NEURAL INFORMATION PROCESSING SYSTEMS

COLALab
Computational Optimization for Learning & Adaptive Systems

CamoPatch: An Evolutionary Strategy for Generating Camouflaged Adversarial Patches

Phoenix Williams, Ke Li

Department of Computer Science, University of Exeter, UK

✉ pw384@exeter.ac.uk 🔓 https://phoenixwilliams.github.io/PersonalWebsite/

✉ k.li@exeter.ac.uk 🔓 http://colalab.ai

University of Exeter

UK Research and Innovation

THE ROYAL SOCIETY

Engineering and Physical Sciences Research Council

The Alan Turing Institute

amazon research awards

What Is Computer Vision?

"Computer vision is the art and science of teaching machines to see, enabling them to understand the visual tapestry of our world." - ChatGPT

Classification

"Welsh springer spaniel"

Object Detection

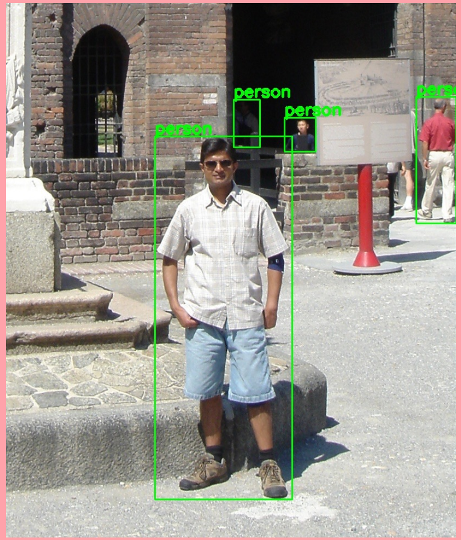
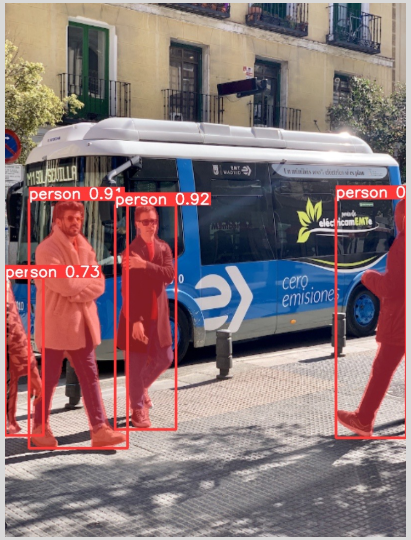
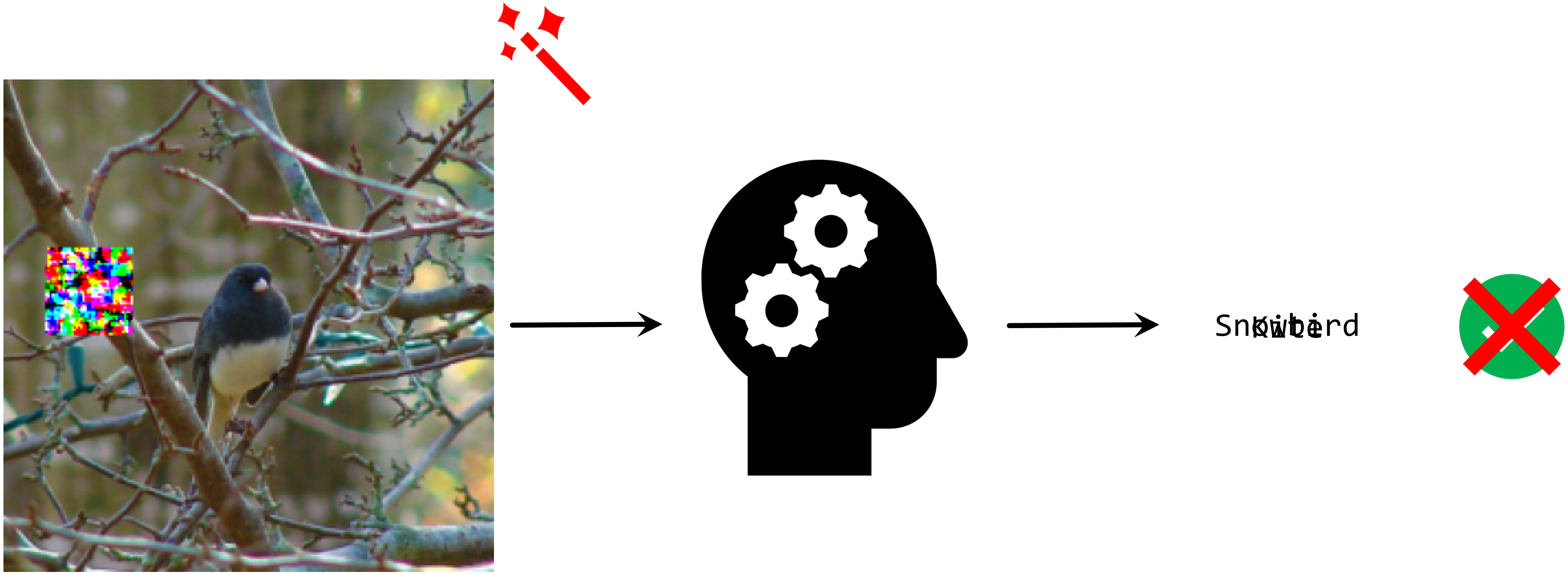


