A Residual Encoder-Decoder Network for Semantic Segmentation in Autonomous Driving Scenarios

Naresh.Y.G[1], Suzanne Little[2] and Noel E. O'Connor[3]

Abstract— **In this paper, we propose an encoder-decoder based deep convolutional network for semantic segmentation in autonomous driving scenarios. The architecture of the proposed model is based on VGG16 [1]. Residual learning is introduced to preserve the context while decreasing the size of feature maps between the stacks of convolutional layers. Also, the resolution is preserved through shortcuts from the encoder stage to the decoder stage. Experiments are conducted on popular benchmark datasets CamVid and CityScapes to demonstrate the efficacy of the proposed model. The experiments are corroborated with comparative analysis with popular encoder-decoder networks such as SegNet and Enet architectures demonstrating that the proposed approach outperforms existing methods despite having fewer trainable parameters.**

I. INTRODUCTION

In recent years, semantic segmentation has been investigated in a variety of different application contexts. In this paper, we consider semantic segmentation in the context of vision-based driver assistance systems. Vision-based driver assistance and other advanced control systems in the car are crucial to enable autonomous driving. Semantic segmentation has a vital role to play in enabling a number of important tasks viz., object localization, tracking, detection and classification, ultimately leading to scene understanding and autonomous decision making in traffic scenarios. Beyond real-time in-vehicle operation for decision making, semantic segmentation has another important application in this context. As is well known, a large amount of high quality annotated training data is vital in order to ensure the output quality of computer vision systems. Important tasks for which annotation is required [2] include pedestrian detection, lane keeping, traffic sign recognition and so on. Semantic segmentation can be applied to legacy footage from existing vehicles to acquire this kind of training data to train ever more sophisticated techniques.

State-of-the-art segmentation models have demonstrated impressive pixel-level classification results. There are several models which have shown encouraging results in pixel-wise classification in road scene understanding [3], [4], [5], [6]. VGG [1] and Alexnet [7] are two popular deep learning models given their performance in the ImageNet Large Scale Visual Recognition Challenge (ILSCRC) [8]. These models have been used as baselines in various categorization and pixel-wise classification models. Fully Convolutional Networks (FCNs) [9] exploit the fully connected layers for prediction. Dilated convolutions [10] were proposed to support exponential expansion of the receptive field without loss of resolution for better pixel-wise classification. ResNet [11] is another popular model based on residual learning. Variants of ResNet are used by various models [5], [12], [4] as a baseline for a variety of applications. Deeplab [13] is a model which is based on convolutions, dilated convolutions, pyramid pooling and fully connected CRF's. The pyramid scene parsing [6] network was proposed based on a pyramid pooling module which exploits global context information for pixel-wise classification. Apart from these models, there are several models which are based on encoder-decoder networks which have drawn attention due to the lower number of trainable parameters compared to the conventional fully convolutional networks. We follow this line of investigation in this paper and propose a novel residual encoder-decoder network which we believe to be particularly suited to vision-based driver assistance.

II. RELATED WORK

Semantic segmentation enables partitioning a scene and recognition of various entities of the scene to understand the context between entities. Machine learning techniques such as random forest and boosting were traditionally used for the task of image segmentation. With the advent of Deep Learning in recent years, many approaches have been proposed for pixel-wise semantic segmentation. Many challenging datasets are available for various purposes. CamVid [14] and CityScapes [15] are popular datasets which are meant for traffic scene understanding. Generally, most of the semantic segmentation models based on an encoder-decoder network utilize popular models like VGG [1] and Residual networks [11].

In the literature, various models can be found based on an encoder-decoder network which are both symmetric and asymmetric in nature. SegNet [16] is one of the most popular pixel-wise classification models based on an encoder-decoder network. SegNet utilizes 13 layers of VGG [1] which is symmetric in nature i.e., the model has an equal number of pooling and unpooling layers in the encoder and decoder

*This work is supported by CloudLSVA project co-funded by the European Unions Horizon 2020 research and innovation programme under grant No. 688099.

