ScaleNet: Scale Invariant Network for Semantic Segmentation in Urban Driving Scenes

Mohammad Dawud Ansari[1,2], Stephan Krauß[1], Oliver Wasenmüller[1] and Didier Stricker[1,2]
[1]*German Research Center for Artificial Intelligence (DFKI), Germany*
[2]*University of Kaiserslautern, Germany*

Keywords: Semantic Segmentation, Autonomous Driving, Labeling, Automotive, Scale.

Abstract: The scale difference in driving scenarios is one of the essential challenges in semantic scene segmentation. Close objects cover significantly more pixels than far objects. In this paper, we address this challenge with a scale invariant architecture. Within this architecture, we explicitly estimate the depth and adapt the pooling field size accordingly. Our model is compact and can be extended easily to other research domains. Finally, the accuracy of our approach is comparable to the state-of-the-art and superior for scale problems. We evaluate on the widely used automotive dataset Cityscapes as well as a self-recorded dataset.

1 INTRODUCTION

One basic technology for Autonomous Driving and Advanced Driver Assistance Systems (ADAS) is a semantic segmentation of the scene in front of the car. This segmentation is used to understand the surrounding of the car and adapt its actions accordingly. Such a segmentation should be able to differentiate between the road and the sidewalk as well as detect pedestrians, cars, vegetation, traffic signs and many more.

The Computer Vision community applies for this task so-called per-pixel semantic labeling algorithms to images captured by a front-facing camera of the car. These algorithms are usually based on Convolutional Neural Network (CNN), and each pixel in the image is labeled with an identifier (ID) indicating the semantic class. The networks are trained with a huge dataset like Cityscapes (Cordts et al., 2016) and achieve reasonable accuracy.

One essential challenge in the context of driving scenarios is the scale of objects. When looking at a street similar objects can appear very close to the camera as well as very far. Since the depth of the objects is inversely proportional to their scale, these objects can cover a very different amount of pixels. The scale difference can be multiple dozens. Thus, these scale differences need to be explicitly considered in the network architecture.

In this paper, we propose a new architecture for semantic scene segmentation called Scale Invariant Network (ScaleNet). Within this network, we explicitly estimate the depth out of the input image and adapt the pooling field size accordingly. With this approach, fine details in high distance can be preserved as well as large objects in short range.

The main contributions of the paper are two-fold: First, a quantized depth network is utilized to estimate sufficient depth information which is further used by the segmentation network to adjust the pooling field size. Second, we describe a scale invariant semantic segmentation neural network model, which can cope with difficult scale changes for similar objects. This model is found to perform robustly in challenging automotive scenarios, like the Cityscapes dataset (Cordts et al., 2016). One can use the same network for much simpler datasets, such as PASCAL VOC (Everingham et al., 2015) and MS COCO (Lin et al., 2014). Finally, we provide a qualitative and quantitative evaluation of our approach on the Cityscapes dataset as well as a self-captured dataset.

2 RELATED WORKS

Semantic segmentation as dense per pixel labeling has been studied from different perspectives for over a decade. The algorithms in the state of the art can be categorized into three groups. The first group is based on using image pyramid methods. It is done by generating a multi-scale input image, feeding it to the CNN and then fusing all the feature maps from every

Ansari, M., Krauß, S., Wasenmüller, O. and Stricker, D.
ScaleNet: Scale Invariant Network for Semantic Segmentation in Urban Driving Scenes.
DOI: 10.5220/0006723003990404
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 399-404
ISBN: 978-989-758-290-5
Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

