CGI2022 manuscript No.
(will be inserted by the editor)

Cost Volume Pyramid Network with Multi-strategies Range Searching

Abstract Multi-view stereo is an important research task in computer vision while still keeping challenging. In recent years, deep learning-based methods have shown superior performance on this task. Most of these methods attempt to reconstruct cost volumes to estimate the plausible depth hypotheses, which, however, require large amount of memory and computation consumption. Cost volume pyramid network–based methods which use the coarse-to-fine strategy to progressively refine the depth in each cost volume computation stage, have yielded promising results. However, these methods fail to take fully consideration of the characteristics of the cost volumes in each stage, leading to adopt similar range search strategies for each cost volume stage. In this work, we present a novel coarse-to-fine deep learning based cost volume pyramid network for multi-view stereo. Our proposed method, denoted by multi-strategies cost volume pyramid multi-view stereo network (MSCVP-MVSNet), combines different depth sampling range estimation strategies for each cost volume stage and make use of multi-dimension uncertainty without extra neural network modules. Furthermore, since the accuracy of the predicted depth map in coarse-to-fine framework is highly dependent on initial low-resolution depth map before refinement, we utilize probability distribution of each pixel as supervision of the initial cost volume to further improve the initial depth estimation. We conducted extensive experiments on both DTU and BlendedMVS datasets, and results show that our method outperforms most state-of-the-art methods.

Keywords Multi-view stereo · 3D reconstruction · Cost volume · Coarse-to-fine

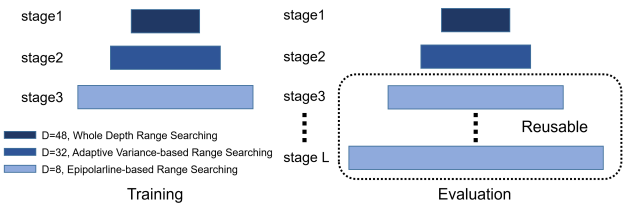

Fig. 1: Our deep MVS method uses different depth sampling range searching strategies in each stage of pyramid during training and evaluation.

1 Introduction

Multi-view stereo is one of the fundamental computer vision tasks which is widely used in augmented reality, 3D modeling and autonomous driving. Given a series of images captured from different views, multi-view stereo aims at reconstructing 3D model of the target by leveraging geometry and photometric information. Traditional methods rely on hand-crafted features and metrics for multi-view stereo matching, but they encounter difficulties in handling non-Lambertian and weakly textured surfaces.

In deep learning era, deep CNNs used for cost regularization and extracting representative image features have achieved promising performance. Yao et al. [2] first proposed an end-to-end MVS pipeline that constructs cost volume based on plane sweeping algorithm and aggregates different views by minimizing differential variance. However, this method consumes huge memory because that 3D CNN used for regularization is cubically proportional to image resolution. As a result, subsequent methods like [2,3] downsample high resolution images to regularize cost volume in a smaller resolution.

To this end, methods designed in coarse-to-fine manner [7–10] are proposed, which iteratively refine depth map based on cost volume pyramid and consume less memory. In coarse-to-fine structure like [9], the entire depth range is uniformly sampled in the first stage to infer a coarsest depth map, and the estimated depth map is iteratively upsampled to a higher resolution and refined in a small depth range.

However, current coarse-to-fine methods suffer from two limitations. First, the accuracy of the predicted depth map is highly dependent on the initial low-resolution depth map, since it is difficult to correct the depth of ill-posed and occluded pixels in the following narrow range. Second, current coarse-to-fine methods use same searching strategies in refinement stages after gaining initial depth map, which, however, not fully considered the characteristics of the cost volumes in each stage. For example, CVP-MVSNet [9] constructs cost volume with same depth searching strategy followed by weights-shared 3D convolution network. By back projecting neighboring pixels on epipolar line, this method leverages contextual information to determine depth searching range in next stage. AACVP-MVSNet [11] introduces a self-attention mechanism to feature extraction module in CVP-MVSNet [9] framework but still retains the same residual depth searching stragtegy. Cas-MVSNet [10] narrows depth range of each stage by hand-crafted range with specific decay ratio, and leverages a non-parameter-sharing cost volume regularization network. CFNet [18] uses an adaptive variance-based disparity range uncertainty estimation for stereo matching, to generate cascade cost volume while keeping the same searching method over three stages pyramid.

All these coarse-to-fine methods mentioned above uses shared depth hypothesis generating strategies in refinement stages, leading to the errors in coarse depth maps propagated into finer depth maps. In this work, we propose a multi-strategies cost volume pyramid multi-view stereo network (MSCVP-MVSNet). Instead of single depth range searching strategy, we utilize multi-dimensional information to calculate depth searching range for each layer. To further utilize the information contained in the cost volume, we introduce unimodal distribution as a training label at second stage during the training process.

Our main contributions can be summarized as follows:

We present multiple depth range searching methods in different stages of pyramid structure, leveraging multi-dimension information. For the second stage of the pyramid, we mine information in probability distribution of each pixel to adaptively determine depth searching range, while for the succeeding refinement stages, we leverage neighboring depth information to refine high resolution depth map in a narrow range iteratively.

To further exploit information in cost volume of deep MVS, we propose unimodal assumption as a training label in second stage and obtain a more accurate low-resolution depth map before iteratively refinement.

Quantitative results show that our method obtains SOTA results on DTU dataset and satisfactory qualitative results on BlendedMVS.

