International Journal of Computer Vision (2021) 129:462–483
https://doi.org/10.1007/s11263-020-01383-2

Benchmarking the Robustness of Semantic Segmentation Models with Respect to Common Corruptions

Christoph Kamann[1] · Carsten Rother[1]

Received: 14 April 2020 / Accepted: 7 September 2020 / Published online: 30 September 2020

Abstract

When designing a semantic segmentation model for a real-world application, such as autonomous driving, it is crucial to understand the robustness of the network with respect to a wide range of image corruptions. While there are recent robustness studies for full-image classification, we are the first to present an exhaustive study for semantic segmentation, based on many established neural network architectures. We utilize almost 400,000 images generated from the Cityscapes dataset, PASCAL VOC 2012, and ADE20K. Based on the benchmark study, we gain several new insights. Firstly, many networks perform well with respect to real-world image corruptions, such as a realistic PSF blur. Secondly, some architecture properties significantly affect robustness, such as a Dense Prediction Cell, designed to maximize performance on clean data only. Thirdly, the generalization capability of semantic segmentation models depends strongly on the type of image corruption. Models generalize well for image noise and image blur, however, not with respect to digitally corrupted data or weather corruptions.

Keywords Semantic segmentation · Corruption robustness · Common image corruptions · Realistic image corruptions

1 Introduction

In recent years, deep convolutional neural networks (DCNNs) have set the state-of-the-art on a broad range of computer vision tasks (Krizhevsky et al. 2012; He et al. 2016; Simonyan and Zisserman 2015; Szegedy et al. 2015; LeCun et al. 1998; Redmon et al. 2016; Chen et al. 2015; Goodfellow et al. 2016). The performance of CNN models is generally measured using benchmarks of publicly available datasets, which often consist of clean and post-processed images (Cordts et al. 2016; Everingham et al. 2010). However, it has been shown that model performance is prone to image corruptions (Zhou et al. 2017; Vasiljevic et al. 2016; Hendrycks and Dietterich 2019; Geirhos et al. 2018; Dodge and Karam 2016; Gilmer et al. 2019; Azulay and Weiss 2019; Kamann and Rother 2020), especially image noise decreases the performance significantly.

Image quality depends on environmental factors such as illumination and weather conditions, ambient temperature, and camera motion since they directly affect the optical and electrical properties of a camera. Image quality is also affected by optical aberrations of the camera lenses, causing, e.g., image blur. Thus, in safety-critical applications, such as autonomous driving, models must be robust towards such inherently present image corruptions (Hasirlioglu et al. 2016; Kamann et al. 2017; Janai et al. 2020).

In this work, we present an extensive evaluation of the robustness of semantic segmentation models towards a broad range of real-world image corruptions. Here, the term *robustness* refers to training a model on clean data and then validating it on corrupted data. We choose the task of semantic image segmentation for two reasons. Firstly, image segmentation is often applied in safety-critical applications, where robustness is essential. Secondly, a rigorous evaluation for real-world image corruptions has, in recent years, only been conducted for full-image classification and object detection, e.g., most recently Geirhos et al. (2018), Hendrycks and Dietterich (2019), and Michaelis et al. (2019).

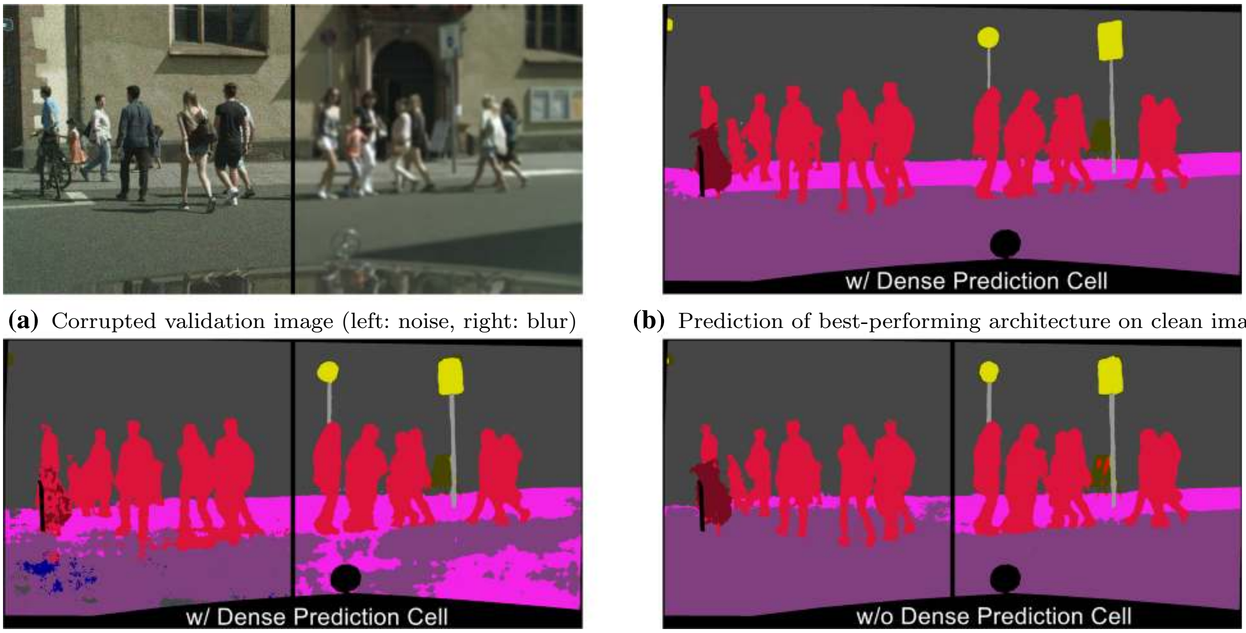
When benchmarking semantic segmentation models, there are, in general, different choices such as: (i) comparing different architectures, or (ii) conducting a detailed ablation study of a state-of-the-art architecture. In contrast to Geirhos et al.

Communicated by Daniel Scharstein.

✉ Christoph Kamann
of182@uni-heidelberg.de

Carsten Rother
carsten.rother@iwr.uni-heidelberg.de

[1] Visual Learning Lab, HCI/IWR, Heidelberg University, Heidelberg, Germany



(a) Corrupted validation image (left: noise, right: blur)

(b) Prediction of best-performing architecture on clean image

(c) Prediction of best-performing architecture on corrupted image

(d) Prediction of ablated architecture on the corrupted image

Fig. 1 Results of our ablation study. Here we train the state-of-the-art semantic segmentation model DeepLabv3+ on clean Cityscapes data and test it on corrupted data. **a** A validation image from Cityscapes, where the left-hand side is corrupted by *shot noise* and the right-hand side by *defocus blur*. **b** Prediction of the best-performing model-variant on the corresponding clean image. **c** Prediction of the same architecture on the corrupted image (*a*). **d** Prediction of an ablated architecture on the corrupted image (*a*). We clearly see that prediction (*d*) is superior to (*c*), hence the corresponding model is more robust with respect to this image corruption, although it performs worse on the clean image. We present a study of various architectural choices and various image corruptions for three datasets: Cityscapes, PASCAL VOC 2012, and ADE20K

(2018) and Hendrycks and Dietterich (2019), which focused on aspect (i), we perform both options. We believe that an ablation study (option ii) is important since knowledge about architectural choices are likely helpful when designing a practical system, where types of image corruptions are known beforehand. For example, Geirhos et al. (2018) showed that ResNet-152 (He et al. 2016) is more robust to image noise than GoogLeNet (Szegedy et al. 2015). Is the latter architecture more prone to noise due to missing skip-connections, shallower architecture, or other architectural design choices? When the overarching goal is to develop robust convolutional neural networks, we believe that it is important to learn about the robustness capabilities of architectural properties.

We use the state-of-the-art DeepLabv3+ architecture (Chen et al. 2018b) with multiple network backbones as reference and consider many ablations of it. Based on our evaluation, we are able to conclude three main findings: (1) Many networks perform well with respect to real-world image corruptions, such as a realistic PSF blur. (2) Architectural properties can affect the robustness of a model significantly. Our results show that atrous (i.e., dilated) convolutions and long-range link naturally aid the robustness against many types of image corruptions. However, an architecture with a Dense Prediction Cell (Chen et al. 2018a), which was designed to maximize performance on clean data, hampers the performance for corrupted images significantly (see Fig. 1). (3) The generalization capability of DeepLabv3+ model, using a ResNet-backbone, depends strongly on the type of image corruption.

In summary, we give the following contributions:

- We benchmark the robustness of many architectural properties of the state-of-the-art semantic segmentation model DeepLabv3+ for a wide range of real-world image corruptions. We utilize almost 4,00,000 images generated from the Cityscapes dataset, PASCAL VOC 2012, and ADE20K.
- Besides DeepLabv3+, we have also benchmarked a wealth of other semantic segmentation models.
- We develop a more realistic noise model than previous approaches.
- Based on the benchmark study, we have several new insights: (1) Models are robust to real-world corruptions, such as a realistic PSF blur. (2) Some architecture properties affect robustness significantly. (3) Semantic

segmentation models generalize well to severe image noise and blur but struggle for other corruption types.
- We propose robust model design rules for semantic segmentation.

This article is an extended version of our recent publication (Kamann and Rother 2020). The additional content of this submission is:

- We provide a model generalization study, where we train models on corrupted data (Sect. 6, Figs. 18–20).
- We provide extensive evaluation for the degradations across severity levels for many image corruptions on each dataset for each ablated variant (Fig. 17)
- We provide more extensive robust model design rules (Sect. 5.7).
- We discuss possible causes of the effects of architectural properties in more detail (Sect. 5.3)
- We provide a detailed evaluation for CD/rCD of non-Deeplab based models, and for the ablation study on ADE20K and PASCAL VOC 2012 (Fig. 11, 15, 16 of Sects. 5.2, 5.4, 5.5).
- We provide much more details of utilized image corruptions in both text and visually (Figs. 2, 3, 5, 6).
- We provide qualitative results for the influence of image properties (Figs. 8, 9, 13, 14).

2 Related Work

Several recent work deals with the robustness towards real-world, common image corruptions. We discuss it in the following and separate the discussion into benchmarking and increasing robustness with respect to common corruptions, respectively.

Benchmarking model robustness w.r.t common image corruptions Michaelis et al. (2019) focus on the task of object detection. The authors benchmarked network robustness and found a significant performance drop when the input data is corrupted.

Dodge and Karam (2016) demonstrate the vulnerability of CNNs against image blur, noise, and contrast variations for image classification, and they demonstrate in Dodge and Karam (2017) further that humans perform better than classification CNNs for corrupted input data (similar to Zhou et al. 2017). Azulay and Weiss (2019) and Engstrom et al. (2019) demonstrate that variations in pixel space may change the CNN prediction significantly.

Vasiljevic et al. (2016) examined the impact of blur on full-image classification and semantic segmentation using VGG-16 (Simonyan and Zisserman 2015). Model performance decreases with an increased degree of blur for both tasks. We also focus in this work on semantic segmentation but evaluate on a much wider range of real-world image corruptions.

Geirhos et al. (2018) compared the generalization capabilities of humans and Deep Neural Networks (DNNs). The ImageNet dataset (Deng et al. 2009) is modified in terms of color variations, noise, blur, and rotation. Models that were trained directly on image noise did not perform well w.r.t other types of more severe noise.

Hendrycks and Dietterich (2019) introduce the "ImageNet-C dataset". The authors corrupted the ImageNet dataset by common image corruptions. Although the absolute performance scores increase from AlexNet (Krizhevsky et al. 2012) to ResNet (He et al. 2016), the relative robustness of the respective models does barely change. They further show that Multigrid and DenseNet architectures (Ke et al. 2017; Huang et al. 2017) are less prone to noise corruption than ResNet architectures. In this work, we use most of the proposed image transformations and apply them to the Cityscapes dataset, PASCAL VOC 2012, and ADE20K (Cordts et al. 2016; Everingham et al. 2010; Zhou et al. 2017, 2016). Recent work deals further with model robustness against night images (Dai and Van Gool 2018), weather conditions (Sakaridis et al. 2019, 2018; Volk et al. 2019), and spatial transformations (Fawzi and Frossard 2015; Ruderman et al. 2018).

Zendel et al. (2017) create a CV model for enabling to apply the hazard and operability analysis (HAZOP) to the computer vision domain and further provides an extensive checklist for image corruptions and visual hazards. This study demonstrates that most forms of image corruptions do also negatively influence stereo vision algorithms. Zendel et al. (2018) propose a fruitful segmentation test-dataset ("WildDash") containing challenging visual hazards such as overexposure, lens distortion, or occlusions.

Robustness of models with respect to adversarial examples is an active field of research (Huang et al. 2017; Boopathy et al. 2019; Cisse et al. 2017; Gu and Rigazio 2014; Carlini and Wagner 2017b; Metzen et al. 2017; Carlini and Wagner 2017a). Arnab et al. (2018) evaluate the robustness of semantic segmentation models for adversarial attacks of a wide variety of network architectures (e.g. Zhao et al. 2017; Badrinarayanan et al. 2017; Paszke et al. 2016; Zhao et al. 2018; Yu and Koltun 2016). In this work, we adopt a similar evaluation procedure, but we do not focus on the robustness with respect to adversarial attacks, which are typically not realistic, but rather on physically realistic image corruptions. We further rate robustness with respect to many architectural properties instead of solely comparing CNN architectures. Our approach modifies a single property per model at a time, which allows for an accurate evaluation.

