ISSN 2464-4617 (print)
ISSN 2464-4625 (DVD)
Computer Science Research Notes
CSRN 2901
WSCG Proceedings Part I

Performance Assessment of Convolutional Neural Networks for Semantic Image Segmentation

Alexander Leipnitz Tilo Strutz Oliver Jokisch
Institute of Communications Engineering
Leipzig University of Telecommunications (HfTL)
Gustav-Freytag-Str. 43-45
04277, Leipzig, Germany
{leipnitz,strutz,jokisch}@hft-leipzig.de

Abstract

Convolutional neural networks are applied successfully for image classification and object detection. Recently, they have been adopted to semantic segmentation tasks and several new network architectures have been proposed. With respect to automotive applications, the Cityscapes dataset is often used as a benchmark. It is one of the biggest datasets in this field and consists of a training, a validation, and a test set. While training and validation allow the optimisation of these nets, the test dataset can be used to evaluate their performance.
Our investigations have shown that while these networks perform well for images of the Cityscapes dataset, their segmentation quality significantly drops when applied to new data. It seems that they have limited generalisation abilities. In order to find out whether the image content itself or other image properties cause this effect, we have carried out systematic investigations with modified Cityscapes data. We have found that camera-dependent image properties like brightness, contrast, or saturation can significantly influence the segmentation quality. This papers presents the results of these tests including eight state-of-the-art CNNs. It can be concluded that the out-of-the-box usage of CNNs in real-world environments is not recommended.

Keywords

Convolutional neural network, Semantic segmentation, Generalisation abilities

1 INTRODUCTION

Recent developments of convolutional neural networks for semantic segmentation led to impressive results on validation and test datasets. However, the datasets for this performance measurement and the dataset on which the training procedure was based share the same image characteristics as, for example, the lighting conditions and the acquisition environment (e.g. camera type and settings). A very prominent and one of the largest datasets in the field of semantic segmentation is the Cityscapes dataset [1]. It shows urban scenes of 50 different cities. Common features between the training, validation and test datasets have been prevented by having no city being doubly represented in one of the sub datasets. Nevertheless, dependencies still exist between them due to the standardised capture settings. Images from real-world scenarios can be much more diverse, especially when using different cameras or settings, and state-of-the-art CNNs are expected to cope with these image variations. So far it was not known how well CNNs trained with the Cityscapes dataset perform under various lighting conditions or in rural areas.

Domain adaptation and transfer-learning are well-known methods to adjust trained models to new conditions or type of scenes. However, they are typically not used to increase the generalisation abilities but to shift the application range. The use of CNNs in real-world applications such as autonomous driving, on the other hand, requires them to function optimally under all kinds of conditions.

In this paper, eight state-of-the-art CNNs are compared using out-of-the-box models available on the Internet in order to assess their generalisation abilities. The investigations have revealed that most of the evaluated nets do not cope well with images having varying characteristics which can be caused by different camera systems. These variations have been simulated by modifying brightness, saturation, or contrast of the images. In a second test, images that do not belong to the Cityscapes dataset have been presented to the CNNs in order to visually evaluate the resulting segmentation masks.

2 RELATED WORK

The focus on a specified dataset and therefore overfitting and limited generalisation abilities of neural net-

works are a known issue and have already been covered in literature.

Adversarial examples are images that can trick a neural network into a false classification by slightly modifying an otherwise correctly classified image. In [2] these cases are described, searched and systematically provoked. Wang et al. presents a theoretical analysis of the functionality of adversarial examples and possible countermeasures [3].

Even small image transformations like object translation can drastically influence CNNs for object recognition [4], which is validated in [5]. Rosenfeld et al. additionally showed that the object position in an image and "object transplanting" from one image to another has not only an influence on its own detection rate but also the detection of other objects in the image.

Global image modifications can also have an effect on the CNN results. In [6], the robustness of classifiers against added random and *semi-random* noise in the samples has been researched and their impact on the classification rate has been proven. In addition, the effects of blur, noise, contrast, JPEG and JPEG2000 related compression artefacts are evaluated for image classification tasks for deep neural networks in [7]. Vasiljevic et al. extend this in [8] by investigating the impact of blurred images to semantic segmentation tasks.

Solutions have already been proposed to overcome these problems. In [9], the robustness of a classification deep neural networks has been improved by including distorted copies of the original images in the training process. They use downsampling, JPEG compression and random cropping for stability training. Random cropping is one form of data augmentation that is used for some CNNs compared in this paper.

An alternative solution to make CNNs robust against these type of modifications is proposed in [10]. Based on the image quality, different paths inside of the network are selected to maximize the classification result.

