Sub-pixel Convolution and Edge Detection for Multi-view Stereo

1st Fanqi Yu
School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Shenzhen, China
fqyu@stu.pku.edu.cn

2nd Ronggang Wang
School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Shenzhen, China
rgwang@pkusz.edu.cn

***Abstract*—The deep multi-view stereo (MVS) approaches generally construct a cost volume pyramid in a coarse-to-fine manner to regularize and regress the depth or disparity, which is often built upon a feature pyramid encoding geometry or an image pyramid. A pyramid is an excellent approach to reducing memory, and many papers said even low-resolution images or features contain enough information for estimating low-resolution depth maps. However, recent papers show that the higher the image resolution, the better the output depth map, which means the resolution of depth maps in each stage cause effect on the final outputs. Therefore, we think the low-resolution depth map may not be enough for the high-resolution depth map. In this paper, we propose a sub-pixel upsampling module for post-processing the cost volume to generate a big resolution depth map at each stage. Besides, we also proposed an edge-weighted loss function for optimizing those inaccurate depth values in the edge regions of objects. Finally, we implement them on CasMVSNet, showing the effectiveness of our proposed method.**

***Index Terms*—multi-view stereo, mvsnet, sub-pixel, edge**

I. Introduction

MVS aims to estimate a depth map for each image and then reconstruct a dense point cloud of the scene. Since MVSNet [1] was proposed, the critical steps of most learning-based MVS methods that rely on the plane sweep algorithm have not changed much. It generally includes feature extraction, cost volume construction, regularization, and depth map inference.

Since MVSNet's proposition [1], the optimization of mainstream networks is pretty clear: the current mainstream feature extraction method FPN (Feature Pyramid Networks) [2] was proposed by Lin in 2017, and then many models have adopted Unet [3] proposed by Ronneberger in 2015. In 2020, Yang and Gu proposed a depth estimation framework based on pyramid structure in CVPMVSNet [4] and CasMVSNet [5], respectively, to effectively reduce memory occupation by using the pyramid structure from small to large resolution. Since then, pyramid-structured MVS networks have become popular, and the balance between memory usage and performance has continued [6]–[9].

Take the pyramid structure model CasMVSNet [5] as an example. The spatial resolution of feature maps gradually increases and is set to 1/16, 1/4 and 1 of the original input image size. Besides, from the first to the third stage of the pyramid, the receptive field decreases from large to small. Although this pyramid structure effectively reduces memory usage, it causes an unavoidable performance degradation. First, with the increase of the feature receptive field, the information contained in the feature is gradually enhanced, but the details are gradually inaccurate, such as the features of boundaries [10]. Second, the depth map obtained by each stage is often up-sampled for the next stage, and each upsampling will inevitably cause errors. Ideally, the size of the depth map obtained at each stage of the pyramid should be the same as the original image size, but this will inevitably increase memory.

This paper proposes two methods to alleviate the above problems effectively. First, we use sub-pixel convolution [11] to upsample and post-process the cost volume to achieve a balance between performance and memory. In most deep MVS, the regularization part uses the 3D convolution [1], [4], [5], which occupies the most significant memory in the entire model. Therefore, the smaller the spatial resolution of the feature extracted by the feature pyramid, the smaller the computation amount and the memory usage during the regularization calculation. So the pyramid structure like CasMVSNet [5] uses down-sampling resolution features and a smaller number of depth hypothesis layers for each stage. In this way, the memory occupation in regularization will be significantly reduced, and so will the final memory. But the accuracy of the depth map obtained by each stage will have to decrease.

Here, we propose a method for upsampling and optimizing the cost volume using sub-pixel convolution [11], which not only adds a minimal number of convolution calculations but also does not increase the number of regularization calculations. Furthermore, it could make the size of each stage's output the same as the next stage. Second, we propose a loss function based on edge detection. This method comes from [12]. In MVS, the depth of edge pixels is more inaccurate than the depth values of non-edge pixels. There could be many reasons accounting for this phenomenon. In addition to the inaccuracy position of convolution itself, the main reason for this problem we think is that the number of edge pixels accounts for a small proportion of the total number, which makes the network not pay enough attention to them. Therefore, we add a loss function for edge pixels to optimize the problem.

