Self-supervised Learning of Depth Inference for Multi-view Stereo Supplementary Material

Jiayu Yang[1], Jose M. Alvarez[2], Miaomiao Liu[1]
[1]Australian National University, [2]NVIDIA
{jiayu.yang, miaomiao.liu}@anu.edu.au, josea@nvidia.com

In this supplementary material, we first provide details of the image synthesis loss functions in section 1. In section 2, we show additional qualitative reconstruction results by our method for different scans on the DTU dataset. In section 3, we provide visualization of the pseudo depth labels generated by our self-supervised learning framework. In section 4, we show the pseudo depth labels generated by each iteration of our self-supervised learning framework. In section 5, we provide more ablation experiments. In section 6, we provide more discussions regarding the proposed method.

1. Image synthesis loss

In our approach, as introduced in section 3.2, we use a weighted combination of four loss functions,

$$l_{syn} = \alpha_1 l_g + \alpha_2 l_{ssim} + \alpha_3 l_p + \alpha_4 l_s, \tag{1}$$

where l_g is the image gradient loss, l_{ssim} is the structure similarity loss, l_p is the perceptual loss, l_s is the depth smoothness loss, and α_i sets the influence of each loss.

Image gradient loss is defined as the L1 distance between the gradient of input reference image $\nabla \mathbf{I}_0^l(\mathbf{x})$ and the synthesized image $\nabla \mathbf{I}_{i\to 0}^l(\mathbf{x})$ for each source view i and each pyramid level l,

$$l_g = \sum_{l=0}^{L} \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{x}\in\Omega} ||\nabla \mathbf{I}_{i\to 0}^l(\mathbf{x}) - \nabla \mathbf{I}_0^l(\mathbf{x})||_1. \tag{2}$$

where Ω is the set of valid pixels of the synthesized image.

Structure similarity loss enforces the contextual similarity between a synthesized image and the input reference image. Specifically, we use the Structure Similarity Index [6] to measure the contextual similarity. This index increases as the structure similarity between the images increases, with a range $[-1, 1]$. We formulate the loss as the negative of $SSIM$ between each synthesized image and input reference image,

$$l_{ssim} = \sum_{l=0}^{L} \frac{1}{N} \sum_{i=1}^{N} 1 - SSIM(\mathbf{I}_{i\to 0}^l, \mathbf{I}_0^l). \tag{3}$$

Perceptual loss also encourages high-level contextual similarity between images [2]. This loss is defined as the L1 distance in the feature space of a shared weight perceptual network taking each image as input [2]. In our experiments, we use a VGG model [4] and extract features from 3_{th},8_{th},15_{th} and 22_{th} layers. Therefore, we formulate the loss as follows,

$$l_p = \sum_{l=0}^{L} \frac{1}{N} \sum_{i=1}^{N} \sum_{j\in[3,8,15,22]} ||VGG(\mathbf{I}_{i\to 0}^l, j) - VGG(\mathbf{I}_0^l, j)||_1. \tag{4}$$

Depth smoothness loss encourages local depth smoothness. This term encourages depth smoothness with respect to the alignment of image and depth discontinuities, which is measured by the gradient of color intensity of input reference image. We define this loss as follows,

$$l_{sm} = \sum_{l=0}^{L} \sum_{\mathbf{x}\in\Omega} |\nabla_u \tilde{D}^l(\mathbf{x})| e^{-|\nabla_u \mathbf{I}_0^l(\mathbf{x})|} + |\nabla_v \tilde{D}^l(\mathbf{x})| e^{-|\nabla_v \mathbf{I}_0^l(\mathbf{x})|} \tag{5}$$

where ∇_u and ∇_v refer to the gradient on x and y direction, and $\tilde{D} = D/\bar{D}$ is the mean-normalized inverse depth [1].

2. Qualitative results on DTU dataset

Fig. 1 shows additional reconstruction results by our self-supervised method on the DTU dataset. As shown, our self-supervised model can achieve similar reconstruction results comparing with the supervised model.

3. Pseudo depth labels visualization

Fig. 2 and Fig. 3 provide visualization of pseudo depth labels generated by our self-supervised learning framework. As shown, our method can generate high quality pseudo depth labels for rich-texture areas. However, our method can not generate pseudo depth labels for severely textureless regions.

Figure 1. **DTU Dataset.** Representative point cloud results. Best viewed on screen.

