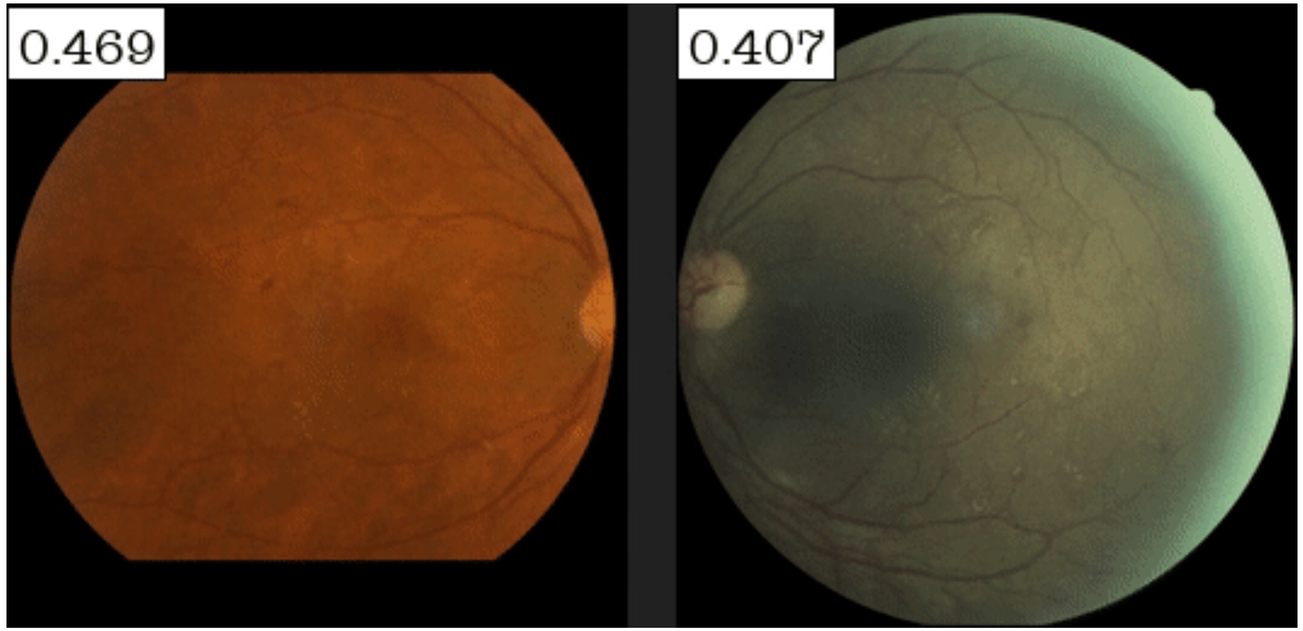


Generative data as a substrate for visual analysis

Phillip Isola

AIM Workshop

Oct 16th 2021

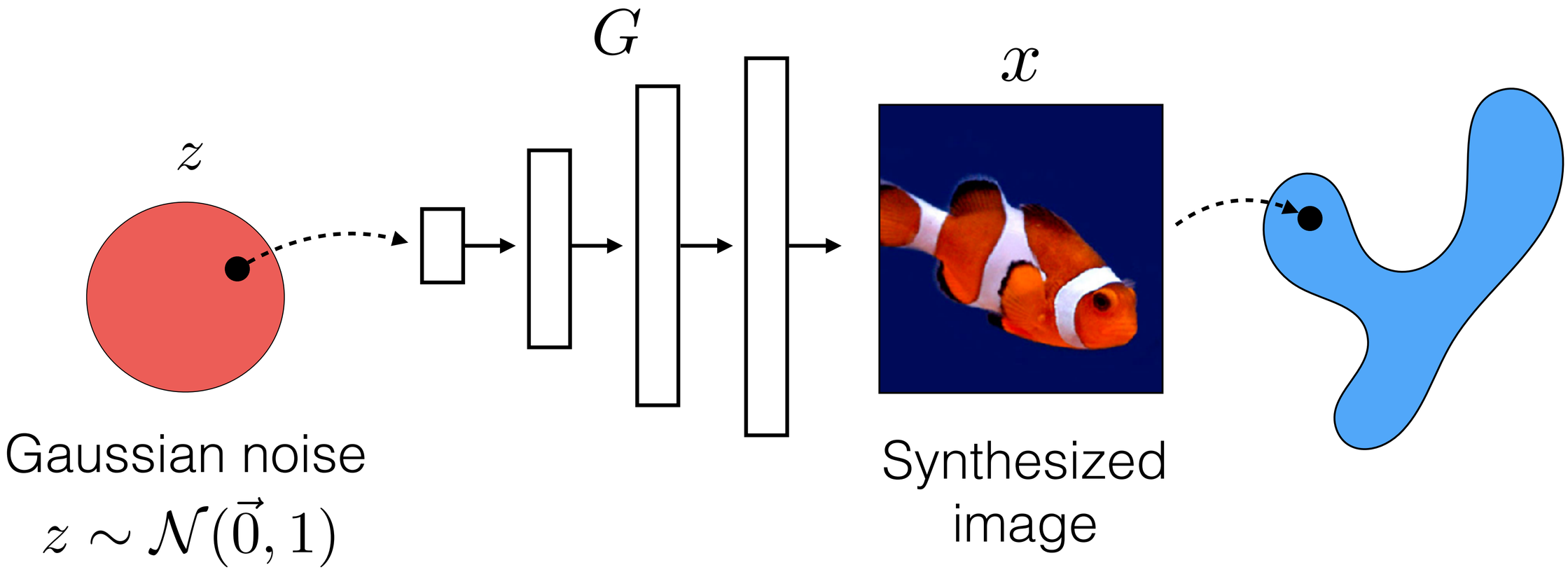
This Cat Does Not Exist

[https://thiscatdoesnotexist.com/]

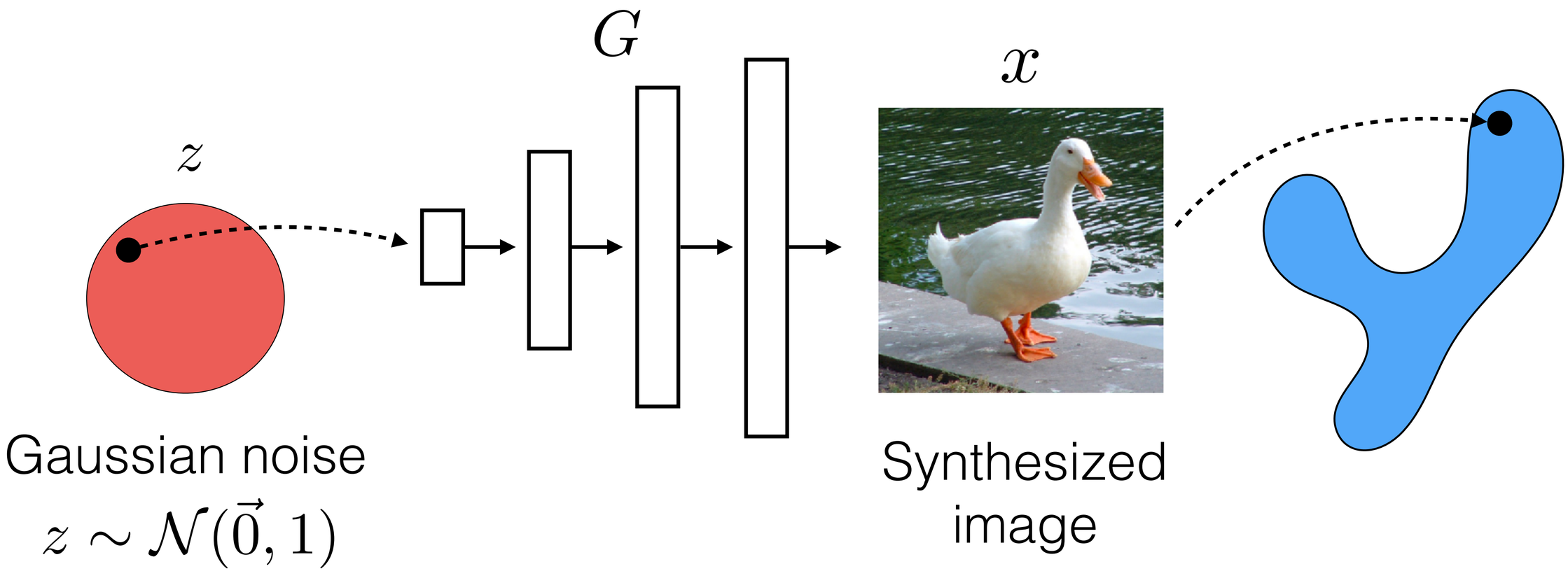
GANs continuously approximate real images

Goodfellow et al. 2014; StyleGAN2. Karras et al. 2020

Generative Models



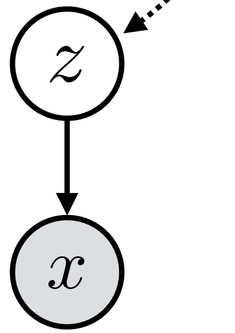
Generative Models



Data++

Datapoint

“DNA” of an image

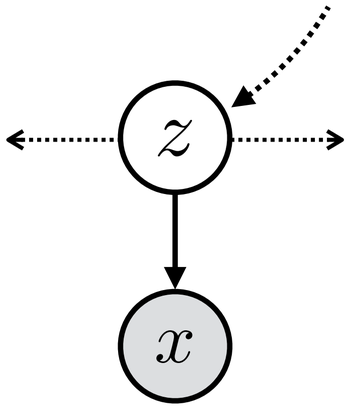


Datapoint++

Data++

Datapoint

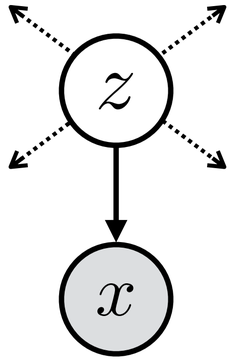
“DNA” of an image



Datapoint++

Generative models as Data++

Datapoint



Datapoint++

Data++ is data with extra functionality

$$\mathbb{X} = \{x, z, G, G^{-1}\}$$

It's data you can navigate, manipulate, and optimize through latent space controls

(Can't you do these things on regular data? No: it takes you "off the manifold")

$\rightarrow$ Graphics, visualization, data aug, counterfactual reasoning, …

Data++ supports counterfactual reasoning

i.e. “What would it have looked like if …?”

