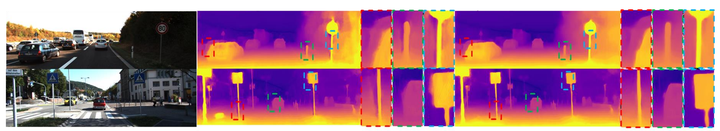
Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes

Juan Luis Gonzalez Bello
juanluisgb@kaist.ac.kr
Munchurl Kim
mkimee@kaist.ac.kr

Introduction

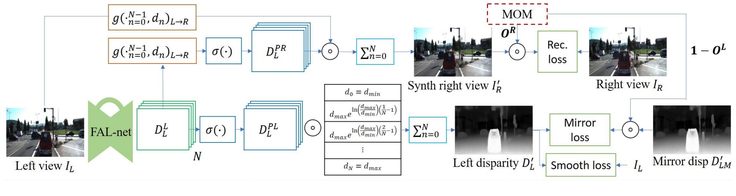
Self-supervised depth estimators have recently shown results comparable to the supervised methods on the challenging single image depth estimation (SIDE) task, by exploiting the geometrical relations between target and reference views in the training data. However, previous methods usually learn forward or backward image synthesis, but not depth estimation, as they cannot effectively neglect occlusions between the target and the reference images. Previous works rely on rigid photometric assumptions or on the SIDE network to infer depth and occlusions, resulting in limited performance.



Depth estimates from our proposed FAL-net with and without our novel Mirror Occlusion Module (MOM)

In this paper, we propose to “Forget About the LiDAR” (FAL), for the supervised training of single image depth estimators (SIDE) 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. Our main contributions are:

- A **novel Mirrored Occlusion Module (MOM),** which is a multi-view occlusion mask generator module. The computed masks are very realistic and are used to filter the invalid image regions due to parallax.
- A **new 2-stage training strategy**: Firstly, we train our network, which we call FAL-net, for plain stereoscopic view synthesis by penalizing the generated synthetic right view 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.
- We **shed light on the effectiveness of the mirrored exponential disparity (MED) representations** for self-supervised SIDE.



Our proposed training strategy with our novel Mirrored Occlusion Module (MOM).

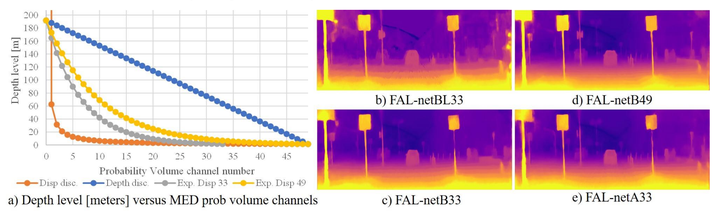
Network architecture

Our FAL-net maps a single left input view I_L to a N-channel “disparity logit” volume D^L_L. Such “disparity logits” can be channel-wise soft-maxed to form a Mirrored Exponential Disparity (MED) “probability volume” D^{PL}_L, which can be quantized and sum-reduced to give rise to the final predicted depth map.

Our disparity logits can also be progressively shifted to the right camera and soft-maxed, generating the right-from-left MED prob. volume D^{PR}_L. The element-wise multiplication of D^{PR}_L 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

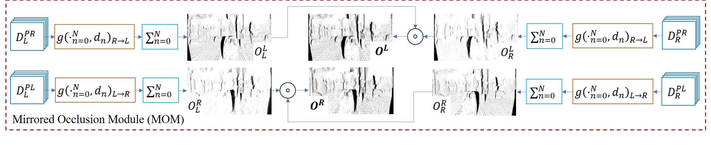
We observed significant improvements by adopting exponential disparity quantization, which is reasonable, as linear quantization of disparity assigns most sampling positions to the very close-by objects due to the inverse relation between disparity and depth.



Comparison between linear quantization and exponential quantization of disparity. Our FAL-netBL33 in (b) with linear quantization struggles to estimate depths beyond a few meters from the camera.

