Supplementary - Deflecting Adversarial Attacks with Pixel Deflection

Results on various classifiers

Original classification accuracy of each classifier on selected 1000 images is reported in the table. However, we omit the images that were originally incorrectly classified, thus the accuracy of clean images without defense is always 100%. Weights for each classifier were obtained from Tensorflow GitHub repository [1].

Model	$\lvert L_2 \rvert$	**No Defense**	**With Defense**	
			Single	Ens-10
ResNet-50, original classification 76%				
Clean	0.00	100	98.3	**98.9**
FGSM	0.05	20.0	79.9	**81.5**
IGSM	0.03	14.1	83.7	**83.7**
DFool	0.02	26.3	86.3	**90.3**
JSMA	0.02	25.5	91.5	**97.0**
LBFGS	0.02	12.1	88.0	**91.6**
C&W	0.04	04.8	92.7	**98.0**
VGG-19, original classification 71%				
Clean	0.00	100	99.8	**99.8**
FGSM	0.05	12.2	79.3	**81.3**
IGSM	0.04	9.79	79.2	**81.6**
DFool	0.01	23.7	83.9	**91.6**
JSMA	0.01	29.1	95.8	**98.5**
LBFGS	0.03	13.8	83.0	**93.9**
C&W	0.04	0.00	93.1	**97.6**
Inception-v3, original classification 78%				
Clean	0.00	100	98.1	**98.5**
FGSM	0.05	22.1	85.8	**87.1**
IGSM	0.04	15.5	**89.7**	89.1
DFool	0.02	27.2	82.6	**85.3**
JSMA	0.02	24.2	93.7	**98.6**
LBFGS	0.02	12.5	87.1	**91.0**
C&W	0.04	07.1	93.9	**98.5**

Table 1: Params: $\sigma = 0.04$, Window=10, Deflections=100
Top-1 accuracy on applying pixel deflection and wavelet denoising across various attack models.

[1] https://github.com/tensorflow/models/tree/master/research/slim#Pretrained

Comparison of small and large perturbations

Model	$\|L_2\|$	No Defense	With Defense	
			Single	Ens-10
Clean	0.00	100	98.3	**98.9**
FGSM	0.05	20.0	79.9	**81.5**
IGSM	0.03	14.1	83.7	**83.7**
DFool	0.02	26.3	86.3	**90.3**
JSMA	0.02	25.5	91.5	**97.0**
LBFGS	0.02	12.1	88.0	**91.6**
C&W	0.04	04.8	92.7	**98.0**
Large perturbations				
FGSM	0.12	11.1	61.5	**70.4**
IGSM	0.09	11.1	62.5	**72.5**
DFool	0.08	08.0	82.4	**88.9**
JSMA	0.05	22.1	88.9	**92.1**
LBFGS	0.04	12.1	77.0	**89.0**

Table 2: Params: $\sigma = 0.04$, Window=10, Deflections=100
Top-1 accuracy on applying pixel deflection and wavelet denoising across various attack models. We evaluate non-efficient attacks at larger $|L_P|$ which leave visible perturbations to show the robustness of our model.

Comparison of various shrinkage

Model	Hard	VISU	SURE	Bayes
Clean	39.5	96.1	92.1	**98.9**
FGSM	35.9	63.8	79.7	**81.5**
IGSM	42.5	67.8	81.1	**83.7**
DFool	37.2	78.4	87.7	**90.3**
JSMA	39.9	93.0	93.0	**97.0**
LBFGS	37.2	81.1	90.4	**91.6**
C&W	36.8	93.4	92.8	**98.0**

Table 3: Params: $\sigma = 0.04$, Window=10, Deflections=100
Comparison of various thresholding techniques, after application of pixel deflection.

In Table 3 we present a comparison of various shrinkage methods on wavelet coefficients after pixel deflection. All the results reported are for applying the given thresholding after pixel deflection. BayesShrink, which learns separate Gaussian parameters for each coefficient, does better than other soft-thresholding techniques. VisuShrink is a faster technique as it uses a universal threshold but that limits its applicability on some images. SUREShrink has been shown to perform well with compression but as evident, in our results, it is less well suited to denoising.