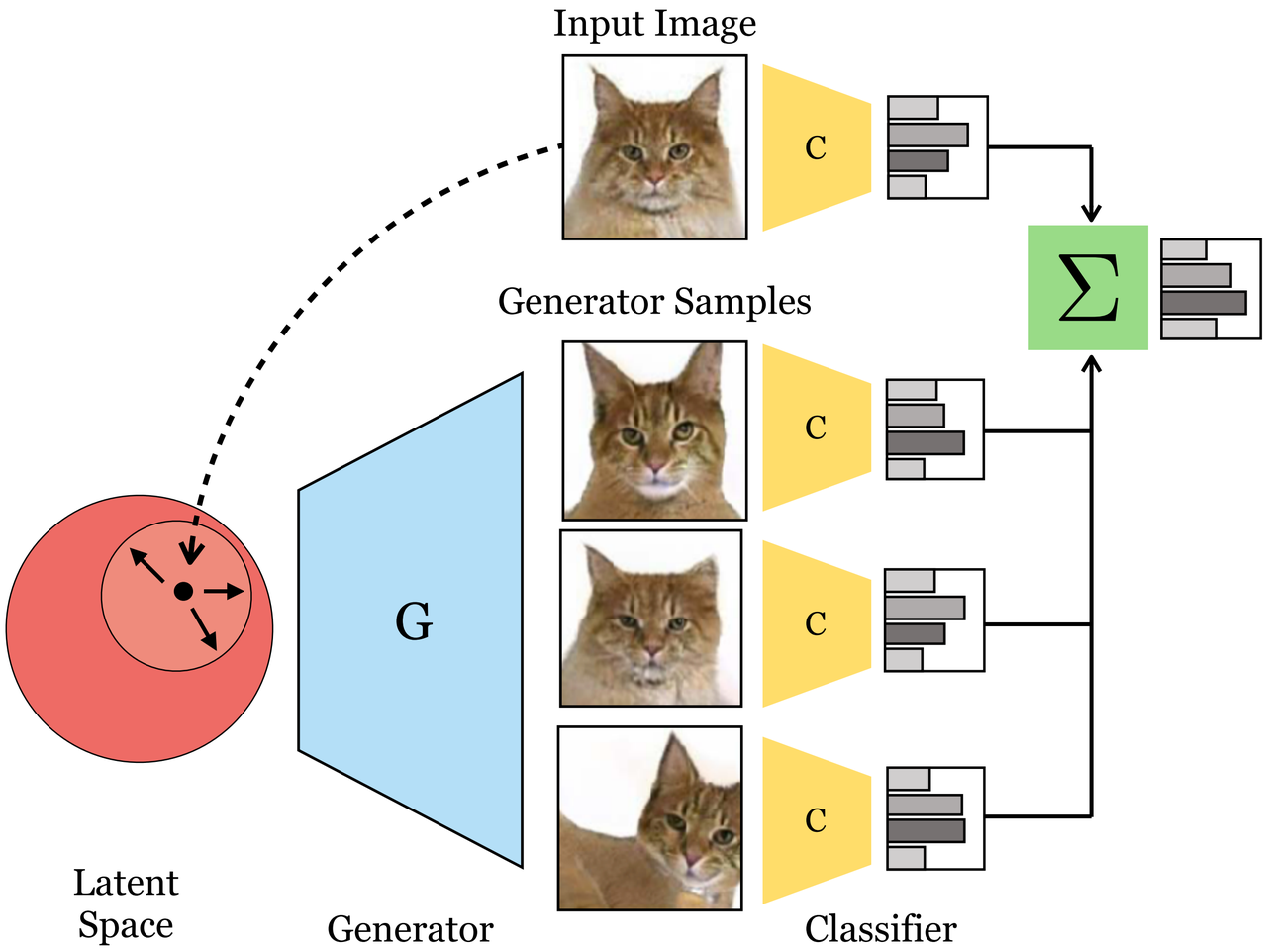
Observation

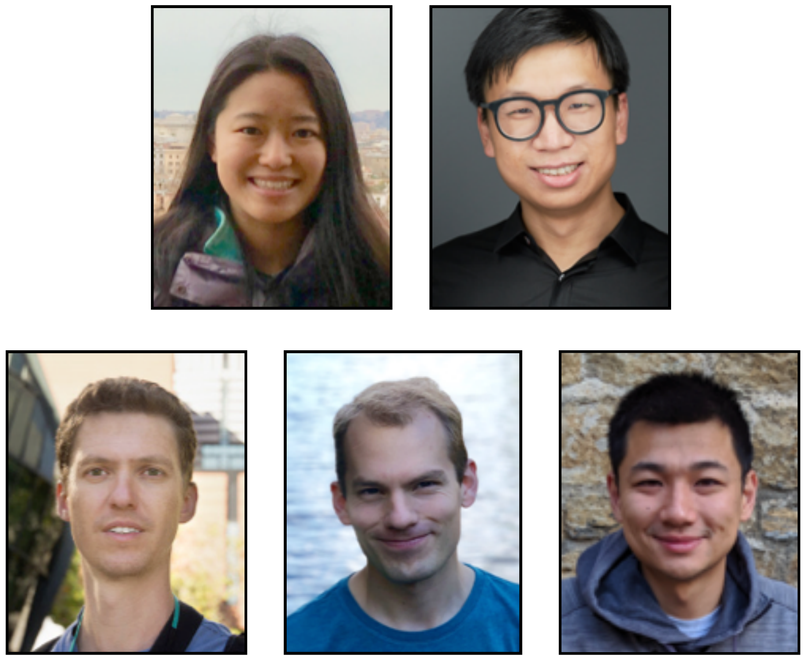
Counterfactual hallucinations

see also: [Mao, Cha, Gupta, Wang, Yang, Vondrick, 2020]
[Sauer & Geiger, 2021]
[Liu, Kailkhura, Loveland, Han, 2019]
[Goetschalckx, Andonian, Oliva, Isola, 2019]
[Oktay, Vondrick, Torralba, 2018]
…

Ensembling with Deep Generative Views

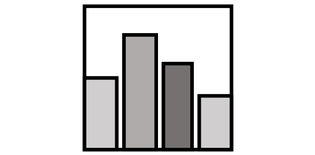
[Chai, Zhu, Shechtman, Isola, Zhang, CVPR 2021]





[Slides credit: Lucy Chai]

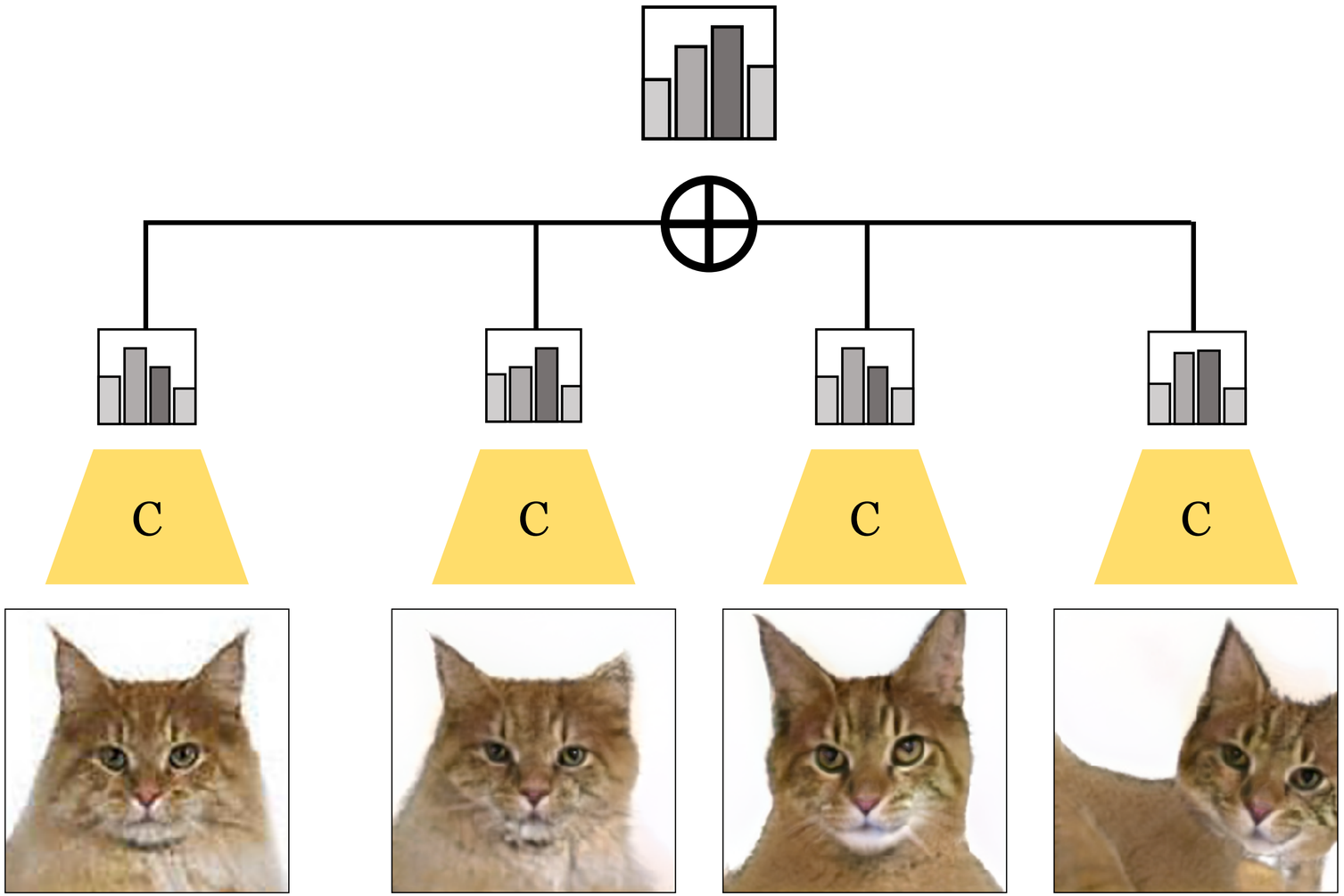
Ensembling GAN views for Classification



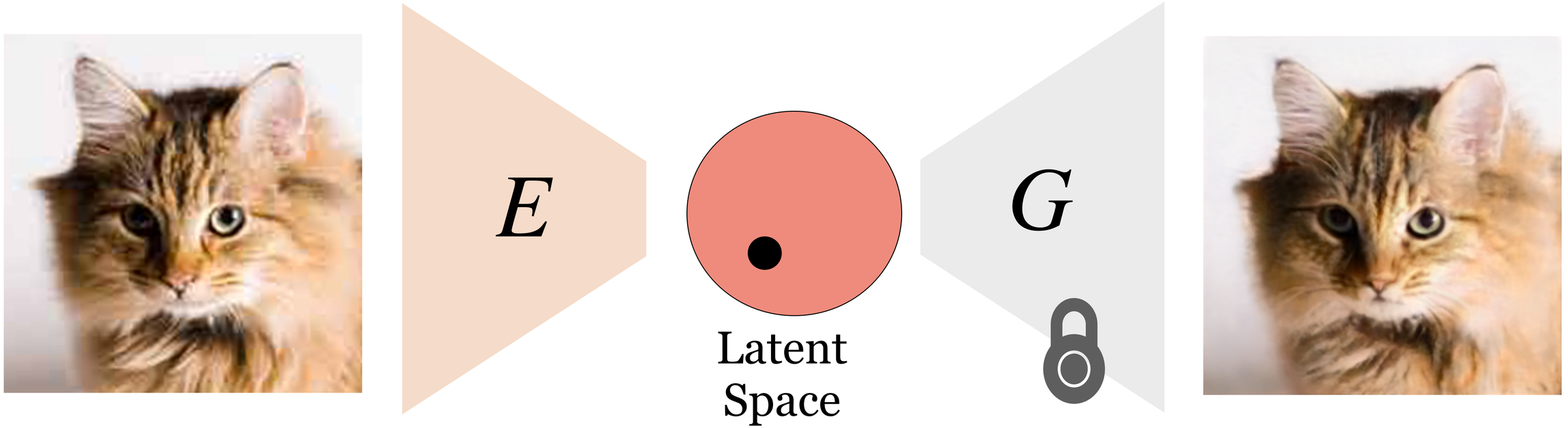
C

Input Image

Ensembling GAN views for Classification



Projecting images into GAN latent space



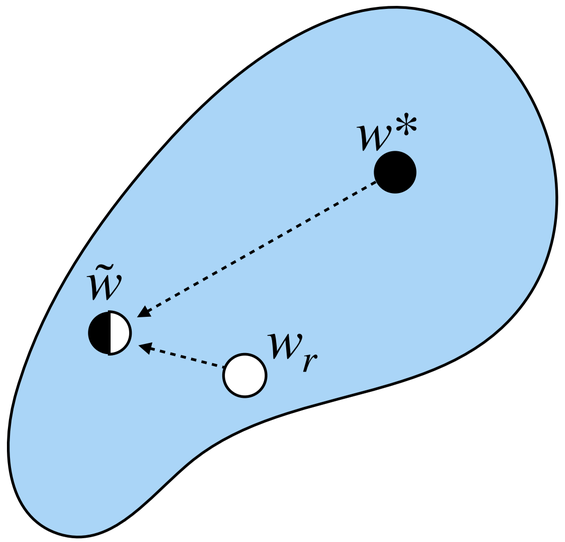
$$w^* = \arg\min_w L_{\text{img}}(x, G(w)) + \lambda L_{\text{latent}}(w, E(x))$$

