Benchmarking the Robustness of Semantic Segmentation Models

Christoph Kamann and Carsten Rother
Visual Learning Lab
Heidelberg University (HCI/IWR)
http://vislearn.de

Abstract

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. While there are recent robustness studies for full-image classification, we are the first to present an exhaustive study for semantic segmentation, based on the state-of-the-art model DeepLabv3+. To increase the realism of our study, we utilize almost 400,000 images generated from Cityscapes, PASCAL VOC 2012, and ADE20K. Based on the benchmark study, we gain several new insights. Firstly, contrary to full-image classification, model robustness increases with model performance, in most cases. Secondly, some architecture properties affect robustness significantly, such as a Dense Prediction Cell, which was designed to maximize performance on clean data only.

1. Introduction

In recent years, Deep Convolutional Neural Networks (DCNN) have set the state-of-the-art on a broad range of computer vision tasks [49, 36, 71, 73, 52, 65, 11, 29, 35, 51]. The performance of DCNN models is generally measured using benchmarks of publicly available datasets, which often consist of clean and post-processed images [18, 24]. However, it has been shown that model performance is prone to image corruptions [87, 75, 38, 27, 23, 28, 3], 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 [33, 47, 45].

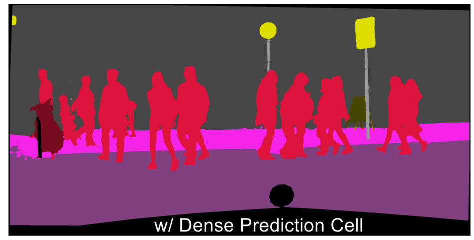
In this work, we present an extensive evaluation of the robustness of semantic segmentation models towards a broad range of real-world image corruptions. Here, the term *robustness* refers to training a model on clean data and then validating it on corrupted data. We choose the task of semantic image 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 [27, 38, 58].

When conducting an evaluation of 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 [27, 38], 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, [27] showed that ResNet-152 [36] is more robust to image noise than GoogLeNet [73]. 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 DCNN models, we believe that it is important to learn about the robustness capabilities of architectural properties.

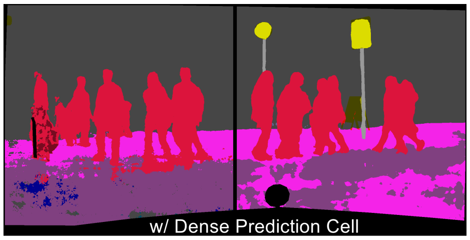
We conduct our study on three popular datasets: Cityscapes [18], PASCAL VOC 2012 [24], and ADE20K [85, 86]. To generate a wide-range of image corruptions, we utilize the image transformations presented by Hendrycks *et al*. [38]. While they give a great selection of image transformations, the level of realism is rather lacking, in our view. Hence we augment their image transformations by additional ones, in particular, intensity-dependent camera noise, PSF blur, and geometric distortions. In total, we employ 19 different image corruptions from the categories of blur, noise, weather, digital, and geometric distortion. We are thus able to validate each DCNN model on almost 400,000 images.

We use the state-of-the-art DeepLabv3+ architecture [14] with multiple network backbones as reference and consider many ablations of it. Based on our evaluation, we are able to conclude two main findings: 1) Contrary to the task of full-image classification, we observe that the ro-

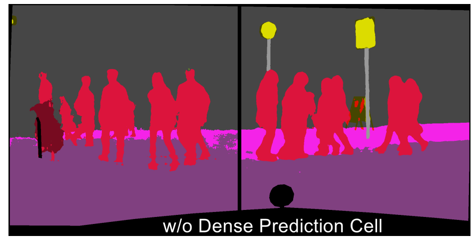
(a) Corrupted validation image (left: noise, right: blur)



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



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



(d) Prediction of ablated architecture on corrupted image

Figure 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. We present a study of various architectural choices and various image corruptions for the three datasets Cityscapes, PASCAL VOC 2012, and ADE20K.

bustness of semantic segmentation models of DeepLabv3+ increases often with model performance. 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 [10], which was designed to maximize performance on clean data, hampers the performance for corrupted images significantly (see Fig. 1).

