Pyramid Multi-view Stereo Net with Self-adaptive View Aggregation

Hongwei Yi[1*], Zizhuang Wei[1*], Mingyu Ding[2], Runze Zhang[3], Yisong Chen[1], Guoping Wang[1], and Yu-Wing Tai[4]

[1] PKU {hongweiyi, weizizhuang, chenyisong, wgp}@pku.edu.cn
[2] HKU myding@cs.hku.hk
[3] Tencent ryanrzzhang@tencent.com
[4] Kwai Inc. yuwing@gmail.com

Abstract. In this paper, we propose an effective and efficient pyramid multi-view stereo (MVS) net with self-adaptive view aggregation for accurate and complete dense point cloud reconstruction. Different from using mean square variance to generate cost volume in previous deep-learning based MVS methods, our **VA-MVSNet** incorporates the cost variances in different views with small extra memory consumption by introducing two novel self-adaptive view aggregations: pixel-wise view aggregation and voxel-wise view aggregation. To further boost the robustness and completeness of 3D point cloud reconstruction, we extend VA-MVSNet with pyramid multi-scale images input as **PVA-MVSNet**, where multi-metric constraints are leveraged to aggregate the reliable depth estimation at the coarser scale to fill in the mismatched regions at the finer scale. Experimental results show that our approach establishes a new state-of-the-art on the ***DTU*** dataset with significant improvements in the completeness and overall quality, and has strong generalization by achieving a comparable performance as the state-of-the-art methods on the ***Tanks and Temples*** benchmark. Our codebase is at https://github.com/yhw-yhw/PVAMVSNet

Keywords: Multi-view Stereo, Deep Learning, Self-adaptive View Aggregation, Multi-metric Pyramid Aggregation

1 Introduction

Multi-view Stereo (MVS) aims to recover dense 3D representation of scenes using stereo correspondences as the main cue given multiple calibrated images [28,23,32,35]. Although they have achieved great success on MVS benchmarks [1,22,31], many of them still have limitations in handling matching ambiguity and usually have a low completeness of 3D reconstruction. Recently, the deep neural network has made tremendous progress in multi-view stereo [17,43,18]. These methods learn and infer the information hardly obtained by stereo correspondences in order to handle matching ambiguity. However, they do not learn and utilize the following important information.

[*] Equal Contribution

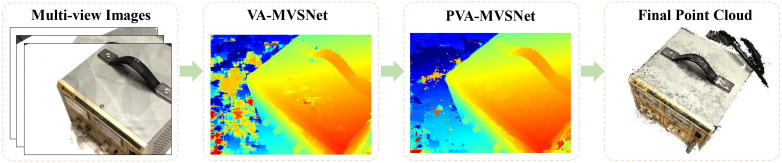

Fig. 1. VA-MVSNet performs an efficient and effective multi-view stereo with self-adaptive view aggregation to generate an accurate depth map. Cast in pyramid images additionally, PVA-MVSNet aggregates multi-scale depth maps with multi-metric constraints to boost the point cloud reconstruction with high accuracy and completeness.

First, the one-stage end-to-end deep MVS architectures [43,44,18] that directly learn from images all follow the philosophy that all view images contribute equally to the matching cost volume [13]. For instance, MVSNet [43] and R-MVSNet [44] both apply the mean square variance operation on multiple cost volumes, and DPSNet [18] selects the mean average operation. However, images from different views lead to heterogeneous image capture characteristics due to different illumination, camera geometric parameters, scene content variability, etc. Based on this observation, we propose a self-adaptive view aggregation module to learn the different significance in multiple matching volumes among images from different views. Our module benefits from the aggregated features by a self-adaptive fusion, where better element-wise matched regions are enhanced while the mismatched ones suppressed.

Second, the multi-scale information is not leveraged well to improve the robustness and completeness of 3D reconstruction. Unlike ACMM [40] where pyramid images are processed progressively to regress the depth map in a coarse-to-fine manner, we propose a novel way to aggregate multi-scale pyramid depth maps which are generated in parallel by multi-metric constraints to a refine depth map. In particular, to correct the mismatched regions at the finer depth map, we progressively aggregate the reliable depth at the coarser level to refine the finer depth map but do not introduce quantization errors benefiting from our multi-metric constraints.

To this end, we propose a novel efficient and effective pyramid multi-view stereo network with self-adaptive view aggregation, denoted as **PVA-MVSNet**. Our method constructs multi-scale pyramid images and processes them in parallel by **VA-MVSNet** to produce pyramid depth maps. To regularize 3D warping feature volumes from different views, we propose two self-adaptive element-wise view aggregation modules to learn different variance of different views in an order-independent manner. Through a depth map estimator, 3D cost volume is utilized to estimate the corresponding depth map. To further improve the robustness and completeness of 3D reconstruction generated by VA-MVSNet, our proposed multi-metric pyramid depth aggregation corrects the mismatched regions at finer depth maps using the reliable depths at coarser depth maps by checking photometric and geometric consistency.

Our main contributions are listed below:

- We propose self-adaptive view aggregation to incorporate the element-wise variances among images from different views, guiding the multiple cost volumes to aggregate a normalized one.
- We investigate to incorporate multi-scale information by our multi-metric pyramid depth maps aggregation in PVA-MVSNet, to further improve the robustness and completeness of 3D reconstruction.
- Our method establishes a new state-of-the-art on the *DTU* and a comparable performance as the state-of-the-art methods on the *Tanks and Temples*.