Ablation studies of various parameters

Attack	$\lvert L_2$	**No Defense**	**With Defense**			
Window=10, Deflections $\longrightarrow$			**10**	**100**	**1K**	**10K**
Clean	0.00	100	**98.4**	98.1	94.7	80.3
FGSM	0.04	19.2	75.7	**79.7**	71.7	69.1
IGSM	0.03	13.8	78.4	**81.7**	75.2	71.2
DFool	0.02	25.0	83.7	**87.7**	81.0	77.0
JSMA	0.02	25.9	91.7	**93.0**	87.7	67.7
LBFGS	0.02	11.6	85.0	**90.3**	82.4	73.0
C&W	0.04	05.2	89.4	**93.1**	86.8	69.7

Table 4: Top-1 accuracy with different deflections.

Attack	**L2**	**No Defense**	**With Defense**			
Deflections=100, Window $\longrightarrow$			**5**	**10**	**50**	**100**
Clean	0.00	100	**98.6**	98.1	96.4	94.4
FGSM	0.04	19.2	**79.7**	**79.7**	78.4	76.7
IGSM	0.03	13.8	81.0	**81.7**	79.7	78.4
DFool	0.02	25.0	86.4	**87.7**	87.7	85.0
JSMA	0.02	25.9	92.3	**93.0**	91.7	90.3
LBFGS	0.02	11.6	89.4	**90.3**	89.0	88.1
C&W	0.04	05.2	91.8	**93.1**	90.5	89.2

Table 5: Top-1 accuracy with different window sizes.