2 Related Work

2.1 Coarse-to-fine MVS methods.

Deep MVS methods [3, 4] based on pipeline of MVSNet [2] build cost volume at the resolution of output images, which usually occupy large memory dealing with high resolution dataset such as DTU [5] or Tanks and Temples [6]. In order to solve this problem, researchers propose coarse-to-fine reconstruction pipeline, which first downsample input images to build a low resolution cost volume and then perform subsequent refinement to obtain a high-resolution depth map. Chen et al. proposed Point-MVSNet [7] which estimates a coarse depth map, back project it into a point cloud and then refine the point cloud iteratively. Fast-MVSNet [8] also infers depth map in a low resolution, which simply uses depth propagation and Gauss-Newton refinement to obtain high-resolution depth map, taking into account speed and accuracy. CVP-MVSNet [9] and Cascade-MVSNet [10] both construct cost volume pyramid in a coarse-to-fine manner. They build cost volume across the entire depth-range in the coarsest resolution, after that they search in the neighbor of the current depth estimation to construct a partial cost volume at higher resolution levels. Based on these works [9, 10], Yu et al [11] propose AACVP-MVSNet, which introduces attention mechanism to CVP-MVSNet [9] framework. Zhang et al. [12] took into account the visibility between different views based on Cascade-MVSNet [10].

2.2 Depth sampling range.

Coarse-to-fine pyramid networks uniformly sample the entire depth range in the first stage. In the following stage, they iteratively narrow depth searching range by various strategies and sample depth hypothesis in this range. CVP-MVSNet [9] determines the local sampling range around the current depth by back projecting the corresponding pixels along epipolar line in source views.

Cas-MVSNet [10] narrows sampling range of each stage by hand-crafted range with specific decay ratio. For the first time, Cheng et al. [16] utilized variance of probability distribution to describe the uncertainty of depth estimation. Mao et al. [17] proposed uncertainty distribution-guided range prediction module to excavate multi-dimension information.

All these methods mentioned above employ identical sampling range searching strategies in each stage of three- or four-layer pyramid. CVP-MVSNet [9] takes advantage of neighbouring depth estimation but only uses the expectation of the probability distribution of each pixel. UCSNet [16] leverages variance of probability distribution but ignore the neighbouring information. In order to leverage both variance and neighbouring contextual information without adding complicated neural network modules, we apply different sampling range calculation strategies in different stage of coarse-to-fine MVS framework.

2.3 Cost volume.

Recently, cost volume is widely used in MVS and stereo matching methods. MVSNet [2] first introduces cost volume for end-to-end MVS pipeline by calculating photometric matching cost of each pixel in different fronto-parallel planes hypothesis. A standard cost volume has a resolution of $H \times W \times D \times F$, where H, W, D, F are height, width, number of plane hypothesis and feature channels, respectively. While cost volume indicates matching cost of each depth hypothesis of each pixel intuitively, it is regularized by 3D UNet to generate an estimated probability value and indirectly supervised as an intermediate layer. In order to integrate multi-scale information of cost volume, Shen et al. [18] proposed cost volume fusion module to obtain better initial disparity map. Like CFNet [18], we further utilize cost volume to obtain better initial depth map before refinement. To achieve this, we modified adaptive unimodal filter proposed by Zhang et al. [13] to our coarse-to-fine framework.

3 Methods

3.1 Overview

In this section,we introduce our multi-strategies cost volume pyramid network for high-resolution MVS reconstruction in details. The overview of the network is shown in Fig. 2. We assume the input reference image denoted by $I_0 \in \mathbb{R}^{H \times W}$, and source images represented by $\{I_i\}_{i=1}^{N-1}$. To build a pyramidal structure, we downsample input images L times to obtain images pyramid $\{I_i^j\}_{j=1}^{L}$, where $i \in \{0, 1, \cdots, N\}$. Feature pyramid $\{F_i^j\}_{j=1}^{L}$ are build by weights-shared feature extraction module.

As shown in Fig. 2, we apply three different strategies in each stage to determine depth sampling ranges in our framework. We use two non-parameter-sharing UNets to regularize cost volume with different depth sampling numbers. Since depth hypothesis sampling strategy 3 (DHS-3) is neural-network-parameters-free, we can build arbitrary number of layers during evaluation, even if we only train a three-layer pyramid. In order to obtain a higher quality low-resolution depth map, we apply adaptive unimodal filtering to put constrains on cost volume in second stage of our coarse-to-fine framework.

Inspired by GwcNet [14], we build cost volume by group-wise correlation instead of calculating feature volume variance over all views proposed by Yao et al. [2].

3.2 Depth sampling range estimation

As introduced in related work, previous methods ([9], [18], [11]) employ single strategy in each stage to calculate depth range, which either ignore statistical properties of each pixel or neighbouring information. However, if we apply additional neural network modules to estimate multi-dimensional uncertainty range on each layer, it will consume more memory and increase the computational complexity. To solve this, we fuse multi-dimensional information by simply combine different uncertainty estimation strategies in different stage and achieve satisfactory results.

In this section, we present our depth hypothesis sampling strategies in details. As shown in Fig. 1, the number of pyramid layers in our framework is flexible, we train 3 different layers while evaluate with arbitrary number of stages.

In the first stage, we uniformly sampled depth hypothesis over the entire range to obtain a coarsest initial depth map. Due to the large sampling range, we sampled more depth hypothesis ($D^1 = 48$) in this stage.

