a sparse-to-dense and coarse-to-fine Fast-MVSNet [28] for fast and accurate MVS depth estimation. Cascade-MVSNet [8] adopted a cascade method in cost volume regression and gradually regressed from coarse to fine to obtain a fine depth map representation. The coarse-to-fine network structure design of cascade-MVSNet greatly reduces the memory usage of MVSNet. Jiayu Yang et al. proposed CVP-MVSNet [25] for depth inference, which built a cost volume pyramid in a coarse-to-fine fashion instead of at a fixed resolution. The above methods greatly reduce memory consumption and improve reconstruction accuracy. However, the reconstruction completeness is still not satisfactory and misses some local areas in large scenes.

The ultimate goal of MVS research is to recover a complete and accurate 3D scenes. At present, many MVS methods are devoted to using novel and efficient network structures to increase their feature extraction and cost volume regression. However, given the aforementioned MVS pipeline, there are two main problems: (a) Local features are well captured by convolutions. The locality of convolution features prevents the perception of global context information, which is essential for robust depth estimation at challenging regions in MVS, such as weak texture, repetitive patterns, and non-Lambertian areas. (b) When decoding

matching costs, the features to be simply added, and potential depth information correspondences are not taken into consideration. Convolution operation has strong ability to extract local feature information, such as texture and color. However, for a whole input image, the correlation degree of the relevant information of the image itself seriously affects the learning of the global features of the object. Recently, [22] is initially proposed for natural language processing, which has been widely employed for its great performance in the computer vision community [29, 30]. Since the self-attention mechanism learns the relationships between elements of a sequence. As opposed to convolution neural networks that process images and can only attend to limited perspective fields, self-attention can attend to complete images thereby learning long-range relationships. Self-attention modules complement convolutions and help model long-range, multi-level dependencies across image regions. With self-attention, the network can capture images in which fine details in each local area are carefully coordinated with fine details in distant parts of the image.

To this end, we propose a method using the Global Context Complementary Network (GCCN) for multi-view 3D object reconstruction. The GCCN consists of three parts, of which a Global Context Interaction Module (GCIM) can strengthen long-range global context aggregation and local details within images. To better adapt GCIM into an end-to-end learning-based MVS pipeline, we introduce a valid skip connection to ensure a smooth transition from locally aggregated features by CNN to features with a global receptive field by GCIM. To fully preserve the underlying depth information in the 3D cost volume, we bridge different depths features with a Contextual-feature Complementary Learning Module (CCLM). Since remaining the global context-aware and more potential depth information within views, GCCN achieves significant improvement in reconstruction completeness and overall simultaneously on DTU dataset [1] (as shown in Figure 1). Moreover, the performance of our GCCN can be generalized to more complex scenes, such as the *intermediate* set of Tanks & Temples benchmark. Consequently, extensive experiments indicate that our method achieves state-of-the-art performance. We also conduct ablation experiments to demonstrate the effectiveness of each proposed module. Our contributions are as follows:

- We propose a novel end-to-end deep neural framework, namely Global Context Complementary Network (GCCN), for robust long-range global context aggregation within images. Moreover, the combination of local and global information contributes to converge network.
- In addition, to better regress the depth map, we introduce a contextual-feature complementary learning module to restore the 3D structure information of the scene.
- Our method achieves state-of-the-art results on the DTU dataset and the Tanks & Temples benchmark.

2 METHODOLOGY

In this section, we present our MVS method, named GCCN. Below, first, we provide a network architecture of GCCN, then we elaborate on the details of its novel depth inference module.

2.1 Network Architecture

As shown in Figure 2. We apply the recent CVP-MVSNet [25] as the backbone network in our framework. Specifically, CVP-MVSNet first takes as input a reference image $I_0 \in R^{H \times W}$,

