Appendices

A. Network Details

Detailed information of the layers of AA-RMVSNet is listed in Tab. 1. Note that the procedure of feature extraction is identical for all N images and the procedure of cost volume processing is identical for all D depth hypotheses.

B. Deformable Sampling in Intra-view AA

In terms of feature extraction for matching, we expect regions with rich texture to be processed by convolutions with smaller receptive fields so that tiny and detailed parts will be preserved during matching. While for low-textured or textureless regions, such as plain surfaces, we prefer a larger receptive field where more context information can get aggregated for more reliable matching.

The proposed intra-view AA module adopts deformable convolution to do the aforementioned job adaptively. For a pixel $\mathbf{p}$ at the object boundary, all sampling points of a deformable convolution kernel tend to be located on the same surface as $\mathbf{p}$. In contrast, for the pixel in textureless regions, sampling points are spread over a larger region and the receptive field is expanded. Fig. 1 visualizes sampling locations of deformable convolution kernels. On the thin cable of the earphone, sampling points tend to be concentrated on the cable itself, while for other low-textured areas of the earphone, the receptive filed is expanded. At boundary regions of objects, sampling points are gathered at the same side of the kernel center.

C. View Reweighting in Inter-view AA

In order to handle an arbitrary number of input views and eliminate the influence of unreliable matching at occluded regions, an inter-view AA module is leveraged to our AA-RMVSNet. The inter-view AA module contains a CNN for yielding pixel-wise attention maps for per-view cost volumes adaptively. For an area in the reference image, if this area is occluded in the source image, lower weights should be assigned to suppress local matching. On the contrary, if an area is well-captured and unoccluded, higher weights are assigned to enhance reliable local matching.

Fig. 2 visualizes two attention maps by gray-scale images. As is clearly framed in red, for areas well-captured in the corresponding source images, attention values are larger. In this way, reliably matched areas of per-view cost volumes are enhanced while those occluded unreliable regions are suppressed by low weights.

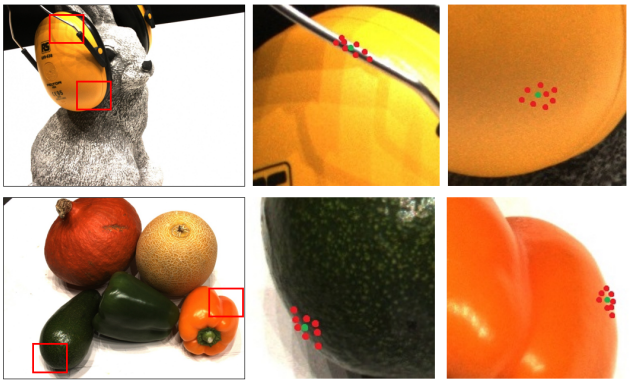
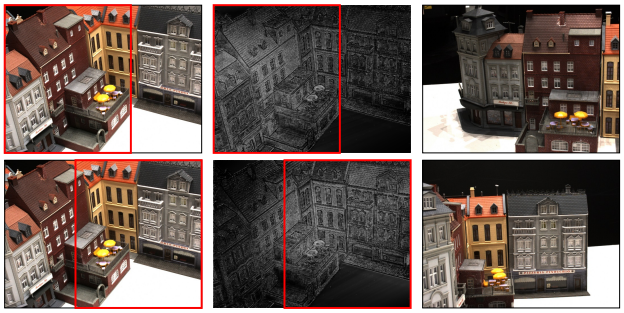


Figure 1. Deformable sampling in different areas, *e.g*. thin object, weak-textured region and object boundary. Green points are centers of convolution kernels and red ones are sampled points with adaptive offsets yielded by sub-networks of deformable convolutions.



(a) Reference image (b) Pixel-wise attention map (c) Neighboring source image

Figure 2. Visualized pixel-wise attention maps yielded by the inter-view AA module. Brighter areas represent higher weights assigned. When matching source images in (c) to the reference image (a), corresponding per-view attention maps are shown as (b).

D. Depth Comparison in Ablation Experiments

To further demonstrate the effectiveness of the proposed intra-view AA module and inter-view AA module, we visualize some representative depth maps for each ablation experiment.

As is shown in Fig. 3, the intra-view AA module manages to eliminate noises at textureless surfaces and boundary areas of objects. At the same time, the inter-view AA module is able to preserve more details for those regions easy to be occluded, *e.g*. the handle. Integrated both AA modules into the proposed network, our AA-RMVSNet benefits from both modules and is capable of predicting accurate and complete depth maps for images under varying conditions.

