Supplementary material for

CIIDefence: Defeating Adversarial Attacks by Fusing Class-specific Image Inpainting and Image Denoising

Successful cases

Successful cases are the cases where the proposed *CIIDefence* has successfully mitigated the adversarial perturbations and enable the classifier to classify correctly. Some examples are shown in the next slides.

Description of these examples from left to right:

a) Adversarial image, I_q.

b) Denoised image obtained after removing the relevant masked image area, i.e., it depicts only that denoised area which is used in the fused image. Mathematically, it denotes $[(1-M)*I_d]$ from Equation (5) of the paper rather than full denoised image, I_d.

c) Image depicting inpainted areas, I_i.

d) Fused Image, I_r.

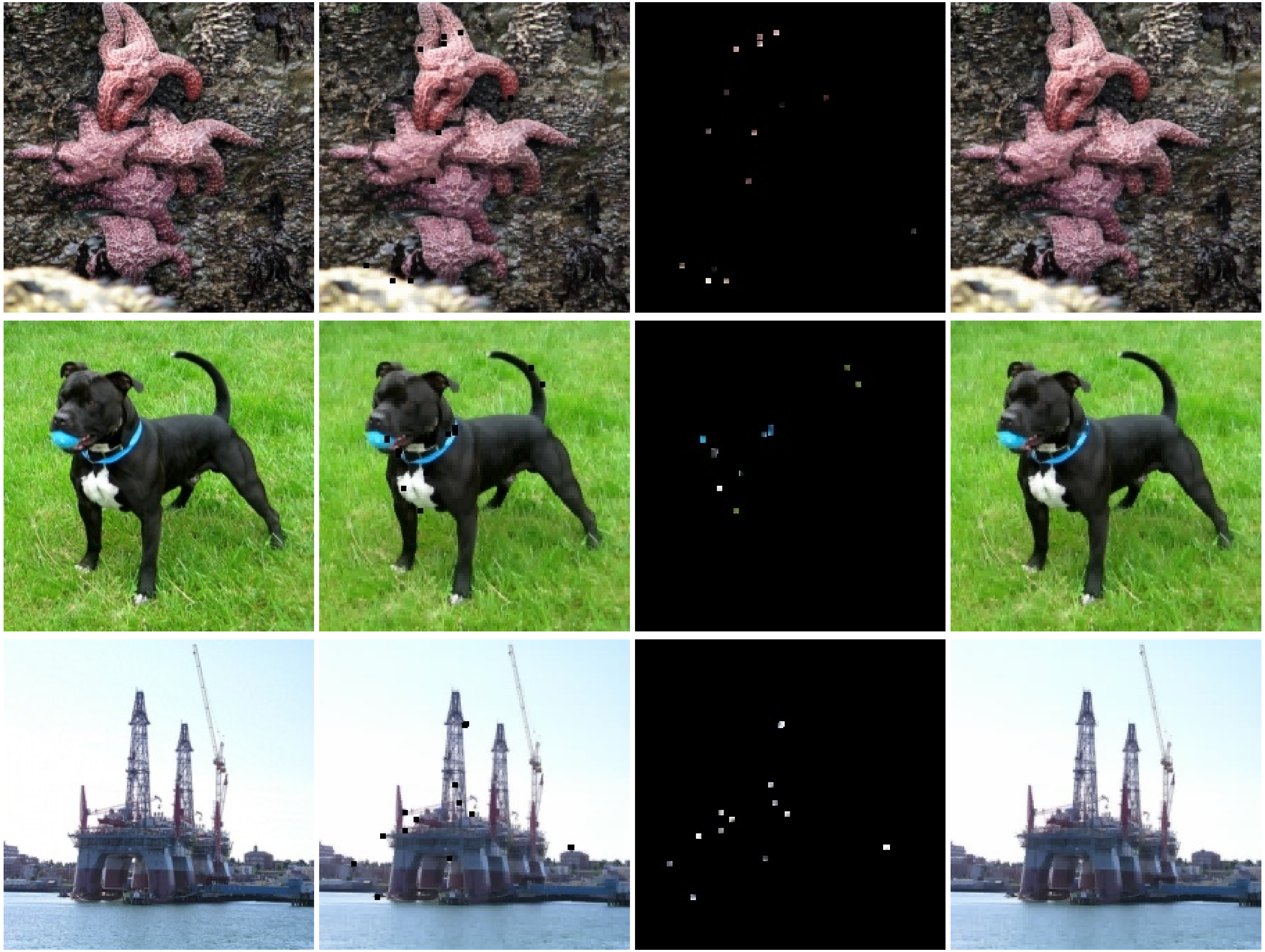
e) Red, green and blue color depict the true classification (i.e., classification of corresponding clean image); classification when adversarial attack is applied, but CIIDefence is not applied; and classification using *CIIDefence* respectively.