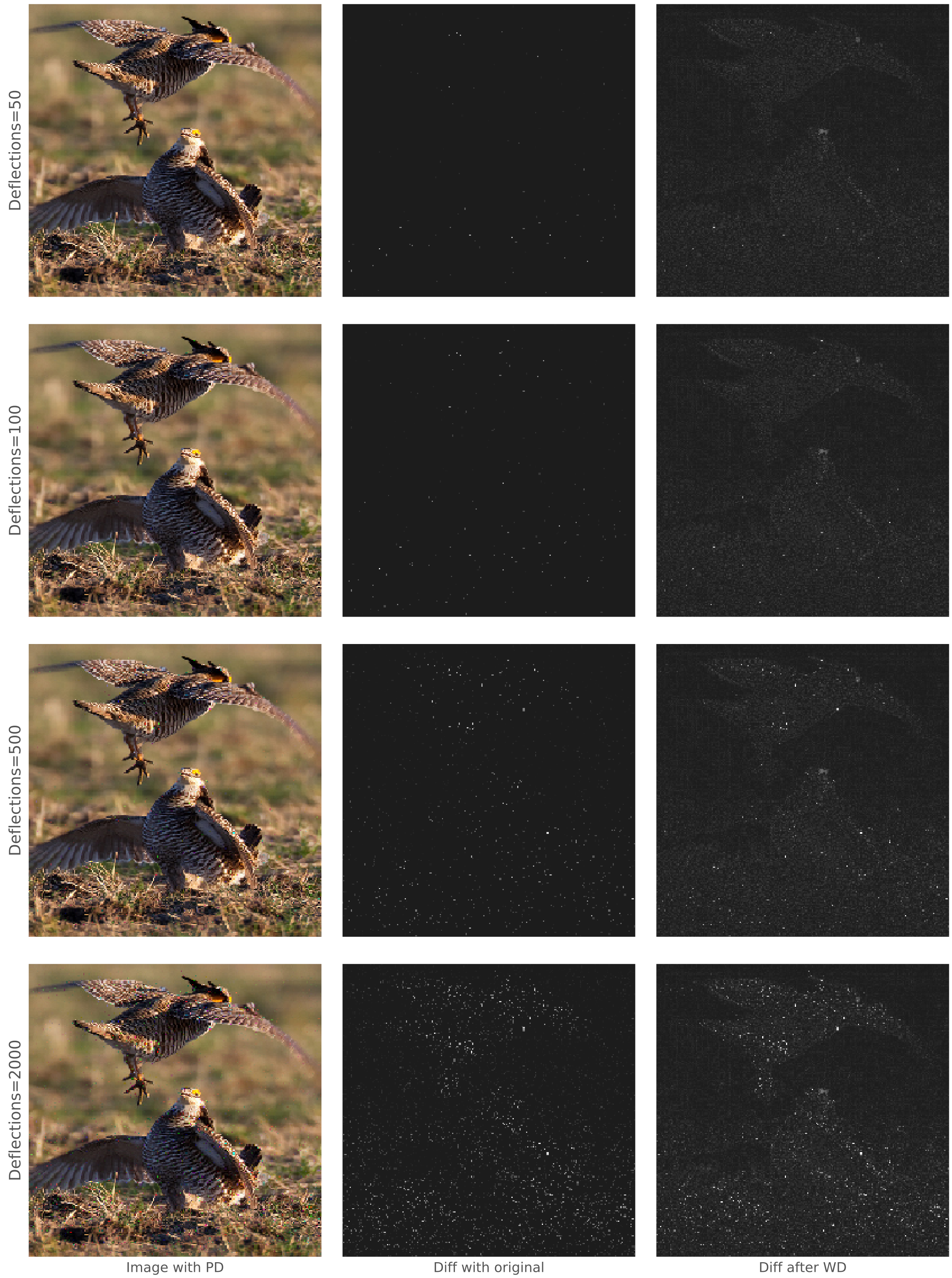


Figure 1: This is same image as Figure 1 on the paper but with a better color scheme. Impact of Pixel Deflection (PD) on a natural image and subsequent denoising using wavelet transform (WD). Left: Image with given number of pixels deflected. Middle: Difference between clean image and deflected image. Right: Difference between clean image and deflected image after denoising. Enlarge to see details.