The impact of image modifications on the task of image classification has already been comprehensively studied in the present literature. The effects are expected to carry over to semantic segmentation tasks. This paper examines the influence of image modifications with real application background (image brightness, contrast, and saturation). Additionally, the segmentation of unknown images is evaluated. This allows a comparison of different network architectures beyond their test dataset segmentation scores.

3 INVESTIGATIONS

CNNs for image classification and object detection use a cascade of convolutional layers and downsampling to compute a vector containing the class scores from the resulting feature maps via fully connected layers. Changes to this classic CNN architecture have been made to cope with the task of image segmentation. The class scores are computed for each pixel on low resolution features maps and different upsampling techniques have been developed to obtain a score map in the original image resolution. This is known as an encoder-decoder network. The basic structure can be seen in **Fig. 1**a) with the feature maps being the possible intermediate result after many different modules or layers (convolution, downsampling or upsampling).

CNN	Test-Set. $mIoU[\%]$	Val.-Set $mIoU[\%]$	Fig. 1
DeepLabv3+	82.1	78.7	a)
PSPNet	81.2	77.0	a)
TuSimple-DUC	77.6	83.7	a)
RefineNet	73.6	75.3	b)
LRR	71.9	72.5	a)
ICNet	69.5	67.7	b)
ESPNet	60.3	59.1	a)
ENet	58.3	53.5	a)

Table 1: Ranking of the CNNs based on their official Cityscapes test dataset results (Test-Set. *mIoU*), the measured results on the validation dataset (Val.-Set. *mIoU*) and the categorization of their network architecture

Every network architecture takes a different approach on this structure with more or less drastic modifications to the encoder or decoder. A major change is the introduction of additional branches on the encoder site that process downsampled versions of the original image in parallel. This is called a multi-path encoder-decoder structure (**Fig. 1**b) and is explained in more detail in the following subsections.

3.1 Selected Convolutional Neural Networks

On the Cityscapes website [11] a lot of CNNs are ranked regarding their segmentation performance on the test dataset. The **Tab. 1** lists some of these network architectures along with their performance ratings (mean Intersection over Union metric [12], *mIoU*) for the Cityscapes test and validation dataset. This selection is based on code availability and covers a broad range of segmentation capabilities (highest, middle and also low *mIoU* scores) and objectives. DeepLabv3+, PSPNet, TuSimple-DUC, RefineNet, and LRR focus on the highest segmentation score possible, while ICNet, ESPNet and ENet also have real-time inference in mind. They all can be categorized as an encoder-decoder (Fig. 1a) or multi-path encoder-decoder (Fig. 1b) network architecture.

 ISBN 978-80-86943-37-4

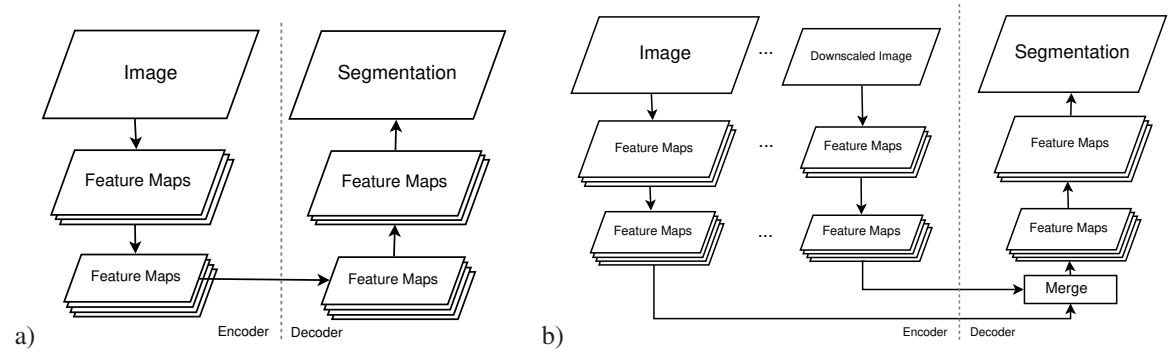

Figure 1: Basic CNN architecture for semantic segmentation; a) encoder-decoder; b) multi-path encoder-decoder

3.1.1 DeepLabv3+

The Deep Labelling Network (DeepLabv3+) [13] is the highest ranked network on the test dataset but only the second best network on the validation dataset of the examined networks (see Table 1). It is based on DeepLabv3 [14] and introduces a modified Xception module [15] as network backbone. In general, the network architecture is based on a complex and powerful encoder that relies on parallel atrous convolution with different rates to enlarge the field-of-view (Atrous Spatial Pyramid Pooling). The decoder module handles the upsampling and combines the final encoder output with low-level features from previous layers with the same spatial size to recover object segmentation details.

