The Bright and Dark Sides of Computer Vision and Machine Learning
Challenges and Opportunities for Robustness and Security

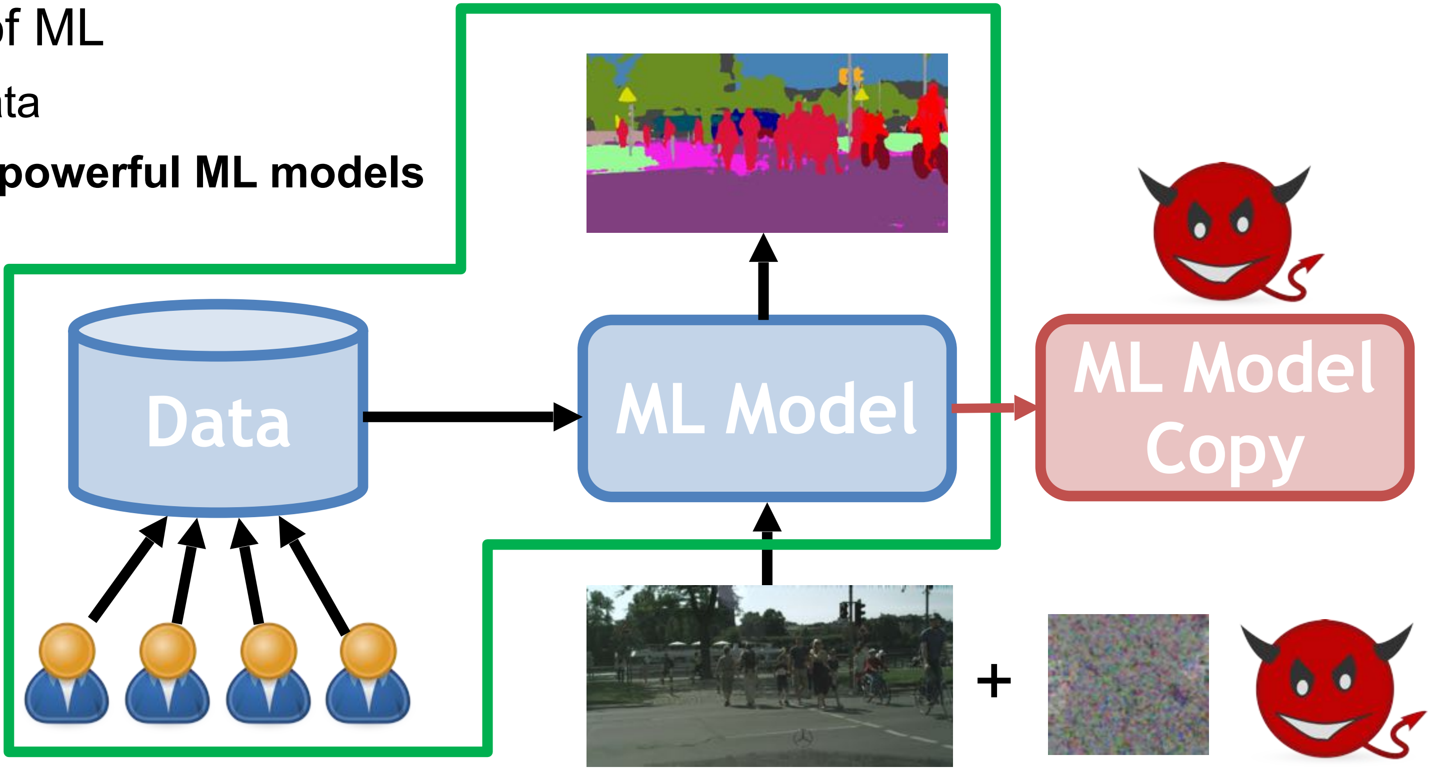
Bernt Schiele

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Robustness & Security in Machine Learning: Towards Trustworthy AI

- **Widespread** deployment of ML
  - future industry is fueled by data
  - **“standard” pipeline** to train **powerful ML models**

- **Security** of **ML-models** is **multi-facetted**:
  - robustness to input variation
  - preventing model “stealing”
  - …



- Membership Inference
- Linkability Attack

Adversarial Perturbations

Overview

- **Robustness** and **Security** of Deep Models
  - **Bright** and **Dark** Side of **Scene Context** — NeurIPS'18, CVPR'19
  - Disentangling **Adversarial Robustness** and **Generalization** — CVPR'19
  - **Reverse Engineering** and **Stealing** Deep Models — ICLR'18, CVPR'19, ICLR'20

max planck institut informatik

Saarland Informatics Campus

Adversarial Scene Editing: Automatic Object Removal from Weak Supervision

@ NeurIPS 2018

Not Using the Car to See the Sidewalk: Quantifying and Controlling the Effects of Context in Classification and Segmentation

@ CVPR 2019

Rakshith Shetty
MPI Informatics

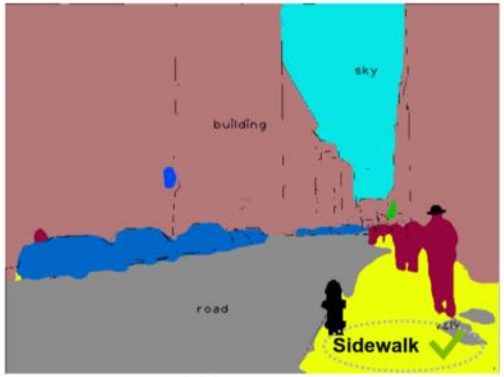
Mario Fritz
CISPA Helmholtz

Bernt Schiele
MPI Informatics

Motivation: The Bright and the Dark Side of Scene Context

- Current models heavily rely on scene context:
  - Original image with cars on the left side:

original ($\mathcal{I}$)



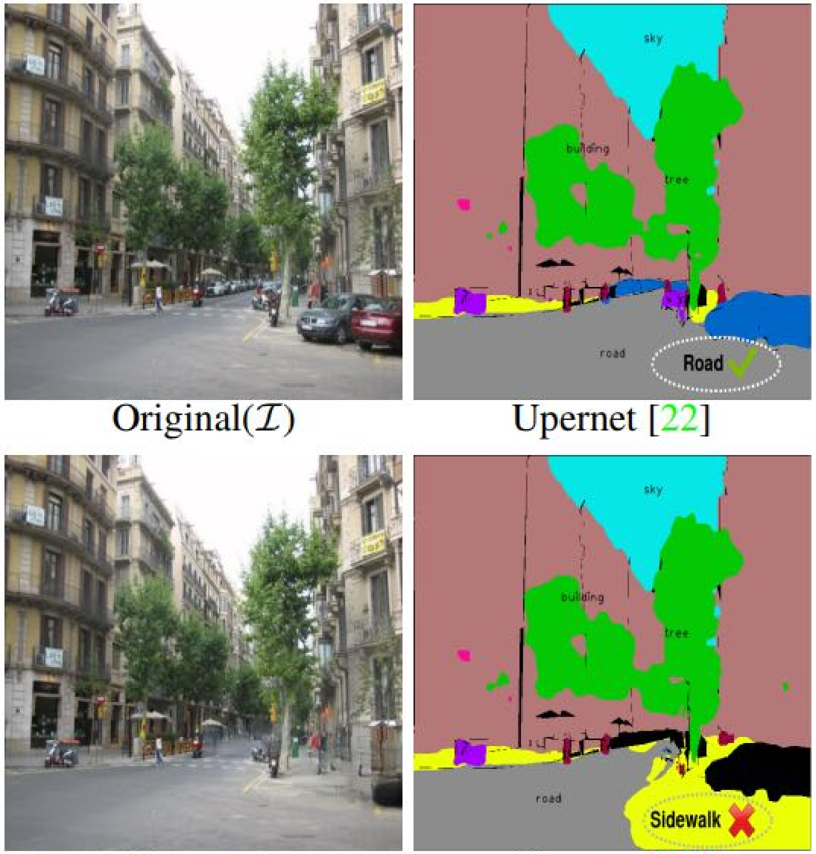
Upernet

  - Same image without those cars:

Question: How Dependent are Current Models on Scene Context?

