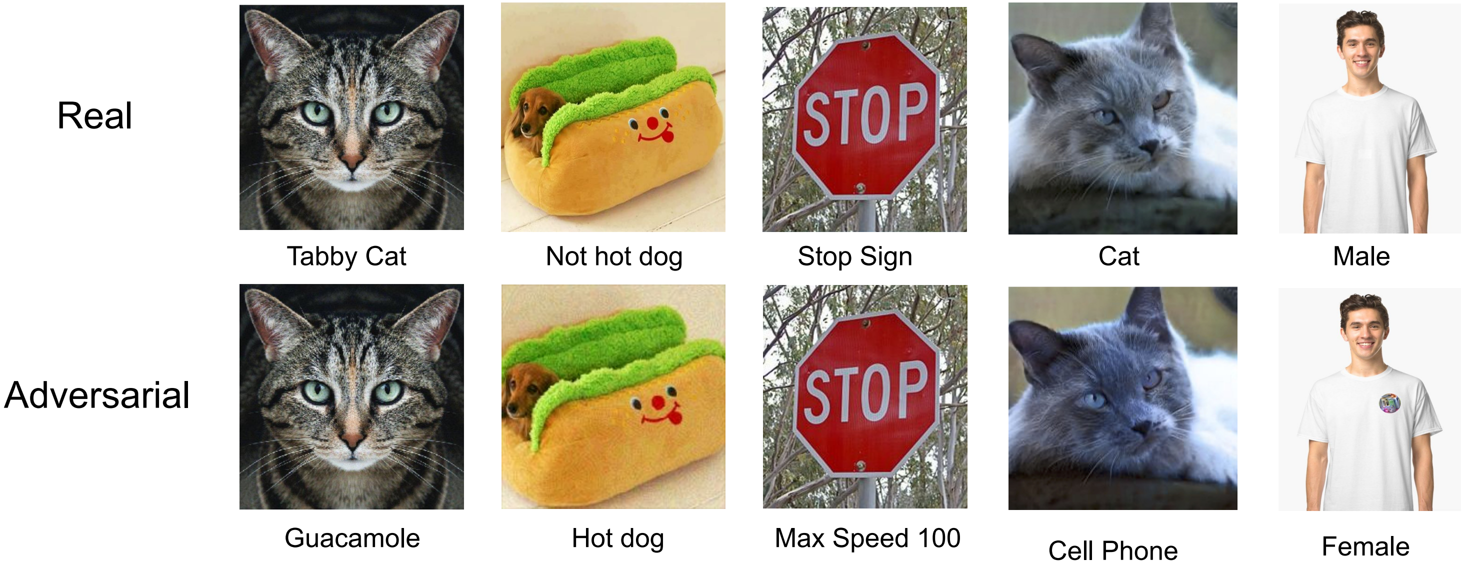
Robustness and Generalization via Generative Adversarial Training