iGAN. Zhu et al. 2016; GAN Inversion: A Survey. Xia et al. 2021

Types of Perturbations in Latent Code

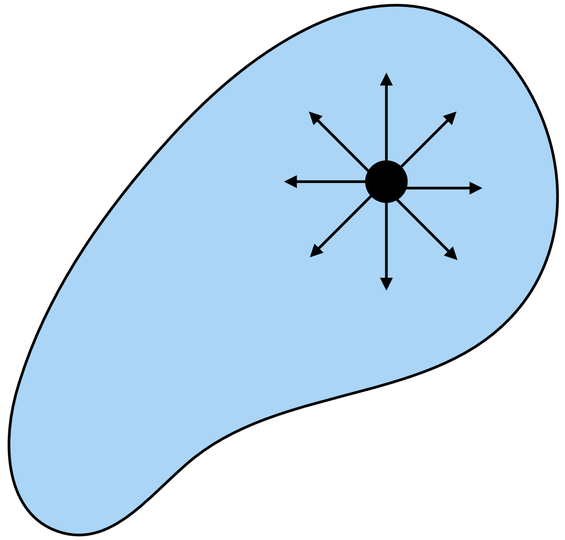
Style-mixing

$\tilde{w} = \text{mix}(w^*, w_r)$



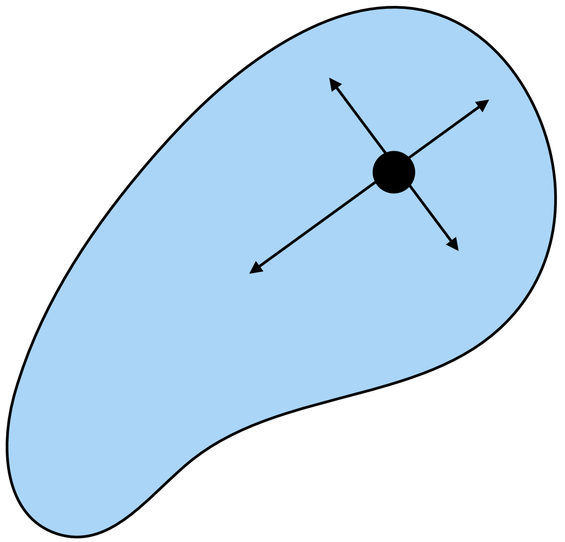
Isotropic

$\tilde{w} \sim \mathcal{N}(w^*, \sigma I)$



PCA Directions

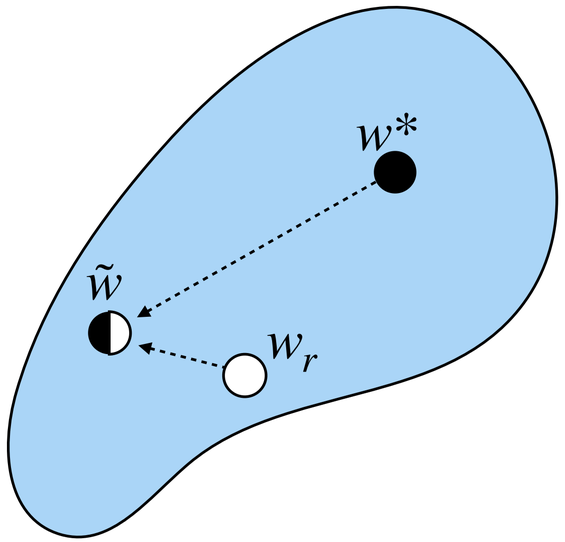
$\tilde{w} = w^* + \beta \tilde{v}_d$



Types of Perturbations in Latent Code

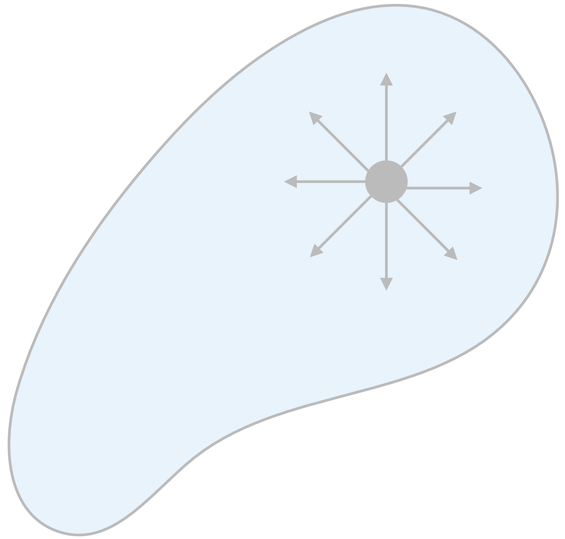
Style-mixing

$$\tilde{w} = \text{mix}(w^*, w_r)$$



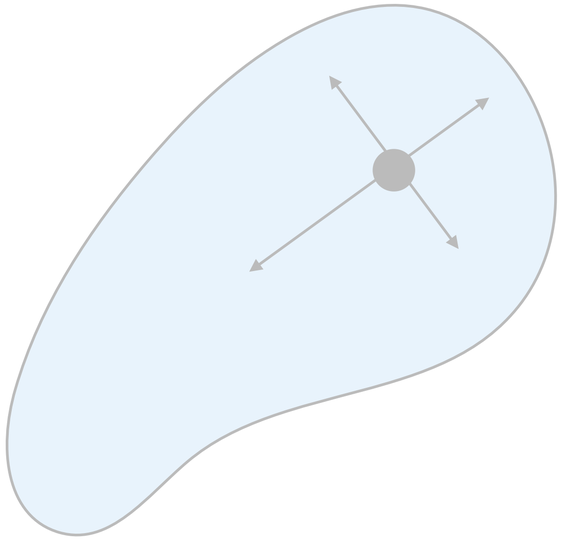
Isotropic

$$\tilde{w} \sim \mathcal{N}(w^*, \sigma I)$$



PCA Directions

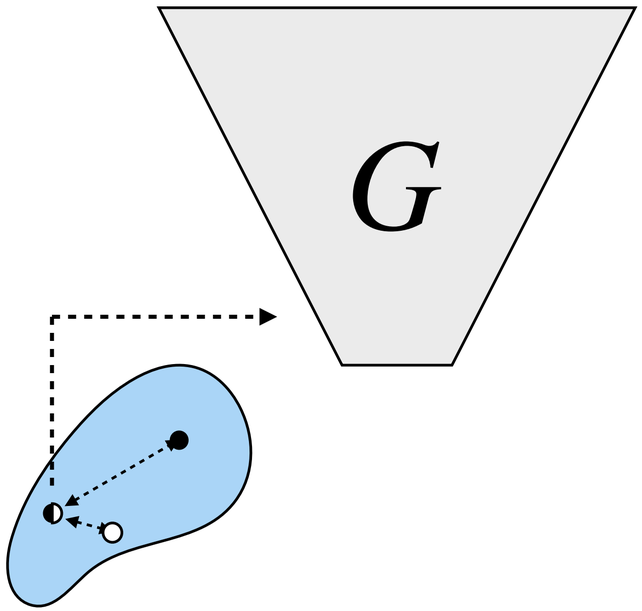
$$\tilde{w} = w^* + \beta \tilde{v}_d$$



Types of Perturbations in Latent Code

Style-mixing

$$\tilde{w} = \text{mix}(w^*, w_r)$$