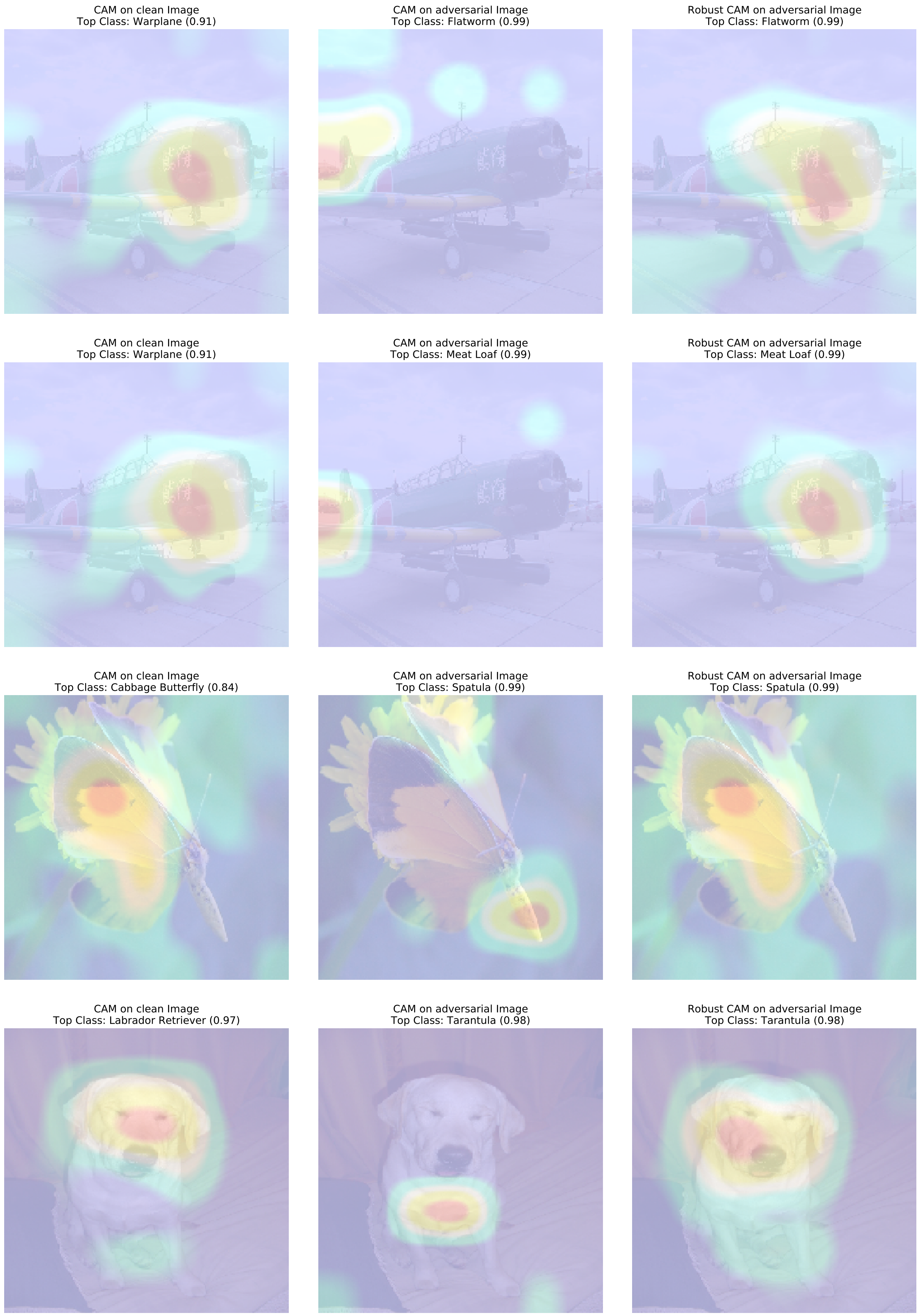
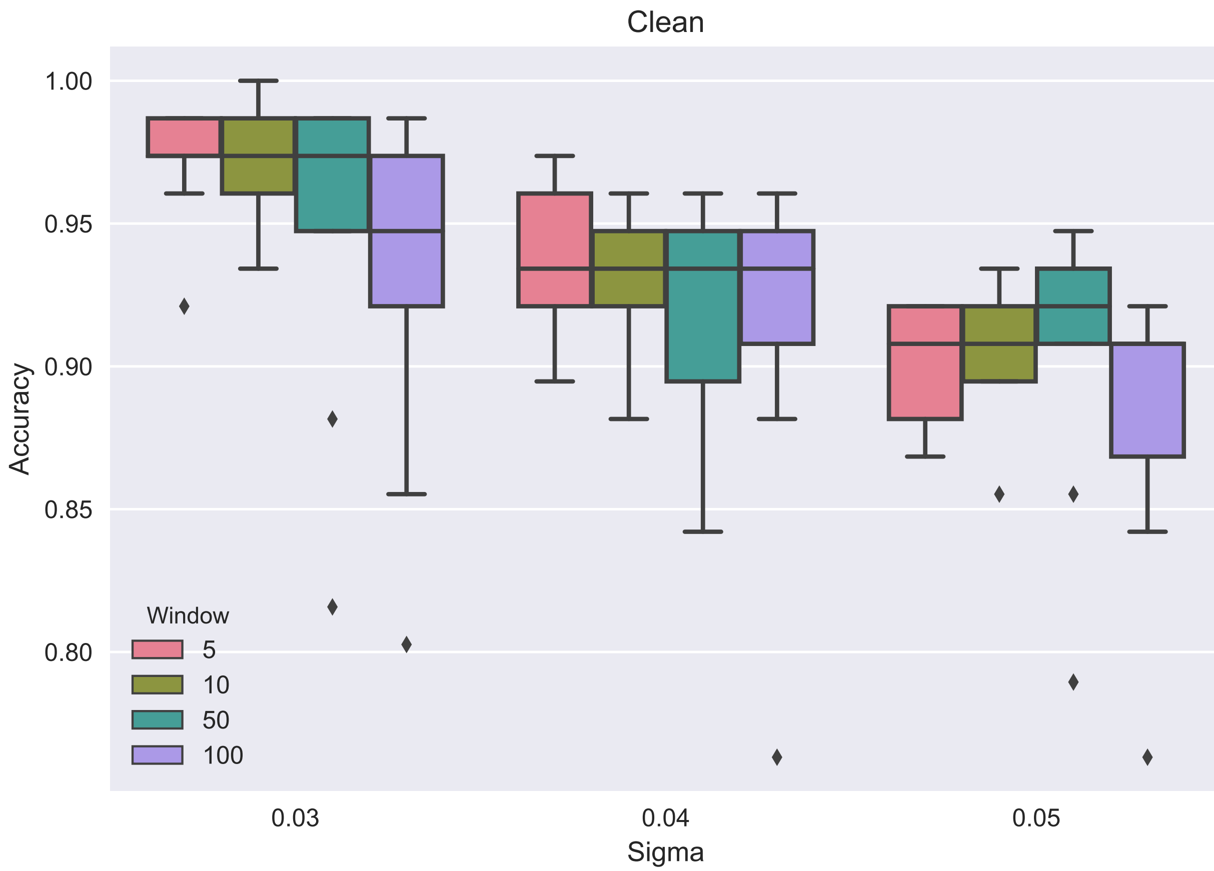
