Supplementary Material for "On the Effects of Data Distortion on Model Analysis and Training"

Anonymous Author(s)
Affiliation
Address
email

A Experimental details

Throughout the paper, we use PreAct-ResNet18 [3] models, trained for 200 epochs with a batch size of 128. For the MSDA parameters we use the same values as Harris et al. [2]. All models are augmented with random crop and horizontal flip and are averaged across 5 runs. We optimise using SGD with 0.9 momentum, learning rate of 0.1 up until epoch 100 and 0.001 for the rest of the training. This is due to an incompatibility with newer versions of the PyTorch library of the official implementation of Harris et al. [2], which we use as a starting point for model training. However, the difference in learning rate schedule between our work and prior art does not affect our findings since we are not introducing a new method to be applied at training time. In our case, it is sufficient to show that the bias exists in at least one configuration. For the analysis we also used adapted code from [1] for patch-shuffling. The models were trained on either one of the following: Titan X Pascal, GeForce GTX 1080ti or Tesla V100. For the analyses, a GeForce GTX 1050 was also used. The average training time was less than two hours, with the exception of model trained on Tiny-ImageNet, which took around 10 hours to run.

Training models

The code for model training is largely based on the open-source official implementation of FMix, which also includes those of MixUp, CutOut, and CutMix. For the experiment where we use the reformulated objective to combine data sets, instead of mixing with a permutation of the batch, as it is done in the original implementation of the mixed-augmentations, we now draw a batch form the desired data set. To ensure a fair comparison, for the basic we also perform inter-batch mixing.

Evaluating robustness

For the CutOcclusion measurement, we modify open-source code to restrict the occluding patch to lie withing the the margins of the image to be occluded. This is to ensure that the mixing factor λ matches the true proportion of the occlusion. For iOcclusion, the implementation of Grad-CAM is again adapted from publicly available code. With both methods, we evaluate 5 instances of the same model and average over the results obtained.

The added computation time of iOcclusion over the regular CutOcclusion for a fixed occlusion fraction is that of performing Grad-CAM on train and test data, as well as evaluating on the latter. With a batch size of 128, this takes under half an hour.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Table B.1: Alternative DI index (%) for PreAct-ResNet18 on grid-shuffled images for four different types of models. Again, a bias can be noted for all considered data sets.

	basic	MixUp	FMix	CutMix
CIFAR-10	$3.52_{\pm 0.56}$	$3.31_{\pm 0.82}$	$0.76_{\pm 0.16}$	$0.43_{\pm 0.13}$
CIFAR-100	$1.40_{\pm 0.38}$	$1.09_{\pm 0.29}$	$0.38_{\pm 0.21}$	$0.16_{\pm 0.08}$
FashionMNIST	$1.56_{\pm 0.39}$	$3.57_{\pm 1.35}$	$1.65_{\pm 0.35}$	$0.82_{\pm 0.13}$
Tiny ImageNet	$3.01_{\pm 0.48}$	$2.24_{\pm 0.30}$	$2.34_{\pm 1.86}$	$11.45_{\pm 10.54}$
ImageNet	0.82	1.49	0.58	–

Table B.2: Shape and texture accuracy of BagNet9 models on the GST data set.

	Shape	Texture
basic	$11.29_{\pm 0.15}$	$18.90_{\pm 0.66}$
MixUp	$11.04_{\pm 0.29}$	$12.56_{\pm 1.26}$
FMix	$11.06_{\pm 0.48}$	$17.47_{\pm 1.74}$
CutMix	$10.76_{\pm 0.27}$	$20.28_{\pm 0.88}$

Table B.3: Di index (%) for alternative grid sizes.