Omid Poursaeed

Adversarial Image Manipulation

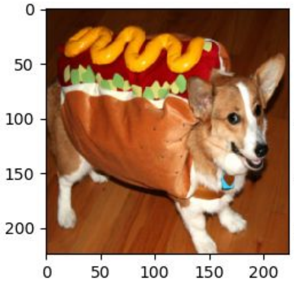


- Look like real images
- Misclassified by the model

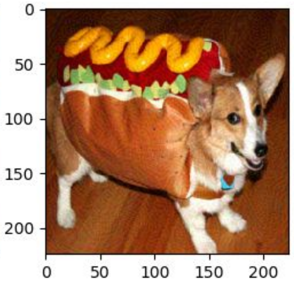
Similarity of Images

L_p similarity: $||x - \hat{x}||_p < \epsilon$

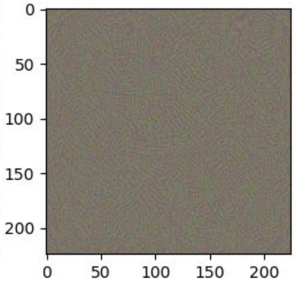
$p = 0, 2, \infty, \ldots$

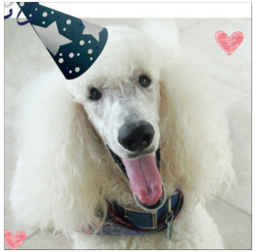


Not hot dog



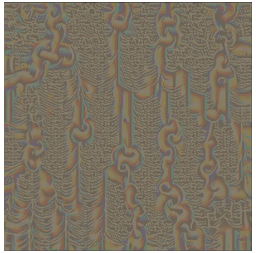
hot dog



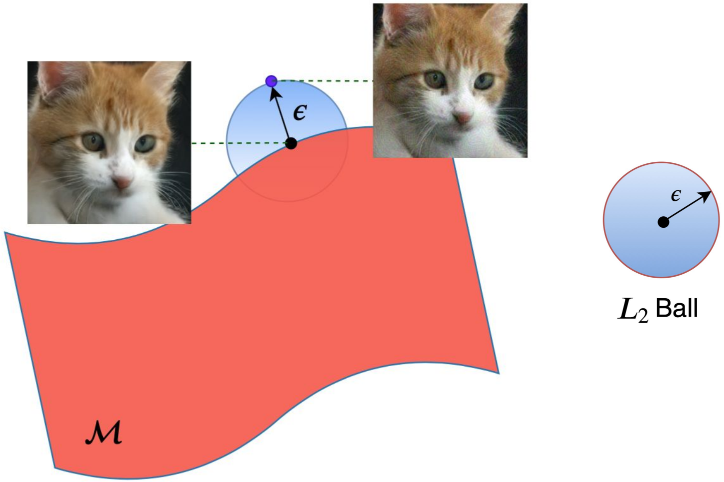


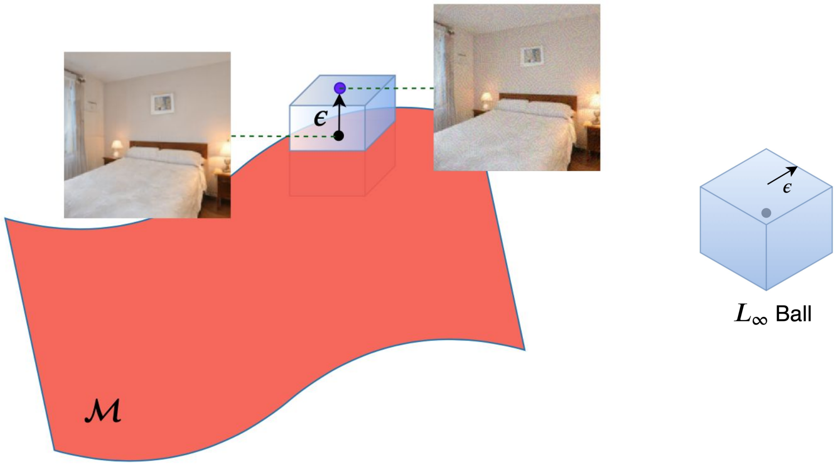
Poodle

Chain



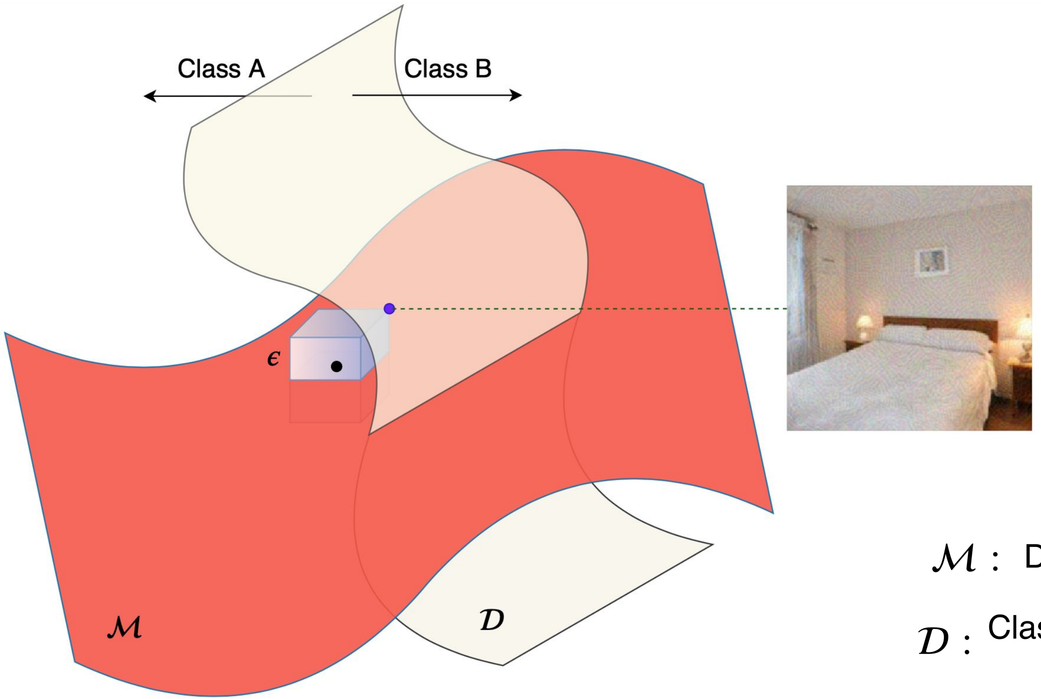
Manifold of Natural Images





$\mathcal{M}$: Data Manifold

Manifold of Natural Images



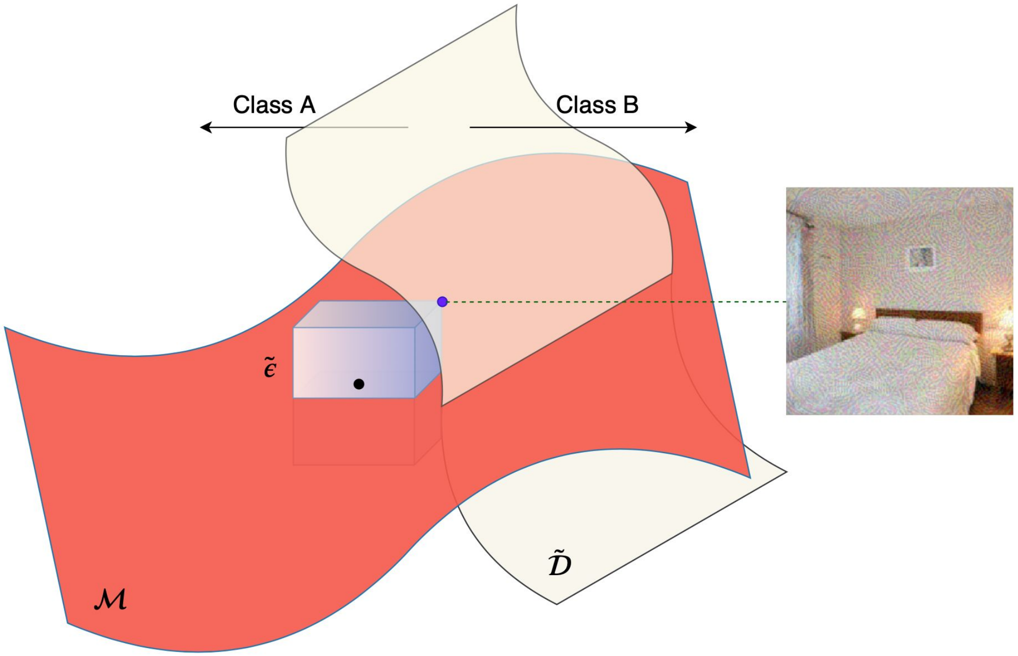