3.1.2 PSPNet

The Pyramid Scene Parsing Network (PSPNet) [16] is the second highest ranked network on the test dataset and third highest network on the validation dataset. Its encoder is ResNet-based [27] while the decoding is performed by a pyramid pooling module. This proposed module uses parallel pooling and convolution with different sized kernels/filters. This aims at a broader receptive field by including local and global context information. The resulting feature maps are then upsampled and concatenated before a final convolutional layer generates the segmentation-output.

3.1.3 TuSimple-DUC

The ResNet-DUC-HDC alias TuSimple-DUC [17] yields the highest segmentation performance on the Cityscapes validation dataset and third highest segmentation performance on the test dataset. It introduces a combination of Dense Upsampling Convolution (DUC) and Hybrid Dilated Convolution (HDC) as an addition to the ResNet-based architecture. The encoder consists of the ResNet and HDC layers while the decoding is performed by the DUC layers.

3.1.4 RefineNet

The Multi-Path Refinement Network (RefineNet) [18] uses parallel processing of the original and downsampled version of the input image. It is a multi-path encoder-cecoder network architecture. The RefineNet-block consists of two ResNet-based Residual Convolution Unit (RCU) for each input, Multi-resolution Fusion, Chained Residual Pooling and final RCU to compute the output feature map.

3.1.5 LRR

The LRR architecture (Laplacian Pyramid Reconstruction and Refinement) [19] introduced the two name-giving techniques. The low-resolution segmentation map is upsampled and refined with the help of higher-resolution feature maps in areas with high uncertainty.

3.1.6 ICNet

The Image Cascade Network (ICNet) [20] is a variation of the PSPNet with focus on real-time inference and is also a multi-path encoder-decoder network architecture with three encoder branches. The PSPNet architecture is only used for a downsampled version of the input image to save computational time. The resulting small spatial sized feature map gets upsampled and merged with feature maps that originated from a higher sampled and later the original sized input image. They both only have passed through a limited number of convolutional layers. After these two Cascade Feature Fusion (CFF) modules, only upsampling and a final convolutional layer is applied to get the final segmentation output.

3.1.7 ESPNet

The Efficient Spatial Pyramid of Dilated Convolutions Network (ESPNet) [21] has introduced an ESP module replacing the standard convolutional layer. It consists of a point-wise convolution and a spatial pyramid of dilated convolutions that result in an computational efficient and bigger receptive field. The network architecture consists of normal convolutional layers, ESP modules and deconvolutional layers [22] for upsampling.

3.1.8 ENet

The Efficient Neural Network (ENet) [23] is especially designed for real-time inference. For this reason, the

 ISBN 978-80-86943-37-4

network architecture is very small compared to the other presented networks by limiting the number of layers and size of the feature maps. It is based on the ResNet architecture.

3.2 Experimental Setup

The generalisation abilities of CNNs can be evaluated in two ways. Firstly, the images of a known dataset can be modified and the change in segmentation performance is measured. Secondly, the segmentation output for unknown images can be visually demonstrated and discussed. The out-of-the-box performance of each network architecture has been tested with the provided and here named models:

- DeepLabv3+: *deeplabv3_cityscapes_train*; model-variant: xception_65;
- TuSimple-DUC: *ResNet_DUC_HDC_CityScapes*
- RefineNet: *refinenet_res101_cityscapes.mat*
- LRR: *LRR4x-VGG16-CityScapes-coarse-and-fine*
- ESPNet: *espnet_p_2_q_8.pth*
- ENet: *cityscapes_weights.caffemodel*.

The official implementation of the PSPNet and ICNet could not be used due to Hard- and Software incompatibilities. Instead, the Tensorflow implementations [24, 25] have been utilized with these models:

- PSPNet: *pspnet101-cityscapes*
- ICNet: *icnet_cityscapes_train_30k.npy*.

The Cityscapes test dataset is not public available so all further tests have to be performed on the validation dataset. Changing the brightness, contrast and saturation of its images represents realistic scenarios in a real-world environment.

3.2.1 Brightness

A brightness modification can be described by an offset b to the R, G and B values of each pixel:

$$R' = \max(\min(R+b,255),0) \quad (1)$$
$$G' = \max(\min(G+b,255),0) \quad (2)$$
$$B' = \max(\min(B+b,255),0) \quad (3)$$

with b ranging between $[-50;50]$ in increments of 10 in our tests. The **Fig. 2** shows the effect of the maximum brightness changes on a Cityscapes validation dataset image. The range of b is chosen so that the resulting images still look realistic and can arise by under- or overexposing the camera sensor. It is to be expected that the influence of this modification on the segmentation results is rather low because the gradients are not affected. Only pixels that have to be clipped to 0 or 255 change their properties in a non-linear way.

