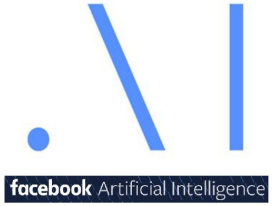
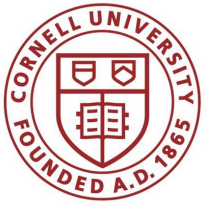




Fine-grained Generation of Unrestricted Adversarial Examples

Omid Poursaeed

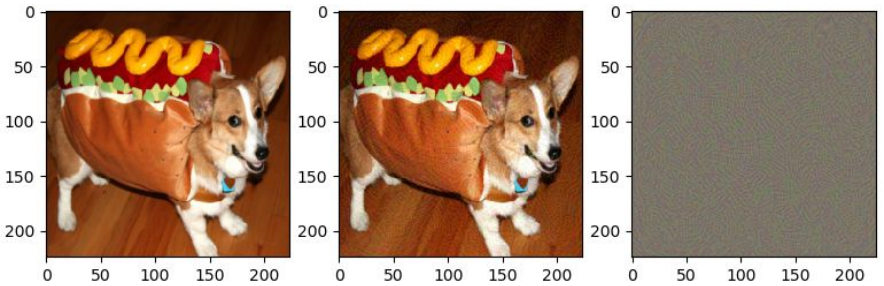
Adversarial Image Manipulation

Real					
	Tabby Cat	Not hot dog	Stop Sign	Cat	Male
Adversarial					
	Guacamole	Hot dog	Max Speed 100	Cell Phone	Female

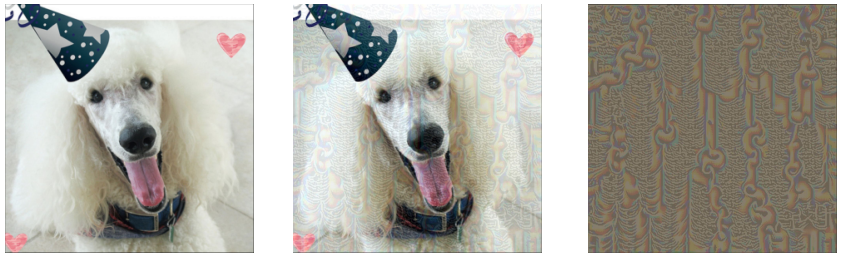
- Look like real images
- Misclassified by the model

