AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network

Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen and Guoping Wang*
Peking University
{weizizhuang,wgp}@pku.edu.cn

Abstract

In this paper, we present a novel recurrent multi-view stereo network based on long short-term memory (LSTM) with adaptive aggregation, namely AA-RMVSNet. We firstly introduce an intra-view aggregation module to adaptively extract image features by using context-aware convolution and multi-scale aggregation, which efficiently improves the performance on challenging regions, such as thin objects and large low-textured surfaces. To overcome the difficulty of varying occlusion in complex scenes, we propose an inter-view cost volume aggregation module for adaptive pixel-wise view aggregation, which is able to preserve better-matched pairs among all views. The two proposed adaptive aggregation modules are lightweight, effective and complementary regarding improving the accuracy and completeness of 3D reconstruction. Instead of conventional 3D CNNs, we utilize a hybrid network with recurrent structure for cost volume regularization, which allows high-resolution reconstruction and finer hypothetical plane sweep. The proposed network is trained end-to-end and achieves excellent performance on various datasets. It ranks 1^{st} among all submissions on Tanks and Temples benchmark and achieves competitive results on DTU dataset, which exhibits strong generalizability and robustness. Implementation of our method is available at https://github.com/QT-Zhu/AA-RMVSNet.

1. Introduction

Multi-view stereo (MVS) aims to obtain 3D dense models of real-world scenes from multiple images, which is one of the core techniques in a variety of applications including virtual reality, autonomous driving and heritage conservation. While traditional MVS methods [5, 9, 10, 26, 4] utilize hand-crafted matching metrics to measure multi-view consistency, recent deep-learning-based methods [33, 34, 21, 37, 31] achieve superior accuracy and completeness on many MVS benchmarks [3, 16, 35, 27] compared with the previous state-of-the-arts, through introducing convolutional neural network (CNN) which makes feature extraction and cost volume regularization more powerful. However, some challenging problems still remain to be solved to further improve reconstruction quality.

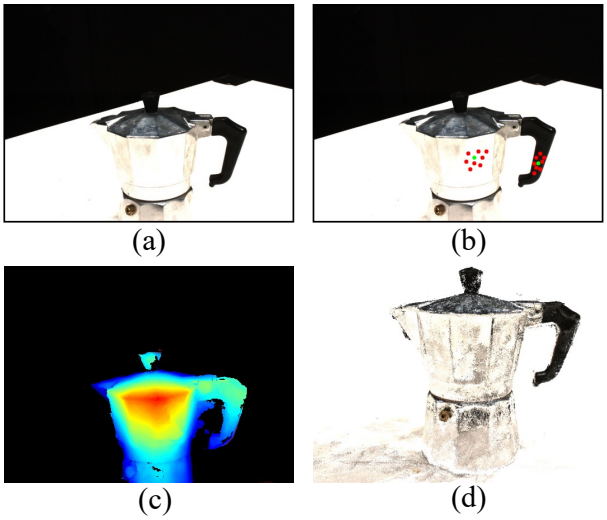


Figure 1. Illustration of multi-view 3D reconstruction of Scan 77 in DTU dataset [3] using the proposed AA-RMVSNet. (a) The reference image; (b) adaptive sampling locations in our intra-view AA approach; (c) the depth map estimated by AA-RMVSNet after filtering; (d) the recovered dense 3D model.

First, general features extracted by 2D CNN in regular pixel grids with fixed receptive fields often have difficulties in handling thin structures or textureless surfaces, which limits the robustness and completeness of 3D reconstruction. Recent MVSNet-based attempts [36, 31, 32] introduce multi-scale information to improve depth estimation. However, context-aware features have not been leveraged well enough for varying richness of texture on different regions.

Second, few works consider pixel-wise visibility issues during multi-view matching cost aggregation, which inevitably deteriorates the final reconstruction quality, especially under severe occlusion. In order to select well-captured views for each pixel, Vis-MVSNet [37] uses

*Corresponding author.

pair-wise matching uncertainties as weighting guidance to attenuate pixels that have difficulties to match. PVA-MVSNet [36] contains a CNN-based voxel-wise view aggregation module to guide multiple cost volume aggregation. However, it is hard to give a perfect solution for the occlusion problem in general case.

Moreover, in order to meet the needs of various real-world applications, memory consumption is also essential for a scalable MVS algorithm. Instead of using 3D CNN, some recent methods [34, 31] apply recurrent convolution structure for cost volume regularization, which is effective and memory efficient to reconstruct scenes with wide ranges of depth.

To tackle the aforementioned problems, we therefore present a novel long short-term memory (LSTM) based recurrent multi-view stereo network with both intra-view and inter-view adaptive aggregation modules, namely AA-RMVSNet. The intra-view scheme is designed for robust feature extraction, where context-aware features are adaptively aggregated for multiple scales and regions with varying richness of texture; the inter-view scheme is used at multi-view cost volume aggregation step, whose aim is to overcome the difficulty of varying occlusion in complex scenarios by allocating higher weights on the well-matched view pairs. As a result, the proposed network is able to obtain accurate and complete depth maps to further generate high quality dense point clouds, as illustrated in Fig. 1.