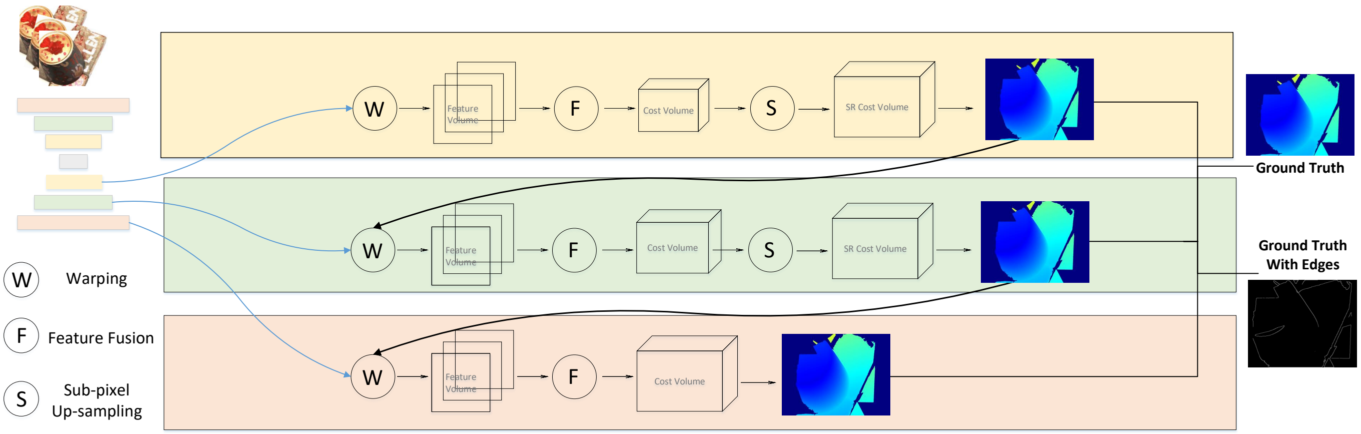

Fig. 1. **Stage-wise pipeline illustration.** This structure depicts the detailed pipeline of our methods, where SR means super resolution.

II. Method

This section will deliver the main contributions of this paper at length. We first review the typical pipeline of the learning-based MVS approach in Sec. A, then introduce the sub-pixel processing in Sec. B and edge loss in Sec. C.

A. Review of Learning-based MVS

As is shown in Fig.1. Most learning-based MVS methods are inherited from MVSNet, and their reasons are all to infer the depth d of the reference image. First, given reference image $\mathbf{I}_0$ and source images $\{\mathbf{I}_i | i = 1 \cdots N\}$, a pyramid feature extraction is applied for generating features at coarse to fine spatial resolution. Second, warp all features into the fronto-parallel planes of reference view to construct the feature volumes utilizing the differentiable homography. The homography between the source features of $\mathbf{i}_{th}$ view and the reference feature is:

$$\mathbf{H}_i(d) = d\mathbf{K}_i\mathbf{T}_i\mathbf{T}_0^{-1}\mathbf{K}_0^{-1} \tag{1}$$

where $\mathbf{K}$ and $\mathbf{T}$ denote as camera intrinsics and extrinsics respectively. Next, all feature volumes are aggregated to one cost volume $\mathbf{C}$ using the variance-based method:

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}(\bar{\mathbf{V}} - \mathbf{V_i})^2 \tag{2}$$

Where $\mathbf{V_i}$ donates the feature volume between reference view and $\mathbf{i^{th}}$ source view , and $\bar{\mathbf{V}}$ denotes the average feature volume.

Then a 3D U-Net is applied to regularize the cost volume and $\bar{\mathbf{V}}$ denotes the average feature volume. Finally, the regularized cost volume is regressed to generate the depth map. If it is in the first two stages of the networks, the corresponding depth map will be sent to the next stage to construct the depth hypothesis else the final output.

B. Processing of Cost Volume

As mentioned before, the depth map of the first two stages needs to be upsampled and then used by the next stage. The conventional upsampling method for depth maps is nearest-neighbor interpolation or bilinear interpolation, which will inevitably cause errors in the downsampling and upsampling process. So ideally, we want to find a method to make the depth map output at each stage the same resolution as the original, even without a significant increase in memory. Given that, we borrow sub-pixel upsampling [11] from the super-resolution field to optimize this network.