$\mathcal{M}$: Data Manifold

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

Manifold of Natural Images

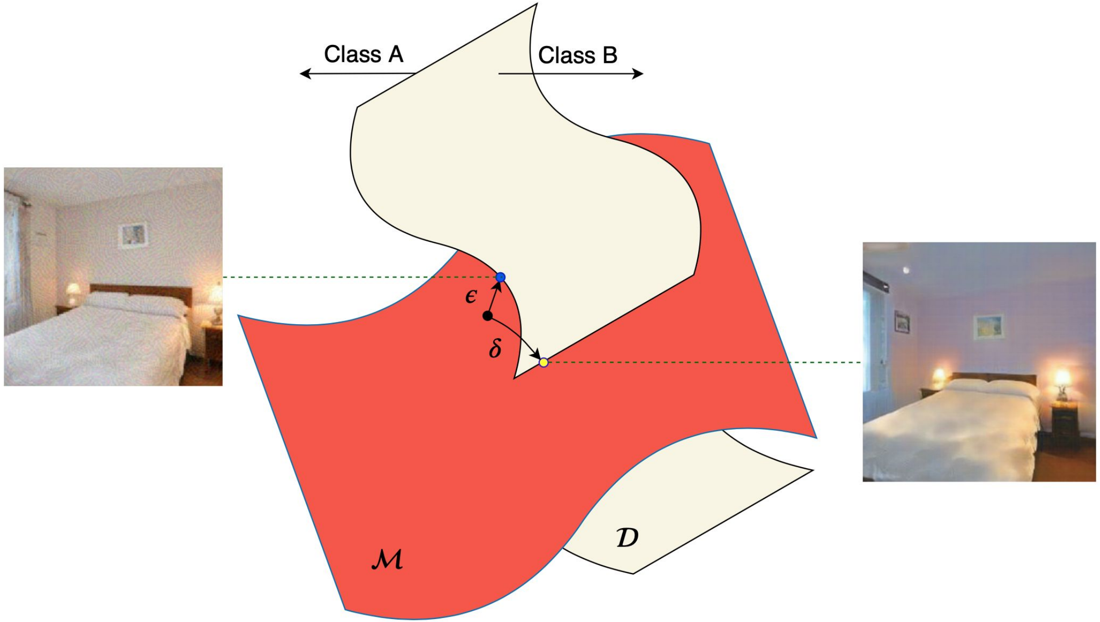
Classifiers equipped with defense

Larger perturbation norms



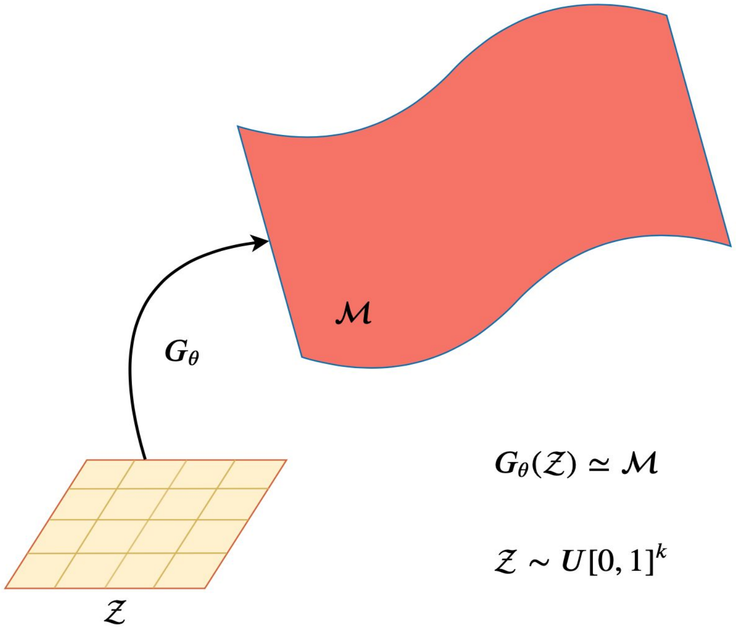
Unrestricted Adversarial Examples

Can we move on the manifold?