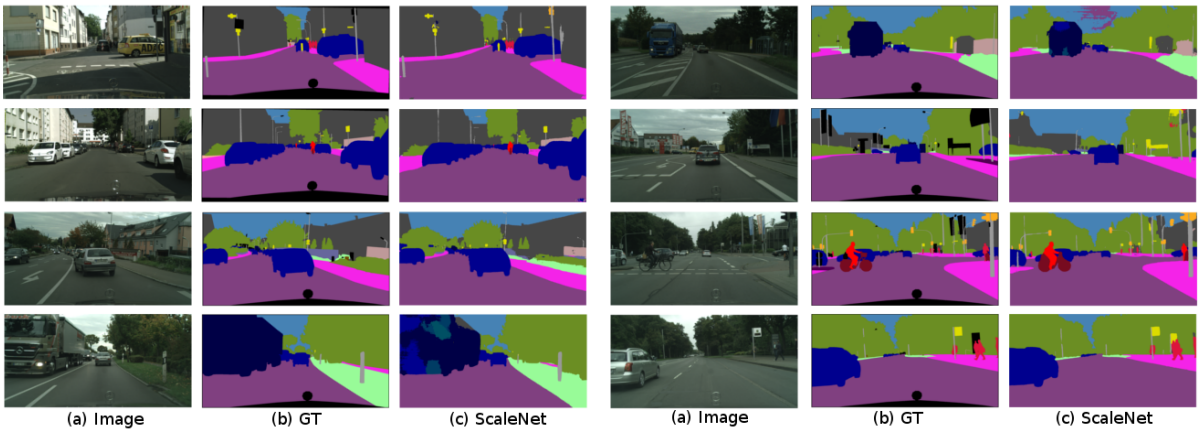

Figure 1: Qualitative evaluation of ScaleNet (ours) compared to the ground truth of the benchmark dataset Cityscapes (validation set).

scale. Small-scale input will encode contextual information while the larger scale will encode small object details. There are several implementation details: (Farabet et al., 2013) converts input images to multi-scale images through a Lapalacian pyramid and feeds them to a CNN. In contrast, (Lin et al., 2016b), (Chen et al., 2016b) and (Chen et al., 2016a) pre-process input images by directly resizing them to different scales. However, the limitation of this type of methods is the amount of memory required for deeper/larger models, especially to the high-resolution datasets (i.e. Cityscapes), hence resizing can be done during inference phase only (Briggs et al., 2000).

The second group of approaches use encoder-decoder networks. In the encoder the features map are reduced by sets of convolution and max/average pooling computation, while the spatial dimension and object details of the input images are recovered in the decoder. Several methods are proposed, for example (Long et al., 2015a; Noh et al., 2015) use deconvolution layers to up-sample low resolution features map. Another approach is to use skip connections (Ronneberger et al., 2015). The effectiveness of this method has been demonstrated by RefineNet (Lin et al., 2016a), which achieves comparable results on several benchmarks of semantic segmentation. Another variant is atrous convolution to up-sample feature maps by using the decoder. DeepLab-v2(Res101) (Chen et al., 2016a) proposes atrous convolution, which explicitly controls the resolution of the feature maps in the CNN layers. It inserts a 'hole' to the convolution kernel to enhance the receptive field in the feature maps. Atrous convolution or dilation convolution has been explored for semantic segmentation. For example, (Wu et al., 2016) are experimenting with different dilation rates to get better results.

The third group of approaches is motivated by spatial pyramid pooling (Grauman and Darrell, 2005; Lazebnik et al., 2006) to capture context from different ranges. DeepLab-v2(Res101) (Chen et al., 2016a) employs atrous convolution layers with different rates in parallel as in spatial pyramid pooling to capture information from multi-scale feature maps. Pyramid Scene Parsing Net (PSPNet) (Zhao et al., 2016) produces outstanding results in several benchmarks by employing spatial pooling at several grid scales.

3 ADAPTIVE POOLING

Since the scale of distant objects is smaller compared to an identical object close by, an ambiguity in the appearance is created. We aim to resolve this shortcoming of segmentation algorithms by utilizing the depth of the scene as an additional modality to cope with the changes in appearance following the work of (Kong and Fowlkes, 2017). This can be stated as depth is inversely proportional to scale. We estimate depth out of the color input image to account for the scale changes within the scene since depth cameras are known to be error proven in automotive context (Yoshida et al., 2017). Each depth estimation corresponding to a pixel serves as an approximation for the scale of the pooling field that covers this pixel. Unlike the conventional depth estimation networks where the accuracy of per-pixel depth affects the overall performance of the algorithm, we quantize the depth range in a small set, e.g., four. These depth sets are then used to adapt the pooling field size in the network during training. One can also use raw depth value. As shown by (Kong and Fowlkes, 2017) using estimated depth boosts the performance of the model. One reason for this increased performance can be that the parameters are tuned in such a