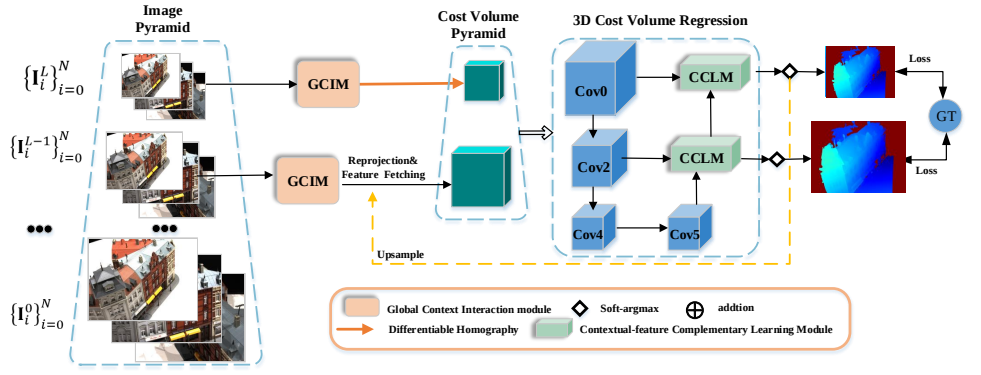

Figure 2: Network architecture of the proposed Global Context Complementary Network (GCCN) on CVP-MVSNet [25], denoted as CVP-MVSNet+Ours. It contains four components: pyramid feature aggregation, Global Context Interaction Module, cost volume, and 3D U-Net regularization, and depth map estimation.

source images $\{I_i^l\}_{i=1}^N$ and the corresponding camera intrinsic and extrinsic parameters for all views $\{K_i, R_i, t_i\}_{i=0}^N$ and infer the depth map D_0 for I_0. The entire architecture of CVP-MVSNet contains four stages. Firstly, build an image pyramid from high to low resolution. Then, extract features using FPN. Next, CVP-MVSNet adopts a cost-volume pyramid structure with weight shared across levels. The operation can be trained with low-resolution images and still handle any high-resolution image during inference. Finally, CVP-MVSNet regresses cost volume by 3D CNN and estimates the final depth map.

A Global Context Interaction Module(GCIM) designed, which involves two key points: local detail extraction and global feature acquisition. This ingenious combination can not only prevent network degradation, but also capture the long-term dependence between the underlying semantic features and the high-level structural features. Therefore, the feature image after contains rich local and global feature information. An effective contextual-feature complementary learning module (CCLM) strategy is introduced in the regression calculation of the 3D cost volume, which can complementarily learn the features of two input cost volumes with different depths. Finally, by constraining depth maps with pixel-level losses, thus the network can acquire more accurate 3D depth information about the objects.

2.2 Global Context Interaction Module (GCIM)

To improve the performance of the network in 2D feature extraction stage, we propose the GCIM module, which can effectively combine the local and holistic features. For multi-view 3D object reconstruction tasks, the different input images correspond to different camera parameters. There are differences in the feature information contained in images from different views, especially for the overlapping region, completeness of features from different viewpoints will directly affect the quality of the final depth map. As shown in Figure 3, we utilize convolutional neural network (CNN) to extract local details in the branch below. Simultaneously, we apply self-attention to learn the holistic information in the above branch. Finally, according to the importance of channels, we employ SE [9] fusion module to better integrate the complementary features of the local and long-term dependence features. Since the attention mechanism attempts to explicitly model the channel interdependencies or spa-

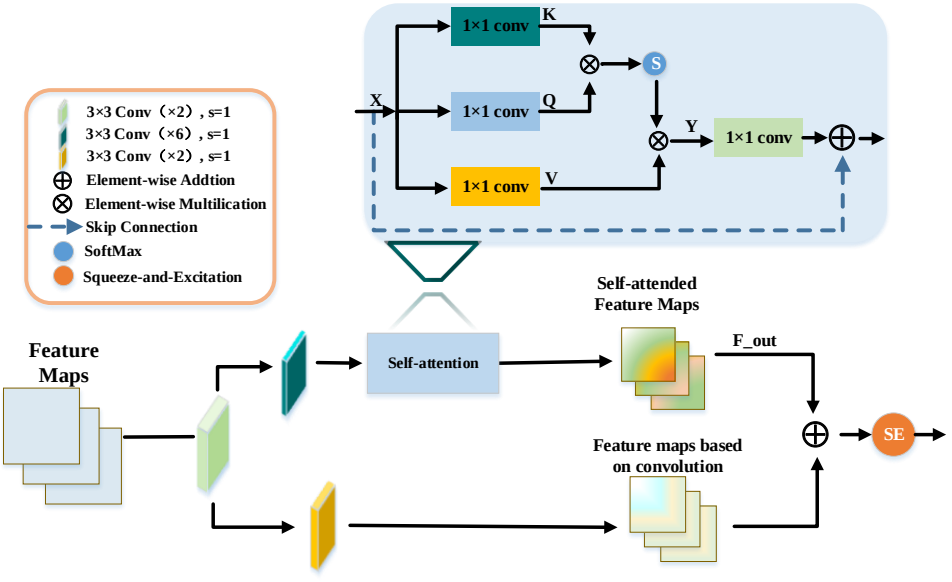

Figure 3: GCIM: The Architecture of Global Context Interaction Module.

