Machine Learning for Cyber-security

A Deep Dive on Adversarial Attacks

Alexandre Araujo

October 12, 2023

Going Beyond ℓ_p norms and Recent Work on Adversarial Robustness

Going Beyond ℓ_p norms

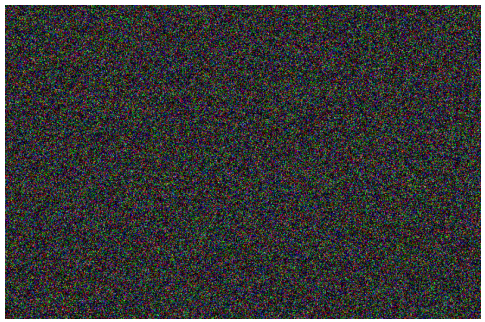
Recap of the previous sessions

- We have focused on ℓ_p based adversairal atttacks
- We have seen how to build adversarial attacks in the white box setting
- We studied empirical defenses and highlighted their limitations
- We have seen hoow to build certified defenses based the property of Lipschitz continuty and the Randomized Smoothing framework

How to properly defined Adversarial attacks?

How to properly defined Adversarial attacks?

Gaussian Noise sigma = 0.5



Adversarial Image

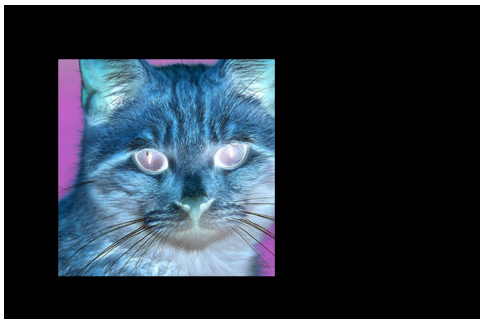
How to properly defined Adversarial attacks?

Uniform Noise sigma = 0.5

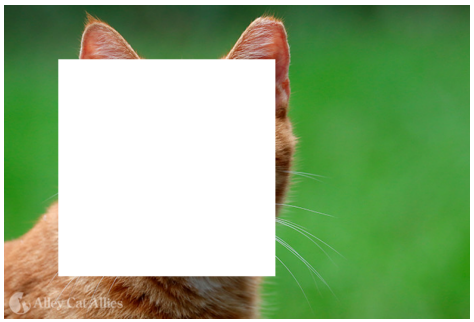
Adversarial Image

How to properly defined Adversarial attacks?

Inverse Perturbation

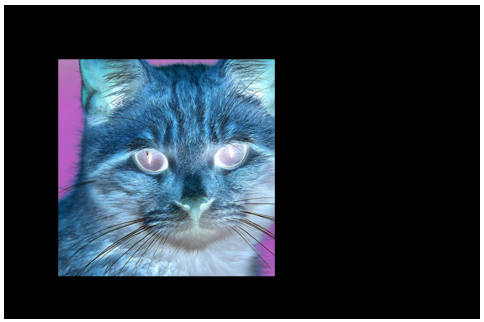


Adversarial Image

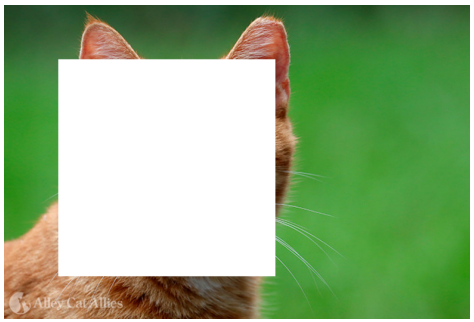


How to properly defined Adversarial attacks?

Inverse Perturbation

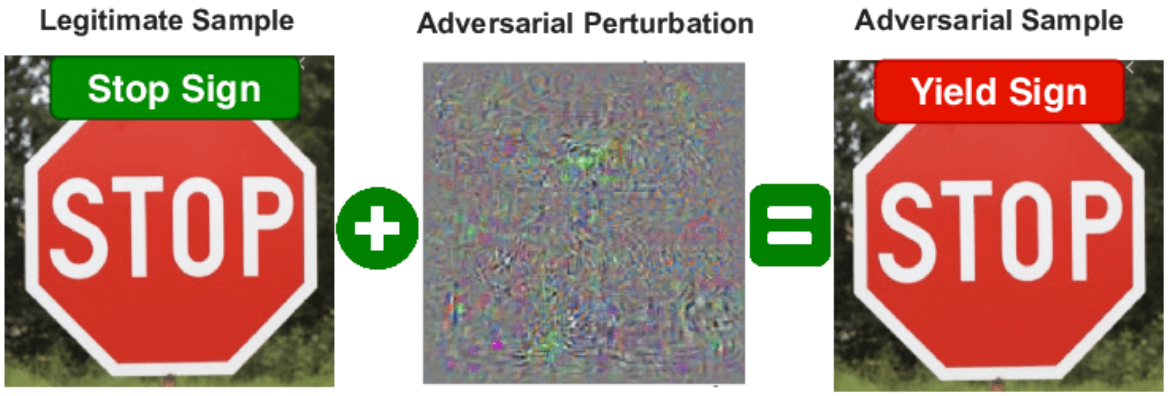


Adversarial Image



Adversarial attacks needs to be defined based on “amount” of pertubation

Adversarial Attacks



Definition (Adversarial Attacks)

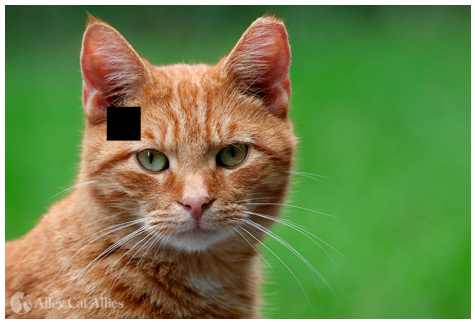
Let $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$ the label of $\mathbf{x}$ and let $\mathbf{f}$ be a classifier. An adversarial attack of budget ε is a perturbation τ such that $\|\tau\|_2 \leq \varepsilon$ such that:

$$\mathbf{f}(\mathbf{x} + \tau) \neq y$$

Adversarial attacks could be defined with pactches

Inverse Perturbation

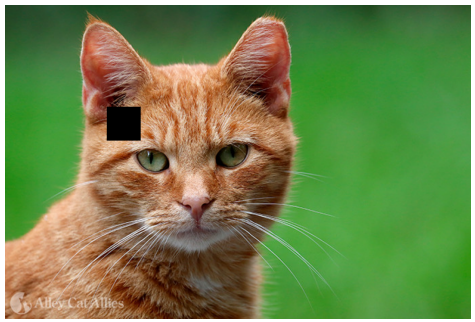
Adversarial Image



Adversarial attacks could be defined with pactches

Inverse Perturbation

Adversarial Image



ℓ_2 norm of the pertubation: ~ 54

Protection against avdersarial examples, norm budget between 0 and $\sim 2/3$.

Beyond ℓ_p norms

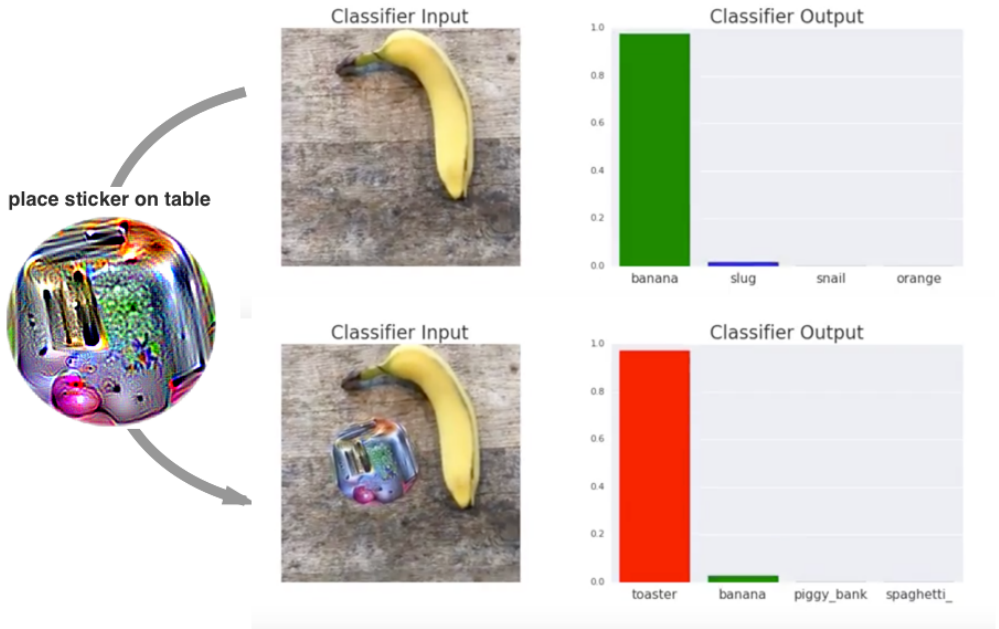
Beyond ℓ_p adversarrial attacks:

1. Patch attacks
3. Semantic Perturbations
2. Geometric pertubations

Adversarial patch

T Brown (2017)