2 Related Work

Traditional MVS Reconstruction: Traditional MVS reconstruction algorithms can be divided into four types: voxel based [33,38], surface based [15,6], patch based [11,9] and depth map based methods [9,2,36,10,45,29]. Among those methods, the depth map based approaches are more concise and flexible. Recently, many advanced MVS algorithms estimate high quality depth maps by view selection, local propagation and multi-scale aggregation strategies. Zheng *et al.* [45] propose a depth map estimation method by solving a probabilistic graphical model. Schönberger *et al.* [29] present a new MVS system named COLMAP where geometric priors are used to better depict the probability of their graphical model. Xu *et al.* [41] propose a multi-scale MVS framework with adaptive checkerboard propagation and multi-hypothesis joint view selection to improve the performance. These works utilize predefined criteria for pixel-wise view selection, which cannot be adaptive for different scenes.

Learning Based Stereo Matching: Recently, the convolutional neural network (CNN) has made tremendous progress in many vision tasks [20,7,42,27,34], including several attempts on multi-view stereo. Early learning-based methods [17,8,19] pre-warp the images to generate plane-sweep volumes as the input. Two promising approaches [43,18] both propose the differential homography warping, which implicitly encodes multi-view camera geometries into the network and enables an end-to-end training fashion. Furthermore, R-MVSNet [44] replaces 3D-CNN in MVSNet [44] by the gated recurrent unit (GRU) to reduce memory consumption during the inference phase. Gu et al. [12] and Cheng et al. [5] both propose a cascaded MVS network through constructing coarse-to-fine cost volume which eases the memory limitation of the volume resolution in comparison with uniformly sampled cost volume [43,44,4]. P-MVSNet [24] proposes a patch-wise matching module to learn the isotropic matching confidence inside the cost volume. Particularly, those methods follow the philosophy that the feature volumes from different view images contribute equally, neglecting heterogeneous image capturing characteristics due to different illumination, camera geometric parameters and scene content variability. PointMVSNet [3] is a two-stage coarse-to-fine method which needs a coarse depth map by a lower-resolution version MVSNet [43].

Based on the above analysis, we propose a self-adaptive view aggregation module to incorporate the different significance in multiple feature volumes

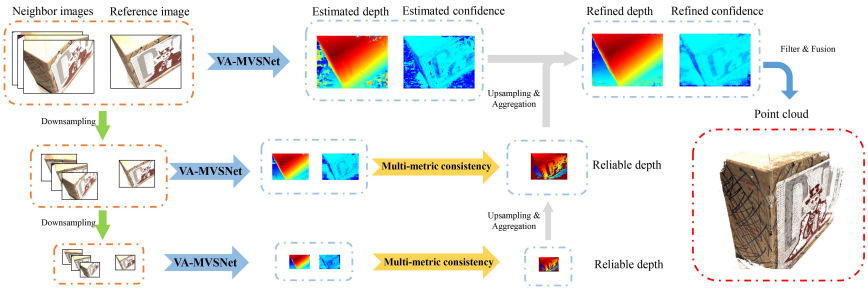

Fig. 2. Overview of PVA-MVSNet. We firstly input multi-scale pyramid images to VA-MVSNet to generate corresponding pyramid depth maps in parallel. Then we progressively replace the mismatched depths in the finer depth map with more reliable depths from a coarser level to achieve a refined depth map. Finally, we reconstruct the point cloud by filter and fusion through all estimated depth maps of the image set.

from different views, where better element-wise matched features can be enhanced while the mismatched errors can be suppressed. To further improve the robustness and completeness of 3D reconstruction point cloud, we propose a multi-metric pyramid depth aggregation to aggregate multi-scale information in pyramid images. The mismatched depth value generated by the original image can be filled-in by the reliable depth value from the downsized image under photometric and geometric consistency.

3 Method

We first describe the overall architecture of PVA-MVSNet in Sec. 3.1. Then, we introduce the details of VA-MVSNet in Sec. 3.2. Finally, we present the multi-metric pyramid depth aggregation in Sec. 3.4.

3.1 Overall

Given a reference image $\boldsymbol{I}_{i=0}$ and $\boldsymbol{I}_{i=1,\cdots,N-1}$ neighboring images and corresponding calibrated camera parameters $\boldsymbol{Q}_{i=0}$ and $\boldsymbol{Q}_{i=1,\cdots,N-1}$, where N represents the number of multi-view images, our goal is to estimate the depth map for each reference image. Afterwards, we filter and fuse all estimated depth maps to reconstruct 3D point clouds.