Figure 2. Pseudo depth labels generated by the self-supervised learning framework. Areas with no pseudo depth labels or no ground-truth depth are marked as blue in the error visualization. Best viewed on screen.

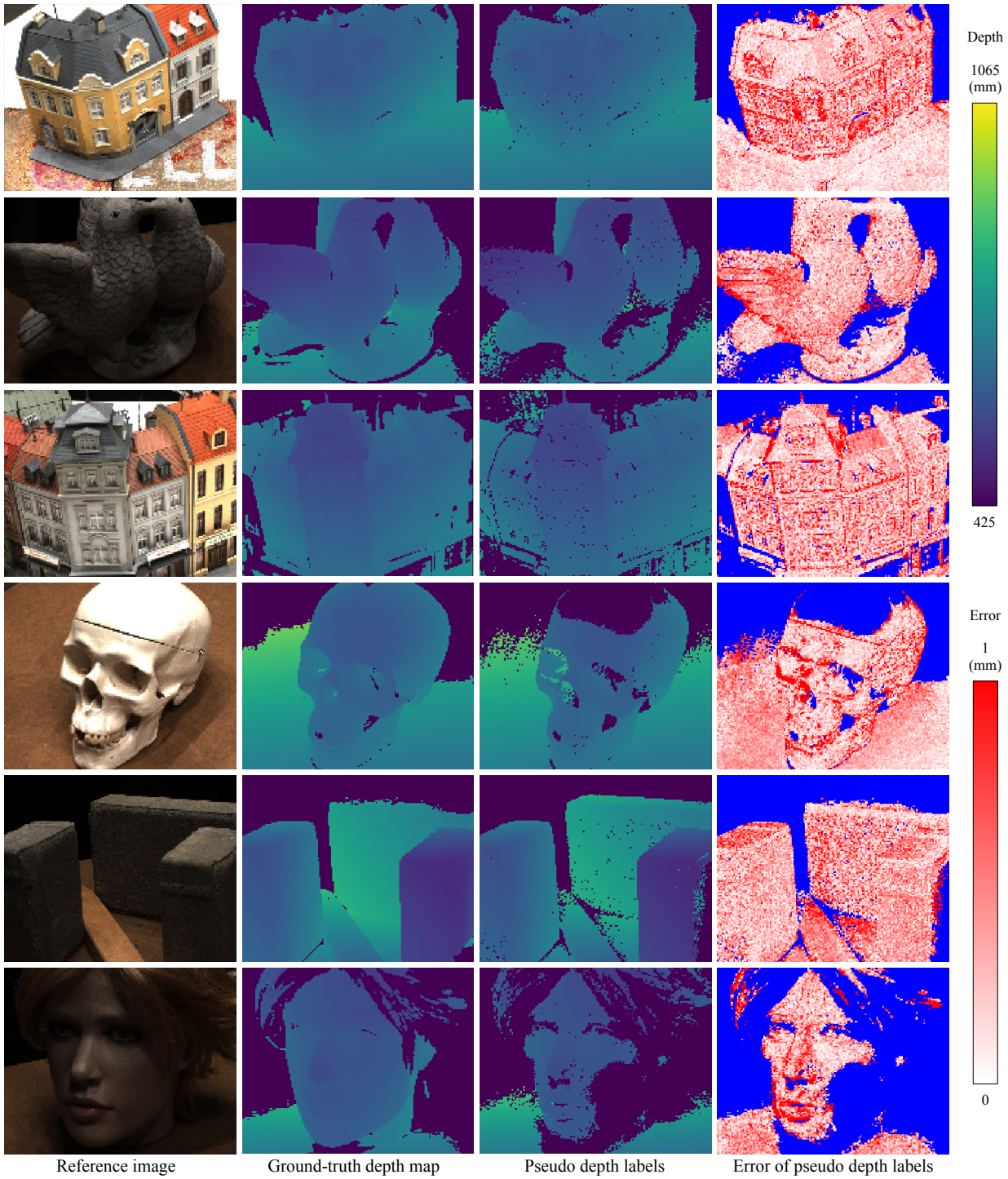
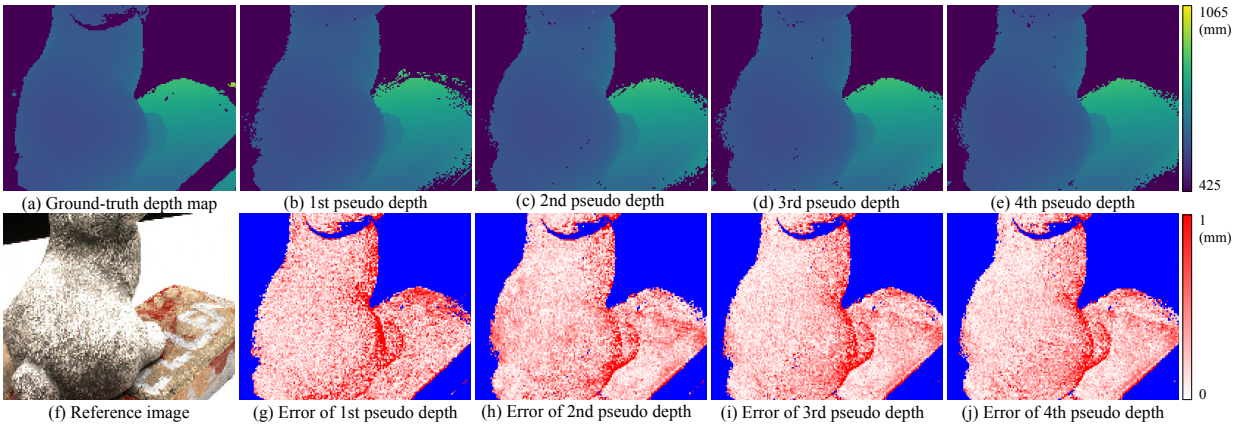


