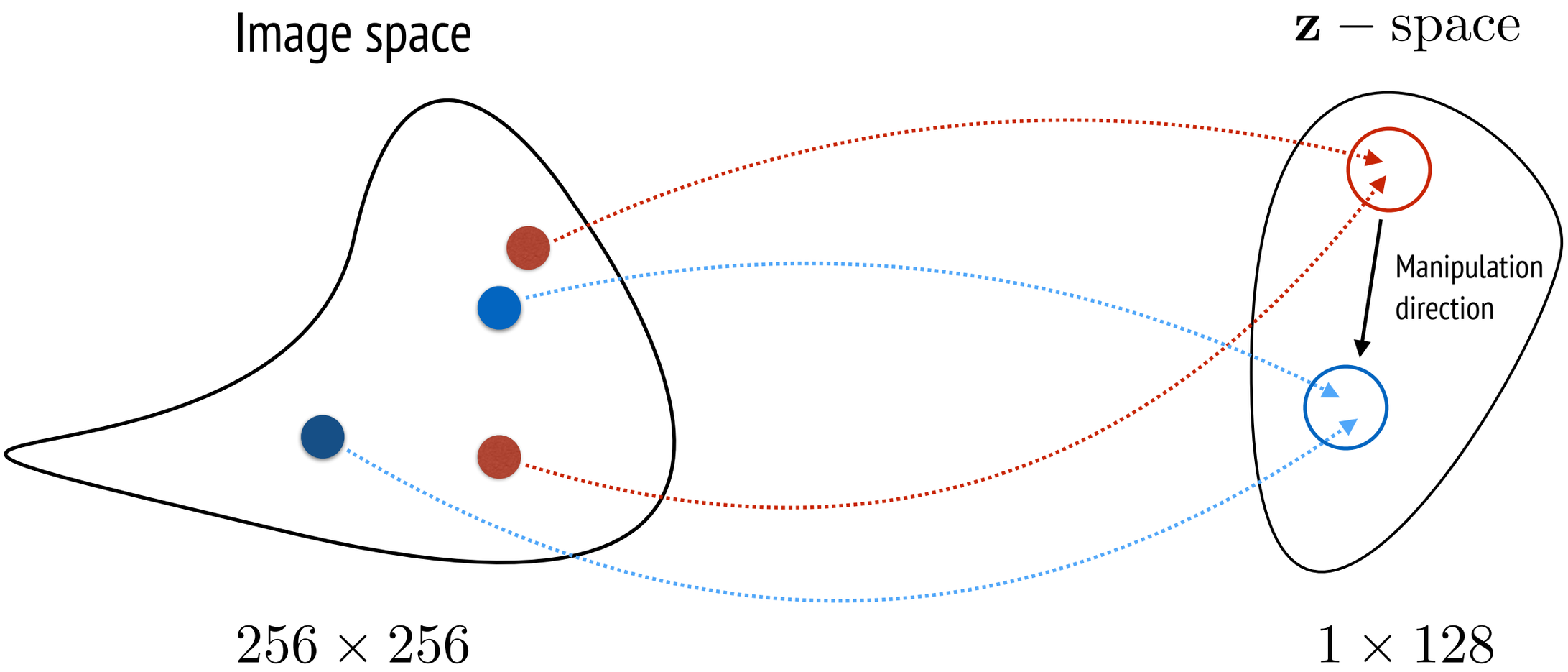
Image-to-image Translation

Sergey Tulyakov

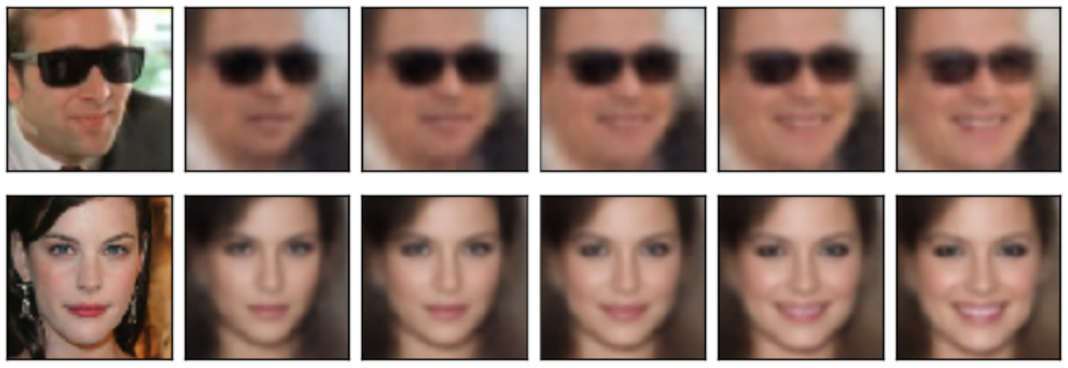
Today

- Image-to-image translation
  - Paired
  - Unpaired
  - Multimodal
- Stacked architectures
- Normalization layers
- Applications

VAE: Interpolation



Not smiling



Smiling

No beard

Has beard

CVAE-GAN: Interpolation

More advanced VAEs can be used but:

- How to pick the right direction?
- How to know where to stop?
- How to change only a single attribute?

Source

Target

Image-to-image Translation

Given two domain the goal is to translate image from one possible representation to another.

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{y})$$

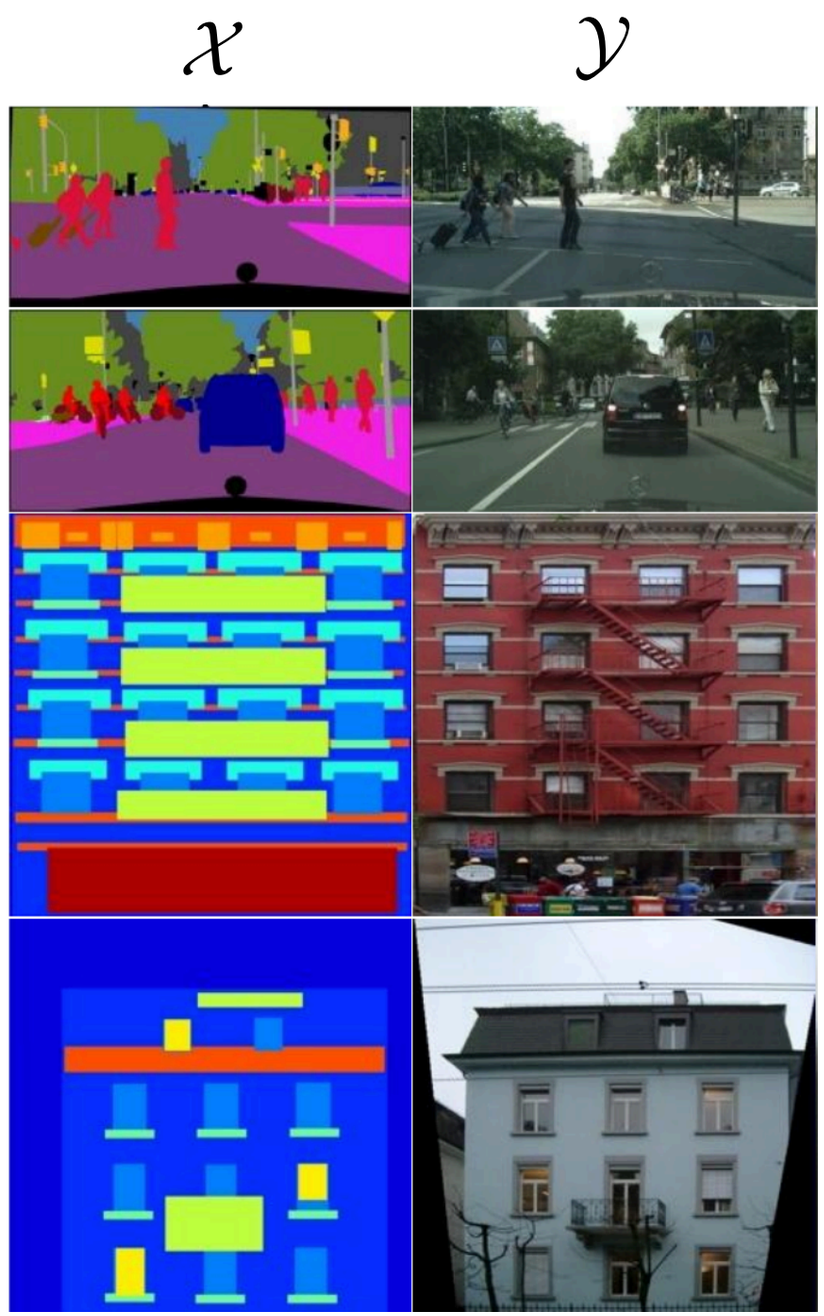
$$\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$$

Paired image-to-image translation

$$\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$$

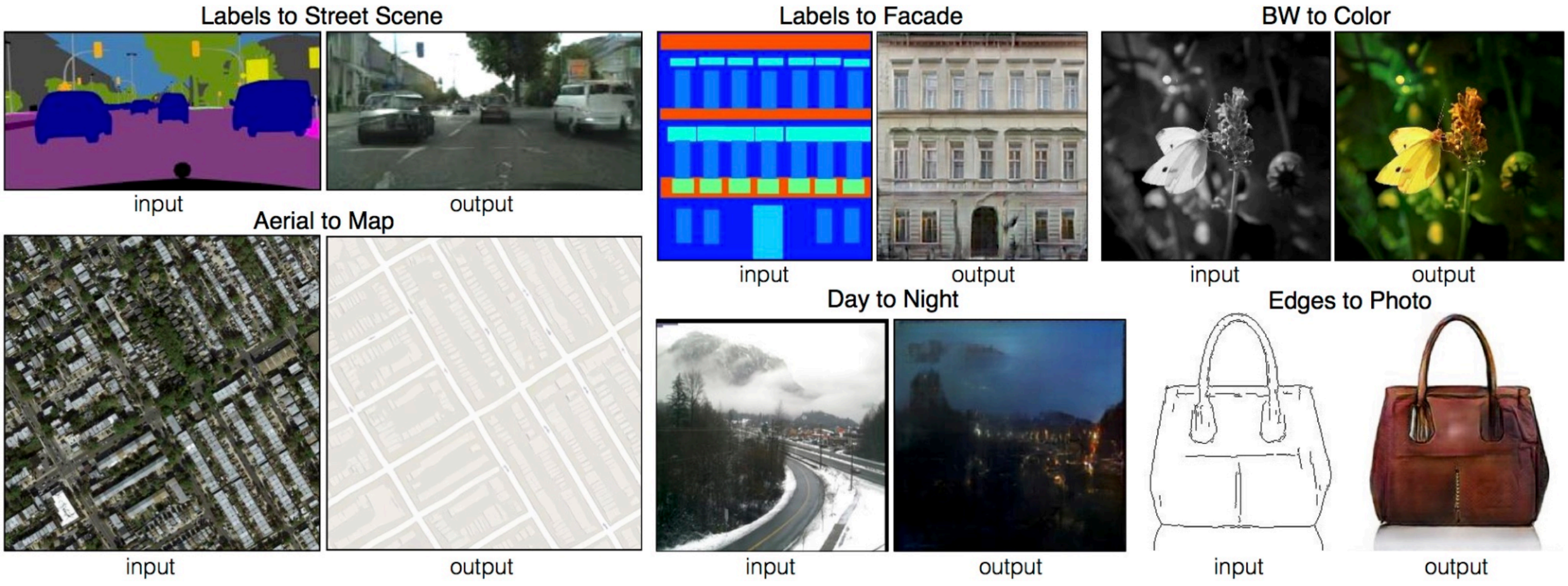
Unpaired

$$\mathbf{x} \sim p(\mathbf{x}), \ \mathbf{y} \sim p(\mathbf{y})$$



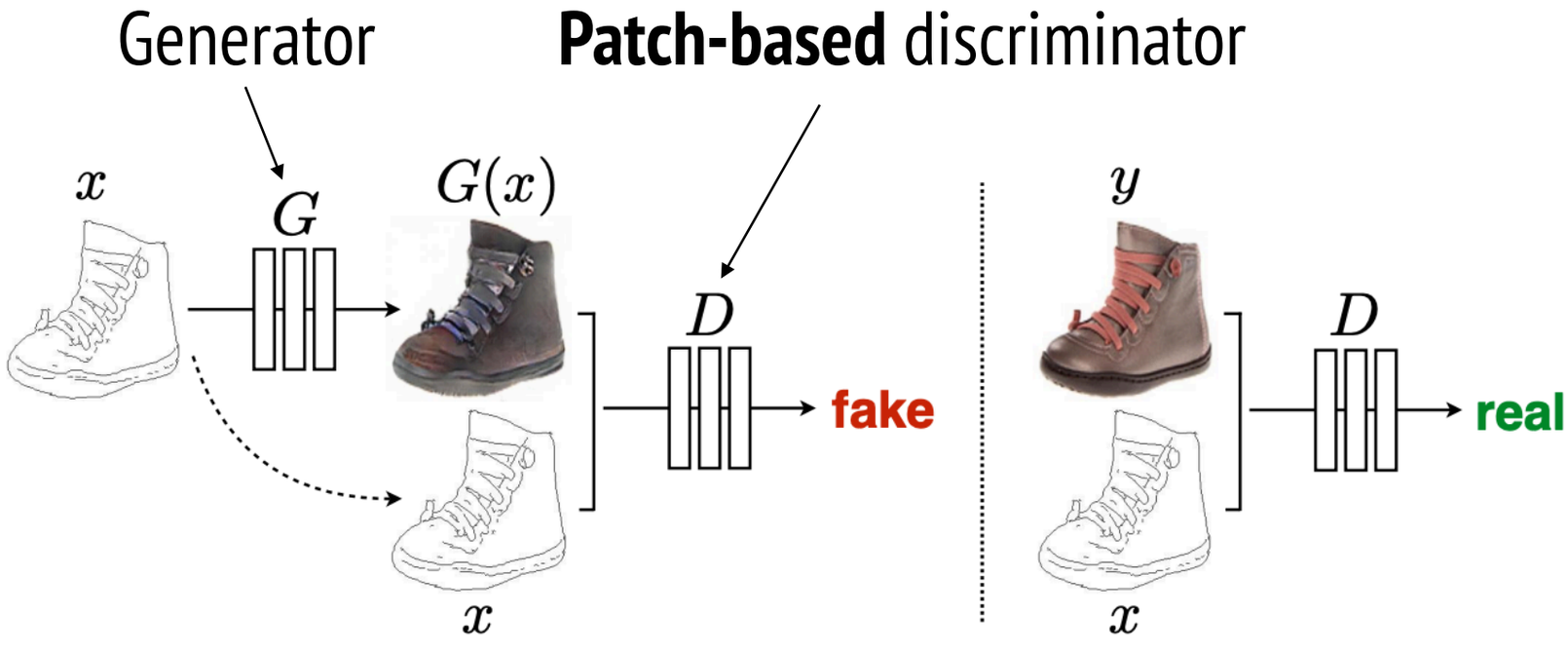
Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Motivation



Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Conditional Architecture



Combined GAN-loss and reconstruction loss:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))]$$

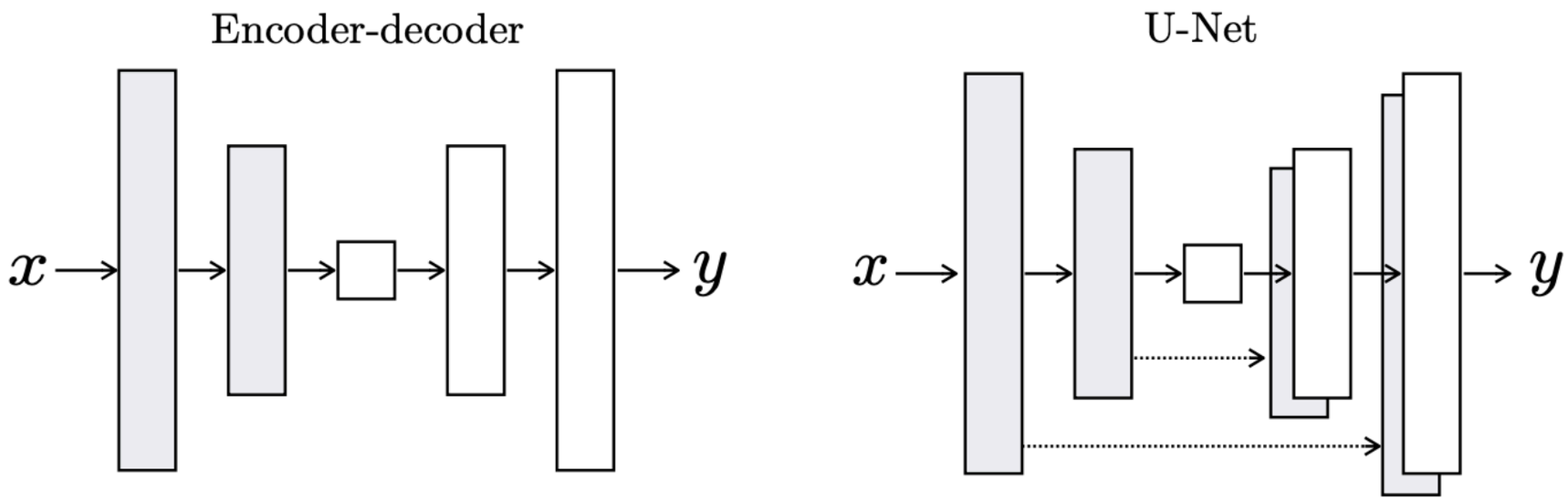
$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1]$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Generator

Skip connections in generator



Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Ablations

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.42	0.15	0.11
GAN	0.22	0.05	0.01
cGAN	0.57	0.22	0.16
L1+GAN	0.64	0.20	0.15
L1+cGAN	**0.66**	**0.23**	**0.17**
Ground truth	0.80	0.26	0.21

Table 1: FCN-scores for different losses, evaluated on Cityscapes labels↔photos.

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Ablations

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	**0.55**	**0.20**	**0.14**

Table 2: FCN-scores for different generator architectures (and objectives), evaluated on Cityscapes labels↔photos. (U-net (L1-cGAN) scores differ from those reported in other tables since batch size was 10 for this experiment and 1 for other tables, and random variation between training runs.)

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

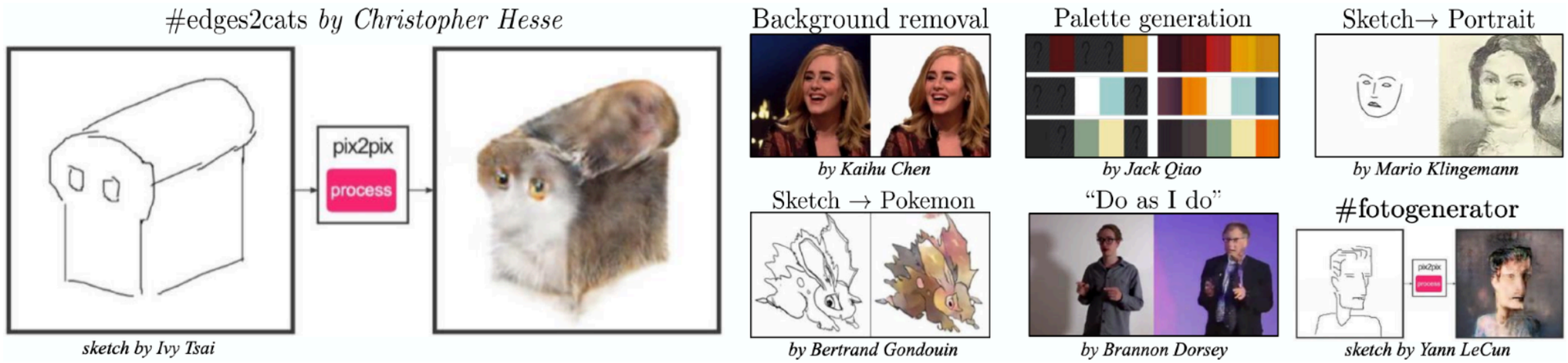
Pix2Pix: Ablations

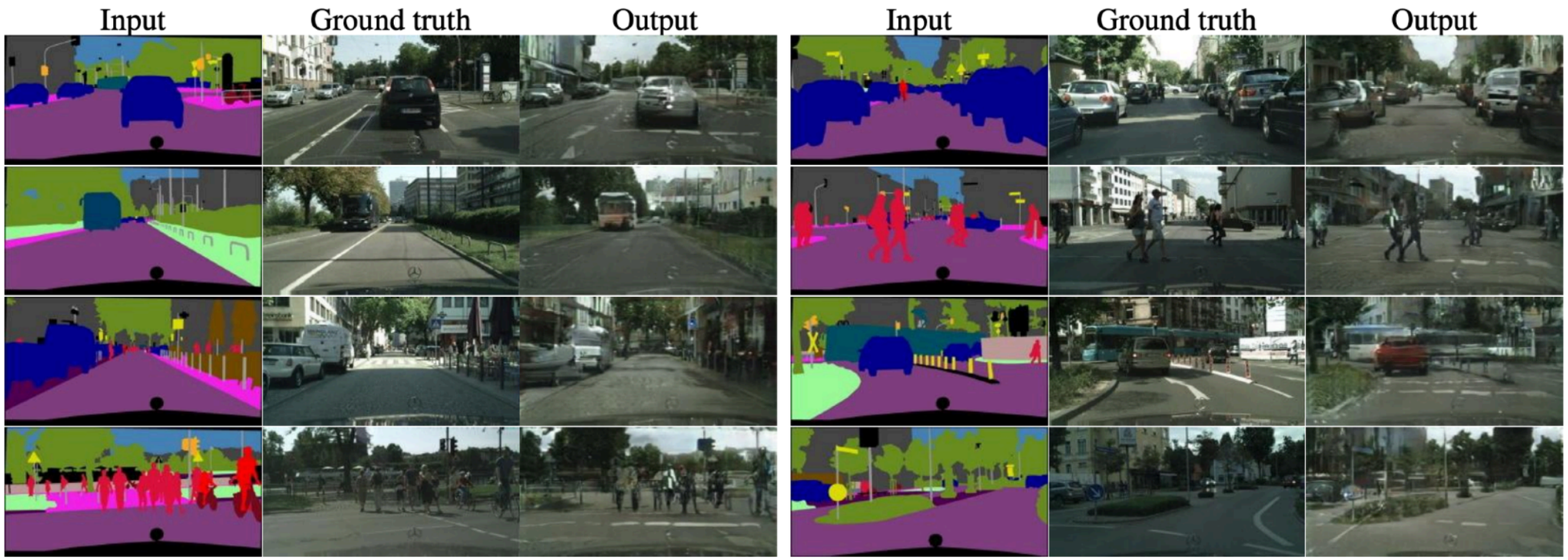
Discriminator receptive field	Per-pixel acc.	Per-class acc.	Class IOU
1×1	0.39	0.15	0.10
16×16	0.65	0.21	**0.17**
70×70	**0.66**	**0.23**	**0.17**
286×286	0.42	0.16	0.11

Table 3: FCN-scores for different receptive field sizes of the discriminator, evaluated on Cityscapes labels→photos. Note that input images are 256 × 256 pixels and larger receptive fields are padded with zeros.

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Results and Applications





Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Results and Applications

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Paired Image-to-image Translation

Given two domains the goal is to translate image from one possible representation to another.

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{y})$$

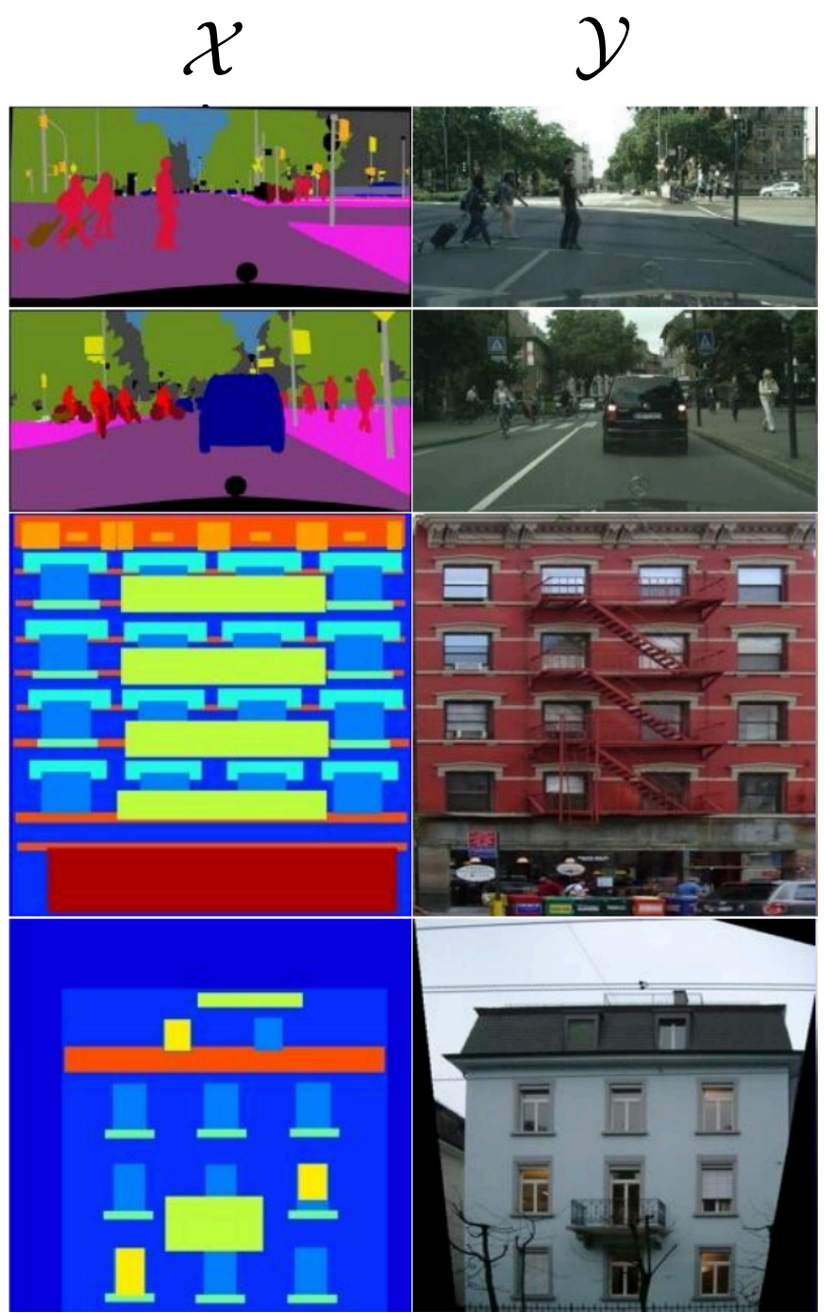
$$\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$$

Paired image-to-image translation

$$\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$$

Unpaired

$$\mathbf{x} \sim p(\mathbf{x}), \ \mathbf{y} \sim p(\mathbf{y})$$



Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Unpaired Image-to-image Translation

