Supplemental Material
Self-distilled Feature Aggregation for Self-supervised Monocular Depth Estimation

Zhengming Zhou[1,2] and Qiulei Dong[1,2,3]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China
`zhouzhengming2020@ia.ac.cn`
`qldong@nlpr.ia.ac.cn`

A Datasets and Metircs

The two datasets used in this work are introduced in detail as follows:

- KITTI [4] contains the rectified stereo image pairs captured from a driving car. We use the Eigen split [2] to train and evaluate the proposed network (called SDFA-Net), which consists of 22600 stereo image pairs for training and 697 images for testing. Additionally, we also evaluate SDFA-Net on the improved Eigen test set, which consists of 652 images and adopts the high-quality ground-truth depth maps generated with the method in [8]. The images are resized into the resolution of 1280×384 at both the training and inference stages, while we assume that the intrinsics of all the images are identical.
- Cityscapes [1] contains the stereo pairs of urban driving scenes, and we take 22972 stereo pairs from it for jointly training SDFA-Net. When SDFA-Net is trained on both the KITTI and Cityscapes datasets, we crop and resize the images from Cityscapes into the resolution of 1280×384. Considering that the baseline length in Cityscapes is different from that in KITTI, we scale the predicted disparities on Cityscapes by the rough ratio of the baseline lengths in the two datasets.

For the evaluation on both the raw and improved KITTI Eigen test set [2], we use the center crop proposed in [3] and the standard depth cap of 80m. The following metrics are used:

- Abs Rel: $\frac{1}{N}\sum_i \frac{\left|\hat{D}_i - D_i^{gt}\right|}{D_i^{gt}}$,
- Sq Rel: $\frac{1}{N}\sum_i \frac{\left|\hat{D}_i - D_i^{gt}\right|^2}{D_i^{gt}}$,
- RMSE: $\sqrt{\frac{1}{N}\sum_i \left|\hat{D}_i - D_i^{gt}\right|^2}$,

- logRMSE: $\sqrt{\frac{1}{N}\sum_i \left|\log\left(\hat{D}_i\right) - \log\left(D_i^{gt}\right)\right|^2}$,
- Threshold (Aj): % $s.t.$ $\max\left(\frac{\hat{D}_i}{D_i^{gt}}, \frac{D_i^{gt}}{\hat{D}_i}\right) < a^j$,

where $\{\hat{D}_i, D_i^{gt}\}$ are the predicted depth and the ground-truth depth at pixel i, and N denotes the total number of the pixels with the ground truth. In practice, we use $a^j = 1.25, 1.25^2, 1.25^3$, which are denoted as A1, A2, and A3 in all the tables.

B Ablation studies on the self-distilled training strategy

We conduct more ablation studies on the KITTI dataset [4] for verifying the effectiveness of the proposed self-distilled training strategy. We firstly train a model that uses the Offset-based Aggregation (OA) modules in the decoder with a straightforward Self-Distilled training strategy ('Swin†+OA (SD)'). Specifically, since the OA module does not have the 'Distilled data path' (described in Section 3.2), this model is trained under the self-distillation manner by using the 'Raw data path' twice in the two steps of each training iteation (described in Section 3.3). The corresponding results are shown in Lines 1-3 of Table A1. The results predicted by the 'Swin†+OA' (without the self-distillation) and the full model 'SDFA-Net' are also reported for comparison. It can be seen that there are only slight improvements under four metrics when the straightforward self-distilled strategy is used. Our full model performs best under all the metrics, which indicates that the proposed self-distilled training strategy with the two data paths is more helpful for predicting accurate depths.

To verify the effectiveness of the principle masks and feature flipping strategy defined under the self-distilled training strategy, we conduct further ablation studies by omitting the visible principle mask ('w/o. M_v^l'), the photometric principle mask ('w/o. M_p^l'), both of the masks ('w/o. Masks'), and the feature flipping ('w/p. Flip'), respectively. The results shown in Lines 4-7 of Table A1 demonstrate that both of the masks could improve the accuracy, and the principle mask has a stronger influence than the visible mask. It also can be seen that the flipped features are more helpful for improving the accuracy. Since the weights of 'Raw & distilled data path' (described in Section 3.2) in SDFA are shared at the self-supervised and self-distilled forward propagation steps, they are trained by minimizing both the image synthesis loss and self-distilled loss. Therefore, the distilled depths predicted in the Self-distilled Forward Propagation are still suffer from occlusions to some extent. Specifically, the model trained with left images of stereo pairs under a self-supervised manner always predicts inaccurate depths on the left side of objects, whether the input features are flipped or not. The feature flipping strategy could alleviate the occlusion problem because the inaccurate depths would occur on the right side of objects in the distilled depth maps when the input features are flipped, which are not on the real occluded regions and could be corrected by the reliable pseudo depths. The visualization results of the depths predicted by the model trained with and without feature flipping shown in Figure A1 also illustrate that our full model predicts sharper and more accurate depths on occluded regions compared to 'SDFA-Net w/o. Flip'. It demonstrates the effectiveness of the feature flipping strategy.

C Ablation study on the backbone

To further explore the effectiveness of the proposed SDFA module on different backbones, we train a new baseline model comprises a ResNet18 [6] as the encoder and a raw decoder proposed in [5] (Res18+Raw). Then we use the ResNet18 [6] to replace the modified Swin-transformer [7] in SDFA-Net (SDFA-Net (Res18)). The corresponding results are reported in Lines 8-9 of Table A1. It can be seen that SDFA also could improve the performance of the ResNet18-based model.

Table A1. Additional quantitative results on the raw KITTI Eigen test set in the ablation study.

Method	Abs Rel↓	Sq Rel↓	RMSE↓	logRMSE↓	A1↑	A2↑	A3↑
Swin†+OA	0.091	0.554	4.082	0.174	0.899	0.967	0.984
Swin†+OA (SD)	0.092	0.549	4.003	0.174	0.900	0.967	0.985
SDFA-Net	0.090	0.538	3.896	0.169	0.906	0.969	0.985
SDFA-Net w/o. M_v^l	0.090	0.544	3.928	0.169	0.905	0.969	0.985
SDFA-Net w/o. M_p^l	0.096	0.566	4.013	0.173	0.902	0.968	0.985
SDFA-Net w/o. Masks	0.095	0.583	4.039	0.175	0.899	0.967	0.985
SDFA-Net w/o. Flip	0.097	0.583	4.078	0.177	0.897	0.966	0.984
Res18+Raw	0.102	0.644	4.341	0.187	0.880	0.960	0.982
SDFA-Net (Res18)	0.101	0.636	4.226	0.180	0.891	0.964	0.984

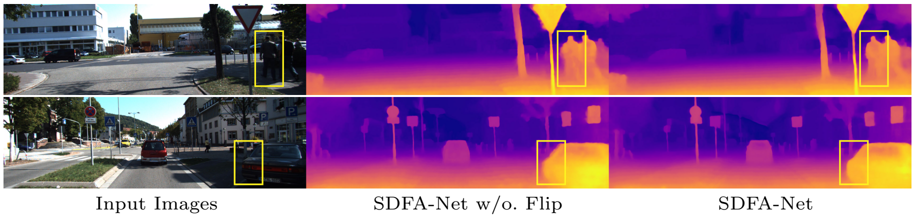

Fig. A1. Visualization results of the SDFA-Net trained without/with the feature flipping strategy on KITTI.

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
2. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
3. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV. pp. 740–756 (2016)
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR. pp. 3354–3361 (2012)
5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV. pp. 3828–3838 (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
8. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 2017 international conference on 3D Vision (3DV). pp. 11–20 (2017)