

# Robust Semantic Segmentation with Superpixel-Mix

Gianni Franchi<sup>1</sup>

gianni.franchi@ensta-paris.fr

Nacim Belkhir<sup>2</sup>

nacim.belkhir@safrangroup.com

Mai Lan Ha<sup>3</sup>

hamailan@informatik.uni-siegen.de

Yufei Hu<sup>1</sup>

yufei.hu.2021@ensta-paris.fr

Andrei Bursuc<sup>4</sup>

andrei.bursuc@valeo.com

Volker Blanz<sup>3</sup>

blanz@informatik.uni-siegen.de

Angela Yao<sup>5</sup>

ayao@comp.nus.edu.sg

<sup>1</sup> U2IS ENSTA Paris

Institut Polytechnique de Paris

<sup>2</sup> Safrantech, Safran Group

<sup>3</sup> Department of Computer  
Science University of Siegen

<sup>4</sup> valeo.ai

<sup>5</sup> School of Computing  
National University of Singapore

---

## Abstract

Along with predictive performance and runtime speed, robustness is a key requirement for real-world semantic segmentation. Robustness encompasses accuracy, predictive uncertainty, stability under data perturbation and distribution shift, and reduced bias. To improve robustness, we introduce Superpixel-mix, a new superpixel-based data augmentation method with teacher-student consistency training. Unlike other mixing-based augmentation techniques, mixing superpixels between images is aware of object boundaries, while yielding consistent gains in segmentation accuracy. Our proposed technique achieves state-of-the-art results in semi-supervised semantic segmentation on the Cityscapes dataset. Moreover, Superpixel-mix improves the robustness of semantic segmentation by reducing network uncertainty and bias, as confirmed by competitive results under strong distributions shift (adverse weather, image corruptions) and when facing out-of-distribution data.

## 1 Introduction

Semantic segmentation is an important task in computer vision with a high potential for practical applications, in particular for autonomous vehicles. Thanks to the predicted 2D segmentation maps by the deep convolutional neural networks (DCNNs), it contributes to an improved understanding of the scenes.

A large body of the recent literature on DCNNs for semantic segmentation focuses on improving predictive performance and run-time through advanced [6, 82] or lighter [24, 45, 81, 86] architectures, better use of multiple resolutions [72, 85] and novel loss functions [9, 47, 71, 87]. However, for real-world deployments, other requirements, e.g., reliability, robustness, must be equally satisfied to avoid any failures. To reach them, there are a few major challenges yet to be fully solved. First, DCNNs have been shown to be overconfident [23] even when their predictions are wrong [29, 56]. In addition, DCNNs struggle to learn when there are few training samples are available, or data and annotations are noisy. In particular, high capacity DCNNs can find “shortcuts” that allow them to exploit spurious correlations in the data (e.g., background information [53, 59], textures and salient patterns [21]) towards minimizing the training error at the cost of generalization. Such DCNNs have been shown to be biased, e.g. contextual bias [80] or texture bias [21]. This type of problem could be addressed by larger and higher quality datasets [42], yet the entire complexity of the world cannot be encompassed in a training dataset with a limited size. Alternative solutions leverage uncertainty estimation for detecting such failures [16, 19, 41]. However the most effective ones are computationally inefficient as they rely on ensembles or multiple forward passes [0, 26, 58].

In this paper, we aim to increase the robustness of semantic segmentation models. For the scope of this work, we define *robust* as follows: *a model is robust if its predictions are accurate and well calibrated when facing typical conditions from the training distribution, but also under distribution shift (epistemic and aleatoric uncertainty) and for unknown object classes which are not seen during training (epistemic uncertainty)*. Our definition extends the scope of *robustness* beyond invariance to different perturbations of the input, e.g., adversarial attacks [4, 73], image corruptions [30, 51], change of style [21], where only prediction accuracy is used as a proxy for robustness. Here, a robust model must not be only accurate, but also well calibrated such that unknown objects or strong input perturbations are designated low confidence scores and easily identified as unreliable and discarded. Although difficult to attain, we argue that both accuracy and calibration are essential for deployment in real world conditions where the data distribution is not identical to the training distribution.<sup>1</sup> To improve the robustness of a DCNN, given the noisy nature of the data, and to address the problem of limited numbers of training labeled images, we propose a technique that combines the teacher-student framework [74] with a novel data augmentation strategy (Fig. 1). Our augmentation method, named *Superpixel-mix*, exchanges superpixels between training images to generate more training samples that preserve object boundaries and disentangle objects’ parts from their frequent contexts. To the best of our knowledge, this is among the first investigations on the use of these techniques for reducing DCNN bias and uncertainty.

**Contributions:** In summary, our contributions are as follows: (i) Superpixel-mix, a new type of data augmentation for creating new unlabeled images to increase DCNNs’ accuracy and robustness. (ii) A theoretical grounding on why mixing augmentation combined with the teacher-student framework can improve robustness. The theory is confirmed by a set of experiments. (iii) A new dataset for quantifying contextual bias of DCNNs.<sup>2</sup>

## 2 Related Work

**Robust Deep Learning.** Robustness of DCNNs has been studied under different perspectives in the past few years, e.g., robustness to adversarial attacks [13, 59]. We focus here rather on

<sup>1</sup>The two metrics are often at odds with each other: a classifier can be accurate but non-calibrated (usually overconfident [23, 49, 56]) and conversely it can be inaccurate yet calibrated, if its predictions are always low-confident.