		basic	MixUp	FMix	CutMix
CIFAR-10	2×2	$0.61_{\pm 0.24}$	$0.56_{\pm 0.33}$	$0.19_{\pm 0.14}$	$0.12_{\pm 0.06}$
	8×8	$6.41_{\pm 0.55}$	$6.95_{\pm 1.96}$	$2.75_{\pm 1.46}$	$1.41_{\pm 1.15}$
CIFAR-100	2×2	$1.03_{\pm 0.29}$	$0.46_{\pm 0.14}$	$0.21_{\pm 0.14}$	$0.12_{\pm 0.07}$
	8×8	$9.16_{\pm 6.15}$	$3.10_{\pm 4.59}$	$1.62_{\pm 0.89}$	$0.65_{\pm 0.50}$
Tiny ImageNet	8×8	$5.76_{\pm 6.61}$	$5.73_{\pm 3.82}$	$2.49_{\pm 1.38}$	$0.60_{\pm 0.69}$
	16×16	$44.01_{\pm 36.47}$	$14.06_{\pm 14.63}$	$11.94_{\pm 17.79}$	$1.86_{\pm 1.98}$
ImageNet	4×4	0.82	1.49	0.58	–
	64×64	4.89	41.16	12.77	

B Analysis of wrong predictions

B.1 Alternative index

Table B.1 the worst-case DI index where we replace $i_{c_{max}}$ in Equation 1 by the maximum increase across the runs. As per the original formulation, we note that the masking methods lead to models which are less sensitive to the artefacts resulted after patch-shuffling.

B.2 Varying the grid size

Table B.3 gives the results obtained when varying the number of image tiles to be randomly rearranged. We observe that data interference appears for different grid sizes.

B.3 Patch-shuffling

We look at the classes which have the highest increase in incorrect predictions and note that their shapes are characterised by strong horizontal and vertical edges. For example, on CIFAR-100, varying the grid size between 2×2, 4×4 and 8×8 gives "Lamp", "Bus" and "Table" as dominant c_{max} classes, while the model trained on Fashion MNIST with the standard procedure tends to predict grid-shuffled images as "Bag". Figure B.1 shows that on ImageNet, the basic model tends to wrongly identify the patch-shuffled images as belonging to class "Envelope".

B.4 CutOcclusion

In this section we experiment with alternative masking methods when computing CutOcclusion. We note that the bias exists when occluding with patches taken from images belonging to different data sets (Table B.4). Figure B.2 gives a visual account of the results obtained for CIFAR-10 when mix-patching. Note that for Fashion MNIST we use MNIST, for Tiny ImageNet we use ImageNet, while for CIFAR-10 we mix with CIFAR-100 and vice versa. Since ImageNet images are significantly larger than those of the other data sets, mixing would imply padding large areas, which would give results very similar to uniform patching. We also experiment with VGG models, where on CIFAR-10 the basic has a DI index of $0.80_{\pm 0.40}$ compared to $0.18_{\pm 0.11}$ of MixUp. We then use masks sampled

