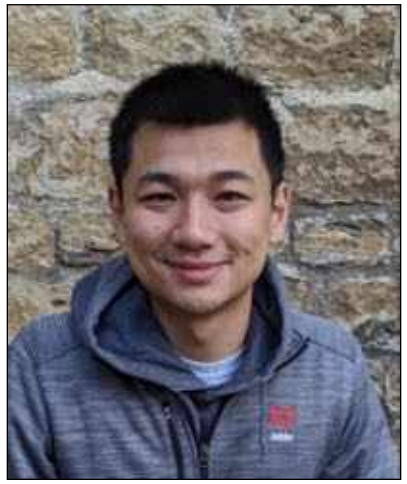
Contrastive Learning for Unpaired Image-to-Image Translation



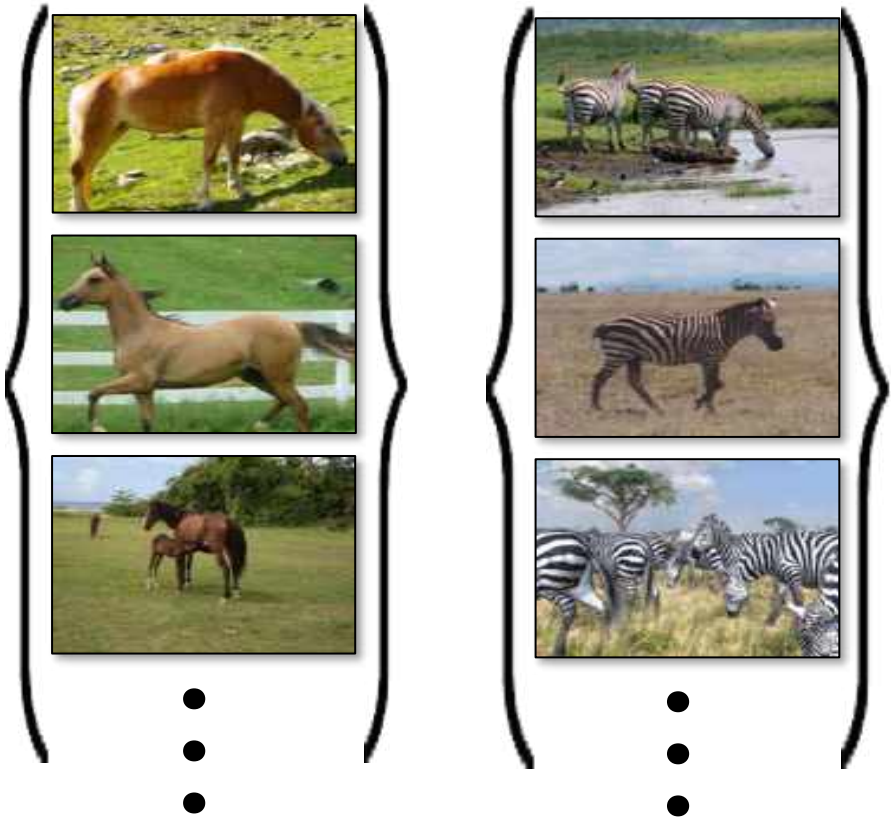



Taesung Park Alexei A. Efros Richard Zhang Jun-Yan Zhu

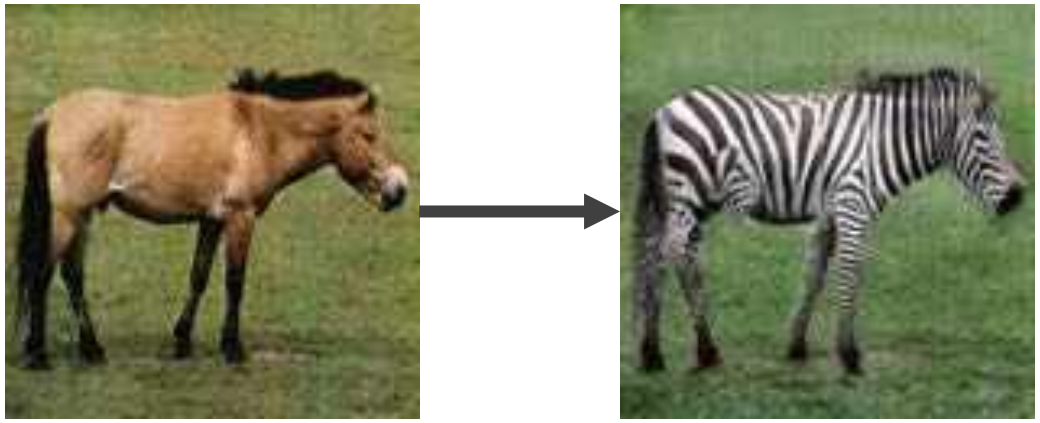
UC Berkeley Adobe Research

ECCV 2020

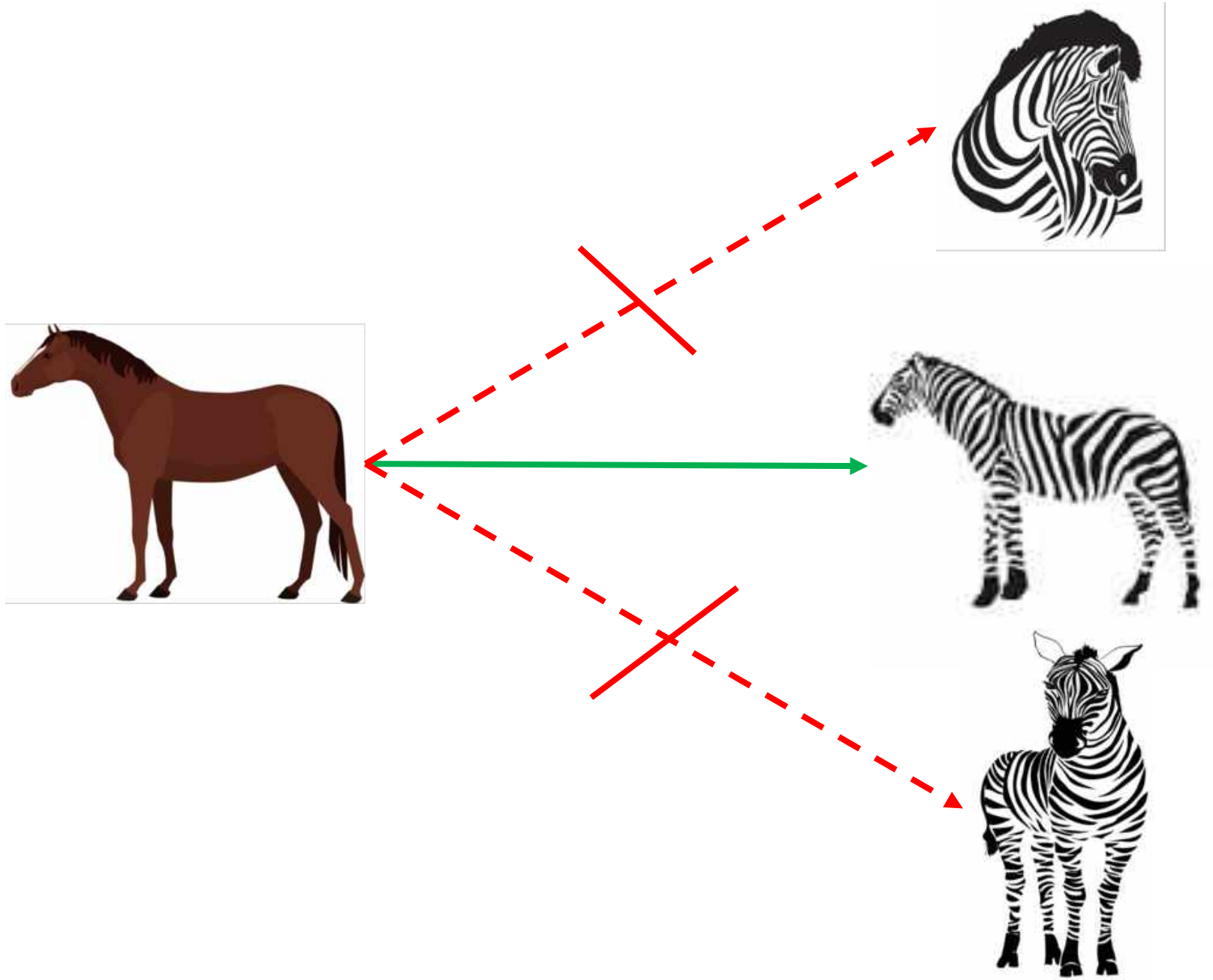
What is Unpaired Image-to-Image Translation?

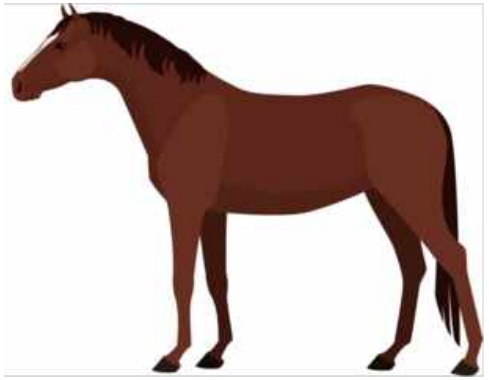


Training Set



Test-time behavior

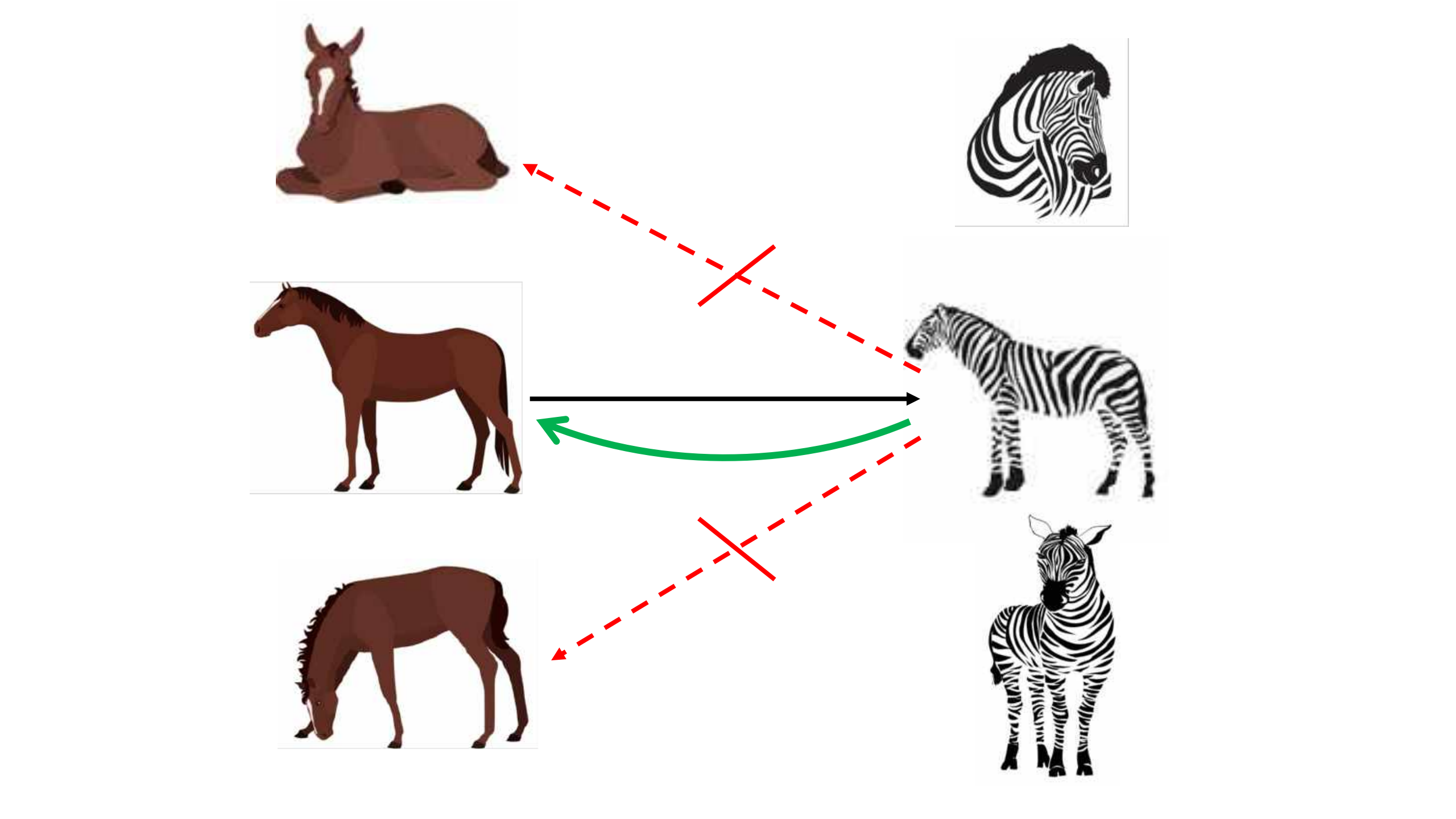


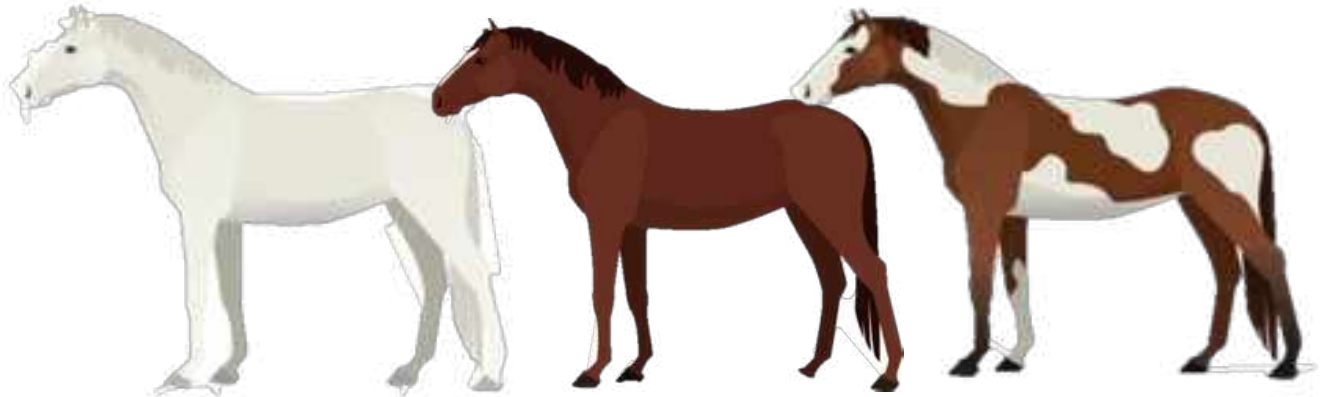


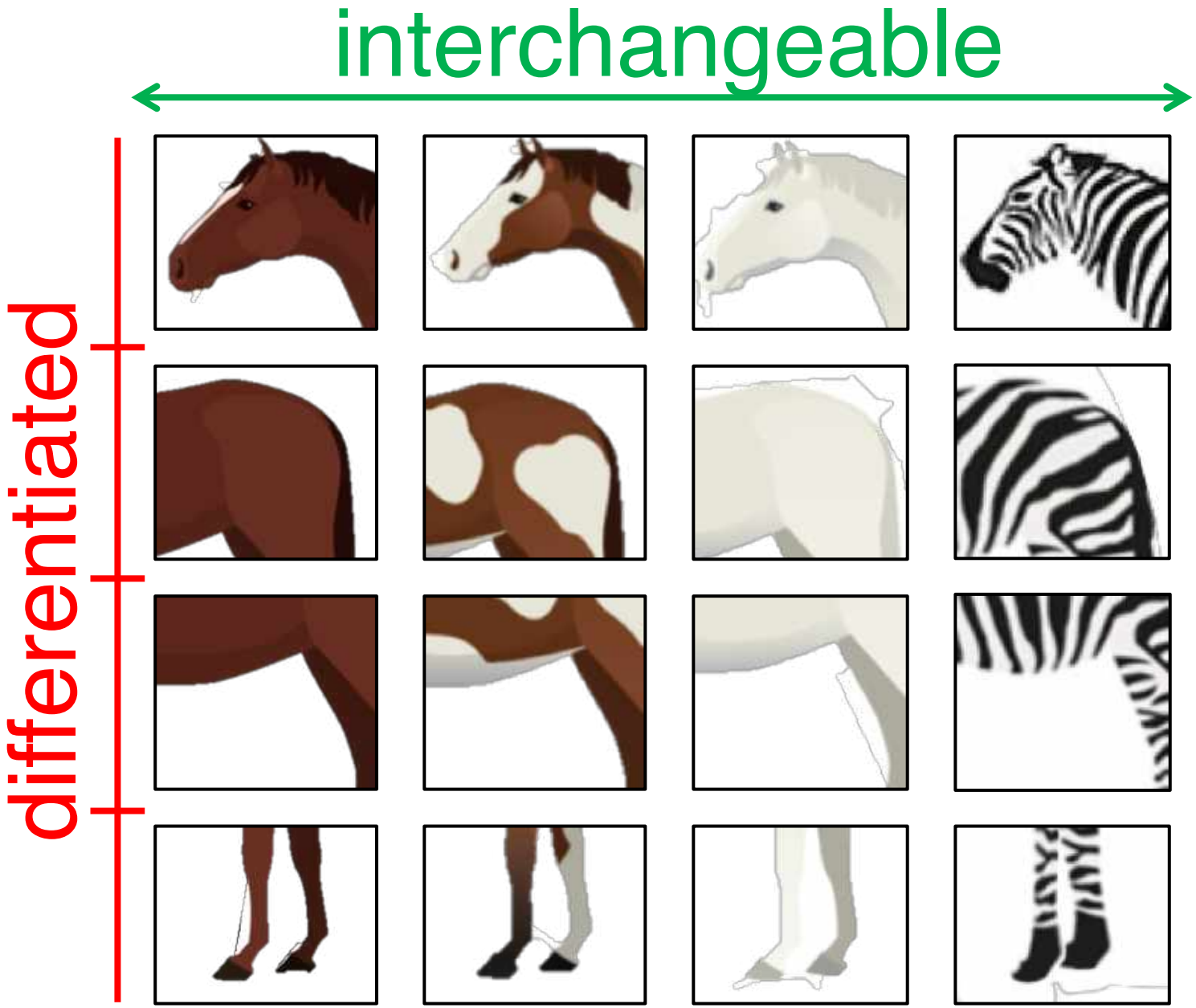
cycle-consistency loss

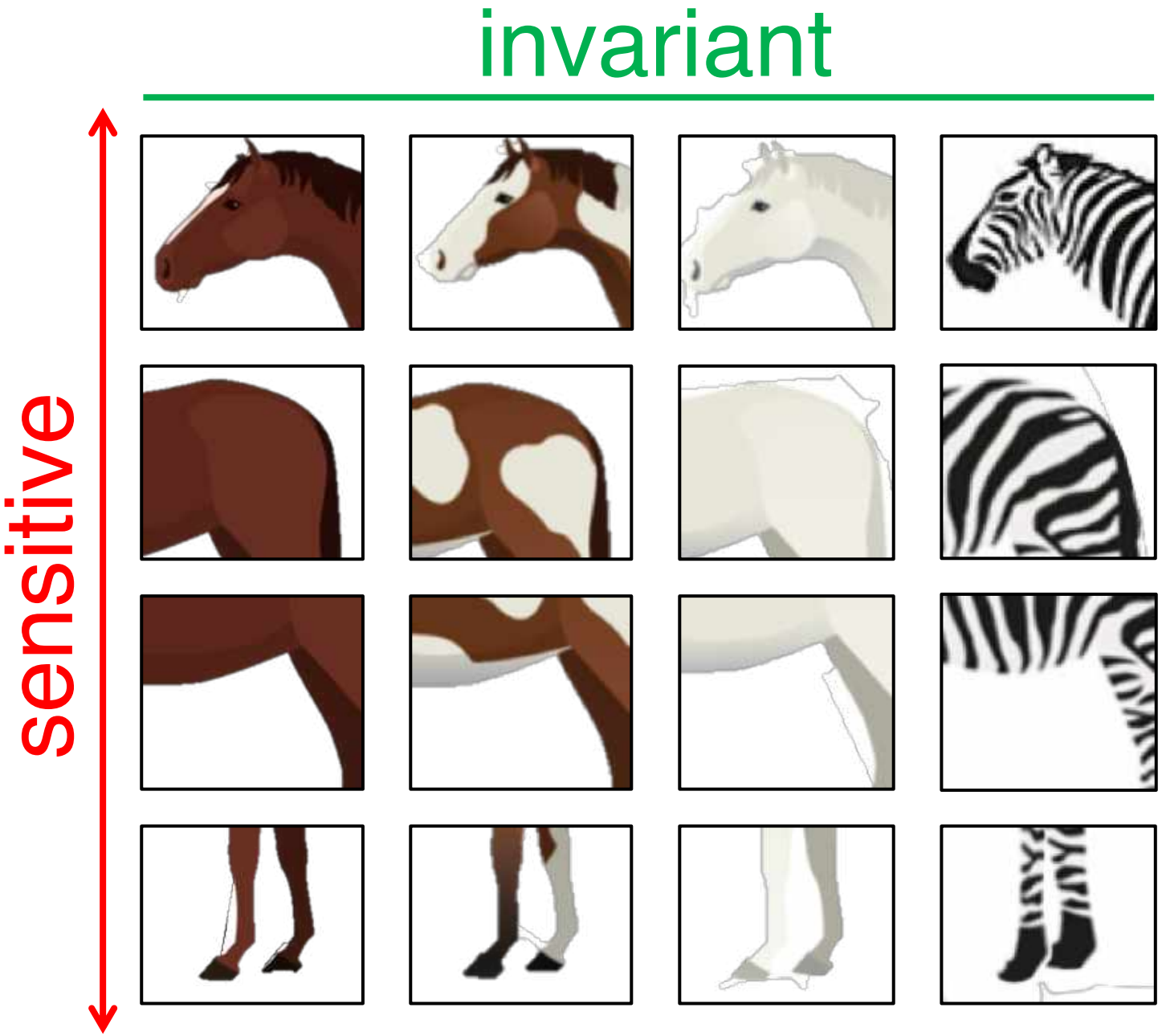
CycleGAN (Zhu et al., ICCV'17)
DiscoGAN (Kim et al., ICML'17)
DualGAN (Yi et al., ICCV'17)

Also used in
MUNIT (Huang et al., ECCV'18)
DRIT (Lee et al., ECCV'18)



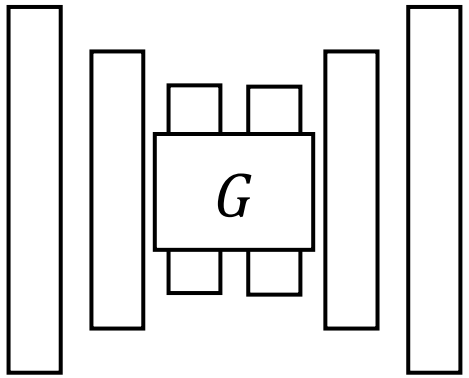



interchangeable
differentiated

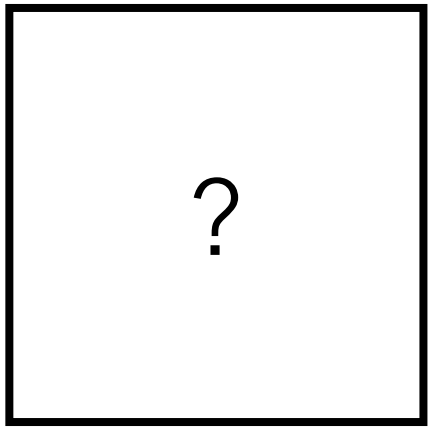

invariant
sensitive

What makes for a good output?