Gilmer et al. (2019) connect adversarial robustness and robustness with respect to image corruption of Gaussian noise. The authors showed that training procedures that increase adversarial robustness also improve robustness with respect to many image corruptions. A parallel work of Rusak et al. (2020) shows, however, that adversarial training reduces corruption robustness.

Increasing model robustness w.r.t common image corruptions The majority of methods have been proposed for full-image classification. Geirhos et al. (2019) showed that humans and DNNs classify images with different strategies. Unlike humans, DNNs trained on ImageNet seem to rely more on local texture instead of global object shape. The authors then show that model robustness with respect to image corruptions increases when CNNs rely more on object shape than on object texture.

Hendrycks and Dietterich (2019) demonstrate that adversarial logit pairing (Kannan et al. 2018) enhances model robustness against adversarial perturbations and common perturbations. Xie et al. (2019) and Mahajan et al. (2018) increase model robustness through increasing the amount of training data. find a similar result for object detection when more complex network backbones are applied. In this work, we find a similar result for the task of semantic segmentation. Zhang (2019) increased the robustness against shifted input. Zheng et al. (2016) and Laermann et al. (2019) applied stability training to increase CNN robustness.

Data augmentation, such as inserting occlusions, cropping and combining images, is a popular technique to increase model robustness (Zhong et al. 2017; DeVries and Taylor 2017; Yun et al. 2019; Zhang et al. 2018; Takahashi et al. 2020). The authors of Gilmer et al. (2019), Lopes et al. (2019) and Rusak et al. (2020) augment the data with noise to reduce model vulnerability against common image corruptions. The work of Hendrycks e al. (2020) and Cubuk e al. (2019), on the other hand, distort the images with (learned) corruptions.

3 Image Corruption Models

We evaluate the robustness of semantic segmentation models towards a broad range of image corruptions. Besides using image corruptions from the ImageNet-C dataset, we propose new and more realistic image corruptions that can be treated as proxy covering the huge diversity of naturally occurring real-world image corruptions.

3.1 ImageNet-C

We employ many image corruptions from the ImageNet-C dataset (Hendrycks and Dietterich 2019). These consist of several types of *Blur:* motion, defocus, frosted glass and Gaussian; *Image Noise:* Gaussian, impulse, shot and speckle; *Weather:* snow, spatter, fog, and frost; and *Digital:* brightness, contrast, and JPEG compression (illustrated in Fig. 2).

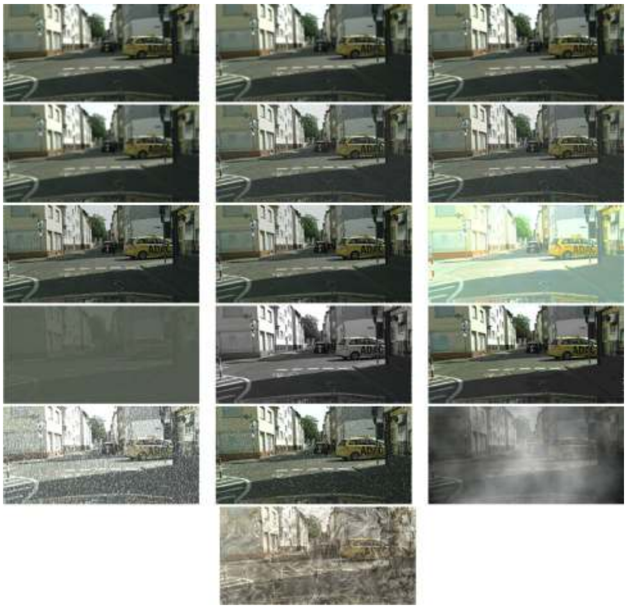


Fig. 2 Illustration of utilized image corruptions of ImageNet-C. First row (severity level 5 each): Motion blur, defocus blur, frosted glass blur. Second row (severity level 4 each): Gaussian blur, Gaussian noise, impulse noise. Third row (severity level 4, 4, 5 respectively): Shot noise, speckle noise, brightness. Fourth row (severity level 4, 2, 4 respectively): Contrast, saturate, JPEG. Fifth row (severity level 3, 3, 5 respectively): Snow, spatter, fog. Sixth row (severity level 5): frost

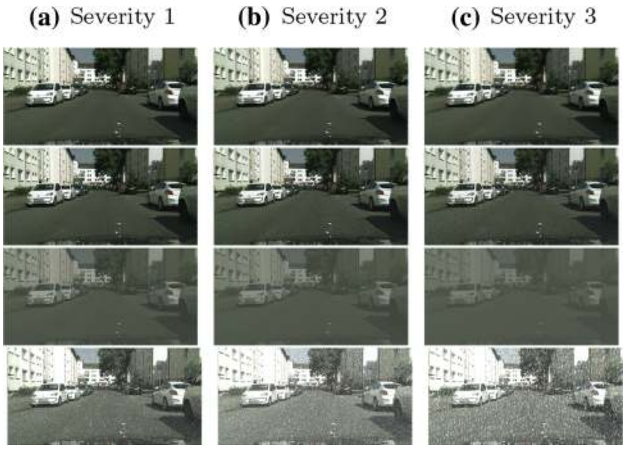


Fig. 3 Illustration of the first three severity levels of Cityscapes-C for a candidate of the categories blur, noise, digital, and weather. First row: Motion blur. Second row: Gaussian noise. Third row: Contrast. Fourth row: Snow

Each corruption is parameterized with five severity levels as illustrated for several candidates in Fig. 3.

3.2 Additional Image Corruptions

Intensity-Dependent Noise Model DCNNs are prone to image noise. Previous noise models are often simplistic, e.g., images are evenly distorted with Gaussian noise. However, *real*

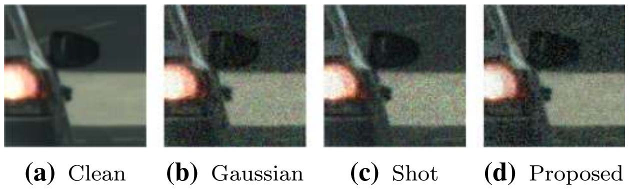

(a) Clean **(b)** Gaussian **(c)** Shot **(d)** Proposed

Fig. 4 A crop of a validation image from Cityscapes corrupted by various noise models. **a** Clean image. **b** Gaussian noise (severity level 1). **c** Shot noise (severity level 1). **d** Our proposed noise model (severity level 3). The amount of noise is high in regions with low pixel intensity

image noise significantly differs from the noise generated by these simple models. Real image noise is a combination of multiple types of noise [e.g., photon noise, kTC noise, dark current noise as described in Healey and Kondepudy (1994), Young et al. (1998), Lukas et al. (2006) and Liu et al. (2008)].

We propose a noise model that incorporates commonly observable behavior of cameras. Our noise model consists of two noise components: (i) a chrominance and luminance noise component, which are both added to original pixel intensities in linear color space. (ii) an intensity-level dependent behavior. Here, the term *chrominance noise* means that a random noise component for an image pixel is drawn for each color channel independently, resulting thus in color noise. *Luminance noise*, on the other hand, refers to a random noise value that is added to each channel of a pixel equally, resulting hence in gray-scale noise. In accordance with image noise observed from real-world cameras, pixels with low intensities are noisier than pixels with high intensities. Shot noise is the dominant noise for dark scenes since the Poisson distribution's mean is not constant but equal to the root of counted photons (Jahne 1997). Since the noise is added in linear color space, that relative amount of noise decreases with increasing intensity in sRGB color gamut. We model the noisy pixel intensity for a color channel c as a random variable $I_{noise,c}$:

$$\begin{aligned} &I_{noise,c}(\Phi_c, N_{luminance}, N_{chrominance,c}; w_s) \\ &\quad = log_2(2^{\Phi_c} + w_s \cdot (N_{luminance} + N_{chrominance,c})) \end{aligned} \tag{1}$$

where Φ_c is the normalized pixel intensity of color channel c, $N_{luminance}$ and $N_{chrominance}$ are random variables following a Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, w_s is a weight factor, parameterized by severity level s.

Figure 4 illustrates noisy variants of a Cityscapes image-crop. In contrast to the other, simpler noise models, the amount of noise generated by our noise model depends clearly on pixel intensity.

PSF blur Every optical system of a camera exhibits aberrations, which mostly result in image blur. A point-spread-function (PSF) aggregates all optical aberrations that result in image blur (Joshi et al. 2008). We denote this type of corruption as *PSF blur*. Unlike simple blur models, such as Gaussian blur, real-world PSF functions are spatially varying. We corrupt the Cityscapes dataset with three different PSF functions that we have generated with the optical design program *Zemax*, for which the amount of blur increases with a larger distance to the image center. Our PSF models correspond to a customary front video automotive video camera with a horizontal field-of-view of 90 degrees. We illustrate the intensity distribution of several PSF kernels for different angles of incidence in Fig. 5.

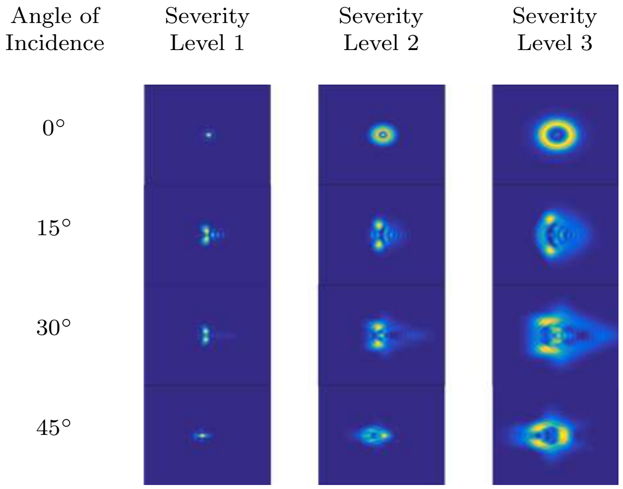


Fig. 5 The intensity distribution of used PSF kernels. The degree of the spatial distribution of intensity increases with the severity level. The shape of the PSF kernel depends on the image region, i.e., the angle of incidence

Geometric distortion Every camera lens exhibits geometric distortions (Fitzgibbon 2001). Distortion parameters of an optical system vary over time, are affected by environmental influences, differ from calibration stages, and thus, may never be fully compensated. Additionally, image warping may introduce re-sampling artifacts, degrading the informational content of an image. It can hence be preferable to utilize the original (i.e., geometrically distorted) image (Hartley and Zisserman 2003, p. 192f). We applied several radially-symmetric barrel distortions (Willson 1994) as a polynomial of grade 4 (Shah and Aggarwal 1996) to both the RGB-image and respective ground truth.

Figure 6 shows examples of our proposed common image corruptions.

4 Models

We employ DeepLabv3+ (Chen et al. 2018b) as the reference architecture. We chose DeepLabv3+ for several reasons. It supports numerous network backbones, ranging from novel state-of-art models [e.g., modified aligned Xception (Chollet

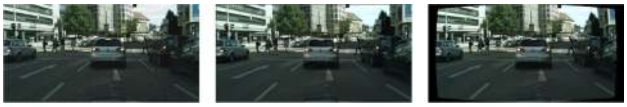


Fig. 6 Illustration of our proposed image corruptions. From left to right: Proposed noise model (severity level 4), PSF blur (severity level 3), and geometric distortion (severity level 1). Best viewed in color (Color figure online)

2017; Chen et al. 2018b], denoted by *Xception*) and established ones [e.g., ResNets (He et al. 2016)]. DeepLabv3+ exhibits architectural properties, which are established in the task of semantic segmentation. For semantic segmentation, DeepLabv3+ utilizes popular architectural properties, making it a highly suitable candidate for an ablation study. Please note that the range of network backbones, offered by DeepLabv3+, represents different execution times since different applications have different demands.

Besides DeepLabv3+, we have also benchmarked a wealth of other semantic segmentation models, such as FCN8s (Long et al. 2015), VGG-16 (Simonyan and Zisserman 2015), ICNet (Zhao et al. 2018), DilatedNet (Yu and Koltun 2016), ResNet-38 (Wu et al. 2019), PSPNet (Zhao et al. 2017), and the recent Gated-ShapeCNN (GSCNN) (Takikawa et al. 2019). In the following, we summarize properties of DeepLabv3+.

4.1 DeepLabv3+

Figure 7 illustrates important elements of the DeepLabv3+ architecture. A network backbone (ResNet, Xception or MobileNet-V2) processes an input image (He et al. 2016; Sandler et al. 2018; Howard et al. 2017). Its output is subsequently processed by a multi-scale processing module, extracting dense feature maps. This module is either Dense Prediction Cell (Chen et al. 2018a) (DPC) or Atrous Spatial Pyramid Pooling (ASPP, with or without global average pooling (GAP)). We consider the variant with ASPP and without GAP as reference architecture. A long-range link concatenates early features from the network backbone with features extracted by the respective multi-scale processing module. Finally, the decoder outputs estimates of the semantic labels.

Atrous convolution Atrous (i.e., dilated) convolution (Chen et al. 2017; Holschneider et al. 1989; Papandreou et al. 2005) is a type of convolution that integrates spacing between kernel parameters and thus increases the kernel field of view. DeepLabv3+ incorporates atrous convolutions in the network backbone.