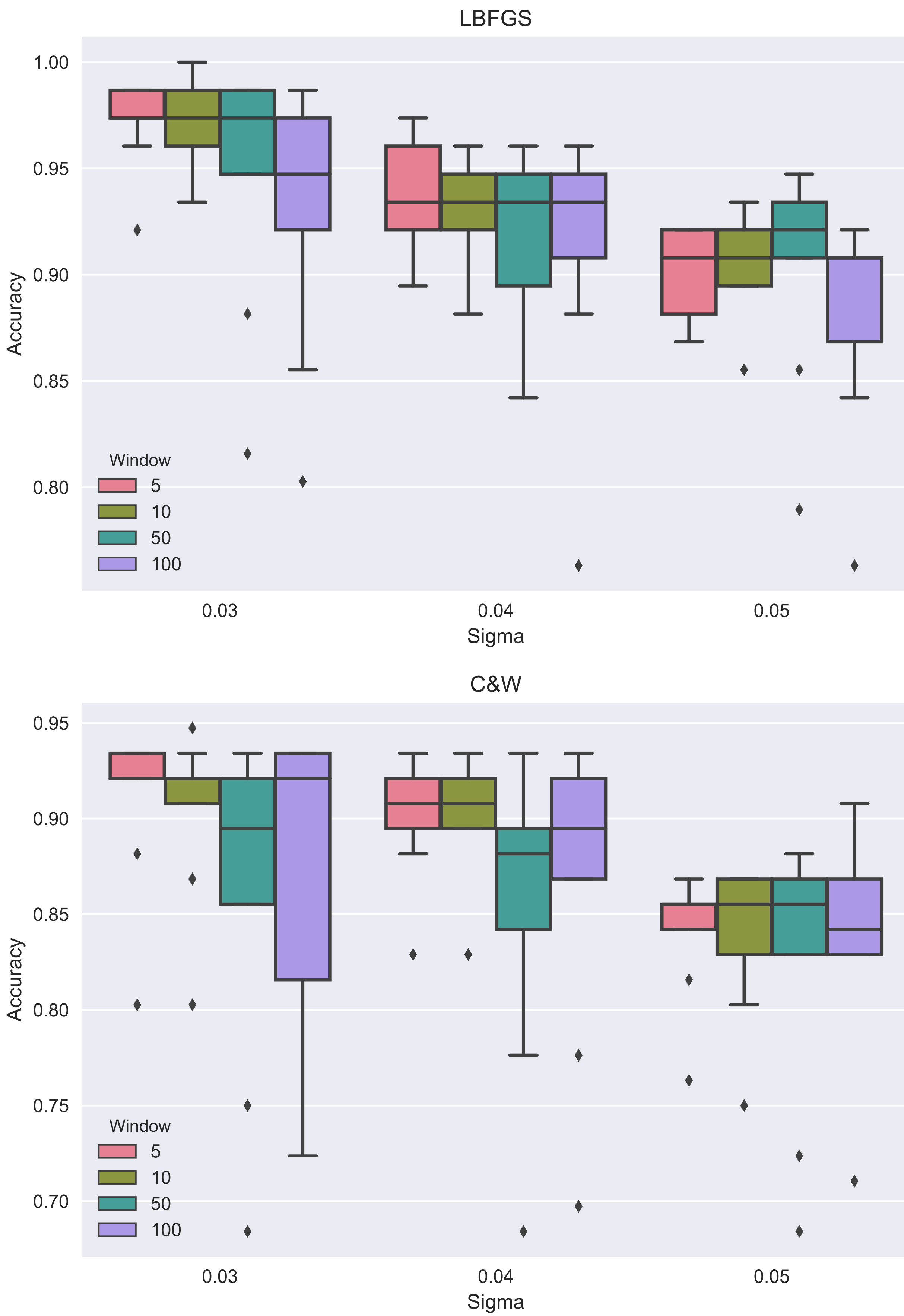


