Attention-Aware Multi-View Stereo

Keyang Luo[1], Tao Guan[1,3], Lili Ju[2], Yuesong Wang[1], Zhuo Chen[1], Yawei Luo[1*]
[1]School of Computer Science & Technology, Huazhong University of Science & Technology, China
[2]University of South Carolina, USA [3]Farsee2 Technology Ltd, China
{kyluo, qd_gt, yuesongw, cz_007, royalvane}@hust.edu.cn, ju@math.sc.edu

Abstract

Multi-view stereo is a crucial task in computer vision, that requires accurate and robust photo-consistency among input images for depth estimation. Recent studies have shown that learning-based feature matching and confidence regularization can play a vital role in this task. Nevertheless, how to design good matching confidence volumes as well as effective regularizers for them are still under in-depth study. In this paper, we propose an attention-aware deep neural network "AttMVS" for learning multi-view stereo. In particular, we propose a novel attention-enhanced matching confidence volume, that combines the raw pixel-wise matching confidence from the extracted perceptual features with the contextual information of local scenes, to improve the matching robustness. Furthermore, we develop an attention-guided regularization module, which consists of multilevel ray fusion modules, to hierarchically aggregate and regularize the matching confidence volume into a latent depth probability volume. Experimental results show that our approach achieves the best overall performance on the DTU dataset and the intermediate sequences of Tanks & Temples benchmark over many state-of-the-art MVS algorithms.

1. Introduction

Multi-view stereo (MVS) is one of the essential topics in computer vision, which aims to recover a 3D scene surface from a group of calibrated 2D images and estimated camera parameters. With the great success of convolutional neural networks (CNNs) in various visual tasks such as semantic segmentation [27, 26], optical flow estimation [14] and stereo matching [6], learning-based MVS methods [16, 41, 25] have been introduced to promote the quality of reconstructed 3D models. A striking characteristic of learning-based MVS methods is that they make use of vector-valued photo-consistency metrics of corresponding pixels among input images, while conventional MVS methods are usually based on scalar-valued metrics, such as the zero-normalized cross-correlation (ZNCC) [9].

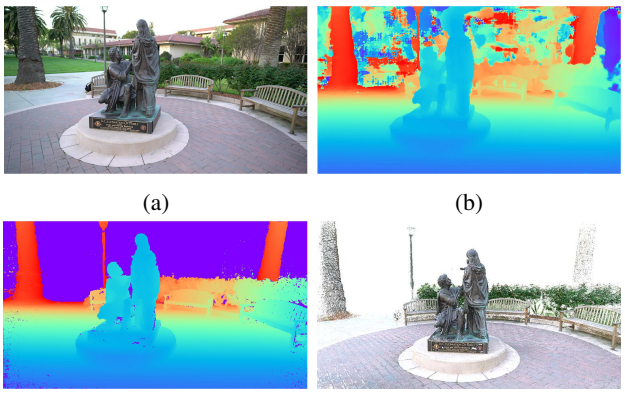


Figure 1: Multi-view 3D reconstruction of the *Family* scene from the Tanks and Temples dataset [20]. (a) The reference image; (b) the inferred depth map from AttMVS; (c) the improved depth map based on (b); (d) the recovered 3D model.

Vector-valued metrics potentially can provide richer matching information for reconstructing high-quality 3D models, but how to fully utilize the perceptual features learned from the images to construct a good matching confidence volume (MCV) is one of the major problems faced by learning-based MVS methods. MVSNet [41] and P-MVSNet [25] use pure photo-consistency information to generate MCVs. However, the feature matching results from different channels are usually not of the same importance since the captured scenes could be significantly different across channels. Inspired by the fact that attention mechanism has achieved great success in natural language processing [36] and visual tasks [12, 8], in this work we combine the photo-consistency information and the contextual information of local scene to construct an attention-enhanced MCV, in which the importance of the matching information from different channels is adaptively adjusted.

Another major problem faced by learning-based MVS methods is how to effectively aggregate and regularize the

*Corresponding author.

matching confidence volume into a latent depth probability volume (LPV), from which the depth/disparity map then can be inferred via some regression or multi-class classification techniques. Inspired by [44], we design a novel attention-guided module to hierarchically aggregate and regularize the matching confidence volume via a top-down/bottom-up manner to achieve a deep regularization.

Quality of the training data also plays a critical role in learning-based MVS methods. High quality data not only can help the target network learn quickly and accurately, but also reflect better the performance of the trained network in the validation stage. The multi-view ground-truth depth maps introduced in [41] for training MVS networks have been widely used, but they still contain quite many wrongly labeled pixels, which could potentially cause some undesired effects on training and validation. In order to avoid this problem, we combine the screened Poisson surface reconstruction method [18] and the visibility-based surface reconstruction approach [37] to improve the quality of existing ground-truth depth maps.

The main contributions of our paper are summarized as follows:

- We design an attention-enhanced matching confidence volume, which takes account of both perceptual information and contextual information of the local scene to improve the matching robustness.
- We propose a novel attention-guided regularization module for hierarchically aggregating and regularizing the matching confidence volume in the top-down/bottom-up manner.
- We develop a simple but effective filtering strategy to improve the quality of multi-view ground-truth depth maps for network training.
- Our method achieves the best overall performance on the DTU benchmark and the *intermediate sequences* of Tanks & Temples benchmark over many state-of-the-art MVS approaches.