Mirrored Occlusion Module (MOM)

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 (more details in our paper).



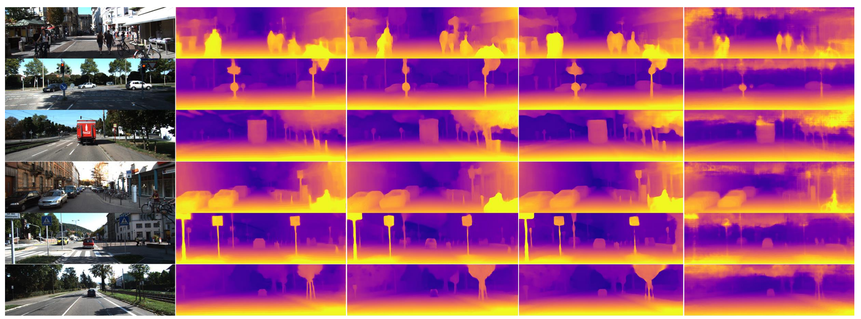
Two-stage Training strategy

In the **first stage**, we train our FAL-net for view synthesis with l_1, perceptual, and smoothness losses and keep a fixed copy of the model.

In the **second stage**, enabled by our Mirrored Occlusion Module, we fine-tune our FAL-net for inverse depth (disparity) estimation with an occlusion-free reconstruction loss, a smoothness loss, and a new “mirror loss”. The mirror loss provides self-supervision for the left-occluded regions which are ignored in the reconstruction loss function.

Results

The depth estimates of our FAL-net are more detailed and consistent than the previous methods. Our method outperforms most recent SOTA methods by a considerable margin on the challenging KITTI dataset with 8x fewer parameters and 3x faster inference speeds.



Performance comparison of existing SIDE methods. K: Eigen train-split. CS: CityScapes. PP: post-processing. Arrows indicate the better metric. **Best** and second-best metrics. Results capped to 80m

Improved Eigen Test Split [28]											
[3]	DORN		D	K	51	0.072	0.307	2.727	0.120	0.932	0.984	0.995
[6]	Monodepth2		V	K	14	0.092	0.536	3.749	0.135	0.916	0.984	0.995
[11]	PackNet (LR)		V	K	120	0.078	0.420	3.485	0.121	0.931	0.986	0.996
[11]	PackNet		V	CS→K	120	0.071	0.359	3.153	0.109	**0.944**	0.990	0.997
[11]	PackNet		V+v	CS→K	120	0.075	0.384	3.293	0.114	0.938	0.984	0.995
[6]	Monodepth2		V+S	K	14	0.087	0.479	3.595	0.131	0.916	0.984	0.996
[6]	Monodepth2		S	K	14	0.084	0.503	3.646	0.133	0.920	0.982	0.994
[31]	DepthHints	✔	S_{SGM}	K	35	0.074	0.364	3.202	0.114	0.936	0.989	0.997
our	FAL-netA33	✔	S	K	**6.6**	0.076	0.335	3.122	0.116	0.934	0.989	0.997
our	FAL-netB33	✔	S	K	17	0.071	0.282	**2.859**	**0.106**	**0.944**	**0.991**	**0.998**
our	FAL-netB49	✔	S	K	17	0.071	0.281	2.912	0.108	0.943	**0.991**	**0.998**
our	FAL-netB49	✔	S	K+CS	17	**0.068**	**0.276**	2.906	**0.106**	**0.944**	**0.991**	**0.998**

Conclusions

- We showed that state-of-the-art SIDE can be achieved by light and straightforward auto-encoder networks that incorporate MED representations in their output layers.
- Our two-stage training strategy with our Mirror Occlusion Module (MOM) aids in learning precise depth instead of plain view-synthesis.
- Our method outperforms the DORN supervised baseline by a large margin in most metrics, which suggests we can “forget about the LiDAR” for the supervision of SIDE networks.

References

[3] 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)

[6] Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3828–3838 (2019)

[11] 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)

[28] 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. IEEE (2017)

[31] 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)