Unrestricted Adversarial Examples

Using a generative model to approximate the manifold



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

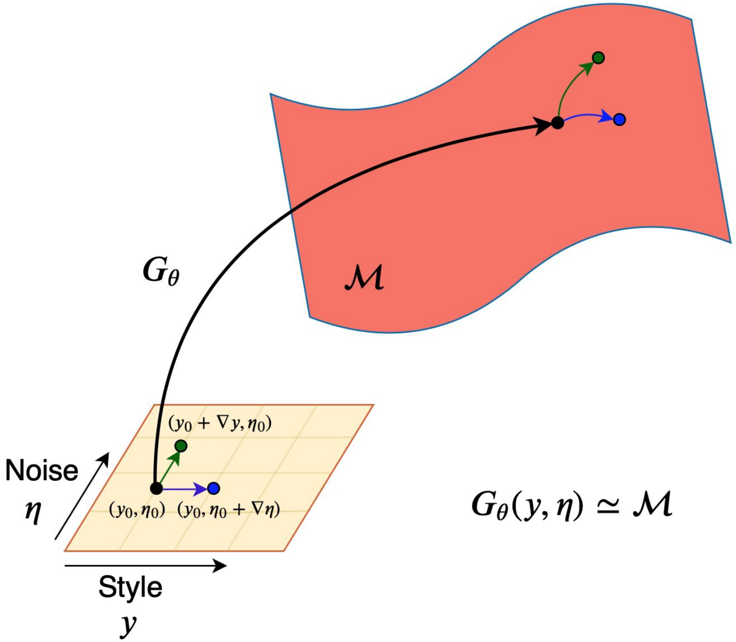
$\mathcal{Z} \sim U[0,1]^k$

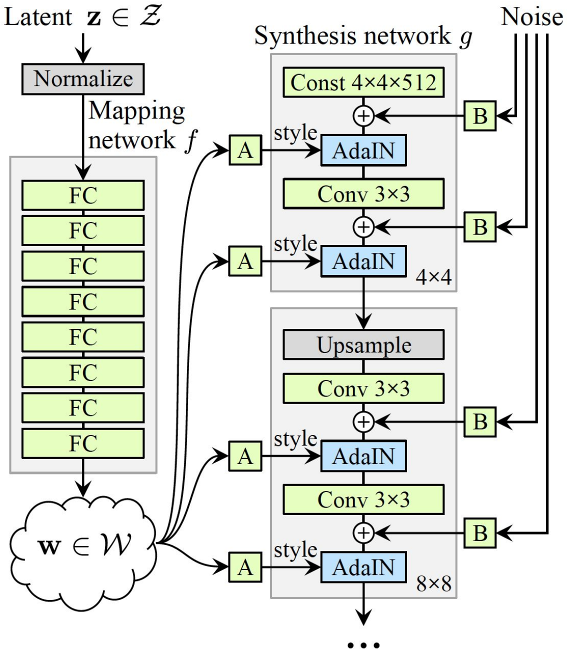
G_θ : Generator

$\mathcal{M}$: Data Manifold

Unrestricted Adversarial Examples

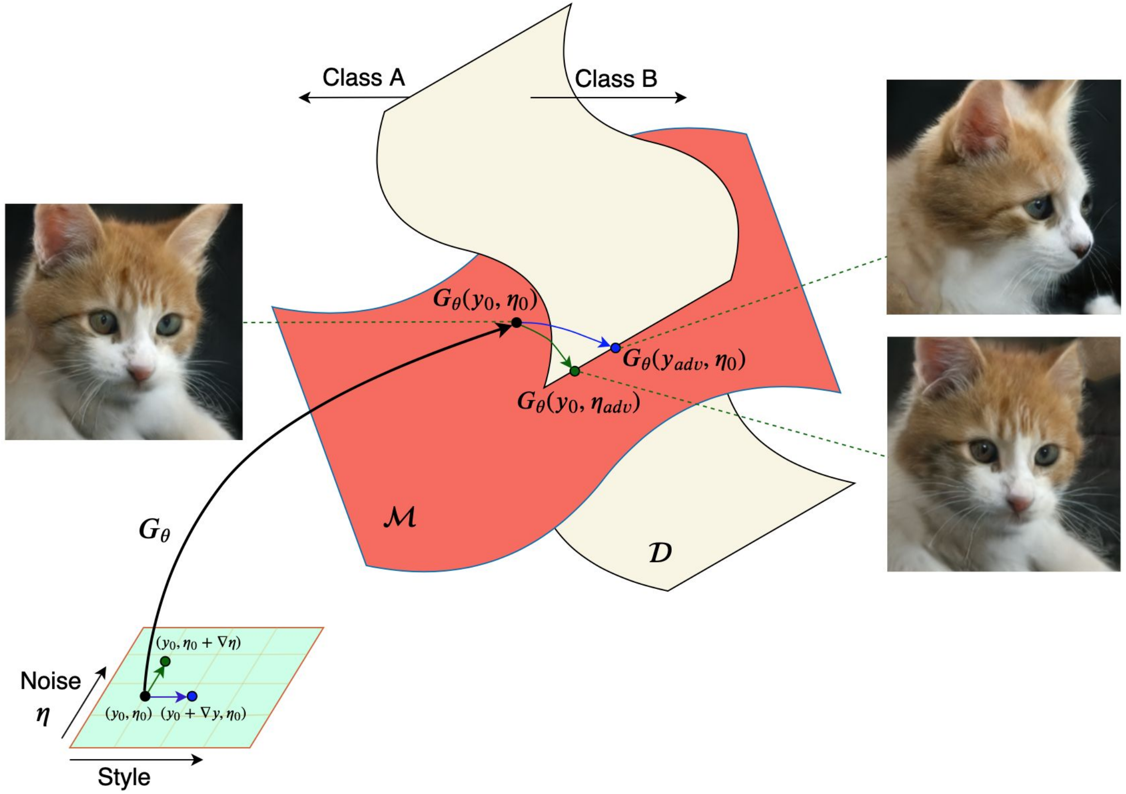
Disentangled Latent Space





Style-GAN

Unrestricted Adversarial Examples

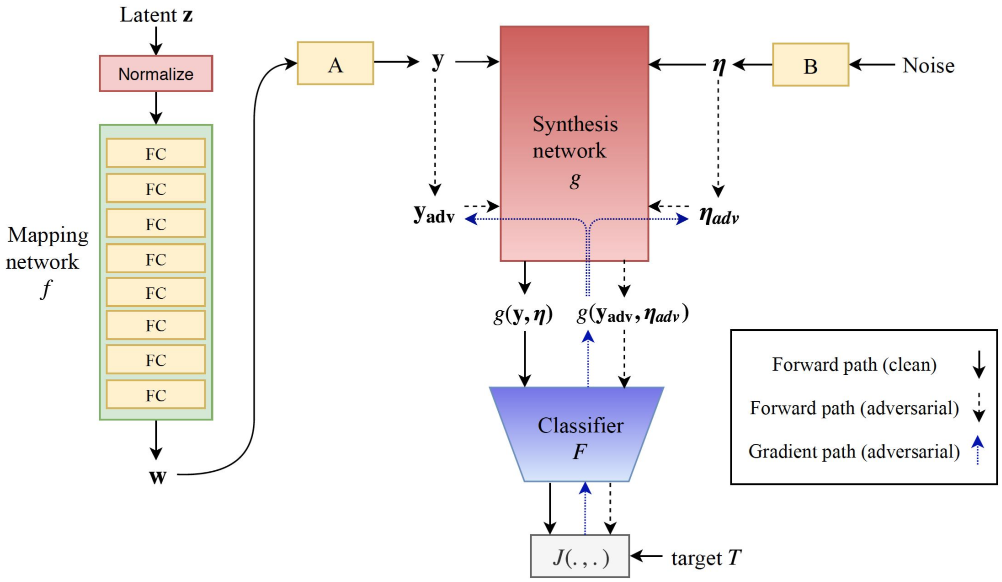


Unrestricted Adversarial Examples

Iteratively updating the variables

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