Results obtained using VGG-16.

Successful cases

a) Adversarial image | b) Denoised image without masked area | c) Inpainted areas | d) Fused Image | e) Classification results

Successful cases



a) Adversarial image	b) Denoised image without masked area	c) Inpainted areas	d) Fused Image	e) Classification results
				Starfish Mitten Starfish
				Staffordshire bull terrier Brabancon griffon Staffordshire bull terrier
				Offshore rig Wreck Offshore rig

Failure cases

Failure cases are the cases where the proposed *CIIDefence* is unsuccessful in mitigated the adversarial perturbations and hence, the classifier provided incorrect classification. Some examples are shown in the next slides.

Description of these examples from left to right:

a) Adversarial image, I_q.

b) Denoised image image obtained after removing the relevant masked image area, i.e., it depicts only that denoised area which is used in the fused image. Mathematically, it denotes $[(1\text{-}M)*I_d]$ from Equation (5) of the paper rather than full denoised image, I_d.

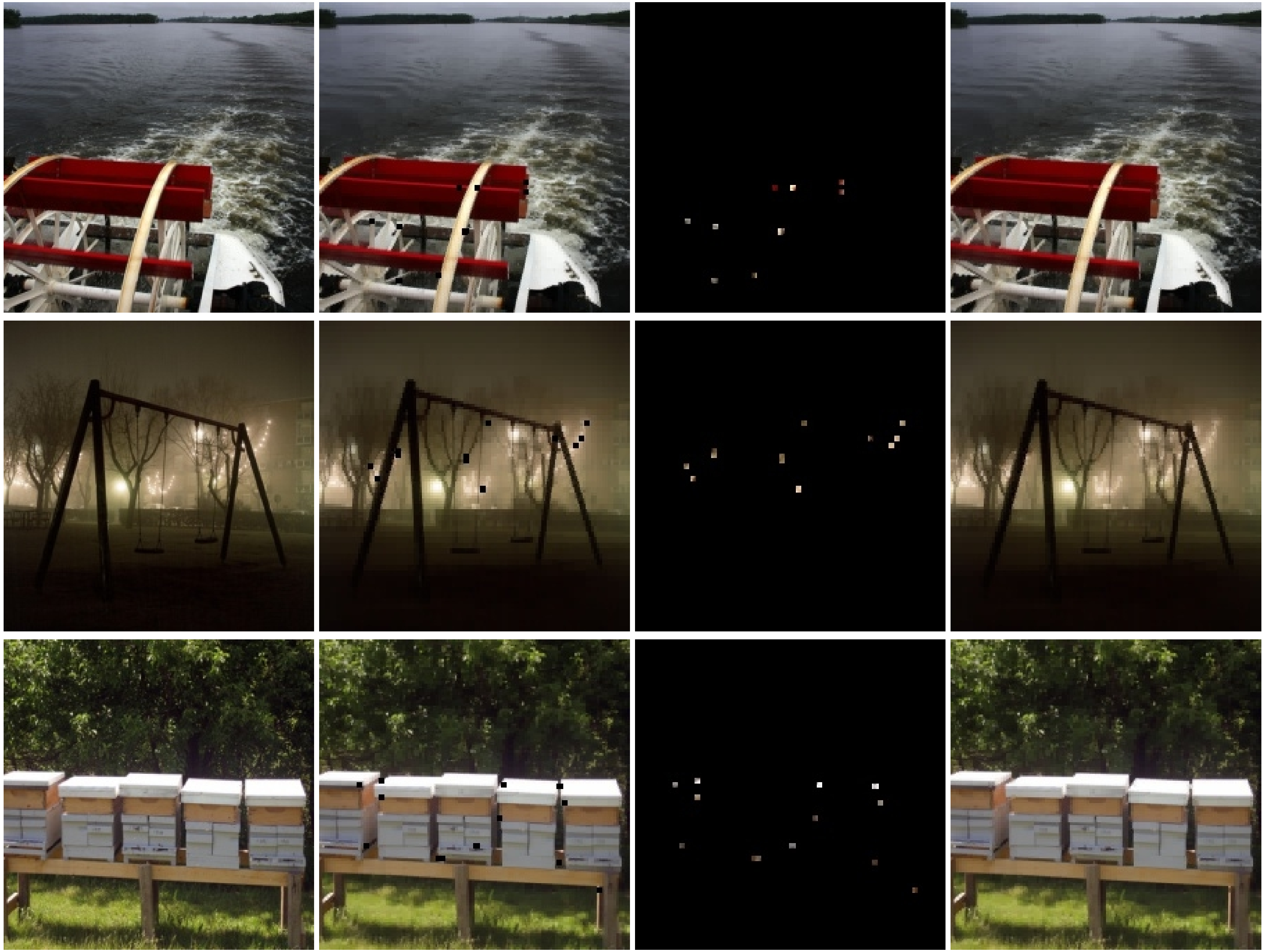
c) Image depicting inpainted areas, I_i.

