Near-field Depth Estimation using Monocular Fisheye Camera: A Semi-supervised learning approach using Sparse LIDAR Data

Varun Ravi Kumar[1], Stefan Milz[1], Christian Witt[1], Martin Simon[1], Karl Amende[1], Johannes Petzold[1], Senthil Yogamani[2] and Timo Pech[3‡]

Abstract

Near-field depth estimation around a self-driving car is an important function that can be achieved by four wide-angle fisheye cameras having a field of view of over 180°. Progress in CNN based depth estimation is hindered as annotation cannot be obtained manually. Synthetic datasets are used commonly but they have limitations due to domain shift. In this work, we explore an alternative approach of training using sparse LIDAR data as ground truth for fisheye camera depth estimation. LIDAR data projected onto the image plane is sparse and hence viewed as semi-supervision for dense depth estimation. To handle the difference in view-points of LIDAR and fisheye camera, an occlusion resolution mechanism is implemented. We built our own dataset using our self-driving car setup which has a 64-beam Velodyne LIDAR and four wide angle fisheye cameras. We started with Eigen's multiscale convolutional network architecture [1] and improved by modifying activation function and optimizer. We obtained promising results on our dataset with RMSE errors better than the state-of-the-art results obtained on KITTI because of vast amounts of training data.

1. Introduction

Depth estimation is a fundamental task for self driving cars to find obstacles. LIDAR and Stereo cameras are mainly used for this purpose but they mainly cover far-field. Wide-angle fisheye cameras are used to cover the near-field region around the car which is essential for urban driving use cases. These cameras have a field of view of over 180° and have a spatially variant distortion as shown in Figure 4 (a). In this paper, we explore depth estimation for these monocular fisheye cameras using sparse LIDAR as semi-supervised ground truth. The other alternative of using synthetic datasets has several limitations. Firstly, there is a large domain shift while performing inference on natural images and this is more pronounced for fisheye cameras. Secondly, synthetic datasets do not capture the extensive variability in the appearance of objects like vehicles present in real datasets.

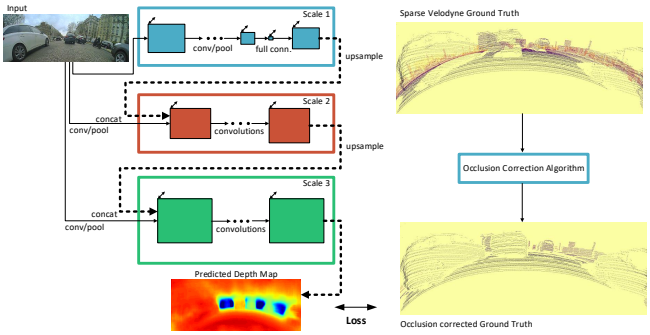

Figure 1. Multi-scale architecture for depth prediction on raw fisheye images with a sparse Velodyne LIDAR ground truth with Occlusion correction (see Section 2.1).

Our motivation is to establish a baseline for single frame depth estimation solely based on sparse Velodyne data as ground truth for training. The main contribution of this paper are building a joint fisheye camera and LIDAR dataset, adapting training data to handle occlusion due to difference in camera and LIDAR viewpoint, tailoring loss function and training algorithm to handle semi-supervision and demonstration of fisheye camera depth estimation using CNN with RMSE scores better than supervised state-of-the-art methods on KITTI.

2. Model Architecture

Our model offers several architectural improvements to [1] which is initially based on Eigen et al. [2]. We adopted a simple architecture for Scale 1 based on AlexNet [4] to achieve real time on an embedded platform. Depending on the whole image area, a multi-scale deep neural network first predicts a coarse global output and refines it using finer-scale local networks. This scheme is

*[1]Valeo und Schalter und Sensoren GmbH, Driving Assistance Advanced Research, Kronach varun-ravi.kumar@valeo.com

†[2]Senthil Yogamani is with Valeo Vision Systems, Ireland senthil.yogamani@valeo.com

‡[3]Timo Pech is with Technische University, Chemnitz, Germany timo.pech@etit.tu-chemnitz.de

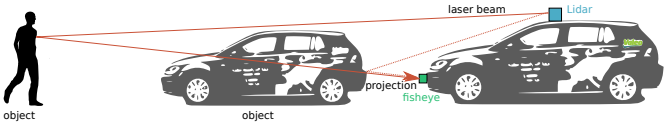

Figure 2. Illustration of occlusion due to LIDAR's viewpoint being higher than that of fisheye camera. 3D points from the object (person) will be mapped even though it is not visible from camera.