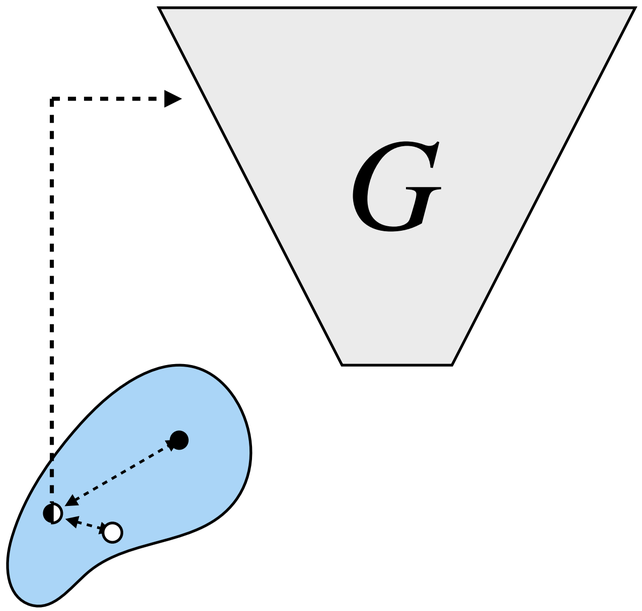
Reconstruction

Coarse Layers

Types of Perturbations in Latent Code

Style-mixing

$$\tilde{w} = \text{mix}(w^*, w_r)$$

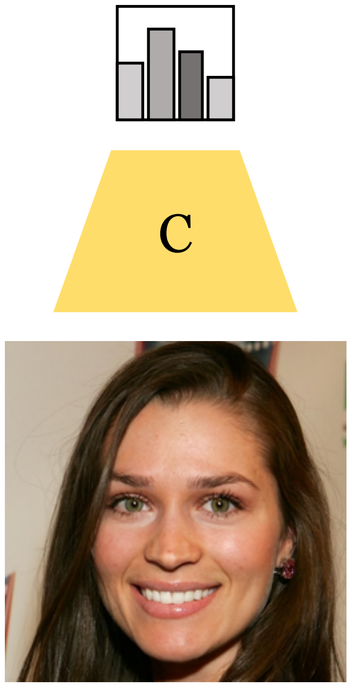


Reconstruction

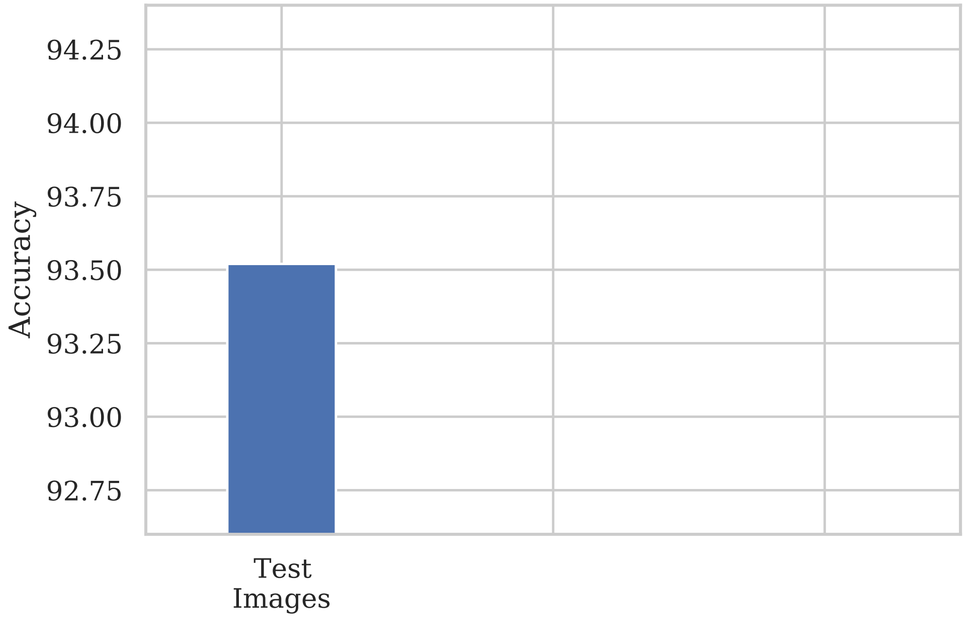
Coarse Layers

Fine Layers

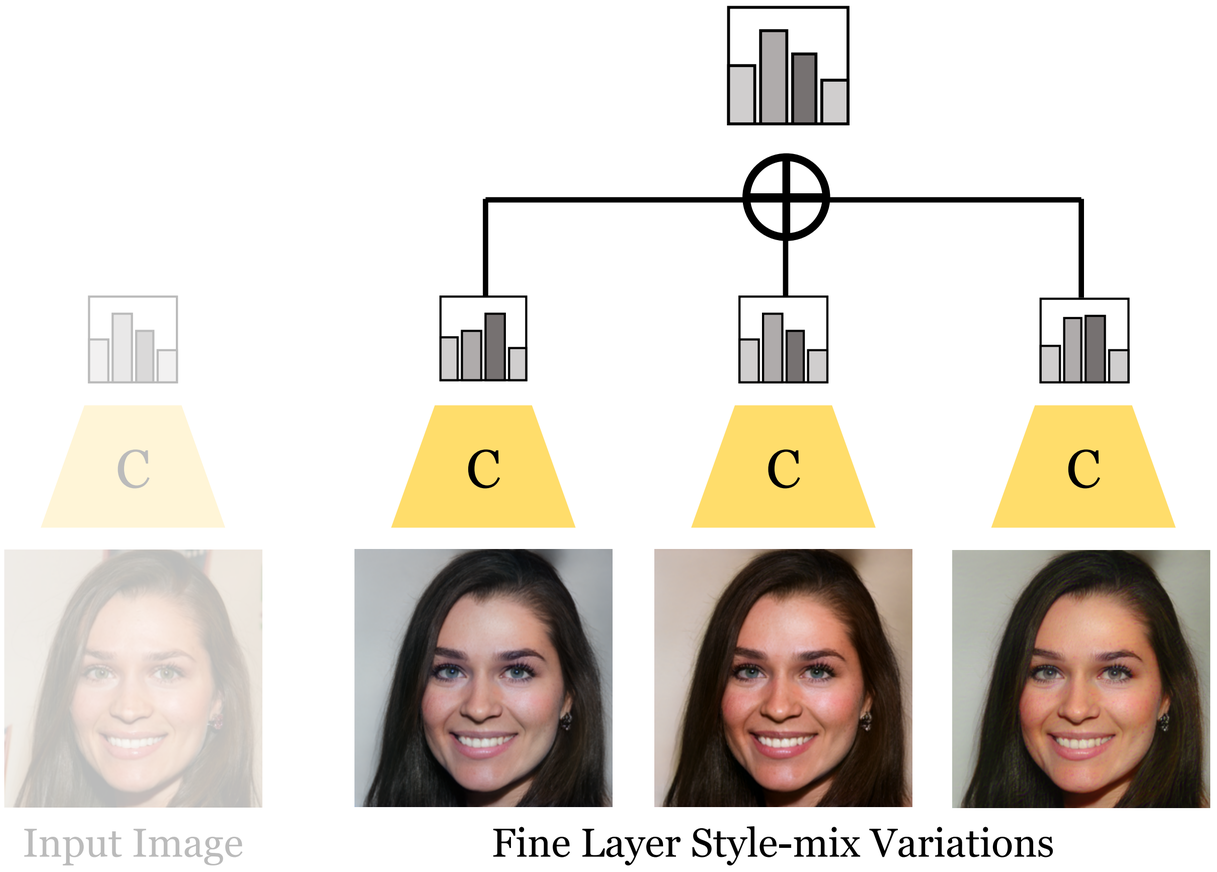
Investigating Ensemble Weight

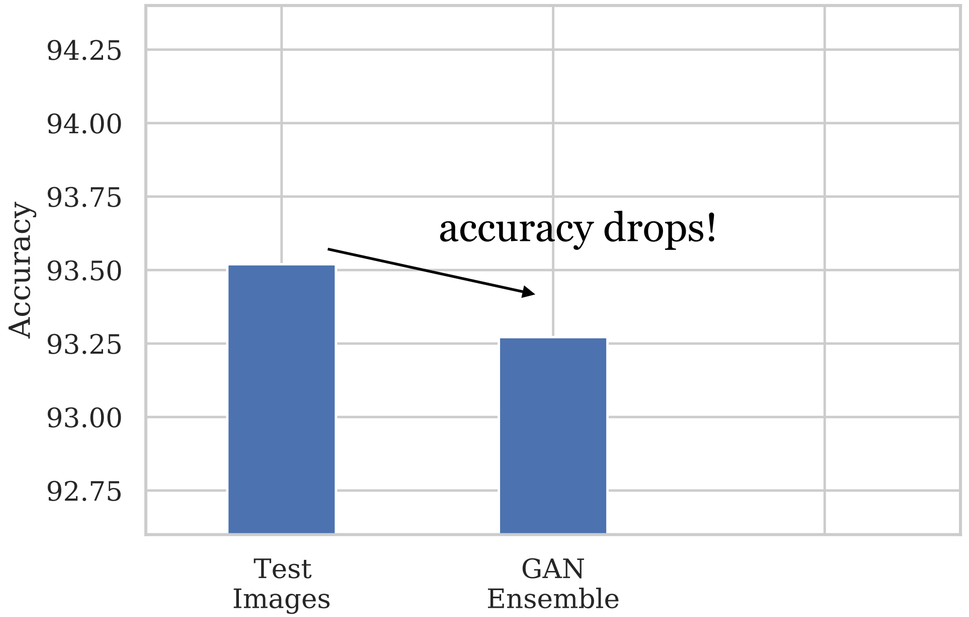


Input Image

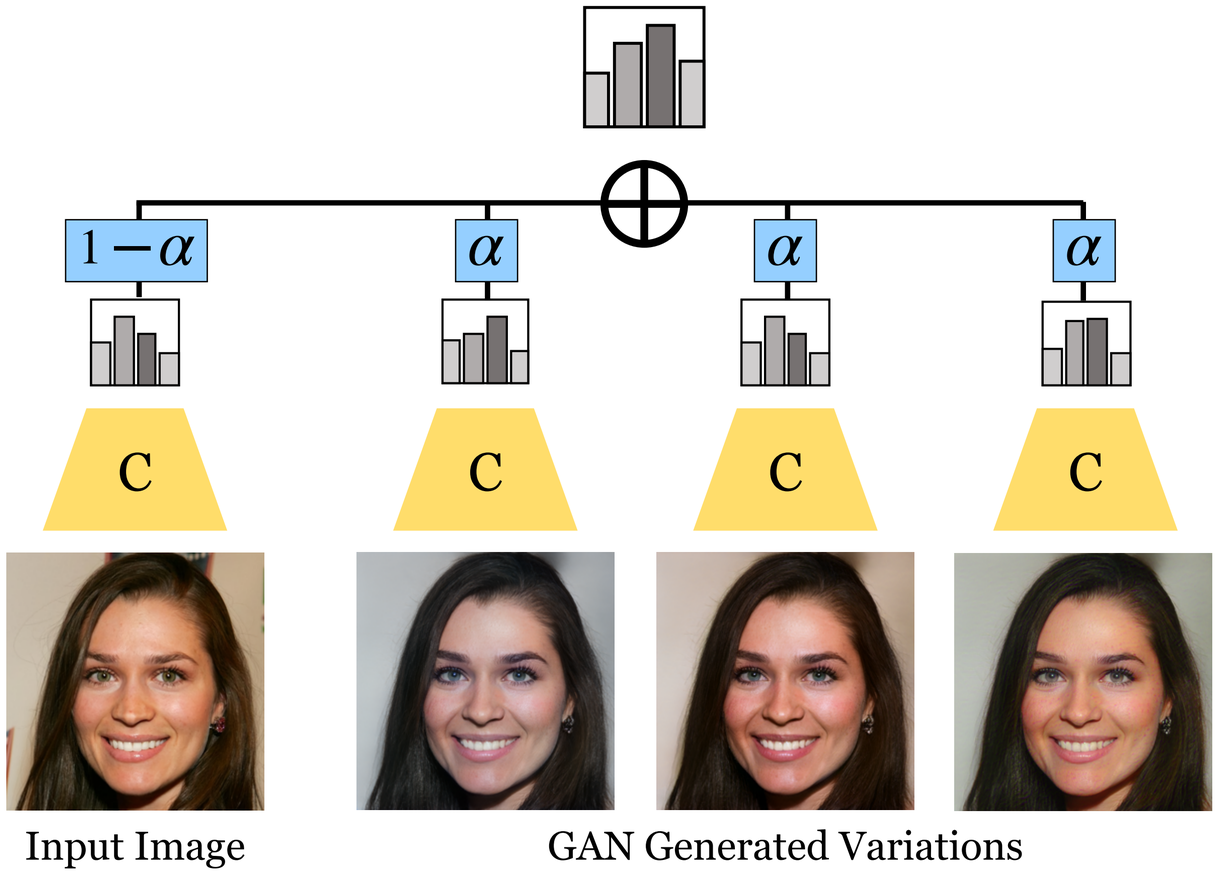


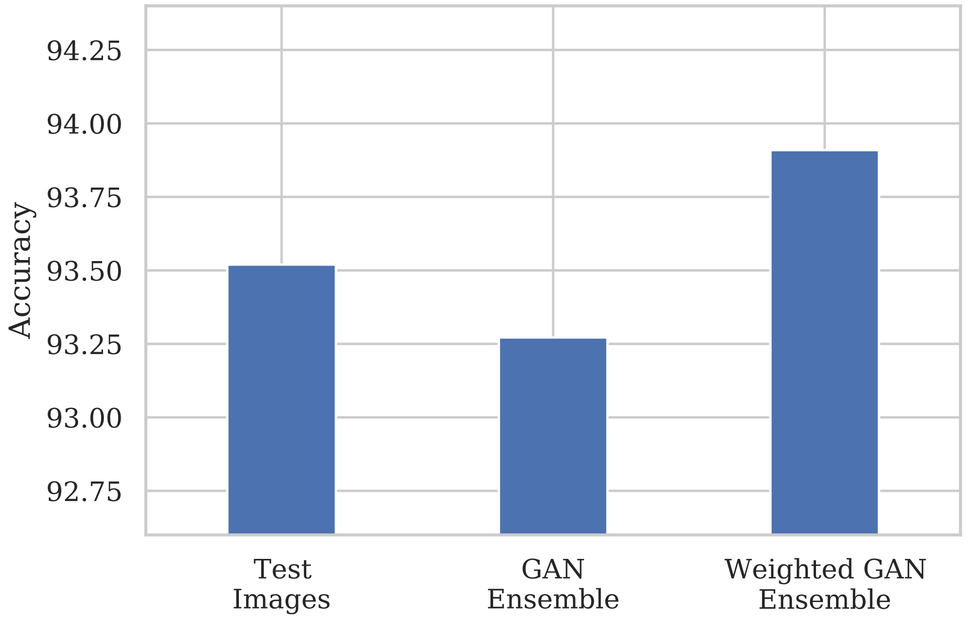
Investigating Ensemble Weight





Investigating Ensemble Weight

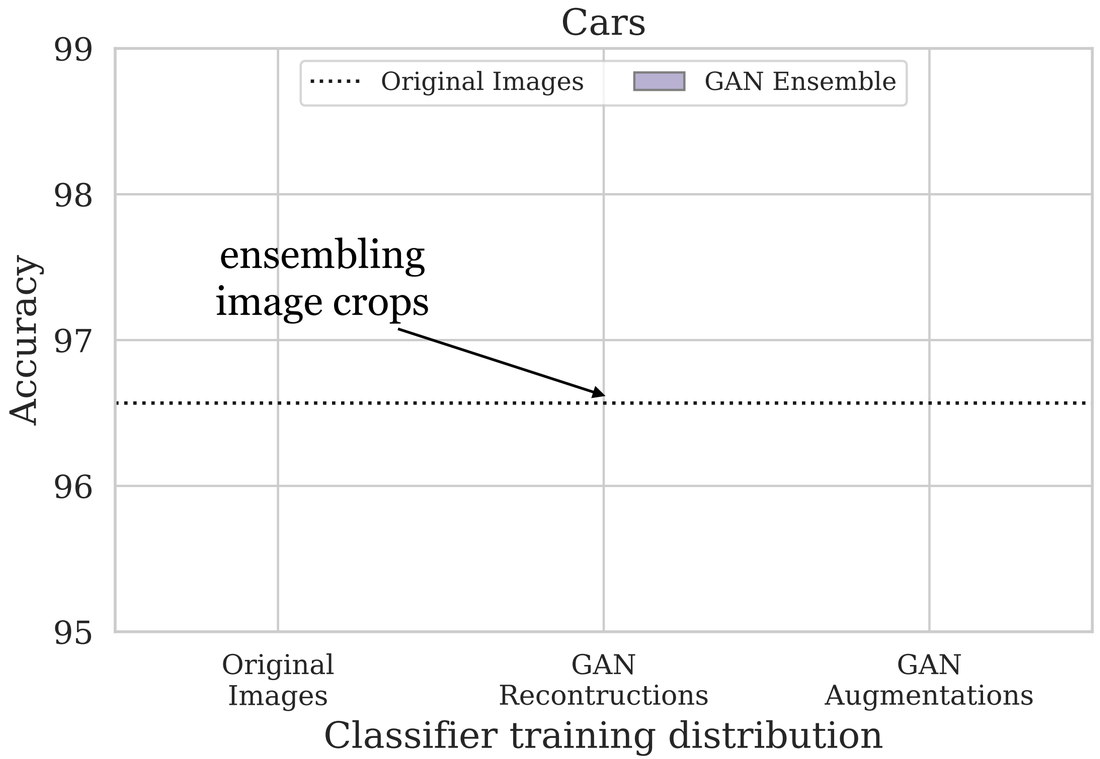




Three-way Cars Domain

Input

Reconstruction

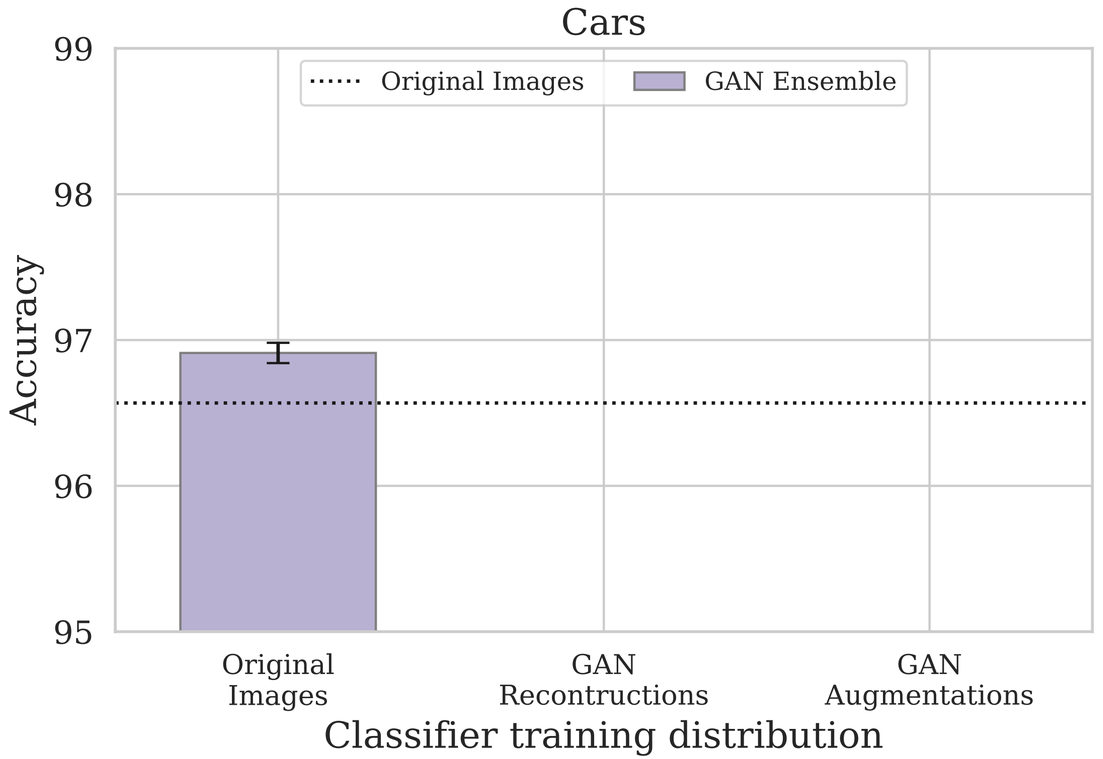


Three-way Cars Domain

Input

Reconstruction

Style-mix Fine Layers

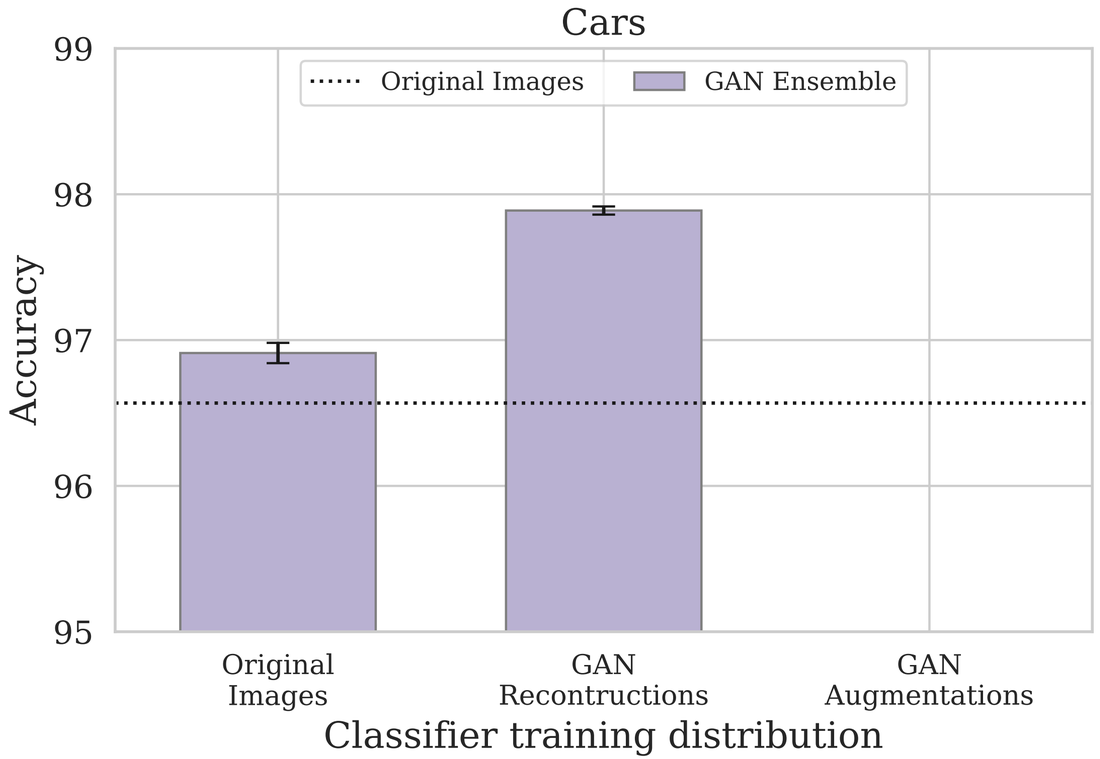