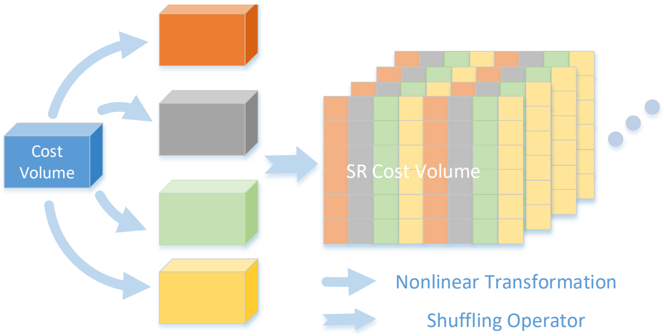

Fig. 2. **Process of Sub-pixel Upsampling.** Sub-pixel upsampling includes two main steps. First, Multiple nonlinear transformations generate multiple volumes. Second, the shuffling operator aggregates the volumes as a big one.

In the first two stages of the pyramid, we get the cost volume that is the same as one from the original method [5]. Then, to allow the network to learn the cost volume of the same width and height as the original image, we add sub-pixel upsampling after the cost volume for optimization. At the first stage of the pyramid, we use sub-pixels upsampling to quadruple the resolution of the regularized cost volume so that the size of the processed regularized cost volume can meet the needs of the second stage. Similarly, the regularized cost volume from the second stage is upsampled with the means of sub-pixels upsampling so that the number of cost volume channels after upsampling remains unchanged, and the height and width are the same as the size required in the third stage, as well. The above operation can be briefly described as:

$$\mathbf{C}^{SR} = \mathbf{PS}(\mathbf{W}_L * \mathbf{f}^{L-1}(\mathbf{C}) + \mathbf{b}_L) \tag{3}$$

where PS is a periodic shuffling operator, SR means super resolution. $\mathbf{f}^i$ is a nonlinear transformation as described by [11]. Fig. 2 shows a more transparent process.

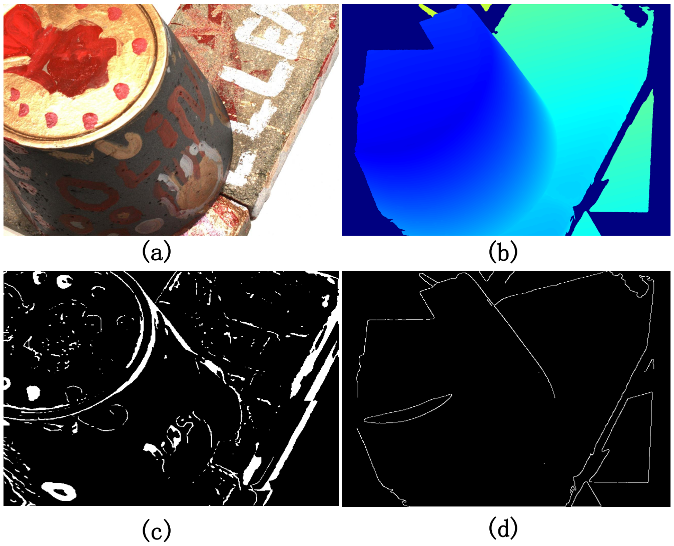

Fig. 3. **Gradient map Comparisons.** (a) and (b) are the image and truth depth of scan1 on DTU. (c) and (d) are gradient maps of (a) and (b).

This allows the network to perform post-processing on the regularized cost volume and avoids the error caused by the traditional upsampling of the pyramid structure. Furthermore, memory has no significant increase since the processing is done after 3D regularization.

C. Edge Loss Function

Inaccurate depth values in the edge regions of objects have become a well-known problem. Many problems cause inaccurate depth values in the edge region. One of the reasons is that the number of pixel samples in the edge part of the object is far less than the number of samples in the non-edge part, so the network cannot thoroughly learn in the process of minimizing the loss of the overall depth value. Based on this, we want to make the network as sensitive as possible to errors in depth values at the edges of objects. However,as is shown in Fig. 3, the gradient map of the picture is not the edge of the object, so it is also a difficult problem to obtain the actual edge of the object. Therefore, we take inspiration from [12] and use the gradient map of truth depth maps as an approximation of the object edge contour maps.