- Here
  - we look at a **particular aspect** of **context** : co-occurring objects

- Goals:
  - **quantify context sensitivity** of classification and segmentation using **object removal** [NeurIPS'18]
  - object removal based **data augmentation** for **better performance**

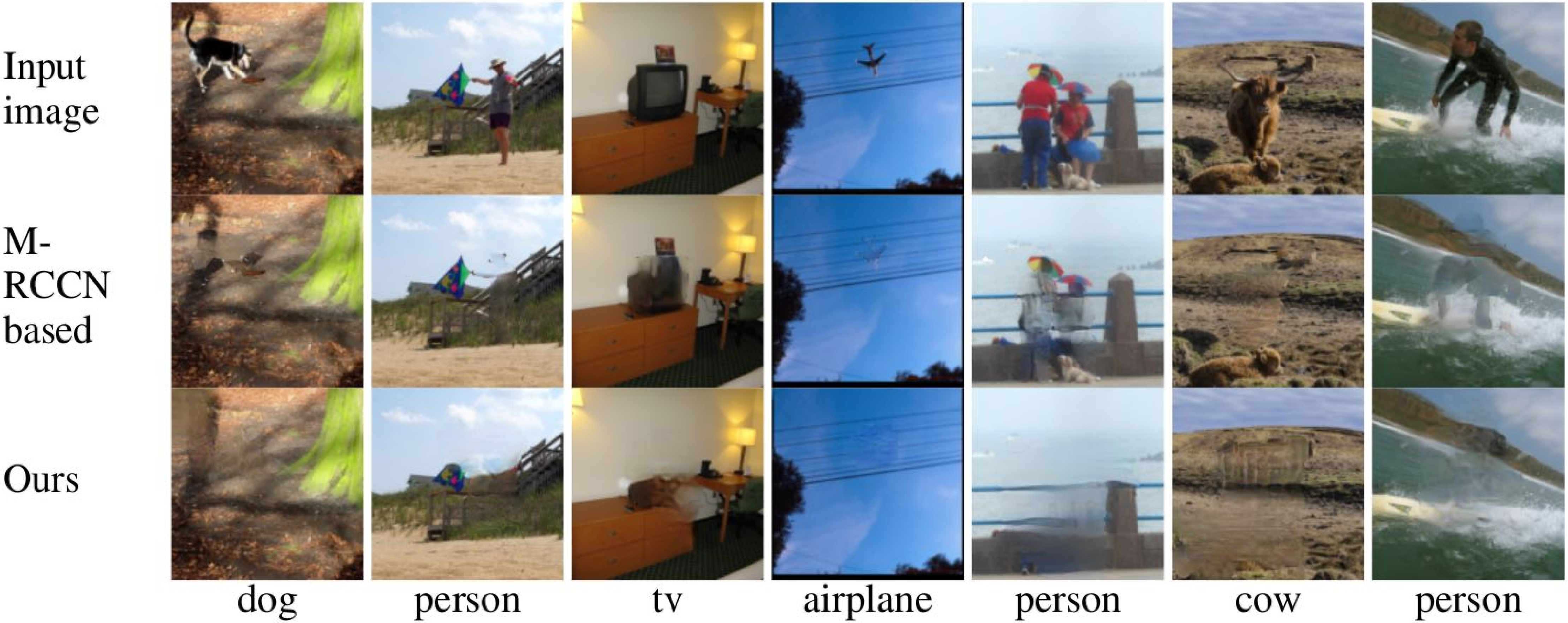


Original($\mathcal{I}$) Upernet [22]

$\mathcal{I} - car$ Upernet [22]

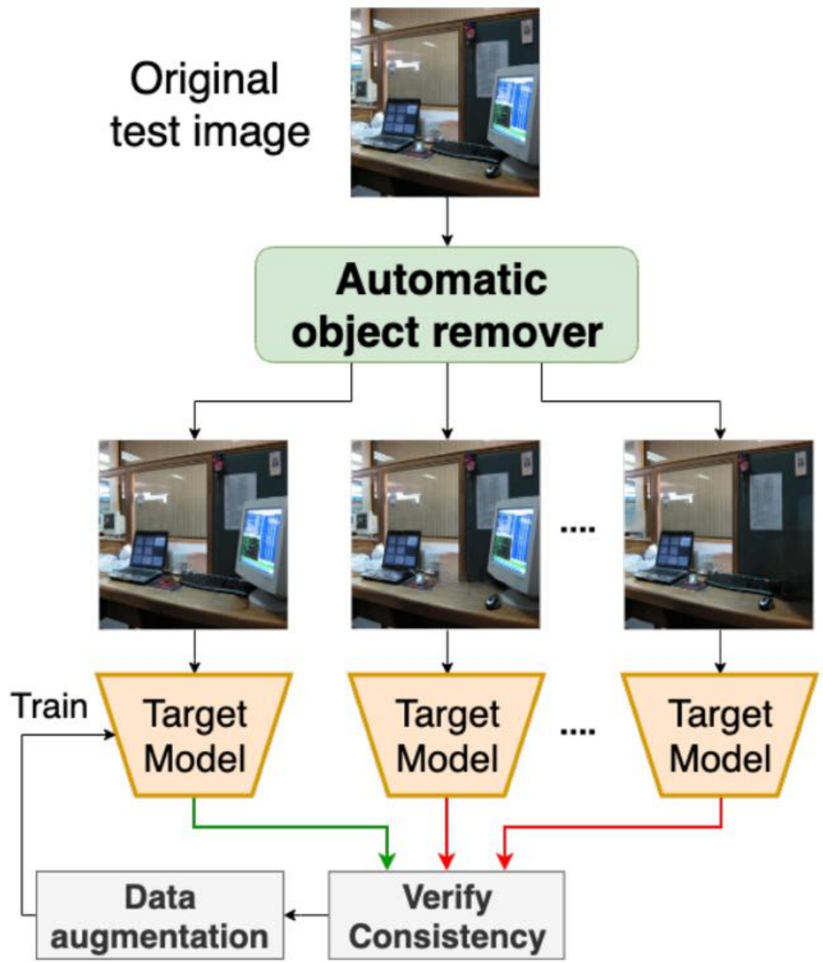
[Shetty, Fritz, Schiele, NeurIPS'18]

Qualitative Results - COCO Dataset



Automated Testing Framework

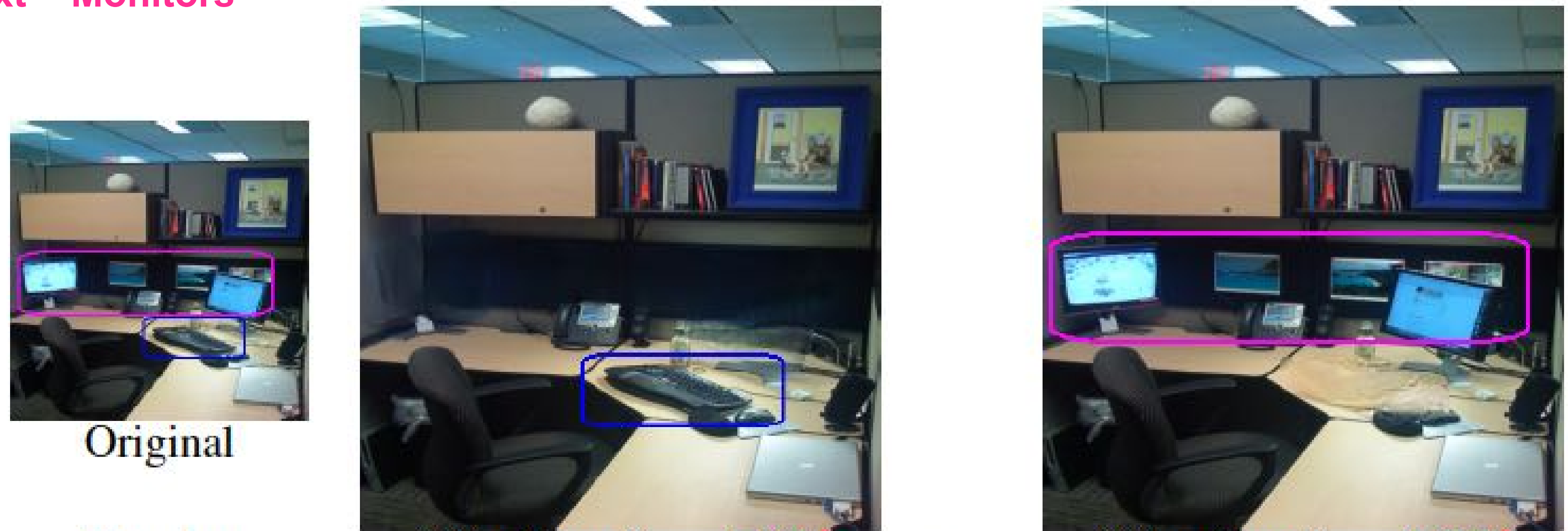
- Idea:
  - create multiple versions of the input image with one object removed in each
- Removal approach: [Shetty, Fritz, Schiele, NeurIPS'18]
  - use ground truth masks + in-painter trained for object removal
- Each image presents new context in the "neighborhood" of the original test image.