For the depth estimation of a reference image, our main architecture is illustrated in Fig. 2. We construct an image pyramid with K multiple scales for all images with a downsampling scale factor η. We denote k-level pyramid images and corresponding camera parameters as $I^k_{i=0,\cdots,N-1}$ and $Q^k_{i=0,\cdots,N-1}$ respectively, where $k = 0, \cdots, K-1$. The scale $k = 0$ of the pyramid represents the original image. We process each level images in the pyramid by VA-MVSNet to obtain depth maps of different scales in parallel. Then we progressively propagate the reliable depths from images with the lower resolution, which satisfy

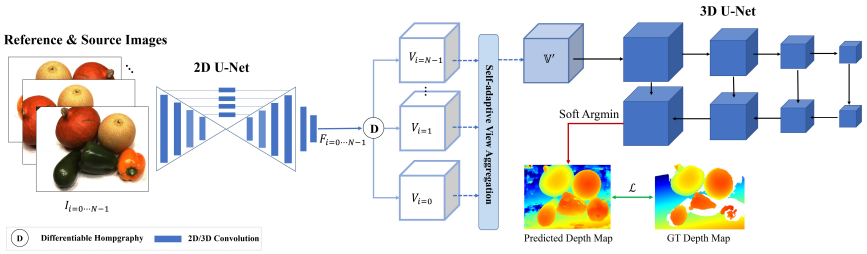

Fig. 3. The network architecture of VA-MVSNet. Multi-view images go through 2D U-Net and differentiable homography warping to generate 3D feature volumes. Cost variances in different views are encoded in self-adaptive view aggregation to aggregate the 3D cost volumes which is regularized by 3D U-Net to regress the depth map.

multi-metric constraints, to correct the mismatched errors of images with the higher resolution by replacements. Finally, we obtain the refined depth map of the raw image. We term our whole method PVA-MVSNet.

3.2 Self-adaptive View Aggregation

In VA-MVSNet in Fig. 3, we first design a 2D U-Net to extract $\{\boldsymbol{F}_i\}_{i=0}^{N-1}$ feature maps with larger receptive fields from the N input images. For efficient computation, the output feature map is downsampled by four to the original image size with 32 channels.

Then each feature map from different views will be warped to the reference camera frustum by the differential homography [43,18] with sampling D_i layers to build 3D plane-sweep feature volumes $\boldsymbol{V}_i$. To handle arbitrary N-view images input and the variances among images from different sources, we propose self-adaptive view aggregation to merge $\boldsymbol{V}_{i=0,\cdots,N-1}$ 3D feature volumes into one cost volume $\boldsymbol{C}$. Let W_i, H_i, D_i, C_i denote the width, height, depth sample number and channel number of the input 3D warping feature volume from image i respectively, the feature volume size can be represented as $S_i = W_i \cdot H_i \cdot D_i \cdot C_i$, the cost volume C_i aggregation can be defined as a function: $M : \underbrace{\mathbb{R}^{S_i} \times \cdots \times \mathbb{R}^{S_{N-1}}}_{N} \rightarrow \mathbb{R}^S$. In previous work [43,44,18], this is a constant function where all views contribute equally, which is the mean square error of all input feature volumes. However, it is not reasonable due to different illumination, camera position, occlusion and image content etc, where a near reference image with no occlusion can provide more accurate geometric and photometric information than a far one with partial occlusion. Thus, we propose to employ self-adaptive view aggregation as this function to flexibly learn the potential different view variance from training data. To achieve this goal, we develop and investigate two different self-adaptive view aggregation modules in Fig. 4, which shows how the self-adaptive view selection incorporates the variance between different views. We introduce the attention mechanism [37,39] for

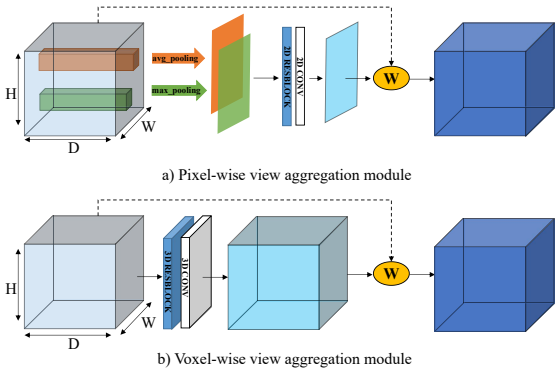

Fig. 4. Illustration of two different self-adaptive element-wise view aggregation modules, a) pixel-wise view aggregation, b) voxel-wise view aggregation.

guiding the network to select important matching information in different views. In the point-wise view selection, similar as ACMM [40], we consider that each pixel in the height and width dimension of 3D cost volume has different saliency but is consistent in the depth dimension. The voxel-wise view selection module is a 3D attention-guided mechanism to guide each voxel in 3D feature volumes to learn its own weight.

Pixel-wise View Aggregation. The pixel-wise view aggregation introduces a selective weighted attention map in the height and width dimension which considers the depth number hypothesis sharing common focusing weight. Given multi-view feature volumes $V_{i=0\cdots N-1}$, our regularized cost volumes are aggregated as $\boldsymbol{c}_{d,h,w}$:

$$\boldsymbol{v}'_{i,d,h,w} = \boldsymbol{v}_{i,d,h,w} - \boldsymbol{v}_{0,d,h,w}, \tag{1}$$

$$\boldsymbol{c}_{d,h,w} = \frac{\sum_{i=1}^{N-1}(1+\boldsymbol{w}_{h,w}) \odot \boldsymbol{v}'_{i,d,h,w}}{N-1}, \tag{2}$$

where $\boldsymbol{w}_{h,w}$ represents a 2D weighted attention map to encode the various pixel-wise saliency among images from different sources and the reference view, and $\odot$ represents element-wise multiply operation.