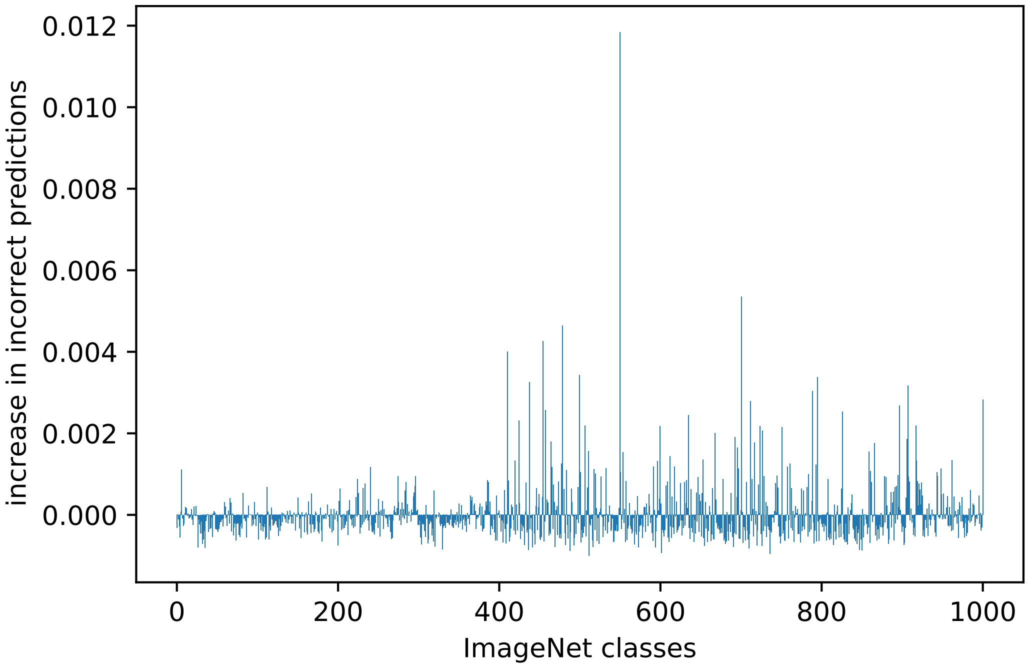


Figure B.1: Difference between the number of times a class was wrongly predicted when presented with regular ImageNet samples and patch-shuffled data.

Table B.4: DI index (%) for occluding with images from another data set.

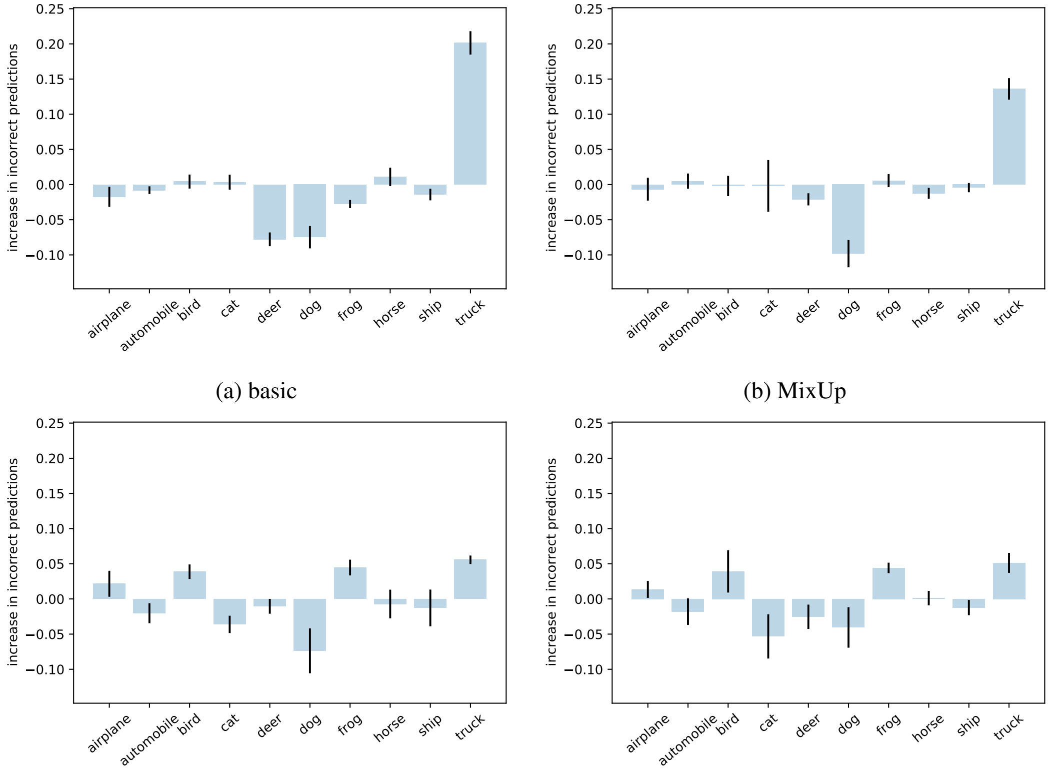
	basic	MixUp	FMix	CutMix
CIFAR-10	$0.18_{\pm 0.05}$	$0.39_{\pm 0.15}$	$0.12_{\pm 0.11}$	$0.08_{\pm 0.06}$
CIFAR-100	$0.48_{\pm 0.09}$	$0.61_{\pm 0.27}$	$0.90_{\pm 0.15}$	$1.25_{\pm 0.25}$
Fashion MNIST	$3.40_{\pm 0.29}$	$3.06_{\pm 1.07}$	$1.81_{\pm 0.55}$	$2.61_{\pm 0.80}$
Tiny ImageNet	$0.25_{\pm 0.12}$	$0.17_{\pm 0.04}$	$0.06_{\pm 0.03}$	$0.12_{\pm 0.04}$