Three-way Cars Domain

Input

Reconstruction

Style-mix Fine Layers

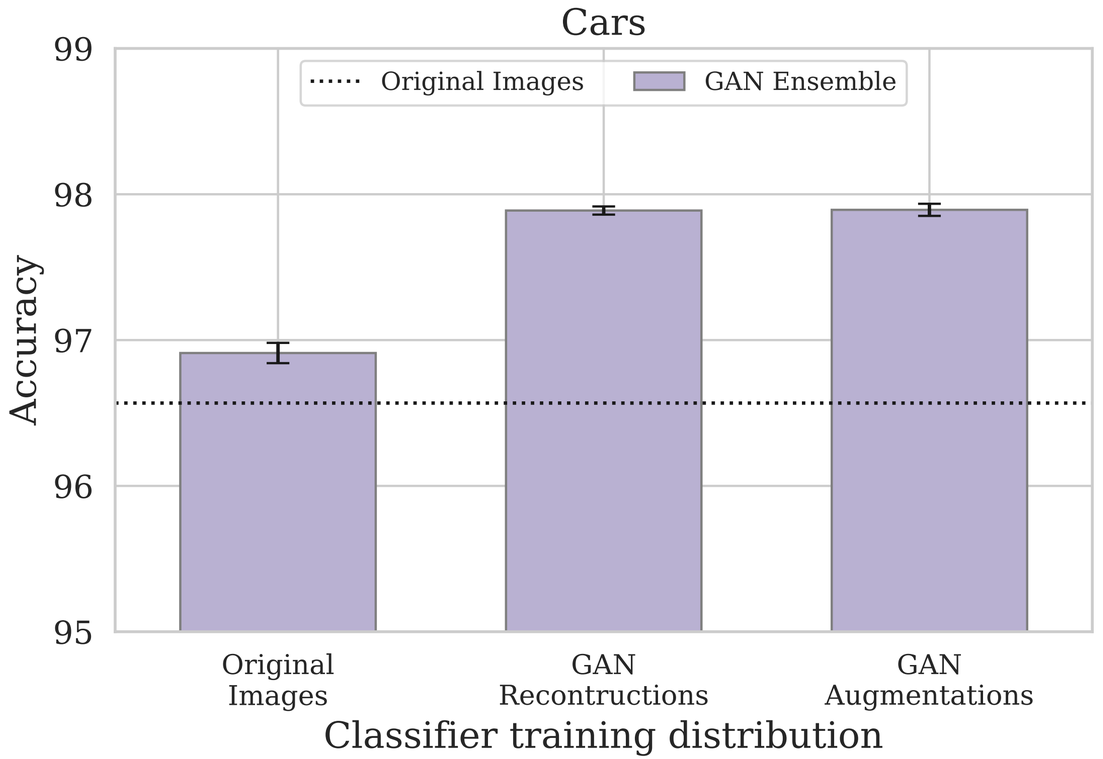


Three-way Cars Domain

Input

Reconstruction

Style-mix Fine Layers

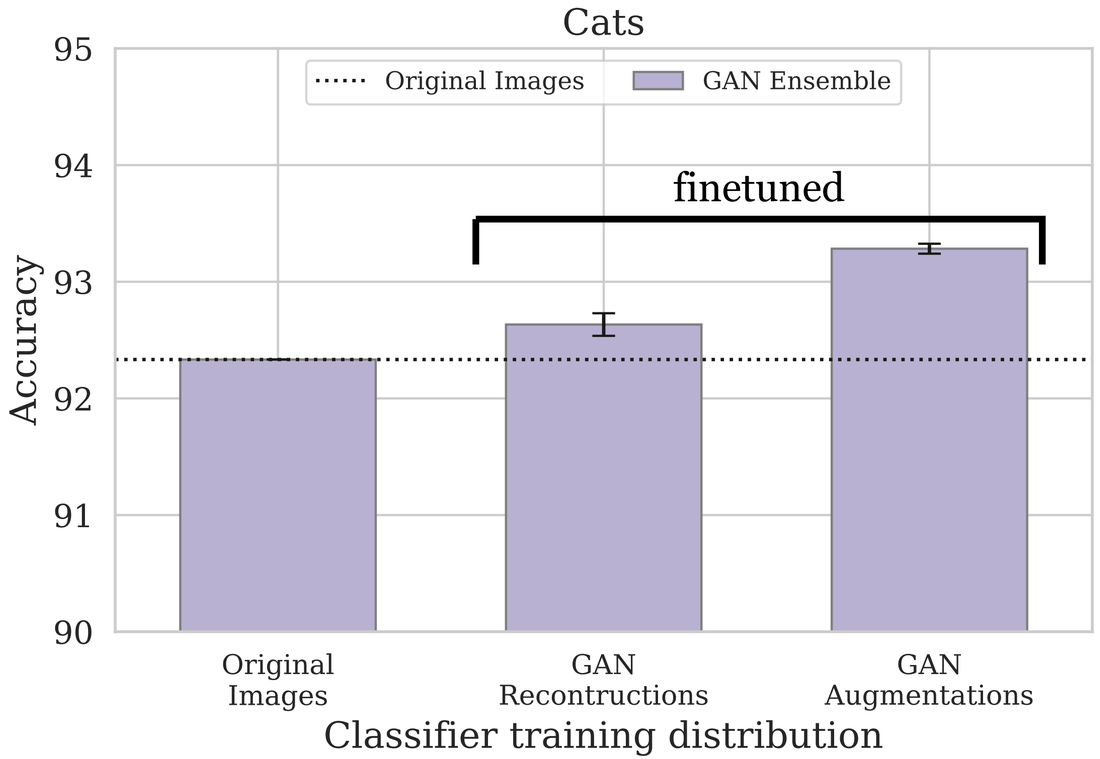


12-way Cats Domain

Input

Reconstruction

Style-mix Coarse Layers



Limitations

- GAN **reconstruction** capability
- GAN inversion **efficiency**
- Classifier **sensitivities** to GAN output
- Currently limited to simple tasks with small, structured datasets…
- But generation and inversion technology is rapidly improving!

Summary

- StyleGAN as a generator of image variations
- Project image into latent space and perturb
- Requires adjustments to mitigate classifier sensitivity to GAN output

Summary

- StyleGAN as a generator of image variations