For second stage, we take advantage of probability distributions to calculate specific depth sampling range of each pixel. Previous methods ([13], [18]) indicate that texture-less and occluded pixels tend to have multiple or wrong matches. In this case, the expectation of the per-pixel distributions can not depict the properties of multimodality and dispersion. To solve this issue, we leverage the variance of the probability distribution as

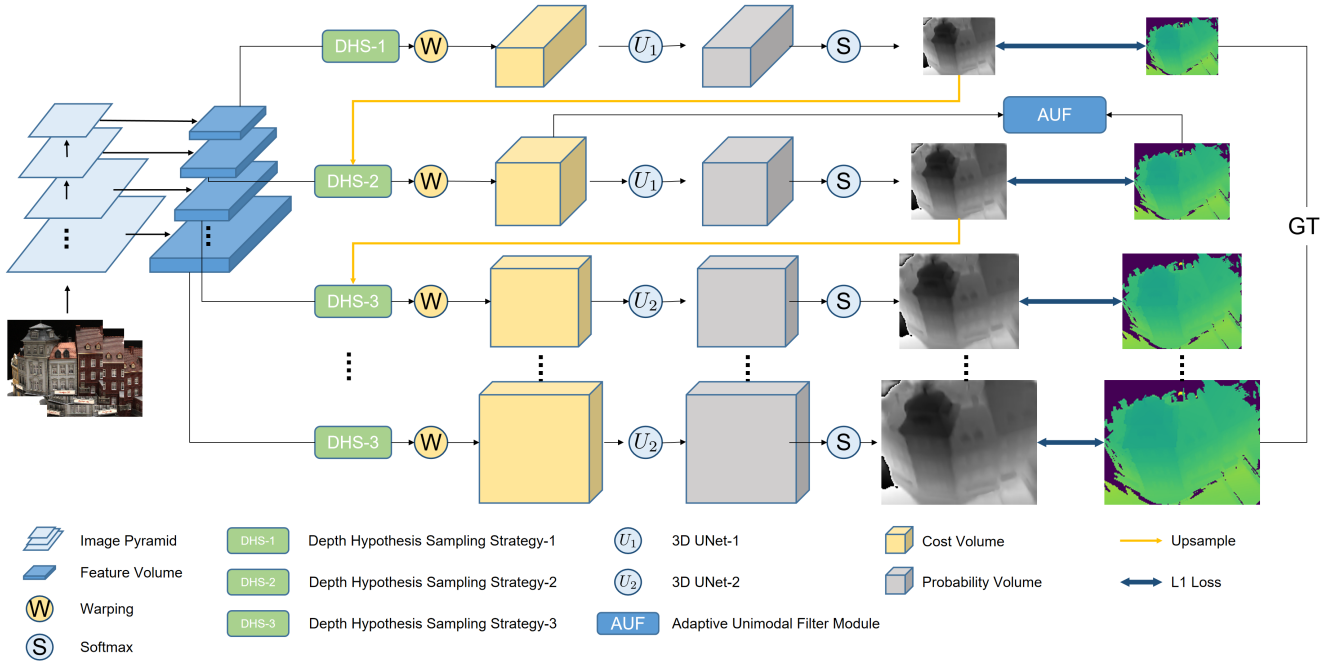

Fig. 2: The network structure of MSCVP-MVSNet. The feature volume pyramid is constructed by a weighted-shared feature-extraction module. In each stage, we estimate uncertainty of each pixel as the depth sampling range of next stage. We apply whole depth range searching in first stage. For second stage, we apply adaptive variance-based range searching method to determine sampling range. From the thrid stage, we back-project neighboring pixels to find depth hypothesis sampling range.

well as adaptive unimodal constraints (Sec. 3.3) to estimate per-pixel uncertainty and reduce local maxima of probabilities. We set the number of depth hypothesis, $D^2 = 32$ in this stage. For stage l, the variance at pixel i is defined as:

$$\hat{V}_i^l = \sum_{j=1}^{D^l} P_{i,j}^l (d_{i,j}^l - \hat{d}_i^l)^2, \tag{1}$$

where $P_{i,j}^l$ is the probability of pixel i at sampled depth j,$d_{i,j}^l$ is the depth of sampled plane j, $\hat{d}_i^l$ is the estimated depth of pixel i at current stage. Different from UCSNet [16], we adopt the idea of CFNet[18] that originally proposed in stereo matching task, which use learned instead of hand-crafted scale parameters to determine confidence interval:

$$\begin{aligned} d_{max}^{l+1}(i) &= \hat{d}_i^l + \alpha^l \sqrt{\hat{V}_i^l} + \beta^l, \\ d_{min}^{l+1}(i) &= \hat{d}_i^l - \alpha^l \sqrt{\hat{V}_i^l} - \beta^l, \end{aligned} \tag{2}$$

where α^l and β^l are learned parameters in stage l. Same as CFNet [18], we initial α^l and β^l as 0 at the beginning of training. In texture-less regions with multi-modal distributions, the variances tend to be large, and adaptive uncertainty range estimation algorithm adjust depth hypothesis to a larger range so as not to miss the truth depth value before small-range refinement. Since we only use this strategy in second stage, $l = 2$ in Eq. (1) and Eq. (2)

Our first two layers have yielded fair results at the low resolution stage, and the depth values of high-resolution depth maps are obtained via upsampling operation. To refine the depth values of the high-resolution depth maps, it is necessary to re-calculate the depth values in a small range from the third stage. Specifically, we apply neural-network-parameters-free method to determine sampling range, which take advantage of contextual information provided by neighboring pixels along epipolar line. Like CVP-MVSNet [9], we calculate depth sapmling range by back projecting neighboring pixels which is 0.5 pixels away along the epipolar line.