To generate a 2D weighted attention map, we design a *PA-Net* in Tab. 1 which consists of several 2D convolutional filters and a ResNet block [14] with the squeezing 2D features from V'_i as input to learn the $\boldsymbol{w}_{h,w}$:

$$\boldsymbol{w}_{h,w} = \textit{PA-Net}(\boldsymbol{f}_{h,w}), \tag{3}$$

$$\boldsymbol{f}_{h,w} = \textit{CONCAT}(\textit{max\_pooling}(\left\|\boldsymbol{v}'_{d,h,w}\right\|_1), \textit{avg\_pooling}(\left\|\boldsymbol{v}'_{d,h,w}\right\|_1)), \tag{4}$$

where both *max_pooling* and *avg_pooling* are used to extract the highest and average cost matching information in the *depth* dimension, and $\textit{CONCAT}(\cdot)$ denotes the concatenation operation.

Voxel-wise View Aggregation. The voxel-wise view aggregation module considers that each pixel with different depth layer hypothesis d is treated differently, where each voxel in 3D feature volume learns its own importance. Based on this, we design a *VA-Net* as shown in Tab. 1 to directly learn the 3D weighted attention map with 3D convolutional filters for selecting useful cost information. The regularized 3D cost volumes $\boldsymbol{c}_{d,h,w}$ are aggregated by $\boldsymbol{v}'_{i,d,h,w}$:

$$\boldsymbol{c}_{d,h,w} = \frac{\sum_{i=1}^{N-1}(1+\boldsymbol{w}_{d,h,w}) \odot \boldsymbol{v}'_{i,d,h,w}}{N-1}. \tag{5}$$

3.3 Depth Map Estimator

We design a 3D convolutional U-Net by leveraging different level information and expanding receptive fields to generate the probability volume P with a *softmax* operation along the depth dimension. The details of 3D U-Net are in the supplementary material.

To produce a continuous depth estimation, we use *soft argmin* operation [16] on the output probability volume P to estimate the depth $\boldsymbol{E}$:

$$\boldsymbol{E} = \sum_{d=d_{\min}}^{d_{\max}} d \times \boldsymbol{P}(d), \tag{6}$$

where $\boldsymbol{P}(d)$ denotes the estimated probability of all pixels for the depth hypothesis d. Following MVSNet [43], the probability map is calculated by the sum over the nearest four hypotheses in the 3D probability volume to measure the estimation quality. Comparing the estimated depth map and confidence map in Fig. 5 with [43], VA-MVSNet generates more reliable and accurate depth map with higher confidence benefiting from self-adaptive view aggregation.

Input	Layer	Output	Output Size
	PA-Net		
$\boldsymbol{f}_{h,w}$	ConvGR,K=3,S=1,F=16	wc_0	$W \times H \times 16$
wc_0	ResBlockGR,K=3,S=1,F=16	wres_1	$W \times H \times 16$
wres_1	Conv,K=3,S=1,F=1	wc_2	$W \times H \times 1$
wc_2	Sigmoid	weight	$W \times H \times 1$
	VA-Net		
$\boldsymbol{v}'$	Conv3DGR,K=3,S=1,F=1	wc3d_0	$D \times W \times H \times 1$
wc3d_0	Conv3D,K=3,S=1,F=1	wc3d_1	$D \times W \times H \times 1$
wc3d_1	Sigmoid	weight3d	$D \times W \times H \times 1$

Table 1. The details of *PA-Net* and *VA-Net*. We denote Conv, Conv3D as 2D and 3D convolution respectively, and use GR to represent the abbreviation of group normalization and the Relu. + and & represent the element-wise addition and concatenation. K, S, F are the kernel size, stride and output channel number. N, H, W, D denote input view number, image height, image height and depth hypothesis number.

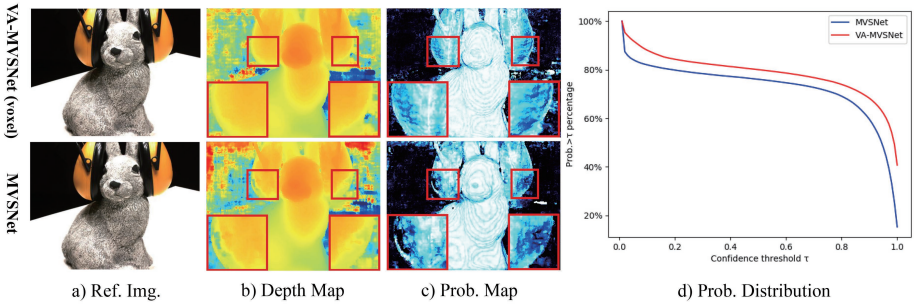

Fig. 5. Comparison on the regressed depth map, probability map and probability distribution with MVSNet [43]. (a) One reference image of Scan 12; (b) the inferred depth map; (c) the probability map; (d) the distribution of the probability map. Our self-adaptive view aggregation enhances the multi-view stereo network to generate more delicate and accurate depth estimations with higher confidence.

Training Loss We use the same training losses in MVSNet [43], which is the mean absolute error defined as $\mathcal{L}$:

$$\mathcal{L} = \sum_{\boldsymbol{x} \in \boldsymbol{x}_{valid}} \left\| \boldsymbol{d}(\boldsymbol{x}) - \hat{\boldsymbol{d}}(\boldsymbol{x}) \right\|_1, \tag{7}$$

where $\boldsymbol{x}_{valid}$ denotes the set of valid pixels in the ground truth, $\boldsymbol{d}(\boldsymbol{x})$ and $\hat{\boldsymbol{d}}(\boldsymbol{x})$ represent the estimated depth map and the ground truth respectively.