2. Related Work

Conventional MVS Conventional MVS methods have achieved excellent performance of predicting depth maps in several recently introduced MVS benchmarks [2, 20, 33]. All of them depend on the PatchMatch algorithm [4] to search the approximate pixel-wise correspondence between images. Galliani *et al.* [11] introduce a GPU-friendly PatchMatch propagation pattern to fully release the parallelization capability of GPUs. Rather than computing the matching confidence based on the image-level view selection, Zheng *et al.* [46] jointly optimize the pixel-level view selection and depth estimation via a probabilistic framework, and Schönberger *et al.* [32] further extend this algorithm to jointly infer pixel-wise depths and normals. Based on [11, 32], Xu and Tao [40] propose a more efficient propagation algorithm guided by the multi-scale geometric consistency and jointly take the views and the depth hypotheses into account. Romanoni *et al.* [30] combine the piecewise planar hypotheses and the EM-based model [32] to estimate the depth in weakly textured planar regions. These attempts have greatly boosted the development of conventional MVS reconstruction algorithms, but how to extend them to manage weak-texture, specular and reactive regions is still a challenging problem.

Learning-based MVS Learning-based MVS methods can be basically categorized into voxel based or depth map based approaches. Voxel based algorithms first compute a bounding box which contains the target object or scene, then divide the bounding box into a three-dimensional volumetric space, and finally estimate whether each voxel belongs to the scene surface or not. SurfaceNet [16] and LSM [17] use generic three-dimensional CNNs while RayNet [29] relies on unrolled Markov random field to estimate the surfaces. These volumetric methods usually are not suitable for large-scale reconstructions. Depth map based methods make use of the plane-sweep stereo algorithm to construct matching confidence volumes, which represent the photo-consistency information coming from the reference and source images. MVSNet [41] and R-MVSNet [42] use the pixel-wise variance-based metric to compute the multi-view photo-consistency of extracted perceptual features while P-MVSNet [25] exploits a confidence metric and learn to aggregate it into a patch-wise matching confidence volume. To regularize the matching confidence volume into a latent depth probability distribution volume, MVSNet [41] uses a generic three-dimensional U-Net, R-MVSNet [42] takes the recurrent neural network to economize the memory usage, and P-MVSNet [25] designs a hybrid three-dimensional U-Net to take the anisotropy of the matching confidence volume into account. Moreover, DeepMVS [13] formulates the depth calculation as a multi-class classification issue and Chen *et al.* [7] introduce a point-based architecture to solve this problem.

Attention-based networks Except natural language processing [36], the attention mechanism have been widely explored in many visual problems including scene segmentation [8, 45, 43], panoptic segmentation [22] and image classification [38]. In particular, SENet [12] adaptively rescales the channel-wise feature responses via an attention-and-gating mechanism. Based on this channel-wise attention mechanism, Zhang *et al.* [45] introduce a context encoding module to improve the feature representation and highlight the class-dependent feature maps selectively. Yu *et al.* [43] present a Smooth Network to enhance the intra-class consistency and select more discriminative features.

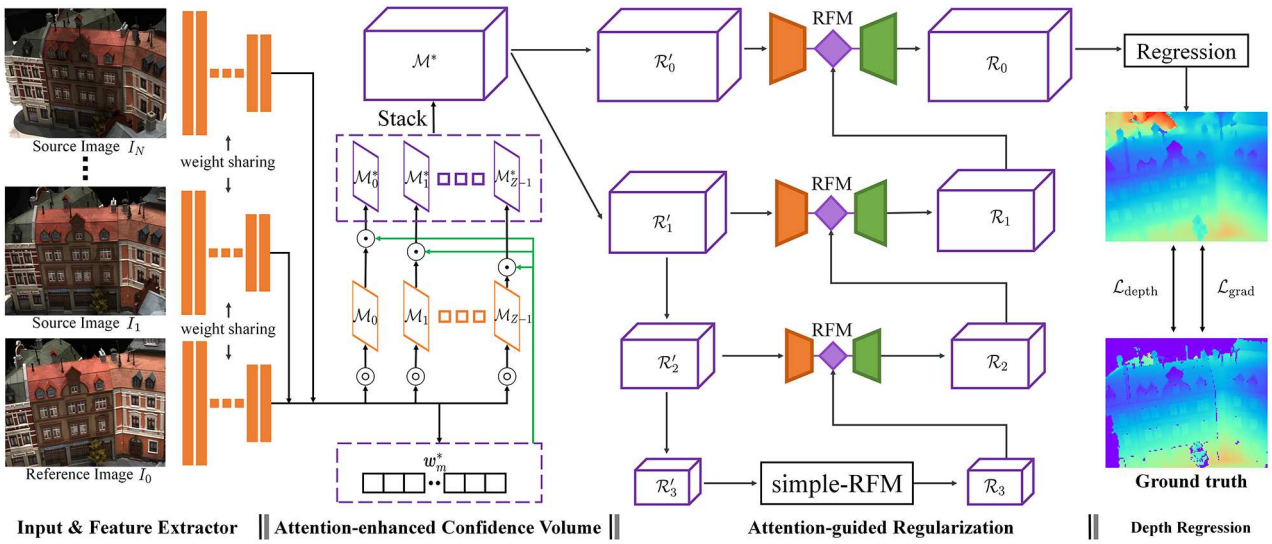


Figure 2: Architecture of the proposed AttMVS for multi-view depth map estimation. The main components include: (a) feature extractor: extracting perceptual features from input images; (b) attention-enhanced confidence volume: constructing the attention-enhanced matching confidence volume for robust and accurate matching; (c) attention-guided regularization: aggregating and regularizing the matching confidence volume hierarchically based on the specially designed ray fusion modules (RFMs); (d) depth regression: estimating the depth map from the regularized confidence volume using 3D convolutions. Here, "$\odot$" denotes the channel-wise multiplication, "$\circledcirc$" represents the homography warping and raw pixel-wise confidence matching, $\mathcal{R}'_i$ and $\mathcal{R}_i$ are the un-regularized and regularized matching confidence volumes on Level i, respectively.