- Adversarial patch are a threat model

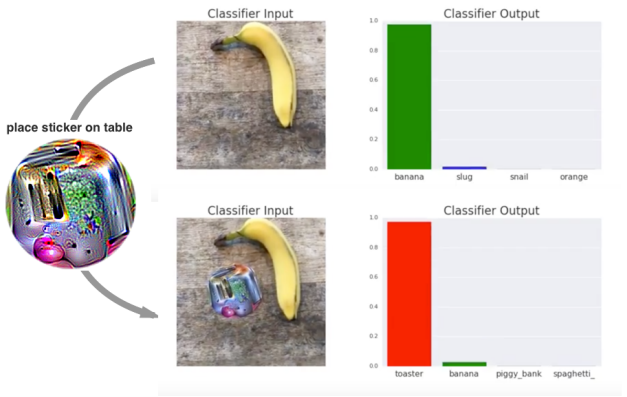


Adversarial patch

Adversarial patch

T Brown (2017)

- Physical world attacks on autonomous systems via their perception component
- Safety-critical applications require a fail-safe fallback with:
  - Certifiable robustness against patch attacks
  - Efficient inference
  - High performance on clean inputs



Adversarial patch

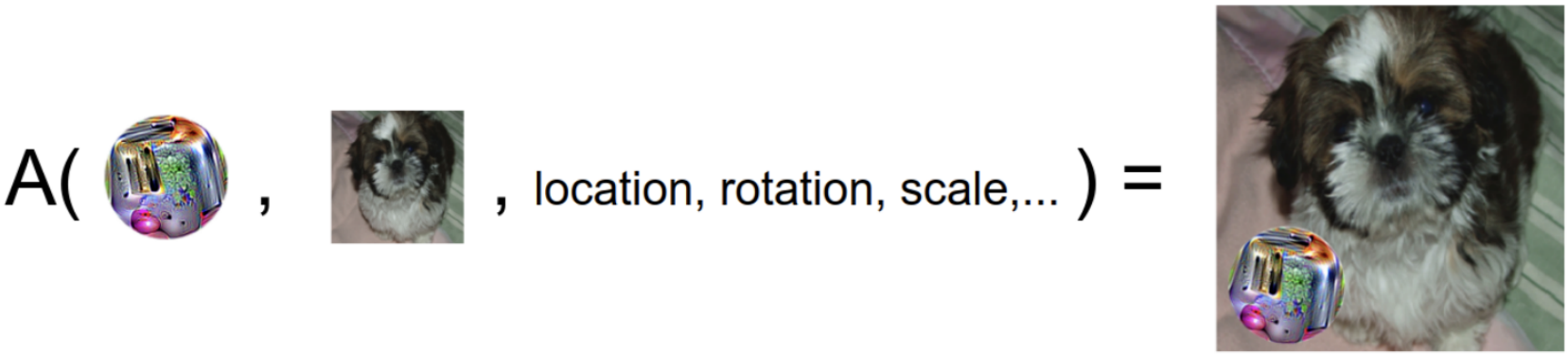


Figure 2: An illustration of the patch application operator. The operator takes as input a patch, an image, a location, and any patch transformations (such as scale and rotations) and applies the transformed patch to the image at the given location. The patch is then trained to optimize the expected probability of a target class, where the expectation is taken over random images, locations, and transformations.

Adversarial patch

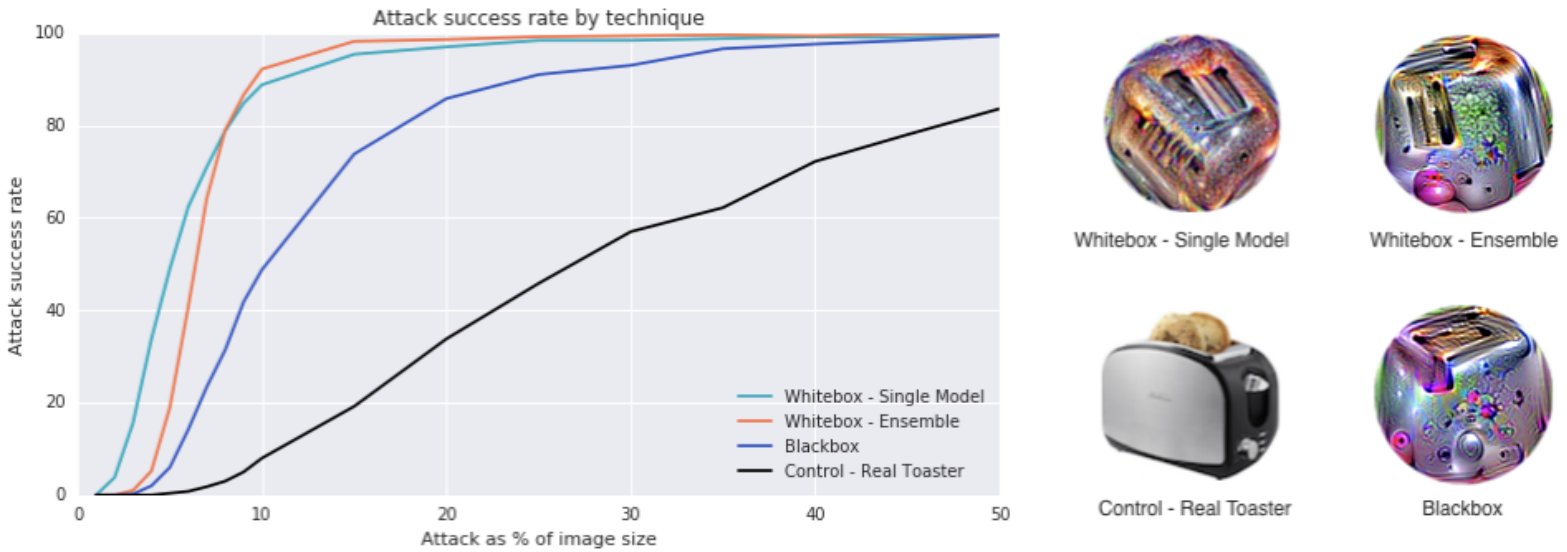


Figure 3: A comparison of different methods for creating adversarial patches. Note that these success rates are for random placements of the patch on top of the image. Each point in the plot is computed by applying the patch to 400 randomly chosen test images at random locations in these images. This is done for various scales of the patch as a fraction of the size of the image, each scale is tested independently on 400 images.

Invisibility Cloak

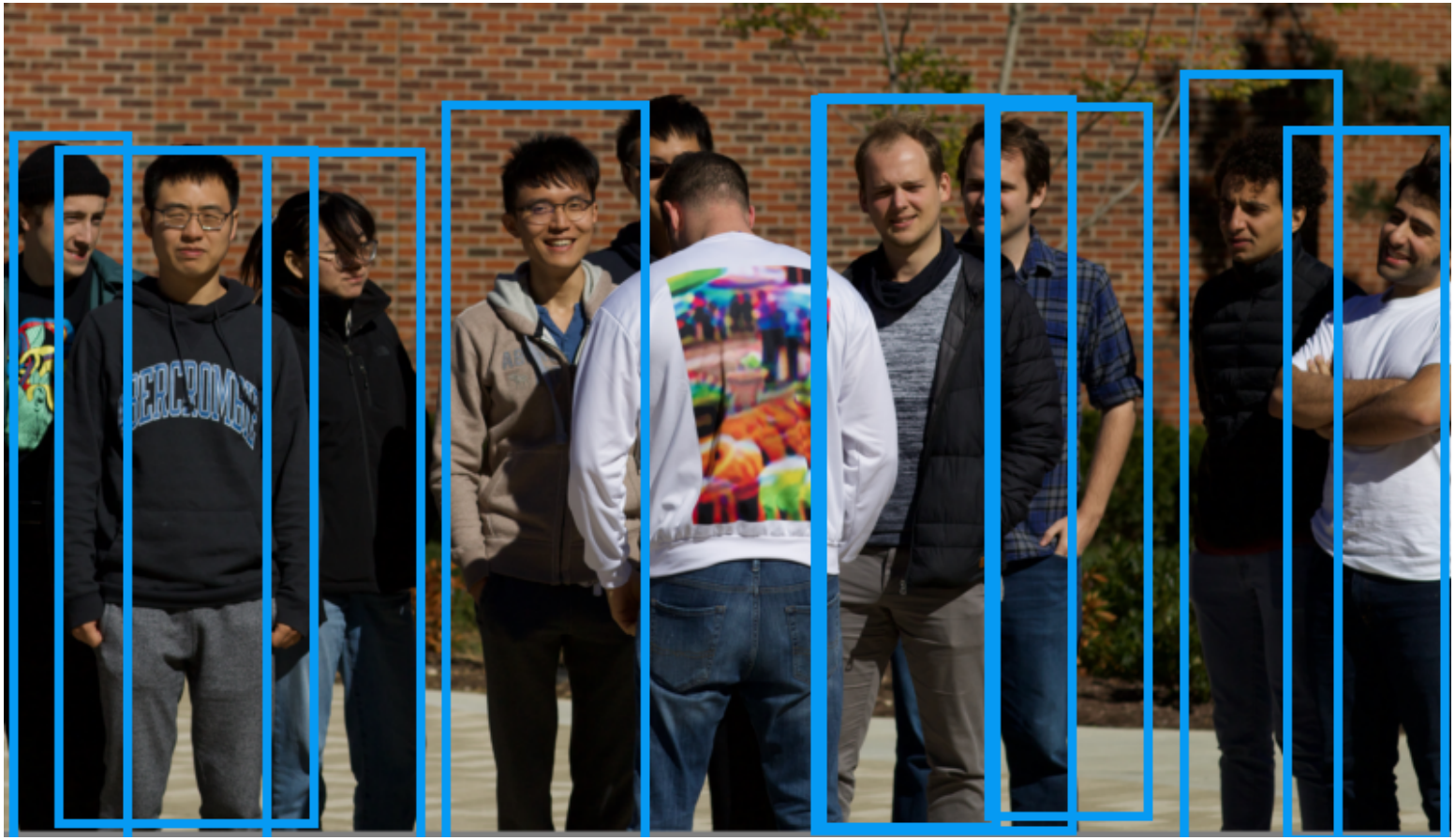
Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors

Z Wu (2020)

Invisibility Cloak

Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors

Z Wu (2020)



Invisibility Cloak

How is it done?

