Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 907–912
DOI: 10.15439/2023F3686
ISSN 2300-5963 ACSIS, Vol. 35

Urban scene semantic segmentation using the U-Net model

Marcin Ciecholewski
Department of Geoinformatics
Faculty of Electronics, Telecommunication and Informatics
Gdańsk University of Technology
Gdańsk, Poland
Email: marcin.ciecholewski@pg.edu.pl

***Abstract*—Vision-based semantic segmentation of complex urban street scenes is a very important function during autonomous driving (AD), which will become an important technology in industrialized countries in the near future. Today, advanced driver assistance systems (ADAS) improve traffic safety thanks to the application of solutions that enable detecting objects, recognising road signs, segmenting the road, etc. The basis for these functionalities is the adoption of various classifiers. This publication presents solutions utilising convolutional neural networks, such as MobileNet and ResNet50, which were used as encoders in the U-Net model to semantically segment images of complex urban scenes taken from the publicly available Cityscapes dataset. Some modifications of the encoder/decoder architecture of the U-Net model were also proposed and the result was named the MU-Net. During tests carried out on 500 images, the MU-Net model produced slightly better segmentation results than the universal MobileNet and ResNet networks, as measured by the Jaccard index, which amounted to 88.85%. The experiments showed that the MobileNet network had the best ratio of accuracy to the number of parameters used and at the same time was the least sensitive to unusual phenomena occurring in images.**

I. Introduction

SEMANTIC segmentation of images is a very important topic in computer vision, and its purpose is to divide the image into regions of different semantic categories. This division is connected with the classification of the image in the sense that it produces per-pixel category prediction instead of image-level prediction [1]. This means that semantic segmentation can be seen as extending image classification from the image level to the pixel level. However, the training data intended for semantic segmentation requires manual labelling at the pixel level, which is much harder and more time-consuming than other vision tasks, such as image classification or object detection.

Much effort has gone into research on image segmentation in recent years and great progress has been made [2], [3], [4], [5], [6]. Despite this, segmentation still remains a difficult problem because of rich intra-class variation, context variation and ambiguities resulting from the low resolution of images.

State-of-the-art approaches used in semantic segmentation adopt a fully convolutional network (FCN) with an encoder/decoder architecture [7], [8]. The encoder generates low-resolution image features and then the decoder upsamples features to segmentation maps and is used for pixel-level classification of the feature representations.

Semantic segmentation has many different applications, notably including: augmented reality, autonomous driving, image editing, medical imaging, robotics, smart cities, and many others [9], [10].

The visual understanding of complex urban street scenes is crucial for problems concerning the smart city, in which autonomous vehicles can drive and certain infrastructure elements can communicate to ensure the greatest comfort of people and reduce the time lost. The use of various large-scale datasets contributed to a great development of research on object detection and a popularisation of methods using deep learning techniques [11], [12]. To use artificial neural networks (ANN) for the semantic segmentation of complex urban scenes, researchers can utilise Cityscapes [13], a benchmark suite and a large-scale dataset to train and test approaches for pixel-level and instance-level semantic labelling. Figure 1 shows example images available in the training subset of the Cityscapes dataset [13]. Images from the training set which can be semantically segmented using specific colours contain 30 different classes describing defined objects found in the city.

This paper presents research on the semantic segmentation of urban scenes using several different convolutional neural networks with an encoder/decoder architecture. For this purpose, MobileNet [14], [15] and ResNet50 [16], [17] were used as encoders in the U-Net model [18]. During the studies, some modifications to the U-Net model were also proposed based on the experiments carried out. The research work was done using the Cityscapes dataset [13]. The purpose of this research was to obtain improved segmentation results, and to assess the proposed solutions in detail, including their advantages and disadvantages.

II. Materials and Methods

A. Data

Research work was carried out using the Cityscapes dataset [13]. This is a collection of 3,475 images from cities in Germany that were recorded during vehicle driving. They are saved in the *png* format and have a resolution of 2048×1024 pixels. This set was divided into a subset designed for training,

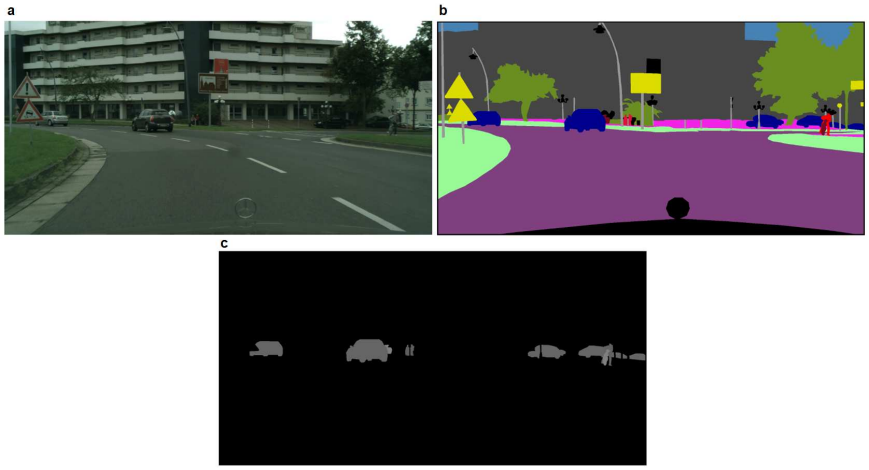

Fig. 1. Sample images from the training subset from the city of Stuttgart. (a) Source image. (b) Semantic segmentation in colour. (c) Segmentation with only vehicles and people marked.

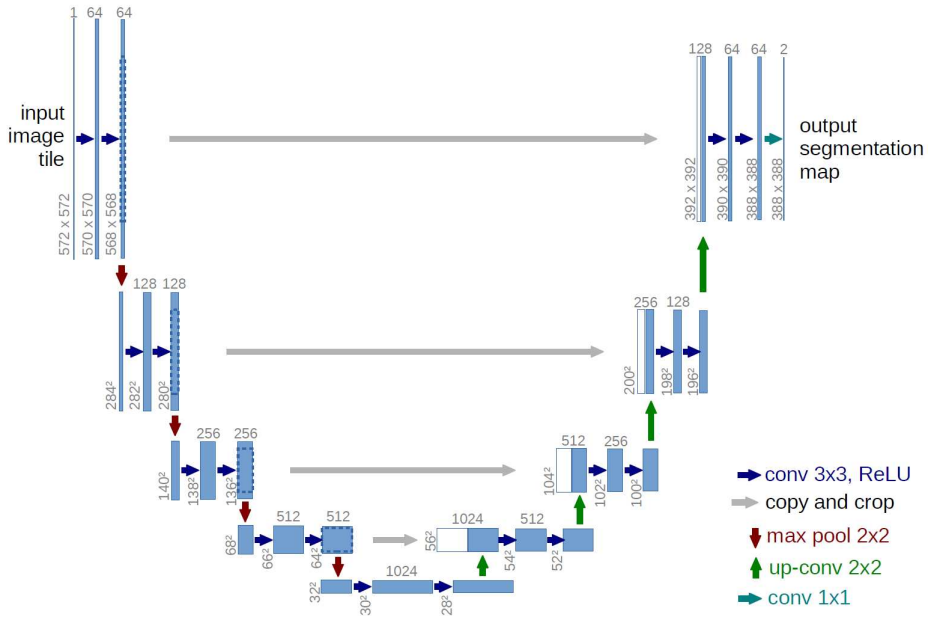

Fig. 2. Diagram of the U-Net model, in which the characteristic letter "U" is visible. Blue rectangles represent multi-channel feature maps. The current size of the maps is written on the left. The current number of channels is written above each rectangle. White rectangles are maps transferred to the decoding part of the model. Blue arrows are convolutional layers, red ones are pooling layers, and green arrows are layers that increase the resolution. Gray arrows connect feature maps obtained during encoding to their counterparts during decoding [18].

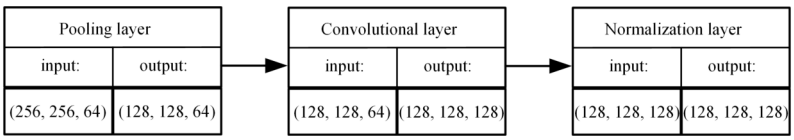

Fig. 3. Proposed encoder block. The first two numerical values are the height and the width of feature maps, and the third is the number of channels. The pooling layer reduces the resolution of feature maps. Then, the convolution layer uses filters to increase the number of channels. At the end, normalization is performed.

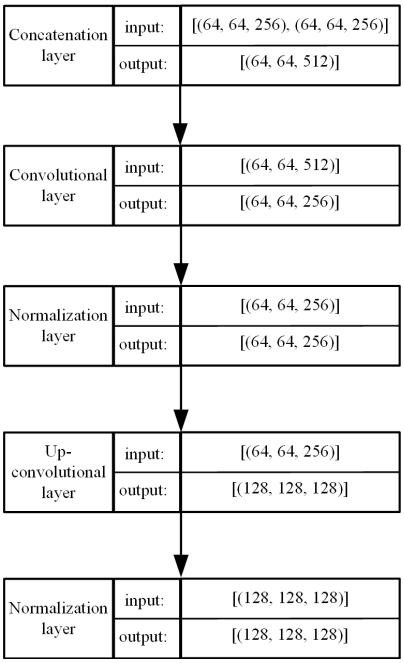

Fig. 4. Proposed decoder block. The first two numerical values are the height and width, the third is the number of channels. The concatenation layer connects the feature maps of the encoder and the decoder that have the same resolution, and then the convolutional layer reduces the number of channels. The next steps are: data normalization and the use of an up-convolutional layer which increases the resolution of feature maps. The last element of the block is the normalization layer whose output forms one of the inputs of the next concatenation layer that begins the next decoder block.

comprising 2975 images, and a subset for testing machine learning models, containing 500 images.

B. Preprocessing

Pre-processing is to shorten the network training time and to properly prepare the images so that the learning process is efficient and the highest possible results of semantic segmentation are obtained on the test set. For this purpose, image resolution change, random cropping and normalization were applied.

1) Resolution change: To improve the training time of ANNs, the original resolution of source images was reduced from 2048 $\times$ 1024 pixels to 600 $\times$ 300 pixels using the nearest neighbour method [19]. Apart from RGB channels, the rescaled images also contained a channel representing the segmentation of individual images.

2) Random cropping: In the next step, random cropping [20] was used to obtain images with the size of 256$\times$ 256 pixels. In addition, every image was mirrored with a probability of 1/2. This produces more diverse input data and reduces the risk that the network will analyse the general features of all images.

3) Normalization: The next step is data normalization. This means changing the value range of image RGB channels to the interval of [0, 1]. The last channel, which contains values representing the semantic segmentation, remains unchanged.

C. Convolutional network models used

During the study, an attempt was made to evaluate two convolutional networks, i.e. MobileNet [14], [15] and ResNet50 [16], [17], used as the encoder in the U-Net [18] model to perform semantic segmentation. Some modifications to the U-Net model were also proposed based on experiments carried out on the training set.

1) U-Net model: U-Net is a neural network model whose original purpose was the semantic segmentation of medical images [18]. The U-Net model consists of two paths which make the model diagram resemble the letter "U", namely the contraction and expansion paths representing the encoder and the decoder, respectively. Both paths are shown in Figure 2.

2) Modified U-Net model: A modified network model based on the standard U-Net model with added normalization layers, abbreviated as MU-Net, was proposed for performing the semantic segmentation. The Rectified Linear Unit (ReLU) [21] was used as the activation function. Example network encoder and decoder blocks are shown in Figures 3 and 4. Convolutional layers use 3$\times$3 filters. A 1$\times$1 filter is used for concatenation layers and at the resolution of 256$\times$256 pixels, when transition to pixel classification occurs. During the convolution, there is a descent to feature maps with the size of 8$\times$8 pixels. This network is configured only for performing the semantic segmentation. This is why the appropriate parameters were selected during many trials to train the network and the possible decrease of the accuracy during the classification of the entire image or the detection of individual objects was not taken into account.

Therefore, during many attempts to teach the network on the training set, appropriate parameters were selected

3) MobileNet: The Mobilenet [14] is a convolutional network that can be used on mobile devices. It is characterized by fewer parameters and a shorter training time than other models of convolutional networks. It has a high ratio of accuracy to the parameter number. The MobileNetV2 network [15] is an extension of the Mobilenet network. The authors mention semantic segmentation as one of the applications of this network. The main changes compared to the previous version are the use of the ReLU6 activation function instead of ReLU, and of the so-called bottleneck [22]. According to the authors' calculations, this network is more accurate than the original version, while the number of parameters is significantly reduced.

4) ResNet50: ResNet50 is a network belonging to the group of so-called residual neural networks [16] introduced in 2015, where the 50 in the name represents the number of network layers. They are characterized by the possibility of skipping some layers during the analysis. The ResNet network has a block-skipping mechanism which transfers to the next layer the parameter value processed only by the activation function. Network blocks use the bottleneck method just like in MobileNetV2. The ResNet50V2 network has a small block consisting of a normalization layer followed by a ReLU activation function. Pre-activation, i.e. the use of blocks

before fully convolutional layers, speeds up the training of the network and improves its accuracy.

5) Decoders used: A decoding part was added to each neural network used to encode image features so that the numbers of encoder and decoder parameters are similar. In addition, in the case of Mobilenet and ResNet50, the same decoder was used for v1 and v2. Because of the similar number of parameters in the MU-Net and ResNet50 encoders, the MU-Net model uses the same decoder as the ResNet50 model.

Data from Table I shows that regardless of using one decoder for the MobileNet network and another for the remaining networks, every U-Net model has a different number of decoder parameters. This is due to the different number of channels in specific encoder layers. The consequence of this is that concatenation layers that follow these layers and have these layers as input also have a different number of channels, resulting in a different number of parameters.

TABLE I
THE NUMBER OF PARAMETERS IN THE ENCODING AND DECODING PARTS OF NETWORKS BASED ON THE U-NET MODEL.

	Number of parameters		
Network	Encoder	Decoder	Entire U-Net model
MobileNet	3 228 864	2 788 834	6 017 698
MobileNetV2	2 257 984	2 288 610	4 546 594
ResNet50	23 587 712	19 483 426	43 071 138
ResNet50V2	23 564 800	15 698 722	39 263 522
MU-Net	25 163 136	19 554 850	44 717 986

III. EXPERIMENTS COMPLETED AND THEIR RESULTS

The accuracy of segmentation performed with CNNs was measured using the Jaccard index. This is the most widespread method of evaluating semantic segmentation. It allows calculating the similarity of the obtained segmentation to the manually labelled by experts. After the process of training on a set of 2,975 images, the results obtained were evaluated on a set of 500 images 256×256 pixels in size, produced by the random cropping of the original test set. It can be said that all ANNs achieved very similar results, as shown in Table II and in Figure 5.

TABLE II
TABLE SHOWING THE ACCURACY OF THE U-NET MODEL NETWORK USING SPECIFIC ENCODERS. THE RESULTS TURNED OUT TO BE VERY SIMILAR DESPITE VERY LARGE DIFFERENCES IN THE NUMBER OF PARAMETERS.

Encoding network	Jaccard index values
MobileNet	86.19%
MobileNetV2	86.20%
ResNet50	86.23%
ResNet50V2	86.27%
MU-Net	88.85%

It is worth noting that the improvements in the new versions of both MobileNet and ResNet50 led to a slight increase in the Jaccard index values of the semantic segmentation, while the number of parameters was reduced by, respectively: 24.4% and 8.9%. For this reason, only the newer versions of both networks were used in subsequent experiments that checked the accuracy of segmentation using the Jaccard index. It can be concluded that increase of performance is not caused by reducing the number of parameters, it is the result of improving the network architecture. The difference in Jaccard index values between the most and least accurate ANNs amounts to 2.7%.

A. Noise in images

The impact of noise on the accuracy of the segmentations performed was checked for 59 images from the test set from the city of Lindau. Noise was introduced in the images using the Hue, Saturation, Value (HSV) colour space and an additional Holdness parameter. The channel values of Hue vary from 0 to 180, and of Saturation and Value from 0 to 255. The Holdness parameter has values from the interval [1, 8] and is inversely proportional to the hue variation. Table III shows the segmentation results measured with the Jaccard index. Figure 6 shows an example source image before and after noise was added, and Table III shows the segmentation results measured with the Jaccard index. Noise with the values of (Hue, Saturation, Value, Holdness) = (10, 22, 22, 1) was added to all images from the test set from the city of Lindau. Even though the noise had been selected so that it would not hinder humans from recognizing any image elements, ANNs encountered a problem and Jaccard index values fell by about 20%.

TABLE III
DIFFERENCE IN ACCURACY OF U-NET MODELS BEFORE AND AFTER NOISE WAS ADDED TO IMAGES.

Encoding network	Original set	Noisy set
MobileNetV2	75.81%	51.90%
ResNet50V2	74.46%	56.52%
MU-Net	79.36%	59.48%

B. Non-standard lighting – shaded images

The 7 most shaded examples were selected from the test image set to test the impact of low light on segmentation accuracy. The results are presented in Table IV.

TABLE IV
A TABLE SHOWING THE ACCURACY OF THE U-NET NETWORK MODEL CHECKED ON IMAGES WITH POOR LIGHTING CAUSED BY SHADE.

Encoding network	Jaccard index value
MobileNetV2	84.22%
ResNet50V2	83.52%
MU-Net	86.39%

The results show that strong image shading does not hinder obtaining positive segmentation results. The approximately 2% drop in accuracy may be due to other features of the selected images.

