KI DELTA LEARNING
Scalable AI for Automated Driving

Self-Supervised Scale-Aware Distance Estimation using Monocular Fisheye Camera for Autonomous Driving

Varun Ravi-Kumar

WP 2.1
WP 3.1

Networks

Use-Case in Delta Learning

Depth estimation models may be learned in a supervised fashion on LiDAR distance measurements. However, setting up the entire rig for such recordings is expensive and time-consuming, limiting the amount of data on which a model can be trained. To overcome this problem, we explore self-supervised learning for depth estimation. Self-supervised learning approaches show promising results and could overcome the problem.

Technical Problem

Obtaining accurate and dense depth supervision is difficult in practice, but We aim to develop a novel self-supervised scale-aware framework for learning Euclidean distance and ego-motion from raw monocular fisheye videos without applying rectification. While it is possible to perform a piece-wise linear approximation of fisheye projection surface and apply standard rectilinear models, it has its own set of issues like re-sampling distortion and discontinuities in transition regions. A limitation of this approach is that both depth and pose are estimated up to an unknown scale factor in the monocular Structure-from-Motion pipeline.

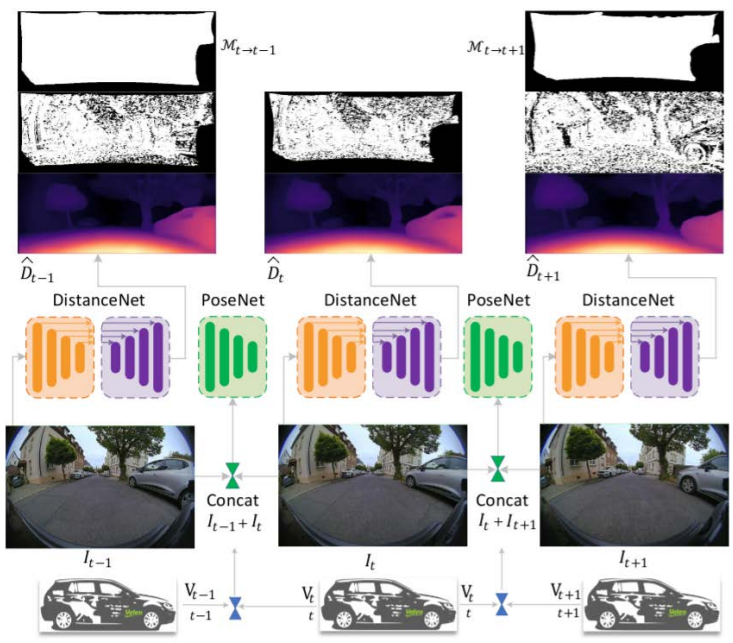

Fig. 1 **Overview of our method.** The first row represents our ego masks, $\mathcal{M}_{t\to t-1}$, $\mathcal{M}_{t\to t+1}$ indicate which pixel coordinates are valid when constructing $\hat{I}_{t-1\to t}$ from I_{t-1} and $\hat{I}_{t+1\to t}$ from I_{t+1} respectively. The second row indicates the masking of static pixels computed after 2 epochs, where black pixels are filtered from the photometric loss (i.e.$\omega = 0$). It prevents dynamic objects at similar speed as the ego car and low texture regions from contaminating the loss. The masks are computed for forward and backward sequences from the input sequence S and reconstructed images. The third row represents the distance estimates corresponding to their input frames. Finally, the vehicle's odometry data is used to resolve the scale factor issue.

Technical Solution

Using view-synthesis as the supervising technique we can train the network using the viewpoint of source images to estimate the appearance of a target image on raw fisheye images. To perform distance estimation on raw fisheye images, we would require metric distance values to warp the source image onto the target frame. To achieve scale-aware distance values, we normalize the pose network's estimate and scale the displacement magnitude relative to target frame which is calculated using vehicle's instantaneous velocity estimates at time.

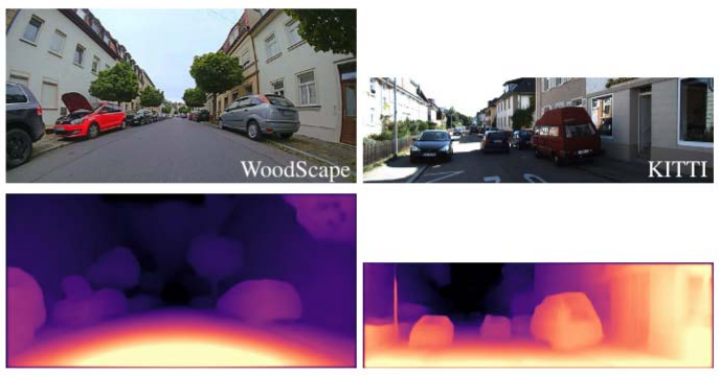

Fig. 1 **Distance and depth derived from a single fisheye image (left) and single pinhole image (right) respectively.** Our self-supervised model, ***FisheyeDistanceNet***, produces sharp, high quality distance and depth maps.

Evaluation

We evaluate the model's depth and distance estimation results using the metrics proposed by Eigen to facilitate comparison. The quantitative results are shown in the Table illustrate that our scale-aware self-supervised approach outperforms all the state-of-the-art monocular approaches. Our fisheye cameras can perform well up to a range of 40m. Therefore, we also report results on a 30m, and a 40m.

Approach	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
	lower is better				higher is better		
KITTI							
Zhou [15]†	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [38]	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Vid2depth [35]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [36]†	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [17]	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [37]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Ranjan [39]	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [33]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth '(M)' [34]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Zhou [27]	0.139	1.057	5.213	0.214	0.831	0.940	0.975
PackNet-SfM [40]	0.120	0.892	4.898	0.196	0.864	0.954	0.980
Monodepth2 [14]	**0.115**	0.903	4.863	0.193	**0.877**	0.959	0.981
FisheyeDistanceNet	0.117	**0.867**	**4.739**	**0.190**	0.869	**0.960**	**0.982**
FisheyeDistanceNet (1024 × 320)	0.109	0.788	4.669	0.185	0.889	0.964	0.982
WoodScape							
FisheyeDistanceNet cap 80 m	0.167	1.108	3.814	0.216	0.794	0.953	0.972
FisheyeDistanceNet cap 40 m	0.152	0.768	2.723	0.210	0.812	0.954	0.974
FisheyeDistanceNet cap 30 m	0.149	0.613	2.402	0.204	0.810	0.957	0.976

TABLE I **Quantitative results of leaderboard algorithms on KITTI dataset [23] and FisheyeDistanceNet on Fisheye dataset part of WoodScape [1].** Single-view depth estimation results using the Eigen Split [41] for depths reported less than 80 m, as indicated in [41] for pinhole model. All the approaches are self-supervised on monocular video sequences. At test-time, all monocular methods excluding our FisheyeDistanceNet, scale the estimated depths using median ground-truth LiDAR depth. For the fisheye dataset, we estimate distance rather than depth. † marks newer results reported on GitHub.

For more information contact:
varun.ravi-kumar@valeo.com

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