Similarity of Images

L_p similarity: $||x - \hat{x}||_p < \epsilon$ $\quad p = 0, 2, \infty, \ldots$

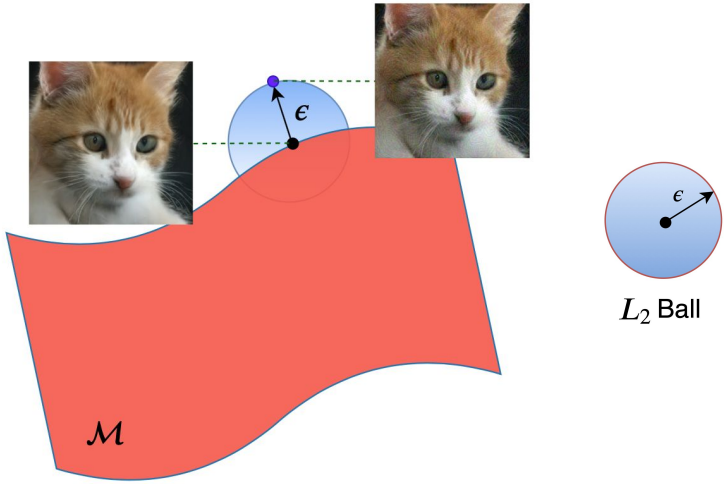


Not hot dog hot dog

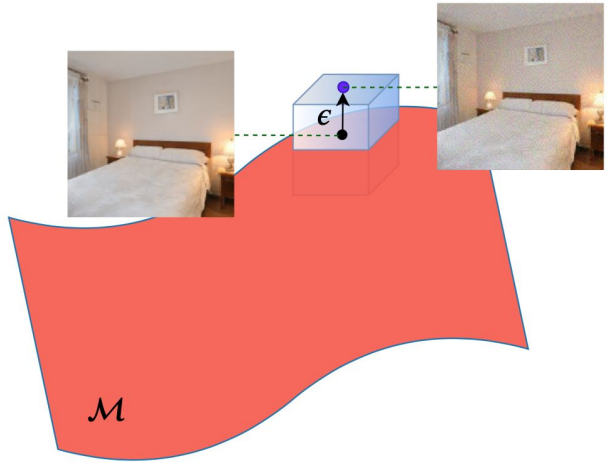


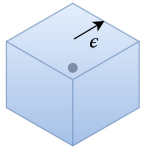
Poodle Chain

Manifold of Natural Images



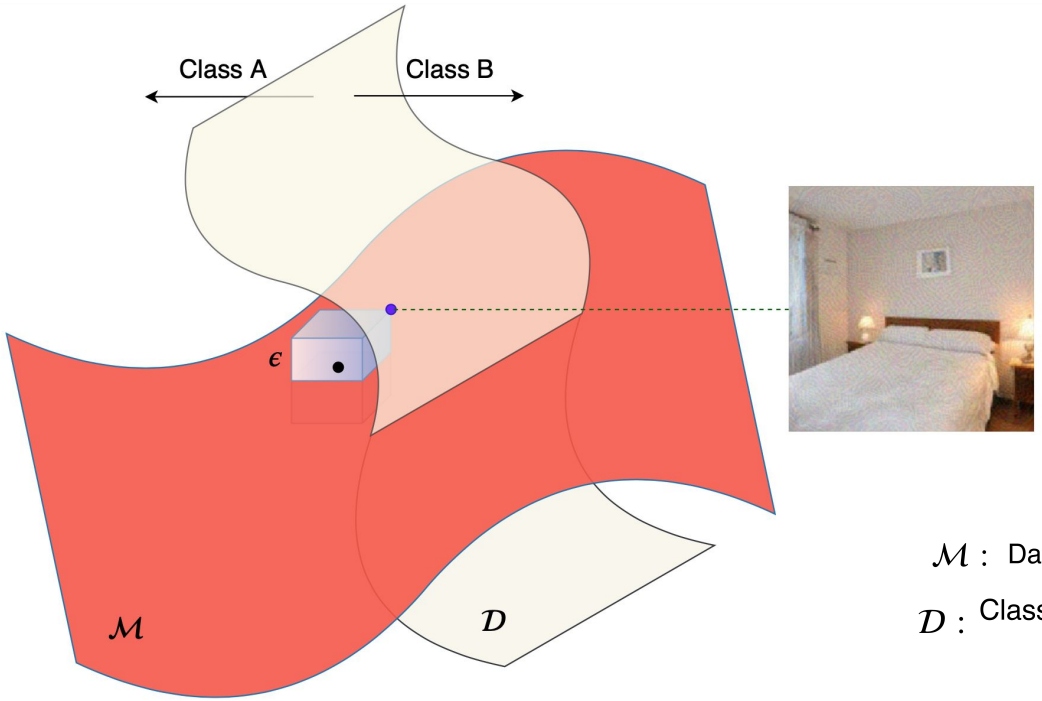
L_2 Ball





L_∞ Ball