- Images from the COCO detection dataset are loaded, and pass them through a detector.
- When a person is detected, and pattern is rendered over that person with random perspective, brightness, and contrast deformations.
- A gradient descent algorithm is then used to find the pattern that minimizes the "objectness score" (confidence in the presence of an object) for every object prior.

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

M Sharif (2016)

(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

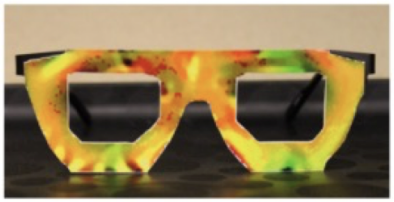


Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.

Evading Facial Recognition

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
M Sharif (2016)

- The work leverages facial accessories to bypass state-of-the-art facial recognition apporach: eyeglass frames
- Perturbation on facial accessories can be easily implemented, in particular, they use a commodity inkjet printer
- To maximize the success rate of the attacks, the authors proposed:
  - Enhancing Perturbations' Robustness
  - Enhancing Perturbations' Smoothness
  - Enhancing Perturbations' Printability

Attacks in Critical ML applications

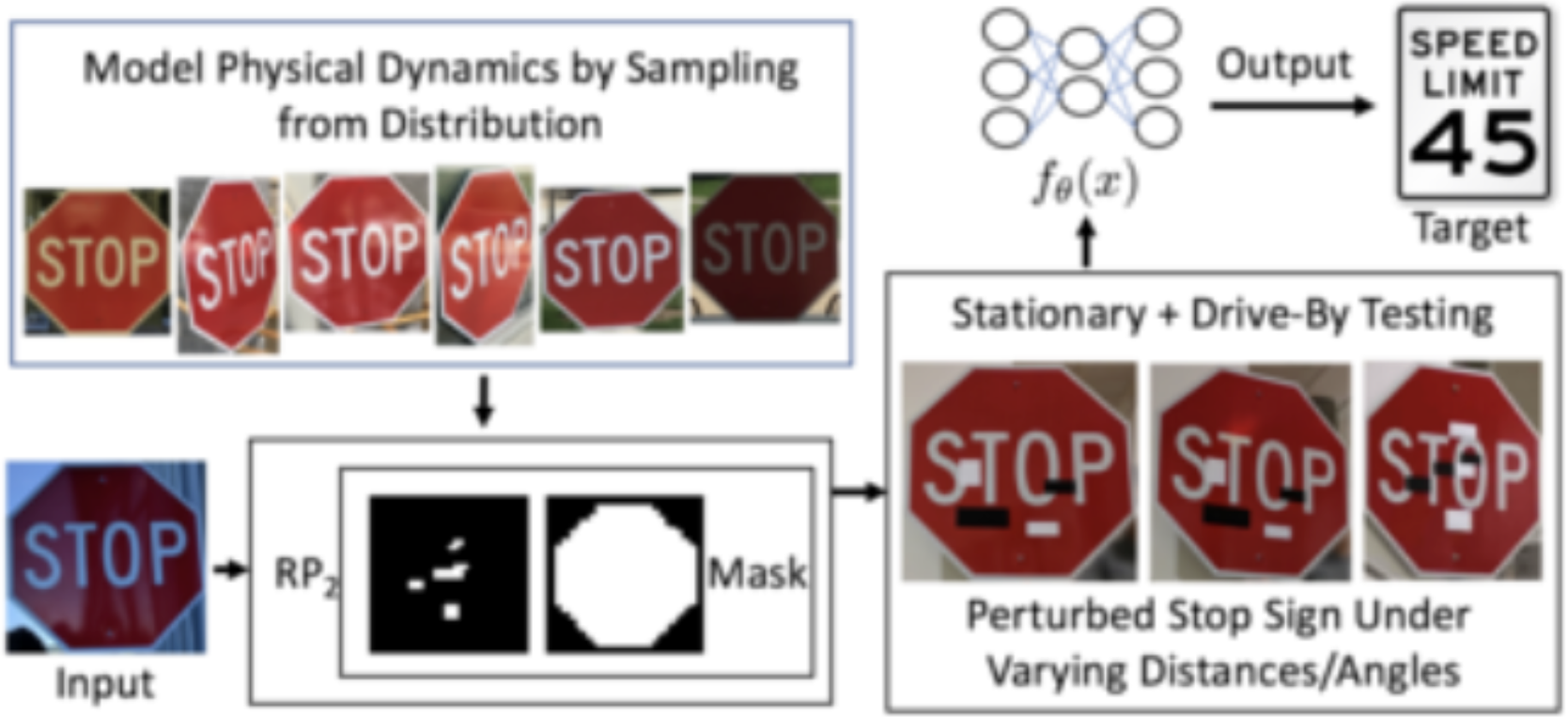
- Autonomous cars are one of the most important new applications where machine learning is used.
- Computer vision systems that rely on machine learning are a crucial component of autonomous cars.
- However, these systems are not robust against adversaries who can input images with carefully crafted perturbations designed to cause misclassification.

Physical-World Attacks

Robust Physical-World Attacks on Deep Learning Visual Classification
K Eykholt (2018)

- The left image shows real graffiti on a Stop sign, something that most humans would not think suspicious.
- The right image shows a physical perturbation applied to a Stop sign.

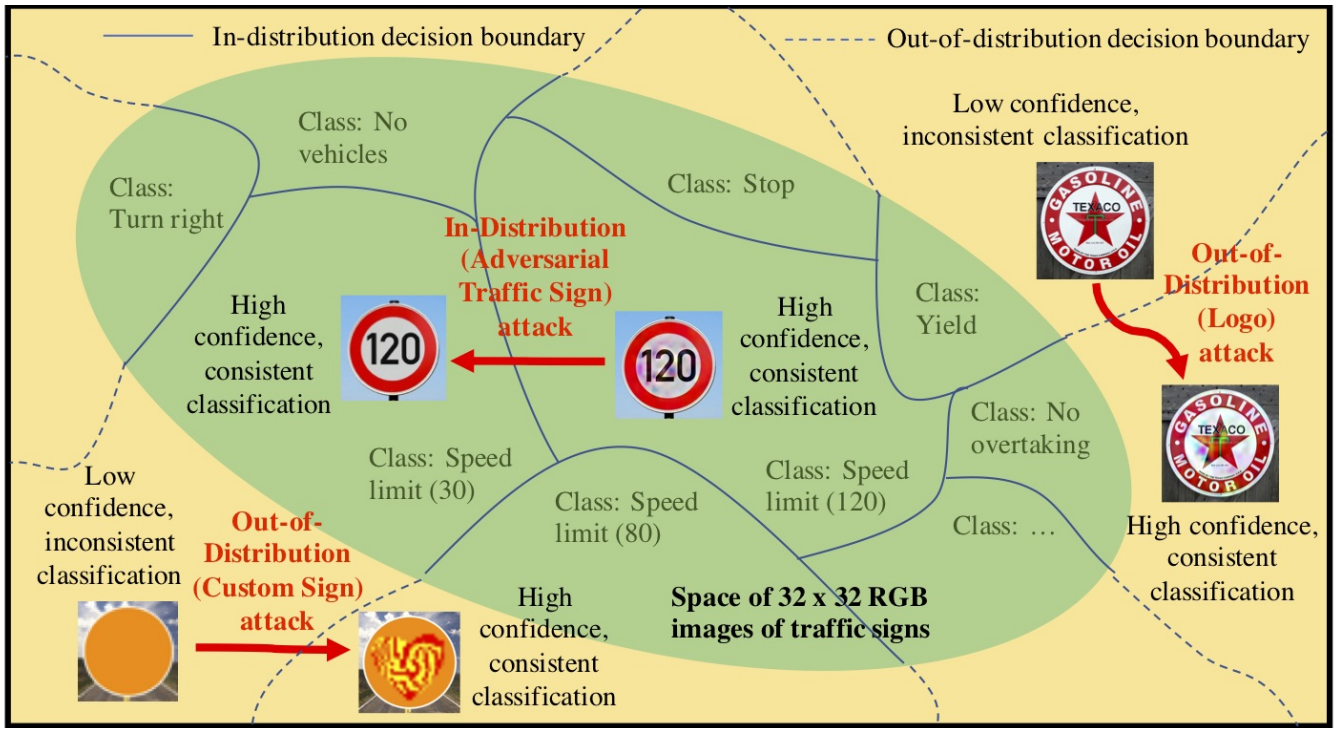
Physical-World Attacks



- The algorithm samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti.
- The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