3.2.2 Contrast

A contrast modification corresponds to the multiplication with a factor c:

$$R' = \max(c \cdot R,255) \quad (4)$$
$$G' = \max(c \cdot G,255) \quad (5)$$
$$B' = \max(c \cdot B,255) \quad (6)$$

with c ranging between $[0.5;2]$ in our tests. We choose the values $c \in \{0.5;0.7;1;1.4;2\}$. The influence of the maximum contrast modification can be seen in **Fig. 3**. This modification is more drastic compared to the brightness change because it is affecting the neighbouring relations between pixels and compresses or stretches the accompanying histogram.

3.2.3 Saturation

The saturation of an image can easily be modified by transforming the image from the RGB to the HSV colour-space first. According to [26], the saturation s_{hsv} is defined by:

$$s_{hsv} = \begin{cases} 0 & \text{if } R=G=B \\ 255 \cdot \frac{max(R,G,B)-min(R,G,B)}{max(R,G,B)} \end{cases} . \quad (7)$$

Analogous to the brightness, the saturation modification can be applied by an offset s:

$$s'_{hsv} = \max(\min(s_{hsv}+s,255),0) \quad (8)$$

with s ranging between $[-40;40]$ in increments of 10 in our tests. **Fig. 4** shows the maximum saturation modification. Due to the transformation, the pixel values change non-linear and the neighbourhood-relations between pixel get distorted the most. It is expected that this modification has the most influence on the segmentation performance.

4 RESULTS AND DISCUSSION

4.1 Modified Cityscapes Validation Dataset

The graphs in **Fig. 5** show the influence of the image modification on the *mIoU* score for each CNN applied to the entire Cityscapes validation dataset.

The images in the Cityscapes dataset are rather dark, which is indicated by the low mean of the colour channels for the images of the training dataset (R: 73.19, G: 82.91 B: 72.39). Therefore, the reduction of brightness can make many objects black. They become indistinguishable resulting in a drastic decrease of the segmentation performance of all nets for a negative b in Fig. 5a). Increasing the brightness does not seem to affect the CNNs output much except for the ICNet.

The influence of the contrast modification in Fig. 5b) surprisingly has the least affect on the segmentation

 ISBN 978-80-86943-37-4

Figure 2: Brightness modification on a Cityscapes validation image; a) original image; b) $b = -50$; c) $b = 50$

Figure 3: Contrast modification on a Cityscapes validation image; a) original image; b) $c = 0.5$; c) $c = 2$

Figure 4: Saturation modification on a Cityscapes validation image; a) original image; b) $s = -40$; c) $s = 40$

performance, although the images in Fig. 3 appear to be the darkest or brightest of all image modifications in their extreme points. Only a very big c leads to a significant decrease probably due to clipping of the pixel values to 255.

As expected, the change in saturation has the greatest impact on the segmentation performance. The curves in Fig. 5c) drop off rather rapidly with an increasing $|s|$. DeepLabv3+ and RefineNet seem to be the least affected by this modification as the flat curves indicate.

Some curves intersect with others. This indicates that some network architectures have a lower segmentation performance on the "default" images but a higher robustness against image modification and therefore less over-fitting. The DeepLabv3+ shows in all three graphs the best generalisation abilities. It has the second highest *mIoU* score and flatter curves compared to TuSimple-DUC which it also intersects. Therefore, the DeepLabv3+ network architecture has the best compromise between *mIoU* score and generalisation abilities in this test.

4.2 Unknown Images Dataset

In a real-world application, a CNN is exposed to various different scenes. The validation and test datasets are usually from the same source and feature the same bias (camera settings, preference by the photographer etc.). This bias is also learned by the CNNs and prevents them from having good generalisation capabilities. To test this further, a visual segmentation evaluation has been performed with unknown images from a completely different source. **Fig. 6** to **Fig. 9** (each a)) show four example images with similar content to the Cityscapes images and exclusively known objects/classes. They only tend to be a bit brighter and originate from a different camera. Even without having ground truth data for these images available, the segmentation outputs produced by the nets in b) - i) show remarkable differences that allow a subjective comparison. The relation between the colours and the classes of the Cityscapes dataset can be seen in **Fig. 10**. The conclusion from the previous section is confirmed with DeepLabv3+ producing the best looking segmentation output and showing the best generalisation performance. The segmentation is almost perfect with only the semantic meaning being wrong in some cases. The biggest problem seems to be the semantic segmentation of the grassland where the border is inexact and the classes "vegetation", "terrain" and "sidewalk" are assigned in an inconsistent way. The second best network architecture in this test appears to be the RefineNet, whose segmentation has the same but more obvious problems. The others CNNs often drastically fail to segment the objects correctly and make fundamental errors regarding the classification. The class "building" stands out by being assigned incorrectly to different areas in the images.