3. Our Method

The architecture of the proposed AttMVS for multi-view depth map estimation is illustrated in Figure 2. Our network first extracts perceptual features from input images using an encoder network (Section 3.1), then use them to construct an attention-enhanced matching confidence volume (Section 3.2). Next, it regularizes the matching confidence volume via an attention-guided hierarchical regularization module (Section 3.3), followed by a depth regression to predict the depth map (Section 3.4).

3.1. Feature extractor

The feature extractor aims to extract perceptual features from input images (the reference image I_0 and N source images $\{I_k\}_{k=1}^N$ of size $H \times W$), which will be used to learn the multi-view photo-consistency. The feature extraction network should possess sufficient capacity, which is essential to obtain accurate and robust feature representations for pixel-level matching. We use the basic architecture of the feature encoder proposed in [25] up to Layer 'conv2_2' with some modifications to build the image feature extractor in our method. In particular, we increase the number of channels for Layers 'conv0_0', 'conv0_1' and 'conv0_2' from 8 to 32, and set that for Layers 'conv1_0', 'conv1_1', 'conv1_2', 'conv2_0', 'conv2_1' and 'conv2_2' to be 64, and finally use a 1×1 convolutional block as the last layer. Thus our feature extractor consists of 10 layers in total and outputs a feature map tensor of size $\frac{1}{4}H \times \frac{1}{4}W \times 16$. Furthermore, the Batch Normalization [15] and ReLU operations used in the original approach are respectively replaced by Instance Normalization [35] and LeakyReLU.

3.2. Attention-enhanced matching confidence

As far as we know, in current learning-based MVS methods, only pixel-wise local perceptual features are used to construct matching confidence volumes. As a result, the overall contextual information of the scene is often neglected in the process. In contrast, in this paper, we combine the photo-consistency information and the contextual cues from the reference and corresponding source image feature maps to construct an attention-enhanced matching confidence volume.

First of all, all extracted image feature maps are squeezed into their individual channel descriptors $\{\boldsymbol{v_i}\}_0^N$ via the global average pooling [24]. From them we calculate the contextual channel-wise statistics $\boldsymbol{w}_v$ of the local scene as follows:

$$\boldsymbol{w}_v = \frac{\sum_{i=0}^{N} (\boldsymbol{v_i} \quad \overline{\boldsymbol{v}})^2}{N}, \tag{1}$$

where $\overline{\boldsymbol{v}}$ is the channel-wise average of $\{\boldsymbol{v_i}\}_0^N$. Next, we calculate the attentional channel-weighted vector $\boldsymbol{w}_v^*$ from $\boldsymbol{w}_v$ via a squeeze-and-excitation block [12] as:

$$\boldsymbol{w}_v^* = \text{Sigmoid}\left(f_2\left(\text{ReLU}\left(f_1\left(\boldsymbol{w}_v, \boldsymbol{s_1}\right)\right), \boldsymbol{s_2}\right)\right) \tag{2}$$

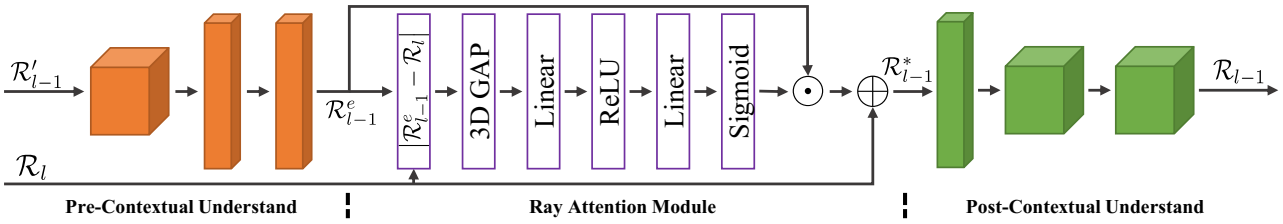


Figure 3: The specially designed ray fusion module (RFM), which includes a pre-contextual understand module, a ray attention module (RAM) and a post-contextual understand module. Here, "$\odot$" denotes the channel-wise multiplication, "$\oplus$" the element-wise sum, and GAP represents Global Average Pooling.

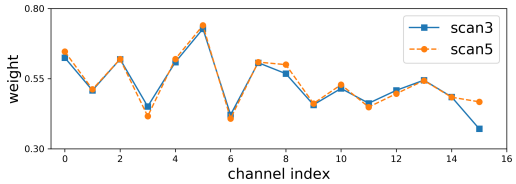


Figure 4: The distribution of the average channel weights of some validation scenes of the DTU dataset.

where $f_1(\cdot,\cdot)$ and $f_2(\cdot,\cdot)$ are two linear transformations, and $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ denote the corresponding transformation parameters. Finally, we obtain the attention-enhanced matching confidence map $\mathcal{M}^*_j$ on the j-th sampled hypothesized depth plane as:

$$\mathcal{M}^*_j = \boldsymbol{w}^*_v \odot \mathcal{M}_j \tag{3}$$

for $j = 0, 1, \cdots, Z \quad 1$, where Z is the total number of sampled hypothesized depth planes, $\odot$ denotes the channel-wise multiplication, and $\mathcal{M}_j$ represents the raw pixel-wise confidence map generated in the way as done in [25] based on the warped feature maps. Figure 4 illustrates an example of the learned weights, from which one can observe that: i) different scenes hold discriminated weights for some channels and similar weights for others; ii) for each scene different channels own different weights.

After computing all attention-enhanced matching confidence maps, we stack them along the depth direction to produce an attention-enhanced matching confidence volume $\mathcal{M}^*$, which will be fed into the regularization module.

3.3. Attention-guided hierarchical regularization

As illustrated in Figure 2, the whole regularization process is described in the following. First, $\mathcal{M}^*$ is encoded into two un-regularized matching confidence volumes $\mathcal{R}'_0$ and $\mathcal{R}'_1$ via two convolutional blocks with stride 1 and 2 respectively. Similarly, $\mathcal{R}'_2$ is then generated by downsampling from $\mathcal{R}'_1$ and $\mathcal{R}'_3$ from $\mathcal{R}'_2$. Thus, we obtain four levels of un-regularized matching confidence volumes $\{\mathcal{R}'_i\}^3_{i=0}$.

Next, the hierarchical regularization process starts with $\mathcal{R}'_3$ on Level 3 (the bottom level) based on multiple ray fusion modules (RFMs) and one simple-RFM. The RFM is used in Levels 0, 1 and 2 and its structure is shown in Figure 3, which consists of a pre-contextual understanding module, a ray attention module (RAM) and a post-contextual understanding module. Both contextual understanding modules are formed by three 3D convolutional blocks, where the second block in the pre-contextual module downsamples the matching confidence volume with increased channels and that in the post-contextual one does the reversed operations. The RAM on Level $l \quad 1$ ($l = 1, 2, 3$) can be explicitly formulated as:

$$\mathcal{R}^*_{l\ \ 1} = \mathcal{R}^e_{l\ \ 1} \odot \boldsymbol{w}^*_r) \oplus \mathcal{R}_l, \tag{4}$$

where $\mathcal{R}^e_{l\ \ 1}$ is the output of the pre-contextual understanding module fed with $\mathcal{R}'_{l\ \ 1}$, $\mathcal{R}_l$ is the regularized matching confidence volume on level l, $\oplus$ denotes element-wise addition, and the ray weighted map $\boldsymbol{w}^*_r$ is calculated from $\boldsymbol{w}_r = |\mathcal{R}^e_{l\ \ 1} \quad \mathcal{R}_l|$ with the same computing structure as Eq. (2). Then, $\mathcal{R}^*_{l\ \ 1}$ is further processed via the post-contextual understand module to obtain the regularized matching confidence volume $\mathcal{R}_{l\ \ 1}$.

The simple-RFM is created by removing the RAM and the upsample and downsample operations from the RFM but keeping a residual connection from the second layer to the fifth one. Note that it is only used on Level 3 to regularize $\mathcal{R}'_3$ into $\mathcal{R}_3$ which can avoid over-cropping of the training and evaluation samples.

3.4. Depth regression and loss function

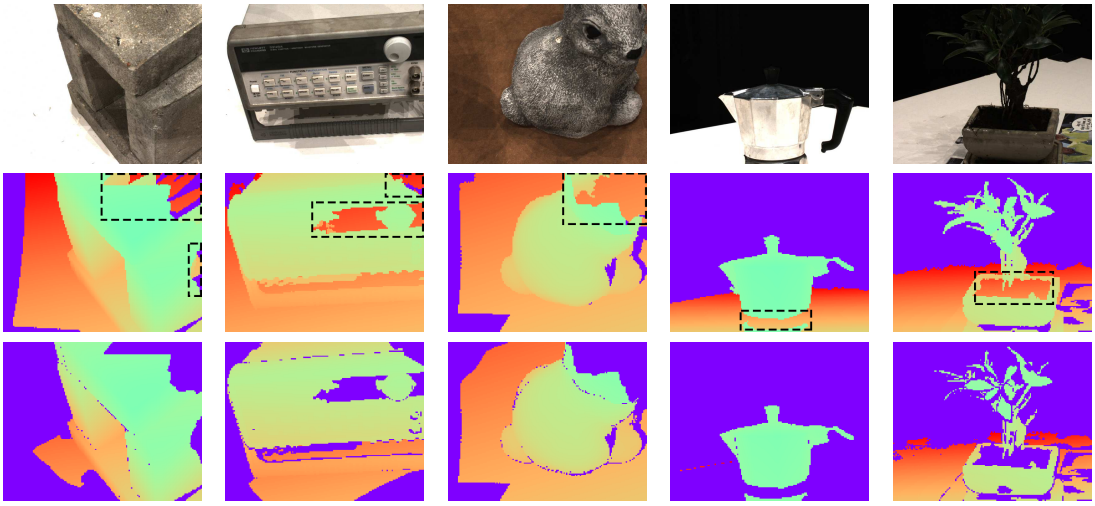
After obtaining the regularized $\mathcal{R}_0$, a three-dimensional convolution layer is first applied to encode it into a depth probability volume $\boldsymbol{V}$. Then we use the depth regression approach introduced in [41] to infer the depth map. The probability of each sampled depth d is computed from $\boldsymbol{V}$ via the Softmax operation $\sigma(\cdot)$. The predicted depth $\hat{d}$ at each labeled pixel is calculated as:

$$\hat{d} = \sum_{d=D_{\min}}^{D_{\max}} d \times \sigma\left(\boldsymbol{V}\right), \tag{5}$$

where $D_{\min}$ and $D_{\max}$ denote respectively the minimum and maximum depth for estimation.