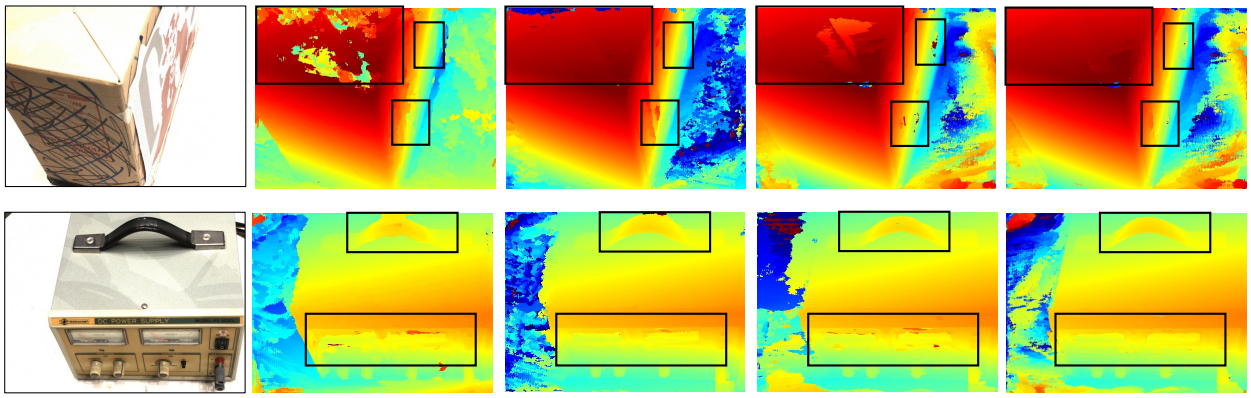
Input	Description	Output	Output Shape
Feature Extraction: $\mathbf{I} \rightarrow \mathbf{f}$ $H \times W \times 3 \rightarrow H \times W \times 32$			
$\mathbf{I}$	Conv(3×3)+GN+ReLU	$\mathbf{x}_0$	$H \times W \times 8$
$\mathbf{x}_0$	Conv(3×3)+GN+ReLU	$\mathbf{x}_1$	$H \times W \times 16$
$\mathbf{x}_1$	Conv(3×3)+GN+ReLU	$\mathbf{x}_2$	$H \times W \times 16$
$\mathbf{x}_2$	Conv(3×3)+GN+ReLU	$\mathbf{x}_3$	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
$\mathbf{x}_3$	Conv(3×3)+GN+ReLU	$\mathbf{x}_4$	$\frac{1}{4}H \times \frac{1}{4}W \times 16$
$\mathbf{x}_2$	DeformConv(3×3)	$\mathbf{x}'_2$	$H \times W \times 16$
$\mathbf{x}_3$	DeformConv(3×3)+BI	$\mathbf{x}'_3$	$H \times W \times 8$
$\mathbf{x}_4$	DeformConv(3×3)+BI	$\mathbf{x}'_4$	$H \times W \times 8$
$[\mathbf{x}'_2, \mathbf{x}'_3, \mathbf{x}'_4]$	Concatenation	$\mathbf{f}$	$H \times W \times 32$
Cost Volume Construction: $\mathbf{f}_{ref, src_{i=1,N-1}} \rightarrow \mathbf{C}^{(d)}$ $N \times H \times W \times 32 \rightarrow H \times W \times 32$			
$\mathbf{f}_{src_i}, d$	Homography	$\mathbf{f}^{(d)}_{src_i}$	$H \times W \times 32$
$\mathbf{f}^{(d)}_{src_i}, \mathbf{f}_{ref}$	$(\mathbf{f}^{(d)}_{src_i} - \mathbf{f}_{ref})^2$	$\mathbf{c}^{(d)}_i$	$H \times W \times 32$
$\mathbf{c}^{(d)}_i$	Conv(3×3)+GN+ReLU	$\mathbf{x}_5$	$H \times W \times 4$
$\mathbf{x}_5$	Conv(1×1)+GN+ReLU	$\mathbf{x}_6$	$H \times W \times 4$
$\mathbf{x}_6$	Conv(1×1)+GN	$\mathbf{x}_7$	$H \times W \times 4$
$\mathbf{x}_5 + \mathbf{x}_7$	ReLU	$\mathbf{x}_8$	$H \times W \times 4$
$\mathbf{x}_8$	Conv(1×1)+Sigmoid	$\boldsymbol{\omega}_i$	$H \times W \times 1$
$(1 + \boldsymbol{\omega}_i), \mathbf{c}^{(d)}_i$	$(1 + \boldsymbol{\omega}_i) \odot \mathbf{c}^{(d)}_i$	$\mathbf{c}'^{(d)}_i$	$H \times W \times 32$
$\mathbf{c}'^{(d)}_{i=1,\ldots,N-1}$	Arithmetic Mean	$\mathbf{C}^{(d)}$	$H \times W \times 32$
Cost Volume Regularization: $\mathbf{C}^{(d)} \rightarrow \mathbf{Y}^{(d)}$ $H \times W \times 32 \rightarrow H \times W \times 1$			
$\mathbf{v}^{(d-1)}_0, \mathbf{C}^{(d)}$	ConvLSTMCell(3×3)	$\mathbf{v}^{(d)}_0$	$H \times W \times 16$
$\mathbf{v}^{(d)}_0$	MaxPooling	$\mathbf{v}^{(d)}_1$	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
$\mathbf{v}^{(d-1)}_2, \mathbf{v}^{(d)}_1$	ConvLSTMCell(3×3)	$\mathbf{v}^{(d)}_2$	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
$\mathbf{v}^{(d)}_2$	MaxPooling	$\mathbf{v}^{(d)}_3$	$\frac{1}{4}H \times \frac{1}{4}W \times 16$
$\mathbf{v}^{(d-1)}_4, \mathbf{v}^{(d)}_3$	ConvLSTMCell(3×3)	$\mathbf{v}^{(d)}_4$	$\frac{1}{4}H \times \frac{1}{4}W \times 16$
$\mathbf{v}^{(d)}_4$	TransConv(3×3)+GN+ReLU	$\mathbf{v}^{(d)}_5$	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
$\mathbf{v}^{(d)}_1, \mathbf{v}^{(d)}_5$	Concatenation	$\mathbf{v}'^{(d)}_5$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
$\mathbf{v}^{(d-1)}_6, \mathbf{v}'^{(d)}_5$	ConvLSTMCell(3×3)	$\mathbf{v}^{(d)}_6$	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
$\mathbf{v}^{(d)}_6$	TransConv(3×3)+GN+ReLU	$\mathbf{v}^{(d)}_7$	$H \times W \times 16$
$\mathbf{v}^{(d)}_0, \mathbf{v}^{(d)}_7$	Concatenation	$\mathbf{v}'^{(d)}_7$	$H \times W \times 32$
$\mathbf{v}^{(d-1)}_8, \mathbf{v}'^{(d)}_7$	ConvLSTMCell(3×3)	$\mathbf{v}^{(d)}_8$	$H \times W \times 8$
$\mathbf{v}^{(d)}_8$	Conv(3×3)	$\mathbf{Y}^{(d)}$	$H \times W \times 1$