3.4 Multi-metric Pyramid Depth Map Aggregation

So far, our proposed network VA-MVSNet generates good-enough depth maps for the point cloud reconstruction. To further improve the robustness and completeness of 3D reconstruction, we propose a novel multi-metric pyramid depth aggregation to aggregate reliable depth estimations in a lower-resolution depth map into a higher-resolution depth map, by replacing corresponding mismatched errors.

In a higher-resolution fine estimated depth map, there are still some inaccurate depths with low confidences due to the matching ambiguity. Note that the same convolutional filter generally extracts less local-wise, but more global information due to a larger receptive field from a downsampled image in comparison to the original image. Quite different from ACMM [40], which casts a image pyramid into VA-MVSNet to generate multi-scale depth maps in parallel, we propose to utilize multi-metric constraints, specifically, geometric and photometric consistency to progressively replace the ambiguous depth estimations at the higher scale by reliable depths at the lower scale. As a result, we optimize both depth and probability maps in Fig. 6.

Considering a pyramid depth map $D^{k=0,\cdots,K-1}$ and a corresponding probability map $P^{k=0,\cdots,K-1}$ from VA-MVSNet, we use the photometric consistency

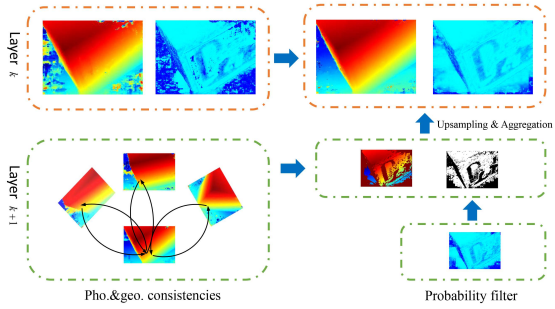

Fig. 6. Illustration of multi-metric pyramid depth map aggregation, where the reliable depth at a lower scale level $k+1$ selected by multi-metric constraints, are used to fill-in the mismatched errors at a higher scale level k by upsampling and aggregation.

to measure the matching quality through the probability map and geometric consistency to measure the depth consistency between multiple images. To select accurate and well-matched depth value in the lower scale $k+1$ depth maps, we only select the estimated depth which satisfies both the photometric and geometric consistency. Firstly, for the photometric consistency, we expect to iteratively replace unreliable depth values with low confidence $P^k(p) < \epsilon_{low}$ at the scale k by reliable depths $P^{k+1}(p) > \epsilon_{high}$ at the downsampling scale $k+1$, where $P^k(p)$ denotes the confidence of pixel p in the probability map P^k and ϵ represent the filtering confidence threshold. After discarding mis-matched errors through the photometric consistency, we project a reference pixel p of image $\mathbf{I}_i$ to the corresponding pixel p_{proj} in the neighbor image $\mathbf{I}_j$ through $D_i(p)$ and camera parameters. In turn, we reproject p_{proj} through $D_j(p_{proj})$ back to the reference image as p_{reproj} with d_{reproj}. We remain the pixeles which satisfy the following geometric constraints in at least three neighbor views:

$$\|p - p_{reproj}\|_2 < \tau_1, \tag{8}$$

$$\|D_i(p) - d_{reproj}\|_1 < \tau_2 \cdot D_i(p). \tag{9}$$

Through our multi-metric pyramid depth map aggregation, the reliable depths at a lower scale $k+1$ can be progressively propagated to replace the mismatched depths at k scale until it leads to a final refinement at $k = 0$ scale, which improves the robustness and completeness of 3D point cloud.

4 Experiments

4.1 Implementation Details

Training We train VA-MVSNet on the DTU dataset [1], which consists of 124 different indoor scenes scanned by fixed camera trajectories in 7 different lighting conditions. Following the common practices [17,19,43,44,3], we train our network

Method	Mean Distance (mm)		
	Acc.	Comp.	*overall*
Colmap [29]	0.400	0.664	0.532
Gipuma [10]	**0.283**	0.873	0.578
MVSNet [43]	0.396	0.527	0.462
R-MVSNet [44]	0.385	0.459	0.422
P-MVSNet [24]	0.406	0.434	0.420
PointMVSNet [3]	0.361	0.421	0.391
PointMVSNet-HiRes [3]	0.342	0.411	0.376
VA-MVSNet	0.378	0.359	0.369
PVA-MVSNet	0.379	**0.336**	**0.357**

Table 2. Quantitative results on the DTU evaluation dataset [1] (lower is better). Our VA-MVSNet and PVA-MVSNet (with voxel-wise view aggregation) outperform all methods in terms of completeness and overall quality with a significant improvement.