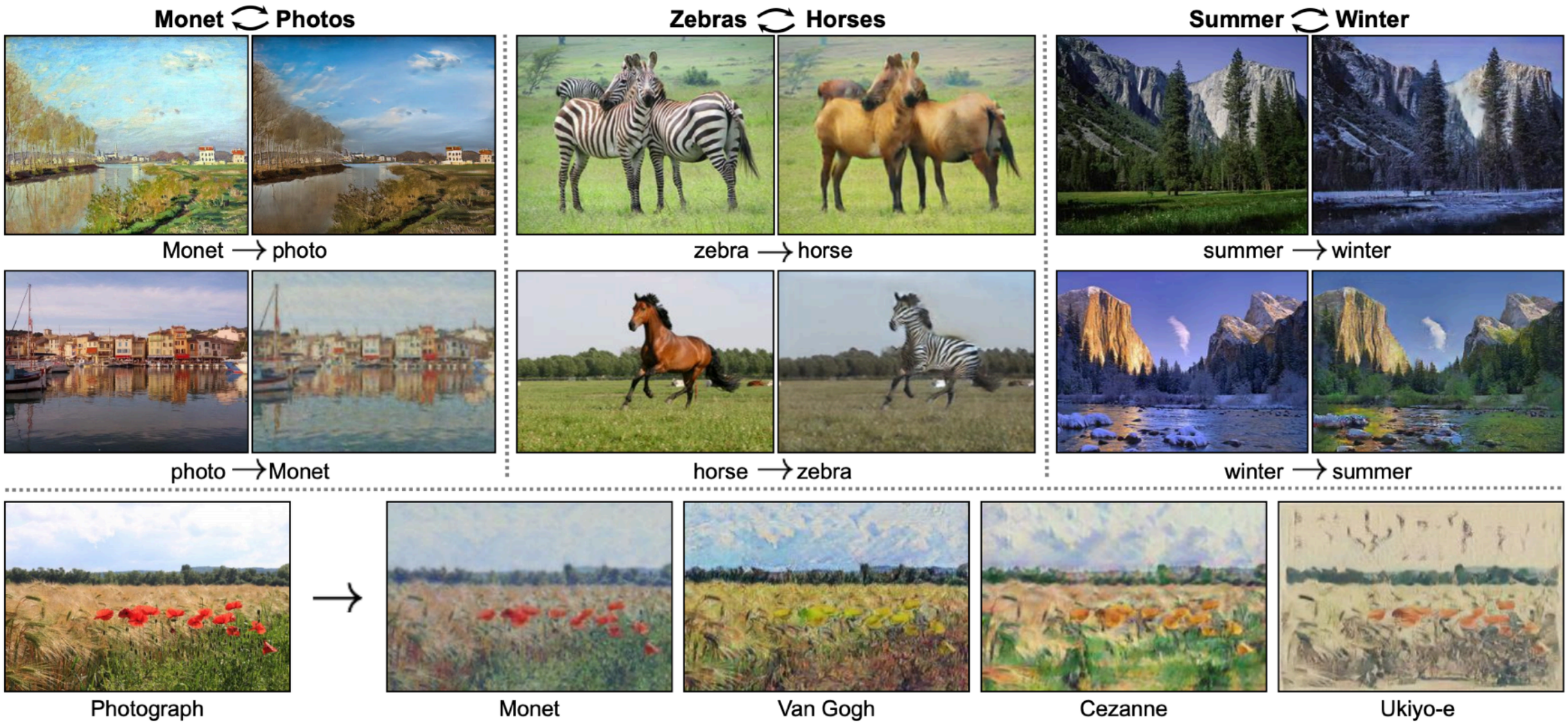
Given inability to sample from joint distribution (i.e. observe paired data), learning conditional distribution (i.e. translation) is an ill-posed problem.

To solve it, constraints are necessary:

- Cycle-consistency constraint
- Weight-sharing constraint
- Equivariance constraint

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV'2017.

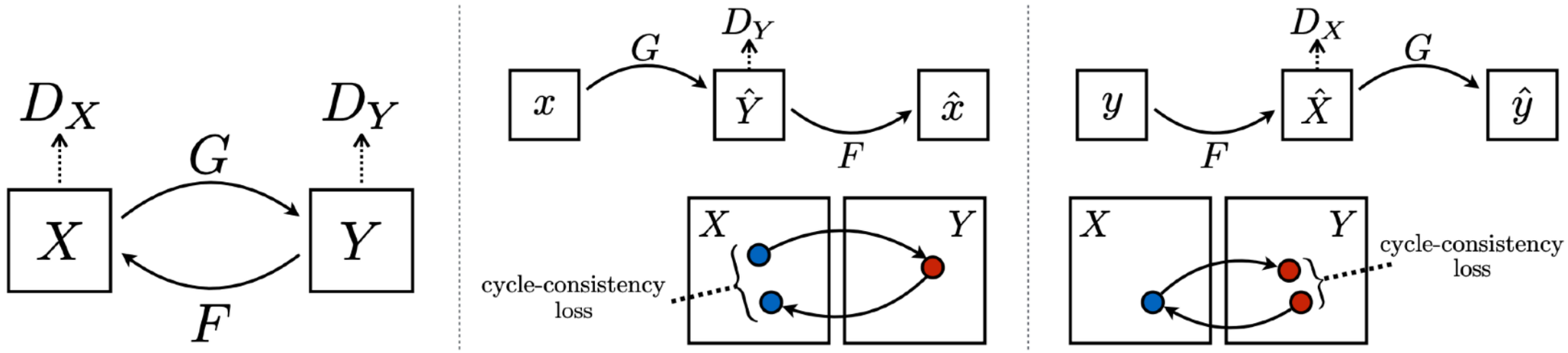
CycleGAN



It is sometimes impossible to get the same image in a different representation

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV'2017.

CycleGAN: Overview



Adversarial loss:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]$$

Cycle-consistency loss:

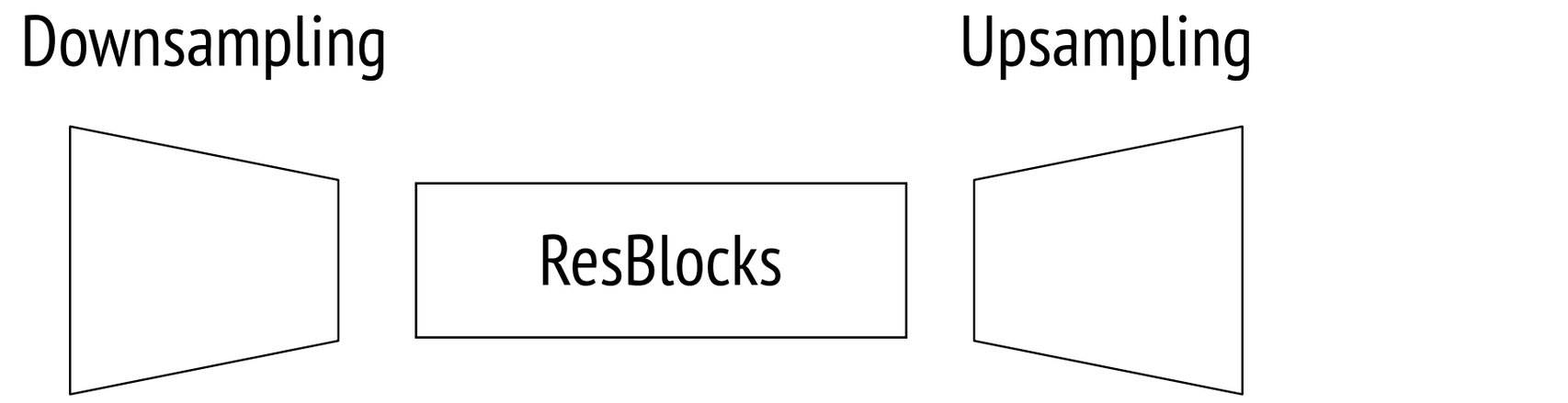
$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$$

Full objective:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F)$$

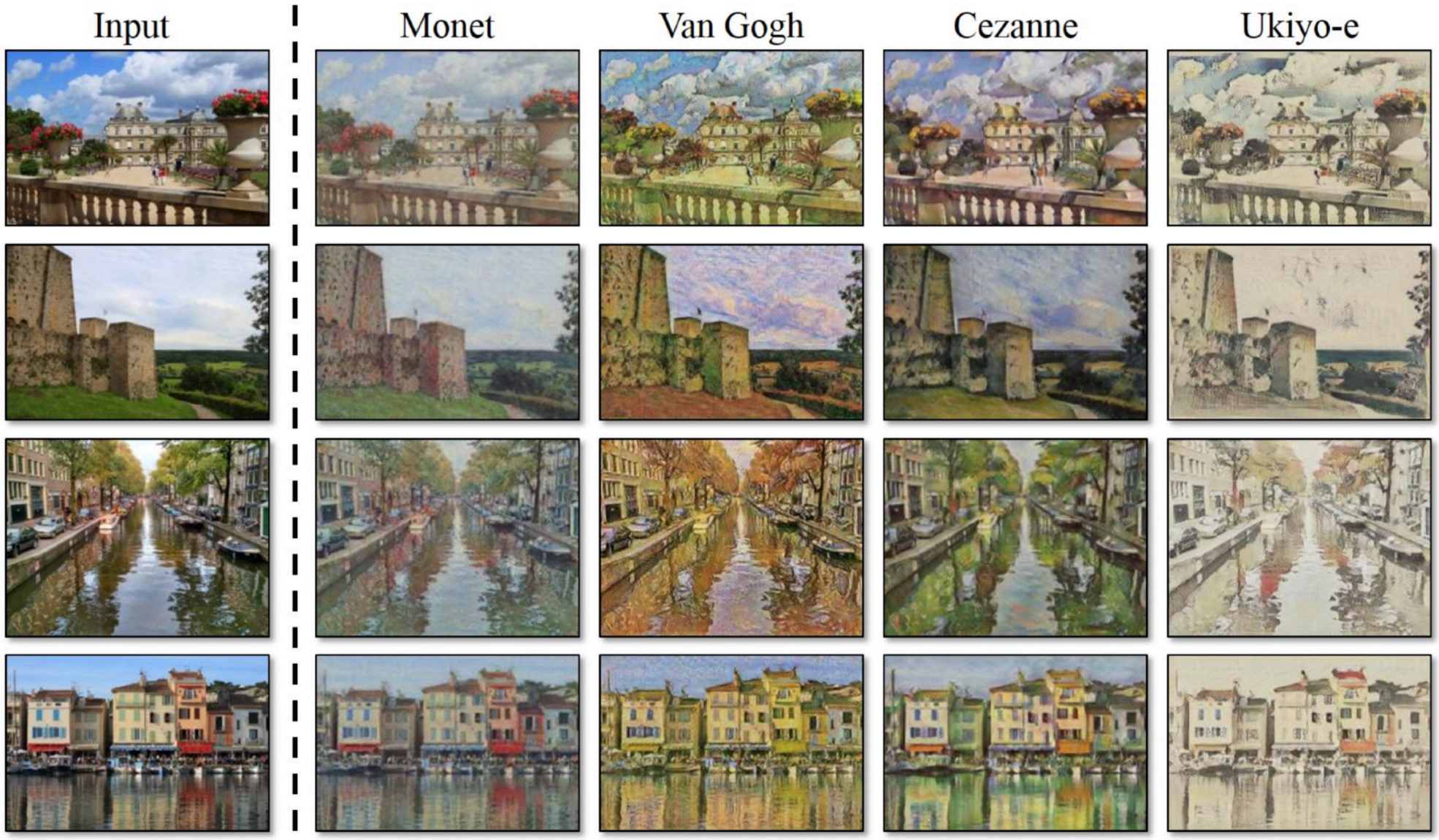
Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV'2017.

CycleGAN: Architecture

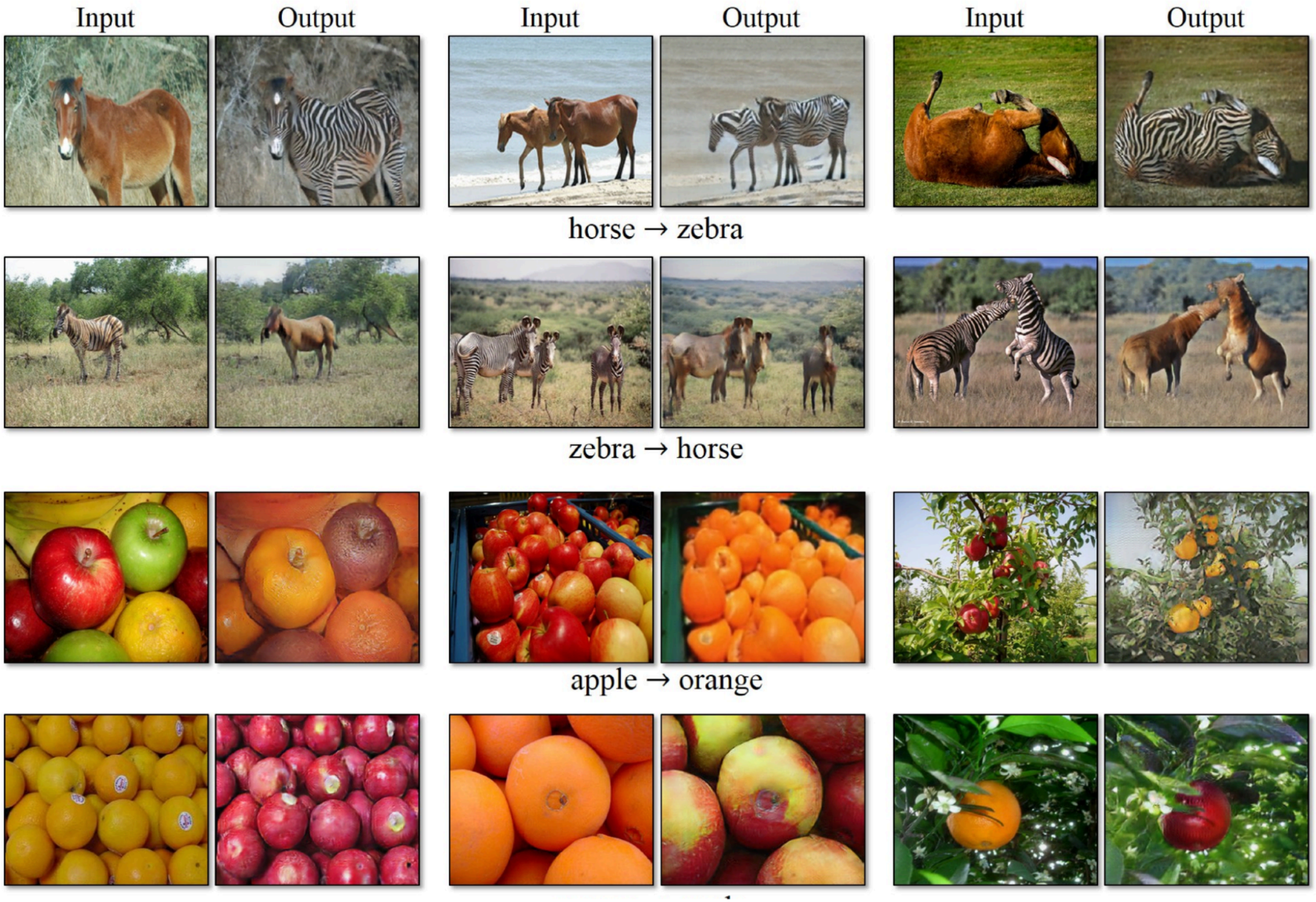


Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." ECCV'2016.