from Fourier space (Table B.5) and note that even for these irregularly shaped distortions, we can identify a gap in most cases. The only exception is in the case of Fashion MNIST. It must be stressed that although all the models we experimented with presented Data Interference for this problem, this does not exclude the possibility of constructing a different model that is insensitive to this distortion. For example, we identify a gap for this problem when mix-masking (DI index of $4.09_{\pm 1.74}$ for the basic model as opposed to $1.87_{\pm 0.27}$ for a model trained on images that were masked out using FMix-like masks). Thus, when occluding with a particular shape we implicitly disfavour models in which learnt representations are related to the features introduced by that shape.

Figure B.4 also gives the results for CutOcclusion and iOcclusion for training with 3 random masks sampled from Fourier space.

Table B.5: DI index (%) for patching using masks sampled from Fourier space.

	basic	MixUp	FMix	CutMix
CIFAR-10	$2.08_{\pm 1.13}$	$1.79_{\pm 1.09}$	$1.32_{\pm 0.99}$	$4.21_{\pm 1.23}$
CIFAR-100	$4.06_{\pm 01.47}$	$3.11_{\pm 02.29}$	$9.90_{\pm 14.32}$	$2.89_{\pm 05.36}$
Fashion MNIST	$49.55_{\pm 20.45}$	$40.69_{\pm 21.63}$	$27.87_{\pm 17.57}$	$61.04_{\pm 17.92}$
Tiny ImageNet	$4.37_{\pm 0.85}$	$6.95_{\pm 1.84}$	$3.60_{\pm 1.73}$	$5.92_{\pm 4.38}$
ImageNet	3.27	2.24	6.08	–



(a) basic

(b) MixUp

(c) CutMix

(d) FMix

Figure B.2: Difference between wrongly predicted classes when testing on original data versus CutMix images. The evaluated models from left to right, top to bottom are trained on CIFAR-10 with: no mixed-data augmentation (basic), MixUp, CutMix, and FMix.

B.4.1 BagNet shape and texture accuracy

We evaluate on the GST data set BagNet9 models trained on Tiny ImageNet and present the results in Table B.2. Despite the basic model displaying a bias towards predicting one of the classes when presented with patch-shuffled images (see Figure B.5), once again it does not have a lower texture or higher shape bias than the masked-based augmentations.