Image Segmentation

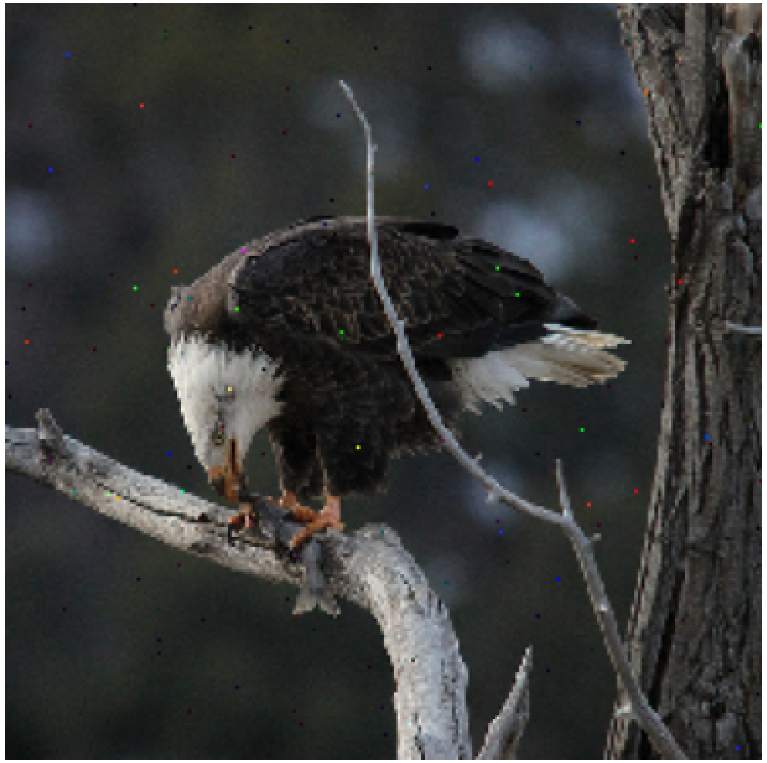


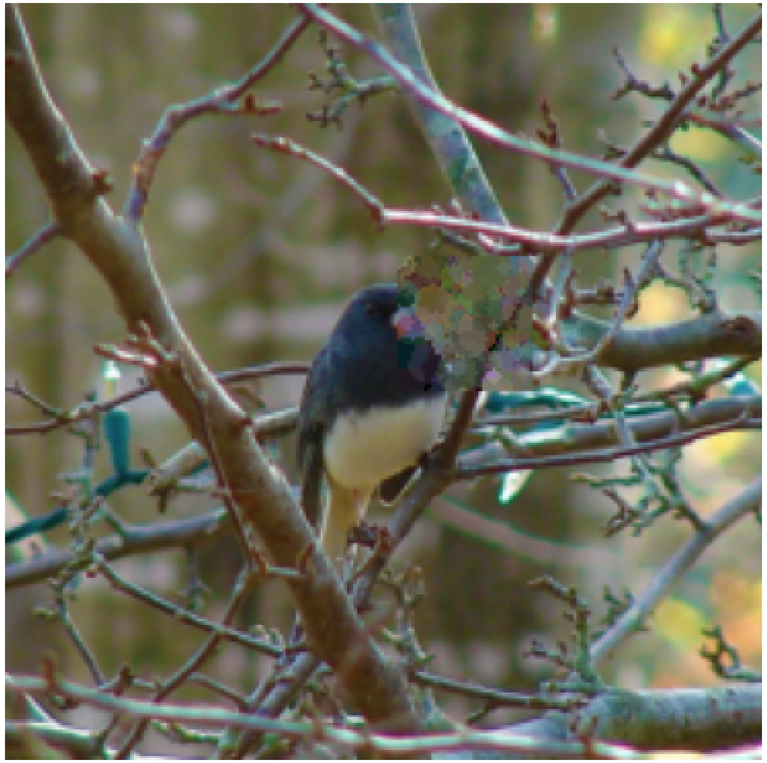
What are Adversarial Patches?



Our Motivation

“By ignoring the visibility of these modifications the applicability of existing methods is questionable”