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 400,000 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) contrary to full-image classification, model robustness of DeepLabv3+ increases with model performance, in most cases; 2) Some architecture properties affect robustness significantly.

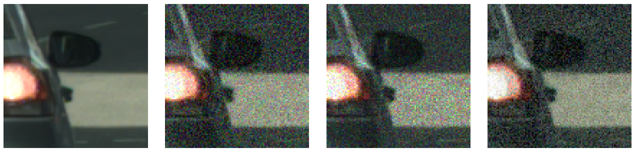
2. Related Work

Robustness studies [87, 75, 38, 27, 22, 23, 60, 58] and robustness enhancement [79, 84, 29, 39, 7, 72, 26] of DCNN architectures [49, 73, 71, 69, 55, 11, 12, 14, 13, 59, 5] have been addressed in various benchmarks [24, 18, 21]. Recent work also dealt with evaluating and increasing the robustness of CNNs against various weather conditions [66, 76, 20, 15, 67]. Vasiljevic *et al.* [75] examined the impact of blur on full-image classification and semantic segmentation using VGG-16 [71]. 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.* [27] compared the generalization capabilities of humans and Deep Neural Networks (DNNs). The ImageNet dataset [21] is modified in terms of color variations, noise, blur, and rotation.

Hendrycks *et al.* [38] introduce the "ImageNet-C dataset". The authors corrupted the ImageNet dataset by common image corruptions. Although the absolute performance scores increase from AlexNet [49] to ResNet [36], the robustness of the respective models does barely change. They further show that Multigrid and DenseNet architectures [48, 42] 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 [18, 24, 85, 86].

Geirhos *et al.* [26] 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 w.r.t. image corruptions increases, when CNNs rely more on object shape than on object texture.



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

Figure 2: A crop of a validation image from Cityscapes corrupted by various noise models. (a) Clean image. (b) Gaussian noise. (c) Shot noise. (d) Our proposed noise model. The amount of noise is high in regions with low pixel intensity.

Robustness of models with respect to adversarial examples is an active field of research [43, 6, 17, 31, 9, 57, 8]. Arnab *et al*. [2] evaluate the robustness of semantic segmentation models for adversarial attacks of a wide variety of network architectures (e.g. [83, 4, 62, 82, 81]). In this work, we adopt a similar evaluation procedure, but we do not focus on the robustness w.r.t. adversarial attacks, which are typically not realistic, but rather on physically realistic image corruptions. We further rate robustness w.r.t. 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.

Ford *et al*. [28] 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.

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.

3.1. ImageNet-C

We employ many image corruptions from the ImageNet-C dataset [38]. These consist of several types of *blur:* motion, defocus, frosted glass and Gaussian; *Noise:* Gaussian, impulse, shot and speckle; *Weather:* snow, spatter, fog, and frost; and *Digital:* brightness, contrast, and JPEG compression. Each corruption is parameterized with five severity levels. We refer to the supplemental material for an illustration of these corruptions.

3.2. Additional Image Corruptions

Intensity-Dependent Noise Model. DCNNs are prone to noise. Previous noise models are often simplistic, *e.g*., images are evenly distorted with Gaussian noise. However, *real* 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 [37, 80, 56, 54]).

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. In accordance with image noise observed from real-world cameras, pixels with low intensities are noisier than pixels with high intensities. Fig. 2 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 [46]. 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.

Geometric distortion. Every camera lens exhibits geometric distortions [25]. We applied several radially-symmetric barrel distortions [77] as a polynomial of grade 4 [70] to both the RGB-image and respective ground truth.

4. Models

We employ DeepLabv3+ [14] 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 [16, 14, 64], denoted by *Xception*) and established ones (*e.g*., ResNets [36]). 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 [55], VGG-16 [71], ICNet [82], DilatedNet [81], ResNet-38 [78], PSPNet [83], and the recent Gated-ShapeCNN (GSCNN) [74].

