KI DELTA LEARNING
Scalable AI for Automated Driving

Attention-Based Self-Supervised Monocular Depth Estimation

Sagar Hanagodimath, Marius Bachhofer

Semi- and Unsupervised Learning

Use-Case in Delta Learning

Relative to traditional supervised approaches, self-supervised depth estimation models allow for extraction of 3D information from raw camera images without relying on additional expensive sensors and labels. This leads to considerable savings in both cost and time and ability to use data from diverse environments and geographies with minimal effort.

Technical Problem

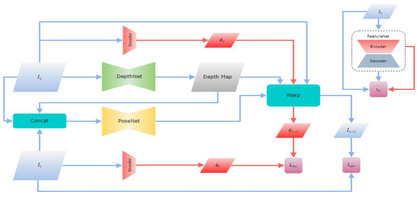

Figure 1: High-level overview of the method

This work attempts to alleviate the problem of disparity between the overall performance of supervised and self-supervised monocular depth estimation models. Though the former performs better, it is relatively restricted due to the need of ground truth data. Additionally, a general strategy to improve such a model's ability during training is also focussed on.

Technical Solution

Most existing monocular depth estimation approaches use the standard encoder-decoder model structure to estimate depth from camera images which may ignore relevant features especially at boundaries and in texture-less regions. Given a consecutive sequence of images I_{t-1}, I_t and I_{t+1}, let I_t be the target image and I_{t-1}, I_{t+1}, the source images. PoseNet provides the camera transformation parameters between the source and target images which in turn is used to inverse warp the source images to synthesise the target image using the output of DepthNet. The photometric difference between the original and synthesised images is used as supervisory signal for the models (Figure 1). Our approach also attempts to study the effects of usage of various attention modules on the model performance. This is achieved by utilising attention blocks in the decoder prior to the upscaling operation. The effects of incorporating depth maps as input to PoseNet and additional encoded feature maps from FeatureNet as input to view synthesis technique was also quantified [1]. This especially helps improve the model's ability to better deal with boundary and texture-less regions.

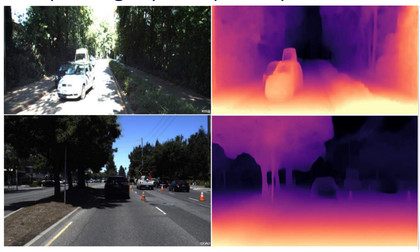

Figure 2: Depth maps for scenes with relatively prominent texture-less (in shadows) regions in both KITTI and DDAD datasets

Evaluation

Experiments with multiple attention techniques were carried out. CBAM and ECA were the front runners implying that channel-based attention techniques perform better for the task of depth estimation [2, 3]. Usage of depth map as additional input to the pose estimation network results in significant improvement in the estimated depth map (Figure 2, Table 1). Limitations include inability to cope with real-world conditions like rain, reflecting surfaces, motion blur etc.

Model	Abs Rel	Sq Rel	RMSE	$\delta_{1.25}$
	Lower is better			Higher is better
KITTI				
DIFFNet	0.0989	0.703	4.317	0.902
DIFFNet + D	**0.0978**	0.694	4.307	0.904
DIFFNet# + D	0.0989	0.728	4.378	0.903
DIFFNet + F	0.0988	0.696	4.295	0.902
DIFFNet + D+ F	**0.0978**	**0.689**	**4.282**	**0.905**
DDAD				
DIFFNet	0.137	3.646	13.717	0.851
DIFFNet + D	0.130	**2.569**	12.896	0.847
DIFFNet# + D	**0.125**	2.652	13.137	**0.857**
DIFFNet + F	0.134	**2.569**	12.896	0.847
DIFFNet + D+ F	0.127	3.174	**12.231**	0.840
Monodepth2	0.198	4.504	16.641	0.781
DepthFormer	0.135	2.953	12.477	0.836

Table 1: Comparison of performance with different models for KITTI and DDAD datasets. DIFFNet – DIFFNet with CBAM, DIFFNet# – DIFFNet with ECA, D - Depth map input to PoseNet, F - FeatureNet [4]. All values in meters.

References

[1] Shu, Chang, et al. "Feature-metric loss for self-supervised learning of depth and egomotion." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*. Cham: Springer International Publishing, 2020.
[2] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[3] Wang, Qilong, et al. "ECA-Net: Efficient channel attention for deep convolutional neural networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[4] Zhou, Hang, et al. "Self-supervised monocular depth estimation with internal feature fusion." *arXiv preprint arXiv:2110.09482* (2021).

Partners

External partners

For more information contact:
hanagodimath.sagar@zf.com
marius.bachhofer@zf.com

KI Delta Learning is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

www.ki-deltalearning.de @KI_Familie KI Familie

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag