Interpreting Adversarial Trained Convolutional Neural Networks

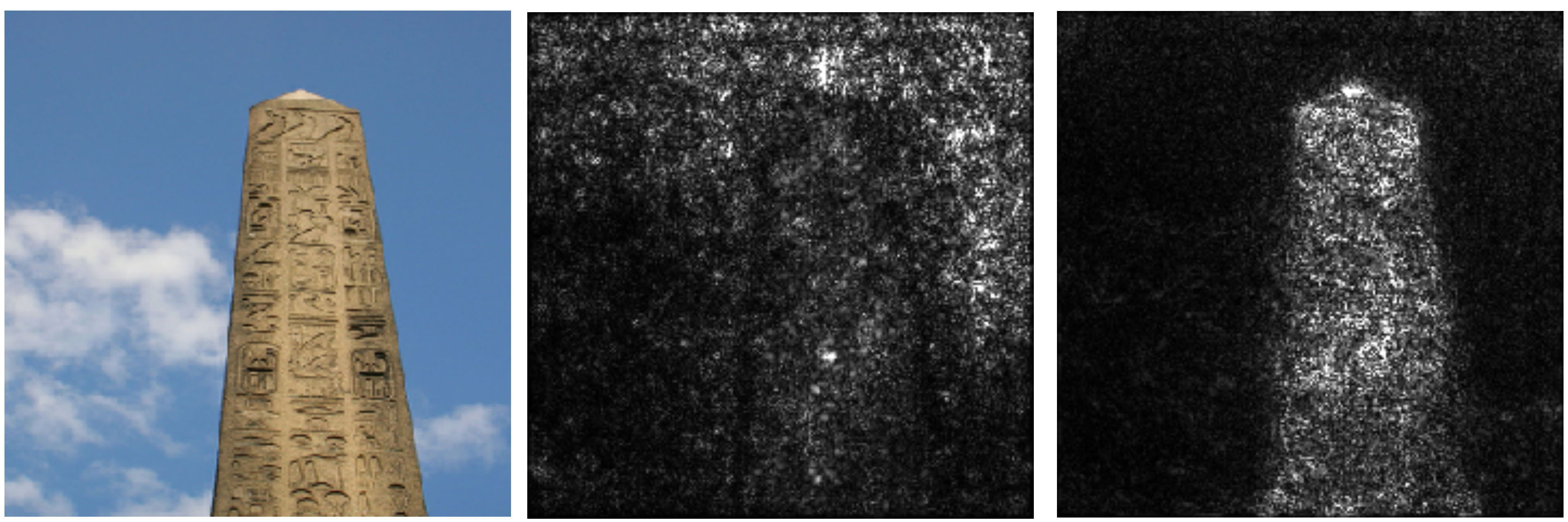
Tianyuan Zhang, Zhanxing Zhu
Peking University
1600012888@pku.edu.cn zhanxing.zhu@pku.edu.cn

Contents

- Normally trained CNNs typically lack of interpretability
    - Biased towards textures
- Adversarially trained CNNs could improve interpretability
    - Capture more semantic features: shapes.
    - Systematic experiments to validate the hypothesis
- Discussions

Sensitivity Map

- **Grad:** input gradient $$E = \frac{\partial S_c(x)}{\partial x} \qquad S_c(x) = \log p_c(x)$$
  - the gradient of the class score function w.r.t. input image
- **SmoothGrad** $$E = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial S_c(x+g_i)}{\partial (x+g_i)}$$
  - Removing the noise by averaging the noise $$g_i \sim \mathcal{N}(0, \sigma^2)$$



