Supplementary Materials

Jingyang Zhang[1]
jzhangbs@cse.ust.hk
Yao Yao[1]
yyaoag@cse.ust.hk
Shiwei Li[2]
sli@altizure.com
Zixin Luo[1]
zluoag@cse.ust.hk
Tian Fang[2]
fangtian@altizure.com

[1] The Hong Kong University of Science and Technology
Hong Kong SAR, China

[2] Everest Innovation Technology
Hong Kong SAR, China

1 Network Architecture

In this section we introduce the detailed architecture of the proposed framework in Tab. 1.

UNet The UNet [4] is an hourglass-shaped encoder-decoder architecture. The encoder downsizes the input as a feature pyramid with number of scales equal to the length of the array F_{enc}. For each level of scale, the feature map is fed into a downsizing residual block [2] and $N_{enc} - 1$ residual blocks. The number of channels in this level of scale is the corresponding entry in the array F_{enc}. The decode upsamples the feature map back to the original size as an inverse pyramid. For each level of scale, the feature map is fed into a transposed convolutional layer with stride 2. The result is then concatenated with the feature map that has the same size in the encoder pyramid along the channel dimension. And the concatenated feature map is further transformed by N_{dec} residual blocks. The number of channels in this level of scale is the corresponding entry in the array F_{dec}.

Groupwise Correlation We adopt a cost volume construction method that is similar to [1]. Given two volumes with 32 channels, we divide all the channels into 8 groups each with 4 channels. Then correlations are computed between each corresponding pair of group, resulting in 8 values for each voxel.

2 Memory and Time Consumption

In this section we discuss the memory and the time consumption of the inference. Tab. 2 shows the results of the inference on the *Tanks and Temples* [3] dataset w.r.t. number of sources. The size of the inputs is $H \times W = 1056 \times 1920$. Note that the memory consumption is not monotonic because of some engineering issues of PyTorch.

3 More Reconstruction Result

In this section, we show the point cloud results on *DTU* dataset in Fig. 1, 2. The number of source images $N_v = 5$.

Name	Layer	Output
input		$H \times W \times 3$
Feature Extraction		
feat-conv0	5x5 conv, stride=2	$1/2H \times 1/2W \times 16$
feat-UNet	$N_{enc} = 2, N_{dec} = 1$ $F_{enc} = [32, 64, 128], F_{dec} = [64, 32]$	$1/2H \times 1/2W \times 32$
feat-out1	conv on $1/8$ scale, w/o BN, ReLU	$1/8H \times 1/8W \times 32$
feat-out2	conv on $1/4$ scale, w/o BN, ReLU	$1/4H \times 1/4W \times 32$
feat-out3	conv on $1/2$ scale, w/o BN, ReLU	$1/2H \times 1/2W \times 32$
Pair-wise Cost Volume		
cost-volume	Groupwise Correlation	$N_d \times H_k \times W_k \times 8$
Pair-wise Regularization		
reg0-UNet	$N_{enc} = 1, N_{dec} = 0$ $F_{enc} = [8, 16], F_{dec} = [8]$	$N_d \times H_k \times W_k \times 8$
Pair-wise Depth and Uncertainty Estimation		
reg0-conv	3D conv w/o BN, ReLU	$N_d \times H_k \times W_k \times 1$
prob-volume	*softmax* along N_d	$N_d \times H_k \times W_k \times 1$
pair-depth*	*soft argmax* along N_d on prob-volume	$(1\times)H_k \times W_k \times 1$
pair-entropy	*entropy* along N_d on prob-volume	$(1\times)H_k \times W_k \times 1$
uncert-res	residual block on pair-entropy	$H_k \times W_k \times 8$
uncertainty	conv w/o BN, ReLU on uncert-res	$H_k \times W_k \times 1$
Volume Fusion		
fused	weighted average on all reg0-UNet	$N_d \times H_k \times W_k \times 8$
Post-fusion Regularization		
reg1-UNet	$N_{enc} = 1, N_{dec} = 0$ $F_{enc} = [8, 16], F_{dec} = [8]$	$N_d \times H_k \times W_k \times 8$
reg1-out	3D conv w/o BN, ReLU	$N_d \times H_k \times W_k \times 1$
Final Depth Estimation		
final-prob-volume	*softmax* along N_d	$N_d \times H_k \times W_k \times 1$
final-depth*	*soft argmax* along N_d	$(1\times)H_k \times W_k \times 1$

Table 1: Detailed network architecture. All the convolutions are without bias, have kernel size 3, stride 1 and are followed by Batch Normalization and ReLU unless otherwise specified. H, W denote the image height and width respectively. H_k, W_k denote the size of the corresponding stage. *The values in the depth maps are the index of the depth hypothesis.

# sources	2	3	4	5	6	7	8	9
VRAM (MB)	6633	6783	7261	7891	7547	8253	8711	8183
Time (s)	2.20	3.00	3.81	4.64	5.52	6.31	7.18	7.98

Table 2: VRAM and time consumption of the inference on *Tanks and Temples* w.r.t. N_v.

Figure 1: Qualitative results of the point clouds on the *DTU* dataset.

Figure 2: Qualitative results of the point clouds on the *DTU* dataset.

References

[1] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Computer Vision and Pattern Recognition (CVPR*, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *computer vision and pattern recognition (CVPR)*, 2016.

[3] 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):78, 2017.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, 2015.