Image Generation Using Disentanglement

Ron Mokady

What is Disentanglement?

No formal definition which is widely accepted.

Our definition: a change in a single underlying factor of variation in the sample x should lead to a change in a single factor in the learned representation r(x).

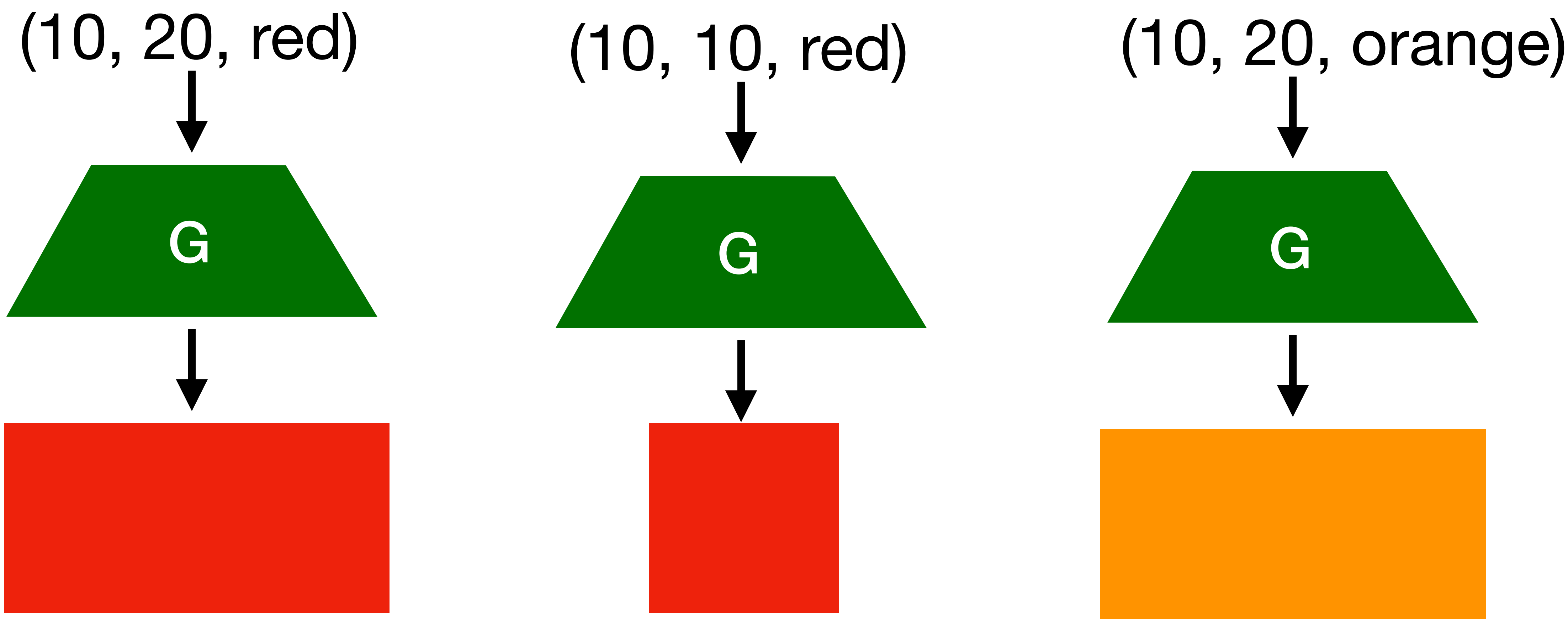
Disentanglement for Generative model

A change in a single underlying factor of variation in the generated image G(z) should lead to a change in a single factor in the learned representation z.

Factors could be style, content, rotation, size, etc.

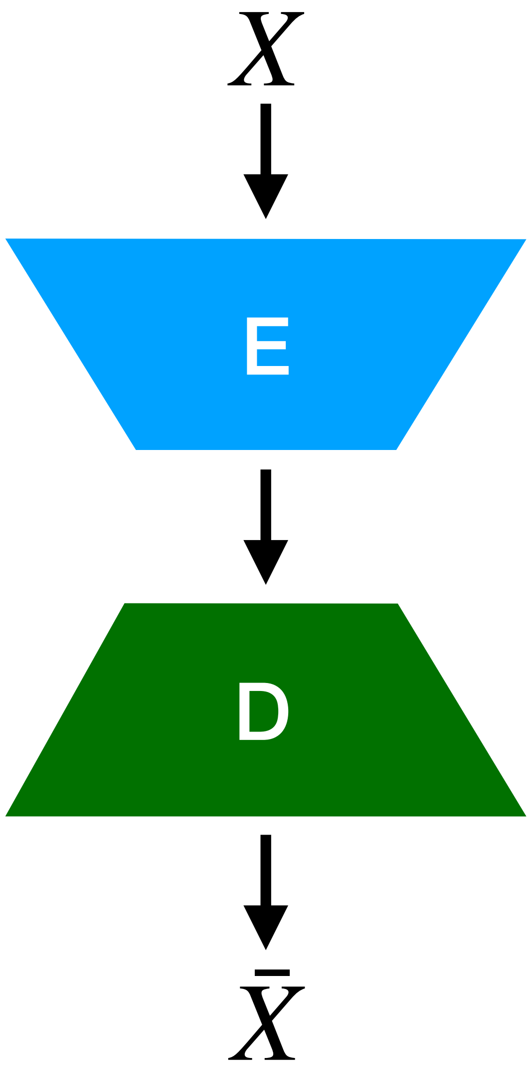
Example - Rectangles

z = (height, width, color)



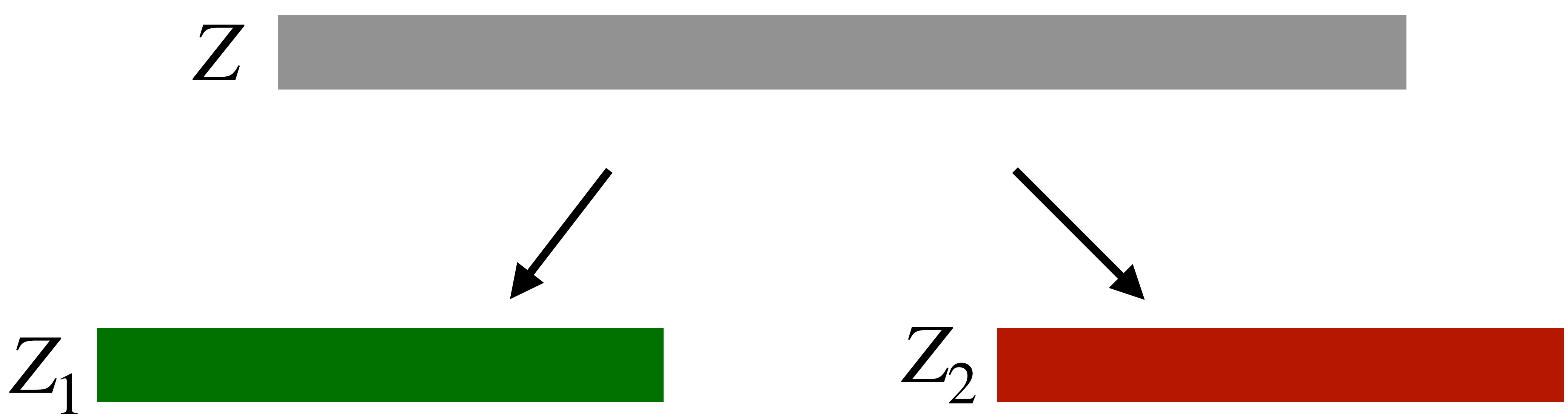
Example - Autoencoder

Using standard Autoencoder, we likely to get entangled latent space.



Separate the Latent Space

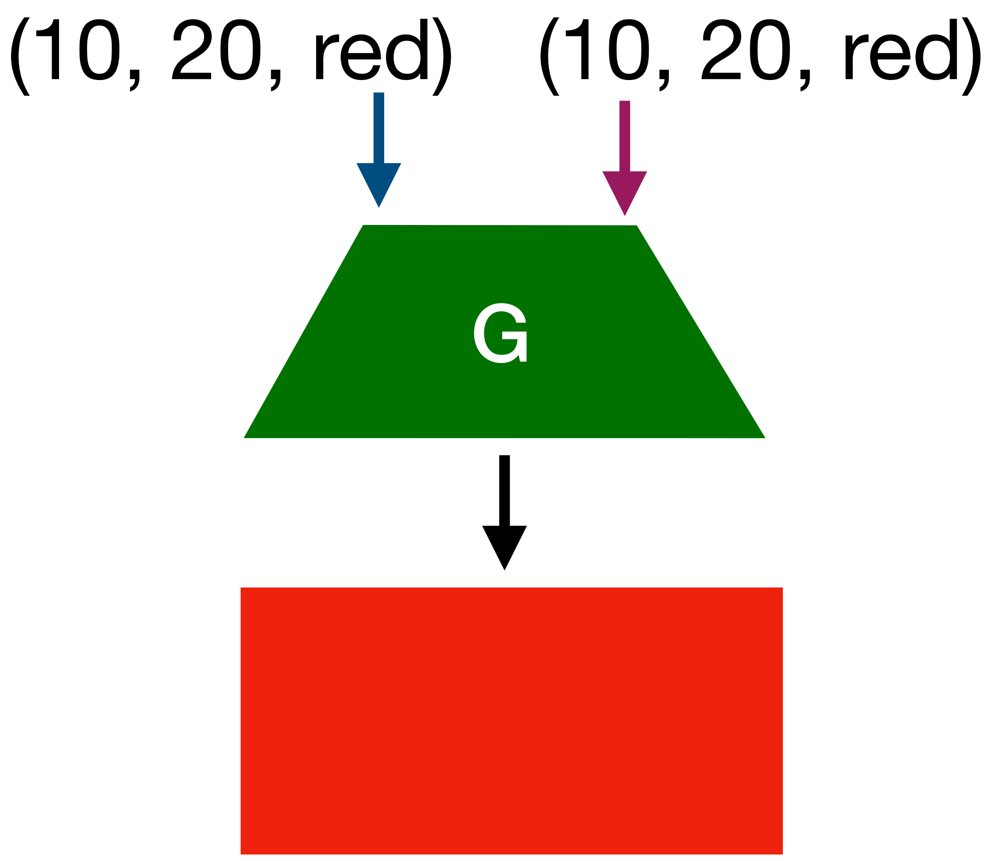
We usually want to separate the latent space, such that each part control one or more factors.

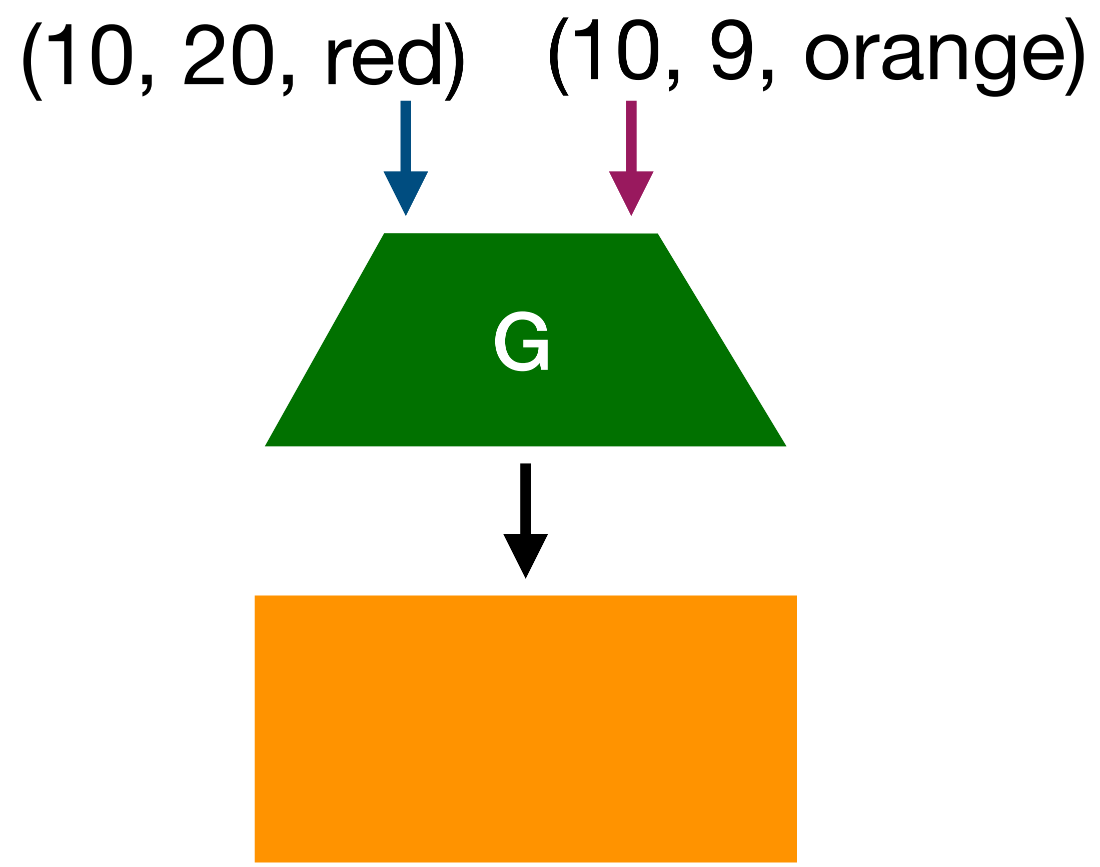


Generative model

The disentanglement is with respect to the generative model.

For example:





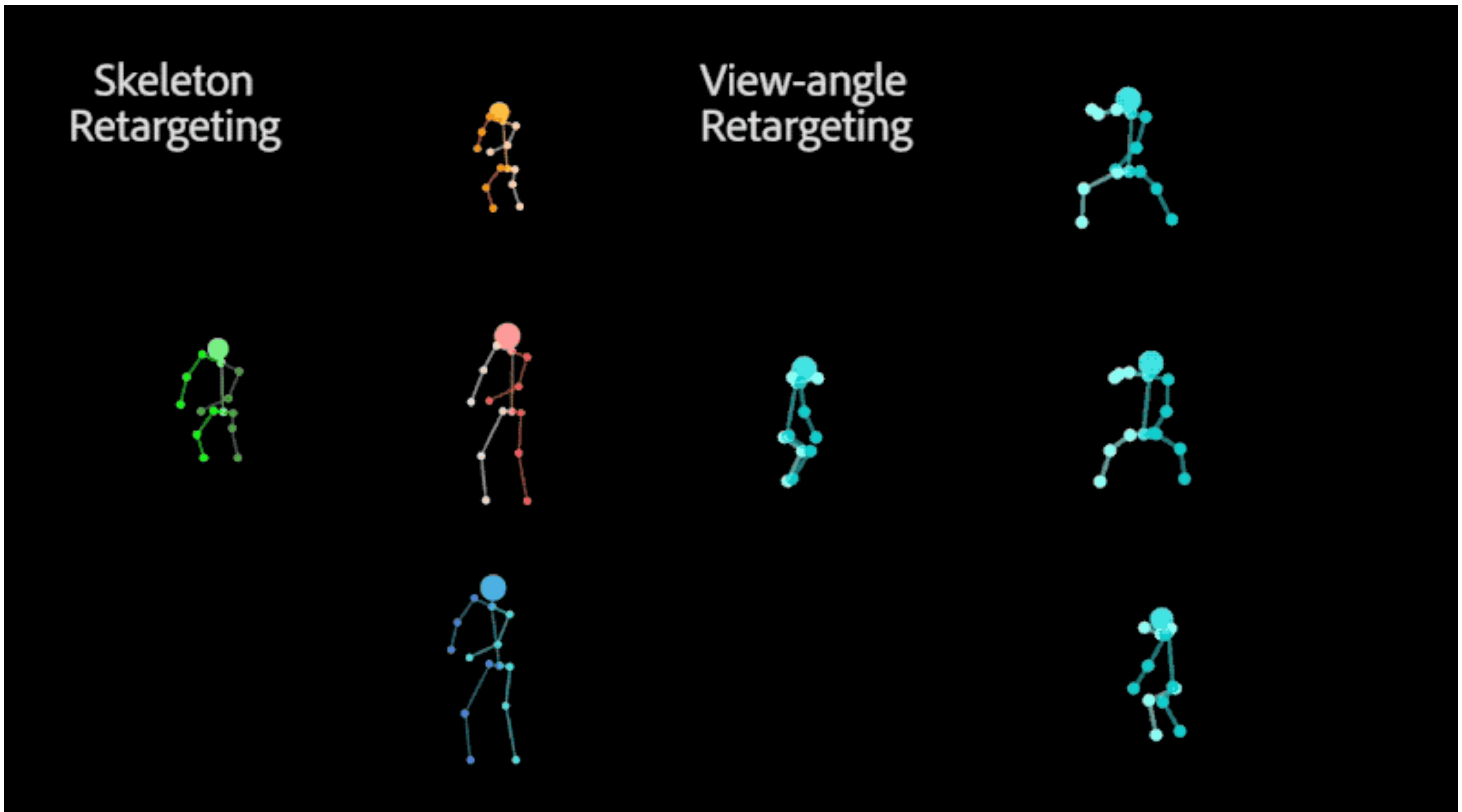
Supervision

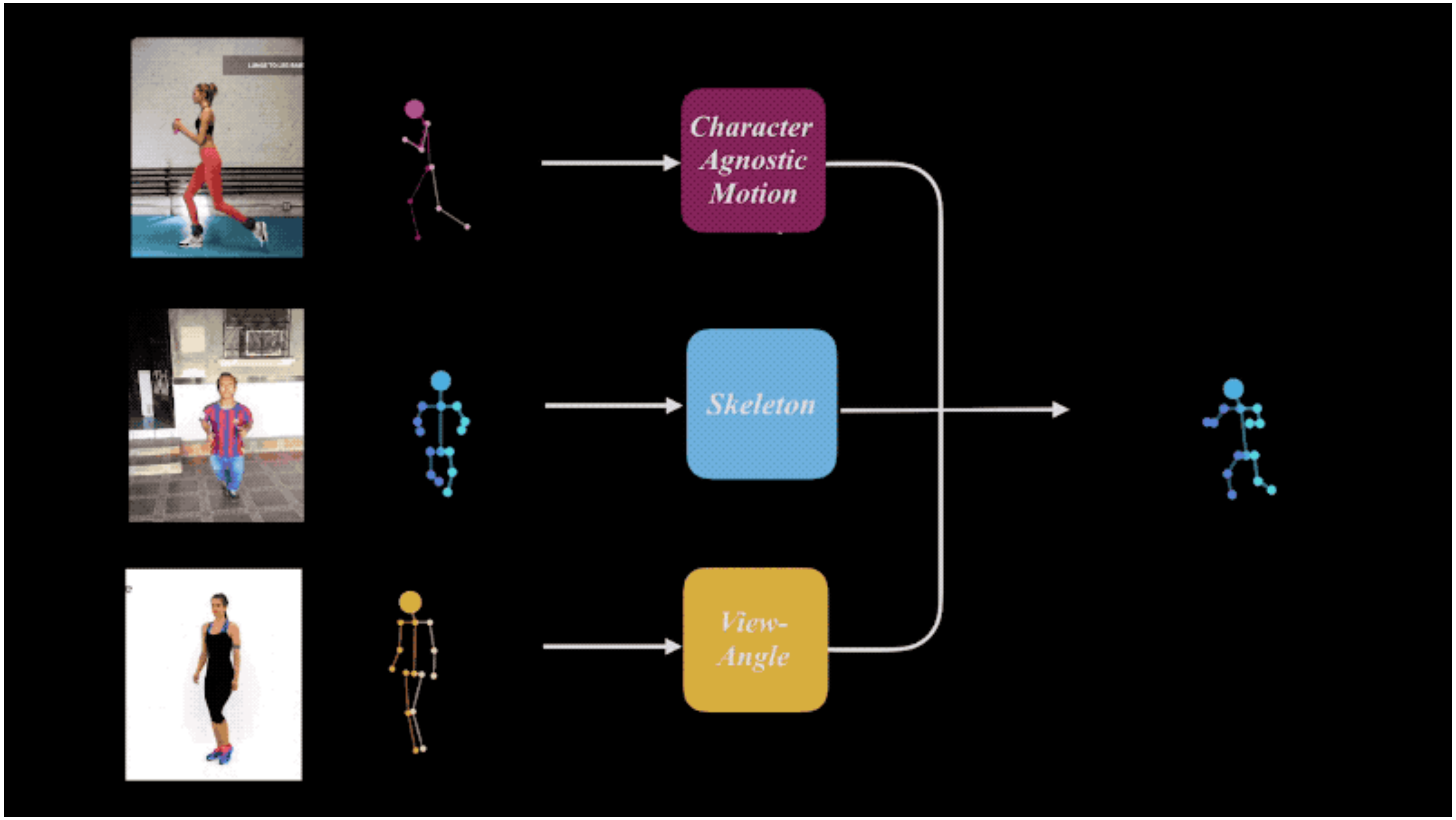
Unsupervised

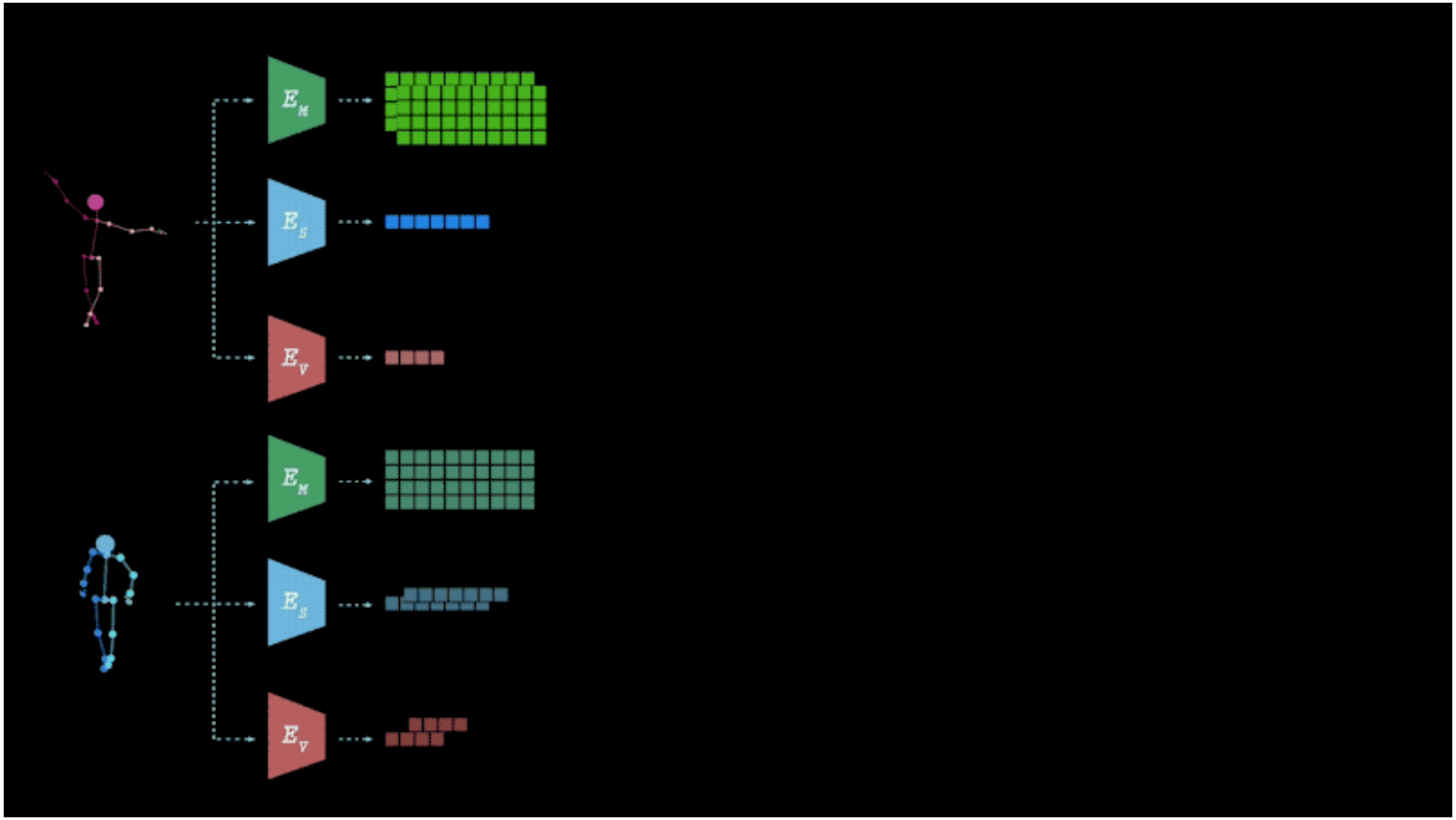
Fully-Supervised

Learning Character-Agnostic Motion for Motion Retargeting in 2D

Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, Daniel Cohen-Or



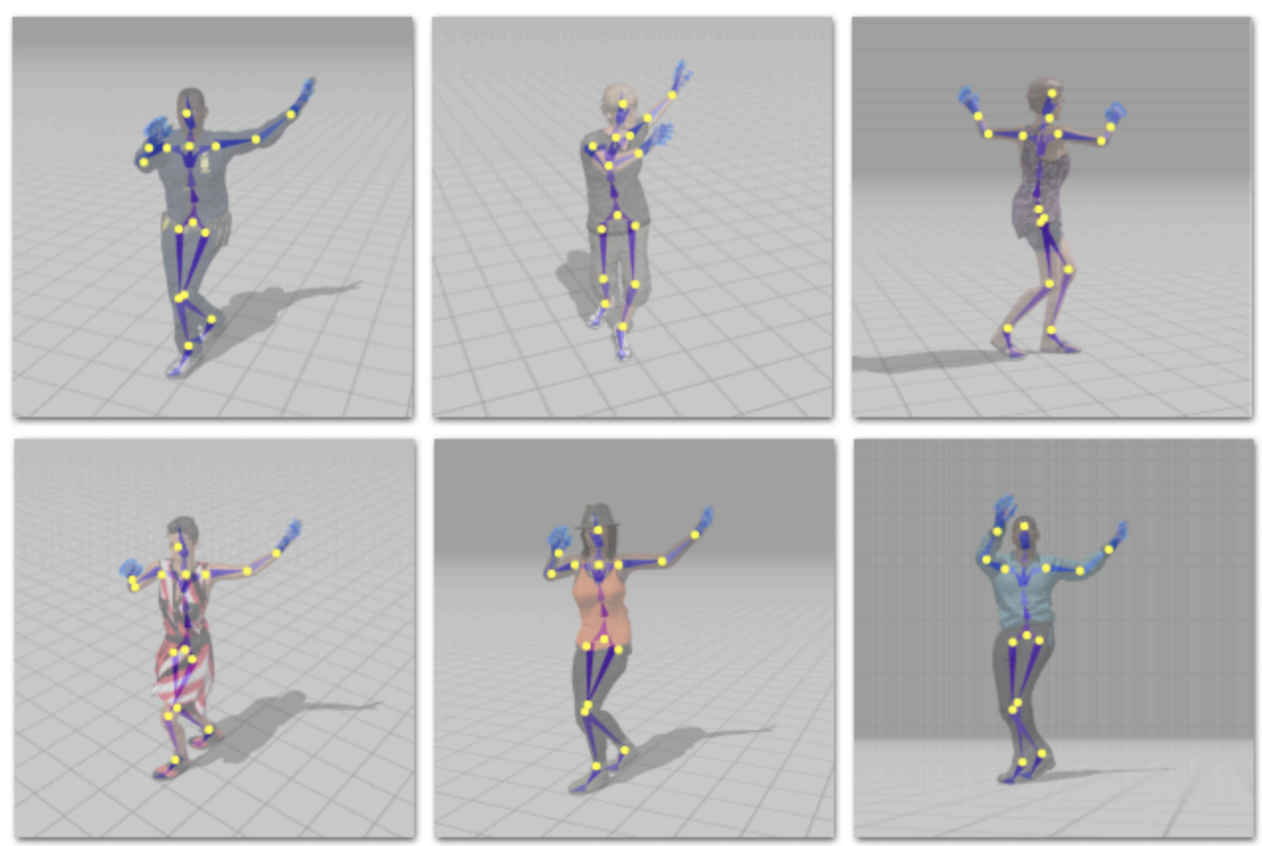

Character Agnostic Motion
Skeleton
View-Angle



E_M
E_S
E_V
E_M
E_S
E_V

Synthetic dataset

Multiple samples of each motion, as performed by the different characters, and these motions can be projected to 2D, from arbitrary view angles.



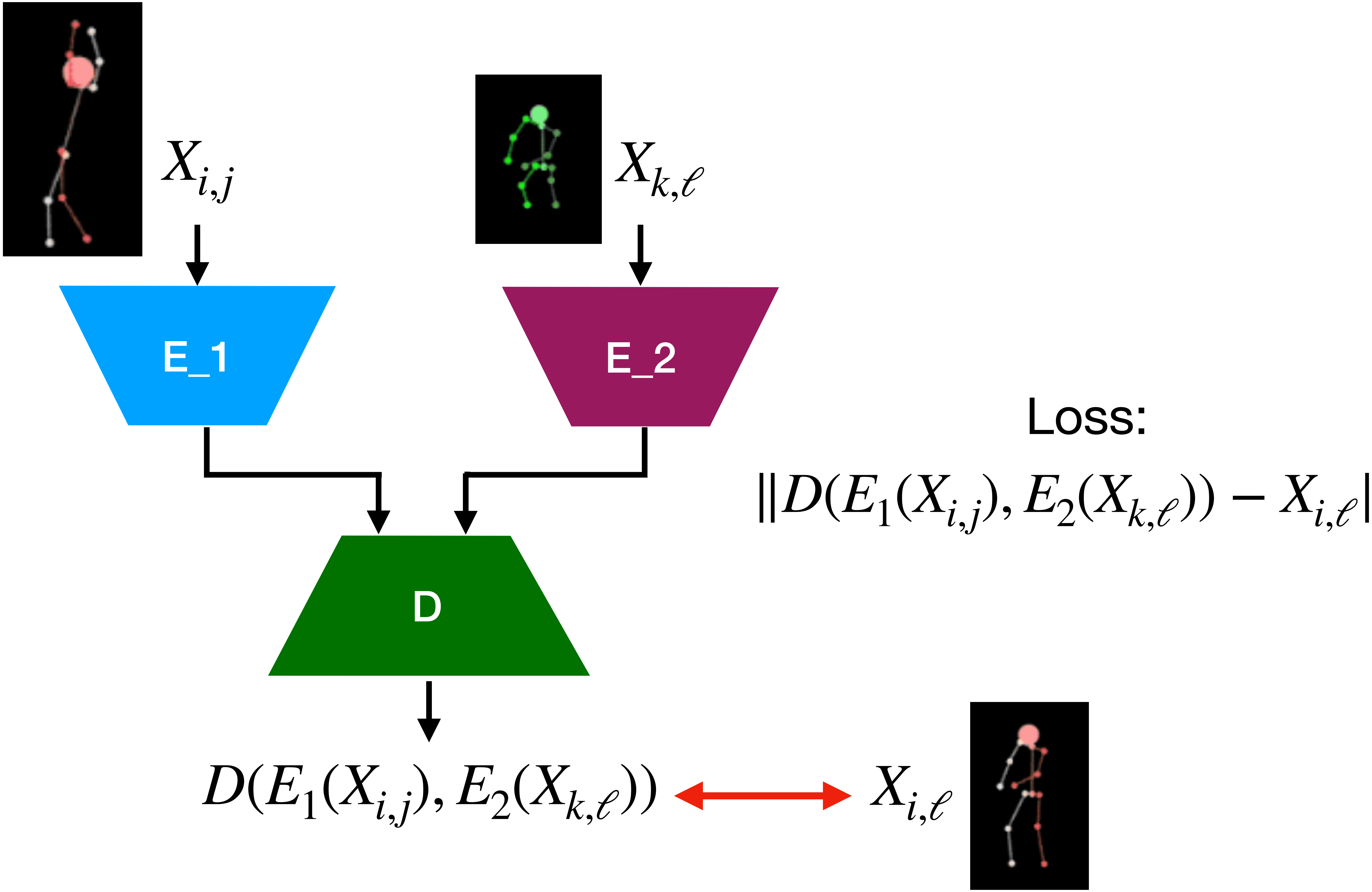
Fully-Supervised

Example: $\{X_{a,b}\}_{a \in A, b \in B}$

Let A and B be two independent factors of variation, then the dataset contains triplets:

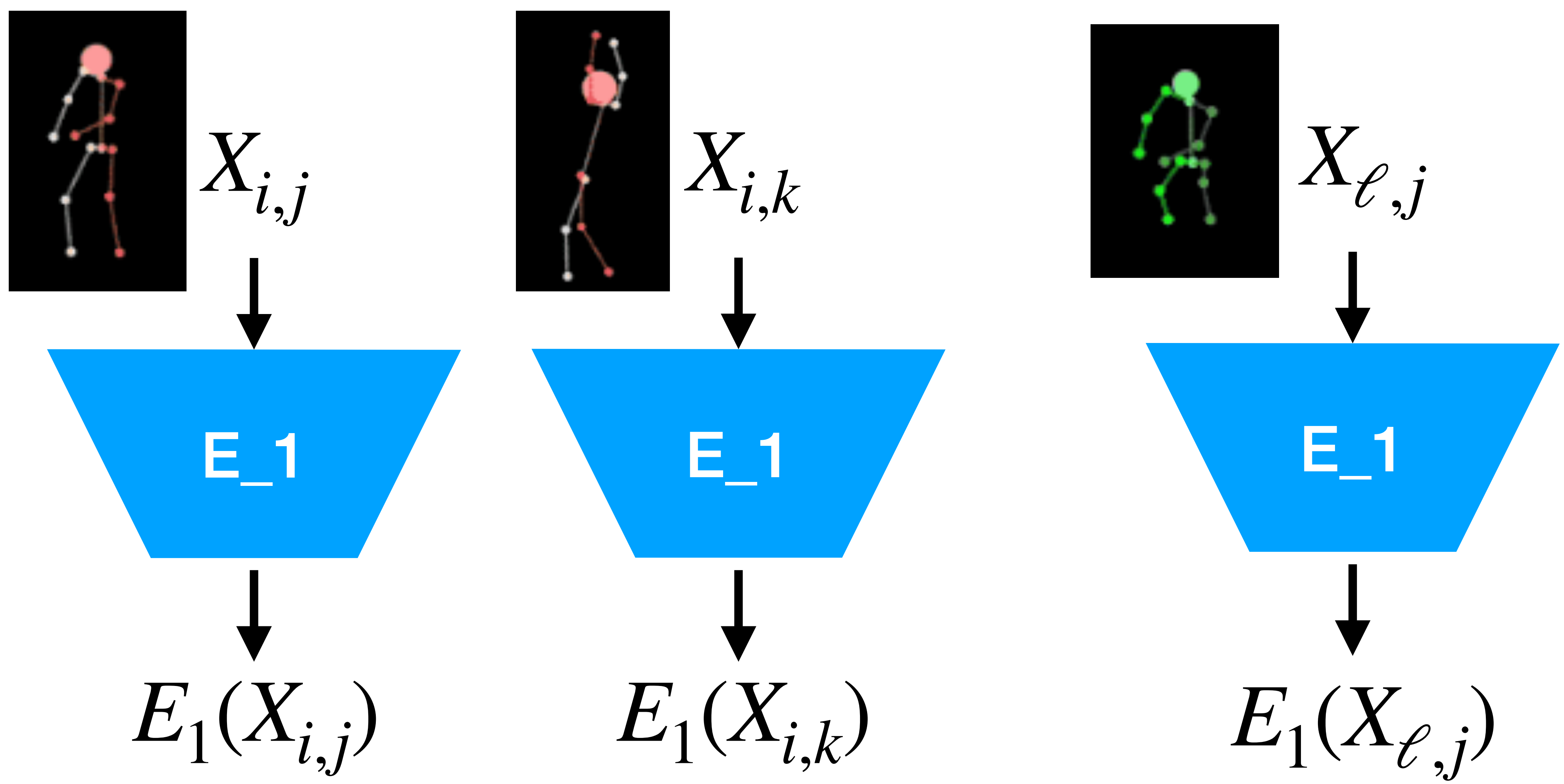
$$(X_{i,j}, X_{i,k}, X_{\ell,j})$$

Cross Loss



$X_{i,j}$
$X_{k,\ell}$
E_1
E_2
D
Loss:
$\|D(E_1(X_{i,j}), E_2(X_{k,\ell})) - X_{i,\ell}\|$
$D(E_1(X_{i,j}), E_2(X_{k,\ell}))$
$X_{i,\ell}$

Triplet Loss



Loss:

$$\max\left(\|E_1(X_{i,j}) - E_1(X_{i,k})\| - \|E_1(X_{i,j}) - E_1(X_{\ell,j})\| + \varepsilon, 0\right)$$

Supervision

Unsupervised

Limited use

?

Fully-Supervised

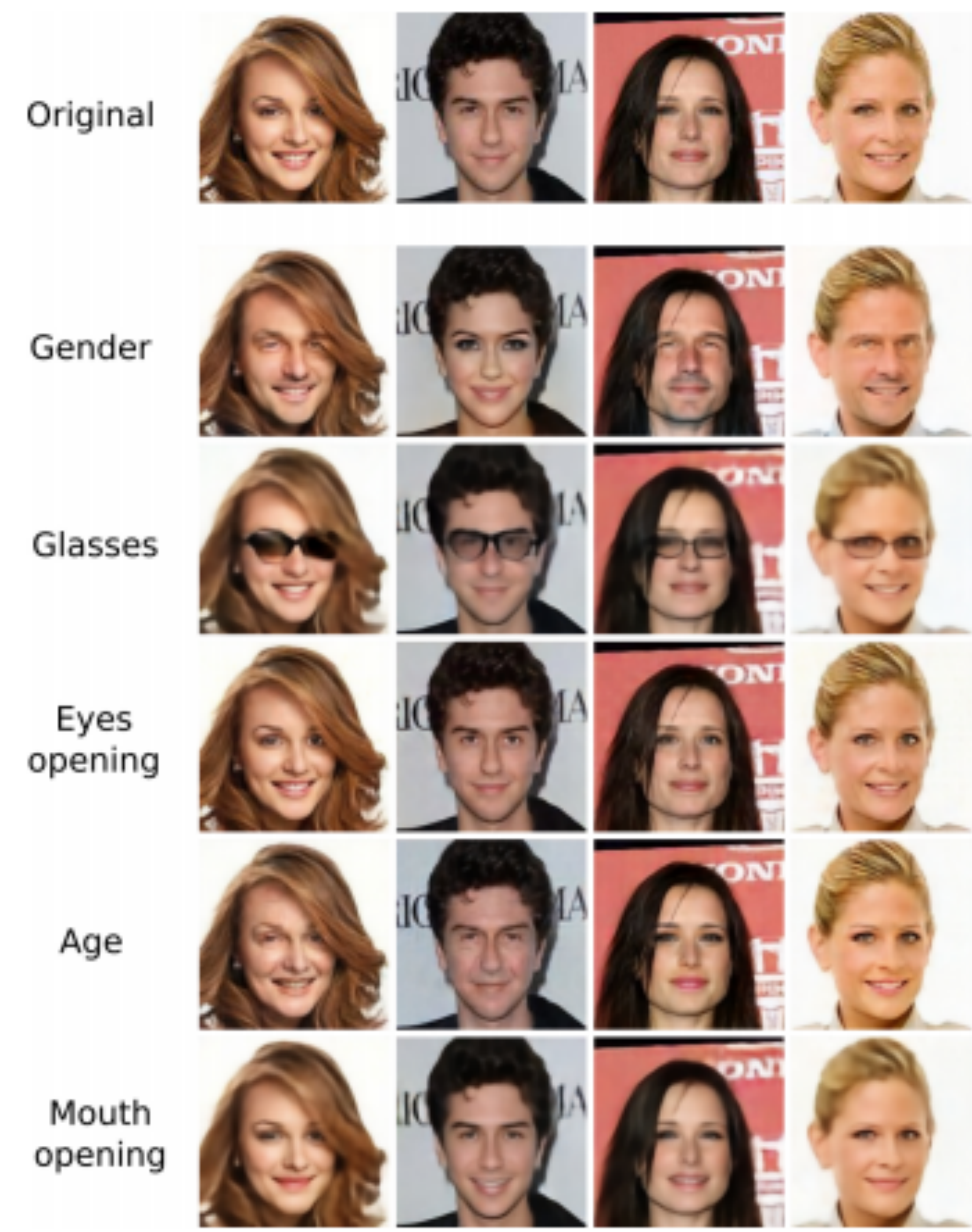
Only synthetic datasets

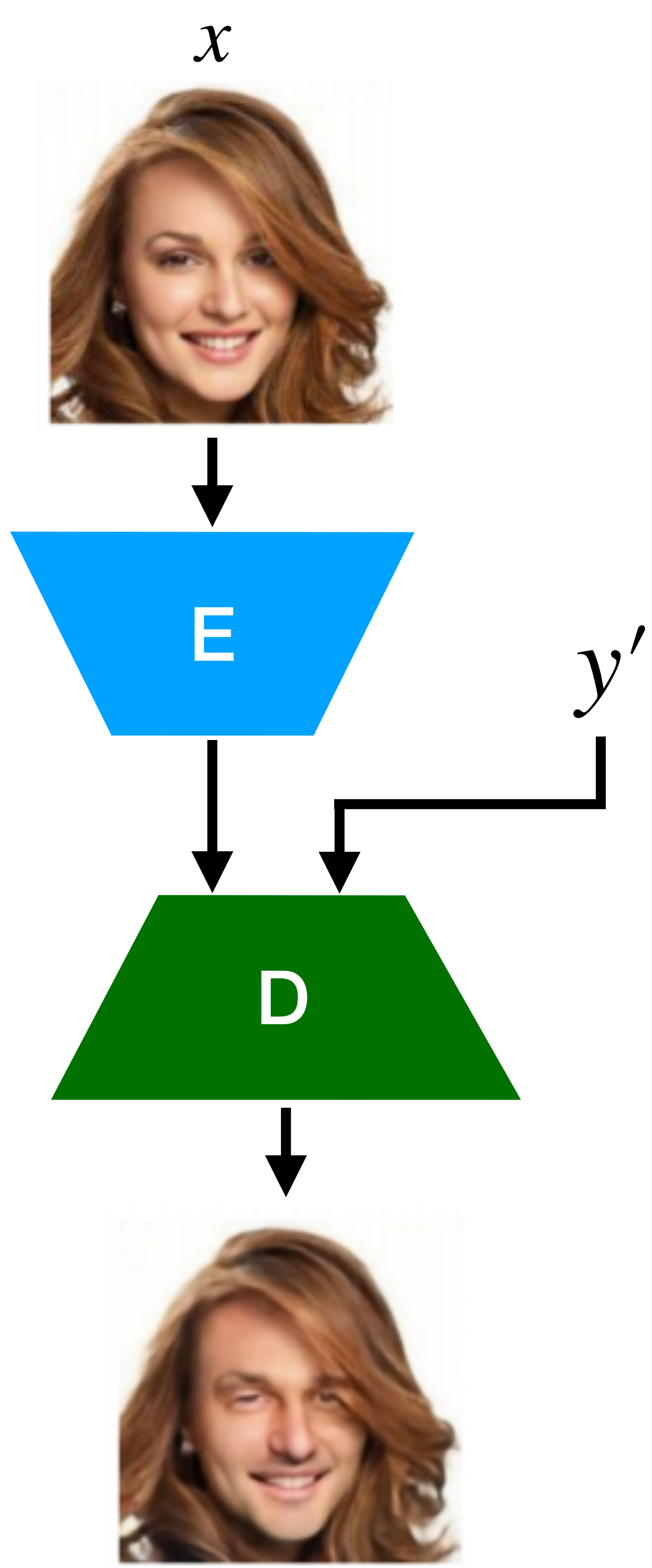
Fader Networks: Manipulating Images by Sliding Attributes

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, Marc'Aurelio Ranzato

Set of images and several attributes for each image.

$$\{(x, y)\}_{x \in X, y \in \{0,1\}^k}$$

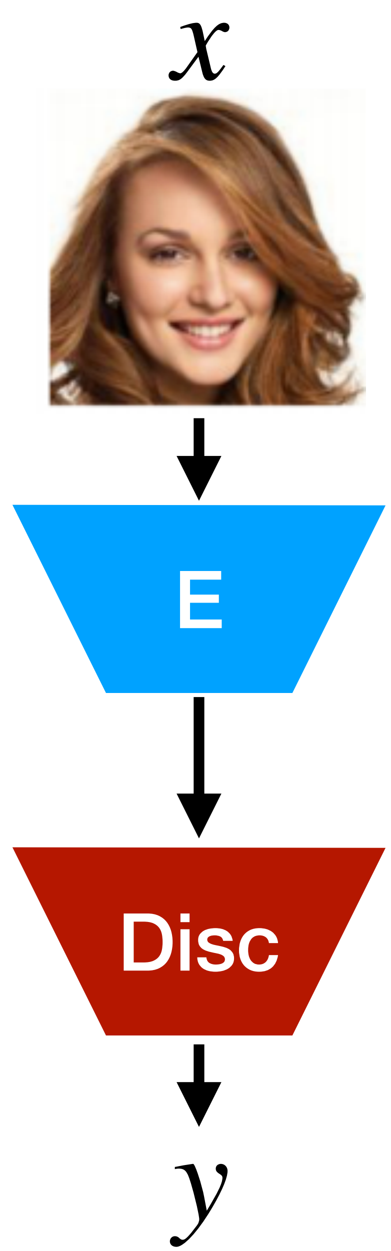




We would like to find representation of the images which is invariant to the image attributes.

Adversarial Loss

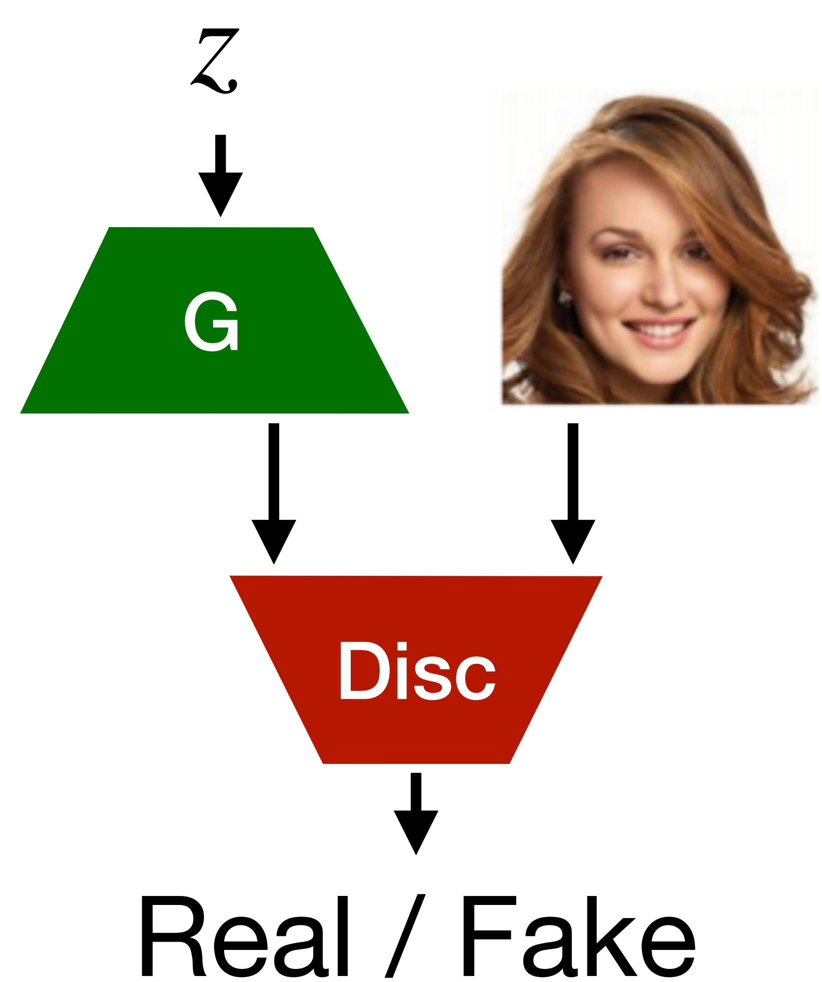
FaderNet:



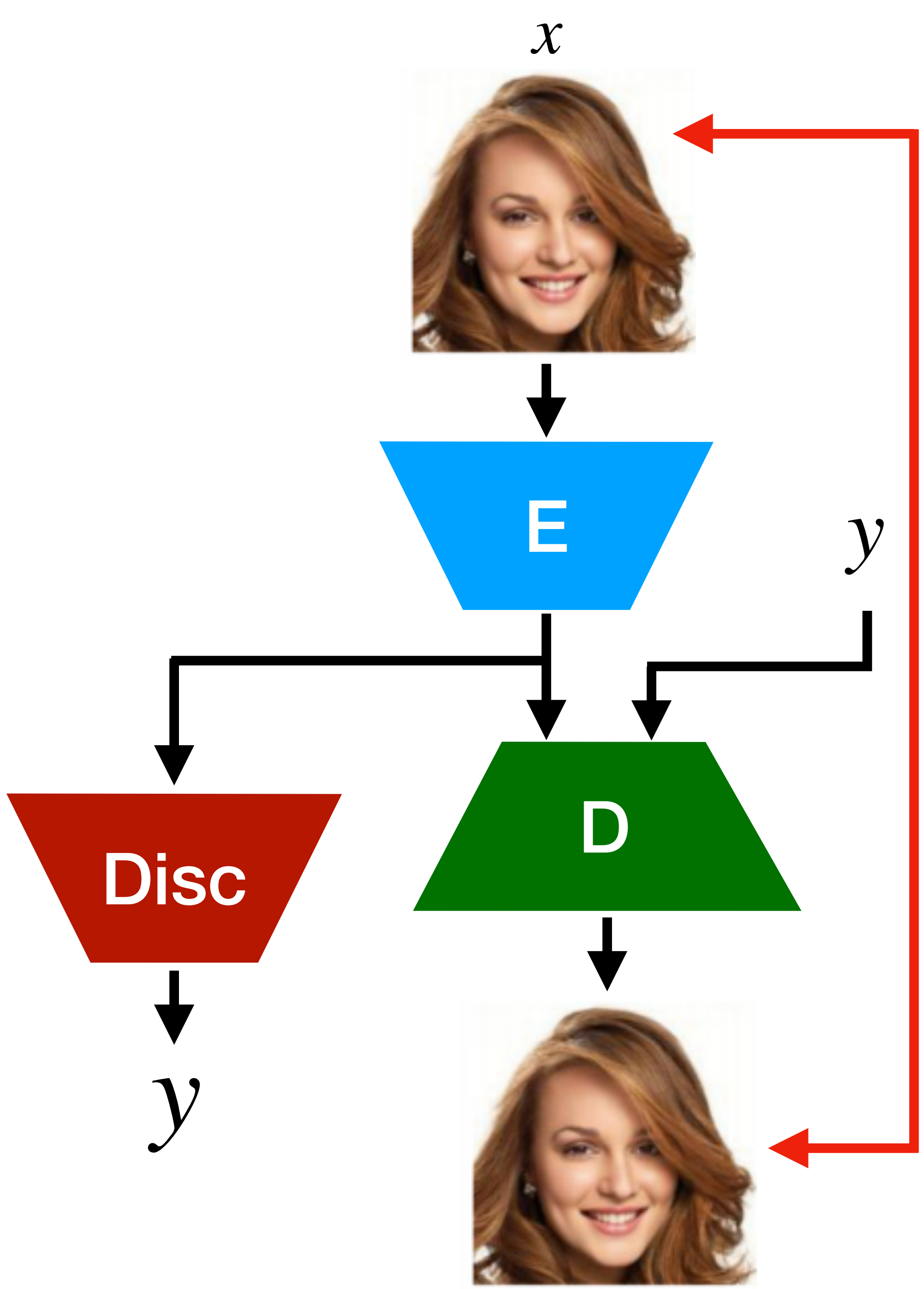
Encoder try to fool the discriminator.

Discriminator try to predict y.

GAN:



Loss Function



1. Adversarial loss:

Discriminator try to predict y. Encoder try to fool the discriminator.

2. Reconstruction loss:

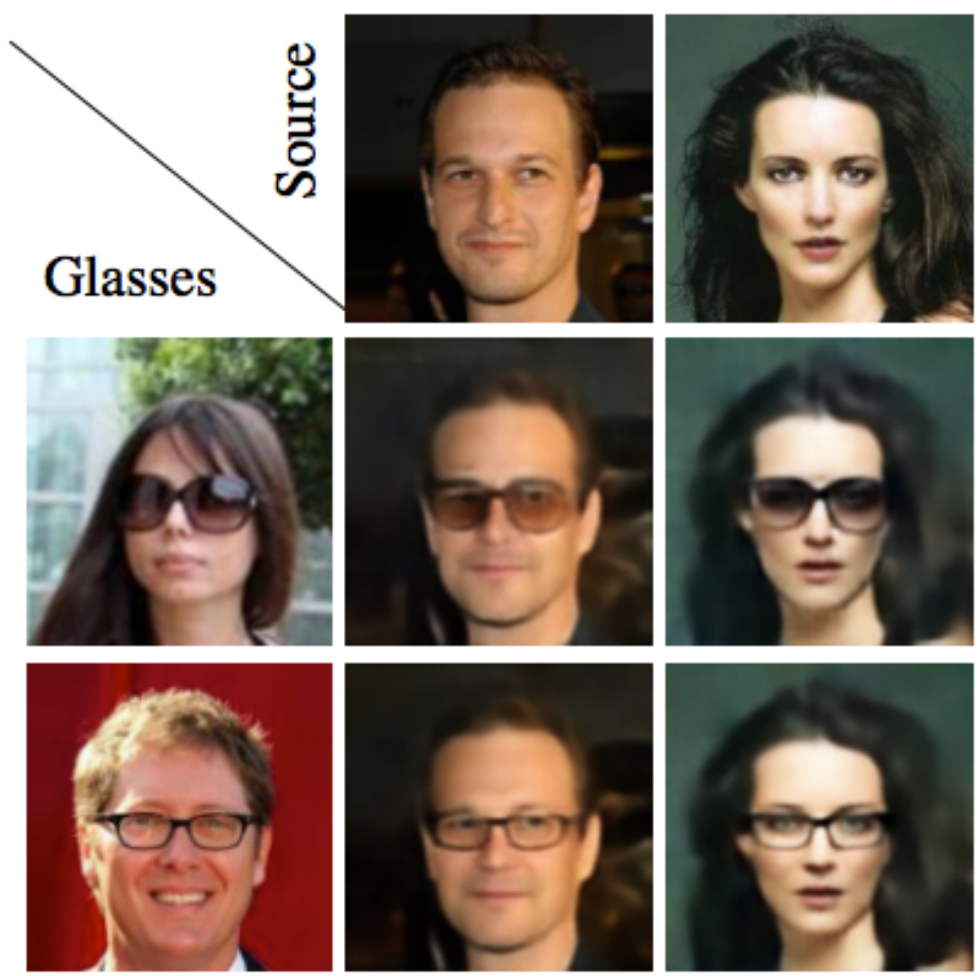
$$\mathscr{L}_R = \|D(E(x), y) - x\|$$

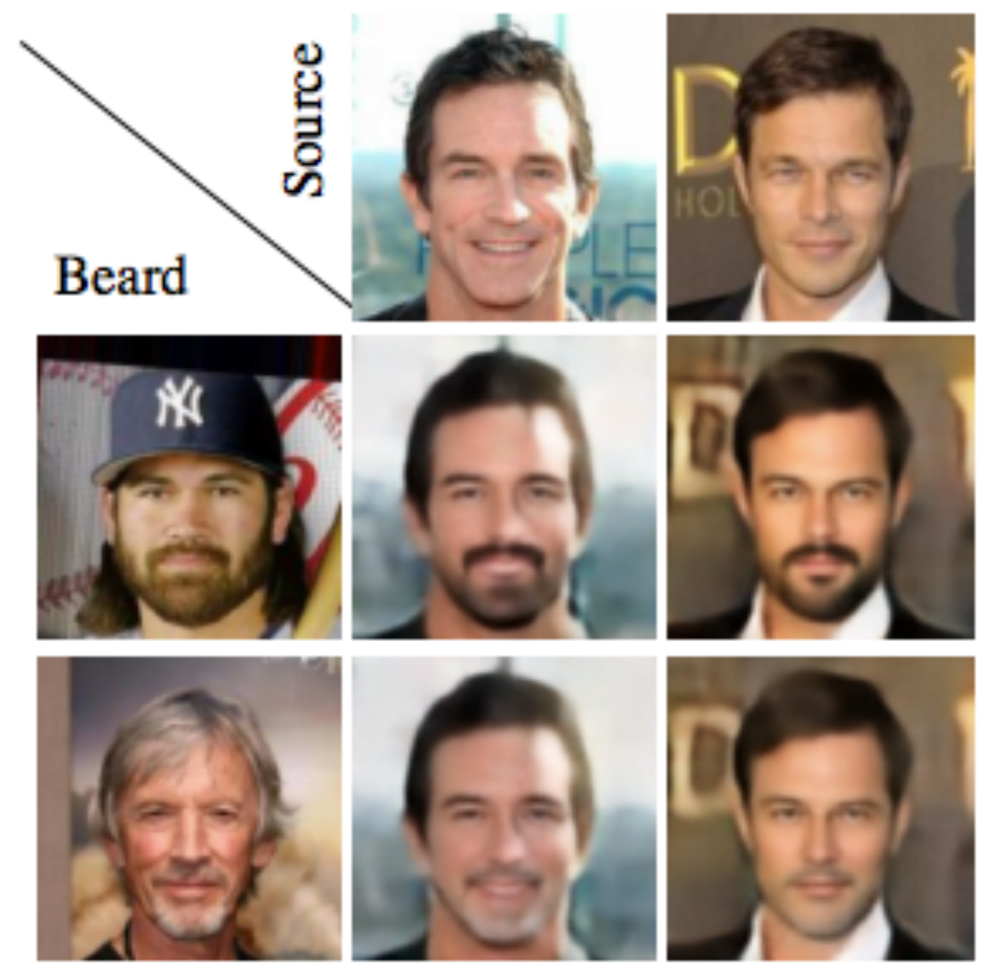
Sliding

EMERGING DISENTANGLEMENT IN AUTO-ENCODER BASED UNSUPERVISED IMAGE CONTENT TRANSFER

Ori Press, Tomer Galanti, Sagie Benaim, Lior Wolf

Content transfer:

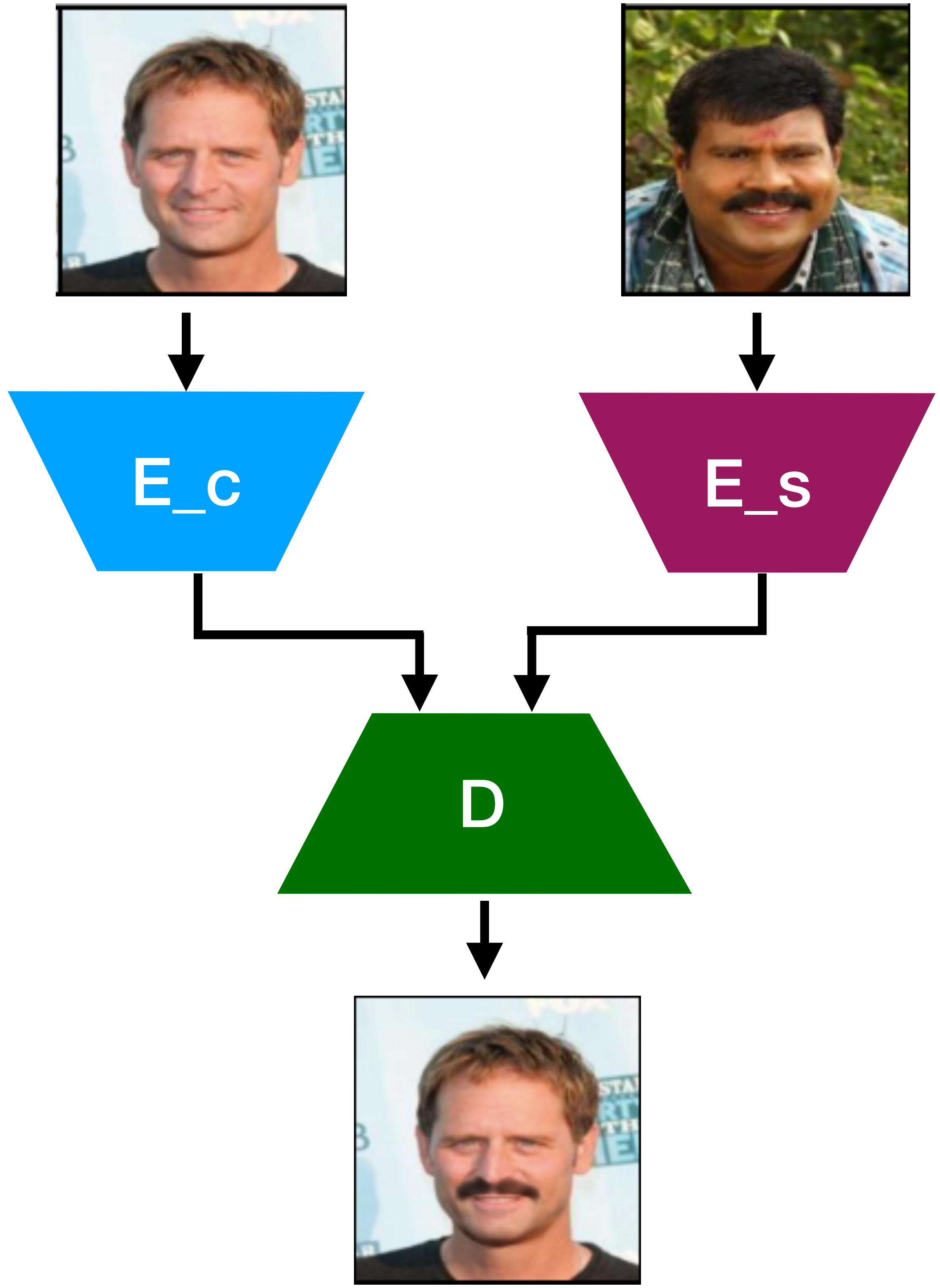




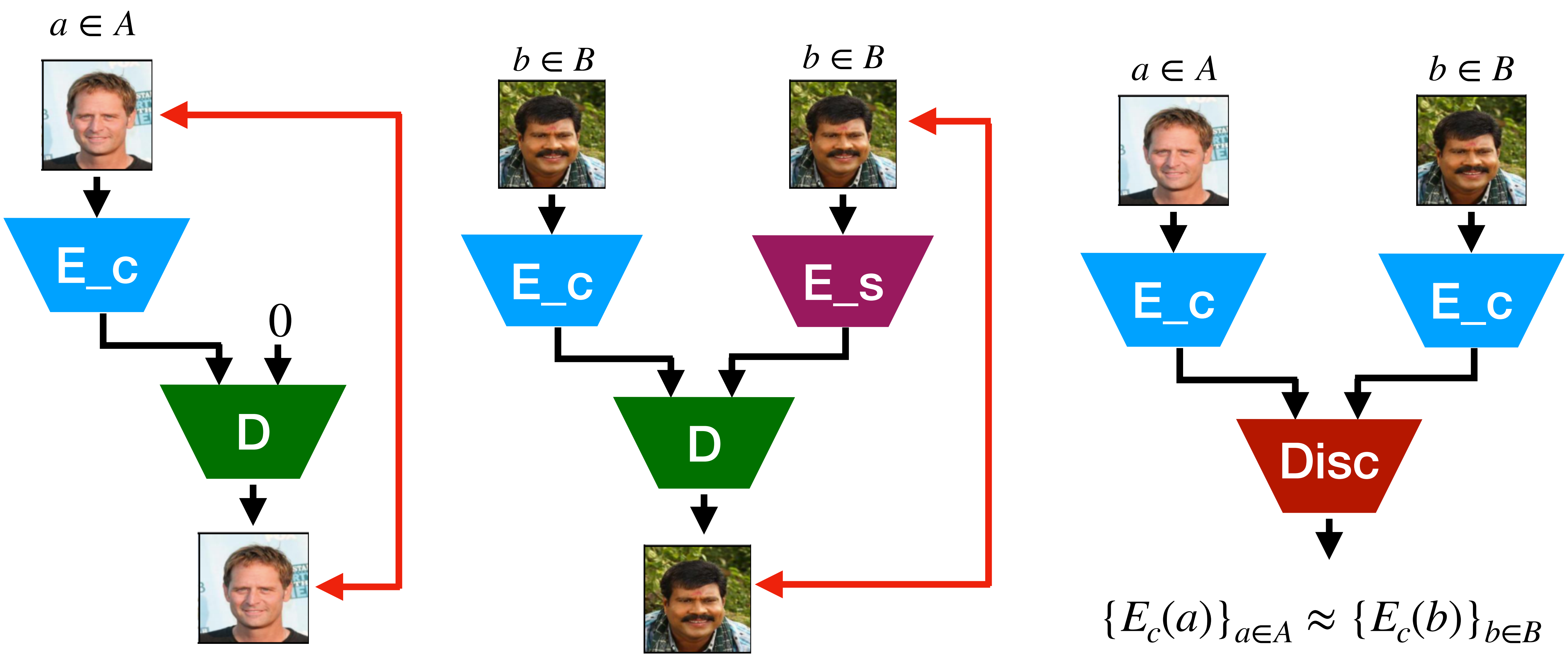
Unsupervised Content Transfer

Domain A contains common content

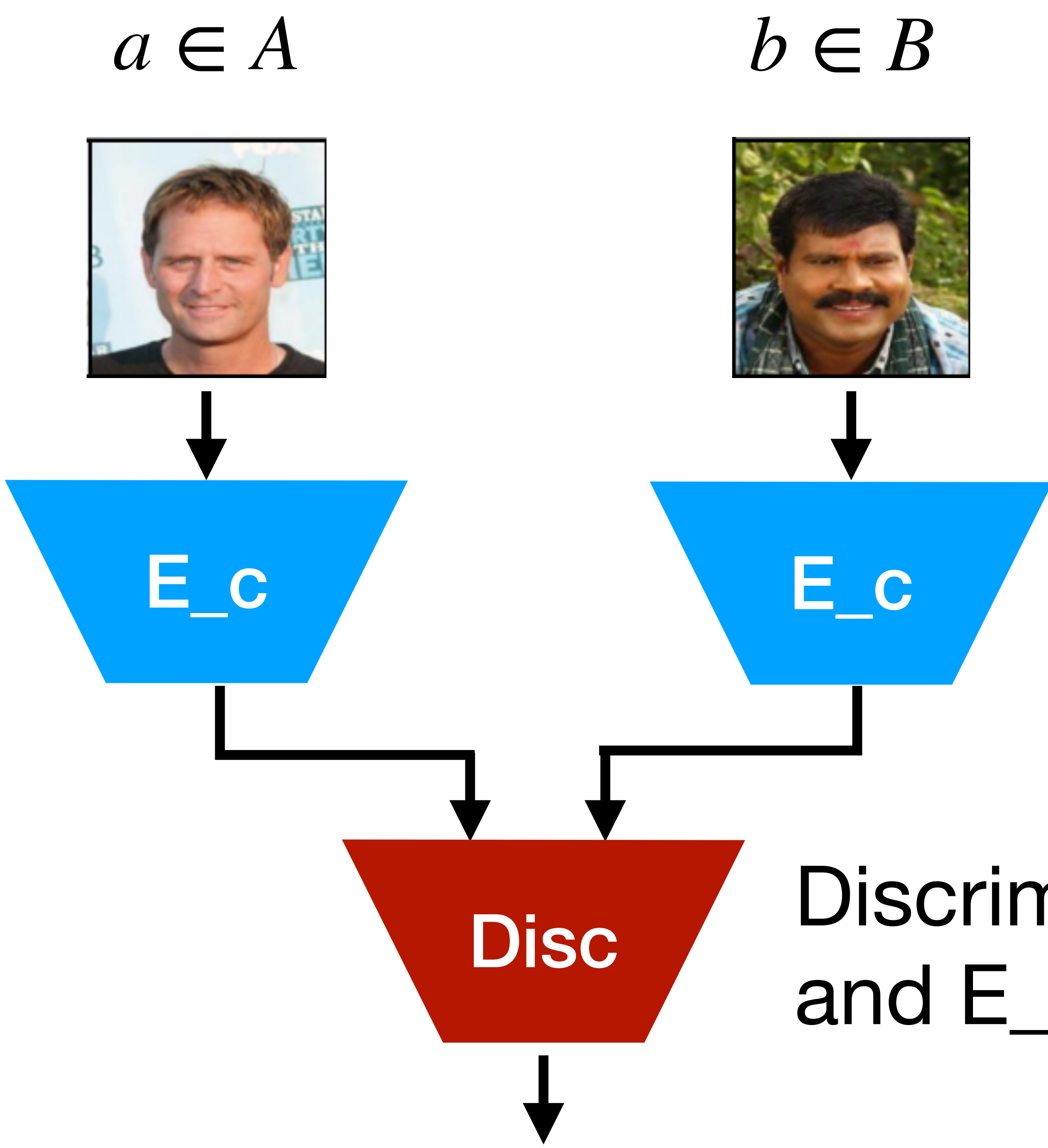
Domain B contains common and separate content



E_c
E_s
D


$a \in A$
E_c
0
D
$b \in B$
$b \in B$
E_c
E_s
D
$a \in A$
$b \in B$
E_c
E_c
Disc
$\{E_c(a)\}_{a \in A} \approx \{E_c(b)\}_{b \in B}$

Domain Confusion



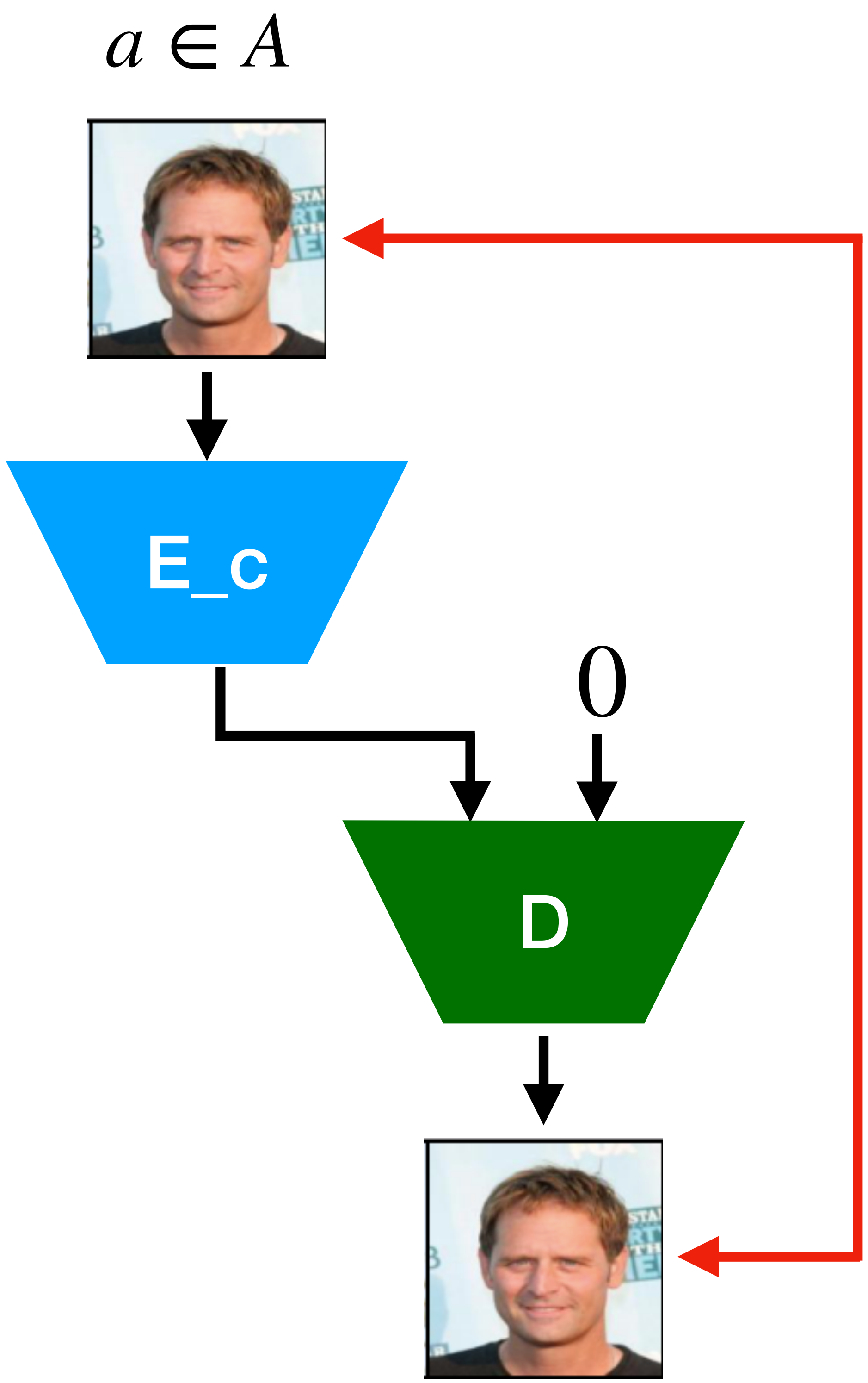
Goal:

$$\{E_c(a)\}_{a \in A} \approx \{E_c(b)\}_{b \in B}$$

Encoder try to fool the discriminator

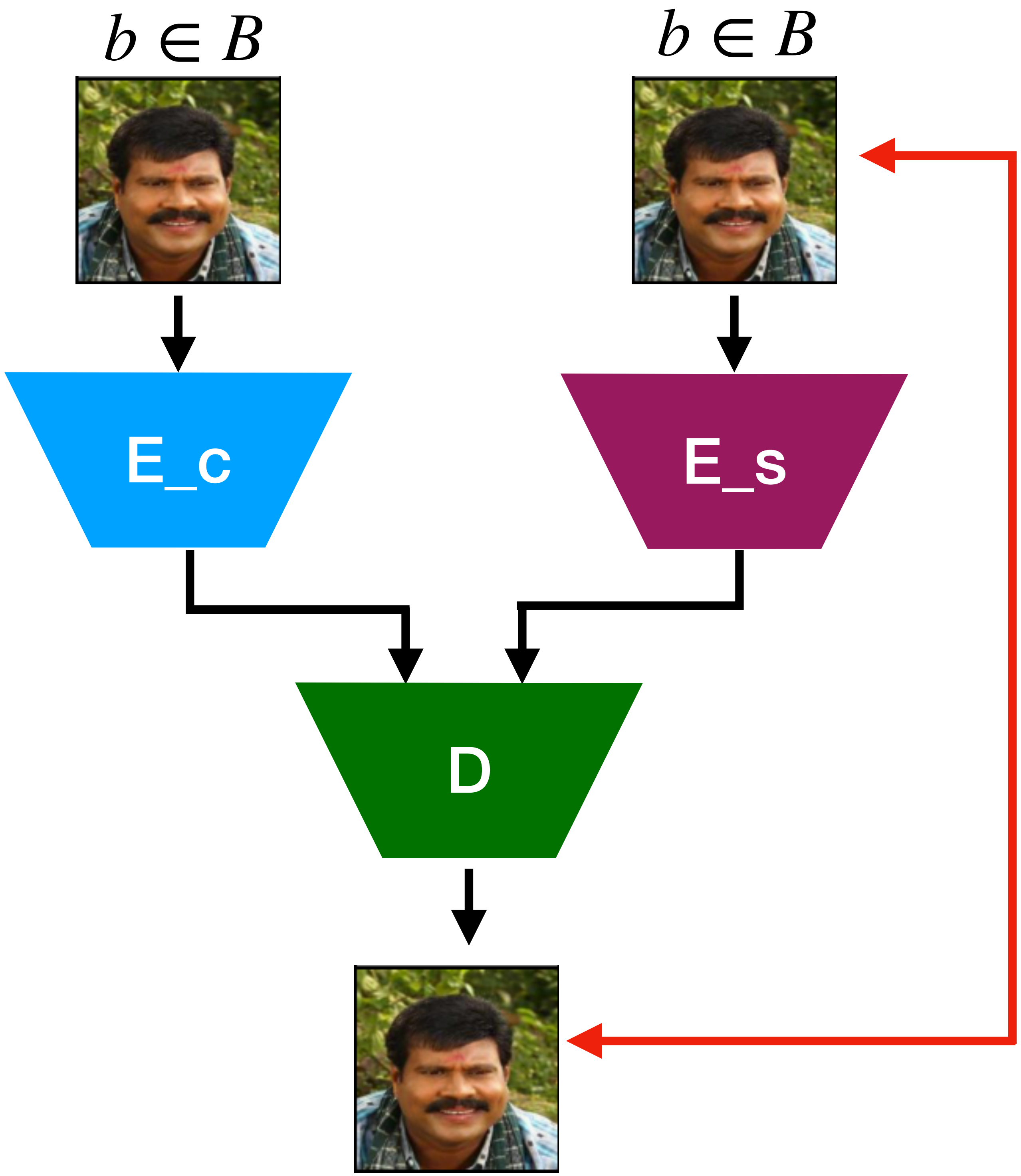
Discriminator try to distinguish E_c(a) and E_c(b)

Reconstruction



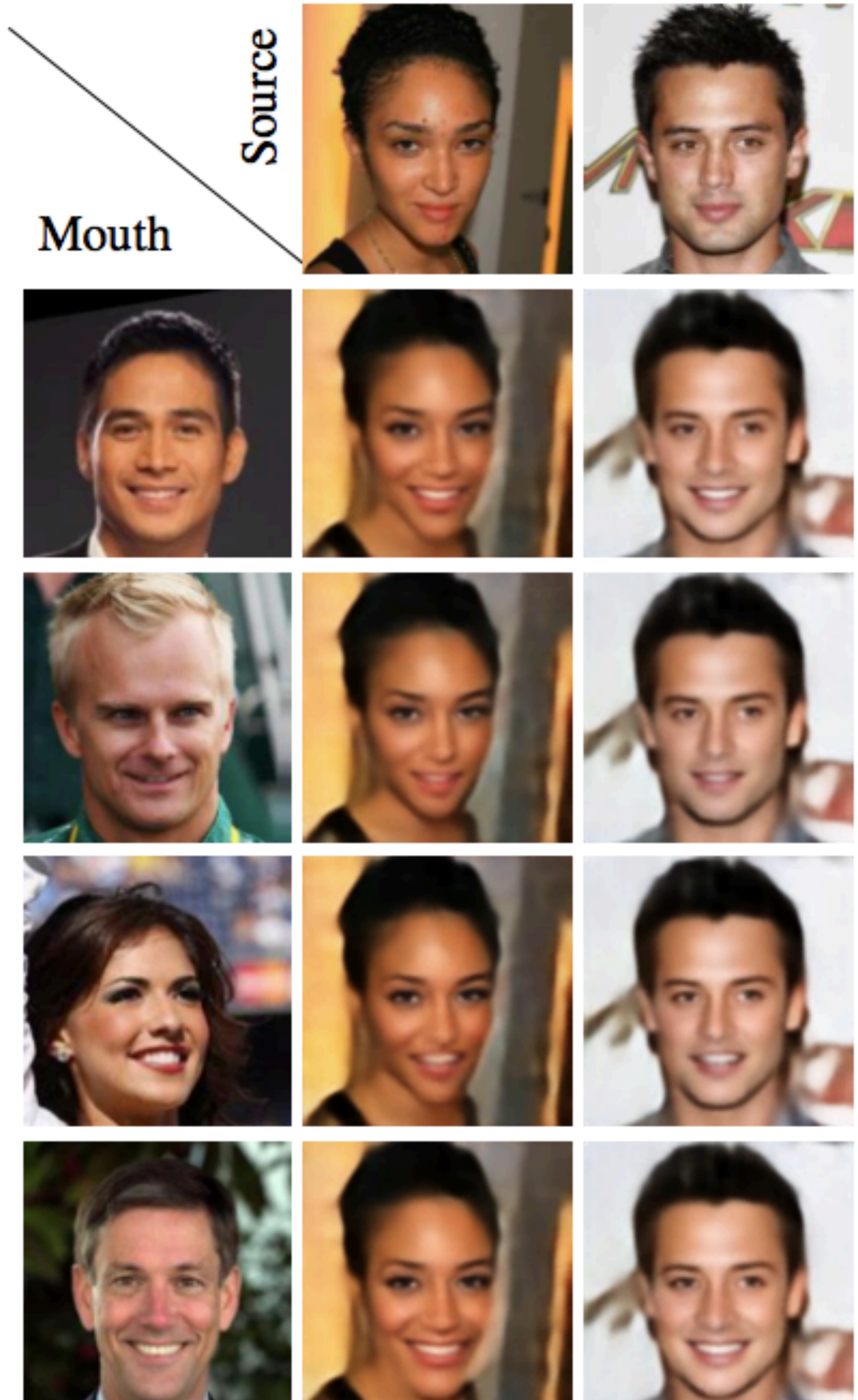
$$\mathscr{L}_R^A = \|D(E_c(a),0) - a\|$$

