Supplementary material for IB-MVS: An Iterative Algorithm for Deep Multi-View Stereo based on Binary Decisions

Christian Sormann[1]
christian.sormann@icg.tugraz.at
Mattia Rossi[2]
mattia.rossi@sony.com
Andreas Kuhn[2]
andreas.kuhn@sony.com
Friedrich Fraundorfer[1]
fraundorfer@icg.tugraz.at

[1] Institute of Computer Graphics and Vision
Graz University of Technology
Austria

[2] Sony Europe B.V.
R&D Center - Stuttgart Laboratory 1
Germany

1 Network architecture hyper-parameters

We provide the network hyper-parameters of D-Net and W-Net in Table 1 and Table 2, respectively. As specified in the main paper, we use three resolution levels $l = (0,1,2)$ at quarter, half and full resolution. For generating the image features Feat_{r_l} and Feat_{s_l}, we utilize the FPN architecture of [7] and set the number of feature channels for each level $F_l = (32,16,8)$. Further, we replace the batch normalization [3] in the FPN [7] with instance normalization [6]. We denote the entropy calculated from the output mask B_{s_l} according to Section 3.4 of the main paper as E_{s_l}. In Tables 1 and 2, we denote 2D convolutions as *2D conv*, deformable 2D convolutions [2] as *2D def. conv* and transposed 2D convolutions as *2D tran. conv.* Further, we denote the leaky ReLU activation function as *LReLU*, the number of input and output channels with $\#C_{\text{in}}$ and $\#C_{\text{out}}$, the stride with *str.* and kernel size with k (we add $\backslash b$ when no bias is used, the padding is set to $\frac{k-1}{2}$).

2 Supplementary qualitative results

In Figure 1, we provide qualitative point cloud results of our method IB-MVS for the DTU [1], Tanks and Temples [4] and ETH3D [5] high and low-res datasets. Additionally, in Figure 3, we provide supplementary qualitative IB-MVS depth map results from the DTU [1] dataset.

In order to provide further insights into IB-MVS, in Figure 2 we provide a visualization of its intermediate outputs at different iterations $t = 0,1,4,8$. For each source image I_s, at iteration t the hypothesis h^t is used to compute the binary decision mask B_s^t and the weight mask W_s^t via D-Net and W-Net, respectively; this permits to compute the new reference

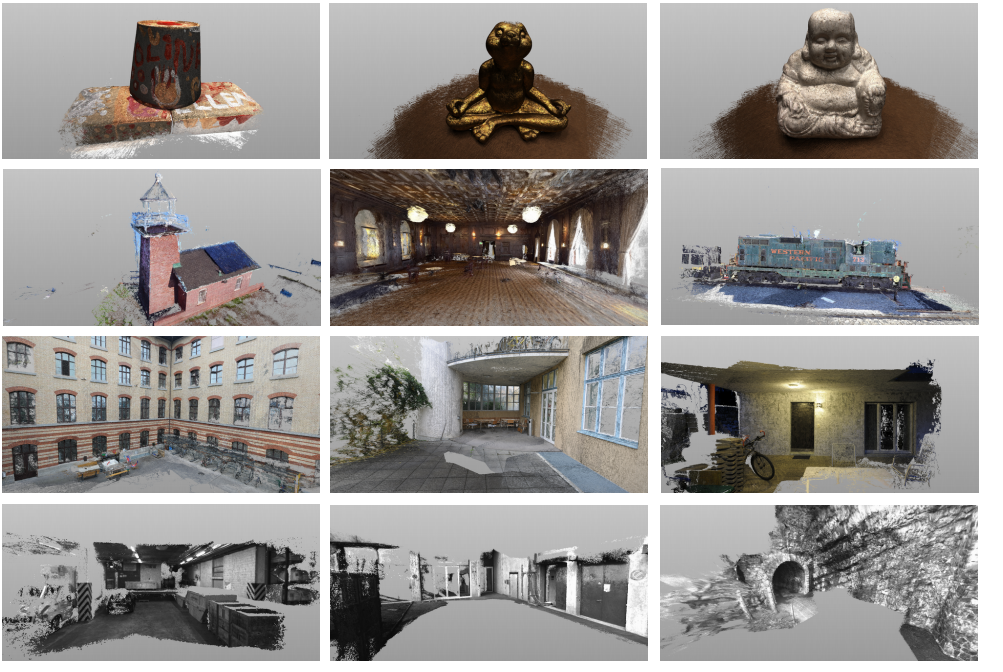

Figure 1: Qualitative point cloud results for DTU [1] (first row), Tanks and Temples [2] (second row) and ETH3D [5] high and low-res (third and fourth row).