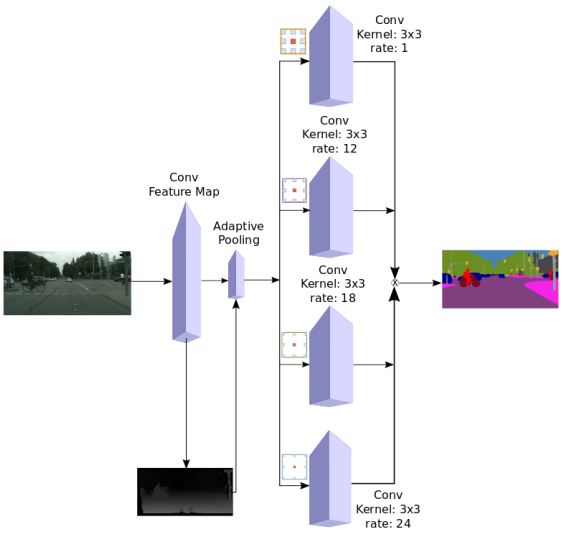

Figure 2: Architecture of the desired model is shown. This model generates depth as the intermediate result, which is used for adaptive pooling.

way that they predict depth which overall elevates the performance of the segmentation network. Following DeepLab-v2(Res101) (Chen et al., 2016a) atrous convolution is used to match the pooling field size. The depth prediction network is a classification network with a fixed number of classes. We used a simple regression model which minimizes the Euclidean distance between the actual and the predicted depth for all the valid pixels available in the ground truth data, which is given as

$$l_{depthReg}(\mathbf{D},\mathbf{D}^*) = \frac{1}{V}\sum_{(i,j)\in V}\left\|\mathbf{D}_{ij}-\mathbf{D}^*_{ij}\right\|_2^2. \quad (1)$$

4 DILATED CONVOLUTION

Semantic segmentation is obtained using CNN in a fully connected fashion and max/average pooling (Long et al., 2015a; Li et al., 2016). A stack of these operations reduces the resolution of the feature map up to 32 times compared to the original size of the image input (Chen et al., 2016a). This issue leads to the diminishing response in a feature map as we go deeper in the network. One remedy for this issue is proposed by (Chen et al., 2016a). We follow a similar method, which adjusts the pooling field size by matching the dilation rate of the atrous convolution (Kong and Fowlkes, 2017). This algorithm allows us to compute the responses of any layer at any desired resolution. This technique has proven to be very intuitive and natural in the sense that we do not need to tune any parameters manually.

Table 1: Quantitative evaluation of several state-of-the-art approaches and ScaleNet (ours) for benchmark dataset Cityscapes (validation set). We use average IoU as the evaluation metric. The metric is multiplied by 100.

Method	IoU
FCN-8s (Long et al., 2015b)	65.3
RefineNet (Chen et al., 2016a)	73.6
RecurrentParsing (Kong and Fowlkes, 2017)	78.2
DeepLab-v2(Res101) (Chen et al., 2016a)	70.4
ScaleNet (ours)	75.1

Initially designed for wavelet transform in the "algorithme átrous" scheme of (Holschneider et al., 1990), the method can also be used to find the response of any feature map. Considering two-dimensional signals, for each location i on the output y and a filter w, atrous convolution is applied over the input feature maps x as

$$y[i] = \sum_k x[i + r\cdot k]w[k], \quad (2)$$

where r is the rate parameter that corresponds to the stride with which the input is sampled (Chen et al., 2016a).

