Counterfactual Explanations on Robust Perceptual Geodesics

Anonymous authors, ICLR 2026 submission

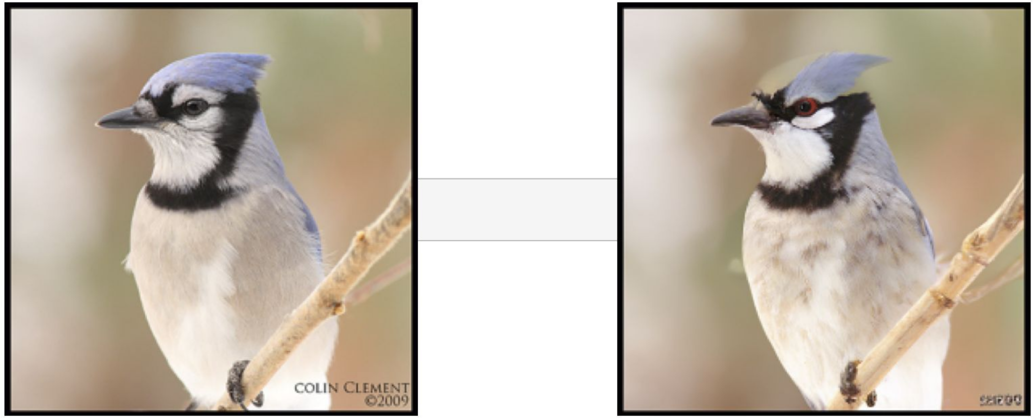
Visual Counterfactual Explanations

Visual Counterfactual Explanations (VCEs)

Identify **minimal semantic** change that alters the decision of a classifier to a certain **target** class

$f(\text{jay} \mid \mathbf{x}^*) = 0.98 \quad f(\text{bulbul} \mid \mathbf{x}_{\text{VCE}}) = 0.97$

VCE



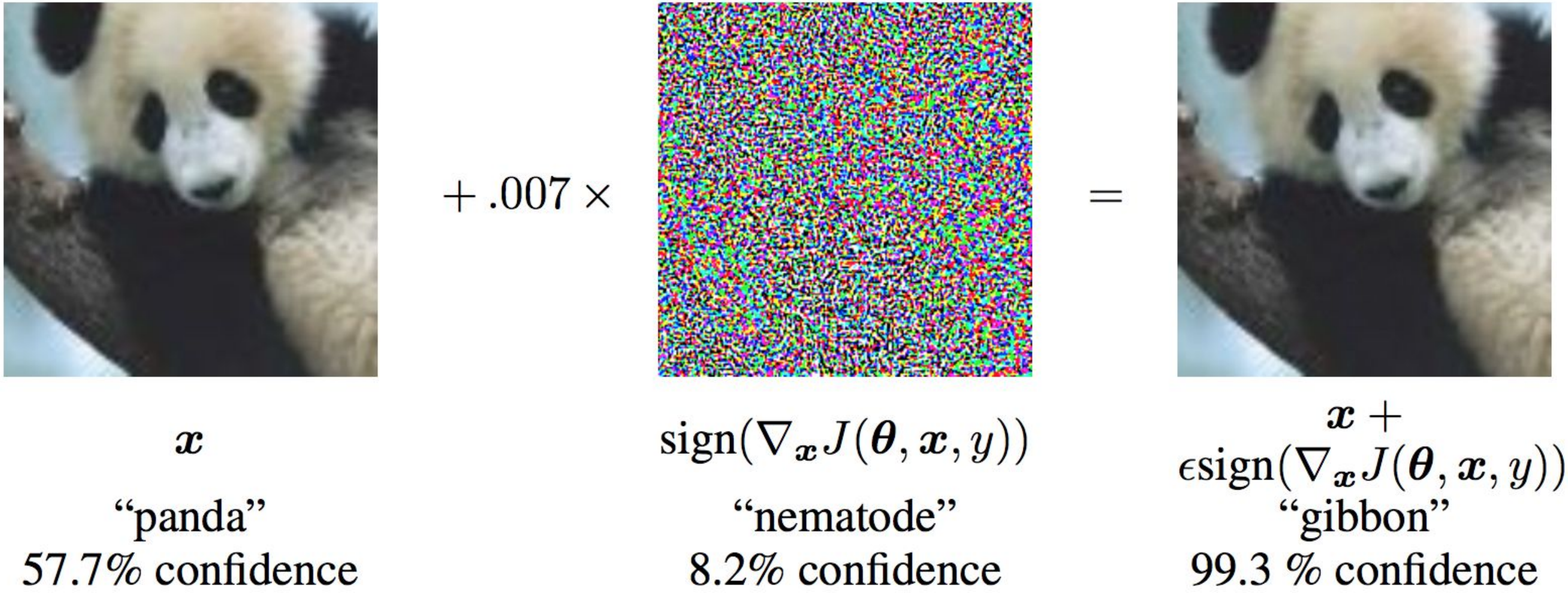
$\mathbf{x}^*$ $\mathbf{x}_{\text{VCE}}$

VCEs formally

$$\min_x \quad \underbrace{r(x^\star, x)}_{\text{Similarity Distance}} \quad + \lambda \;\; \underbrace{\ell(f(x), y')}_{\text{Classification Loss}}$$

Adversarial attacks/examples

Adversarial attacks/examples



Adversarial attacks/examples formally

$$\min_{x} \quad \underbrace{r(x^\star, x)}_{\text{Similarity Distance}} \quad + \lambda \; \underbrace{\ell(f(x), y')}_{\text{Classification Loss}}$$