Input image **Grad** **SmoothGrad**

Smilkov et.al (2017) SmoothGrad: removing noise by adding noise

Normally Trained CNN

- Interpreting normally trained CNN: **texture bias**

Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

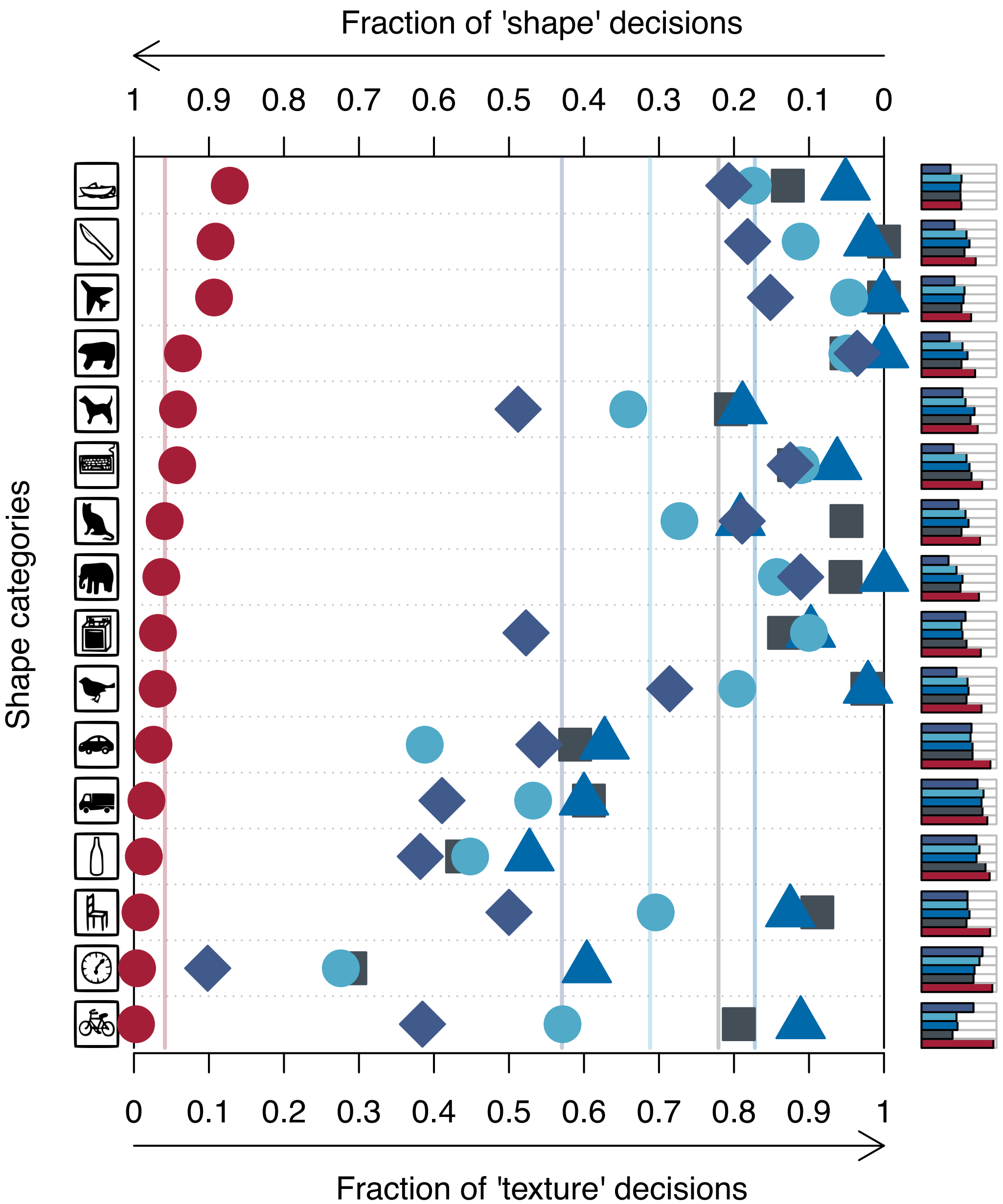
Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

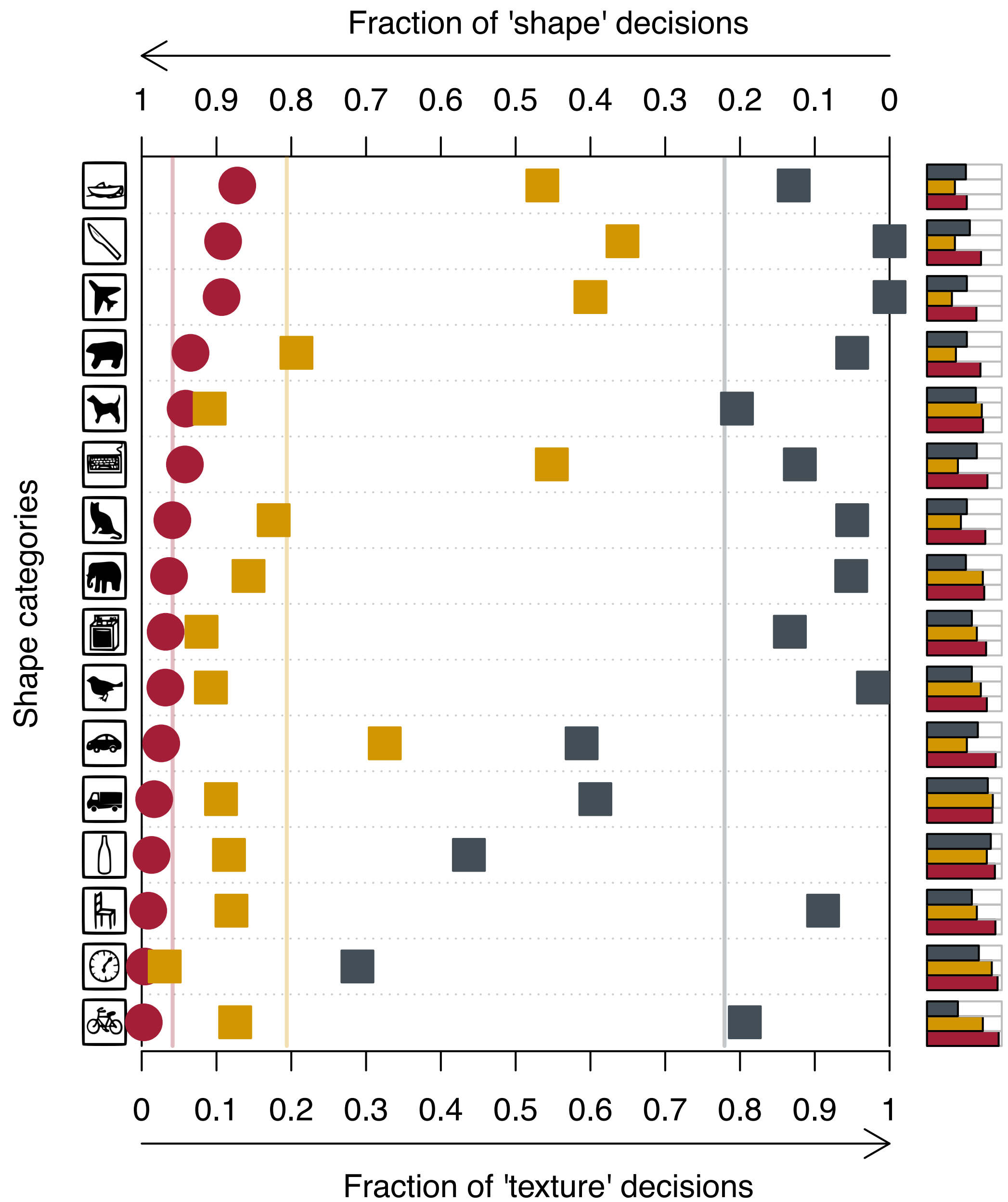
Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan