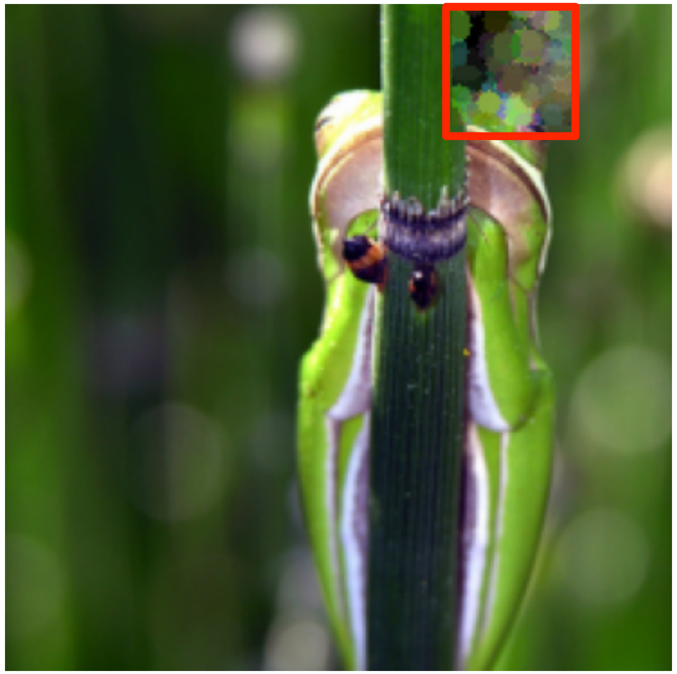


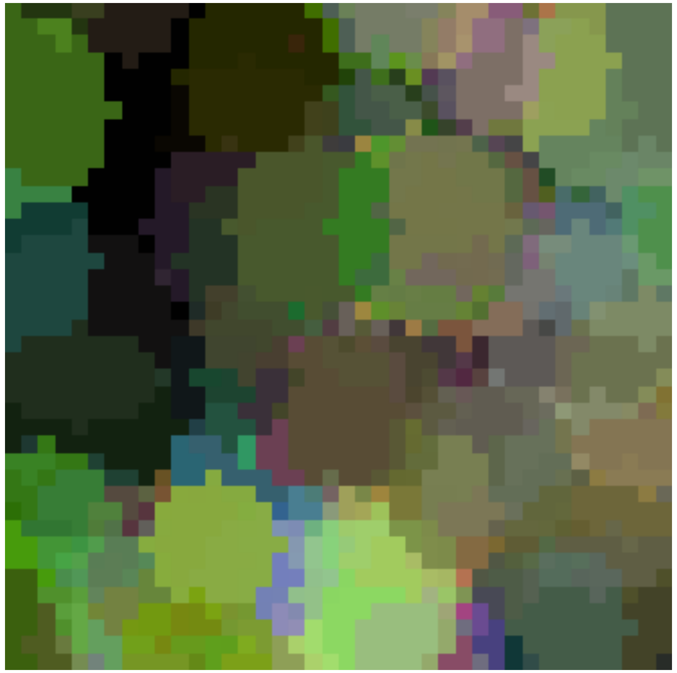


“Minimising the size of the modification increases its applicability within the real-world whilst improving the evaluation of an AI’s robustness”