Example Result:

- Here:
  - Object = Keyboard
  - Context = Monitors

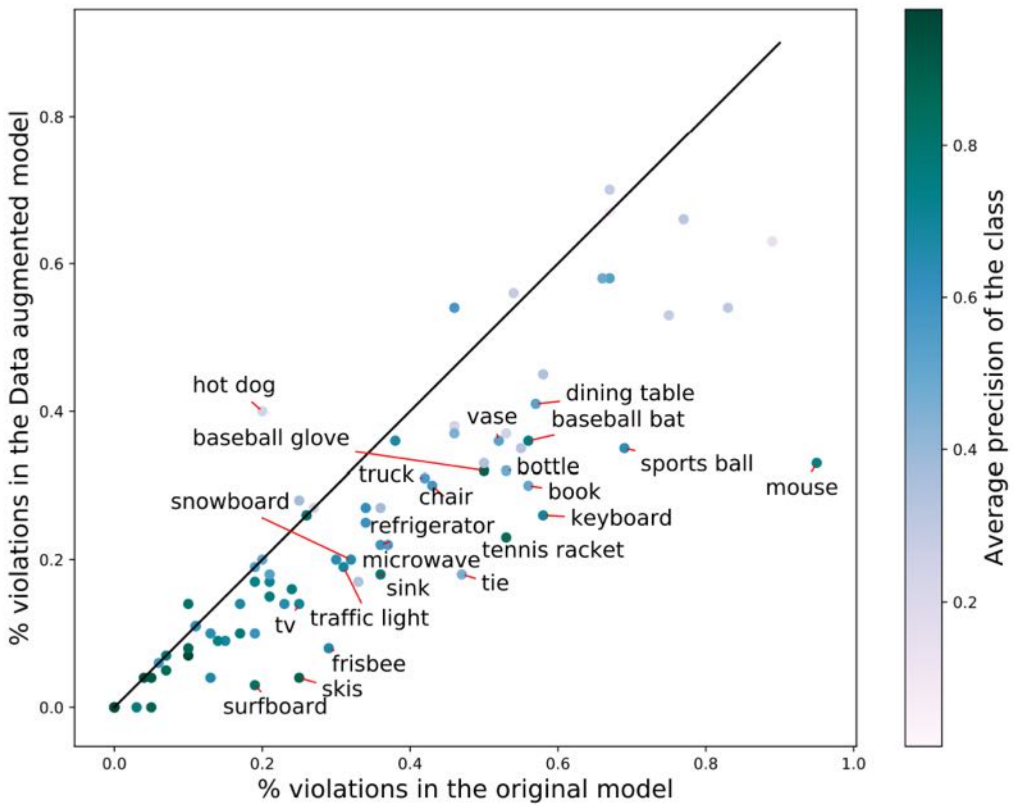
Object without Context **Context without Object**