The main contributions of this work are listed below:

- We introduce an intra-view feature aggregation module to adaptively extract image features by using deformable convolution and multi-scale aggregation.
- We propose an inter-view cost volume aggregation module to adaptively aggregate cost volumes of different views by yielding pixel-wise attention maps for each view.
- Our method ranks 1^{st} among all submissions on Tanks and Temples online benchmark and obtains competitive results on DTU dataset.

2. Related Work

2.1. Traditional MVS

According to output scene representations, traditional MVS reconstruction methods can be categorized into three types: volumetric [18, 28], point-based [19, 9] and depth-based [5, 10, 26, 29]. Volumetric methods first discretize the whole 3D space into regular cubes and then decide whether a voxel belongs to the surface or not with the photometric consistency metric. The space discretization is memory intensive, thus these methods are not scalable to large-scale scenarios. Point-based methods focus on the 3D points, usually start from a sparse set of matched key points and use the propagation strategy to gradually densify the reconstruction, which limits the capacity of parallel data processing. In contrast, depth-based methods have shown more flexibility in modeling the 3D geometry of scene. It reduces the MVS reconstruction into relatively small problems of per-view depth map estimation, and can be further fused to point cloud [22] or the volumetric reconstructions [23]. Many successful traditional MVS algorithms yielding depth maps have been proposed. Schönberger *et al*. present COLMAP [26], which uses hand-crafted features and jointly estimates pixel-wise view selection, depth map and surface normal to utilize the photometric and geometric priors. Xu *et al*. propose ACMM [29] with multi-scale geometric consistency, adaptive checkerboard sampling and multi-hypothesis joint view selection. Although traditional MVS methods yield impressive results, they utilize hand-crafted features which are not suitable for non-Lambertian surfaces, low-textured and texture-less regions where photometric consistency is unreliable.

2.2. Learning-based MVS

Rather than using traditional hand-crafted image features, recent studies on MVS apply deep learning for better reconstruction accuracy and completeness. Volumetric methods SurfaceNet [12] and LSM [13] are first proposed. They construct a cost volume using multi-view images and use 3D CNNs to regularize and infer the voxel. However, SurfaceNet and LSM are restricted to only small-scale reconstructions due to the common drawback of the volumetric representation. Depth-based method MVSNet [33] improves the MVS reconstruction performance a lot compared with SurfaceNet and LSM. MVSNet takes one reference image and several source images as input and extracts deep image features, then encodes camera geometries in the network to build the 3D cost volumes via differentiable homography. To reduce the huge memory consumption of MVSNet, some variants of MVSNet have been proposed recently and can be divided into: multi-stage methods and recurrent methods. Multi-stage methods, such as CasMVSNet [11], CVP-MVSNet [32], UCS-Net [7], Vis-MVSNet [37], use the coarse-to-fine strategy that first predict a low resolution depth map with large depth interval and iteratively upsamples and refines the depth map with a narrow depth range. Though the coarse-to-fine architectures successfully reduce memory consumption, they are not suitable for high-resolution depth reconstructions as the depth prediction of coarse stage may be wrong with a large depth interval. To this end, recurrent methods, such as R-MVSNet [34] and D^2HC-RMVSNet [31], are proposed. They sequentially regularize cost maps along the depth dimension with recurrent networks to avoid using memory-intensive 3D CNNs; thus they can infer depth maps within a very large depth range. R-MVSNet regularizes cost volumes in a sequen-