ESPNet and ENet have the biggest problems with the images. The segmentation of objects is mostly wrong and often the segments are classified into the wrong class. Especially Fig. 7 is negatively noticeable here.

4.3 Comparison of Results Between Both Datasets

In our tests, DeepLabv3+, first place on the test dataset but only second place on the validation dataset, seems to be the least influenced by the image modifications and showed the best segmentation output for the four unknown images.

ISBN 978-80-86943-37-4

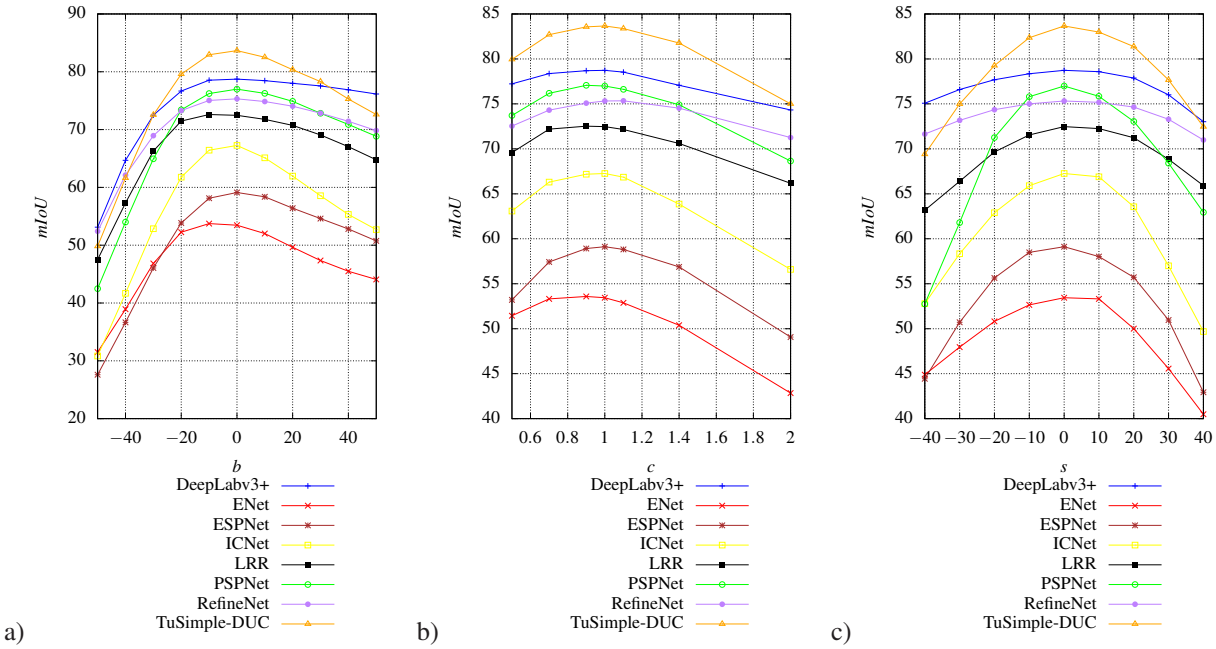

Figure 5: Influence on the *mIoU* score a) brightness; b) contrast; c) saturation