depth map hypothesis h_s^{t+1} using Eq. (2) of the main paper. The rows 2-5 of Figure 2 show B_s^t, W_s^t and h_s^{t+1} for the 4 source images and different values of t. The new reference image depth hypothesis h_s^{t+1} are then fused into a single depth map h^{t+1} using the weights W_s^t. The first row of Figure 2 shows the reference image along with h^{t+1} for different values of t. We conclude by observing that, as desired, the weight masks in Figure 2 assign a low confidence to those areas of the reference image that are occluded in the source image, as these areas cannot be matched. This can be appreciated in the weight masks W_s^t depicted in rows 2 and 3, where the area below the sofa and the left-most region (highlighted in yellow) of the reference image are dark because they are occluded in the respective source images.

in name	out name	$\#C_{\text{in}}$	$\#C_{\text{out}}$	operation
Feat_{r_l}	Conv1	F_l	F_l	2D conv., k=3, str.=1, act.=LReLU
Feat_{s_l}	DConv1	F_l	F_l	2D def. conv., k=5, str.=1, act.=LReLU
Conv1 ∣ DConv1	Conc1	$2F_l$	$2F_l$	concatenate along channel dim.
Conc1	Conv2	$2F_l$	$2F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv2	Sc1	$2F_l$	$2F_l$	2D conv., k=3, str.=2, act.=LReLU
Feat_{r_l}	Feat_{r_l} half	F_l	F_l	bilinear interp. downscale to half
Feat_{s_l}	Feat_{s_l} half	F_l	F_l	bilinear interp. downscale to half
Feat_{r_l} half	Conv3	F_l	F_l	2D conv., k=3, str.=1, act.=LReLU
Feat_{s_l} half	DConv2	F_l	F_l	2D def. conv., k=5, str.=1, act.=LReLU
Conv3 ∣ DConv2	Conc2	$2F_l$	$2F_l$	concatenate along channel dim.
Conc2	Conv4	$2F_l$	$2F_l$	2D conv., k=3, str.=1, act.=LReLU
Sc1 ∣ Conv4	Conc3	$4F_l$	$4F_l$	for $l = 0$ concatenate along channel dim.
Conc3	Conv5	$4F_l$	$4F_l$	for $l = 0$ 2D conv., k=3, str.=1, act.=LReLU
Fo_{l-1} ∣ Sc1 ∣ Conv4	Conc3	$4F_l + 4F_{l-1}$	$4F_l + 4F_{l-1}$	for $l > 0$ concatenate along channel dim.
Conc3	ConvPr	$4F_l + 4F_{l-1}$	$4F_l$	for $l > 0$ 2D conv., k=3, str.=1, act.=LReLU
ConvPr	Conv5	$4F_l$	$4F_l$	for $l > 0$ 2D conv., k=3, str.=1, act.=LReLU
Conv5	Sc2	$4F_l$	$4F_l$	2D conv., k=3, str.=2, act.=LReLU
Feat_{r_l}	Feat_{r_l} quar.	F_l	F_l	bilinear interp. downscale to quarter
Feat_{s_l}	Feat_{s_l} quar.	F_l	F_l	bilinear interp. downscale to quarter
Feat_{r_l} quar.	Conv6	F_l	F_l	2D conv., k=3, str.=1, act.=LReLU
Feat_{s_l} quar.	DConv3	F_l	F_l	2D def. conv., k=5, str.=1, act.=LReLU
Conv6 ∣ DConv3	Conc4	$2F_l$	$2F_l$	concatenate along channel dim.
Conc4	Conv7	$2F_l$	$2F_l$	2D conv., k=3, str.=1, act.=LReLU
Sc2 ∣ Conv7	Conc5	$6F_l$	$6F_l$	concatenate along channel dim.
Conc5	Conv8	$6F_l$	$6F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv8	Conv9	$6F_l$	$6F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv9	Conv10	$6F_l$	$6F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv10	UConv1	$6F_l$	$6F_l$	2D tran. conv.,k=4\b, str.=2, act.=LReLU
Conv5 ∣ UConv1	Conc6	$10F_l$	$10F_l$	concatenate along channel dim.
Conc6	Conv11	$10F_l$	$4F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv11	Conv12	$4F_l$	$4F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv12	UConv2	$4F_l$	$4F_l$	2D tran. conv.,k=4\b, str.=2, act.=LReLU
Conv2 ∣ UConv2	Conc7	$6F_l$	$6F_l$	concatenate along channel dim.
Conc7	Fo_l	$6F_l$	$4F_l$	2D conv., k=3, str.=1, act.=LReLU
Fo_l	B_{s_l}	$4F_l$	1	2D conv., k=3\b, str.=1, act.=sigmoid

