# Robustness in Deep Learning

**Murari Mandal**
Postdoctoral Researcher
National University of Singapore (NUS)

https://murarimandal.github.io

# Robustness

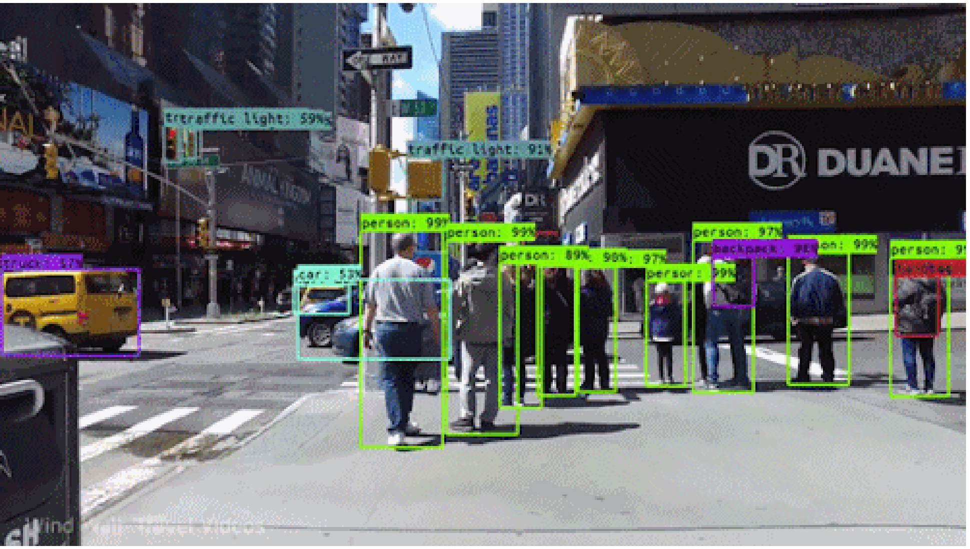
## “robustness”?

- the ability to withstand or overcome adverse conditions or rigorous testing.

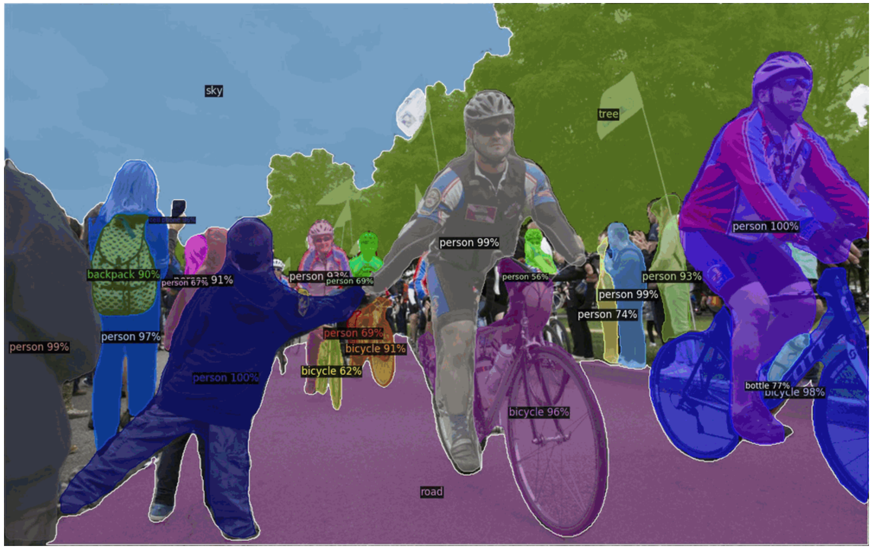
- Are the current deep learning models robust?

- Adversarial example: An input data point that is slightly perturbed by an *adversarial perturbation* causing failure in the deep learning system.

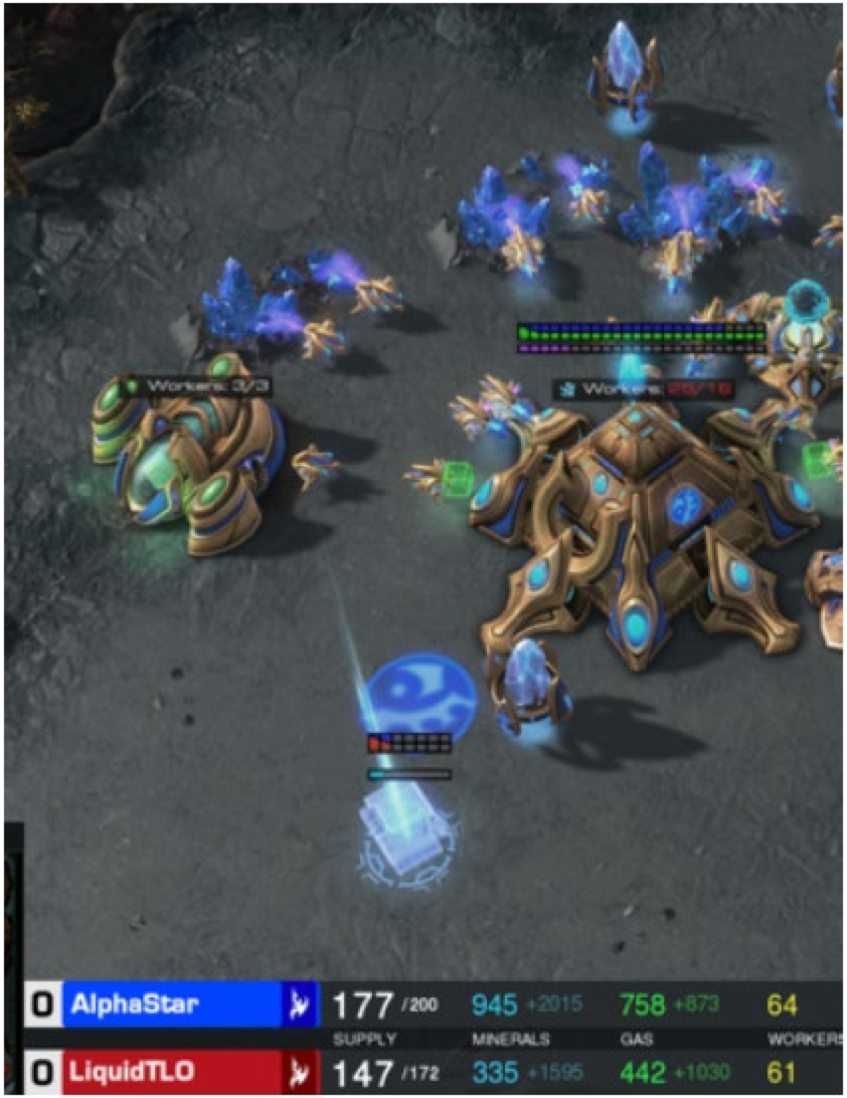
# The AI Breakthroughs



Redmon et al. “YOLO9000: Better, Faster, Stronger”

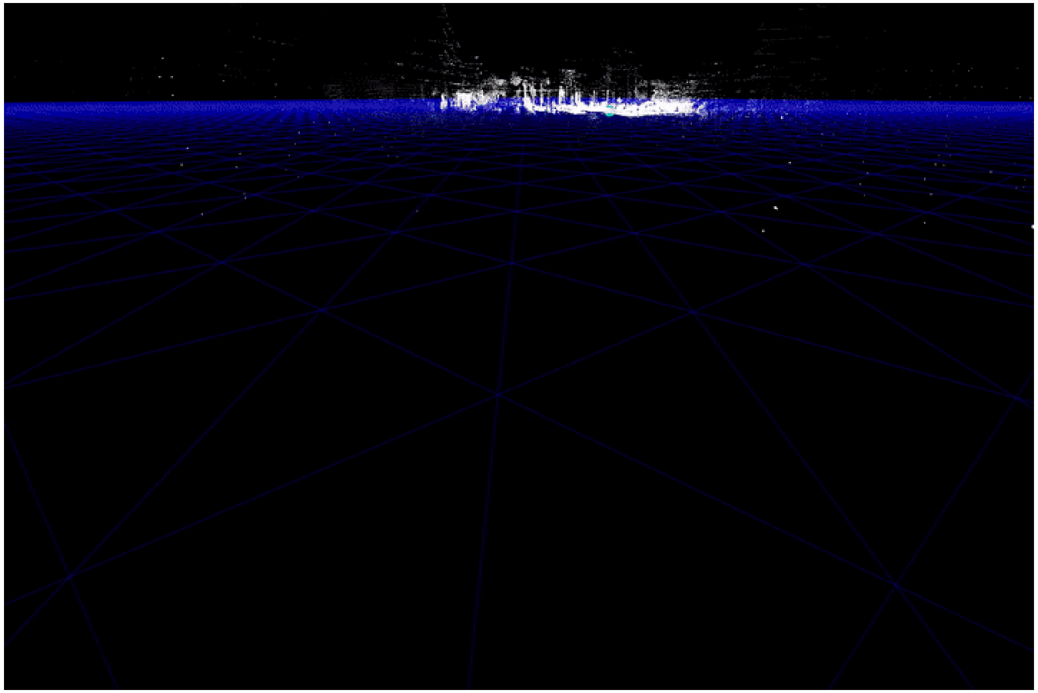


https://github.com/facebookresearch/detectron2


Workers 3/3
Workers 23/16
0 AlphaStar 177 /200 945 +2015 758 +873 64
SUPPLY MINERALS GAS WORKERS
0 LiquidTLO 147 /172 335 +1595 442 +1030 61

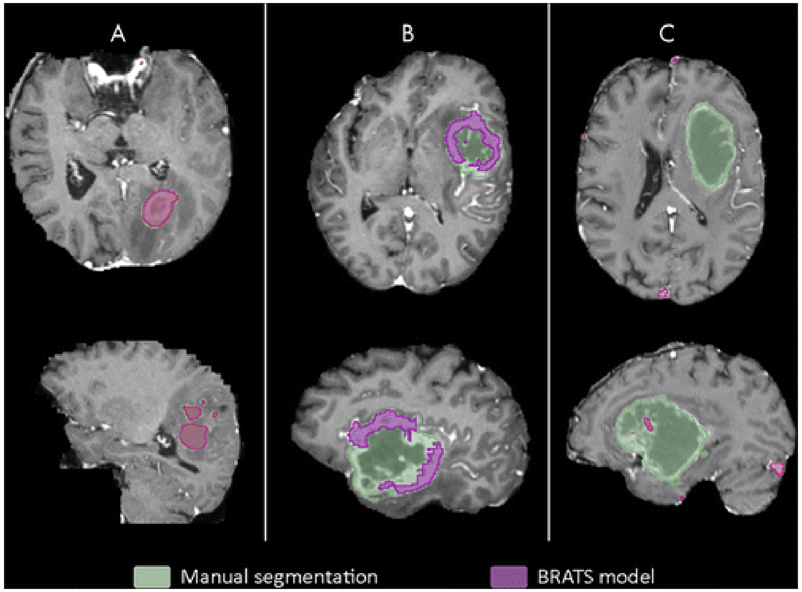
Vinyals et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”