Manifold of Natural Images



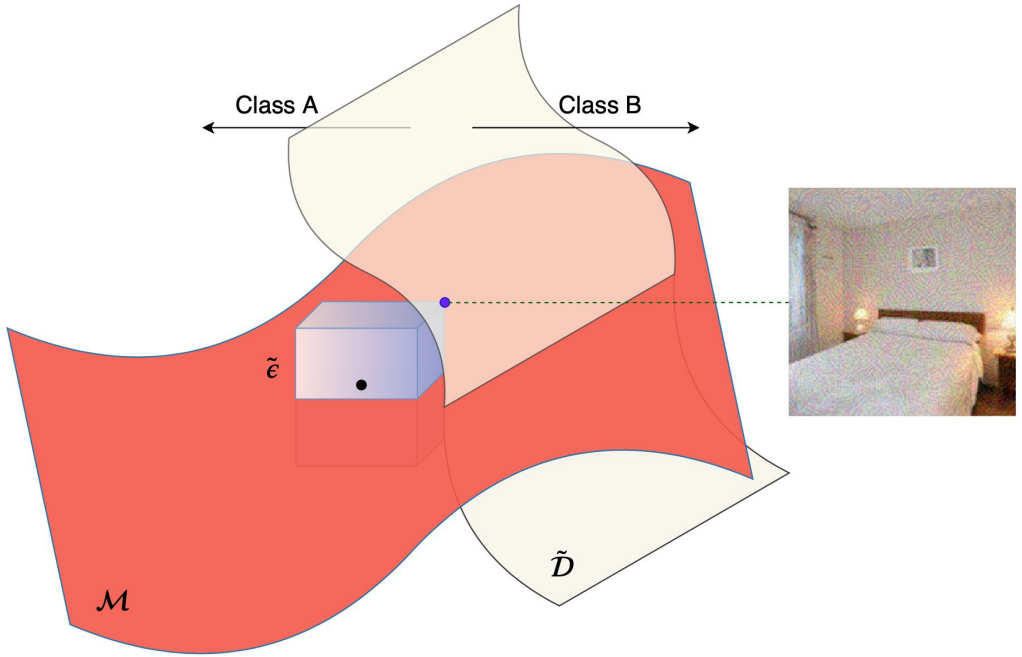
$\mathcal{M}$: Data Manifold

$\mathcal{D}$: Classifier's decision boundary

Manifold of Natural Images

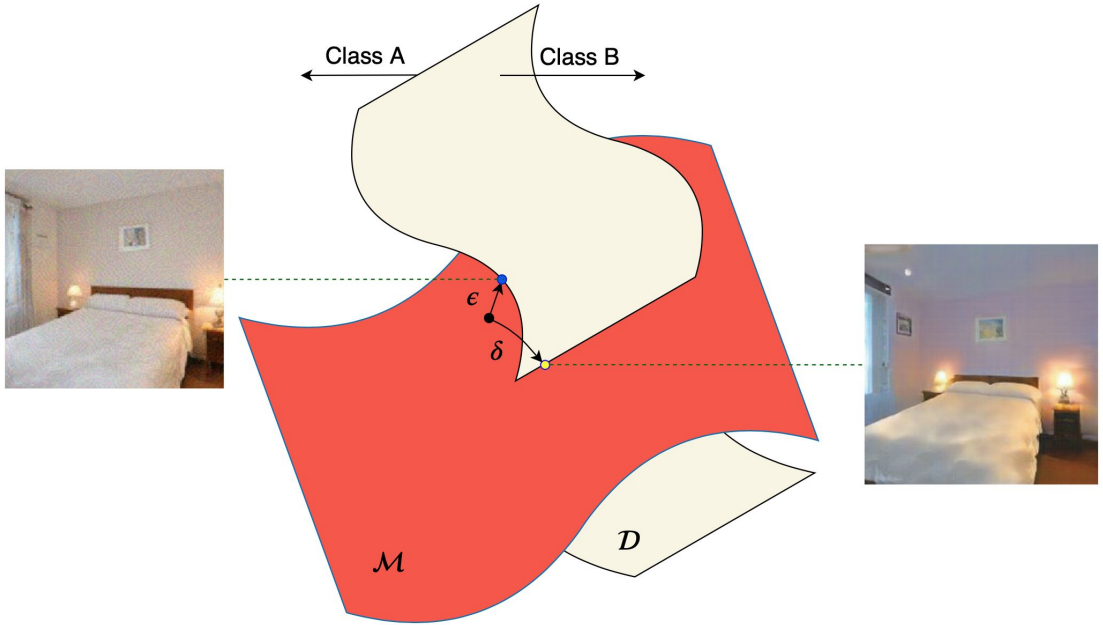
Classifiers equipped with defense

Larger perturbation norms



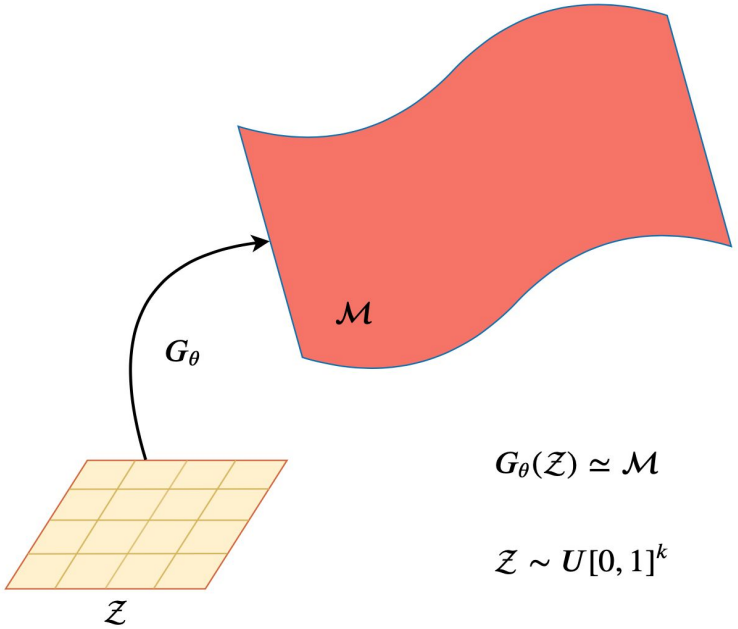
Unrestricted Adversarial Examples

Can we move on the manifold?