Augmented Stylized-ImageNet could improve shape bias.

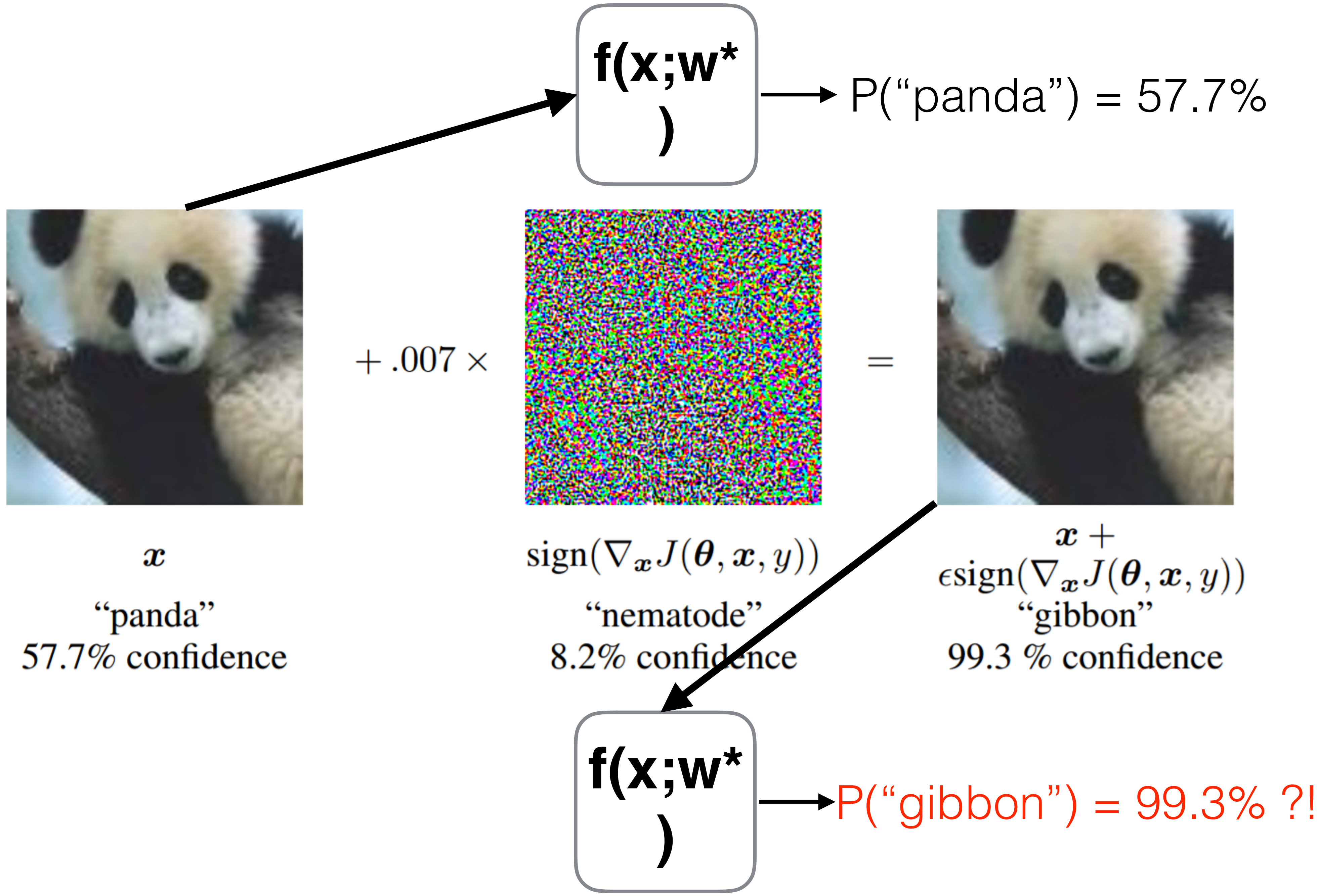
Are there any other models that could improve shape bias?

↓

Adversarially trained CNNs!