Crucial difference

Choice of similarity distance

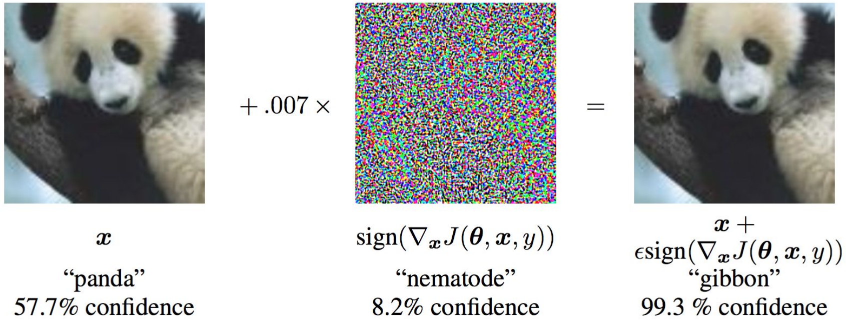
Adversarial attacks

- e.g., l_1, l_2
- unrelated to human perception

Choice of similarity distance

Adversarial attacks

- e.g., l_1, l_2
- unrelated to human perception



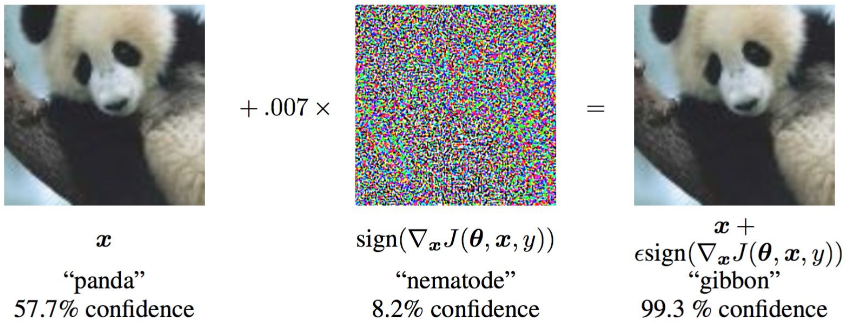
Choice of similarity distance

Adversarial attacks

- e.g., l_1, l_2
- unrelated to human perception

Counterfactual explanations

- generative-model-based
- closely related to human perception



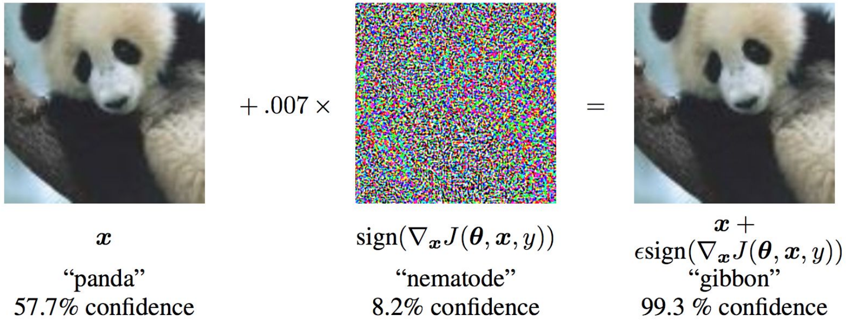
Choice of similarity distance

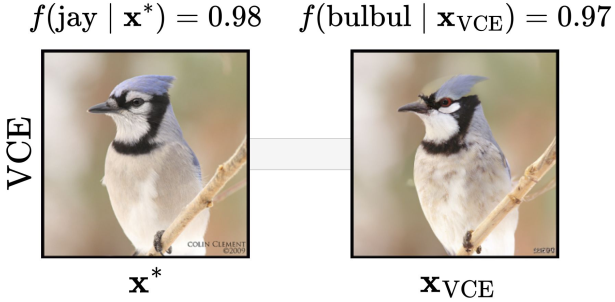
Adversarial attacks

- e.g., l_1, l_2
- unrelated to human perception

Counterfactual explanations

- generative-model-based
- closely related to human perception

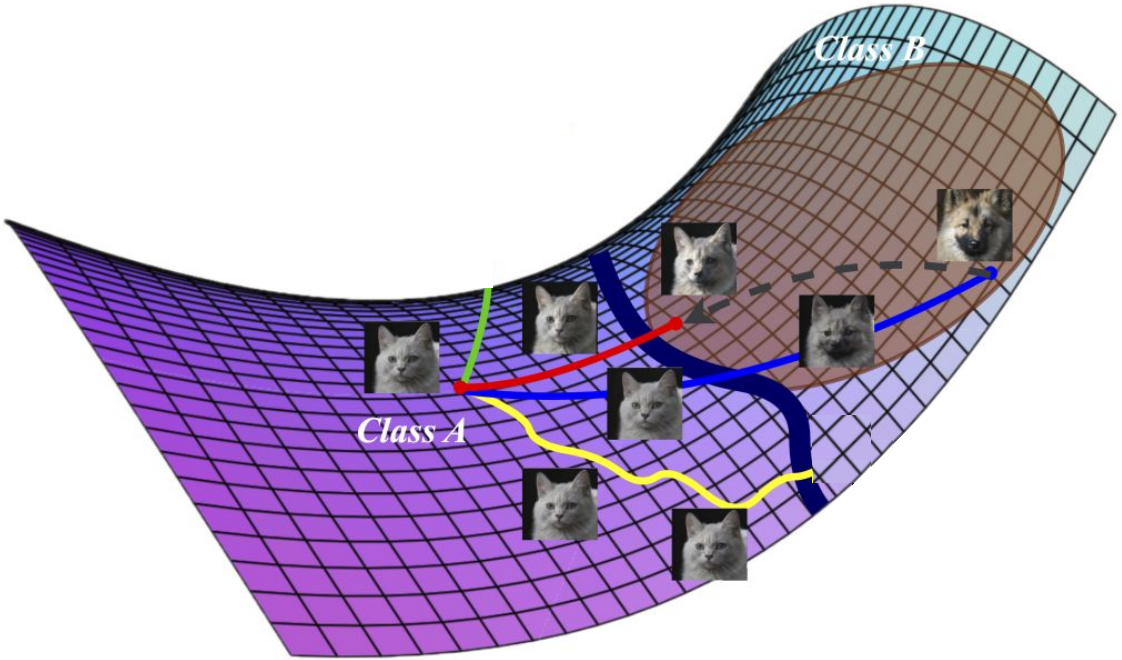




But is that enough?

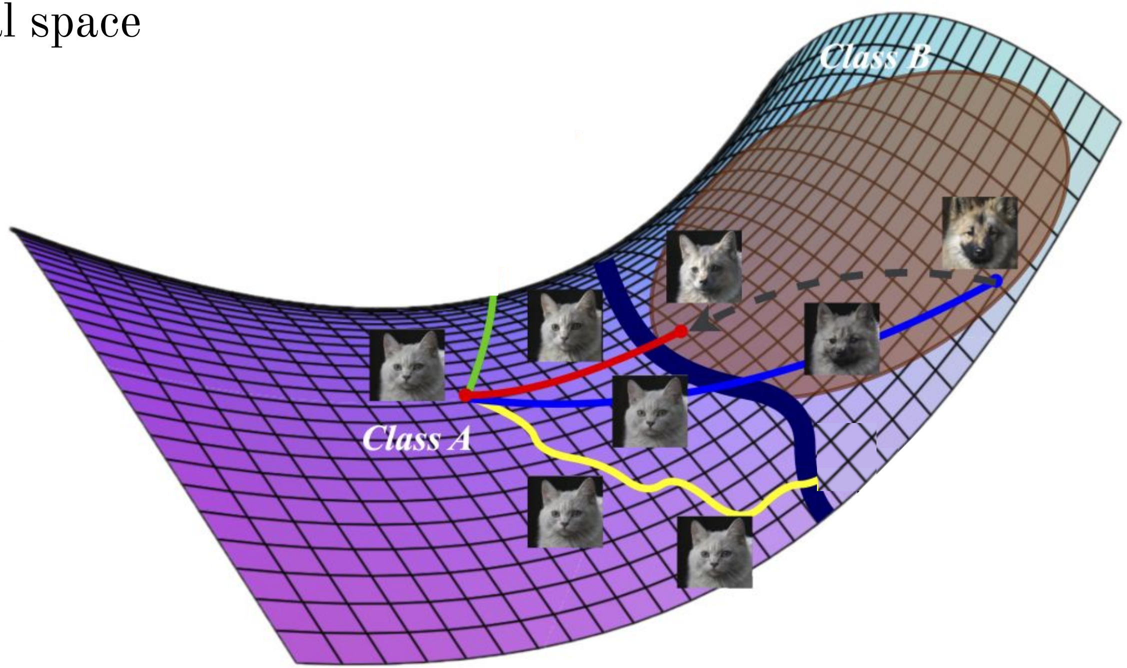
Main claim of this paper: not really

Manifold hypothesis



Manifold hypothesis

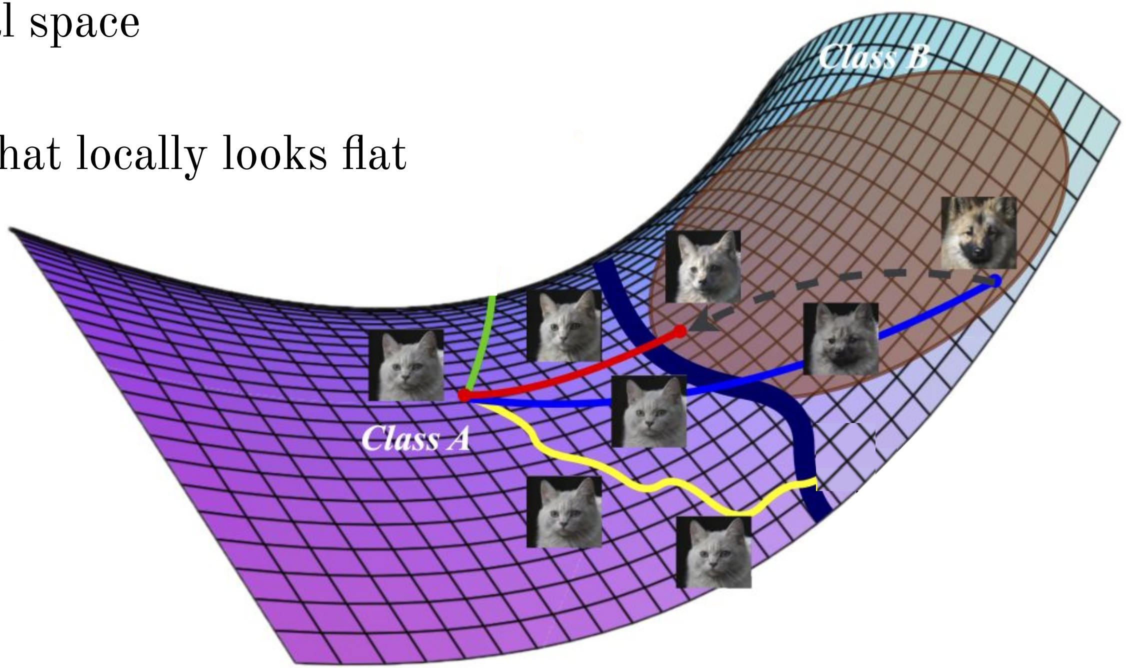
Real high-dimensional data lies on a low-dimensional manifold embedded in a high-dimensional space



Manifold hypothesis

Real high-dimensional data lies on a low-dimensional manifold embedded in a high-dimensional space