- Project image into latent space and perturb
- Requires adjustments to mitigate classifier sensitivity to GAN output

Project Website + Code + Colab:
https://chail.github.io/gan-ensembling/

Research at Google

Explaining in Style:
Training a GAN to explain a classifier
ICCV 2021

Oran Lang

Yossi Gandelsman

Michal Yarom

Yoav Wald

Gal Elidan

Avinatan Hassidim

Bill Freeman

Philip Isola

Amir Globerson

Michal Irani

Inbar Mosseri

Explaining a classifier

Dog vs. Cat classifier

Why was this image classified as “Cat”?

Prior Work

Dog vs. Cat classifier

Attention Maps

Output important regions

Counterfactual Examples

"Input X w[illegible] X' → output Y would change to Y' "

Only spatially localized attributes

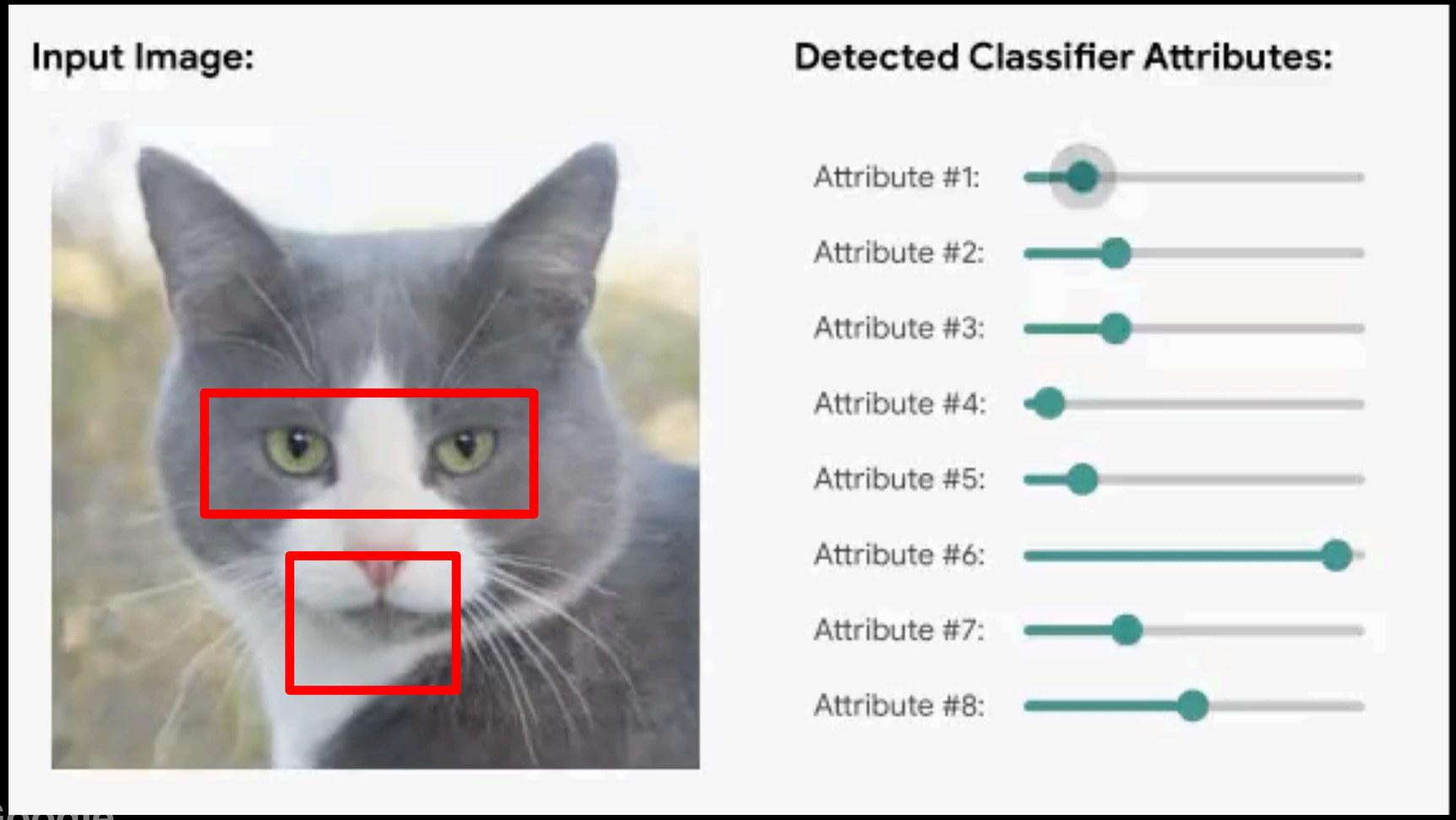
Changes all attributes at once

e.g., [Selvaraju et al. 2017]