Adversarial Examples

- Deep neural networks are easily fooled by adversarial examples. **Not robust!**



Adversarial Training

- Adversarial training for defensing adversarial examples:
  - A robust optimization problem

Projected Gradient Descent

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in S} \ell(f(x+\delta;\theta), y) \right]$$

→ $\|\delta\| \leq \varepsilon$

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(f(x;\theta), y) \right]$$

→ Standard training

- Interpreting adversarially trained CNNs (**AT-CNNs**)
  - What have AT-CNNs learned to make them robust?
  - **Compared with standard CNNs, AT-CNNs tend to be more shape-biased.**

Two ways for interpreting AT-CNNs

- Qualitative method
  - Visualizing sensitivity maps
- Quantitative method
  - Evaluate the generalization performance on either **shape or texture preserved data sets**

Constructing Datasets

1. Stylizing: shape preserved, texture destroyed

2. Saturating: shape preserved, texture destroyed

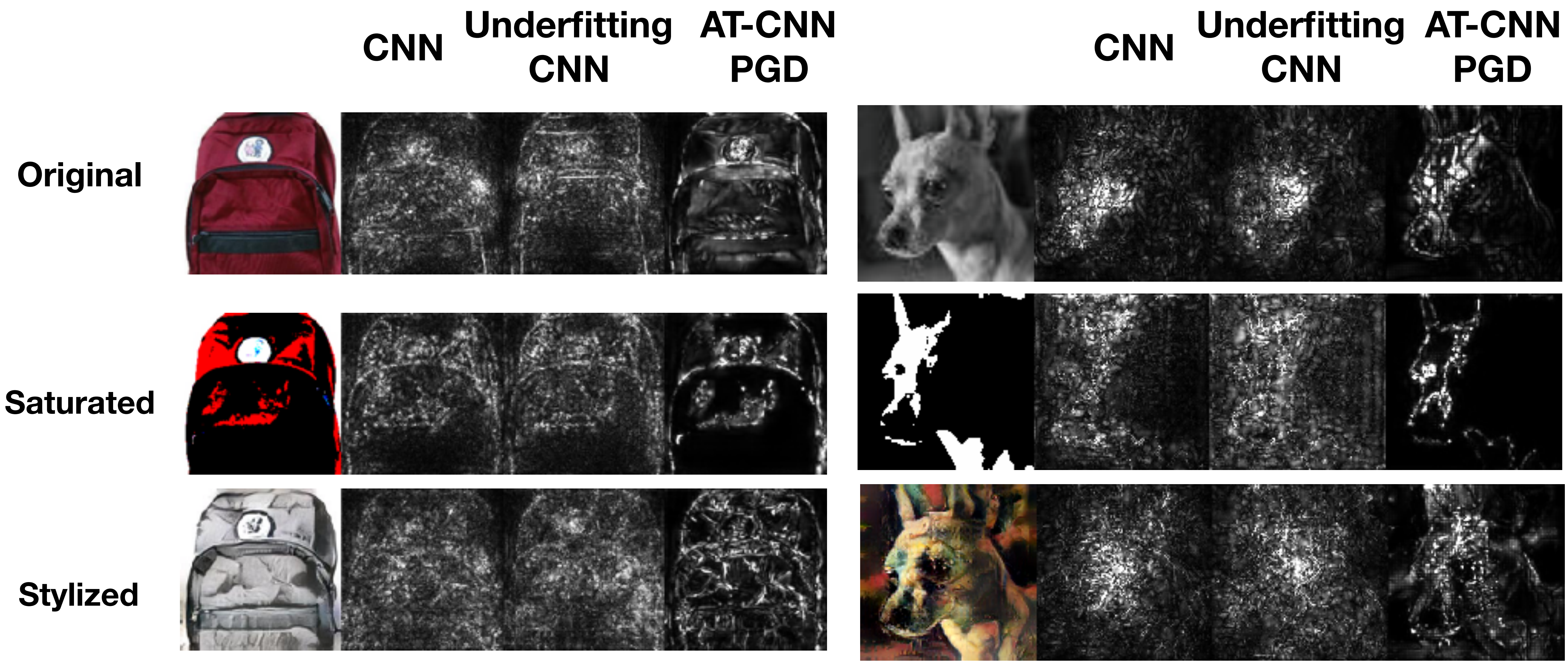
3. Patch-shuffling: shape destructed, texture preserved

(a) Original (b) Stylized (c) Saturated 8 (d) Saturated 1024 (e) patch-shuffle 2 (f) patch-shuffle 4

Figure 1. Visualization of three transformations. Original images are from Caltech-256. From left to right, original, stylized, saturation level as 8, 1024, 2×2 patch-shuffling, 4×4 patch-shuffling.