<sup>2</sup>The dataset will be made publicly available after the anonymity period

the robustness of the perception functions, and less on security aspects. Geirhos *et al.* [20] observe that classification models trained on ImageNet are biased towards textures and blind to shapes. They mitigate this by augmenting the training set with stylized images [20], yet this can be detrimental for semantic segmentation as object boundaries are distorted. Shetty *et al.* [64] counter contextual bias with a data augmentation strategy that removes random objects from images. Some other works focus on evaluating the robustness under different image perturbations, e.g., blur [77], brightness [60]. A more systematic study of robustness of classification DCNNs to image perturbations over varying levels of corruption is proposed in [60]. This idea is extended to autonomous driving datasets [63], where robustness of object detection methods is evaluated. New datasets with challenging weather conditions, e.g., rain [54], fog [64], low light [65] are created to assess and improve robustness of visual perception models. Most approaches simply evaluate evolution of accuracy under such distribution shifts, but ignore other metrics, e.g., calibration that is essential for robustness (calibrated predictions facilitate thresholding for low-confidence predictions and detection of distribution shift). We evaluate our proposed method on multiple shifted datasets [60, 54, 63, 64] and show robustness improvements, beyond accuracy.

**DCNN Uncertainty Estimation.** Uncertainty estimation, i.e., knowing when a model does not “know” the answer, is a crucial functionality for robust DCNNs. Most DCNN approaches for uncertainty estimation are inspired from Bayesian Neural Networks [60, 65]. Deep Ensembles (DE) [40] train multiple instances of a DCNN with different random initializations, while MC-Dropout [19] mimics an ensemble through multiple forward passes with active Dropout [70] layers. In-between them, some methods generate ensembles with lower training cost (by analyzing weight trajectories during optimization [76, 60]) or with lower forward cost (by generating ensembles from lower dimensional weights [17, 79] or multiple network heads [43]). Other works prioritize computational efficiency to compute uncertainties from a single forward pass [6, 67, 62, 62, 66, 75], but become specialized to a single type of uncertainty [83, 88, 62], e.g., Out-Of-Distribution (OOD). DE methods are top-performers across benchmarks [76, 68]. Yet, their computational costs make them unfeasible for complex vision tasks, e.g., semantic segmentation. With Superpixel-mix we aim to attain most properties of ensembles, e.g., predictive uncertainty and calibration, in a cost-effective manner.

**Augmentation by mixing samples.** Initially seen as a mere heuristic to address overfitting, data augmentation is now an essential part of recent supervised [10, 83, 84], semi-supervised [9, 18, 68] and self-supervised learning methods [8, 23]. Mixing techniques, among the most powerful augmentation strategies, generate new “virtual” samples (and labels) from pairs of training samples. Mixup[84] interpolates two images, while Manifold Mixup [78] interpolates hidden activations. CutMix [18, 83] replace a random square inside an image with a patch from another image. Classmix [67] cuts and mixes object classes. Puzzle-Mix [69] and Co-Mixup [40] mix salient areas. For semantic segmentation, mixing by square blocks is agnostic to object boundaries and is likely to increase contextual bias as objects or object parts can be recognized via their context, i.e., learning shortcuts [22]. Superpixel-mix mitigates this by mixing within object boundaries.

## 3 Proposed method

### 3.1 Overview

In this paper, we propose a novel superpixel-mix method to generate new training data and leverage the results of this data augmentation technique in an existing teacher-student

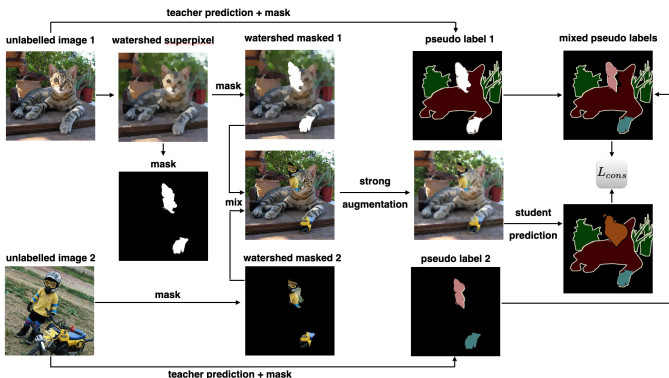


Figure 1: Consistency training with a student-teacher framework using unlabeled images (Task 2). To mix two unlabeled images, superpixels are randomly sampled to create a mixing mask. This mask is used to merge the two images and their pseudo-label outputs from the teacher network. A cross-entropy loss is applied to the student network to encourage consistency between the mixed pseudo-labels and the student network labels.

framework [174]. The combination of our mixing technique and the teacher-student framework forms an optimization component that serves as a consistency constraint in our DCNN training system for semantic segmentation. We conduct experiments for evaluating our trained DCNNs on both types of uncertainties: epistemic and aleatoric. At the end, we also assess our proposed approach in semi-supervised learning.