Atrous spatial pyramid pooling To extract features at different scales, several semantic segmentation architectures (Chen et al. 2017, 2015; Zhao et al. 2017) perform Spatial Pyramid Pooling (He et al. 2015; Grauman and Darrell 2005; Lazebnik et al. 2006). DeepLabv3+ applies *Atrous spatial pyramid pooling* (ASPP), where three atrous convolutions with large atrous rates (6, 12 and 18) process the DCNN output.

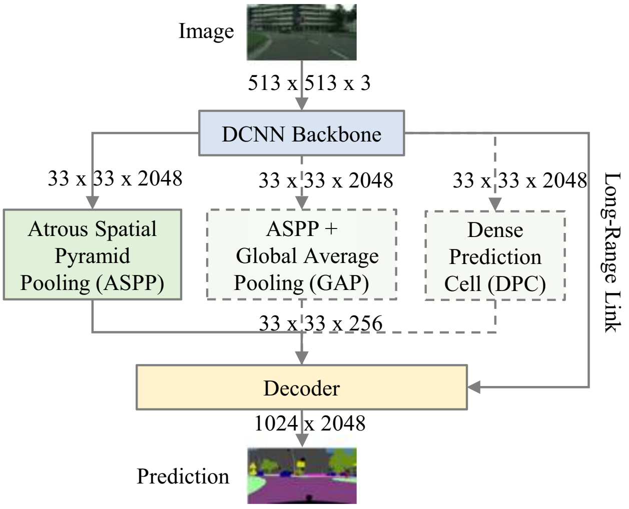


Fig. 7 Building blocks of DeepLabv3+. Input images are firstly processed by a network backbone, containing atrous convolutions. The backbone output is further processed by a multi-scale processing module (ASPP or DPC). A long-range link concatenates early features of the network backbone with encoder output. Finally, the decoder outputs estimates of semantic labels. Our reference model is shown by regular arrows (i.e., without DPC and GAP). The dimension of activation volumes is shown after each block

Dense prediction cell Chen et al. (2018a) is an efficient multi-scale architecture for dense image prediction, constituting an alternative to ASPP. It is the result of a neural-architecture-search with the objective to maximize the performance for clean images. In this work, we analyze whether this objective leads to overfitting.

Long-range link A long-range link concatenates early features of the encoder with features extracted by the respective multi-scale processing module (Hariharan et al. 2015). In more detail, for Xception (MobileNet-V2) based models, the long-range link connects the output of the second or the third Xception block (inverted residual block) with ASPP or DPC output. Regarding ResNet architectures, the long-range link connects the output of the second residual block with the ASPP or DPC output.

Global average pooling A global average pooling (GAP) layer (Lin et al. 2014) averages the feature maps of an activation volume. DeepLabv3+ incorporates GAP in parallel to the ASPP.

4.2 Architectural Ablations

In the next section, we evaluate various ablations of the DeepLabv3+ reference architecture. In detail, we remove atrous convolutions (AC) from the network backbone by transforming them into regular convolutions. We denote this ablation in the remaining sections as w\ o AC. We further removed the long-range link (LRL, i.e., w\ o LRL) and Atrous Spatial Pyramid Pooling (ASPP) module (w\ o ASPP). The removal of ASPP is additionally replaced by Dense Prediction Cell (DPC) and denoted as w\ o ASPP+w\ DPC. We also examined the effect of global average pooling (w\ GAP).

5 Experiments

We present the experimental setup and report results of benchmarking numerous network backbones, the effect of architectural properties on robustness towards common image corruptions and the generalization behavior of semantic segmentation models.

We firstly benchmark multiple neural network backbone architectures of DeepLabv3+ and many other semantic segmentation models (Sect. 5.2). While this procedure gives an overview of the robustness across several architectures, no conclusions about which architectural properties affect the robustness can be drawn. Hence, we modify multiple architectural properties of DeepLabv3+ (as described in Sect. 4.2) and evaluate the robustness for the re-trained ablated models with respect to image corruptions (Sects. 5.3–5.5). Our findings show that architectural properties can have a substantial impact on the robustness of a semantic segmentation model with respect to image corruptions. We derive robust model design rules in Sect. 5.7.

Finally, instead of training a model on clean data only, we add corrupted data to the training set. We demonstrate the generalization capability for severe image noise and show that DeepLabv3+ generalizes considerably well to various types of image noise (Sect. 6).

5.1 Experimental Setup

Network backbones We trained DeepLabv3+ with several network backbones on clean and corrupted data using TensorFlow (Abadi et al. 2016). We utilized MobileNet-V2, ResNet-50, ResNet-101, Xception-41, Xception-65 and Xception-71 as network backbones. Every model has been trained with batch size 16, crop-size 513 × 513, fine-tuning batch normalization parameters (Ioffe and Szegedy 2015), initial learning rate 0.01 or 0.007, and random scale data augmentation.

Datasets We use PASCAL VOC 2012, the Cityscapes dataset, and ADE20K for training and validation. The training set of PASCAL VOC consists of 1,464 train and 1,449 validation images. We use the high-quality pixel-level annotations of Cityscapes, comprising of 2975 train and 500 validation images. We evaluated all models on original image dimensions.

ADE20K consists of 20,210 train, 2,000 validation images, and 150 semantic classes.

Evaluation metrics We apply mean Intersection-over-Union as performance metric (mIoU) for every model and average over severity levels. In addition, we use, and slightly modify, the concept of Corruption Error and relative Corruption Error from Hendrycks and Dietterich (2019) as follows.

We use the term *Degradation* D, where $D = 1 - mIoU$ in place of *Error*. Degradations across severity levels, which are defined by the ImageNet-C corruptions (Hendrycks and Dietterich 2019), are often aggregated. To make models mutually comparable, we divide the degradation D of a trained model f through the degradation of a reference model ref. With this, the *Corruption Degradation* (CD) of a trained model is defined as

$$CD_c^f = \left(\sum_{s=1}^{5} D_{s,c}^f \right) \Big/ \left(\sum_{s=1}^{5} D_{s,c}^{ref} \right) \qquad (2)$$

where c denotes the corruption type (e.g., Gaussian blur) and s its severity level. Please note that for *category noise*, only the first three severity levels are taken into account.

For comparing the robustness of model architectures, we also consider the degradation of models relative to clean data, measured by the *relative Corruption Degradation* (rCD).

$$rCD_c^f = \left(\sum_{s=1}^{5} D_{s,c}^f - D_{clean}^f \right) \Big/ \left(\sum_{s=1}^{5} D_{s,c}^{ref} - D_{clean}^{ref} \right) \qquad (3)$$

We predominantly use the Corruption Degradation (CD) to rate model robustness with respect to image corruptions, since the CD rates model robustness in terms of absolute performance. The relative Corruption Degradation (rCD), on the other hand, incorporates the respective model performance on clean data. The degradation on clean data is for both models (i.e., the model for which the robustness is to be rated, and the reference model) subtracted, resulting hence in a measure that gives a ratio of the absolute performance decrease in the presence of image corruption.

5.2 Benchmarking Network Backbones

We trained various network backbones (MobileNet-V2, ResNets, Xceptions) on the original, clean training-sets of PASCAL VOC 2012, the Cityscapes dataset, and ADE20K. Table 1 shows the average mIoU for the Cityscapes dataset, and each corruption type averaged over all severity levels.

Table 1 Average mIoU for clean and corrupted variants of the Cityscapes validation set for several network backbones of the DeepLabv3+ architecture (*top*) and non-DeepLab based models (*bottom*)

Architecture	Clean	Blur					Noise				
		Motion	Defocus	Frosted glass	Gaussian	PSF	Gaussian	Impulse	Shot	Speckle	Intensity
MobileNet-V2	72.0	53.5	49.0	45.3	49.1	70.5	6.4	7.0	6.6	16.6	26.9
ResNet-50	76.6	58.5	56.6	47.2	57.7	74.8	6.5	7.2	10.0	31.1	30.9
ResNet-101	77.1	59.1	56.3	47.7	57.3	75.2	13.2	13.9	16.3	36.9	39.9
Xception-41	77.8	61.6	54.9	51.0	54.7	76.1	**17.0**	**17.3**	**21.6**	**43.7**	48.6
Xception-65	78.4	63.9	59.1	**52.8**	59.2	**76.8**	15.0	10.6	19.8	42.4	46.5
Xception-71	**78.6**	**64.1**	**60.9**	52.0	**60.4**	76.4	14.9	10.8	19.4	41.2	**50.2**
ICNet	65.9	45.8	44.6	**47.4**	44.7	65.2	8.4	8.4	10.6	27.9	29.7
FCN8s-VGG16	66.7	42.7	31.1	37.0	34.1	61.4	6.7	5.7	7.8	24.9	18.8
DilatedNet	68.6	44.4	36.3	32.5	38.4	61.1	**15.6**	14.0	**18.4**	32.7	35.4
ResNet-38	77.5	54.6	45.1	43.3	47.2	74.9	13.7	**16.0**	18.2	**38.3**	**35.9**
PSPNet	78.8	**59.8**	53.2	44.4	53.9	76.9	11.0	15.4	15.4	34.2	32.4
GSCNN	**80.9**	58.9	**58.4**	41.9	**60.1**	**80.3**	5.5	2.6	6.8	24.7	29.7

Architecture	Digital				Weather				
	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric Distortion
MobileNet-V2	51.7	46.7	32.4	27.2	13.7	38.9	47.4	17.3	65.5
ResNet-50	58.2	54.7	41.3	27.4	12.0	42.0	55.9	22.8	69.5
ResNet-101	59.2	54.5	41.5	37.4	11.9	47.8	55.1	22.7	69.7
Xception-41	63.6	56.9	**51.7**	38.5	18.2	46.6	57.6	20.6	**73.0**
Xception-65	65.9	**59.1**	46.1	31.4	**19.3**	**50.7**	63.6	**23.8**	72.7
Xception-71	**68.0**	58.7	47.1	**40.2**	18.8	50.4	**64.1**	20.2	71.0
ICNet	41.0	33.1	27.5	**34.0**	6.3	30.5	27.3	11.0	56.5
FCN8s-VGG16	53.3	39.0	36.0	21.2	11.3	31.6	37.6	19.7	62.5
DilatedNet	52.7	32.6	38.1	29.1	12.5	32.3	34.7	19.2	63.3
ResNet-38	60.0	50.6	46.9	14.7	**13.5**	45.9	52.9	22.2	73.2
PSPNet	60.4	51.8	30.6	21.4	8.4	42.7	34.4	16.2	73.6
GSCNN	**75.9**	**61.9**	**70.7**	12.0	12.4	**47.3**	**67.9**	**32.6**	**76.4**

Every mIoU is averaged over all available severity levels, except for corruptions of category noise where only the first three (of five) severity levels are considered. Xception based network backbones are usually most robust against each corruption. Most models are robust against our realistic PSF blur. Highest mIoU per corruption is bold

As expected, for DeepLabv3+, Xception-71 exhibits the best performance for clean data with an mIoU of 78.6%.[1] The bottom part of Table 1 shows the benchmark results of non-DeepLab based models.

Network backbone performance Most Xception based models perform significantly better than ResNets and MobileNet-V2. GSCNN is the best performing architecture on clean data of this benchmark.

Performance w.r.t blur Interestingly, all models (except DilatedNet and VGG16) handle PSF blur well, as the respective mIoU decreases only by roughly 2%. Thus, even a lightweight network backbone such as MobileNet-V2 is hardly vulnerable against this realistic type of blur. The number of both false positive and false negative pixel-level classifications increases, especially for far-distant objects. With respect to Cityscapes this means that persons are simply overlooked or confused with similar classes, such as rider (see Fig. 8).

Performance w.r.t noise Noise has a substantial impact on model performance (see Fig. 9). Hence we only averaged over the first three severity levels. Xception-based network backbones of DeepLabv3+ often perform similar or better than non-DeepLabv3+ models. MobileNet-V2, ICNet, VGG-16, and GSCNN handle the severe impact of image noise significantly worse than the other models.

Performance w.r.t digital The first severity levels of corruption types contrast, brightness, and saturation are handled

[1] Note that we were not able to reproduce the results from Chen et al. (2018b). We conjecture that this is due to hardware limitations, as we could not set the suggested crop-size of 769 × 769 for Cityscapes.