Sensitivity maps of AT-CNNs

SmoothGrad

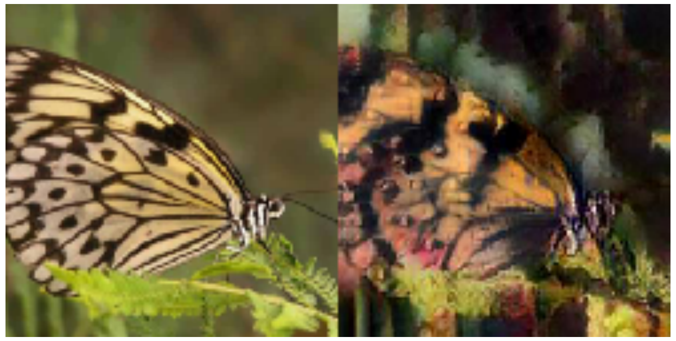


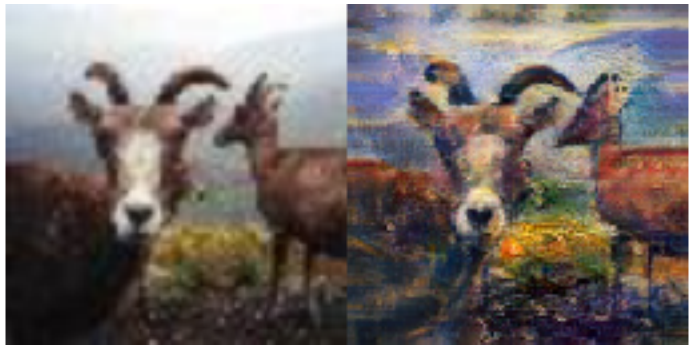
Generalization on Constructed Datasets

- **Stylized data**

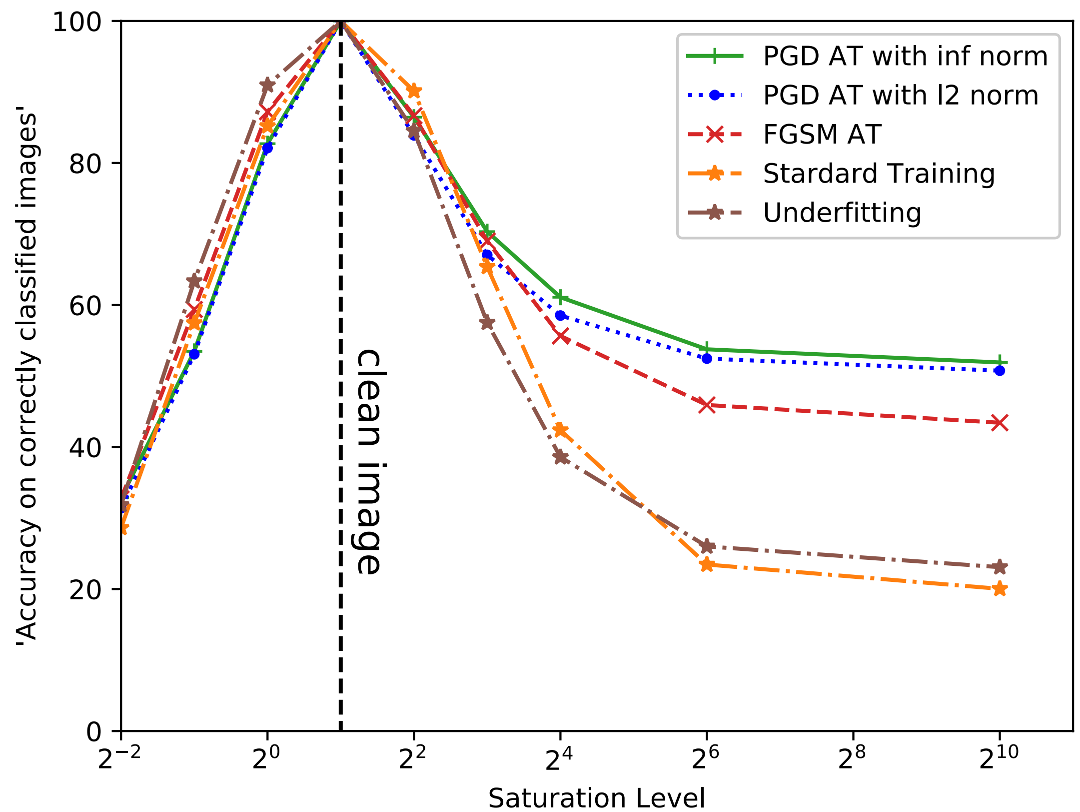
Accuracy on correctly classified images

Dataset	Cal-256	Stylized Cal-256	TinyINt	Stylized TinyIN
Standard	**83.32**	16.83	**72.02**	7.25
Underfit	69.04	9.75	60.35	7.16
PGD-l_2: 4	74.12	**22.53**	64.24	**21.05**