Manifold - a topological space that locally looks flat

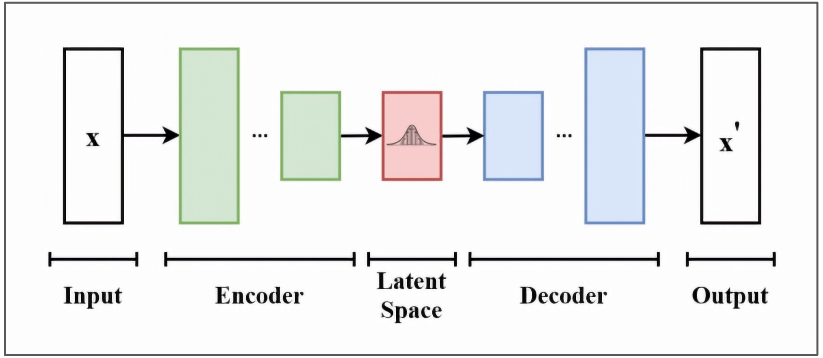


‘Seeing’ the manifold

through latent-space generative models

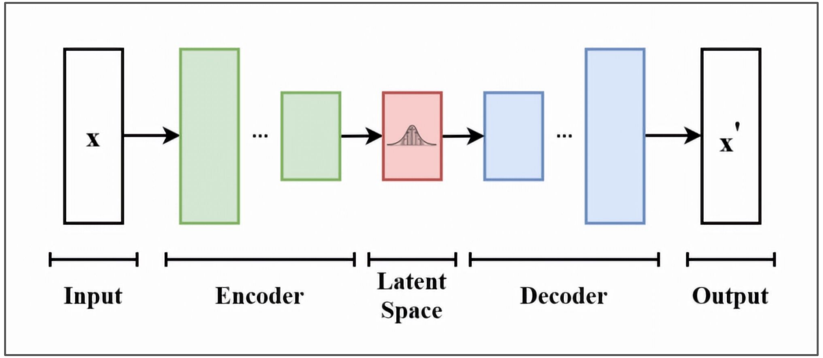
‘Seeing’ the manifold

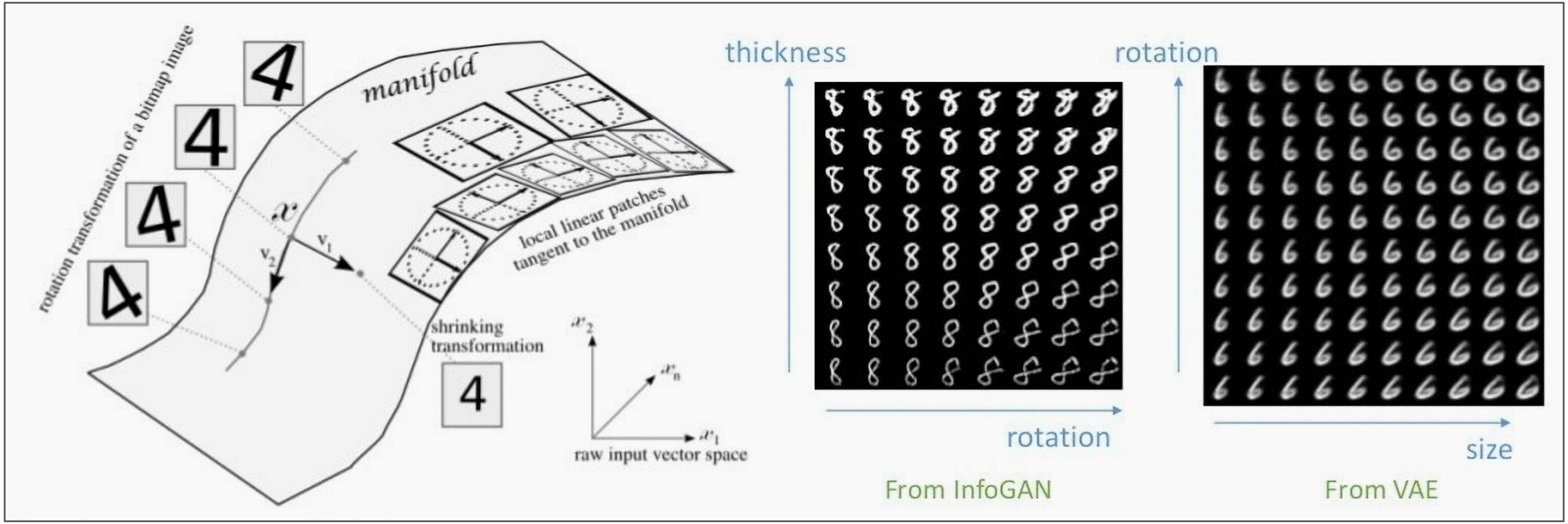
through latent-space generative models



‘Seeing’ the manifold

through latent-space generative models





So why not enough?

Informal definitions

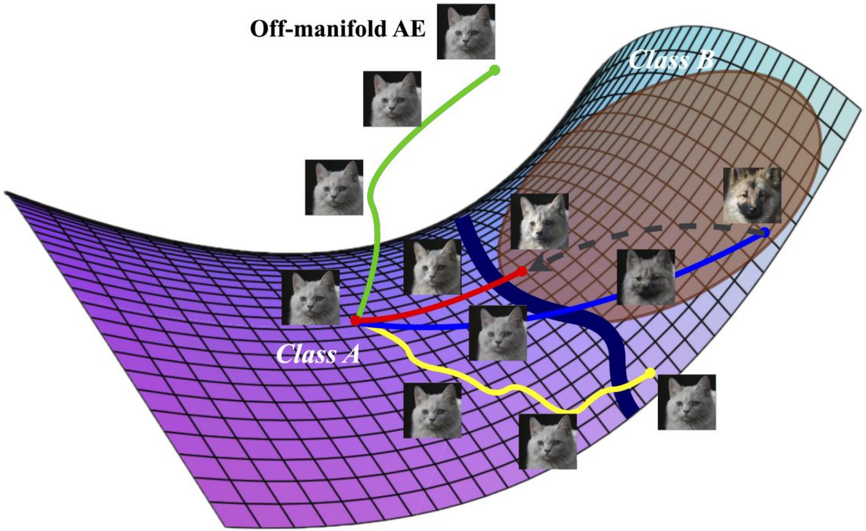
- on-manifold
  - **perceptually** indistinguishable from true data

Informal definitions

- on-manifold
  - **perceptually** indistinguishable from true data
- off-manifold
  - **perceptually** distinguishable from true data

Informal categorization

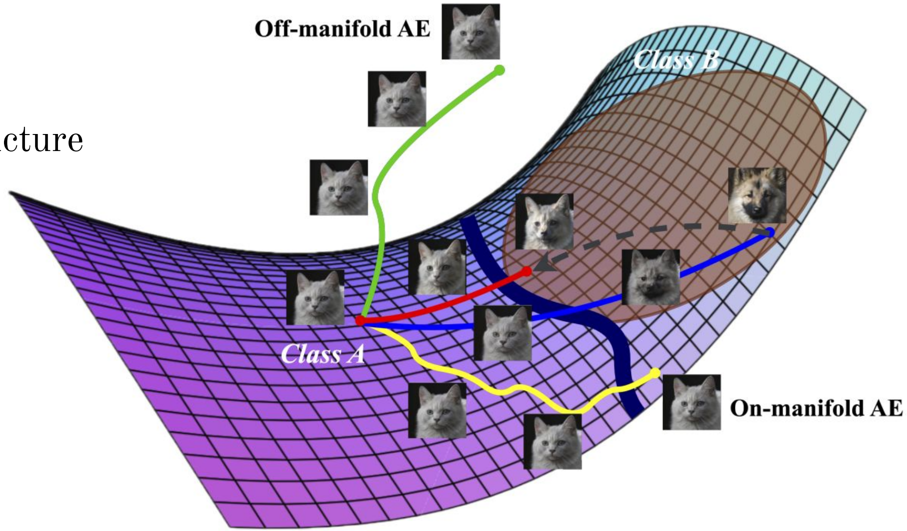
- off-manifold adversarial examples
  - easy to distinguish
  - a failure to generate semantically coherent content



Informal categorization

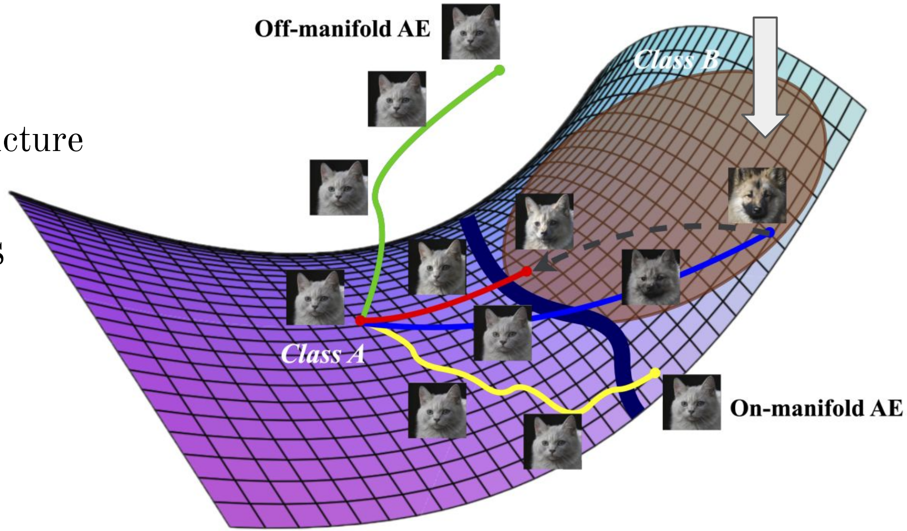
- off-manifold adversarial examples
  - easy to distinguish
  - a failure to generate semantically coherent content

- on-manifold adversarial examples
  - images that “look” like on-manifold samples
  - but contain (adversarial) noise or similar structure