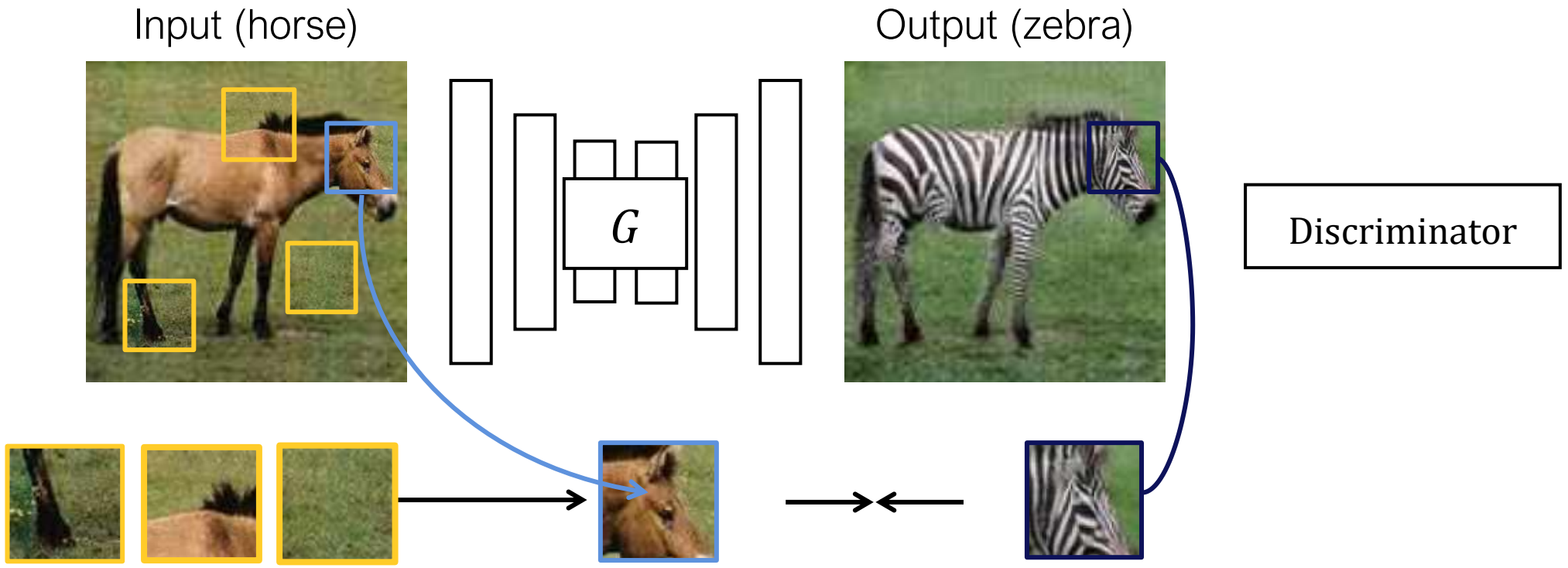
Input (horse)



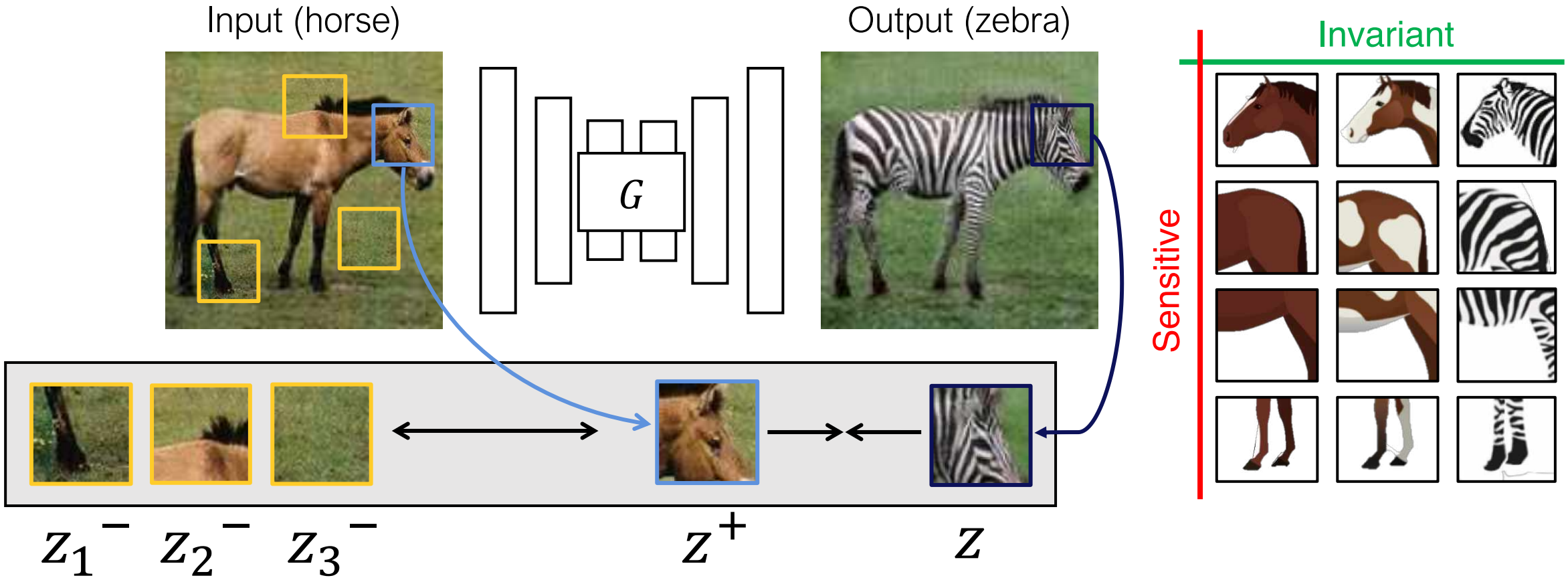
Output (zebra)



Retaining input **content**

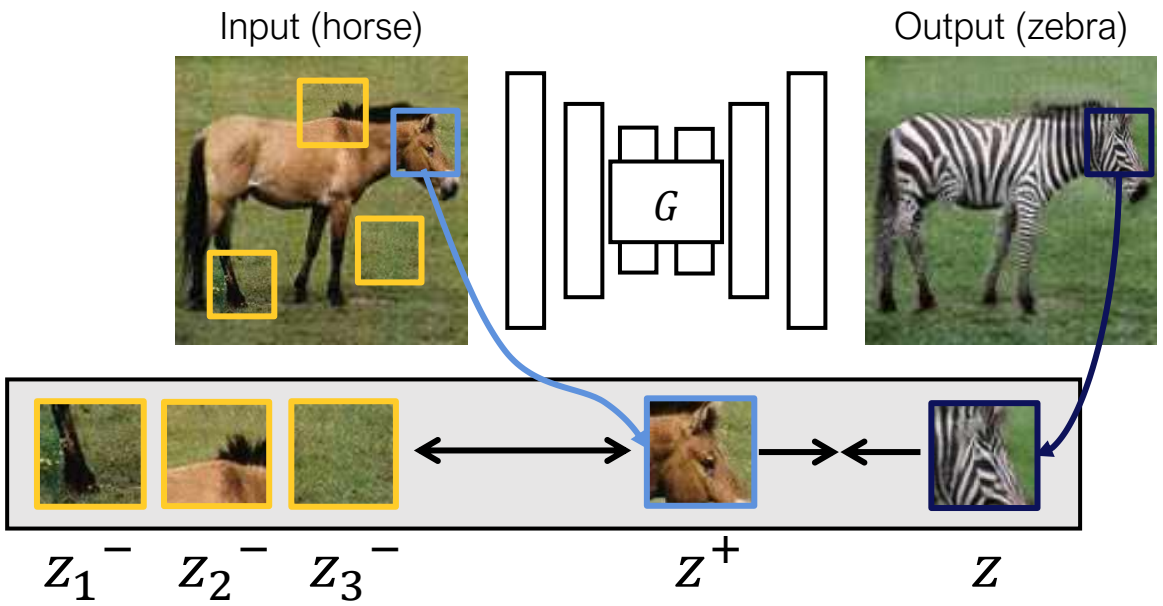


Retaining input **content**



Corresponding patches should have **high similarity**

Patch-based Contrastive Loss



$$\text{softmax}\begin{pmatrix} \uparrow z \cdot z^+ / \tau \\ \downarrow z \cdot z_1^- / \tau \\ \downarrow z \cdot z_2^- / \tau \\ \vdots \\ \downarrow z \cdot z_N^- / \tau \end{pmatrix} \rightarrow \begin{matrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{matrix}$$

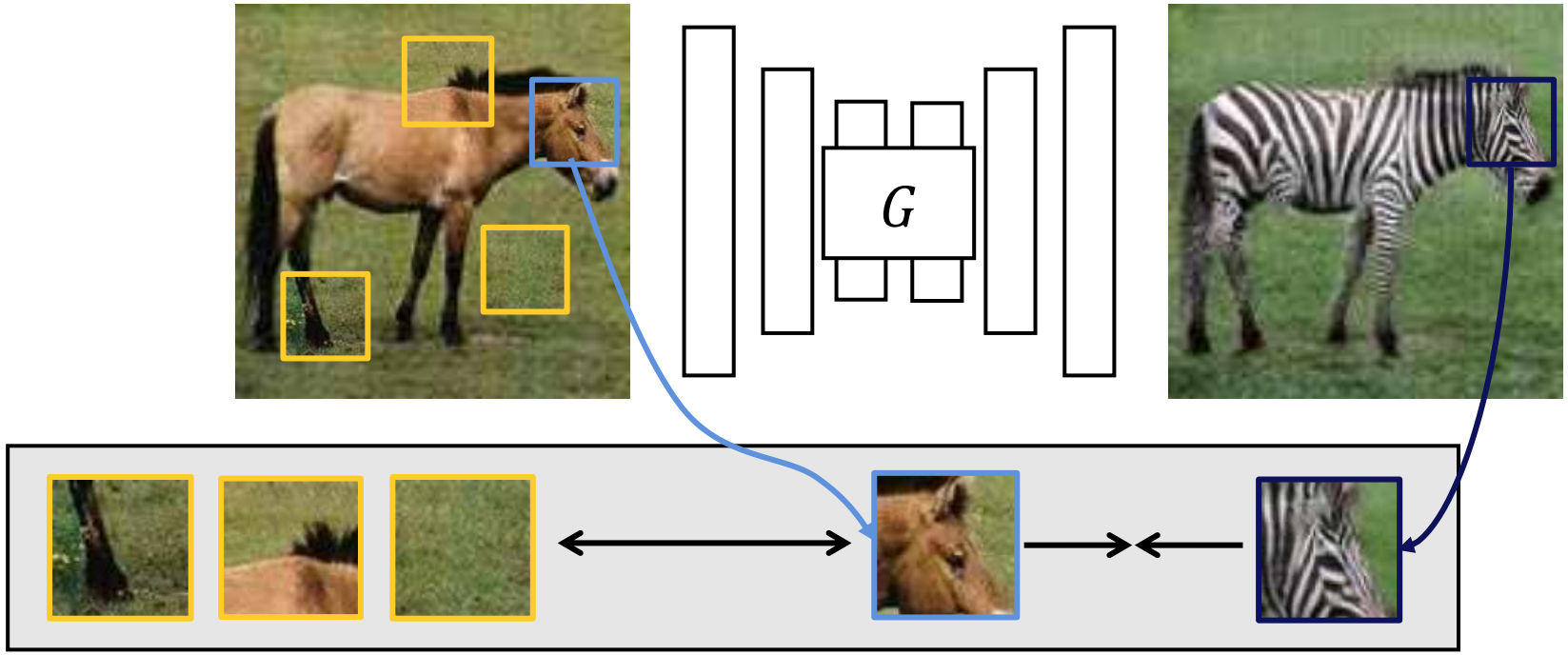
softmax (cosine similarities $/\tau$)

τ=0.07

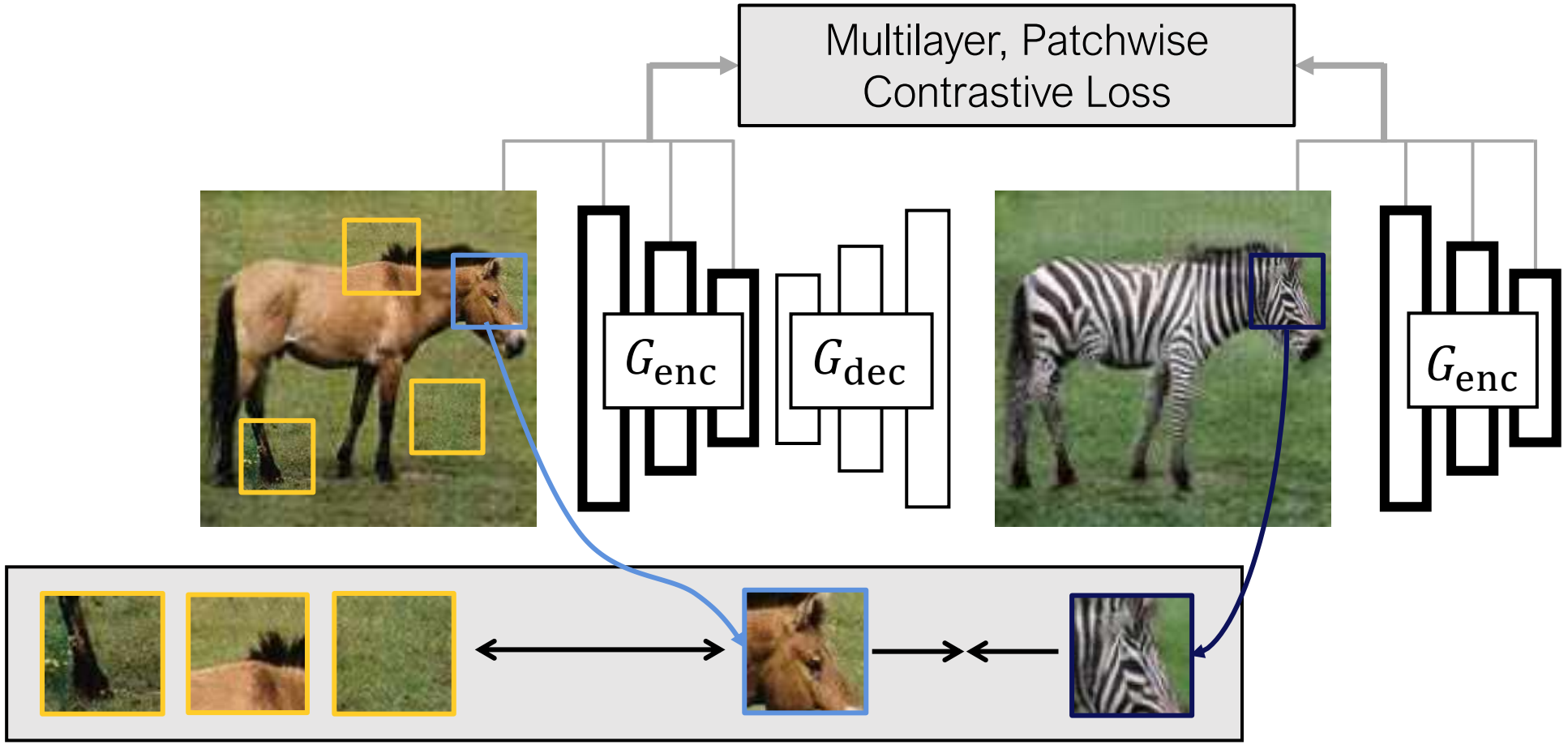
- InfoNCE loss (Gutmann et al., AISTATS18 , van den Oord et al., 2018) used in MoCo and SimCLR
- To produce positive pairs:
  - Handcrafted data augmentation (MoCo, SimCLR, etc.)
  - Input and synthesized image (ours)

MoCo: He et al., CVPR20, SimCLR: Chen et al., ICML20

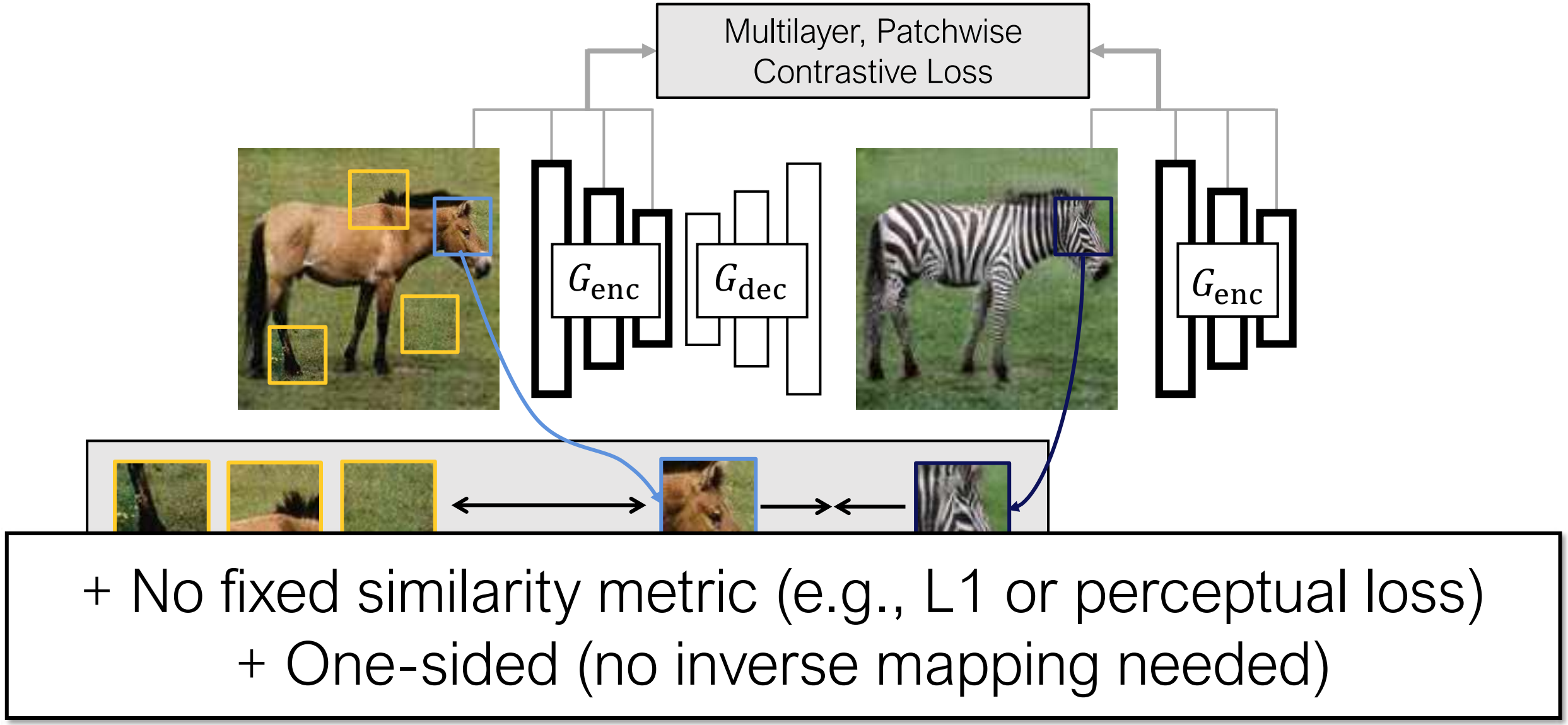
Patchwise contrastive loss



Patchwise contrastive loss

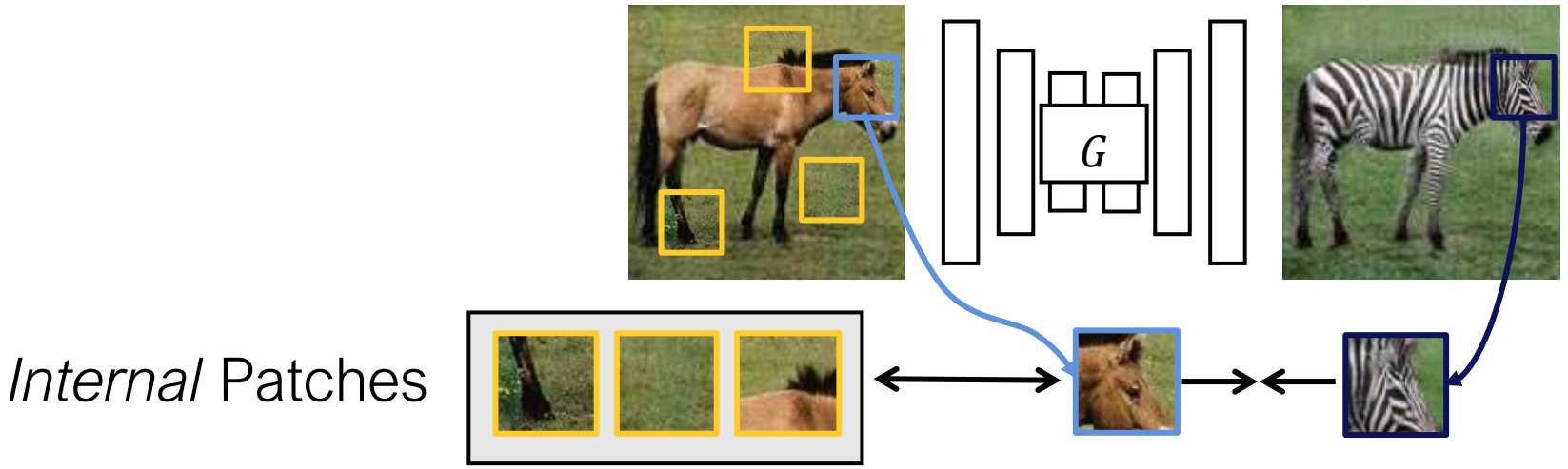


Patchwise contrastive loss

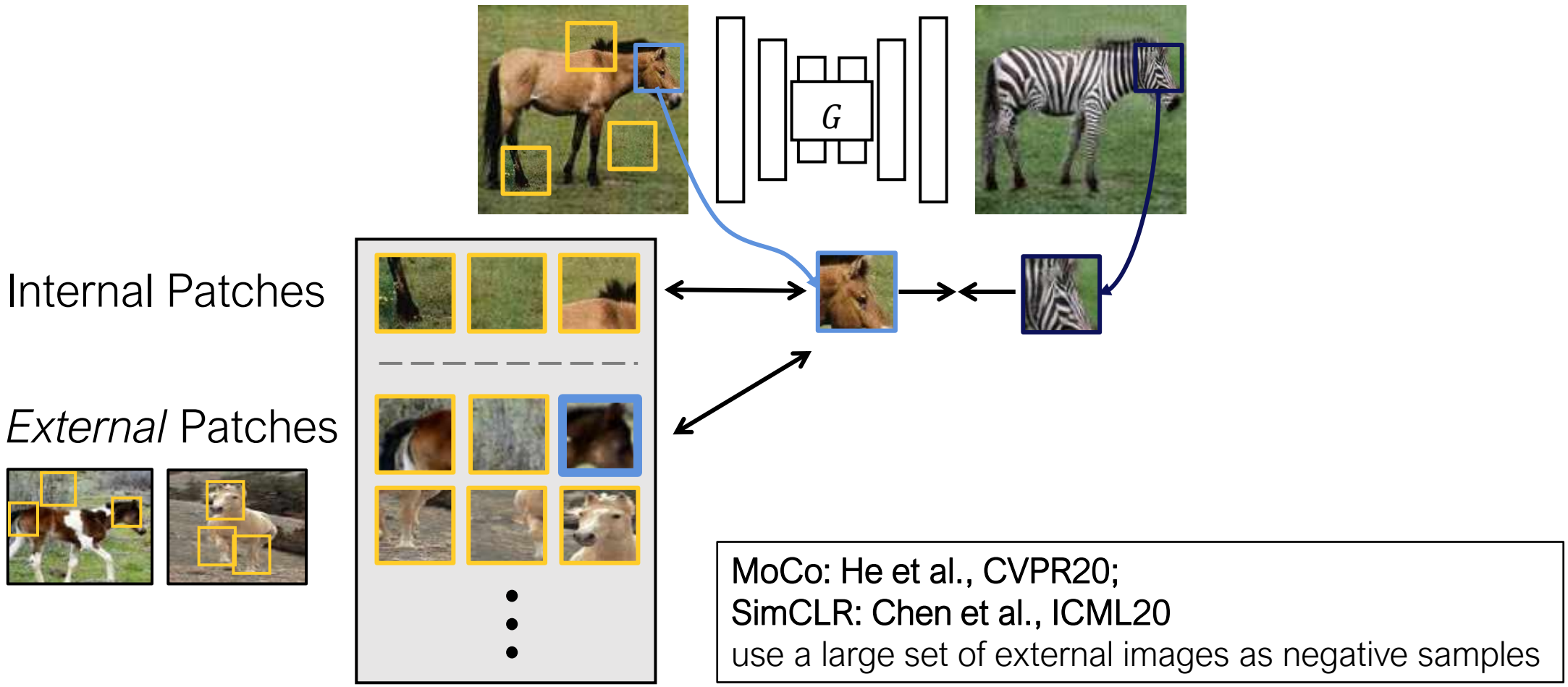


+ No fixed similarity metric (e.g., L1 or perceptual loss)
+ One-sided (no inverse mapping needed)

Internal vs External Patches



Internal vs External Patches

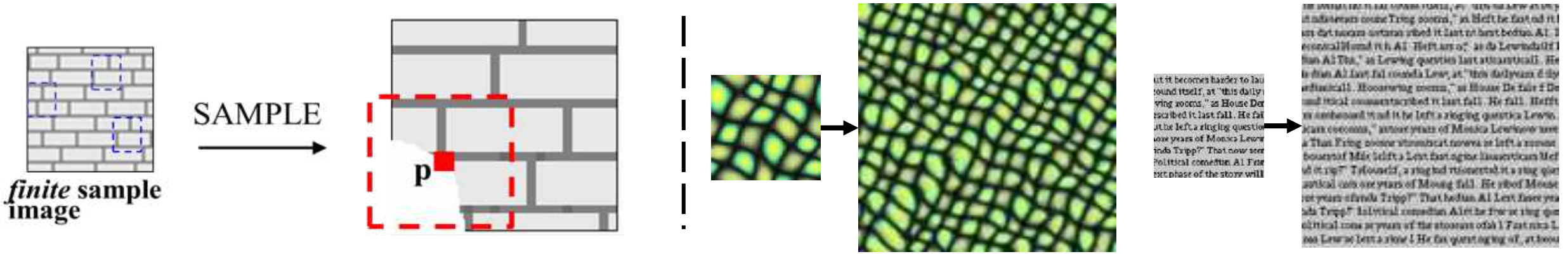


MoCo: He et al., CVPR20;
SimCLR: Chen et al., ICML20
use a large set of external images as negative samples