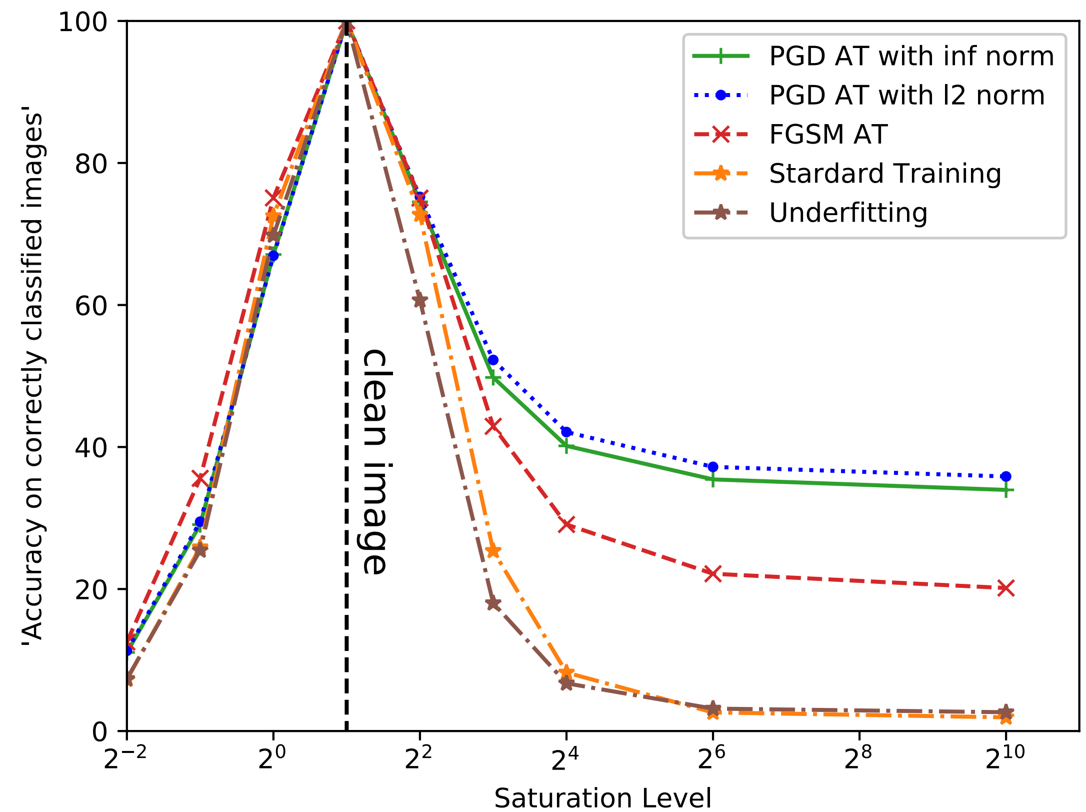




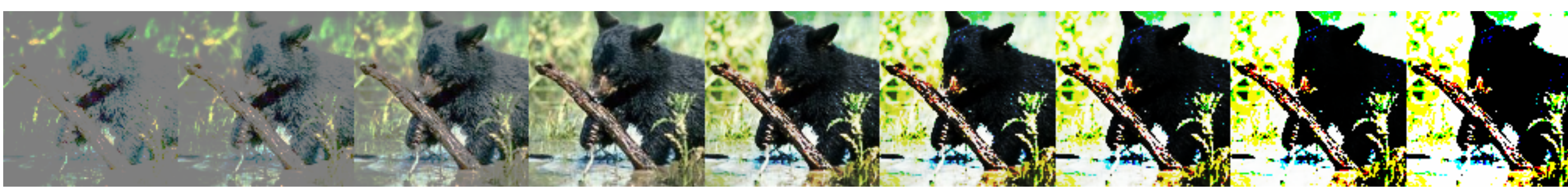
- # Saturated data



Caltech-256

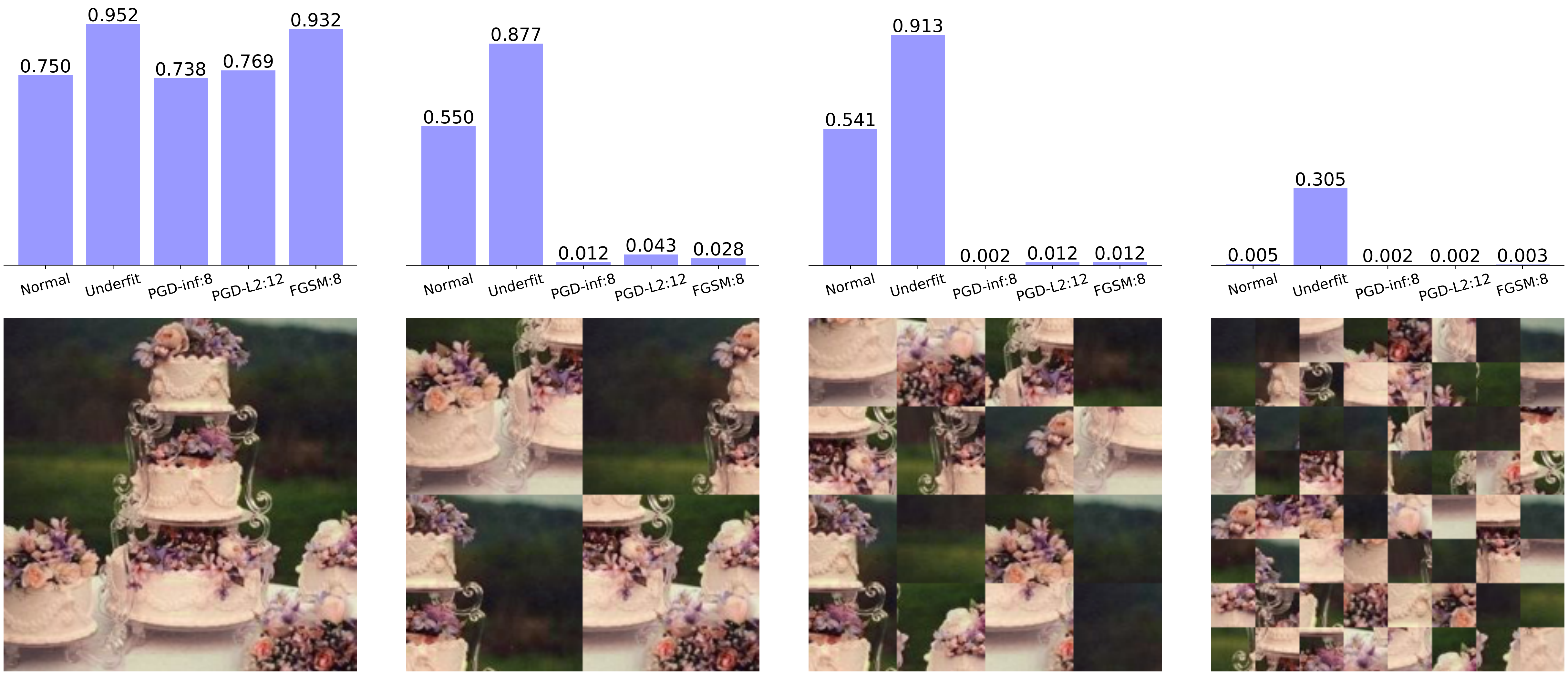