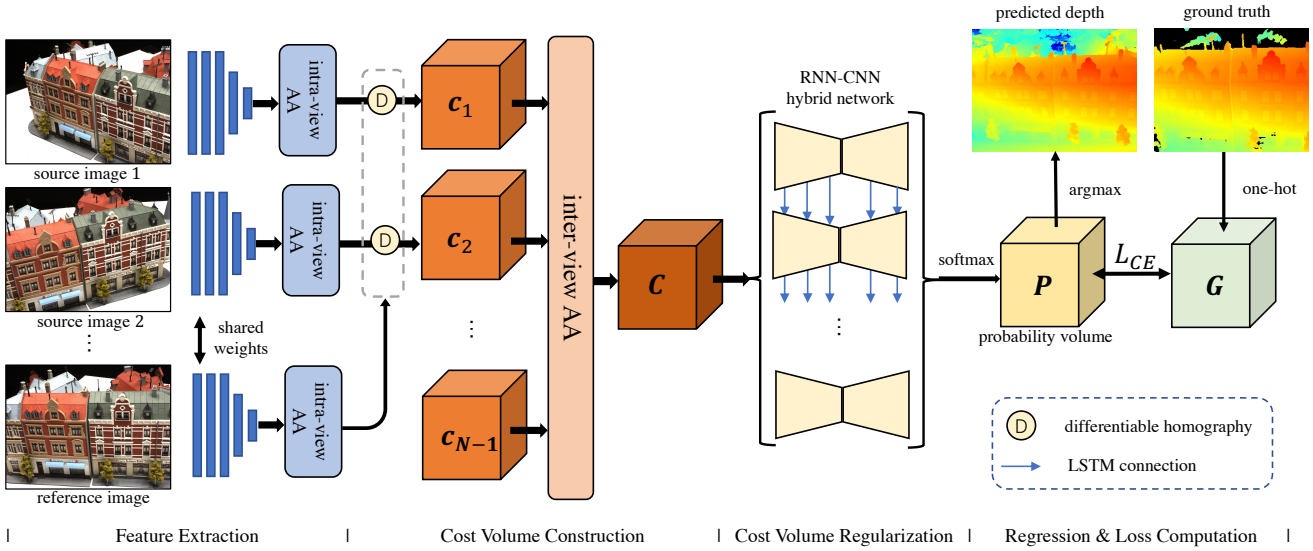


Figure 2. Overall architecture of AA-RMVSNet that consists of 4 stages. Intra-view AA module aims to aggregate context-aware features for multiple scales and regions with varying richness of texture. Inter-view AA module adaptively aggregates cost volumes of different views by yielding pixel-wise attention maps for each view. A RNN-CNN hybrid network is adopted to regularize cost volumes in a recurrent slice-by-slice pattern. At last, cross entropy for pixel-wise classification is adopted to calculate the loss for back propagation.

tial manner using convolutional gated recurrent unit (GRU). D^2HC-RMVSNet improves R-MVSNet with more powerful recurrent convolutional cells, ConvLSTMCells, and a dynamic consistency checking strategy.

Though achieving promising results, most of the aforementioned learning-based MVS methods still have difficulties in handling challenging regions and severe occlusion problems in MVS.

3. Methodology

The overall architecture of our AA-RMVSNet follows the typical pattern of a learning-based MVS pipeline, which is illustrated in Fig. 2. Input images are separated into 1 reference image and $N-1$ source images. Image features ($H \times W \times F$) of all N images are extracted by an encoder with shared weights and a 3D cost volume ($H \times W \times D \times F$) is constructed via differentiable homography by warping features of source images to the reference camera frustum. Then the cost volume is regularized to obtain a probability volume $H \times W \times D$ which generates the prediction of depth map. Feature maps of all images are filtered and fused to obtain the dense point cloud of the scene.

Particularly for AA-RMVSNet, per-view matching cost volumes are computed by matching features of the $N-1$ warped source images and the reference image with D depth hypotheses. The pixel-wise mapping relation between the reference image and the i-th source image with depth hypothesis d is described by the differentiable homography as

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

where $\mathbf{T}$ and $\mathbf{K}$ denote camera extrinsics and intrinsics respectively. Then per-view cost volumes are calculated by

$$\mathbf{c}_i^{(d)} = (\mathbf{f}_{src_i}^{(d)} - \mathbf{f}_{ref})^2, \tag{2}$$

where $\mathbf{f}_{src_i}^{(d)}$ represents extracted features of the i-th source image and $\mathbf{f}_{ref}$ represents features of the reference image. All $N-1$ cost volumes are aggregated and cost volume regularization is then carried out by a hybrid neural network to obtain depth maps and corresponding probability distribution.

AA-RMVSNet further improves the pipeline by leveraging the idea of adaptive aggregation at two stages, namely intra-view adaptive aggregation (intra-view AA) at feature extraction (Sec. 3.1) and inter-view adaptive aggregation (inter-view AA) at cost volume construction (Sec. 3.2). Besides, a RNN-CNN hybrid neural network (Sec. 3.3), which is memory efficient and robust for varying scenes, is adopted to commit cost volume regularization recurrently.

3.1. Intra-view Adaptive Aggregation

As have been covered, 3D cost volumes are constructed by matching 2D feature maps so extracting recognizable and reliable features is of great significance in MVS. As for 3D reconstruction, it is universally acknowledged that reflective surfaces and low-textured or texture-less areas are

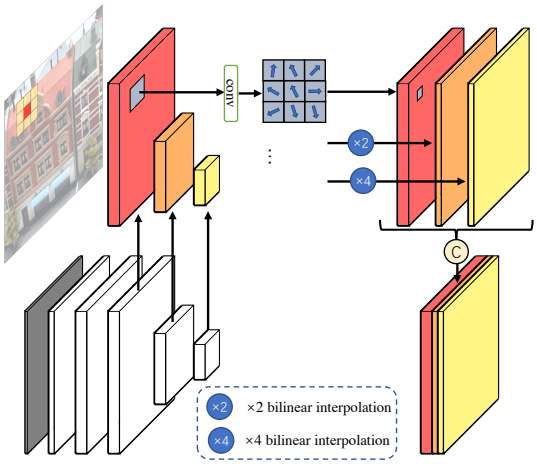


Figure 3. Intra-view AA module. All convolution kernels are 3×3. Feature channels at the encoder (colored white) are 8, 16, 16, 16, 16. Multi-scale feature maps are sent into three deformable convolutions respectively, whose parameters are not shared. By bilinear interpolation and concatenation, a feature map of $H \times W \times (16+8+8)$ is built.