5 MODEL ARCHITECTURE

The model architecture is shown in Figure 2 consists of two sub-networks: An adaptive pooling, which is controlled by a depth regression network, and a network which fuses the responses of the atrous convolution with different self-adjusting dilation rates. We use a pre-trained ResNet101 model for feature generation with some modifications for maintaining the resolution of the output score maps according to (Chen et al., 2016a; Kong and Fowlkes, 2017). The top global 7×7 pooling layer and the last two 2×2 pooling layers are removed, and an atrous convolution layer with dilation rate of two is inserted. This will generate the $\frac{1}{8}$ size of the output score map compared to the input image size. Finally, a bilinear interpolation is applied on the output score map to obtain the identical resolution as input. As a result, the feed-forward network remains the same as DeepLab-v2(Res101), but with two additional 3×3 kernel (without atrous convolution) added above the generated feature maps from the ResNet101 feature generation model. After this, one 3×3 atrous convolution layer is added whose rate is adjusted using adaptive pooling module.

Implementation. For the implementation we reused parts of the open-source software [1] of (Kong and Fowlkes, 2017), who uses the MatConvNet (Vedaldi and

[1] https://github.com/aimerykong/Recurrent-Scene-Parsing-with-Perspective-Understanding-in-the-loop

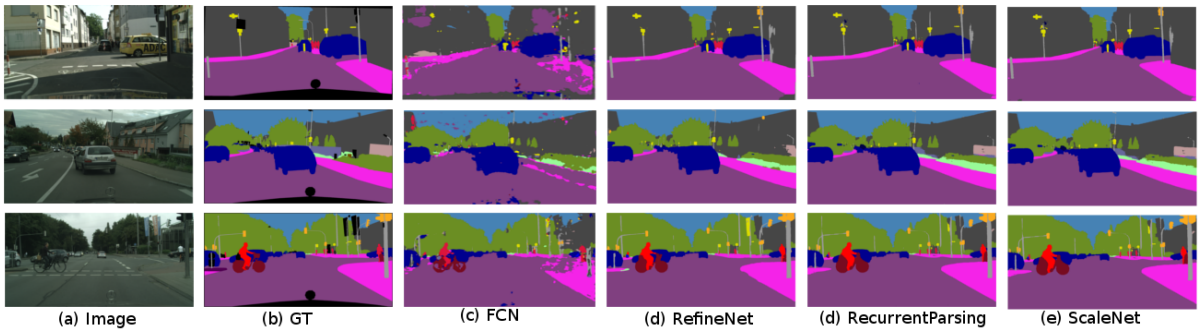

Figure 3: Qualitative evaluation of FCN-8s (Long et al., 2015b), RefineNet (Lin et al., 2016a), RecurrentParsing (Kong and Fowlkes, 2017) and ScaleNet (our) for benchmark dataset Cityscapes (validation set).

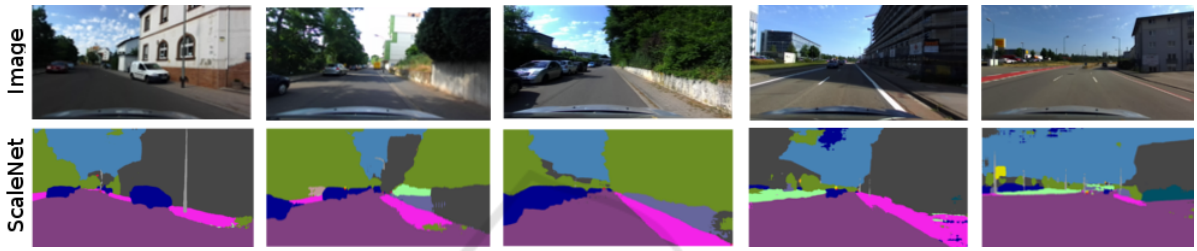

Figure 4: Qualitative evaluation of ScaleNet (ours) for self captured ZED dataset.

