Supplemental Materials of Constraining Depth Map Geometry for Multi-View Stereo: A Dual-Depth Approach with Saddle-shaped Depth Cells

Xinyi Ye[1] Weiyue Zhao[1] Tianqi Liu[1] Zihao Huang[1] Zhiguo Cao[1*] Xin Li[2]
[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Department of Computer Science, University of Albany, Albany NY 12222
{xinyiye, zhaoweiyue, tianqiliu, zihaohuang, zgcao}@hust.edu.cn
xli48@albany.edu

1. Mathematic Expectations of Interpolated Biases

In the motivation section, we present the effect of depth structures both qualitatively and quantitatively. Here, we report the details about mathematic expectations of different interpolated biases.

Bilinear interpolation. Bilinear interpolation is considered as a multiplexing of the linear interpolation. Due to the facts that the interpolated result is affected by surrounding four pixels and the pixel grid can be viewed as 1, the bilinear interpolation of the estimated bias map $\triangle \boldsymbol{D}$ can be expressed as

$$\begin{aligned}\triangle \boldsymbol{D}(x,y) = &\triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y \rfloor) * (\lfloor x+1 \rfloor - x) * (\lfloor y+1 \rfloor - y) \\ &+ \triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y \rfloor) * (x - \lfloor x \rfloor) * (\lfloor y+1 \rfloor - y) \\ &+ \triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y+1 \rfloor) * (\lfloor x+1 \rfloor - x) * (y - \lfloor y \rfloor) \\ &+ \triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y+1 \rfloor) * (x - \lfloor x \rfloor) * (y - \lfloor y \rfloor)\end{aligned}, \tag{1}$$

where $\lfloor x \rfloor$ indicates the function `floor(x)`. The expectation of absolute interpolated biases can be calculated as

$$\int_{\lfloor x \rfloor}^{\lfloor x+1 \rfloor} \int_{\lfloor y \rfloor}^{\lfloor y+1 \rfloor} |\triangle \boldsymbol{D}(x,y)| * \frac{1}{S} dy dx \,, \tag{2}$$

where $|\cdot|$ denotes the absolute operarion and $S = (\lfloor x+1 \rfloor - \lfloor x \rfloor) * (\lfloor y+1 \rfloor - \lfloor y \rfloor) = 1$.

Expectation of the one-sided cell. One-sided cell indicates all depths are on the same side (beyond or below) of the ground truth. To briefly express the expectation, we assume the estimated bias are all "1", which suggests $\triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y \rfloor) = \triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y \rfloor) = \triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y+1 \rfloor) = \triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y+1 \rfloor) = 1$. Bringing them into Eqs. (1) and (2).The expectation of the one-sided cell is

$$\int_{\lfloor x \rfloor}^{\lfloor x+1 \rfloor} \int_{\lfloor y \rfloor}^{\lfloor y+1 \rfloor} |1| dy dx = 1 \,. \tag{3}$$

Expectation of the saddle-shaped cell. Saddle-shaped cell suggests depths of two adjacent pixels are not on the same side of the ground truth. To briefly express the expectation, we assume $\triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y \rfloor) = \triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y+1 \rfloor) = 1$ and $\triangle \boldsymbol{D}(\lfloor x+1 \rfloor, \lfloor y \rfloor) = \triangle \boldsymbol{D}(\lfloor x \rfloor, \lfloor y+1 \rfloor) = -1$. The expectation of the saddle-shaped cell is

$$\begin{aligned}&\int_{\lfloor x \rfloor}^{\lfloor x+1 \rfloor} \int_{\lfloor y \rfloor}^{\lfloor y+1 \rfloor} \\ &\qquad |(\lfloor x+1 \rfloor - x) * (\lfloor y+1 \rfloor - y) \\ &\qquad - (x - \lfloor x \rfloor) * (\lfloor y+1 \rfloor - y) \\ &\qquad - (\lfloor x+1 \rfloor - x) * (y - \lfloor y \rfloor) \\ &\qquad + (x - \lfloor x \rfloor) * (y - \lfloor y \rfloor)| dy dx \\ &= 0.25 \,.\end{aligned} \tag{4}$$

According to the results in Eqs. (3) and (4), the mathematic expectation of absolute interpolation error for the "one-sided cell" is four times larger than that for the"saddle-shaped cell".