In general, our training process for all of our experiments comprises two steps that are optimized simultaneously: (1) supervised learning where we train the DCNNs using images with ground-truth labels, and (2) using teacher-student optimization with superpixel-mix data augmentation on images as a consistency constraint that does not use ground-truth labels. In the uncertainty experiments such as OOD in § 4.2, DCNNs’ bias studying in § 4.3, and aleatoric uncertainty in § 4.4, we use datasets that contain ground-truth labels for all images. We first train the DCNNs in fully supervised learning fashion as in step 1. We then remove all those labels and use only the images to optimize for the consistency constraint as in step 2. In the semi-supervised learning experiment in § 4.5, we use the training dataset that consists of two parts: labelled images and unlabelled images. We train the DCNN using the labelled data for step 1 and the unlabelled data for step 2.

**Step 1 - supervised learning with labelled images:** We use a standard pixel-wise cross-entropy loss denoted by  $\mathcal{L}_{sup}$  and apply it to all the labelled images. In addition, we use a weak data augmentation (WDA) that consists of *horizontal flipping* and/or *random cropping*.

**Step 2 - consistency constraint with unlabelled images:** We apply two transformations on an unlabelled image: one is WDA and the other is a strong data augmentation (SDA). Consistency training encourages predictions of the DCNNs to be consistent in the results of the two transformations. For SDA, we merge WDA and a superpixel mixing technique (see § 3.3). The consistency loss for optimizing this constraint is denoted as  $\mathcal{L}_{cons}$ .

For every experiment, we optimize the loss in the overall framework as the **joint loss**  $\mathcal{L} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{cons}$ , where  $\lambda$  is a weighting hyper-parameter and is set to 1.

## 3.2 Teacher-Student Framework

To learn from unlabeled images, we follow the teacher-student framework established in [174], where the teacher network produces pseudo-labels learned from the labeled data and the

student network is encouraged to be consistent with the teacher. Consistency is encouraged via cross-entropy loss between the two outputs. In our case, we encourage consistency between the mixed output labels of a teacher network corresponding to the two unlabelled inputs and the output label of a student network predicted from the input that resulted by mixing the two unlabelled images. We explain this approach in details in the following paragraph.

Let  $g_\phi$  represent the teacher network with weights  $\phi$  and  $f_\theta$  be the student network with weights  $\theta$ . For two unlabeled images  $x_u^1$  and  $x_u^2$ , we can use the teacher network to generate pseudo-labels  $y_u^1$  and  $y_u^2$ :  $y_u^1 = g_\phi(x_u^1)$  and  $y_u^2 = g_\phi(x_u^2)$ . Assume now we are given some mixing function  $\text{mix}$  with a mixing parameter  $m$ . Without any assumption on the mixing itself, we denote a mixed output label  $y_m$ , where  $y_m = \text{mix}(y_u^1, y_u^2, m)$ . For the same mixing parameter  $m$ , we can also mix the inputs  $x_u^1$  and  $x_u^2$ , *i.e.*  $x_m = \text{mix}(x_u^1, x_u^2, m)$ . Applying  $x_m$  to the student network, we expect  $f_\theta(x_m)$  to be the same as the mixed output  $y_m$ . This is enforced by minimizing the consistency loss  $\mathcal{L}_{\text{cons}} = \text{CE}(y_m, f_\theta(\text{mix}(x_u^1, x_u^2, m)))$ , where CE is the pixel-wise cross-entropy and  $y_m$  is the mixed pseudo-labels from the teacher.

During training, both the teacher and student networks evolve together. Similarly to [74], we update the teacher network weights  $\phi$  after each iteration with an Exponential Moving Average (EMA), *i.e.*  $\phi = \alpha\phi + (1 - \alpha)\theta$  where  $\alpha = 0.99$  is a momentum-like parameter.

### 3.3 Superpixel-mix for semantic segmentation

To mix two unlabeled images, we use masks generated by sampling superpixels. Superpixels are local clusters of visually similar pixels. Therefore, a group of pixels belonging to the same superpixel are likely to be in the same object or the same part of the object. There are several superpixel variants, including SEEDS [76], SLIC [8] or Watershed superpixels [43]. We opt to use Watershed superpixels as their boundaries retain more salient object edges [49]. We refer the reader to the Supplementary Material for details on how we use the watershed transformation to produce superpixels.

Given an unlabeled image  $x_u^1$ , we apply the watershed superpixel algorithm, which results in a set of  $n$  superpixels  $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ . A mixing mask  $m$  (which is a binary mask)<sup>3</sup> is created from a sampled subset of superpixels  $\mathcal{S}$ :  $m = \cup_{j \in \sigma(k, n)} S_j$  where  $\sigma(k, n)$  is a subset of size  $k$  of the  $n$  indices, and  $k$  is the number of superpixels we want to keep.

The mixing mask  $m$  defines the pixels in  $x_u^1$  which will be replaced by pixels from the unlabeled image  $x_u^2$  to form the mixed input  $x_m$ , *i.e.*  $x_m = \text{mix}(x_u^1, x_u^2, m) = (1 - m) \odot x_u^1 + m \odot x_u^2$ , where  $\odot$  is a pixel-wise multiplication. Superpixels are uniformly sampled given a fixed proportion of selected superpixels. Contrary to existing regularization techniques such as Cutout [14] or Cutmix [18], our superpixel mixing strategy enforces each set of selected pixels in the unlabeled image  $x_u^1$  to belong to the same object. However, the algorithm is allowed to select a set of superpixel clusters from different objects as illustrated in Figure 1.

We use 200 superpixels per image for all the evaluated datasets. Studies on the number and proportions of superpixels used in mixing are shown in Section 4.6 and the Supplement.

