Deflecting Adversarial Attack with Pixel Deflection

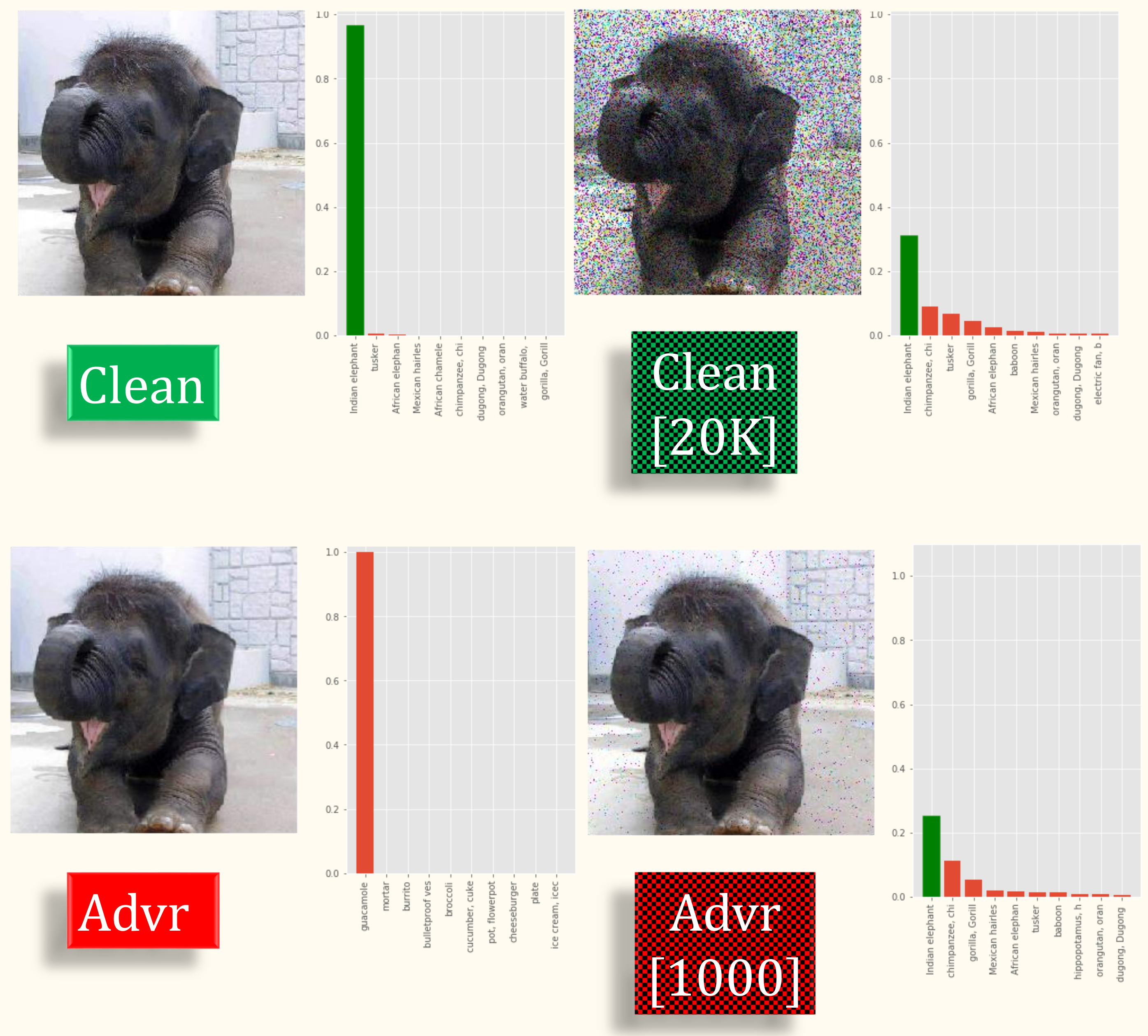
A. Prakash, N. Moran, S. Garber, A. DiLillo & J. Storer

aprakash@brandeis.edu, iamaaditya.github.io

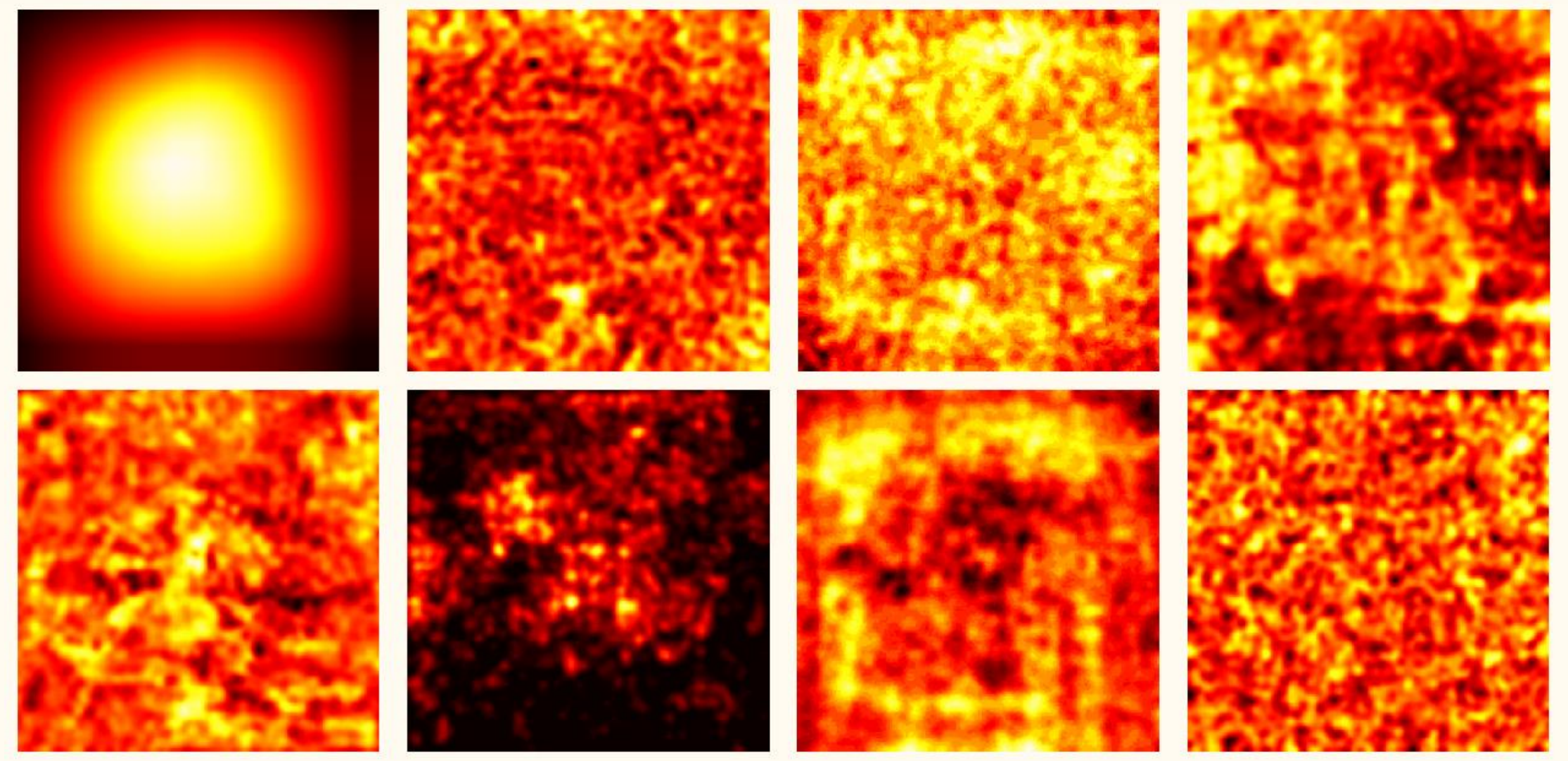
www.github.com/iamaaditya/pixel-deflection