[1] Naresh.Y.G is with Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland naresh.yarlapati@insight-centre.org

[2] Suzanne Little is with Insight Centre for Data Analytics and School of Computing, Dublin City University, Dublin, Ireland suzanne.little@dcu.ie

[3] Noel E.O'Connor is with Insight Centre for Data Analytics and School of Electronic Engineering, Dublin City University, Dublin, Ireland noel.oconnor@dcu.ie

ISBN 978-90-827970-1-5 © EURASIP 2018

networks respectively. Max-pooling indices are saved while downsampling in the encoder network and utilized in upsampling layers in the decoder network. The U-Net [17] model was proposed for biomedical image segmentation which transfers the feature maps from the encoder network by concatenating them to the upsampled feature maps of the decoder network. LinkNet [12] is also a encoder-decoder network for pixel-wise segmentation. It uses ResNet18 in the encoder and full-convolution layers in the decoder. Furthermore, spatial information is bypassed from encoder blocks to corresponding decoder blocks. Bypassing preserves the information lost due to downsampling in the encoder network. An asymmetric encoder-decoder based model known as a residual deconvolutional network [18] was proposed for brain electron microscopy image segmentation. The model is capable of preserving contextual and resolution information.

Despite the fact that several models are available for pixel-wise classification or semantic segmentation, there is still scope for exploring new models which are efficient in terms of trainable parameters whilst producing high quality segmentation results. Thus in this work, we propose a novel residual encoder-decoder based network which preserves the spatial and resolution information whilst reducing the number of parameters. Our model shares similarities with LinkNet [12] and residual deconvolutional network [18]. However, LinkNet is based on ResNet18 [11] and residual deconvolutional network [18] has 23 convolutional layers, 20 deconvolutional layers and 4 pooling and unpooling layers. Our method uses the first 13 layers in VGG 16 to exploit its strength to capture shape information of the objects. However, instead of using unpooling layers in decoder network, we exploit deconvolutional layers to reconstruct the input from the encoder network. This reduces the overhead of storing the max-pooling indices as in SegNet [16].

III. Method

The proposed model has an encoder network and a corresponding mirrored decoder network, followed by a final 1×1 convolution layer with softmax activation to represent categorical probability distribution. The architecture of the proposed model is shown in Fig. 1.