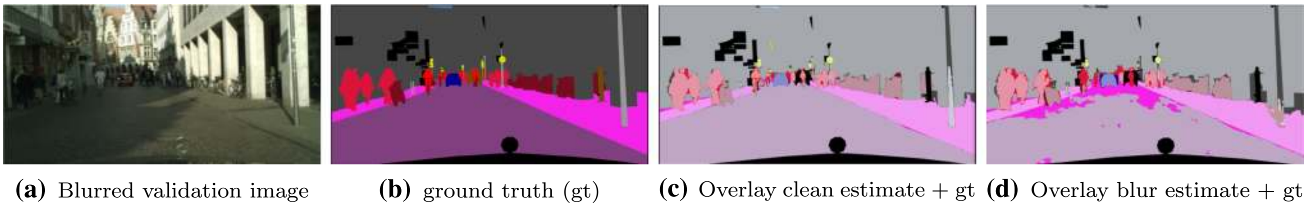

(a) Blurred validation image
(b) ground truth (gt)
(c) Overlay clean estimate + gt
(d) Overlay blur estimate + gt

Fig. 8 Prediction of the reference architecture (i.e., original DeepLabv3+) on blurred input, using Xception-71 as network backbone. **a** A blurred validation image (Gaussian blur, severity level 3) of the Cityscapes dataset and corresponding ground truth (**b**). **c** Prediction on the clean image overlaid with the ground truth. True-positives are alpha-blended, false-positives and false-negatives remain unchanged. Hence, wrongly classified pixels can be easier spotted. **d** Prediction on the blurred image overlaid with the ground truth (*b*). Whereas the *riders* are mostly correctly classified in (*c*), they are in (*d*) miss-classified as *person*. Extensive areas of *road* are miss-classified as *sidewalk*

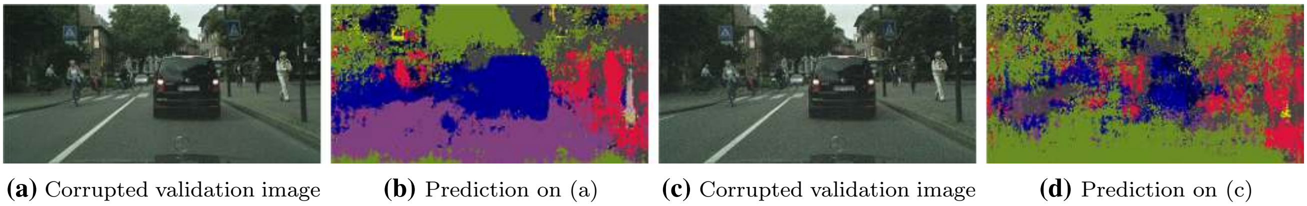

(a) Corrupted validation image
(b) Prediction on (a)
(c) Corrupted validation image
(d) Prediction on (c)

Fig. 9 Drastic influence of image noise on model performance. **a** A validation image of Cityscapes is corrupted by the second severity level of Gaussian noise and respective prediction (**b**). **c** A validation image of Cityscapes is corrupted by the third severity level of Gaussian Noise and respective prediction (**d**). Predictions are produced by the reference model, using Xception-71 as the backbone

well. However, JPEG compression decreases performance by a large margin. Notably, PSPNet and GSCNN have for this corruption halved or less mIoU than Xception-41 and -71, though their mIoU on clean data is similar.

Performance w.r.t weather Texture-corrupting distortions as snow and frost degrade mIoU of each model significantly.

Performance w.r.t geometric distortion All models perform similarly with respect to geometric distortion. The GSCNN is the most robust model against this image corruption. Whereas most models withstand the first severity level (illustrated in Fig. 6) well, the mIoU of GSCNN drops only by less than 1%.

This benchmark indicates, in general, a similar result as in Geirhos et al. (2019), that is image distortions corrupting the texture of an image (e.g., image noise, snow, frost, JPEG), often have a distinctly negative effect on model performance compared to image corruptions preserving texture to a certain point (e.g., blur, brightness, contrast, geometric distortion).

To evaluate the robustness w.r.t image corruptions of proposed network backbones, it is also interesting to consider Corruption Degradation (CD) and relative Corruption Degradation (rCD). Figure 10 illustrates the mean CD and rCD with respect to the mIoU for *clean* images (lower values correspond to higher robustness than the reference model). Each dot depicts the performance of one network backbone, averaged over all corruptions except for PSF blur.[2]

Discussion of CD Subplot a−c illustrates respective results for PASCAL VOC 2012, Cityscapes, and ADE20K, and subplot d illustrates the results for the non-DeepLab-based networks evaluated on Cityscapes. On each of the three datasets, the CD for Xception-71 is the lowest for DeepLabv3+ architecture, which decreases, in general, with increasing mIoU on clean data.

A similar trend can be observed for the non-DeepLab models, except for PSPNet and FCN8s (VGG16). The Gated-Shape-CNN (GSCNN) is among them clearly the overall most robust architecture. The CD scores for models evaluated on Cityscapes (subplot b and d) are in a similar range, even though the utilized reference models are different architectures (but the respective mIoU on clean data is similar).

Discussion of rCD The rCD, on the other hand, behaves contrary between subplot a–c (where it usually decreases such as CD, except for ResNets on ADE20K and Xception-65 on PASCAL VOC 2012) and subplot d. The authors of Hendrycks and Dietterich (2019) report the same result for the task of full-image classification: The rCD for established networks stays relatively constant, even though model performance on clean data differs significantly, as Fig. 10d indicate. When we, however, evaluate within a semantic seg-

[2] Due to the considerably smaller impact of PSF blur on model performance, small changes in mIoU of only tenths percentage can have a significant impact on the corresponding rCD.

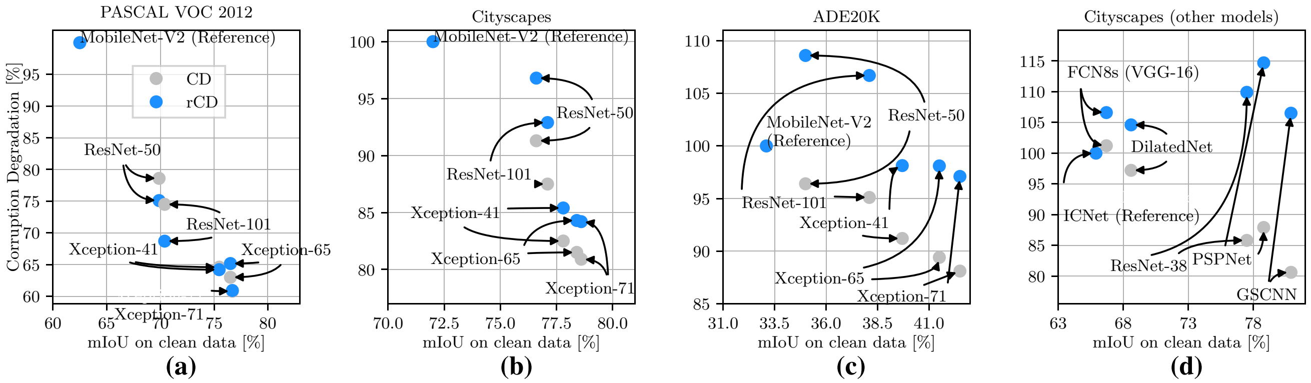


Fig. 10 **a–c** CD and rCD for several network backbones of the DeepLabv3+ architecture evaluated on PASCAL VOC 2012, the Cityscapes dataset, and ADE20K. MobileNet-V2 is the reference model in each case. rCD and CD values below 100% represent higher robustness than the reference model. In almost every case, model robustness increases with model performance (i.e., mIoU on clean data). Xception-71 is the most robust network backbone on each dataset. **d** CD and rCD for non-DeepLabv3+ based models evaluated on Cityscapes. While CD decreases with increasing performance on clean data, rCD is larger than 100%

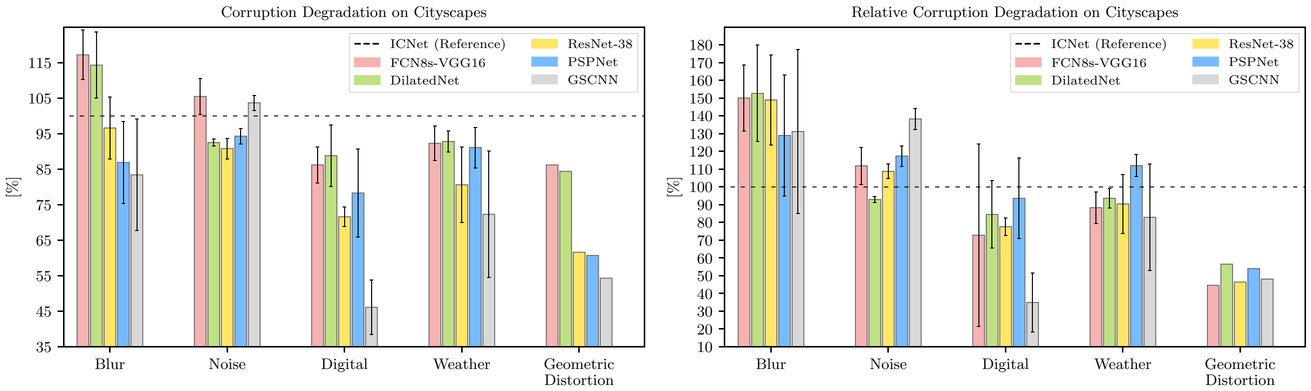


Fig. 11 CD (left) and rCD (right) evaluated on Cityscapes for ICNet (set as reference architecture), FCN8s-VGG16, DilatedNet, ResNet-38, PSPNet, GSCNN w.r.t. image corruptions of category blur, noise, digital, weather, and geometric distortion. Each bar except for geometric distortion is averaged within a corruption category (error bars indicate the standard deviation). The CD of image corruption “jpeg compression” of category digital is not included in this barplot, since, contrary to the remaining image corruptions of that category, the respective CDs range between 107 and 133%. Bars above 100% represent a decrease in performance compared to the reference architecture. Best viewed in color (Color figure online)

mentation architecture, as DeepLabv3+, the contrary result (i.e., decreasing rCD) is generally observed, similar to Orhan (2019) and Michaelis et al. (2019) for other computer vision tasks.

The following speculation may give further insights. Geirhos et al. (2019) stated that (i) DCNNs for full-image classification examine local textures, rather than global shapes of an object, to solve the task at hand, and (ii) model performance w.r.t image corruption increases when the model relies more on object shape (rather than object texture).

Transferring these results to the task of semantic segmentation, Xception-based backbones, and the GSCNN might have a more pronounced shape bias than others (e.g., ResNets), resulting hence in a higher rCD score image corruption.

Figure 11 illustrates the CD and rCD averaged for the proposed image corruption categories for non-DeepLabv3+ based models. Please note that the CD of image corruption “jpeg compression” of category digital is not included in this barplot. CD (left) and rCD (right) evaluated on Cityscapes for ICNet (set as reference architecture), FCN8s-VGG16, DilatedNet, ResNet-38, PSPNet, GSCNN w.r.t. image corruptions of category blur, noise, digital, weather, and geometric distortion. Each bar except for geometric distortion is averaged within a corruption category (error bars indicate the standard deviation).

FCN8s-VGG16 and DilatedNet are vulnerable to corruptions of category blur. DilatedNet is more robust against

Table 2 Average mIoU for clean and corrupted variants of the Cityscapes validation dataset for Xception-71 and five corresponding architectural ablations

		Blur					Noise				
Deeplab-v3+ backbone	Clean	Motion	Defocus	Frosted glass	Gaussian	PSF	Gaussian	Impulse	Shot	Speckle	Intensity
Xception-71	78.6	64.1	60.9	52.0	60.4	76.4	14.9	10.8	19.4	41.2	**50.2**
w/o ASPP	73.9	60.7	59.5	51.5	58.4	72.8	**18.5**	**14.7**	**22.3**	39.8	44.7
w/o AC	77.9	62.2	57.9	51.8	58.2	76.1	7.7	5.7	11.2	32.8	43.2
w/o ASPP+w/ DPC	**78.8**	62.8	59.4	52.6	58.2	76.9	7.3	2.8	10.7	33.0	42.4
w/o LRL	77.9	64.2	**63.2**	50.7	**62.2**	76.7	13.9	9.3	18.2	**41.3**	49.9
w/ GAP	78.6	**64.2**	61.7	**55.9**	60.7	**77.8**	9.7	8.4	13.9	36.9	45.6

	Digital				Weather				
Deeplab-v3+ backbone	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric distortion
Xception-71	**68.0**	58.7	47.1	40.2	**18.8**	50.4	64.1	20.2	71.0
w/o ASPP	63.4	56.2	42.7	39.9	17.6	49.0	58.3	21.8	69.3
w/o AC	67.6	55.6	46.0	**40.7**	18.2	50.1	61.1	**21.6**	71.1
w/o ASPP+w/ DPC	64.8	59.4	45.3	32.0	14.4	48.6	64.0	20.8	72.1
w/o LRL	64.5	59.2	44.3	36.1	16.9	48.7	**64.3**	21.3	71.3
w/ GAP	68.0	**60.2**	**48.4**	40.6	16.8	**51.0**	62.1	20.9	**73.6**

Based on DeepLabv3+ we evaluate the removal of atrous spatial pyramid pooling (**ASPP**), atrous convolutions (**AC**), and long-range link (**LRL**). We further replaced ASPP by Dense Prediction Cell (**DPC**) and utilized global average pooling (**GAP**). Mean-IoU is averaged over severity levels. The standard deviation for image noise is 0.2 or less. Highest mIoU per corruption is bold

corruptions of category noise, digital, and weather than the reference. ResNet-38 is robust against corruptions of category weather. The rCD of PSPNet is oftentimes higher than 100% for each category. GSCNN is vulnerable to image noise. The rCD is considerably high, indicating a high decrease of mIoU in the presence of this corruption. The low scores for geometric distortion show that the reference model is vulnerable to this corruption. GSCNN is the most robust model of this benchmark with respect to geometric distortion, and overall mostly robust except for image noise.

5.3 Ablation Study on Cityscapes

Instead of solely comparing robustness across network backbones, we now conduct an extensive ablation study for DeepLabv3+. We employ the state-of-the-art performing Xception-71 (XC-71), Xception-65 (XC-65), Xception-41 (XC-41), ResNet-101, ResNet-50 and, their lightweight counterpart, MobileNet-V2 (MN-V2) (width multiplier 1, 224 × 224), as network backbones. XC-71 is the best performing backbone on clean data, but at the same time, computationally most expensive. The efficient MN-V2, on the other hand, requires roughly ten times less storage space. We ablated for each network backbone of the DeepLabv3+ architecture the same architectural properties as listed in Sect. 4.2. Each ablated variant has been re-trained on clean data of Cityscapes, PASCAL VOC 2012, and ADE20K, summing up to over 100 trainings. Table 2 shows the averaged mIoU for XC-71, evaluated on Cityscapes. In the following sections, we discuss the most distinct, statistically significant results.