tial correlation to enhance the learning of convolutional features, which aims to increase the sensitivity of the network to informative features. The self-attention mechanism is to further analyze the essential information from long-term dependence so that the network can focus on these features. Convolution and self-attention are complementary learning from local and global features, respectively, which preserve richer detailed features, thereby achieving high-completeness 3D scene reconstruction. To prevent the network degradation, our GCIM is based on the skip connection and self-attention mechanism, we adopt residual connections at two points in the module. The feature maps $X \in R^{C\times H\times W}$. Self-attention is defined as follows:

$$Y = Att(X) = \text{SoftMax}\left(QK^T\right)V \tag{1}$$

where Q, K, and V represent query, key, and value respectively. Compared with the traditional convolution calculation, the self-attention calculation can be divided into three steps: First calculate the Q, K, and value V. Then, the product KQ^T to measure their similarity; Finally, it is weighted according to the calculated similarity and all steps are repeated for each pixel in the image patch. In this paper, a residual connection is added after the GCIM operation, and the feature map weighted by the learned mixed weight is added to the input feature map to prevent the network from overfitting. The operation is defined as follows:

$$F\_out = \mathrm{W}\cdot Y + X \tag{2}$$

where W consists of the learned parameters, that is, the weight matrix after mixing. $F\_out \in R^{C\times H\times W}$ represents the feature map weighted by the GCIM and the input feature map after mutual parameterization.

2.3 Contextual-feature Complementary Learning Module (CCLM)

Most current learning-based MVS methods follow the MVSNet [26] approach to construct 3D cost volumes. As a core step connecting the 2D feature extraction and 3D regularization network, the deformation of the source image viewpoint to the reference image viewpoint is achieved in a differentiable case, and this depth map inference approach is trained in an end-to-end manner. The benchmarking network CVP-MVSNet [25], its 3D cost volume regression network adopts the standard 3D CNN U-shape network, the process is divided into the encoding (down-sampling) part and decoding (up-sampling) part. The 3D cost volume structure of the same depth in the encoding stage and the decoding stage adopts a direct connection. Due to the limitations of the U-shaped structure itself, downsampling followed by upsampling will lead to the dilution of high-level semantic information and the loss of spatial information, thus affecting the final integrity of the depth map.

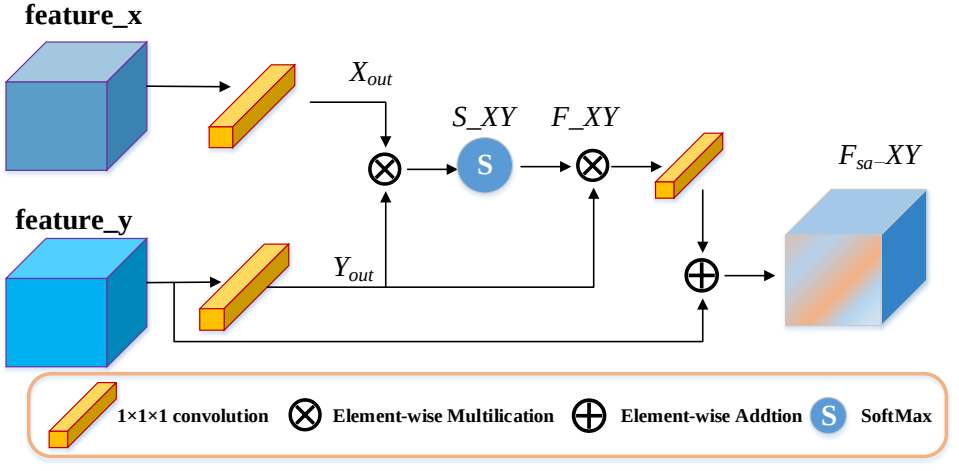

Figure 4: CCLM: A Diagram of the Contextual-feature Complementary Learning Module.