3.3 Supervise on Cost Volume

We further utilize the information in cost volume at 2nd stage to obtain better low-resolution depth map. To avoid increasing the computing time and complexity during evaluation, we modified adaptive unimodal filter

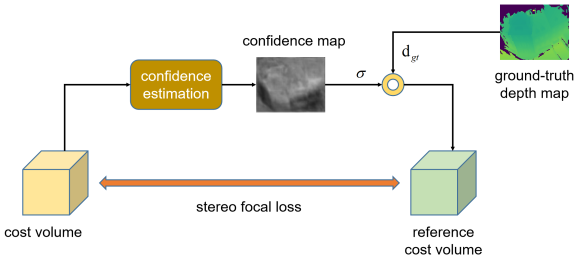

Fig. 3: Adaptive unimodal filtering.

proposed by Zhang et al. [13], which is only activated during training to constrain the cost volume.

Cost volume is defined to reflect the similarity between different views, where the true depth value should have the lowest cost, which means the probability distribution should be unimodal and peaked at the true depth hypothesis under ideal circumstances. Based on this assumption, we construct unimodal distributions as reference distributions which directly constraint on the cost volume to reduce errors introduced by multi-modal distributions. Following [13], we defined reference unimodal distribution as:

$$P^l(i) = softmax(-\frac{|d^l(i) - d^l_{gt}(i)|}{\sigma_i}), \tag{3}$$

where σ_i is variance of reference distribution for pixel i, which controls the sharpness of peak and it is defined as:

$$\sigma^l_i = \alpha^l_c(1 - f^l_i) + \beta^l_c, \tag{4}$$

where f^l_i is confidence value of pixel i in stage l. We estimate confidence value for each pixel by a 2D confidence estimation network. α^l_c and β^l_c are scale factor and lower bound, respectively. Different from [13], we use learned neural network parameters instead of handcrafted factors to adapt different properties of probability distributions for different datasets. Large σ indicates low confidence of pixel, which usually caused by mismatch in textureless regions.

We leverage stereo focal loss proposed by AcfNet [13] to guide network to generate unimodal distributions for each pixel. The stereo focal loss is defined as:

$$\mathcal{L}_{SF} = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}(\sum_{d=0}^{D-1}(1 - P_i(d))^{-\gamma} \cdot (-P_i(d) \cdot \log\hat{P}_i(d))), \tag{5}$$

where $P_i(d)$ is probability value of reference unimodal distribution at depth d of pixel i, and $\hat{P}_i(d)$ is estimated probability of pixel i at depth d given by our UNet.

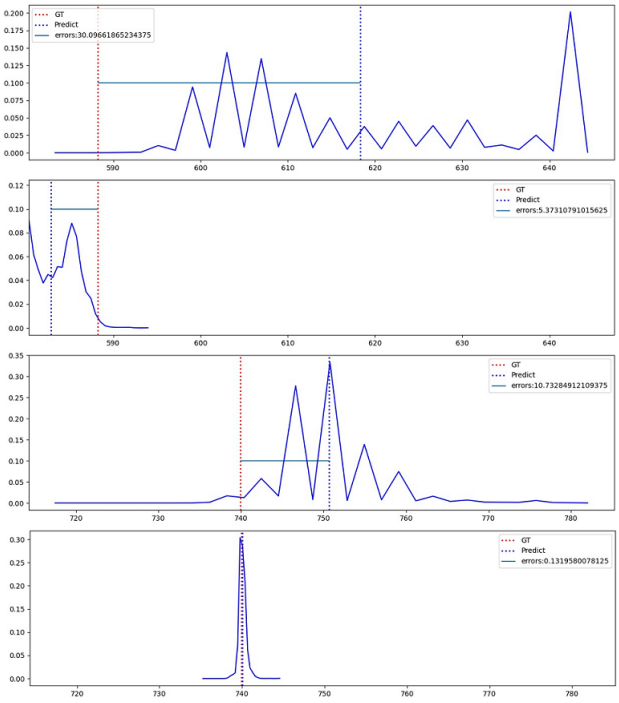

Fig. 4: Probability distributions of two pixels with and without our adaptive unimodal filtering. The first and third rows show the probability distributions of two pixels in stage 2 without adaptive unimodal filtering. The second and fourth rows are corresponding distributions after adaptive unimodal filtering. The estimated and ground-truth depth values are indicated by the vertical dashed lines in red and blue respectively. The errors between the estimated and ground-truth depth values are marked by the green horizontal line.

Instead of simple cross entropy loss, we set $\gamma \geq 0$ to force unimodal guidance to focus on high-confidence regions.

As shown in Fig. 4, after adaptive unimodal filtering (AUF module), some local maximas are eliminated, and the errors in stage 2 are decreased. Different from [13] which construct cost volume in a fixed resolution, we apply this filter in coarse-to-fine framework, which constructs more than one cost volume during the whole process. We apply adaptive unimodal filter only in 2nd stage instead of each stage of our pyramid. There are two reasons for this, firstly, appling AUF module for each stage consumes more time and computing resource. Second, for stages with small sampling numbers D, it is difficult to fit their distributions to unimodal distributions.

3.4 Loss Function

Our total loss consists of three parts: regression loss in each stage, stereo focal loss and confidence loss, which is denoted as:

$$\begin{aligned}\mathcal{L} =&\lambda_{SF}\mathcal{L}_{SF} + \lambda_C\mathcal{L}_C \\ &+\sum_{l=1}^{L}\omega^l\mathcal{L}^l_{regression}\end{aligned} \tag{6}$$

Where λ_{SF} and λ_C are two factors to balance stereo focal loss and confidence loss on second stage. The confidence loss $\mathcal{L}_C$ is defined as:

$$\mathcal{L}_C = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} -\log f_i \tag{7}$$

We apply negative log-likelihood function as confidence loss to encourage confidence estimation network to predict high confidence values for each pixel.