Figure 3. Pseudo depth labels generated by the self-supervised learning framework. Areas with no pseudo depth labels or no ground-truth depth are marked as blue in the error visualization. Best viewed on screen.



(a) Ground-truth depth map (b) 1st pseudo depth (c) 2nd pseudo depth (d) 3rd pseudo depth (e) 4th pseudo depth

(f) Reference image (g) Error of 1st pseudo depth (h) Error of 2nd pseudo depth (i) Error of 3rd pseudo depth (j) Error of 4th pseudo depth

Figure 4. Pseudo depth labels generated by each iteration of the iterative self-supervised learning framework. (a) Ground-truth depth map. (b-e) Pseudo depth labels generated by 1-4 iteration of the self-supervised learning framework. (f) Reference image. (g-j) Error of each iteration of the pseudo depth labels. Areas with no pseudo depth labels are marked as blue in the error visualization. Best viewed on screen.

4. Pseudo labels generated on each iteration

Fig. 4 shows visualization of pseudo depth labels generated by the different iterations of the self-supervised learning framework. As shown, pseudo depth labels tend to become stable after the second iteration.

5. Additional Ablation Experiments

Use geometric MVS methods as initialization.

We use a traditional MVS method OpenMVS to generate the pseudo labels to replace our unsupervised learning process on DTU dataset. Specifically, we render a depth map from the mesh generated by OpenMVS and treat it as the label for the first iteration of the iterative training. As shown in Tab. 1, using OpenMVS outputs as pseudo labels can achieve similar performance as our proposed initialization method after 3 iterations of self-supervised learning.

Method\f-score	Init	Iter. 1	Iter. 2	Iter. 3
OpenMVS Init.	73.70%	76.86%	87.95%	88.02%
Ours Init.	77.06%	88.16%	88.42%	88.49%

Table 1. Performance with different initialization method.

6. Discussions

Runtime Despite the limitation on texture-less area mentioned in main paper, another limitation appears on the runtime of proposed self-supervised learning method. Each iteration of the self-training process takes around 15 hours on our machine, which adds up to days for several iterations of self-training or fine-tuning on novel data. For comparison, the supervised CVP-MVSNet takes around 10 hours. A classical MVS method such as the OpenMVS even does not need any training to achieve compromised results. Improving the efficiency of the proposed learning method can be an direction of future research.

Limitation of geometric processing Another concern appears on the geometric filtering and fusion methods we used to refine the pseudo depth label. The traditional geometric methods such as consistency check and Screened Poission Surface Reconstruction have their limitations. Specifically, the SPSR is a non-learning, indifferentiable and time-consuming step, which might be too heavy for fine-tuning on novel data. It also have limited performance on very complex scenes. Improving pseudo label processing methods can be a direction of future research.

References

[1] Rui Zhu Simon Lucey Chaoyang Wang, Jose Miguel Buenaposada. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 1

[2] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. 1

[3] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *ArXiv*, 2019. 2

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[5] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, 2020. 2

[6] H. R. Sheikh Z. Wang, A. C. Bovik and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 1