4.1. DeepLabv3+

Fig. 3 illustrates important elements of the DeepLabv3+ architecture. A network backbone (ResNet, Xception or MobileNet-V2) processes an input image [36, 68, 41]. Its output is subsequently processed by a multi-scale processing module, extracting dense feature maps. This module is either Dense Prediction Cell [10] (DPC) or Atrous Spatial Pyramid Pooling (ASPP, with or without global average pooling (GAP)). We consider the variant with ASPP and

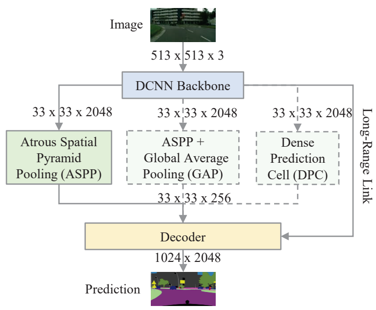


Figure 3: Building blocks of DeepLabv3+. Input images are firstly processed by a network backbone, containing atrous convolutions. The backbone output is further processed by a multi-scale processing module (ASPP or DPC). A long-range link concatenates early features of the network backbone with encoder 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.

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 [12, 40, 61] 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 [12, 11, 83] perform Spatial Pyramid Pooling [34, 30, 50]. DeepLabv3+ applies *Atrous* Spatial Pyramid Pooling (ASPP), where three atrous convolutions with large atrous rates (6,12 and 18) process the DCNN output.

Dense Prediction Cell. [10] 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 [32]. 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 [53] 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 (sec. 5.1) and then the results of two different experiments. We firstly benchmark multiple neural network backbone architectures of DeepLabv3+ and other semantic segmentation models (sec. 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+ (sec. 4.2) and evaluate the robustness for re-trained ablated models w.r.t. image corruptions (sec. 5.3, 5.4, 5.5). Our findings show that architectural properties can have a substantial impact on the robustness of a semantic segmentation model w.r.t. image corruptions.

5.1. Experimental Setup

Network backbones. We trained DeepLabv3+ with several network backbones on clean and corrupted data using TensorFlow [1]. 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 [44], 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, 2000 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 [38] as follows.

We use the term *Degradation D*, where $D = 1 - mIoU$ in place of *Error*. Degradations across severity levels,

		Blur					Noise					Digital				Weather				
Architecture	Clean	Motion	Defocus	Frosted Glass	Gaussian	PSF	Gaussian	Impulse	Shot	Speckle	Intensity	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric Distortion
MobileNet-V2	72.0	53.5	49.0	45.3	49.1	70.5	6.4	7.0	6.6	16.6	26.9	51.7	46.7	32.4	27.2	13.7	38.9	47.4	17.3	65.5
ResNet-50	76.6	58.5	56.6	47.2	57.7	74.8	6.5	7.2	10.0	31.1	30.9	58.2	54.7	41.3	27.4	12.0	42.0	55.9	22.8	69.5
ResNet-101	77.1	59.1	56.3	47.7	57.3	75.2	13.2	13.9	16.3	36.9	39.9	59.2	54.5	41.5	37.4	11.9	47.8	55.1	22.7	69.7
Xception-41	77.8	61.6	54.9	51.0	54.7	76.1	**17.0**	**17.3**	**21.6**	**43.7**	48.6	63.6	56.9	**51.7**	38.5	18.2	46.6	57.6	20.6	**73.0**
Xception-65	78.4	63.9	59.1	**52.8**	59.2	**76.8**	15.0	10.6	19.8	42.4	46.5	65.9	**59.1**	46.1	31.4	**19.3**	**50.7**	63.6	**23.8**	72.7
Xception-71	**78.6**	**64.1**	**60.9**	52.0	**60.4**	76.4	14.9	10.8	19.4	41.2	**50.2**	**68.0**	58.7	47.1	**40.2**	18.8	50.4	**64.1**	20.2	71.0
ICNet	65.9	45.8	44.6	**47.4**	44.7	65.2	8.4	8.4	10.6	27.9	29.7	41.0	33.1	27.5	**34.0**	6.3	30.5	27.3	11.0	35.7
FCN8s-VGG16	66.7	42.7	31.1	37.0	34.1	61.4	6.7	5.7	7.8	24.9	18.8	53.3	39.0	36.0	21.2	11.3	31.6	37.6	19.7	36.9
DilatedNet	68.6	44.4	36.3	32.5	38.4	61.1	**15.6**	14.0	**18.4**	32.7	35.4	52.7	32.6	38.1	29.1	12.5	32.3	34.7	19.2	38.9
ResNet-38	77.5	54.6	45.1	43.3	47.2	74.9	13.7	**16.0**	18.2	**38.3**	**35.9**	60.0	50.6	46.9	14.7	**13.5**	45.9	52.9	22.2	43.2
PSPNet	78.8	**59.8**	53.2	44.4	53.9	76.9	11.0	15.4	15.4	34.2	32.4	60.4	51.8	30.6	21.4	8.4	42.7	34.4	16.2	**43.4**
GSCNN	**80.9**	58.9	**58.4**	41.9	**60.1**	**80.3**	5.5	2.6	6.8	24.7	29.7	**75.9**	**61.9**	**70.7**	12.0	12.4	**47.3**	**67.9**	**32.6**	42.7