Regression loss $\mathcal{L}^l_{regression}$ is defined to reflect the difference between the predicted depth map and ground-truth at stage l. We use hand-crafted weight ω at each stage. For stage l, the $L1$ norm is defined as:

$$\mathcal{L}^l_{regression} = \sum_{i\in\mathcal{P}}\|d^l_i - \hat{d}^l_i\|_1 \tag{8}$$

4 Experiment

4.1 Dataset

DTU Dataset. We train and evaluate our network on DTU dataset [5] to obtain quantitative results. DTU dataset [5] consists of 124 large scale of scenes in 49 or 64 different views and 7 different light conditions, with the evaluation reference obtained by a structured light scanner. We use the same splited training and evaluation sets with [3,9,11]. While the original size of evaluation image is 1600×1200, we crop it to 1600×1184 to fit the upsample process.

BlendedMVS. BlendedMVS [22] is a collection of images captured from different views of 113 various scenarios. It contains 17K training samples in low-resolution (768×576) as well as high-resolution (2048×1536). Following the official training and validation list given by the released dataset files, we divided 106 scenes for training and the other 7 for validation. We choose low-resolution BlendedMVS for our training set and evaluate our method on both low-resolution and high-resolution datasets.

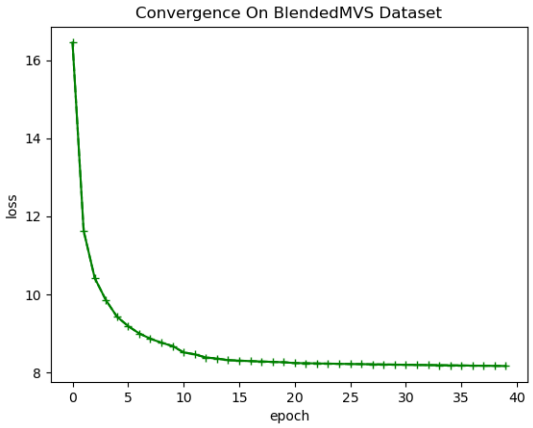

Fig. 5: Convergence on BlendedMVS dataset.

4.2 Implementation details

Training. We train and evaluate our model on DTU dataset and low-resolution BlendedMVS. As stated above, we construct a 3-layers pyramid during training, and applying whole range sampling, adaptive variance-based sampling and neighboring back-projection sampling on 1st, 2nd, and 3rd stage, respectively. For first stage, we uniformly sample the whole depth range $[425, 1065]$ with $D^1 = 48$, while for 2nd and 3rd stage, we choose $D^2 = 32$ and $D^3 = 8$, respectively. As the training process with high-resolution inputs is memory and time consuming, we downsample the training set into a size of 320×256, and the coarsest resolution is 40×32 in the first stage. We set hyperparameters $\lambda_{SF} = 10$, $\lambda_C = 80$ in equation (6). We set $\omega^1 = 0.5$, $\omega^2 = 1$ and $\omega^3 = 2$ to balance $L1$ loss in each stage. As for the reference unimodal distribution, we initialize the neural network parameters as $\alpha_c^2 = 13$ and $\beta_c^2 = 9$ based on empirical evidence from [13]. We use 3 different views as inputs and Adam [19] as optimizer in the training stage of the proposed network. We set batch size as 16 and train our model on 2 Nvidia GeForce RTX 3090 for 40 epoches with initial learning rate 0.001 multiplied by 0.5 at 10th, 12th, 14th, 20th epoch.

Evaluation. For DTU dataset,we crop the original images to 1600×1184 for evaluation. We set $L = 5$ for image feature pyramid to maintain a similar size with training stage at the coarsest stage (50×37). Similar to [2,3,9], we choose 5 views in evaluation for fair comparison. We set the same sampling numbers D in each stage as training process. As for BlendedMVS, we evaluate our proposed method on both low-resolution and high-resolution dataset.

Fig. 6: 3D models constructed by CVP-MVSNet [9], AACVP-MVSNet [11] and our method on DTU dataset.

Post processing and Metrics. After estimating the depth map, we fuse all views into a dense point cloud model for each scene. For fair comparison, we follow the common post processing method used by [2, 3, 9], which is a fusion method provided by Galliani et al. [20]. We run the official evaluation code provided by DTU dataset [5] to obtain quantitative results in terms of mean accuracy (acc.), mean completeness (com.) and overall score (overall). The evaluation results are listed in Tab. 1.

4.3 Results on DTU dataset

We train and evaluate our method on DTU dataset to conduct quantitative results in comparison with with other learning based methods. As shown in Tab. 1, our method achieves state-of-the-art results in overall score, which is comparable to PVSNet [21]. Especially, our method outperforms all methods in Tab. 1 in terms of completeness. As shown in Fig. 6, We visualize several

Table 1: Quantitative results on DTU dataset

Methods	acc.(mm)	comp.(mm)	overall(mm)
MVSNet[2]	0.396	0.527	0.462
R-MVSNet[3]	0.383	0.452	0.418
MVSCRF[4]	0.371	0.426	0.398
PointMVSNet[7]	0.361	0.421	0.391
CVP-MVSNet[9]	**0.296**	0.406	0.351
AACVP-MVSNet[11]	0.357	0.326	0.341
Vis-MVSNet[12]	0.369	0.361	0.365
USCNet[16]	0.338	0.349	0.344
PVSNet[21]	0.337	0.315	**0.326**
Ours	0.379	**0.278**	0.328