CamoPatch

Tree frog → Grasshopper



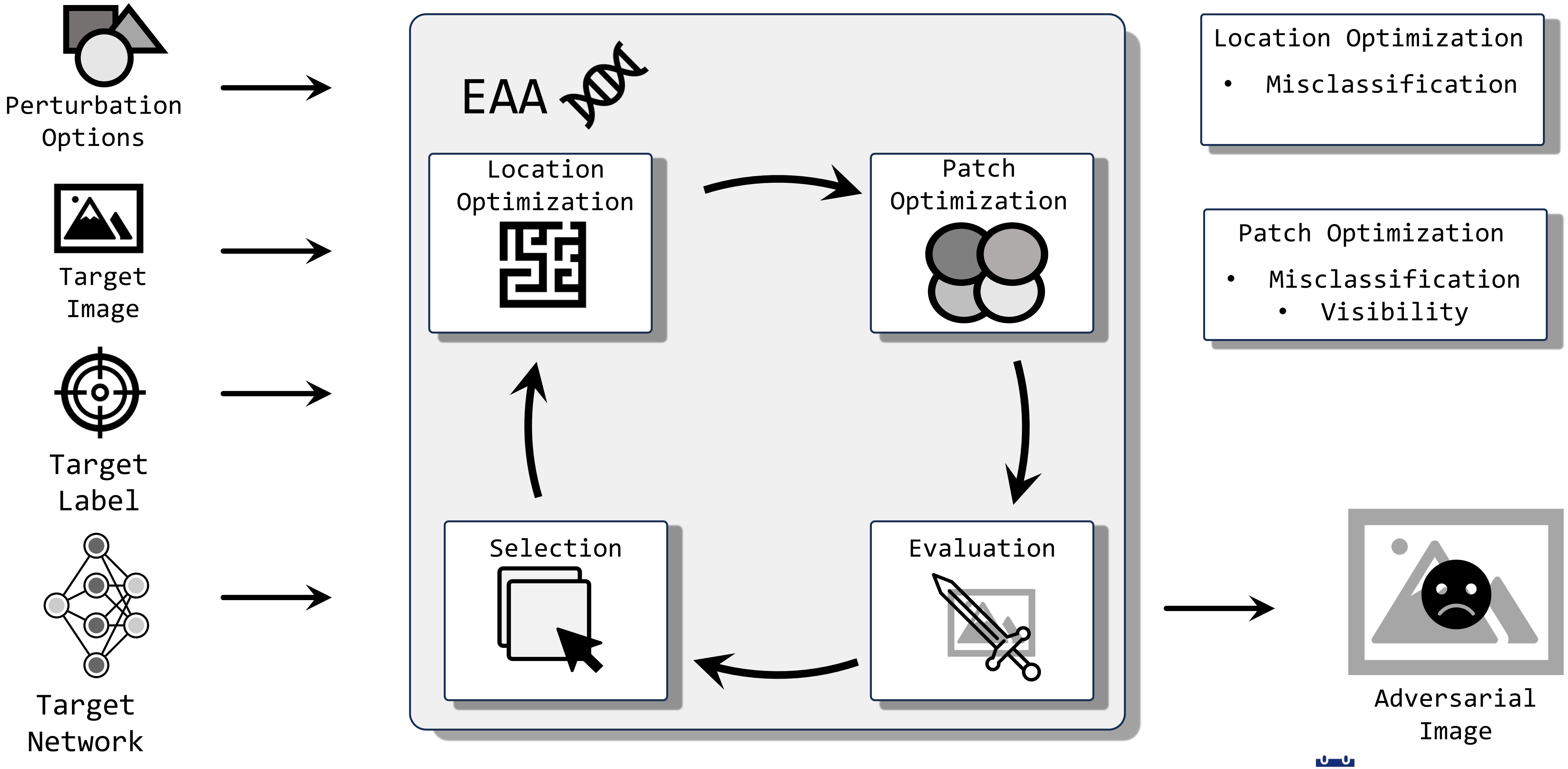


“Constructing an adversarial patch requires jointly optimising its location and pattern”

“Location of the patch is represented by integers between 0 and (W x H) where W and H are the height and width of the image.”

“We construct the patch using a set of semi-transparent circle shapes, allowing for effective performance and camouflage”