### 3.4 From empirical risk to teacher-student mixup

In this section, we show that the training loss of the teacher-student framework in combination with superpixel-mix data augmentation is bounded by the accuracy of the teacher network and the quality of the data augmentation. Let  $\mathcal{D} = \{(x_i, y_i)\}_i \sim \mathcal{P}$  be the labelled dataset which follows the joint distribution  $\mathcal{P}$  and  $l$  be a loss between the target  $y$  and the prediction  $f_\theta(x)$

<sup>3</sup>For simplicity, we overload the notation of the mixing parameter  $m$  simply as the mixing mask itself.

of the DCNN  $f_\theta$ . Typically, in deep learning, the objective is to learn  $\theta$  that minimizes the expected risk defined by:  $\mathbf{R}_\mathcal{P}(f_\theta) = \int l(f_\theta(x), y) d\mathcal{P}(x, y)$ . As we do not have access to the distribution  $\mathcal{P}$ , we optimize the loss function that is formed by the empirical risk on  $\mathcal{D}$ :

$$\hat{\mathbf{R}}_{\mathcal{P}_\delta}(f_\theta) = \frac{1}{n} \sum_{i=1}^n l(f_\theta(x_i), y_i) = \int l(f_\theta(x), y) d\mathcal{P}_\delta(x, y), \quad (1)$$

where the summation is converted back to the integral based on  $\mathcal{P}_\delta(x, y) = \frac{1}{n} \sum_{i=1}^n \delta(x = x_i, y = y_i)$ , as shown by [84].

Therefore, we optimize the parameters of the DCNN using the empirical risk. However, the representation of the discretized data is likely to be sparse, Zhang *et al.* [84] proposed to work with  $\mathcal{D}_{\text{mix}} = \{(x_{m,i}, y_{m,i})\}_i \sim \mathcal{P}_{X,Y}^{\text{mix}}$  where  $x_{m,i}$ , and  $y_{m,i}$  are the data of  $\mathcal{D}$  where a mixing procedure has been applied. The hypothesis in [84] is that the mixing procedure helps to better approximate the dataset distribution. Let  $\mathcal{P}_\delta^{\text{mix}}$  denote the discrete distribution of this augmented dataset. The risk to fit the teacher prediction on  $\mathcal{P}_\delta^{\text{mix}}$  can then be defined as:  $\hat{\mathbf{R}}_{\mathcal{P}_\delta^{\text{mix}}}(f_\theta, g_\phi) = \int l(f_\theta(x), g_\phi(x)) d\mathcal{P}_\delta^{\text{mix}}(x, y)$ . Therefore, our training loss for the overall framework is defined in detail as the following:  $\mathcal{L}(\theta) = \hat{\mathbf{R}}_{\mathcal{P}_\delta}(f_\theta) + \hat{\mathbf{R}}_{\mathcal{P}_\delta^{\text{mix}}}(f_\theta, g_\phi)$ . As the loss  $l$  is a norm that satisfies the triangle equality, we can prove that the training loss  $\mathcal{L}(\theta)$  is bounded by the following:

$$\mathcal{L}(\theta) \leq 2\mathbf{R}_\mathcal{P}(f_\theta) + M(\|\mathcal{P}_\delta^{\text{mix}} - \mathcal{P}\|_1 + \|\mathcal{P}_\delta - \mathcal{P}\|_1) + \hat{\mathbf{R}}_{\mathcal{P}_\delta^{\text{mix}}}(g_\phi), \quad (2)$$

where the four terms are linked to the true error, approximation error, mixing distribution error and teacher error, respectively. This implies that the quality of the DCNN is bounded by the accuracy of the teacher. It is also bounded by how much the mixing strategy can sample the true distribution of the dataset. Finally, the distribution of the training data with respect to the true data distribution also plays an important role. This finding can be applied to all teacher-student frameworks, such as those used in SSL, self supervised training, and domain adaptation. The detailed proof and analysis is given in the Supplementary Material.

$\|\mathcal{P}_\delta^{\text{mix}} - \mathcal{P}\|_1$  and  $\|\mathcal{P}_\delta - \mathcal{P}\|_1$  reflect the capacity of the two distributions  $\mathcal{P}_\delta$  and  $\mathcal{P}_\delta^{\text{mix}}$  to approximate the true dataset distribution. This bound allows us to control the variation of the risk. To reduce the risk, we can increase the number of training data or improve the quality of the data augmentation and so reduce  $\|\mathcal{P}_\delta^{\text{mix}} - \mathcal{P}\|_1$ . This motivates our research on data augmentation strategies to approximate the true distribution in a data-efficient way. We can also improve the quality of the teacher using EMA training that stabilizes the training loss.

### 3.5 Uncertainty and Deep Learning

Consider a joint distribution  $\mathcal{P}$  over input  $x$  and labels  $y$  over a set of labels  $\mathcal{Y}$ . When a DCNN performs inference, it predicts  $f_\theta = \mathcal{P}(y|x, \theta)$ , where  $\theta$  is optimized to minimize the loss over the training set  $\mathcal{D}$ . This likelihood typically suffers from two kinds of uncertainty [83, 88]. First is aleatoric uncertainty, linked to the unpredictability of the data acquisition process. During inference, instead of working with  $x$ , we may have access to  $x + n$ , where  $n$  represents noise on the input data. Second is epistemic uncertainty, linked to the lack of knowledge of the model, i.e., the weights  $\theta$  of the network. In addition, the epistemic uncertainty can be subdivided into two sub-types: one linked to the OOD [62] and the other one linked to networks' bias. Epistemic uncertainty models the uncertainty associated with limited sizes of training datasets. Most works focus only on the ability to detect OOD. In this paper, we

conduct experiments for all types of uncertainties: aleatoric (i.e., testing models on noisy data) and two sub-types of epistemic (i.e., OOD detection and models' bias).