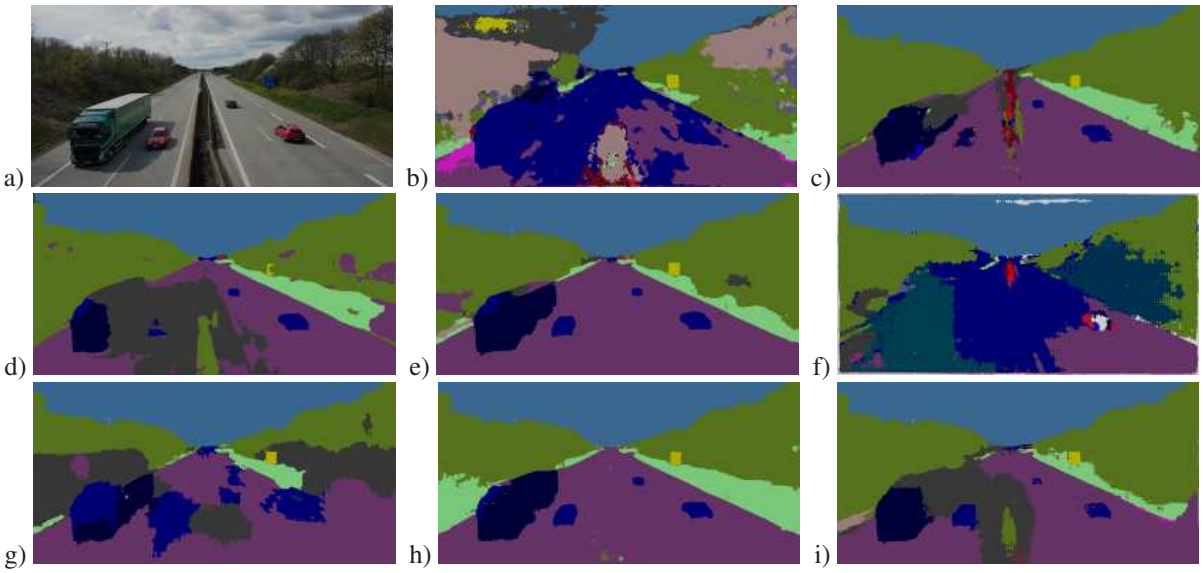

Figure 6: Segmentation-output; a) original image; b) ENet; c) ICNet; d) PSPNet; e) RefineNet; f) ESPNet; g) LRR; h) DeepLabv3+; i) TuSimple-DUC

The investigated shifts of brightness, saturation, and contrast are realistic modifications that can occur under various practical conditions, and convolutional neural networks should be able to cope with them.

The surveyed CNNs use a variety of different preprocessing steps but there does not seem to be a correlation between them and the results in this paper. ICNet, PSPNet, TuSimple-DUC and LRR subtract a fixed value for each colour channel to distribute the pixel values around zero, while ESPNet normalizes the input image with the Cityscapes dataset mean and its standard deviation. DeepLabv3+ also normalizes the pixel values x to $[-1;1]$ by $x' = (2/255) \cdot x - 1.0$. The other network architectures (ENet and RefineNet) do not use any image preprocessing steps. The DeepLabv3+ and RefineNet showed the best generalisation abilities in both tests despite their fundamentally different network architectures and preprocessing methods. The reasons for the divers robustness against varying image characteristics could not be clarified with our experimental set-up yet.

ISBN 978-80-86943-37-4

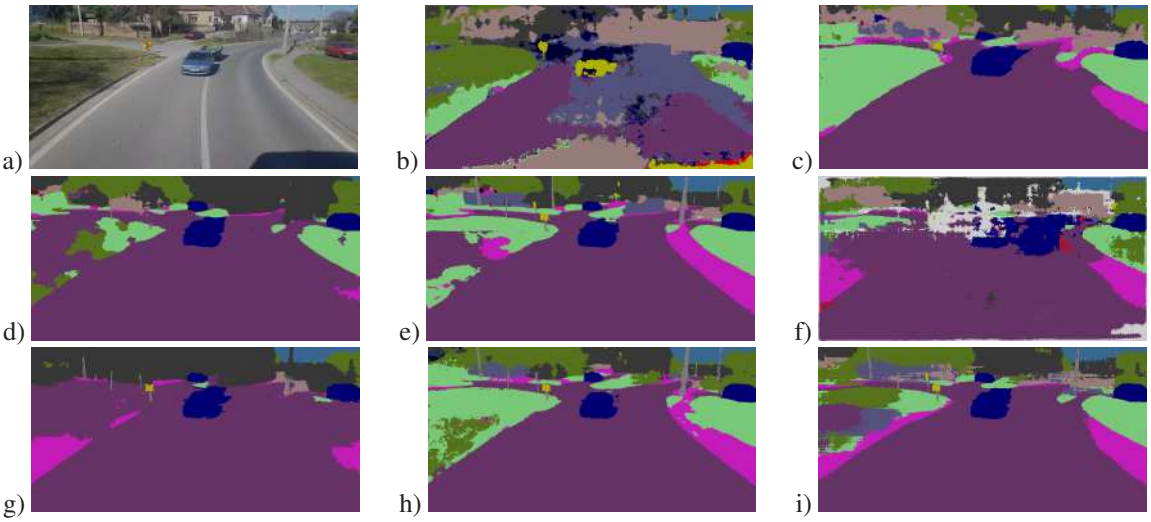

Figure 7: Segmentation-output; a) original image; b) ENet; c) ICNet; d) PSPNet; e) RefineNet; f) ESPNet; g) LRR; h) DeepLabv3+; i) TuSimple-DUC