We combine a relative depth loss $\mathcal{L}_{\text{depth}}$ and an inter-gradient regularization loss $\mathcal{L}_{\text{grad}}$ to jointly optimize:

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + \lambda \mathcal{L}_{\text{grad}}, \tag{6}$$



(a) Scan 40 (b) Scan 47 (c) Scan 56 (d) Scan 77 (e) Scan 102

Figure 5: Top row: the reference images from the DTU dataset; middle row: the raw ground-truth depth maps which contain many outliers due to incorrect occlusion relationship; bottom row: the improved ground-truth depth maps by our filtering method.

where $\lambda > 0$ is a weighting coefficient. The relative depth loss function is defined as

$$\mathcal{L}_{\text{depth}}(d^*, \hat{d}) = \frac{1}{\delta \mathcal{N}_d} \sum_{(i,j)} \left| d^*_{i,j} \quad \hat{d}_{i,j} \right| \tag{7}$$

where $\mathcal{N}_d$ denotes the total number of labeled pixels (i, j), $\delta = (D_{max} \quad D_{min})/(Z \quad 1)$ is the length of the sampling interval between hypothesized depth planes, and d^* is the ground-truth depth. To enforce the consistency of depth gradient between the predicted depth map and the ground-truth depth map, the inter-gradient regularization loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{grad}}(d^*, \hat{d}) = \sum_{(i,j)} \Bigg(& \frac{1}{\mathcal{N}_x} \left| \varphi_x(d^*_{i,j}) \quad \varphi_x(\hat{d}_{i,j}) \right| \\ & + \frac{1}{\mathcal{N}_y} \left| \varphi_y(d^*_{i,j}) \quad \varphi_y(\hat{d}_{i,j}) \right| \Bigg), \end{aligned} \tag{8}$$

where $\mathcal{N}_x$ denotes the number of labeled pixels whose neighboring pixels along the x-direction are also labeled, φ_x is the corresponding depth derivative in the x-direction, and $\mathcal{N}_y$ and φ_y represent the similar information along the y-direction.

4. Point Cloud Reconstruction

After obtaining all depth maps, we could directly use the depth map filtering and fusing methods developed in [25] to reconstruct a complete 3D point cloud. On the other hand, for high-resolution scenes with large depth ranges, due to the limitation of the GPU memory, it may be impossible to sample sufficient hypothesized planes for estimating depth map with satisfactory accuracy. To alleviate this issue, we propose to further refine the produced depth maps by maximizing the multi-view photometric consistency with pixel-level view selection. Denote by D_0 the predicted depth map of the reference image I_0 from AttMVS and by $\theta_{i,j}$ the correct depth associated with each pixel (i, j) of I_0. The refinement process can be defined as:

$$\theta^{\text{opt}}_{i,j} = \underset{\theta_{i,j}}{\text{argmin}} \sum_{k=1}^{N} P(k) \quad 1 \quad \rho^k_{i,j} \Big)^2, \tag{9}$$

where $\rho^k_{i,j}$ is the ZNCC measurement and $P(k)$ represents the probability of the source image I_k being the best for depth refinement of pixel (i, j) as defined in [46]. The computation of $P(k)$ needs $\theta_{i,j}$ and that of $\rho^k_{i,j}$ involves $P(k)$, thus we use the GEM algorithm [46] with D_0 as the initial guess to iteratively solve the problem (9).

5. Experimental Results

5.1. Improving ground-truth depth maps

The DTU benchmark [2] is a popular large-scale MVS benchmark, which contains 124 diverse scenes captured in varying lighting conditions. Each scan consists of a reference point cloud, 49 or 64 captured images and their corresponding camera parameters. Unfortunately, this dataset can not be directly used by the depth map based methods for network training, and one need to generate the corresponding depth maps from the provided reference point clouds. The scheme proposed in [41] has been widely used for this purpose. Specifically, for each scan, it first produces and trims the mesh surface based on the screened Poisson surface reconstruction algorithm [18], then renders the corresponding depth maps according to the camera parameters from different viewpoints, which we regard as the *raw depth maps* in this paper. However, the *raw depth maps* could contain many outliers due to incomplete mesh information and

Table 1: Comparison results of the proposed AttMVS with different model variants on the DTU validation set.

Models	Settings					MADE	Pred. prec. ($\tau = \delta$)	Pred. prec. ($\tau = 3\delta$)
	Mod. fea. extr.	Att MCV	Simple-RFM	RFMs	Joint loss			
Baseline						2.14	83.11	95.77
Model-A	√					1.96	84.57	96.25
Model-B	√	√				1.91	84.98	96.36
Model-C	√	√	√			1.89	85.64	96.45
Model-D	√	√	√	√		1.82	87.08	96.84
Full	√	√	√	√	√	**1.79**	**87.61**	**97.04**

incorrect occlusion relationship, which seriously hinder us from training high-performance networks.

To address this issue, we propose an efficient depth filtering method to improve quality of the depth maps. First of all, for each scan, we estimate the mesh surface using a reconstruction system similar to [37] based on the ground-truth camera settings, which produces a highly complete water-tight mesh but may not be accurate enough. Then, we render the *visibility depth maps* based on this mesh using the same rendering procedure for the *raw depth maps*. Finally, for a *raw depth map* D^r and its corresponding *visibility depth map* D^v, the filtered ground-truth depth map D^* is finally generated by

$$D^*_{i,j} = \begin{cases} D^r_{i,j} & \left|D^r_{i,j} \quad D^v_{i,j}\right| < \eta, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where η is a threshold to control completeness of the filtered depth map ($\eta = 5mm$ is set in all experiments). Figure 5 illustrates the effectiveness of our filtering strategy in improving the quality of ground-truth depth maps.