main difficulties for a common CNN to handle which is operated on regular 2D grids with fixed receptive fields. For those challenging regions that are generally lacking in texture, we expect the receptive fields of convolutions to be larger while smaller receptive fields are favored for regions with rich texture. We introduce an intra-view AA module illustrated as Fig. 3 for adaptive aggregating features of different scales and regions with varying richness of texture. In the intra-view AA module, 3 feature maps of different spatial scales, whose sizes are $H \times W \times 16$, $\frac{H}{2} \times \frac{W}{2} \times 16$ and $\frac{H}{4} \times \frac{W}{4} \times 16$ respectively, are processed by 3 one-stride deformable convolutions [8, 38] with exclusive parameters. The definition of a deformable convolution is defined as

$$\mathbf{f}'(\mathbf{p}) = \sum_k w_k \cdot \mathbf{f}(\mathbf{p} + \mathbf{p}_k + \Delta\mathbf{p}_k) \cdot \Delta m_k, \qquad (3)$$

where $\mathbf{f}(\mathbf{p})$ denotes the feature value pixel $\mathbf{p}$; w_k and $\mathbf{p}_k$ represent the kernel parameter and fixed offset defined in a common convolution operation; $\Delta\mathbf{p}_k$ and Δm_k are the offset and modulation weight yielded adaptively by learnable sub-networks of deformable convolution. By interpolating smaller feature maps to $H \times W$, we obtain 3 feature maps with 16, 8, 8 channels respectively and these features are concatenated to construct a feature map of $H \times W \times 32$.

3.2. Inter-view Adaptive Aggregation

After per-view cost volumes have been constructed, the next step is to aggregate all cost volumes into one for regularization.

A common practice is to average $N-1$ cost volumes, whose underlying principle is that all views should be of equal importance. However, this is not reasonable enough since varying shooting angles may lead to problems such as occlusion and different lighting conditions of non-Lambertian surfaces that make depth estimation more difficult.

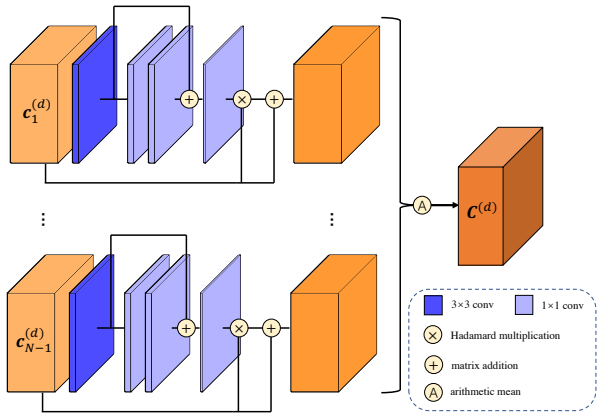


Figure 4. Inter-view AA module. For an input cost volume of $H \times W \times 32$, the following intermediate channel numbers are 4, 4, 4, 1. After reweighted by a $H \times W \times 1$ attention map, all cost volumes are summed and divided by $N-1$.

Therefore, as illustrated in Fig. 4, we design an inter-view AA module to handle unreliable matching costs, which is defined as

$$\mathbf{C}^{(d)} = \frac{1}{N-1} \sum_{i=1}^{N-1} [1 + \boldsymbol{\omega}(\mathbf{c}_i^{(d)})] \odot \mathbf{c}_i^{(d)}, \qquad (4)$$

where $\odot$ denotes Hadamard multiplication and $\boldsymbol{\omega}(\cdot)$ is pixel-wise attention maps adaptively yielded according to per-view cost volumes. In this way, pixels that are likely to be confusing for matching will be suppressed while those with crucial context information will be assigned with larger weights. $1+\boldsymbol{\omega}(\cdot)$ better prevents over-smoothness than $\boldsymbol{\omega}(\cdot)$ alone.

3.3. Recurrent Cost Regularization

Cost regularization is to leverage spatial context information and to turn matching costs into a probability distribution of D depth hypotheses. The regularization network adopts a RNN-CNN hybrid fashion where a cost volume ($H \times W \times D \times 32$) is sliced at the dimension of D. As is illustrated in Fig. 2, feature passing in the regularization network has both horizontal direction and vertical direction. Horizontally, each slice of 3D cost volume is regularized by a CNN with encoder-decoder architecture; on the vertical direction, there are 5 parallel RNNs to deliver intermediate outputs of former ConvLSTMCells to later ones.

Considering a cost volume slice of depth hypothesis d to be processed by the j-th convolution layer, denoted as

$\mathbf{v}_{j-1}^{(d)}$, the output of this layer with depth hypothesis $d-1$ as $\mathbf{v}_{j}^{(d-1)}$ and memory maintained (or hidden state) as $\mathbf{m}_{j}^{(d-1)}$, operations within a ConvLSTMCell are as follows.

Firstly, $\mathbf{v}_{j-1}^{(d)}$ and $\mathbf{v}_{j}^{(d-1)}$ are concatenated and after being processed by a convolution layer, the tensor is split into 4 tensors from the feature dimension, namely $\mathbf{w}$, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. The 4 signals within a LSTM cell are defined as

$$\begin{cases} \mathbf{i} = \sigma(\mathbf{w}) \\ \mathbf{f} = \sigma(\mathbf{x}) \\ \mathbf{o} = \sigma(\mathbf{y}) \\ \mathbf{g} = \tanh(\mathbf{z}) \end{cases} \tag{5}$$

where all signals are two-dimensional in space and Sigmoid function $\sigma(\cdot)$ and hyperbolic tangent function $\tanh(\cdot)$ are all element-wise operations. Then the memory of LSTM is updated by

$$\mathbf{m}_{j}^{(d)} = \mathbf{m}_{j}^{(d-1)} \odot \mathbf{f} + \mathbf{i} \odot \mathbf{g}, \tag{6}$$

while the output of the cell is

$$\mathbf{v}_{j}^{(d)} = \mathbf{o} \odot \tanh(\mathbf{m}_{j}^{(d)}). \tag{7}$$

3.4. Loss Function

Since cost volume regularization turns matching costs into a pixel-wise probability distribution of depth hypothesis, the task of depth estimation is now similar to a pixel-wise classification problem. Therefore, by encoding the ground truth with one-hot pattern, we adopt cross entropy to calculate the training loss, defined as