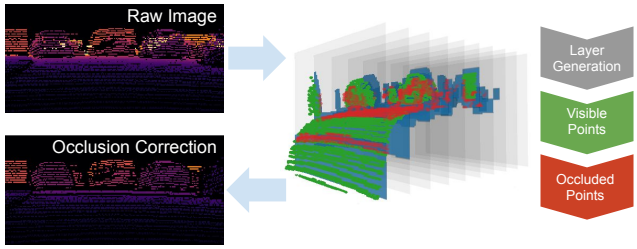

Figure 3. Visualization of the distance based segmentation technique with morphological filters for occlusion correction.

described in Fig. 1. The model is deeper with more convolutional layers compared to [2]. Second, with the added third scale from [1] at higher resolution, bringing the final output resolution up to half the input, or $284\,\mathrm{px} \times 80\,\mathrm{px}$ for our sparse LIDAR fisheye camera dataset. In addition, we use swish [7] as the activation function rather than the mostly preferred ReLU [6]. Finally, we adopt Adam optimizer [3] which yields better training error/loss instead of the Stochastic gradient descent (SGD) used by Eigen et al. [1, 2]. Multi channel feature maps were passed similarly to [1] avoiding the flow of output predictions from the coarse scale to the refine scale.

2.1. Occlusion Correction

In our vehicle, the fisheye cameras are positioned in the front for near-field sensing and LIDAR is placed at the top as seen in Fig. 2. LIDAR perceives the environment behind objects that occlude the view for the camera. This problem of occlusion results in wrong mapping of depth-points that are not visible to the camera. To solve this problem, we adapted a distance based segmentation technique with morphological filters as shown in the Fig. 3. Instead of directly projecting points from the LIDAR into the image plane of the fisheye camera, we introduce I layers within the camera view located at a distance d_i^{img}, $i = 1, \ldots, I$. Each LIDAR point will be projected onto the layer next to it. We apply a morphological filter that dilates points within each layer to fill the sparse regions (in Fig. 3 dilated parts of the layers are colored blue). A point at a distance d is regarded as occluded, if a layer i exists with $d_i^{\text{img}} < d$. Otherwise the valid point is projected onto the camera plane.

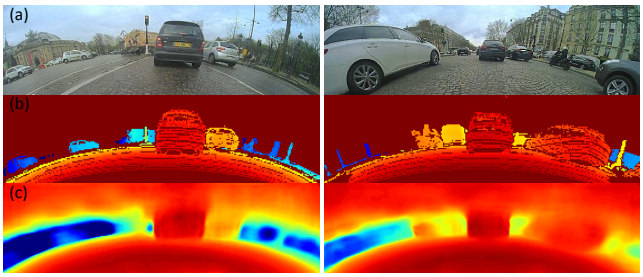

Figure 4. Qualitative results: (a) RGB Input (b) LIDAR Ground Truth (c) Predicted CNN Depth Map

3. Results

The model is completely trained on our internal dataset obtained from four fisheye cameras and sparse Velodyne HDL-64E rotating 3D laser scanner as ground truth. The training set contains $50\,000$ images out of $55\,000$ images collected by driving around Paris, France and Bavaria, Germany. The remaining 5000 images are used for testing. The training set includes scenes from the *city*, *residential* and *sub-urban* categories of our raw dataset. Points without depth value are left unfilled without any post-processing. Eigen's model [1] handles missing values by eliminating them in the loss function. A protocol evaluation is applied and results are shown by discarding ground-truth depth below $0\,\mathrm{m}$ and above $50\,\mathrm{m}$ while capping the predicted depths into $0\,\mathrm{m} - 50\,\mathrm{m}$ depth interval. This implies, we set predicted depths to $0\,\mathrm{m}$ and $50\,\mathrm{m}$ if they are below $0\,\mathrm{m}$ or above $50\,\mathrm{m}$, respectively. In Table 1, we show better accuracy than the state-of-the-art algorithm on KITTI [5].

Table 1. Quantitative Comparison on Kuznietsov et al. [5] depth estimation on KITTI dataset and Valeo's Fisheye dataset.

	Kuznietsov [5]	**Ours**
RMSE (linear)	3.531	**1.717**
RMSE (log)	0.183	**0.236**
RMSE (log, scale-invariant)	-	**0.226**
Absolute Relative Difference	0.117	**0.160**
Squared Relative Difference	0.597	**0.397**
$\delta < 1.25$	0.861	**0.816**
$\delta < 1.25^2$	0.964	**0.934**
$\delta < 1.25^3$	0.989	**0.969**

4. Conclusion

We have demonstrated that semi-supervision using sparse LIDAR can provide reliable dense depth estimations for fisheye cameras. To our knowledge, this is the first attempt at CNN based depth estimation on fisheye images using sparse Velodyne ground truth. Even though the camera/LIDAR setups are different, the results provide a reasonable comparison to KITTI on performance of monocu-

lar depth regression using sparse LIDAR input. In future work, we aim to improve the results by using more consecutive frames which can exploit the motion parallax and better CNN encoders.

References

[1] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *CoRR*, abs/1411.4734, 2014.

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. . Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. *CoRR*, abs/1702.02706, 2017.

[6] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress.

[7] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017.