Lenc, 2015) framework for designing the model. For training, the batch size is set to one, and the learning rate has a base value of $2.0 \cdot 10^{-5}$ which is scaled as $(1 - \frac{iter}{maxiter})^{0.7}$. We trained the model by augmenting extra datasets. Different data augmentation is performed for model training. A small image per object is generated out of Cityscapes (Cordts et al., 2016) images with an additional padding of 50px, and later we create rotating inputs with additional zero-padded values to keep the input image a rectangular shape. The rotation is performed in the range $[-15°, 15°]$. For additional training coarse annotations including the train-extra data is also available in Cityscapes. We start the training with the fine annotated images, followed by the coarse annotated images of Cityscapes dataset. Later, augmented datasets are used in a sequence of small images followed by the rotated images. Finally, the model is fine-tuned for ten more epochs using fine annotation of Cityscapes. We trained our model for 180 epochs.

6 EXPERIMENTS

We perform an experimental evaluation on the recognized and widely used automotive semantic segmentation dataset Cityscapes (Cordts et al., 2016). It contains 2975 training, 1525 test and 500 validation set images, which compose the fine annotation part of the dataset. This dataset contains images from several German cities captured from a camera integrated into a moving car. This dataset has been a standard for benchmarking an automotive semantic segmentation algorithm, which is the reason why we also choose this benchmark for our evaluation. We provide the quantitative results of our architecture on the complete test set and provide our comparison with other well-known approaches such as FCN-8s (Long et al., 2015b), RefineNet (Lin et al., 2016a), RecurrentParsing (Kong and Fowlkes, 2017) and DeepLab-v2(Res101) (Chen et al., 2016a).

Additionally, we recorded a dataset ourselves. This dataset is captured using a ZED stereo camera installed on top of a car. We drive on multiple road types with single and double lanes. For qualitative evaluation, we take the images captured from the left camera only.

For measuring the accuracy of our results we utilize the intersection-over-union (IoU) metric, which is well-known for the Cityscapes (Cordts et al., 2016) and PASCAL VOC (Everingham et al., 2015) dataset. The metric is defined as

$$IoU = \frac{TP}{TP + FP + FN}, \tag{3}$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels, respectively, determined over the whole test set (Cordts et al., 2016). In Table 1 we compare our ScaleNet against other approaches listed in Cityscapes. With an IoU of 75.1% we achieve state-of-the-art performance. In

Table 2: Quantitative evaluation of FCN-8s (Long et al., 2015a), RefineNet (Lin et al., 2016a), RecurrentParsing (Kong and Fowlkes, 2017), DeepLab-v2(Res101) (Chen et al., 2016a) and ScaleNet (ours) for benchmark dataset Cityscapes (validation set). We use IoU as the evaluation metric. The metric is multiplied by 100.

Object	FCN-8s	RefineNet	RecurrentParsing	DeepLab-v2(Res101)	ScaleNet (ours)
road	97.40	98.20	98.50	97.86	98.32
sidewalk	78.40	83.21	85.44	81.32	84.82
building	89.21	91.28	92.51	90.35	92.25
wall	34.93	47.78	54.41	48.77	50.05
fence	44.23	50.40	60.91	47.36	59.61
pole	47.41	56.11	60.17	49.57	62.75
traffic light	60.08	66.92	72.31	57.86	71.79
traffic sign	65.01	71.30	76.82	67.28	76.68
vegetation	91.41	92.28	93.10	91.85	93.16
terrain	69.29	70.32	71.58	69.43	71.35
sky	93.86	94.75	94.83	94.19	94.62
person	77.13	80.87	85.23	79.83	83.63
rider	51.41	63.28	68.96	59.84	65.15
car	92.62	94.51	95.70	93.71	95.05
truck	35.27	64.56	70.11	56.50	56.01
bus	48.57	76.07	86.54	67.49	71.64
train	46.54	64.27	75.49	57.45	59.88
motorcycle	51.56	62.20	68.30	57.66	66.28
bicycle	66.76	69.95	75.47	68.84	73.62

Table 2 we show the IoU for each class separately. One can see that we achieve superior accuracy for objects with fine geometry compared to other state-of-the-art algorithms. This can be seen especially in classes like *fence*, *pole*, *traffic light*, *vegetation*, *bicycle* and many more. While we achieve superior accuracy for these classes, the remaining classes still have comparable accuracies. Thus, we verified the scale invariance of our network and showed the proper performance of the adaptive pooling.

