SynDistNet: Self-Supervised Monocular Fisheye Camera Distance Estimation Synergized with Semantic Segmentation for Autonomous Driving Supplementary Material

Varun Ravi Kumar[1,4] Marvin Klingner[3] Senthil Yogamani[2]
Stefan Milz[4] Tim Fingscheidt[3] Patrick Mäder[4]
[1]Valeo DAR Kronach, Germany [2]Valeo Vision Systems, Ireland
[3]Technische Universität Braunschweig, Germany [4]Technische Universität Ilmenau, Germany

1. Additional Method Details

Edge-Aware Distance Smoothness Loss: In order to regularize distance and avoid divergent values in occluded or texture-less low-image gradient areas, we add a geometric smoothing loss. We adopt the edge-aware term similar to [1]. The regularization term is imposed on the inverse distance map. The loss is weighted for each of the image pyramid levels and is decayed by a factor of 2 on each downsampling.

$$\mathcal{L}_s(\hat{D}_t) = |\partial_u \hat{D}_t^*| e^{-|\partial_u I_t|} + |\partial_v \hat{D}_t^*| e^{-|\partial_v I_t|} \quad (1)$$

To discourage shrinking of distance estimates [8], mean-normalized inverse distance of D_t is considered, i.e. $\hat{D}_t^* = \hat{D}_t^{-1}/\overline{D}_t$, where $\overline{D}_t$ denotes the mean of $\hat{D}_t^{-1} := 1/\hat{D}_t$.

Cross-Sequence Distance Consistency Loss: Following FisheyeDistanceNet [5], we enforce the cross-sequence distance consistency loss (CSDCL) for the training sequence S:

$$\mathcal{L}_{dc} = \sum_{t=1}^{N-1} \sum_{t'=t+1}^{N} \left(\sum_{p_t} \left| D_{t\to t'}(p_t) - \hat{D}_{t\to t'}(p_t) \right| + \sum_{p_{t'}} \left| D_{t'\to t}(p_{t'}) - \hat{D}_{t'\to t}(p_{t'}) \right| \right) \quad (2)$$

Eq. 2 contains one term for which pixels and point clouds are warped forwards in time (from t to t') and one term for which they are warped backwards in time (from t' to t), where $\hat{D}_{t'}$ and $\hat{D}_t$ are the estimates of the images $I_{t'}$ and I_t respectively for each pixel $p_t \in I_t$.

Additional Considerations: In all the previous works [10, 1, 2], networks are trained to recover inverse depth $g_d : p \mapsto g_D^{-1}(I_t(p))$. A limitation of these approaches is that both depth or distance and pose are estimated up to an unknown scale factor. We incorporate the scale recovery technique from FisheyeDistanceNet [5] and obtain scale-aware depth and distance directly for pinhole and fisheye images. We also incorporate the clipping of the photometric loss values, which improves the optimization process and provides a way to strengthen the photometric error. Additionally, we include the backward sequence training regime, which helps to resolve the unknown distance estimates in the image border.