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*). 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.

which are defined by the ImageNet-C corruptions [38], 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

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

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. While we predominately use CD 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). We highlight the difference between CD and rCD in more detail in the supplement.

$$rCD_c^f = \left(\sum_{s=1}^{5} D_{s,c}^f - D_{clean}^f \right) \Big/ \left(\sum_{s=1}^{5} D_{s,c}^{ref} - D_{clean}^{ref} \right) \tag{2}$$

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. We refer to the supplement for the respective results for other datasets and individual severity levels.

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. Please find some result images in the supplement.

Performance w.r.t. noise. Noise has a substantial impact on model performance. 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 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. Models of DeepLabv3+ handle geometric distortion significantly better than non-DeepLabv3+ based models.

This benchmark indicates, in general, a similar result as in [26], that is image distortions corrupting the texture of an image (*e.g.*, image noise, snow, frost, JPEG), often have distinct 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). Fig. 4 illustrates the mean CD and rCD with re-

[1]Note that we were not able to reproduce the results from [14]. We conjecture that this is due to hardware limitations, as we could not set the suggested crop-size of 769×769 for Cityscapes.

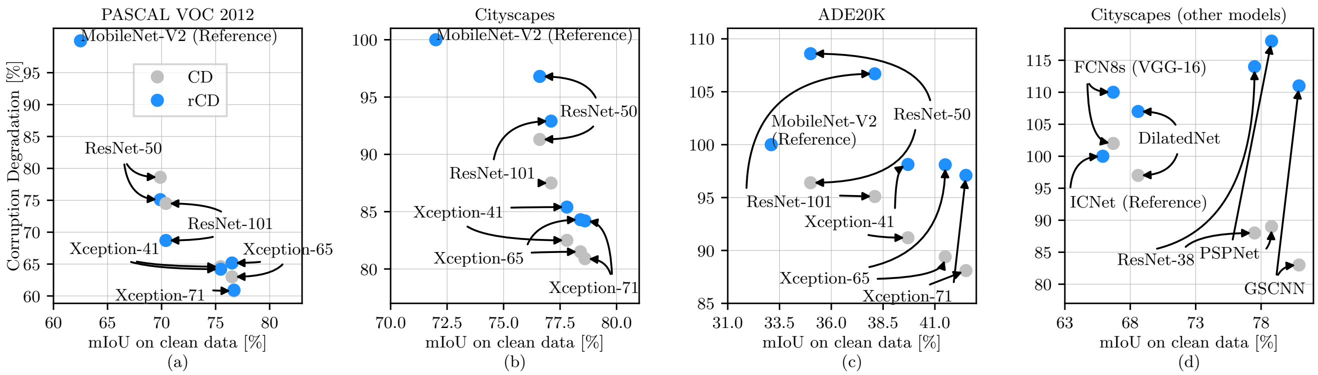


Figure 4: (a−c) CD and rCD for several network backbones of the DeepLabv3+ architecture evaluated on PASCAL VOC 2012, the Cityscapes dataset, and ADE20K. MobileNet-V2 is the reference model in each case. rCD and CD values below 100 % represent higher robustness than the reference model. In almost every 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 %.