We see that with Dense Prediction Cell (DPC), we achieve the highest mIoU on clean data followed by the reference model. We also see that removing ASPP reduces mIoU significantly.

To better understand the robustness of each ablated model, we illustrate the average CD within corruption categories (e.g., blur) in Fig. 12 (bars above 100% indicate reduced robustness compared to the respective reference model).

Effect of ASPP Removal of ASPP reduces model performance significantly (Table 2 first column).

Effect of AC Atrous convolutions (AC) generally show a positive effect w.r.t corruptions of type blur for most network backbones, especially for XC-71 and ResNets (see Fig. 13). For example, without AC, the average mIoU for defocus blur decreases by 3.8% for ResNet-101 (CD=109%). Blur reduces high-frequency information of an image, leading to similar signals stored in consecutive pixels. Applying AC can hence increase the amount of information per convolution filter, by skipping direct neighbors with similar signals. Regarding XC-71 and ResNets, AC clearly enhance robustness on noise-based corruptions (see Fig. 14). The mIoU for the first severity level of Gaussian noise are 12.2% (XC-71), 10.8% (ResNet-101), 8.0% (ResNet-50) less than respective reference. In summary, AC often increase robustness against a broad range of network backbones and image corruptions.

Effect of DPC When employing Dense Prediction Cell (DPC) instead of ASPP, the models become clearly vulnera-

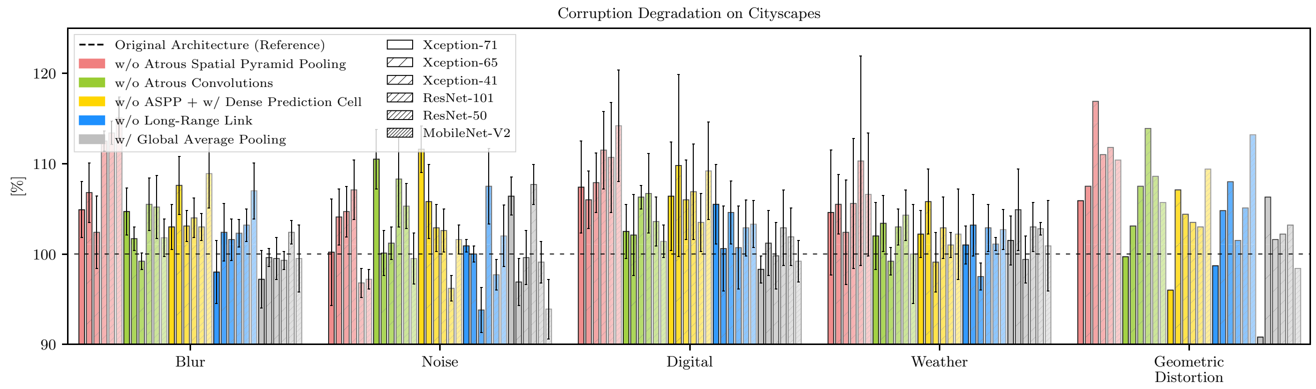
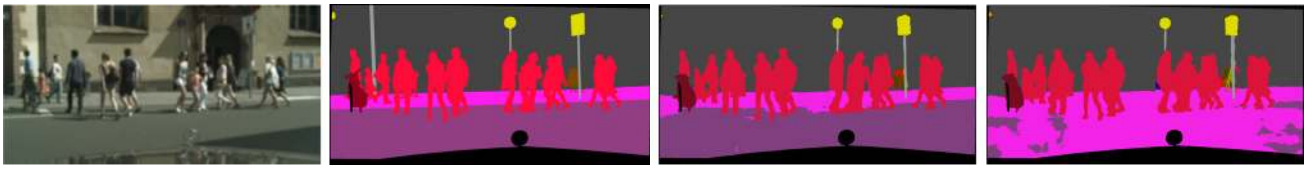
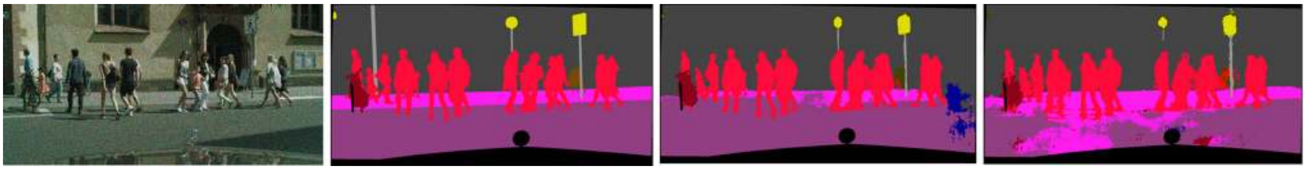


Fig. 12 CD evaluated on Cityscapes for the proposed ablated variants of the DeepLabv3+ architecture w.r.t image corruptions, employing six different network backbones. Bars above 100% represent a decrease in performance compared to the respective reference architecture. Each ablated architecture is re-trained on the original training dataset. Removing ASPP reduces the model performance significantly. Atrous convolutions increase robustness against blur. The model becomes vulnerable against most effects when Dense Prediction Cell is used. Each bar is the average CD of a corruption category, except for geometric distortion (error bars indicate the standard deviation)



(a) corrupted image **(b)** ground truth **(c)** prediction of ref. model **(d)** prediction w/o AC

Fig. 13 Predictions of reference architecture and the ablated variant without atrous convolutions (AC), which is especially vulnerable to blur. Validation image is corrupted by defocus blur (severity level 2)



(a) corrupted image **(b)** ground truth **(c)** prediction of ref. model **(d)** prediction w/o AC

Fig. 14 Predictions of reference architecture and the ablated variant without atrous convolutions (AC), which is especially vulnerable to noise. Validation image is corrupted by shot noise (severity level 1)

ble against corruptions of most categories. While this ablated architecture, (i.e., w\ DPC) reaches the highest mIoU on clean data for XC-71, it is less robust to a broad range of corruptions. For example, CD for defocus blur on MN-V2 and XC-65 are 113 and 110%, respectively. Average mIoU decreases by 6.8 and by 4.1%. For XC-71, CD for all corruptions of category noise are within 109 and 115%. The average mIoU of this ablated variant is least for all, but one type of noise (Table 2). Similar behavior can be observed for other corruptions and backbones.

DPC has been found through a neural-architecture-search (NAS, e.g. Zoph et al. 2018; Zoph and Le 2017; Pham et al. 2018) with the objective of maximizing performance on clean data. This result indicates that such architectures tend to overfit on this objective, i.e., clean data. It may be an interesting topic to evaluate robustness w.r.t image corruptions for other NAS-based architectures as future work, however, is beyond the scope of this paper. Consequently, performing NAS on corrupted data might deliver interesting findings of robust architectural properties–similar as in Cubuk et al. (2019) w.r.t adversarial examples.

We further hypothesize that DPC might learn less multi-scale representations than ASPP, which may be useful for common image corruptions (e.g., Geirhos et al. 2019 shows that classification models are more robust to common corruption if the shape bias of a model is increased). Whereas ASPP processes its input in parallel by three atrous convolution (AC) layers with large symmetric rates (6, 12, 18), DPC firstly processes the input by a single AC layer with small rate (1 × 6) (Chen et al. 2018a, Fig. 5). When we test DPC on corrupted data, it cannot hence apply the same beneficial multi-scale cues (due to the comparable small atrous convolution with rate 1×6) as ASPP and may, therefore, perform worse.

Effect of LRL A long-range link (LRL) appears to be very beneficial for ResNet-101 against image noise. The model struggles especially for our noise model, as its CD equals 116%. For XC-71, corruptions of category digital as *brightness* have considerably high CDs (e.g., $\mathrm{CD}^{\text{XC-71}} = 111\%$). For MN-V2, removing LRL decreases robustness w.r.t defocus blur and geometric distortion as average mIoU reduces by 5.1% (CD $= 110\%$) and 4.6% (CD $= 113\%$).

Effect of GAP Global average pooling (GAP) increases robustness w.r.t blur slightly for most Xceptions. Interestingly, when applied in XC-71 (ResNet-101), the model is vulnerable to image noise. Corresponding CD values range between 103 and 109% (106 and 112%). ResNet-101 shows similar behavior.

5.4 Ablation Study on Pascal VOC 2012

We generally observe that the effect of the architectural ablations for DeepLabv3+ trained on PASCAL VOC 2012 is not always similar to previous results on Cityscapes (see Fig. 15). Since this dataset is less complex than Cityscapes, the mIoU of ablated architectures differ less.

We do not evaluate results on MN-V2, as the model is not capable of giving a comparable performance. Table 3 contains the mIoU of each network backbone for clean and corrupted data.

Effect of ASPP Similar to the results on Cityscapes, removal of ASPP reduces model performance of each network backbone significantly.

Effect of AC Unlike on Cityscapes, atrous convolutions show no positive effect against blur. We explain this with the fundamentally different datasets. On Cityscapes, a model without AC often overlooks classes covering small image-regions, especially when far away. Such images are hardly present in PASCAL VOC 2012. As on Cityscapes, AC slightly helps performance for most models with respect to geometric distortion. For XC-41 and ResNet-101, we see a positive effect of AC against image noise.

Effect of DPC As on Cityscapes, DPC decreases robustness for many corruptions. Generally, CD increases from XC-41 to XC-71. The impact on XC-71 is especially strong as indicated by the CD score, averaged over all corruptions, is 106%. A possible explanation might be that the neural-architecture-search (NAS) e.g.

Zoph et al. (2018), Zoph and Le (2017) and Pham et al. (2018) has been performed on XC-71 and enhances, therefore, the overfitting effect additionally, as discussed in Sect. 5.3.

Effect of LRL Removing LRL increases robustness against noise for XC-71 and XC-41, probably due to discarding early features. Removing the Long-Range Link (LRL) discards early representations. The degree of, e.g., image noise is more pronounced on early CNN levels. Removing LRL tends hence to increase the robustness for a more shallow backbone as Xception-41 on PASCAL VOC 2012 and Cityscapes, as less corrupted features are linked from encoder to decoder. For a deeper backbone like ResNet-101, this behavior cannot be observed. However, this finding does not hold for XC-65. As reported in Sect. 5.2, on PASCAL VOC 2012, XC-65 is also the most robust model against noise.

Effect of GAP When global average pooling is applied, the overall robustness of every network backbone increases, particularly significant. The mIoU on clean data increases for every model (up to 2.2% for ResNet-101, probably due to the difference between PASCAL VOC 2012 and the remaining dataset. Global average pooling (GAP) increases performance on clean data on PASCAL VOC 2012, but not on the Cityscapes dataset or ADE20K. GAP averages 2048 activations of size 33×33 for our utilized training parameters. A possible explanation for the effectiveness of GAP on PASCAL VOC 2012 might be that the Cityscapes dataset and ADE20K consist of both a notably larger number and spatial distribution of instances per image. Using GAP on these datasets might, therefore, not aid performance since important features may be lost due to averaging.

5.5 Ablation Study on ADE20K

The performance on clean data ranges from MN-V2 (mIoU of 33.1%) to XC-71 using DPC, as best-performing model, achieving an mIoU of 42.5%. The performance on clean data for most Xception-based backbones (Res-Nets) is highest when Dense Prediction Cell (global average pooling) is used. Our evaluation shows that the mean CD for each ablated architecture is often close to 100.0%, see Fig. 16. Table 4 contains the mIoU of each network backbone for clean and corrupted data.

Effect of ASPP The removal of ASPP can reduce model performance significantly.

Effect of AC The removal of AC decreases the performance slightly for most backbones against corruptions of category digital and weather.

Effect of DPC As on PASCAL VOC 2012 and Cityscapes, applying DPC oftentimes decreases the robustness, especially for Xception-71 against most image corruptions. As on Cityscapes, using DPC along Xception-71, results in the best-performing model on clean data.

Effect of LRL The removal of LRL impacts, especially Xception-71, against image noise.

Effect of GAP When GAP is applied, the models generally perform most robust.