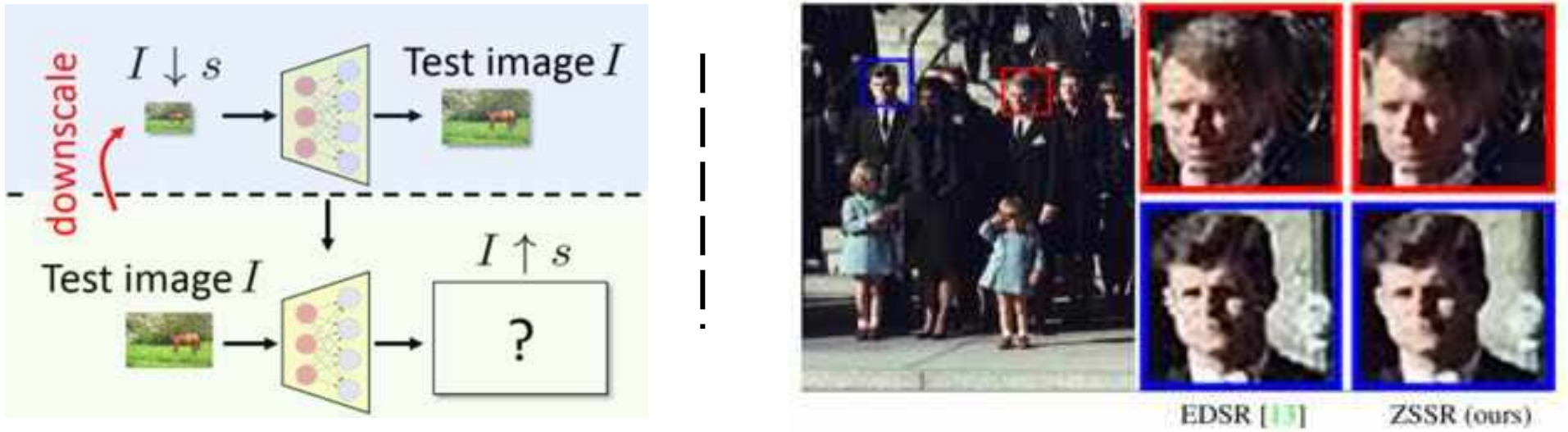
External patches make things worse

Power of *Internal* patches

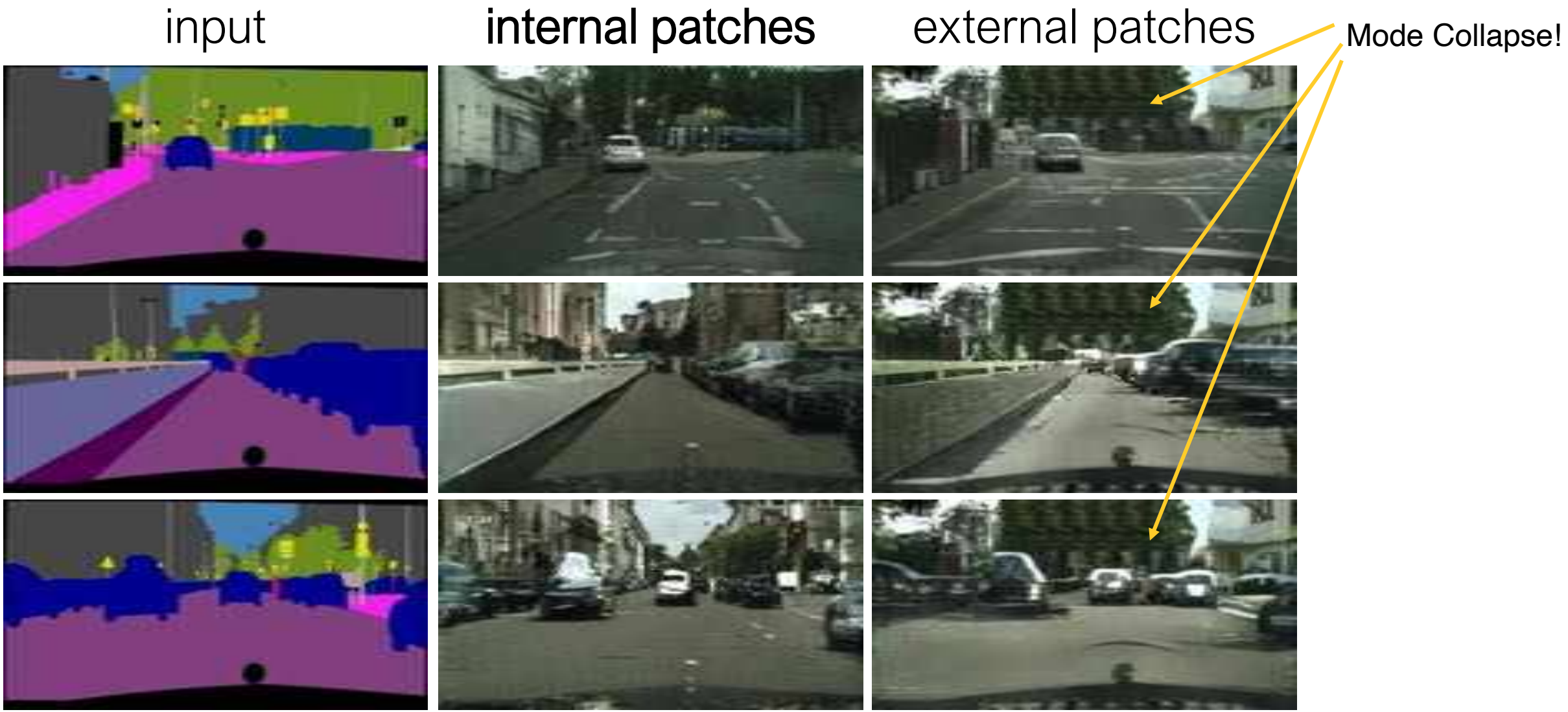
Texture Synthesis by Non-parametric Sampling (Efros & Leung, ICCV'99, Efros & Freeman, SIGGRAPH'01)



"Zero-Shot" Super-resolution using Deep Internal Learning (Shocher, Cohen & Irani CVPR'18)

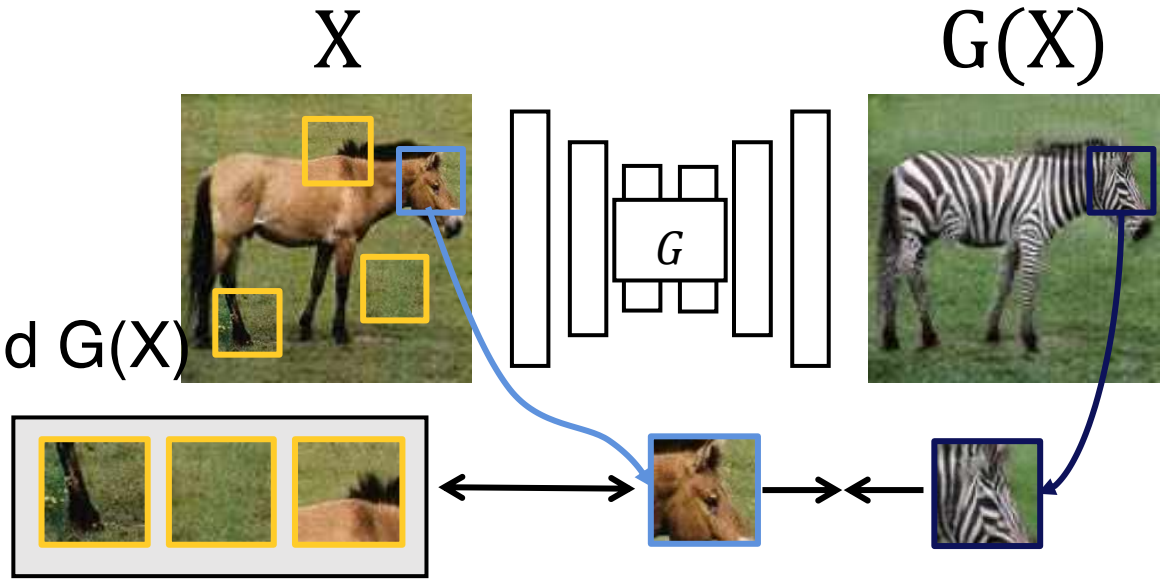


Internal vs External Patches



Identity Loss Regularization

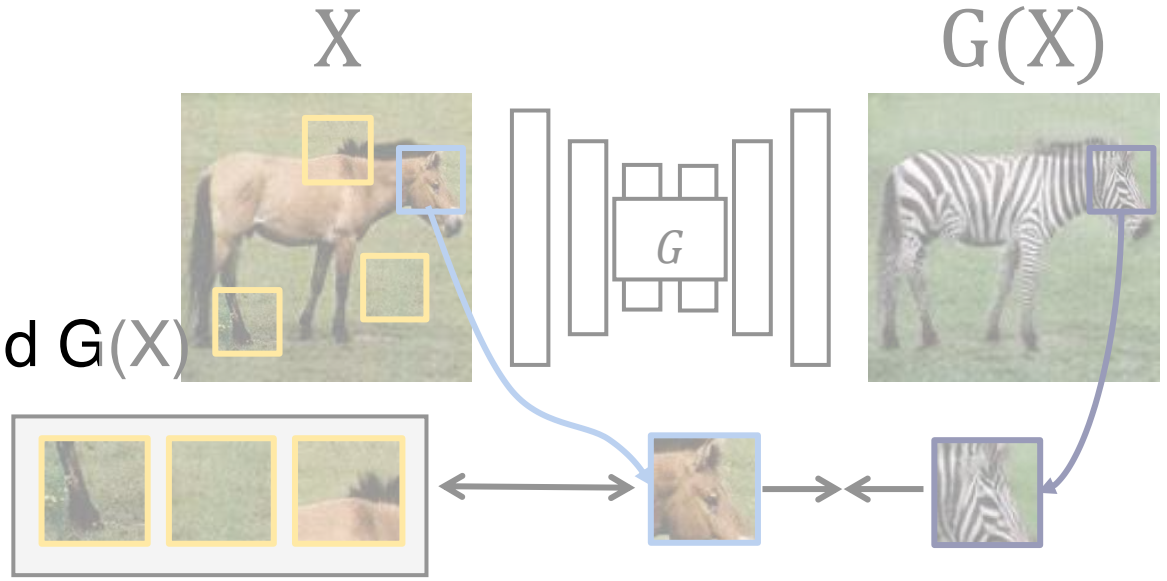
Normally,
Contrastive Loss between X and G(X)



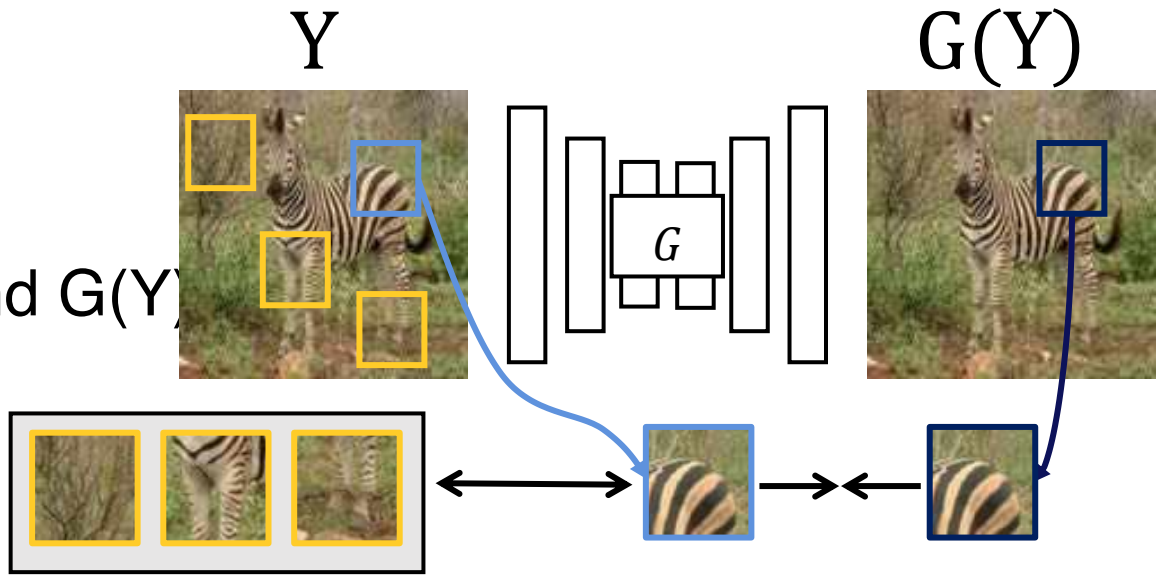
DTN (Taigman et al., ICLR'17), CycleGAN (Zhu et al., ICCV'17)

Identity Loss Regularization

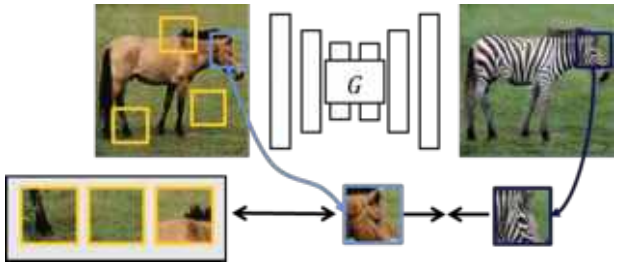
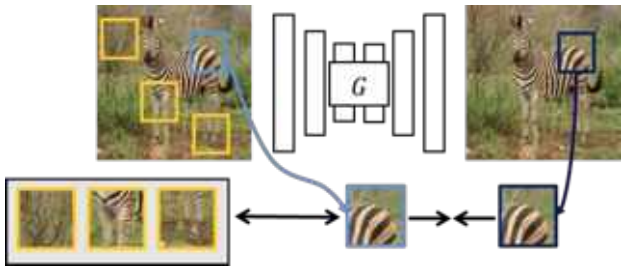
Normally,
Contrastive Loss between X and G(X)



Identity loss regularization
Contrastive Loss between Y and G(Y)

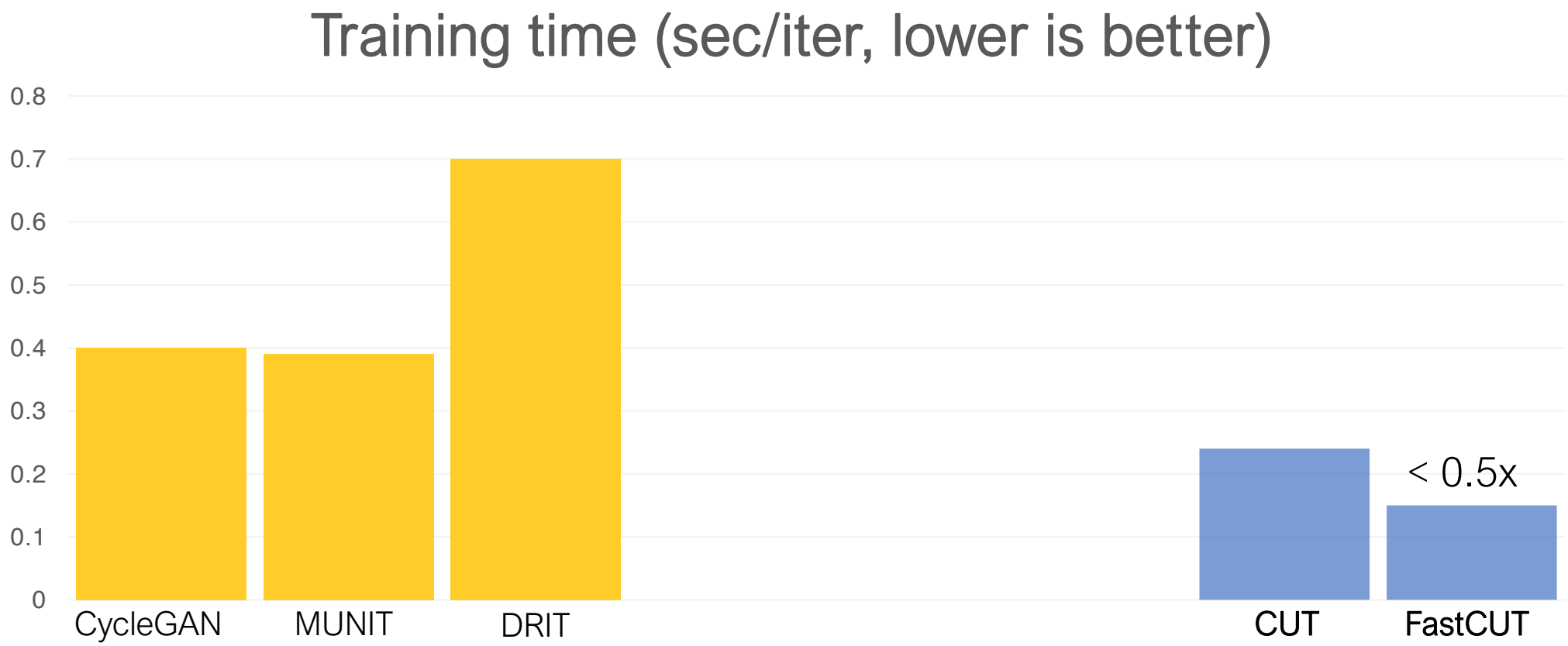


DTN (Taigman et al., ICLR'17), CycleGAN (Zhu et al., ICCV'17)

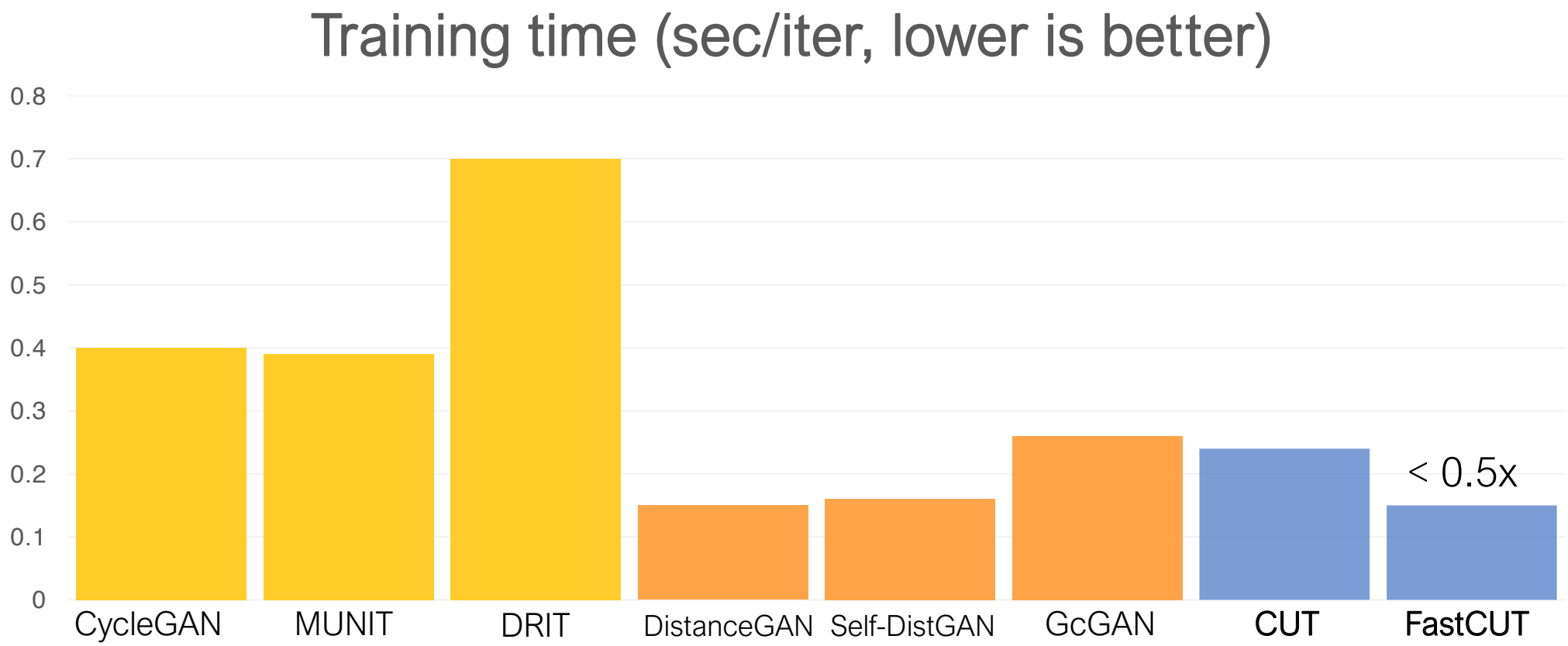
	CUT Contrastive Unpaired Translation	FastCUT
Contrastive Loss  	$\lambda = 1$	$\lambda = 10$
Identity Loss Regularization  	✓	✗
	Flexible, Faster than CycleGAN	Conservative, Even Faster than CUT