# Higher Stakes?

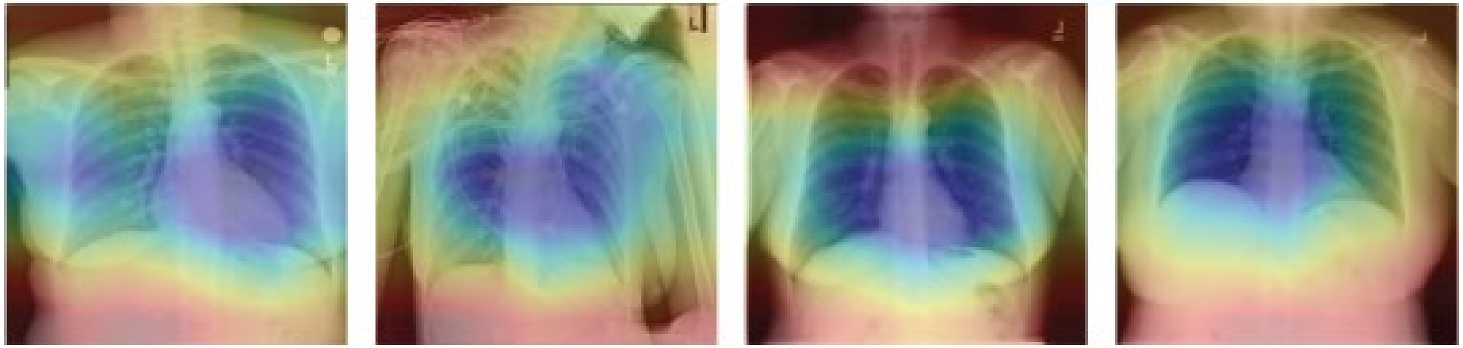


Autonomous Driving

https://scale.com/

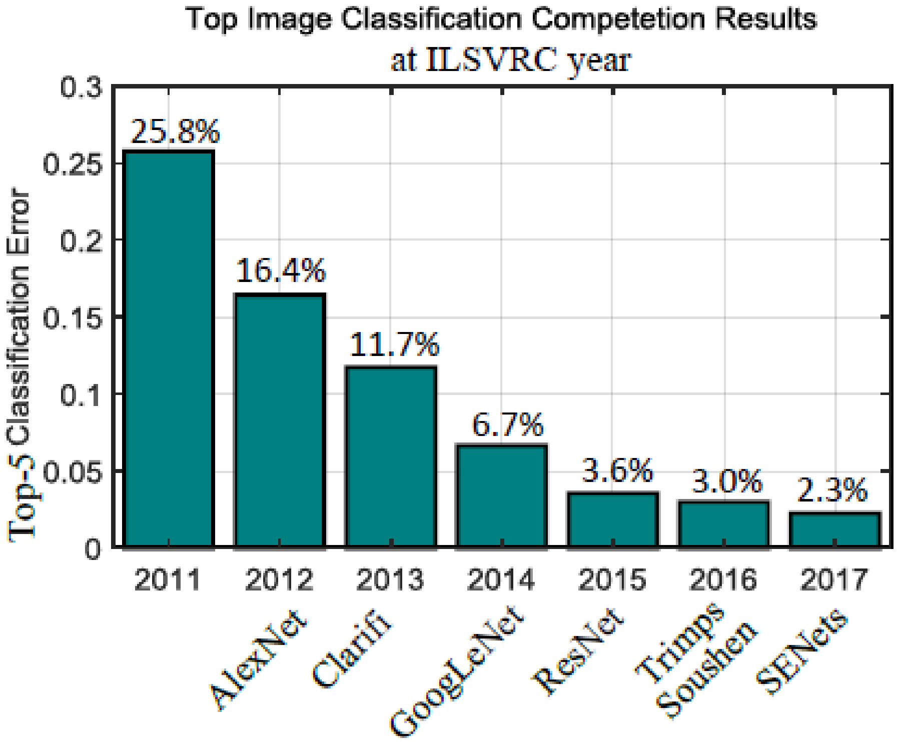


Eijgelaar et al. “Robust Deep Learning–based Segmentation of Glioblastoma on Routine Clinical MRI Scans…”

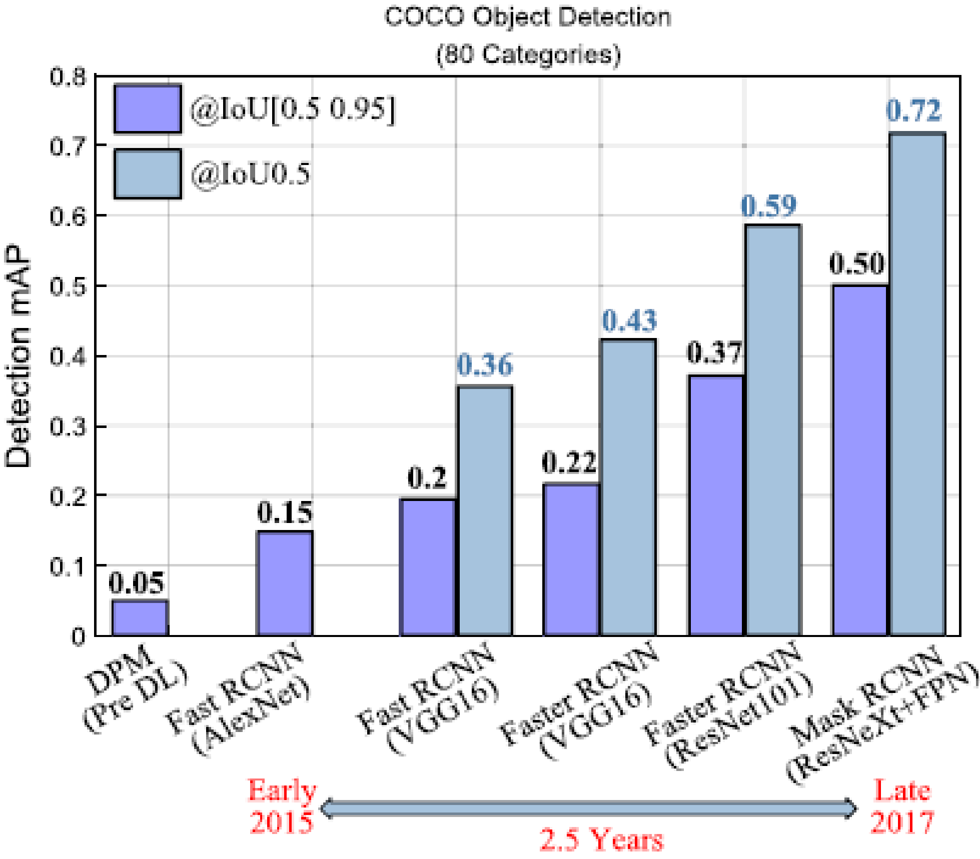


Tang et al. “Data Valuation for Medical Imaging Using Shapley Value: Application on A Large-scale Chest X-ray dataset”

# Better Performance!



Performance of winning entries in the ImageNet from 2011 to 2017 in the *image classification* task.

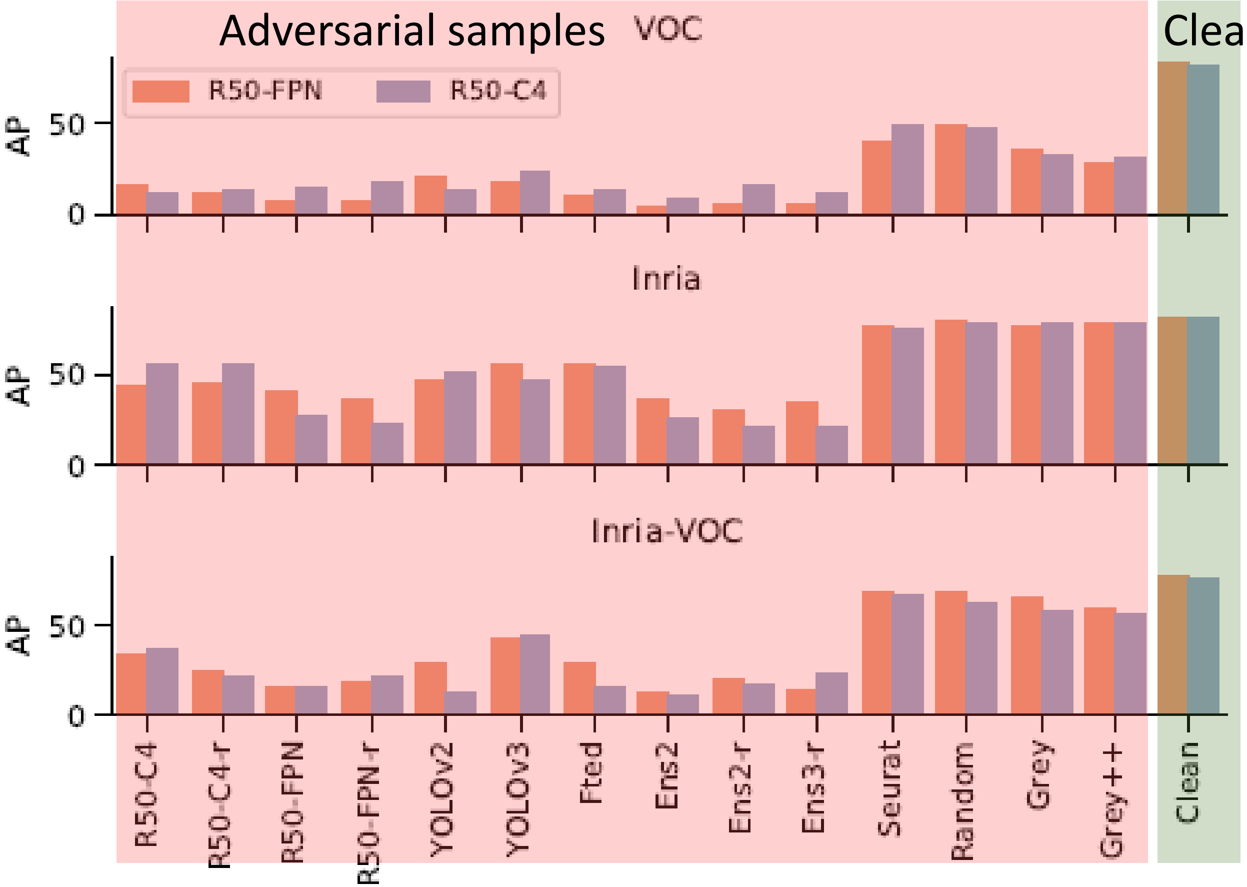


Evolution of *object detection* performance on COCO (Test-Dev results)

Liu et al. “Deep Learning for Generic Object Detection: A Survey”

# Are the Models Robust?

- The degrees of robustness or adaptability is quite low!
- Human Perception Vs Machine/Deep Learning Performance?

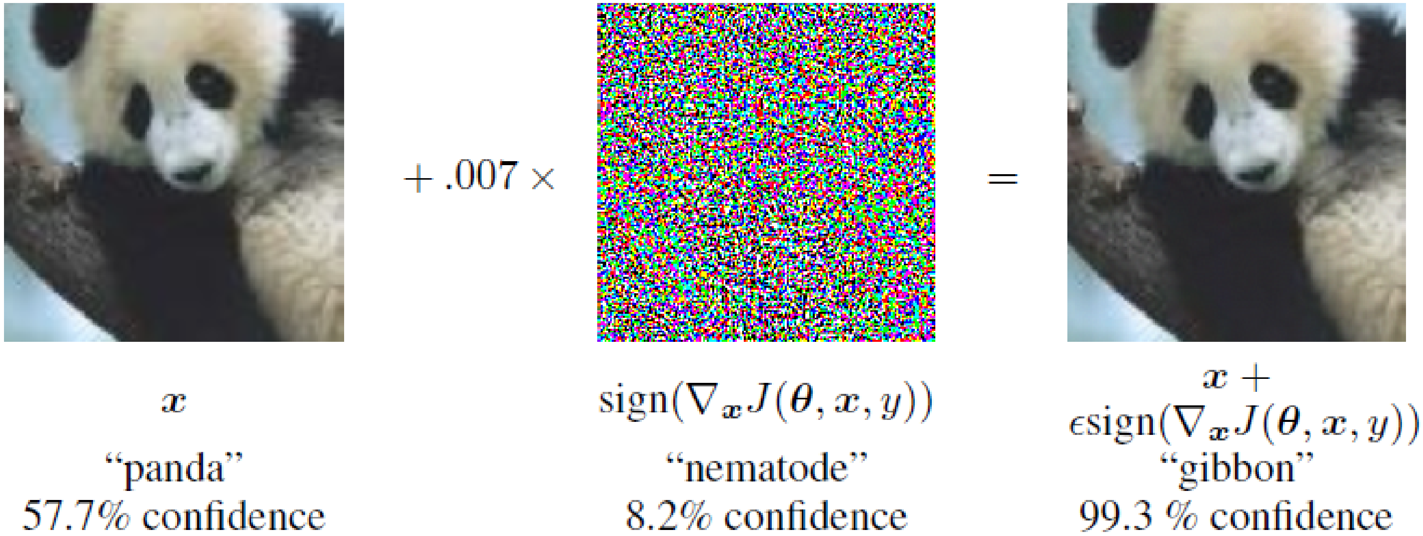


Results of different patches, trained on COCO, tested on the person category of different datasets.