The specific method is as follows: we first use the canny operator to extract the edge map ($\mathbf{E}$) of the actual depth. Then, we set a threshold α to edge maps to judge whether a pixel belongs to edges, resulting in an edge mask(EM). Finally, we add an edge loss function(EL) as follows:

$$EL(\mathbf{D}_{gt}, \mathbf{D}) = \begin{cases} 0.5(\mathbf{D}_{gt} - \mathbf{D})^2 * EM, & \text{if } |\mathbf{D}_{gt} - \mathbf{D}| < 1 \\ (|\mathbf{D}_{gt} - \mathbf{D}| - 0.5) * EM, & \text{else} \end{cases} \tag{4}$$

where $\mathbf{D}_{gt}$ is ground-truth depth, $\mathbf{D}$ is estimated depth map.

III. EXPERIMENT RESULT AND DISCUSSION

This section demonstrates the effectiveness of our methods. We first introduce the datasets and implementation and then analyze our results.

Method	ACC.(mm)	Comp.(mm)	Overall(mm)
Baseline [5]	0.325	0.385	0.355
Baseline+EL	0.328	0.370	0.349
Baseline+SP	0.329	0.358	0.344

TABLE I
Quantitative results on DTU evaluation set. EL MEANS EDGE LOSS AND SP MEANS SUB-PIXEL CONVOLUTION.

A. Datasets

We evaluate our model on DTU [13]. DTU is an indoor MVS dataset captured in laboratory conditions with structured light scanner ground truth. It has 124 different scenes scanned from 49 or 64 views under 7 different lighting conditions with fixed camera trajectories. We adopt the same training, validation, and evaluation split as defined in [1].

B. Implementation Details

We use CasMVSNet as our baseline and implement our methods on it respectively, then train our model on DTU training set and evaluate it on DTU evaluation set. The input view selection and data pre-processing strategies are the same as [1]. We also follow the same training strategies and model configuration at each stage as [5] in training and evaluation. When training on DTU, the number of input images is 3; the image resolution is resized to 640 × 512, and α is set to 20. During the evaluation of DTU, we also resize the input image size to 1152×864 and set the number of the input images to 5. We still use the official evaluation protocol [13] to evaluate our methods' performances.

Fusibile [14] is our post-processing consisting. However, unlike previous methods [5], [15], we only introduce geometric constraints for depth map filtering instead of photometric filtering. It means the probability threshold is 0. Furthermore, the number of consistent views is set to 3.

C. Experiment Result and Discussion

We implement our methods on the baseline. The quantitative results on the DTU evaluation set is outlined in Tab. I, which exhibits that our method can make significant progress in performance.

Furthermore, in addition to the objective benchmark evaluation on DTU dataset, we also compare the performance of our method with the baseline on relatively complex data in DTU dataset. As Fig. 4 shows, to estimate the accurate depths of scan13 and scan48 is challenging due to weak texture and lighting issues. In CasMVSNet, the estimation results for these two scenes do not perform well. After adding EL to casmvsnet, it can be seen that due to the extra effort to some difficult-to-predict edge data, it is possible to estimate some depth that was difficult to predict. After adding SP, the subjective results can also explain that it benefits the cost volume to a certain degree.

IV. Conclusion

In this paper, we propose two methods for learning-based multi-view stereo. We first propose a sub-pixel upsampling module for post-processing the cost volume to generate a big-resolution depth map. Then we proposed an edge loss function