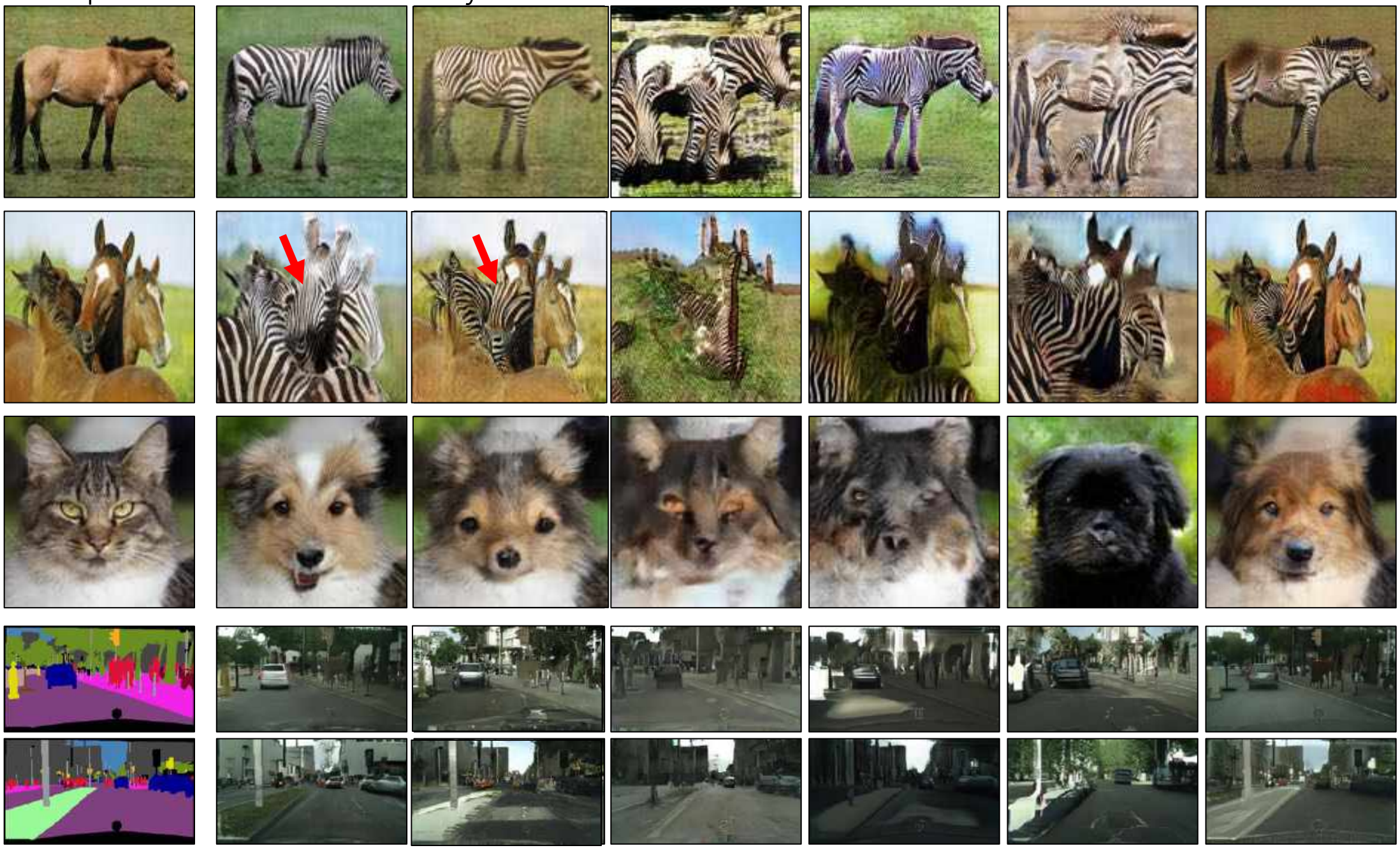
Lighter Footprint

Lighter Footprint



Lighter Footprint




Input
CUT
FastCUT
MUNIT
DRIT
DistanceGAN
GcGAN

Dealing with Dataset Bias

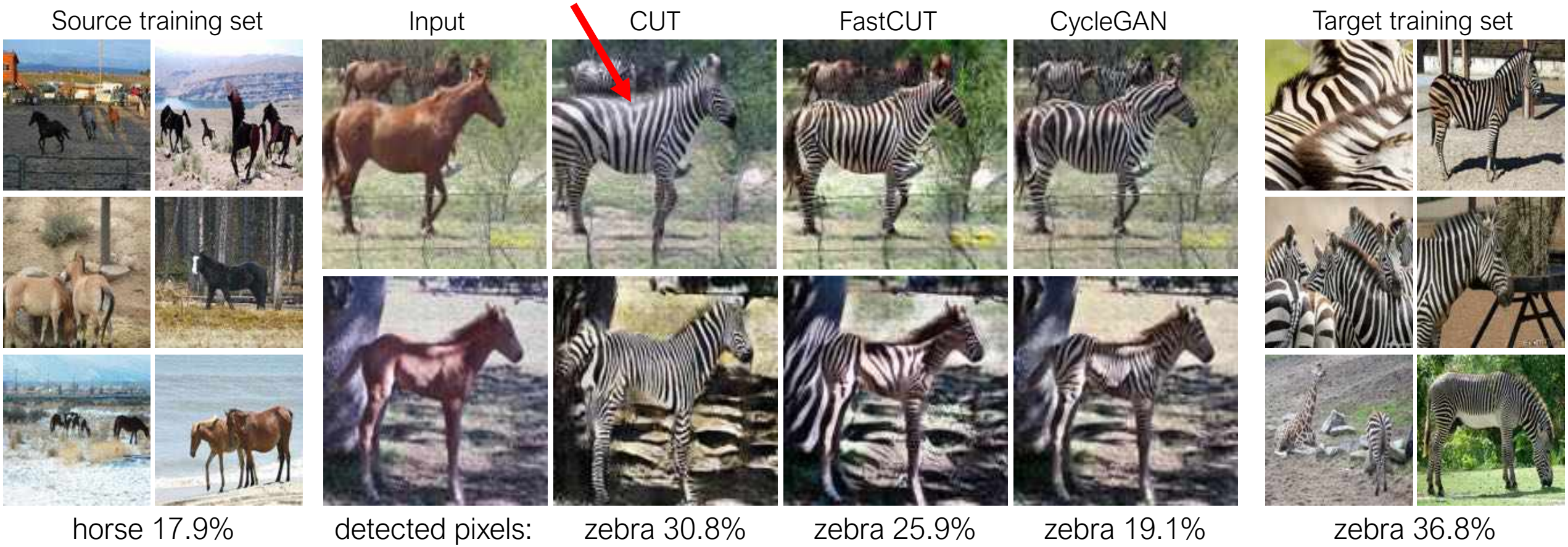
Source training set

horse 17.9%

Target training set

zebra 36.8%

Dealing with Dataset Bias

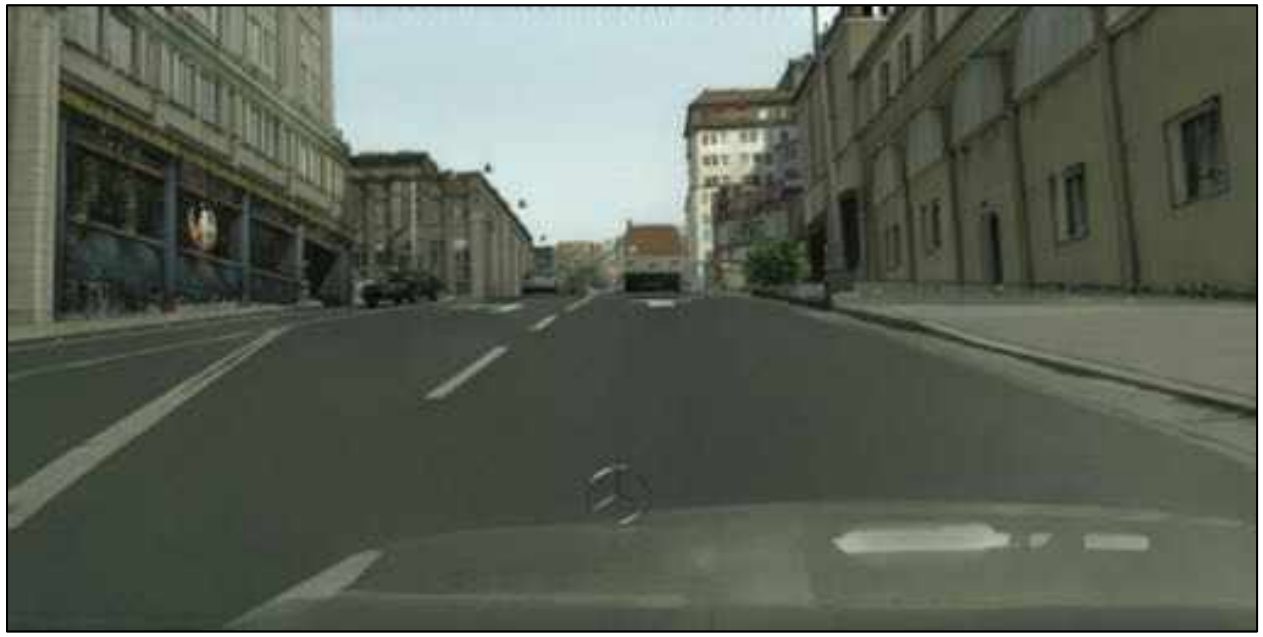


Cat

Yosemite Summer -

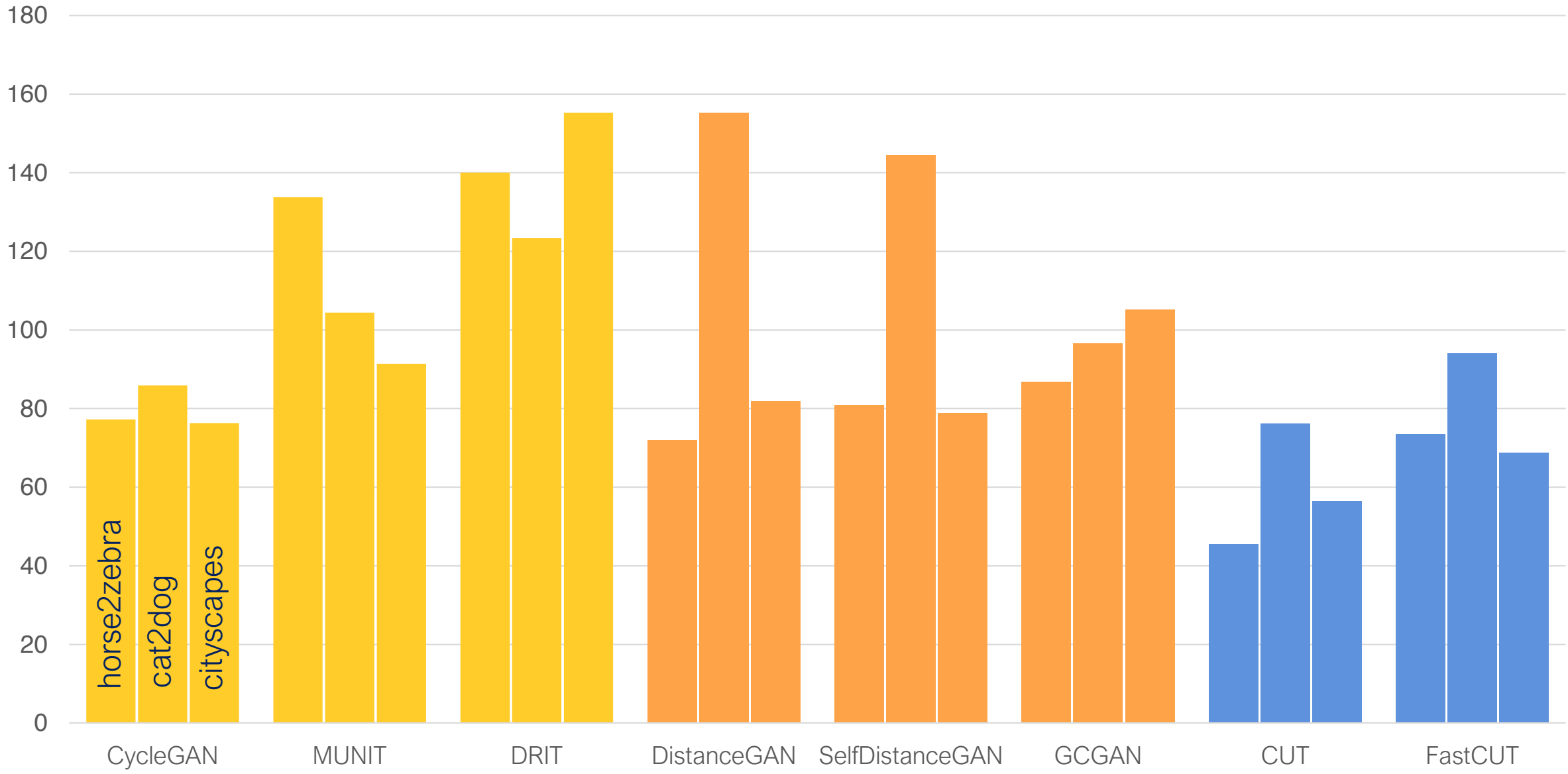
Apple ·

Paris ·

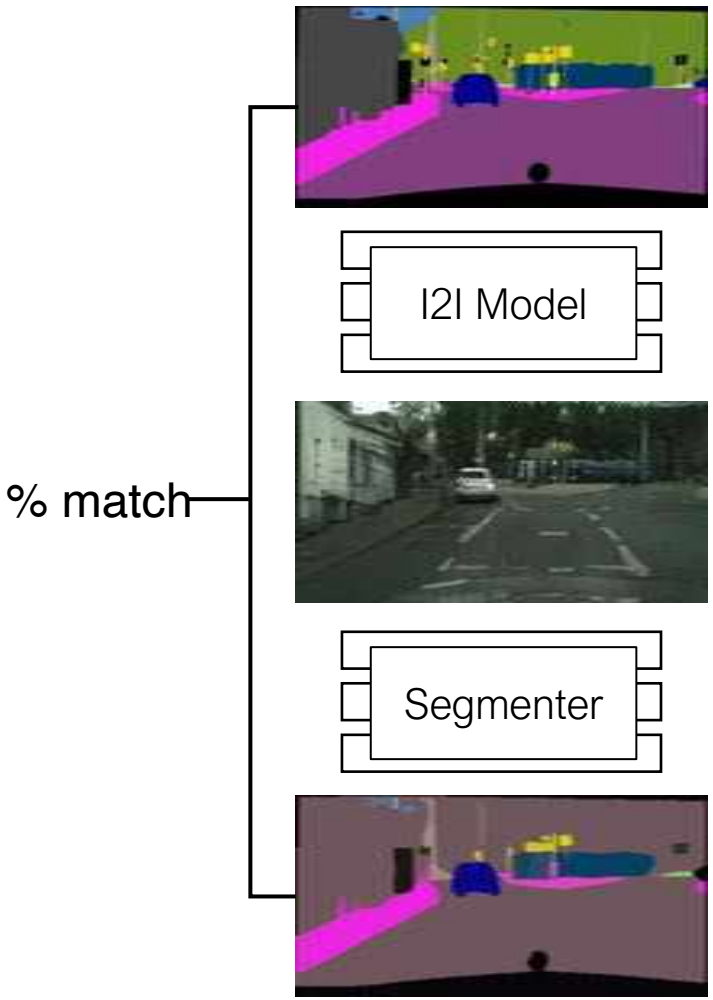


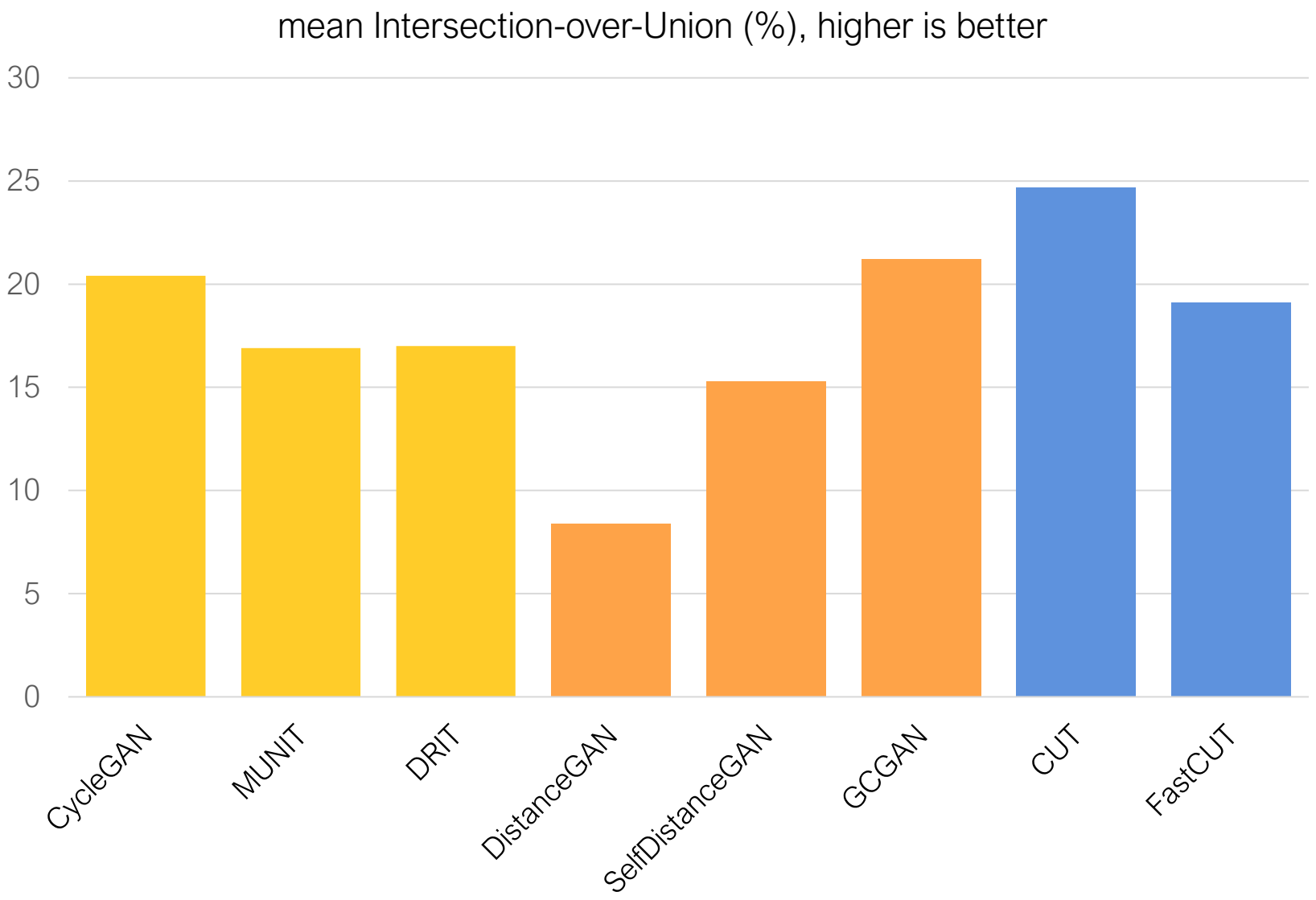
GTA

FID evaluating the realism of output images (lower is better)

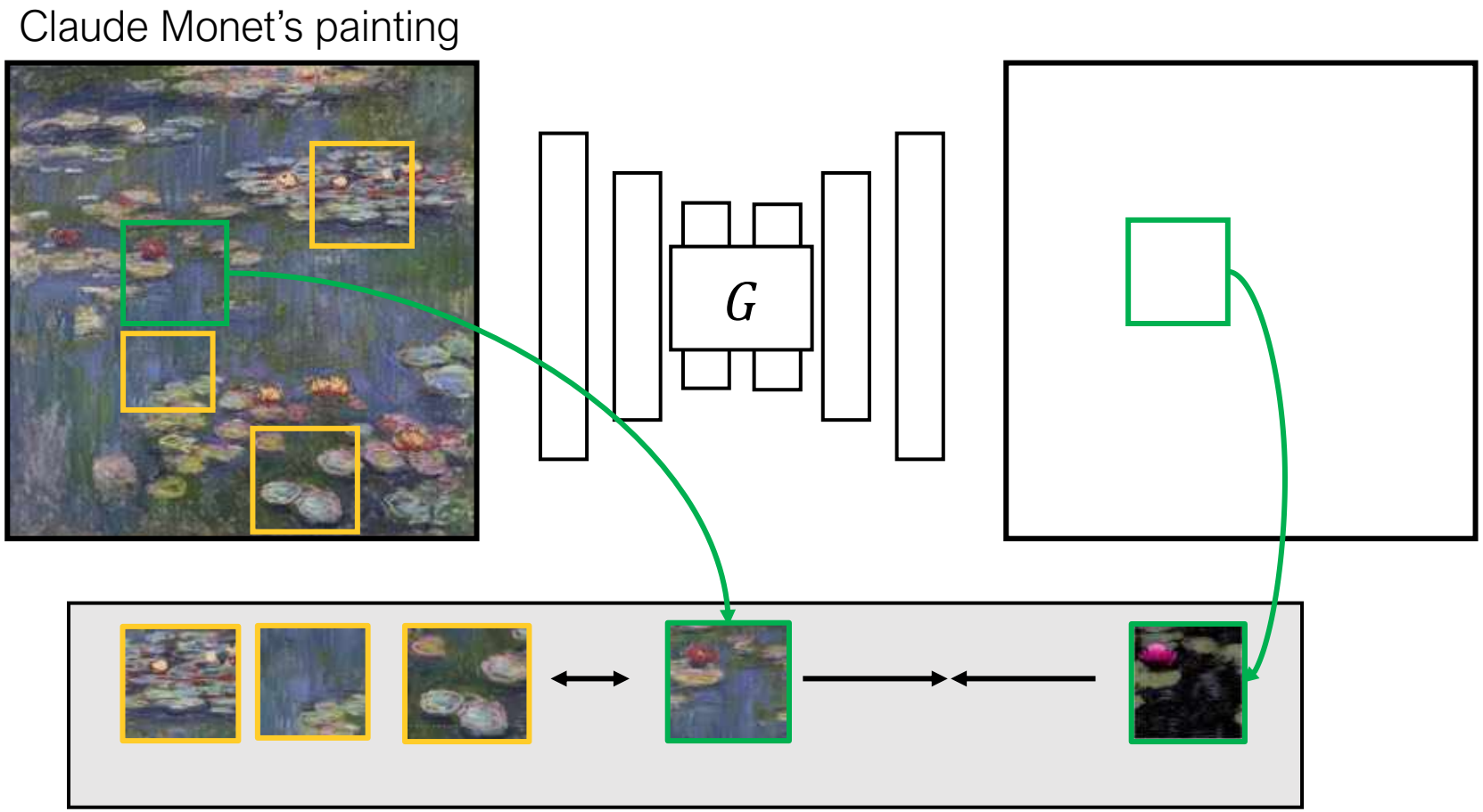


Segmentation Score evaluating correspondences





Single Image Translation



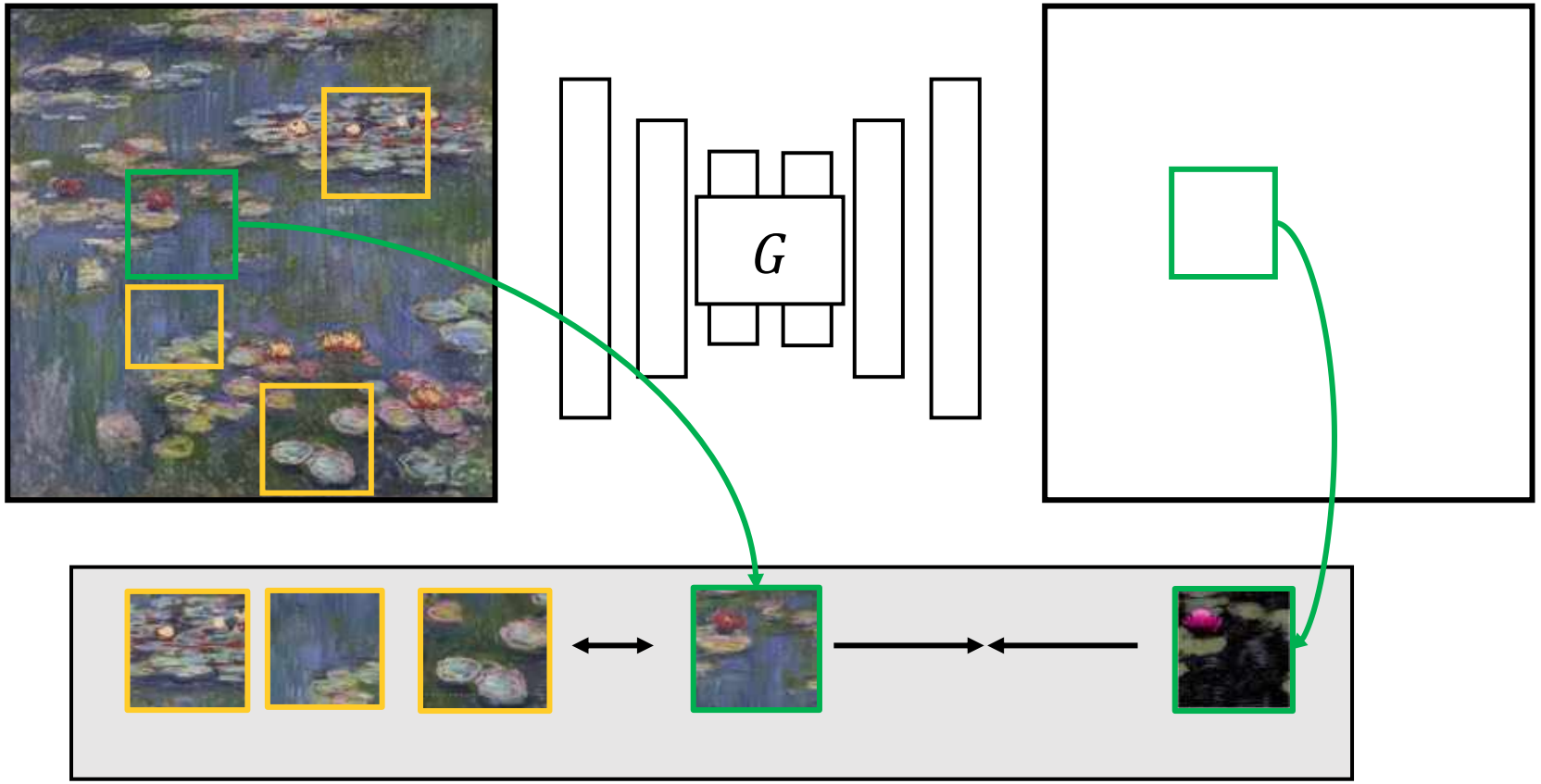
Internal contrastive loss is well-suited for single image translation.

Also see InGAN (Shocher et al., ICCV'19), SinGAN (Shaham et al., ICCV'19)

Single Image Translation

Reference photo

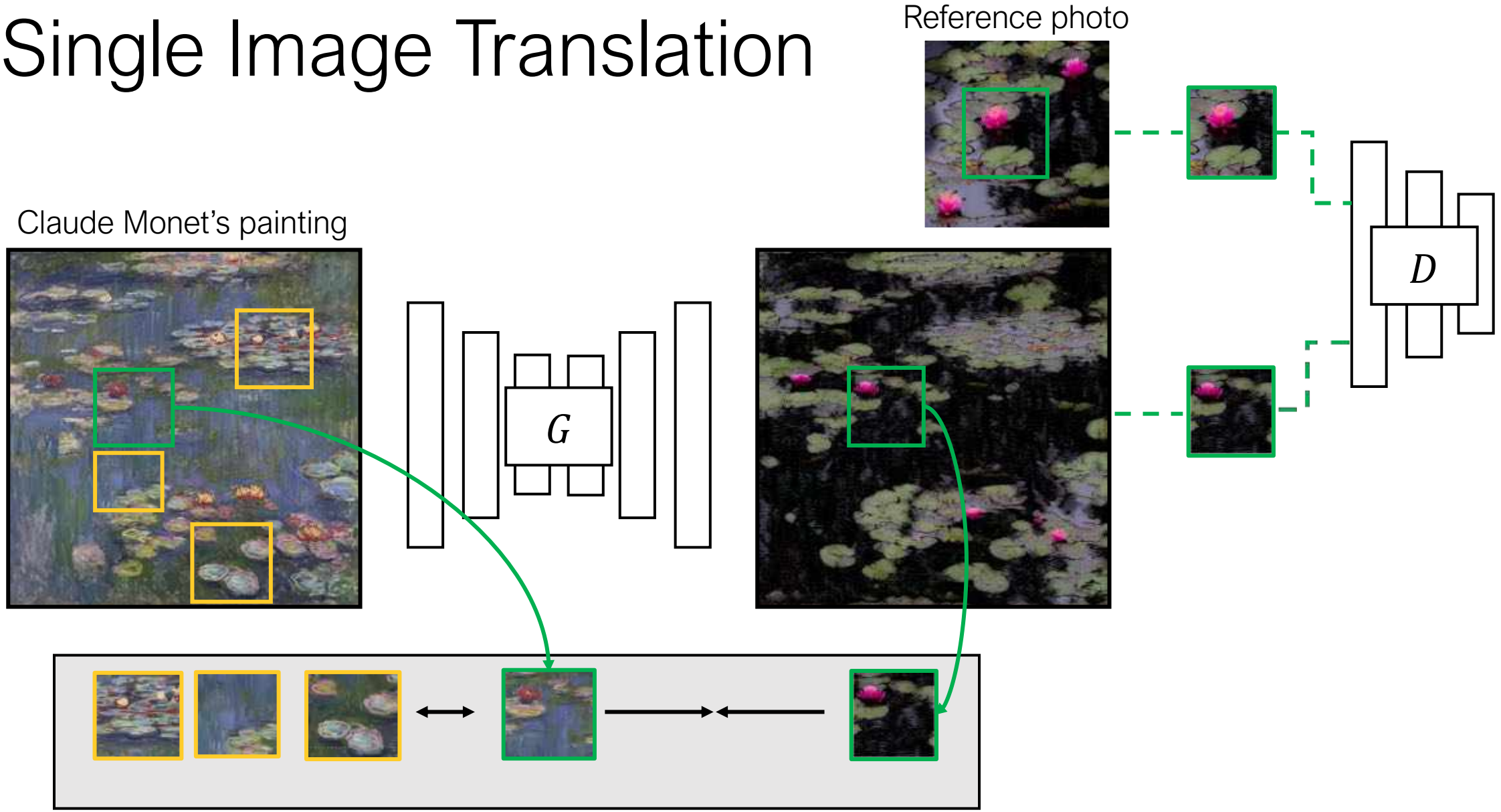
Claude Monet's painting



Internal contrastive loss is well-suited for single image translation.

Also see InGAN (Shocher et al., ICCV'19), SinGAN (Shaham et al., ICCV'19)

Single Image Translation



Internal contrastive loss is well-suited for single image translation.