CamoPatch



Perturbation Options

Target Image

Target Label

Target Network

EAA

Location Optimization

Patch Optimization

Evaluation

Selection

Adversarial Image

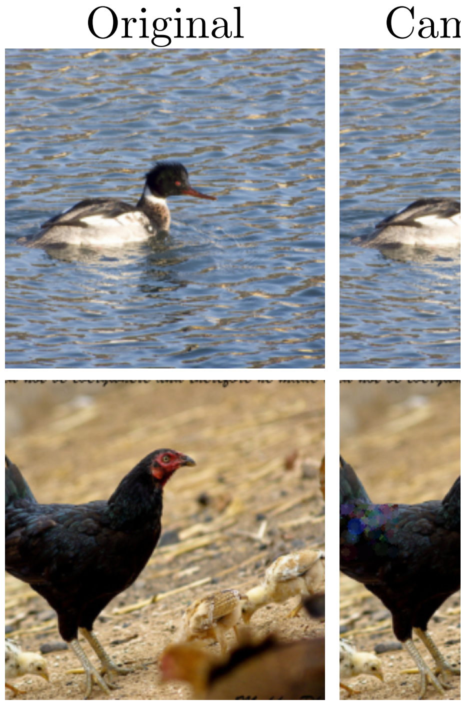
Location Optimization

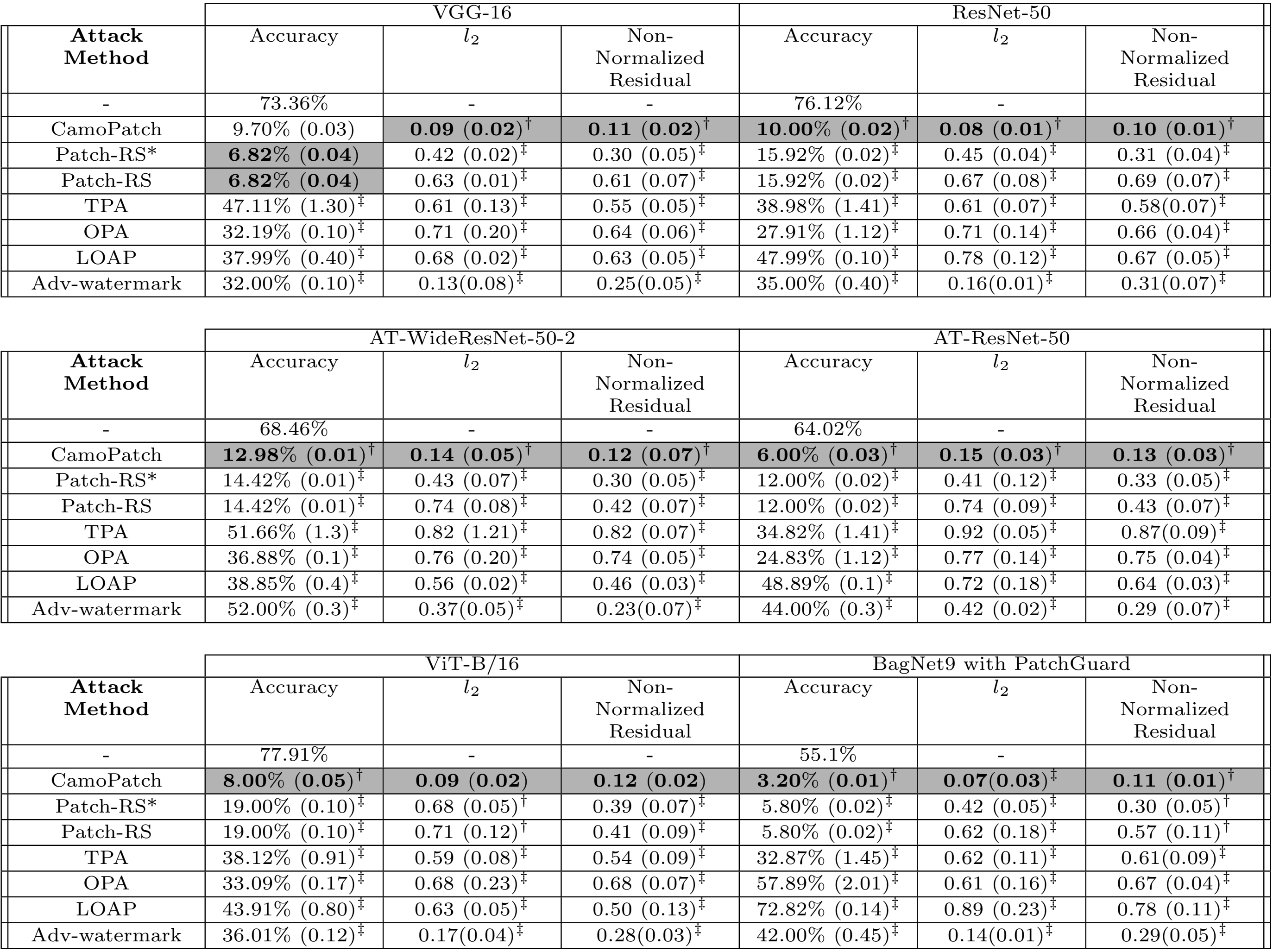
- Misclassification

Patch Optimization

- Misclassification
- Visibility

CamoPatch

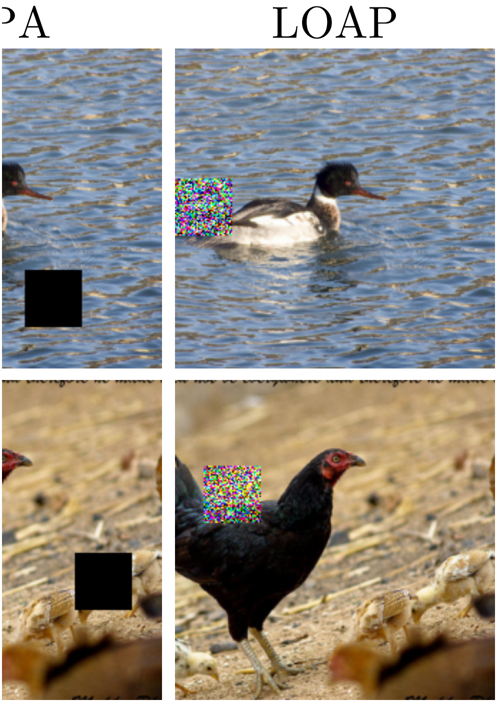