2. Qualitative Results of Our Saddle-shaped Depth Geometry

As a supplementary of Fig. 9 in the main text, we display more visualizations of the depth geometries of our estimated depth maps in Fig. 2. The Dual-Depth Prediction presents depth predictions at the pixel level in an oscillating pattern, such that there are many saddle-shaped cells in the depth maps. As discussed in main text and Sec. 1, the saddle-shaped cells are more helpful to decrease the interpolated bias under the same level of depth estimated quality

*Corresponding author
https://github.com/DIVE128/DMVSNet

Settings	Param. (M)	Macs (G)	Time (ms)	Overall
w.o. Dual-Depth	0.94	268	220	0.345
w.o. Cascade Dual-Depths	1.80	394	373	0.320
DMVSNet	2.67	502	424	0.313

Table 1. Inference time, parameters, and reconstruction quality.

and contribute to a better quality of 3-D reconstruction. The depth geometry with more saddle-shaped cells is the main factor that DMVSNet achieve top results in both indoor and outdoor datasets.

3. Reliable Confidence Maps

In the main text, we demonstrate the superiority of the confidence map defined with the interval of two depths against that of variance. We additionally show the paired depth map and confidence map in Fig. 1. Generally, it is more difficult for network to predict accurate depths for weak-texture, reflective, and edge regions. DMVSNet predicts an greatly oscillating pattern on these regions, results in depths with higher uncertainties. However, thanks to the confidence maps generated by Dual-Depth, the confidence level for these high-uncertainty regions is much lower, such that we can distinguish them.

4. Inference Time

The structural designed for saddle-shaped cells introduces additional parameters because of the doubled 3-D CNNs, which would cost more inference time. As shown in Line 2 $\sim$ 3 of Table 1, the Dual-Depth and Cascade Dual-Depths process more parameters and computational overheads. Since the Dual-Depth requires double branches of 3-D CNNs to produce two depths for each pixel, the increasing cost is inevitable. Desipte more cost, considering the improvements in the quality of 3-D reconstructions, our structural design for saddle-shaped is necessary and effective.

5. More Qualitative Results

We visualize more qualitative results of DTU [1] and Tanks-and-Temples [4] benchmarks. In Figs. 4 and 5, the 3-D point reconstructions of our approach are both dense and accurate in indoor and outdoor scenes, which demonstrates the robustness and scalability of DMVSNet on well-controlled scenarios and complex reality scenarios. Besides, we visualize the completeness error maps of scenarios in Tanks-and-Temples and report their F-scores on the top-right of visualizations in Fig. 3. Compared with Cas-MVSNet [3],TransMVSNet [2], and UniMVSNet [5], our 3-D point reconstructions are more complete as there are fewer dark red areas in the completeness error maps.

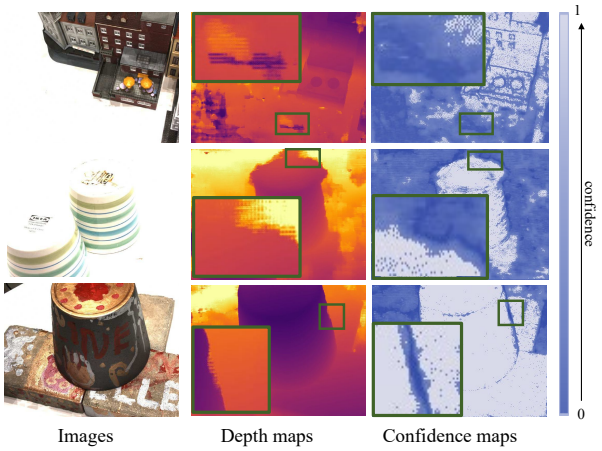

Figure 1. **Visualization of depths and their confidence maps.** We list some images with weak-texture, reflective, and edge regions and visualize their depth maps and confidence maps.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.*, 120:153–168, 2016.

[2] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8585–8594, 2022.

[3] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2495–2504, 2020.

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):1–13, 2017.

[5] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8645–8654, 2022.

Figure 2. **More visualizations of our depth geometry.** We visualize the depth geometry by coloring the pixel whose estimated depth is beyond the ground truth with orange and the others with blue.