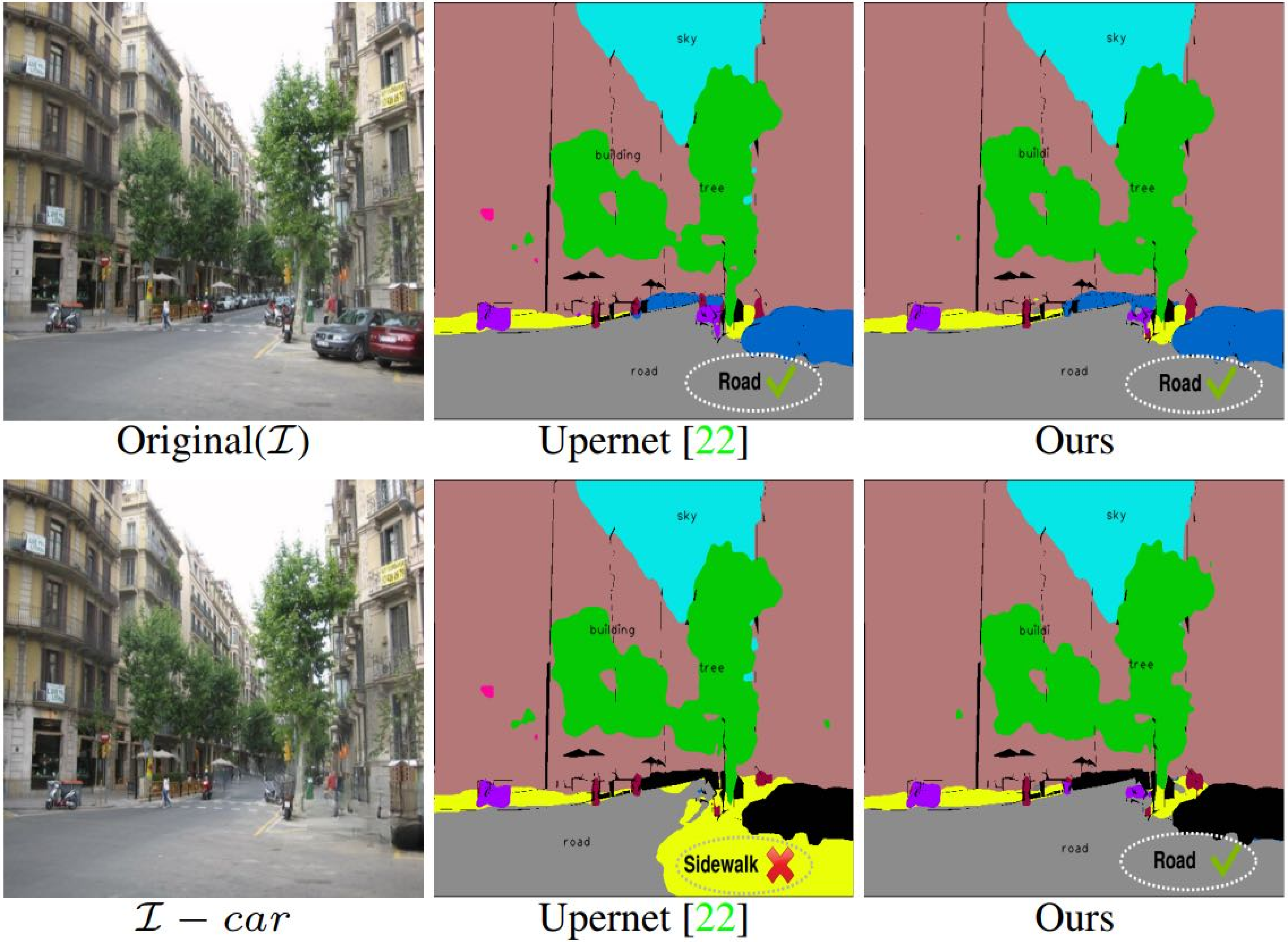
Original

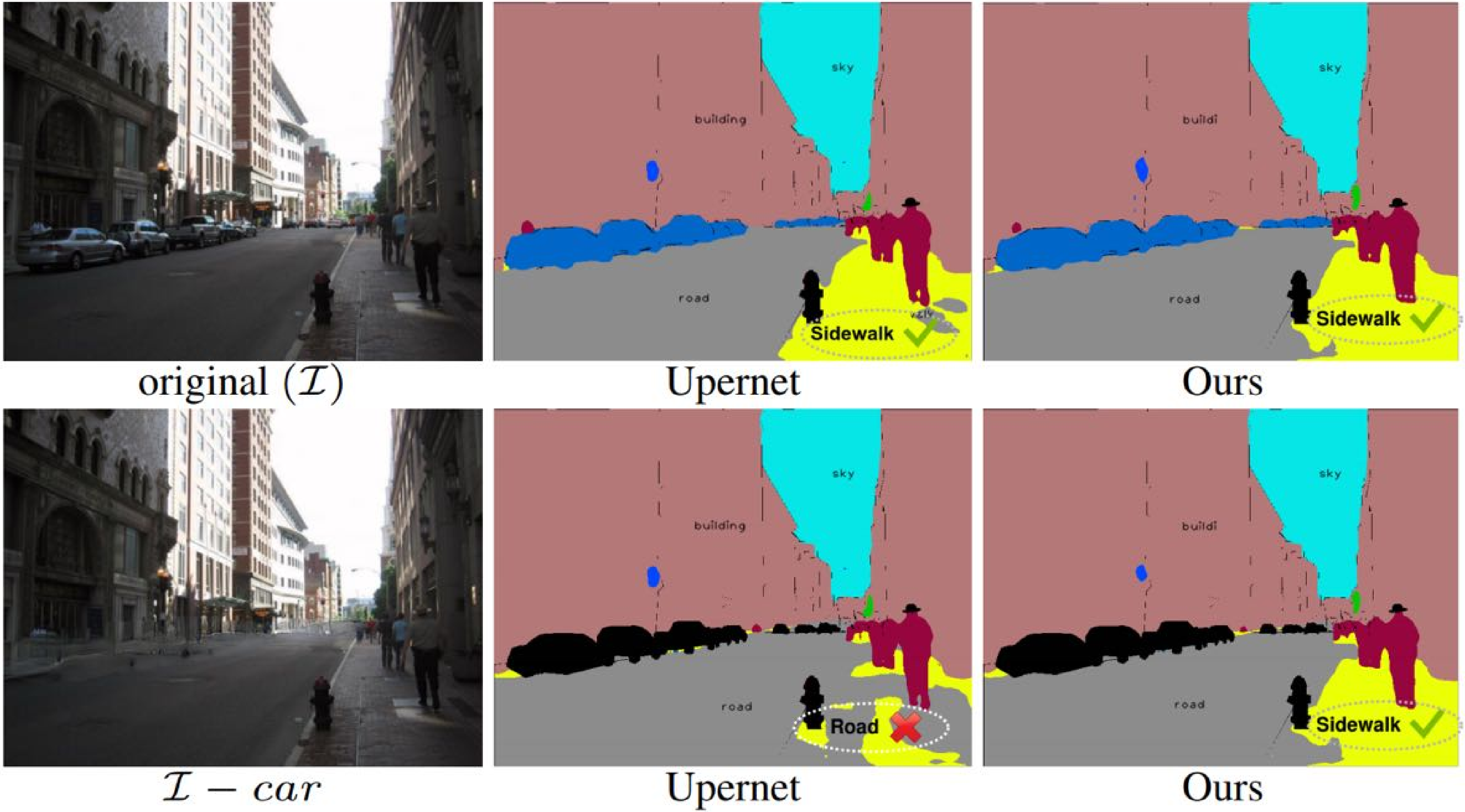
	Object without Context		Context without Object
Regular	$S(keyboard) = 1.99$ ↯	$\geq$	$S(keyboard) = 4.67$ ↯
Ours	$S(keyboard) = 3.40$		$S(keyboard) = 1.39$

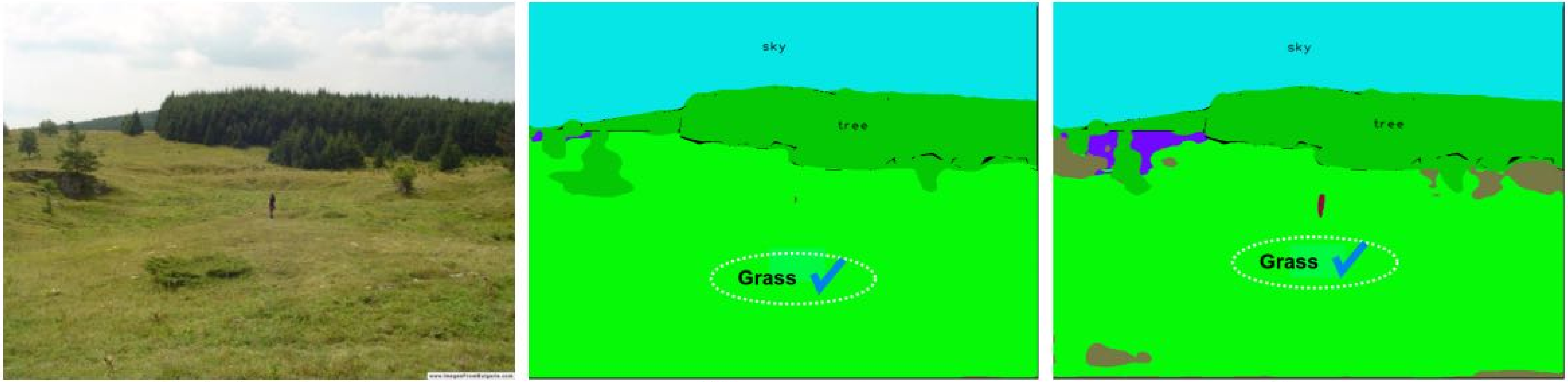
Effect of Data Augmentation on Robustness of Different Classes in Classification



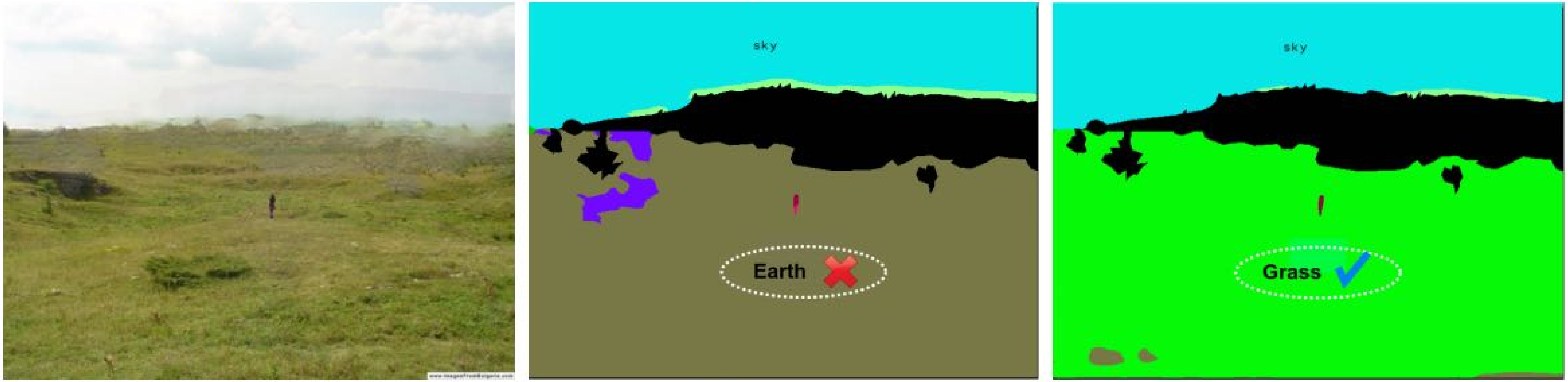
- Observations:
  - many well-performing classes are not robust to scene context changes
- Example:
  - mouse AP = 0.84, violations = 90%
  - training with data augmentation reduces this (90% drops to 36%)
- Improves performance on out of context dataset (Unrel)


sky
building
tree
road
Road
buildi
Sidewalk
Original($\mathcal{I}$)
Upernet [22]
Ours
$\mathcal{I} - car$
Upernet [22]
Ours

