KI DELTA LEARNING
Scalable AI for Automated Driving

Self-supervised Learning for Object Detection

Sebastian Wirkert, Jan Huenermann

WP 3.1

Networks

Use-Case in Delta Learning

Especially in the automotive environment, a large amount of data can be produced in a relatively simple way without labels. This data can be exploited using unsupervised or semi-supervised methods. Unsupervised learning can be used as pre-training to extract descriptive features from existing data.
This poster focuses on self-supervised learning for monocular depth prediction. This setting is the basis for exploring self-supervision as a generalization method on time domain.

Technical Problem

Recent approaches to monocular self-supervised depth estimation are capable of estimating relative camera position and pixel-wise depth from image sequences and camera intrinsic. These methods must rely on changes in camera position due to vehicle motion and thus time. This violates the underlying assumption of a static scene and prevents the correct prediction and handling of motion in environments with dynamic objects.
In this approach we relax the static scene assumption and add motion estimation to the self-supervised stack.

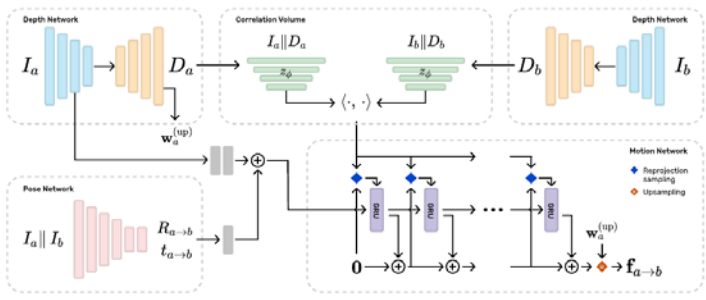

Figure 1: Our approach consists of a depth and pose network as well as a correlation feature extractor and a motion network.

	Abs	Sq	RMSE
EPC++	0.141	1.029	5.350
GeoNet	0.155	1.296	5.857
Li et al.	0.130	0.950	5.138
Ours	**0.122**	**0.816**	**4.885**

Table 1: Depth estimation performance of current methods for self-monitored estimation of depth, pose, and scene flow (lower is better).

Technical Solution

We decompose the scene flow into (1) first-person motion and (2) object motion. To account for object motion, we modify the point transformation using the relative pose to the reference frame of b by including the motion field predicted by the model. Our method follows conventional approaches by using a recurrent all-pairs field transform (RAFT)[1] for iterative motion prediction. For 3D motion prediction, we change the output dimension of the recurrent neural network to predict a 3D residual motion field.
Reprojection sampling. At each iteration, we compute dense correspondences for each pixel by reprojecting points in image a onto image b using the depth map, relative pose, and motion estimation. Feature sharing. Instead of a context network, we run the features of our depth coder through a two-layer convolutional network and add a linear projection of the predicted pose.3D-aware correlation volume. To enable spatial inference in the recurrent unit, we link the predicted depth maps of images a and b to the input of the feature prediction NN for each image.

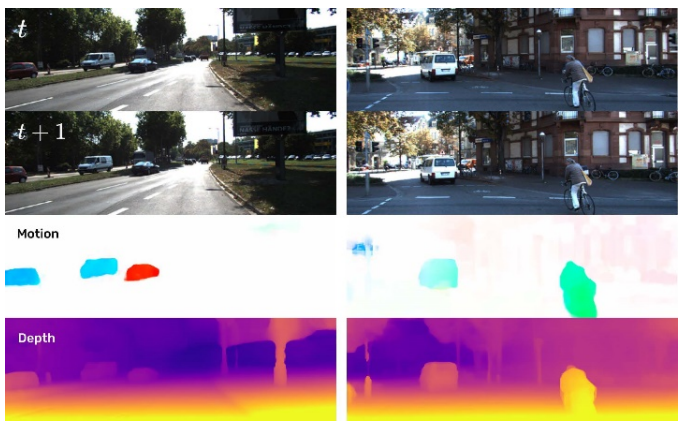

Figure 2: Depth and motion maps generated by our method

Evaluation

We train our method on three datasets: KITTI [3]. All our tests are performed on the KITTI and KITTI Scene Flow [2] datasets.
For depth evaluation, we use the eigen test split and report the proposed metrics. We perform per-image median scaling with the bottom half of the depth map, as is common to solve the depth ambiguity problem in unsupervised depth estimation.

[1] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, In ECCV, 2020
[2] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015
[3] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. In *IJRR*, 2013

BMW GROUP

For more information contact:
sebastian.wa.wirkert@bmw.de,
artem.savkin@bmw.de

Partners

External partners

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 Energy.

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

www.ki-deltalearning.de @KI_Familie KI Familie