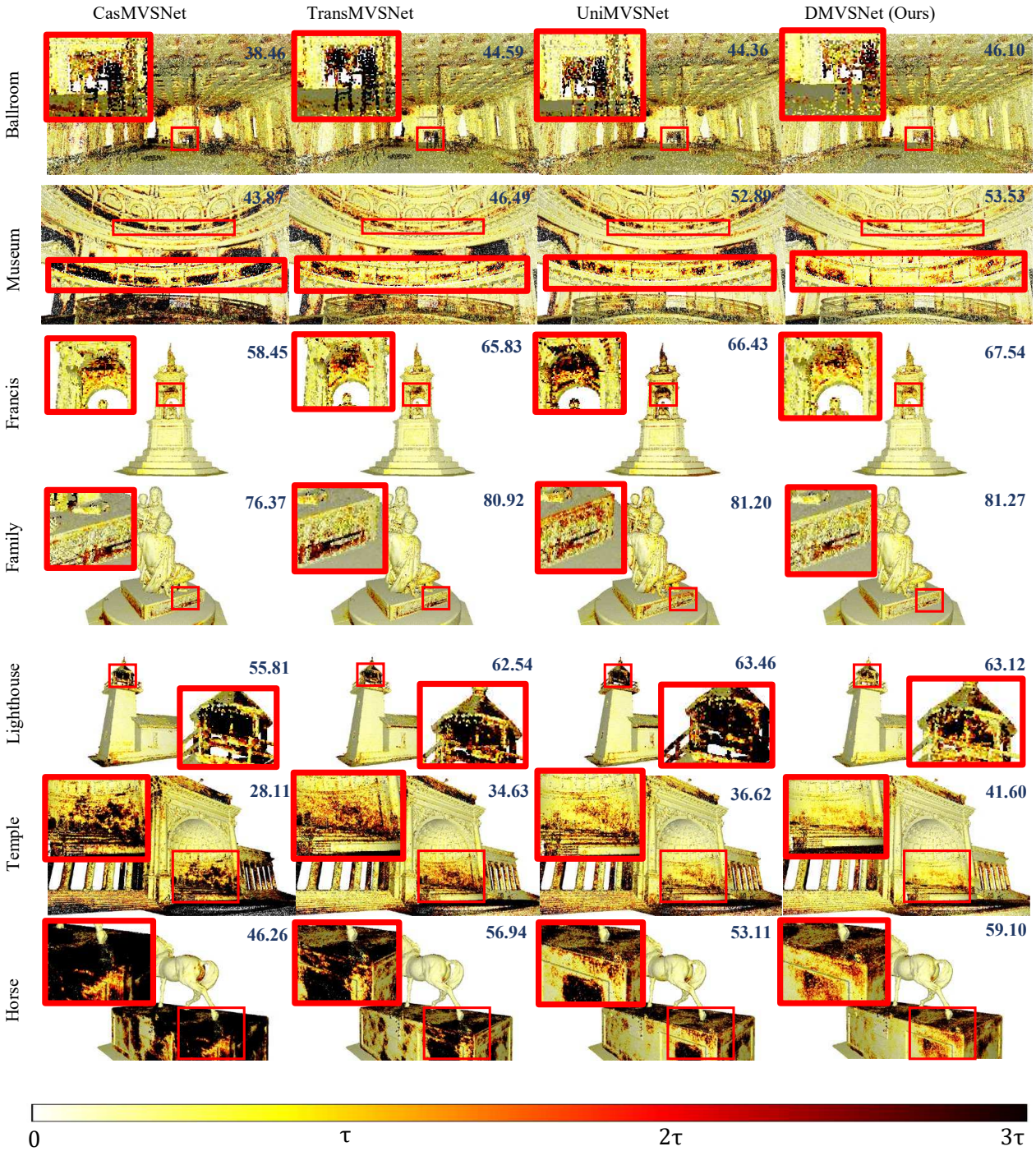

Figure 3. **Qualitative comparison on Tanks-and-Temples.** We visualize the completeness error of some sc as well as their F-scores. The τs of them are 10mm, 10mm, 5mm, 3mm, 5mm, 15mm, and 3mm respectively.

Figure 4. **3-D Point clouds.** We visualize our 3-D Point clouds of scenes in DTU evaluation set.

Figure 5. **3-D Point clouds.** We visualize our 3-D Point clouds of scenes in Tanks-and-Temples.