DeepMVS: Learning Multi-view Stereopsis –Supplementary Material–

Po-Han Huang[1] Kevin Matzen[2] Johannes Kopf[2] Narendra Ahuja[1] Jia-Bin Huang[3]

[1]University of Illinois, Urbana-Champaign
{phuang17,n-ahuja}@illinois.edu

[2]Facebook
{matzen,jkopf}@fb.com

[3]Virginia Tech
jbhuang@vt.edu

In this supplementary document, we present additional implementation details and results to complement the main paper. We first describe in detail the formulation for applying DenseCRF to disparity refinement. We then present additional experimental results, including evaluation using scale-invariant errors, results on DeMoN's testing dataset, and the effect of input image resolutions. Finally, we show additional qualitative comparisons with conventional MVS algorithms.

A. DenseCRF

In DenseCRF [3], the pixel-wise classification problem is modeled as a Markov random field characterized by a Gibbs distribution with its energy term being the summation of 1) the unary energy, ϕ_u and 2) the pairwise energy, ϕ_p, over all the pixels in an image, denoted as

$$P(\mathbf{D}|\mathbf{I}) = \frac{1}{Z}\exp\left(-\left(\sum_i \phi_u(d_i) + \sum_{i<j}\phi_p(d_i,d_j|\mathbf{I})\right)\right),$$

where $\mathbf{D}$ is a labeling of the entire image, $\mathbf{I}$ is the reference image, and Z is the normalization factor.

We use the negative logarithm of the probability estimated by the network for each disparity level as the unary energy for pixel-wise priors,

$$\phi_u(d_i) = -\log y_{d_i},$$

and model the pairwise energy as the square of the difference in disparity levels between two predictions multiplied by a bilateral kernel:

$$\phi_p(d_i,d_j|\mathbf{I}) = \frac{|d_i-d_j|^2}{2\sigma_d^2}\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma_{xy}^2} - \frac{\|\mathbf{c}_i-\mathbf{c}_j\|^2}{2\sigma_{rgb}^2}\right),$$

where d_i is the predicted label for i-th pixel, $\mathbf{x}_i$ is the x-y image coordinates, and $\mathbf{c}_i$ is the color of pixel i. Our model encourages the pixels which are spatially close and with similar in color to have closer disparity predictions. In all of our experiments, we use the parameters $\sigma_{xy} = 80$, $\sigma_{rgb} = 15$, $\sigma_d = 10$, and number of iterations $= 5$.

Table 1. Geometric errors and scale-invariant errors on the ETH3D Dataset and the DeMoN's testing dataset.

Method	ETH3D Dataset		DeMoN Dataset	
	Geo.	Sc.-Inv.	Geo.	Sc.-Inv.
DeepMVS	**0.036**	**0.291**	**0.074**	0.364
DeMoN [6]	0.045	0.309	0.096	**0.232**

B. Evaluation using Scale-Invariant Errors

Since DeMoN [6] tends to predict inaccurate scaling factors, the authors in [6] quantify the performance using the *scale-invariant error*, which is defined as

$$\text{sc-inv} = \sqrt{\frac{1}{n}\sum_i(\log\hat{d}_i - \log d_i)^2 - \frac{1}{n^2}\left(\sum_i(\log\hat{d}_i - \log d_i)\right)^2}$$

, where $\hat{d}_i$ and d_i are the estimated disparity and the ground truth disparity at pixel i, respectively. Table 1 shows that our algorithm produces lower scale-invariant errors than DeMoN on the ETH3D dataset.

C. Results on DeMoN's Testing Dataset

We also evaluate our algorithm on the testing dataset proposed in DeMoN [6] consisting of 354 image pairs. In Table 1, we show that our algorithm results in lower geometric errors but higher scale-invariant errors than DeMoN on this dataset. While our algorithm is designed for handling *multiple* images, our method still generates high-quality disparity maps with competitive performance using only *image pairs*. As our method relies on the construction of cost volume rather than direct prediction (e.g., as in [6]), our method has difficulty in hallucinating the disparity values for non-overlapping or the occluded regions. Figure 1 shows several results from the DeMoN testing dataset.

Figure 1. Qualitative comparisons between DeMoN and our algorithm on the testing dataset used in [6].

Table 2. Results on the ETH3D dataset using different image resolutions.