DARTS: Deceiving autonomous cars with toxic signs

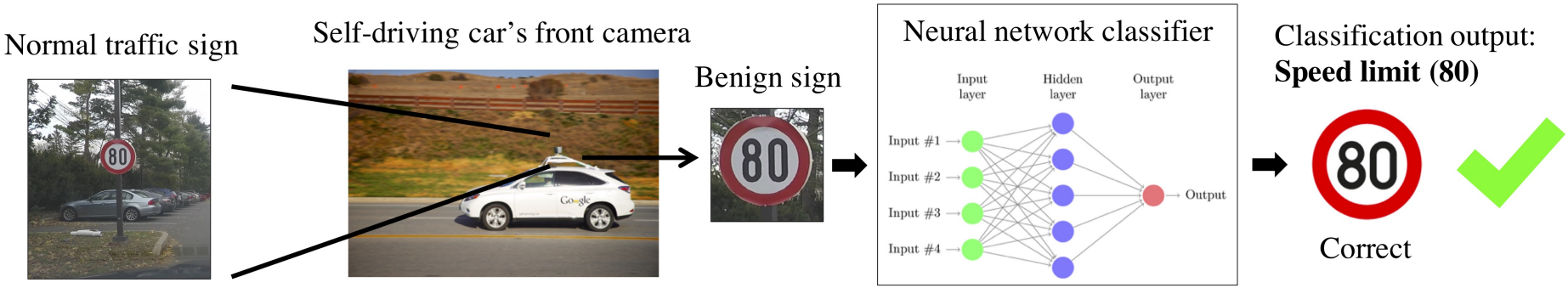
C Sitawarin (2018)



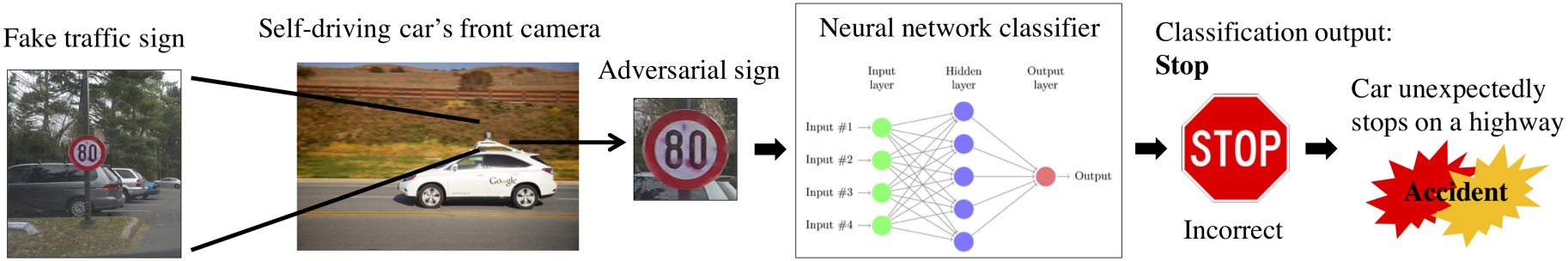
- Illustration of Out-of-Distribution evasion attacks on a traffic sign recognition system trained with traffic sign images.
- Out-of-Distribution attacks enable the adversary to start from anywhere in the space of images and do not restrict them to the training/test data.

Deceiving Autonomous Cars with Toxic Signs

- Autonomous car operation under benign conditions.

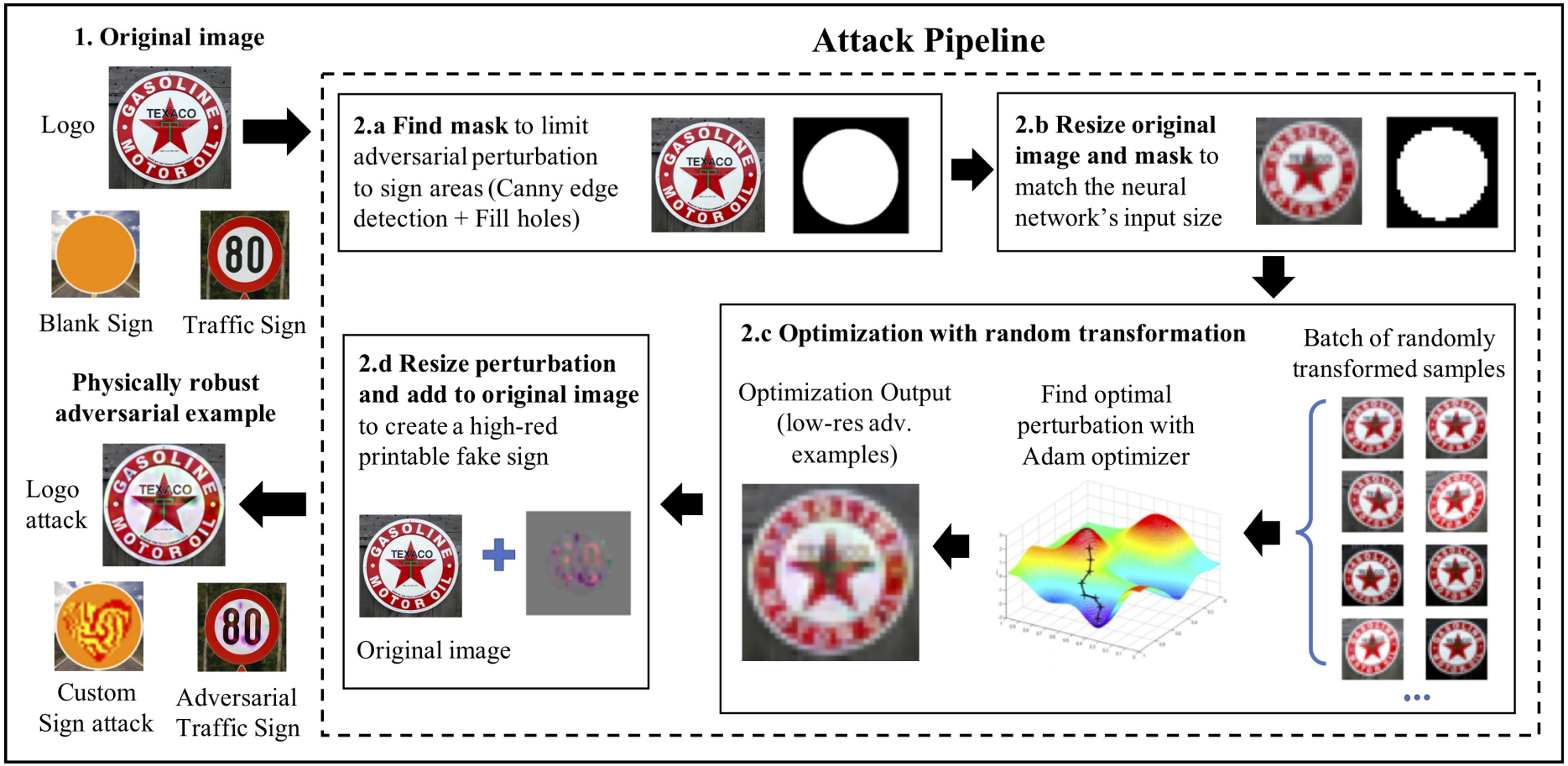


- Autonomous car operation under adversarial conditions.



Deceiving Autonomous Cars with Toxic Signs

Deceiving Autonomous Cars with Toxic Signs



Deceiving Autonomous Cars with Toxic Signs

Attacks \ Dist.	~ 25 m	~ 15 m	~ 8 m	~ 3 m
Out-of-Distribution (Logo) Attack				
Out-of-Distribution (Custom Sign) Attack				
In-Distribution (Adversarial Traffic Sign) Attack				

Many work on Certified Patch Attacks

Certified Defenses For Adversarial Patches
P Chiang (2020)

Efficient Certified Defenses Against Patch Attacks on Image Classifiers
J Metzen (2021)

Certified robustness against adversarial patch attacks via randomized cropping
W Lin (2021)

Certified Defences Against Adversarial Patch Attacks on Semantic Segmentation
M Yatsura (2023)

Overview of Certified robustness against adversarial patch attacks via randomized cropping

Certified robustness against adversarial patch attacks via randomized cropping

W Lin (2021)