Reconstruction 2

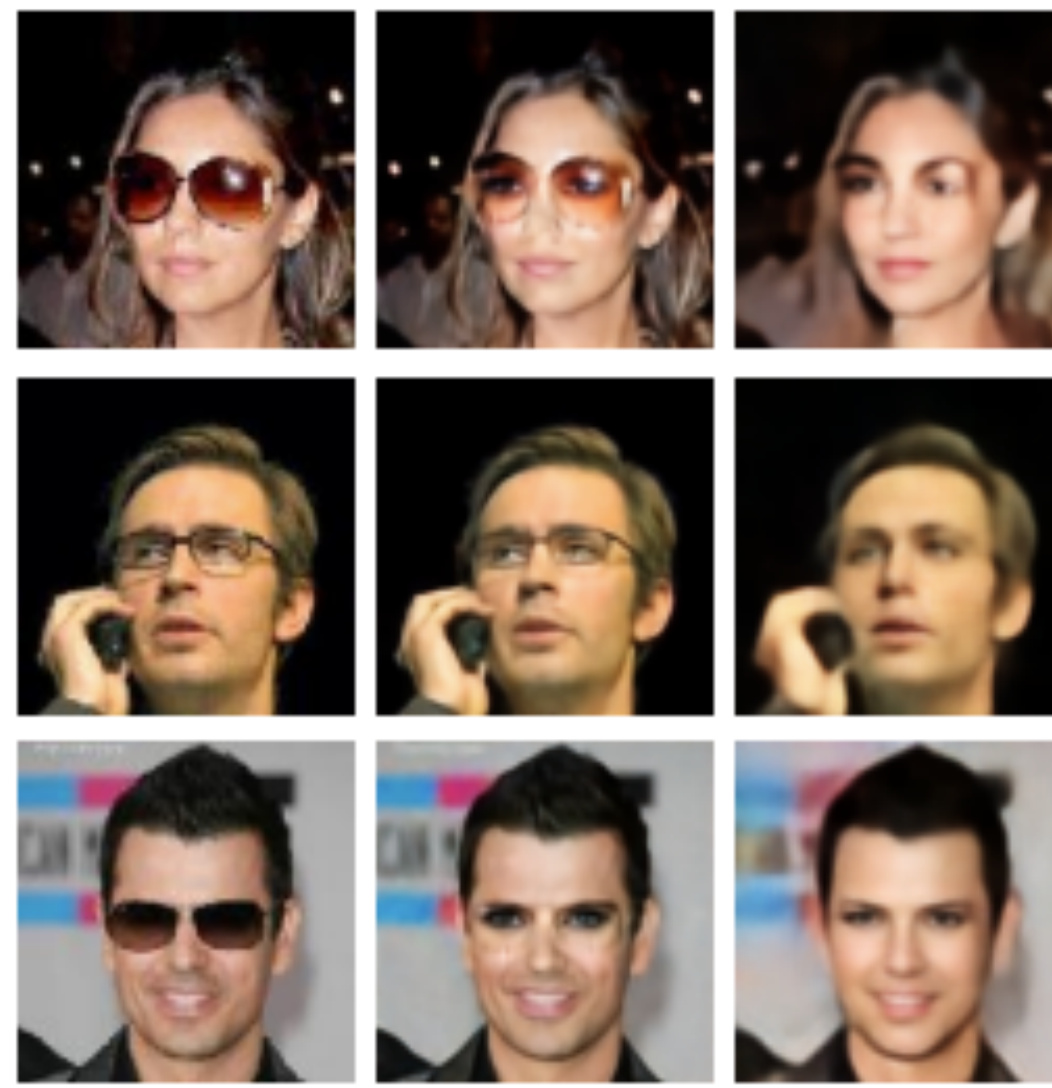


$$\mathscr{L}_R^A = \|D(E_c(b), E_c(b)) - b\|$$


Source
Beard


Source
Mouth

Original Fader Our



Original Fader Our

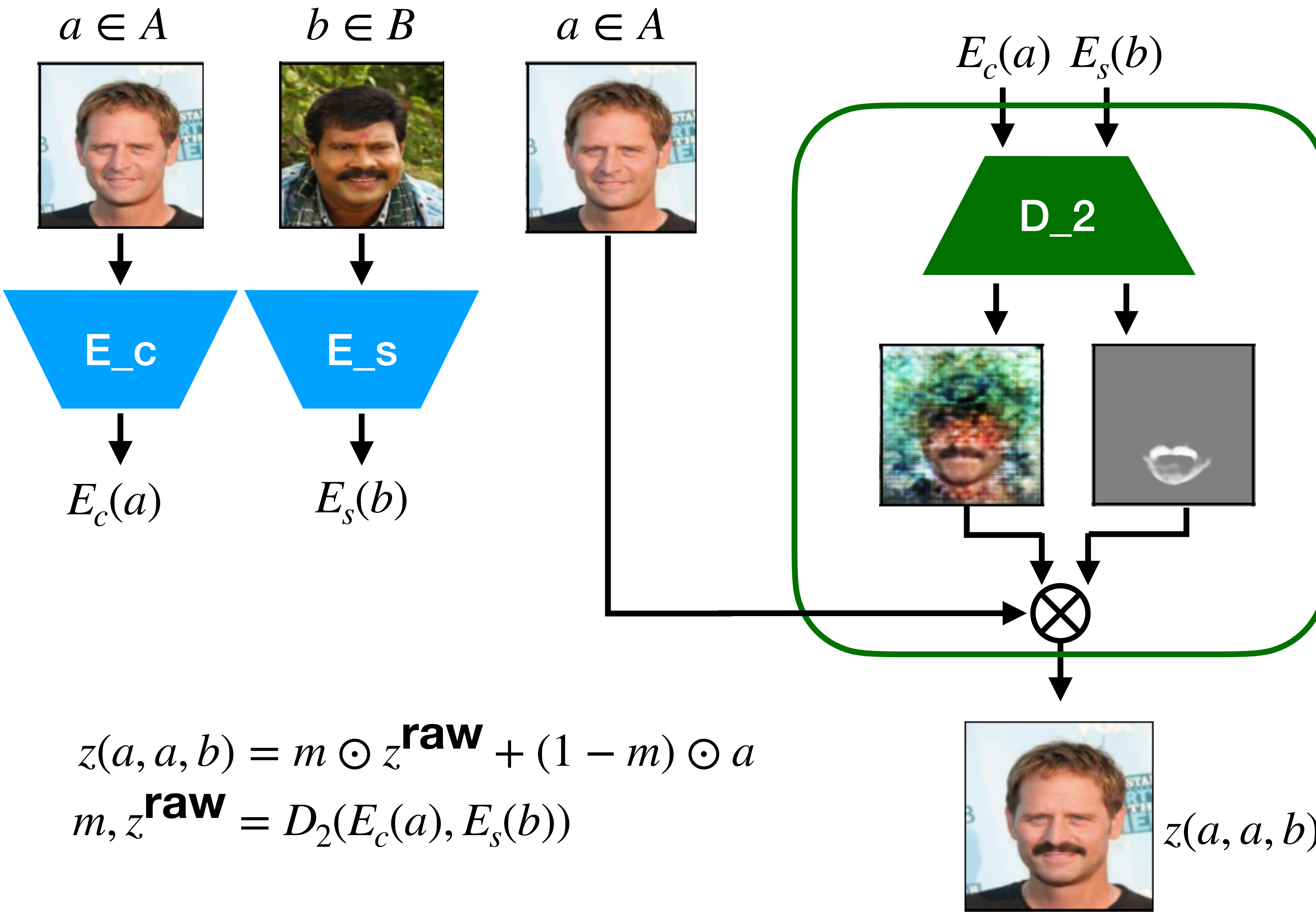
MASK BASED UNSUPERVISED CONTENT TRANSFER

Ron Mokady, Sagie Benaim, Lior Wolf, Amit Bermano

- Transfer content using a mask for state-of-the-art quality.
- Use the mask as a weakly-supervised semantic segmentation.