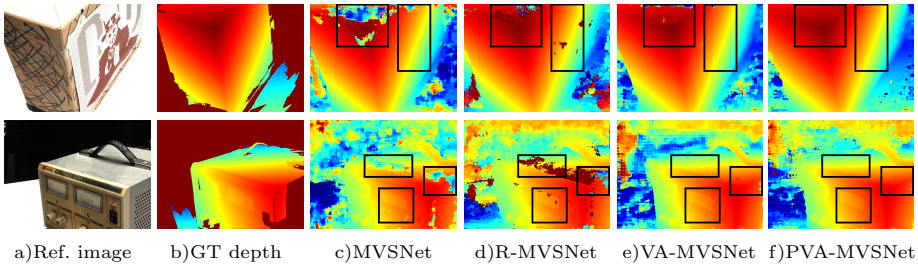

Fig. 7. Comparison of depth map estimations of *Scan* 13 *and* 11 in the *DTU* [1]. Our VA-MVSNet and PVA-MVSNet achieve more accurate, continuous and complete depth map in comparison to [43,44] methods [43,44].

on the training split and evaluate on the evaluation part and use the same depth maps provided by MVSNet [43]. During training, the input image size is set to $W \times H = 640 \times 512$ and the number of input images $N = 5$. The depth hypotheses are sampled from $425mm$ to $935mm$ with depth plane number $D = 192$ in an inverse manner as illustrated in R-MVSNet [44]. We implement our network on **PyTorch** [26] and train it end-to-end for 16 epochs using *Adam* [21] with an initial learning rate 0.001 which is decayed by 0.9 every epoch. Batch size is set to 4 on 4 NVIDIA TITANX graphics cards.

Evaluation For testing, we use $N = 7$ image views and $D = 192$ for depth plane sweeping in an inverse depth setting. We evaluate our methods on *DTU* with an original input image resolution: 1600×1184. For *Tanks and Temples*, the camera parameters are computed by OpenMVG [25] following MVSNet [43] and the input image resolution is set to 1920×1056. We use the same multi-metric constraint parameters, where $\epsilon_{low} = 0.5$, $\epsilon_{high} = 0.9$, $\tau_1 = 1$ and $\tau_2 = 0.01$.

Filtering and Fusion We fuse all depth maps into a complete point cloud as in [10,43]. In our experiments, we only consider the reliable depth values with

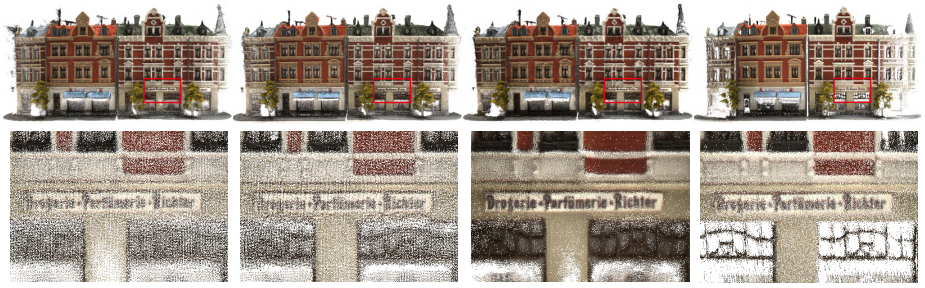

Fig. 8. Comparison of reconstruction point clouds for the model *Scan 15* in the benchmark *DTU* [1]. Our method generates denser, smoother and more complete point cloud compared with other methods [43,44].

confidence larger than $\epsilon = 0.9$ and utilize the aforementioned geometric consistency to select those pixels occurring in more than three neighbor views. Finally, the depths are projected to 3D space and fused to produce a 3D point cloud.

4.2 Benchmarks Results

DTU Dataset We evaluate our proposed method on the *DTU* [1] evaluation set. Quantitative results are shown in Tab. 2. The accuracy and completeness are calculated using the official matlab script provided by the *DTU* [1] dataset. The overall reconstruction quality is evaluated by calculating the average of the accuracy and completeness, as mentioned in [43,1]. While Gipuma [10] performs the best regarding to accuracy, our PVA-MVSNet and VA-MVSNet establish a new state-of-the-art both in completeness and overall quality with a significant margin compared with all previous methods [36,43,44,3]. We compare our depth maps with [43,44] in Fig. 7. VA-MVSNet predicts a more accurate, delicate and complete depth map by introducing different variances in multi-views through our proposed self-adaptive view aggregation. Moreover, PVA-MVSNet further fill-in the mismatched errors with reliable depths in the pyramid depth maps by our multi-metric pyramid depth map aggregation. Benefiting from more accurate, smooth and complete depth map estimation, our method can generate denser and more complete and delicate point clouds in Fig. 8.

Tanks and Temples Benchmark To explore the generalization of PVA-MVSNet, we compare our method **without any fine-tuning** with other baselines [43,44,3,4] on the *Tanks and Temples*, which is a more complicated outdoor dataset. Tab. 3 summarizes the results. The mean *f-score* increases from 43.48 to 54.46 (larger is better, date: Mar. 5, 2020) compared with MVSNet [43], which demonstrates the efficacy and strong generalization of PVA-MVSNet. Our method outperforms Point-MVSNet [3] significantly with a higher 13% mean *f-score*, which is the best baseline on *DTU* dataset. And we achieve a comparable result with

Method	Rank	Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
MVSNet [43]	52.75	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet [44]	42.62	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
Point-MVSNet [3]	40.25	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
P-MVSNet [24]	**17.00**	**55.62**	**70.04**	44.64	40.22	**65.20**	55.08	**55.17**	**60.37**	**54.29**
PVA-MVSNet	21.75	54.46	69.36	**46.80**	**46.01**	55.74	**57.23**	54.75	56.70	49.06

Table 3. Quantitative Results on the *Tanks and Temples* benchmark [22]. The evaluation metric is *f-score* which higher is better. (L.H. and P.G. are the abbreviations of *Lighthouse* and *Playground* dataset respectively.)