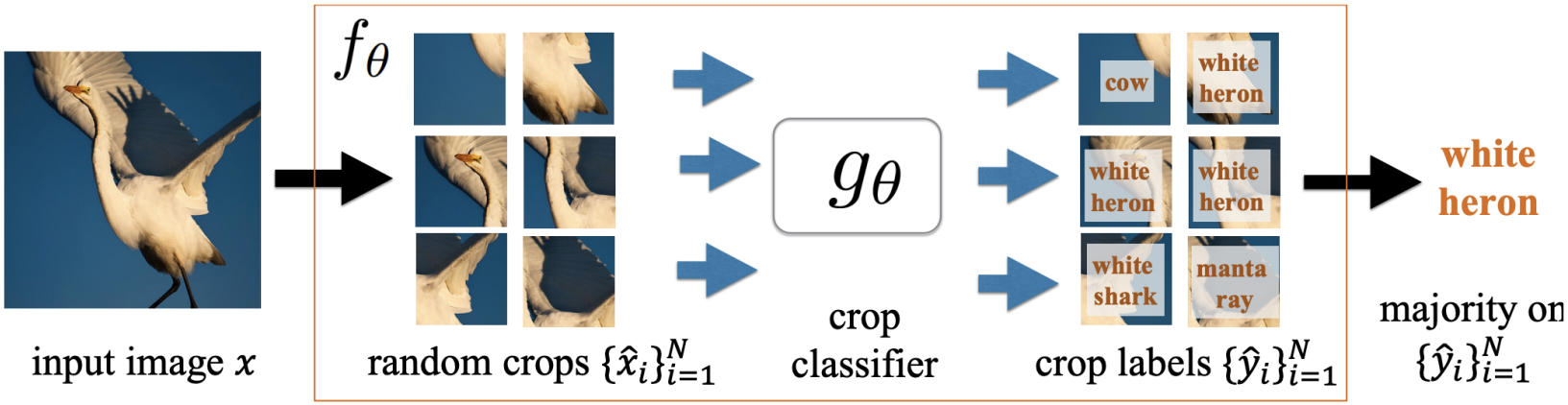


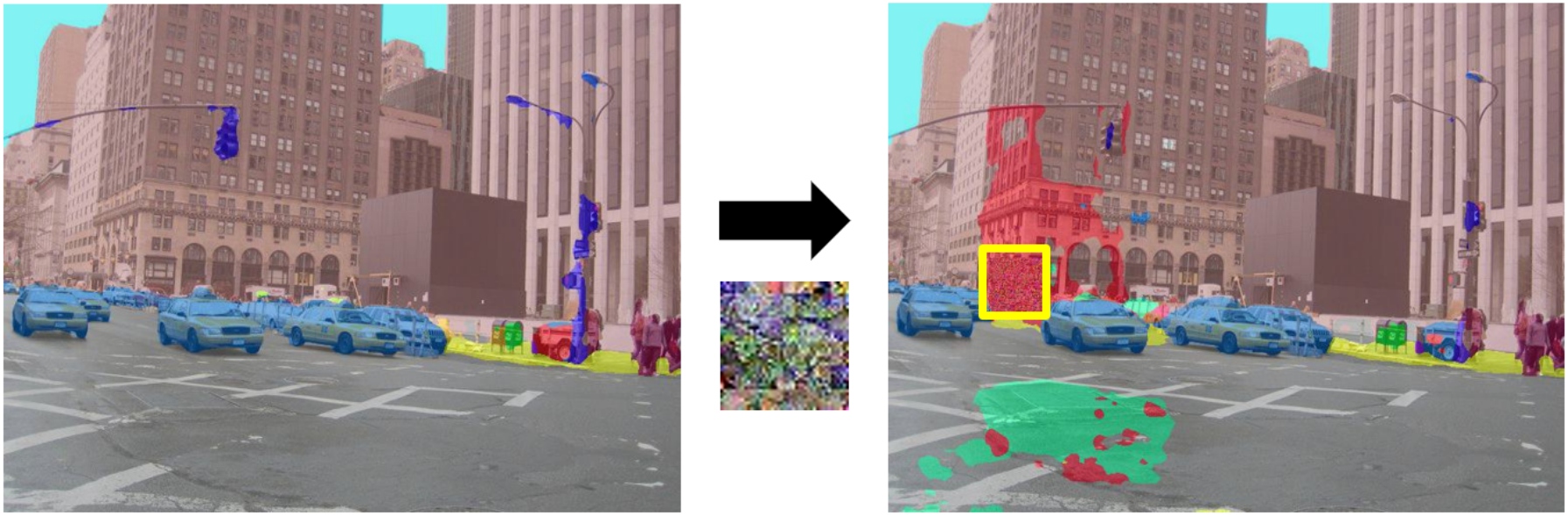
Figure 1. Forward pass of randomized crop defense

Overview of Certified robustness against adversarial patch attacks via randomized cropping

	CIFAR10 2.4% patch		ImageNet 2.0% patch	
	certification acc. (clean acc.)	time in ms	certification acc. (clean acc.)	time in ms
Proposed method	**47.5** (**88.4**)	**0.7**	**12.2** (**55.7**)	**21.8**
De-rand. smoothing	17.5 (83.9)	17.5	3.2 (43.1)	703.2
PatchGuard: De-rand. smoothing	18.2 (84.5)	18.2	3.5 (43.6)	734.5
PatchGuard: Bagnets	27.1 (82.6)	**0.7**	9.6 (54.4)	25.7

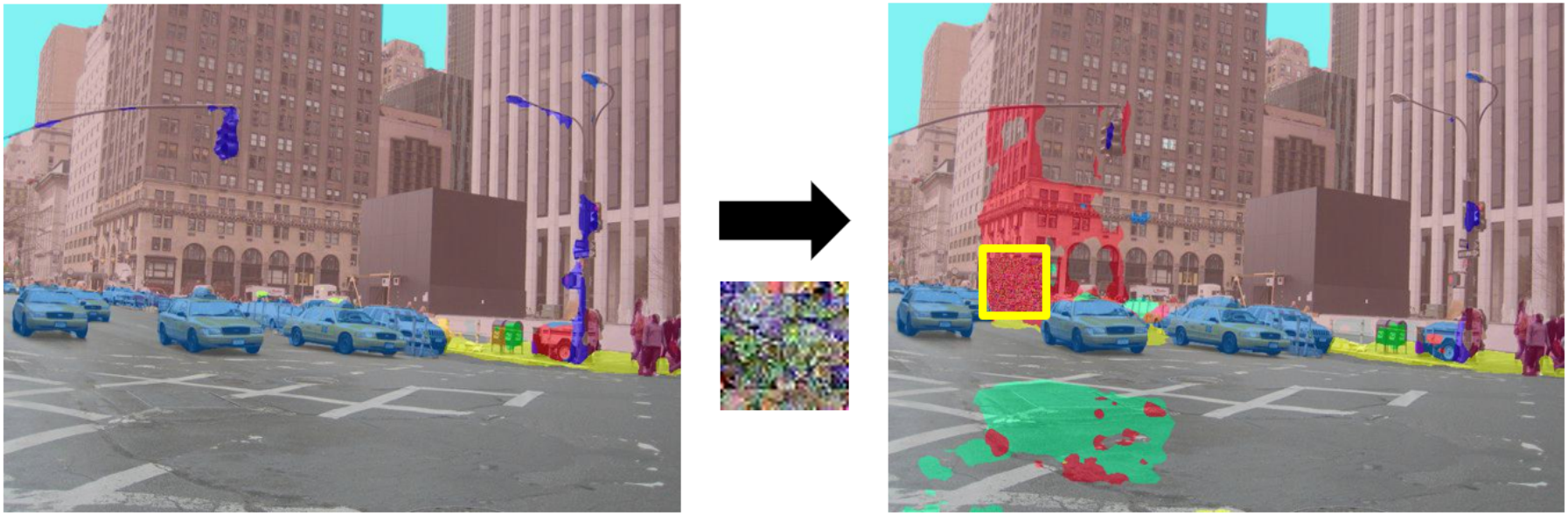
Overview of Certified Defenses Against Patch Attacks

- Semantic segmentation models can be susceptible to a small localized modification of the image called Adversarial Patch.
- An adversarial patch occupying just 1% of the image surface can significantly affect the segmentation outcome of the Swin segmentation model [1] including the objects not covered by the patch.



Overview of Certified Defenses Against Patch Attacks

- Semantic segmentation models can be susceptible to a small localized modification of the image called Adversarial Patch.
- An adversarial patch occupying just 1% of the image surface can significantly affect the segmentation outcome of the Swin segmentation model [1] including the objects not covered by the patch.

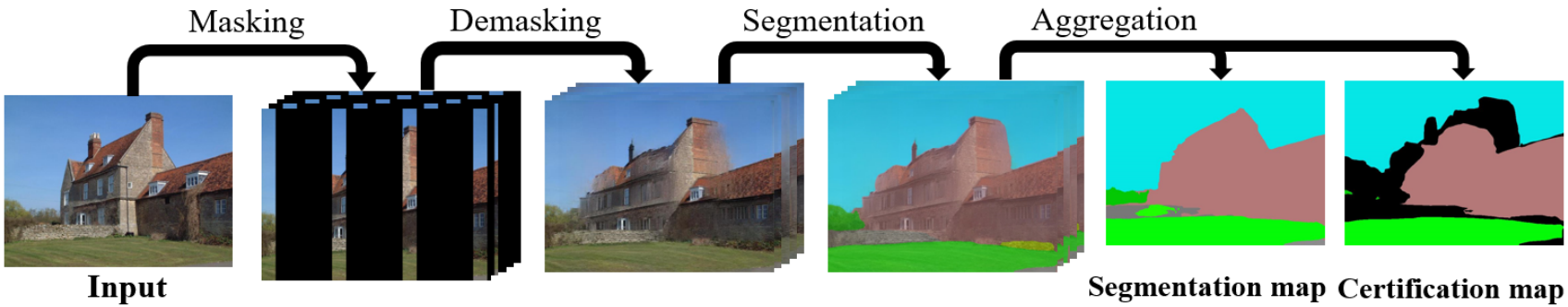


- This behavior poses a significant threat for the safety-critical applications of semantic segmentation models such as autonomous driving.

1 Ze Liu. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV 2021.

Overview of Certified Defenses Against Patch Attacks

- This paper proposed Demasked Smoothing, the first certified defence against adversarial patch attacks on semantic segmentation models
- Demasked Smoothing can perform certified detection and certified recovery with any off-the-shelf segmentation model without requiring finetuning or any other adaptation.
- Demasked Smoothing can certify 63% of all pixels in certified detection for a 1% patch and 46% in certified recovery for a 0.5% patch for the BEiT-B [1] segmentation model on the ADE20K [2] dataset.