CycleGAN: Results

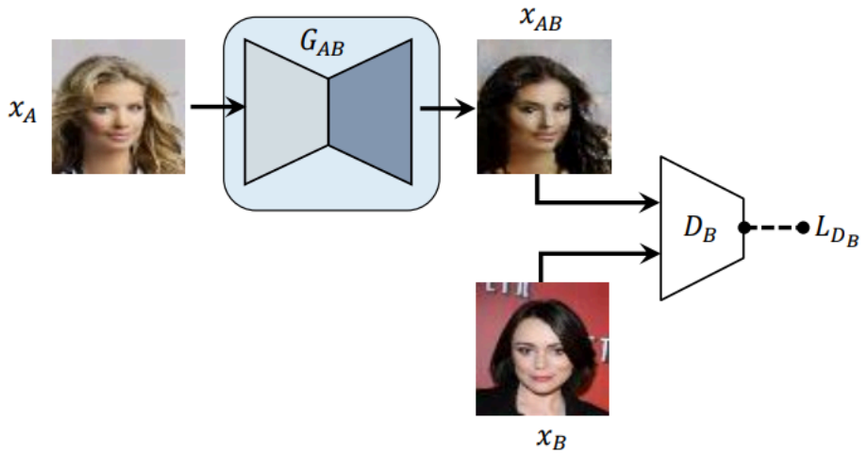


CycleGAN: Results

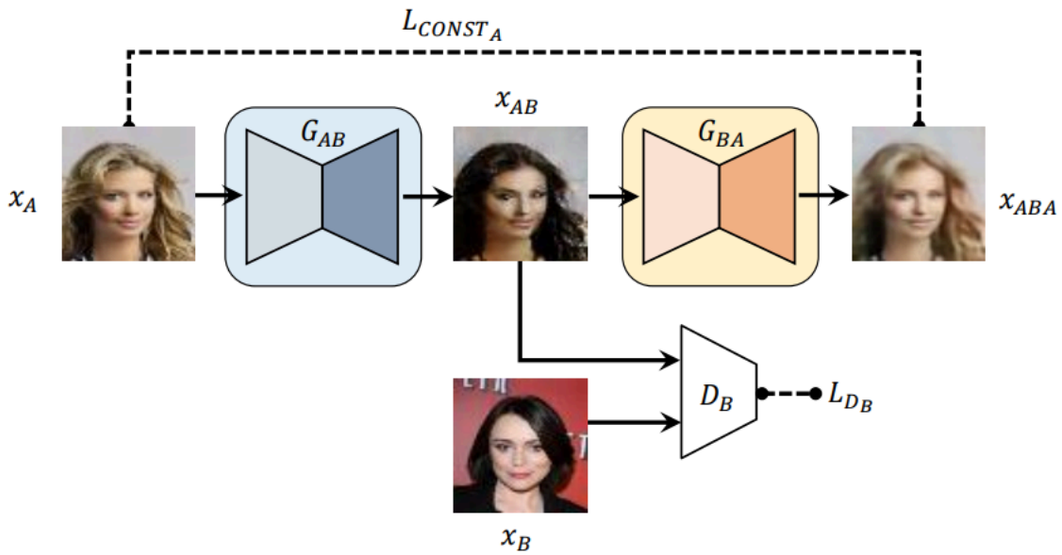


DiscoGAN

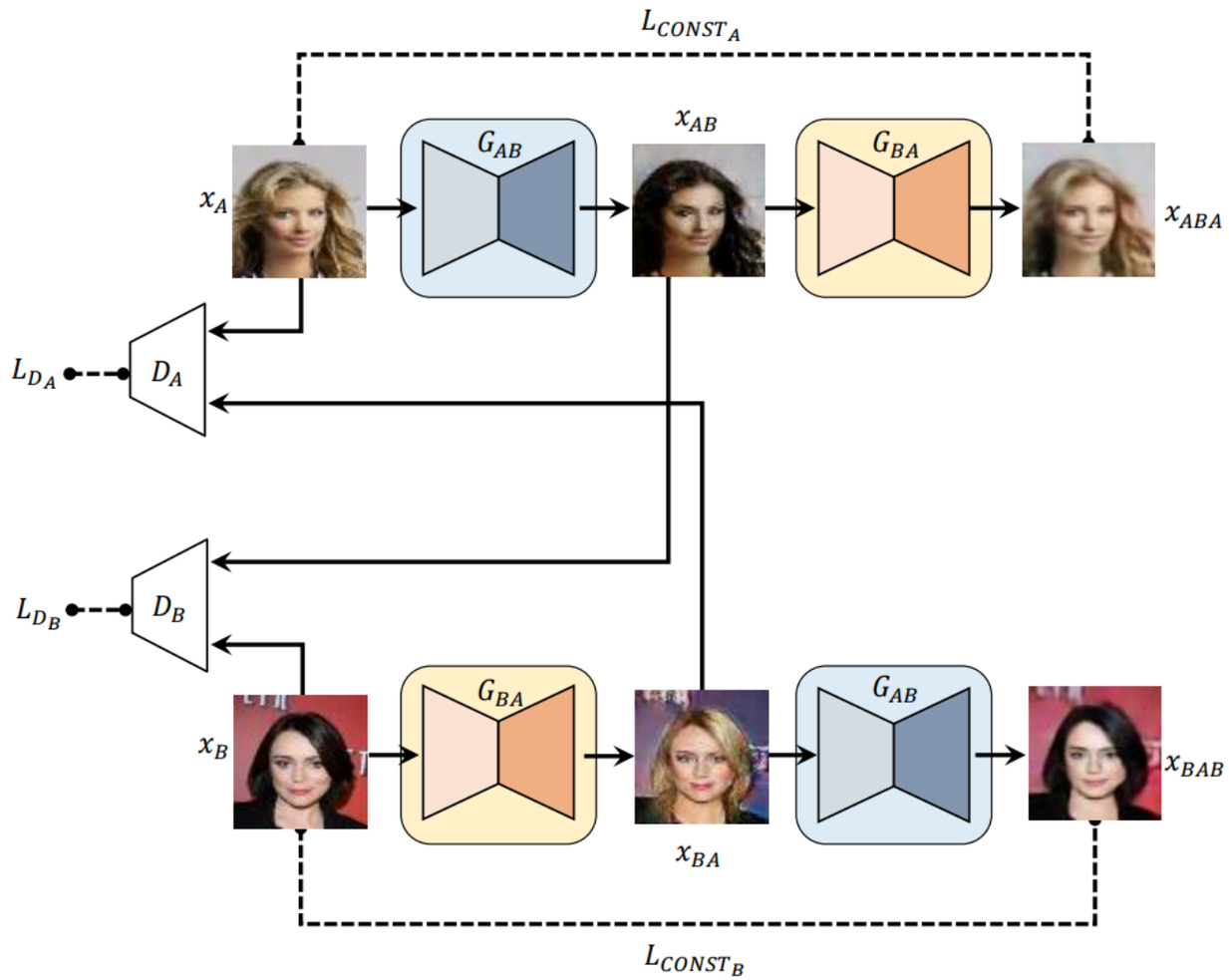
Standard GAN



Gan w/ reconstruction loss

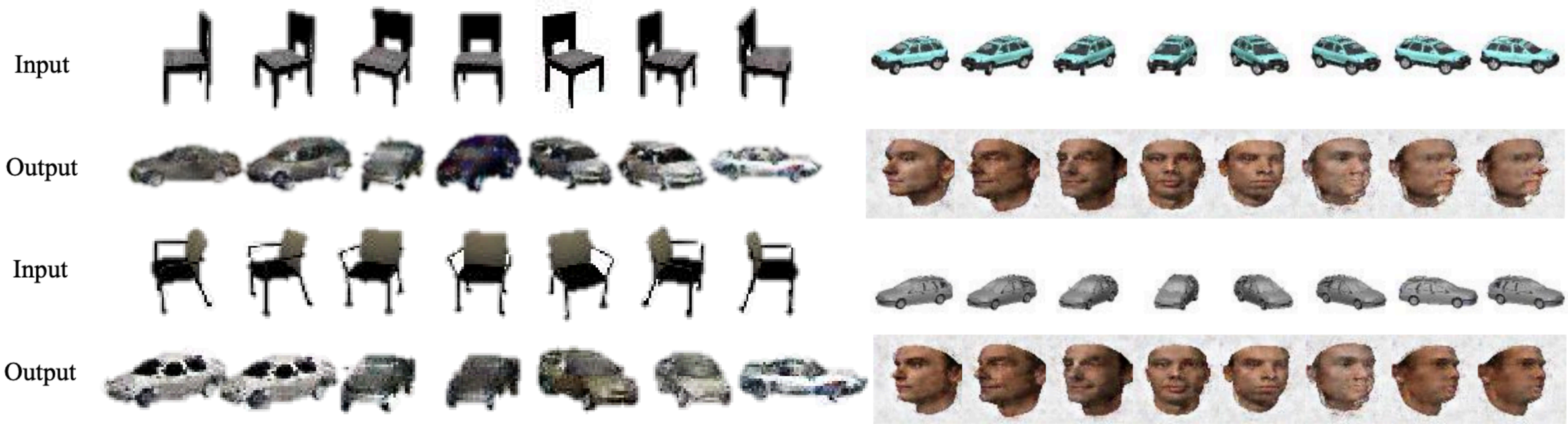


DiscoGAN



Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." ICML'2017.

DiscoGAN



Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." ICML'2017.

Image-to-image Translation

Given inability to sample from joint distribution (i.e. observe paired data), learning conditional distribution (i.e. translation) is an ill-posed problem.

To solve it, constraints are necessary:

- Cycle-consistency constraint ←
- Weight-sharing constraint
- Geometry-consistency constraint

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV'2017.

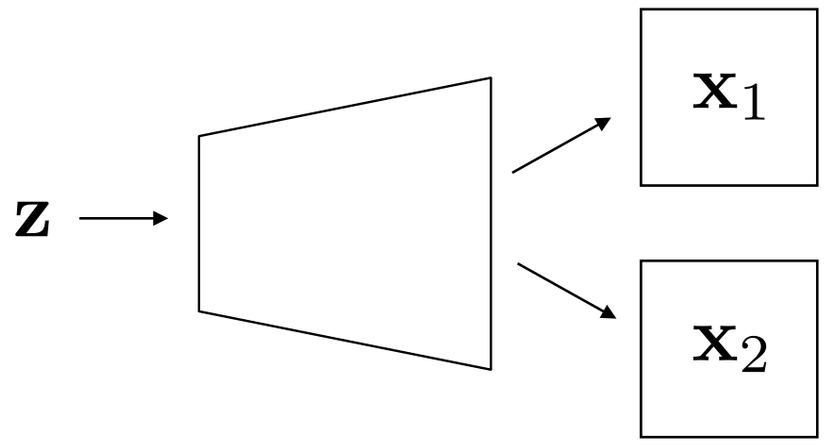
CoGAN: Coupled Generative Adversarial Networks

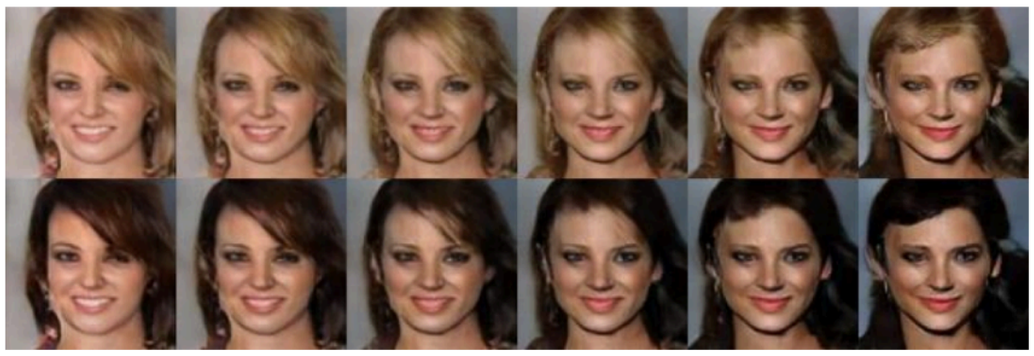
Problem: generate unconditional samples from a joint distribution:

$$\mathbf{x}_1, \mathbf{x}_2 \sim p(\mathbf{x}_1, \mathbf{x}_2)$$

having access to marginal distributions:

$$\mathbf{x}_1 \sim p(\mathbf{x}_1), \;\; \mathbf{x}_2 \sim p(\mathbf{x}_2)$$



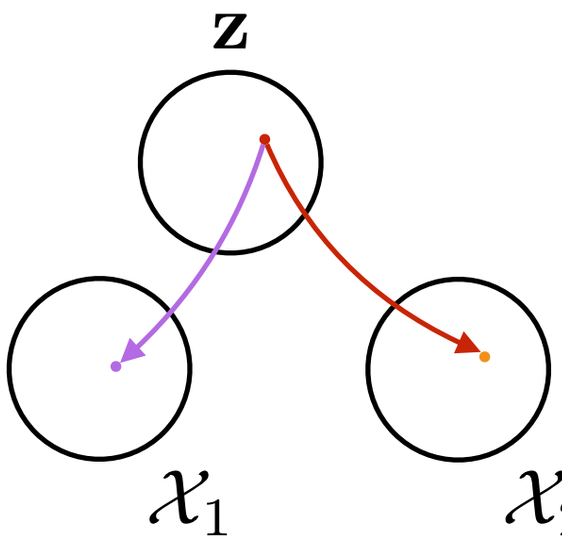


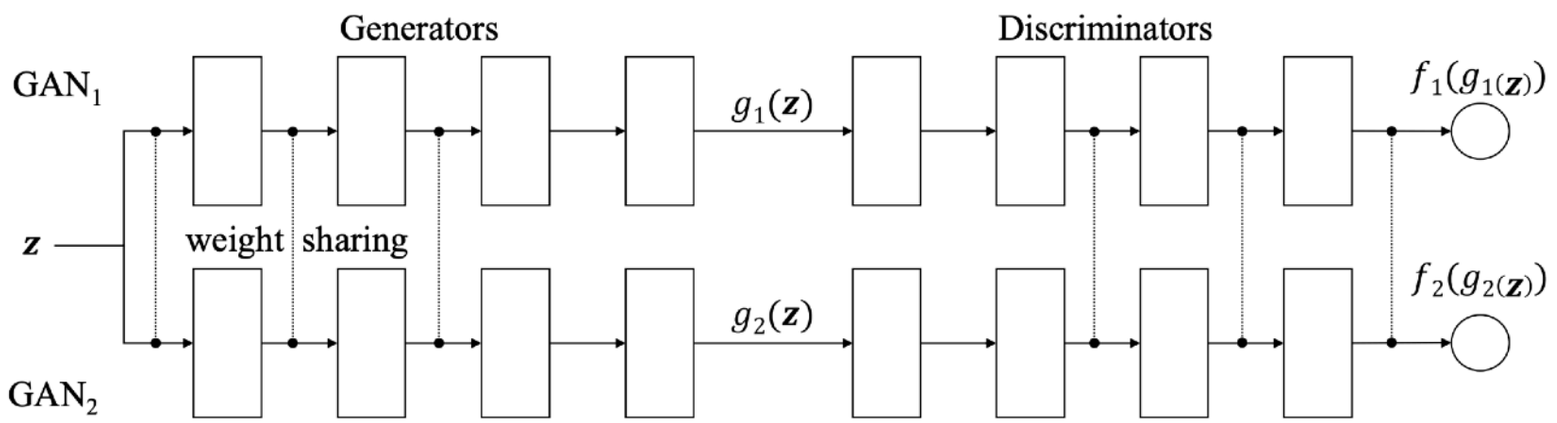
Observation: images in both domain share structure but not the style

Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." NIPS'2016

Constraint: Shared Latent Space

Observation: images in both domain share structure but not the style

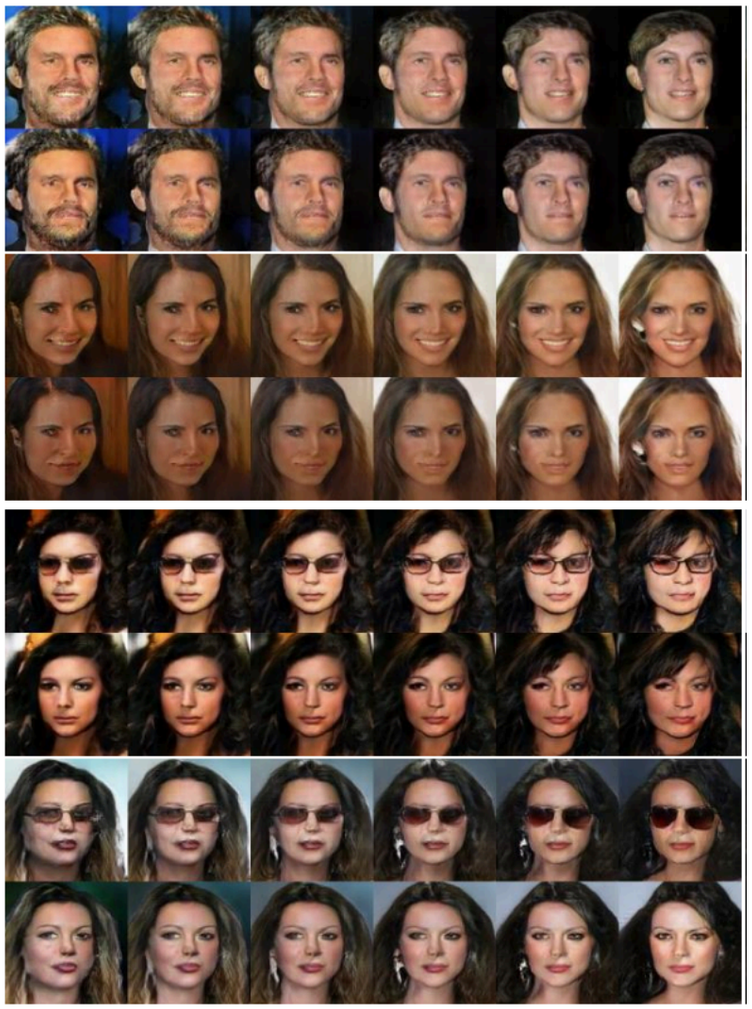


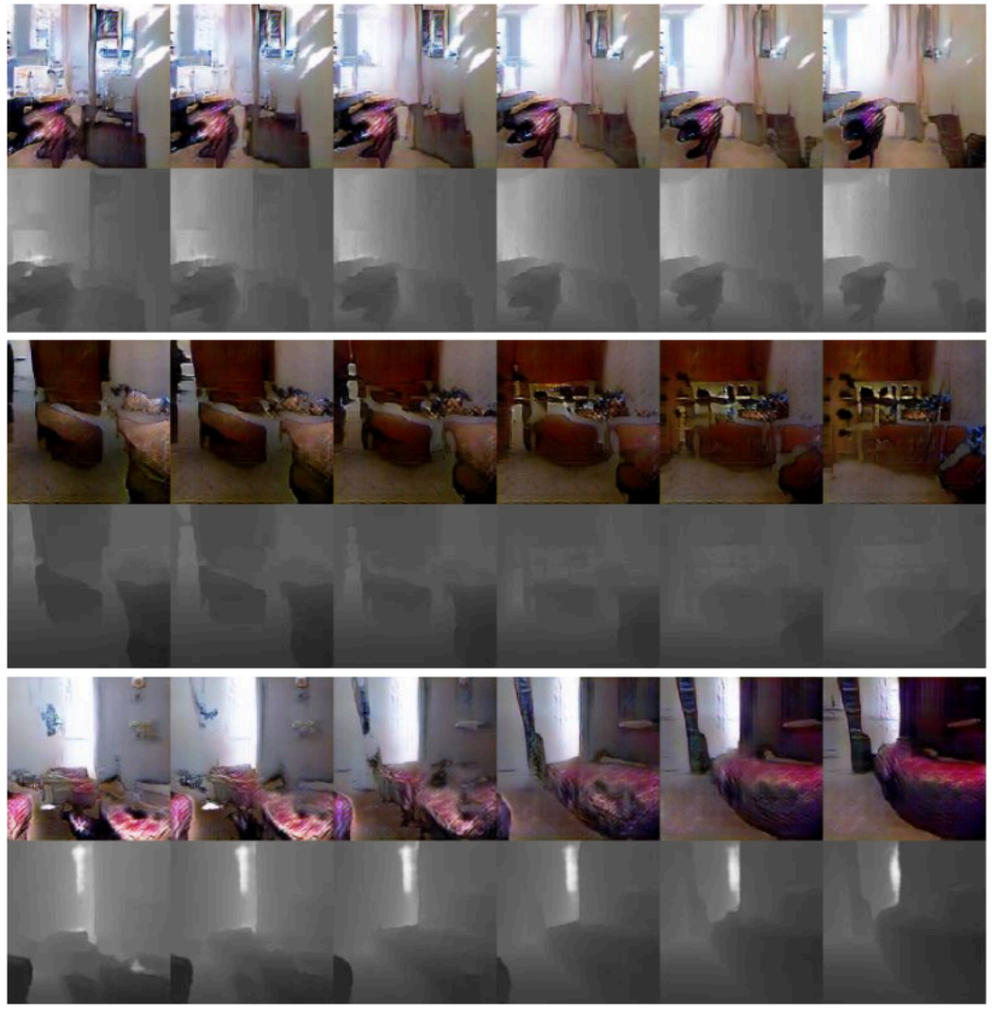


Implementation: shared latent space via weights sharing. Initial layers render low-frequency structure, while last layers encode the style

Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." NIPS'2016

CoGAN: Results

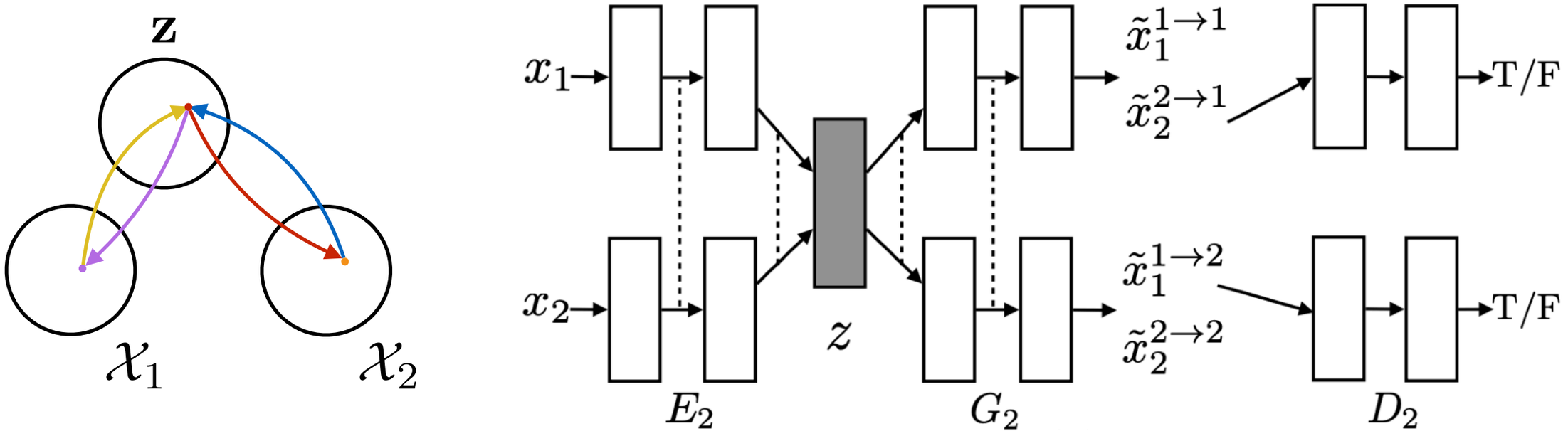




Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." NIPS'2016

Extending CoGAN: MUNIT

Observation: images in both domain share structure but not the style

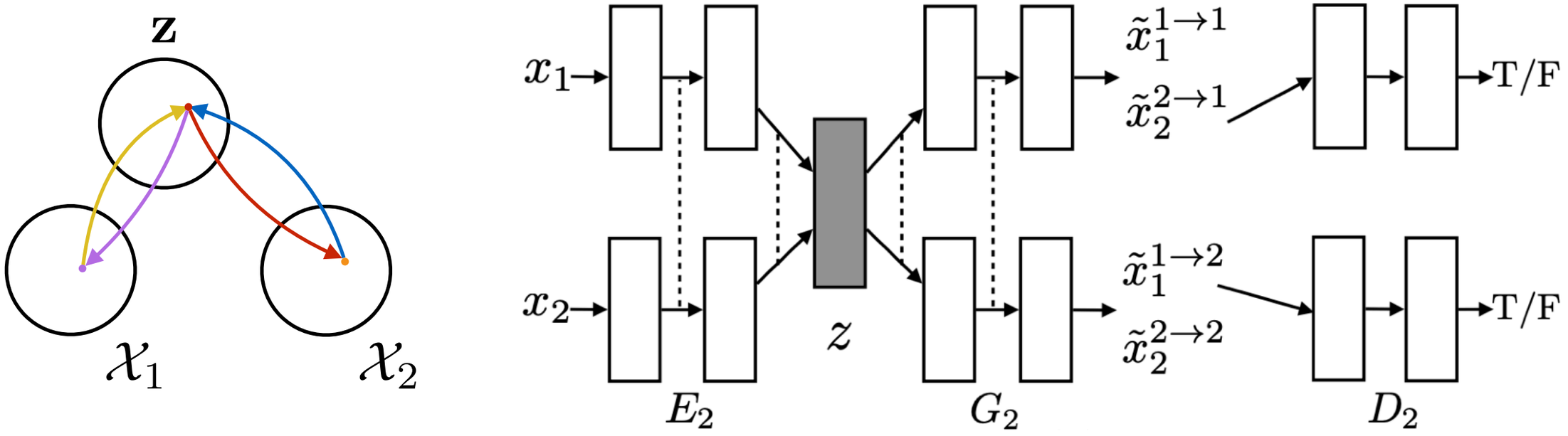


Implementation: train encoders with shared weights to go from image space to shared latent space

Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." NIPS'2017

Extending CoGAN: MUNIT

Observation: images in both domain share structure but not the style



VAE

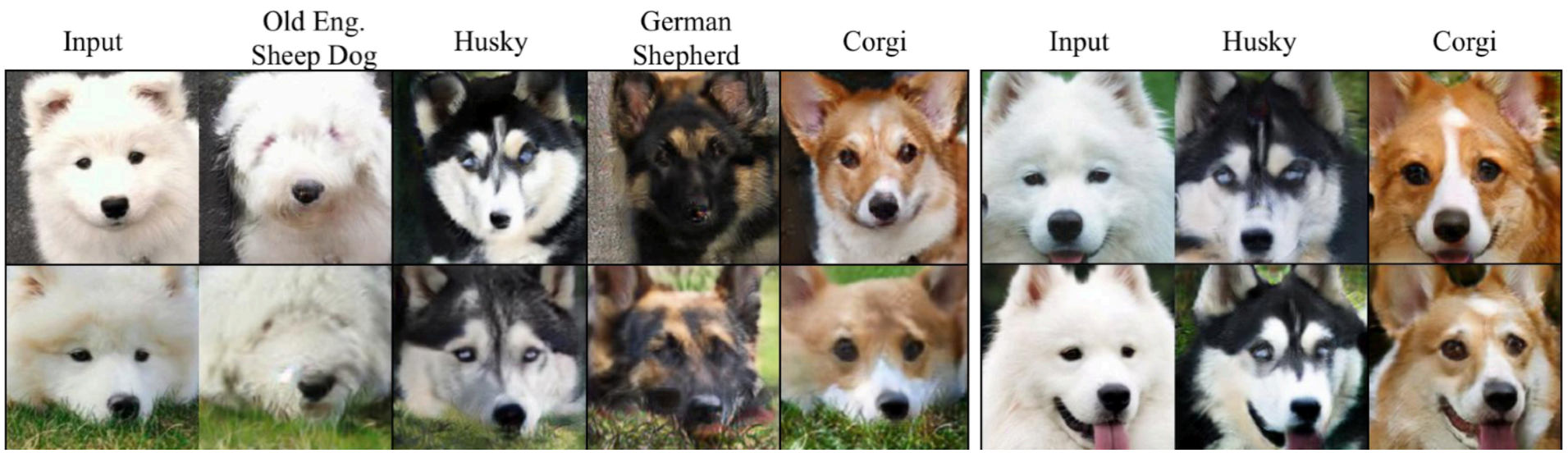
Issue: it is not a VAE, since it doesn't support sampling and cannot compute likelihood. Latent representation is a tensor, not a vector