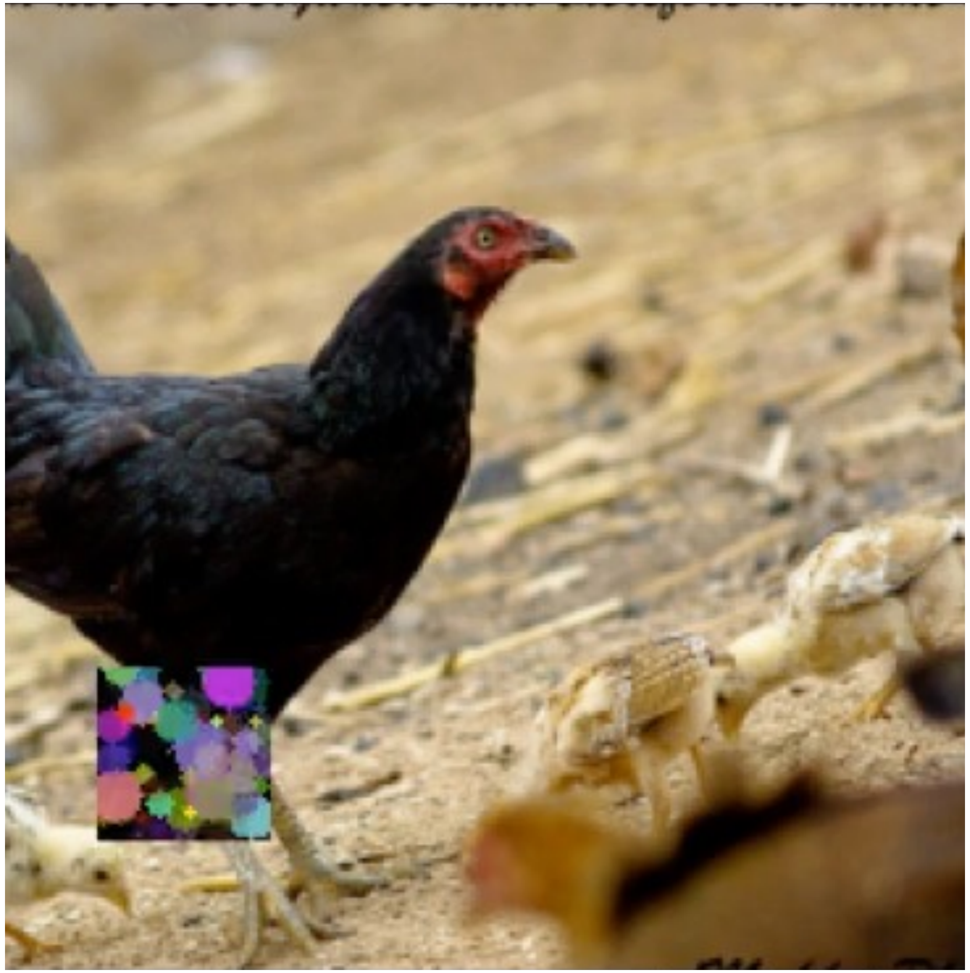
	VGG-16			ResNet-50		
Attack Method	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	73.36%	-	-	76.12%	-	
CamoPatch	9.70% (0.03)	**0.09 (0.02)**†	**0.11 (0.02)**†	**10.00% (0.02)**†	**0.08 (0.01)**†	**0.10 (0.01)**†
Patch-RS*	**6.82% (0.04)**	0.42 (0.02)‡	0.30 (0.05)‡	15.92% (0.02)‡	0.45 (0.04)‡	0.31 (0.04)‡
Patch-RS	**6.82% (0.04)**	0.63 (0.01)‡	0.61 (0.07)‡	15.92% (0.02)‡	0.67 (0.08)‡	0.69 (0.07)‡
TPA	47.11% (1.30)‡	0.61 (0.13)‡	0.55 (0.05)‡	38.98% (1.41)‡	0.61 (0.07)‡	0.58(0.07)‡
OPA	32.19% (0.10)‡	0.71 (0.20)‡	0.64 (0.06)‡	27.91% (1.12)‡	0.71 (0.14)‡	0.66 (0.04)‡
LOAP	37.99% (0.40)‡	0.68 (0.02)‡	0.63 (0.05)‡	47.99% (0.10)‡	0.78 (0.12)‡	0.67 (0.05)‡
Adv-watermark	32.00% (0.10)‡	0.13(0.08)‡	0.25(0.05)‡	35.00% (0.40)‡	0.16(0.01)‡	0.31(0.07)‡