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

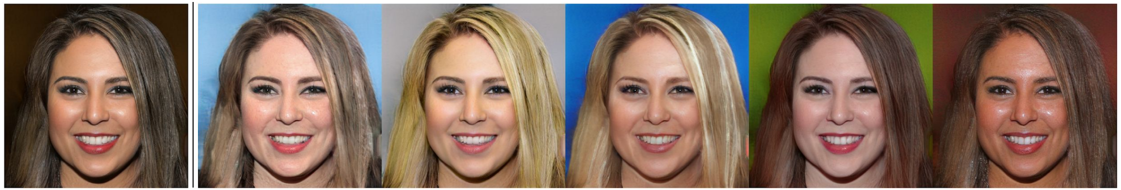


Fine-grained Unrestricted Adversarial Examples

Only manipulating specific layers

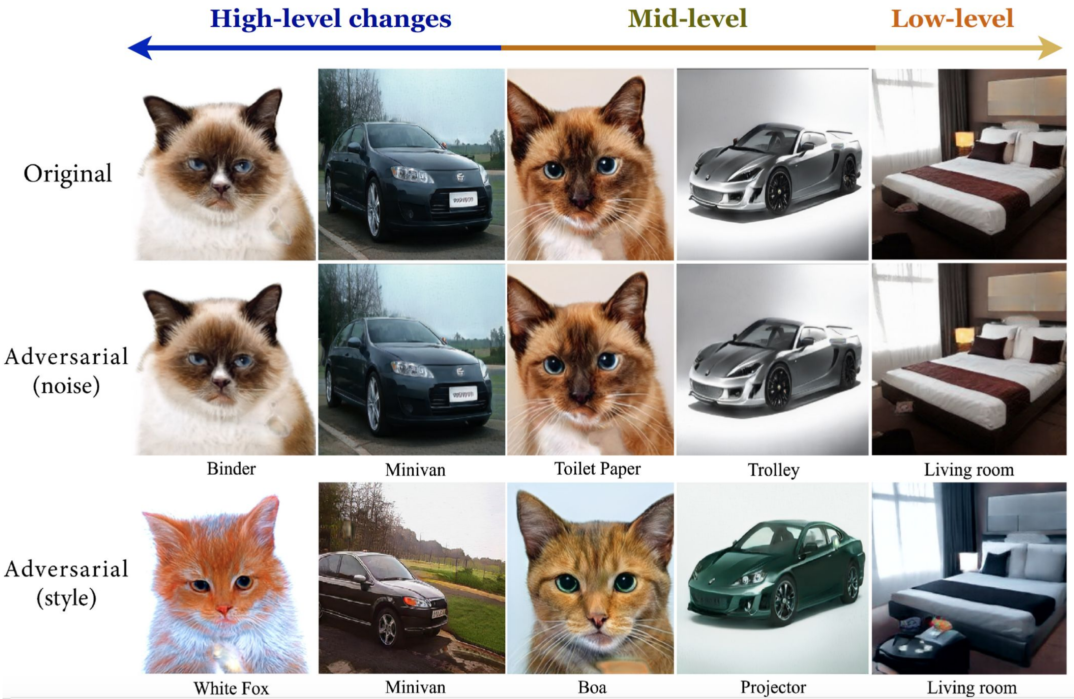
Top layers: high-level changes

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

Adversarial Training

Including adversarial images in training the classifier

- Effective as a defense
- Improves performance on clean images

	Classification (LSUN)		Classification (CelebA-HQ)		Segmentation		Detection	
	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
Adv. Trained	**89.5%**	78.4%	**96.2%**	83.6%	**69.1%**	60.2%	**40.2%**	33.7%
Original	88.9%	0.0%	95.7%	0.0%	67.9%	2.7%	39.0%	2.0%

Adversarial Training

Model	Attack						Mean
	Clean	GAT	PGD	Spatial	Recolor	Perceptual	
GAT (Ours)	**89.5%**	78.4%	39.4%	47.8%	52.3%	28.9%	**42.1%**
AT PGD [27]	81.2%	6.3%	56.7%	5.1%	37.9%	2.8%	13.0%
AT AdvProp [37]	89.4%	7.8%	57.6%	6.0%	38.5%	3.5%	22.7%
AT Spatial [36]	76.3%	5.4%	3.1%	66.0%	4.1%	2.2%	3.7%
AT Recolor [24]	88.6%	4.7%	7.3%	0.4%	60.7%	1.7%	3.5%
PAT [25]	72.4%	18.3%	40.1%	46.3%	42.5%	30.1%	36.5%

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%

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.