5.2. Model training

The proposed AttMVS is implemented in PyTorch and trained using the DTU dataset using the Adam optimizer [19] with batch-size equal to 2. We refer to [41, 25] for partition of the DTU dataset. The learning rate is initialized to be 10^{-3}, then decays every epoch with the rate of 0.85 and we fix it as $10^{-3} \times 0.85^{10}$ from the $11th$ epoch. Each training sample consists of one reference image and two source images, and a set of $Z = 256$ fronto-parallel hypothesized depth planes are uniformly sampled from $D_{\min} = 425mm$ to $D_{\max} = 935mm$. All images are resized and cropped to height $H = 512$ and width $W = 640$ as done in [41]. In the training process, we observe that the computation efficiency of homography transformation on GPU is very low. We also notice that all scanned scenes share the same set of camera parameters and the adjacent relationships between the cameras are also fixed. Therefore, we pre-calculate all possible homography transformations in advance and directly use them during training of the network, which reduces the training time of each mini-batch from around $1.8s$ to $1.2s$ (saves about one-third of the training time). The whole model is trained from scratch for 20 epochs in total with an NVIDIA Titan RTX GPU, which costs about four days.

5.3. Ablation study

In this section, we perform an ablation study to verify the performance of feature extractor, attention-enhanced matching confidence volume and attention-guided regulation module in the proposed AttMVS. The evaluation metrics we use are *mean absolute depth error* (MADE) and *prediction precision* [25] (i.e., the percentage of the number of pixels where the absolute error of the predicted depth is less than an error threshold τ to the total number of valid pixels in the ground-truth depth map).

The baseline model is created by using the original feature extractor, the raw matching confidence volume and the generic 3D U-Net regularizer (without any attention-mechanism), and trained with the relative depth loss only. Based on the baseline model, we then start to employ the modified feature extractor, the attend-enhanced matching confidence volume, the attention-guided regularization module, and finally the joint loss training step-by-step. All model variants are trained with the same procedure as described in Section 5.2 and then tested on the DTU validation set. The settings of the validation samples are the same as those for the training samples. The performance results are reported in Table 1, which clearly demonstrate the effectiveness of these specially designed components in our method. The full AttMVS model decreases the *mean absolute depth error* from $2.14mm$ to $1.79mm$, and increase the *prediction precision* with $\tau = \delta$ from 83.11% to 87.61% and with $\tau = 3\delta$ from 95.77% to 97.04%.

5.4. Comparison with other methods

5.4.1 On the DTU benchmark

We will compare the performance of proposed AttMVS with many existing state-of-the-art methods, including conventional algorithms [5, 10, 34, 11] and recently introduced learning-based approaches [41, 42, 25, 7]. We infer the depth maps for all the images of each scan from the DTU evaluation set firstly, then we fuse all related depth maps to recover the corresponding three-dimensional point cloud for each scan. We adopt the popularly used *accuracy* and *completeness* of reconstructed three-dimensional

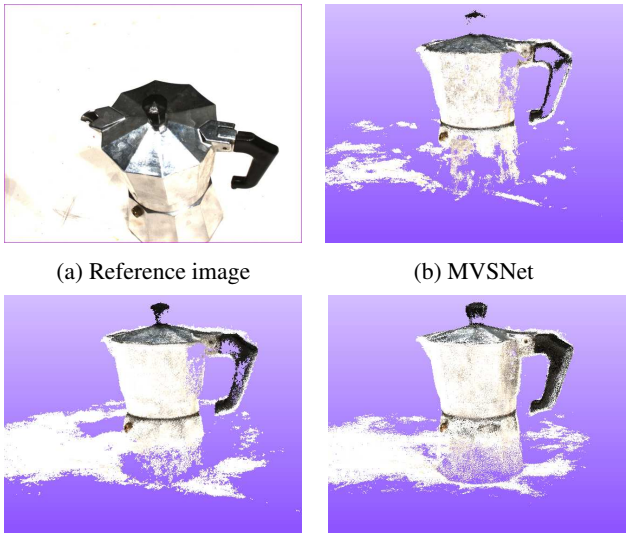
Table 2: Comparisons on the recovered three-dimensional models for the DTU evaluation scenes by different methods. AttMVS* denotes inclusion of the refinement of the depth maps by (9).

Method	Mean accuracy	Mean completeness	Overall
Gipuma [11]	**0.274**	1.193	0.734
tola [34]	0.343	1.190	0.767
furu [10]	0.612	0.939	0.776
camp [5]	0.836	0.555	0.696
SurfaceNet [16]	0.450	1.043	0.746
MVSNet [41]	0.396	0.527	0.462
R-MVSNet [42]	0.385	0.459	0.422
Point-MVSNet [7]	0.342	0.411	0.376
P-MVSNet [25]	0.406	0.434	0.420
AttMVS ($Z = 256$)	0.412	0.394	0.403
AttMVS ($Z = 384$)	0.391	**0.345**	**0.368**
AttMVS* ($Z = 384$)	0.383	**0.329**	**0.356**

point clouds as the evaluation measures, and the evaluation is conducted via the MATLAB code [2] with the default configuration. The quantitative comparison results are presented in Table 2, which shows that AttMVS outperforms all comparison methods in *completeness* and keeps quite competitive in *accuracy*, and as a consequence, AttMVS achieves the best overall performance. Another observation is that the quality of AttMVS reconstruction can be greatly improved with the increase of the number of hypothesized depth planes. In addition, we specially add the depth map refinement process (9) for the case of AttMVS with $Z = 384$ and it is found from Table 2 that such a refinement step can further improve the reconstruction quality. Figure 6 shows the qualitative comparison of *scan 77* (often regarded as the most challenging scene in the DTU evaluation set) among MVSNet [41], P-MVSNet [25] and our AttMVS.