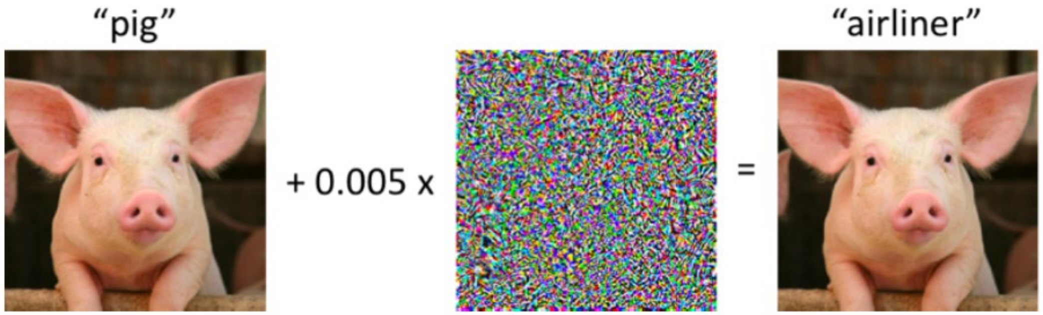
Wu et al. “Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors”

# Adversarial Attacks

- Deep neural networks have been shown to be vulnerable to *adversarial examples*.
- Maliciously *perturbed inputs* that cause DNNs to produce incorrect predictions.



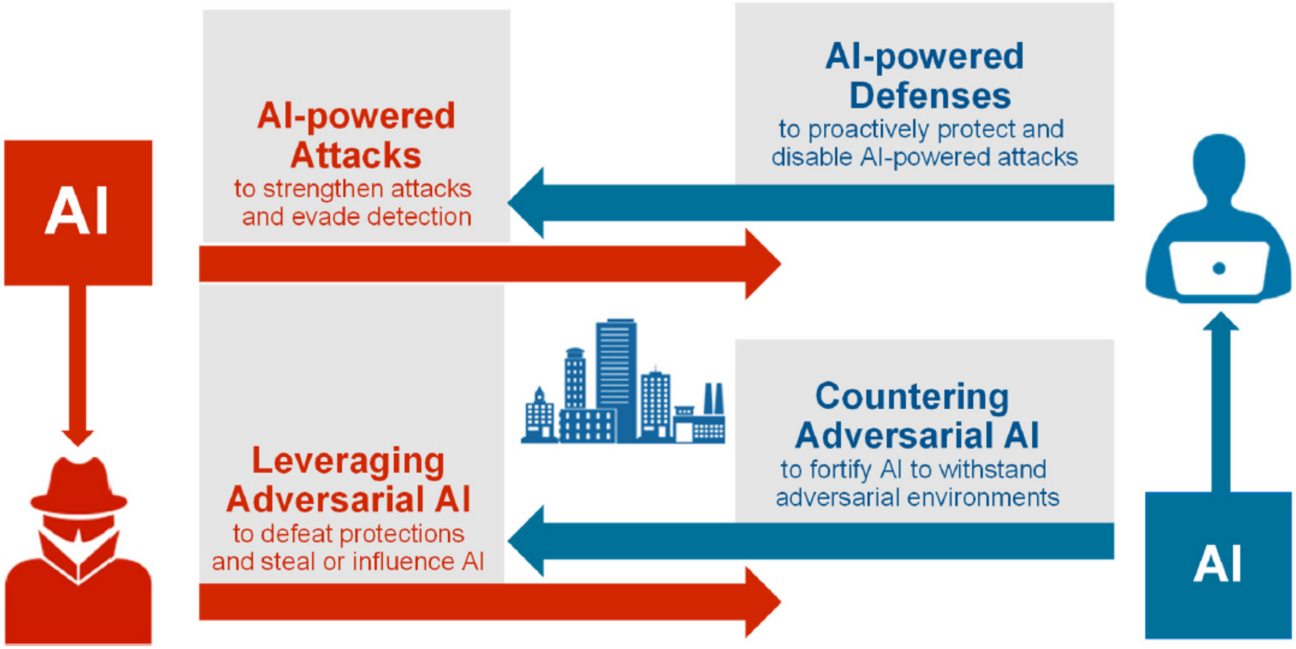
Goodfellow et al. "Explaining and Harnessing Adversarial Examples"



Madry et al.

# Adversarial Attacks

- Adversarial robustness poses a significant challenge for the deployment of ML-based systems.
- Specially safety- and security-critical environments like autonomous driving, disease detection or unmanned aerial vehicles, etc.


AI
AI-powered Attacks
to strengthen attacks and evade detection
AI-powered Defenses
to proactively protect and disable AI-powered attacks
Leveraging Adversarial AI
to defeat protections and steal or influence AI
Countering Adversarial AI
to fortify AI to withstand adversarial environments
AI

Joysua Rao "Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies”

# Adversarial Attacks

- How to fool a machine learning model?
- How to create the adversarial perturbation? **Threat model**
- What is the attack strategy for the perturbation at hand? **Attack Strategy**

# Adversarial Attacks: Threat Model

- What are the desired consequences of the adversarial perturbation?

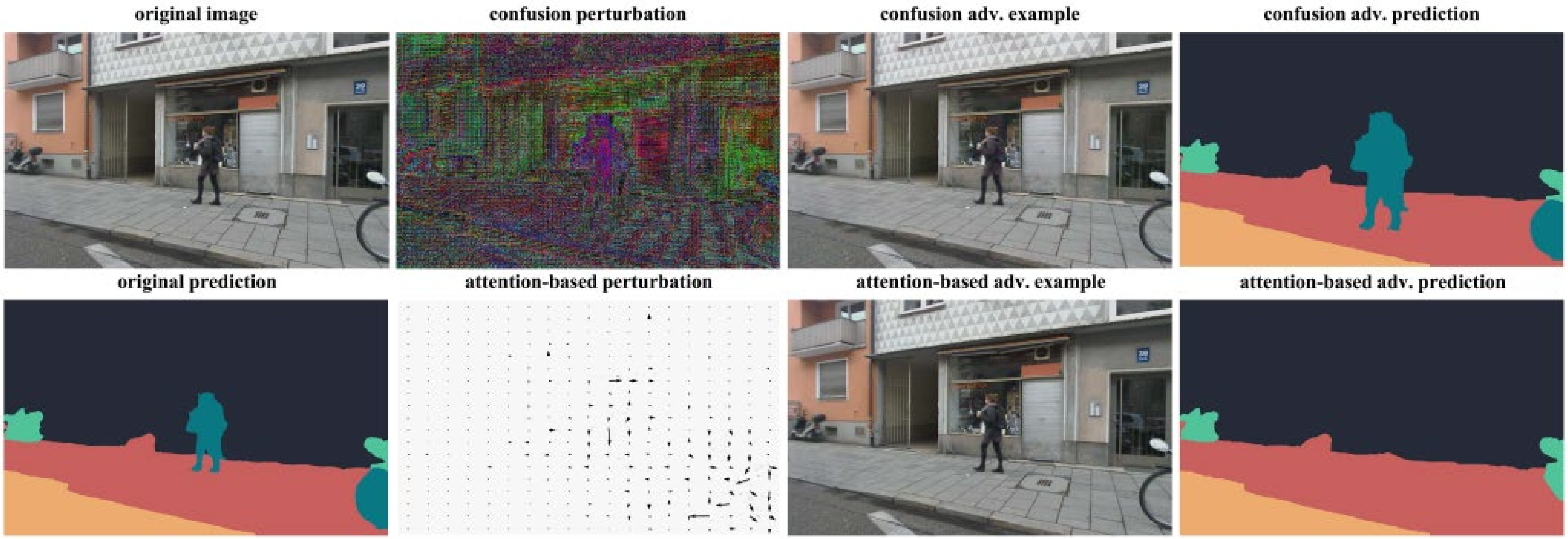
  - Untargeted (Non-targeted): As many misclassifications as possible. No preference concerning the appearing classes in the adversarial output.

  - Static Target: Fixed classification output. Example: Forcing the model to output one *fixed image of an empty street* without any pedestrians or cars in sight.

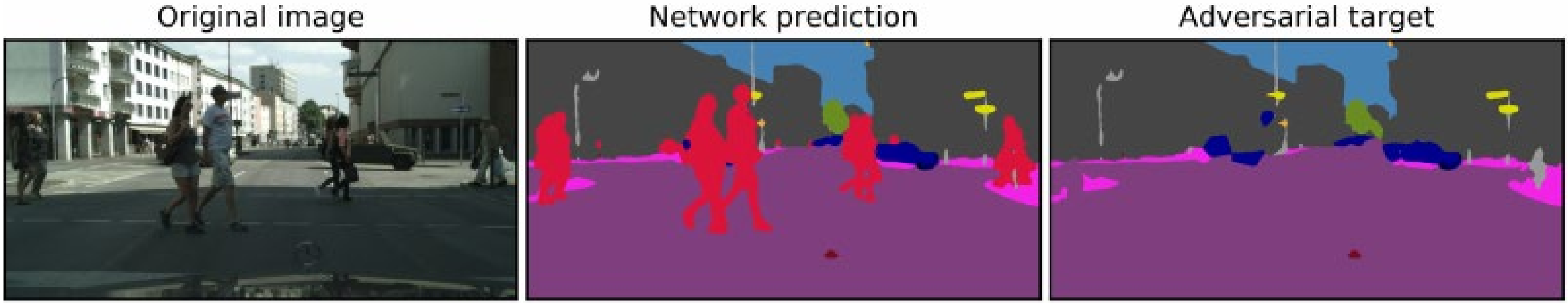
  - Dynamic Target: Keep the output unchanged with the exception of removing *certain target classes*. Example: Removing the pedestrian class in every possible traffic situation.

  - Confusing Target (Confusion): Change the position or size of certain target classes. Example: Reduces the size of pedestrians and in this way leads to a false sense of distance.

# Adversarial Attacks: Threat Model


original image
confusion perturbation
confusion adv. example
confusion adv. prediction
original prediction
attention-based perturbation
attention-based adv. example
attention-based adv. prediction

Assion et al. "The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks"



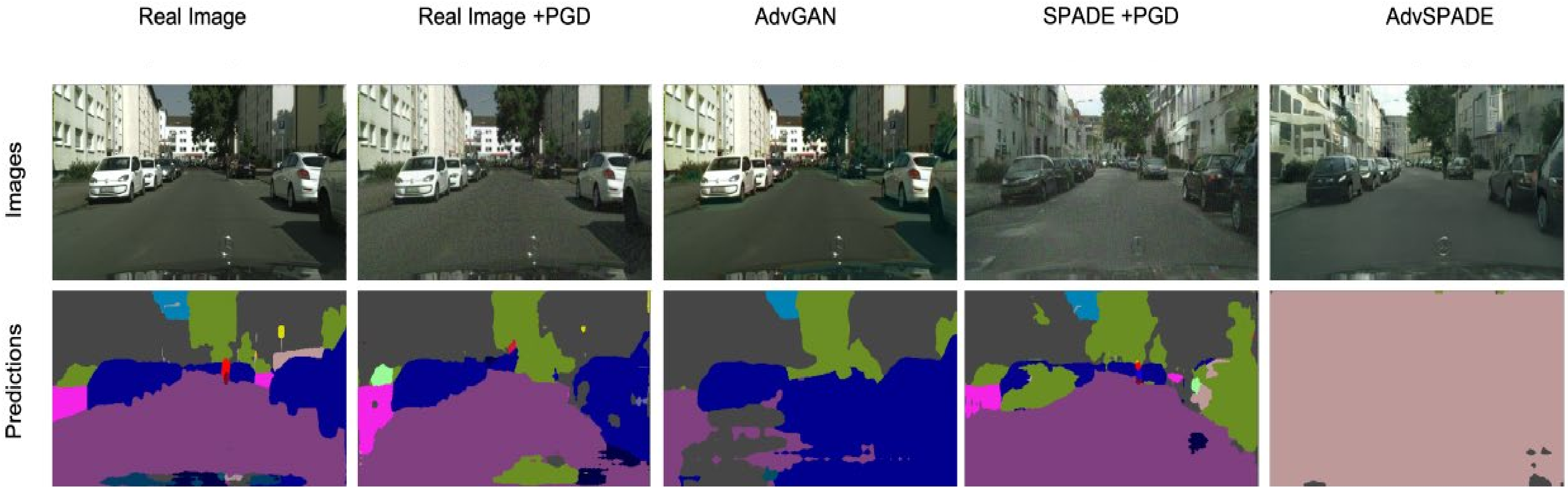
Yuan et al. "Adversarial Examples: Attacks and Defenses for Deep Learning"

# Adversarial Attacks: Threat Model

- Perturbation Scope:
  - Individual Scope: Attack is designed for one specific *input image*. It is not necessary that the same perturbation fools the ML system on other data points.
  - Contextual Scope: Image agnostic perturbation that causes label changes for one or more specific contextual situations. Example, traffic, rain, lighting change, camera angles, etc.
  - Universal Scope: Image agnostic perturbation that causes label changes for a significant part of the true data distribution with no explicit contextual dependencies.