Note: We evaluate our model on DTU dataset in original image resolution and compare it with other methods. Quantitative results show that our network achieves best performance on completeness. (lower is better)

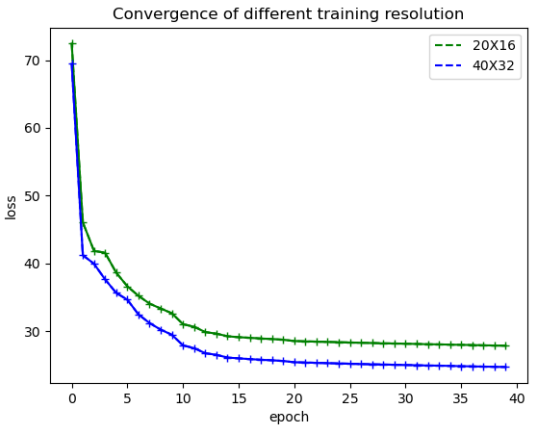

Fig. 7: Convergence of different training resolution.

reconstructed 3D models constructed by CVP-MVSNet [9], AACVP-MVSNet [11] and our proposed method.

4.4 Results on BlendedMVS

As stated above, we train our modal on low-resolution sets and evaluate it on both low- and high- resolution sets of BlendedMVS. Fig. 5 shows convergence of our model on BlendedMVS dataset. As BlendedMVS dataset does not provide any reference point clouds for quantitative evaluation, we conduct the visual comparison with CVP-MVSNet [9]. We set $L = 3$ in training process. During evalution, we set $L = 5$ and $L = 6$ for low and high resolution evaluation sets, respectively. The reconstruction results of low-resolution sets are shown in Fig. 8. We can clearly see that our approach is better than CVP-MVSNet [9] in completeness. To demonstrate the ability of our method to reconstruct large scenes in high resolution, we evaluate our method on several scenes of high-resolution BlendedMVS dataset. In the same way, we compare our method with CVP-MVSNet [9] and the results of high-resoluion dataset are shown in Fig. 9. On high-resolution data sets, the superiority of our method in terms of completeness is even more evident.

4.5 Ablation study

In this section, we perform ablation experiments on DTU dataset to validate the effectiveness of each component of our proposed network. Results are shown in Tab. 2. Below we analyse each component in details.

• **Non-parameter-sharing UNet.** 3D UNet is designed for cost volume regularisation and explore cost volume information in three dimensions. We replace our two separated UNets in proposed model with a parameter-sharing UNet (denoted as CVP-MS-Auf in Tab. 2). Quantitative results on DTU dataset show that our two parameter-separating UNets gain better results (0.328 vs.0.360) than parameter-sharing UNet. This indicates that former stages which search in a wider range have different characteristics with refinement stages in the cost volume regularization process.

• **Depth range estimation strategies.** We choose CVP-MVSNet [9] which apply epipolar line-based range estimation at each stage as baseline model. Although simply applying the proposed depth sampling range estimation strategy does not give a better result (CVP-MS in Tab. 2), we demonstrate that our reconstruction results are significantly improved when combining the proposed searching strategy with two non-parameter-sharing UNets.

• **Supervise on cost volume.** While our multi-strategies with two non-parameter-sharing UNet framework has achieve promising results (CVP-MS-U^2Net in Tab. 2), we obtain even better results when further adding adaptive unimodal filtering on 2nd stage, which