High-completeness feature representations are essential for 3D scene reconstruction, which could be obtained by capturing long-range contextual information and local feature information. In order to effectively fuse 3D cost volumes and retain rich contextual relationships of different depths, we employ the contextual-feature complementary learning module (CCLM) to easily achieve high-completeness 3D cost volume regression. The CCLM decodes a wider range of contextual information between two-layer features into local features, thus enhancing their representation capability. The specific operation steps of the CCLM are shown in Figure 4. Given two-layer 3D cost volumes $feature\_x, feature\_y \in R^{C\times D\times H\times W}$ with inconsistent depths, we first feed them into a convolution layer with a stride $1\times 1\times 1$ to generate two new feature maps X_{out} and Y_{out}, respectively. After that, we perform a matrix multiplication between the transpose of X_{out} and Y_{out}, and apply a softmax layer to calculate the spatial attention map S_XY, which is defined as:

$$\begin{aligned} X_{out} &= \mathrm{Conv}_{1\times1\times1}(feature\_x) \\ Y_{out} &= \mathrm{Conv}_{1\times1\times1}(feature\_y) \\ S\_XY &= \mathrm{Softmax}(X_{out} \otimes Y_{out}) \end{aligned} \tag{3}$$

Meanwhile, we perform a matrix multiplication between Y_{out} and the transpose of S_XY

to obtain the features F_XY. Finally, we perform an element-wise sum operation with the $feature\_y$ and F_XY, which is defined as:

$$F\_XY = (S\_XY)^T \otimes Y_{out} \tag{4}$$

$$F_{sa}\_XY = \mathrm{Conv}_{1\times1\times1}(F\_XY) + feature\_y \tag{5}$$

where F_XY represents the correlation coefficient matrix of the input 3D cost volume $feature\_x$ and $feature\_y$, and F_{sa}_XY represents the 3D cost volume output after the contextual-feature complementary learning module.

2.4 Training Loss

We utilize a loss function to constrain the error between the depth map and the ground truth predicted by the 3D cost volume regression. The calculation formula adopts mean absolute error (MAE) as the training loss of the network (Formula 6):

$$Loss = \sum_{l=0}^{L} \sum_{p\in P_{\text{valid}}} \|d(p) - d^p(p)\|_1 \tag{6}$$

where P_{valid} represents the valid pixel set in the ground truth depth map, $d(p)$ represents the depth value of pixel p in the ground truth depth map and $d^p(p)$ donates the depth estimation value predicted by the network. l represents the first layer of the pyramid image, as shown in Figure 2, the predicted initial depth map is upsampled back to the same size as the input image pyramid, and then its error values are accumulated layer by layer. During training, l is set to 2.

3 Experiment

In this section, firstly we describe the datasets and training procedure adopted for GCCN, then we compare it to state-of-the-art works on popular MVS benchmarks.

3.1 Datasets

DTU [1] Dataset: The DTU dataset is a large-scale MVS dataset that has 124 different scenes with 49 scans using the robotic arms [1, 10]. Each scan has seven known pose conditions as lighting changes. DTU provides 3D point clouds acquired using structured-light sensors. Each view contains calibrated camera parameters and the corresponding depth maps. We follow the same train-test split as in MVSNet [26] and other methods based on MVS [8, 15, 25]. We select scenes {1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110, 114, 118} as the testing set and other scenes as the training set.

Tanks & Temples [14]: The dataset is a large online benchmark that captures more complex real-world indoor and outdoor scenes. It mainly contains pictures of tanks and temples, which are used for 3D reconstruction tasks. Different scenes have different scales, surface reflections, and exposure conditions. The dataset includes a training dataset and a test dataset. The test dataset is divided into the intermediate group and advanced group. In this paper, the intermediate and advanced groups are used for testing, and the data includes sculptures, trains, playgrounds, temples, and some buildings with appearance camera trajectories.

	Methods	DTU test set		
		Acc.(mm)	Comp.(mm)	Overall(mm)
Traditional	Camp [2]	0.835	0.554	0.695
	Furu [6]	0.613	0.941	0.777
	Tola [20]	0.342	0.190	0.766
	Gipuma [7]	**0.283**	0.873	0.578
Learning-based	SurfaceNet [11]	0.450	1.040	0.745
	MVSNet [26]	0.456	0.646	0.551
	R-MVSNet [27]	0.383	0.452	0.417
	P-MVSNet [16]	0.406	0.434	0.420
	MVSCRF [24]	0.371	0.426	0.398
	EF-MVS [15]	0.402	0.375	0.388
	AA-RMVSNet[23]	0.376	0.339	0.357
	Cascade-MVS[8]	0.325	0.385	0.355
	CVP(Baseline) [25]	0.296	0.406	0.351
	Ours	0.371	**0.303**	**0.337**

Table 1: Multi-view Stereo Quantitative Results of Different Methods on DTU [1] Dataset (Lower is Better) Our Method Outperforms All Methods on Completeness and Overall Reconstruction Quality.