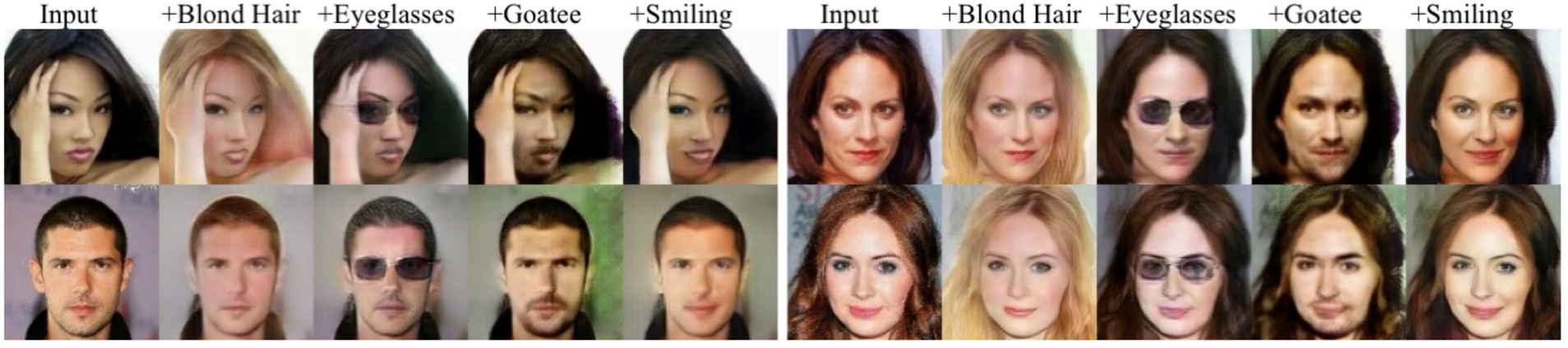
Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." NIPS'2017

MUNIT: Results

Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." NIPS'2017

MUNIT: Results





Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." NIPS'2017

Image-to-image Translation

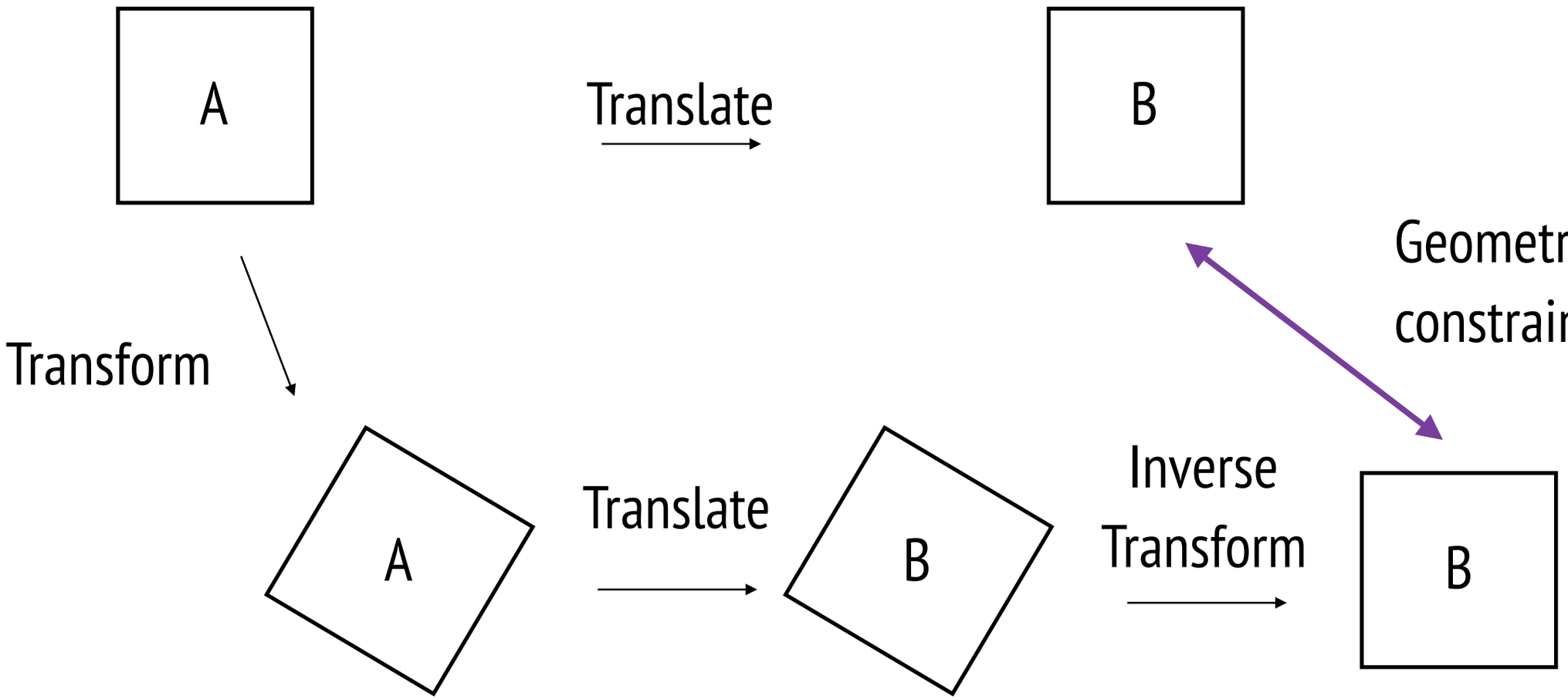
Given inability to sample from joint distribution (i.e. observe paired data), learning conditional distribution (i.e. translation) is an ill-posed problem.

To solve it, constraints are necessary:

- Cycle-consistency constraint
- Weight-sharing constraint
- Geometry-consistency constraint

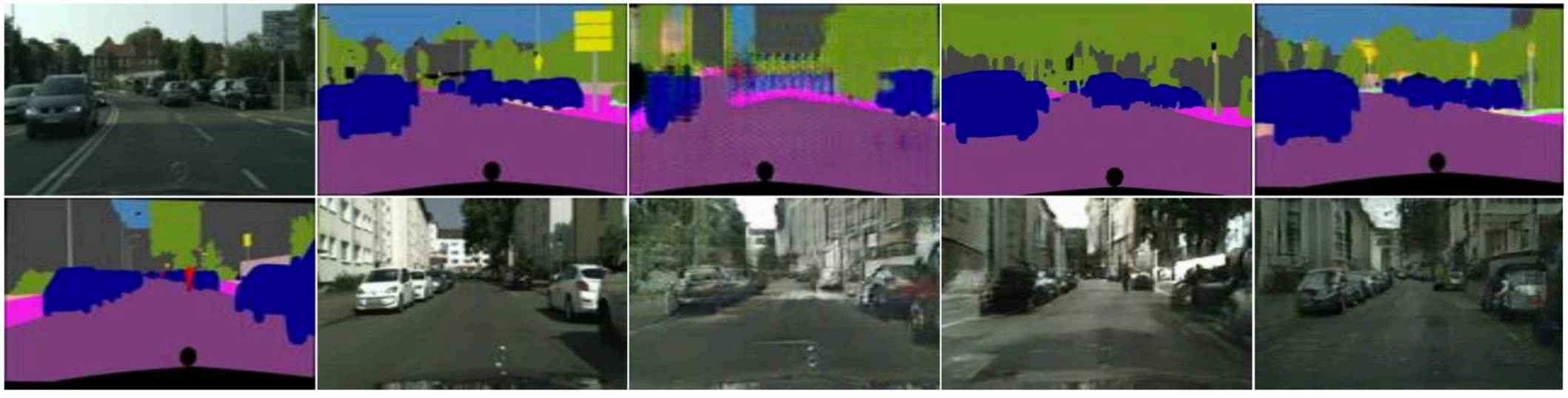
Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV'2017.

Geometry-consistency Constraint



Fu, Huan, et al. "Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping." CVPR'2019

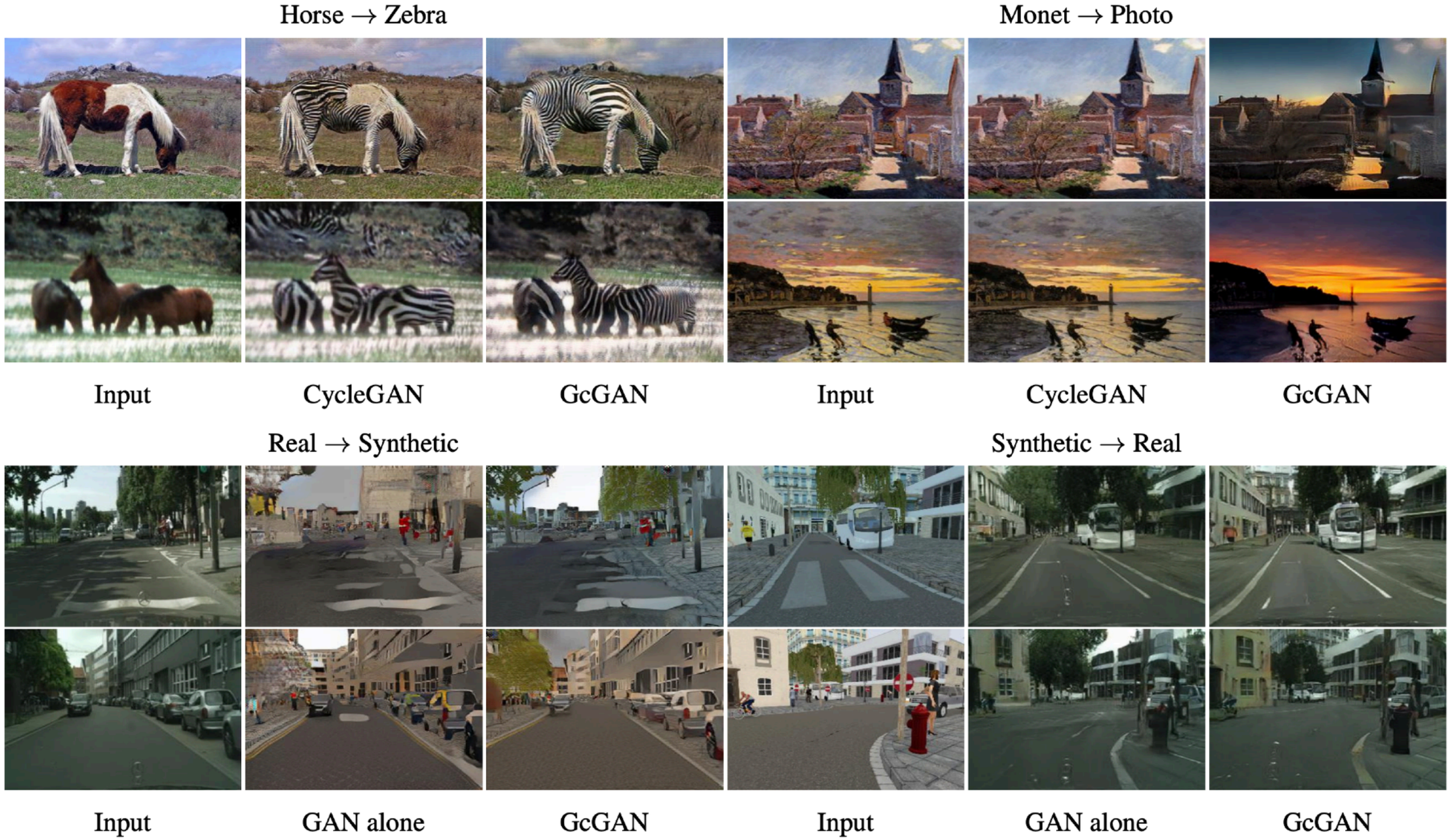
Geometry-consistency Constraint



Input | Ground Truth | GAN alone | CycleGAN | GcGAN

Fu, Huan, et al. "Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping." CVPR'2019

Geometry-consistency Constraint: Results



Fu, Huan, et al. "Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping." CVPR'2019

Paired Image-to-image Translation

Given two domains the goal is to translate image from one possible representation to another.

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{y})$$

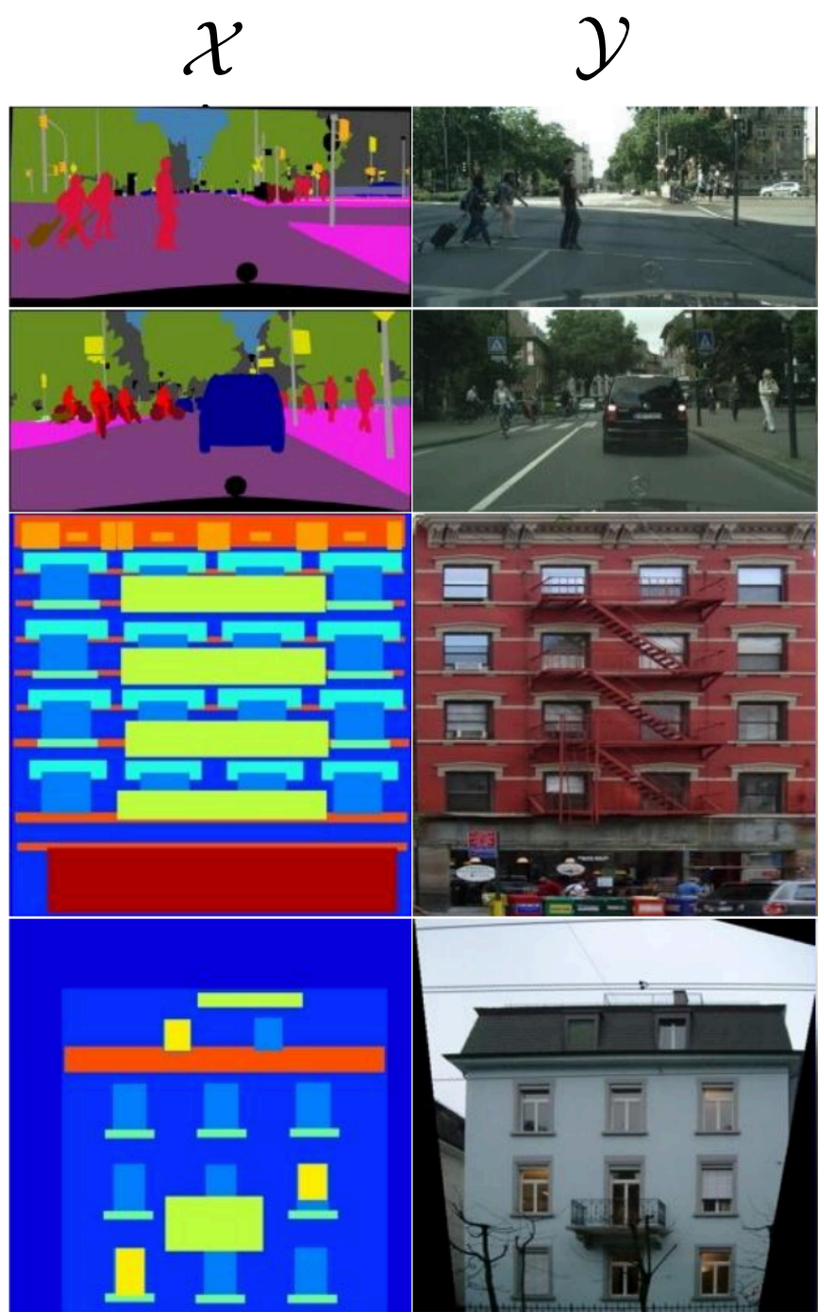
$$\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$$