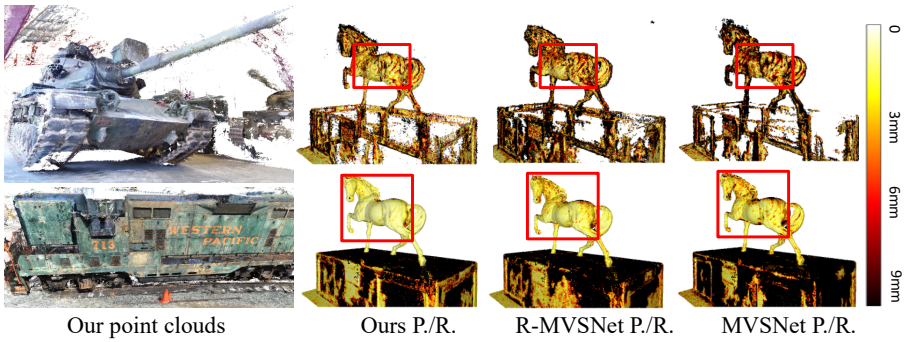

Fig. 9. The visualization of our partial point cloud results and the comparison with [43,44] on the Precision and Recall of *Horse* dataset on the *Tanks and Temples* [22] benchmark. The darker means the bigger error.

P-MVSNet [24]. The simple fusion process we adopted achieves a comparable result with P-MVSNet [24], which uses an extra *refinenet* and more depth filtering process to pursue better performance. Our partial reconstructed point clouds are shown in Fig. 9, and we compare the Precision and Recall of the *Horse* dataset with [43,44], which is provided by the *Tanks and Temples* [22] benchmark. Our method generates more accurate and complete point clouds with higher precision and recall than the others [43,44], due to the enhanced accuracy from self-adaptive view aggregation and the increased completeness and robustness from our multi-metric pyramid depth map aggregation.

4.3 Ablation Studies

In this section, we provide ablation experiments to analyze the strengths of the key components of our architecture. For following studies, to eliminate the non-learning influence, all experiments use the same consistency-check parameters in Sec. 4.1 and are tested on the *evaluation* and *validation DTU* [1] dataset.

Self-adaptive View Aggregation As shown in Tab. 4, compared with our baseline method which is using the same mean square error as cost volume aggregation

Fig. 10. Validation results of the mean average depth error with different components in VA-MVSNet during training.

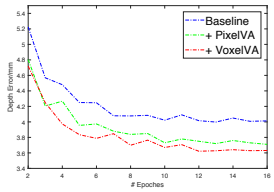

Table 4. Contributions of different components in our architecture on the evaluation DTU [1].

Components	Acc.	Comp.	*Overall*
baseline	0.454	0.372	0.413
+PixelVA	0.390	0.369	0.379
+VoxelVA	0.378	0.359	0.369
+PixelVA+MMP	0.392	0.341	0.366
+VoxelVA+MMP	**0.379**	**0.336**	**0.357**

Number of views		Number of	Acc.	Comp.	*Overall*
training	test	pyramid	(mm)	(mm)	(mm)
$N = 3$	$N = 2$	\	0.415	0.467	0.441
$N = 3$	$N = 3$	\	0.380	0.379	0.380
$N = 3$	$N = 5$	\	0.381	0.361	0.371
$N = 3$	$N = 7$	\	0.380	0.361	0.370
$N = 4$	$N = 7$	\	0.380	0.359	0.370
$N = 5$	$N = 7$	\	0.378	0.359	0.369
$N = 5$	$N = 7$	$K = 1$	**0.372**	0.350	0.361
$N = 5$	$N = 7$	$K = 2$	0.378	0.341	0.360
$N = 5$	$N = 7$	$K = 3$	0.379	**0.336**	**0.357**

Table 5. Ablation study on different number of views N in training and testing phase and different numbers of image pyramid on *DTU* [1] evaluation dataset.

in MVSNet [43], both PixelVA and VoxelVA can improve the results of 3D reconstruction point cloud with a significant margin, especially on the accuracy of reconstruction quality. Specifically, the VoxelVA provides a 16.7% increase on *accuracy*, which is better than the PixelVA 14.1% due to the learning variance of the depth wise hypothesis. Besides, the VoxelVA has more parameters but less operations compared with PixelVA as denoted in Tab. 1. During training, as shown in Fig. 10, the depth error on *validation* dataset drops significantly by introducing our proposed novel self-adaptive view aggregation.

Number of Views We investigate the influence of variant numbers of views N in different phases on *DTU evaluation* dataset. VA-MVSNet can process an arbitrary number of views and well leverage the variant importance in multi-views due to our proposed self-adaptive view aggregation. In the test phase, we use the model trained on 3 views to compare the reconstruction results with different numbers of views $N = 2, 3, 5, 7$. As shown in Tab. 5, the result with $N = 5$ achieves a great improvement compared with $N = 2, 3$, but the influence from two more extra views in $N = 7$ is quite small which can be ignored. It demonstrates that our proposed self-adaptive view aggregation can well enhance the valid information in the good neighbor views and eliminate bad information in farrer views (the neighbor views are ranked by the matching quality with the reference view in SfM [30]). In the training phase, we compare the results on the input view $N = 7$ using the models trained on $N = 3, 4, 5$. The model trained on $N = 5$ is slightly better than $N = 3$ but with more training time.