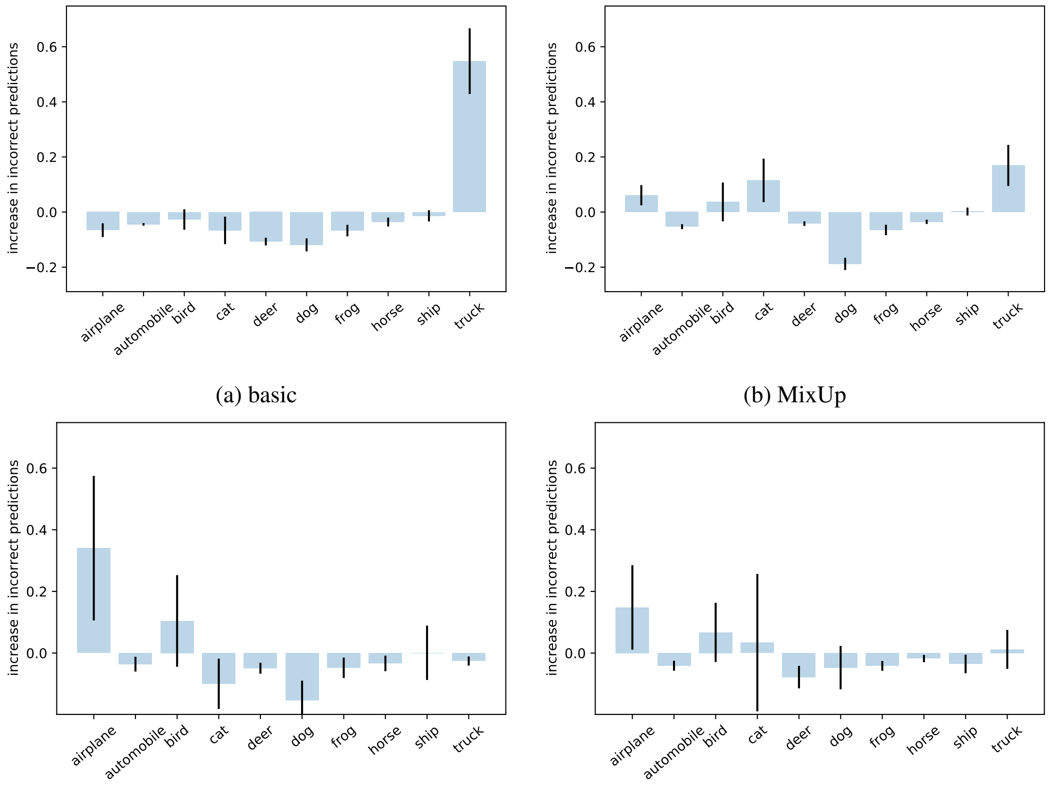
C Further results on iOcclusion experiments

C.1 Alternative CutOcclusion

Table C.1 gives the DI index when forcing the occluding patch to lie within image boundaries. Note that in the case of Tiny ImageNet the bias is more visibly present for larger occluders. As such, uniformly sampling the patch size from the interval [0.3, 1] results in a DI index of $13.46_{\pm 5.74}$ for the basic model, while the level of data interference from MixUp is only $4.75_{\pm 1.93}$. However, this does not change the conclusions of our experiments since as mentioned in the main paper, robustness studies are usually carried out with large occluder sizes.

C.2 Occluding with images from another data set

Since CutOcclusion does not account for the bias introduced by the occluding method, it is expected that changing the patch to a non-uniform one would greatly affect the results. For CIFAR-10 models, Figure C.1 presents the results of occluding with CIFAR-100 images. iOcclusion better rules out the specifics of the occluding patch, its uniform version giving similar results to the non-uniform one, whereas CutOcclusion pushes everything together.



(a) basic

(b) MixUp

(c) CutMix

(d) FMix

Figure B.3: Difference between wrongly predicted classes when testing on original data versus CutOut images. The evaluated models from left to right, top to bottom are trained on CIFAR-10 with: no mixed-data augmentation (basic), MixUp, CutMix, and FMix.