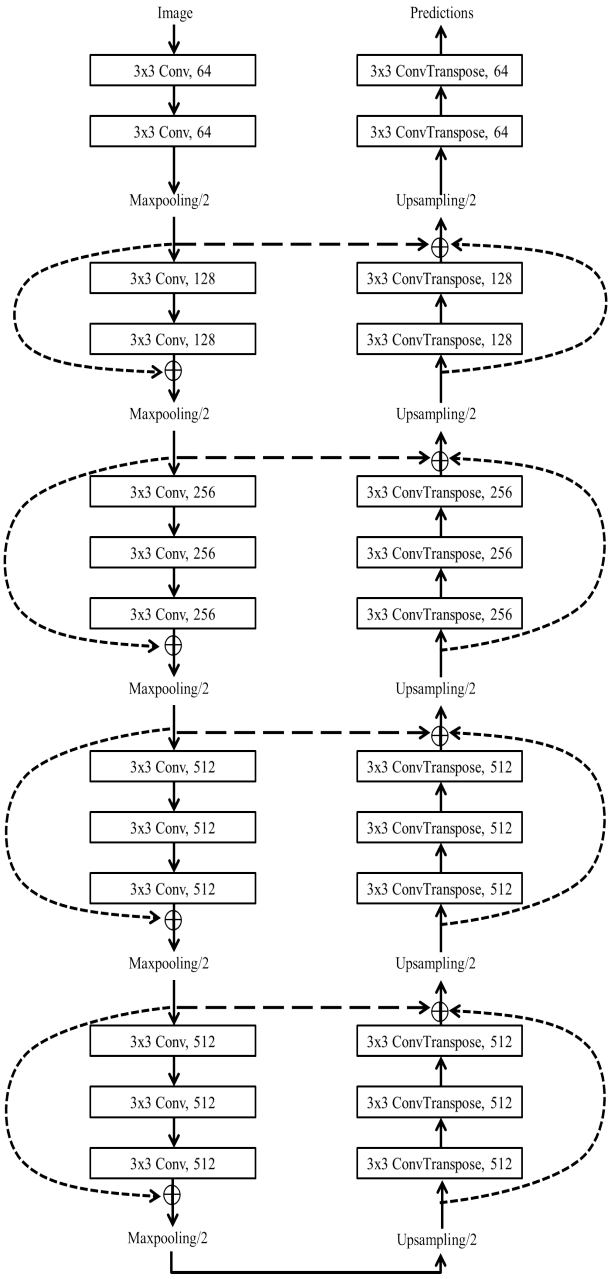

Fig. 1. Block diagram of the proposed model.

The encoder network consists of 13 convolutional layers similar to that of the VGG16 network [1]. Filters with a receptive field of 3×3 are used in the convolutional layers. The convolution stride is fixed to 1 pixel and the spatial resolution is preserved with spatial padding of 1 pixel after each convolution layer. A batch normalization layer is applied to the feature maps yielded by each convolution layer. Element-wise rectified non-linearity is applied as in [1], [16] after each convolutional layer. Max-pooling layers are used to decrease the size of feature maps and also to achieve translation invariance. Further, we use residual learning [11] by a shortcut connection[19], [20], [21] and element-wise addition to preserve spatial and context information. Importantly, this introduces neither extra parameters nor computation complexity. The shortcut connections are performed as shown in the Fig. 1. The shortcut connections shown in Fig. 1 are projection shortcuts which perform linear projection [11], [18] to match the dimensions of input vectors and the residual mapping [11] to be learned.

The Decoder network possesses the same structure as the encoder, but the convolutional layers are replaced by deconvolutional [22] layers (Transpose convolution) and also the ordering of the layers is mirrored. The Max-pooling layers are replaced by Upsampling layers.

Some spatial information is lost during multiple downsampling operations in encoder stage. To overcome this loss of

ISBN 978-90-827970-1-5 © EURASIP 2018

information, the resolution-preserving paths [18], [12] are incorporated to transfer the missing resolution information from the encoding stage to the mirrored decoding stage. The input for the stack of convolutional layers in the encoder network is bypassed to the output of its mirrored stack of convolutional layers in the decoder network addressed via the resolution-preserving path with a dropout of 0.3 to prevent overfitting as shown in Fig. 1. The resolution-preserving paths are also projection shortcuts.

IV. Experiments

In order to validate the efficacy of the proposed model, experiments have been conducted on CamVid [14] and CityScapes [15] datasets which are designed for urban scene understanding suitable for autonomous driving scenarios. Our model is implemented using the keras neural network API [23] with tensorflow backend [24]. To train the model, we used the Adam optimizer with its default learning rate and other parameters as provided in [25]. We use the cross-entropy loss [9], [16] function for training the network. The classes present in the dataset are highly imbalanced. Hence, we perform class balancing through the *median frequency balancing* method [26]. The training set is shuffled before every epoch. We trained our model until the training loss converges. Accuracy results are presented using the Intersection-over-Union (IoU) metric [15]. The experiments are conducted on an Nvidia Titan X GPU. We compare the performance of our method with Enet [4] and SegNet [16], since these methods are known to use fewer parameters but with good performance in pixel-wise semantic segmentation. Our model has 32.95M parameters and 32.93M trainable parameters. The model is thus light weight when compared to models based on fully convolutional networks.