5.4.2 On the Tanks & Temples benchmark

The Tanks & Temples is a widely used large-scale MVS benchmark and consists of two sequences: *intermediate sequences* and *advanced sequences*. All of them are acquired in realistic environments under different weather conditions and only the captured images are provided for subsequent evaluation. *F-score* is the only evaluation metric, which takes both *accuracy* and *completeness* into account to measure the quality of reconstruction comprehensively. This dataset is used to evaluate and compare the generalization ability of our method. For evaluation, we first recover the camera poses and calibration parameters of the provided image set based on the revised SfM pipeline of COLMAP [31] and compute the prediction depth range of each reference image based on the SfM result. Next, we infer each depth map using the corresponding reference and source images with $Z = 384$ uniformly sampled hypothesized planes. Finally, we upsample the inferred depth maps back to input image resolution and refine them by (9), then fuse them into a unified point cloud for each scene. Note that all sequences of Tanks & Temples provide many images and overlaps of the images are very large, thus we apply stricter fusion thresholds to suppress possible outliers than DTU dataset.



(a) Reference image (b) MVSNet

(c) P-MVSNet (d) AttMVS

Figure 6: Qualitative comparison of three-dimensional models of *scan* 77 on the DTU benchmark.

The evaluation and comparison results are reported in Table 3 (for *intermediate sequences*) and Table 4 (for *advanced sequences*), and Figure 5 visually illustrates the reconstructed point clouds for some scenes. It is observed that our AttMVS achieves the best overall performance (#1 in rank and mean) among all comparison methods on the *intermediate sequences*, and specifically the reconstructed point clouds for *Francis*, *Playground* and *Train* by our method obtain the best quality. The performance of our method on the *advanced sequences* is still competitive but is worse than that on the *intermediate sequences* when compared with some conventional MVS methods. We think the main reason is that for great majority part of images in the *advanced sequences*, the interested depth ranges are very large, but due to the GPU memory limitation, our method could not sample sufficient hypothesized depth planes to assure the quality of predicted depth maps even though the depth map refinement has been used. Thus, our method suits better to reconstruct the scenes with the interested depth range of the captured images being concentrated, which is also the common restriction of current learning-based MVS algorithms.

6. Conclusion

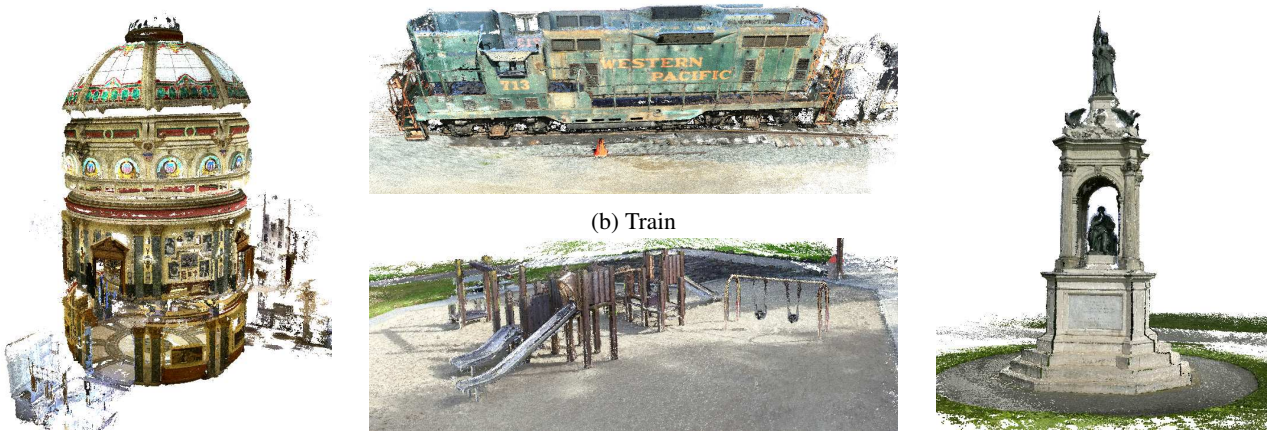
In this paper we have proposed a novel attention-aware MVS network (AttMVS) for multi-view depth map estimation. Specifically, the matching robustness is improved

Table 3: Performance comparisons of various reconstruction algorithms on the *intermediate sequences* of the Tanks & Temples benchmark. Our AttMVS ranks 1st among all of the submissions.

