Supplementary Material for KD-MVS: Knowledge Distillation Based Self-supervised Learning for Multi-view Stereo

Yikang Ding[1,2] Qingtian Zhu[1] Xiangyue Liu[1] Wentao Yuan[1]
Haotian Zhang[1⋆] Chi Zhang[1]

[1]Megvii Research [2]Tsinghua University

1 More Insights of KD-MVS

In this section, we discuss the potential reasons why self-supervised methods can obtain comparable (even better) results compared to supervised methods. Here we elaborate this from the following two perspectives.

(a) Self-supervised methods are able to generate accurate pseudo labels with cross-view check. According to Eq. (4) of the paper, only the inliers (whose depth prediction is accurate) can be kept given a strict threshold. We also visualize the depth error of the pseudo depth in Fig. 8, which is relevant to the overall error of point cloud results. The first and the second row indicate the pseudo depth has already been to some extent accurate in most scenes. Similar conclusion can also be inferred from the Tab. 7(2)(3) of the paper.

(b) The pseudo labels have advantages over GT in some aspects. Generally speaking, the datasets contain many pixels (namely training samples) that are normally textureless and considered as "toxic" for training. For example, if we force the network to estimate depth values for a purely white region, which should be inherently unpredictable, the network will be confused since there are simply no valid features extracted. Fig. 1 shows visualized comparisons of GT depth and pseudo depth, the red boxes highlight the textureless regions, which are nearly unpredictable. The pseudo depth maps contain fewer misleading regions by filtering the outliers with cross-view check. By reducing the toxic samples, the training process will be more stable and the performance will be improved. Fig. 2 shows the comparison of several metrics in the training phase. Compared with using the original depth (the orange curve), reducing the toxic samples by using the mask of the pseudo depth makes the training phase more stable and helps the student model converge faster.

Besides, we propose to leverage the probabilistic knowledge, which is verified to be effective in classification task [8,6,10,11]. Some works [5,13] call this probabilistic knowledge as dark knowledge and believe they contain inter-class information. As MVS can also be handled as a classification task, it can benefit from probabilistic knowledge too.

This work is done by the first four authors as interns at Megvii Research.

⋆ Corresponding author (zhanghaotian@megvii.com).