CVPR 2018
Salt Lake City

1. Classifiers are robust to noise but Adversarial systems are not

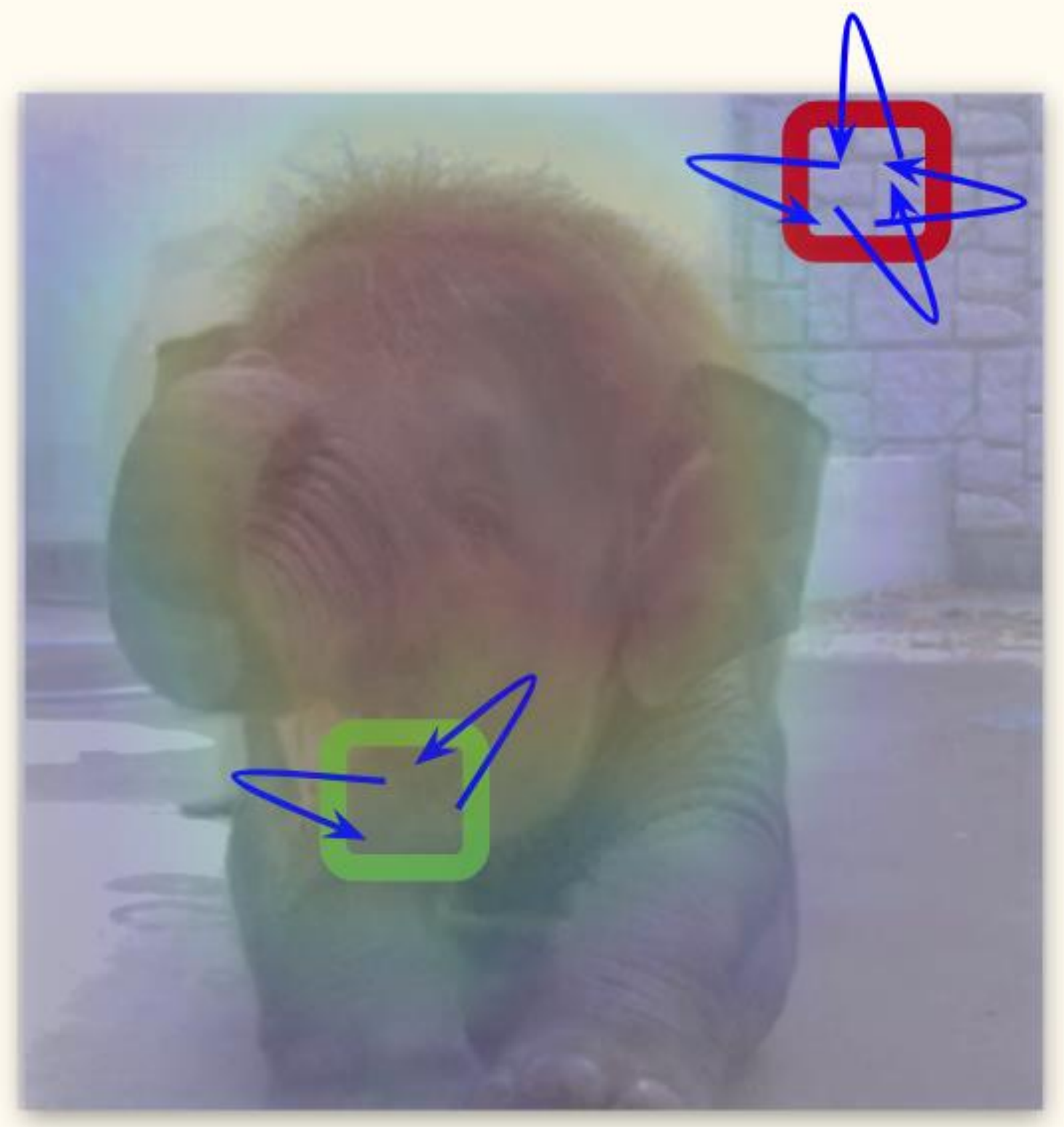


2. Classifiers look for semantic regions but Adversarial systems are content agnostic



Visualization showing average location in the image of semantic content (TL), and various adversarial systems.