Table 3 Average mIoU for clean and corrupted variants of the PASCAL VOC 2012 validation set for several network backbones of the DeepLabv3+ architecture

		Blur				Noise				
Deeplab-v3+ backbone	Clean	Motion	Defocus	Frosted glass	Gaussian	Gaussian	Impulse	Shot	Speckle	Intensity
ResNet-50	69.6	38.7	43.5	31.1	45.5	43.2	40.7	44.2	50.9	59.8
ResNet-101	70.3	45.8	45.6	33.2	46.6	49.4	48.3	50.1	55.4	61.3
Xception-41	75.5	52.9	54.7	35.5	53.9	55.8	53.3	56.7	62.8	67.6
Xception-65	76.5	53.5	58.3	37.7	57.2	56.6	54.7	57.4	62.5	69.3
Xception-71	**76.7**	**56.5**	**59.1**	**40.2**	**59.5**	**56.6**	**57.8**	**57.6**	**63.2**	**69.9**

	Digital				Weather				
Deeplab-v3+ backbone	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric distortion
ResNet-50	63.5	50.3	63.8	58.2	31.3	47.0	56.9	39.8	67.2
ResNet-101	64.5	50.6	65.3	59.7	31.4	50.4	57.6	41.2	67.6
Xception-41	70.8	51.9	70.9	64.6	42.5	59.0	63.1	48.4	73.0
Xception-65	71.8	55.9	72.1	66.7	40.2	58.5	64.0	47.5	73.6
Xception-71	**72.1**	**57.1**	**72.6**	**68.1**	**43.9**	**60.9**	**66.1**	**50.9**	**73.6**

Every mIoU is averaged over all available severity levels, except for corruptions of category noise where only the first three (of five) severity levels are considered. Highest mIoU per corruption is bold

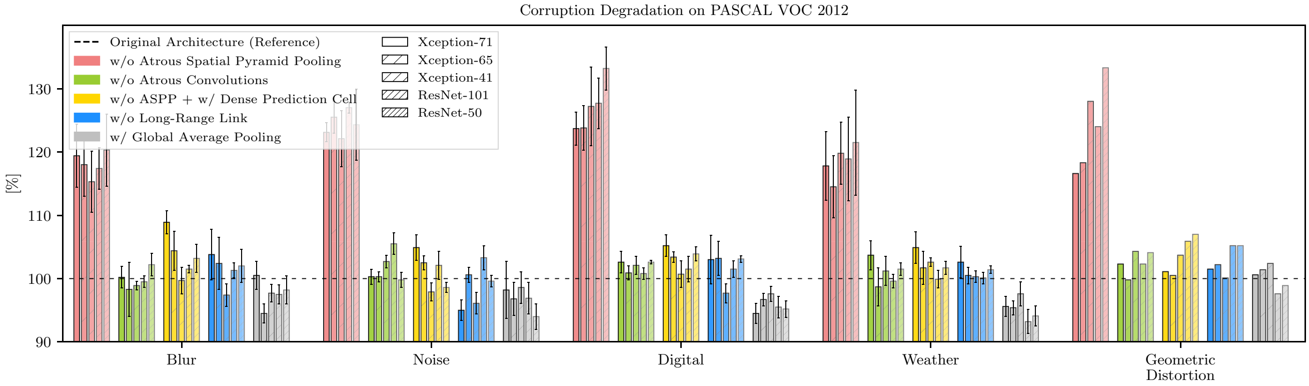


Fig. 15 CD evaluated on PASCAL VOC 2012 for the proposed ablated variants of the DeepLabv3+ architecture w.r.t image corruptions, employing five different network backbones. Each bar except for geometric distortion is averaged within a corruption category (error bars indicate the standard deviation). Bars above 100% represent a decrease in performance compared to the respective reference architecture. Each ablated architecture is re-trained on the original training dataset. Removing ASPP reduces the model performance significantly. AC and LRL decrease robustness against corruptions of category *digital* slightly. Xception-71 is vulnerable against many corruptions when DPC is used. GAP increases performance against many corruptions. Each backbone performs further best on clean data when GAP is used. Best viewed in color (Color figure online)

5.6 Performance for Increasing Severity Levels

We illustrate in Fig. 17 the model performance evaluated on every dataset with respect to individual severity levels. The figure shows the degrading performance with increasing severity level for some candidates of category blur, noise, digital, and weather of a reference model and all corresponding architectural ablations.

The ablated variant without ASPP oftentimes has the lowest mIoU. However, it performs best on speckle noise for severity level 3 and above. The mIoU of the ablated variant without AC is relatively low for defocus blur and contrast. The mIoU of the ablated variant without ASPP and with DPC is relatively low for speckle noise, shot noise (for severity level 4 and 5), spatter. The mIoU of the ablated variant without LRL is relatively high for speckle noise and shot noise. The mIoU of the ablated variant with GAP is high for PASCAL VOC 2012 on clean data and low for speckle noise.

5.7 Robust Model Design Rules

We presented a detailed, large-scale evaluation of state-of-the-art semantic segmentation models with respect to

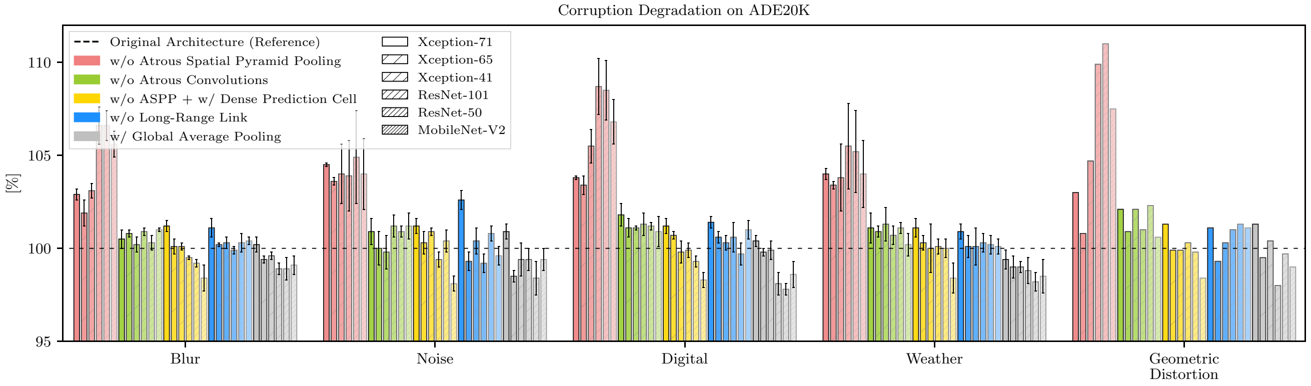


Fig. 16 CD evaluated on ADE20K for the proposed ablated variants of the DeepLabv3+ architecture with respect to image corruptions, employing six different network backbones. Each bar except for geometric distortion is averaged within a corruption category (error bars indicate the standard deviation). Bars above 100% represent a relative decrease in performance compared to the respective reference architecture. Each ablated architecture is re-trained on the original training dataset. Removing ASPP decreases performance oftentimes. AC increase performance slightly against most corruptions. DPC and LRL hamper the performance for Xception-71 with respect to several image corruptions. GAP increases the robustness for most backbones against many image corruptions. Best viewed on screen

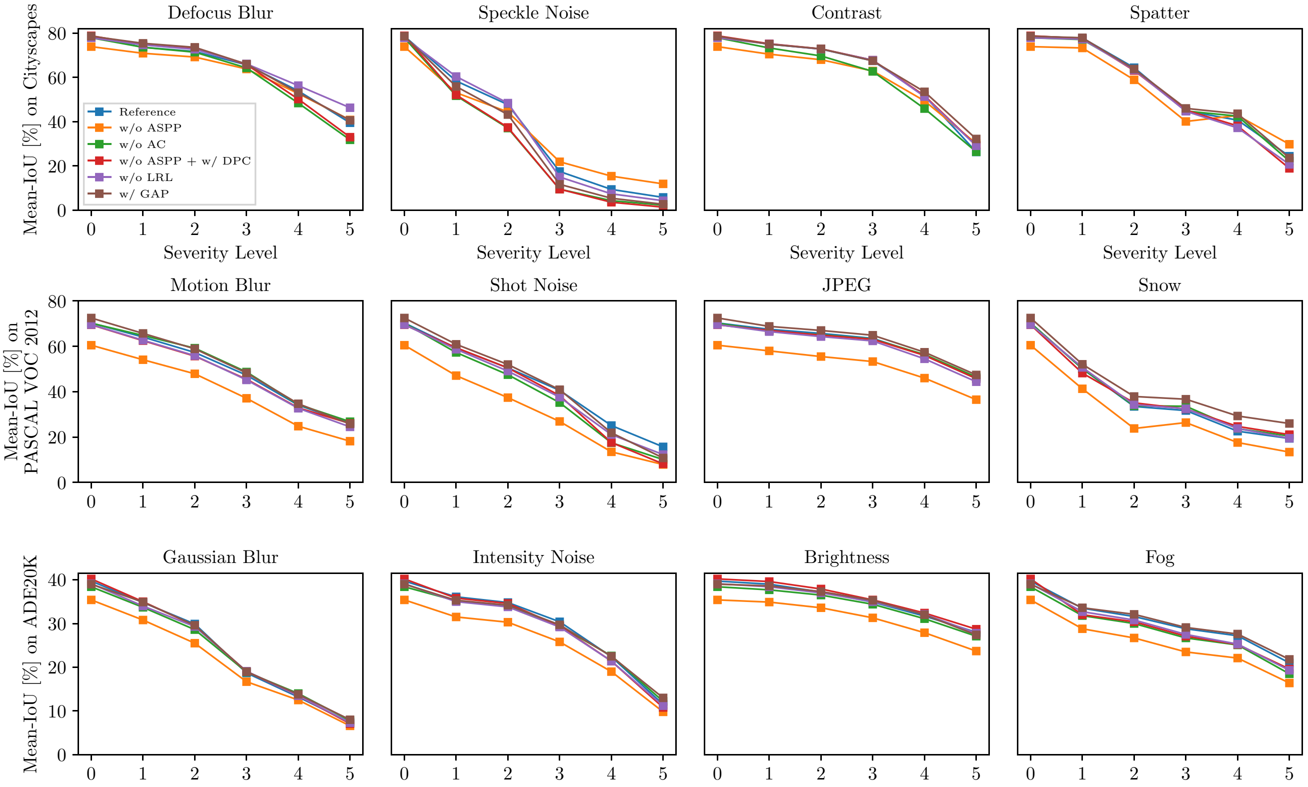


Fig. 17 Model performance (mIoU) for many candidates with respect to the image corruption categories blur (first column), noise (second column), digital (third column), and weather (fourth column) for a reference model and all corresponding architectural ablated variants, evaluated for every severity levels on Cityscapes, PASCAL VOC 2012, and ADE20K. Severity level 0 corresponds to clean data. *First row:* Xception-71 evaluated on the Cityscapes dataset for defocus blur, speckle noise, contrast, and spatter. *Second row:* ResNet-101 evaluated on PASCAL VOC 2012 for motion blur, shot noise, JPEG, and snow. *Third row:* Xception-41 evaluated on ADE20K for Gaussian blur, intensity noise, brightness, and fog

Table 4 Average mIoU for clean and corrupted variants of the ADE20K validation set for several network backbones of the DeepLabv3+ architecture

	Blur					Noise				
Deeplab-v3+ backbone	Clean	Motion	Defocus	Frosted glass	Gaussian	Gaussian	Impulse	Shot	Speckle	Intensity
MobileNet-V2	33.1	16.1	16.6	14.9	16.5	12.1	11.5	12.4	17.0	24.7
ResNet-50	37.4	18.0	19.7	16.9	19.2	14.1	12.8	14.4	19.4	28.5
ResNet-101	38.1	19.1	20.6	17.3	19.8	15.4	14.6	15.7	20.7	28.8
Xception-41	39.7	22.1	22.7	17.4	20.8	20.8	18.1	20.5	24.8	33.7
Xception-65	41.4	23.4	25.2	18.9	22.7	23.2	19.8	22.9	27.1	35.4
Xception-71	**42.4**	**24.4**	**26.4**	**19.5**	**23.9**	**24.0**	**20.3**	**23.3**	**27.5**	**36.8**

	Digital				Weather				
Deeplab-v3+ backbone	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric distortion
MobileNet-V2	27.2	14.8	26.5	25.1	7.8	18.5	20.1	10.7	28.3
ResNet-50	31.1	18.0	30.1	29.5	8.8	21.5	23.9	13.6	32.9
ResNet-101	31.6	19.7	31.2	31.4	10.2	22.9	25.6	14.0	32.8
Xception-41	34.2	20.9	32.5	32.6	13.0	25.0	28.4	17.0	34.4
Xception-65	36.1	23.5	34.8	34.2	14.8	27.7	30.0	18.4	35.6
Xception-71	**37.2**	**25.3**	**35.7**	**34.7**	**16.1**	**29.4**	**31.3**	**19.8**	**37.1**

Every mIoU is averaged over all available severity levels, except for corruptions of category noise where only the first three (of five) severity levels are considered. Highest mIoU per corruption is bold

real-world image corruptions. Based on the study, we can introduce robust model design rules.

Network backbones and architectures Regarding DeepLabv3+, Xception-41 has, in most cases, the best price-performance ratio. It performs especially with respect to Cityscapes and ADE20K close to Xception-71 (the most robust network backbone overall), for a similar performance on clean data but approx. 50% less storage space and less complex architecture. Xception-based backbones are generally more robust than ResNets, however, for a less severe degree of image corruption, this difference is low. MobileNet-V2 is vulnerable to most image corruptions, also for a low severity, however, it is capable of handling blurred data well.

For non-DeepLab-based models, the GSCNN, a model that incorporates shape information, is overall robust against most weather and digital corruptions, and geometrically distorted input, but is also vulnerable against image noise.

Atrous Spatial Pyramid Pooling A multi-scale feature extracting module, like ASPP, is important for geometrically distorted input. Removing ASPP decreases the mIoU, especially for PASCAL VOC 2012, considerably. The relative decrease, when no ASPP is used, is less for the remaining datasets.

Atrous convolutions On Cityscapes, atrous convolutions are generally recommended since they increase robustness against many common corruptions. For such a dataset, atrous convolutions increase the robustness against image blur and noise for many network backbones. With respect to ADE20K, similar tendencies can be observed.

Dense Prediction Cell Models using the DPC instead of ASPP is throughout the datasets vulnerable to many types of image corruptions, especially image noise. This should hence be considered in applications, such as low-light scenarios, where the amount of image noise may be considerably high.