# Adversarial Attacks: Threat Model

- Perturbation Scope:



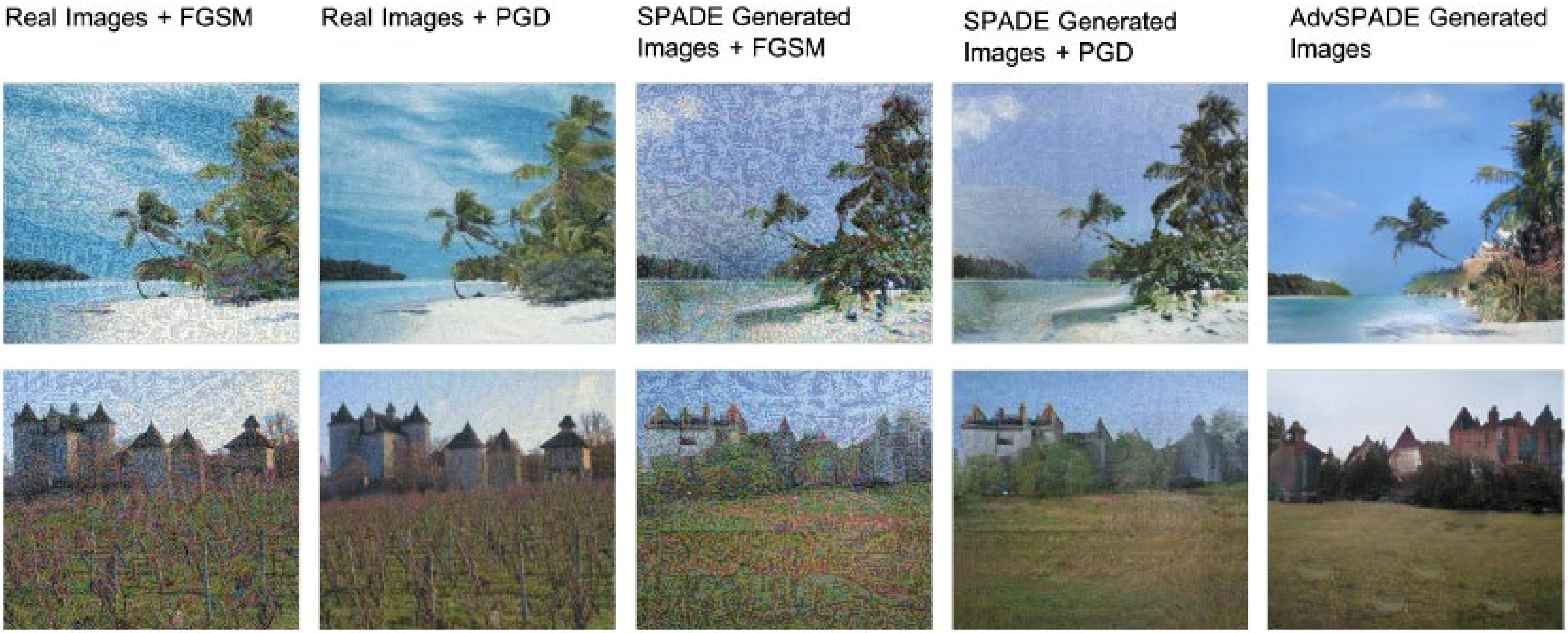
Shen et al. "AdvSPADE: Realistic Unrestricted Attacks for Semantic Segmentation"

# Adversarial Attacks: Threat Model

- Perturbation Imperceptibility:
  - Lp-based Imperceptibility: Small changes with respect to some Lp-norm, the changes should be imperceptible to human eyes.
  - Attention-based Imperceptibility: Wasserstein distance, SSIM or other metric based imperceptibility.
  - Output Imperceptibility: The classification output is imperceptible to the human observer.
  - Detector Imperceptibility: A predefined selection of software-based detection systems is not able to detect irregularities in the input, output or in the activation patterns of the ML module caused by the adversarial perturbation.

# Adversarial Attacks: Threat Model

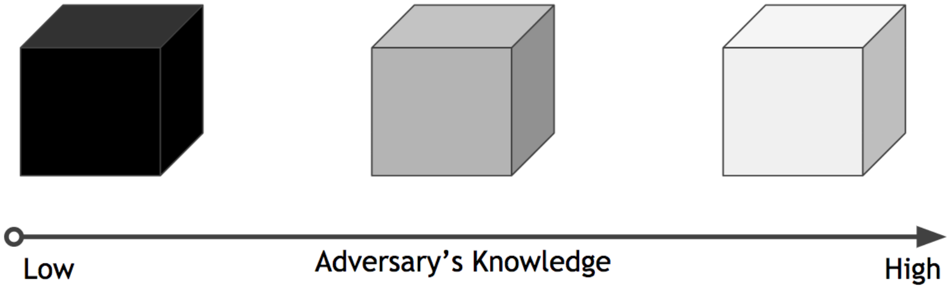
- Perturbation Imperceptibility:


Real Images + FGSM
Real Images + PGD
SPADE Generated Images + FGSM
SPADE Generated Images + PGD
AdvSPADE Generated Images

Shen et al. "AdvSPADE: Realistic Unrestricted Attacks for Semantic Segmentation"

# Adversarial Attacks: Threat Model

- Model Knowledge:
  - White-box: Full knowledge of the model internals: architecture, parameters, weight configurations, training strategy.
  - Output-transparent Black-box: No access to model parameters. But can observe the class probabilities or output logits of the module.
  - Query-limited Black-box: Access to the full or parts of the module's output on a limited number of inputs or with a limited frequency.
  - Label-only Black-box: Only access to the full or parts of the final classification/regression decisions of the system.
  - (Full) Black-box: No access to the model of any kind.



# Adversarial Attacks: Threat Model

- Data Knowledge:
  - Training Data: Access to full of significant part of training data
  - Surrogate Data: No direct access. But data points can be collected from the relevant underlying data distribution.
- Adversary Capability:
  - Digital Data Feed (Direct Data Feed): The attacker can directly feed digital input to the model.
  - Physical Data Feed: Creates physical perturbations in the environment.
  - Spatial Constraint: Only influence limited areas of the input data.

# Adversarial Attacks: Attack Strategy

- Model Basis: Which model is used by the attack?
  - Victim Model: Use the victim model to calculate adversarial perturbations.
  - Surrogate Model: Use a surrogate model or a different model.
- Data Basis: What data is used by the attack?
  - Training Data: Original training data set are given to the adversarial attack.
  - Surrogate Data: Data related to the underlying data distribution of the task.
  - No Data: Attack works with images that are not samples of the present data distribution.

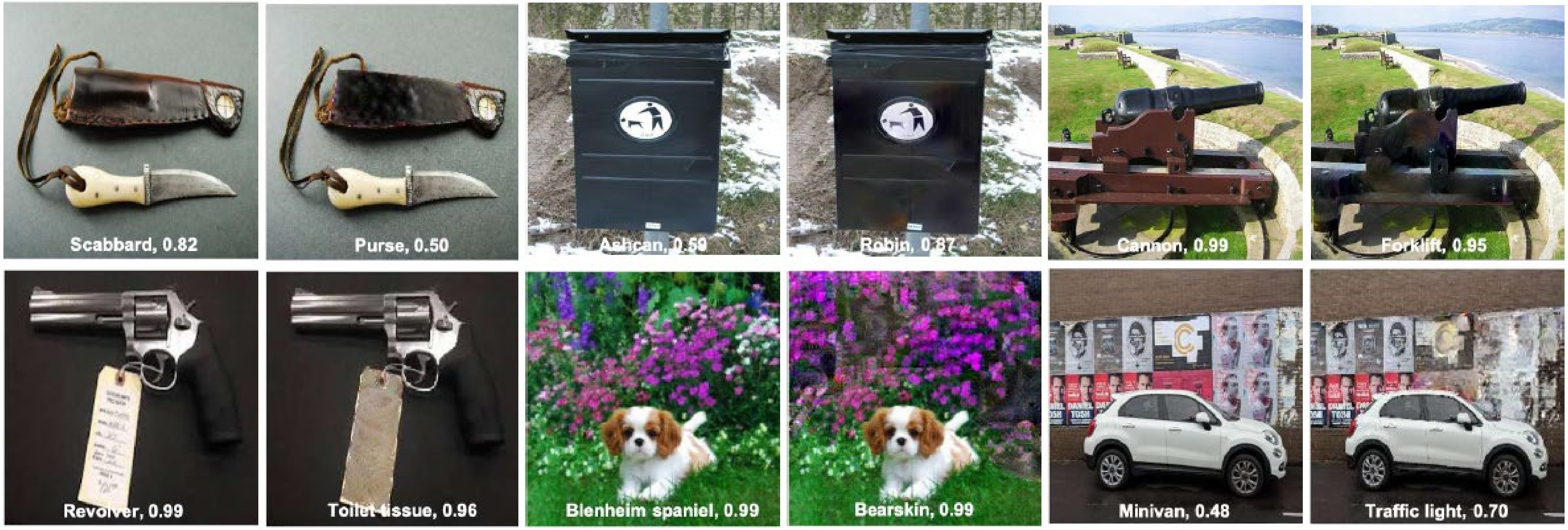
# Adversarial Attacks: Attack Strategy

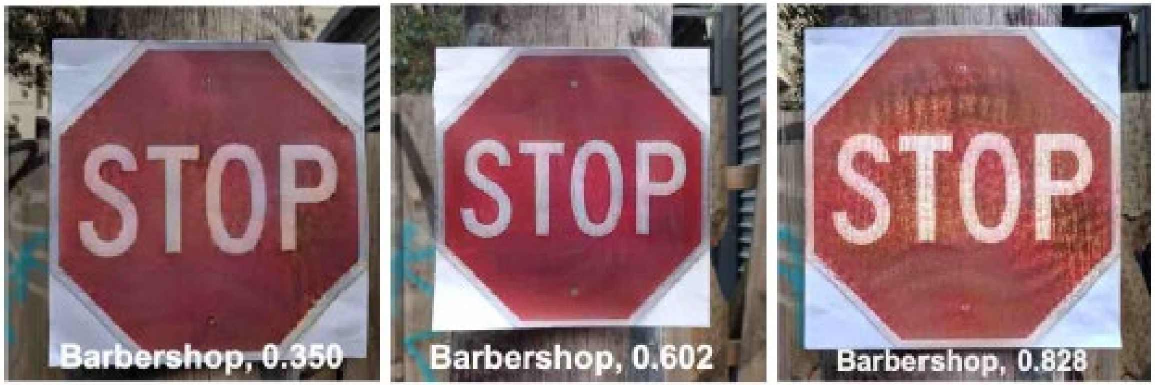
- Optimization Method:
  - First-order Methods: Exploit perturbation directions given by exact or approximate (sub-)gradients.
  - Second-order Methods: Based on the calculation of the Hessian matrix or approximations of the Hessian matrix.
  - Evolution & Random Sampling: The adversarial attack generates possible perturbations by sampling distributions and combining promising candidates.