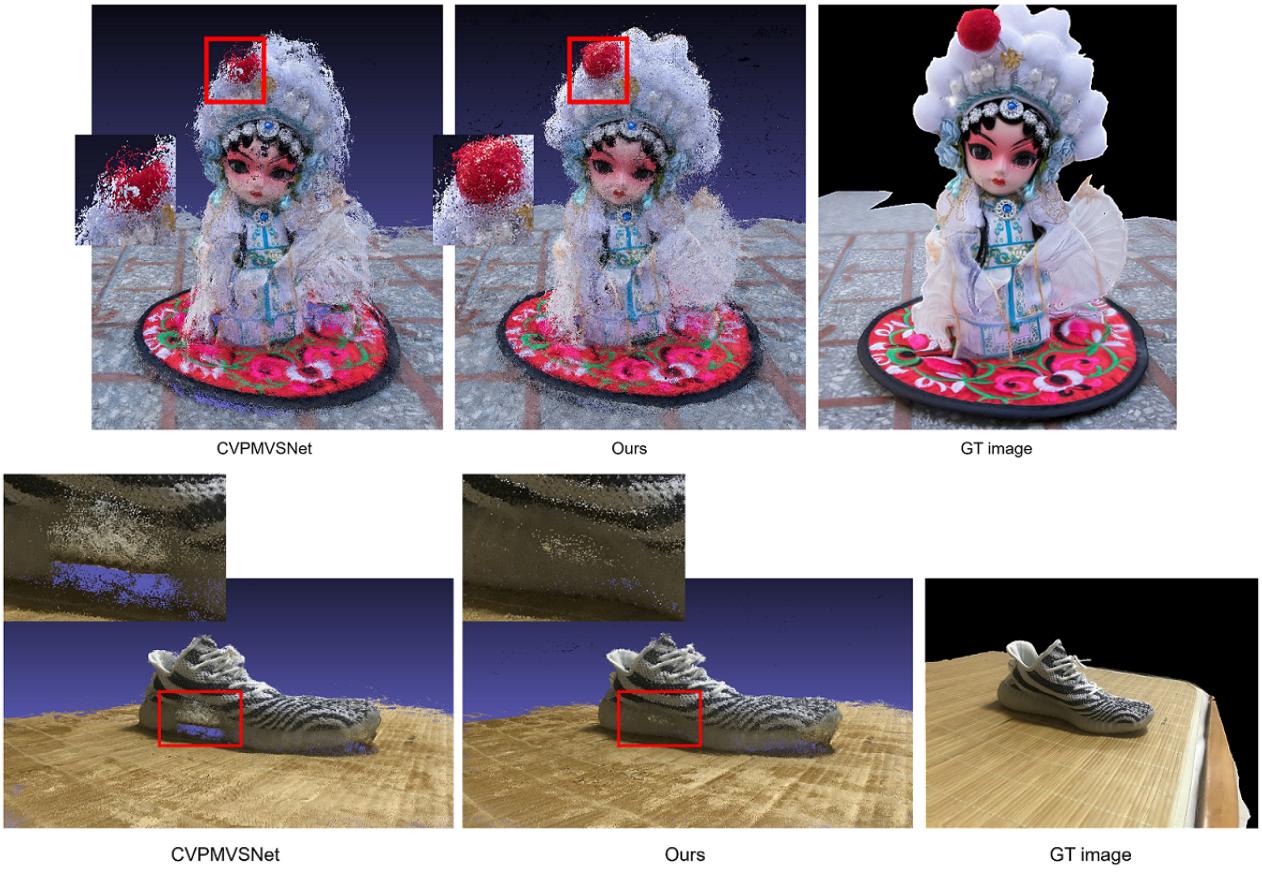

Fig. 8: Results on low-res BlendedMVS, compared with CVP-MVSNet [9].

Table 2: Ablation study on DTU dataset

Methods	Variance range	Epipolar line range	U^2Net	Auf	acc.(mm)	comp.(mm)	overall(mm)
CVP(baseline)	✗	✓	✗	✗	**0.313**	0.394	0.354
CVP-MS	✓	✓	✗	✗	0.343	0.439	0.391
CVP-U^2Net	✗	✓	✓	✗	0.330	0.379	0.355
CVP-MS-Auf	✓	✓	✗	✓	0.321	0.398	0.360
CVP-MS-U^2Net	✓	✓	✓	✗	0.389	0.279	0.334
Ours	✓	✓	✓	✓	0.379	**0.278**	**0.328**

Note: Our baseline model, denoted as CVP, is CVP-MVSNet [9] with epipolar line-based range searching. CVP-MS denotes multi-strategies which combines baseline model with variance-based range searching. U^2Net indicates two parameter-separated UNets, while our baseline uses one parameter-sharing UNet. Auf denotes adaptive unimodal filtering in 2nd stage.

is our final model (the last row of Tab. 2). Interestingly, quantitative results of CVP-MS-Auf and CVP-MS in Tab. 2 show that adaptive unimodal filtering gives a greater boost when parameter-sharing UNet is adopted.

• **Image resolution during training and evaluation.** Fig. 7 shows convergence of different resolutions on DTU dataset during training. "40×32" and "20×16" denote different coarsest resolution of pyramid in training process. As it is shown in Fig. 7, although both of them are converging, the final loss of higher res-