Method	540×360	810×540	1080×720
DeepMVS	0.038	0.036	0.037
COLMAP [4]	0.047	0.046	0.047

D. Effect of Image Resolutions

To show the capability of our algorithm for handling different image resolutions, we compare our method with COLMAP [4] on the ETH3D dataset [5] using three different image resolutions: 540×360, 810×540, and 1080×720 pixels. Table 2 shows that the performance of our approach is not affected by the image resolutions. Under all image resolutions, our method produces lower geometric errors than COLMAP.

E. Additional Qualitative Comparisons

We compare our algorithm with three conventional MVS algorithms, PMVS [2], MVE [1], and COLMAP [4], on the ETH3D dataset in Figure 2. As PMVS generates the 3D point cloud directly, the disparity maps are generated by projecting all the 3D points in the predicted point cloud back to each view with a splatting kernel size of 3×3 pixels. Our algorithm generates complete disparity maps and is able to estimate better disparities in near-textureless regions and reflective regions than the conventional algorithms.

References

[1] S. Fuhrmann, F. Langguth, and M. Goesele. Mve-a multi-view reconstruction environment. In *Eurographics Workshop on Graphics and Cultural Heritage*, 2014. 2, 3

[2] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 2, 3

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *NIPS*, pages 109–117. 2011. 1

[4] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 3

[5] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 2

[6] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 1, 2

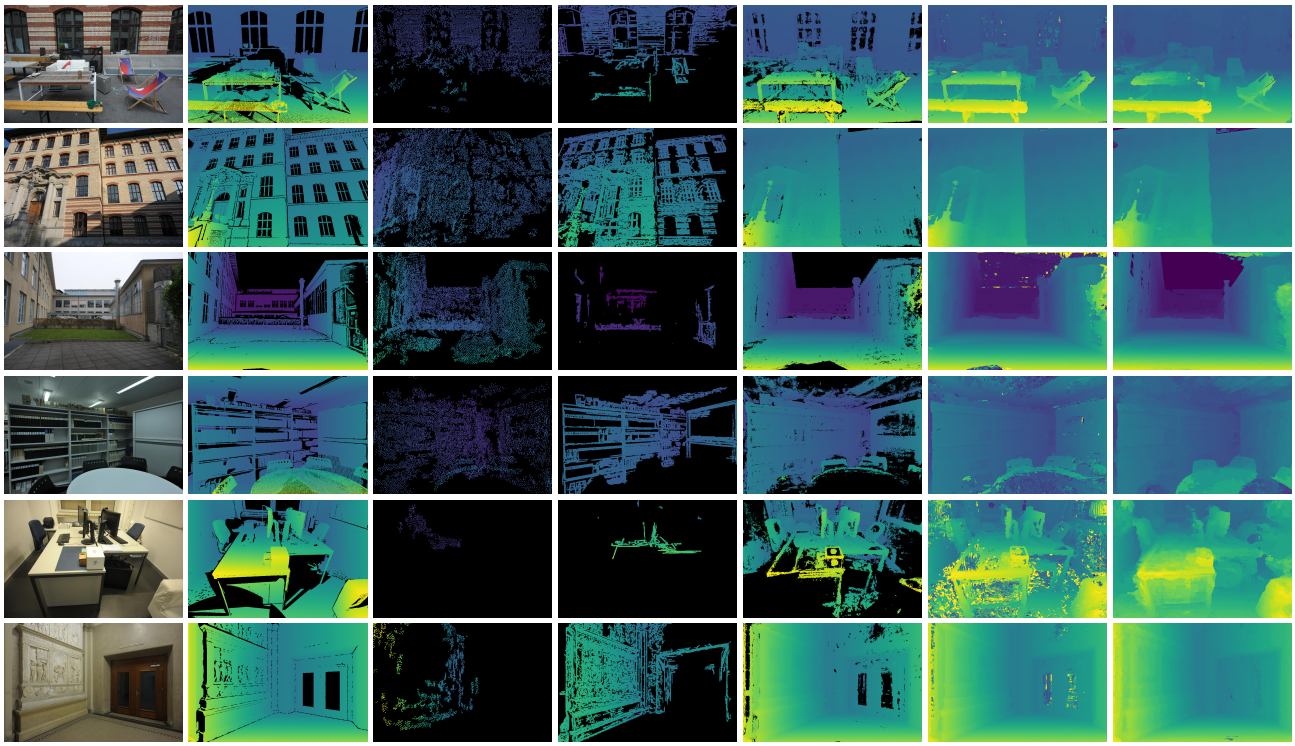

Figure 2. Qualitative comparisons between our approach and conventional MVS algorithms on ETH3D dataset.