A. CamVid Dataset

This dataset has 367 training and 233 testing RGB images at 360×480 resolution. The training images are annotated for 11 classes such as road, building, cars, etc. The classes other than these 11 classes are annotated as an unlabeled class. Further details can be obtained from [14]. In our experiments, 20% of the training images are considered as validation set. We consider Enet and SegNet for our experiments. TABLE. I shows the Classwise IoU and MeanIoU on CamVid [14] dataset of our model. The Enet and SegNet results are taken from [4].

B. Cityscapes Dataset

This dataset has 2,975 training, 500 validation and 1,525 testing RGB images at 1024×2048 resolution. The training images are annotated for 19 classes with some of the additional classes in the category of vehicle when compared to the CamVid dataset. In our experiments, the images are downsampled by a factor of 4. Fig. 3 shows the pixel-wise classification of the proposed model. Table. II shows the Classwise IoU and MeanIou on CityScapes [15] dataset of our model.

TABLE I
PER-CLASSIOU ON CAMVID TEST SET IN (%)

Classes	SegNet	ENet	Our Model
Building	88.8	74.7	86.4
Tree	87.3	77.8	76.6
Sky	92.4	95.1	91.2
Car	82.1	82.4	**82.9**
Sign	20.5	51.0	22.3
Road	97.2	95.1	95.6
Pedestrian	57.1	67.2	58.4
Fence	49.3	51.7	49.1
Pole	27.5	35.4	24.4
Sidewalk	84.4	86.7	**87.1**
Bicyclist	30.7	34.1	32.3
MeanIoU	**65.6**	65.3	64.2

TABLE II
PER-CLASSIOU ON CITYSCAPES TEST SET IN (%)

Classes	SegNet	ENet	Our Model
Road	96.4	96.3	93.12
Sidewalk	73.2	74.2	63.76
Building	84.0	75.0	**84.72**
Wall	28.4	32.2	**39.55**
Fence	29.0	33.2	**36.58**
Pole	35.7	43.4	**47.73**
Traffic Light	39.8	34.1	**49.27**
Traffic Sign	45.1	44.0	**57.11**
Vegetation	87.0	88.6	88.49
Terrain	63.8	61.4	46.84
Sky	91.8	90.6	89.52
Person	62.8	65.5	65.06
Rider	42.8	38.4	38.44
Car	89.3	90.6	88.29
Truck	38.1	36.9	**44.02**
Bus	43.1	50.5	**56.50**
Train	44.1	48.1	33.75
Motorcycle	35.8	38.5	26.56
Bicycle	51.9	55.4	**60.19**
MeanIoU	56.95	58.28	**58.39**

V. DISCUSSION

In the above experiments, it can be observed that our model performs well for some of the classes in cityscapes dataset when compared to SegNet [16] and Enet [4]. The results on CamVid dataset of the proposed are competitive with the above mentioned model. The results on CityScapes dataset outperform the two other approaches in terms of MeanIoU.

Classes like traffic sign, traffic light are specifically challenging classes which are misclassified due to their proximity in appearance. Similarly, bus and truck classes are also misclassified due to their similarities. However, our method performs well on these classes.

VI. CONCLUSIONS

In this work, we exploit the deconvolutional layers in decoder network to reconstruct shape of the input image without