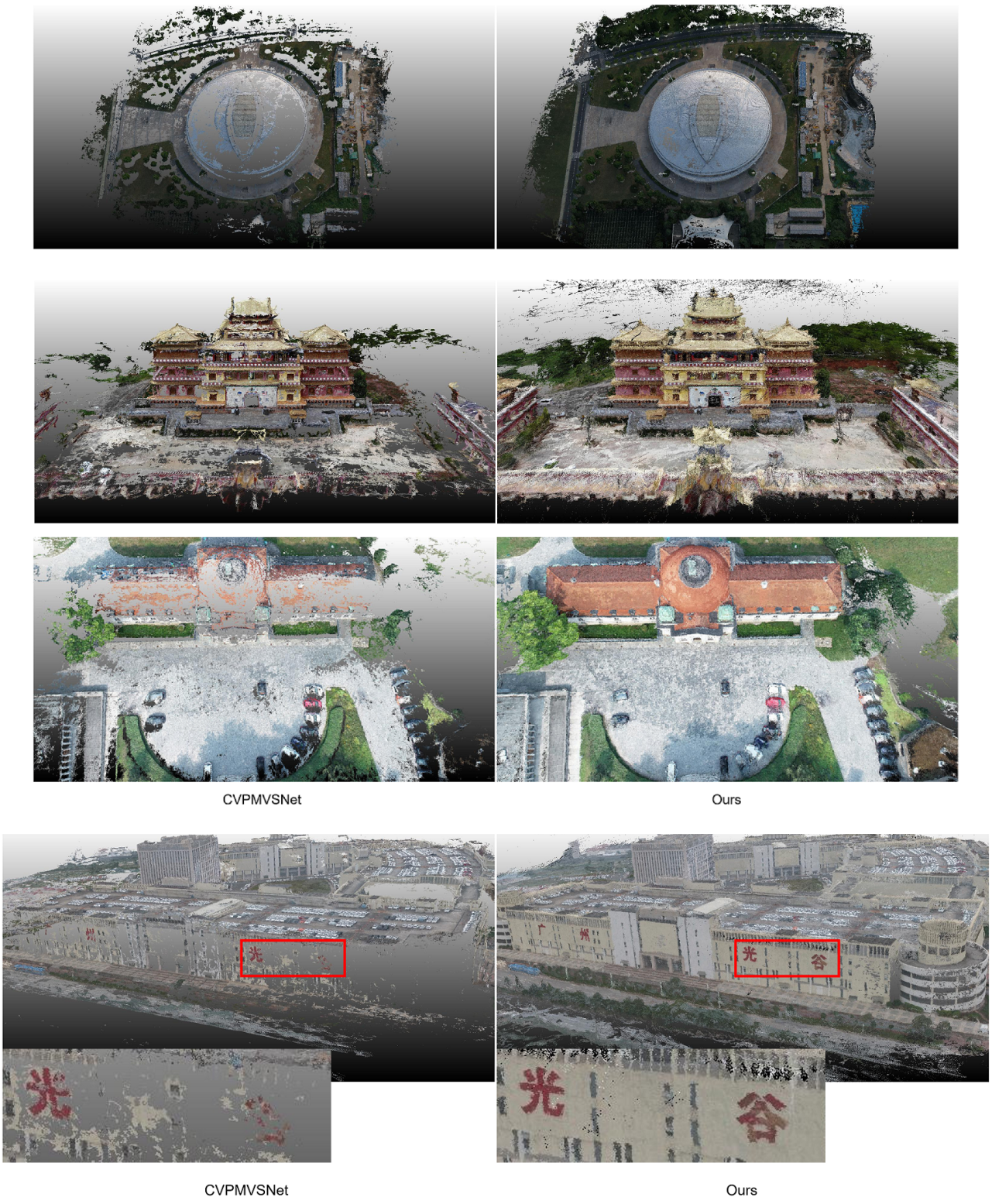

Fig. 9: Results on high-res BlendedMVS, compared with CVP-MVSNet [9]

Table 3: Quantitative results on DTU dataset with different training and evaluation resolution.

Coarsest Res_T	Coarsest Res_E	$Levels_E$	acc.(mm)	comp.(mm)	overall(mm)	mem.(M)	runtime(s)
40×32	25×18	6	0.372	0.292	0.332	**6809**	2.543
20×16	25×18	6	0.382	0.324	0.353		
40×32	50×37	5	0.379	**0.278**	**0.328**	7863	2.550
20×16	50×37	5	0.371	0.328	0.349		
40×32	100×74	4	0.360	0.311	0.335	6935	2.483
20×16	100×74	4	**0.349**	0.478	0.413		
40×32	200×148	3	0.375	0.530	0.452	7861	**2.366**
20×16	200×148	3	0.531	1.959	1.245		

Note: Quantitative results on DTU dataset with different training and evaluation resolution. Coarsest Res_T and Coarsest Res_E denote the coarsest image resolution in the pyramid during training and evaluation, respectively. $Levels_E$ means the number of pyramid layers during evaluation.

olution is smaller. Tab. 3 shows that the performance of the model trained with higher resolution input is better than that with lower resolution input. This is probably due to high-resolution images contain more discriminative features that are helpful for high-quality reconstruction.

To discover the relationship between pyramid levels and quality of output depth map, we also evaluate our method with different pyramid levels (the minimum level is 3) on DTU dataset. As shown in Tab. 3, we achieve the best overall score with 5 pyramid stages in evaluation process.

5 Conclusion

In this paper, we present an efficient deep-learning based cost volume pyramid network for MVS. By combining different sampling range estimation strategies for each stage, we integrate multi-dimensional information without additional neural network modules. Then, we apply adaptive unimodal filters to further improve the low-resolution depth map before refinement, and proves its effectivenes in coarse-to-fine cost volume pyramid framework. Results on different datasets show the effectiveness and generalisability of our method.

References

1. Ji, M., Gall, J., Zheng, H., Liu, Y., & Fang, L. (2017). Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2307-2315).
2. 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).
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. Xue, Y., Chen, J., Wan, W., Huang, Y., Yu, C., Li, T., & Bao, J. (2019). Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4312-4321).
5. Aanas, 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(2), 153-168.
6. 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.
7. Chen, R., Han, S., Xu, J., & Su, H. (2019). Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1538-1547).
8. Yu, Z., & Gao, S. (2020). 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* (pp. 1949-1958).
9. Yang, J., Mao, W., Alvarez, J. M., & Liu, M. (2020). Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4877-4886).
10. 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).
11. Yu, A., Guo, W., Liu, B., Chen, X., Wang, X., Cao, X., & Jiang, B. (2021). Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 448-460.
12. Zhang, J., Yao, Y., Li, S., Luo, Z., & Fang, T. (2020). Visibility-aware Multi-view Stereo Network.

arXiv:2008.07928, 2020. https://arxiv.org/abs/2008.07928, Aug. 2020.
13. Zhang, Y., Chen, Y., Bai, X., Yu, S., Yu, K., Li, Z., & Yang, K. (2020, April). Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 12926-12934).
14. Guo, X., Yang, K., Yang, W., Wang, X., & Li, H. (2019). Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3273-3282).
15. Chang, J. R., & Chen, Y. S. (2018). Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5410-5418).
16. 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).
17. Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y. T., & Lee, H. S. (2021). UASNet: Uncertainty Adaptive Sampling Network for Deep Stereo Matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6311-6319).
18. Shen, Z., Dai, Y., & Rao, Z. (2021). Cfnet: Cascade and fused cost volume for robust stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13906-13915).
19. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
20. 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).
21. Xu, Q., & Tao, W. (2020). Pvsnet: Pixelwise visibility-aware multi-view stereo network. arXiv preprint arXiv:2007.07714.
22. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., ... & Quan, L. (2020). Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1790-1799).