d) Fused Image, I_r.

e) Red, green and blue color depict the true classification (i.e., classification of corresponding clean image); classification when adversarial attack is applied, but CIIDefence is not applied; and classification using *CIIDefence* respectively.

Results obtained using VGG-16.

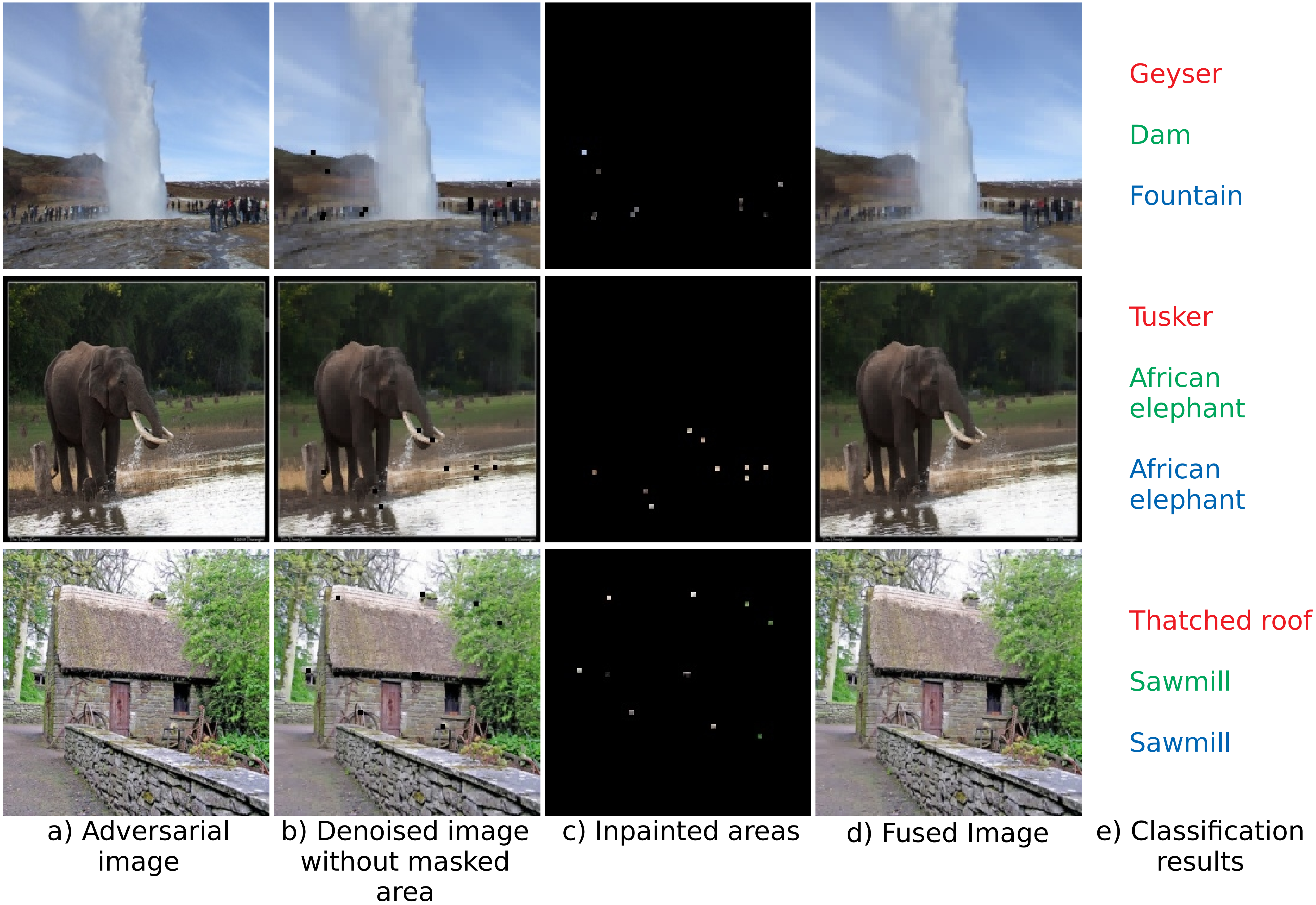
Failure cases



Paddlewheel	Swing	Apiary
Sandbar	Bannister	Mobile home
Trimaran	Tripod	Mobile home

a) Adversarial image b) Denoised image without masked area c) Inpainted areas d) Fused Image e) Classification results

Failure cases



Importance of *CIIDefence* over Inpainting