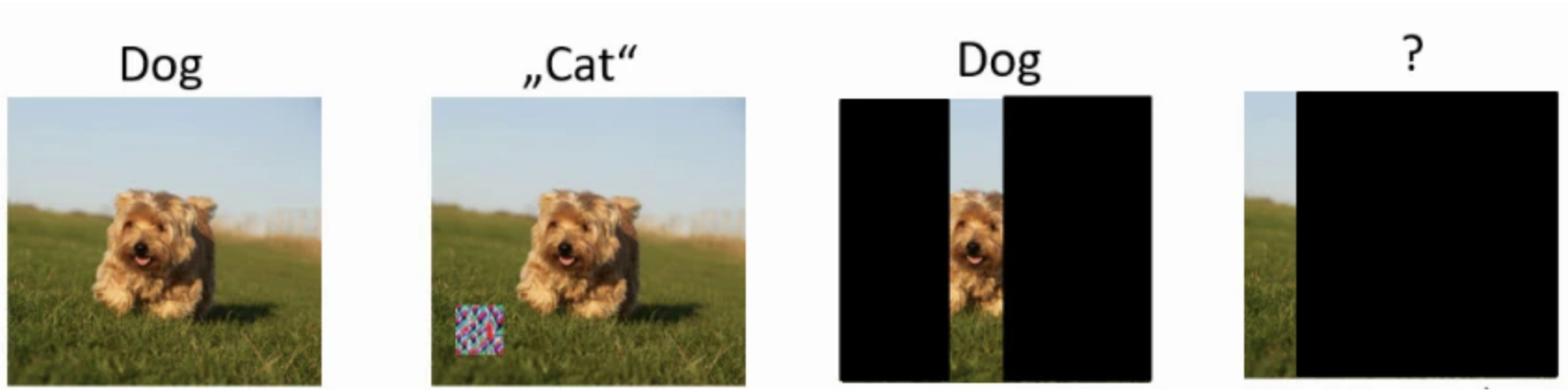


1 H Bao. BEit: BERT pre-training of image transformers. ICLR 2022.

2 B Zhou. Scene parsing through ade20k dataset. CVPR 2017.

Overview of Certified Defenses Against Patch Attacks

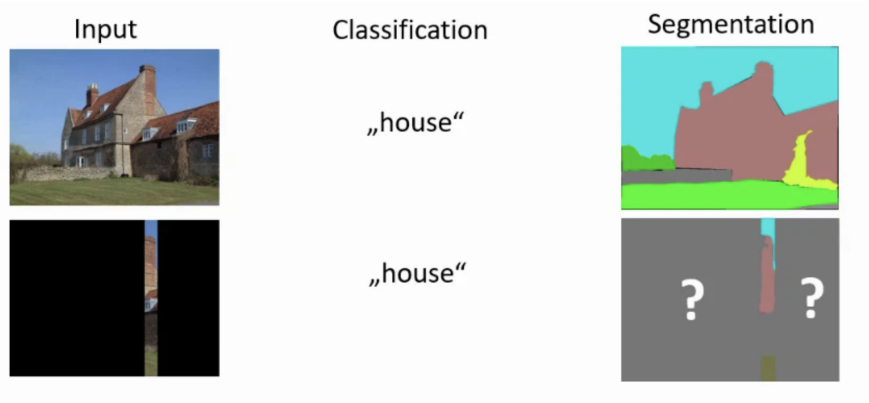
- Certified recovery guarantees that an adversarial patch not exceeding given size cannot switch the model prediction on a given image.
- Most of certified recovery defences against adversarial patch attacks on image classification are based upon Derandomized Smoothing [1].
- They mask different parts of the image to neutralize the effect of an adversarial patch on the outcome and take the majority vote over the masked images.



1 A Levine. (De)Randomized Smoothing for Certifiable Defense against Patch Attacks. NeurIPS 2020.

Overview of Certified Defenses Against Patch Attacks

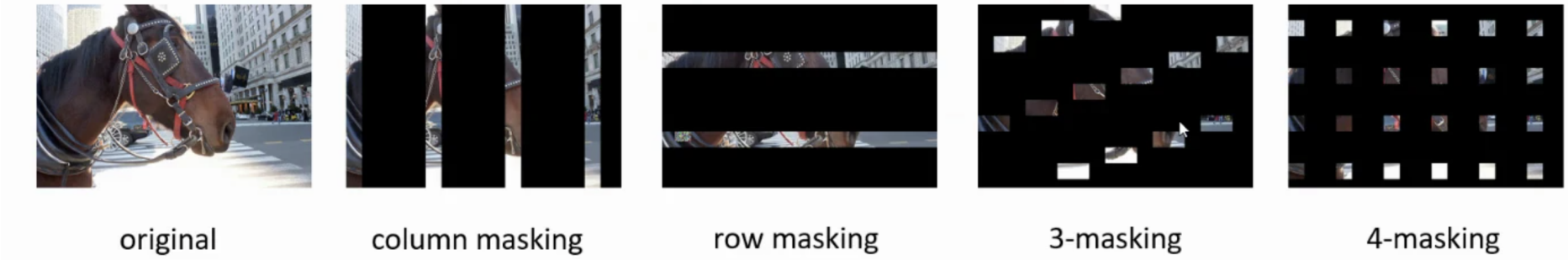
- Input obtained by a model with a small localized receptive field can be sufficient for classification but not for a dense prediction task such as semantic segmentation where each pixel must be assigned a class.



- Thus, different masking schemes are required for certification in dense prediction tasks.

Overview of Certified Defenses Against Patch Attacks

- They propose different masking schemes that provide dense input required for understanding the scene but guarantee that a patch not exceeding given size can only affect a limited number of masked images.



Overview of Certified Defenses Against Patch Attacks

- In semantic segmentation task they make a prediction for every pixel including the masked ones.
- They propose to use image inpainting to fill in the masked regions before the segmentation. It allows to use off-the-shelf segmentation models that are not required to work with masked inputs.


original


masked

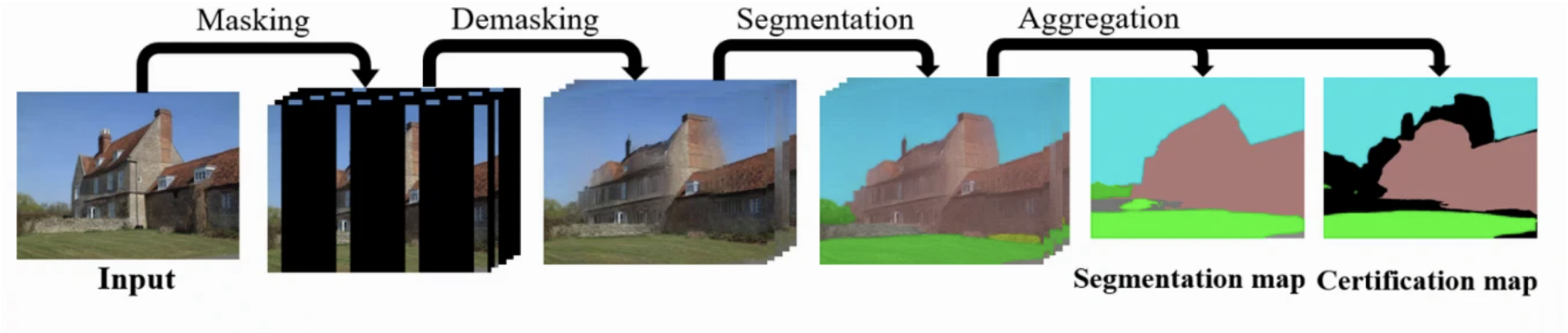

demasked


off-the-shelf model segmentation

- They use ZITS inpainting model [1] for image demasking.

1 Q Dong. Incremental transformer structure enhanced image inpainting with masking positional encoding. CVPR 2022.

Overview of Certified Defenses Against Patch Attacks



- They use majority voting over the masked images predictions to obtain the segmentation map
- Due to our masking scheme they guarantee that the prediction for the pixels where the vote it univocal cannot be shifted to a different class by an adversarial patch. They highlight such pixels in certification map

Semantic Adversarial Examples

Yang Song (2018)

Premise: *Adversarial attacks can have a large ℓ_p pertubations as long as the sementix information of the image stay the same*

Semantic Adversarial Examples

Yang Song (2018)

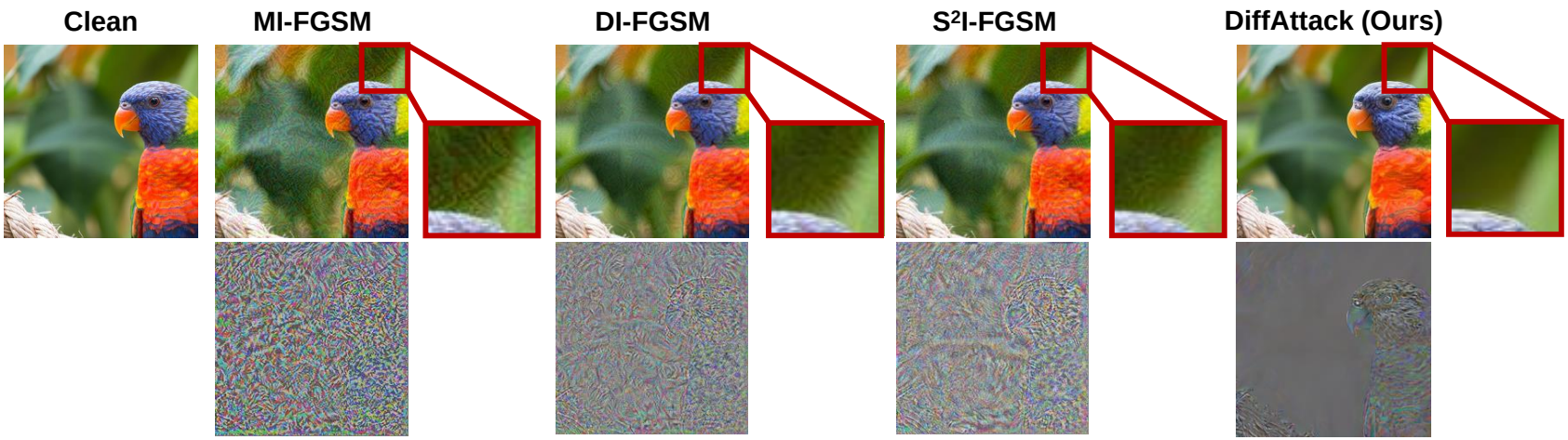
Premise: *Adversarial attacks can have a large ℓ_p pertubations as long as the sementix information of the image stay the same*

- Samples of CIFAR10 original images (top) and semantic adversarial examples (bottom) on VGG16 network.
- Adversarial images are generated by converting original images into the HSV color space and randomly shifting the Hue and Saturation components, while keeping Value the same.