Methods	H,W,D	Mem.	Time.	Overall
MVSNet	1600, 1184, 256	15.4GB	1.18s	0.462
R-MVSNet	1600, 1184, 512	**6.7GB**	2.35s	0.422
PointMVSNet	1280, 960, 96	7.2GB	1.69s	0.391
PointMVSNet-HiRes	1600, 1152, 96	8.7GB	5.44s	0.376
VA-MVSNet	1600, 1184, 192	18.1GB	**0.91s**	0.369
PVA-MVSNet	1600, 1184, 192	24.87GB	1.01s	**0.357**

Table 6. Comparisons on the time and memory cost on the *evaluation* DTU [1] dataset. MVSNet and R-MVSNet are implemented in TensorFlow while others in PyTorch.

Multi-metric Pyramid Depth Aggregation As shown in Tab. 4, The completeness can be averagely improved by 7.0% while introduce a negligible drop 0.39%, benefiting from "MMP" (multi-metric pyramid depth aggregation). As denoted in *e)* and *f)* in Fig. 7, PVA-MVSNet improves VA-MVSNet by generating more delicate and complete depth maps. To better analyse the improvement from different pyramid level image, we explore the influence by different numbers of image pyramid in Tab. 5. The $K = 1$ level pyramid image improves both accuracy and completeness with a big margin. A trade-off between accuracy and completeness is achieved by using more pyramid images $k = 2$ and $k = 3$, it leads to reconstructed 3D point cloud with better overall quality.

4.4 Runtime and Memory Performance

Given time and memory performance in Tab. 6, all methods are tested on GeForce RTX 2080 Ti. VA-MVSNet runs fast at a speed of $0.91s$ / view, even if it runs with the biggest memory consumption. Unlike PointMVSNet [3], multi-scale pyramid images can be processed independently in parallel. Therefore, with little extra time about $0.1s$ for multi-metric pyramid depth aggregation, the performance of 3D point cloud reconstruction increases significantly from 0.369 to 0.357 in PVA-MVSNet on *DTU* [1] dataset.

5 Conclusion

We present a novel pyramid multi-view stereo network with the self-adaptive view aggregation. The proposed VA-MVSNet dynamically selects the element-wise feature importance while suppresses the mismatching cost, which is quite efficient and effective. Casting in multi-scale pyramid images, benefiting from utilizing multi-metric constraint, PVA-MVSNet estimates a refined depth map for further improving the robustness and completeness of 3D reconstruction. Experimental results demonstrate that our proposed method PVA-MVSNet establishes a new state-of-the-art on the *DTU* dataset and shows great generalization by achieving a comparable performance as other state-of-the-art methods on *Tanks and Temples* benchmark without any fine-tuning.

Acknowledgements This project was supported by the National Key R&D Program of China (No.2017YFB1002705, No.2017YFB1002601) and NSFC of China (No.61632003, No.61661146002, No.61872398).

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. IJCV **120**(2), 153–168 (2016)
2. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: ECCV (2008)
3. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: ICCV (2019)
4. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. arXiv preprint arXiv:1908.04422 (2019)
5. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: CVPR (2020)
6. Cremers, D., Kolev, K.: Multiview stereo and silhouette consistency via convex functionals over convex domains. PAMI **33**(6), 1161–1174 (2010)
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (2015)
8. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. In: CVPR (2016)
9. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. PAMI **32**(8), 1362–1376 (2009)
10. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV (2015)
11. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: ICCV (2007)
12. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR (2020)
13. Hartmann, W., Galliani, S., Havlena, M., Van Gool, L., Schindler, K.: Learned multi-patch similarity. In: ICCV (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Hiep, V.H., Keriven, R., Labatut, P., Pons, J.P.: Towards high-resolution large-scale multi-view stereo. In: CVPR (2009)
16. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: CVPR (2018)
17. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR (2018)
18. Im, S., Jeon, H.G., Lin, S., Kweon, I.S.: Dpsnet: End-to-end deep plane sweep stereo. In: ICLR (2019)
19. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In: ICCV (2017)
20. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV (2017)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)
22. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. TOG **36**(4), 78 (2017)

23. Lhuillier, M., Quan, L.: A quasi-dense approach to surface reconstruction from uncalibrated images. PAMI **27**(3), 418–433 (2005)
24. Luo, K., Guan, T., Ju, L., Huang, H., Luo, Y.: P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In: ICCV (2019)
25. Moulon, P., Monasse, P., Marlet, R., et al.: Openmvg. an open multiple view geometry library (2014)
26. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NeurIPS Autodiff Workshop (2017)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
28. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
29. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016)
30. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
31. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
32. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
33. Sinha, S.N., Mordohai, P., Pollefeys, M.: Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In: ICCV (2007)
34. Song, X., Zhao, X., Hu, H., Fang, L.: Edgestereo: A context integrated residual pyramid network for stereo matching. In: ACCV (2018)
35. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR (2008)
36. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications **23**(5), 903–920 (2012)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
38. Vogiatzis, G., Esteban, C.H., Torr, P.H., Cipolla, R.: Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. PAMI **29**(12), 2241–2246 (2007)
39. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
40. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: CVPR (2019)
41. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: CVPR (2019)
42. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: ECCV (2018)
43. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
44. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: CVPR (2019)
45. Zheng, E., Dunn, E., Jojic, V., Frahm, J.M.: Patchmatch based joint view selection and depthmap estimation. In: CVPR (2014)