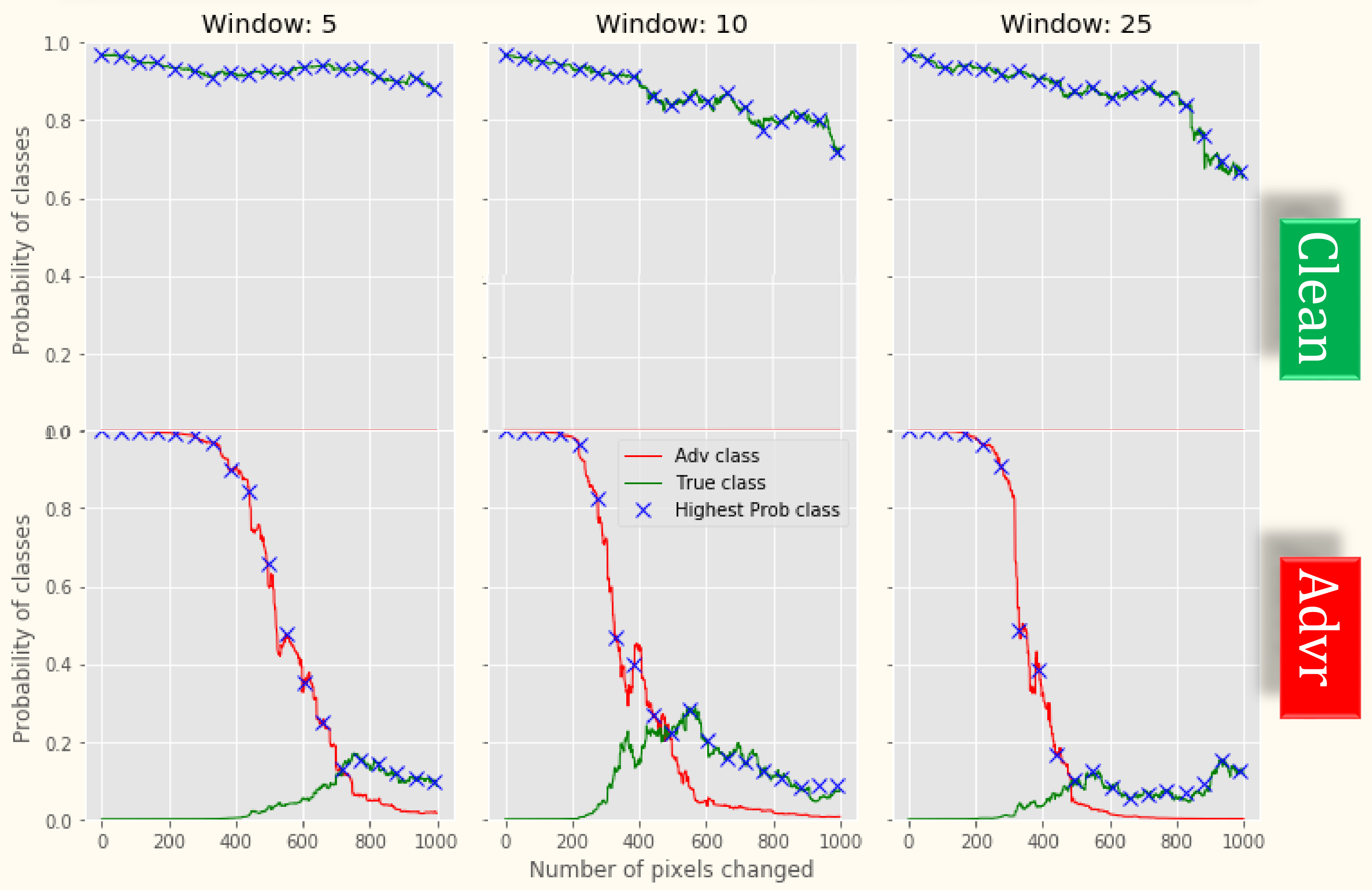
Pixel Deflection



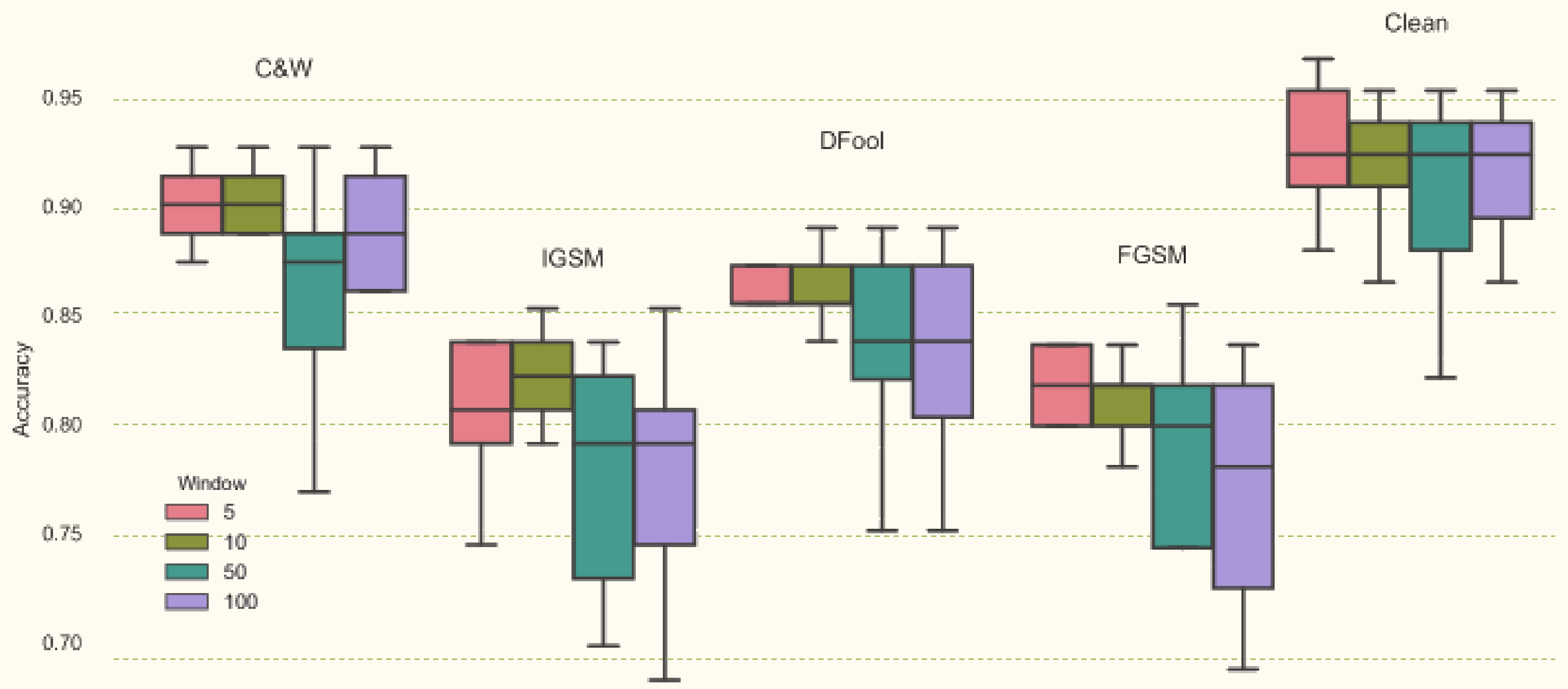
Algorithm

```
Input  : Image $\mathbb{I}$, deflections $\mathbf{K}$,
         window $\mathbf{w}$, activation map $\mathbf{M}$
Output: Image $\mathbb{I}'$
$\mathbb{I}' \leftarrow \mathbb{I}$
for $i \leftarrow 0$ to $\mathbf{K}$ do
    Let $p_i \sim \mathcal{U}(\mathbb{I})$
    if $\mathbf{M}[p_i] < \mathcal{U}(0:1)$
        Let $n_i \sim \mathcal{U}(R_w[p_i] \cap \mathbb{I})$
        $\mathbb{I}'[p_i] = \mathbb{I}[n_i]$
```

Performance



Parameters



Destruction Rate Image⇢Adversary⇢Pixel Deflection

Model	$\lvert L_2 \rvert$	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**

Comparison with SOTA defenses

Defense	**FGSM**	**IGSM**	**DFool**	**C&W**
Feature Squeezing (Xu et al [49])				
(a) Bit Depth (2 bit)	0.132	0.511	0.286	0.170