ISBN 978-90-827970-1-5 © EURASIP 2018

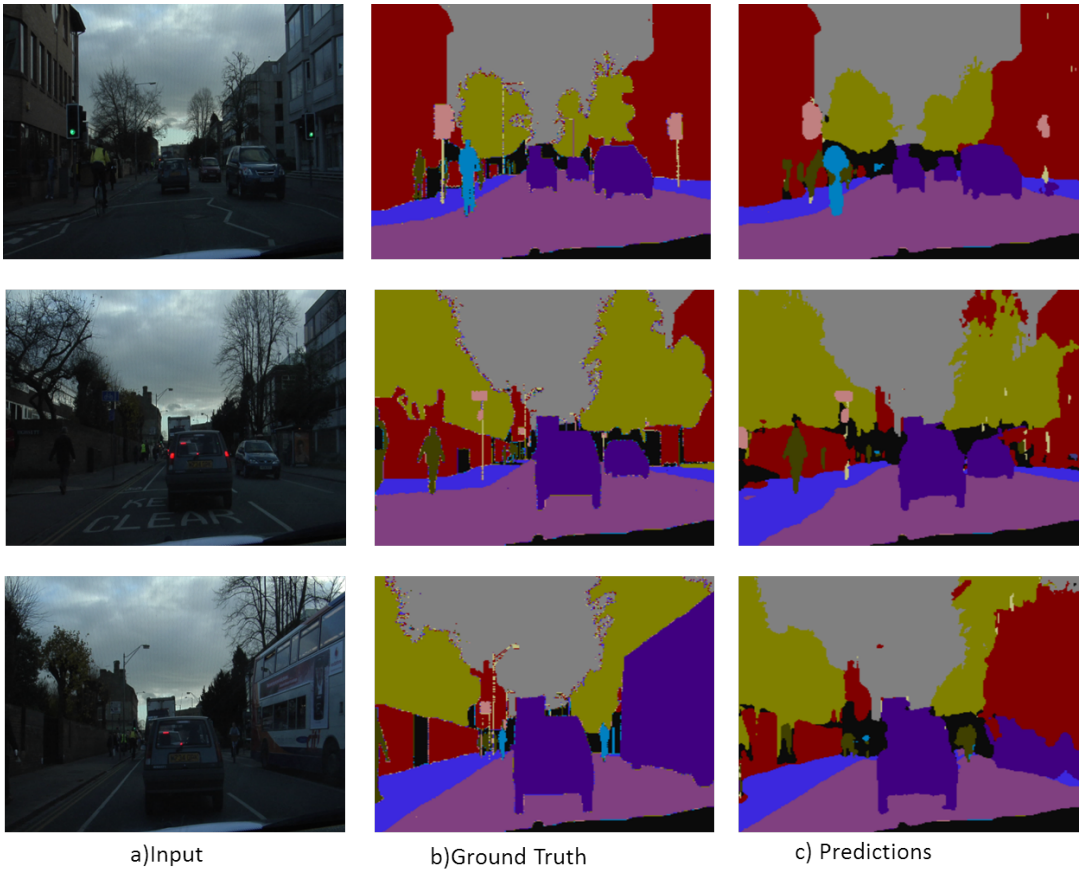

Fig. 2. Results of our model on Test set in CamVid dataset.

unpooling layers [16]. The results of our method is quite competitive for both the datasets as shown in Table. I and Table. II. Further, the effects due to ablation of some of the stacks of layers in the network shall be explored. Fusion of convolutional features from various models shall be explored to achieve efficient and effective pixel-wise classification. To explore the complexity of this encoder-decoder architecture when embedded in a typical autonomous vehicle hardware.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] J. Horgan, C. Hughes, J. McDonald, and S. Yogamani, "Vision-based driver assistance systems: Survey, taxonomy and advances," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pp. 2032–2039, IEEE, 2015.

[3] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, *et al.*, "Speeding up semantic segmentation for autonomous driving," *NIPSW*, vol. 1, no. 7, p. 8, 2016.

[4] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[5] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," *arXiv preprint arXiv:1611.08323*, 2016.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[12] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *arXiv preprint arXiv:1707.03718*, 2017.

