Top-GAP: Integrating Size Priors in CNNs for more Interpretability, Robustness, and Bias Mitigation

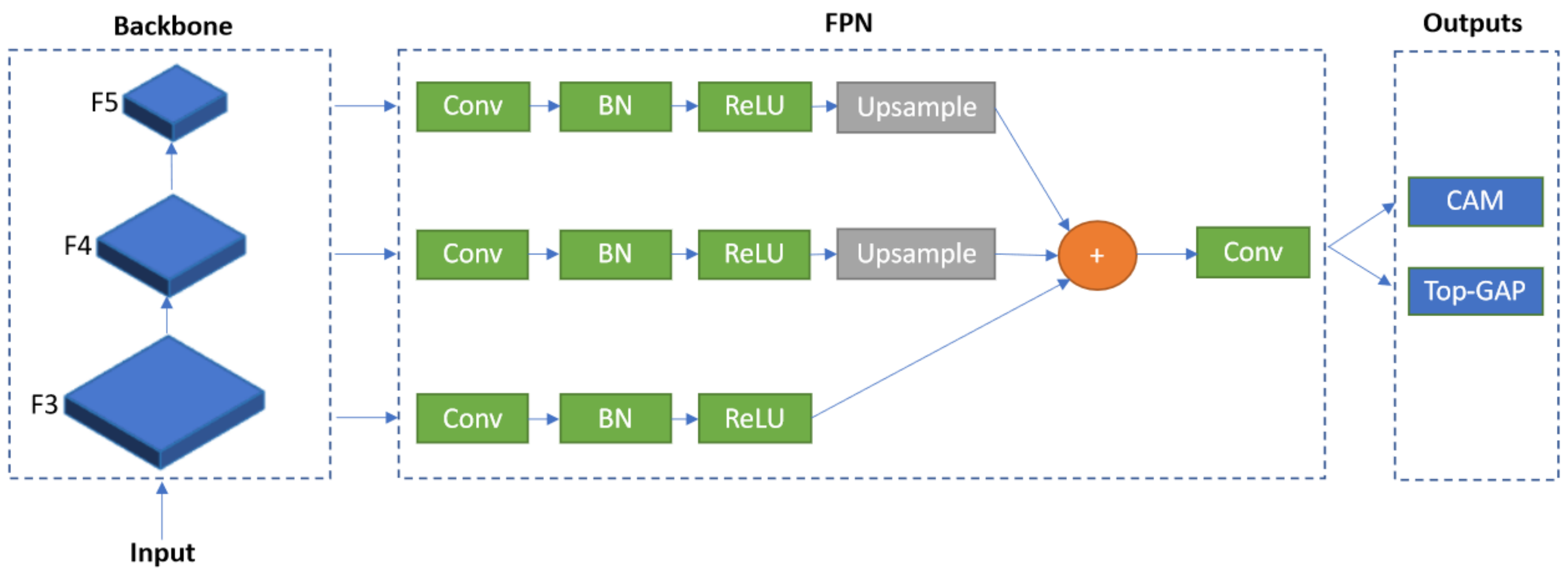
Lars Nieradzik[1]**,** Henrike Stephani[1], Janis Keuper[2]

We introduce Top-GAP, a novel regularization technique that enhances the explainability and robustness of CNNs. By constraining the spatial size of the learned feature representations, our method forces the network to focus on the most salient image regions, effectively reducing background influence.

Method

Our approach consists of a modification of the standard CNN architecture, a top-k pooling and a sparseness loss.