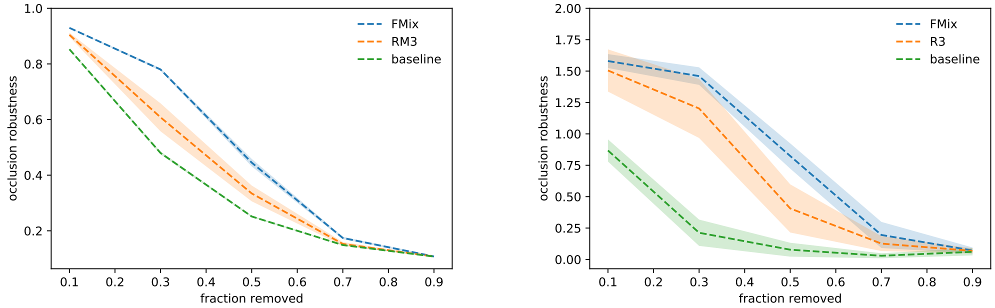


Figure B.4: CutOcclusion (left) and iOcclusion (right). Note that there is a difference in scale and the two should not be directly compared. We are rather interested in how the methods situate the different augmentation with respect to each other. It is important to notice that when measuring the robustness with CutOcclusion, RM3 appears significantly less robust than FMix due to its sensitivity to patching with rectangles. On the other hand, iOcclusion highlights the robustness specific to FMix.

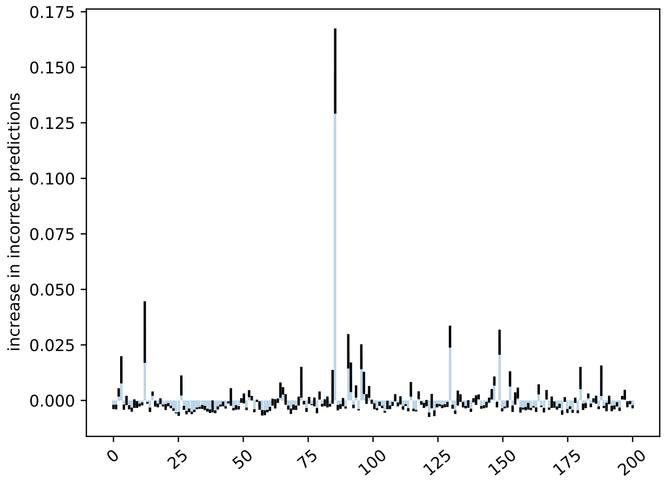


Figure B.5: Difference between wrongly predicted classes when testing on original Tiny ImageNet data versus CutOut images.

Table C.1: DI index (%) for sampling occluder size from a uniform distribution when the patch is restricted to lying within image boundaries.

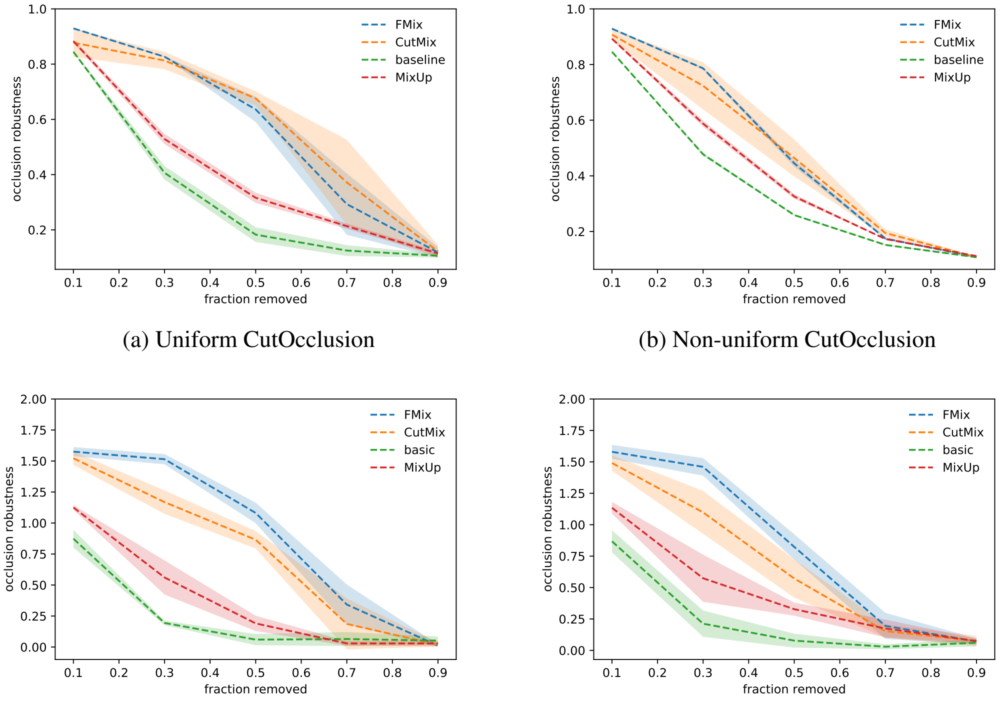
	basic	MixUp	FMix	CutMix
CIFAR-10	$5.88_{\pm 1.82}$	$0.76_{\pm 0.69}$	$1.30_{\pm 1.27}$	$3.68_{\pm 3.66}$
CIFAR-100	$29.71_{\pm 10.19}$	$6.80_{\pm 7.55}$	$6.08_{\pm 6.28}$	$13.19_{\pm 25.85}$
Fashion MNIST	$0.67_{\pm 0.38}$	$3.51_{\pm 1.38}$	$1.87_{\pm 3.33}$	$1.78_{\pm 2.93}$
Tiny ImageNet	$15.25_{\pm 4.84}$	$6.38_{\pm 4.03}$	$5.87_{\pm 6.07}$	$13.97_{\pm 24.02}$
ImageNet	9.93	28.72	11.52	−