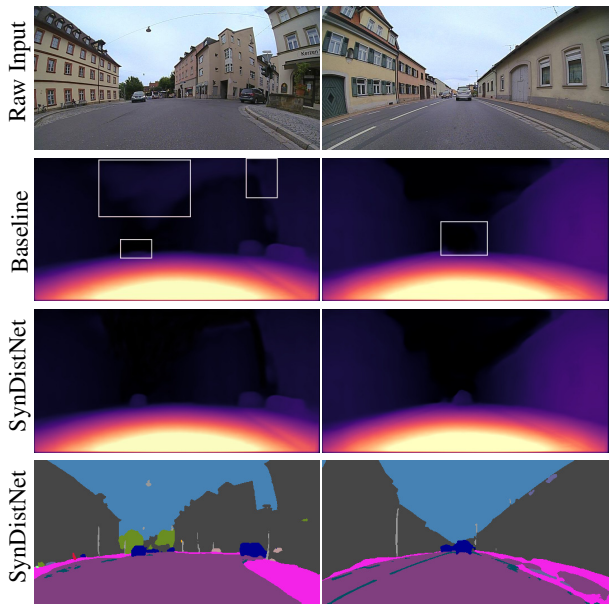

Figure 1: **Qualitative result comparison on the Fisheye WoodScape dataset** between the baseline model without our contributions and the proposed SynDistNet. Our SynDistNet can recover the distance of dynamic objects (left images) which eventually solves the infinite distance issue. In the 3rd and 4th columns, we can see that semantic guidance helps us to obtain curbs and resolve the distance of homogeneous areas outputting sharp distance maps on raw fisheye images. The final row indicates the semantic segmentation predictions.

2. Implementation Details

The distance estimation network is mainly based on FisheyeDistanceNet [5], an *encoder-decoder* network with skip connections. After testing different variants of ResNet family, we chose ResNet18 [3] as the encoder as it provides a high-quality distance prediction, and improvements in higher complexity encoders were incremental. It would also aid in obtaining real-time performance on low-power embedded systems. We also incorporate self-attention layers in the encoder and drop the deformable convolutions used in the baseline model. We could leverage the usage of a more robust loss function over L_1 to reduce training times on ResNet18 by performing a single-scale image depth prediction than the multi-scale in [5]. The semantic segmentation is trained in a supervised fashion with Cross-Entropy loss and is jointly optimised along with the distance estimation. We use Pytorch [7] and employ Ranger (RAdam [6] + LookAhead [9]) optimizer to minimize the training objective function than the previously employed Adam [4]. RAdam leverages a dynamic rectifier to adjust Adam's adaptive momentum based on the variance and effectively provides an automated warm-up custom-tailored to the current dataset to ensure a solid start to training. LookAhead "lessens the need for extensive hyperparameter tuning" while achieving "faster convergence across different deep learning tasks with minimal computational overhead." Hence, both provide breakthroughs in different aspects of deep learning optimization, and the combination is highly synergistic, possibly providing the best of both improvements for the results.

We train the model for 17 epochs, with a batch size of 20 on 24GB Titan RTX with an initial learning rate of 10^{-4} for the first 12 epochs, then drop to 10^{-5} for the last 5 epochs. A significant decrease in training time of 8 epochs over the previous training of the model for 25 epochs in FisheyeDistanceNet [5]. The sigmoid output σ from the distance decoder is converted to distance with $D = a \cdot \sigma + b$. For the pinhole model, depth $D = 1/(a \cdot \sigma + b)$, where a and b are chosen to constrain D between 0.1 and 100 units. The original input resolution of the fisheye image is 1280×800 pixels; we crop it to 1024×512 to remove the vehicle's bumper, shadow, and other artifacts of the vehicle. Finally, the cropped image is downscaled to 512×256 before feeding to the network. For the pinhole model on KITTI, we use 640×192 pixels as the network input.

3. Qualitative Results

Figure 1 and Figure 2 provides qualitative results of SynDistNet on WoodScape and KITTI test dataset for segmentation and depth estimation tasks respectively. Figure 3 illustrates the qualitative comparison of depth estimation with the recent state of the art methods.