Table 1: D-Net architecture hyper-parameters, specifying the convolution type, number of input and output channels, kernel size, stride and activation function.

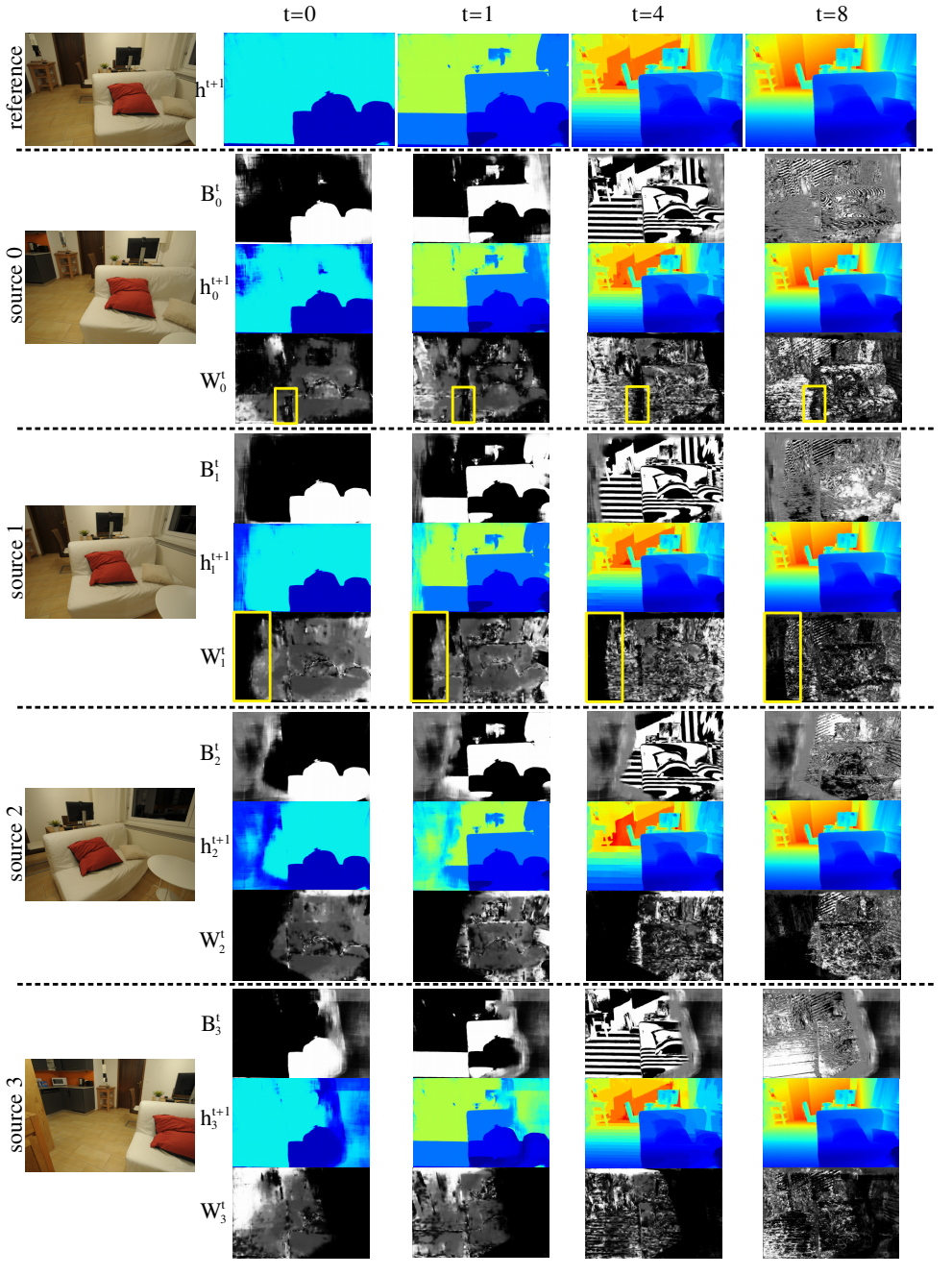

Figure 2: We visualize intermediate results of IB-MVS for a view of the ETH3D [5] high-res `living room` scene. The top row shows the depth hypothesis h^{t+1} predicted at iteration t by fusing the depth hypothesis h_s^{t+1} from different source images according to W_s^t. We also show the binary decision masks B_s^t. We color code low to high depth values from blue to red. For B_s^t and W_s^t black represents the value 0 and white represents the value 1.

in name	out name	$\#C_{\text{in}}$	$\#C_{\text{out}}$	operation
E_{s_l}	Conv1	1	$2F_l$	for $l = 0$ 2D conv., k=3, str.=1, act.=LReLU
E_{s_l}	Conv0	1	F_l	for $l > 0$ 2D conv., k=3, str.=1, act.=LReLU
Fo_{l-1}	Fo_{Up}	$\frac{F_{l-1}}{2}$	$\frac{F_{l-1}}{2}$	for $l > 0$ bilinear interp. upscale to double res.
Fo_{Up}	ConvPr	$\frac{F_{l-1}}{2}$	F_l	for $l > 0$ 2D conv., k=3, str.=1, act.=LReLU
Conv0 | ConvPr	Conc1	$2F_l$	$2F_l$	for $l > 0$ concatenate along channel dim.