Also see InGAN (Shocher et al., ICCV'19), SinGAN (Shaham et al., ICCV'19)

Painting

Reference Photo

Painting

Reference Photo

Painting

Gatys et al. CVPR'16

Reference Photo

Painting

STROTSS (Kolkin et al., CVPR'19)

Reference Photo

Painting

WCT^2 (Yoo et al., ICCV'19)

Reference Photo

Painting

Our translation result

Reference Photo

Painting

CycleGAN

Painting

Reference Photo

Painting

Gatys et al. CVPR'16

Reference Photo

Painting

STROTSS (Kolkin et al., CVPR'19)

Reference Photo

Painting

WCT^2 (Yoo et al., ICCV'19)

Reference Photo

Painting

Ours

Reference Photo

Painting

CycleGAN

Reference Photo

Painting

Our translation result

Reference Photo

Painting

Our translation result

Reference Photo

Painting

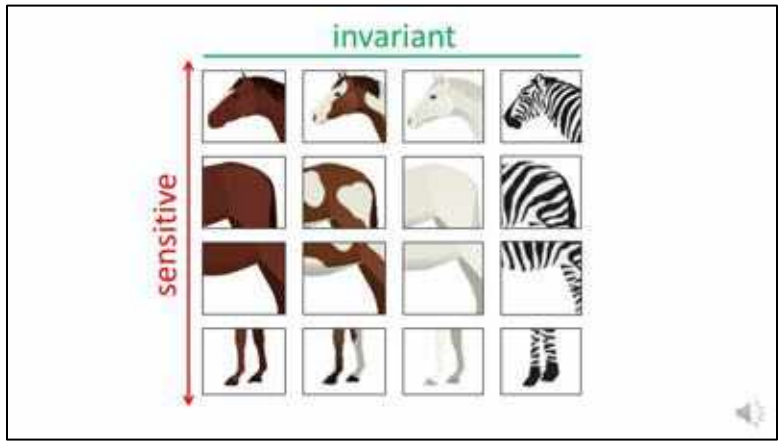
Our translation result

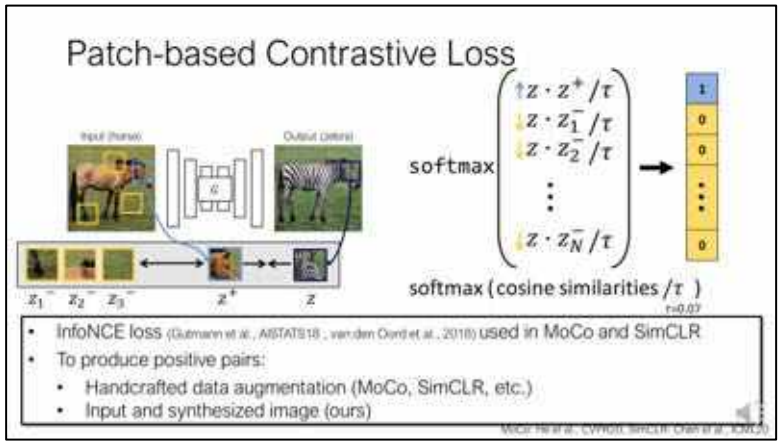
Reference Photo

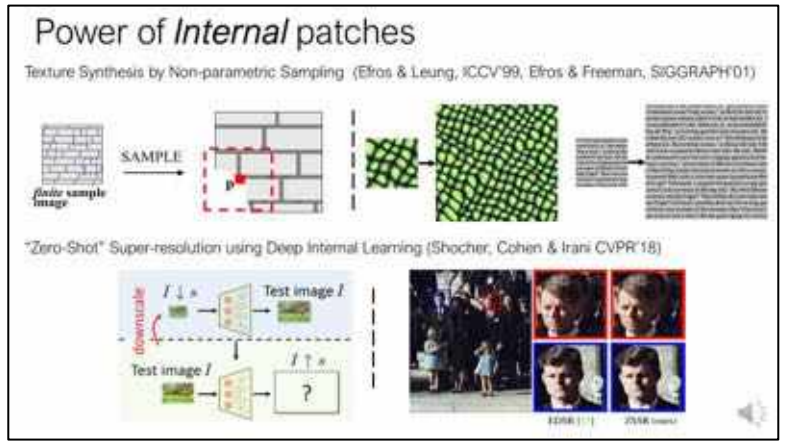
Painting

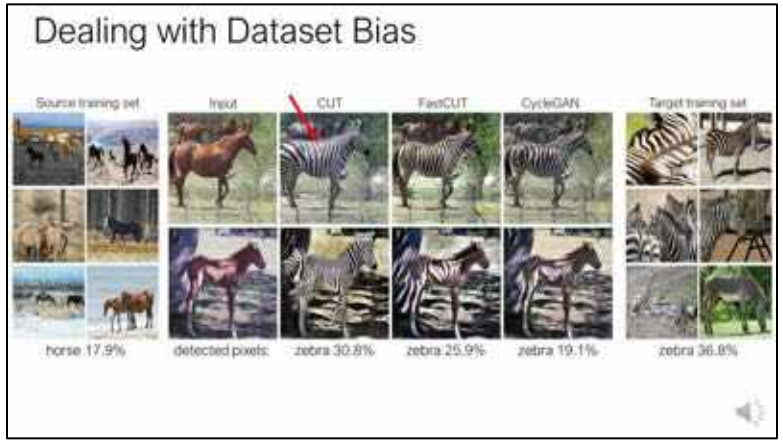
Our translation result

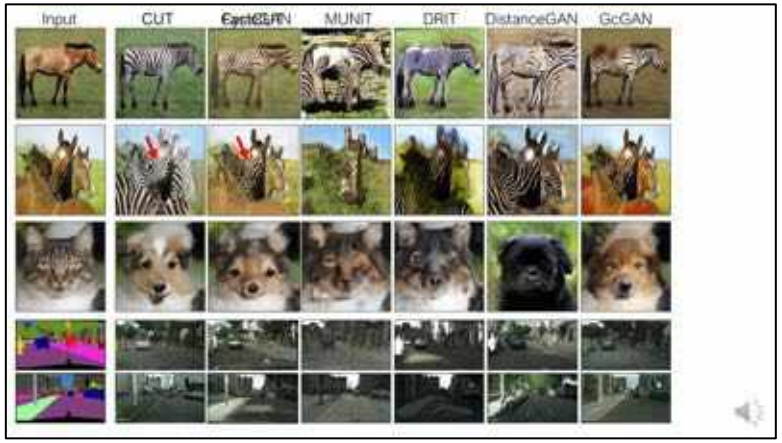
Questions or Comments?

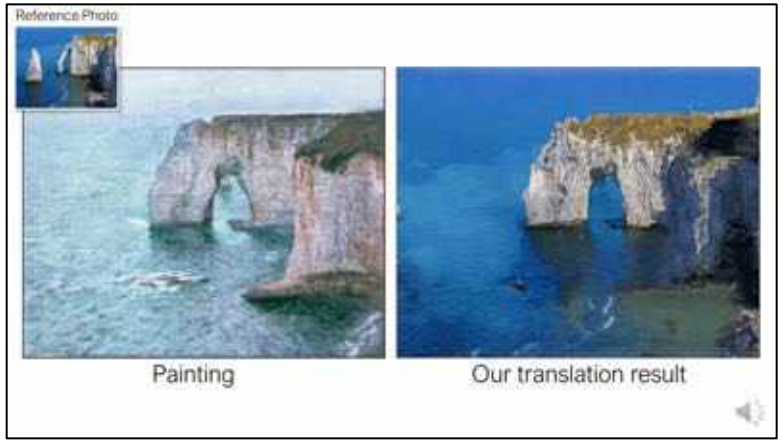








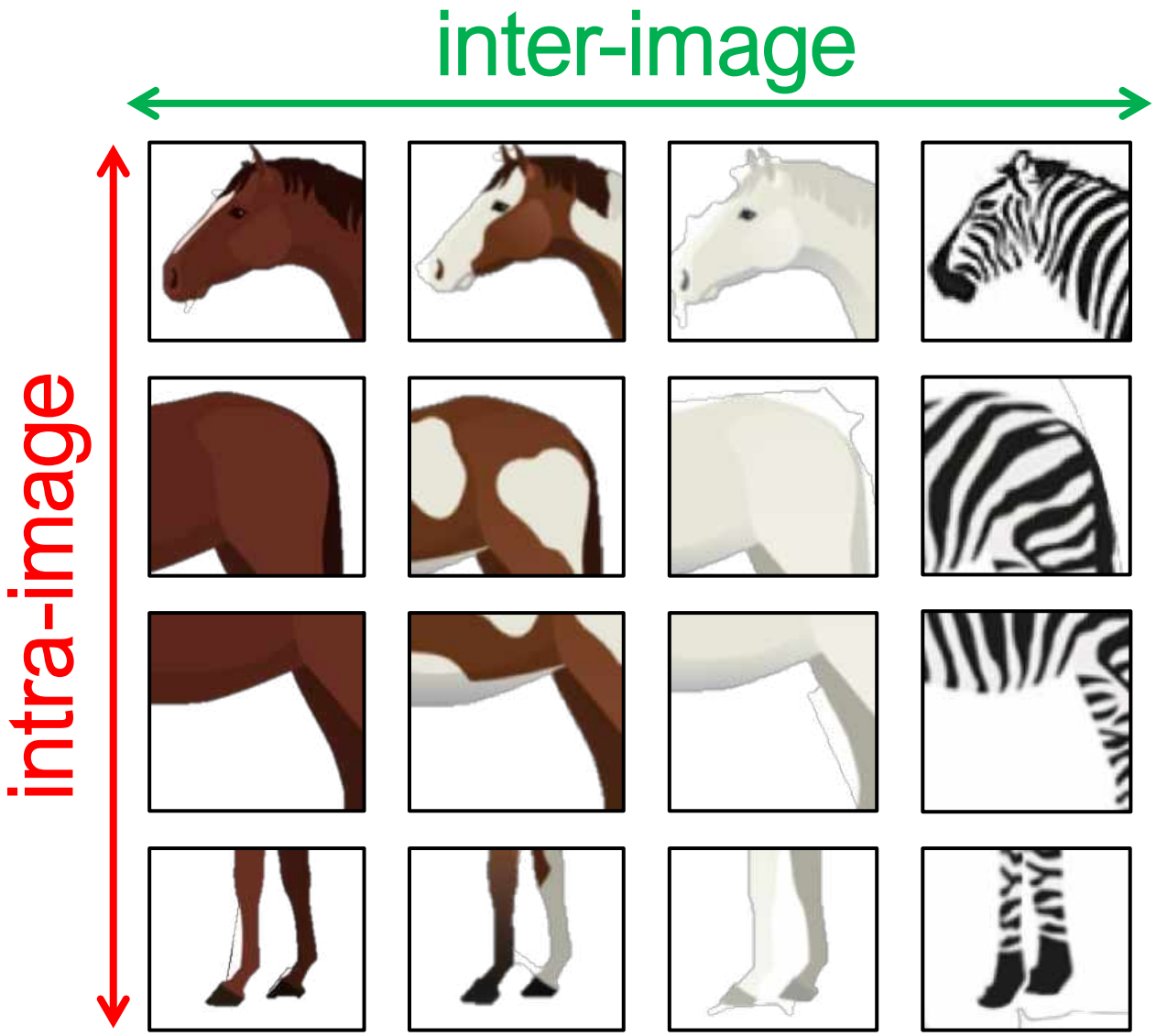




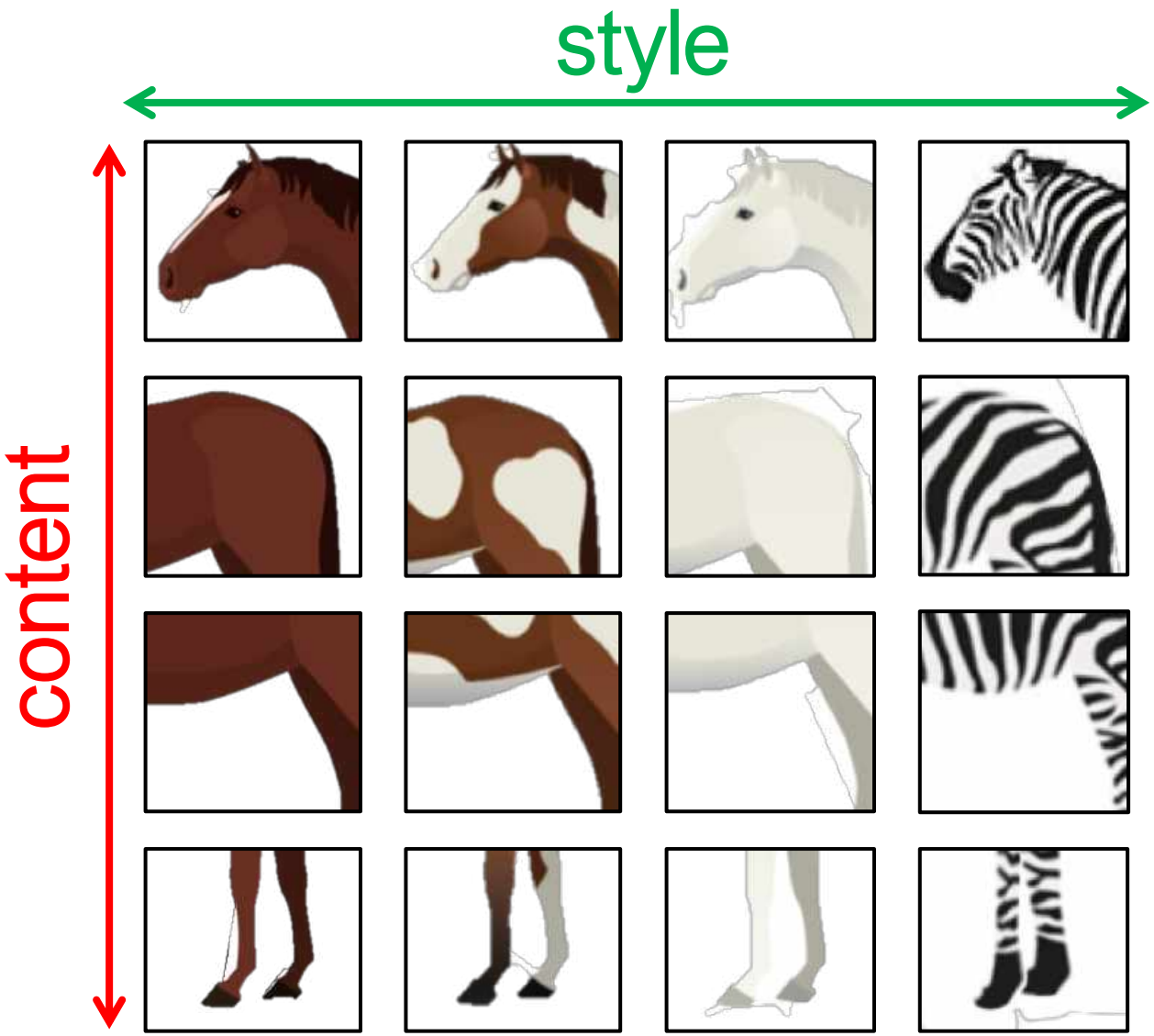
Swapping Autoencoder for Deep Image Manipulation

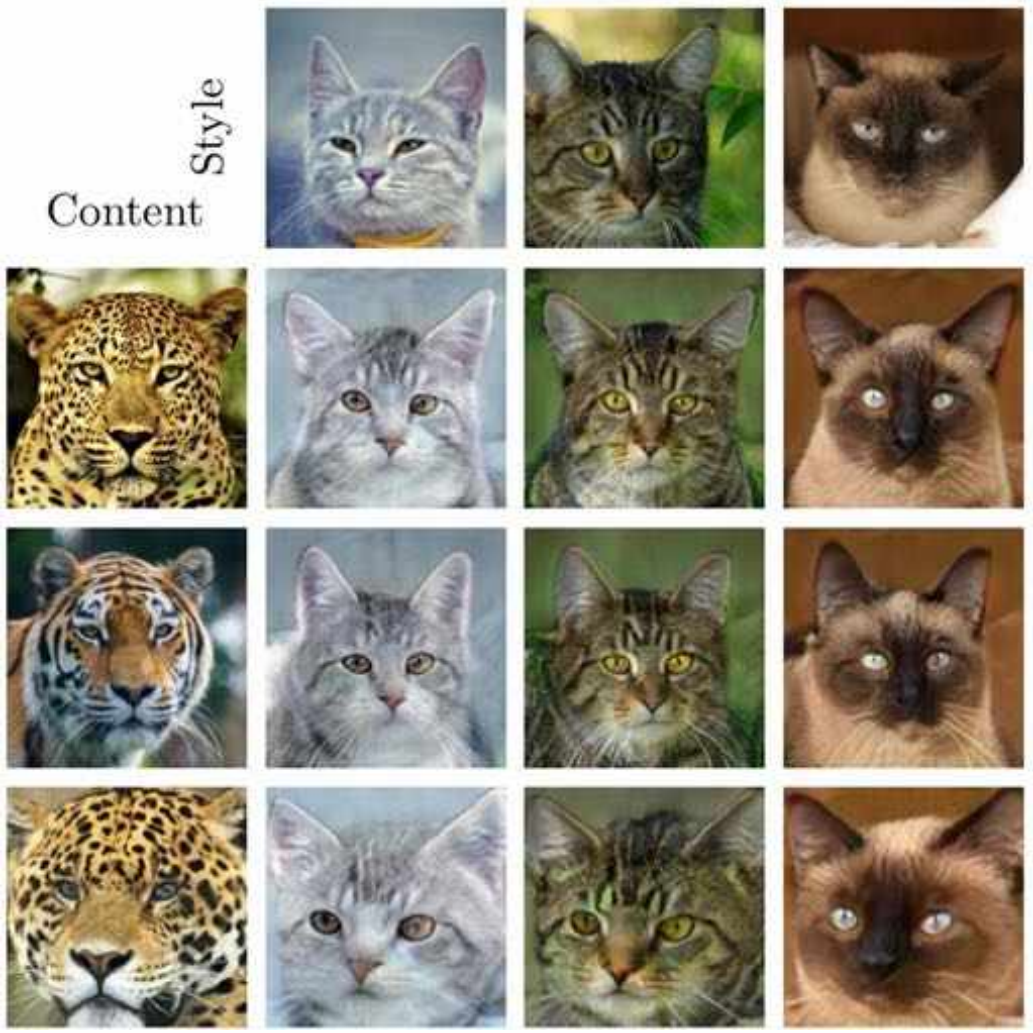
Taesung Park[1], Jun-Yan Zhu[2], Oliver Wang[2], Jingwan Lu[2], Eli Shechtman[2], Alexei Efros[1,2], Richard Zhang[2]

[1]UC Berkeley, [2]Adobe Research



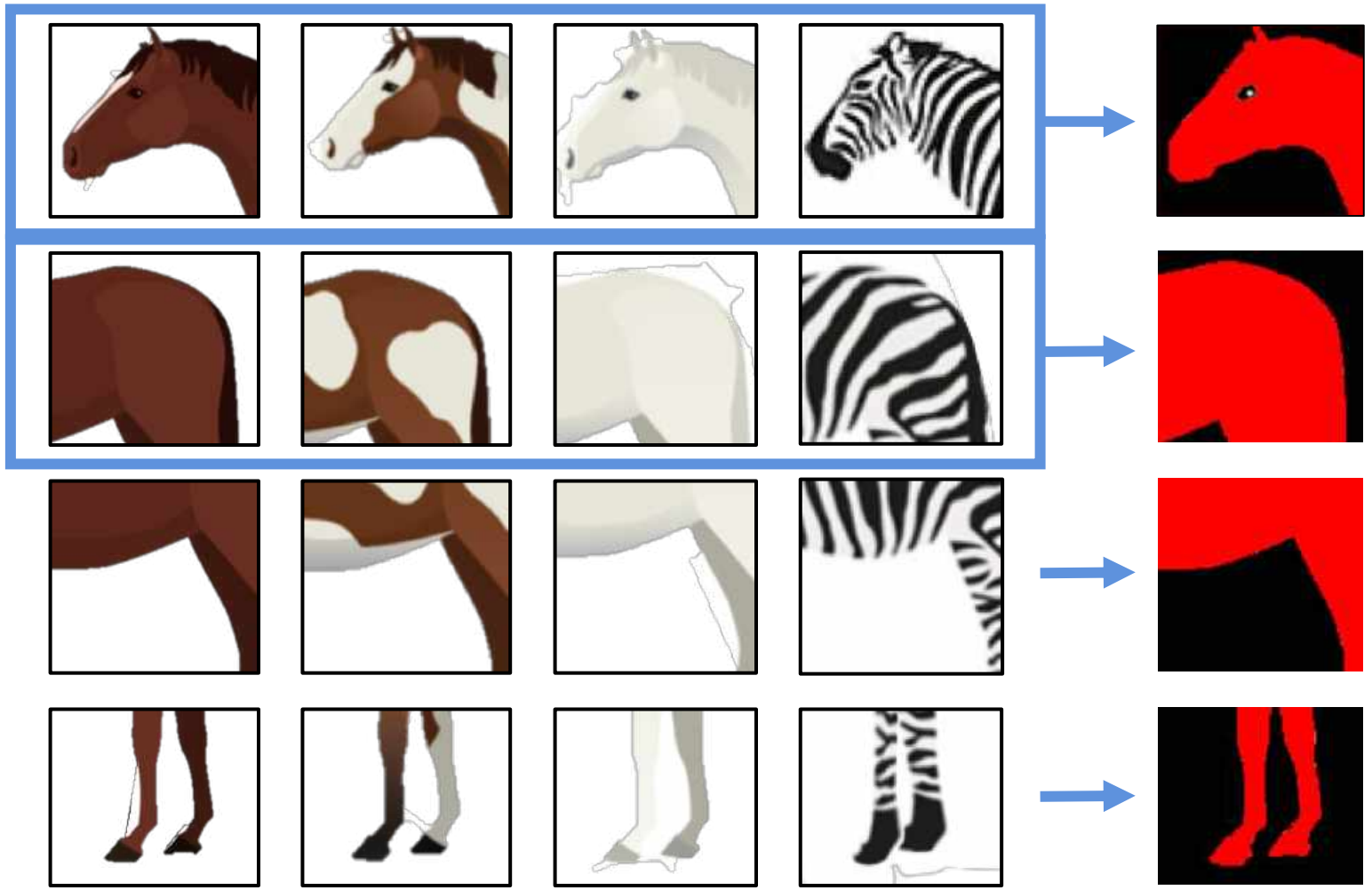
Disentanglement?



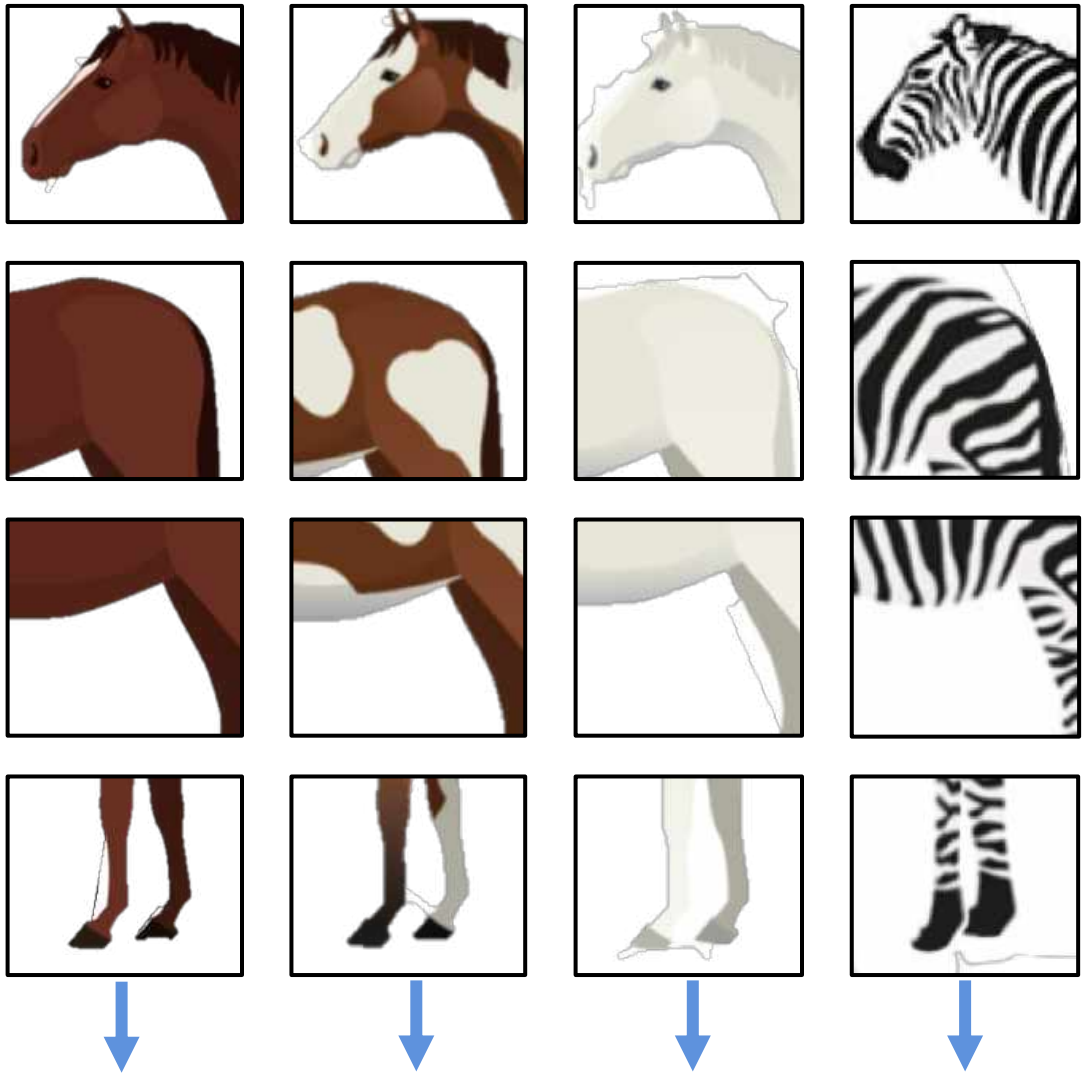


MUNIT (Huang, Liu, Belongie, Kautz, ECCV'18)

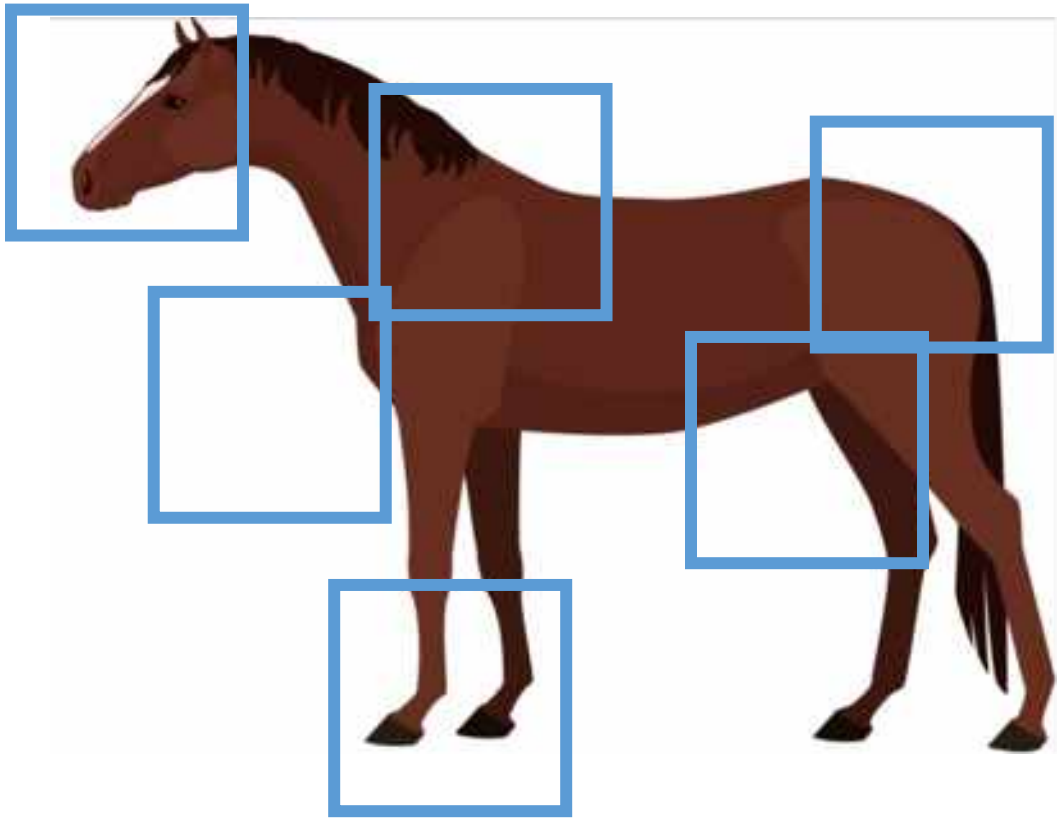
Structure for each row

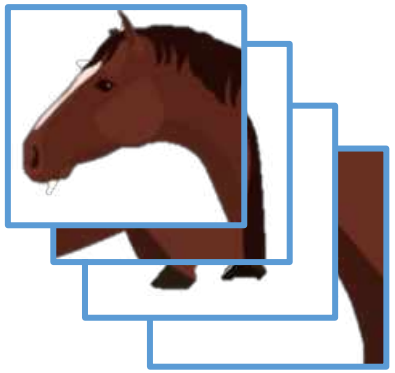


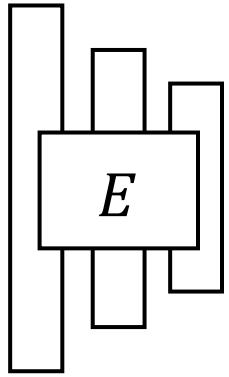
Style for each column



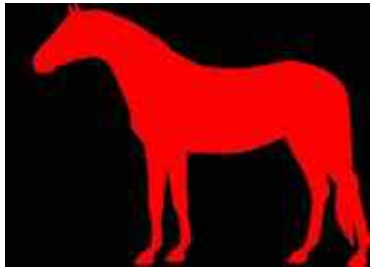
Extracting style and structure from an image