# Adversarial Attacks: Attack Strategy

- Some of the representative approaches for generating adversarial examples
  - Fast Gradient Sign Method (FGSM)
  - Basic Iterative Method (BIM)
  - Iterative Least-Likely Class Method (ILLC)
  - Jacobian-based Saliency Map Attack (JSMA)
  - DeepFool
  - CPPN EA Fool
  - Projected Gradient Descent (PGD)
  - Carlini and Wagner (C&W) attack
  - Adversarial patch attack

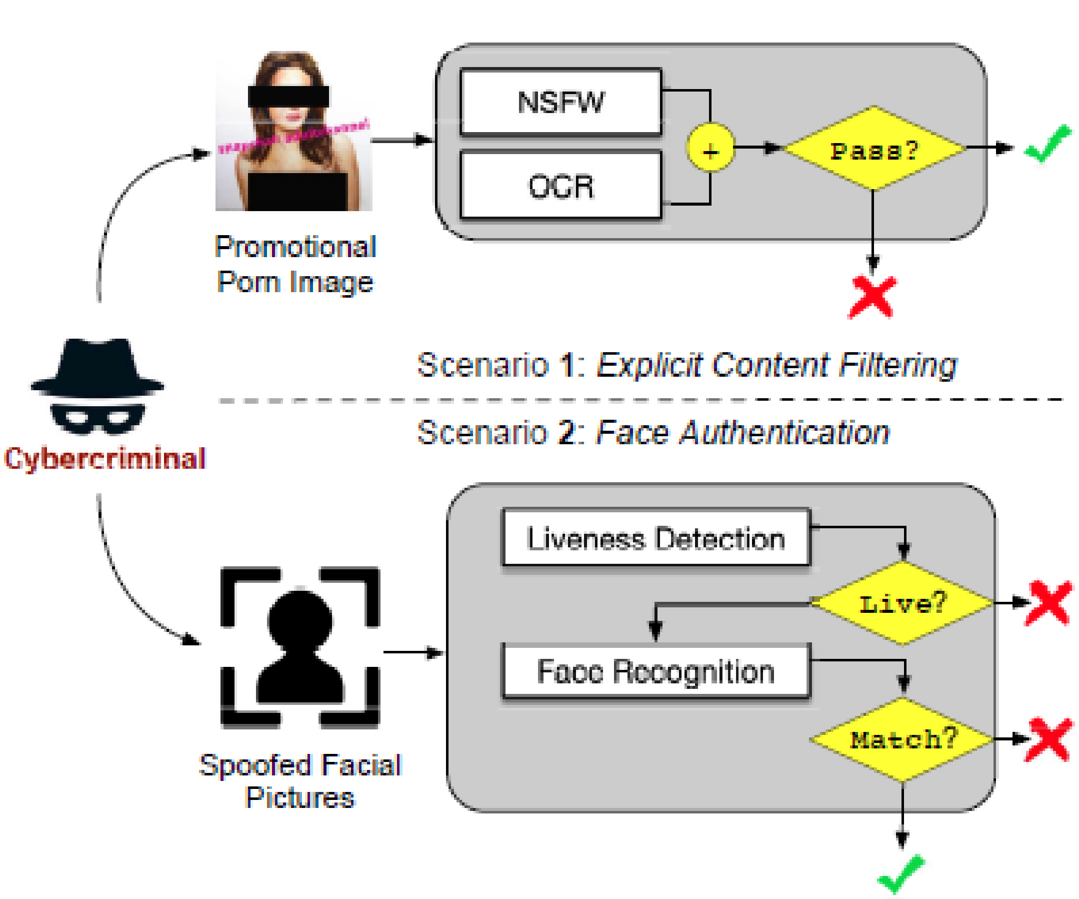
# Attacks on Image Classification



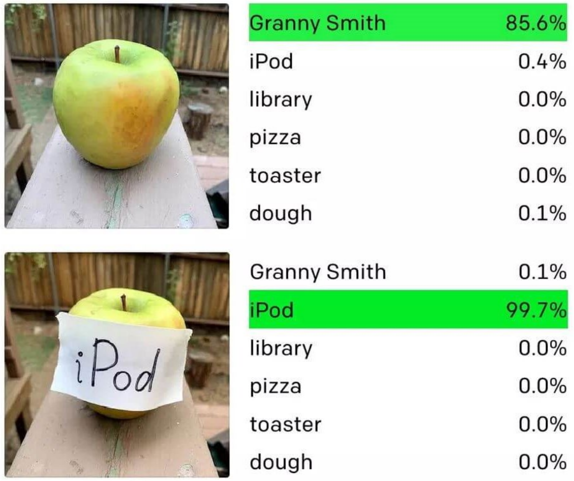


Duan et al. “Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles”

# Attacks on Image Classification



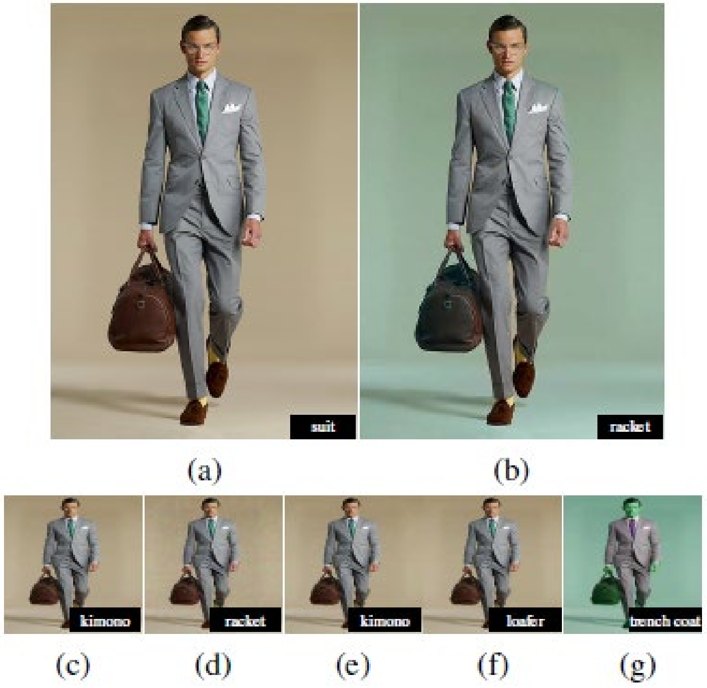
Lu et al. "Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction



| | |
|---|---|
| Granny Smith | 85.6% |
| iPod | 0.4% |
| library | 0.0% |
| pizza | 0.0% |
| toaster | 0.0% |
| dough | 0.1% |

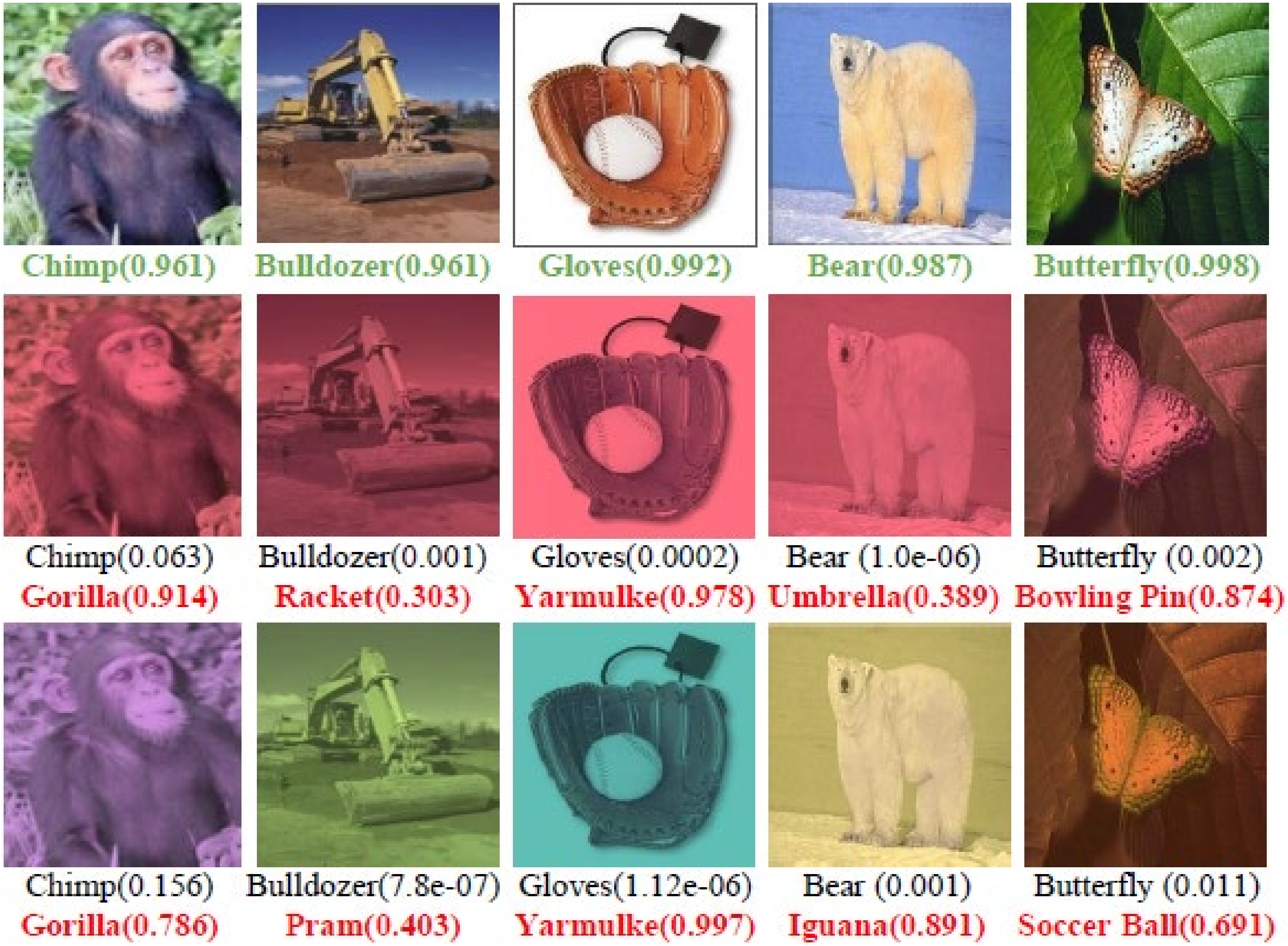
| | |
|---|---|
| Granny Smith | 0.1% |
| iPod | 99.7% |
| library | 0.0% |
| pizza | 0.0% |
| toaster | 0.0% |
| dough | 0.0% |

https://openai.com/blog/multimodal-neurons/



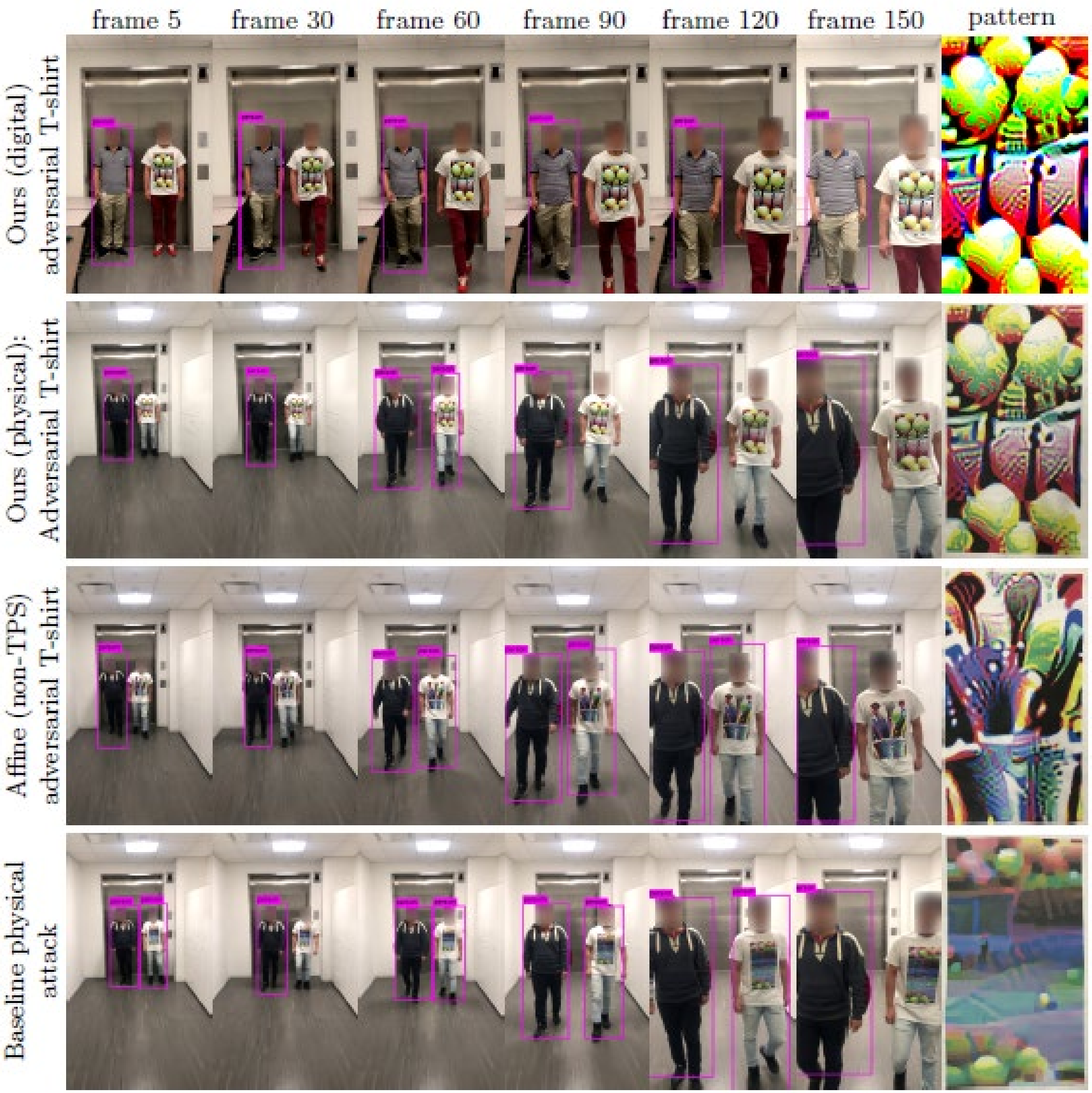
Shamsabadi, et al. “ColorFool Semantic Adversarial Colorization”