Tiny ImageNet

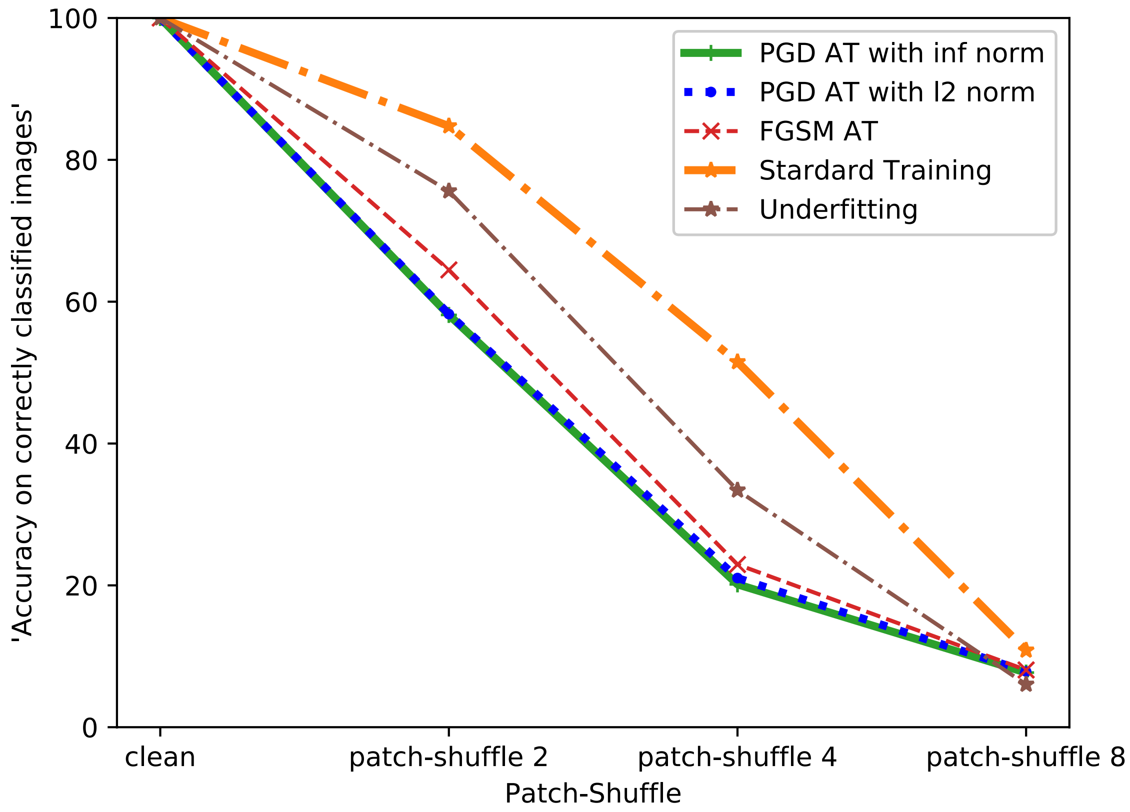


Loosing both texture and shape info. → **Loosing texture and preserve shape info.**