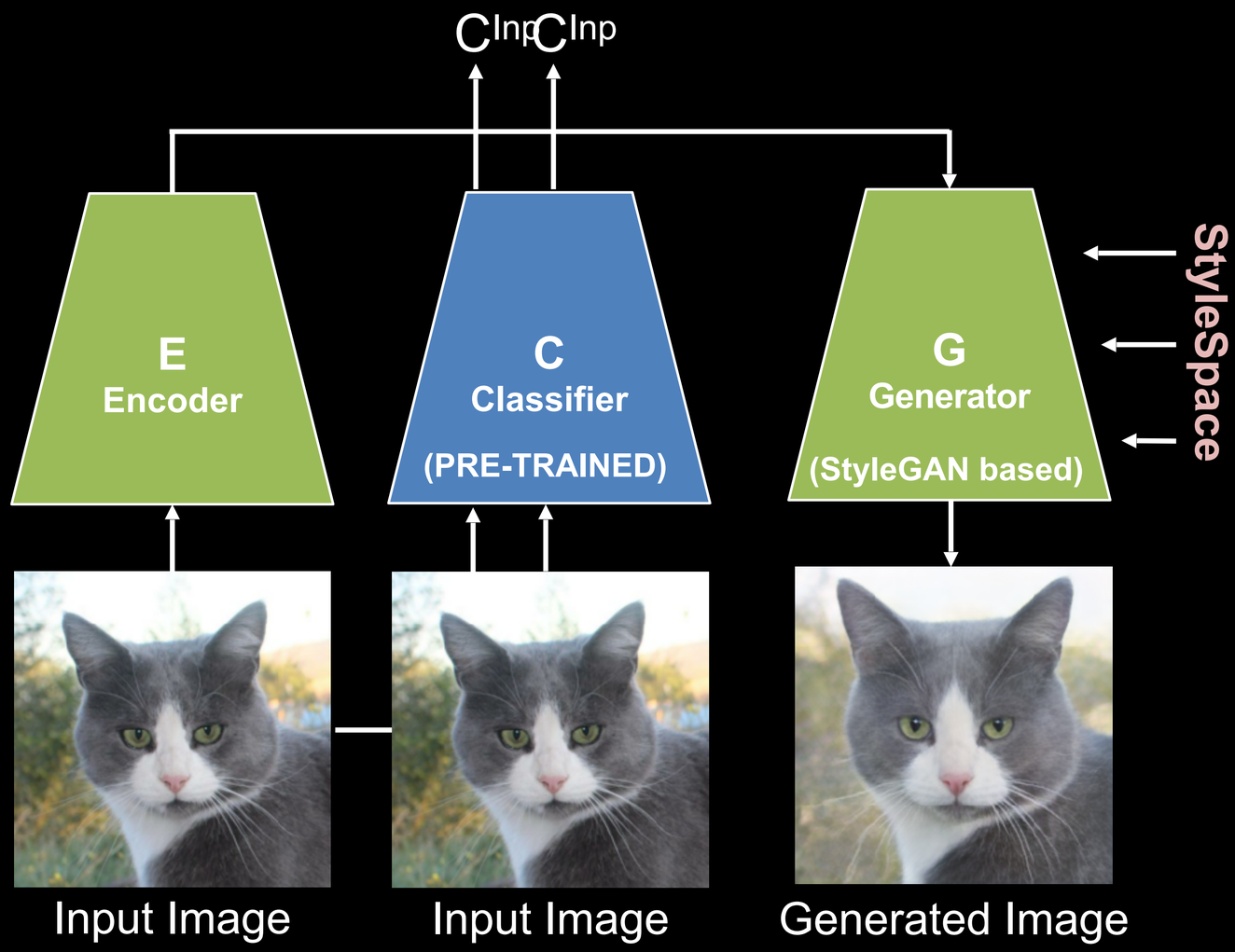
*e.g., "**Ganalyze**" [Goetschalckx et al. 2019]*

Our Approach

Automatically discover **disentangled** attributes ➔ generate counterfactual examples



Method

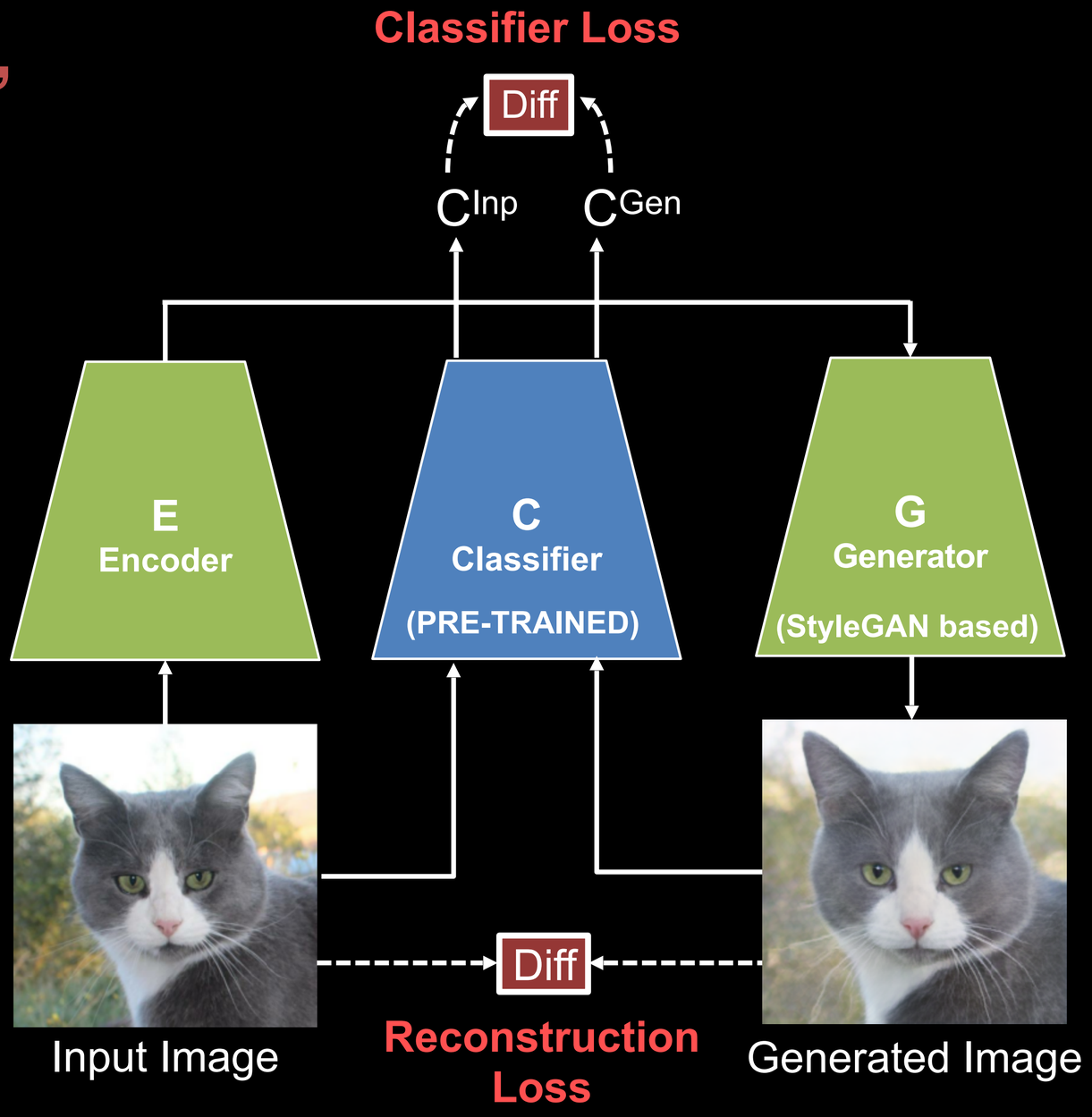


Google

["StyleSpace": Wu, Lischinski, Shechtman 2020]

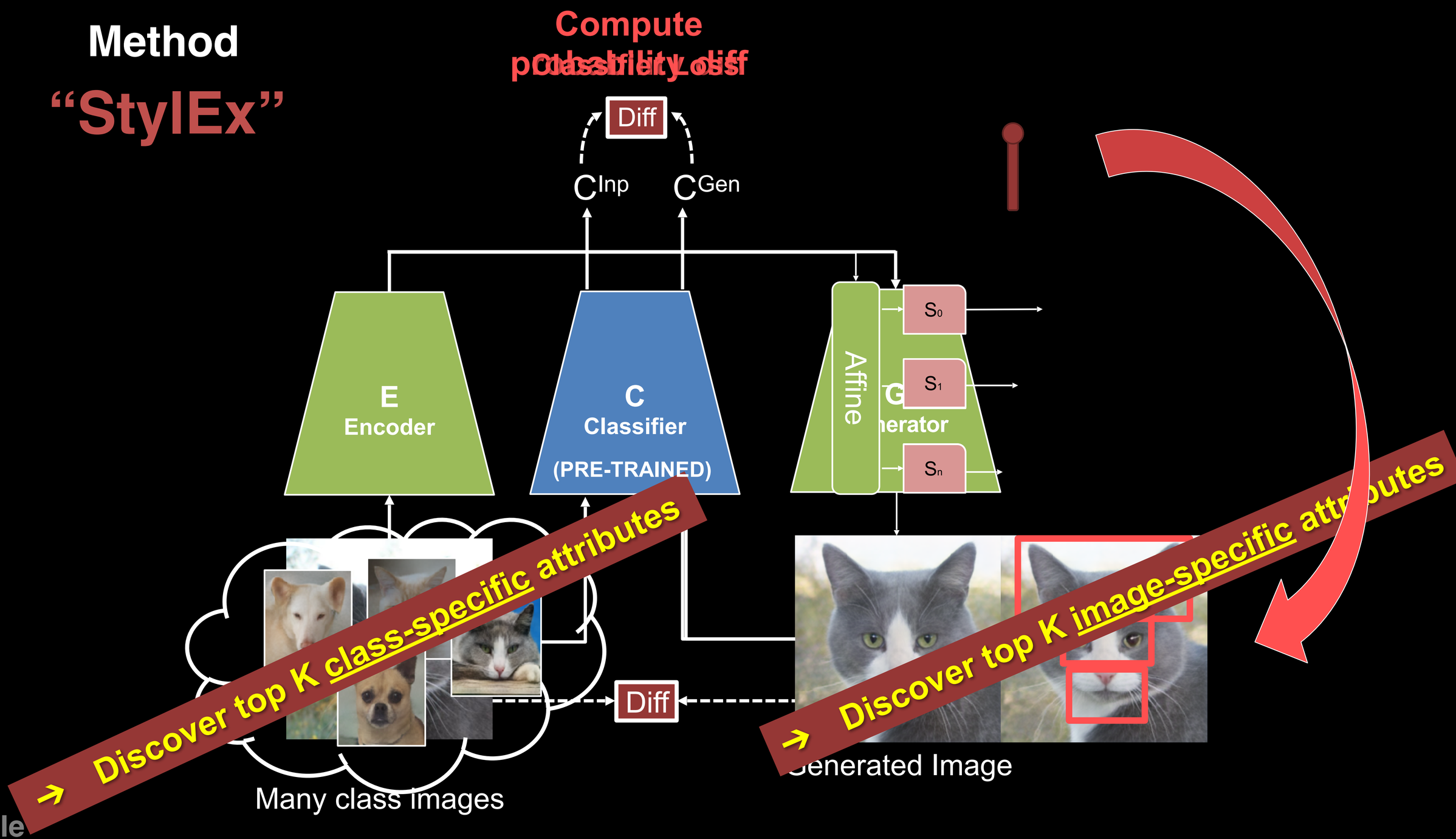
Method

“StylEx”



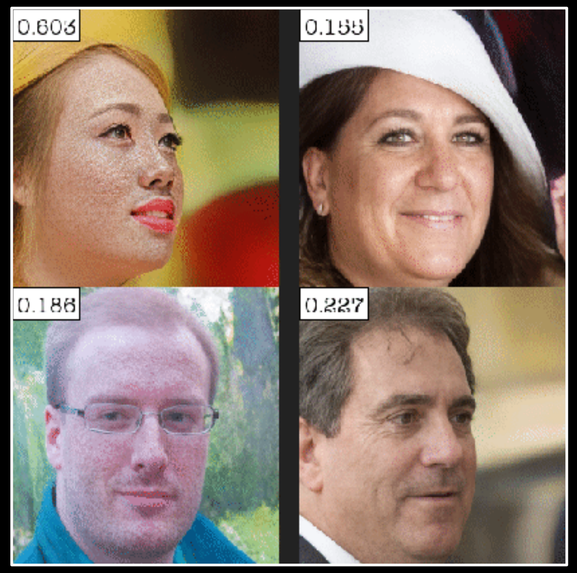
Method

“StylEx”