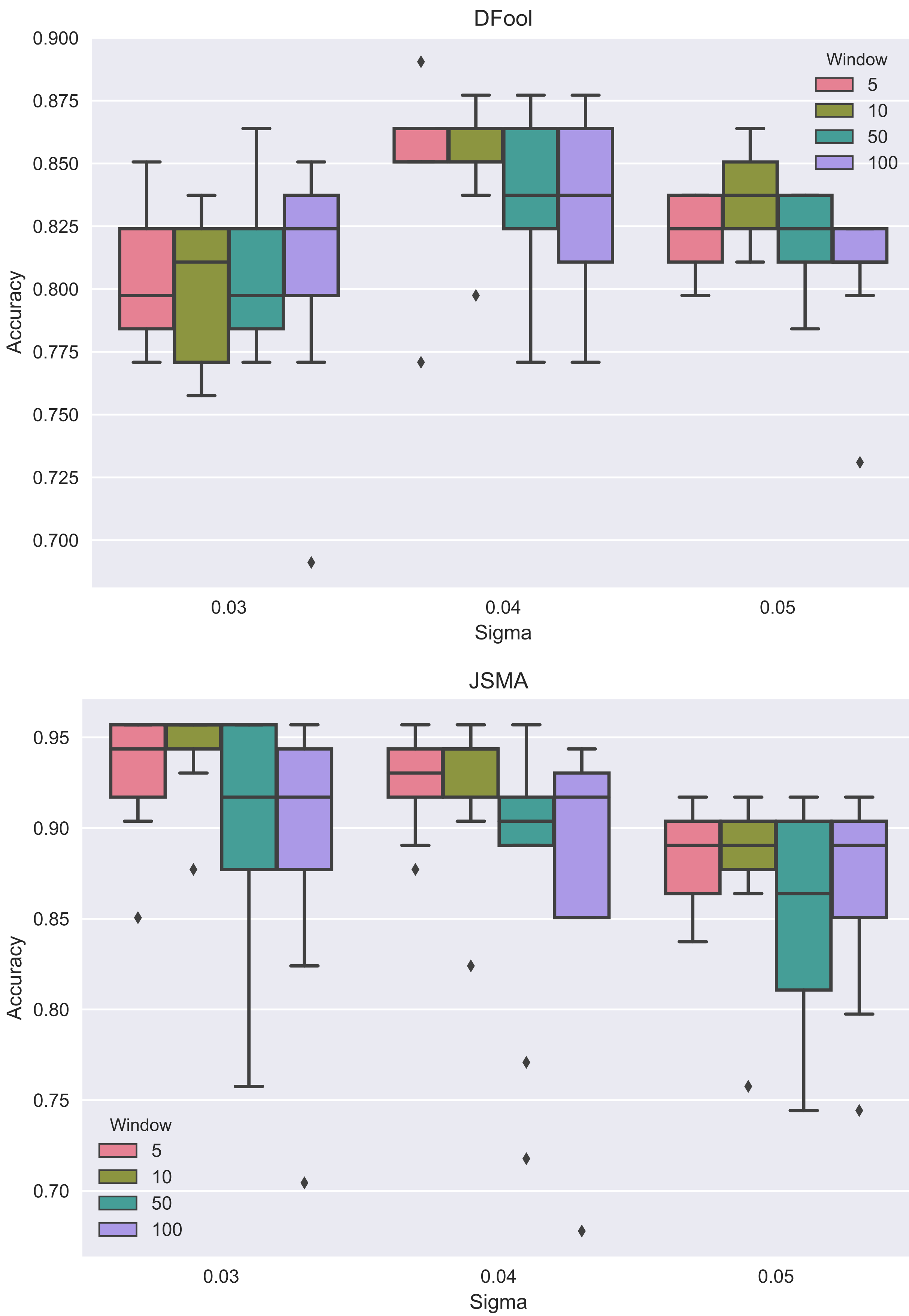
Figure 2: Comparison of Class activation maps and Robust Activation maps