3.2 Implementation Details

During the training phase, we only use the training set of the DTU. Following the CVP-MVSNet [25] work, this paper also uses a smaller resolution, 160×128, and estimates a depth map that is the same size as the input image. For training, we downsample the high-resolution image to a smaller size, then build the image and set the real depth pyramid level to 2. To build the cost volume pyramid, we uniformly sample M = 48 depth hypotheses across the entire depth range at the coarsest (level 2). Each pixel has M = 8 depth residual hypotheses at the next level for refining the depth estimate. Models were trained using a single NVIDIA Quadro RTX 6000 series graphics card. Its GPU with about 15G of available memory can process multiple batches. The training batch size is set to 8. Training for a total of 27 epochs, the learning rate for the initial epoch is set to 0.001, and at the 10th, 12th, 14th, and 20th epochs the learning rate is divided by 2. For a fair comparison with other MVS methods, we use one reference image and two source images. In the testing phase, the resolution of the input image is 1600×1184, and the pyramid has 5 layers. We implemented our network using the popular deep learning framework Pytorch [17] and applied ADAM [13] to optimize our model.

3.3 Comparison with State-of-the-Art Methods

Evaluation on DTU dataset We compare our method with both traditional methods and recent learning-based methods. The quantitative results are shown in Table 1. Although Gipuma [7] achieves the best performance in terms of *accuracy*, our method outperforms all competing methods in both *completeness* and *overall* quality. In particular, the reconstruction *completeness* and baseline methods have significantly improved. Figure 5 shows the qualitative comparison with the result of CVP-MVSNet [25] and AA-RMVSNet [23].

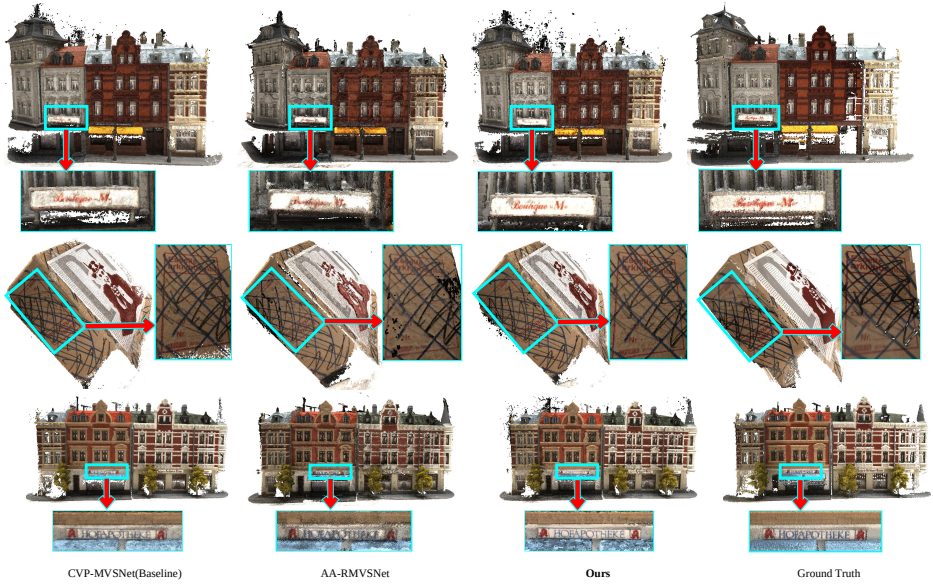

Figure 5: Comparison of reconstructed results with state-of-the-art methods [8, 25] on DTU evaluation set. Qualitative results of *scan*9,*scan*13, and *scan*15. The zoomed local areas in the point cloud are shown with a yellow rectangle. Our reconstruction contains more fine detailed structures, which demonstrates the effectiveness of our method.