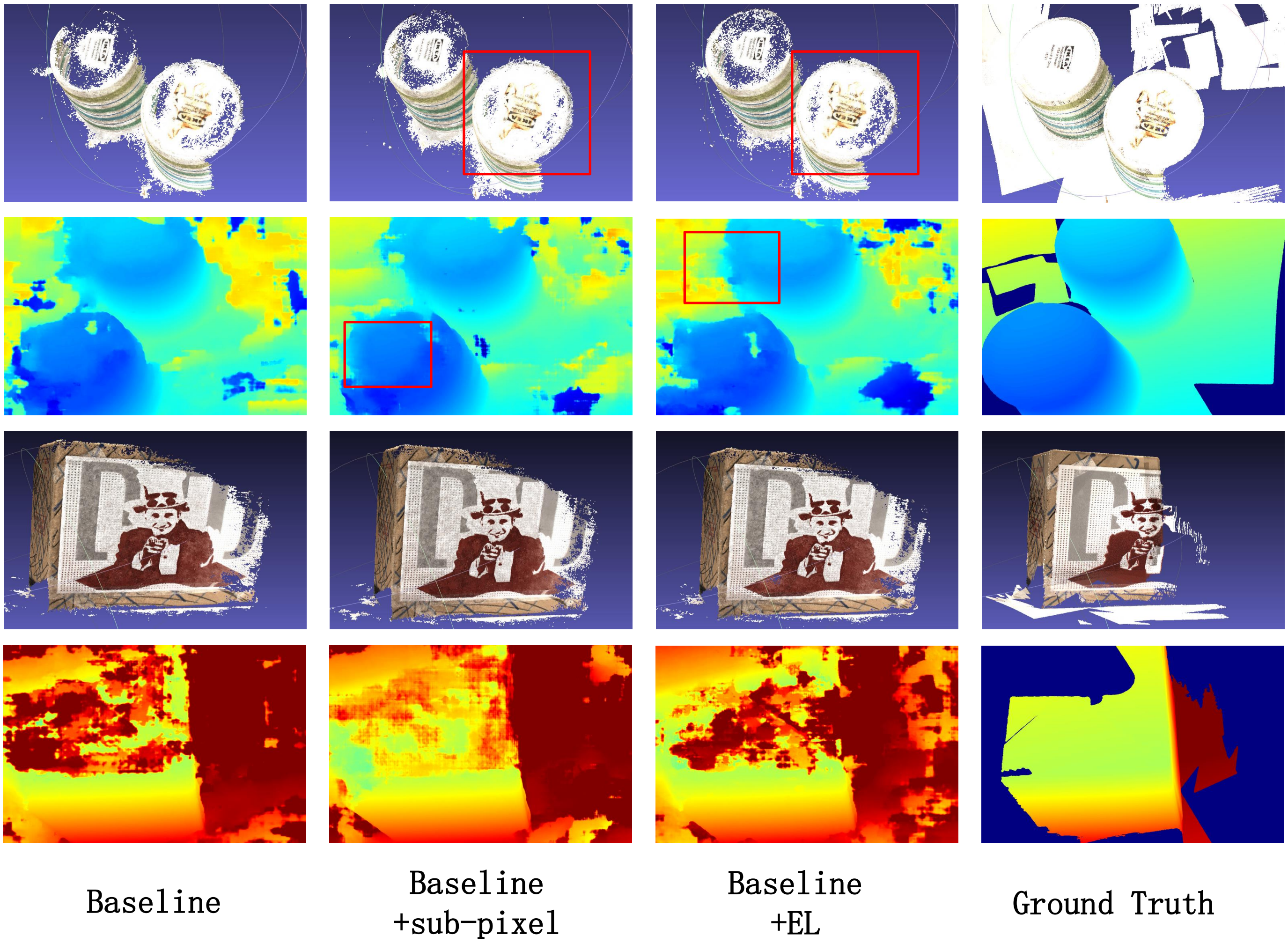

Fig. 4. **Results of Scan13 and Scan48.** The first and third lines are 3D point cloud. The second and forth lines are depth map.

to optimize those inaccurate depth values in the edge regions. The results show that our methods can get better accuracy and completeness. In the future, we will investigate more valuable strategies to optimize the MVS models.

References

[1] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *European Conference on Computer Vision (ECCV)*, 2018.

[2] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2015, pp. 234–241.

[4] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4877–4886.

[5] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. 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*, 2020, pp. 2495–2504.

[6] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1590–1599.

[7] S. Gao, Z. Li, and Z. Wang, "Cost volume pyramid network with multi-strategies range searching for multi-view stereo," *arXiv preprint arXiv:2207.12032*, 2022.

[8] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8626–8634.

[9] Z. Mi, C. Di, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 981–12 990.

[10] S. Xie and Z. Tu, "Holistically-nested edge detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.

[11] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[12] J. Chu, Y. Chen, W. Zhou, H. Shi, Y. Cao, D. Tu, R. Jin, and Y. Xu, "Pay more attention to discontinuity for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2020, pp. 166–175.

[13] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 153–168, 2016.

[14] S. Galliani, K. Lasinger, and K. Schindler, “Massively parallel multiview stereopsis by surface normal diffusion,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 873–881.

[15] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, “Rethinking depth estimation for multi-view stereo: A unified representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.