Informal categorization

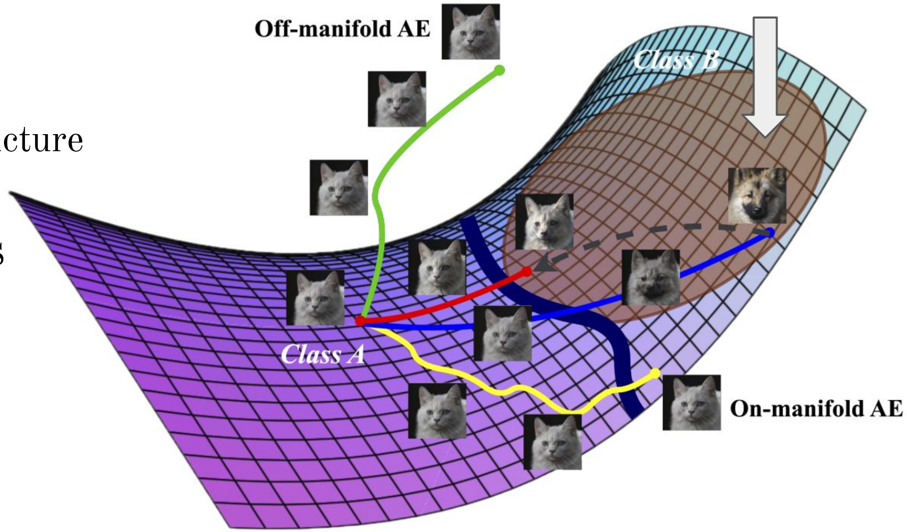
- off-manifold adversarial examples
  - easy to distinguish
  - a failure to generate semantically coherent content
- on-manifold adversarial examples
  - images that “look” like on-manifold samples
  - but contain (adversarial) noise or similar structure
- on-manifold counterfactual explanations
  - the sample stays on the manifold
  - the change is purely semantic



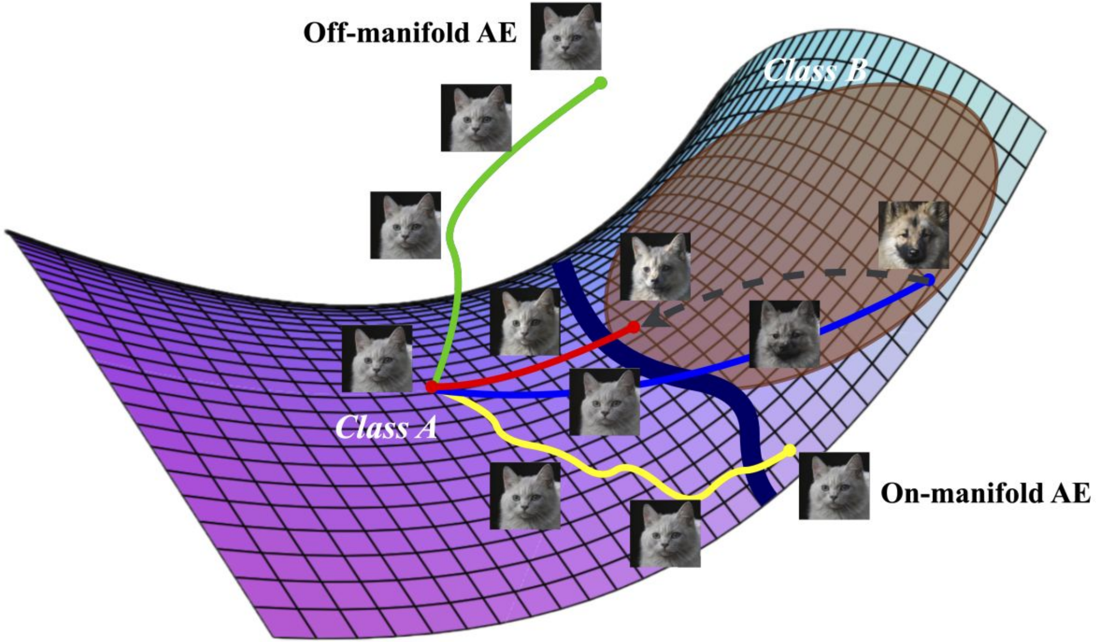
Informal categorization

- off-manifold adversarial examples
  - easy to distinguish
  - a failure to generate semantically coherent content
- on-manifold adversarial examples
  - images that “look” like on-manifold samples
  - but contain (adversarial) noise or similar structure
- on-manifold counterfactual explanations
  - the sample stays on the manifold
  - the change is purely semantic

How to enforce on-manifold VCEs?

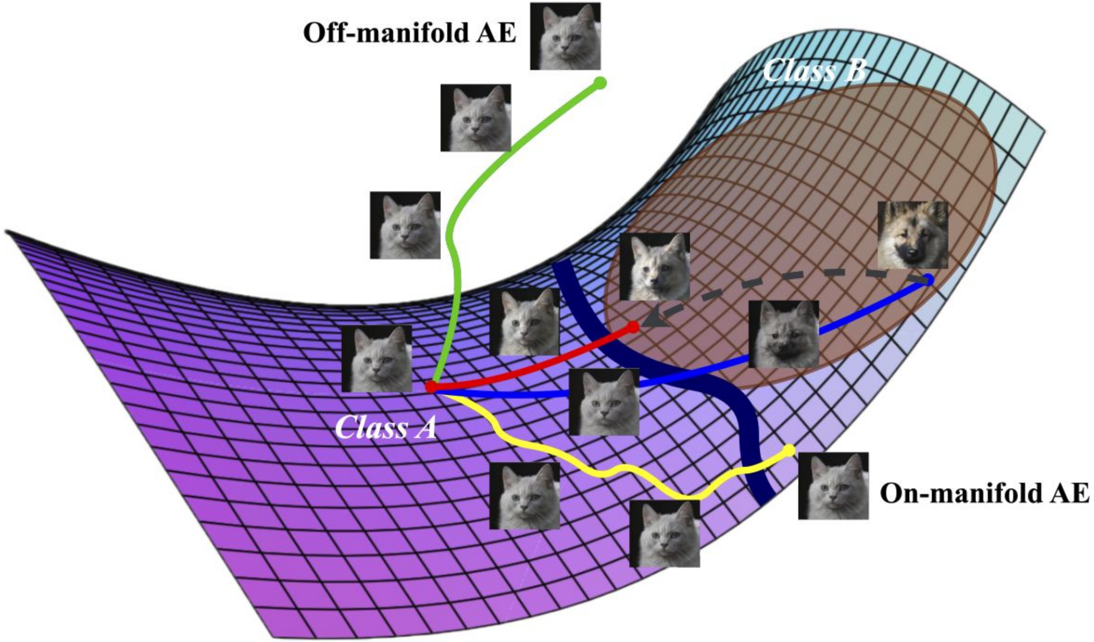


Problem restatement



Problem restatement

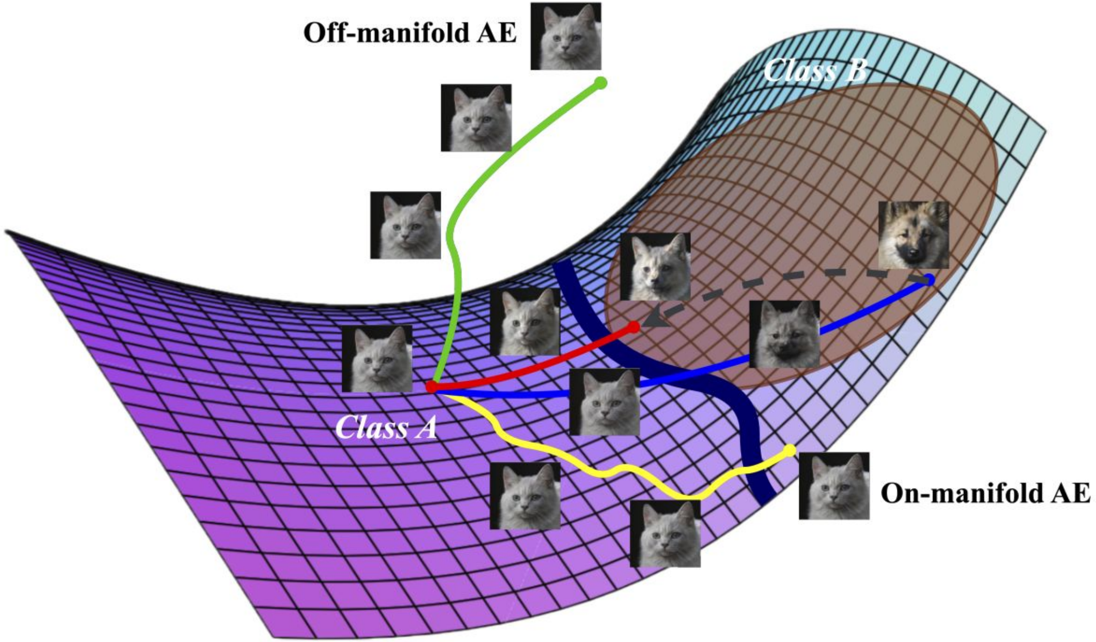
We are walking in a correct space



Problem restatement

We are walking in a correct space

We just don’t know how to walk



(Small) primer on differential geometry

(Small) primer on differential geometry

to find out how to walk (= to enforce a proper geometry)

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

$X \subset \mathbb{R}^D$ data space

$d \ll D$

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

$X \subset \mathbb{R}^D$ data space

$d \ll D$

$g : Z \to X$ generator

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

$X \subset \mathbb{R}^D$ data space

$d \ll D$

$g : Z \to X$ generator

$\mathcal{M} = g(Z) \subset X$ manifold

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

$X \subset \mathbb{R}^D$ data space

$d \ll D$

$g : Z \to X$ generator

$\mathcal{M} = g(Z) \subset X$ manifold

$J_g \triangleq \dfrac{\partial g}{\partial z} : Z \to \mathbb{R}^{D \times d}$ jacobian

How to borrow geometry?