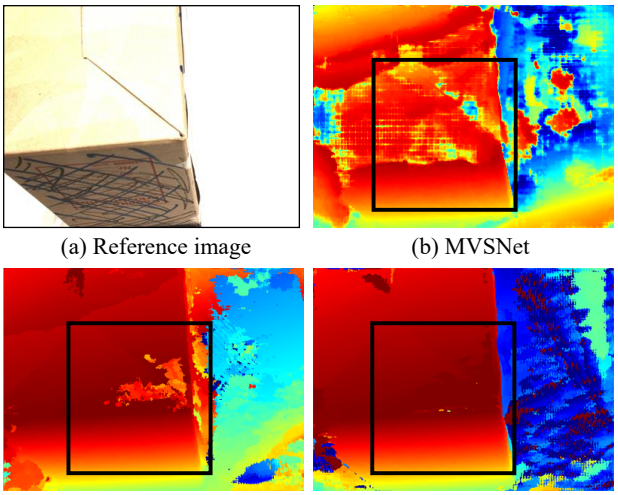
$$L = \sum_{\mathbf{p} \in \{\mathbf{p}_v\}} \sum_{d=d_0}^{d_{D-1}} -G^{(d)}(\mathbf{p}) \ \log[P^{(d)}(\mathbf{p})], \tag{8}$$

where $G^{(d)}(\mathbf{p})$ and $P^{(d)}(\mathbf{p})$ denote ground truth probability and predicted probability of depth hypothesis d at pixel $\mathbf{p}$. $\{\mathbf{p}_v\}$ is the set of valid pixels with reliable depth.

4. Experiments

4.1. Datasets

DTU dataset DTU dataset [3] is an indoor MVS dataset collected under well-controlled laboratory conditions with fixed camera trajectory. It contains 128 scans with 49 views under 7 different lighting conditions and is split into 79 training scans, 18 validation scans and 22 evaluation scans. By setting each image as reference, there are 27097 training samples in total. Following common configurations, we apply DTU dataset for network training and evaluation.



(a) Reference image (b) MVSNet

(c) D²HC-RMVSNet (d) **AA-RMVSNet**

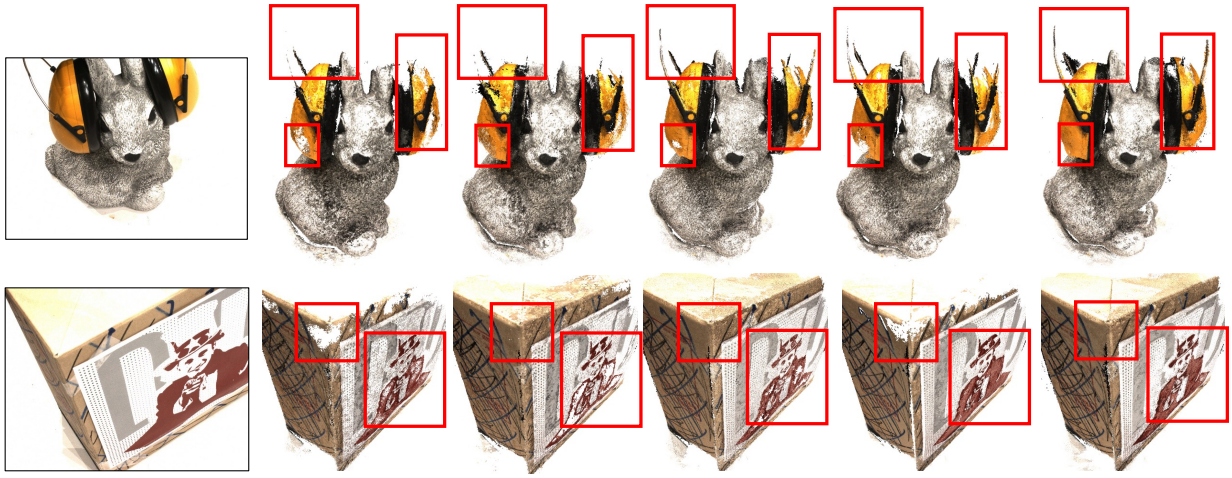
Figure 5. Comparison of depth map estimation of Scan 13 in DTU evaluation set [3]. Our AA-RMVSNet's prediction is more accurate, continuous and complete in contrast to [33, 31].

BlendedMVS dataset BlendedMVS dataset [35] is a recently published large-scale synthetic dataset for multi-view stereo training containing a variety of indoor and outdoor scenes, such as cities, architectures, sculptures and shoes. The dataset consists of over 17k high-resolution images and is split into 106 training scenes and 7 validation scenes. However, this dataset does not officially provide evaluation tools, so we utilize BlendedMVS dataset for network fine-tuning and qualitative evaluation.