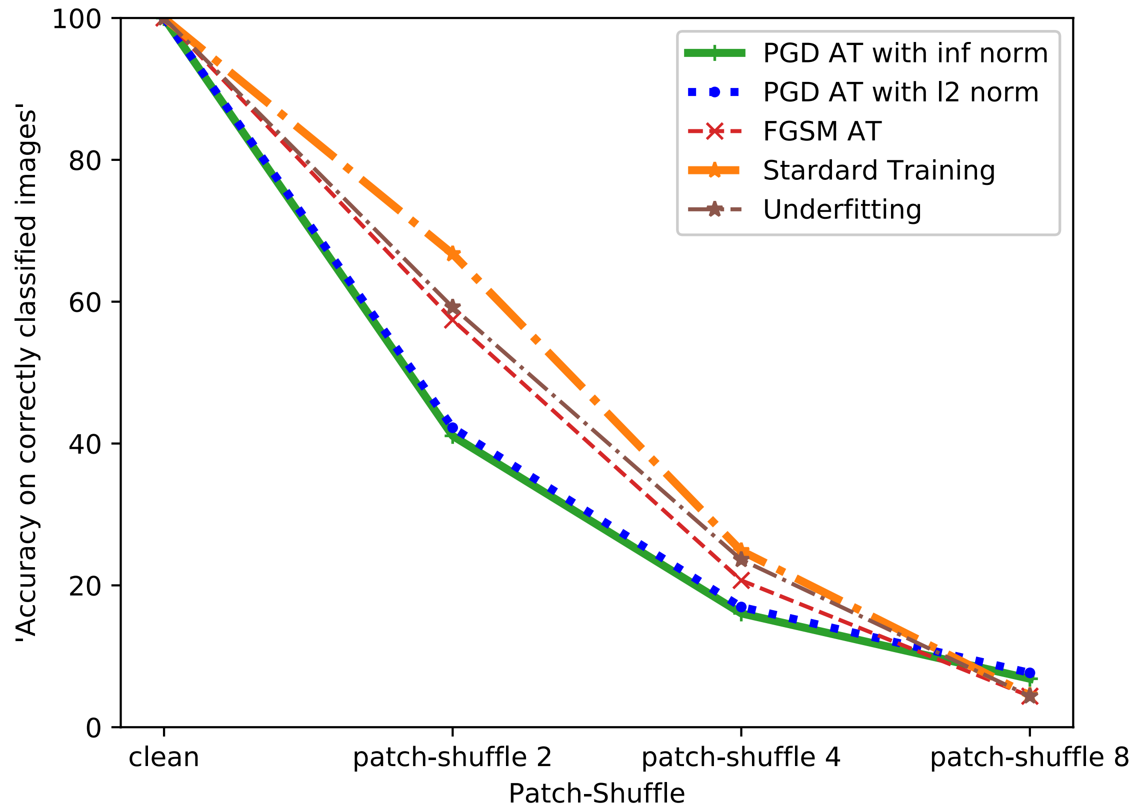
- **Patch-shuffled data**



(a) Original Image (b) Patch-Shuffle 2 (c) Patch-Shuffle 4 (d) Patch-Shuffle 8



Caltech-256



Tiny-ImageNet

Discussions

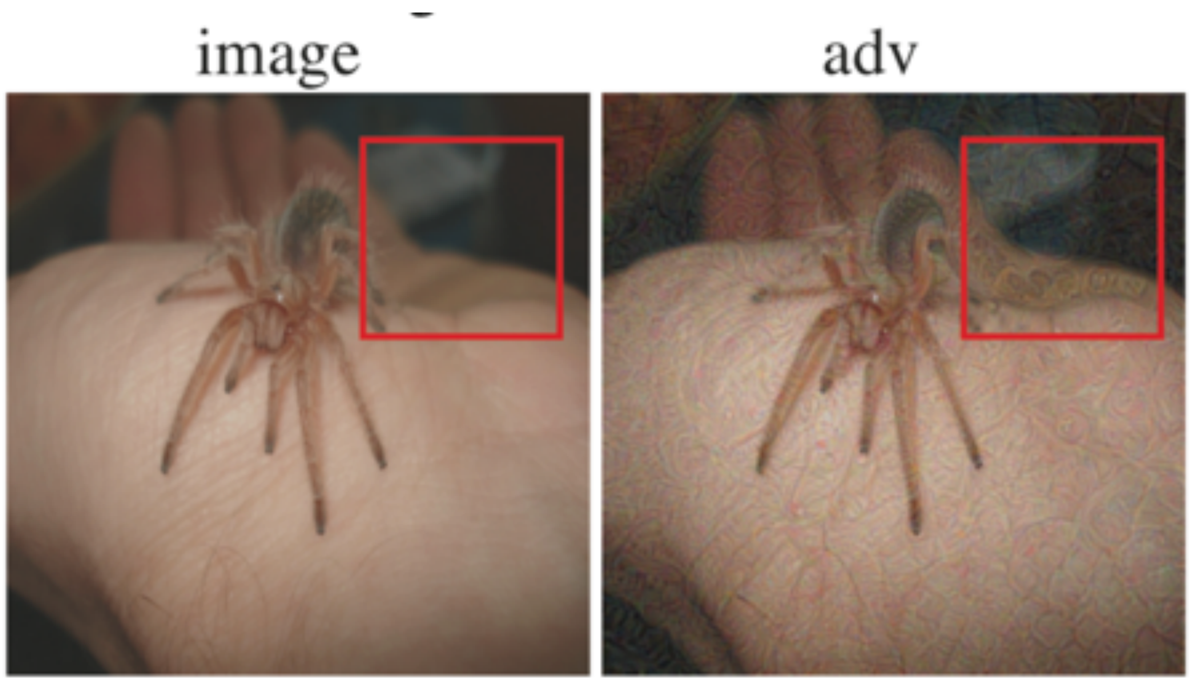
- Interpreting adversarially trained CNNs
    - Adversarial training helps capturing global structures, a more shape-based representation
    - We provide both qualitative and quantitive ways for model interpretation.

Discussions

- Insights for defensing adversarial examples
  - Whether models better capturing long-range representation tend to be more robust (e.g, non-local, Xie, et al 2018) ?
- Interpreting AT-CNNs based on other types of adversarial attacks
  - Spatially transformed adv. examples (Xiao et.al 2018)
  - GAN-based adv. examples (Song et.al 2018)

Why?

- PGD attack often change local features



- Adversarial training acts like data augmentation, which can effectively increase invariance against corruptions of local features

Thanks!
Q & A