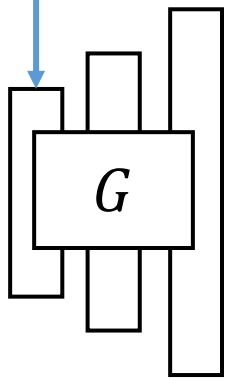


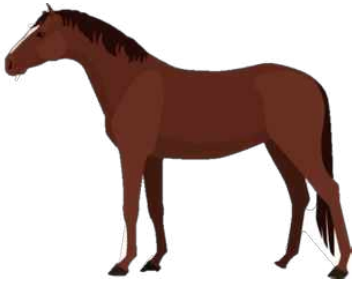


style code

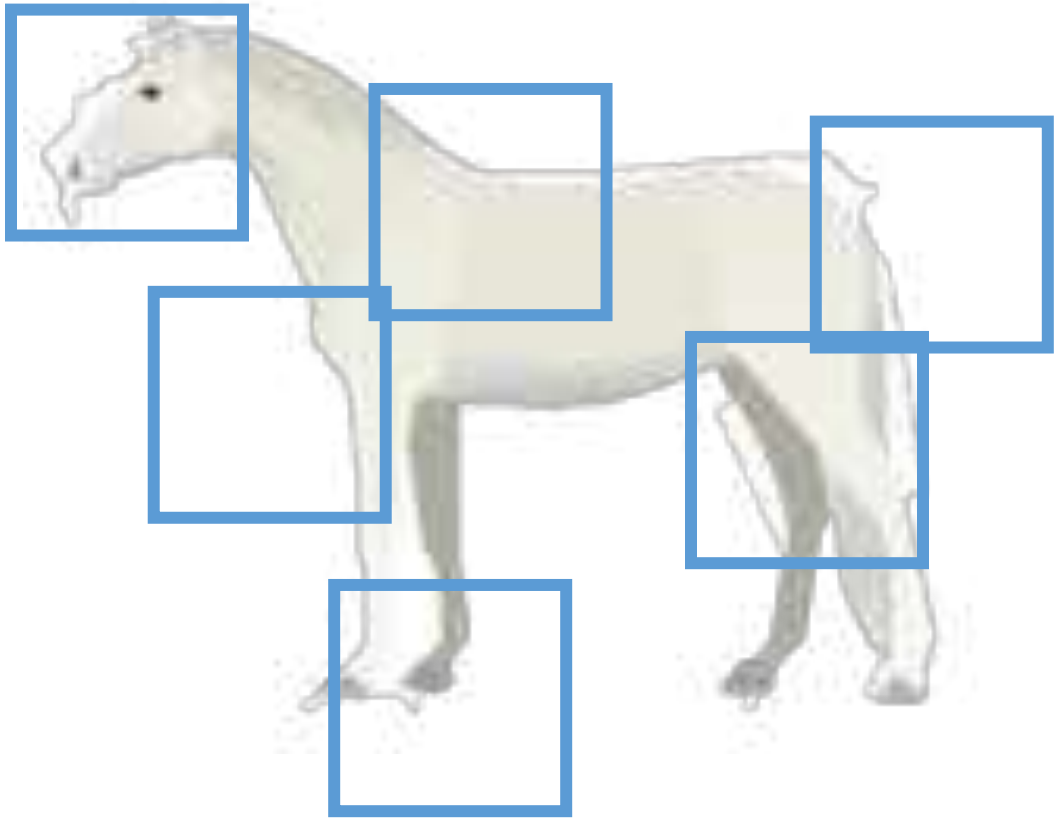


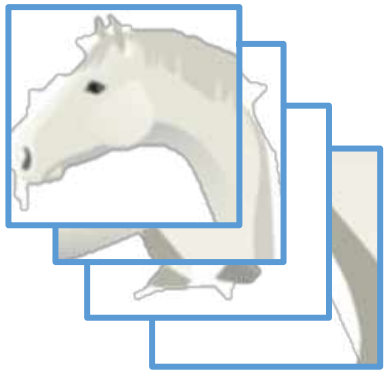
structure code

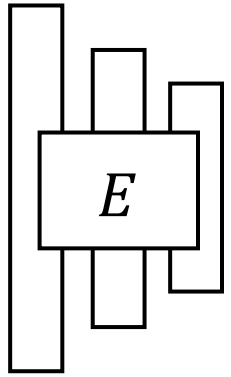




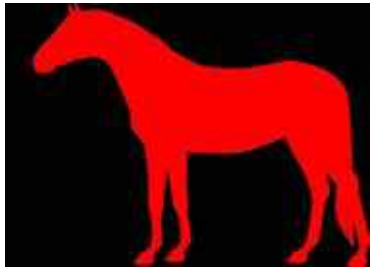
Extracting style and structure from an image



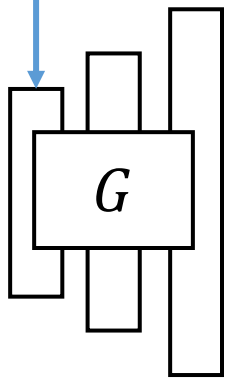


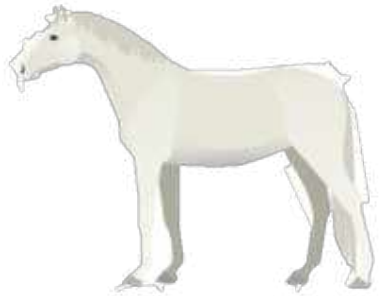


style code

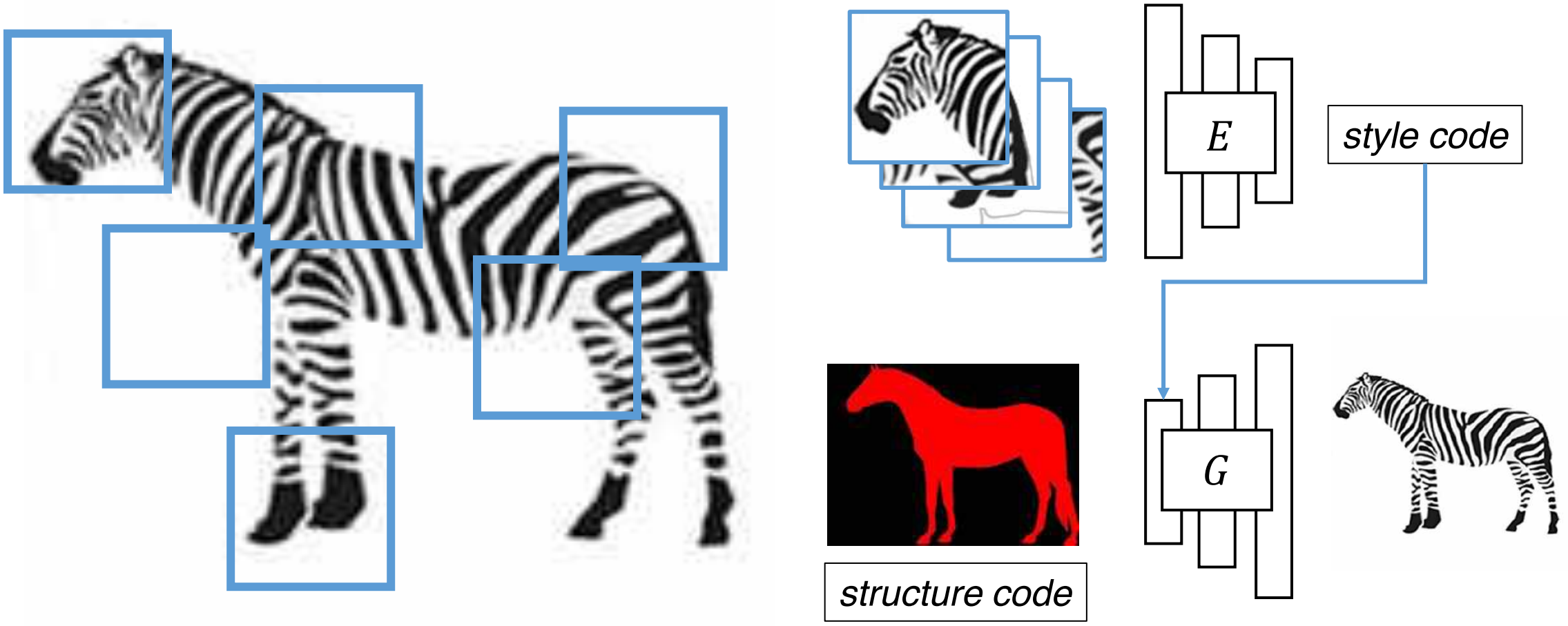


structure code

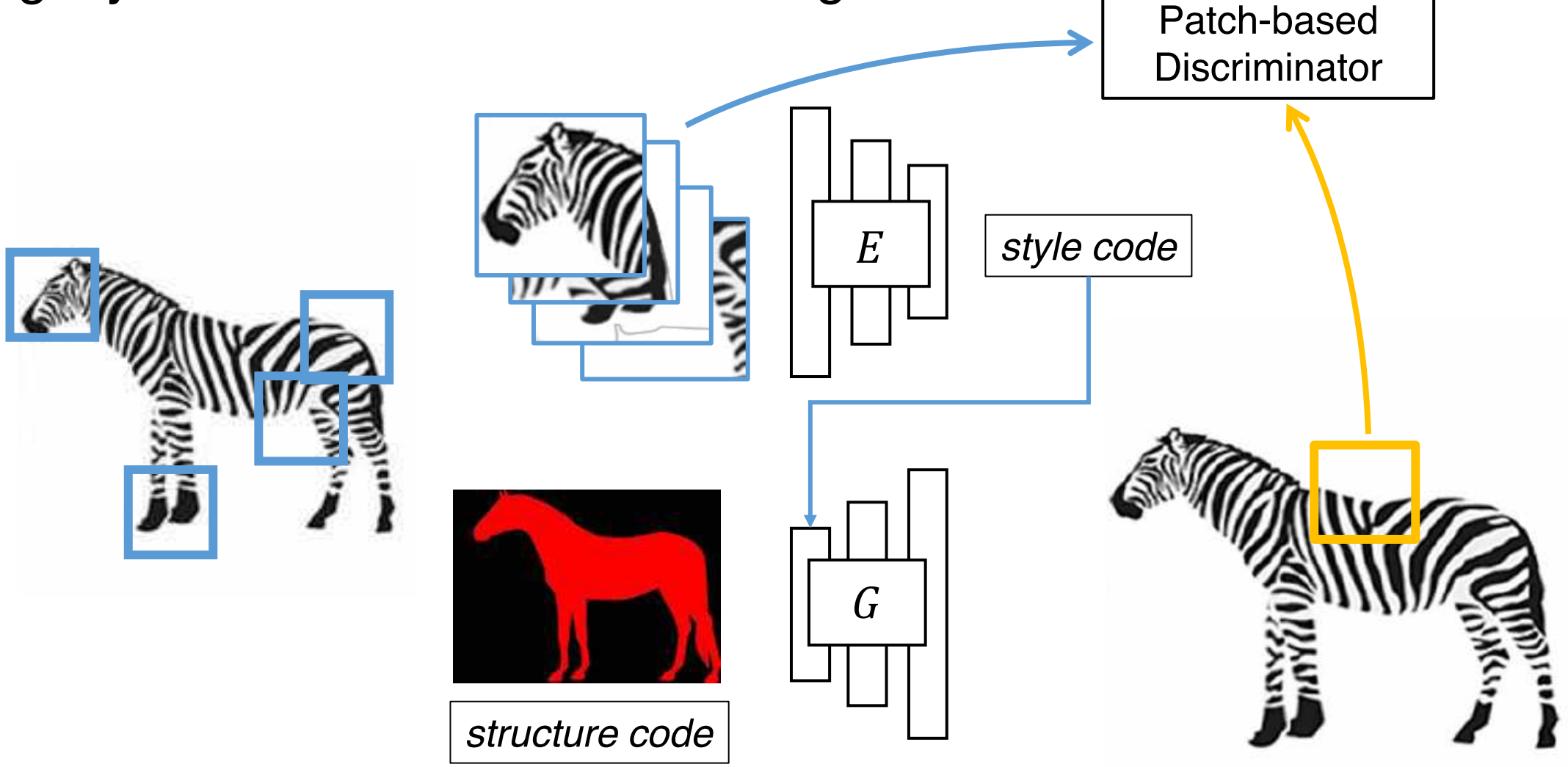




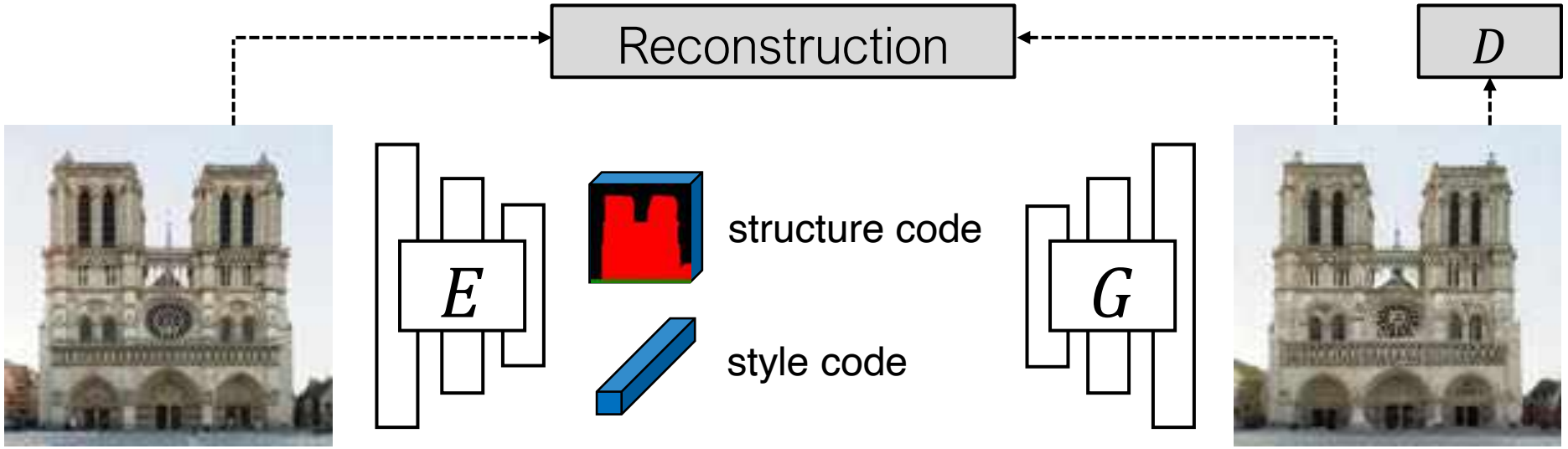
Extracting style and structure from an image

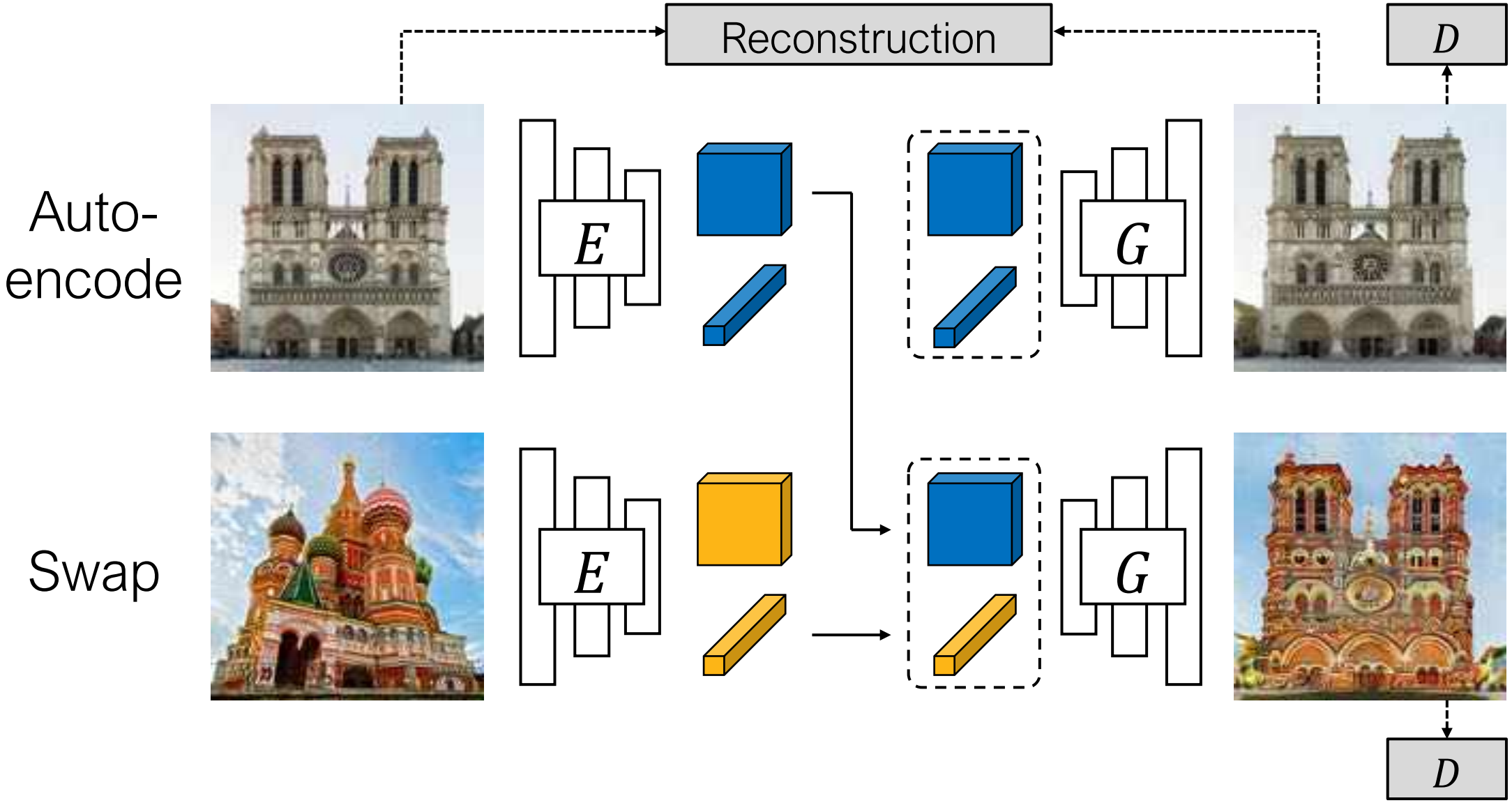


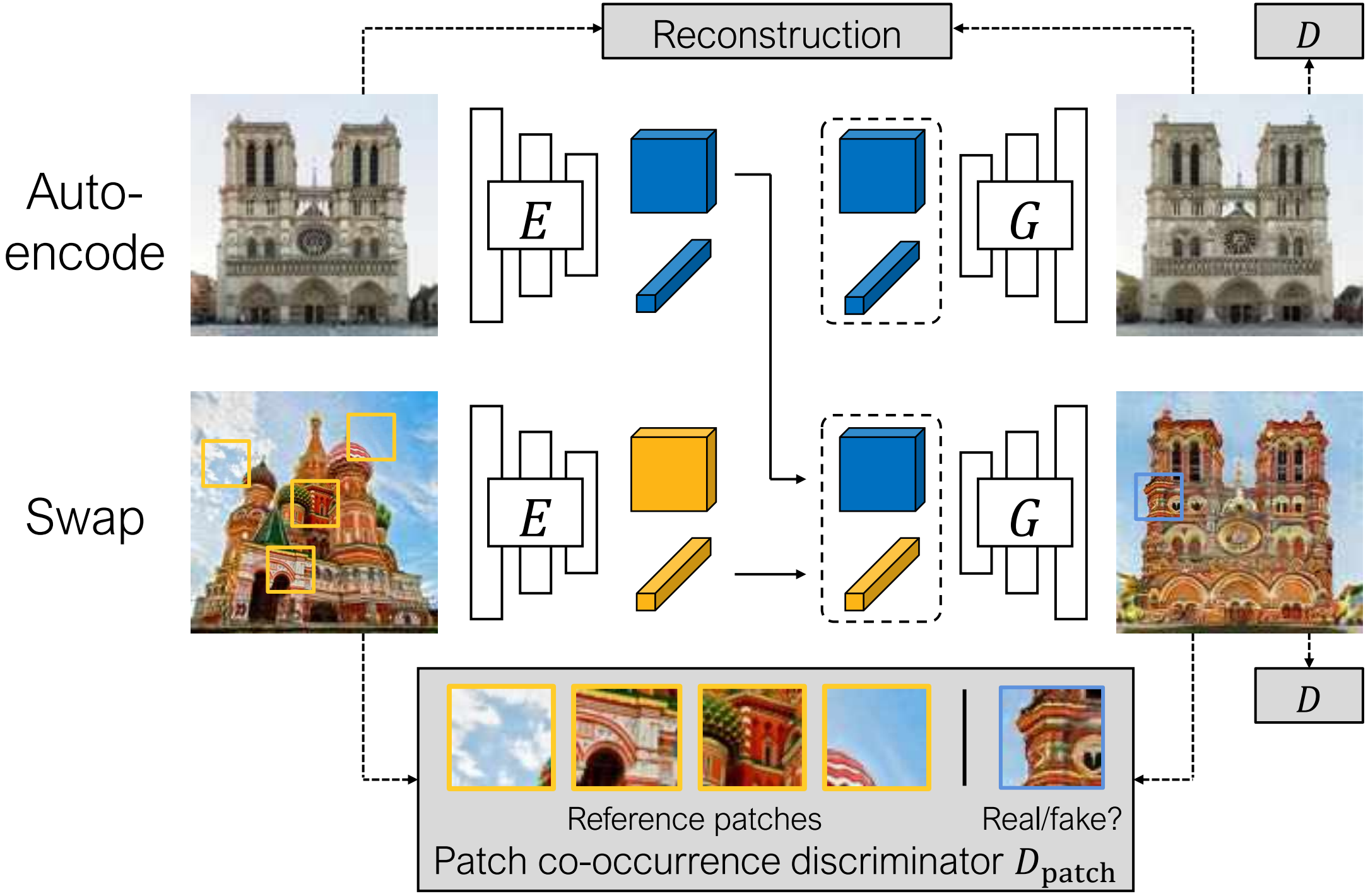
Extracting style and structure from an image



Auto-encode




Reconstruction
D
Auto-
encode
E
G
Swap
E
G
D


Reconstruction
D
Auto-
encode
E
G
Swap
E
G
D
Reference patches
Real/fake?
Patch co-occurrence discriminator D_{patch}


style
structure

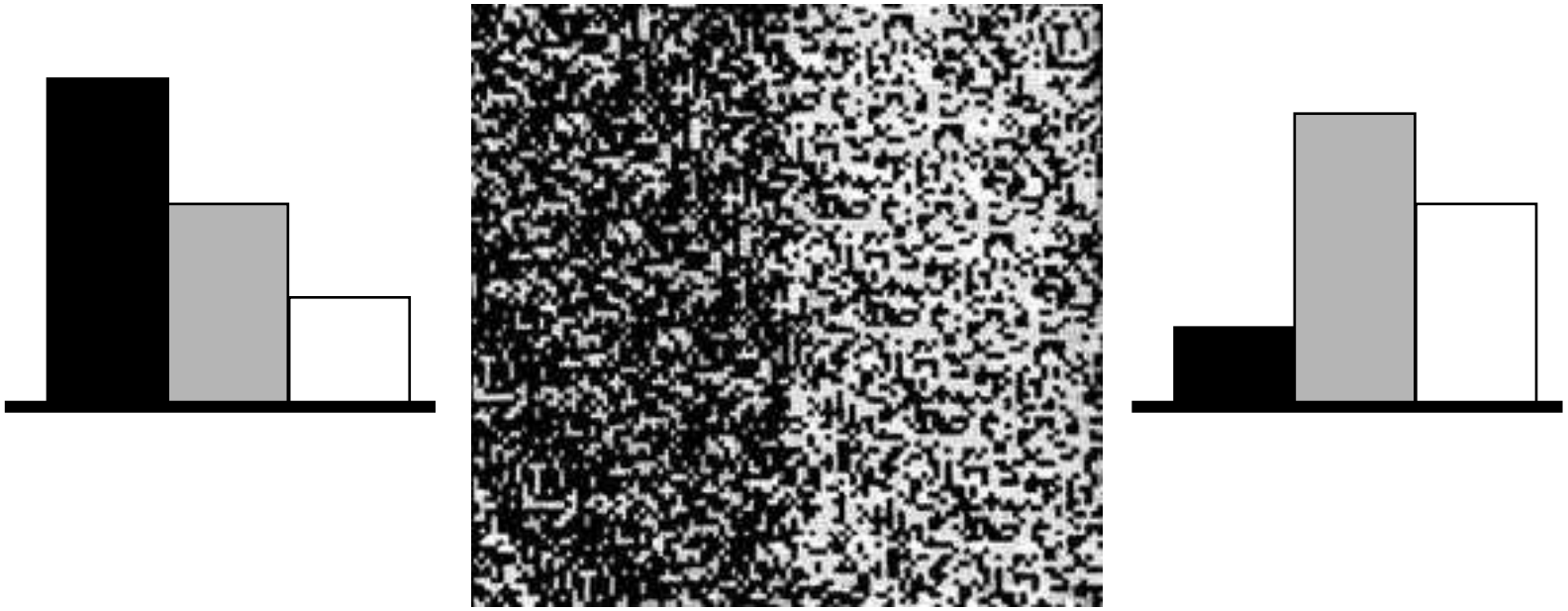

style
structure

Patch Co-Occurrence Discriminator is a Texture Discriminator

What is Texture?