spect to the mIoU for *clean* images (lower values correspond to higher robustness regarding the reference model). Each dot depicts the performance of one network backbone, averaged over all corruptions except for PSF blur[2]. Subplot a−c illustrates respective results for PASCAL VOC 2012, Cityscapes, and ADE20K. On each dataset, Xception-71 is the most robust network backbone for DeepLabv3+ architecture. Interestingly, rCD decreases often with increasing model performance, except for Xception-65 on PASCAL VOC 2012 (Fig. 4 a) and ResNets on ADE20K (Fig. 4 c). The latter result indicates that ResNet-based backbones are vulnerable when applied for a large-scale dataset as ADE20K. Fig. 4 d presents the respective result for several non-DeepLabv3+ based segmentation models. The rCD for these models increases slightly. On the other hand, CD decreases mostly with increasing model performance on clean data. The authors of [38] 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. 4 d indicate. When we, however, evaluate within a semantic segmentation architecture, as DeepLabv3+, the contrary result (*i.e.*, decreasing rCD) is generally observed. The following speculation may also give further insights. Geirhos *et al*. [26] stated recently 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 might have a more pronounced shape bias than others (*e.g.*, ResNets), resulting hence in a higher rCD score w.r.t. image corruption. This may be an interesting topic for future work, however, beyond the scope of this paper.

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 section 4.2. Each ablated variant has been retrained 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. We refer to the supplement for the results of the remaining backbones. 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. 5 (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). We refer to the supplement for an evaluation.

Effect of AC. Atrous convolutions (AC) generally show a positive effect w.r.t. corruptions of type blur for most net-

[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.

		Blur					Noise					Digital				Weather				
Deeplab-v3+ Backbone	Clean	Motion	Defocus	Frosted Glass	Gaussian	PSF	Gaussian	Impulse	Shot	Speckle	Intensity	Brightness	Contrast	Saturate	JPEG	Snow	Spatter	Fog	Frost	Geometric Distortion
Xception-71	78.6	64.1	60.9	52.0	60.4	76.4	14.9	10.8	19.4	41.2	**50.2**	**68.0**	58.7	47.1	40.2	**18.8**	50.4	64.1	20.2	71.0
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	63.4	56.2	42.7	39.9	17.6	49.0	58.3	21.8	69.3
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	67.6	55.6	46.0	**40.7**	18.2	50.1	61.1	**21.6**	71.1
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	64.8	59.4	45.3	32.0	14.4	48.6	64.0	20.8	72.1
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	64.5	59.2	44.3	36.1	16.9	48.7	**64.3**	21.3	71.3
w/ GAP	78.6	**64.2**	61.7	**55.9**	60.7	**77.8**	9.7	8.4	13.9	36.9	45.6	68.0	**60.2**	**48.4**	40.6	16.8	**51.0**	62.1	20.9	**73.6**

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

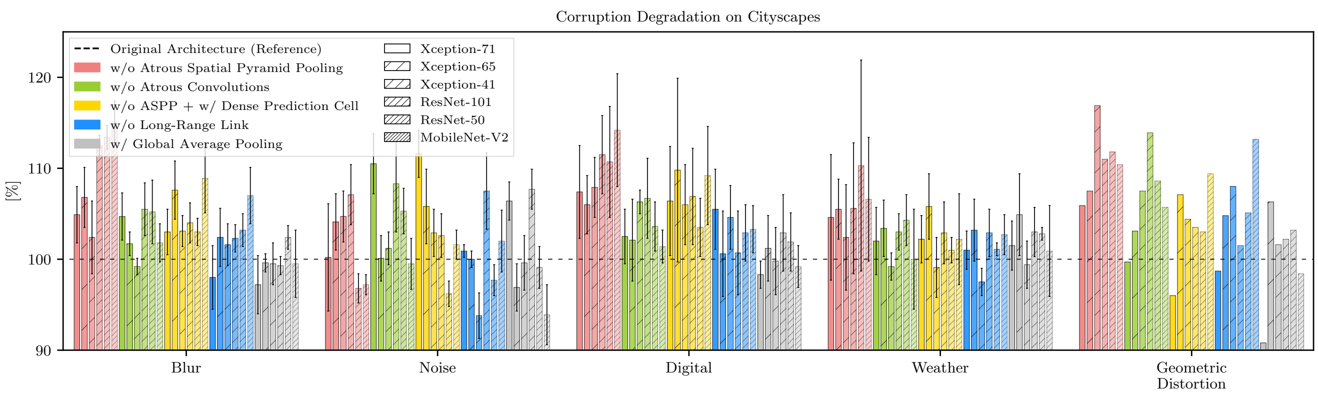


Figure 5: CD evaluated on Cityscapes for the proposed ablated variants of the DeepLabv3+ architecture w.r.t. image corruptions, employing six different network backbones. Bars above 100 % represent a decrease in performance compared to the respective reference architecture. Each ablated architecture 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).