Unrestricted Adversarial Examples

Using a generative model to approximate the manifold



$G_\theta(\mathcal{Z}) \simeq \mathcal{M}$

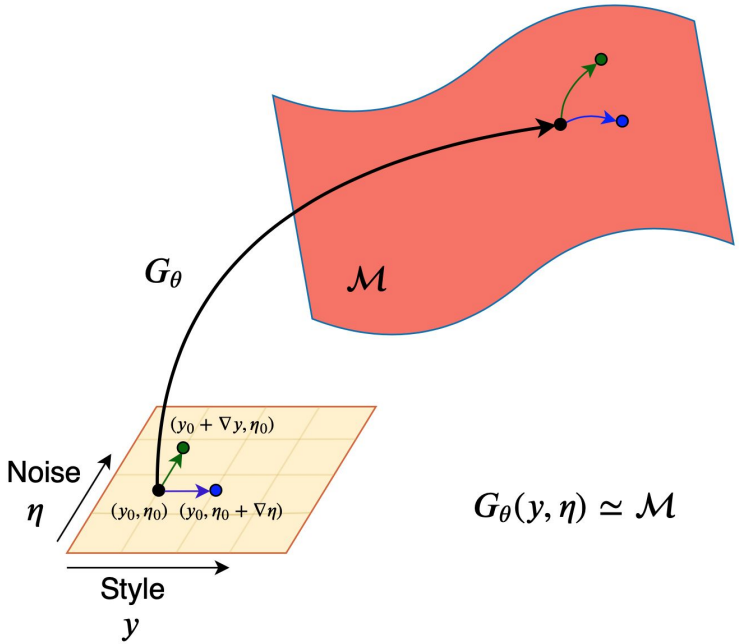
$\mathcal{Z} \sim U[0,1]^k$

G_θ : Generator

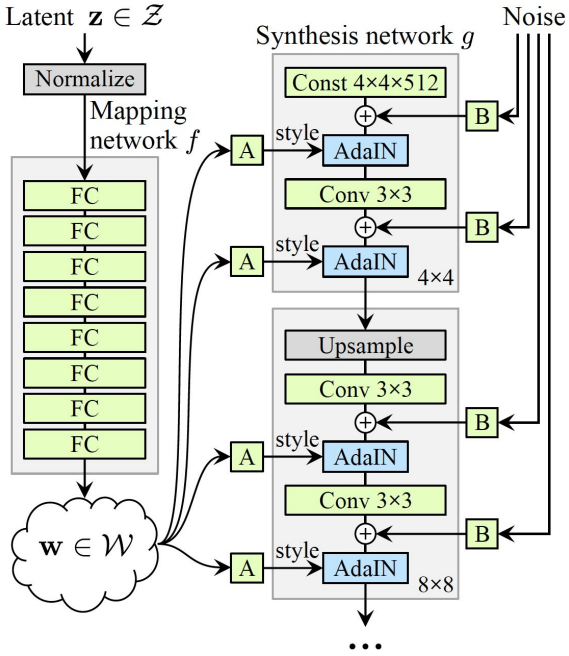
$\mathcal{M}$: Data Manifold

Unrestricted Adversarial Examples

Disentangled Latent Space

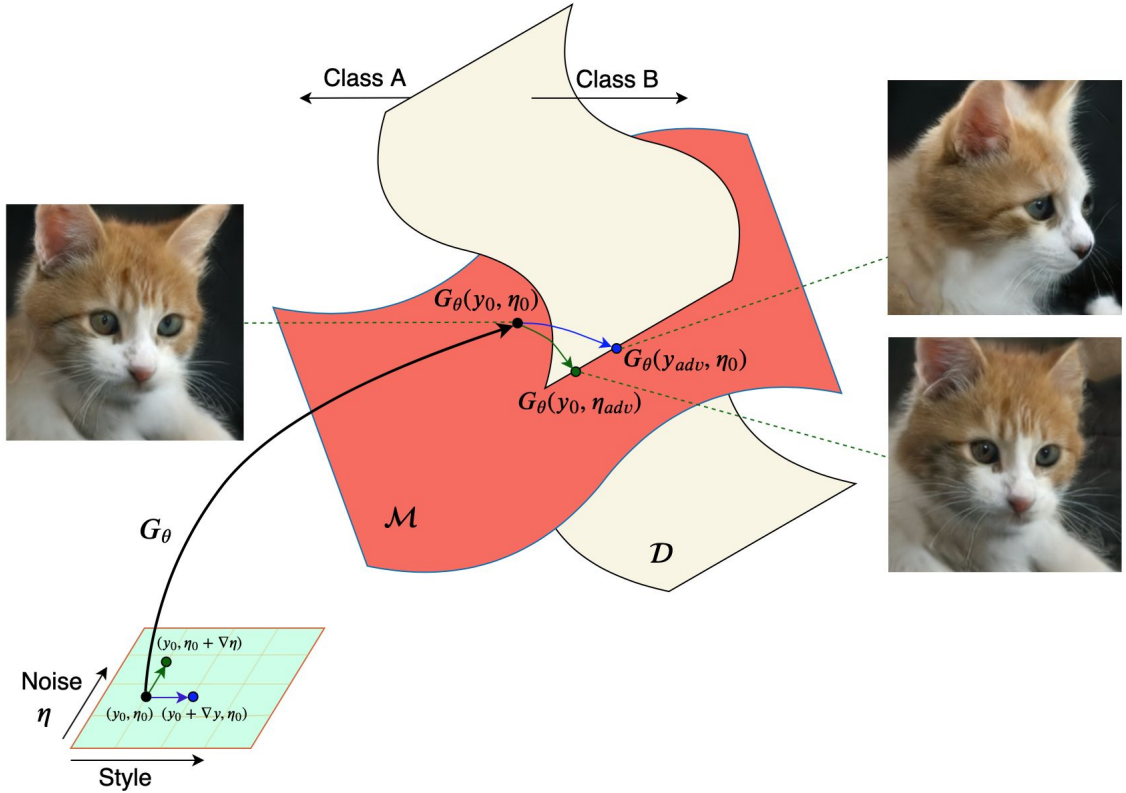


$$G_\theta(y, \eta) \simeq \mathcal{M}$$



Style-GAN

Unrestricted Adversarial Examples

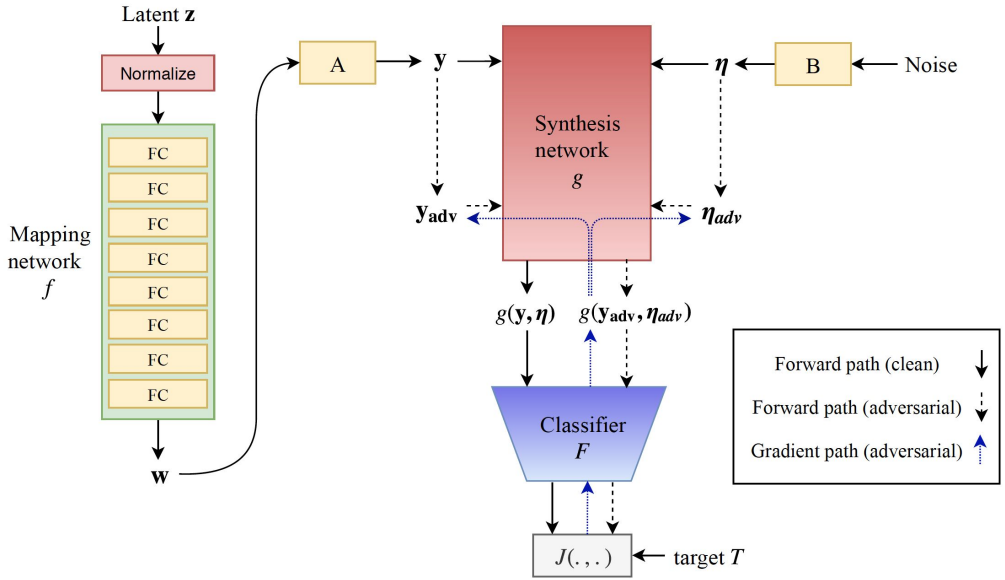


Unrestricted Adversarial Examples

Iteratively updating the variables

$$\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t+1})} = \mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})} - \epsilon \cdot \mathrm{sign}(\nabla_{\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), T))$$

$$\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t+1})} = \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})} - \delta \cdot \mathrm{sign}(\nabla_{\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), T))$$