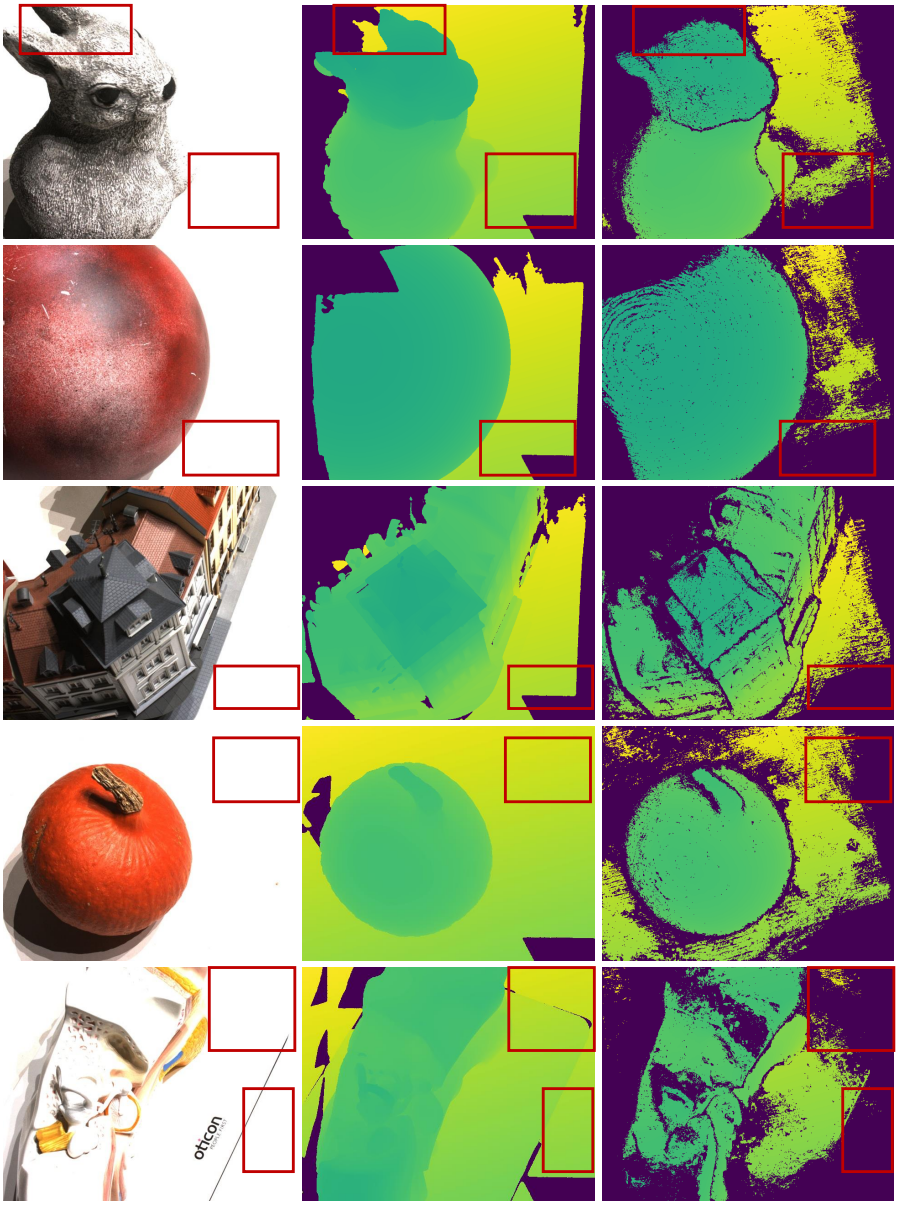

Fig. 1: Visualized comparisons between GT depth and pseudo depth. Red boxes indicate textureless regions, which are nearly unpredictable. Pseudo depth maps contain less textureless (misleading) regions by filtering the outliers with cross-view check.

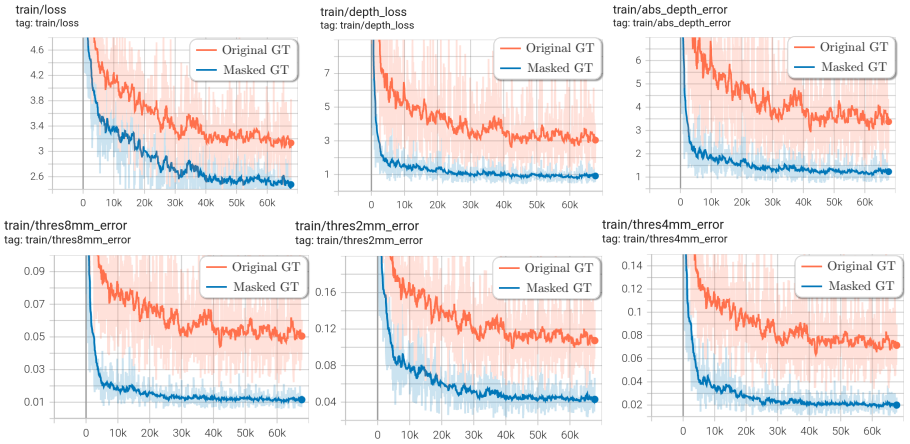

Fig. 2: Comparisons of training processes in several metrics. Original GT indicates training with the original ground-truth depth of DTU dataset [1]. Masked GT is obtained by masking the original GT with the mask of pseudo depth. Reducing the toxic samples makes the training phase more stable.

2 More Details of Experiment Settings

2.1 Fine-tuning on BlendedMVS

As done in many supervised methods [12,15,3,2,9], we train the student model on BlendedMVS dataset [14] for better performance and fair comparison. Concretely, we use the student model trained on DTU dataset [1] as the teacher model and generate pseudo probabilistic labels for BlendedMVS dataset. A new student model is then trained using the pseudo labels of both BlendedMVS and DTU from scratch. The process of fine-tuning still follows the self-supervised scheme of KD-MVS and leverages no extra manual annotation.

2.2 Ablation Study on the Insights of effectiveness

We perform ablation study to verify the insights of effectiveness in Sec. 5.1 of the paper, and the results are listed in Tab. 7. The *Mask* indicates whether use the validated masks, which are generated by performing cross-view check on the raw depth maps of the teacher model. For (2) in Tab. 7, since the GT depth itself has a mask, we take the intersection of the two masks as the final mask. The *Depth* indicates which kind of depth is used to train the student model. Comparing the results of (2) and (3), it can be found that the pseudo depth label is accurate. The *Loss* indicates which kind of loss function is used to train the student model. When the GT depth is used, we can only use the ℓ-1 loss. When the distillation loss is used, we use the probabilistic encoding to generate the pseudo probability distribution. Comparing the results of (3) and (4), we can find that the distillation loss brings a big performance gain.