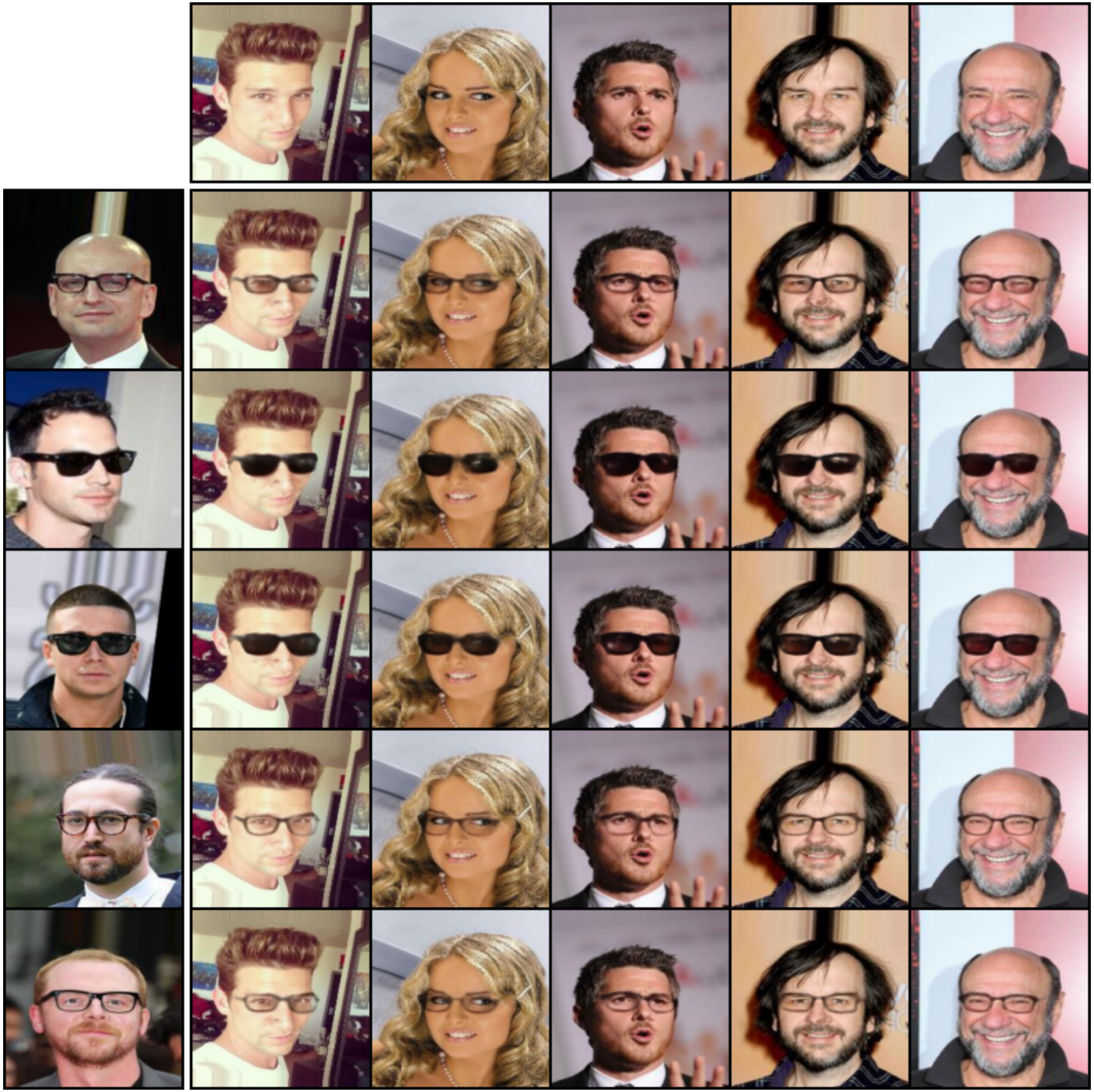
Comparison

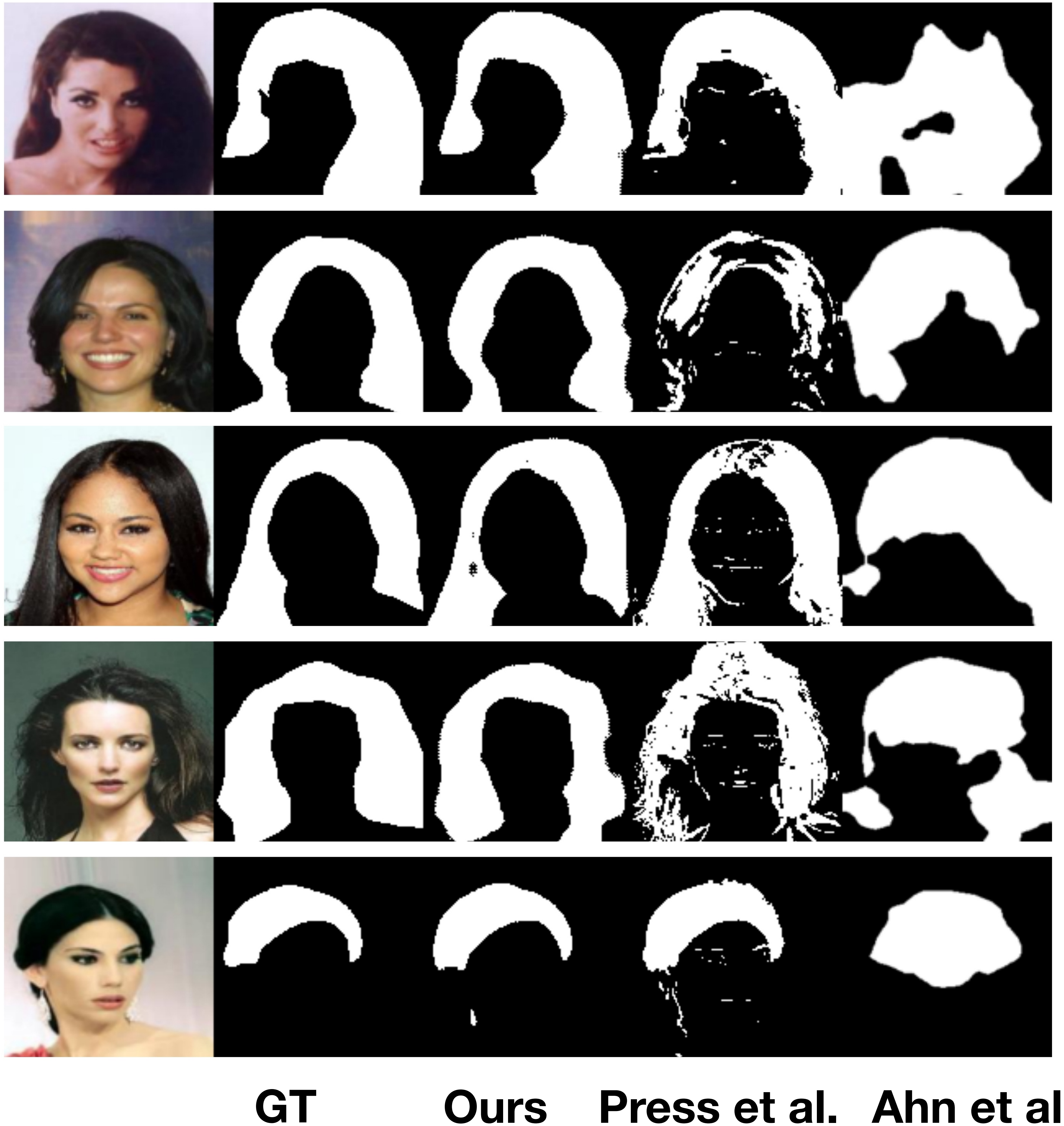
Using a mask

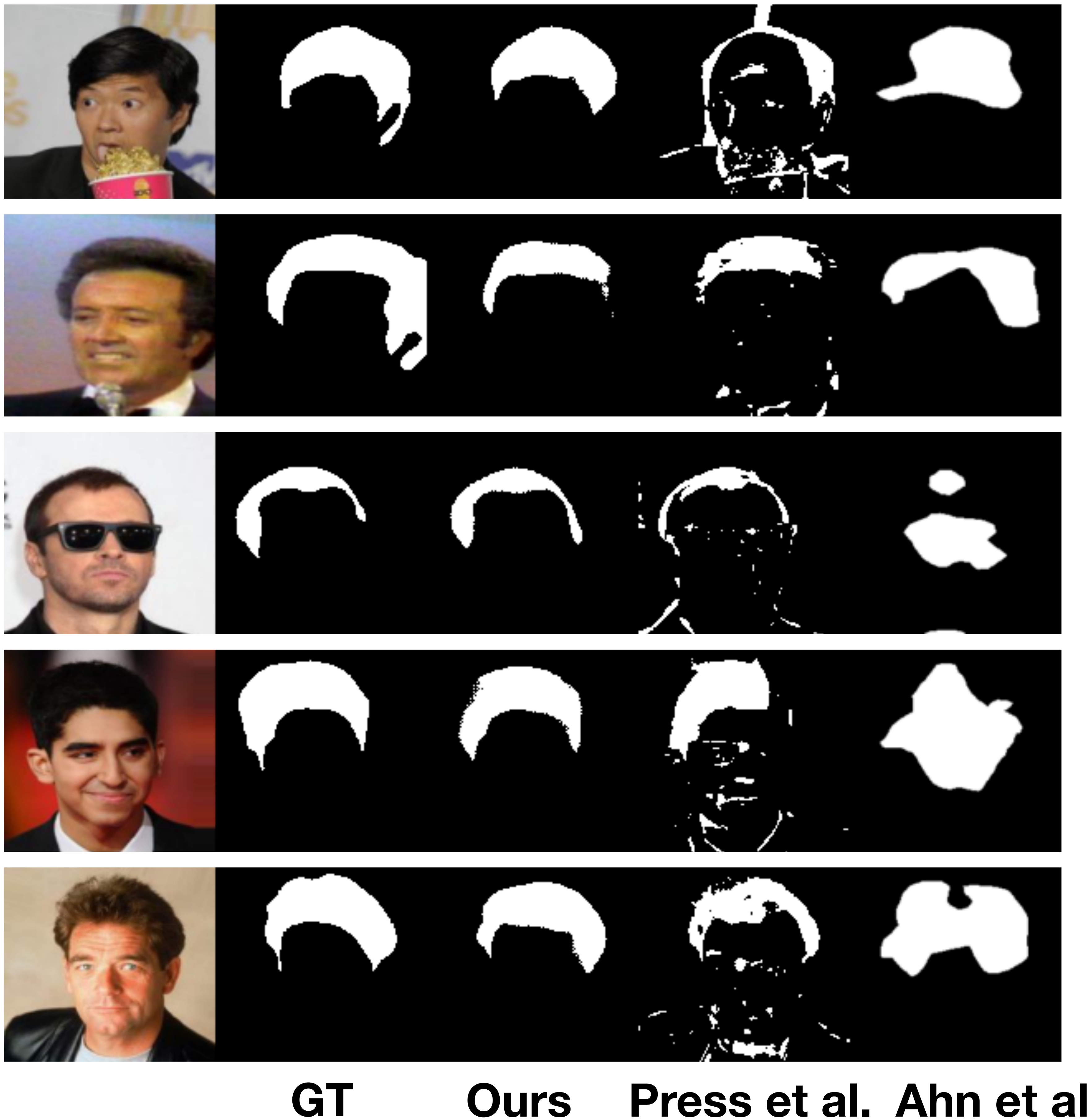


$$z(a, a, b) = m \odot z^{\textbf{raw}} + (1 - m) \odot a$$

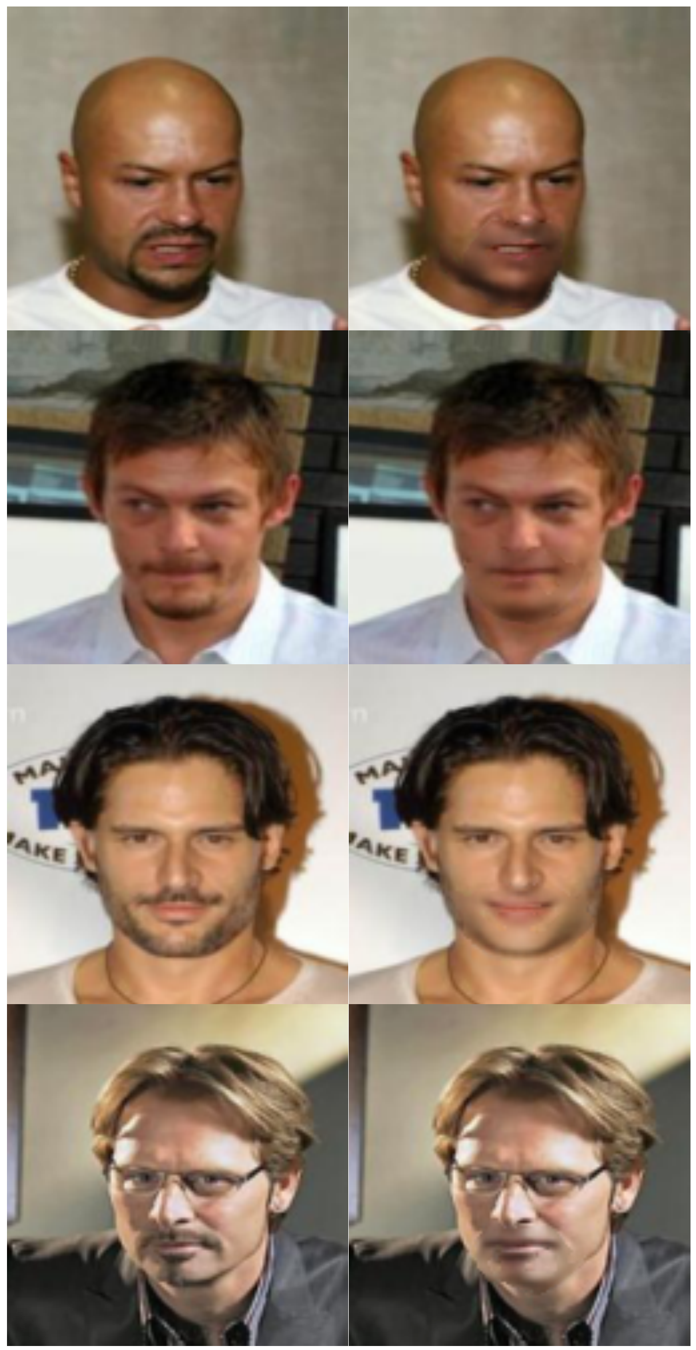
$$m, z^{\textbf{raw}} = D_2(E_c(a), E_s(b))$$




GT
Ours
Press et al.
Ahn et al.


GT
Ours
Press et al.
Ahn et al.

Removal



Generate Unseen Images

Domain A: glasses without mustache.

Domain B: mustache without glasses.

Can we generate face with both mustche and glasses? Or one without mustache and glasses?

Domain Intersection and Domain Difference

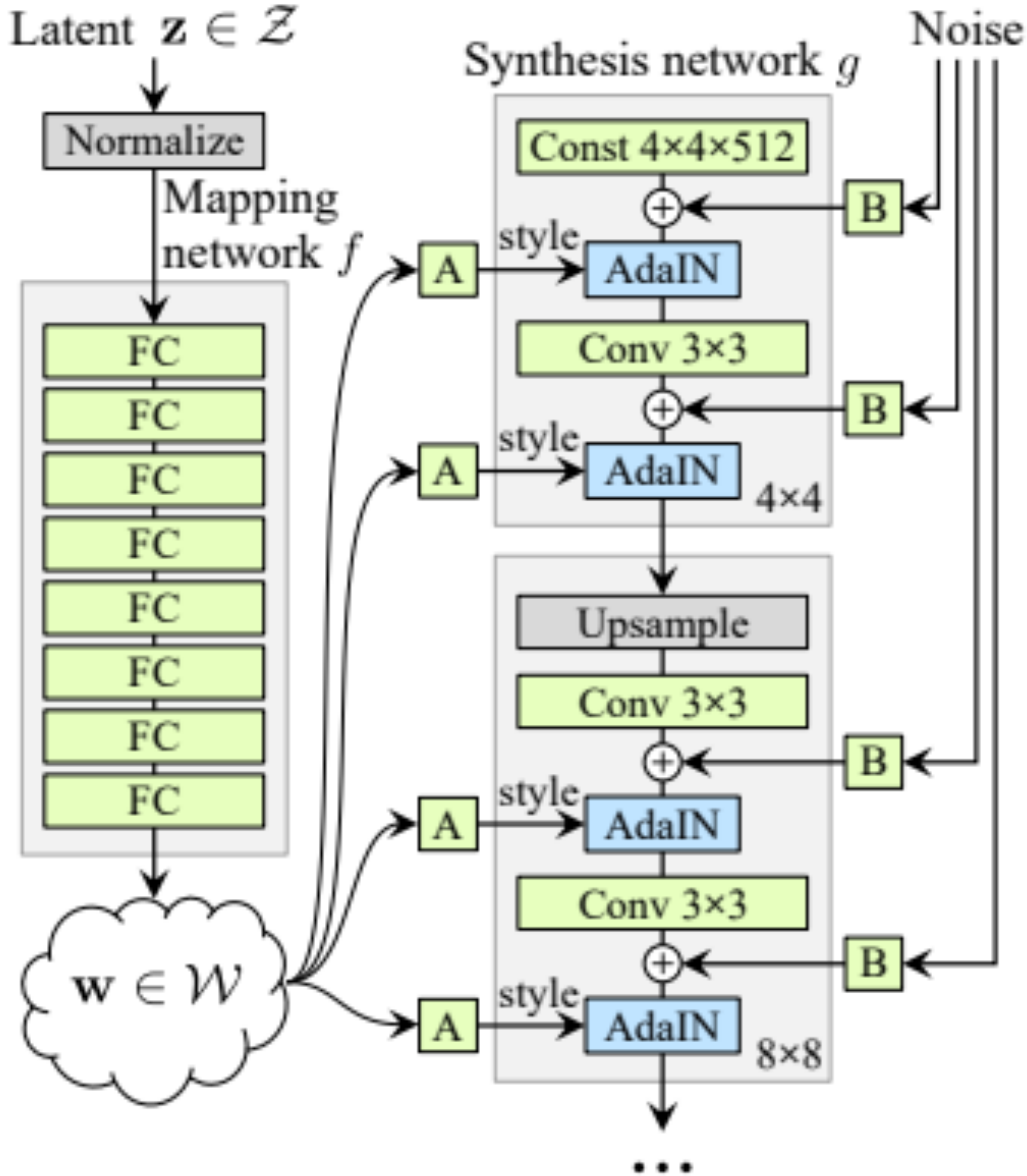
Sagie Benaim, Michael Khaitov , Tomer Galanti, Lior Wolf

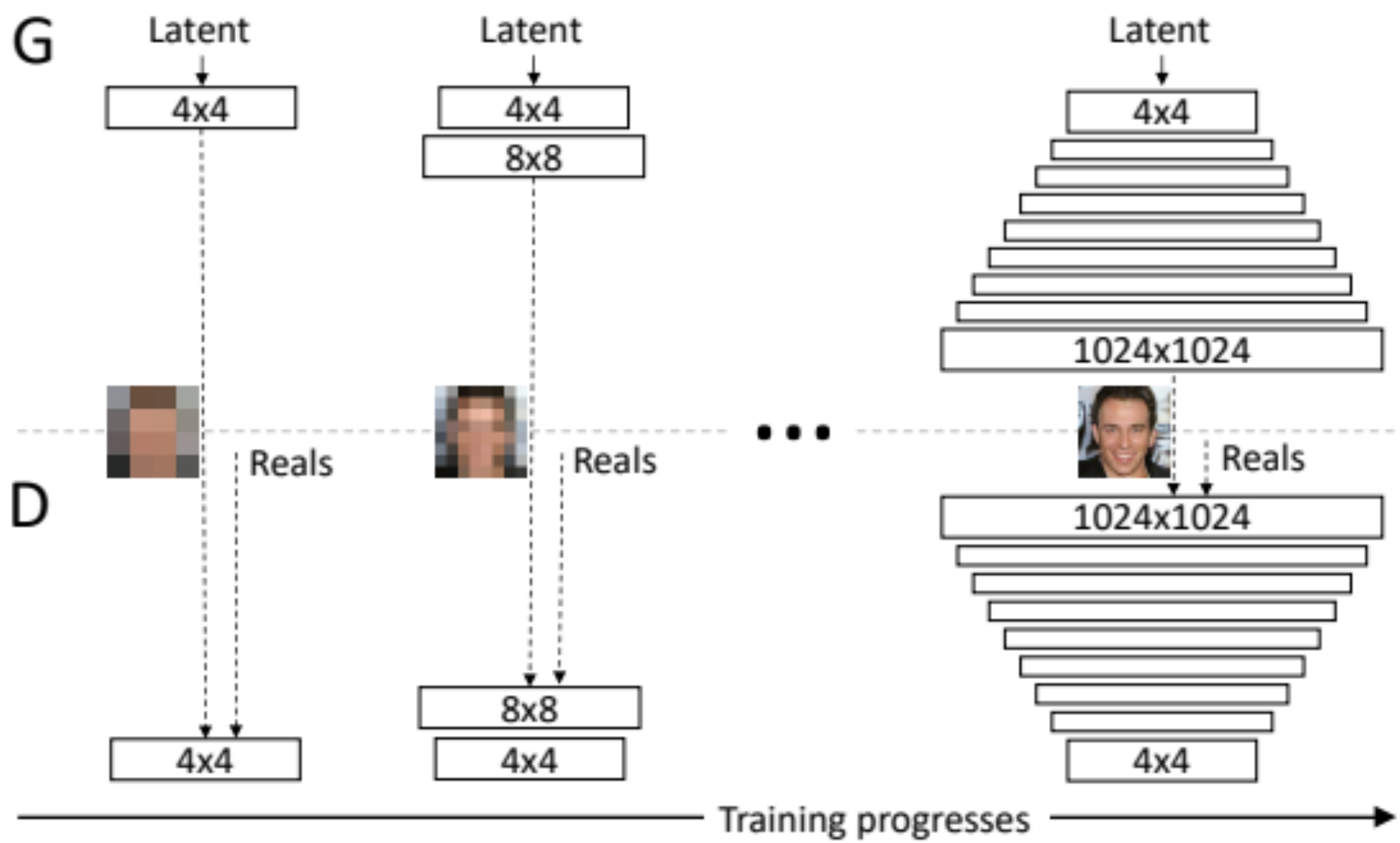
Three encoders, one for the common and two for the attributes.

Same Reconstruction and Domain-Confusion loss as Press et el. with Additional zero loss.

A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras, Samuli Laine, Timo Aila


Latent $\mathbf{z} \in \mathcal{Z}$
Normalize
Mapping network f
FC
FC
FC
FC
FC
FC
FC
FC
$\mathbf{w} \in \mathcal{W}$
Synthesis network g
Noise
Const 4×4×512
B
A
style
AdaIN
Conv 3×3
B
A
style
AdaIN
4×4
Upsample
Conv 3×3
B
A
style
AdaIN
Conv 3×3
B
A
style
AdaIN
8×8
...

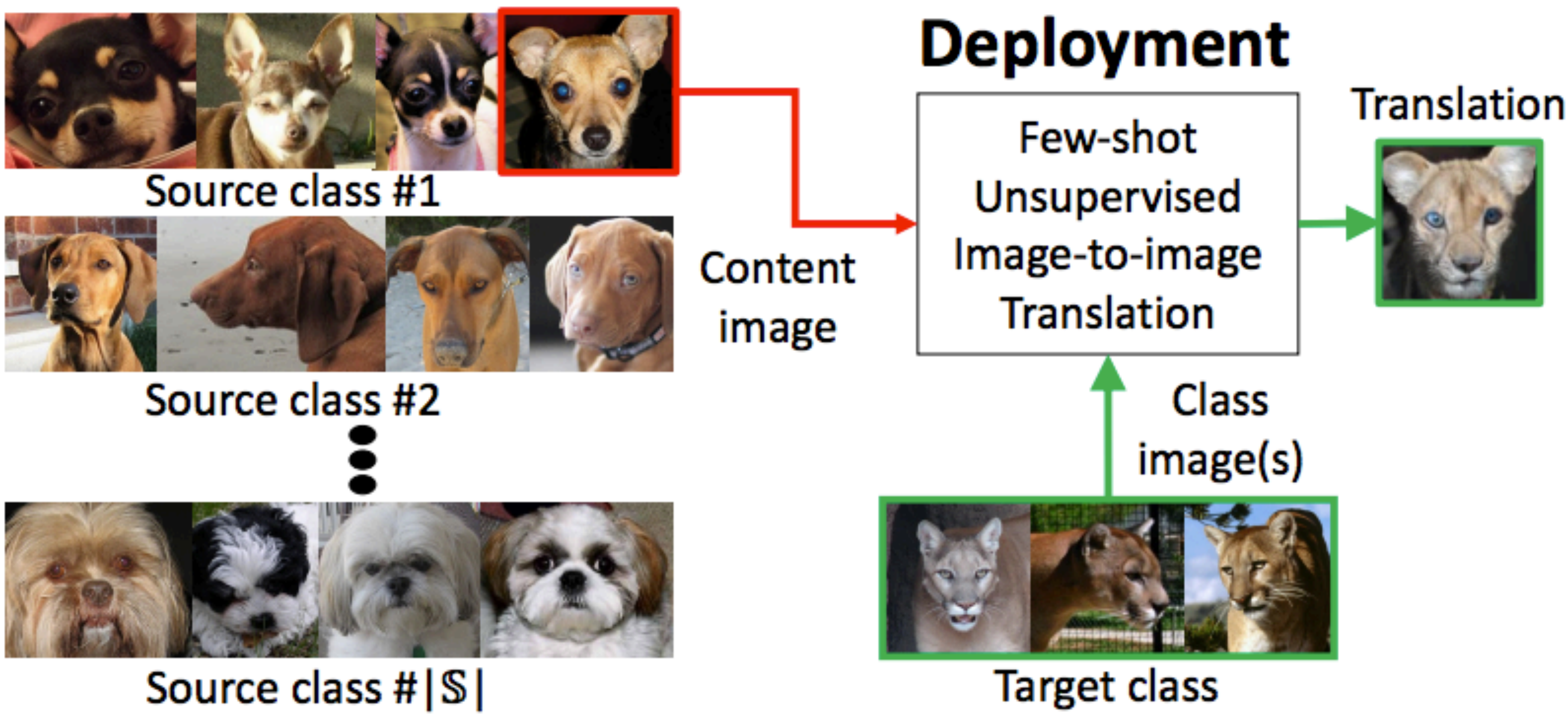

G
Latent
4x4
Latent
4x4
8x8
Latent
4x4
1024x1024
Reals
Reals
Reals
D
4x4
8x8
4x4
1024x1024
4x4
Training progresses