$Z \subset \mathbb{R}^d$ latent space

$X \subset \mathbb{R}^D$ data space

$d \ll D$

$g : Z \to X$ generator

$\mathcal{M} = g(Z) \subset X$ manifold

$J_g \triangleq \frac{\partial g}{\partial z} : Z \to \mathbb{R}^{D \times d}$ jacobian

$\operatorname{rank}(J_g) = d$ full-rank

How to borrow geometry?

$\langle \cdot, \cdot \rangle_x$ smooth (wrt x) inner product

How to borrow geometry?

$\langle \cdot, \cdot \rangle_x$ smooth (wrt x) inner product

$G(x)$ induces a Riemannian metric

How to borrow geometry?

$\langle \cdot, \cdot \rangle_x$ smooth (wrt x) inner product

$G(x)$ induces a Riemannian metric

$(\mathcal{M}, G)$ which leads to a Riemannian manifold

How to borrow geometry?

$$\langle u, v \rangle_z := \langle J_g(z)u, J_g(z)v \rangle_{G_X(g(z))}$$

How to borrow geometry?

comparing u and v from tangent space of z

$$\langle u, v\rangle_z := \langle J_g(z)u, J_g(z)v\rangle_{G_X(g(z))}$$

How to borrow geometry?

comparing u and v from tangent space of z

$$\langle u, v \rangle_z := \langle J_g(z)u, J_g(z)v \rangle_{G_X(g(z))}$$

equivalent to comparing in X space by pushing with g

How to borrow geometry?

comparing u and v from tangent space of z

$$\langle u, v \rangle_z := \langle J_g(z)u, J_g(z)v \rangle_{G_X(g(z))}$$

equivalent to comparing in X space by pushing with g

$$\langle u, v \rangle_z = u^\top J_g(z)^\top G_X(g(z)) J_g(z) v$$

spelled out

How to borrow geometry?

comparing u and v from tangent space of z

$$\langle u, v\rangle_z := \langle J_g(z)u, J_g(z)v\rangle_{G_X(g(z))}$$

equivalent to comparing in X space by pushing with g

$$\langle u, v\rangle_z = u^\top J_g(z)^\top G_X(g(z)) J_g(z) v$$

spelled out

$$G_Z(z) = J_g(z)^\top G_X(g(z)) J_g(z)$$

explicit metric on Z pulled from X

What did we gain?

- Clear relationship between the geometries of Z and X

What did we gain?

- Clear relationship between the geometries of Z and X
- Even more, between Z and *some other space*

What did we gain?

- Clear relationship between the geometries of Z and X
- Even more, between Z and *some other space*
- X's geometry is suboptimal - we already know

Picking *the right* geometry

$h \Leftarrow$ robustly trained model

Picking *the right* geometry

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

h ⇦ robustly trained model

Picking *the right* geometry

metric on X

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

h

robustly trained model

Picking *the right* geometry

metric on X

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

robust

h robustly trained model

Picking *the right* geometry

sum over K layers

metric on X

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

robust

h robustly trained model

Picking *the right* geometry

sum over K layers

metric on X

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \qquad w_k = \frac{1}{N_k}$$

robust

weight for kth layer

h robustly trained model

Picking *the right* geometry

sum over K layers

metric on X

Jacobian of kth layer

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

robust

weight for kth layer

h robustly trained model

Picking *the right* geometry

sum over K layers

metric on X

Jacobian of kth layer

$$G_R(x) = \sum_{k=1}^{K} w_k J_{h_k}(x)^\top J_{h_k}(x), \qquad w_k = \frac{1}{N_k}$$

robust

weight for kth layer

kth layer number of elements

h robustly trained model

Why?

Robustness via curvature regularization, and vice versa

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, Pascal Frossard, CVPR 2018

Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness

Suraj Srinivas, Sebastian Bordt, Hima Lakkaraju, NeurIPS 2023

Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry, NeurIPS 2019

Robustness May Be at Odds with Accuracy

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry, ICLR 2019

Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness

Suraj Srinivas, Sebastian Bordt, Hima Lakkaraju,

NeurIPS 2023

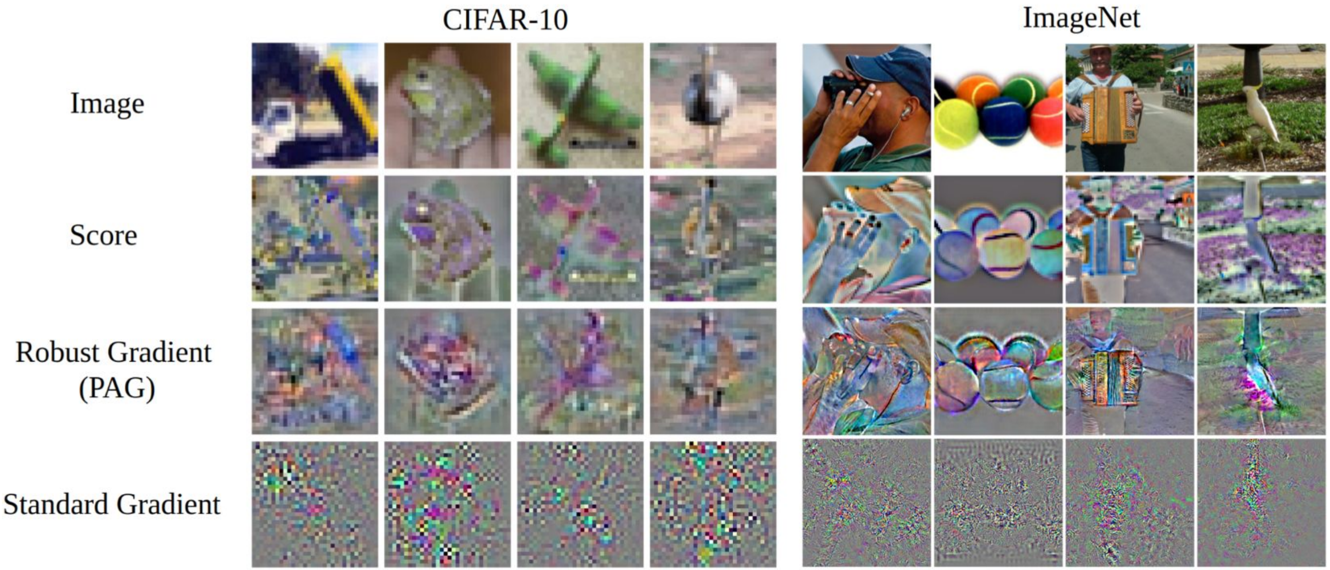
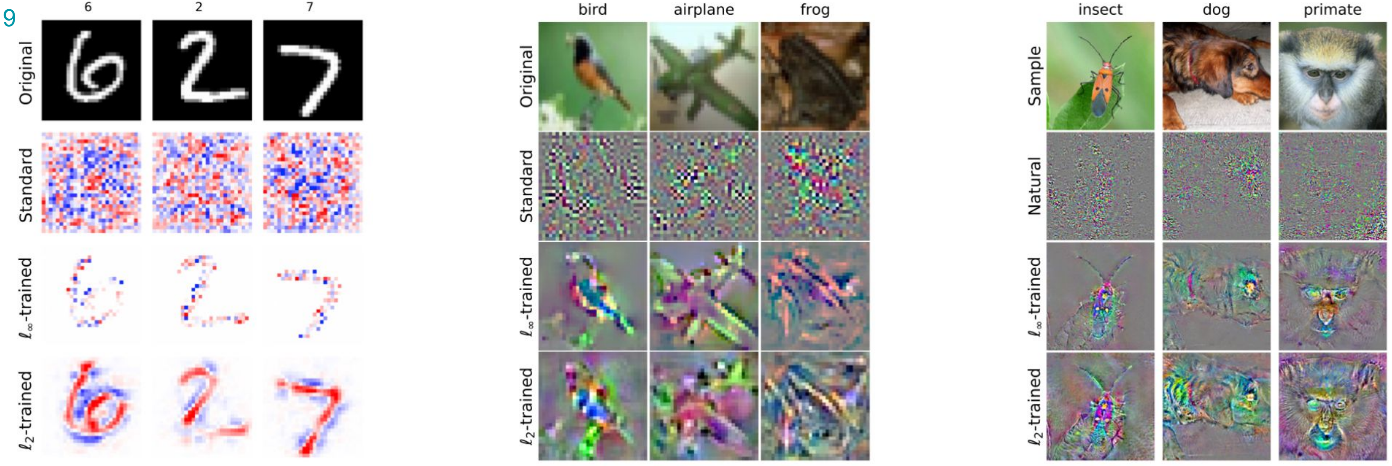


Figure 1: A demonstration of the perceptual alignment phenomenon. The input-gradients of robust classifiers ("robust gradient") are perceptually similar to the score of diffusion models [10], while being qualitatively distinct from input-gradients of standard models ("standard gradient"). Best viewed in digital format.