Table 1: Ablation study on number of input views N and image resolution $H \times W$ on DTU evaluation set [1] (**lower is better**). We use the same student model and keep the other settings fixed.

N	$H \times W$	Acc.(mm)	Comp.(mm)	Overall(mm)
3	864×1152	0.366	0.316	0.341
5	864×1152	**0.359**	**0.295**	**0.327**
7	864×1152	**0.359**	0.299	0.329
9	864×1152	0.360	0.301	0.331
5	512×640	0.405	0.313	0.359

Table 2: Ablation study on different settings of the threshold parameters in cross-view check. We use the same raw depth of the self-supervised teacher model to genrate the pseudo labels, and show the results of the student model on DTU evaluation set [1] (**lower is better**).

	τ_{conf}	τ_{reproj}	τ_{geo}	Acc.(mm)	Comp.(mm)	Overall(mm)
(a)	0.30	2.0	0.020	0.410	0.382	0.396
(b)	0.20	1.5	0.015	0.384	0.320	0.352
(c)	0.15	1.0	0.010	0.359	**0.295**	**0.327**
(d)	0.10	0.5	0.005	**0.350**	0.306	0.328

3 Ablation Study on Hyper-parameters

3.1 Number of Views & Input Resolution

We perform ablation study against the number of input views N and input resolution $H \times W$ on DTU evaluation set [1], and the results are shown in Tab. 1.

3.2 Thresholds in Cross-view Check

As introduced in Sec. 5.3 of the paper and Sec. 1 of the supp. materials, the quality of pseudo probability distribution depends on the cross-view check strategy and the relevant hyper-parameters. We perform ablation study on the three threshold parameters in cross-view check, and the results are shown in Tab. 2.

4 More Point Cloud Results

We visualize more point cloud results of KD-MVS (applied with CasMVSNet [4]) on DTU evaluation set [1], Tanks and Temples benchmark [7] respectively in Fig. 3 and Fig. 4.

5 Failure Cases

As discussed in Sec. 5.3 of the paper, KD-MVS may face challenges when training student models under the following situations:

(a) When KD-MVS is applied on a relatively small-scale dataset. We attempted to generate the pseudo probability distribution on the intermediate set of the Tanks and Temples dataset [7] ($\sim 2K$ samples) for training student models. However, when trained on the dataset alone, the performance of the student model degrades significantly compared to the model trained on BlendedMVS dataset [14] ($\sim 17K$ samples) alone. The potential reason is that small-scale datasets cannot provide sufficient data diversity for knowledge distillation, so the student model is not able to learn robust feature representations and performs unsatisfactorily.

(b) When the thresholds in cross-view check are inappropriate. The performance of the student model relies on the quality of the generated pseudo labels, which is greatly affected by the thresholds set in cross-view check. Tab. 2 shows when these thresholds are inappropriate, the performance of the student model will degrade.

Fig. 3: Point cloud results on DTU dataset [1].

Fig. 4: Point cloud results on Tanks and Temples benchmark [7].

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**(2), 153–168 (2016) 3, 4, 6
2. Ding, Y., Li, Z., Huang, D., Li, Z., Zhang, K.: Enhancing multi-view stereo with contrastive matching and weighted focal loss. arXiv preprint arXiv:2206.10360 (2022) 3
3. Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X.: Transmvsnet: Global context-aware multi-view stereo network with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8585–8594 (2022) 3
4. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020) 4
5. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015) 1
6. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017) 1
7. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017) 4, 5, 7
8. Li, T., Li, J., Liu, Z., Zhang, C.: Few sample knowledge distillation for efficient network compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14639–14647 (2020) 1
9. Liao, J., Ding, Y., Shavit, Y., Huang, D., Ren, S., Guo, J., Feng, W., Zhang, K.: Wt-mvsnet: Window-based transformers for multi-view stereo. arXiv preprint arXiv:2205.14319 (2022) 3
10. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 268–284 (2018) 1
11. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019) 1
12. Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G.: Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6187–6196 (2021) 3
13. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020) 1
14. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1790–1799 (2020) 3, 5
15. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. British Machine Vision Conference (BMVC) (2020) 3