Examples depicting the importance of *CIIDefence* over inpainting are presented in the next slide. We fuse the inpainted areas with the adversarial image and present some examples where *CIIDefence* provides correct classification but fusion of image inpainting and adversarial images provides incorrect classification. It indicates that denoising plays a crucial role in *CIIDefence.*

Description of the examples from left to right:

a) Adversarial image, I_q.

b) Image depicting inpainted areas, I_i.

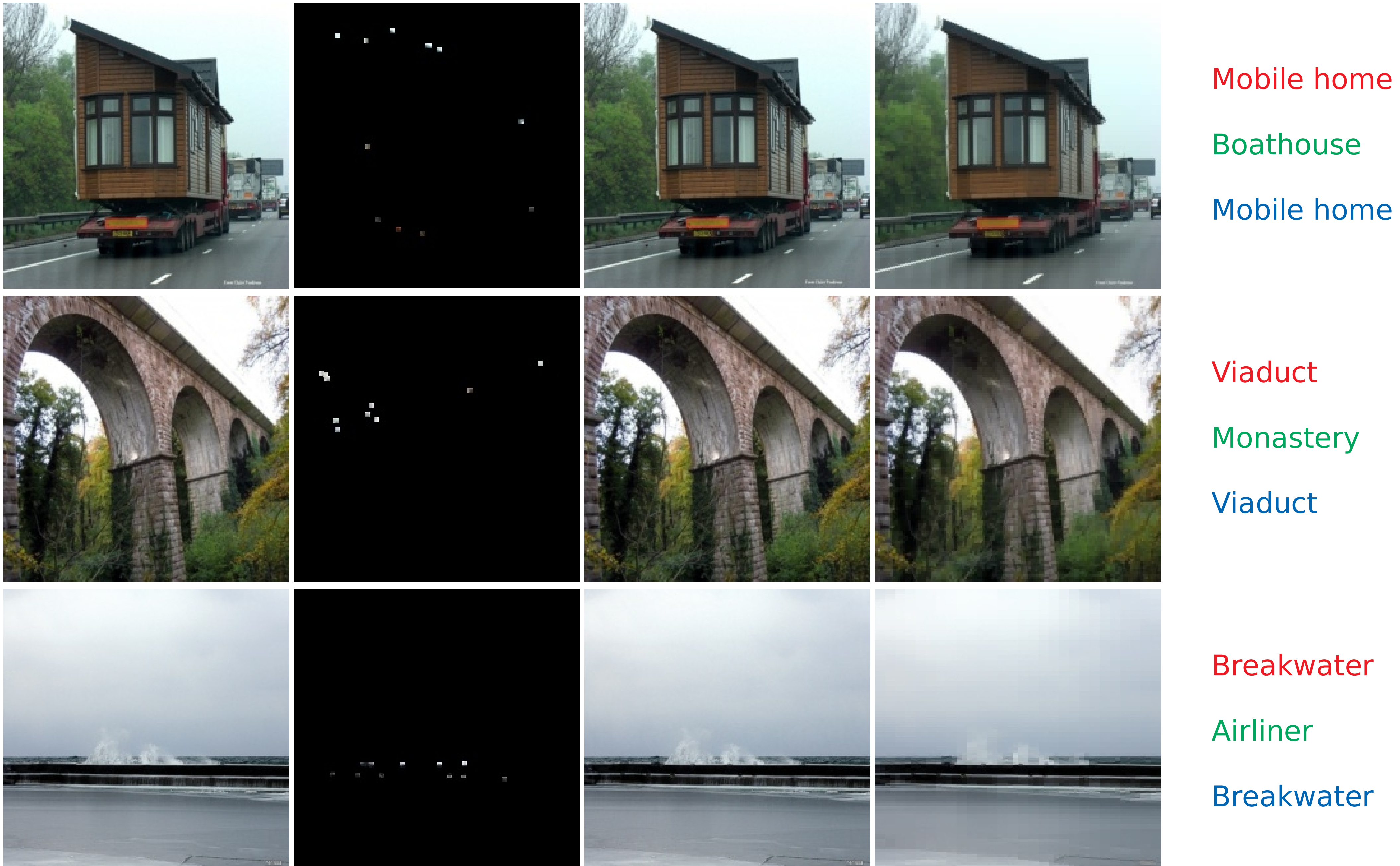
c) Image obtained by fusing inpainted and adversarial images. That is, it denotes $[M*I_i + (1-M)*I_q]$ rather than the Equation (5) of the paper.