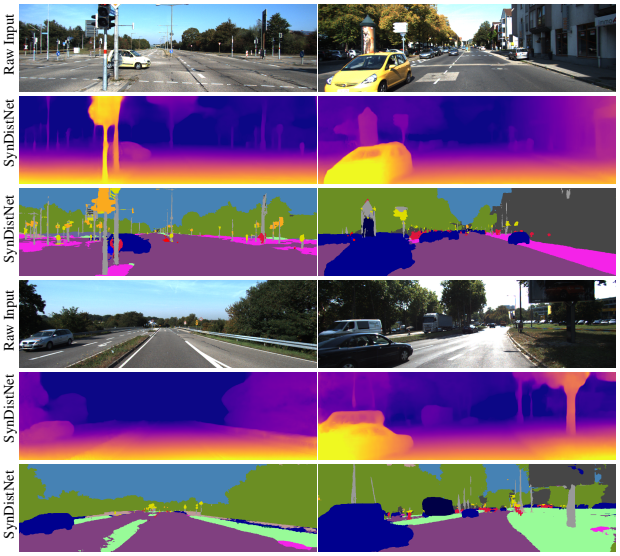

Figure 2: **Qualitative results on the KITTI dataset.** We showcase depth estimation as well as semantic segmentation outputs on the KITTI dataset using our SynDistNet.

References

[1] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *Proc. of ICCV*, 2019. 1, 3

[2] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. 3D Packing for Self-Supervised Monocular Depth Estimation. In *Proc. of CVPR*, 2020. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, 2016. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014. 2

[5] Varun Ravi Kumar, Sandesh Athni Hiremath, Markus Bach, Stefan Milz, Christian Witt, Clément Pinard, Senthil Yogamani, and Patrick Mäder. Fisheyedistancenet: Self-supervised scale-aware distance estimation using monocular fisheye camera for autonomous driving. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 1, 2

[6] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *arXiv preprint arXiv:1908.03265*, 2019. 2

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 2

[8] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning Depth From Monocular Videos Using Direct Methods. In *Proc. of CVPR*, 2018. 1

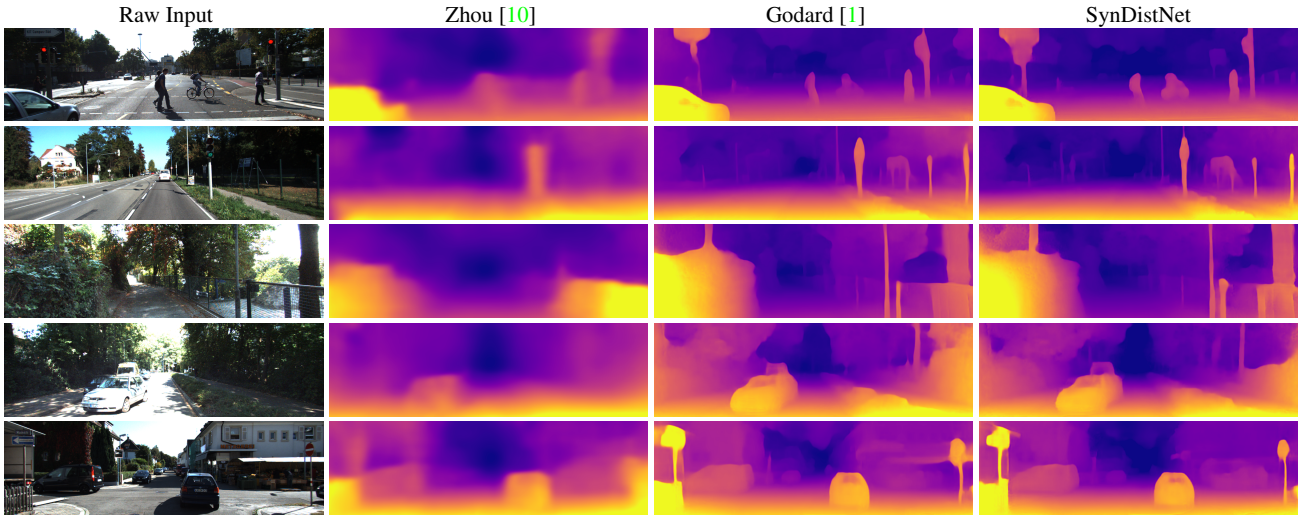

Figure 3: **Qualitative results on the KITTI dataset.** Our SynDistNet produces sharp depth maps on raw pinhole camera images and can recover the distance of dynamic objects.

[9] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, 2019. 2

[10] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 3