Just a Few Points are All You Need for Multi-view Stereo: A Novel Semi-supervised Learning Method for Multi-view Stereo

Taekyung Kim[1], Jaehoon Choi[2], Seokeon Choi[1], Dongki Jung[3], Changick Kim[1]
[1]Korea Advanced Institute of Science and Technology
[2]University of Maryland
[3]NAVER LABS
{tkkim93, seokeon, changick}@kaist.ac.kr, kevchoi@umd.edu, dongki.jung@naverlabs.com

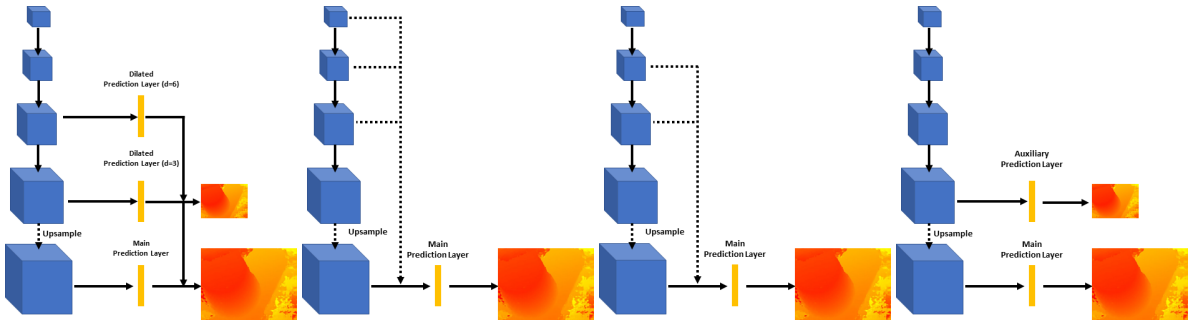


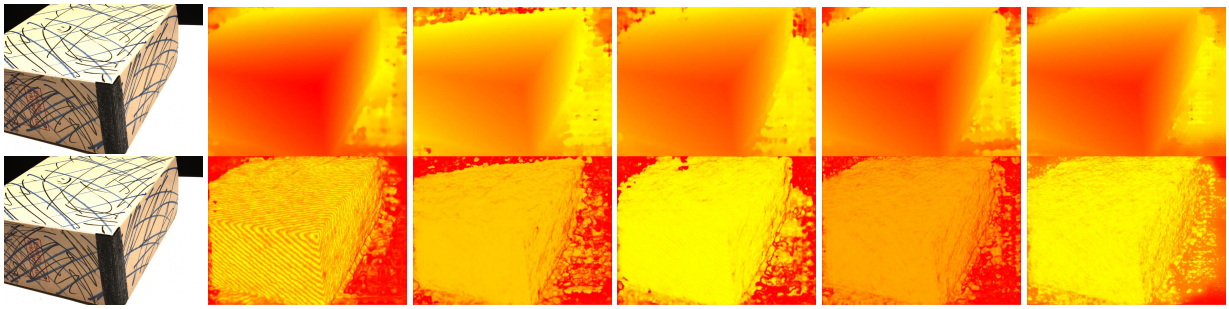
Figure 1: Description of the structures of the failed prediction layers Dilated-PL, CPL, SCPL, and DPL represent for dilated prediction layer, contextual prediction layer, shallow contextual prediction layer, and double prediction layer.

1. Failures in Prediction Layer Design

We introduce some representatives of the failed prediction layer structures. We tried several types of prediction layers and their variants to address the errors in occluded regions through richer contextual information, and we share some representative failure cases among them: dilated prediction layer (Dilated-PL), contextual prediction layer (CPL), shallow contextual prediction layer (SCPL). Each structures are described in Fig. 1. Figure 2 shows some examples of their prediction results with confidence maps and Fig. 3 shows some examples of their 3D reconstruction results. As shown in Fig. 2(c) and (d), utilizing context vectors from higher-level cost volumes rather hampers details near the boundary. In contrast, since Dilated-PL is designed to refer larger receptive field through dilated convolution layers, each predicted depth value is harmonious with its neighbors and seems to be predicted accurately overall. However, some less confident regions show slightly inaccurate depth values compare to their neighbors. Moreover, Dilated-PL also shows inaccurate predictions in textureless regions during 3D reconstruction, as shown in Fig. 3. All these results mainly originate from the phenomena that MVS networks try to overuse dilated convolution layers or contextual vectors during prediction. Specifically, MVS networks enforce to learn how to reduce the depth regression loss for every pixel in the receptive field of the ground-truth position, and it eventually causes blurred or erroneous predictions near the edges and inaccurate predictions on the textureless regions. These results verify that enforcing to learn context encoding can contaminate the discriminability of the mvs network, so it is necessary to separate erroneous regions like edges and occlusions and address them through the propagation of the accurate depth values, which backs up our strategy.