Robustness May Be at Odds with Accuracy

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry, ICLR 2019

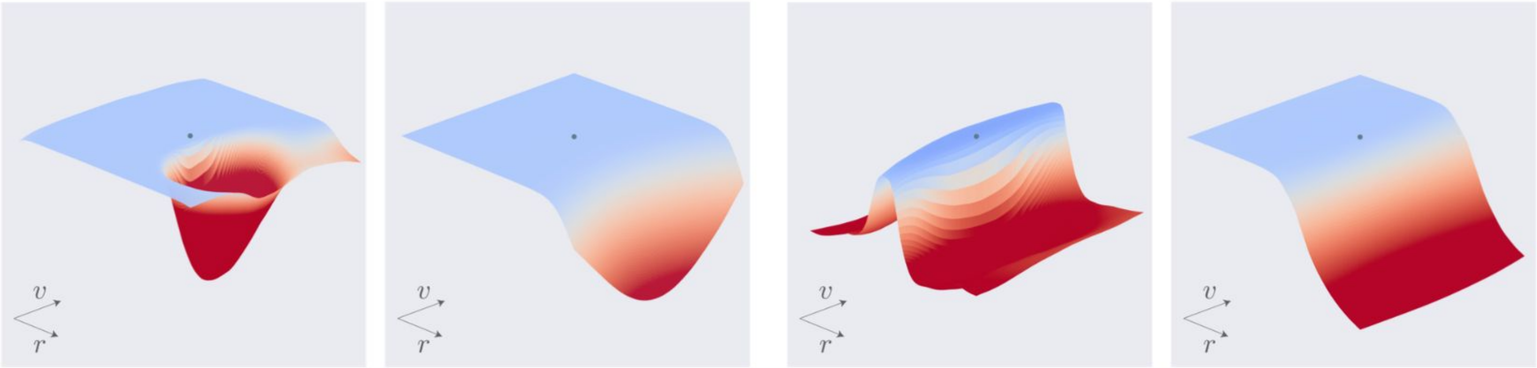


(a) MNIST (b) CIFAR-10 (c) Restricted ImageNet

Figure 2: Visualization of the loss gradient with respect to input pixels. Recall that these gradients highlight the input features which affect the loss most strongly, and thus the classifier's prediction. We observe that the gradients are significantly more human-aligned for adversarially trained networks – they align well with perceptually relevant features. In contrast, for standard networks they appear very noisy. (For MNIST, blue and red pixels denote positive and negative gradient regions respectively. For CIFAR-10 and ImageNet, we clip gradients to within ± 3 standard deviations of their mean and rescale them to lie in the $[0, 1]$ range.) Additional visualizations are presented in Figure 10 of Appendix G.

Robustness via curvature regularization, and vice versa

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, Pascal Frossard, CVPR 2018



(a) Original (CIFAR-10) (b) Fine-tuned (CIFAR-10) (c) Original (SVHN) (d) Fine-tuned (SVHN)

Figure 3: Illustration of the negative of the loss function; i.e., $-\ell(s)$ for points s belonging to a plane spanned by a normal direction r to the decision boundary, and random direction v. The original sample is illustrated with a blue dot. The light blue part of the surface corresponds to low loss (i.e., corresponding to the classification region of the sample), and the red part corresponds to the high loss (i.e., adversarial region).

Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry,

NeurIPS 2019

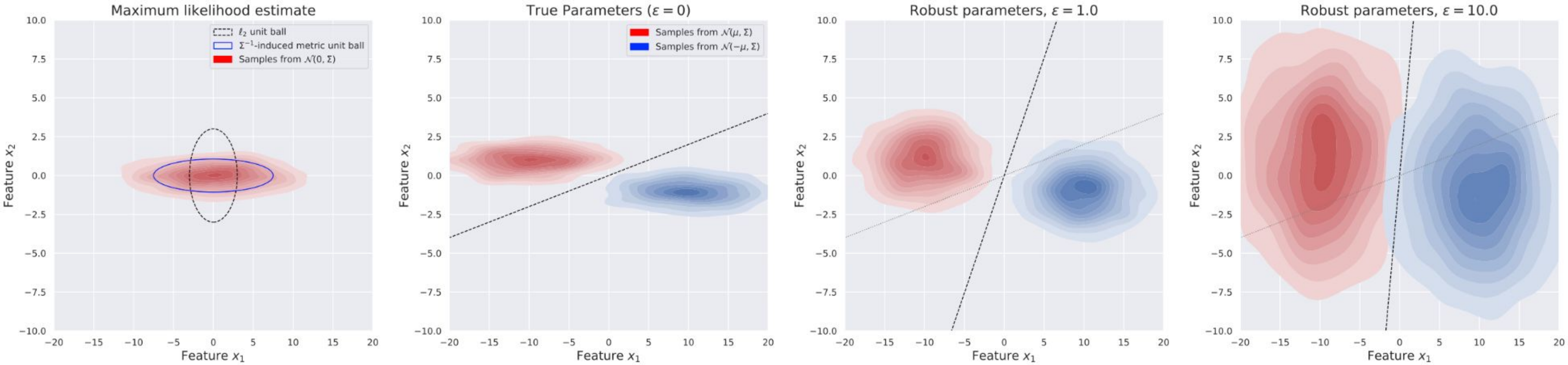


Figure 4: An empirical demonstration of the effect illustrated by Theorem 2—as the adversarial perturbation budget ε is increased, the learned mean μ remains constant, but the learned covariance "blends" with the identity matrix, effectively adding more and more uncertainty onto the non-robust feature.

Induced geometry

$$G_Z(z) = J_g(z)^\top G_R(g(z)) J_g(z)$$

Induced geometry

$$G_Z(z) = J_g(z)^\top G_R(g(z)) J_g(z)$$

robust metric

Perceptual Counterfactual Geodesics (PCG)

PCG

1. find a *geodesic* between the true x and some target class example

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

2. geodesic found by minimizing the robust perceptual energy

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

2. geodesic found by minimizing the robust perceptual energy

$$E(g(\gamma)) = \frac{1}{2}\int_0^1 \gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)dt$$

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

2. geodesic found by minimizing the robust perceptual energy

$$E(g(\gamma)) = \frac{1}{2}\int_0^1 \gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)dt \quad E_{\text{robust}}(z) = \frac{1}{2}\sum_{i=0}^{T-1}\sum_{k=1}^{K} w_k \delta t \left\| h_k\left(g(z_{i+1})\right) - h_k\left(g(z_i)\right)\right\|_2^2$$

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

2. geodesic found by minimizing the robust perceptual energy

$$E(g(\gamma)) = \frac{1}{2}\int_0^1 \gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)dt \quad E_{\text{robust}}(z) = \frac{1}{2}\sum_{i=0}^{T-1}\sum_{k=1}^{K} w_k \delta t \left\| h_k\left(g(z_{i+1})\right) - h_k\left(g(z_i)\right)\right\|_2^2$$

3. then, release the endpoint and also optimize for prediction

PCG

1. find a *geodesic* between the true x and some target class example

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt$$

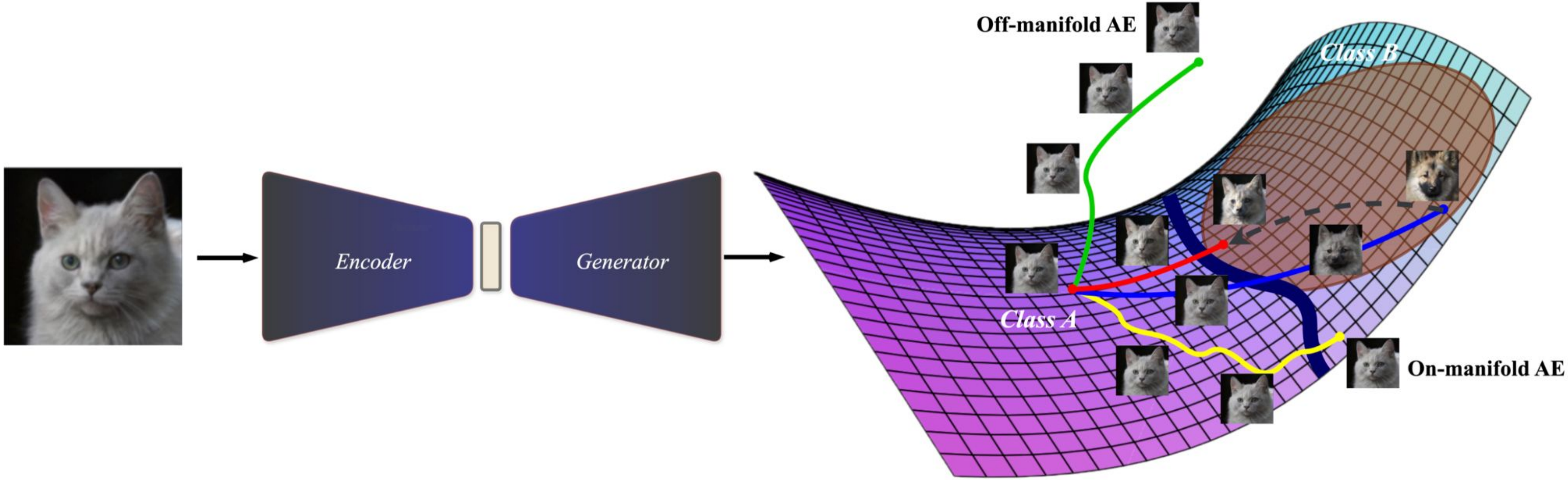
2. geodesic found by minimizing the robust perceptual energy

$$E(g(\gamma)) = \frac{1}{2}\int_0^1 \gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)dt \quad E_{\text{robust}}(z) = \frac{1}{2}\sum_{i=0}^{T-1}\sum_{k=1}^{K} w_k \delta t \left\| h_k\left(g(z_{i+1})\right) - h_k\left(g(z_i)\right)\right\|_2^2$$