Table 1. Details information of network layers of AA-RMVSNet. Conv, TransConv and DeformConv denote 2D convolution, 2D transposed convolution (also known as deconvolution) and 2D deformable convolution, respectively. GN represents group normalization while BI represents bilinear interpolation.

E. Ablation Study on Experiment Settings

As is showed in Tab. 2, we investigate the influence of variant numbers of input views N, numbers of depth hypotheses D and resolutions of input images W and H.

Number of Views Our AA-RMVSNet is capable of processing an arbitrary number of views and leveraging the variant importance in multiple views due to the proposed inter-view AA module. With fixed D and image resolution, we compare reconstruction results under $N = 3, 5, 7$. As is shown in Tab. 2, the larger N turns, the better the recon-



(a) Reference image (b) Baseline (c) +intra-view AA (d) +inter-view AA (e) AA-RMVSNet

Figure 3. Comparison between depth maps predicted by the network with and without the two proposed AA modules and full AA-RMVSNet.

N	D	Resolution	Acc.(mm)	Comp.(mm)	O.A.(mm)
3	256	480×360	0.424	0.387	0.405
5	256	480×360	0.414	0.358	0.386
7	256	480×360	0.408	0.351	0.380
7	512	480×360	0.387	0.356	0.372
7	512	640×480	0.381	0.352	0.366
7	512	800×600	**0.376**	**0.339**	**0.357**

Table 2. Ablation study on number of input views N and number of depth hypotheses D on DTU evaluation set (lower is better).

struction results are in terms of all metrics. It demonstrates that our proposed inter-view AA module can well enhance the valid information in the good neighboring views and eliminate bad information in occluded views.

Number of Depth Hypotheses In AA-RMVSNet, cost volumes are regularized recurrently by a RNN-CNN hybrid network. In this way, memory usage is reduced considerably and more room is left for finer division of depth space (or known as plane sweep). We compare reconstruction quality when $D = 256$ and when $D = 512$ with fixed $N = 7$ and image resolution 480×360. As a result, finer depth division lowers reconstruction error.