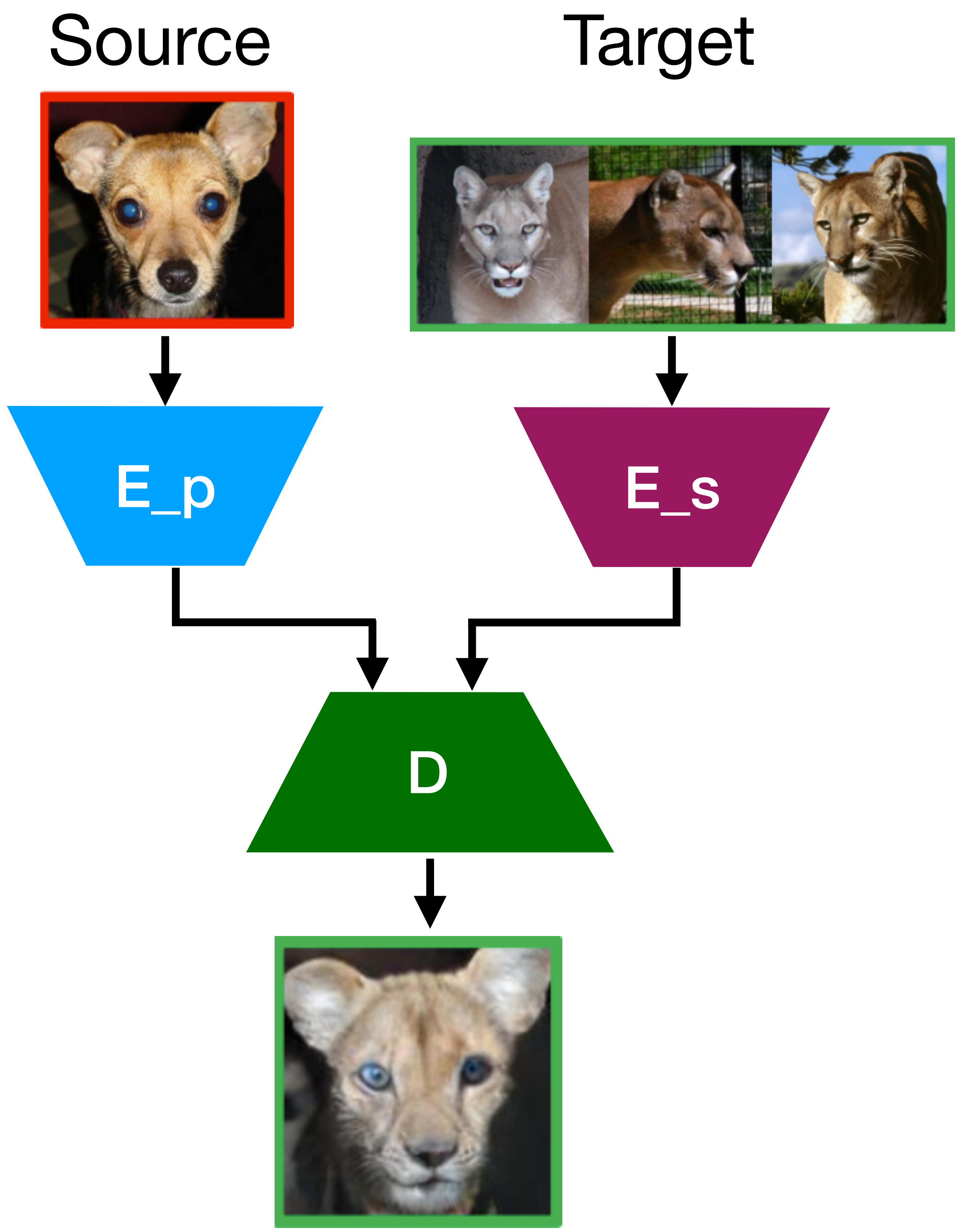
A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras, Samuli Laine, Timo Aila

Few-Shot Unsupervised Image-to-Image Translation

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, Jan Kautz




Source
Target
E_p
E_s
D

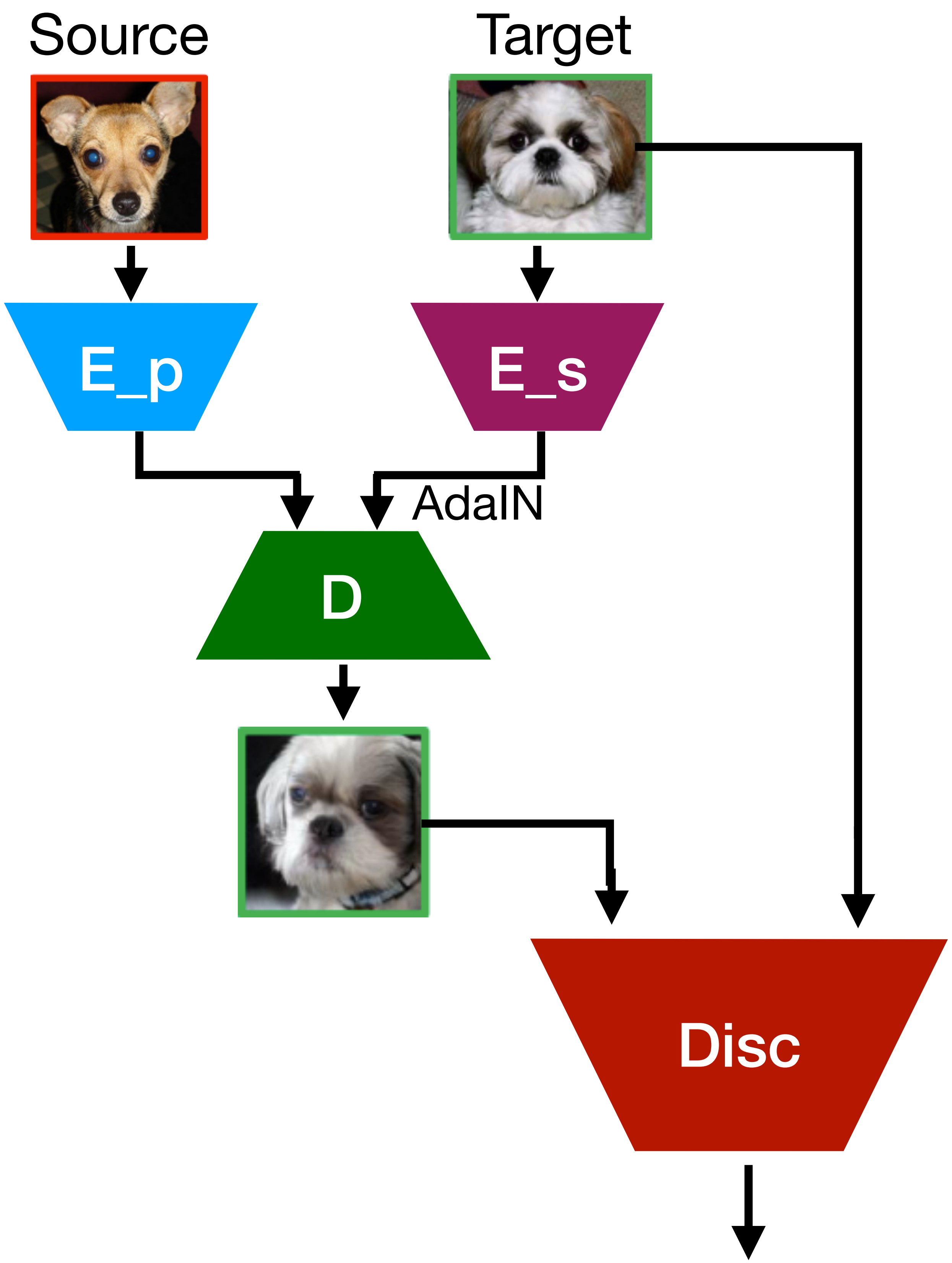
Style Encoder

Map each image to latent space then computes the mean.

Decoder

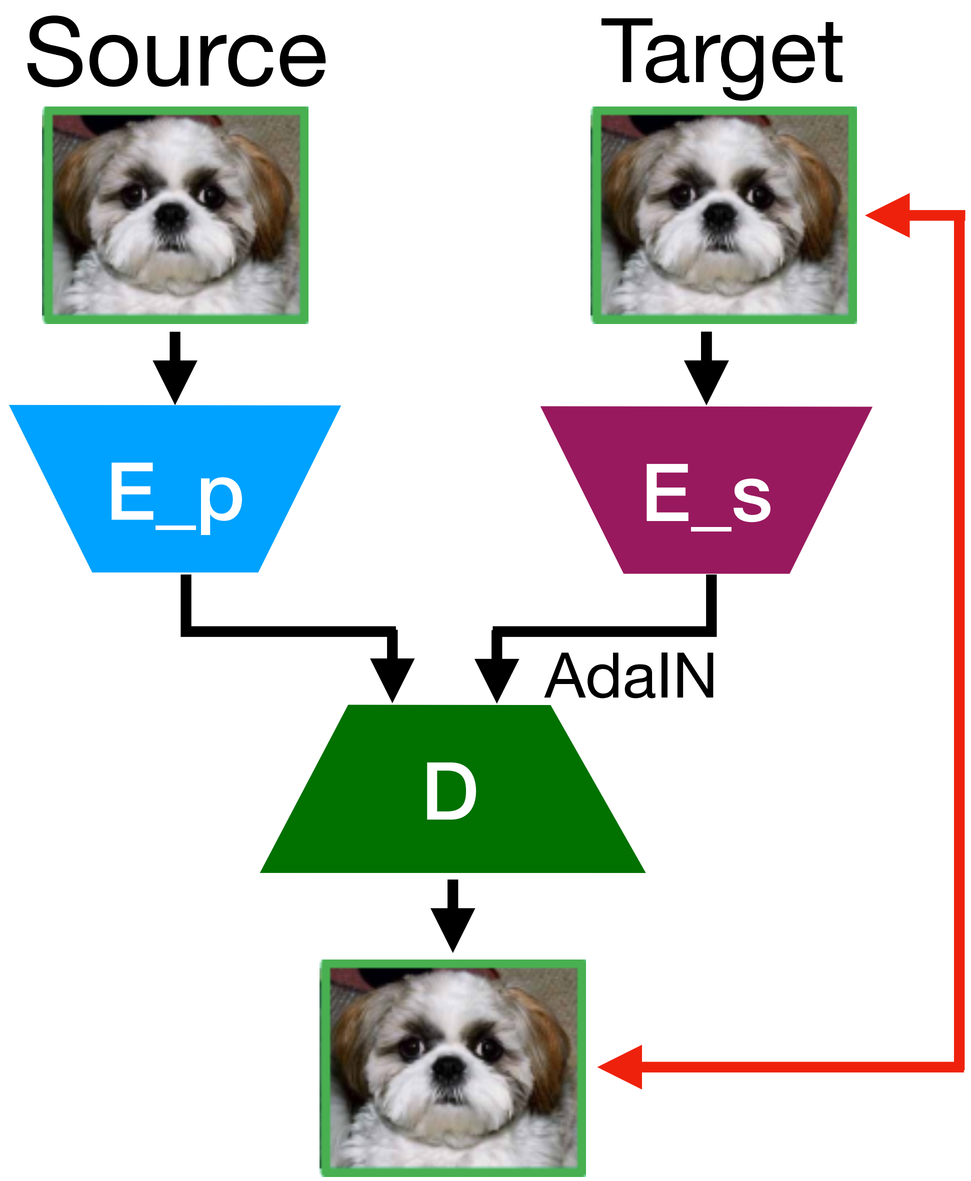
Since target domain used only for style, we feed the target encoding (style encoding) to the decoder via the AdaIN layers. We let the target images control the global look (e.g., object appearance), while the content image determines the local structure.

GAN Loss

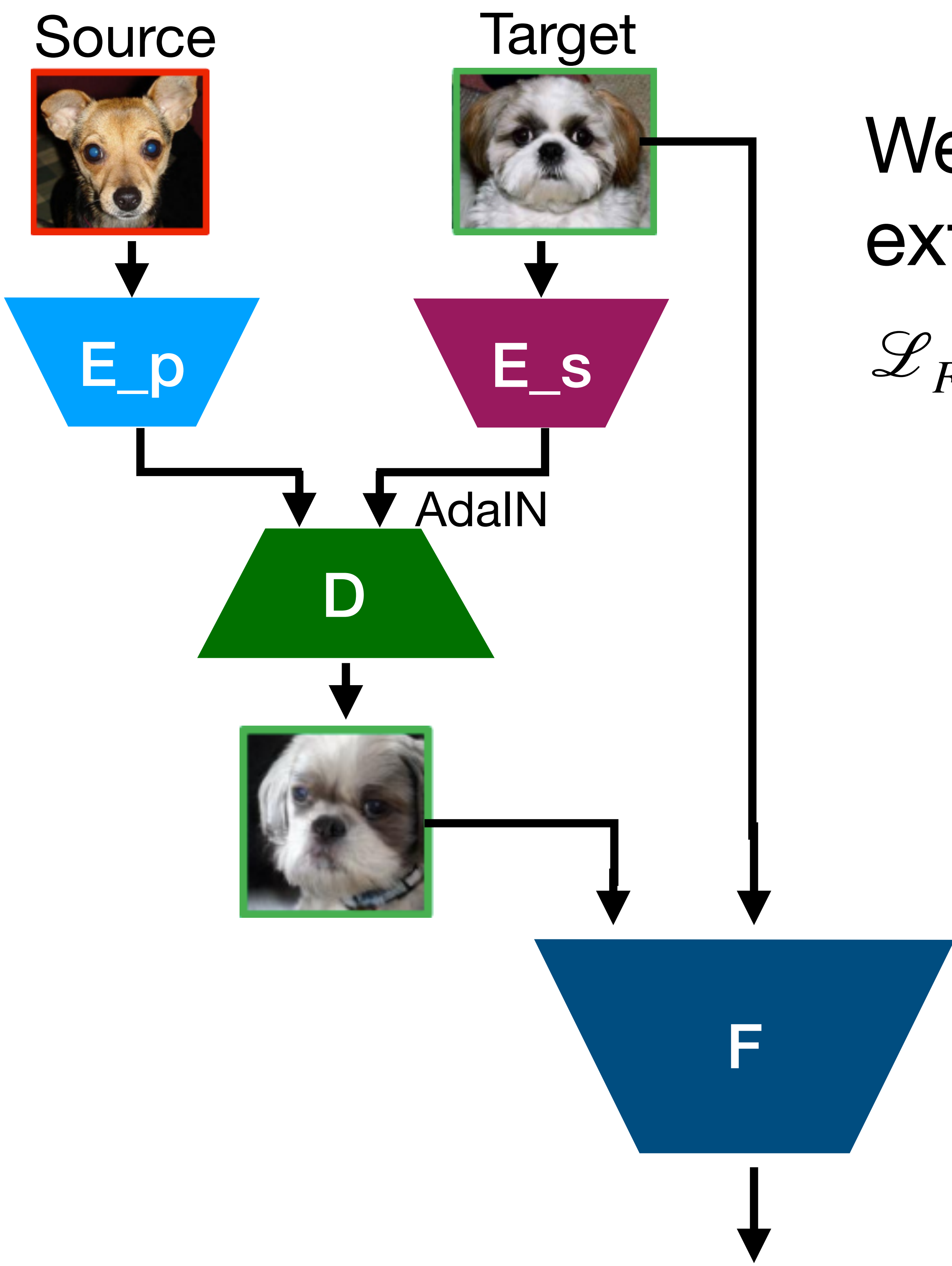


Discriminator output
Real / Fake per class.

Reconstruction



Feature Matching Loss



We use feature extractor F.

$$\mathscr{L}_F = \|F(D(E_c(x), E_s(y))) - F(y)\|$$

Disentanglement?

Network architecture. Style encoding can’t change the local structure.


$\mathbf{y}_1$
$\mathbf{y}_2$
$\mathbf{x}$
$\bar{\mathbf{x}}$

Conclusions

- Transfer some factors from image to image (content, style, rotation, etc.).
- Easier to work over latent space (e.g. adversarial loss).
- Enables to generate out-of-domain images.

Questions?