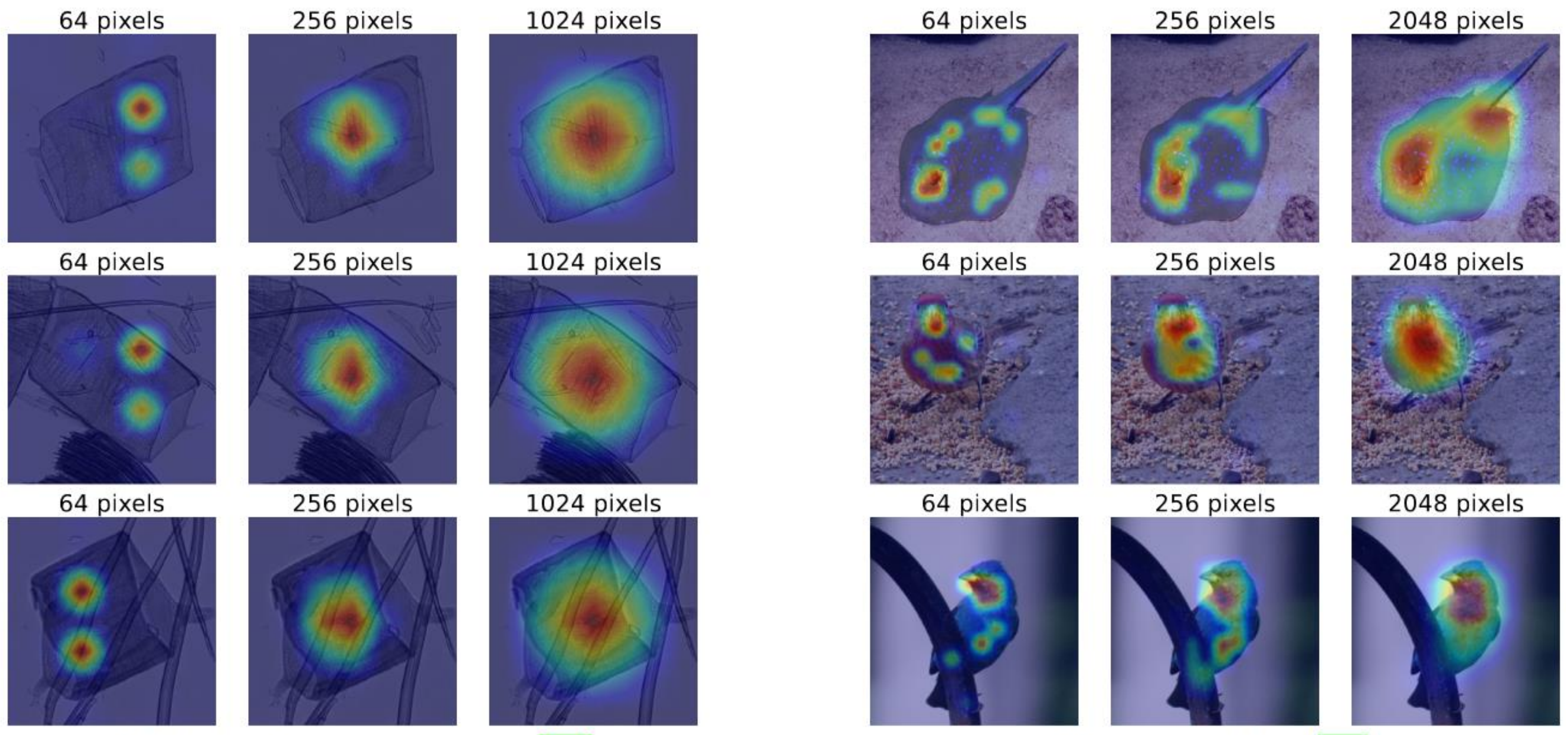
CNN structure



We add skip connections from previous layers to increase the resolution of the CAM. The output of the CAM is 56 x 56 x C for a 224x224 image, where C are the classes. We note that this does not always lead to an increase of the number of parameters.

Top-GAP and sparseness loss

Instead of averaging all pixels, we force the network to only use the most important k locations in the feature map. The „importance“ stems from an additional sparsity loss that forces the network to output an empty feature map. Part of the loss tries to increase k locations, while another part tries to set all of them to zero.