C.3 Randomising labels

To assess the sensitivity of CutOcclusion and iOcclusion to the overall performance of the model, we also experiment with randomising all the labels of the CIFAR-10 data set. When evaluated on the unaugmented training data, all the basic models achieve 100% accuracy, while the FMix models reach $99.99_{\pm 0.01}$. Since all labels are corrupted, the accuracy on the test set before and after occlusion is no greater than random. However, the robustness of the augmentation-trained model can be seen on the training data, as captured by our metric (See Figure C.2). On the other hand, CutOcclusion makes no distinction between learning with regular and augmented data (Table C.2). Despite being such a peculiar case, it shows the comprehensiveness gained by accounting for the degradation on test data in relation to that on train.

C.4 Approximating iOcclusion

As alternative methods for computing iOcclusion we experiment both with masks sampled from Fourier space and randomly positioned square patches. Although using this type of random masking methods for computing $\mathcal{D}^i_{train}$ and $\mathcal{D}^i_{test}$ in Equation (1) gives less precise results, it has the advantage of incurring less computation and can be used for rapid model analysis. For assessing a model across 5 runs for 6 different levels of occlusion, this method leads to a carbon footprint of 0.05 kgCO$_2$eq as opposed to 1.04 using Grad-CAM. In Figure C.3 we present the results obtained with these alternatives. We expect both methods to provide overoptimistic results for small patches, while Fourier sampling is expected to give more truthful scores as the size of the patch increases. On the other hand, the contiguity of CutOut-based occlusion comes at the cost of not determining the robustness to multiple simultaneous occluders. This seems to play a role especially in the case of CutMix augmentation. Indeed, when superimposing a rectangular patch, it is difficult to differentiate



(a) Uniform CutOcclusion

(b) Non-uniform CutOcclusion

(c) Uniform iOcclusion

(d) Non-uniform iOcclusion

Figure C.1: Comparison of metric sensitivity to textured occlusion. Uniform occlusion refers to superimposing uniform patches over CIFAR-10 images, while nonuniform refers to superimposing part of CIFAR-100 samples. Nonuniform CutOcclusion provides significantly different results to its regular counterpart.

Table C.2: Robustness to occluding with patches covering 50% of each image. The models are trained with and without masking augmentation on data with randomised labels. CutOcclusion makes no difference between regular and augmented training.

	basic random	FMix random	FMix clean
CutOcclusion	$10.24_{\pm0.27}$	$9.78_{\pm0.18}$	$63.63_{\pm4.54}$
iOcclusion	$14.63_{\pm1.12}$	$47.94_{\pm19.84}$	$82.36_{\pm10.06}$