C. Class imbalance

Class imbalance is a phenomenon in which the analysed classes are not equally represented. A dominant number of pixels belonging to one or several classes may occur in the

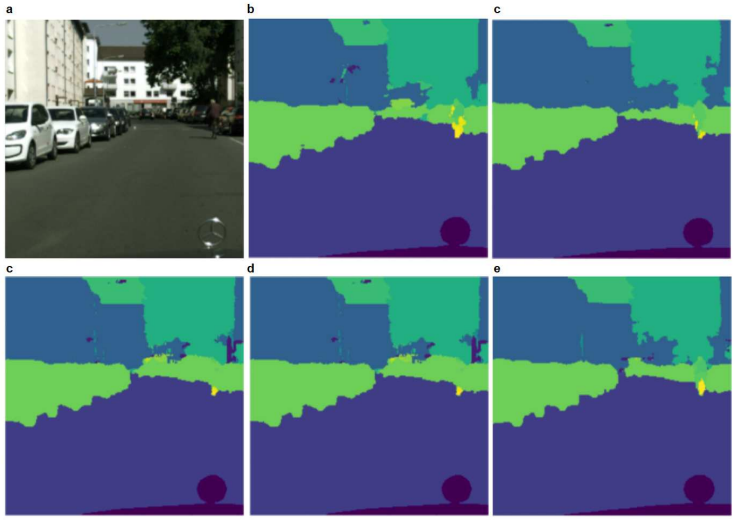

Fig. 5. Example results of a semantic segmentation on a sample image from the test set. (a) Original image (b) MobileNet (c) MobileNetV2 (d) ResNet50 (e) ResNet50V2 (f) MU-Net.

Fig. 6. Example test images from the city of Lindau (a) Original source image (b) Image with added noise with values of (Hue, Saturation, Value, Holdness) =(10, 22, 22, 1).

semantic segmentation. An example is shown in Figure 7, in which the road and vegetation are darker, and bright sunlight penetrates only to a small extent. As a result, two dominant classes are visible, namely the road and vegetation. ANNs frequently do not receive images with strongly dominant classes during training, or receive too few such images to later produce correct results when the classifier is tested. To check the segmentation results, 20 images with strongly dominating classes were selected from the test set, and the results obtained are presented in Table V. The results from Table V demonstrate a certain advantage of the ResNetV2 network in this test. The MobileNetV2 network also achieved a better result than the proposed MU-Net model, which may indicate some overtraining of this network, which produced the worst result this time.

TABLE V
A TABLE SHOWING THE ACCURACY OF THE U-NET MODEL NETWORK USING SPECIFIC ENCODERS, CHECKED ON 20 IMAGES WITH DOMINANT CLASSES.

Encoding network	Jaccard index value
MobileNetV2	76.67%
ResNet50V2	79.36%
MU-Net	74.99%

IV. CONCLUSIONS

This paper describes the practical properties of neural network models, namely MobileNet, ResNet, U-Net, and the MU-Net model, used for the semantic segmentation of images from the Cityscapes dataset [13]. The U-Net model is a very interesting approach to the problem of semantic segmentation, which is an extremely difficult area of digital image analysis. However, this model has some accuracy limitations and the constant increase of the number of parameters will not ensure satisfactory results, which is one of the conclusions. During the research, the author was able to propose an MU-Net model, i.e. an ANN dedicated to semantic segmentation, which produced results slightly better than universal networks like MobileNet or ResNet. However, the MobileNetV2 network turned out to be the most interesting and promising ANN used. It has a

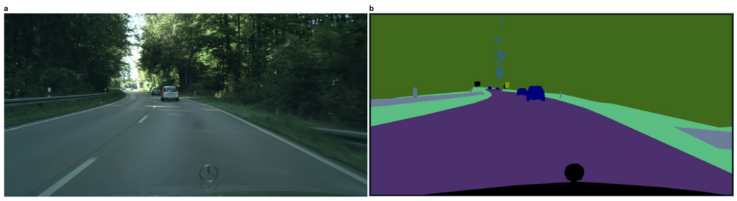

Fig. 7. Example image with semantic segmentation showing non-standard lighting and class imbalance. Most of the image is covered by the road and vegetation, while bright sunlight and moving cars occupy a small fragment of the image. (a) Original image. (b) Semantic segmentation containing mainly two classes.

very good ratio of accuracy to the number of parameters and, at the same time, is less affected by non-standard phenomena in images. Due to the constantly increasing computing power of mobile devices, neural networks designed for analysing images on mobile devices with even better parameters can be expected in the near future. In future research, it is definitely worth investigating improving the accuracy of the semantic segmentation of noisy images and the issue of class imbalance. There are also other interesting directions of research, e.g. performing a semantic segmentation that simulates autonomous vehicle driving using recorded videos, and carrying out a three-dimensional semantic segmentation of urban scenes. It is also worth trying to supplement training sets using various augmentation methods, but keeping in mind the need to prevent learning the wrong patterns.