In Figure 1 we compare the labeling results of ScaleNet with the ground truth of Cityscapes validation set. It can be seen that it highly matches the ground truth labels in different driving scenarios. Except for a few misclassifications ScaleNet provides a decent accuracy comparable to state-of-the-art segmentation algorithms. Homogeneous regions are correctly classified as well as objects with fine geometry. In Figure 3 we qualitatively compare ScaleNet with other state-of-the-art methods. First of all, ScaleNet provides comparable accuracy since most labels are classified correctly. When looking at objects with fine structure – like poles or traffic lights – one can see that we achieve higher accuracy there. This confirms again the effectiveness of the scale invariant architecture.

In Figure 4 we apply ScaleNet on the dataset we recorded using the ZED camera. One can see that the network performs well also on this dataset, although it was trained on Cityscapes. Fine structures as well as homogeneous regions are classified correctly. With that we can verify that the network generalized well to unknown images and scenes.

One remaining challenge in our approach is sharp boundaries in the segmentation results – like for many algorithms in the state-of-the-art. We tested several approaches like edge-aware filtering (Wasenmüller et al., 2015) with limited accuracy improvement.

7 CONCLUSION

In this paper, we proposed ScaleNet – a network architecture for semantic scene segmentation – to address scale differences. These scale differences occur especially in driving scenarios, since objects can appear at very different distances. We showed that this effect could be handled by an adaptive pooling depending on the depth of the respective pixel. In our evaluation, we verified the state-of-the-art performance of ScaleNet and showed the effect of the featured scale handling.

Future work could be a combination of scene flow (Schuster et al., 2018) or ego-motion (Wasenmüller et al., 2016) estimation together with semantic segmentation to support each estimation jointly while computation.

ACKNOWLEDGEMENTS

This work was partially funded by the European project *Eyes of Things* under contract number

GA643924. Furthermore, we want to thank Alwi Husada for the fruitful discussions.

REFERENCES

Briggs, W. L., Henson, V. E., and McCormick, S. F. (2000). *A multigrid tutorial*. SIAM.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., and Yuille, A. L. (2016b). Attention to scale: Scale-aware semantic image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3640–3649.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR)*.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136.

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1915–1929.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1458–1465. IEEE.

Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. (1990). A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer.

Kong, S. and Fowlkes, C. (2017). Recurrent scene parsing with perspective understanding in the loop. *arXiv preprint arXiv:1705.07238*.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE.

Li, Y., Qi, H., Dai, J., Ji, X., and Wei, Y. (2016). Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*.

Lin, G., Milan, A., Shen, C., and Reid, I. (2016a). Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. *arXiv preprint arXiv:1611.06612*.

Lin, G., Shen, C., van den Hengel, A., and Reid, I. (2016b). Efficient piecewise training of deep structured models for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3194–3203.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, pages 740–755. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015a). Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Long, J., Shelhamer, E., and Darrell, T. (2015b). Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *International Conference on Computer Vision (ICCV)*, pages 1520–1528.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Schuster, R., Wasenmüller, O., Kuschk, G., Bailer, C., and Stricker, D. (2018). Sceneflowfields: Dense interpolation of sparse scene flow correspondences. In *IEEE Winter Conference on Computer Vision (WACV)*.

Vedaldi, A. and Lenc, K. (2015). Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*.

Wasenmüller, O., Ansari, M. D., and Stricker, D. (2016). Dna-slam: Dense noise aware slam for tof rgb-d cameras. In *Asian Conference on Computer Vision Workshop (ACCV workshop)*. Springer.

Wasenmüller, O., Bleser, G., and Stricker, D. (2015). Combined bilateral filter for enhanced real-time upsampling of depth images. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 5–12.

Wu, Z., Shen, C., and Hengel, A. v. d. (2016). Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*.

Yoshida, T., Wasenmüller, O., and Stricker, D. (2017). Time-of-flight sensor depth enhancement for automotive exhaust gas. In *IEEE International Conference on Image Processing (ICIP)*.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2016). Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*.