## 4 Experiments

We conduct experiments on five datasets. First, we study network robustness to epistemic uncertainty experiments on StreetHazards [32]. The test set contains some object classes that are not available in the training set. The goal is to detect these out-of-distribution (OOD) classes. We also evaluate the performance of the DCNNs on an contextually unbiased dataset. Furthermore, we investigate the networks' robustness for the aleatoric uncertainty. To this end, we train a DCNN on Cityscapes [9] and evaluate the performances on Rainy [64] and Foggy Cityscapes [64]. Finally, we evaluate on the semi-supervised learning task on Cityscapes [9] and Pascal [17]. We implement the experiments using PyTorch (see Supplementary Material).

### 4.1 Evaluation criteria

The first criterion we use is the mIoU [36], reflecting the predictive performance of segmentation models. Second, similarly to [4] we use the negative log-likelihood (NLL), a proper scoring rule [24], which depends on the aleatoric uncertainty and can assess the degree of overfitting [25]. In addition, we use the expected calibration error (ECE) [25] that measures how the confidence score predicted by a DCNN is related to its accuracy. Finally, we use the AUPR, AUC, and the FPR-95-TPR defined in [31] that evaluate the quality of a DCNN to detected OOD data. With multiple metrics, we can get a clearer picture on the performance of the DCNNs with regards to accuracy, calibration error, failure rate, OOD detection. Even though it is difficult to achieve top performance on all metrics, we argue that it is more pragmatic and convincing to evaluate on multiple metrics [15, 53] than optimizing for a single metric, potentially at the expense of many others. For example, a DCNN with a low accuracy and a low confidence score is well-calibrated. Therefore, evaluating a DCNN on a single metric such as ECE or mIoU alone is not enough. We aim to have a good compromise between accuracy and calibration.

### 4.2 Epistemic uncertainty: Out-Of-Distribution (OOD) Detection

One cause of the epistemic uncertainty in deep learning is the limited training data that does not cover all possible object classes. The evaluation of this type of epistemic uncertainty is often linked to OOD detection. Therefore, this experiment is designed to evaluate the epistemic uncertainty using StreetHazards [32]. StreetHazards is a large-scale dataset that consists of different sets of synthetic images of street scenes. This dataset is composed of 5,125 images for training and 1,500 test images. The training dataset contains pixel-wise annotations for 13 classes. The test dataset comprises 13 training classes and 250 OOD classes, unseen in the training set, making it possible to test the robustness of the algorithm when facing a diversity of possible scenarios. For this experiment, we use DeepLabv3+ [7] with the experimental protocol from [22]. We use ResNet50 encoder [28]. For this experiment, we compare our algorithm to Deep Ensembles [21], BatchEnsemble [49], LP-BNN [12], TRADI [16], MIMO [27] achieving state-of-the-art results on epistemic uncertainty. We also compare our model with MCP which is the baseline DCNN where we consider the maximum probability class as a confidence score, and with Cutmix [18] strategy. The results in Table 1 show that our method is the only one to have the best results in three out of five measures. Running up, Deep Ensemble and LP-BNN achieve best results on one measure

Dataset	OOD method	mIoU $\uparrow$	AUC $\uparrow$	AUPR $\uparrow$	FPR-95-TPR $\downarrow$	ECE $\downarrow$	# Forward passes $\downarrow$
StreetHazards DeepLabv3+ ResNet50	Baseline (MCP) [14]	53.90%	0.8660	0.0691	0.3574	0.0652	1
	TRADI [14]	52.46%	0.8739	0.0693	0.3826	0.0633	4
	Cutmix [14]	56.06%	0.8764	0.0770	0.3236	0.0592	1
	MIMO [14]	55.44%	0.8738	0.0690	0.3266	0.0557	4
	BatchEnsemble [14]	56.16%	0.8817	0.0759	0.3285	0.0609	4
	LP-BNN [14]	54.50%	0.8833	0.0718	0.3261	<b>0.0520</b>	4
	Deep Ensembles [14]	55.59%	0.8794	<b>0.0832</b>	0.3029	0.0533	4
	Superpixel-mix (ours)	<b>56.39%</b>	<b>0.8891</b>	0.0778	<b>0.2962</b>	0.0567	1

Table 1: Comparative results on the OOD task for semantic segmentation. Results are averaged over three seeds.

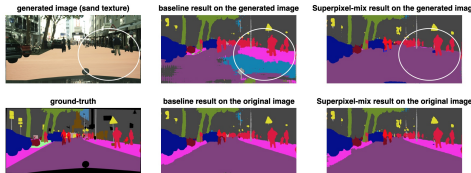


Figure 2: A qualitative example of the network bias study. When the road and pavements are replaced with sand texture, the baseline supervised network makes wrong segmentations. There is still pavement segmentation due to the association with people. Superpixel-mix produces better results without the wrong pavement segmentation, also provides a clean segmentation for the sand texture.

only. Moreover, our method achieves the best results faster than Deep Ensemble and LP-BNN, we only need one inference pass compared to 4 inference passes for the others.