work backbones, especially for XC-71 and ResNets. 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. 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. AC generally exhibit also a positive effect w.r.t. geometric distortion. For MN-V2 and ResNets, the averaged mIoU reduces by up to 4.2 % ($\text{CD}^{\text{ResNet-50}} = 109\,\%$, $\text{CD}^{\text{ResNet-101}} = 114\,\%$, $\text{CD}^{\text{MN-V2}} = 106\,\%$). 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 vulnerable against corruptions of most categories. While this ablated architecture 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.*, [89, 88, 63]) with the objective of maximizing performance on clean data. This result indicates that such architectures tend to over-fit 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 [19] w.r.t. adversarial examples. We further hypothesize that DPC learns less multi-scale representations than ASPP, which may be useful against common corruptions. We discuss this hypothesis in the supplement.

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.*, $\text{CD}^{\text{XC-71}} = 111\,\%$). 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

Ablation	**w/o ASPP**					**w/o AC**					**w/o ASPP w/ DPC**					**w/o LRL**					**w/ GAP**				
Network	ResNet–		Xception–			ResNet–		Xception–			ResNet–		Xception–			ResNet–		Xception–			ResNet–		Xception–		
Backbone	50	101	41	65	71	50	101	41	65	71	50	101	41	65	71	50	101	41	65	71	50	101	41	65	71
Blur	**120**	117	115	118	**119**	**102**	99	99	98	**100**	**103**	101	100	104	**109**	**102**	101	97	102	**104**	**98**	98	98	95	**101**
Noise	124	**127**	122	**126**	123	100	**106**	103	100	101	99	**102**	98	103	**105**	100	**103**	96	**101**	95	94	**97**	**99**	97	98
Digital	**133**	128	**127**	124	124	**103**	101	102	101	**103**	**104**	102	101	103	**105**	**103**	102	98	**103**	103	95	**96**	**98**	97	94
Weather	**121**	119	**120**	114	118	**101**	100	101	99	**104**	**102**	100	103	102	**105**	**101**	100	100	101	**103**	**94**	93	**98**	95	96
Geometric Distortion	**133**	124	**128**	118	117	**104**	102	**104**	100	102	**107**	106	**104**	100	101	**105**	105	100	**102**	102	**99**	98	**102**	101	101

Table 3: CD evaluated on PASCAL VOC 2012 for ablated network backbones of the DeepLabv3+ architecture w.r.t. image corruptions.

slightly robustness w.r.t. blur 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 $112\,\%$).

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 Table 3). 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. Please see the supplement corresponding mIoU scores.

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 w.r.t. 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.*, [89, 88, 63] has been performed on XC-71 and enhances, therefore, the over-fitting effect additionally, as discussed in section 5.3.

Effect of LRL. Removing LRL increases robustness against noise for XC-71 and XC-41, probably due to discarding early features (we refer to the supplement for discussion). However, this finding does not hold for XC-65. As reported in section 5.2, on PASCAL VOC 2012, XC-65 is also the most robust model against noise. Regarding ResNets, the LRL affects the image corruption of category geometric distortion the most.

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 (we refer to supplement).

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 $42.5\,\%$ (detailed results listed in the supplement). The performance on clean data for most Xception-based backbones (ResNets) 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 $100.0\,\%$. The impact of proposed architectural properties on model performance is thus on the large-scale dataset ADE20K hardly present. A possible explanation is probably that the effect of architectural design choices becomes more decisive, and respective impacts are more pronounced when models perform well, *i.e.* have large mIoU. DeepLabv3+ performs much poorer on ADE20K than, *e.g.*, on the Cityscapes dataset.