(b) Bit Depth (5 bit)	0.057	0.022	0.310	0.957
(c) Median Smoothing (2x2)	0.358	0.422	0.714	0.894
(d) Median Smoothing (3x3)	0.264	0.444	0.500	0.723
(e) Non-local Mean (11-3-2)	0.113	0.156	0.357	0.936
(f) Non-local Mean (13-3-4)	0.226	0.444	0.548	0.936
Best model (b) + (c) + (f)	0.434	0.644	0.786	0.915
Random resizing + padding (Xie *et al.* [48])				
Pixel padding	0.050	-	0.972	0.698
Pixel resizing	0.360	-	0.974	0.971
Padding + Resizing	0.478	-	**0.983**	0.969
Quilting + TVM (Guo *et al.* [19])				
Quilting	0.611	0.862	0.858	0.843
TVM + Quilting	0.619	0.866	0.866	0.841
Cropping + TVM + Quilting	0.629	0.882	0.883	0.859
Our work: PD - Pixel Deflection, R-CAM: Robust CAM				
PD	0.735	0.880	0.914	0.931
PD + R-CAM	0.746	0.912	0.911	0.952
PD + R-CAM + DCT	0.737	0.906	0.874	0.930
PD + R-CAM + DWT	**0.769**	**0.927**	0.948	**0.981**