Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes

Juan L. Gonzalez
juanluisgb@kaist.ac.kr

Munchurl Kim
mkimee@kaist.ac.kr

School of Electrical Engineering
Korea Advanced Institute of Science and Technology

1 Introduction

We present a self-supervised two-stage method for the accurate learning of single image depth estimation (SIDE) from stereo pairs with novel Mirrored Exponential Disparity (MED) representations and Mirrored Occlusion Modules (MOM). We propose to "Forget About the LiDAR" (FAL), or 3D laser scanning, for the supervised training of SIDE DCNNs and show that our self-supervised method achieves superior performance than the state-of-the-art (SOTA) self-, semi- and fully supervised methods on the challenging KITTI dataset [3]. We recognize that instead of focusing our efforts on developing unnecessary complex (and large) DCNN architectures, it is more worthwhile to focus on loss functions and training strategies that can better exploit the geometrical dependencies in the data for effective self-supervision. Our network, called FAL-net, incorporates our proposed MED representations and outperforms the most recent SOTA methods of [2, 3], with almost 8x fewer model parameters and 3x faster inference times for full-resolution depth maps. The main contributions of our work are summarized as follows:

1. A novel Mirrored Occlusion Module (MOM), which is a multi-view occlusion mask generation module. The generated masks are very realistic and are used to filter the invalid image regions due to parallax for two views with known (or estimated) camera poses.
2. A new 2-step training strategy: Firstly, we train our FAL-net for plain stereoscopic view synthesis, that is, the synthetic right view is penalized in all image regions; Secondly, we train our FAL-net for SIDE using our MOM to remove the burden of learning the synthesis of right-occluded contents (not related to depth) and provide self-supervision signals for the left-occluded regions which are ignored in the reconstruction losses.
3. We shed light on the effectiveness of the exponential disparity representations for self-supervised SIDE. This small change from the linear to the exponential domain makes our FAL-net, even without MOM, perform surprisingly well, compared to the recent SOTA methods.

2 Method

Self-supervised learning of SIDE is commonly carried out by minimizing reconstruction errors between depth-guided synthetic images and reference views. However, reducing such objective loss functions involves estimating the occluded regions' contents, which degrades the networks' performance on the depth estimation task. Previous methods fail in effectively making the occluded regions transparent to the networks, as the geometrical dependencies of the given views are not effectively considered. Moreover, previous works methods become overloaded with the task of predicting uncertainty masks or occluded contents, leading to the waste of their learning capacities for SIDE during training time. To solve this issue, our FAL-net with MED, MOM, and our new two-step training strategy is proposed.

Network architecture. Our simple, yet effective FAL-net architecture, which incorporates a MED representation in its output, is a 6-stage auto-encoder with one residual block after each strided

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1: Comparison of existing SIDE methods on the improved KITTI Eigen Test Split [1]

Methods	Sup	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	$a^1 \uparrow$	Methods	Sup	abs rel	sq rel	rmse	$\text{rmse}_{\log}$	$a^1 \uparrow$
DORN [2]	D	0.072	0.307	2.727	0.120	0.932	PackNet [4]	V	0.071	0.359	3.153	0.109	**0.944**
DepthHints [5]	S_{SGM}	0.074	0.364	3.202	0.114	0.936	*Our FAL-net*	S	**0.068**	**0.276**	2.906	**0.106**	**0.944**

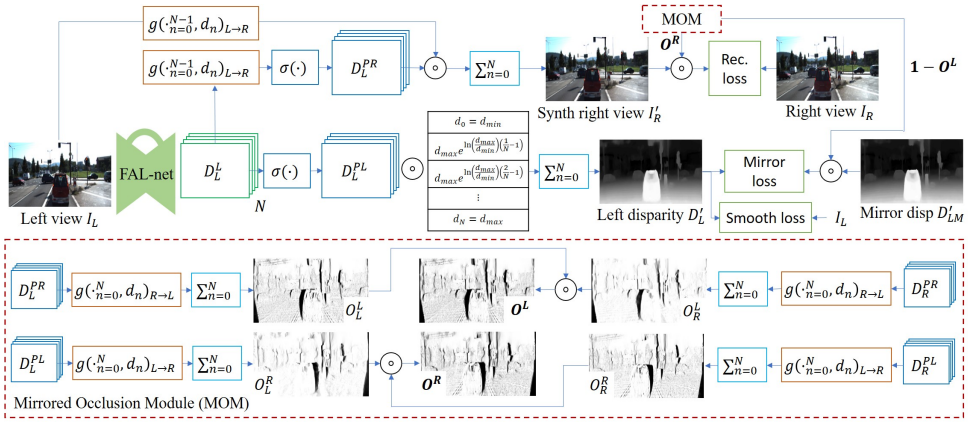

Figure 1: Our proposed training strategy and novel Mirrored Occlusion Module (MOM).

convolution in the encoder side and skip-connections between the encoder and decoder. Our FAL-net maps a single left-view input I_L to a N-channel "disparity logit" volume D_L^L, as shown in Fig. 1. Such "disparity logits" can be channel-wise soft-maxed to form a mirrored exponential disparity (MED) "probability volume" D_L^{PL}, which can be quantized and sum-reduced to give rise to the final predicted depth map as shown in Fig. 1. However, our disparity logits can also be progressively shifted to the right camera and soft-maxed, generating the right-from-left MED prob. volume D_L^{PR}. The element-wise multiplication of D_L^{PR} with equally warped N versions of I_L, followed by a sum-reduction operation, produces a synthetic right view I'_R, which we use to train our network for stereoscopic synthesis.

Exponential quantization. We observed significant improvements by adopting exponential disp. quantization, which is reasonable, as linear quantization of disparity assigns most sampling positions to the very close-by objects due to the inverse relation between disparity and depth.

Mirrored Occlusion Module. Our novel Mirrored Occlusion Module (MOM) is a multi-view occlusion mask generation module that allows our FAL-net to directly learn SIDE by cross-computing occlusion maps from the MED probability distributions of two training images with known (or estimated) camera positions. The process for computing the occlusion masks O_L and O_R is illustrated at the bottom of Fig. 1.

Training strategy. We define a two-step training strategy. In the first step, we train our FAL-net for view synthesis with l_1, perceptual, and smoothness losses and keep a fixed copy of the trained model. In the second step, enabled by our Mirrored Occlusion Module, we fine-tune our FAL-net for inverse depth (disparity) estimation with an occlusion-free reconstruction loss, smoothness loss, and a "mirror loss". Our mirror loss uses a mirrored disparity prediction D'_{LM}, generated by the fixed copy of the model, to provide self-supervision only to the regions that are occluded in the right view but visible in the left view, as shown in Fig. 1.

3 Results and Conclusion

In Table 1, we show that state-of-the-art SIDE can be achieved by light and straightforward auto-encoder networks that incorporate MED representations in their output layers. Qualitatively, our FAL-net generates sharper and more consistent depth estimates. Quantitively, our FAL-net outperforms all previous methods in most (if not all) metrics. Our two-stage training strategy with our Mirror Occlusion Module (MOM) aids in making the network learn precise depth instead of just view-synthesis. Furthermore, our method outperforms the DORN [2] supervised baseline by a large margin, which suggests we can "forget about the LiDAR" for the supervision of SIDE networks. Moreover, any task that requires proper handling of occluded regions caused by rigid motions such as learning of stereo disparity, SIDE from videos, and optical flow can benefit from our method.

References

[1] 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)

[2] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)

[3] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)

[4] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[5] Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2162–2171 (2019)