The tendencies of the previous findings are nevertheless present. Regarding XC-71, for example, the corresponding means of both CD and rCD for DPC are respectively $101\,\%$ and $107\,\%$, showing its robustness is again less than the reference model. ASPP, on the other hand, affects segmentation performance also significantly.

6. 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 can introduce robust model design rules: Atrous convolutions are generally recommended since they increase robustness against many corruptions. The vulnerability of Dense Prediction Cell to many corruptions must be considered, especially in low-light and safety-critical applications. The ASPP module is important for decent model performance, especially for digitally and geometrically distorted input. Global average pooling should always be used on PASCAL VOC 2012. Our detailed study may help to improve on the state-of-the-art for robust semantic segmentation models.

References

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. In *CVPR*, 2018.

[3] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR*, abs/1805.12177, 2018.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In *PAMI*, 2017.

[5] Chris H Bahnsen, David Vázquez, Antonio M López, and Thomas B Moeslund. Learning to Remove Rain in Traffic Surveillance by Using Synthetic Data. In *VISIGRAPP*, 2019.

[6] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks. In *AAAI*, Jan. 2019.

[7] Tejas S. Borkar and Lina J. Karam. DeepCorrect: Correcting DNN models against Image Distortions. arXiv:1705.02406 [cs.CV], 2017.

[8] Nicholas Carlini and David Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pages 3–14, New York, NY, USA, 2017. ACM.

[9] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[10] Liang-Chieh Chen, Maxwell D. Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jonathon Shlens. Searching for Efficient Multi-Scale Architectures for Dense Image Prediction. In *NIPS*, 2018.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*, volume abs/1412.7062, 2015.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. In *TPAMI*, 2017.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation, 2017.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*, 2018.

[15] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018.

[16] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*, 2017.

[17] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval Networks: Improving Robustness to Adversarial Examples. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, International Convention Centre, Sydney, Australia, 2017. PMLR.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[19] Ekin Dogus Cubuk, Barret Zoph, Samuel Stern Schoenholz, and Quoc V. Le. *Intriguing Properties of Adversarial Examples*. 2018.

[20] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, pages 3819–3824. IEEE, 2018.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[22] Samuel Dodge and Lina Karam. 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)*, pages 1–7. IEEE, 2017.

[23] Samuel F. Dodge and Lina J. Karam. Understanding how image quality affects deep neural networks. In *Quomex*, 2016.

[24] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. In *IJCV*, 2010.

[25] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR*, volume 1, pages I–I, Dec. 2001.

[26] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, May 2019.

[27] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. *NIPS*, abs/1808.08750, 2018.

[28] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2280–2289, Long Beach, California, USA, June 2019. PMLR.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[30] K. Grauman and T. Darrell. The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In *ICCV*, 2005.

[31] Shixiang Gu and Luca Rigazio. Towards Deep Neural Network Architectures Robust to Adversarial Examples. *NIPS Workshop on Deep Learning and Representation Learning*, abs/1412.5068, 2014.

[32] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, pages 447–456, 2015.

[33] S. Hasirlioglu, A. Kamann, I. Doric, and T. Brandmeier. Test methodology for rain influence on automotive surround sensors. In *ITSC*, pages 2242–2247, Nov. 2016.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *ECCV*, 2014.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *ICCV*, pages 1026–1034, 2015.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[37] Glenn E Healey and Raghava Kondepudy. Radiometric CCD camera calibration and noise estimation. *PAMI*, 16(3):267–276, 1994.

[38] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[39] João F. Henriques and Andrea Vedaldi. Warped Convolutions: Efficient Invariance to Spatial Transformations. In *ICML*, 2017.

[40] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform. In J.-M. Combes, A. Grossmann, and P. Tchamitchian, editors, *Wavelets. Time-Frequency Methods and Phase Space*, page 286, 1989.

[41] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, abs/1704.04861, 2017.

[42] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *CVPR*, pages 2261–2269, 2017.

[43] Xiaowei Huang, Marta Z. Kwiatkowska, Sen Wang, and Min Wu. Safety Verification of Deep Neural Networks. In *CAV*, 2017.

[44] Ioffe, Sergey and Szegedy, Christian. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015.

[45] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. *Arxiv*, 2017.