Resolution of Images Since our AA-RMVSNet regularizes cost volumes in a memory-efficient fashion, we are able to use images of larger resolution for reconstruction. We fix $N = 7$ and $D = 512$ and compare reconstruction results with image resolution of 480×360 and 800×600. Experimental results demonstrate that larger resolution is beneficial for reconstruction.

F. More Point Cloud Results

We visualize all results of DTU evaluation set, the intermediate set of Tanks and Temples benchmark and BlendedMVS validation set respectively in Fig. 4, Fig. 5 and Fig. 6. Our AA-RMVSNet demonstrates its robustness and scalability on scenes with varying depth ranges.

Figure 4. All point clouds results of DTU evaluation set.

Figure 5. All point clouds results of the intermediate set of Tanks and Temples benchmark.

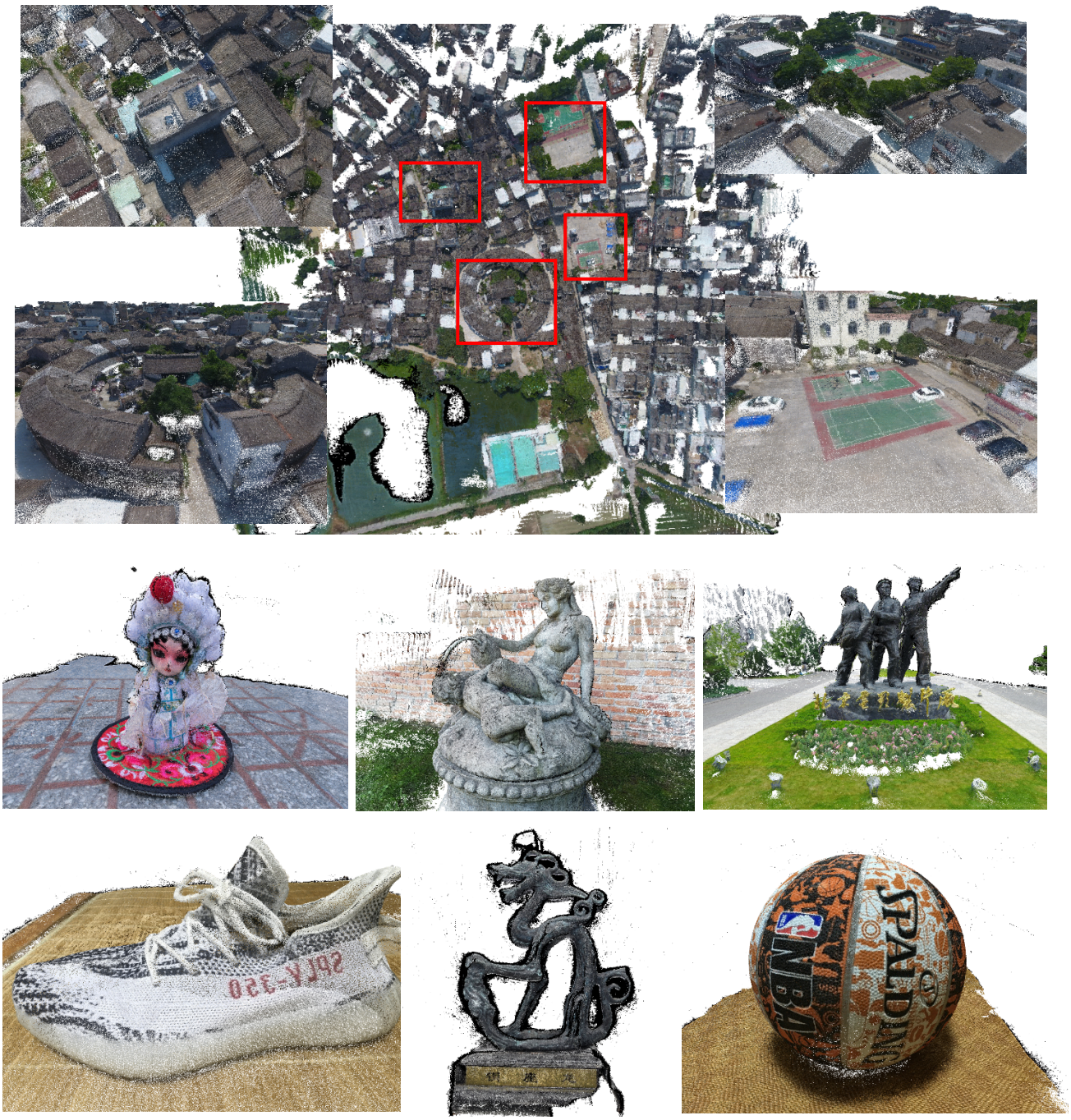


Figure 6. All point clouds results of BlendedMVS validation set.