(a) Wood identification [32] **(b)** ImageNet [11]

Results

We systematically analyze the influence by evaluating the behavior with respect to receptive field, adversarial attacks, and segmentation overlap.

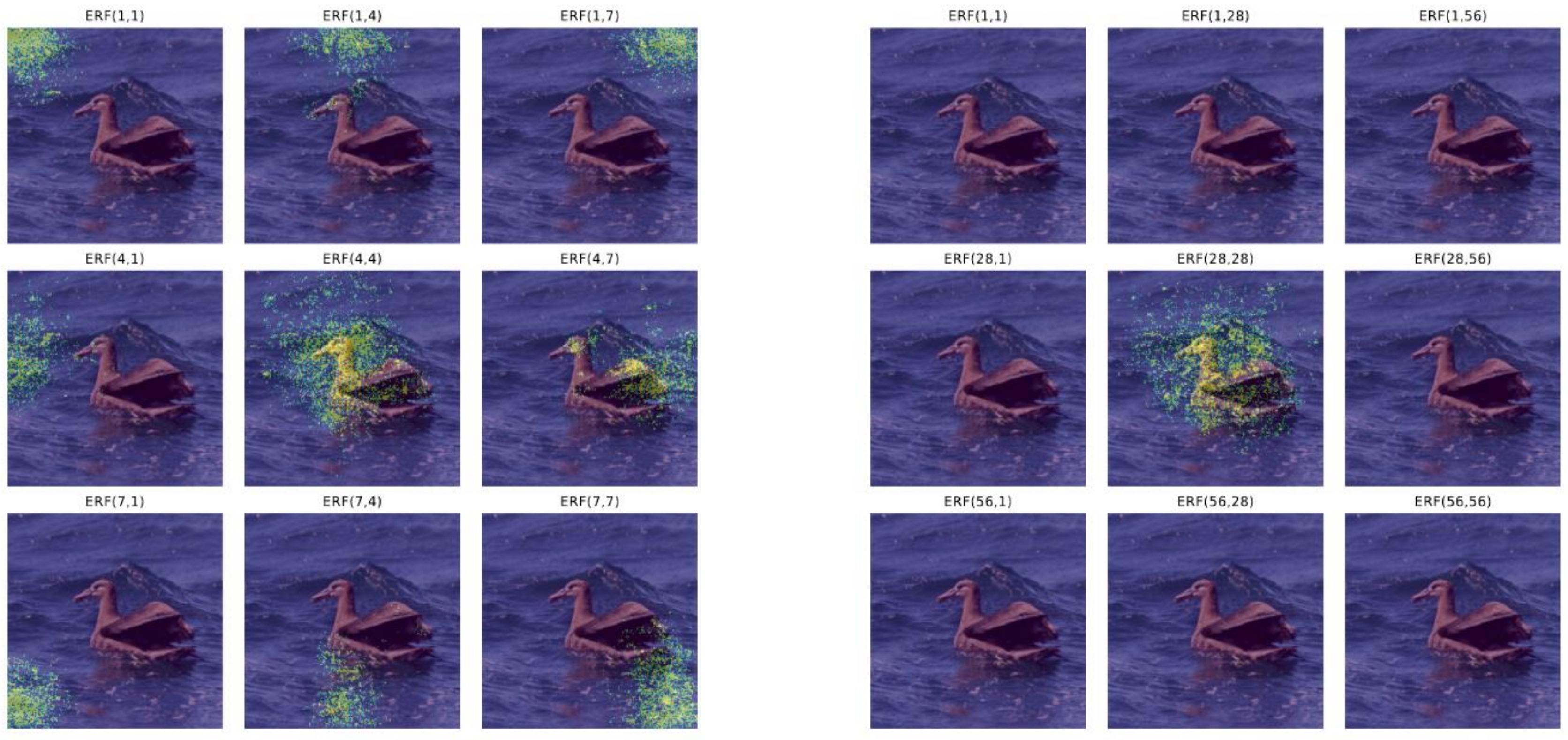
Background is less susceptible to adversarial attacks

The following table shows the results for CIFAR-10. We also found improvements for other datasets such as ImageNet.