Conc1	Conv1	$2F_l$	$2F_l$	for $l > 0$ 2D conv., k=3, str.=1, act.=LReLU
Conv1	Conv2	$2F_l$	$2F_l$	2D conv., k=3, str.=1, act.=LReLU
Conv2	Conv3	$2F_l$	F_l	2D conv., k=3, str.=1, act.=LReLU
Conv3	Fo_l	$2F_l$	$\frac{F_l}{2}$	2D conv., k=3, str.=1, act.=LReLU
Fo_l	w_{s_l}	$\frac{F_l}{2}$	1	2D conv., k=3\b, str.=1, act.=identity

Table 2: W-Net architecture hyper-parameters, specifying the convolution type, number of input and output channels, kernel size, stride and activation function.

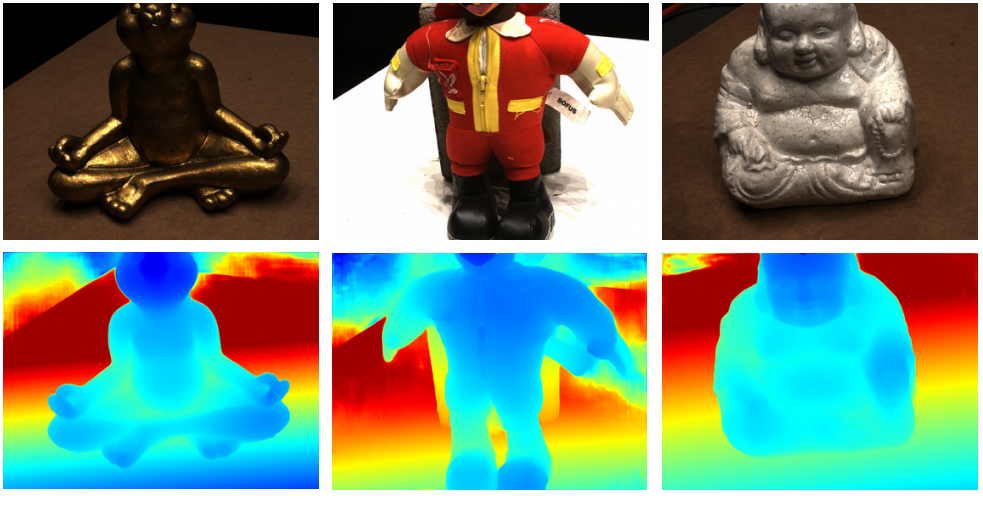

Figure 3: Qualitative depth map results for DTU [1]. For each column, the reference image is at the top and the corresponding IB-MVS depth map at the bottom. Low to high depth values are color coded from blue to red.

References

[1] 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] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

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

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arxiv:1607.08022 [cs.CV]*, 2016.

[7] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.