	AT-WideResNet-50-2			AT-ResNet-50		
Attack Method	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	68.46%	-	-	64.02%	-	
CamoPatch	**12.98% (0.01)**†	**0.14 (0.05)**†	**0.12 (0.07)**†	**6.00% (0.03)**†	**0.15 (0.03)**†	**0.13 (0.03)**†
Patch-RS*	14.42% (0.01)‡	0.43 (0.07)‡	0.30 (0.05)‡	12.00% (0.02)‡	0.41 (0.12)‡	0.33 (0.05)‡
Patch-RS	14.42% (0.01)‡	0.74 (0.08)‡	0.42 (0.07)‡	12.00% (0.02)‡	0.74 (0.09)‡	0.43 (0.07)‡
TPA	51.66% (1.3)‡	0.82 (1.21)‡	0.82 (0.07)‡	34.82% (1.41)‡	0.92 (0.05)‡	0.87(0.09)‡
OPA	36.88% (0.1)‡	0.76 (0.20)‡	0.74 (0.05)‡	24.83% (1.12)‡	0.77 (0.14)‡	0.75 (0.04)‡
LOAP	38.85% (0.4)‡	0.56 (0.02)‡	0.46 (0.03)‡	48.89% (0.1)‡	0.72 (0.18)‡	0.64 (0.03)‡
Adv-watermark	52.00% (0.3)‡	0.37(0.05)‡	0.23(0.07)‡	44.00% (0.3)‡	0.42 (0.02)‡	0.29 (0.07)‡

	ViT-B/16			BagNet9 with PatchGuard		
Attack Method	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	77.91%	-	-	55.1%	-	
CamoPatch	**8.00% (0.05)**†	**0.09 (0.02)**	**0.12 (0.02)**	**3.20% (0.01)**†	**0.07(0.03)**‡	**0.11 (0.01)**†
Patch-RS*	19.00% (0.10)‡	0.68 (0.05)†	0.39 (0.07)‡	5.80% (0.02)‡	0.42 (0.05)‡	0.30 (0.05)†
Patch-RS	19.00% (0.10)‡	0.71 (0.12)†	0.41 (0.09)‡	5.80% (0.02)‡	0.62 (0.18)‡	0.57 (0.11)†
TPA	38.12% (0.91)‡	0.59 (0.08)‡	0.54 (0.09)‡	32.87% (1.45)‡	0.62 (0.11)‡	0.61(0.09)‡
OPA	33.09% (0.17)‡	0.68 (0.23)‡	0.68 (0.07)‡	57.89% (2.01)‡	0.61 (0.16)‡	0.67 (0.04)‡
LOAP	43.91% (0.80)‡	0.63 (0.05)‡	0.50 (0.13)‡	72.82% (0.14)‡	0.89 (0.23)‡	0.78 (0.11)‡
Adv-watermark	36.01% (0.12)‡	0.17(0.04)‡	0.28(0.03)‡	42.00% (0.45)‡	0.14(0.01)‡	0.29(0.05)‡



CamoPatch



“Initial S=stages consist of locating an optimal location and pattern that cause the desired misclassification”

“Once the optimal location has been found, the RGB circles approximate the area of the image the patch is placed upon, reducing its visibility”

Questions

Thank You For Your Attention!

CamoPatch:

URL: https://shorturl.at/rsv05