Fine-grained Unrestricted Adversarial Examples

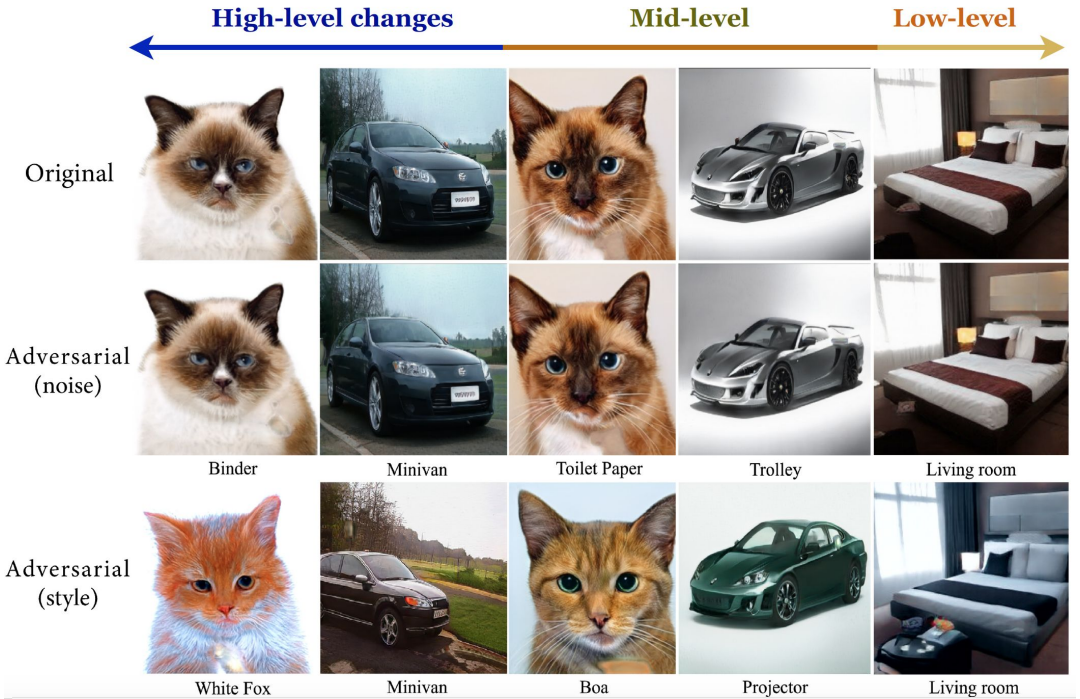
Only manipulating specific layers

Top layers: high-level changes

Bottom layers: low-level changes

Fine-grained Unrestricted Adversarial Examples

Results on LSUN: Non-targeted



Fine-grained Unrestricted Adversarial Examples

Results on LSUN: Targeted

Fine-grained Unrestricted Adversarial Examples

Results on CelebA-HQ Gender Classification

Evaluation on Certified Defenses

Certified defenses exist on norm-bounded attacks

- Vulnerable to our unrestricted attack

	Accuracy
Clean	63.1%
Adversarial (style)	21.7%
Adversarial (noise)	37.8%

Table 1: Accuracy of a certified classifier equipped with randomized smoothing on adversarial images.

Adversarial Training

Including adversarial images in training the classifier

- Effective as a defense
- Does not decrease performance on clean images

	Adv. Trained	Original
Clean	87.6%	87.7%
Adversarial (noise)	81.2%	0.0%
Adversarial (style)	76.9%	0.0%

Table 2: Accuracy of adversarially trained and original classifiers on clean and adversarial test images.

User Study

Real or Fake?

- Accuracy on un-adversarial generated images: 74.7%
- Accuracy on style-based adversarial images: 70.8%
- Accuracy on noise-based adversarial images: 74.3%

Correct category?

- Accuracy on style-based images: 98.7%
- Accuracy on noise-based images: 99.2%