Method	Arch	PGD^{20} ↑	PGD^{50} ↑	Square ↑	Clean ↑
Standard	PRN18	0.0	0.0	0.0	0.945
Top-GAP (ours)	PRN18	0.517	0.313	0.343	**0.951**
FGSM-AT [2]	PRN18	-	**0.476**	-	0.81
SAT [36]	RN50	**0.552**	-	-	0.849

Gradient of object pixels becomes more important

To measure the influence of the background using our model, we measure the absolute change of the output with all input pixels. We call our metric„ERF distance“, where ERF is the abbreviation for the effective receptive field. The following plots visualize the effective receptive field, comparing our method against a standard ResNet-18 at different coordinates.



(a) Standard ResNet-18 **(b)** ResNet-18 with our approach

Increased interpretability due to pixel constraints

We measure the Intersection-Over-Union (IOU) with the ground truth segmentation masks and compare our method against Grad-CAM (GC) and Recipro-CAM (RC).

Dataset	Arch	IOU (GC) ↑	IOU (RC) ↑	IOU (ours) ↑
COCO	EN	0.309	0.245	**0.348**
	CN	0.103	0.268	**0.361**
	RN	0.371	0.359	**0.391**
CUB	EN	0.323	0.337	**0.414**
	CN	0.125	0.279	**0.389**
	RN	0.268	0.34	**0.435**

Apart from the segmentation performance, we also have found increases in accuracy on Waterbirds, ImagetNet-Sketch and ImageNet-C (left table). These increases come without a decrease in regular accuracy (right table).

Dataset	Arch	Acc ↑	Acc ↑ (ours)
CUB → Waterbirds	EN	0.521	**0.564**
	CN	0.722	**0.737**
	RN	0.468	**0.520**
ImageNet → Sketch	VG	0.179	**0.200**
	RN	0.206	**0.236**
ImageNet → ImageNet-C	VG	0.494	**0.498**
	RN	0.513	**0.535**

Dataset	Arch	Accuracy ↑	Accuracy (ours) ↑
COCO	EN	0.801 ± 0.009	**0.803** ± 0.006
	CN	0.939 ± 0.006	**0.940** ± 0.005
	RN	0.853 ± 0.004	**0.868** ± 0.005
Wood	EN	0.672 ± 0.037	**0.681** ± 0.041
	CN	0.721 ± 0.030	**0.724** ± 0.033
Oxford	EN	0.854 ± 0.008	**0.863** ± 0.010
	RN	0.861 ± 0.007	**0.862** ± 0.007
CUB	EN	0.76 ± 0.01	**0.77** ± 0.005
	RN	**0.69** ± 0.014	0.685 ± 0.006
	CN	**0.862** ± 0.007	0.854 ± 0.005
ImageNet	VG	**0.704**	0.699
	RN	**0.698**	0.697

An improvement in accuracy can be observed for almost all datasets. Our proposed method works for any CNN architecture and comes „without cost“: no more GPU resources are needed and the training time stays the same.

This is unlike many methods in both the adversarial robustness and standard classification literature, which make use of specialized training techniques and more data to improve the accuracy.

Our approach provides an efficient form of network regularization that incorporates human knowledge into the training process. This is particularly useful for applications that have a single centered object in the image, such as in the biomedical field.

Contact

Lars Nieradzik
Fraunhofer ITWM
Department Image Processing
lars.nieradzik@itwm.fraunhofer.de

[1] Image Processing department, Fraunhofer ITWM
[2] Institute for Machine Learning and Analysis, Offenburg University

Gefördert durch:
Bundesministerium
für Ernährung
und Landwirtschaft
aufgrund eines Beschlusses
des Deutschen Bundestages
THÜNEN