3. then, release the endpoint and also optimize for prediction

$$\mathcal{L}(z) = E_{\text{robust}}(z) + \lambda \cdot \ell\left(f(g(z_T)), y'\right)$$

PCG



Encoder
Generator
Off-manifold AE
Class B
Class A
On-manifold AE

Prior approaches

PCG

1. Consider a semantically meaningful space

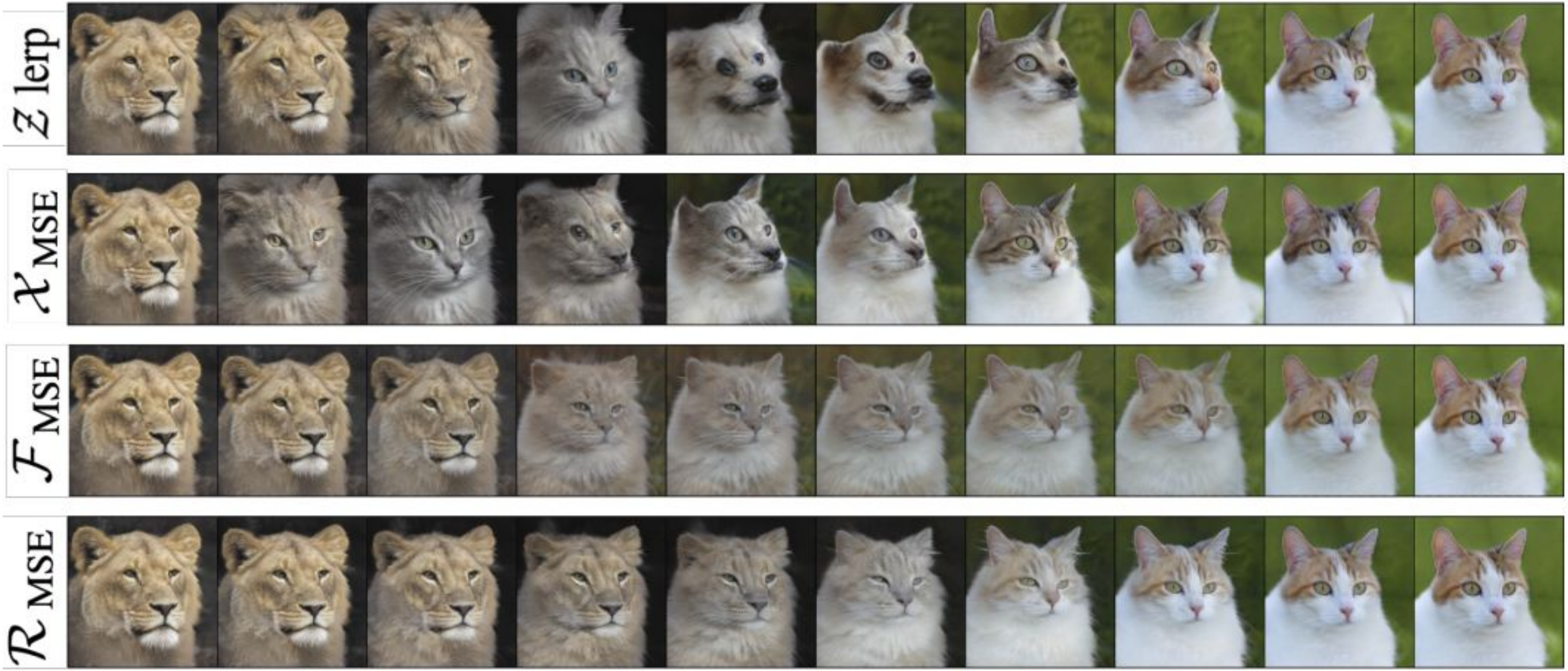
2. But **move with perception in mind**

Question: what replicates human perception well?

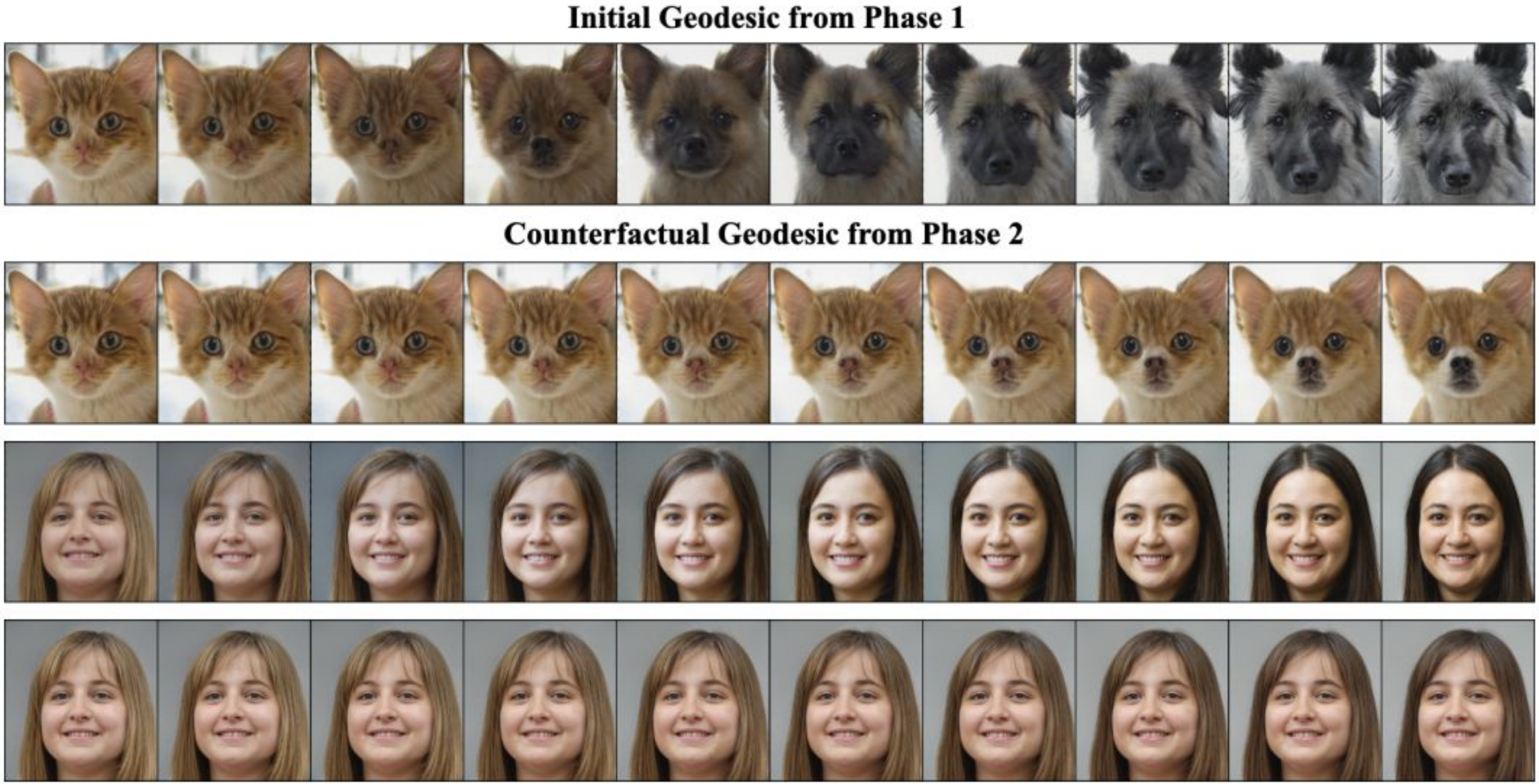
   1. quite good: LPIPS, since it aggregates convolution-based, often human-interpretable features
   2. even better: adversarial models, as they additionally incorporate robustness to perturbations, which improves representation geometry