Long-Range Link The previously discussed results indicate that more shallow networks as Xception-41 and ResNet-50 are more robust to corruptions of category image noise, and we recommend hence to omit an LRL for these networks if the respective application comes along with image noise.

Global average pooling Global average pooling should always be used on PASCAL VOC 2012, as its mIoU and robustness are often increased. For Cityscapes, utilizing GAP in Xception-71 is clearly vulnerable to image noise.

6 Image Degradation Study on Cityscapes

In the previous sections, we evaluated the robustness of semantic segmentation models when we train the models on clean data only and evaluated on corrupted data. In this section, we present results when corrupted data is added to the training set.

We train DeepLabv3+ using the ResNet-50 backbone and add a corrupted variant of each image corruptions category (i.e., blur, noise, digital, and weather). This results in four trainings where, compared to a regular training, the amount of training data is doubled.

The results are presented in Fig. 18. Each plot shows the performance degradation for increasing severity level, for

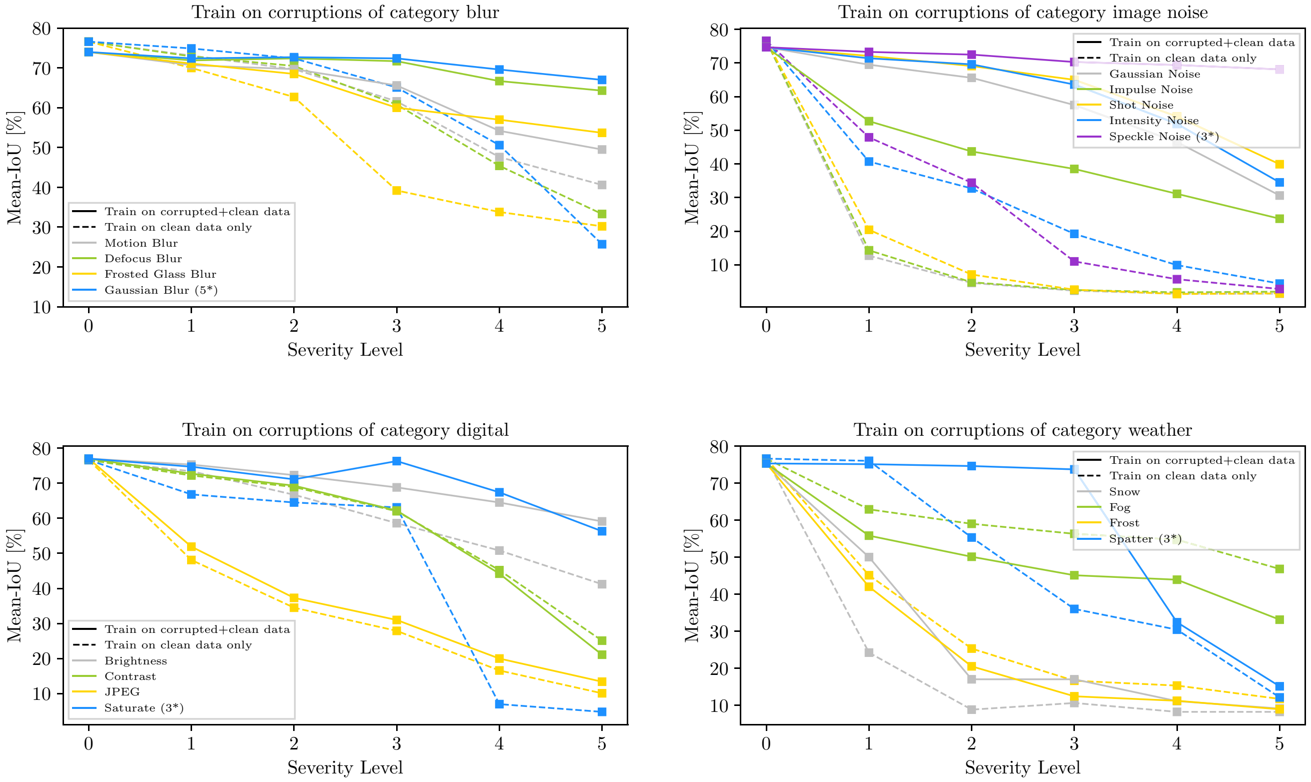


Fig. 18 Model performance when corrupted data is added to the training set. We train four models of DeepLabv3+ using the ResNet-50 backbone and add a corrupted variant of each image corruptions category (i.e., blur, noise, digital, and weather). Each plot shows the performance degradation for increasing severity level, for either a model that is trained on clean data only (dashed lines) or both clean and corrupted data (solid lines). Severity level 0 corresponds to clean data. The last element of each legend is used as training data, marked by an asterisk, and the scalar value indicates the utilized severity level. When the model is trained on corruptions of category blur, noise, and digital, it can generalize to unseen types of respective image corruptions. The model is able to generalize significantly up to a certain severity level well to a wide range of noise models. The model is not able to perform well on every unseen image corruption of category digital

either a trained model on clean data only (dashed lines) or both clean and corrupted data (solid lines). Each legend's last element is used as training data, marked by an asterisk, and the scalar value indicates the utilized severity level.

Study on blur The performance on clean data decreases by 2.6% when image data corrupted by Gaussian blur is added to the clean training data. The model performance further increases for the remaining types of blur. The performance is especially high for defocus blur, probably due to similarity to Gaussian blur.

Study on image noise The performance on clean data decreases by 1.9% when image noise is added to the training data. Interestingly, the model is able to generalize quite well to a wide range of noise models (similar to Rusak et al. (2020) for full-image classification). The model performs well for severity level 4 and 5 of speckle noise, though it is trained on severity level 3. The signal-to-noise ratio of severity level 5 is more than 3dB less than of severity level 3, which corresponds to doubling the degree of noise for that severity level. Whereas the mIoU for Gaussian, impulse, and shot noise is already below 10% for severity level 2, when the model is trained on clean data only, it is significantly increased for the model that is trained on image noise. The model performance decreases significantly for higher severity levels for image noise types that are not part of the training data.

Study on digital corruptions The performance on clean data increases slightly by 0.4% when image corruption "saturate" is added to the training data. Besides for "saturate", the mIoU increases only for "brightness" compared to the model that is trained on clean data only. The image corruptions of this category are quite diverse. "Brightness" and "saturate" have a contrary effect as "contrast". The "JPEG compression", on the other hand, posterizes large areas of the image.

Study on weather corruptions The performance on clean data decreases by 1.9% when image corruption "spatter" is added to the training data. Unlike for image noise, the model cannot generalize to a more severe degree of the corrupted data the model is trained on (i.e., its performance on the fourth and fifth severity level of "spatter" is hardly increased). The mIoU for image corruption "snow" increases significantly

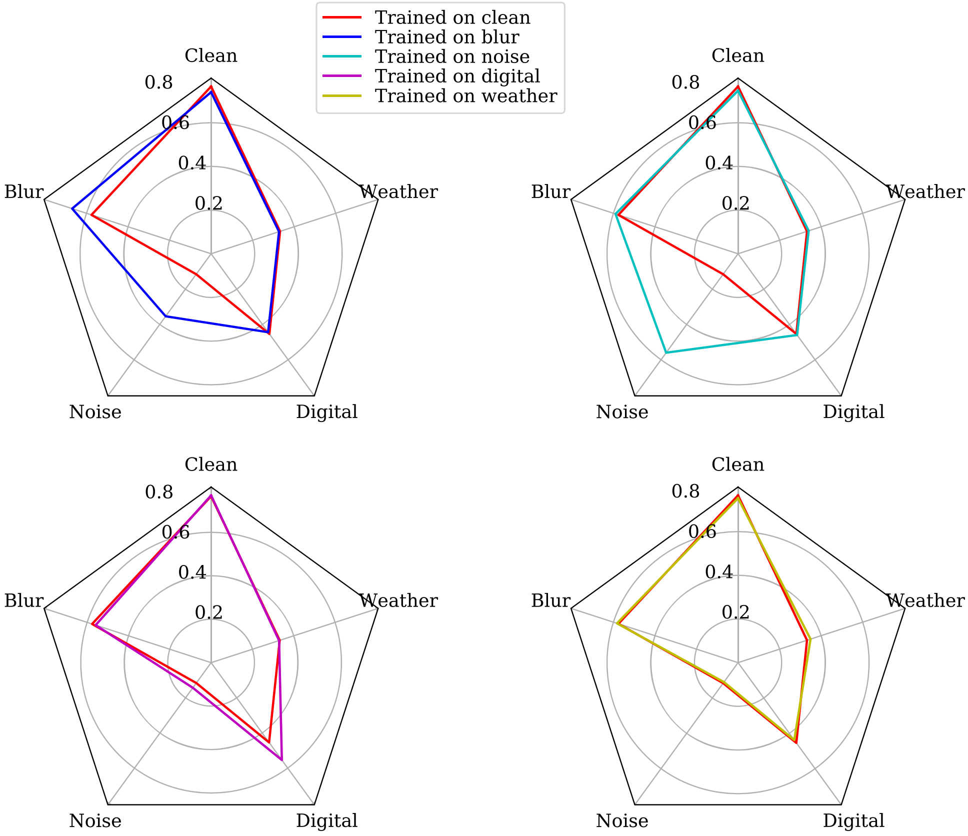


Fig. 19 Averaged mIoU for clean data and the four image corruption categories blur (Gaussian blur, severity level 5), noise (speckle noise, severity level 3), digital (saturate, severity level 3), and weather (spatter, severity level 3). Each radar plot illustrates the performance of a model that is trained on clean data only and a model that is additionally trained on one image corruption category. The models which are trained on a noise, digital, or weather corruption increase the performance in general solely for the respective image corruption category. However, the model that is trained on blur increases the performance also on image noise significantly

severity level 1. Interestingly, this model does not generalize with respect to "fog" and "frost", and performs even worse than the reference model, which is trained on clean data only.

We previously discussed the model performance solely within an image corruption category. In our final evaluation, we illustrate the performance of the remaining image corruption categories (see Fig. 19) as averaged mIoU. Please note that the results in this Figure are based on the same experiments as conducted for Fig. 18. When blurred data is added to the clean training data, the model increases the performance also for noisy data. When noisy data is added to the clean training data, the performance on the remaining image corruption categories is barely affected. Similar results can be observed when data of category digital is added to the training data. For image corruptions of category weather, the average mIoU is only slightly increased when the model is trained on that corruption category.

6.1 Influence of Realistic Image Corruptions

This section focuses on evaluating model robustness with respect to our proposed, more realistic image corruptions. Figure 20 shows the model performance of the ResNet-50 model again when corrupted data is added to the training set. To make severity levels mutually comparable, we average their Signal-to-Noise ratio (SNR) in this Figure over the validation set, i.e., each abscissa represents the averaged SNR of a severity level.[3]

PSF blur We observe that our modeled PSF blur (purple, Fig. 20 left) is in terms of SNR by considerably less severe than the severity levels of the remaining image corruptions of category blur. The mIoU with respect to PSF blur of the first two severity levels is considerably higher than for the remaining types of blur with a similar SNR (i.e., severity level 1 of defocus blur and motion blur), which is observed not only for the ResNet-50 (as illustrated in this Figure) backbone but also for all remaining backbones.

These results might indicate that a CNN could learn, to a certain extent, real PSF blur, which is inherently present in the training data. The fact that the mIoU with respect to PSF blur and Gaussian blur (i.e., the weakest blurs regarding their SNR) decreases when Gaussian blur is added to the training data might also support this hypothesis. However, the performance quickly degrades similarly to an mIoU score that is comparable to the remaining blur types.

Intensity noise The model performs significantly worse for our proposed noise model than for speckle noise, when

[3] Contrary to Fig. 18, where each abscissa corresponds to a severity level in terms of integer values.

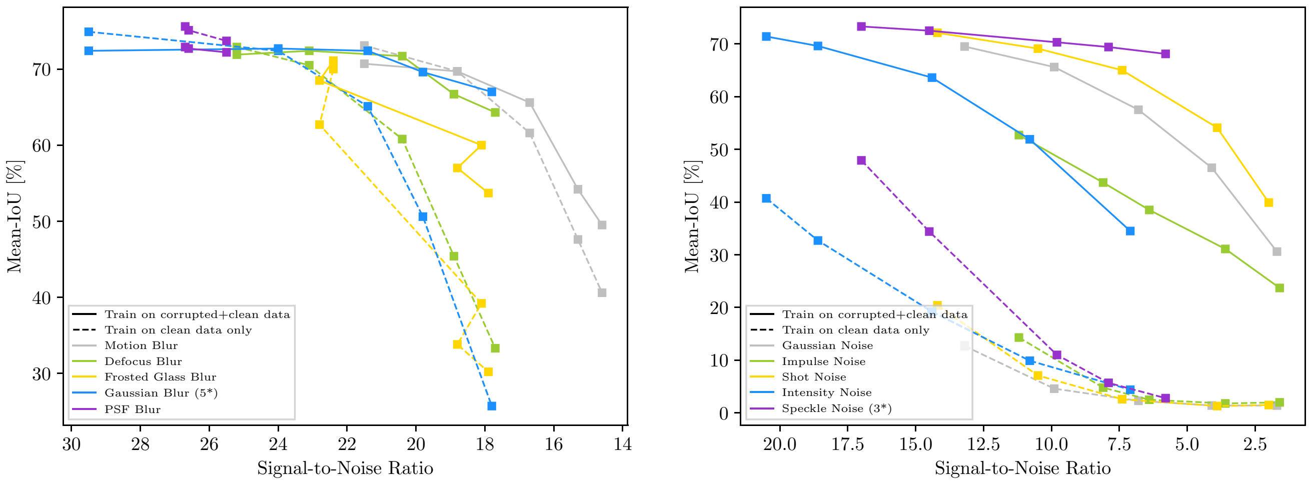


Fig. 20 Model performance when corrupted data is added to the training set. We train four models of DeepLabv3+ using the ResNet-50 backbone and add a corrupted variant of blur and image noise. To make the image corruptions mutually more comparable, each abscissa corresponds to the averaged Signal-to-Noise ratio of the respective image corruption. The models are trained on Gaussian blur (severity level 5, left) or speckle noise (severity level 3, right) (Color figure online)

the model is trained with clean data only (purple, Fig. 20 right, dashed lines). The model's mIoU tends to converge to a common value for each image corruption of category noise. When noisy data is added to the training data, the model performs clearly superior to this particular image corruption. The mIoU of the fifth severity level of speckle noise and third severity level of impulse noise has a similar SNR, but the mIoU differs by approx. 30%.