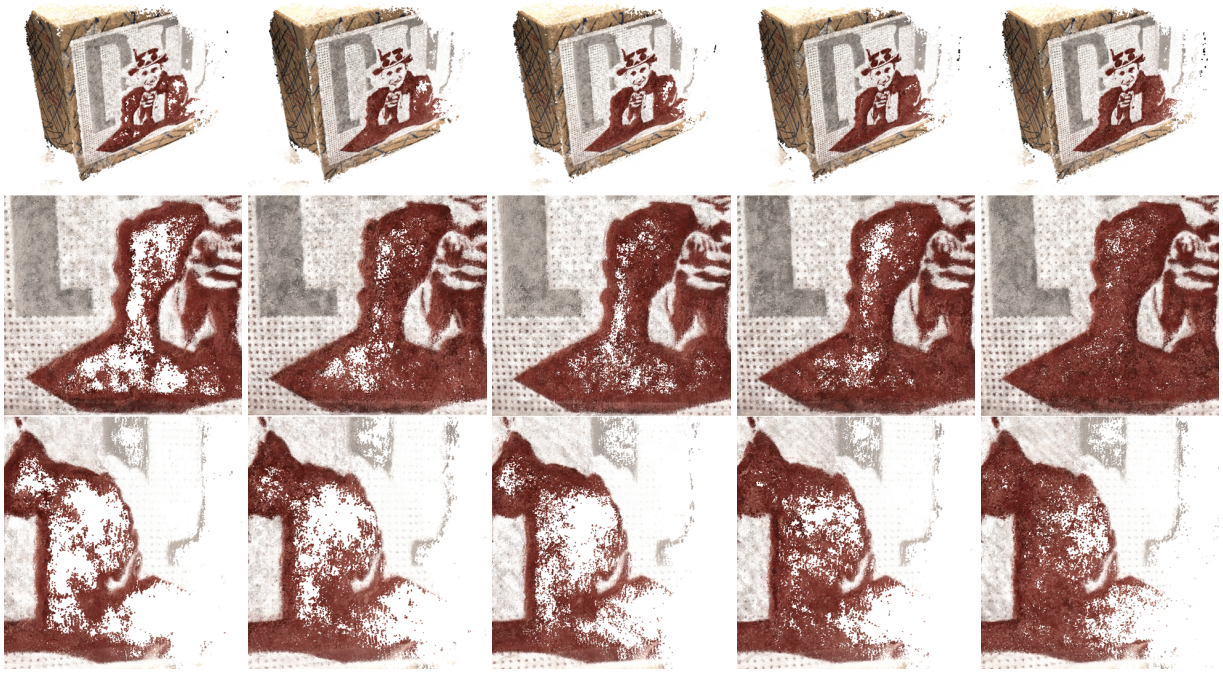
2. 3D reconstruction results on the DTU dataset

We visualized the 3D reconstruction results of the SGT-MVSNet trained with sparse ground-truth of the DTU dataset [1] with sampling ratio 1×10^{-5} in Fig. 4. Though our network is trained only with tens of ground-truth points for each 3D structures, it shows reasonable reconstruction results. Also, unlike other MVS networks, we can easily refine our model by collecting additional multi-view scenes



(a) Reference (b) Dilated-PL (c) CPL (d) SCPL (e) DPL (f) Ours

Figure 2: Visualization of the prediction results and their confidence maps of the failed prediction layer structures. We used the 25th scene of *scan10* in the DTU dataset [1]. Dilated-PL, CPL, SCPL, and DPL represent for dilated prediction layer, contextual prediction layer, shallow contextual prediction layer, and double prediction layer.



(a) Dilated-PL (b) CPL (c) SCPL (d) DPL (e) Ours

Figure 3: Visualization of the 3D reconstruction results of the failed prediction layer structures. We used *scan13* of the DTU dataset [1] for visualization. Dilated-PL, CPL, SCPL, and DPL represent for dilated prediction layer, contextual prediction layer, shallow contextual prediction layer, and double prediction layer.

with sparse ground-truths that can be easily collected. Thus, our method can be easily extended to autonomous driving and robot perceptions.

References

[1] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014.

[2] 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.

Figure 4: Qualitative results of the intermediate set in the tank and temples dataset [2]. The tested SGT-MVSNet is trained with sparse ground-truth of the DTU dataset [1] sampled with 1×10^{-5} ratio.