### 4.3 Epistemic uncertainty : Unbiased experiment

The second aspect of epistemic uncertainty is related to network biased caused limited samples and diversity of scenarios, e.g., certain objects always co-occur in the training data. Here, we evaluate the epistemic uncertainty under the lens of model bias. In urban datasets such as Cityscapes, road and car pixels appear most of the time together, raising the question of co-occurrence dependency between objects. If two objects are dependent, then the network will likely fail when the car object is encountered in different context other than roads. Shetty *et al.* [14] study object dependency and suggest using GANs to remove one object by inpainting and training the network for the new contexts. Their results show that a network that is less biased towards the co-occurrence dependency yields better accuracy for segmentation.

To measure the bias of our superpixel-mix method, we create a new dataset dubbed Out-of-Context Cityscapes (OC-Cityscapes), by replacing roads in the validation data of Cityscapes with various textures such as water, sand, grass, etc. Example images are shown in the Supplementary Material. Studies in [14] show that DCNNs are biased towards texture. By replacing different textures for roads, we test the trained networks on these new context images and evaluate the bias level for each network.

In Table 2, we show the performances of the fully supervised network (baseline) and our network trained using the superpixel-mix method for the Cityscapes validation set and our generated dataset. The results are mIoU, ECE, and NLL scores averaged over 3 runs for classes that are not road. On our experimental dataset, the baseline’s performance drops by 21.97% while Superpixel-mix drops by 19.83% for the mIoU metric. The results also show that Superpixel-mix produces a less biased model for co-occurring objects with the best scores on mIoU and NLL measures. Superpixel-mix changes the image context while preserving object shapes, effectively regularizing the model to address shortcut learning, i.e., overfitting on the co-occurrence of objects and their typical contexts. For a visual example, see Figure 2.



Evaluation data	Cityscapes mIoU		OC-Cityscapes mIoU		Cityscapes NLL		OC-Cityscapes NLL		Cityscapes ECE		OC-Cityscapes ECE	
	mIoU	NLL	mIoU	NLL	NLL	NLL	ECE	ECE	ECE	ECE	ECE	ECE
Baseline (MCP) [10]	76.51%	0.1303	54.54%	-0.9456	-0.9456	-0.7565	0.1303	0.1303	0.1303	0.1303	0.2162	0.2162
Cutmix [11]	78.37%	0.1365	54.78%	-0.955	-0.955	-0.7435	0.1365	0.1365	0.1365	0.1365	0.2587	0.2587
MIMO [12]	77.13%	0.1398	55.87%	-0.9516	-0.9516	-0.7431	0.1398	0.1398	0.1398	0.1398	0.2359	0.2359
Deep Ensembles [13]	77.48%	0.1274	57.09%	-0.9469	-0.9469	-0.7613	0.1274	0.1274	0.1274	0.1274	0.1968	0.1968
Superpixel-mix (ours)	<b>78.99%</b>	0.1348	<b>59.16%</b>	<b>-0.9563</b>	<b>-0.9563</b>	<b>-0.7768</b>	0.1348	0.1348	0.1348	0.1348	0.2244	0.2244

Table 2: **Comparative results for network biases on OC-Cityscapes.** The results are segmentation mIoU, NLL and ECE, for classes that are not road. The baseline is the result from supervised training.

Evaluation data	Cityscapes			Rainy Cityscapes			Foggy Cityscapes			Cityscapes-C		
	mIoU ↑	ECE ↓	NLL ↓	mIoU ↑	ECE ↓	NLL ↓	mIoU ↑	ECE ↓	NLL ↓	mIoU ↑	ECE ↓	NLL ↓
Baseline (MCP) [10]	76.51%	0.1303	-0.9456	58.98%	0.1395	-0.8123	69.89%	0.1493	-0.9001	40.85%	0.2242	-0.7389
Cutmix [11]	78.37%	0.1365	-0.9550	61.86%	0.1559	-0.8200	73.57%	0.1484	-0.9289	39.16%	0.3064	-0.6865
MIMO [12]	77.13%	0.1398	-0.9516	59.27%	0.1436	-0.8135	70.24%	0.1425	-0.9014	40.73%	0.2350	-0.7313
BatchEnsemble [14]	77.99%	0.1129	-0.9472	60.29%	0.1436	-0.7820	72.19%	0.1425	-0.9132	40.93%	0.2270	-0.7082
LP-BNN [15]	77.39%	<b>0.1105</b>	-0.9464	60.71%	0.1338	-0.7891	72.39%	<b>0.1358</b>	-0.9131	<b>43.47%</b>	0.2085	-0.7282
Deep Ensembles [13]	77.48%	0.1274	-0.9469	59.52%	<b>0.1078</b>	-0.8205	71.43%	0.1407	-0.9070	43.40%	<b>0.1912</b>	-0.7509
Superpixel-mix (ours)	<b>78.99%</b>	0.1348	<b>-0.9563</b>	<b>61.87%</b>	0.1583	<b>-0.8207</b>	<b>74.39%</b>	0.1411	<b>-0.9266</b>	42.58%	0.2338	<b>-0.7513</b>

Table 3: Aleatoric uncertainty study on Cityscapes-C, Foggy Cityscapes [62] and Rainy Cityscapes [62].