This result indicates that semantic segmentation models generalize on image noise since a clear mIoU increase is observable; however, it depends strongly on the similarity of image noise models. Based on this assumption, the poor performance with respect to our proposed intensity noise (blue line) indicates that training a model with unrealistic image noise models, is not a reasonable choice for increasing model robustness towards real-world image noise.

Geometric distortion As stated in Sect. 5.2, the model performance with respect to geometric distortion is comparable among the benchmarked architectures (see the last column of Table 1). In general, the GSCNN is the most robust network against geometric distortion. The mIoU of GSCNN decreases for the first severity level by less than 1%. The Xception-based backbones are for the DeepLabv3+ architecture the best-performing networks.

7 Conclusion

We have presented a detailed, large-scale evaluation of state-of-the-art semantic segmentation models with respect to real-world image corruptions. Based on the study, we report various findings of the robustness of specific architectural choices and the generalization behavior of semantic segmentation models. On the one hand, these findings are useful for practitioners to design the right model for their task at hand, where the types of image corruptions are often known. On the other hand, our detailed study may help to improve on the state-of-the-art for robust semantic segmentation models. When designing a semantic segmentation module for a practical application, such as autonomous driving, it is crucial to understand the robustness of the module with respect to a wide range of image corruptions.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *12th*

USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265–283). https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Arnab, A., Miksik, O., & Torr, P. H. S. (2018). On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*.

Azulay, A., & Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, *20*(184), 1–25. http://jmlr.org/papers/v20/19-519.html.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*. https://doi.org/10.1109/TPAMI.2016.2644615. http://ieeexplore.ieee.org/document/7803544/.

Boopathy, A., Weng, T. W., Chen, P. Y., Liu, S., & Daniel, L. (2019). CNN-Cert: An efficient framework for certifying robustness of convolutional neural networks. In *AAAI*.

Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security, AISec '17* (pp. 3–14). New York: ACM. https://doi.org/10.1145/3128572.3140444.

Carlini, N., & Wagner, D. A. (2017). Towards evaluating the robustness of neural Networks. In *2017 IEEE symposium on security and privacy (SP)*.

Chen, L. C., Collins, M. D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., & Shlens, J. (2018a). Searching for efficient multi-scale architectures for dense image prediction. In *Advances in neural information processing systems*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected CRFs. *ICLR*. arXiv:1412.7062.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision - ECCV 2018* (pp. 833–851). Cham: Springer. https://doi.org/10.1007/978-3-030-01234-2_49.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Cisse, M, Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning, Proceedings of machine learning research*. Sydney: PMLR, International Convention Centre. http://proceedings.mlr.press/v70/cisse17a.html.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). AutoAugment: Learning augmentation policies from data. https://arxiv.org/pdf/1805.09501.pdf. Accessed 15 Feb 2020.

Cubuk, E. D., Zoph, B., Schoenholz, S. S., & Le, Q. V. (2018). Intriguing properties of adversarial examples. https://openreview.net/forum?id=rk6H0ZbRb.

Dai, D., & Van Gool, L. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International conference on intelligent transportation systems* (pp. 3819–3824). IEEE.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Conference on computer vision and pattern recognition (CVPR)*.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)* (pp. 1–7). IEEE.

Dodge, S. F., & Karam, L. J. (2016). Understanding how image quality affects deep neural networks. In *Quomex*.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., & Madry, A. (2019). Exploring the landscape of spatial robustness. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proceedings of the 36th international conference on machine learning, *Proceedings of machine learning research* (Vol. 97, pp. 1802–1811). PMLR, Long Beach. http://proceedings.mlr.press/v97/engstrom19a.html.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. In *IJCV*. https://doi.org/10.1007/s11263-009-0275-4.

Fawzi, A., & Frossard, P. (2015). Manitest: Are classifiers really invariant? In *British Machine Vision Conference*.

Fitzgibbon, A. W. (2001). Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR* (Vol. 1, pp. I–I). https://doi.org/10.1109/CVPR.2001.990465.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*. https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Conference on neural information processing systems*. arXiv:1808.08750.

Gilmer, J., Ford, N., Carlini, N., & Cubuk, E. (2019). Adversarial examples are a natural consequence of test error in noise. In K. Chaudhuri, R. Salakhutdinov (Eds.) *Proceedings of the 36th international conference on machine learning, Proceedings of machine learning research* (Vol. 97, pp. 2280–2289). Long Beach: PMLR. http://proceedings.mlr.press/v97/gilmer19a.html.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.

Grauman, K., & Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE international conference on computer vision*.

Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. In *NIPS workshop on deep learning and representation learning*. arXiv:1412.5068.

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. In *CVPR* (pp. 447–456).

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. New York: Cambridge University Press.

Hasirlioglu, S., Kamann, A., Doric, I., & Brandmeier, T. (2016). Test methodology for rain influence on automotive surround sensors. In *ITSC* (pp. 2242–2247). https://doi.org/10.1109/ITSC.2016.7795918.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Healey, G. E., & Kondepudy, R. (1994). Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(3), 267–276.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the international conference on learning representations*.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). AugMix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the international conference on learning representations (ICLR)*.

Holschneider, M., Kronland-Martinet, R., Morlet, J., & Tchamitchian, P. (1989). A real-time algorithm for signal analysis with the help of the wavelet transform. In J.-M. Combes, A. Grossmann, & P. Tchamitchian (Eds.), *Wavelets* (pp. 286–297). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-75988-8_28.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. In *CoRR*. arXiv:1704.04861.

Huang, G., Liu, Z., Maaten, L. V. D., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 2261–2269).

Huang, X., Kwiatkowska, M. Z., Wang, S., & Wu, M. (2017). Safety verification of deep neural networks. In *International conference on computer aided verification*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Jahne, B. (1997). *Digital image processing: Concepts, algorithms, and scientific applications* (4th ed.). Berlin, Heidelberg: Springer.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *Foundations and Trends in Computer Graphics and Vision*, *12*(1–3), 1–308. https://doi.org/10.1561/0600000079.

Joshi, N., Szeliski, R., & Kriegman, D. J. (2008). PSF estimation using sharp edge prediction. In *CVPR* (pp. 1–8). https://doi.org/10.1109/CVPR.2008.4587834.

Kamann, A., Hasirlioglu, S., Doric, I., Speth, T., Brandmeier, T., & Schwarz, U. T. (2017). Test methodology for automotive surround sensors in dynamic driving situations. In *2017 IEEE 85th vehicular technology conference (VTC Spring)* (pp. 1–6). https://doi.org/10.1109/VTCSpring.2017.8108194.

Kamann, C., & Rother, C. (2020). Benchmarking the robustness of semantic segmentation models. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Kannan, H., Kurakin, A., & Goodfellow, I. (2018). Adversarial logit pairing. arXiv preprint arXiv:1803.06373.

Ke, T. W., Maire, M., & Yu, S. X. (2017). Multigrid neural architectures. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 4067–4075).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Laermann, J., Samek, W., & Strodthoff, N. (2019). Achieving generalizable robustness of deep neural networks by stability training. In *German conference on pattern recognition* (pp. 360–373). Springer.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR, Washington, DC, USA*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. In *International conference on learning representations*. arXiv:1312.4400.

Liu, C., Szeliski, R., Kang, S. B., Zitnick, C. L., & Freeman, W. T. (2008). Automatic estimation and removal of noise from a single image. *PAMI*, *30*(2), 299–314. https://doi.org/10.1109/TPAMI.2007.1176.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Lopes, R. G., Yin, D., Poole, B., Gilmer, J., & Cubuk, E. D. (2019). Improving robustness without sacrificing accuracy with patch Gaussian augmentation. arXiv preprint arXiv:1906.02611.

Lukas, J., Fridrich, J., & Goljan, M. (2006). Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, *1*(2), 205–214. https://doi.org/10.1109/TIFS.2006.873602.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., & van der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 181–196).

Metzen, J. H., Genewein, T., Fischer, V., & Bischoff, B. (2017). On detecting adversarial perturbations. In *International conference on learning representations*. arXiv:1702.04267.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., & Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. In *Machine learning for autonomous driving workshop, NeurIPS* (Vol. 190707484). arXiv:1907.07484.

Orhan, A. E. (2019). Robustness properties of Facebook's ResNeXt WSL models. arXiv preprint arXiv:1907.07640.

Papandreou, G., Kokkinos, I., & Savalle, P. A. (2015). Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 390–399).

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). ENet: A deep neural network architecture for real-time semantic segmentation. In *CoRR*. arXiv:1606.02147.

Pham, H., Guan, M. Y., Zoph, B., Le, Q. V., & Dean, J. (2018). Efficient neural architecture search via parameter sharing. In J. Dy & A. Krause (Eds.), *Proceedings of machine learning research* (pp. 4095–4104). Stockholm, Sweden: PMLR. http://proceedings.mlr.press/v80/pham18a.html.

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *CVPR* (pp. 779–788).

Ruderman, A., Rabinowitz, N. C., Morcos, A. S., & Zoran, D. (2018). Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs. arXiv preprint arXiv:1804.04438.

Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., & Brendel, W. (2020). Increasing the robustness of DNNs against image corruptions by playing the game of noise. arXiv:2001.06057.

Sakaridis, C., Dai, D., & Gool, L. V. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 7374–7383).

Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, *126*(9), 973–992.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shah, S., & Aggarwal, J. (1996). Intrinsic parameter calibration procedure for a (high-distortion) fish-eye lens camera with distortion model and accuracy estimation. *Pattern Recognition*, *29*(11), 1775–1788.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*. arXiv:1409.1556.

Szegedy, C., Liu, Wei, Jia, Yangqing, Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *CVPR*. https://doi.org/10.1109/CVPR.2015.7298594. http://ieeexplore.ieee.org/document/7298594/.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., et al. (2013). Intriguing properties of neural networks. In *CoRR*. arXiv:1312.6199.

Takahashi, R., Matsubara, T., & Uehara, K. (2020). Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 2917–2931.

Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019). Gated-SCNN: Gated shape CNNs for semantic segmentation. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 5228–5237).

Vasiljevic, I., Chakrabarti, A., & Shakhnarovich, G. (2016). Examining the impact of blur on recognition by convolutional networks. arXiv:1611.05760 [cs.CV].

Volk, G., Stefan, M., von Bernuth, A., Hospach, D., Bringmann, O. (2019). Towards robust CNN-based object detection through augmentation with synthetic rain variations. In *International conference on intelligent transportation systems*.

Willson, R. G. (1994). Modeling and calibration of automated zoom lenses. In S. F. El-Hakim (Ed.), *Videometrics III* (Vol. 2350, pp. 170–186). International Society for Optics and Photonics, SPIE. https://doi.org/10.1117/12.189130

Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, *90*, 119–133.

Xie, Q., Hovy, E., Luong, M. T., & Le, Q. V. (2019). Self-training with noisy student improves imagenet classification. arXiv preprint arXiv:1911.04252.

Young, I. T., Gerbrands, J. J., & Van Vliet, L. J. (1998). *Fundamentals of image processing* (Vol. 841). Delft: Delft University of Technology.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International conference on learning representations*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE international conference on computer vision* (pp. 6023–6032).

Zendel, O., Honauer, K., Murschitz, M., Steininger, D., & Fernandez Dominguez, G. (2018). Wilddash-creating hazard-aware benchmarks. In *The European conference on computer vision (ECCV)*.

Zendel, O., Murschitz, M., Humenberger, M., & Herzner, W. (2017). How good is my test data? Introducing safety analysis for computer vision. *International Journal of Computer Vision*, *125*(1–3), 95–109.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*. https://openreview.net/forum?id=r1Ddp1-Rb

Zhang, R. (2019). Making convolutional networks shift-invariant again. In *International conference on machine learning*.

Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). ICNet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*. arXiv:1612.01105.

Zheng, S., Song, Y., Leung, T., & Goodfellow, I. J. (2016). Improving the robustness of deep neural networks via stability training. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 4480–4488).

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). Random erasing data augmentation. arXiv preprint arXiv:1708.04896.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641).

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., et al. (2016). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, *127*, 302–321. https://doi.org/10.1007/s11263-018-1140-0.

Zhou, Y., Song, S., & Cheung, N. M. (2017). On classification of distorted images with deep convolutional neural networks. In *ICASSP*.

Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. arXiv:1611.01578.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *CVPR* (pp. 8697–8710).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.