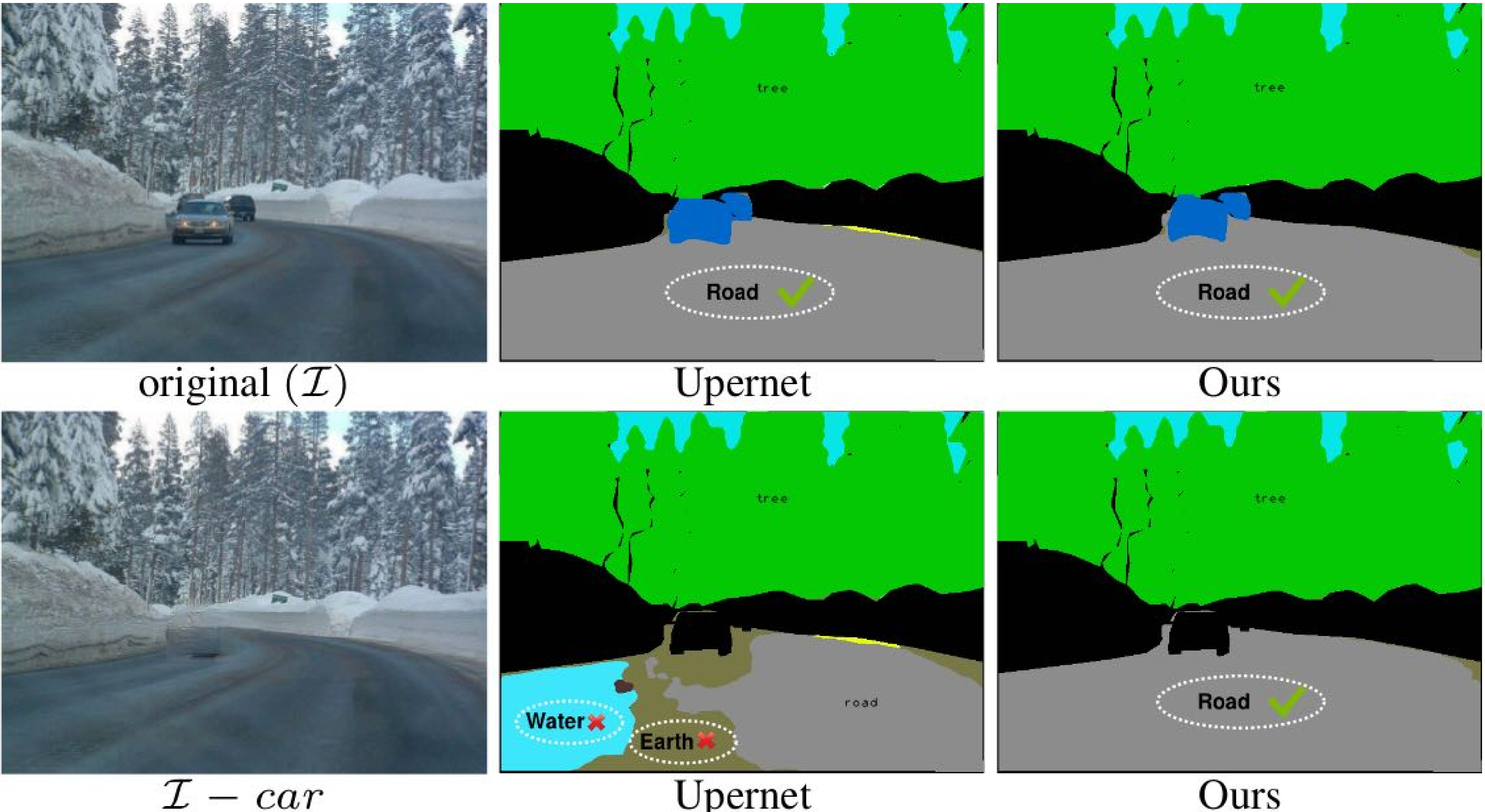

sky
building
road
Sidewalk
buildi
Road
original ($\mathcal{I}$)
Upernet
Ours
$\mathcal{I} - car$
Upernet
Ours



Original: $\mathcal{I}$ Upernet Ours



Edited: $\mathcal{I} - tree$ Upernet Ours


tree
Road
original ($\mathcal{I}$)
Upernet
Ours
Water
Earth
road
$\mathcal{I} - car$
Upernet
Ours

Take Home Message - Towards more Robust Models

- The **bright and dark sides of scene context**
  - scene context helps to achieve better performance - however **current models** are **too dependent** on **scene context**
- Proposed **new testing framework**
  - **automatically generate diverse** set of **scene context** (via object removal)
  - reveals weakness of current models
- Proposed **new data augmentation framework**
  - allows to **overcome some** of the **context dependencies**
- **More work required** !

Overview

- **Robustness** and **Security** of Deep Models
  - Bright and Dark Side of Scene Context — NeurIPS'18, CVPR'19
  - Disentangling **Adversarial Robustness** and **Generalization** — CVPR'19
  - **Reverse Engineering** and **Stealing** Deep Models — ICLR'18, CVPR'19, ICLR'20