## 4.4 Aleatoric uncertainty experiments

Aleatoric uncertainty is associated with unpredictability of the data acquisition process that causes various noises in the data. In the following experiments we evaluate the aleatoric uncertainty of DCNNs trained on normal images (e.g., normal weather images) when facing test images with various types of noise (e.g., rainy or foggy environments). In semantic segmentation, the DCNN must be robust to aleatoric uncertainty. To check that, we use the rainy [62] and foggy Cityscapes [62] datasets, which are built by adding rain or fog to the Cityscapes validation images. The goal is to evaluate the performance of DCNNs to resist these perturbations. We further generate an additional Cityscapes variant with images modified with different perturbations and intensities to mimic a distribution shift [60]. We apply the following perturbations: Gaussian noise, shot noise, impulse noise, defocus blur, frosted, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, pixelate, JPEG. For more information, please refer to [60]. We call this dataset Cityscapes-C.

To measure the robustness under aleatoric uncertainty, we compute ECE, mIoU and NLL scores averaged over 3 runs. Table 3 shows results close to the state of the art. DE reaches good results, yet this approach needs to train several DCNNs, so it is more time-consuming for training and inference. In the Supplementary Material, we report mIoU scores of different approaches for the different levels of noise. Overall, our experiments indicate that Superpixel-mix tends to be robust to high level of noise.

We note that our method does not achieve the top ECE scores across the various perturbations and weather conditions in these experiments. However, none of the considered strong baselines based on ensembles outperforms the others consistently either due to the difficulty and diversity of the considered test sets. Superpixel-mix achieves competitive ECE scores and is a top performer on mIoU and NLL.

## 4.5 Semi-Supervised Learning experiments

To evaluate the robustness of our method to missing annotation, we tested our approach on a semi-supervised learning task on two datasets: Cityscapes [9] and Pascal VOC 2012 [17]. We follow the common practice for this task from prior works [18, 65] and use DeepLab-V2 [9] model with ResNet101 [28] encoder pre-trained on ImageNet [10] and MS-COCO [66]. The weights of both the teacher and student models are initialized the same manner. We evaluate our method and compare with existing methods using three sets of labeled data: 1/30 (100 images), 1/8 (372 images) and 1/4 (744 images). Our results are reported as average mIoU of

Labeled samples	1/30 (100)	1/8 (372)	1/4 (744)
Adversarial [15]	-	58.80%	62.30%
s4GAN [15]	-	59.30%	61.90%
Cutout [15]	47.21% $\pm$ 1.74	57.72% $\pm$ 0.83	61.96% $\pm$ 0.99
Cutmix [15]	51.20% $\pm$ 2.29	60.34% $\pm$ 1.24	63.87% $\pm$ 0.71
Classmix [15]	54.07% $\pm$ 1.61	61.35% $\pm$ 0.62	63.63% $\pm$ 0.33
Classmix [15]	54.07% $\pm$ 1.61	61.35% $\pm$ 0.62	63.63% $\pm$ 0.33
Baseline(*)	43.84% $\pm$ 0.71	54.84% $\pm$ 1.14	60.08% $\pm$ 0.62
Superpixel-mix (ours)	<b>54.11% <math>\pm</math> 2.88</b> ( $\uparrow$ 7.27%)	<b>63.44% <math>\pm</math> 0.88</b> ( $\uparrow$ 8.60%)	<b>65.82% <math>\pm</math> 1.78</b> ( $\uparrow$ 5.74%)

Table 4: **Evaluation in the semi-supervised learning regime on Cityscapes. We report mIoU scores as mean  $\pm$  std.dev computed over 12 runs.** The ( $\uparrow$ ) shows the improvement of our methods over the baselines. (\*) The baselines are from [15] as we use a similar base procedure.

superpixels algorithm	Watershed [15]	SLIC [11]	Felzenszwalb [14]
mIoU	78.99%	78.89%	77.99%

Table 5: **Ablation study results over influence of different superpixel techniques. We report mIoU scores for semantic segmentation on Cityscapes.**

12 runs (4 times on each of 3 official splits) as well as the standard deviation. The results are shown in Table 4. We present results on Pascal dataset in the Supplementary Material.

## 4.6 Ablations and analysis

We perform various ablations to understand the influence of different hyper-parameters and choices in our algorithm. First, we vary the number of superpixels extracted per image from 20 to 1,000. The results (in Table 5, Supplement) show that our method achieves the highest mIoU on Cityscapes when the number of superpixels per image is from 100 to 200. Secondly, we study the proportion of chosen superpixels for mixing over the total number of superpixels per image. The proportion ranges from 0.1 to 0.9. We find that the results vary little across all the proportion values (Table 6, Supplement). The best mIoU is obtained at the proportion value of 0.6. Finally, we study the influence of different superpixel techniques to generate mixing masks: Watershed [15], SLIC [11], and Felzenszwalb [14]. The results in Table 5 show that on Cityscapes segmentation the performance of Superpixel-mix is relatively stable across superpixel methods, with Watershed yielding the best mIoU score.