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

ISBN 978-90-827970-1-5 © EURASIP 2018

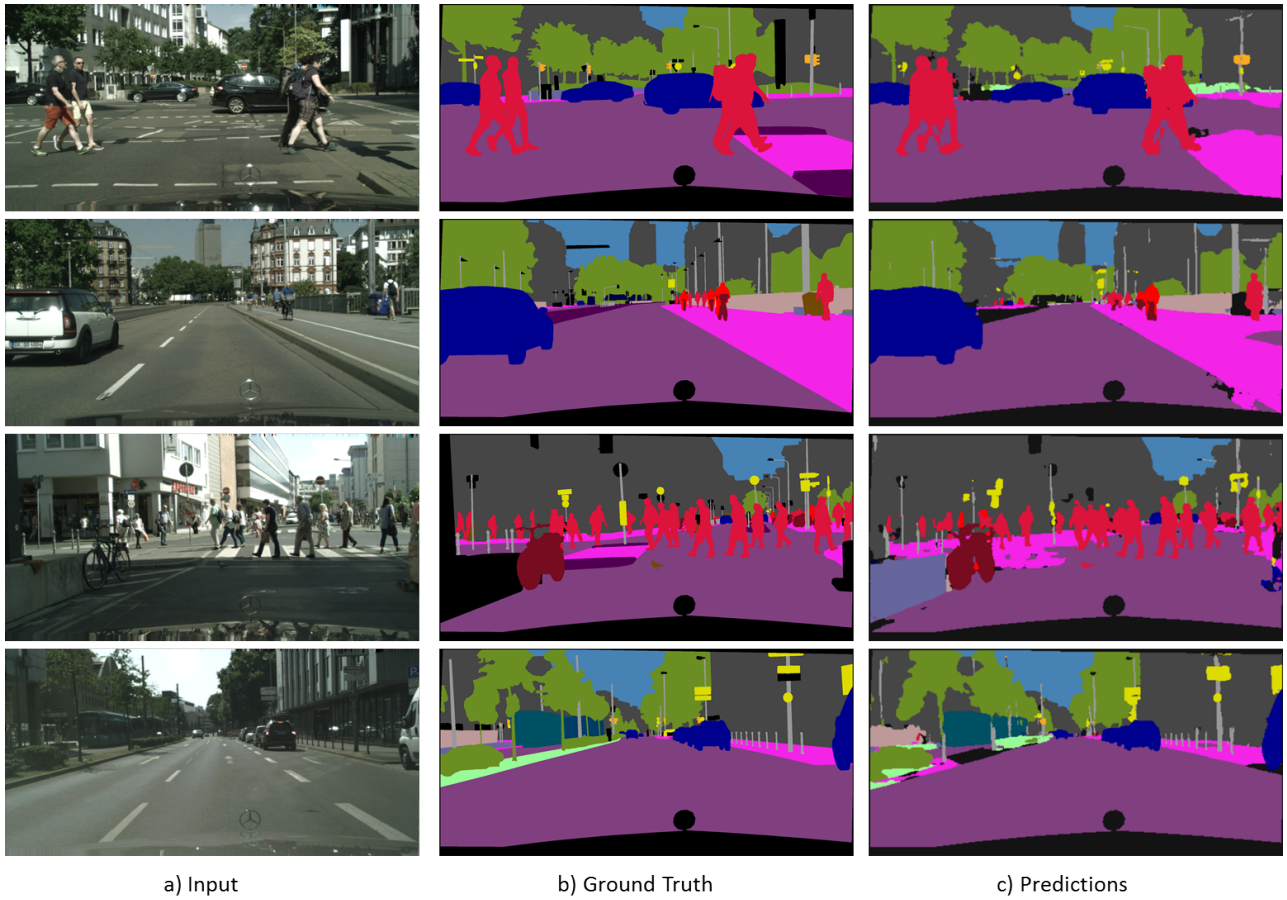

Fig. 3. Results of our model on validation set in Cityscapes dataset.

[14] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, pp. 44–57, 2008.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[18] A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 447–456, 2017.

[19] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[20] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 2007.

[21] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528, 2015.

[23] F. Chollet, "Keras https://github. com/fchollet/keras," 2017.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

ISBN 978-90-827970-1-5 © EURASIP 2018