d) Fused Image, I_r.

e) Red, green and blue color depict the true classification (i.e., classification of corresponding clean image); classification when image in c) is used; and classification using *CIIDefence* respectively.

Results obtained using VGG-16.

Importance of *CIIDefence* over Inpainting



a) Adversarial image

b) Inpainted areas

c) Inpainted image fused with adversarial image

d) Fused Image

e) Classification results

Importance of *CIIDefence* over Denoising

Examples depicting the importance of *CIIDefence* over denoising are presented in the next slide. They provide correct classification when *CIIDefence* is used but incorrect classification when denoised image is used. It indicates that inpainting plays a crucial role in *CIIDefence.*

Description of the examples from left to right:

a) Adversarial image, I_q.

b) Denoised image, I_d.

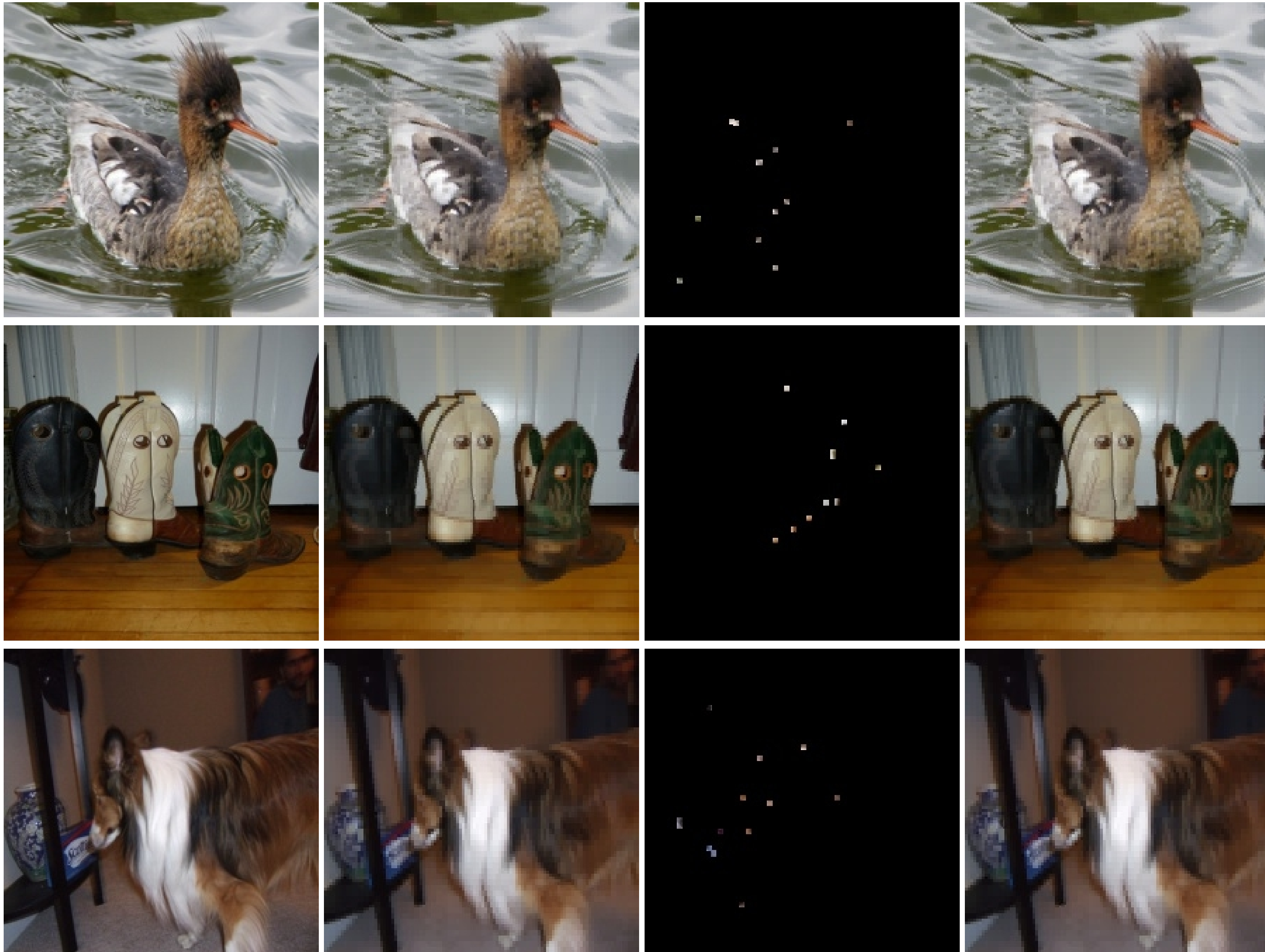
c) Image depicting inpainted areas, I_i.

d) Fused Image, I_r.

e) Red, green and blue color depict the true classification (i.e., classification of corresponding clean image); classification when image in b) is used; and classification using *CIIDefence* respectively.

Results obtained using VGG-16.

Importance of *CIIDefence* over Denoising



New Ablation Study: Comparision to PD [1]

Table description:

- Here, per class CAM is replaced with an averaging CAM used in [1].
- It uses the same test setup as in Section 5.5 of the paper.

	Original	FGSM	IGSM	DFool	C&W
PD [1]	96.9	69.4	81.8	82.7	85.8
Our + avg. CAM	99.1	87.1	93.4	97.2	98.1
Our+ per class CAM	99.2	87.6	93.8	97.8	98.4

It can be observed from the table that per class CAM has positive impact on the results. However, the performance gap to PD [1] is mainly due to global inpainting and non-differentiable operation for gradient masking.

[1]: Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8571–8580, 2018.

New Ablation Study: Generalization of Hyperparameter

In the paper, we used a fixed set of parameter values for all attacks. These values are chosen based on the average performance over the five attacks in the training set (see Section 5.1). In this new experiment, the parameter values are determined with one attack type and then tested with other attacks.

The Table indicate that: 1) performance increases slightly for the selected attack; 2) decreases for the others; and 3) the mean performance does not change more than 1%. The optimal values for $\hat{p}$ and n were found to be equal in all cases, while w changed slightly.

	FGSM	IGSM	DFool	C&W	*w*	*ACC*
FGSM	88.0%	92.4%	96.8%	97,0%	2	94.3%
IGSM	86.8%	94.2%	96.4%	97.4%	4	94.4%
DFool	87.6%	93.8%	97.8%	98.4%	3	95.2%
C&W	87.6%	93.8%	97.8%	98.4%	3	95.2%