Method	Rank	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
AttMVS (**Ours**)	**2.38**	**60.05**	73.90	**62.58**	44.08	64.88	56.08	59.39	**63.42**	**56.06**
Altizure-HKUST-2019 [3]	4.00	59.03	**77.19**	61.52	42.09	63.50	59.36	58.20	57.05	53.30
3Dnovator [1]	4.62	58.37	73.43	52.51	37.08	64.55	**59.58**	**62.88**	62.88	51.40
ACMM [40]	6.12	57.27	69.24	51.45	**46.97**	63.20	55.07	57.64	60.08	54.48
Altizure-SFM, PCF-MVS [21]	7.38	55.88	70.99	49.60	40.34	63.44	57.79	58.91	56.59	49.40
OpenMVS [28]	7.75	55.11	71.69	51.12	42.76	58.98	54.72	56.17	59.77	45.69
P-MVSNet [25]	7.75	55.62	70.04	44.64	40.22	**65.20**	55.08	55.17	60.37	54.29
ACMH [39]	9.75	54.82	69.99	49.45	45.12	58.86	52.64	52.37	58.34	51.61
PLC_ [23]	10.62	54.56	70.09	50.30	41.94	59.04	49.19	55.53	56.41	54.13
Point-MVSNet [7]	18.25	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
Dense R-MVSNet [42]	18.38	50.55	73.01	54.46	43.42	43.88	46.80	46.69	50.87	45.25
R-MVSNet [42]	21.50	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
MVSNet [41]	27.88	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
COLMAP [31, 32]	30.12	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04

Table 4: Performance comparisons of various reconstruction approaches on the *advanced sequences* of the Tanks & Temples benchmark.

Method	Rank	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
Altizure-HKUST-2019 [3]	**3.17**	**37.34**	24.04	**44.52**	36.64	49.51	**30.23**	**39.09**
Altizure-SFM, PCF-MVS [21]	4.33	35.69	**28.33**	38.64	35.95	48.36	26.17	36.69
OpenMVS [28]	5.50	34.43	24.49	37.39	38.21	47.48	27.25	31.79
3Dnovator [1]	5.67	34.51	18.61	40.77	37.17	50.30	27.60	32.61
PLC_ [23]	5.83	34.44	23.02	30.95	42.50	49.61	26.09	34.46
COLMAP-SFM, PCF-MVS [21]	6.17	34.59	26.87	31.53	**44.70**	47.39	24.05	32.97
ACMM [40]	6.33	34.02	23.41	32.91	41.17	48.13	23.87	34.60
AttMVS (**Ours**)	8.00	31.93	15.96	27.71	37.99	**52.01**	29.07	28.84
Dense R-MVSNet [42]	11.83	29.55	19.49	31.45	29.99	42.31	22.94	31.10
R-MVSNet [42]	15.67	24.91	12.55	29.09	25.06	38.68	19.14	24.96



(a) Museum (b) Train (d) Playground (c) Francis

Figure 7: Visual results of Tanks & Temples benchmark. The *Francis*, *Train* and *Playground* scenes are from the *intermediate sequences* while the *Museum* scene is from the *advanced sequences*.

by the attention-enhanced matching confidence volume, which combines the contextual information of the scene with the raw pixel-wise matching volume through an adaptive weighting approach, and the corresponding attention-guided regularization module can hierarchically aggregate and regularize the matching confidence volume in a deep manner. In addition, we also have proposed a simple but effective filtering strategy to enhance the quality of ground-truth depth maps for network training. Comprehensive experiments on the Tanks & Temples and DTU benchmarks qualitatively and quantitatively demonstrate the excellent performance of the proposed AttMVS.

References

[1] 3Dnovator. http://www.3dnovator.com/. 8

[2] 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*, pages 1–16, 2016. 2, 5, 7

[3] Altizure. https://www.altizure.com/. 8

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM SIGGRAPH 2009 Papers*, SIGGRAPH '09, New York, NY, USA, 2009. Association for Computing Machinery. 2

[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. 6, 7

[6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1

[7] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1538–1547, 2019. 2, 6, 7, 8

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1, 2

[9] Luigi Di Stefano, Stefano Mattoccia, and Federico Tombari. Zncc-based template matching using bounded partial correlation. *Pattern recognition letters*, 26(14):2129–2134, 2005. 1

[10] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 6, 7

[11] 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. 2, 6, 7

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, June 2018. 1, 2, 3

[13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 2

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *international conference on machine learning*, pages 448–456, 2015. 3

[16] 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. 1, 2, 7

[17] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2

[18] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013. 2, 5

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4), July 2017. 1, 2

[21] Andreas Kuhn, Shan Lin, and Oliver Erdler. Plane completion and filtering for multi-view stereo reconstruction. In *German Conference on Pattern Recognition*, pages 18–32. Springer, 2019. 8

[22] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 2

[23] Jie Liao, Yanping Fu, Qingan Yan, and Chunxia Xiao. Pyramid multi-view stereo with local consistency. In *Pacific Graphics*, 2019. 8

[24] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3

[25] 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 International Conference on Computer Vision*, pages 10452–10461, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[26] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019. 1

[27] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 1

[28] OpenMVS. open multi-view stereo reconstruction library. https://github.com/cdcseacave/openMVS. 8

[29] Despoina Paschalidou, Ali Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2018. 2

[30] Andrea Romanoni and Matteo Matteucci. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. *arXiv preprint arXiv:1903.10929*, 2019. 2

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 4104–4113, 2016. 7, 8

[32] 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. 2, 8

[33] Thomas Schöps, 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, 2017. 2

[34] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 6, 7

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv: Computer Vision and Pattern Recognition*, 2016. 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 1, 2

[37] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):889–901, 2011. 2, 6

[38] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 2

[39] Qingshan Xu and Wenbing Tao. Multi-view stereo with asymmetric checkerboard propagation and multi-hypothesis joint view selection. *CoRR*, abs/1805.07920, 2018. 8

[40] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 8

[41] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *The European Conference on Computer Vision (ECCV)*, pages 767–783, September 2018. 1, 2, 4, 5, 6, 7, 8

[42] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *arXiv preprint arXiv:1902.10556*, 2019. 2, 6, 7, 8

[43] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1857–1866, June 2018. 2

[44] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2403–2412. IEEE, 2018. 2

[45] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, June 2018. 2

[46] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 2, 5