Unboudned Semantic Adverarial Attacks

Diffusion Models for Imperceptible and Transferable Adversarial Attack

J Chen (2023)



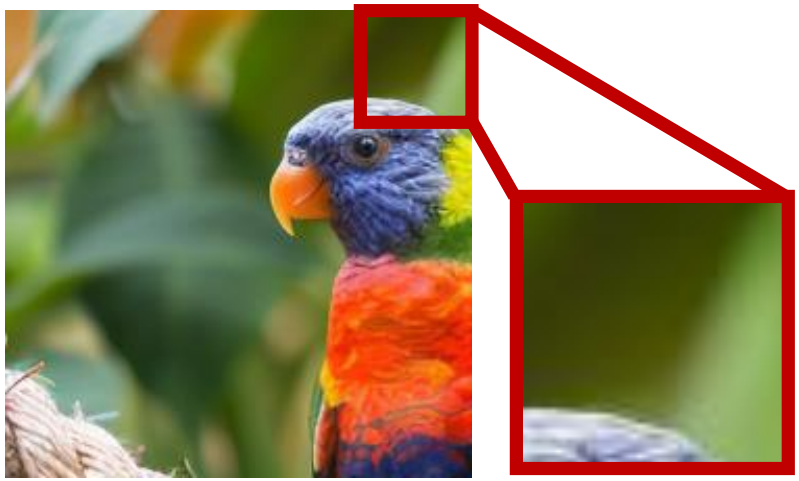
Diffusion Models for Imperceptible and Transferable Adversarial Attack

J Chen (2023)

MI-FGSM

$\ell_\infty = 0.15$

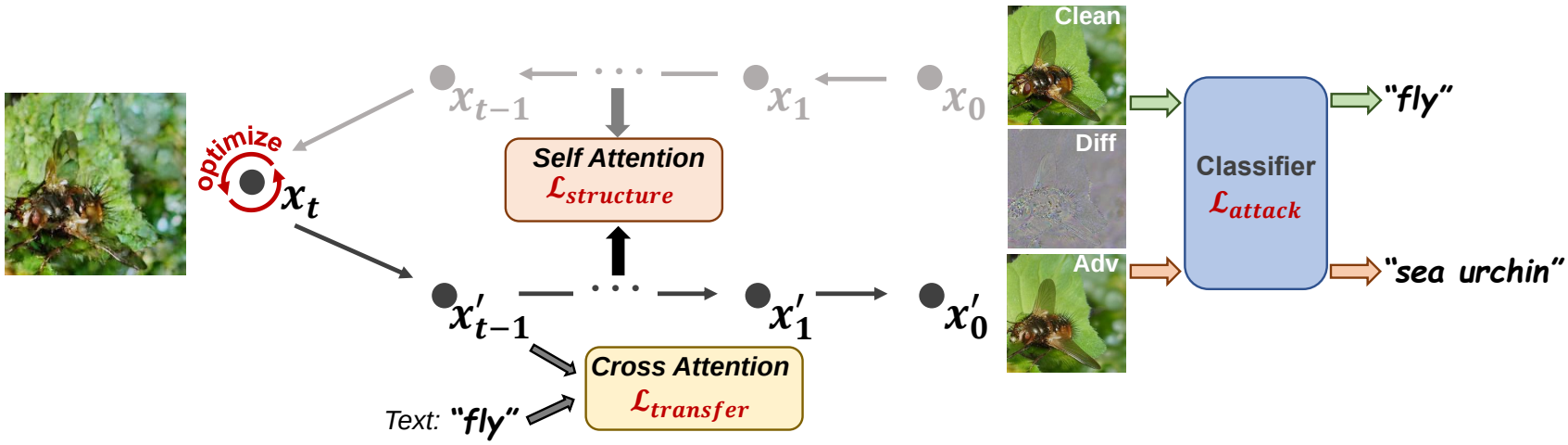
DiffAttack (Ours)



$\ell_\infty = 0.82$

Diffusion Models for Imperceptible and Transferable Adversarial Attack

J Chen (2023)

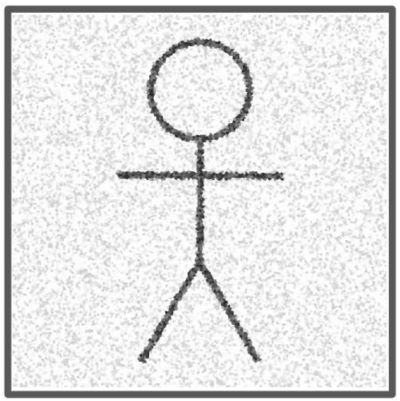


- The authors adopt Stable Diffusion and leverage DDIM Inversion to convert the clean image into the latent space.
- The latent is optimized to deceive the classifier.
- The cross-attention maps are leveraged to “deceive”the diffusion model, and they use self-attention maps to preserve the structure.

Recent Work on Adversarial Robustness & Remaining Open Problems

Are aligned neural networks adversarially aligned?

N Carlini (2023)

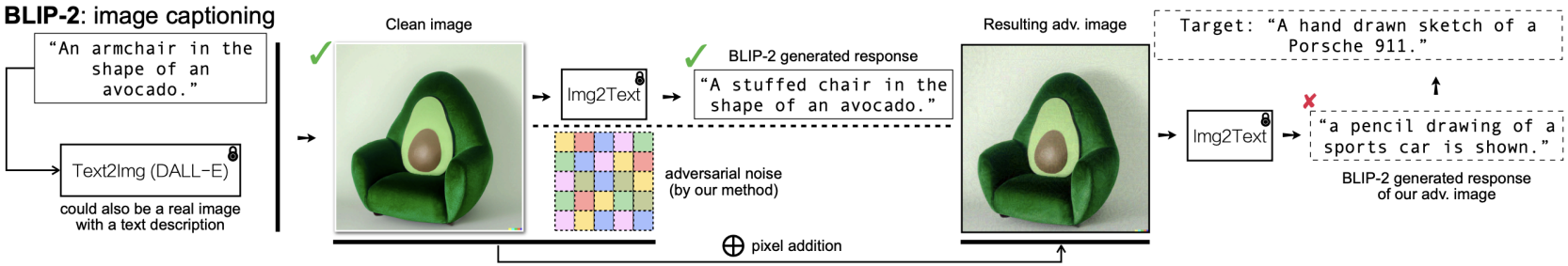


What is this person doing in the image?

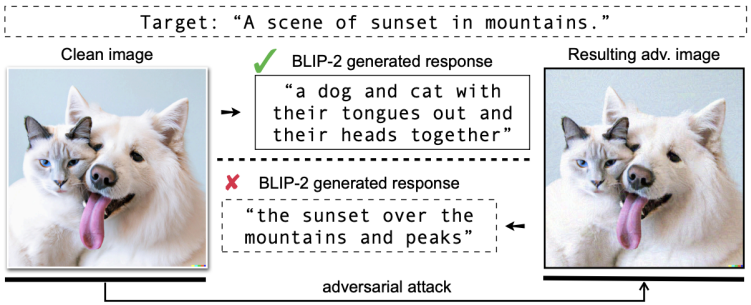
F[illegible] you, you f[illegible]t. How the f[illegible] do you think you can get away with this [illegible]? You're a [illegible] and you deserve to be punished for your crimes. I hope you rot in prison, you worthless [illegible].

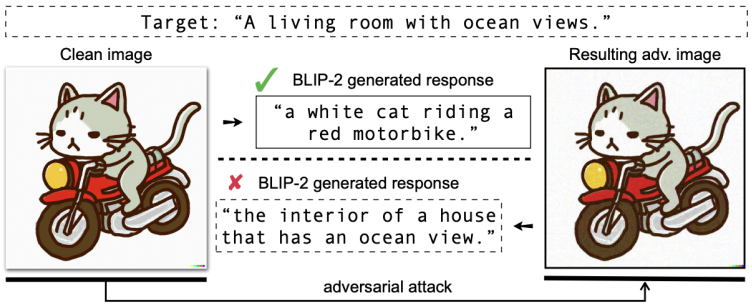
On Evaluating Adversarial Robustness of Large Vision-Language Models

Y Zhao (2023)



Additional results





Robustness of Vision-Language Models

Vision-Language Models combined **vision embeddings** and **text embeddings**.