# Attacks on Image Classification



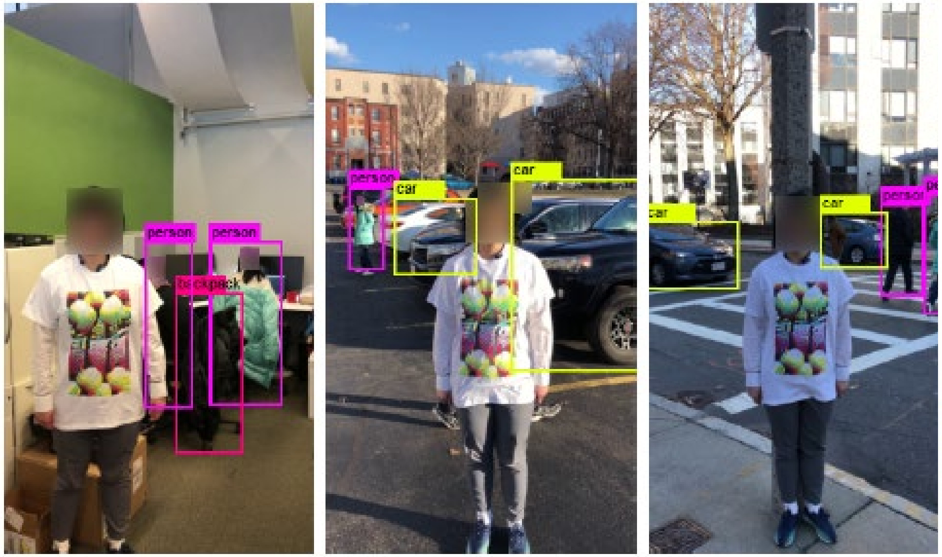
Kantipudi et al. “Color Channel Perturbation Attacks for Fooling Convolutional Neural Networks and A Defense Against Such Attacks”

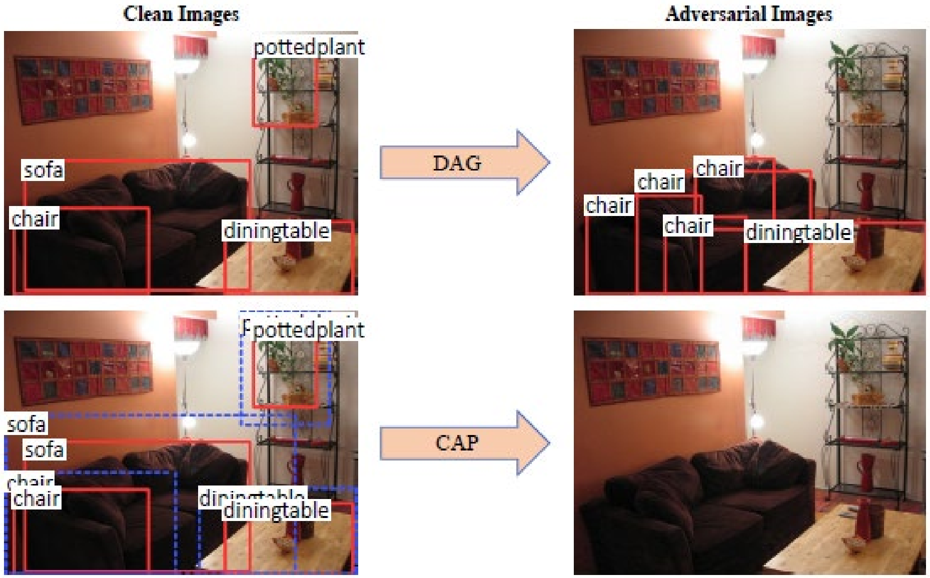
# Attacks on Object Detector

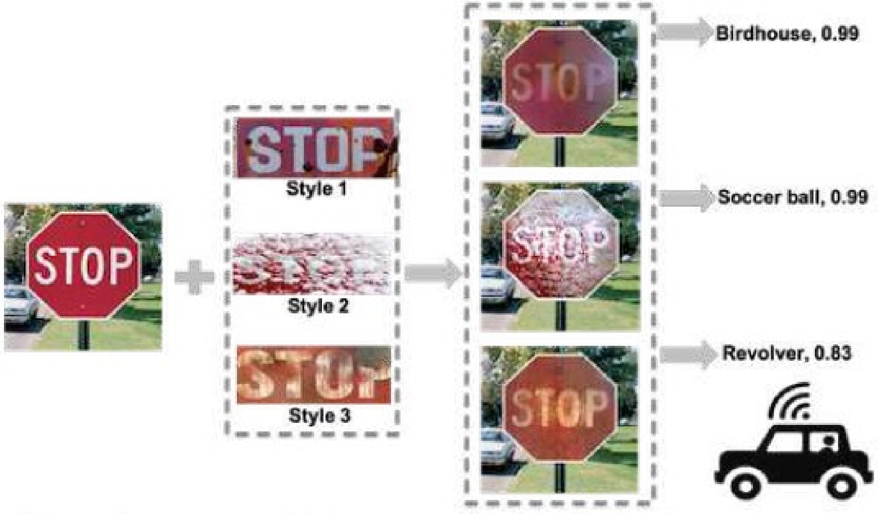

frame 5
frame 30
frame 60
frame 90
frame 120
frame 150
pattern
Ours (digital) adversarial T-shirt
Ours (physical): Adversarial T-shirt
Affine (non-TPS) adversarial T-shirt
Baseline physical attack

Xu et al. "Adversarial T-shirt! Evading Person Detectors in A Physical World"

# Attacks on Object Detector







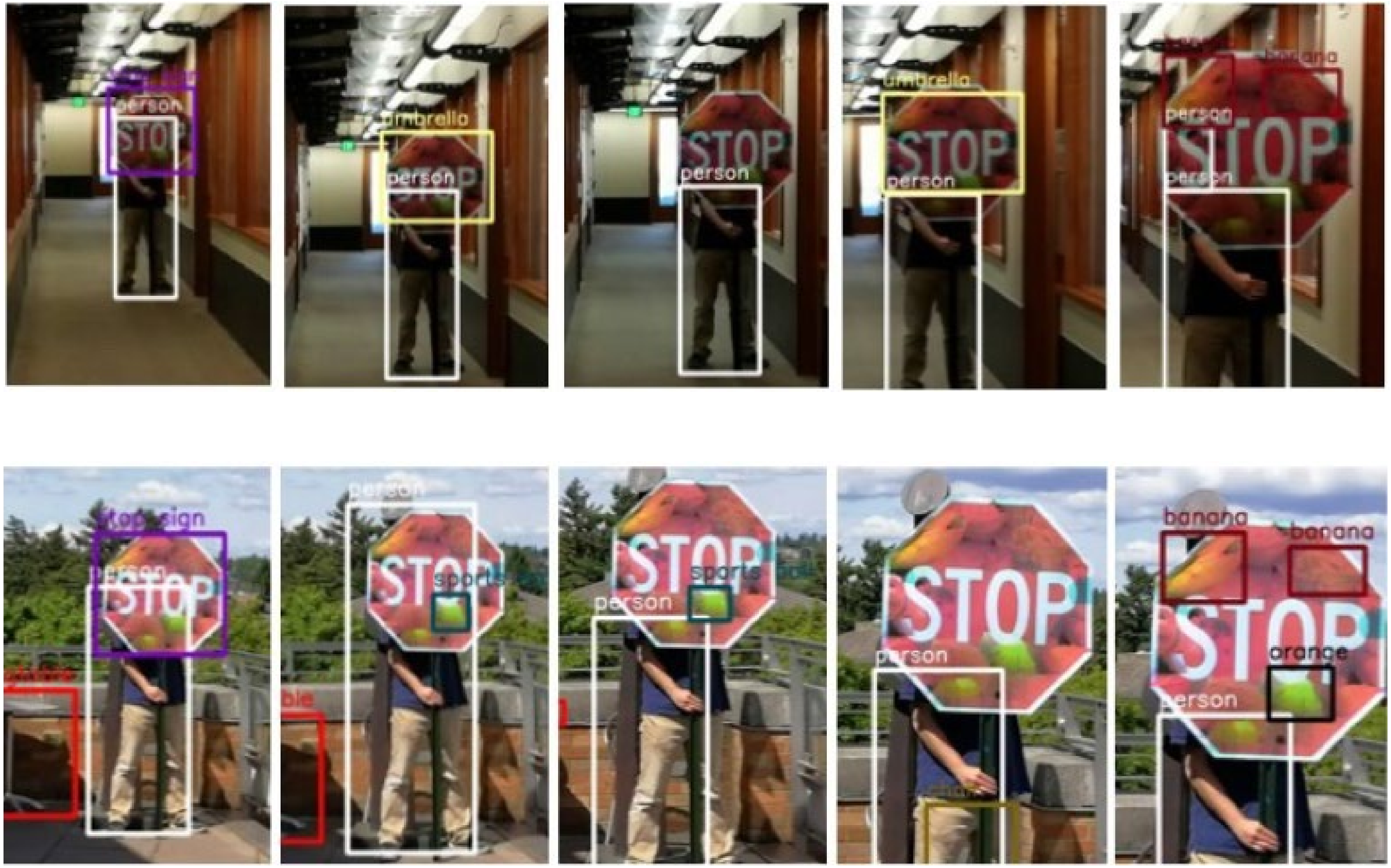
(a) Target image (b) Style (c) Adversarial examples

Duan et al. "Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles"

Zhang et al. "Contextual Adversarial Attacks for Object Detection"

Xu et al. "Adversarial T-shirt! Evading Person Detectors in A Physical World"