[46] N. Joshi, R. Szeliski, and D. J. Kriegman. PSF estimation using sharp edge prediction. In *CVPR*, pages 1–8, June 2008.

[47] A. Kamann, S. Hasirlioglu, I. Doric, T. Speth, T. Brandmeier, and U. T. Schwarz. Test Methodology for Automotive Surround Sensors in Dynamic Driving Situations. In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pages 1–6, June 2017.

[48] Tsung-Wei Ke, Michael Maire, and Stella X. Yu. Multigrid Neural Architectures. In *CVPR*, pages 4067–4075, 2017.

[49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[50] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, Washington, DC, USA, 2006.

[51] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. In *Nature*, 2015.

[52] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[53] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014.

[54] C. Liu, R. Szeliski, S. Bing Kang, C. L. Zitnick, and W. T. Freeman. Automatic Estimation and Removal of Noise from a Single Image. *PAMI*, 30(2):299–314, Feb. 2008.

[55] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, volume abs/1411.4038, 2015.

[56] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, June 2006.

[57] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On Detecting Adversarial Perturbations. In *ICLR*, 2017.

[58] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. In *Machine Learning for Autonomous Driving Workshop, NeurIPS 2019*, volume 190707484, July 2019.

[59] Jashojit Mukherjee, K Praveen, and Venugopala Madumbu. Visual Quality Enhancement Of Images Under Adverse Weather Conditions. In *ITSC*, pages 3059–3066. IEEE, 2018.

[60] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning. In *NIPS*, 2017.

[61] George Papandreou, Iasonas Kokkinos, and Pierre-André Savalle. Modeling local and global deformations in Deep Learning: Epitomic convolution, Multiple Instance Learning, and sliding window detection. In *CVPR*, pages 390–399, 2015.

[62] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *CoRR*, abs/1606.02147, 2016.

[63] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *ICML*, 2018.

[64] Haozhi Qi, Zheng Zhang, Bin Xiao, Han Hu, Bowen Cheng, Yichen Wei, and Jifeng Dai. Deformable convolutional networks–coco detection and segmentation challenge 2017 entry. In *ICCV COCO Challenge Workshop*, volume 15, 2017.

[65] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, pages 779–788, 2016.

[66] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018.

[67] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation. In *ICCV*, 2019.

[68] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018.

[69] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Robert Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[70] Shishir Shah and JK Aggarwal. Intrinsic parameter calibration procedure for a (high-distortion) fish-eye lens camera with distortion model and accuracy estimation. *Pattern Recognition*, 29(11):1775–1788, 1996.

[71] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[72] Zhun Sun, Mete Ozay, Yan Zhang, Xing Liu, and Takayuki Okatani. Feature Quantization for Defending Against Distortion of Images. In *CVPR*, June 2018.

[73] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[74] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. *ICCV*, 2019.

[75] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the Impact of Blur on Recognition by Convolutional Networks. *arXiv:1611.05760 [cs.CV]*, abs/1611.05760, 2016.

[76] Georg Volk, Mueller Stefan, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards Robust CNN-Based Object Detection through Augmentation with Synthetic Rain Variations. In *ITSC*, 2019.

[77] Reg G Willson. Modeling and calibration of automated zoom lenses. In *Videometrics III*, volume 2350, pages 170–187. International Society for Optics and Photonics, 1994.

[78] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[79] Jonghwa Yim and Kyung-Ah Sohn. Enhancing the Performance of Convolutional Neural Networks on Quality Degraded Datasets. *DICTA*, 2017.

[80] Ian T Young, Jan J Gerbrands, and Lucas J Van Vliet. *Fundamentals of image processing*, volume 841. Delft University of Technology Delft, 1998.

[81] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*, 2016.

[82] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *ECCV*, 2018.

[83] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017.

[84] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the Robustness of Deep Neural Networks via Stability Training. In *CVPR*, pages 4480–4488, 2016.

[85] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[86] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, pages 1–20, 2016.

[87] Yiren Zhou, Sibo Song, and Ngai-Man Cheung. On Classification of Distorted Images with Deep Convolutional Neural Networks. *ICASSP*, 2017.

[88] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. 2017.

[89] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018.