“An image that can be represented by first and second-order statistics”

Conjecture by Bela Julesz, 1962



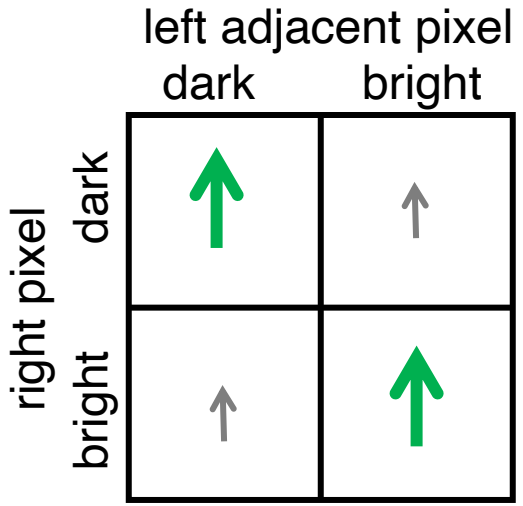
Two textures that differ by first-order statistics

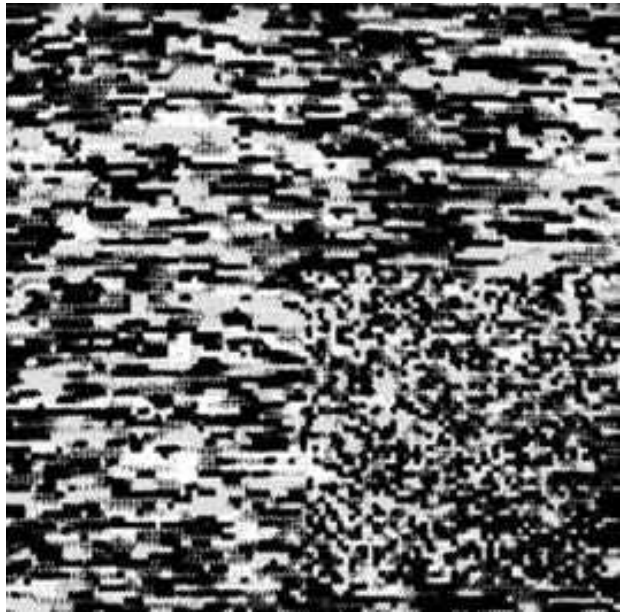
Patch Co-Occurrence Discriminator is a Texture Discriminator

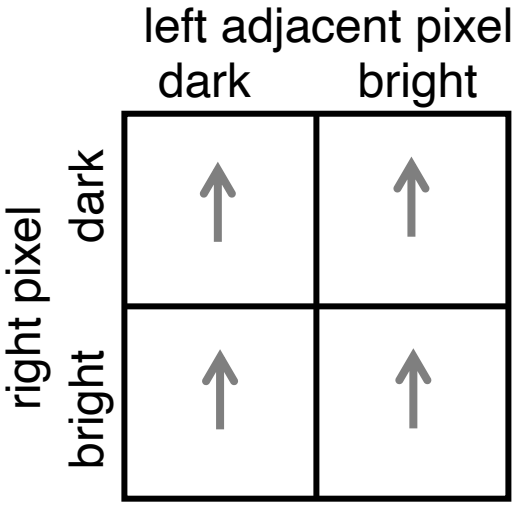
What is Texture?

“An image that can be represented by first and second-order statistics”

Conjecture by Bela Julesz, 1962







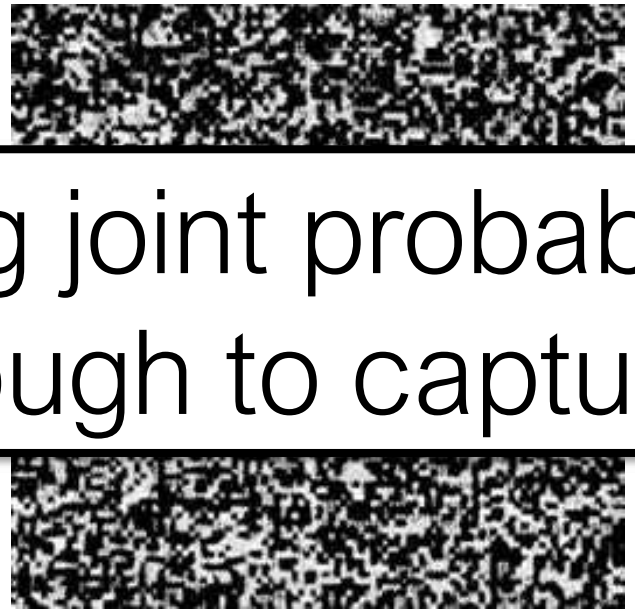
Two textures that differ by second-order statistics

Patch Co-Occurrence Discriminator is a Texture Discriminator

What is Texture?

“An image that can be represented by first and second-order statistics”

Conjecture by Bela Julesz, 1962



Modeling joint probability is (almost) enough to capture texture

Two textures that differ by third-order statistics

Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) =

Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) = Different Image

Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) =

Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) = Same Image

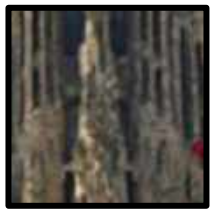
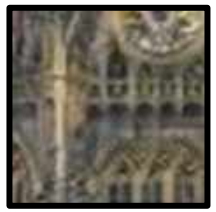
Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) =

Patch Co-Occurrence Discriminator is a Texture Discriminator

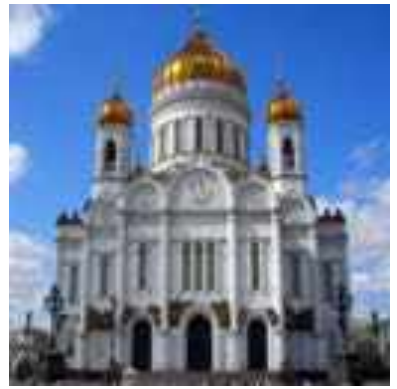
D( , ) = Different Image

Patch Co-Occurrence Discriminator is a Texture Discriminator

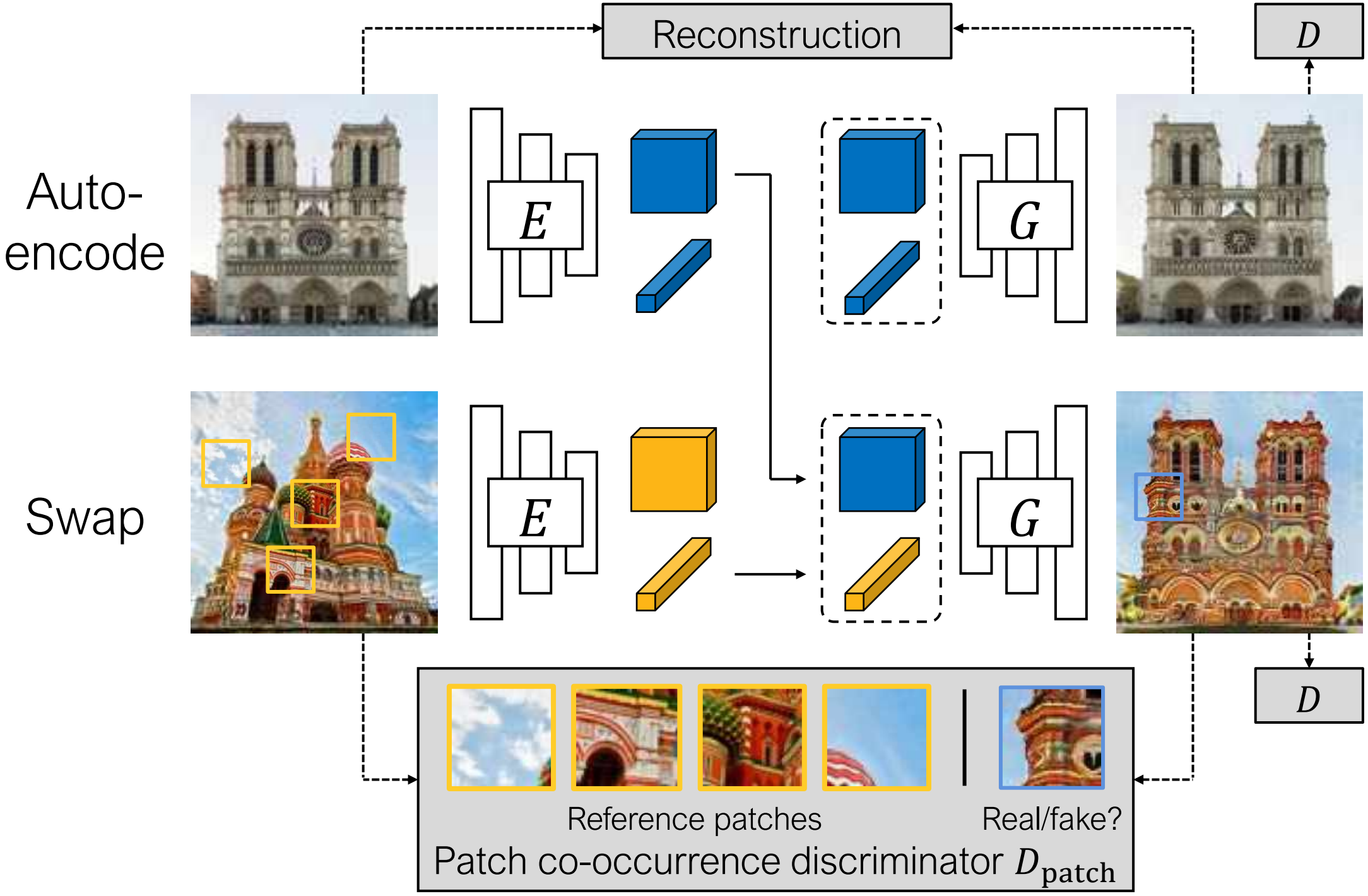
D( , ) =

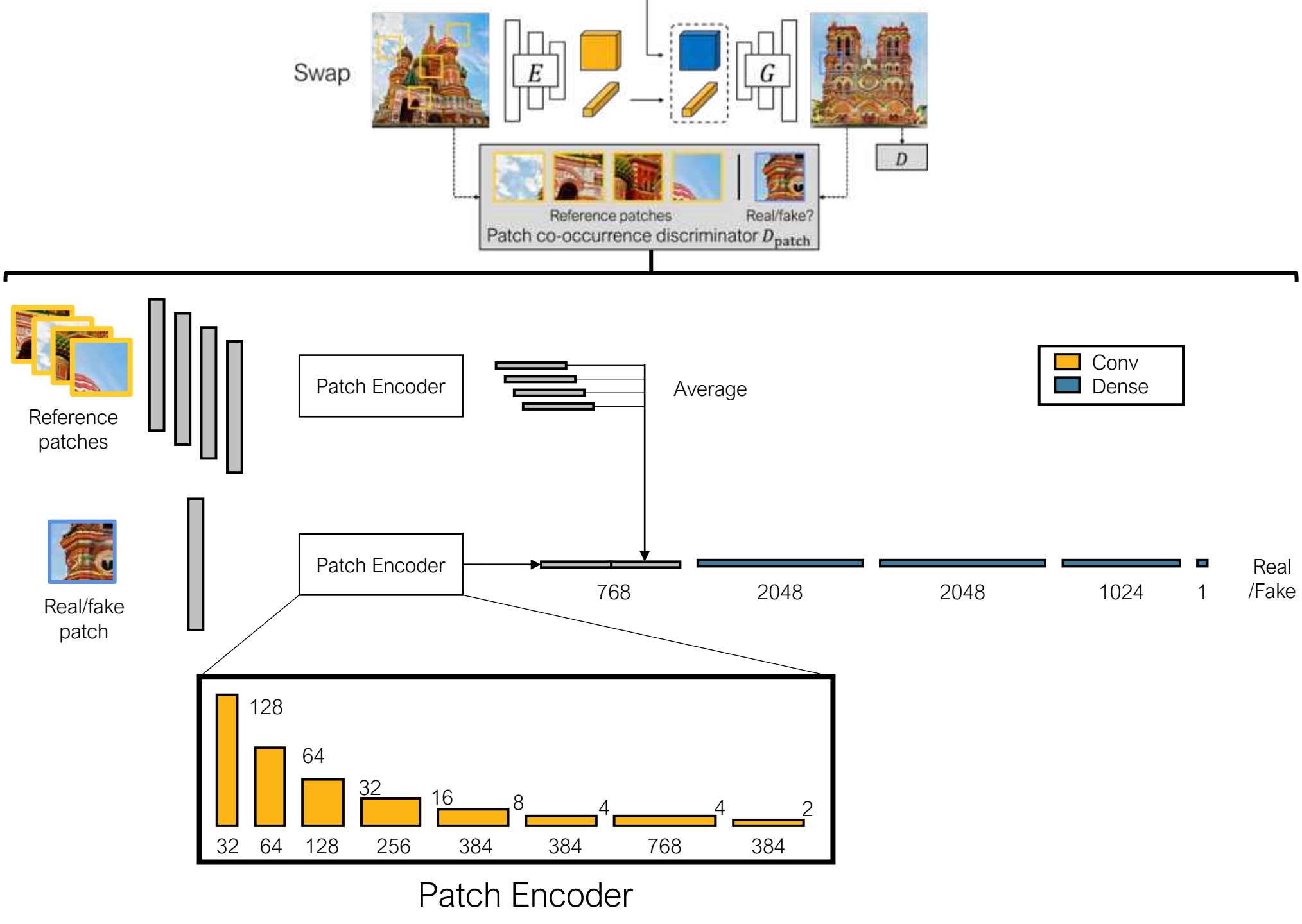
Patch Co-Occurrence Discriminator is a Texture Discriminator

D( , ) = Different (fake) Image



structure image


Reconstruction
D
Auto-
encode
E
G
Swap
E
G
D
Reference patches
Real/fake?
Patch co-occurrence discriminator D_{patch}


Swap
E
G
D
Reference patches
Real/fake?
Patch co-occurrence discriminator D_{patch}
Reference patches
Real/fake patch
Patch Encoder
Patch Encoder
Average
Conv
Dense
768
2048
2048
1024
1
Real /Fake
128
64
32
16
8
4
4
2
32
64
128
256
384
384
768
384
Patch Encoder

Embedding a Real Input Image

Fast

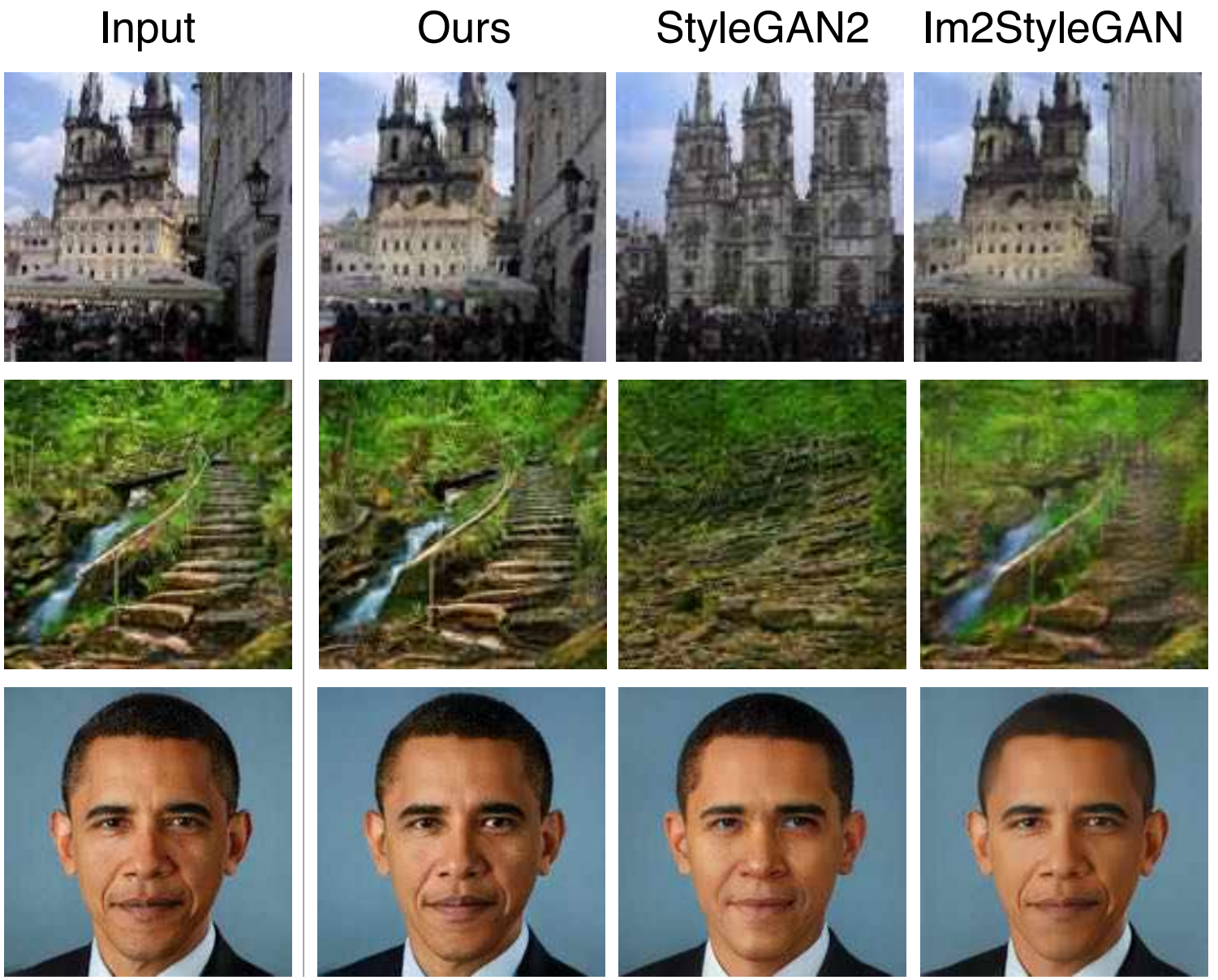
feed-forward pass of the encoder.

Magnitudes faster than baselines.

Accurate

Prior distribution not enforced on the latent space.

Spatial resolution is retained.



StyleGAN2 (Karras et al., CVPR'20), Im2StyleGAN (Abdal et al., ICCV'19)

Embedding a Real Input Image

Fast

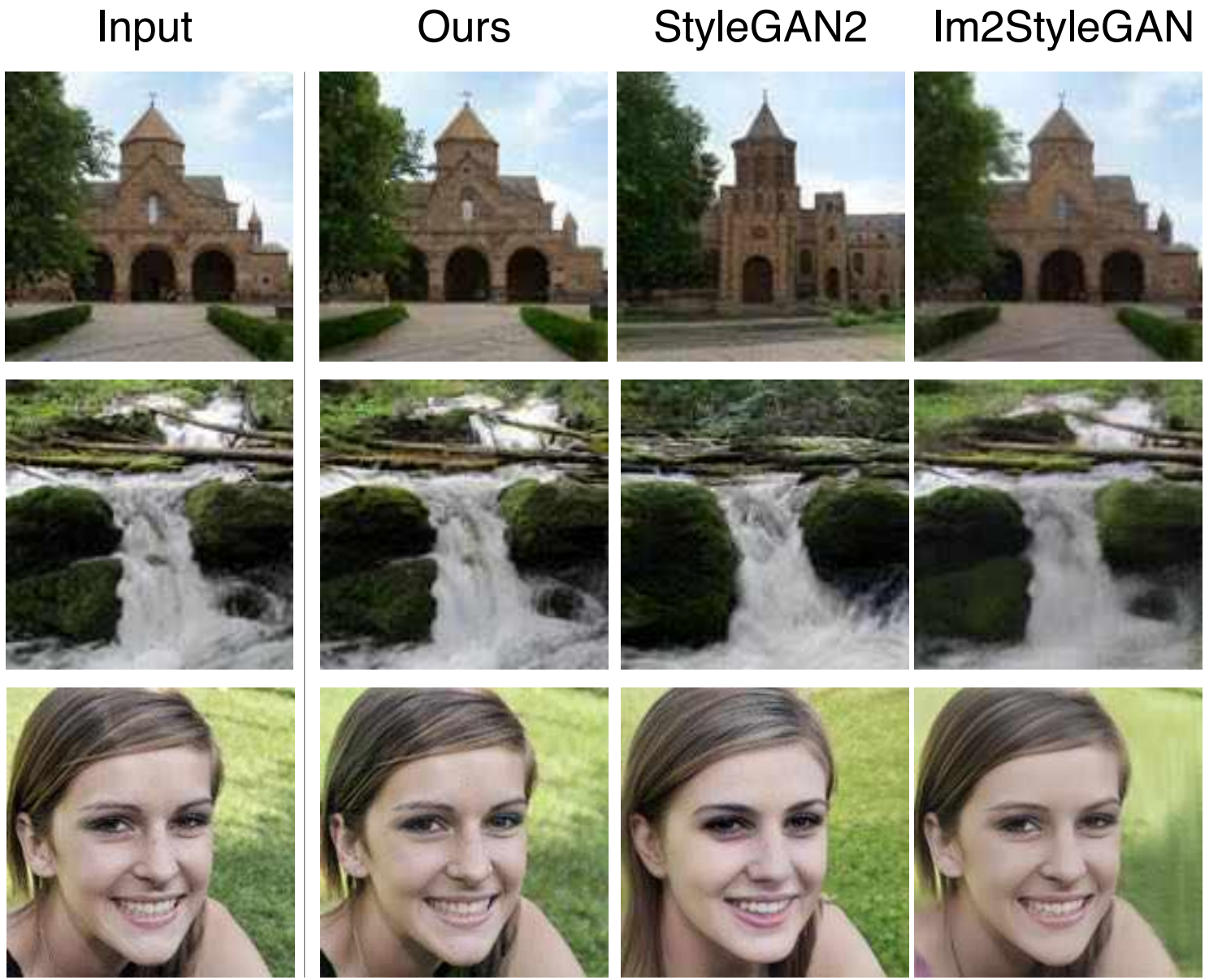
feed-forward pass of the encoder.

Magnitudes faster than baselines.

Accurate

Prior distribution not enforced on the latent space.

Spatial resolution is retained.