Full size figures of Figure 5 (Linear search for model parameters on training data)

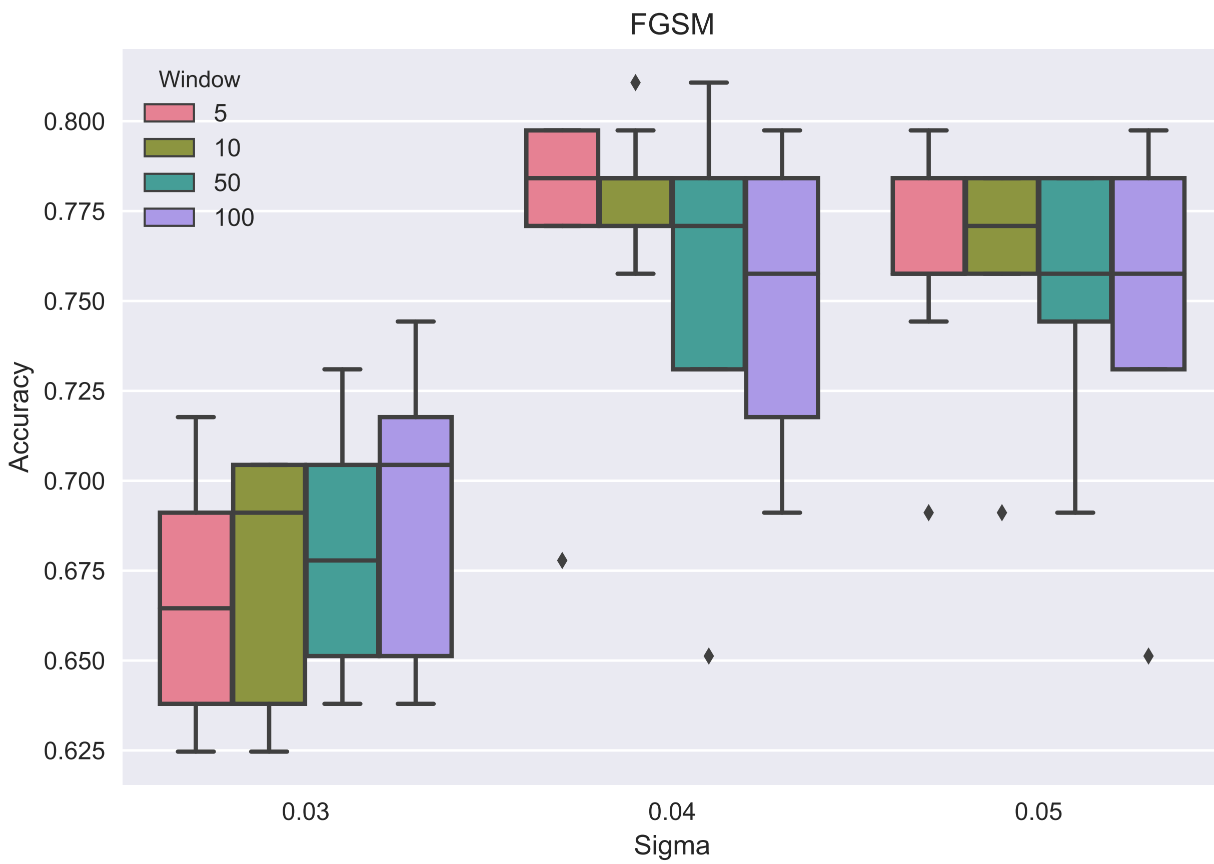


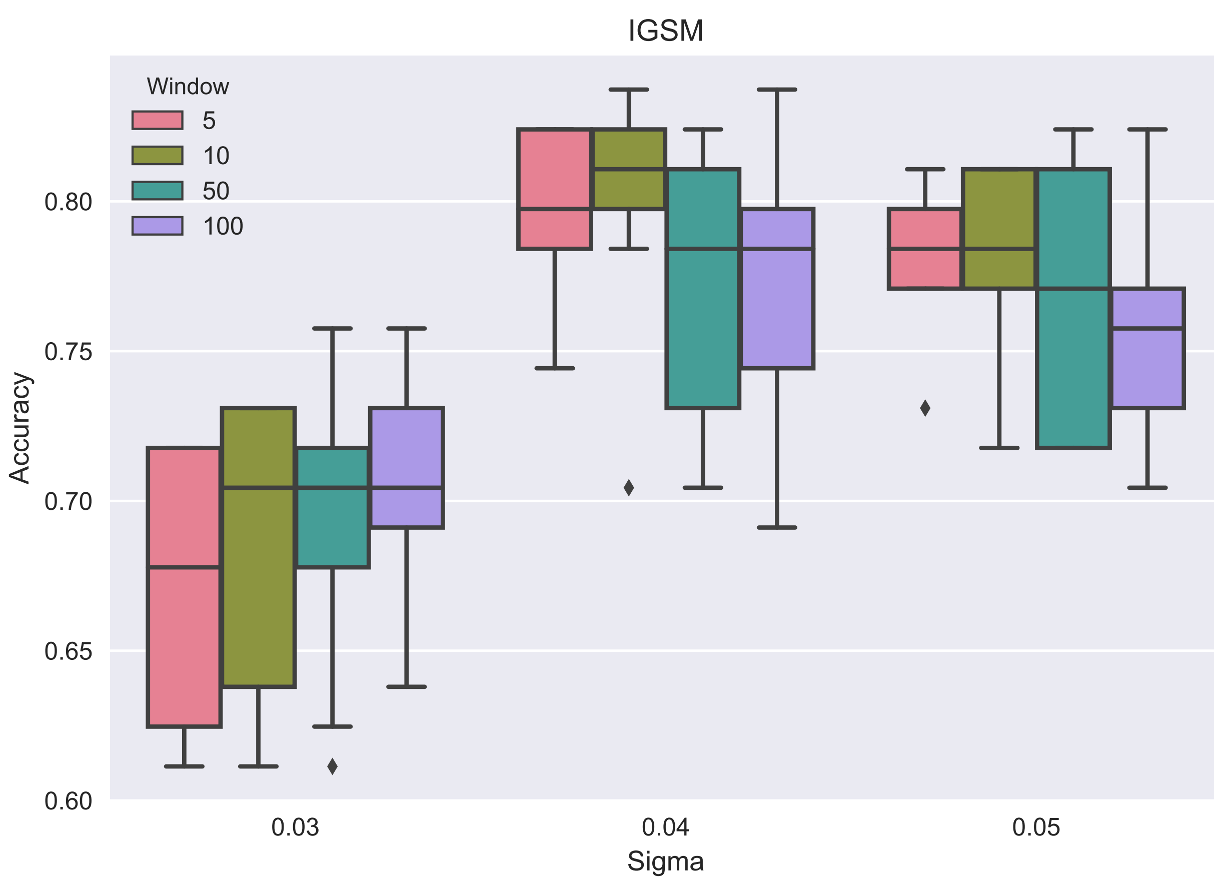


LBFGS
1.00
0.95
0.90
0.85
0.80
Accuracy
Window
5
10
50
100
0.03
0.04
0.05
Sigma
C&W
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
Window
5
10
50
100
0.03
0.04
0.05
Sigma



DFool
Accuracy
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0.725
0.700
0.03
0.04
0.05
Sigma
Window
5
10
50
100
JSMA
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.03
0.04
0.05
Sigma
Window
5
10
50
100


FGSM
Window
5
10
50
100
Accuracy
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.03
0.04
0.05
Sigma


IGSM
Window
5
10
50
100
Accuracy
0.80
0.75
0.70
0.65
0.60
0.03
0.04
0.05
Sigma