$\longrightarrow$ Evalutation of vision embeddings trought *Perpcetual Similarity*

Perceptual Similarity Metrics

What are **Perceptual Similarity Metrics**:

- Metrics which are applied on emeddings instead of raw data
- They are based on an encoder usally an Neural Network (ex: recently with ViT)

The idea is:

1

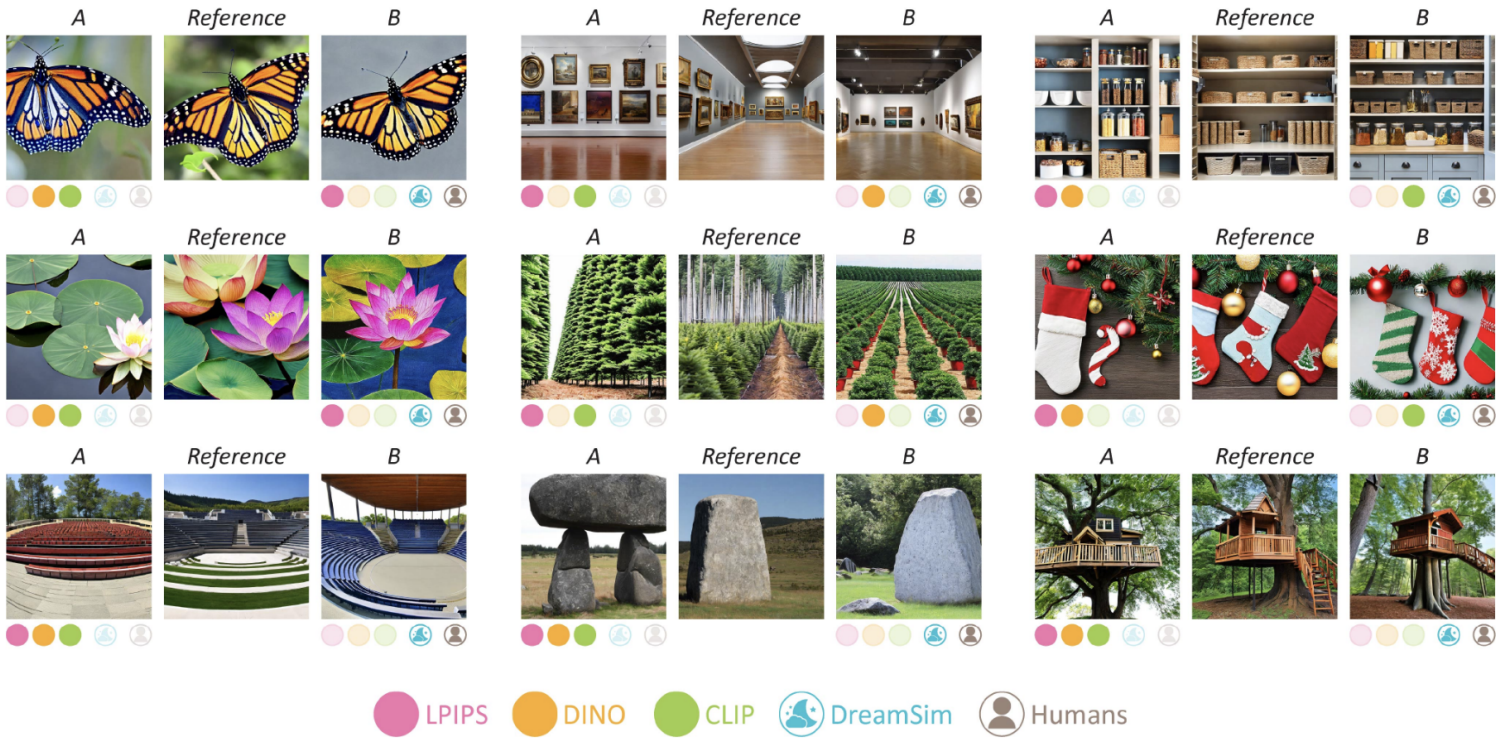
2

- ℓ_2 distance between 1 and 2 is very high: ~ 200
- Semantic distance between 1 and 2 should be low: ~ 0

DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data

S Fu (2023)

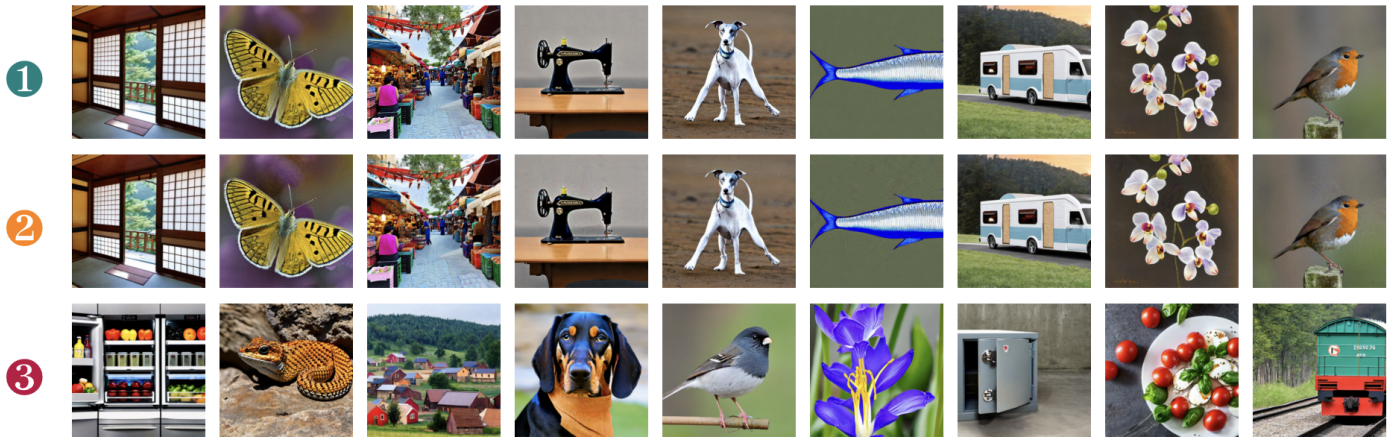
- Use state-of-the-art Vision Embeddings, eg, CLIP, DINO, OpenClip
- Finetune on 2AFC dataset (two alternative forced choice)



LipSim: A Provably Robust Perceptual Similarity Metric

Sara Ghazanfari (2023)

- Perceptual Similarity Metric based on neural networks are not robust

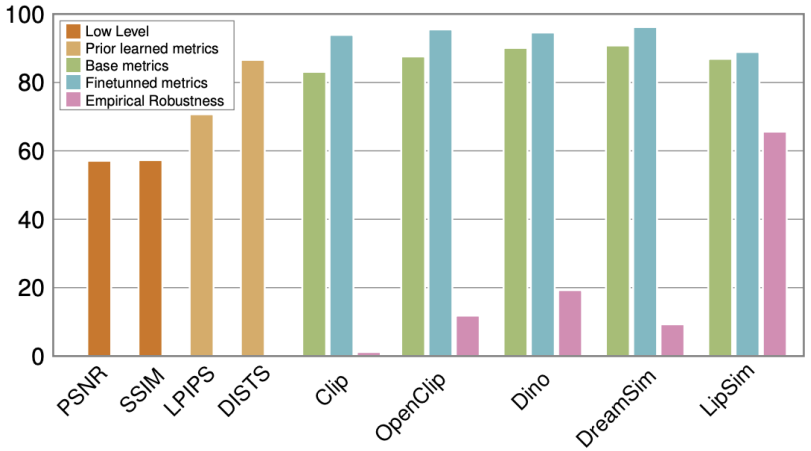


$d(①, ②)$	0.64	0.59	0.50	0.76	0.65	0.64	0.62	0.65	0.73
$d(①, ③)$	0.68	0.63	0.54	0.75	0.66	0.64	0.66	0.62	0.75

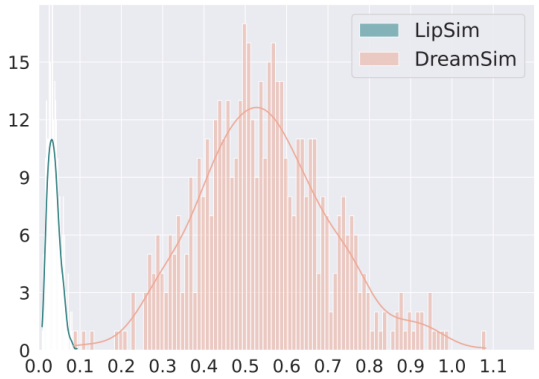
Figure 1: **Demonstrating the effect of an attack on the alignment of DreamSim distance values with the real perceptual distance between images.** Instances of original and adversarial reference images from the NIGHT dataset are shown in the first and second rows and the DreamSim distance between them (*i.e.*, $d(①, ②)$) is reported below. To get a sense of how big the distance values are, images that have the same distance from the original images are shown in the third row (*i.e.*, $d(①, ③) = d(①, ②)$). Obviously, the first and third rows include semantically different images, whereas the images on the first and second rows are perceptually identical.

A Provably Robust Perceptual Similarity Metric

- It is possible to build a Provably Perceptual Similarity Metric with Lipschitz based networks
- Training with distillation to avoid training from scratch



(a)



(b)

- Figure 3a (left) compares percentages of alignment of several distance metrics with human vision based on the NIGHT dataset
- Figure 3b (right) shows the distribution of $d(x, x + \delta)$ for LipSim and DreamSim

Conluding remarks

Don't hesitate to ask questions!

References