Tanks and Temples benchmark Tanks and Temples [16] is a large-scale outdoor benchmark captured in more complex real scenarios. It contains an intermediate set and an advanced set. Specifically, the intermediate set has eight scenes: Family, Francis, Horse, Lighthouse, M60, Panther, Playground and Train. Different scenes have different scales, surface reflection and exposure conditions. Evaluation of Tanks and Temples benchmark is done online by uploading reconstructed point clouds to its official website [2]. Until now, there have been hundreds of submissions on Tanks and Temples leaderboard including almost all recent state-of-the-art methods.

4.2. Implementation Details

Training We train our AA-RMVSNet on DTU training set [3] consisting of 79 different scenes. Since DTU dataset only provides laser ground truth point clouds, in order to obtain ground truth depth maps for network training, we follow the previous MVS methods [33, 34, 37, 11] to generate coarse ground truth depth maps by screened Poisson surface reconstruction algorithm [14] and depth rendering. After that, we improve the reliability of original depth maps by



(a) Reference image (b) MVSNet (c) PVA-MVSNet (d) R-MVSNet (e) D²HC-RMVSNet (f) **Ours**

Figure 6. Qualitative comparisons with [33, 36, 34, 31] of Scan 33 and Scan 13 in DTU dataset [3]. Our method provides more complete 3D dense point clouds with details preserved.

Method	Acc.(mm)	Comp.(mm)	*Overall*(mm)
Furu [9]	0.613	0.941	0.777
Gipuma [10]	**0.283**	0.873	0.578
COLMAP [26]	0.400	0.664	0.532
MVSNet [33]	0.396	0.527	0.462
R-MVSNet [34]	0.385	0.459	0.422
P-MVSNet [20]	0.406	0.434	0.420
PointMVSNet [6]	0.361	0.421	0.391
D^2HC-RMVSNet [31]	0.395	0.378	0.386
PointMVSNet [6]	0.342	0.411	0.376
Vis-MVSNet [37]	0.369	0.361	0.365
CasMVSNet [11]	0.325	0.385	0.355
CVP-MVSNet [32]	0.296	0.406	**0.351**
AA-RMVSNet	0.376	**0.339**	0.357

Table 1. Quantitative results on DTU evaluation set [3] (lower is better). Our method AA-RMVSNet exhibits a competitive *overall* score compared with other state-of-the-art methods. Specially, our method outperforms all methods mentioned in terms of *completeness*.

cross-filtering with their neighboring views, which is similar to [21]. We resize the original images to the size of $W \times H = 160 \times 128$ which is equal to the resolution of the refined ground truth depth maps. The number of input images is set to $N = 7$ while the number of depth hypotheses is set to $D = 192$, which is uniformly sampled from $425mm$ to $935mm$. We implement our AA-RMVSNet by PyTorch [24] and train the proposed network end-to-end using Adam [15] with an initial learning rate of 0.001, which decays by 0.9 each epoch. The total training phase costs 20.16GB memory and takes about 3 days. Batch size is set to 4 on 4 NVIDIA TITAN RTX GPUs.

Testing Since the training phase needs extra memory to save intermediate gradients for back propagation, the testing phase of AA-RMVSNet is relatively memory efficient so that it could deal with higher resolution images and finer depth plane sweep. We set $N = 7$ and $D = 512$ in the testing phase to obtain depth maps with finer details. In order to fit the network, the height and width of input images must be a multiple of 8. We use input images of 800×600 resolution for DTU evaluation. Before testing on BlendedMVS, we fine-tune our network on the training set of BlendedMVS to boost the performance of various scenarios. We test our network on the validation set of BlendedMVS using original images of 768×576 with inverse depth setting. For benchmarking on Tanks and Temples, we apply COLMAP-SfM [25] to estimate depth ranges and camera parameters. Different from the image cropping methods in [33, 34, 36, 31, 11], we resize and pad images to the size of 1024×544 or 960×544 to fit our network, so context information near image boundary is preserved in this way.

Filtering and Fusion Similar to the previous MVS methods [33, 34, 36, 11, 21], we introduce photometric and geometric constraints for depth map filtering. The photometric constraint measures the multi-view matching quality, where depth with low confidence value is considered as an outlier. In our experiments, we discard pixels whose probability of estimated depth is lower than 0.3. The geometric constraint measures multi-view depth consistency, where depth inconsistent with its neighboring views should also be discarded. We follow the dynamic geometric consistency checking method presented in [31] to cross-filter original depth maps. After that, we utilize a visibility-based depth fusion method proposed by [22] with a mean average fusion approach [33] to produce final 3D point clouds.

Method	**Rank**	**Mean**	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
CIDER [30]	95.00	46.76	56.79	32.39	29.89	54.67	53.46	53.51	50.48	42.85
Point-MVSNet [6]	93.88	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
Dense R-MVSNet [34]	83.50	50.55	73.01	54.46	43.42	43.88	46.80	46.69	50.87	45.25
PVA-MVSNet [36]	56.62	54.46	69.36	46.80	46.01	55.74	57.23	54.75	56.70	49.06
CVP-MVSNet [32]	55.12	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
P-MVSNet [20]	43.12	55.62	70.04	44.64	40.22	65.20	55.08	55.17	60.37	54.29
CasMVSNet [11]	40.38	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
ACMM [29]	34.25	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48
DeepC-MVS [17]	24.62	59.79	71.91	54.08	42.29	**66.54**	55.77	**67.47**	60.47	**59.83**
Altizure-HKUST-2019 [1]	24.00	59.03	77.19	61.52	42.09	63.50	59.36	58.20	57.05	53.30
AttMVS [21]	19.00	60.05	73.90	**62.58**	44.08	64.88	56.08	59.39	**63.42**	56.06
D^2HC-RMVSNet [31]	18.38	59.20	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92
Vis-MVSNet [37]	15.38	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07
AA-RMVSNet	**6.38**	**61.51**	**77.77**	59.53	**51.53**	64.02	**64.05**	59.47	60.85	54.90