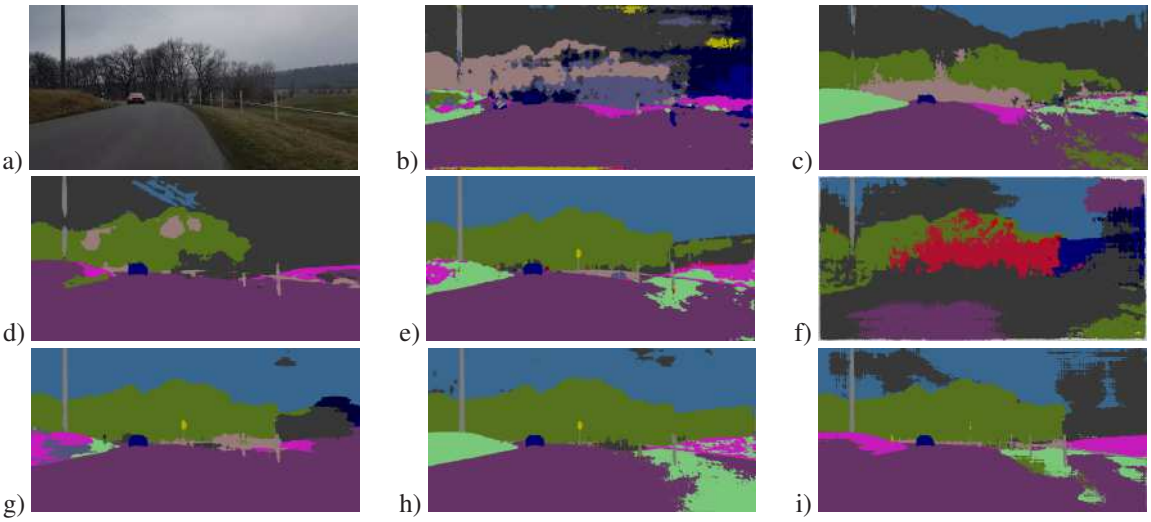

Figure 8: Segmentation-output; a) original image; b) ENet; c) ICNet; d) PSPNet; e) RefineNet; f) ESPNet; g) LRR; h) DeepLabv3+; i) TuSimple-DUC

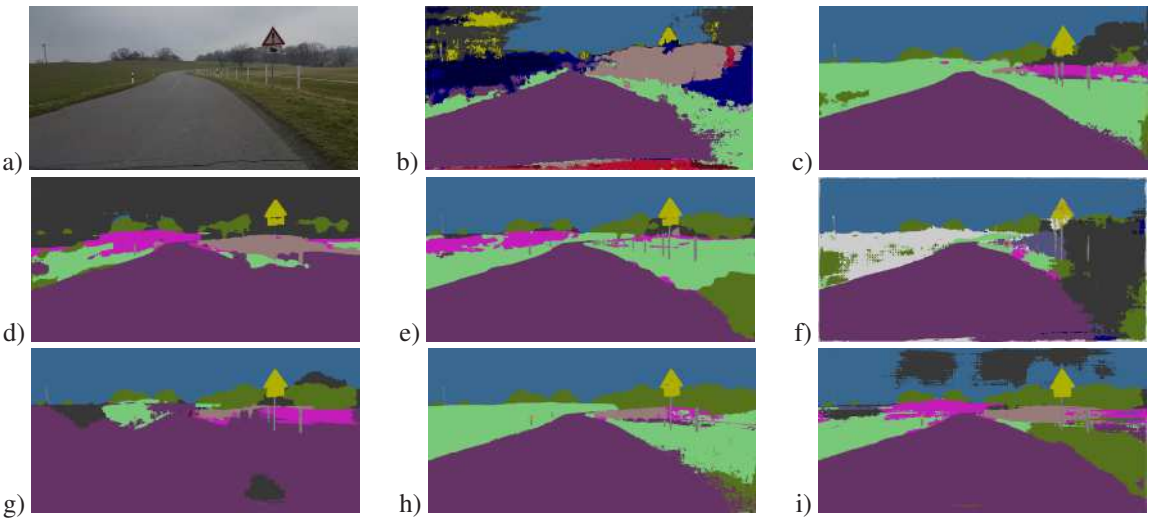

Figure 9: Segmentation-output; a) original image; b) ENet; c) ICNet; d) PSPNet; e) RefineNet; f) ESPNet; g) LRR; h) DeepLabv3+; i) TuSimple-DUC