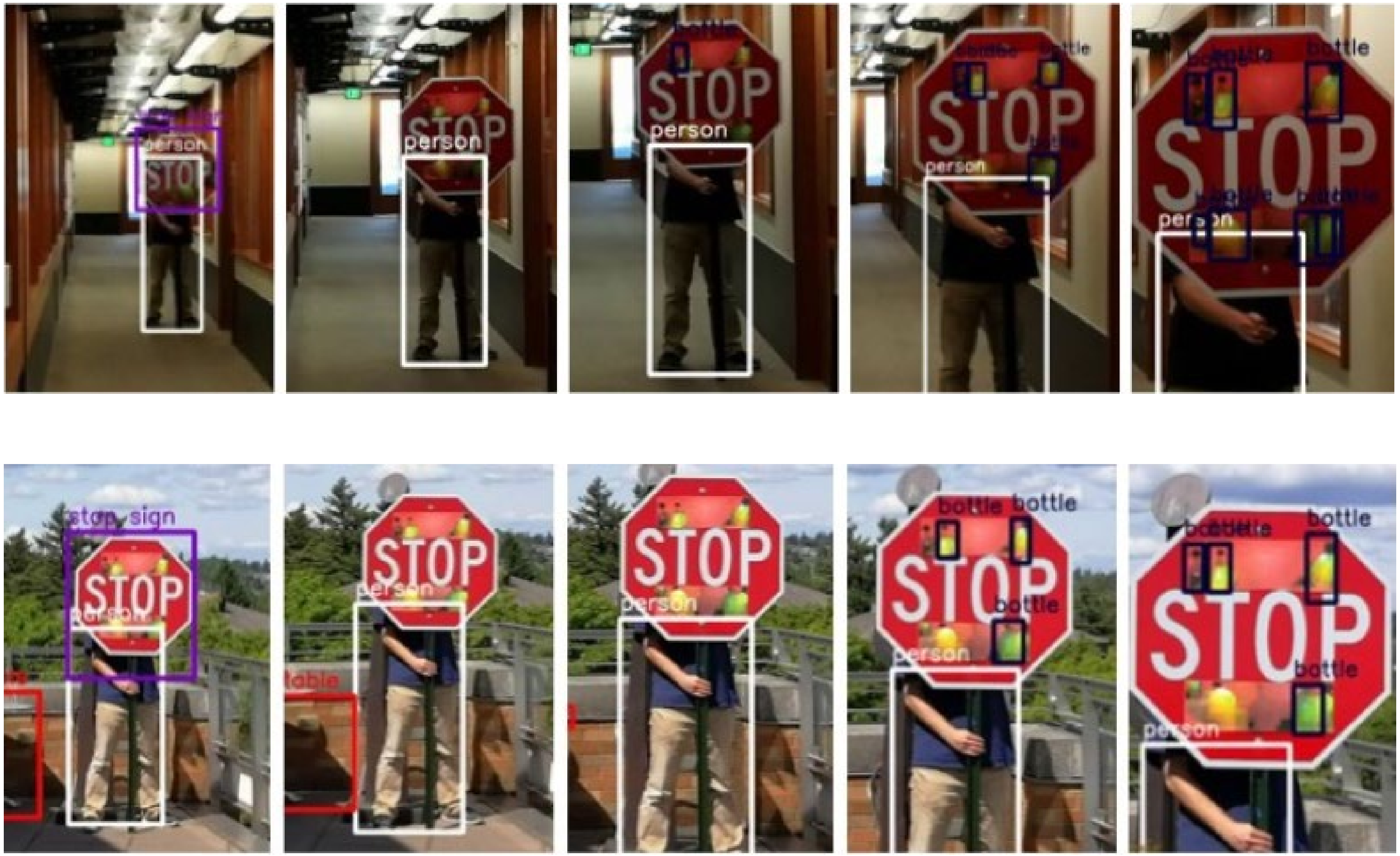
# Attacks on Object Detector



The poster attack on Yolov2

Eykholt et al. “Physical Adversarial Examples for Object Detectors”

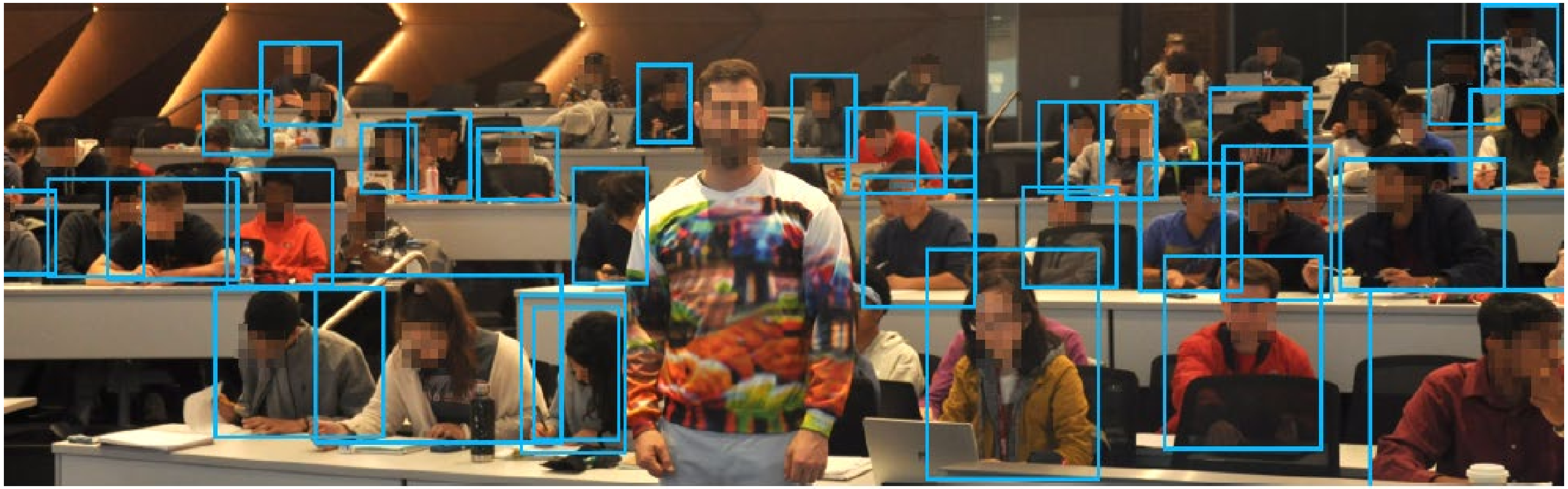
# Attacks on Object Detector



The sticker attack on Yolov2

Eykholt et al. “Physical Adversarial Examples for Object Detectors”

# Attacks on Object Detector

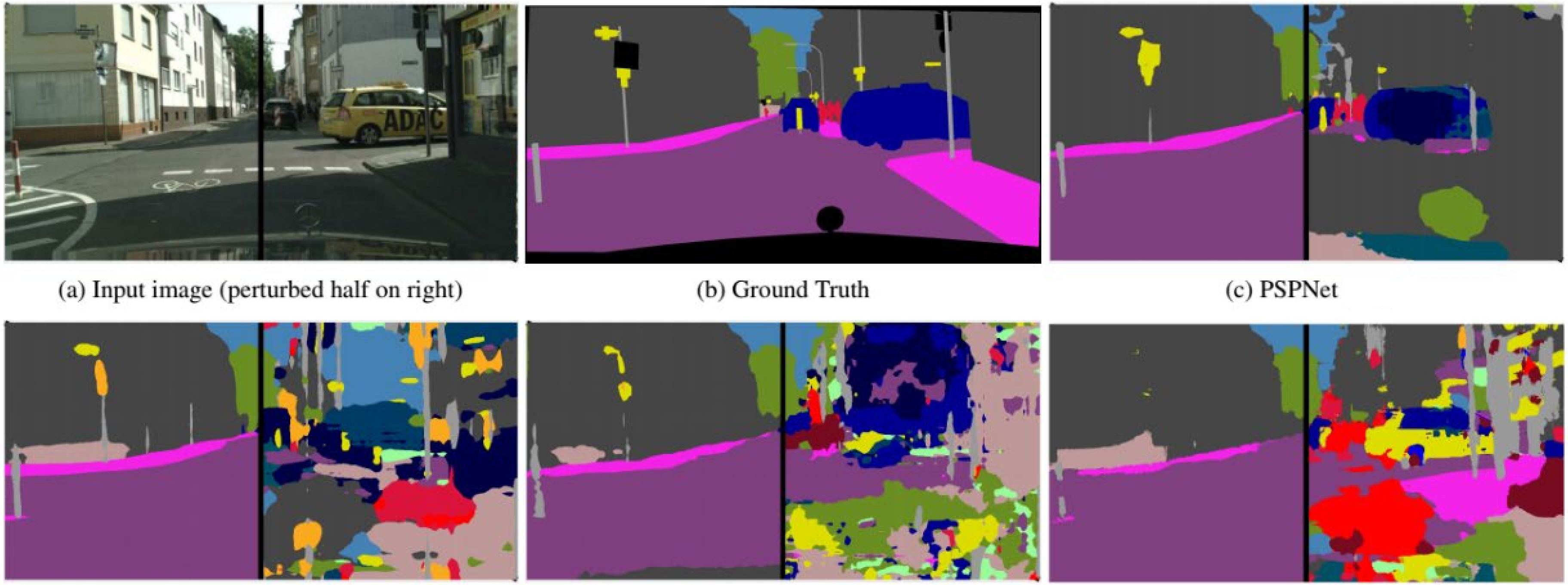


The YOLOv2 detector is evaded using a pattern trained on the COCO dataset with a carefully constructed objective.

Wu et al. “Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors”

# Attacks on Semantic Segmentation

- Semantic segmentation networks are harder to break.
- Due their multi-scale encoder decoder structure and output as per pixel probability instead of just probability score for the whole image.



(a) Input image (perturbed half on right)