Table 2. Benchmarking results on the Tanks and Temples [16]. The evaluation metric is mean *F-score* (higher is better). AA-RMVSNet outperforms all existing MVS methods with a significant margin and ranks 1^{st} on Tanks and Temples leaderboard (Mar. 15, 2021). The **Rank** is a metric representing the average rank of all 8 scenes and is the basis for final ranking.

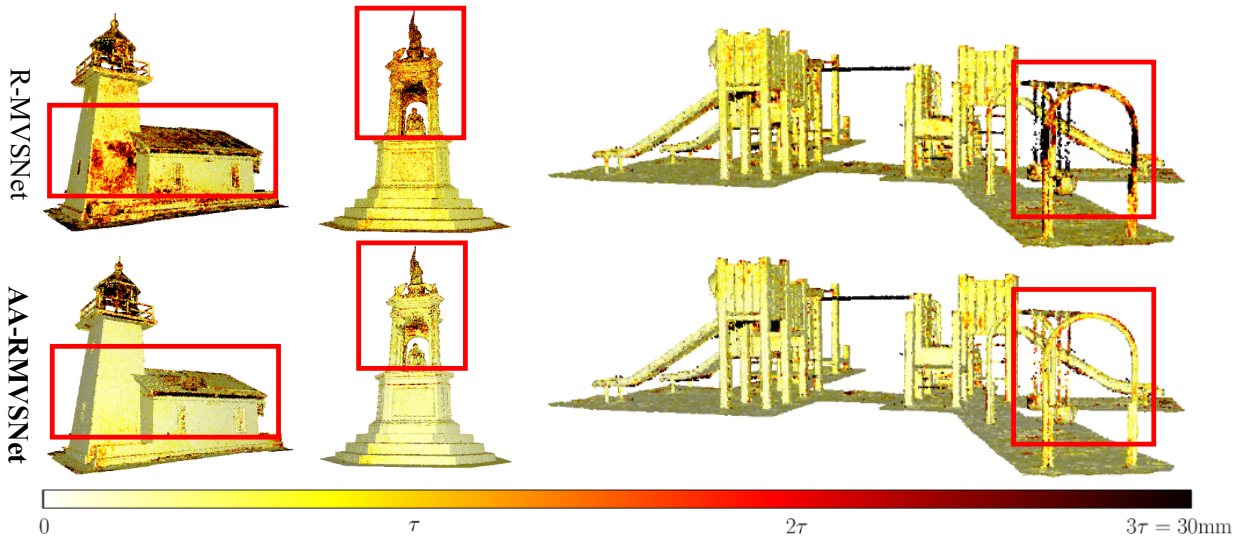


Figure 7. Error visualization of Lighthouse, Francis and Playground in Tanks and Temples benchmark [16] calculated according to corresponding ground truth point clouds, in contrast to R-MVSNet [34].

4.3. Experimental Results

Results on DTU dataset We firstly evaluate AA-RMVSNet on DTU evaluation set [3]. The depth comparison of Scan 13 with [33, 31] is shown in Fig. 5. Benefited from the intra-view AA module which integrates multi-scale and context-aware features, our method is able to estimate more complete and continuous depths for the low-textured surface of the paper box. Some qualitative results compared with other methods are shown in Fig. 6. Due to the improvement of depth map estimation, our method obtains more complete 3D dense point clouds with details reserved. The quantitative results of the whole DTU evaluation set are shown in Tab. 1, where *accuracy* and *completeness* are two absolute distances calculated by the official MATLAB evaluation code [3]. *Overall* is the mean average of the two metrics. Compared with the advanced methods, our method achieves best *completeness* and competitive *overall* performance. Through the comparison with two previous recurrent MVS networks R-MVSNet and D^2HC-RMVSNet, our method significantly improves both *accuracy* and *completeness* on DTU dataset.

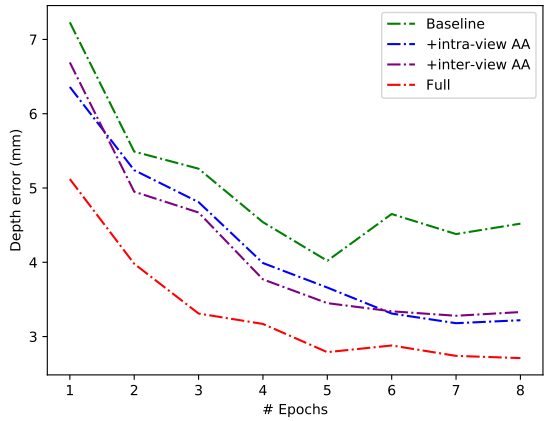


Figure 8. Validation results of the mean average depth error with different network architectures during training.