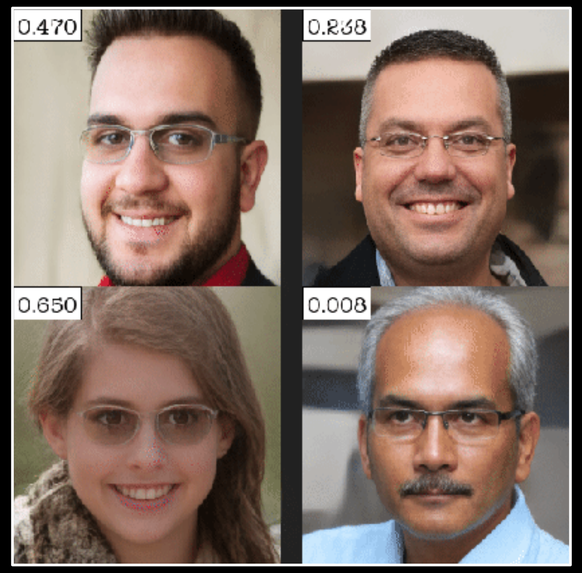
Class-specific explanation

Perceived Age Classifier:



Attribute #1
"Skin Pigmentation"

Attribute #2
"Eyebrow Thickness"

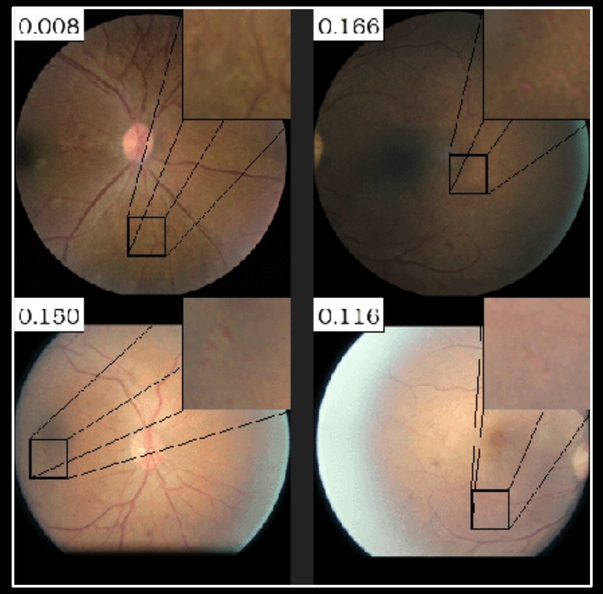


Attribute #3
"Glasses"

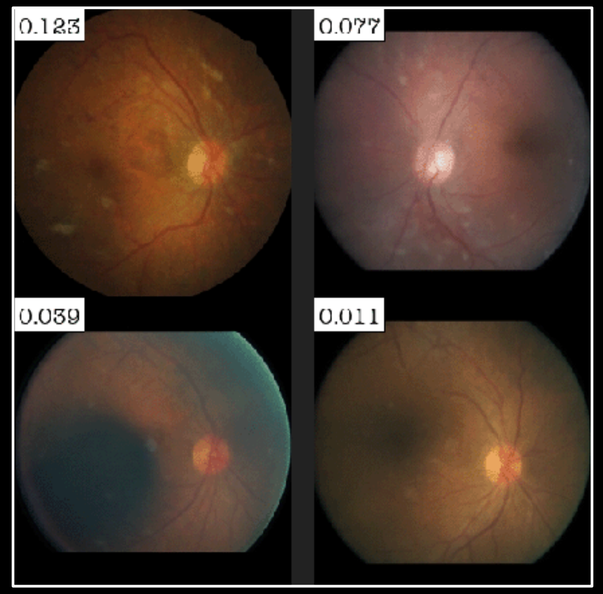
Attribute #4
"Dark / White Hair"

Class-specific explanation

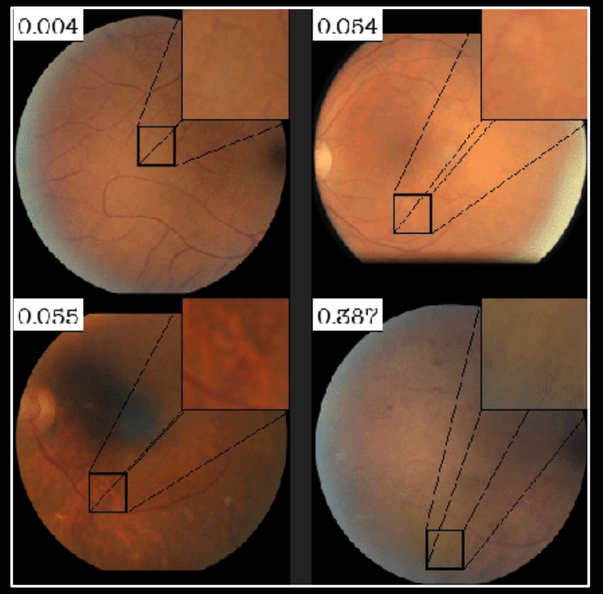
Retinal Fundus Classifier:



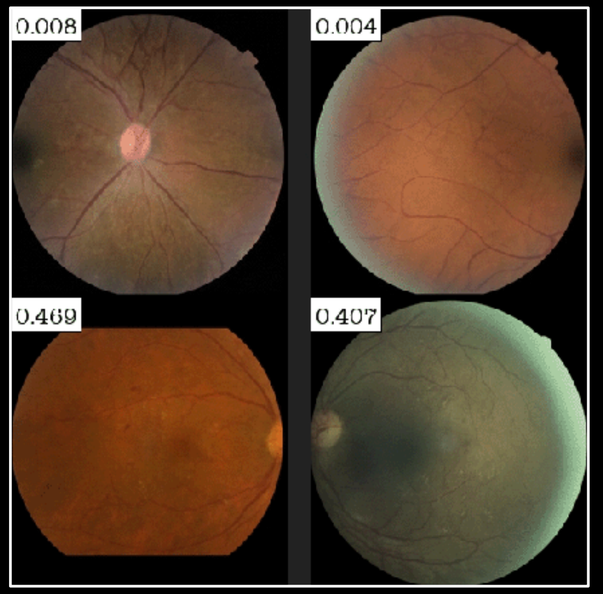
Attribute #1
“Exudates”



Attribute #2
“Cotton Wool”



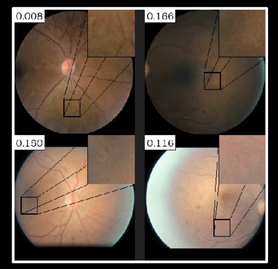
Attribute #3
“Hemorrhages”



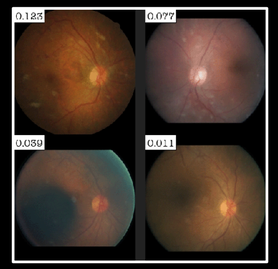
Attribute #4
“Clustered Exudates”

Works on many domains

Retinal Fundus Classifier

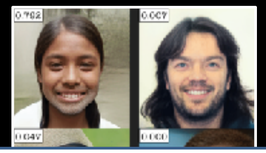


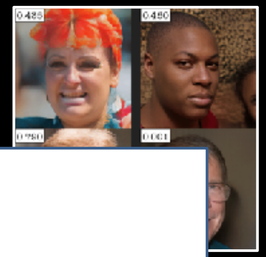
Attribute #1
"Exudates"



Attribute #2
"Cotton Wool"

Perceived Gender Classifier



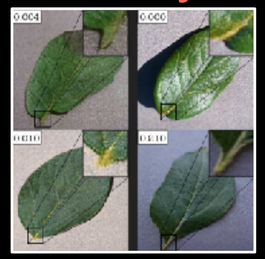


Attribute #4
"Eyebrow Thickness"

"StylEx" can be used for:

- ❖Explain what a classifier has learned
- ❖Detect unknown features in medical images
- ❖Detect classifier/dataset biases

Healthy / Sick Lea



Attribute #1
"Leaf Base Color"

Attribute #
"Rotten Ap

Attribute #4
"Dark / White Hair"

Caltech UCSD bir

Attribute #1
"Black Belly"

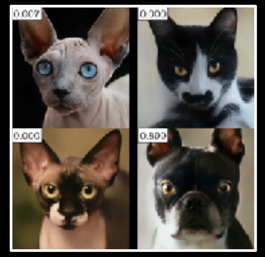
Attribute #2
"Black Upperparts"

Attribute #3
"Solid Wing Pattern"

Attribute #1
"Open / Close Mouth"

Attribute #2
"Ear Dropped / Pointed"

Attribute #3
"Eye Circumference"



Attribute #4
"Eye Shape"

Is this a cat or a dog?

C

“I think it’s a cat but let me get a better angle”

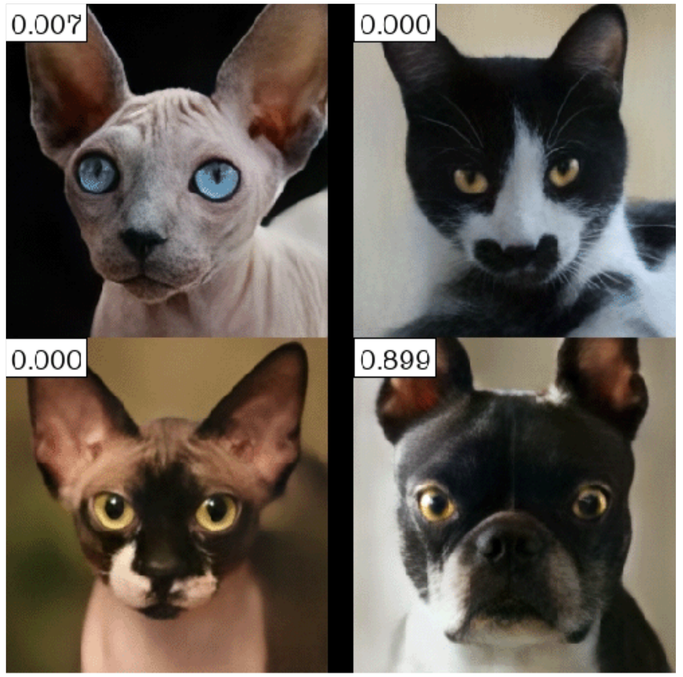
Is this a cat or a dog?

C

"Yes, definitely a cat"

"It's a cat because:"

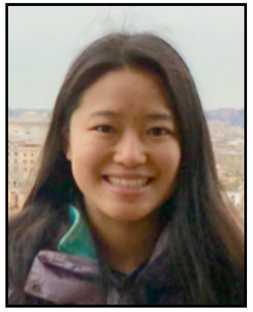
C



"The pupils are narrow"

"The mouth is closed"

Thanks!



Lucy Chai

Jun-Yan Zhu

Oran Lang

Yossi Gandelsman

Michal Yarom

Yoav Wald

Gal Elidan

Eli Shechtman

Richard Zhang

Avinatan Hassidim

Bill Freeman

Amir Globerson

Michal Irani

Inbar Mosseri

Massachusetts Institute of Technology

Adobe

Carnegie Mellon University

Research at Google