Disentangling
Adversarial Robustness and Generalization

@ CVPR 2019

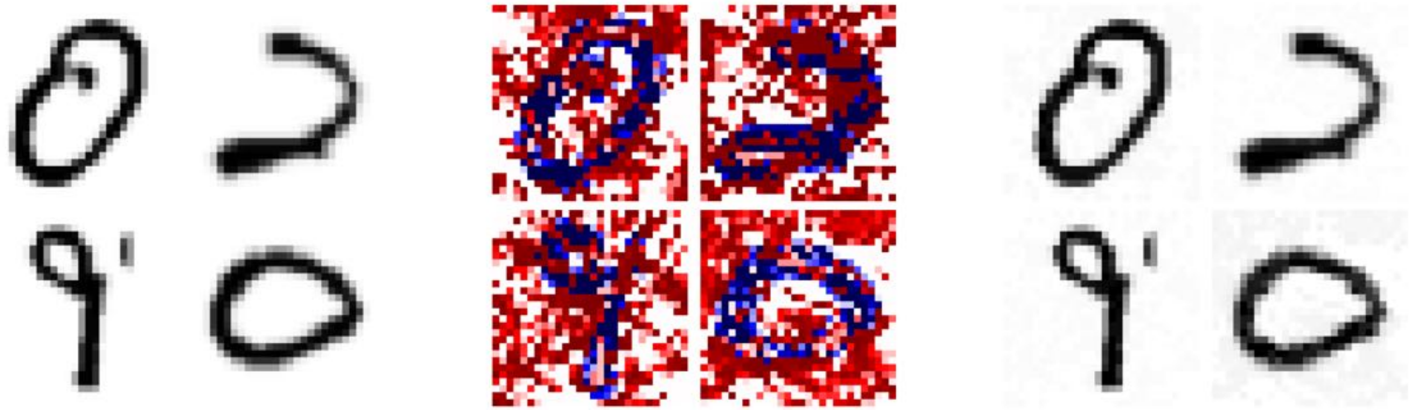


David Stutz
MPI Informatics

Matthias Hein
U Tübingen

Bernt Schiele
MPI Informatics

Adversarial Examples

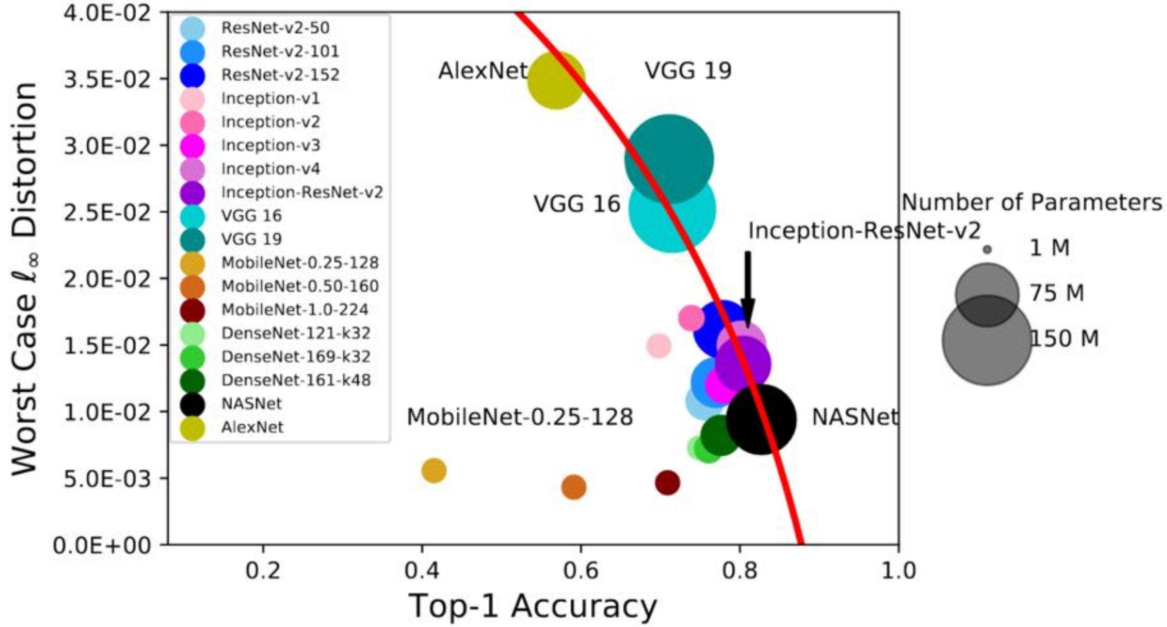


Original Images
“0”, “2”, “9” and “0”

Perturbation
$(\times 0.02 - 0.09)$

Perturbed Image
“9”, “3”, “8” and “6”

Sacrifice Robustness for Accuracy?

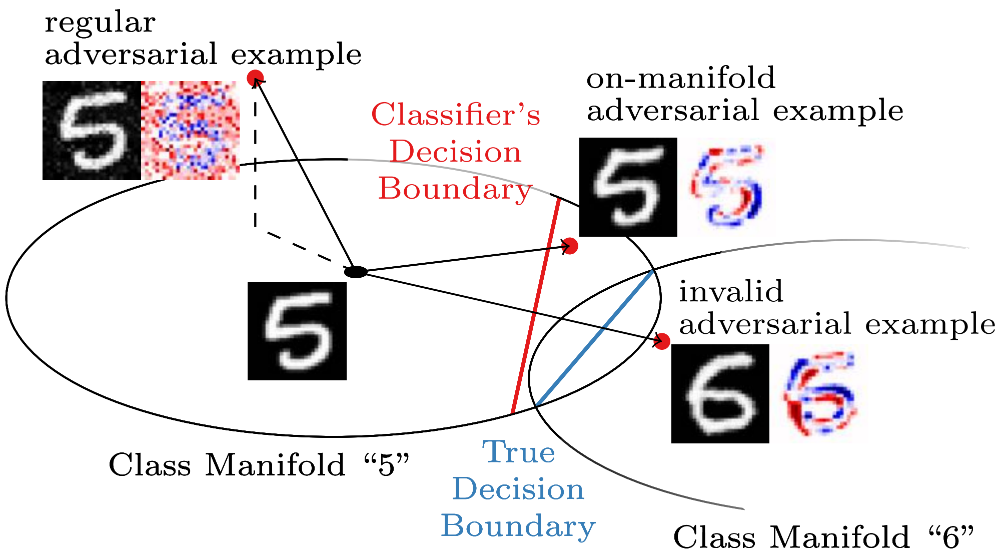


Hypothesis: Accuracy needs to be sacrified for robustness.

Su et al. *Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models*. arXiv:1808.01688.

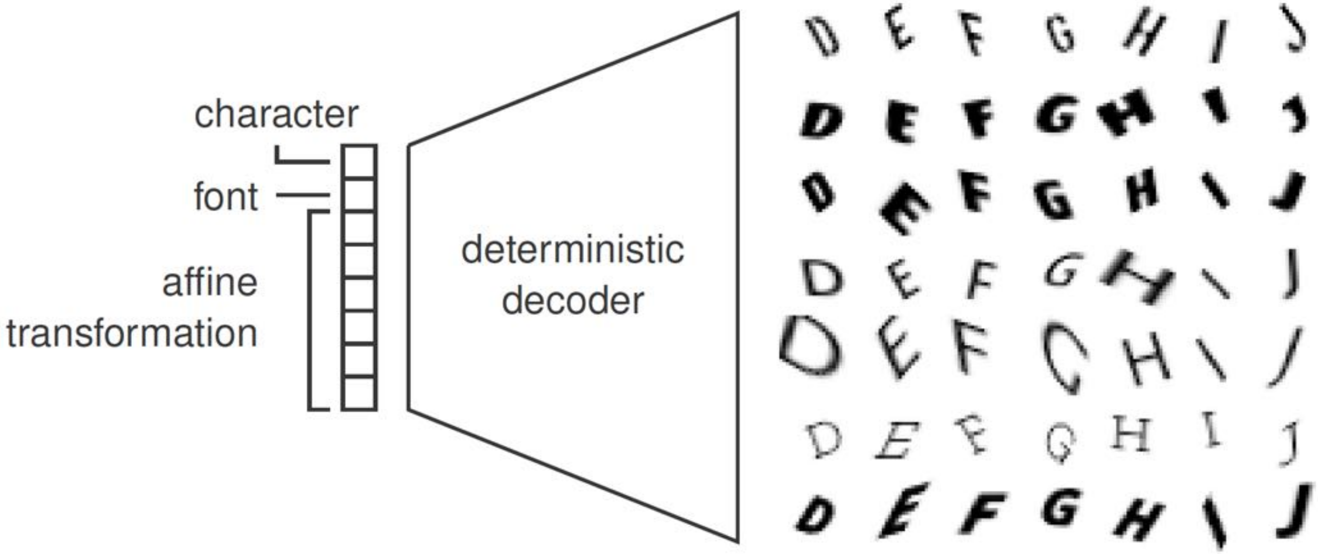
Distinction Required Between…

- **“regular”** adversarial examples
  - no constraints to be on or off the class manifold
- **“on-manifold”** adversarial examples
  - adversarial example has to be a correct instance of the class
- **“invalid”** adversarial examples
  - example is a “proper” instance of another class

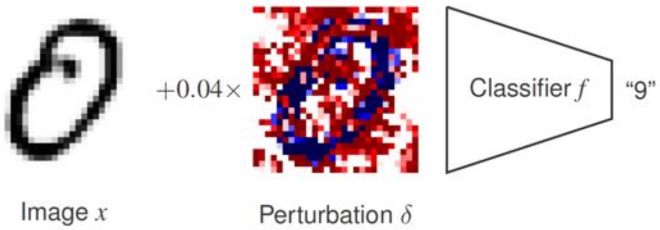


Data and Class Manifolds in the Following

- New synthetic dataset:
  FONTS: synthetic data generation with **known class manifold**
  - known manifold with perfect, deterministic generator
  - font and character are discrete; affine transformation continuous



Adversarial Examples:
Regular (Off-Manifold) Adversarial Examples



Obtain a perturbation δ for image x with true label y:

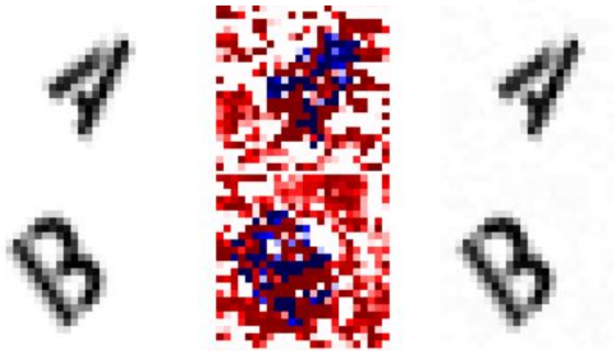
Classifier

$$\underset{\delta}{\operatorname{argmax}}\ \mathcal{L}(f(x+\delta; w), y)$$

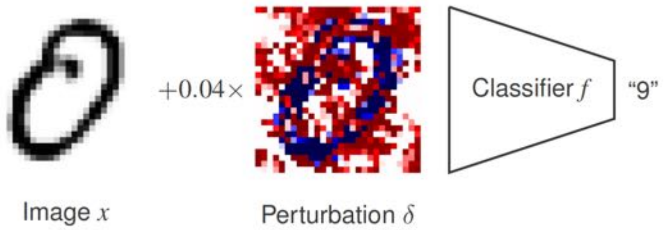
Training Loss

$$\text{s.t. } \|\delta\|_p \leq \epsilon$$

Perceptual Similarity



Adversarial Examples:
Regular (Off-Manifold) vs. On-Manifold



Obtain a perturbation δ for image x with true label y:

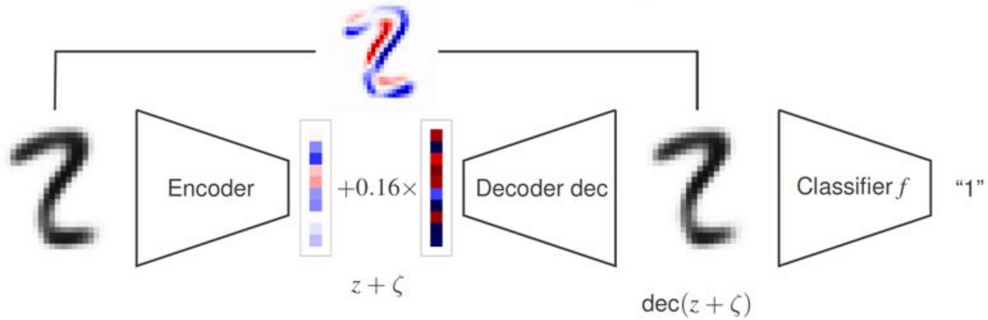
Classifier

$$\underset{\delta}{\text{argmax}}\ \mathcal{L}(f(x+\delta; w), y)$$

Training Loss

$$\text{s.t. } \|\delta\|_p \leq \epsilon$$

Perceptual Similarity



Obtain a perturbation ζ for latent code z:

Classifier

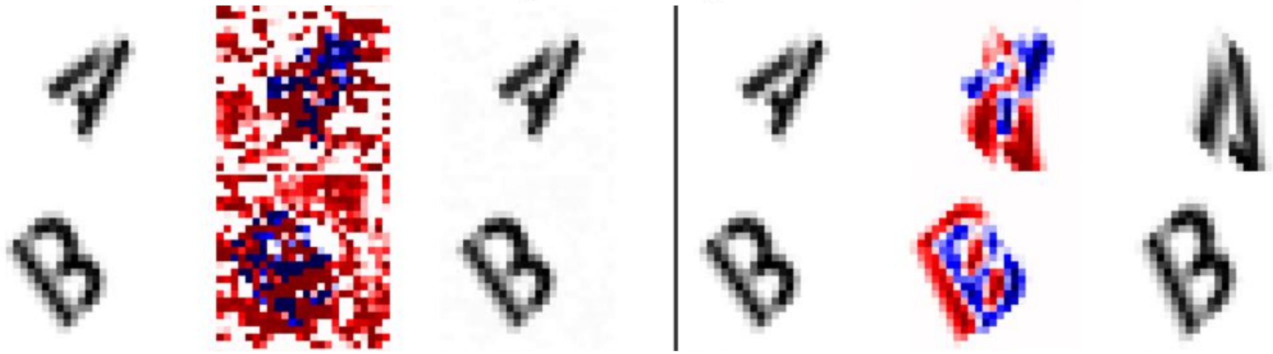
Decoder

$$\underset{\zeta}{\text{argmax}}\ \mathcal{L}(f(\text{dec}(z+\zeta); w), y)$$

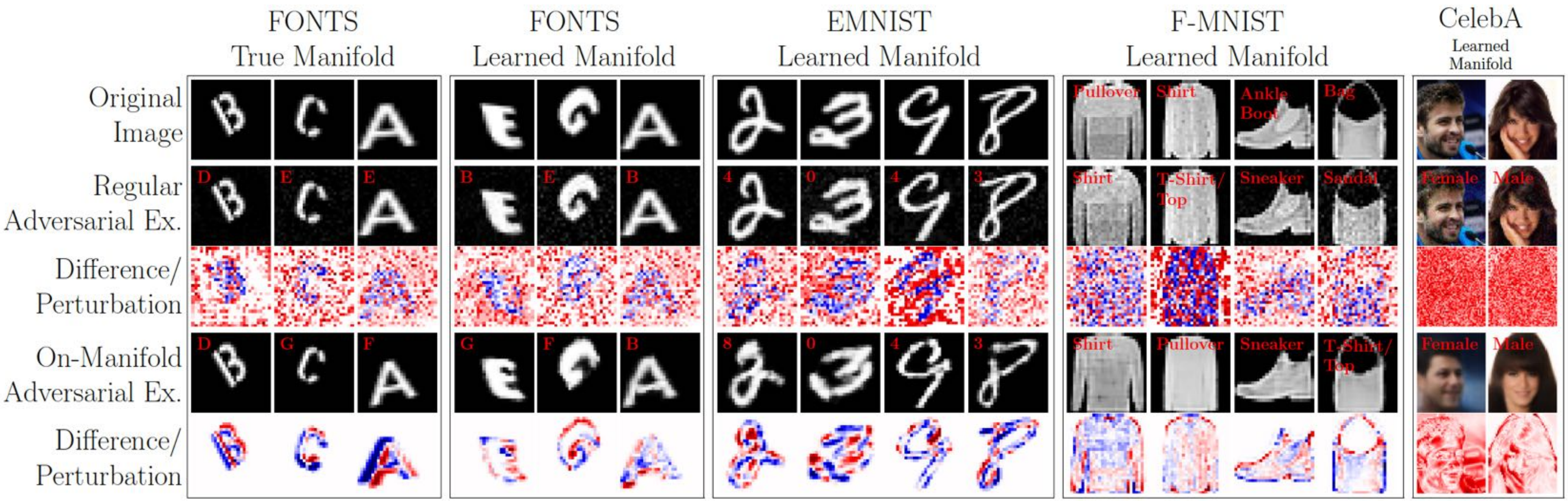
Training Loss

$$\text{s.t. } \|\zeta\|_p \leq \epsilon$$

Perceptual Similarity



Regular (Off-Manifold) vs. On-Manifold

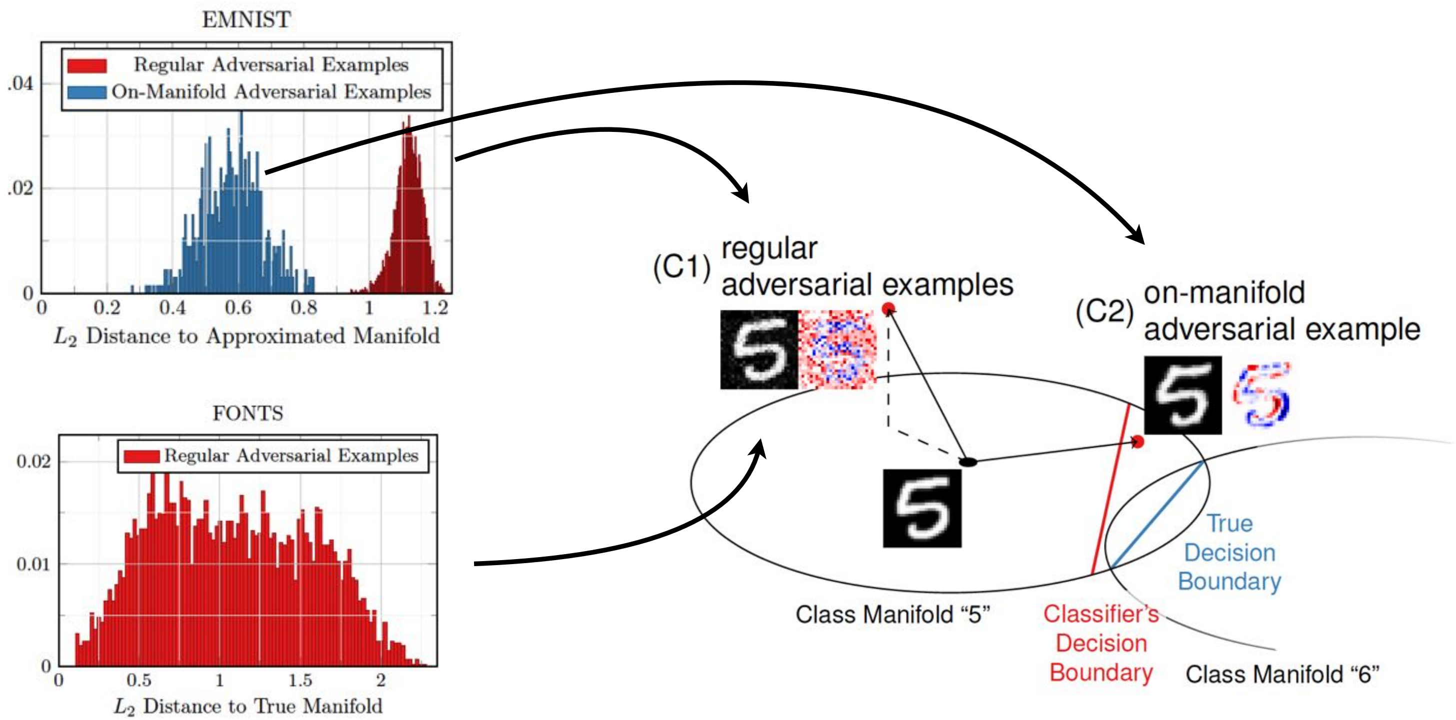


Main Findings:

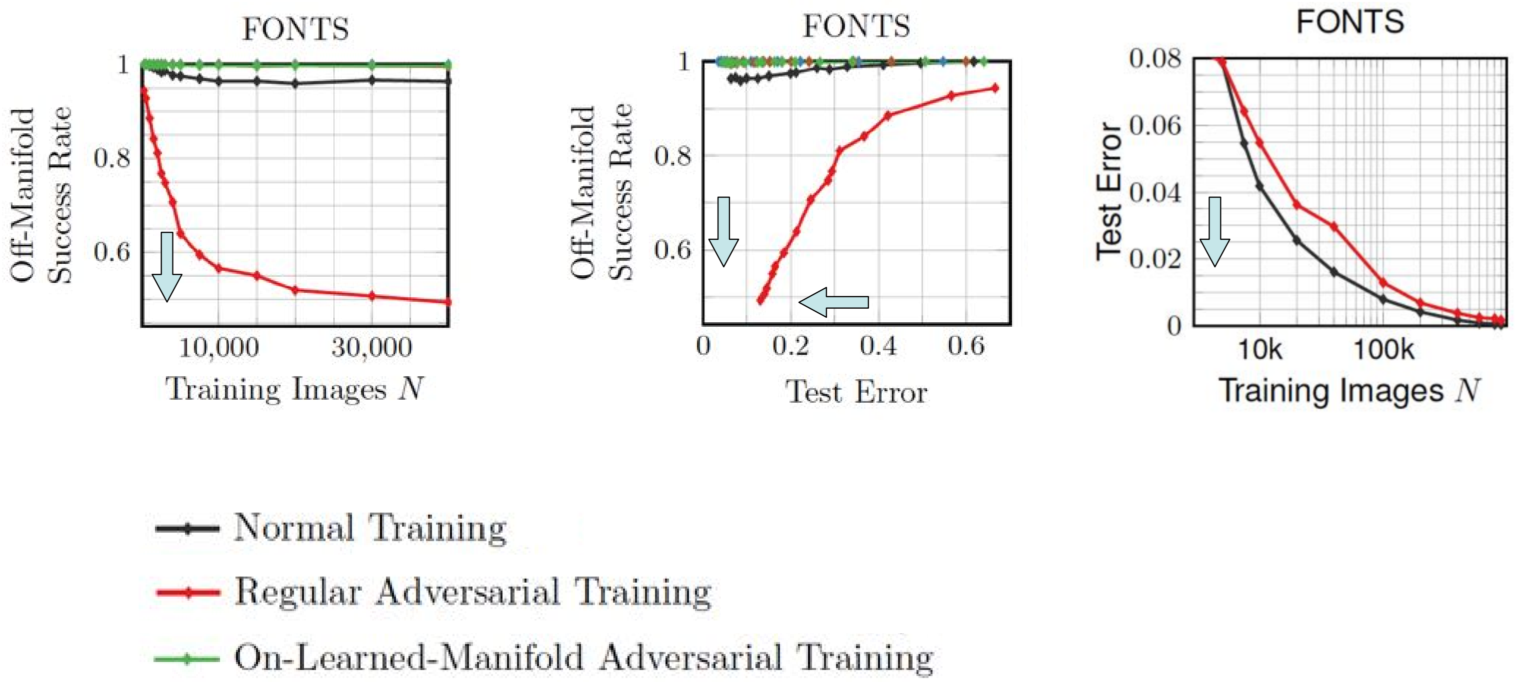
- “**Regular**” adversarial examples **leave the manifold**

manifold
learned
(VAE)

manifold
known

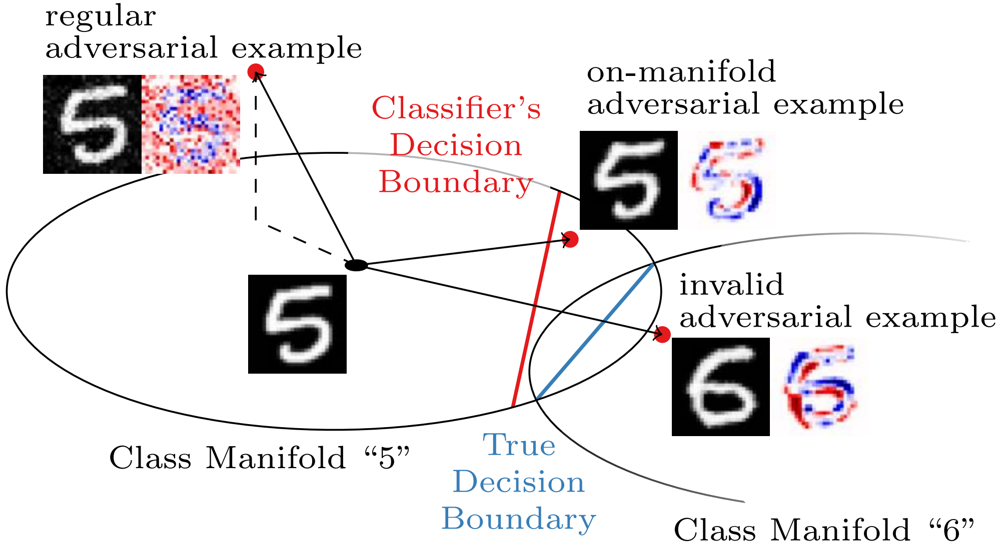


“Regular” Robustness and Generalization are NOT Contradicting



Take Home Message - Adversarial Robustness vs. Generalization

- **Adversarial robustness not well understood**
  - distinction between **“regular”**, **“on-manifold”**, and **“invalid”** adversarial examples
  - currently very active area — not all work is great :)
  - **“regular”** adversarial examples **leave the manifold** (= “off-manifold”)
  - **“regular” robustness** and **generalization** are **not contradicting**
    - but sample efficiency is an issue
  - **“on-manifold”** adversarial examples **exist**
    - **“on-manifold” robustness is generalization**



Final Words…

- Embrace the **"Bright and the Dark Side"**
  - let's **better understand** and **control robustness & security (& privacy)**

- We need a **lot more research** in the area
  - keep **knowledge** in the **public** domain to build **trust**

- Responsibility in **education**
  - **educate students** about **both opportunities** and **potential dangers**
  - **distinguish** between **"what can be done"** and "**what should be done"**