StyleGAN2 (Karras et al., CVPR'20), Im2StyleGAN (Abdal et al., ICCV'19)

Photorealistic and Disentangled Swapping Quality

Structure

Texture

StyleGAN2

Im2StyleGAN

STROTSS

WCT^2

Ours

StyleGAN2 (Karras et al., CVPR'20), Im2StyleGAN (Abdal et al., ICCV'19)
STROTSS (Kolkin et al., CVPR'19), WCT^2 (Yoo et al., ICCV'19)

Photorealistic and Disentangled Swapping Quality

Structure

Texture

StyleGAN2

Im2StyleGAN

STROTSS

WCT^2

Ours

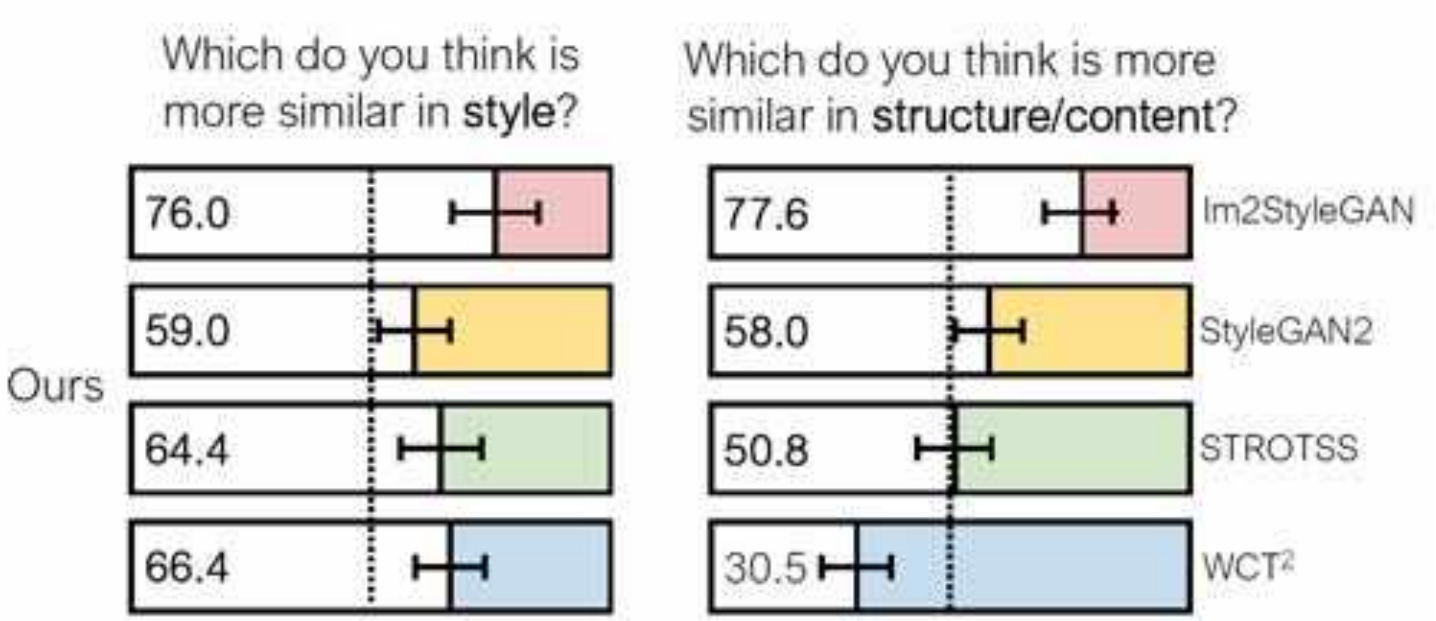
StyleGAN2 (Karras et al., CVPR'20), Im2StyleGAN (Abdal et al., ICCV'19)
STROTSS (Kolkin et al., CVPR'19), WCT^2 (Yoo et al., ICCV'19)

Realism of generated images

Method	Runtime (sec) (↓)	Human Perceptual Study (AMT Fooling Rate) (↑)			
		Church	FFHQ	Waterfall	**Average**
Swap Autoencoder (Ours)	**0.113**	**31.3±2.4**	**29.8±2.5**	**41.8±2.2**	**34.4±1.4**
Im2StyleGAN [1, 37]	990	8.5±2.1	11.6±2.3	12.8±2.4	11.0±1.3
StyleGAN2 [38]	192	24.3±2.2	22.8±2.5	35.3±2.4	27.5±1.4
STROTSS [41]	166	13.7±2.2	13.6±2.5	23.0±2.1	16.7±1.3
WCT^2 [69]	1.35	***27.9±2.3***	***26.6±2.4***	35.8±2.4	30.1±1.4

StyleGAN2 Im2StyleGAN STROTSS WCT^2 **Ours**

Structure Texture



StyleGAN2 Im2StyleGAN STROTSS WCT^2 **Ours**

Structure Texture

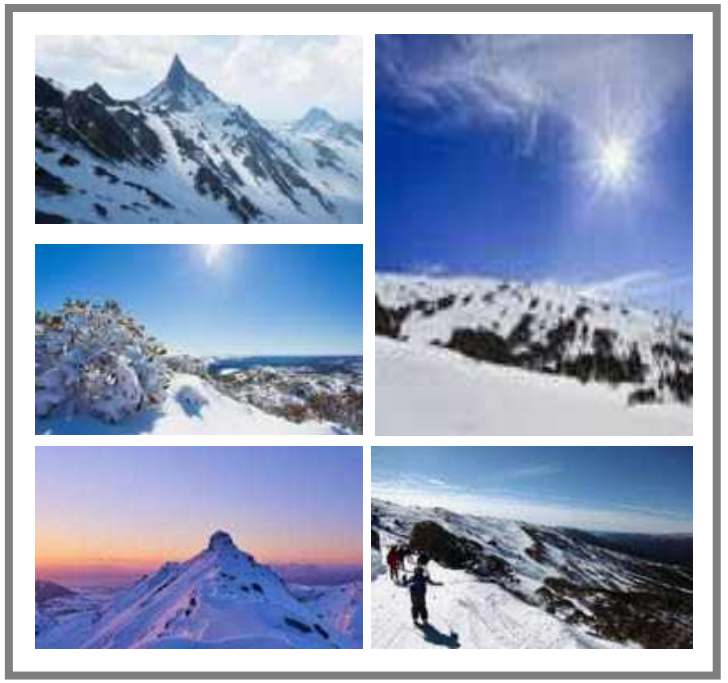
Smooth Latent Space

Encoder

$z_1, z_1, \ldots, z_N$

average

z_{summer}



Encoder

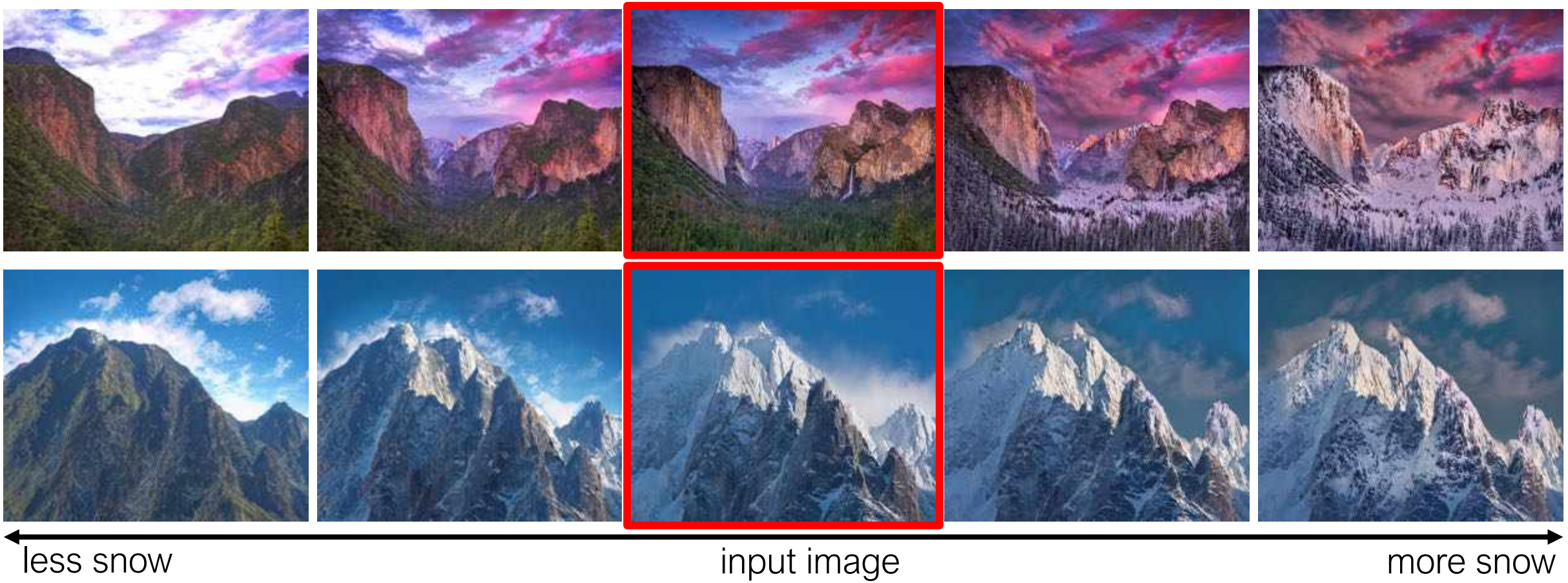
$z_1, z_1, \ldots, z_N$

average

z_{snow}

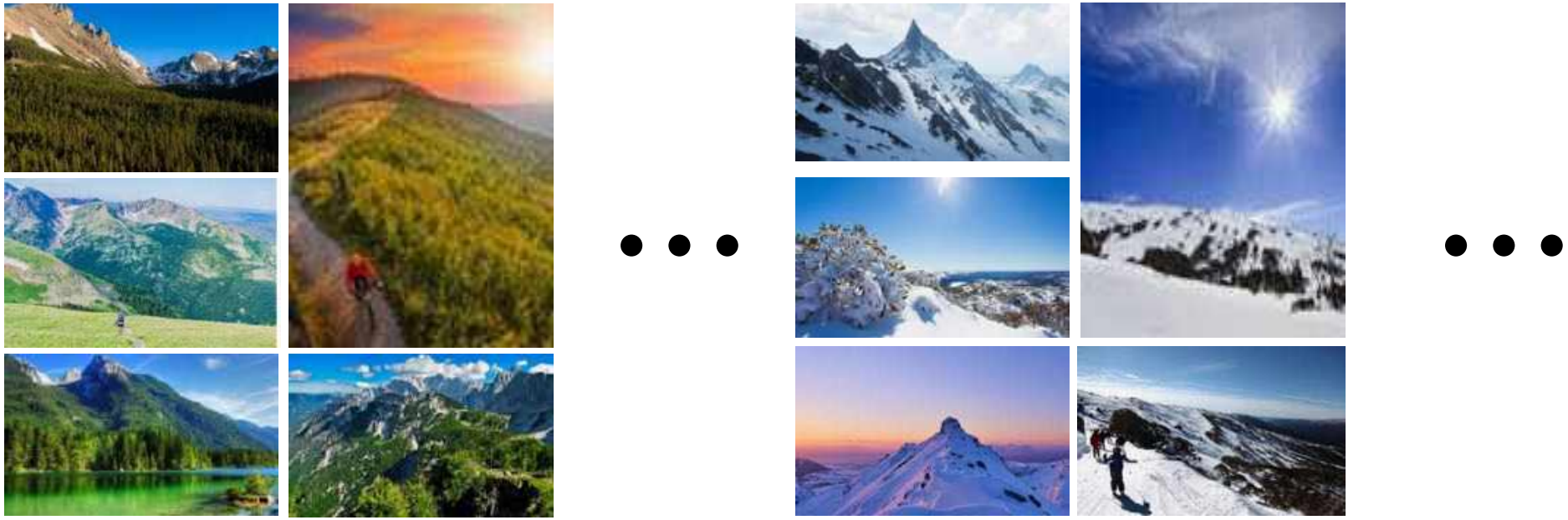
$$Z_{\text{add\_snow}} = z_{snow} - z_{summer}$$

Smooth Latent Space



PCA on the Latent Space

All kinds of mountains (summer, snow, desert, canyon, …)



Encoder

$z_1, z_1, \ldots, z_N$

PCA

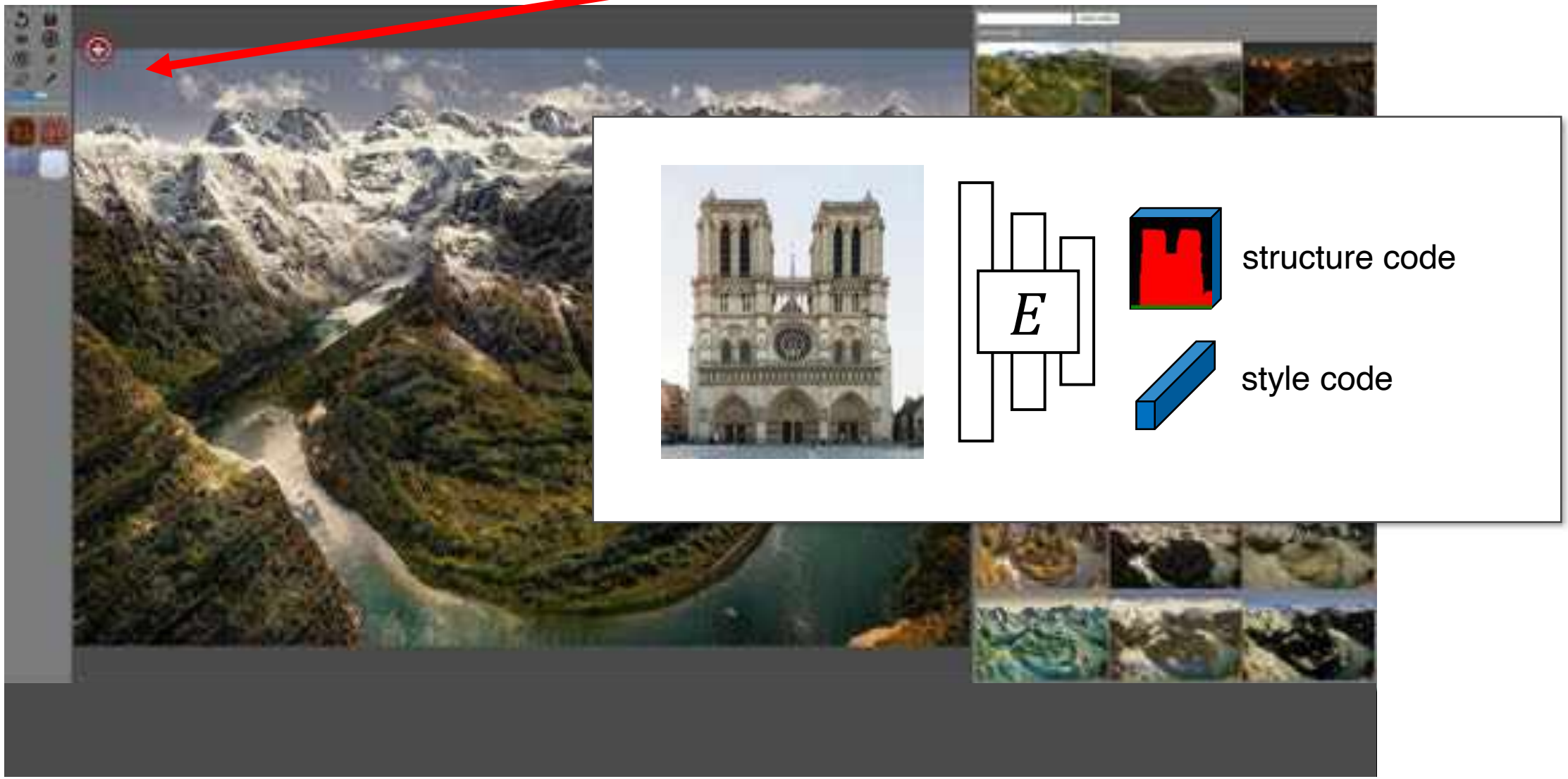
discovered edit vectors

GANSpace (Harkonen et al., 2020)

Interactive UI

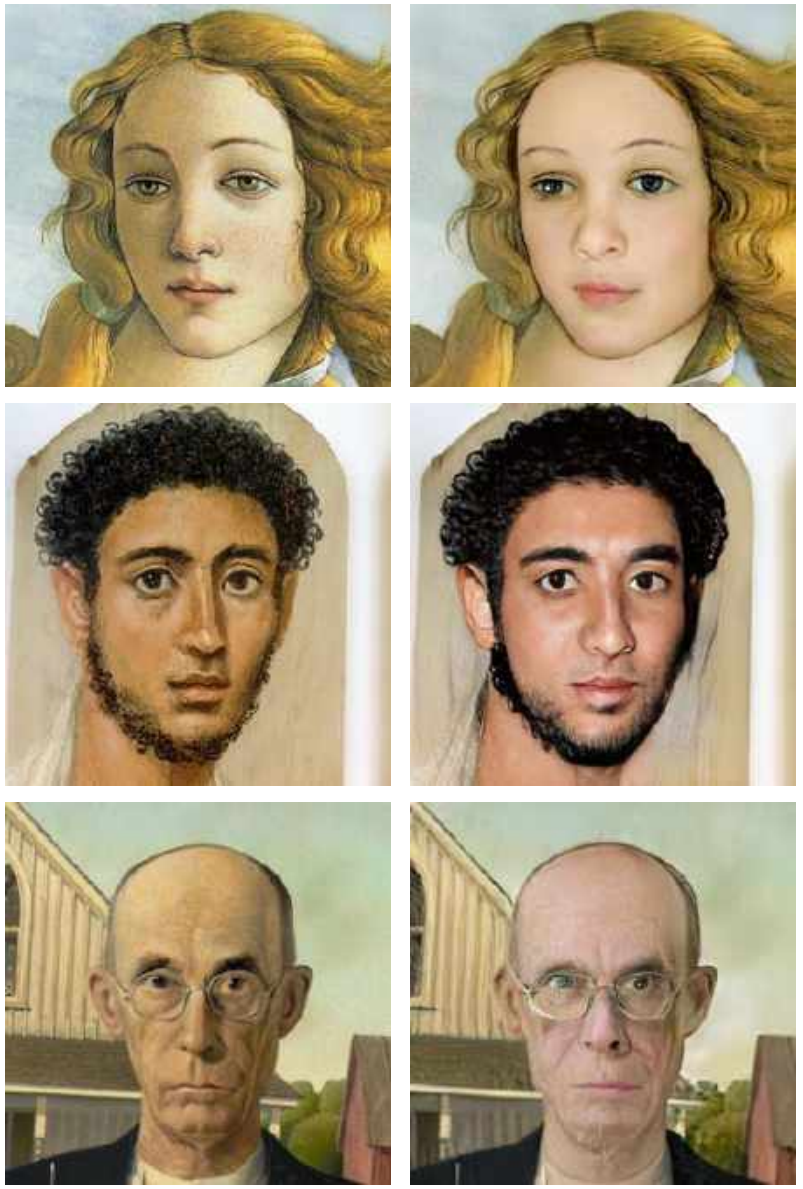
"Preview"
of each principal axis

Structural Editing

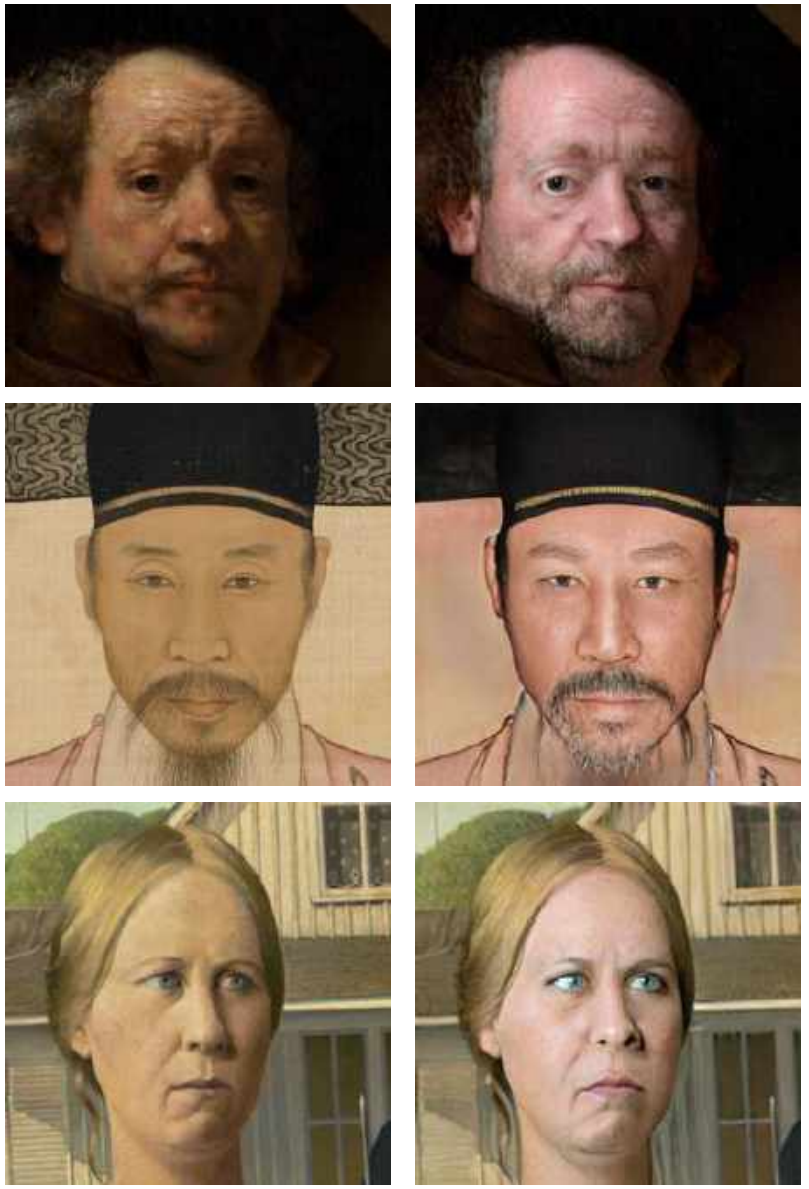


Structural Editing

User-Guided Portrait Painting to Photo



input output



input output

User-Guided Animal Face Transformation

input

same pose, different styles

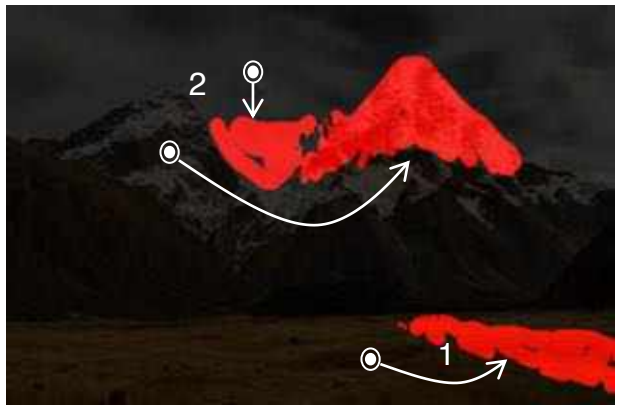
PCA on the Structure Code

input | bigger eyes | gaze direction | more smile | 5 o'clock shadow

Editing Landscape Images

Editing Landscape Images

UI with input image



brush stroke visualization

1. remove road

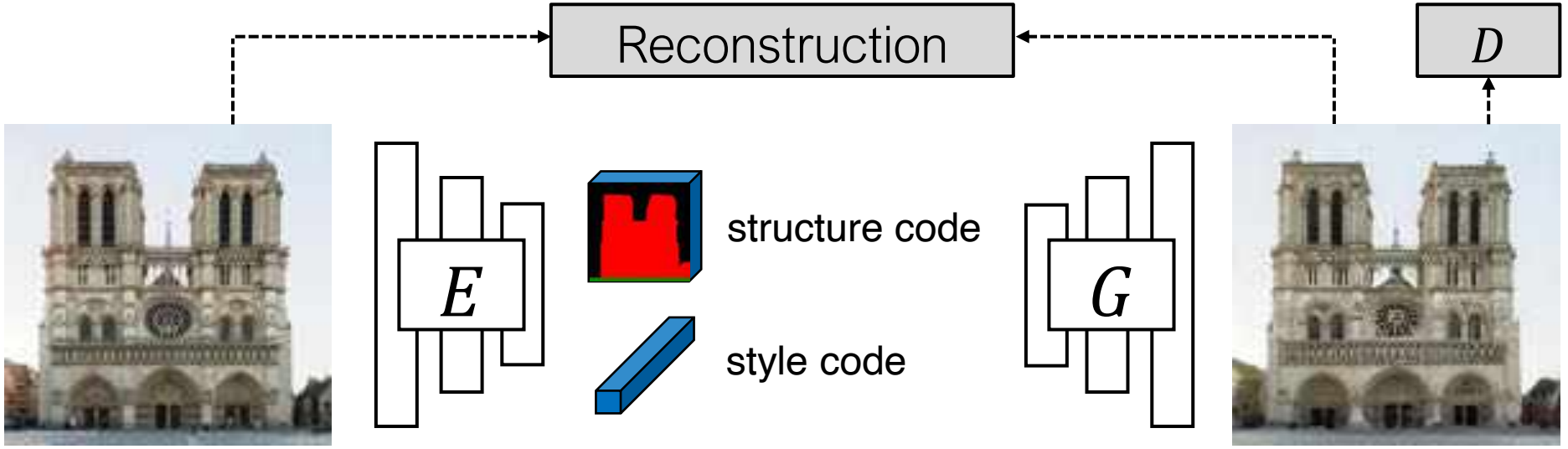
2. draw mountain

PCA on the structure code,
with user-drawn mask

PCA with the style (texture) code

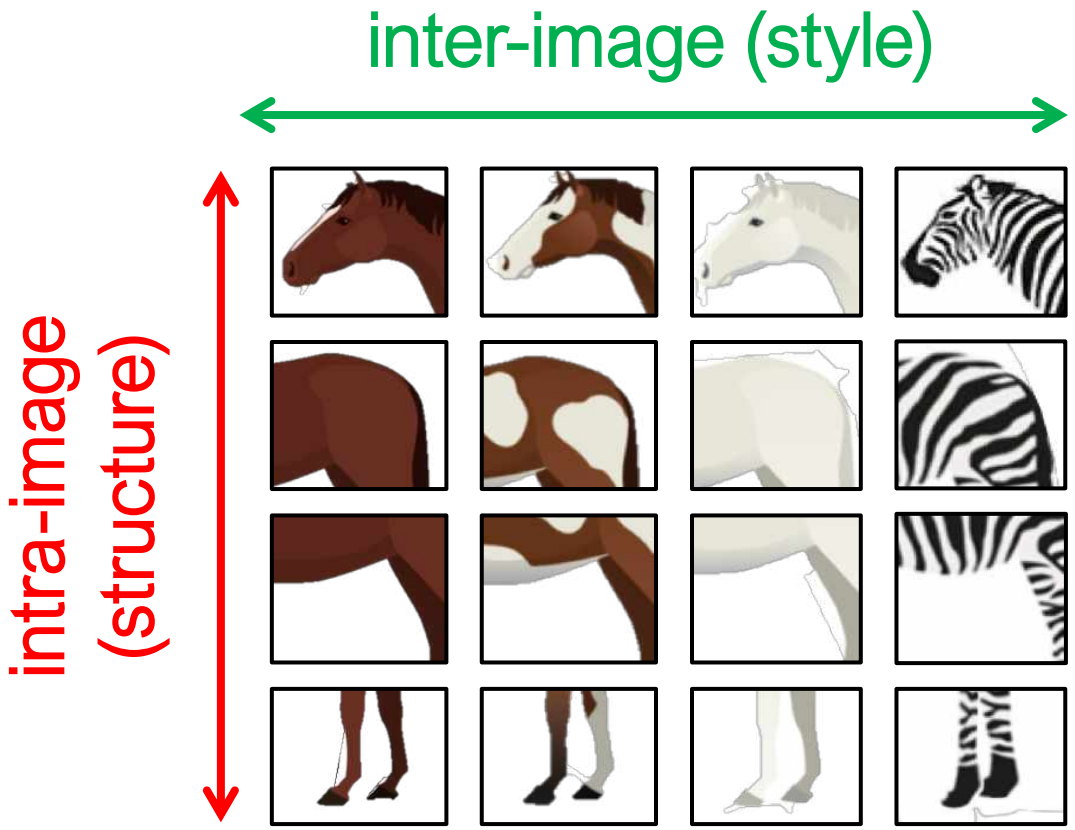
Summary

Auto-encode



GAN that can embed images

Summary



structure / style disentanglement

Summary

interactive user editing

Thank you!

https://taesung.me/ContrastiveUnpairedTranslation
https://taesung.me/SwappingAutoencoder