Model Architecture		Mean Distance(mm)		
GCIM	CCLM	Overall	Comp.	Acc.
✗	✗	0.351	0.406	**0.296**
✓	✗	0.353	0.307	0.400
✗	✓	0.345	0.310	0.380
✓	✓	**0.337**	**0.303**	0.371

Table 2: Performance Comparison When With or Without the Module We Propose.

Our method produces the most complete point clouds, especially in those textureless and reflected areas. Our reconstruction is cleaner around finely detailed structures, which validates the effectiveness of our methods.

Generalization on Tanks & Temples dataset To further demonstrate the generalization ability of our method, we test the proposed method on more complex outdoor *Tanks and Tepmple* [14] dataset, using the model trained on DTU. The qualitative point clouds results of the *intermediate* set and *advanced* set are visualized in Figure 6.

3.4 Ablation Experiments

To verify the effectiveness of our method in the baseline [25] network structure, the completeness and accuracy of the 3D object reconstruction results and the average comparison of the two are performed on the DTU test dataset by us. And the accuracy and completeness

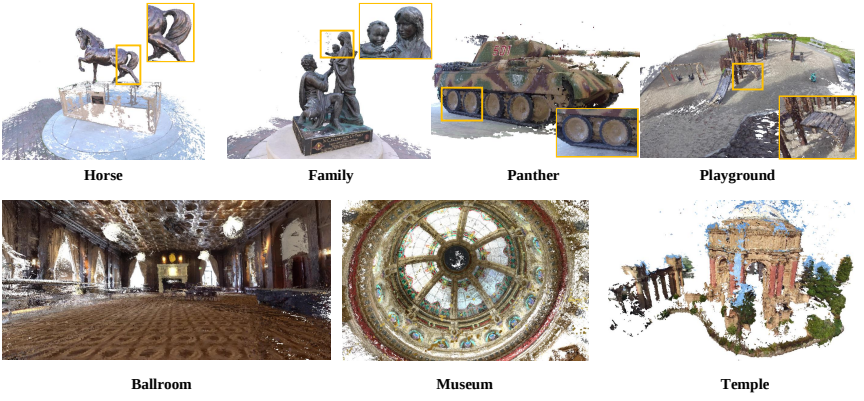

Figure 6: Point cloud reconstruction of Tanks and Temples dataset [14]. The first row shows the qualitative results for the *intermediate* set, and the second row shows the qualitative results for the *advanced* set. The zoomed local areas in the whole point cloud are shown with a yellow rectangle. Best viewed on screen.

describe the error between the reconstructed 3D point clouds model and the ground truth point cloud model, so the lower value means the better the model. As shown in Table 2, the best results are shown in bold. The control variable method is adopted to verify each module of the method in this paper. The ablation experiments show that the proposed method and network architecture have robustness in multi-view 3D object reconstruction tasks.

4 Conclusion and Future work

In this paper, we have designed a Global Context Complementary Network (GCCN) for high completeness MVS reconstruction, which focuses on long-distance dependencies, thereby maximizing the preservation of on global information of the scene. Specifically, our global context interaction module comprises effective context-aware information within images, which focuses on global structure information of objects. In addition, we utilize the contextual-feature complementary learning module to effectively fuse the cost volume features of different depths, and regress a higher completeness depth map.

For future works, we are interested in employing fast and efficient MVS structures such as tensor decomposition, ensuring the completeness of large-scale scene reconstruction while improving its accuracy.

5 Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62062056, in part by the Ningxia Graduate Education and Teaching Reform Research and Practice Project 2021, and in part by the National Natural Science Foundation of China under Grant 61662059.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

[2] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008.

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019.

[4] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1212–1221, 2017.

[5] J. M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, and M. Pollefeys. Building rome on a cloudless day. *Springer, Berlin, Heidelberg*, 2010.

[6] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[7] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016.

[8] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 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*, pages 2495–2504, 2020.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[11] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

[12] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[15] Kui Lin, Lei Li, Jianjun Zhang, Xing Zheng, and Suping Wu. High-resolution multi-view stereo with dynamic depth edge flow. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[16] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[18] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[19] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006.

[20] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012.

[21] Roberto Tron, Xiaowei Zhou, and Kostas Daniilidis. A survey on rotation optimization in structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 77–85, 2016.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6187–6196, 2021.

[24] Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, and Jiayu Bao. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4312–4321, 2019.

[25] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

[26] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[27] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[28] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020.

[29] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[30] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.