Benchmarking on Tanks and Temples In order to evaluate the performance of our method under complex outdoor scenes, we test our method on Tanks and Temples benchmark as demonstrated in Tab. 2. Our proposed AA-RMVSNet outperforms all existing MVS methods with a significant margin and ranks 1^{st} on Tanks and Temples leardboard (Mar. 15, 2021) with 61.51 mean *F-score*. Compared with the state-of-the-art methods on DTU dataset, such as CasMVSNet and CVP-MVSNet, our method exhibits stronger robustness and generalizability for varying scenarios. Fig. 7 visualizes the error maps calculated according to the corresponding ground truth point clouds. In contrast to the original recurrent MVS network R-MVSNet, our method significantly improves overall reconstruction quality, especially at challenging regions such as low-textured planes, occluded areas and thin objects, which is benefited from our robust feature extraction and view aggregation methods.

Results on BlendedMVS dataset To further demonstrate the generalizability and scalablility of our method, we also test it on BlendedMVS validation set [35]. Our method successfully reconstructs whole wide-range aerial scenes as well as the small objects. Please check the appendices for results.

4.4. Ablation Study

In this section, we provide ablation experiments to quantitatively analyze the effectiveness and memory cost of each adaptive aggregation method. The following ablation studies are performed on DTU dataset using the same parameters as Sec. 4.2. We compare four different network architectures with or without the proposed adaptive aggregation modules. *Baseline* applies general 2D CNN for feature extraction and the same hybrid LSTM structure for cost volume regularization without any additional modules.

Model	Acc.	Comp.	*O.A.*(mm)	Mem.(GB)
Baseline	0.408	0.374	0.391	**2.41**
+intra-view AA	0.396	0.346	0.371	4.15
+inter-view AA	0.377	0.363	0.370	2.52
Full	**0.376**	**0.339**	**0.357**	4.25
MVSNet [33]	0.396	0.527	0.462	15.4
R-MVSNet [34]	0.385	0.459	0.422	6.7

Table 3. Quantitative and memory performance with different components on DTU evaluation dataset [3].

Validation results of the mean average depth error with different components during training are shown in Fig. 8. It is clear that each individual module can significantly lower the depth error, and the two modules are complementary in full AA-RMVSNet to achieve the best performance.

We also test the point cloud results generated by different network models as shown in Tab. 3. Both intra-view AA and inter-view AA can improve the accuracy and completeness of 3D reconstruction results. Specifically, intra-view AA takes about 1.74GB additional memory and improves *completeness* by 0.28, while inter-view AA only costs extra 0.11 GB and gains 0.31 more in *accuracy*. The *overall* error drops from 0.391 to 0.357 with both two modules. Full AA-RMVSNet only takes 4.25GB to obtain dense and accurate depth maps with 800×600 resolution, indicating that our method is fairly memory efficient.

Regarding ablation study for different experiment settings, please refer to the appendices for detailed results.

5. Conclusion

We have presented a novel recurrent multi-view stereo network with adaptive aggregation modules, denoted as AA-RMVSNet. The intra-view feature aggregation module efficiently improves the performance on thin objects and large low-textured surfaces, by integrating multi-scale and context-aware features adaptively. The inter-view cost volume aggregation module successfully handles the problem of varying occlusion in complex scenes by adaptive pixel-wise view aggregation. The two modules are lightweight, effective and complementary. As a result, our method achieves competitive results on DTU dataset and outperforms other submissions with a significant margin on Tanks and Temples benchmark, showing great generalizability and scalability.

Acknowledgements

This research is supported by National Key Technology Research and Development Program of China, grant number 2017YFB1002601; National Natural Science Foundation of China (NSFC), grant number 61632003; PKU-Baidu Fund, grant number 2019BD007.

References

[1] Altizure. https://github.com/altizure. 7

[2] Tanks and temples benchmark. https://www.tanksandtemples.org. 5

[3] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 1, 5, 6, 7, 8

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1

[5] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 1, 2

[6] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019. 6, 7

[7] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1, 2, 6

[10] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 1, 2, 6

[11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2, 5, 6, 7

[12] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017. 2

[13] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NIPS*, 2017. 2

[14] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 5

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2014. 6

[16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 5, 7

[17] Andreas Kuhn, Christian Sormann, Mattia Rossi, Oliver Erdler, and Friedrich Fraundorfer. Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 404–413. IEEE, 2020. 7

[18] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 2

[19] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005. 2

[20] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019. 6, 7

[21] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020. 1, 6, 7

[22] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2, 6

[23] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011. 2

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. *NeurIPS Autodiff Workshop*, 2017. 6

[25] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 6

[26] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 1, 2, 6

[27] Thomas Schops, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 1

[28] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999. 2

[29] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5483–5492, 2019. 2, 7

[30] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020. 7

[31] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020. 1, 2, 5, 6, 7

[32] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 1, 2, 6, 7

[33] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1, 2, 5, 6, 7, 8

[34] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019. 1, 2, 5, 6, 7, 8

[35] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 1, 5, 8

[36] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Pyramid multi-view stereo net with self-adaptive view aggregation. In *European Conference on Computer Vision*, pages 766–782. Springer, 2020. 1, 2, 6, 7

[37] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *British Machine Vision Conference (BMVC)*, 2020. 1, 2, 5, 6, 7

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 4