REFERENCES

[1] 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* 2015, pp. 3431-3440, https://doi.org/10.1109/CVPR.2015.7298965.

[2] L.C. Chen, Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *In Proceedings of the European conference on computer vision (ECCV)* 2018, pp. 801-818, https://doi.org/10.1007/978-3-030-01234-2_49.

[3] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32(6), 2020, pp. 2547-2560, https://doi.org/10.1109/TNNLS.2020.3006524.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image segmentation using deep learning: A survey, " *IEEE transactions on pattern analysis and machine intelligence*, vol. 44(7), 2021, pp. 3523-3542, 10.1109/TPAMI.2021.3059968.

[5] P. Malík, Š. Krištofík K. Knapová, "Instance segmentation model created from three semantic segmentations of mask, boundary and centroid Pixels verified on GlaS dataset, " *In 2020 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 569-576*, http://dx.doi.org/10.15439/2020F175.

[6] L. Ming, Y. Qingbo, L. Mingyu, "Retinal blood vessel segmentation based on multi-scale deep learning, " *In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1-7*, http://dx.doi.org/10.15439/2018F127

[7] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.I. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, " *IEEE transactions on pattern analysis and machine intelligence*, vol. 40(4), 2017, pp. 834-848, https://doi.org/10.1109/TPAMI.2017.2699184.

[8] V. Badrinarayanan, A. Kendall and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation, " *IEEE transactions on pattern analysis and machine intelligence*, vol. 39(12), 2017, pp. 2481-2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[9] M. Siam, S. Elkerdawy, M. Jagersand and S. Yogamani, "Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges, " *In 2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp. 1-8, https://doi.org/10.1109/ITSC.2017.8317714.

[10] Z. W. Hong, C. Yu-Ming, S. Y. Su, T. Y. Shann, Y. H. Chang, H. K. Yang, *ldots* & C. Y. Lee, "Virtual-to-real: Learning to control in visual semantic segmentation, " *arXiv preprint*, 2018, 1802.00285, https://doi.org/10.48550/arXiv.1802.00285.

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks, " *Communications of the ACM*, 2017, vol. 60(6), pp. 84-90, https://doi.org/10.1145/3065386.

[12] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation, " *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440, https://doi.org/10.1109/CVPR.2015.7298965.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele, "The cityscapes dataset for semantic urban scene understanding, " *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213-3223, https://doi.org/10.1109/CVPR.2016.350.

[14] A. G. Howard, Z. Menglong, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications, " *CoRR*, 2017, abs/1704.04861, https://doi.org/10.48550/arXiv.1704.04861.

[15] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation, " *CoRR*, 2018, abs/1801.04381.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition, " *CoRR*, 2015, abs/1512.03385, https://doi.org/10.1109/CVPR.2016.90.

[17] K. He, X. Zhang, S. Ren, J. Sun, "Identity mappings in deep residual networks, " *CoRR*, 2016, abs/1603.05027, https://doi.org/10.1007/978-3-319-46493-0_38.

[18] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation, " *In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany*, Part III 18, pp. 234-241, https://doi.org/10.1007/978-3-319-24574-4_28.

[19] O. Rukundo, H. Cao, "Nearest neighbor value interpolation, " *arXiv preprint*, 2012, 3:25:30, https://doi.org/10.14569/IJACSA.2012.030405.

[20] R. Takahashi, T. Matsubara, K. Uehara, "Data augmentation using random image cropping and patching for deep CNNs, " *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, vol. 30(9), pp. 2917-2931, https://doi.org/10.1109/TCSVT.2019.2935128.

[21] V. Nair, G. E. Hinton, "Rectified linear units improve restricted boltzmann machines, " *In Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807-814.

[22] E. R. De Rezende, G. C. Ruppert, A. Theophilo, E. K. Tokuda, T. Carvalho, "Exposing computer generated images by using deep convolutional neural networks. Signal Processing, " *Image Communication*, 2018, vol. 66, pp. 113-126, https://doi.org/10.1016/j.image.2018.04.006.