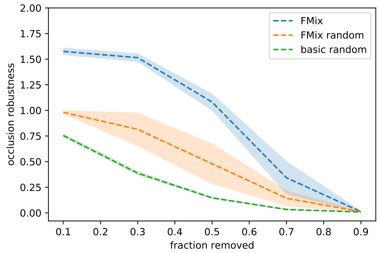
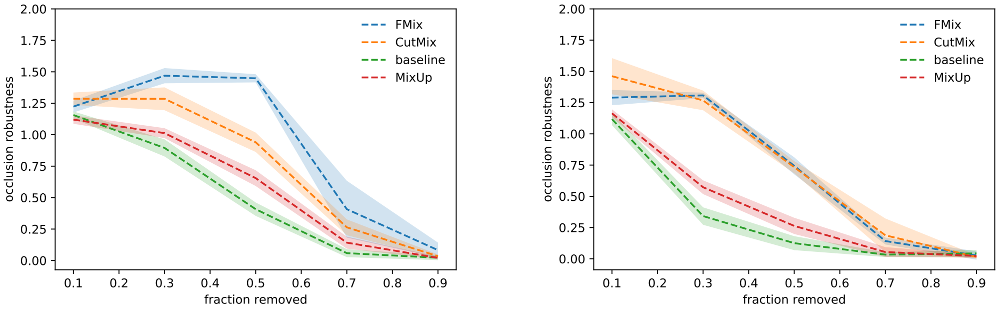


Figure C.2: iOcclusion results for training with clean and corrupted labels for basic and FMix augmentation.

CutMix from FMix-trained models. To confirm that this is caused by the granularity of the occluders and not the shape, we also experiment with occluding using multiple rectangular patches. We split the images in a 4×4 grid and occlude i% of the tiles, obtaining results that are more similar to those obtained when occluding with Fourier-sample patches. Thus, while significantly noisier, using randomly positioned occluders can provide an alternative for computing iOcclusion given that one takes into account the number of occluders.

D Removing the dominant class

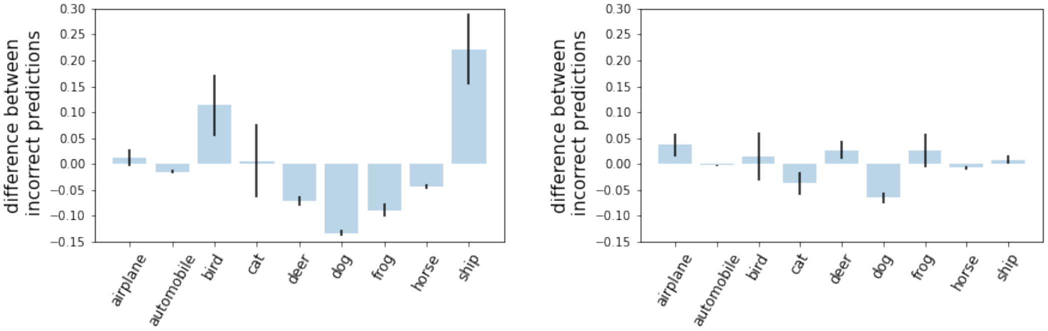
We remove the 10th class from the CIFAR-10 data set and retrain on the remaining classes. In the main paper we give the results for occluding with non-uniform patches. When using black patches to



(a) Fourier-sampled patches

(b) Rectangular patches

Figure C.3: Approximating iOcclusion with random masking can provide a first intuition, but is significantly noisier than using a saliency method.



(a) Fourier-sampled patches

(b) Rectangular patches

Figure D.1: Difference in incorrect predictions for the basic (left) and CutOut(right) models.

obstruct images, we again identify a gap, but this time with respect to a CutOut-trained model (see Figure D.1). The basic model has a DI index of $1.23_{\pm 0.72}$, while CutOut $0.13_{\pm 0.10}$. Thus, in both cases a model that is less affected by the artefacts than the basic model can be found. Thus, when measuring CutOcclusion, the basic model will be disadvantaged.

References

[1] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[2] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Understanding and enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.