ISBN 978-80-86943-37-4

Figure 10: Colourmap of the Cityscapes classes

5 CONCLUSIONS

Our investigations show that (i) modern CNNs are sensitive to simple image modifications in the validation dataset and that (ii) a high segmentation score on the validation or test dataset is not necessarily an indicator for a good generalisation capability of network architectures. We assume that the compared neural networks did not primarily learn the structural properties of objects in the scene, but some colour properties which coincide with objects. Consequently, segmentation scores on validation and test data are not sufficient as a benchmark test. To select a powerful network architecture, also the generalisation capability in a real-world application need to be considered.

To support reproducible research, all scripts, CNN models and images are provided in [29].

6 ACKNOWLEDGMENTS

We acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the framework of the Era.Net HARMONIC project (project number 01DJ18011). Further thank goes to the unknown reviewers, whose comments allowed us to improve the initial manuscript.

7 REFERENCES

[1] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R.: Intriguing properties of neural networks. *arXiv preprint* arXiv: 1312.6199, 2013.

[3] Wang, B.; Gao, J.; Qi, Y.: A Theoretical Framework for Robustness of (Deep) Classifiers against Adversarial Examples. *arXiv preprint* arXiv: 1612.00334, 2016.

[4] Azulay, A.; Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations?. *arXiv preprint* arXiv: 1805.12177, 2018.

[5] Rosenfeld, A.; Zemel, R.; Tsotsos, J. K.: The Elephant in the Room. *arXiv preprint* arXiv: 1808.03305, 2018.

[6] Fawzi, A.; Moosavi-Dezfooli, S.M.; Frossard, P.: Robustness of classifiers: from adversarial to random noise. *Advances in Neural Information Processing Systems*, 2016, 1632–1640.

[7] Dodge, S.; Karam, L.: Understanding how image quality affects deep neural networks. *Eighth International Conference on Quality of Multimedia Experience, QoMEX*, Lisbon, Portugal, Jun. 2016, 1–6.

[8] Vasiljevic, I.; Chakrabarti, A.; Shakhnarovich, G.: Examining the Impact of Blur on Recognition by Convolutional Networks. *arXiv preprint* arXiv: 1611.05760, 2016.

[9] Zheng, S.; Song, Y.; Leung, T.; Goodfellow, I.: Improving the robustness of deep neural networks via stability training. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 4480–4488.

[10] Ghosh, S.; Shet, R.; Amon, P.; Hutter, A.; Kaup, A.: Robustness of Deep Convolutional Neural Networks for Image Degradations. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, 2916–2920.

[11] Cityscapes Dataset: https://www.cityscapes-dataset.com/benchmarks/#scene-labeling-task/ last visited 25th April 2019.

[12] Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111.1, 2015, 98–136.

[13] Chen, L. C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *European Conference on Computer Vision (ECCV)*, 2018.

[14] Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint* arXiv:1706.05587, 2017.

[15] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 1800–1807.

[16] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J.: Pyramid Scene Parsing Network. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, 2881–2890.

[17] Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G.: Understanding Convo-

ISBN 978-80-86943-37-4

lution for Semantic Segmentation. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, 1451–1460.

[18] Lin, G.; Milan; A., Shen, C.; Reid, I.D.: RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, 5168–5177.

[19] Ghiasi, G.; Fowlkes, C.C.: Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation. *Proc. of the European Conference on Computer Vision (ECCV)*, 2016, 519–534.

[20] Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J.: ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *Proc. of the European Conference on Computer Vision (ECCV)*, 2017, 405–420.

[21] Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H.: ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. *Proc. of the European Conference on Computer Vision (ECCV)*, 2018.

[22] Noh, H.; Hong, S.; Han, B.: Learning Deconvolution Network for Semantic Segmentation. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, 1520–1528.

[23] Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E.: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv preprint* arXiv: 1606.02147, 2016.

[24] PSPNet: https://github.com/hellochick/PSPNet-tensorflow last visited 25th April 2019.

[25] ICNet: https://github.com/hellochick/ICNet-tensorflow last visited 25th April 2019.

[26] Smith, A.R.: Color gamut transform pairs. *ACM Siggraph Computer Graphics*, 12.3, 1978, 12–19.

[27] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 770–778.

[28] Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88.2, 2010, 303–338.

[29] http://www1.hft-leipzig.de/leipnitz/papers/CNNrobustness-resources/ last visited 25th April 2019.

 ISBN 978-80-86943-37-4