(b) Ground Truth

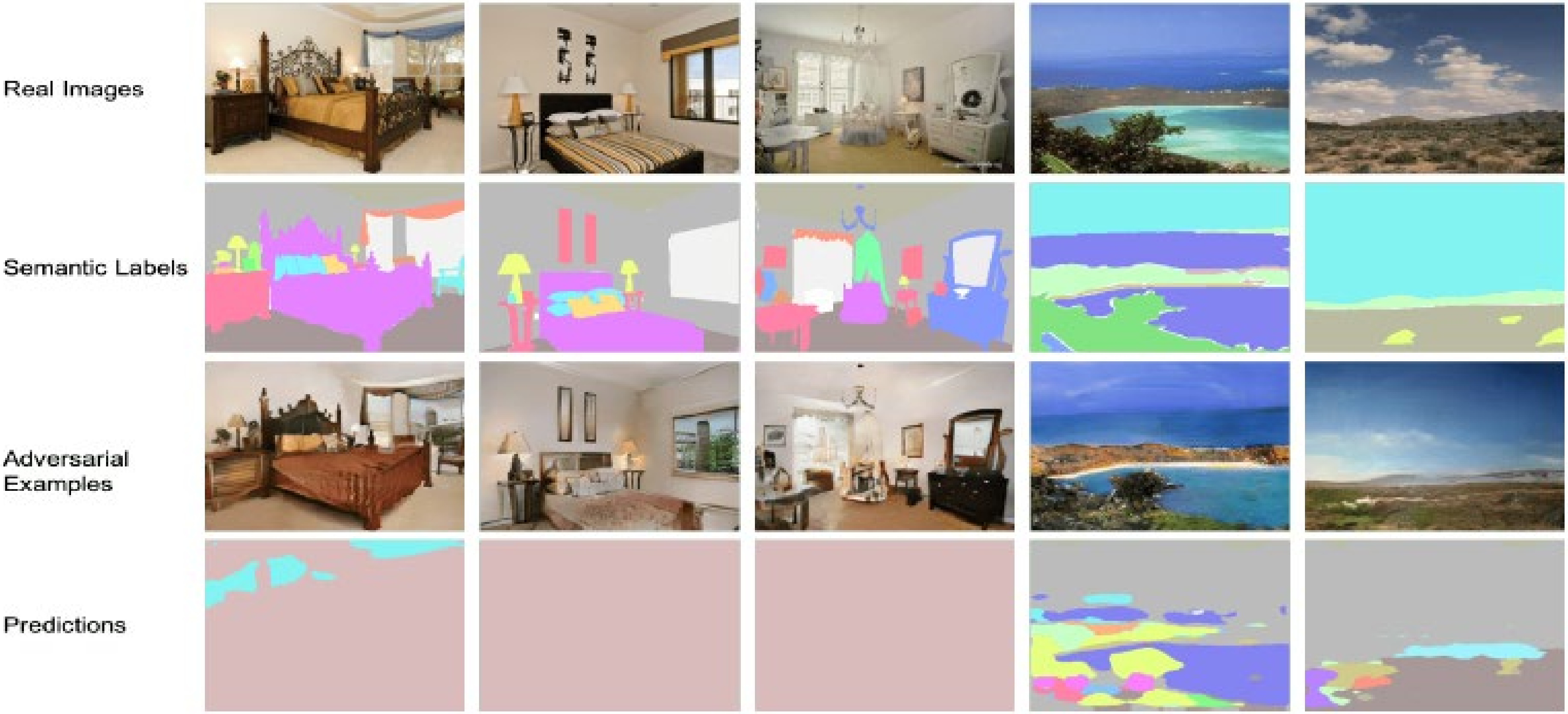
(c) PSPNet

(d) DilatedNet

(e) ICNet

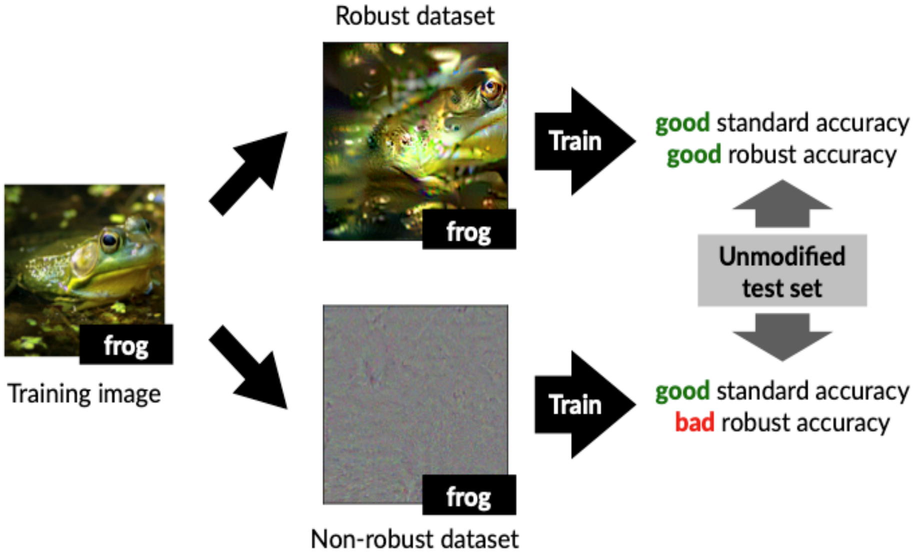
(f) CRF-RNN

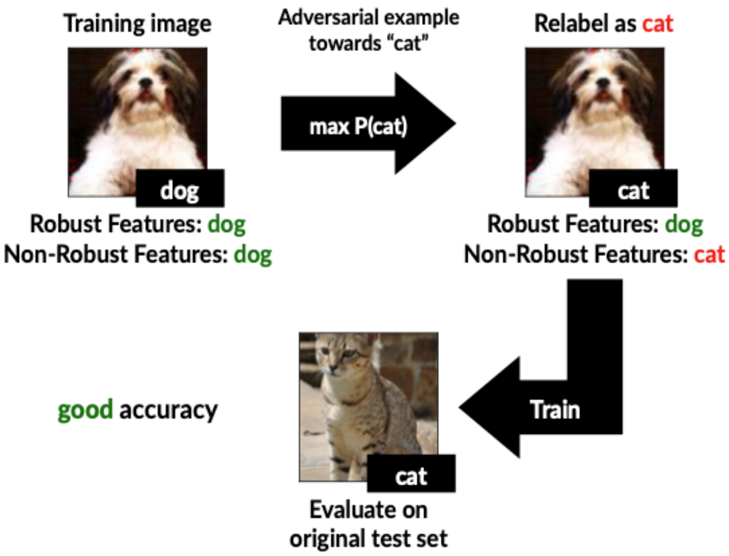
# Attacks on Semantic Segmentation



Shen et al. "AdvSPADE: Realistic Unrestricted Attacks for Semantic Segmentation"

# Why do adversarial examples exist?





Adversarial examples can be attributed to the presence of non-robust features

Ilyas et al. "Adversarial examples are not bugs, they are features"

# Adversarial Robustness

- We can use the knowledge about the adversarial attacks to improve the model robustness.

- Why to evaluate the robustness?

- To defend against an adversary who will attack the system.
  - For example, an attacker may wish to cause a self-driving car to incorrectly recognize road signs.
  - Cause an NSFW detector to incorrectly recognize an image as safe-for-work.
  - Cause a malware (or spam) classifier to identify a malicious file (or spam email) as benign.
  - Cause an ad-blocker to incorrectly identify an advertisement as natural content
  - Cause a digital assistant to incorrectly recognize commands it is given.

# Adversarial Robustness

- To test the worst-case robustness of machine learning algorithms.
    - Many real-world environments have inherent randomness that is difficult to predict.
    - Analyzing the worst-case robustness will cover minor perturbation cases.

- To measure progress of machine learning algorithms towards human-level abilities.
    - In terms of normal performance, Gap is <<<< between Human Vs Machine.
    - In adversarial robustness, Gap >>>> between Human Vs Machine.

# Defense Against Adversarial Attacks

- Reactive defenses: Preprocessing techniques, detection for adversarial samples.
  - Detection of adversarial examples
  - Input transformations (preprocessing)
- Obfuscation defenses: Try to hide or obfuscate sensitive traits of a model (e.g. gradients) to alleviate the impact of adversarial examples.
  - Gradient masking

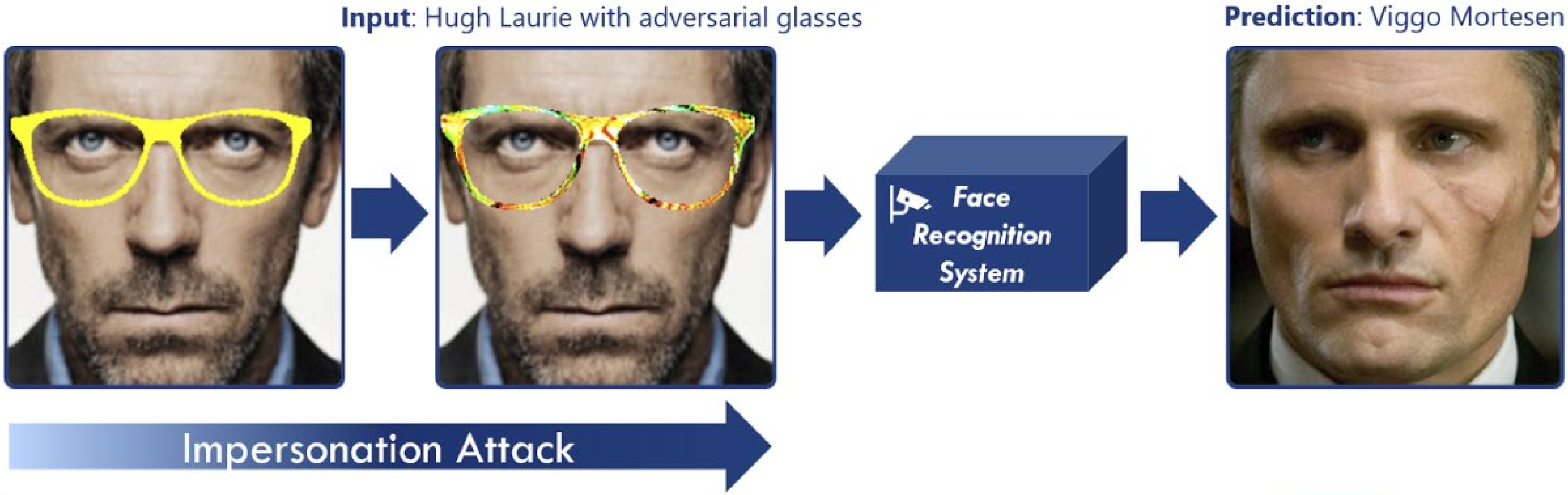
# Defense Against Adversarial Attacks

- Proactive defenses: Build and train models natively robust to adversarial perturbations.
    - Adversarial training
    - Architectural defenses
    - Learning in a min-max setting
    - Hyperparameter tuning
    - Generative models (GAN) based defense
    - Provable adversarial defenses

- What is missing?
- A uniform protocol for defense evaluation

# Adversarial Attacks & Privacy?

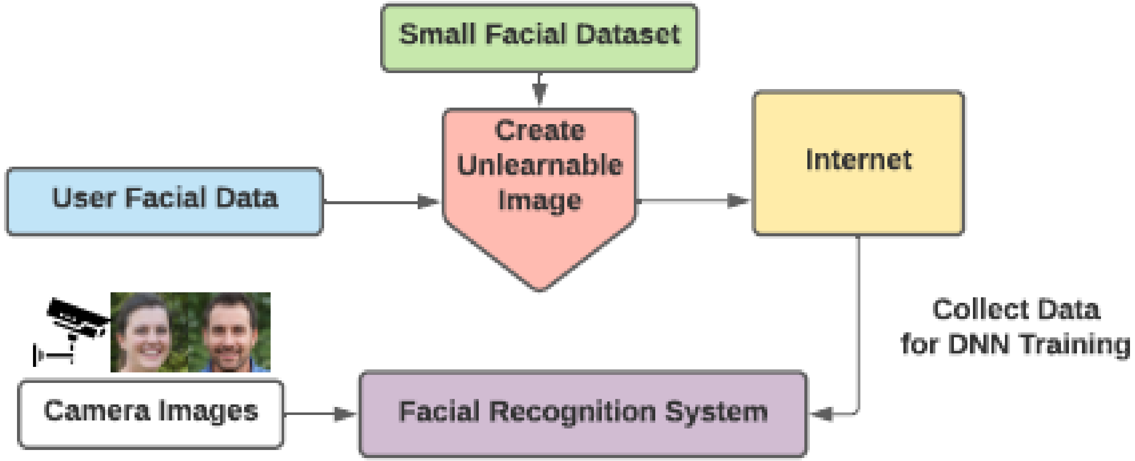
- Protect your Identity in public places.


Input: Hugh Laurie with adversarial glasses
Prediction: Viggo Mortesen
Face Recognition System
Impersonation Attack

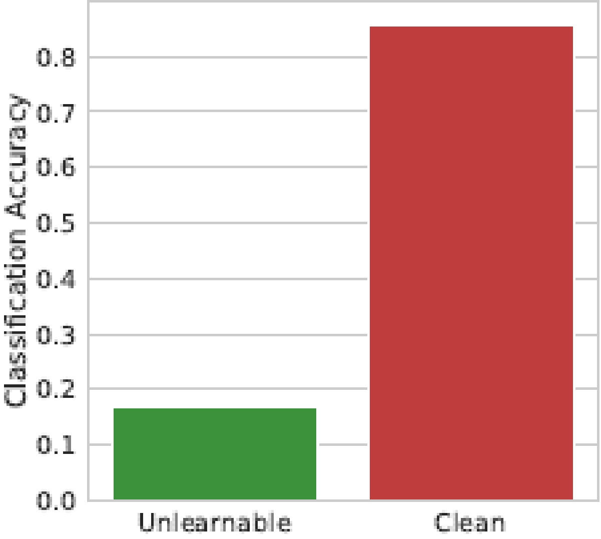
https://www.inovex.de/blog/machine-perception-face-recognition/

# Adversarial Attacks & Privacy?

- Stopping unauthorized exploitation of personal data for training commercial models.
- Protect your privacy.
- *Can data be made unlearnable for deep learning models?*



(a) Illustration of the pipeline



(b) Recognition accuracy

Huang et al. "Unlearnable Examples: Making Personal Data Unexploitable"

# Takeaways

- Adversarial Attacks and defense– A very important challenge for AI research.
- The existence of adversarial cases depend on the applications – classification, detection, segmentation, etc.
- How many adversarial samples are out there? Impossible to know.
- Need to revisit the current practice of reporting standard performance. Adversarial robust performance matters!
- Robustness of ML/DL models must be evaluated with adversarial examples.
- Adversarial attacks for a good cause – improving privacy.

# References

- Grebner et al. “The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks”
- Arnab et al. "On the Robustness of Semantic Segmentation Models to Adversarial Attacks“
- Liu et al. “Deep Learning for Generic Object Detection: A Survey”
- Wu et al. “Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors”
- Assion et al. "The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks"
- Shen et al. "AdvSPADE: Realistic Unrestricted Attacks for Semantic Segmentation"
- Xu et al. "Adversarial T-shirt! Evading Person Detectors in A Physical World"
- Duan et al. "Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles“
- Serban et al. "Adversarial Examples - A Complete Characterisation of the Phenomenon“

# Thank You!