Experiments

Effect of latent geometry on interpolation



Two-phase visualization

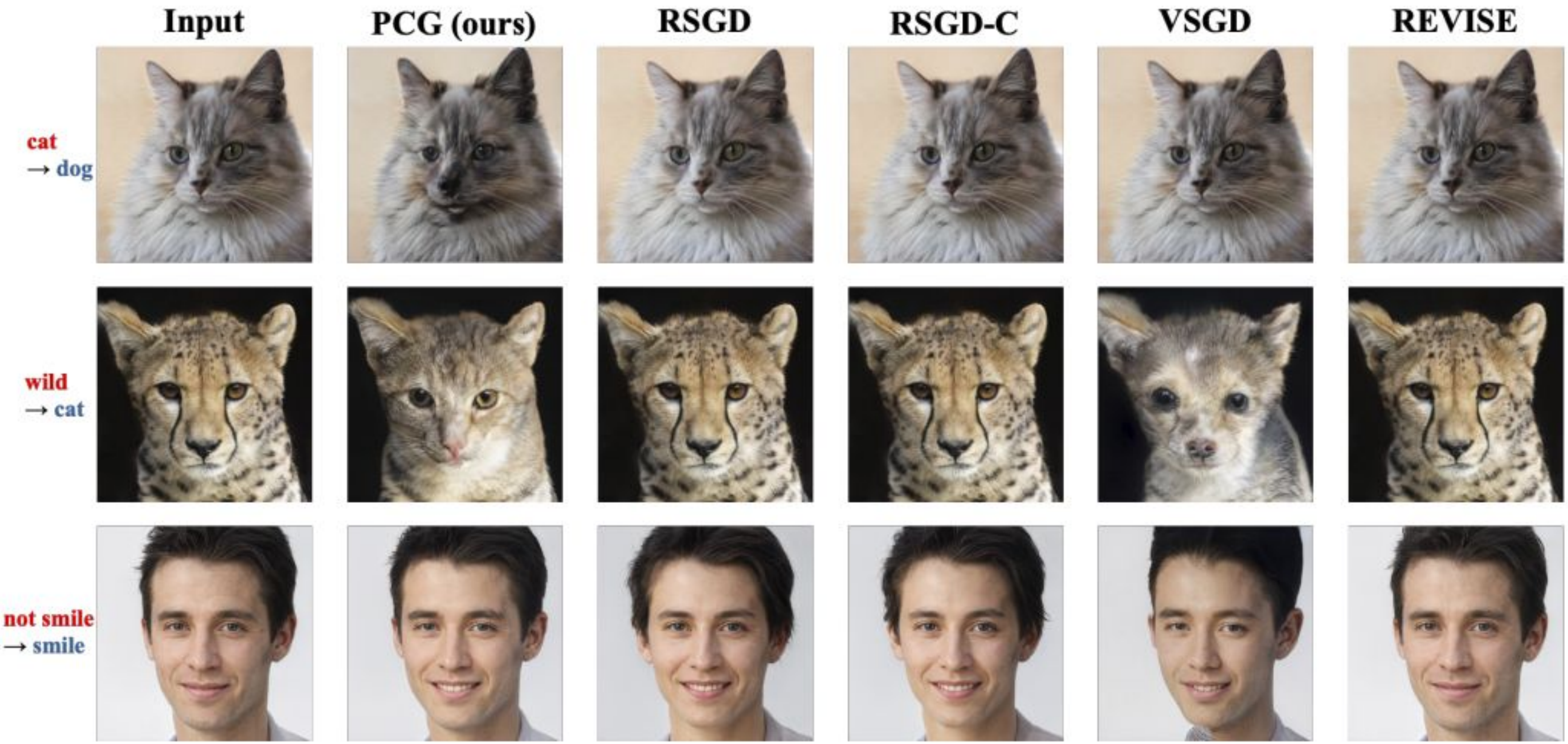


Quantitative comparison

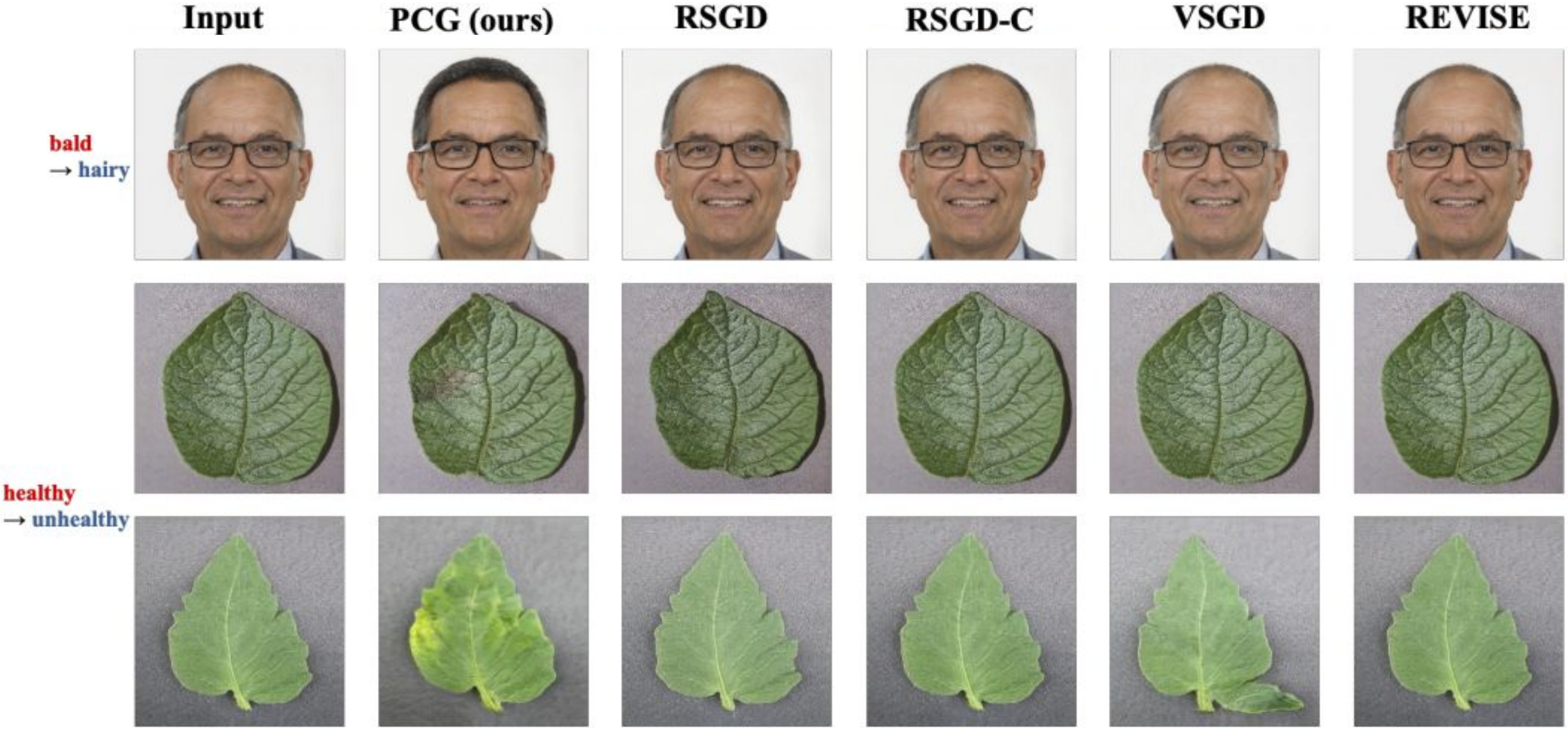
1. fair comparison is difficult
2. baselines adapted from tabular data

Method	AFHQ				FFHQ				PlantVillage			
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$
REVISE	1.20±0.12	**0.73**±0.18	1.08±0.10	2.70±0.05	0.82±0.08	**0.32**±0.13	0.82±0.08	2.78±0.06	0.50±0.13	**0.38**±0.15	0.96±0.06	2.87±0.07
VSGD	1.31±0.11	1.49±0.15	1.60±0.09	2.90±0.08	0.79±0.11	0.96±0.10	1.50±0.12	2.86±0.07	0.83±0.13	0.94±0.17	1.18±0.07	3.01±0.09
RSGD	0.85±0.08	1.32±0.09	0.70±0.07	1.85±0.05	0.61±0.05	0.84±0.07	0.61±0.04	2.41±0.05	0.78±0.08	0.82±0.11	0.54±0.05	2.28±0.04
RSGD-C	0.93±0.10	1.45±0.17	0.65±0.08	1.75±0.06	0.68±0.06	0.93±0.09	0.48±0.04	2.11±0.04	0.80±0.10	0.86±0.13	0.45±0.05	2.03±0.06
PCG (ours)	**0.79**±0.07	1.14±0.10	**0.53**±0.06	**0.31**±0.02	**0.42**±0.03	0.72±0.09	**0.39**±0.05	**0.22**±0.06	**0.36**±0.03	0.56±0.05	**0.34**±0.04	**0.20**±0.05

Qualitative comparison



Qualitative comparison



What if I reviewed it?

Strengths

1. Currently, one of the most important problems in VCEs

Strengths

1. Currently, one of the most important problems in VCEs

2. Actually a novel, well thought-out approach

Strengths

1. Currently, one of the most important problems in VCEs
2. Actually a novel, well thought-out approach
3. Adaptation of tabular baselines

Weaknesses

1. Robust models are never “infinitely” robust

Weaknesses

1. Robust models are never “infinitely” robust

2. Lack of theory for the metric’s definition

Weaknesses

1. Robust models are never “infinitely” robust

2. Lack of theory for the metric’s definition

3. Lack of explanandum-related metrics

Weaknesses

1. Robust models are never “infinitely” robust

2. Lack of theory for the metric’s definition

3. Lack of explanandum-related metrics

4. Some claims based purely on single examples

Weaknesses

1. Robust models are never “infinitely” robust
2. Lack of theory for the metric’s definition
3. Lack of explanandum-related metrics
4. Some claims based purely on single examples
5. (they don’t cite us)

Thank YOU