## 5 Conclusions

Superpixel-mix data augmentation is a promising new training technique for semantic segmentation. This strategy for creating diverse data, combined with a teacher-student framework, leads to better accuracy and to more robust DCNNs. We are the first, to the best of our knowledge, to successfully apply the watershed algorithm in data augmentation. What sets our data augmentation technique apart from existing methods is the ability to preserve the global structure of images and the shapes of objects while creating image perturbations. We conduct various experiments with different types of uncertainty. The results show that our strategy achieves state-of-the-art robustness scores for epistemic uncertainty. For aleatoric uncertainty, our approach produces state-of-the-art results in Foggy Cityscapes and Rainy Cityscapes. In addition, our method needs just one forward pass.

Previous research in machine learning has established that creating more training data using data augmentation may improve the accuracy of the trained DCNNs substantially. Our work not only confirms that, but also provides evidence that some data augmentation methods, such as our Superpixel-mix, help to improve the robustness of DCNNs by reducing both epistemic and aleatoric uncertainty.

## Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (Grant 2020-AD011011970) and (Grant 2021-AD011011970R1).

## References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34, 2012.
- [2] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *ICLR*, 2020.
- [3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018.
- [4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.
- [5] Nicolas Brosse, Carlos Riquelme, Alice Martin, Sylvain Gelly, and Éric Moulines. On last-layer algorithms for classification: Decoupling representation from uncertainty estimation. *arXiv preprint arXiv:2001.08049*, 2020.
- [6] 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. *TPAMI*, 40, 2017.
- [7] 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.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [9] 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.
- [10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
- [11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- [12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, 2012.

- [13] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. In *CVPR*, 2018.
- [14] PF Felzenswalb and DP Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004.
- [15] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. In *arXiv*, 2019.
- [16] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. Tradi: Tracking deep neural network weight distributions. In *ECCV*, 2020.
- [17] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch. Encoding the latent posterior of bayesian neural networks for uncertainty quantification. *arXiv preprint arXiv:2012.02818*, 2020.
- [18] Geoff French, S. Laine, Timo Aila, M. Mackiewicz, and G. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.
- [19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.
- [20] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.
- [21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture: Increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.
- [22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.
- [23] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *CVPR*, 2020.
- [24] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *JASA*, 102, 2007.
- [25] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.
- [26] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *CVPRW*, 2020.
- [27] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *arXiv*, 2020.
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

- [29] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.
- [30] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.
- [31] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.
- [32] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. In *arXiv*, 2019.
- [33] Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.
- [34] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *CVPR*, 2019.
- [35] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018.
- [36] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2): 37–50, 1912.
- [37] Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. In *ICML*, 2020.
- [38] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.
- [39] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*, 2020.
- [40] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021.
- [41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.
- [42] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: a composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020.
- [43] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.
- [44] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *CVPR*, 2019.
- [45] Xin Li, Yiming Zhou, Zheng Pan, and Jiashi Feng. Partial order pruning: for best speed/accuracy trade-off in neural architecture search. In *CVPR*, 2019.

- [46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.
- [47] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017.
- [48] V. Machairas, E. Decencière, and T. Walter. Waterpixels: Superpixels based on the watershed transformation. In *ICIP*, 2014.
- [49] Vaia Machairas, Matthieu Faessel, David Cárdenas-Peña, Théodore Chabardes, Thomas Walter, and Etienne Decencière. Waterpixels. *TIP*, 24, 2015.
- [50] David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.
- [51] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS*, 2019.
- [52] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.
- [53] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *CoRR*, abs/1907.07484, 2019.
- [54] S. Mittal, M. Tatarchenko, and T. Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *TPAMI*, 2019.
- [55] Radford M Neal. Bayesian learning for neural networks. *PhD thesis, University of Toronto*, 1995.
- [56] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.
- [57] Viktor Olsson, Wilhelm Truhedden, Juliano Pinto, and Lennart Svensson. ClassMix: segmentation-based data augmentation for semi-supervised learning. *arXiv preprint arXiv:2007.07936*, 2020.
- [58] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.
- [59] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACCCS*, 2017.
- [60] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

- [61] Zachary Pezzementi, Trenton Tabor, Samuel Yim, Jonathan K Chang, Bill Drozd, David Guttendorf, Michael Wagner, and Philip Koopman. Putting image manipulations in context: Robustness testing for safe perception. In *SSRR*, 2018.
- [62] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *ICCV*, pages 2931–2940, 2019.
- [63] Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room. *CoRR*, abs/1808.03305, 2018.
- [64] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018.
- [65] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019.
- [66] Murat Sensoy, Lance M Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018.
- [67] R. Shetty, B. Schiele, and M. Fritz. Not using the car to see the sidewalk — quantifying and controlling the effects of context in classification and segmentation. In *CVPR*, 2019.
- [68] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.
- [69] M. Srivastava, T. Hashimoto, and P. Liang. Robustness to spurious correlations via human annotations. In *ICML*, 2020.
- [70] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15, 2014.
- [71] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.
- [72] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.
- [73] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.
- [74] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

- [75] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICMLR*, 2020.
- [76] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *ECCV*, 2012.
- [77] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.
- [78] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, 2019.
- [79] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2020.
- [80] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.
- [81] C Yu, J Wang, C Peng, C Gao, G Yu, and N Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.
- [82] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017.
- [83] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.
- [84] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2017.
- [85] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- [86] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnets for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.
- [87] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *CVPR*, 2019.