Paired image-to-image translation

$$\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$$

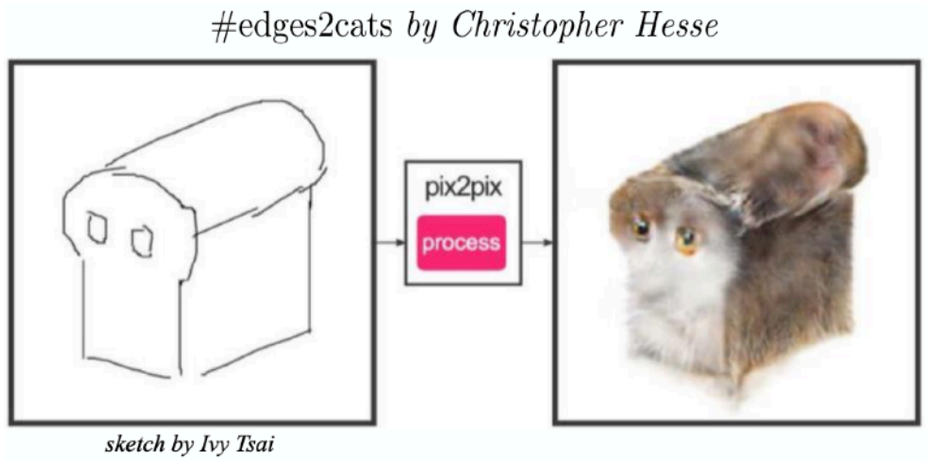
Unpaired

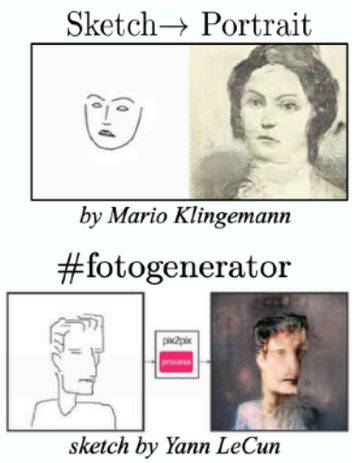
$$\mathbf{x} \sim p(\mathbf{x}), \ \mathbf{y} \sim p(\mathbf{y})$$

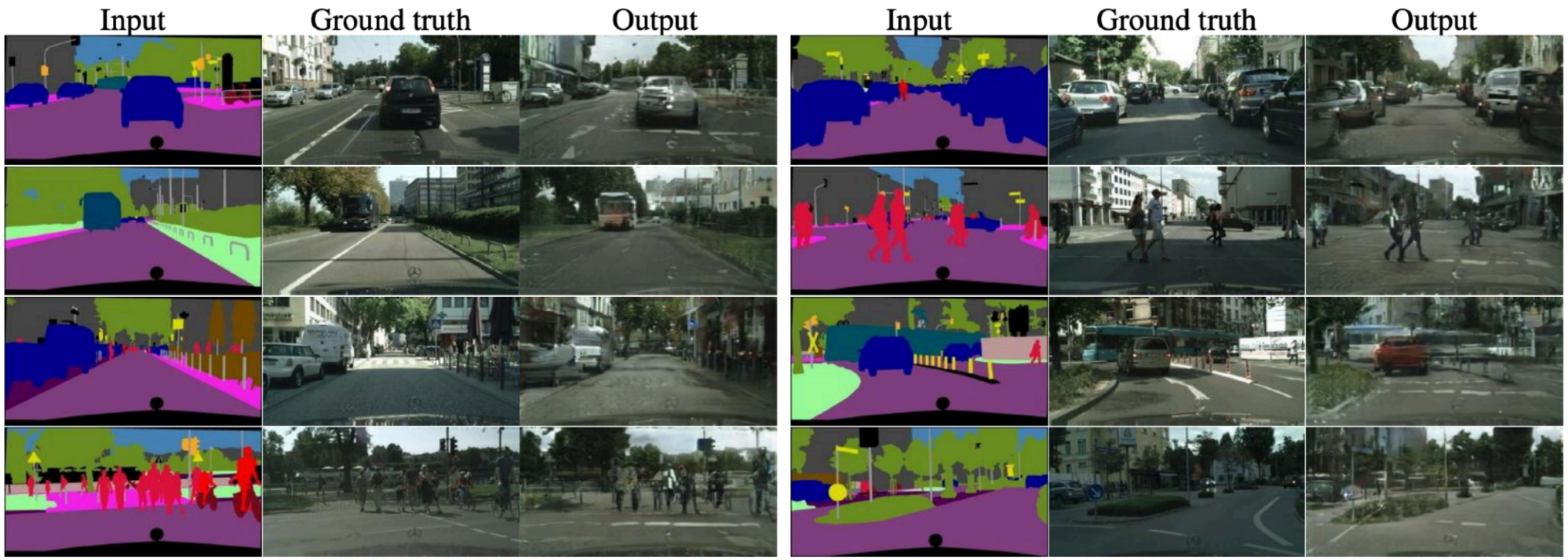


Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Pix2Pix: Results and Applications

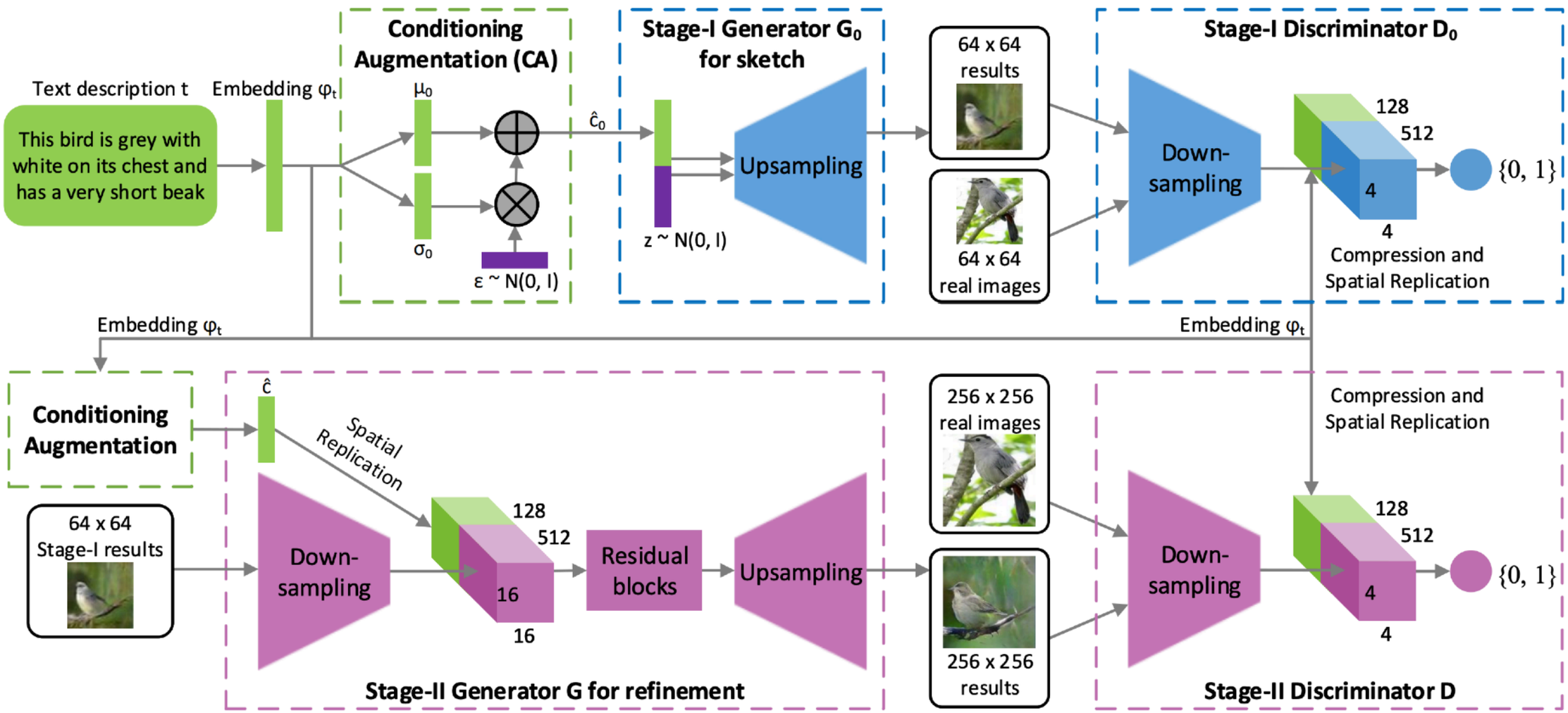






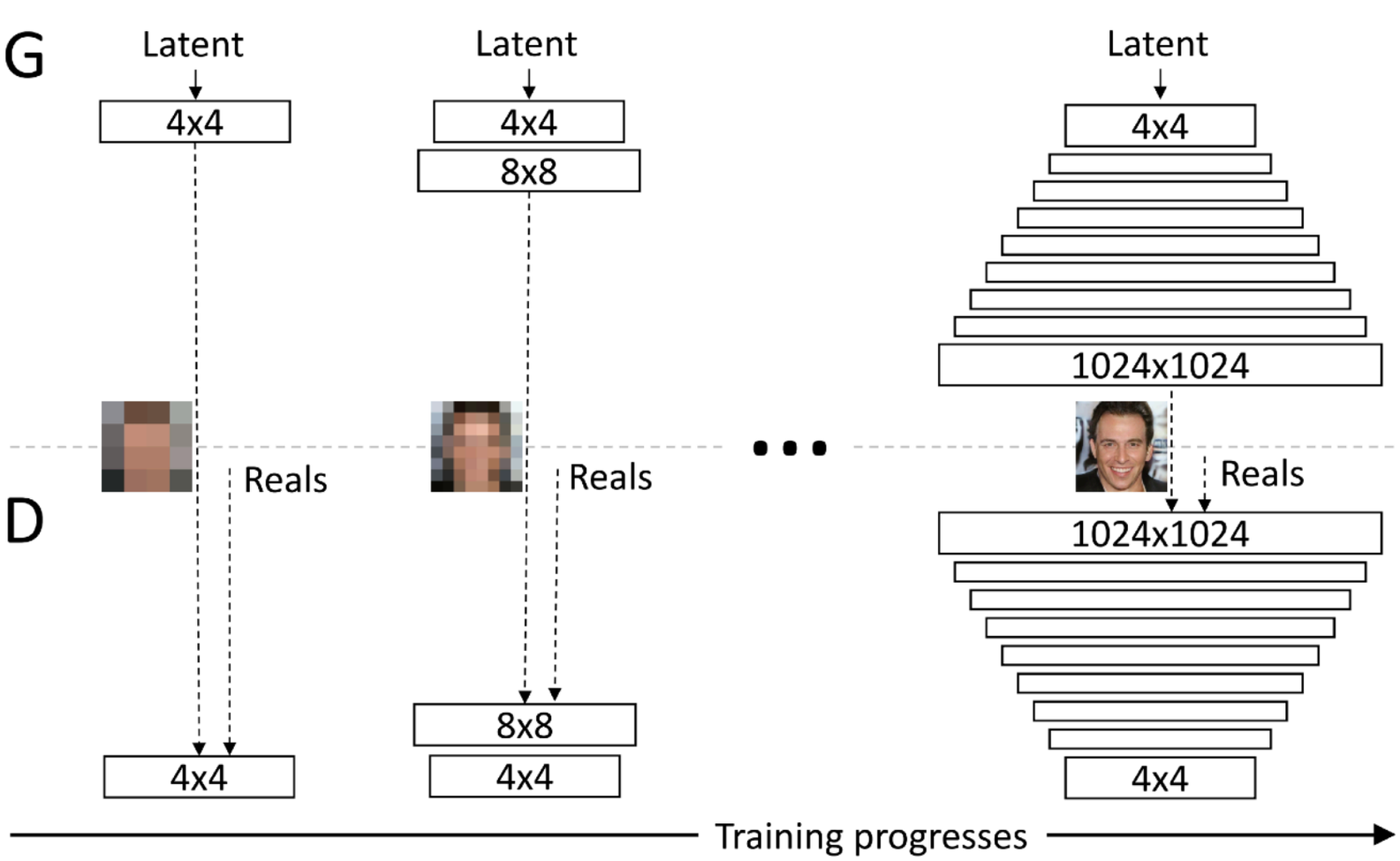
Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Multi-stage Architectures: StackGAN



Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." ICCV'2017

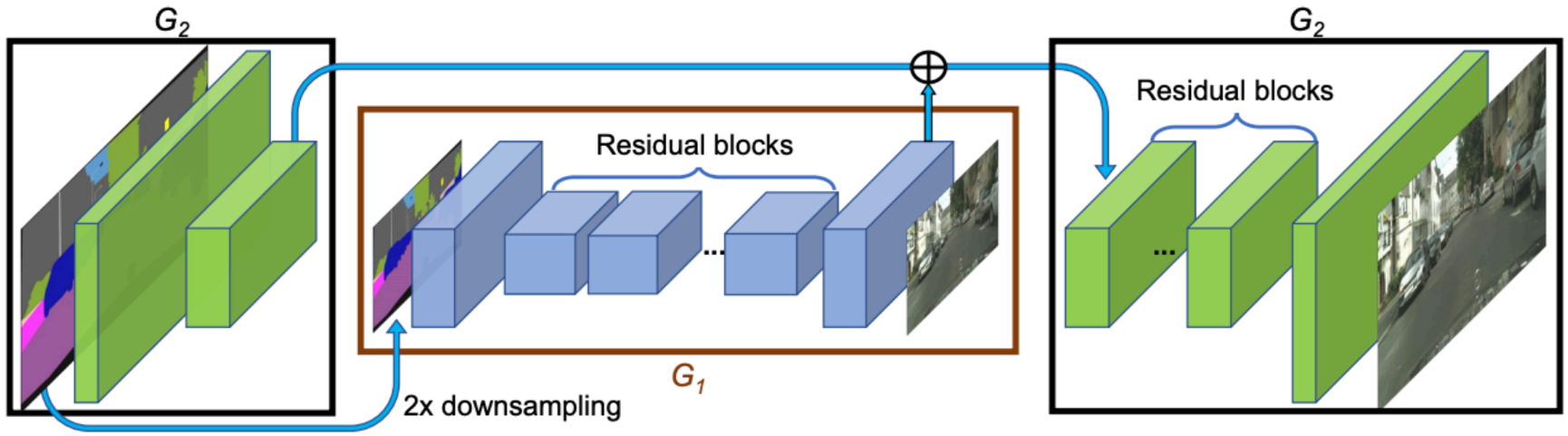
Multi-stage Architectures: ProgressiveGAN



Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." ICLR'2017

Pix2PixHD: Multi-stage for I2I-translation

Idea: two-stage coarse-to-fine generation of HD images



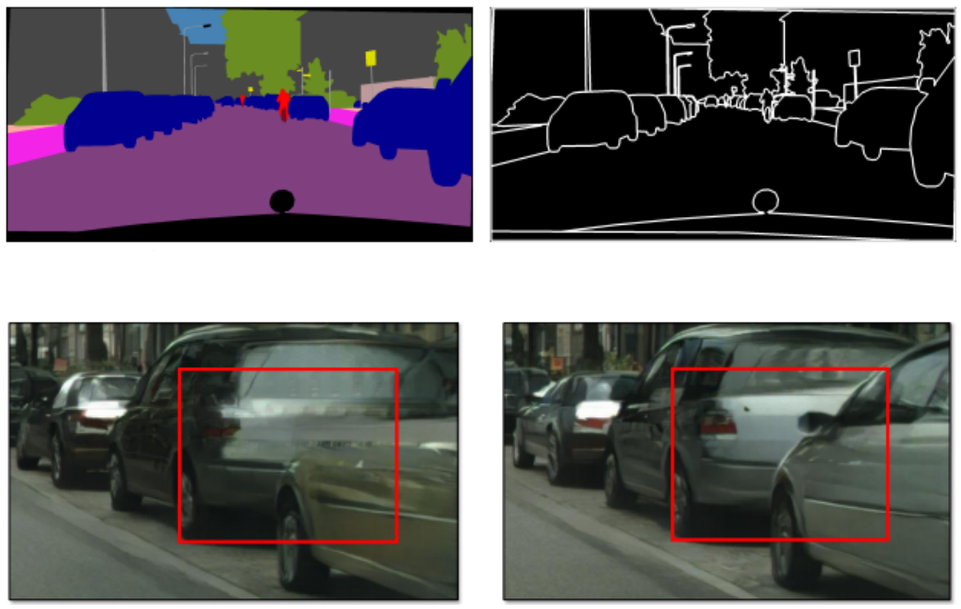
Other contributions:

- Instance segmentation information
- Instance-wise feature embeddings

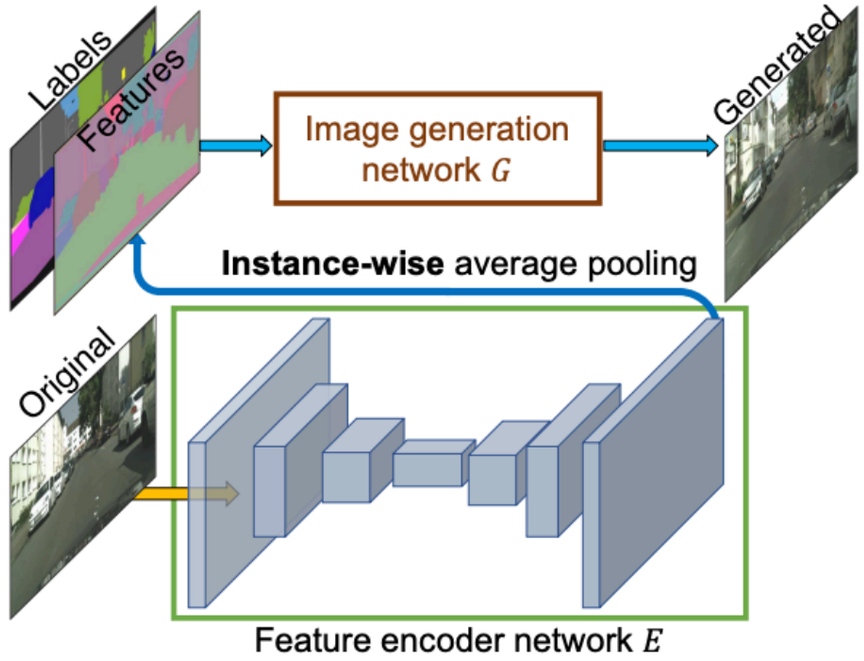
Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Pix2PixHD: Multi-stage for I2I-translation

Instance segmentation



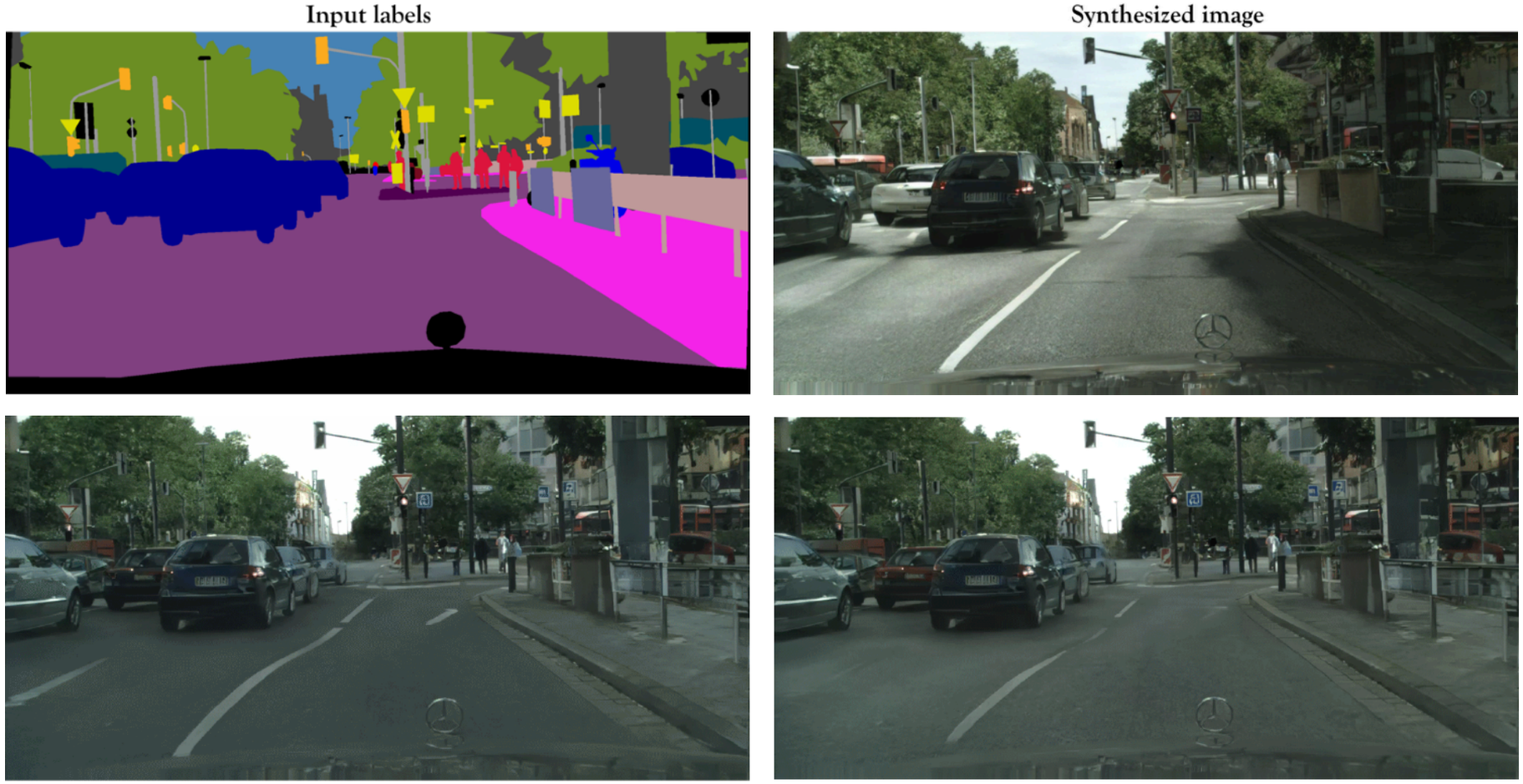
Instance-wise features



Features are then average-pooled for each instance

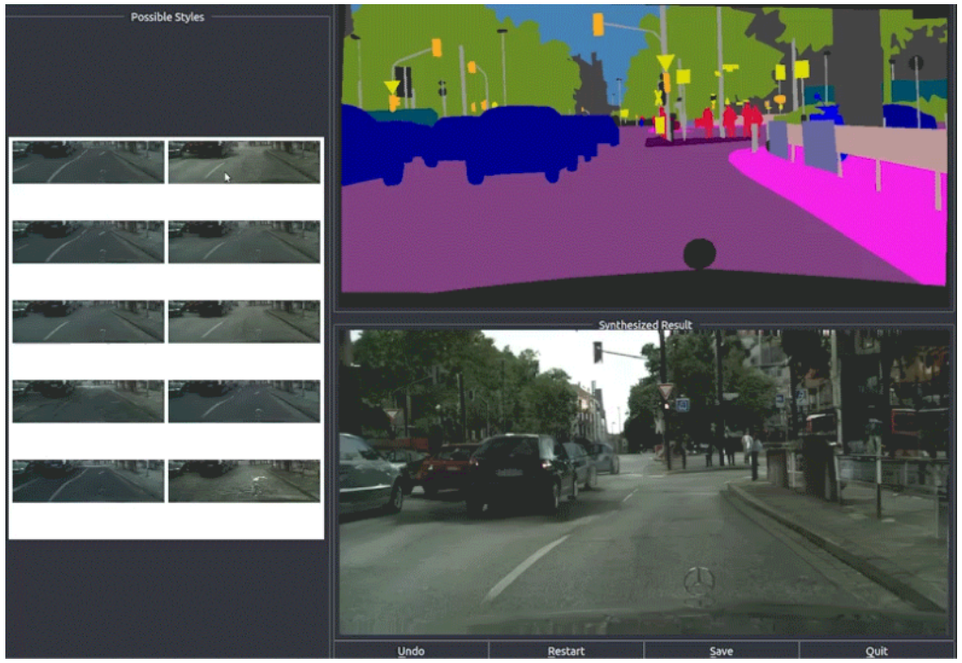
Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Pix2PixHD: Results



Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Pix2PixHD: Results



Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Normalization Layers

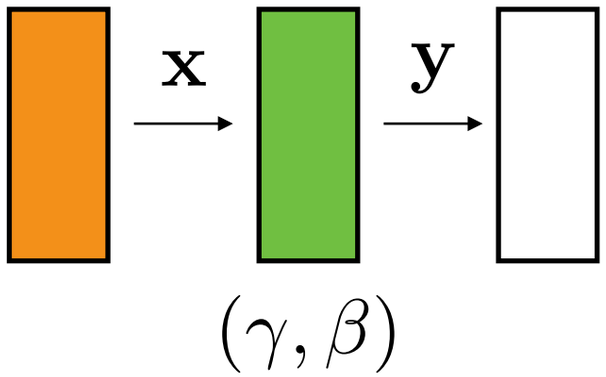
We will cover the following:

- Batch Normalization (BN)
- Instance Normalization (IN)
- Adaptive Instance Normalization (AdaIN)
- SPatially Adaptive DEnormalization (SPADE)

There are many, many more….

Batch Normalization

Problem: small changes in the initial layers result in significant changes in the deeper layers. This way deeper layers have to learn to adapt to different distributions of their inputs.



Compute:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$$

Normalize:

$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{(k)^2} + \epsilon}}$$

For stability

Learned

Transform:

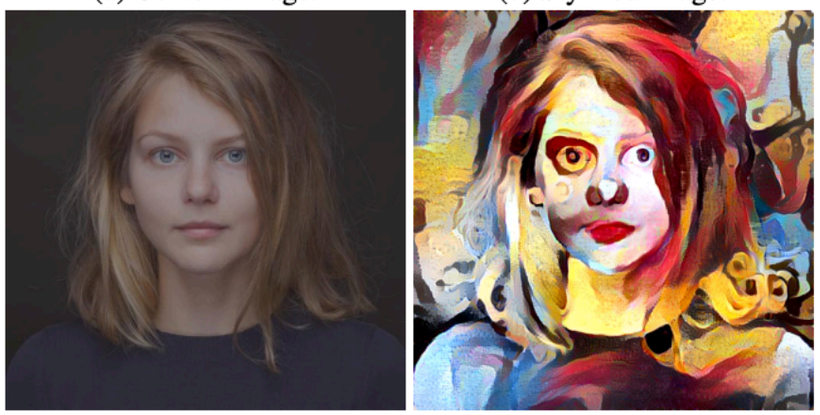
$$y_i^{(k)} = \gamma^{(k)}\hat{x}_i^{(k)} + \beta^{(k)}$$

Ioffe, and Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." 2015

Instance Normalization

Problem: Contrast of a stylized image does not depend on the contrast of content image

(a) Content image. (b) Stylized image.



(c) Low contrast content image. (d) Stylized low contrast image.

Compute:

$$\mu_{ti} = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} x_{tilm}$$

$$\sigma_{ti}^2 = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} (x_{tilm} - mu_{ti})^2$$

Ulyanov, Vedaldi, and Lempitsky. "Instance normalization: The missing ingredient for fast stylization." 2016

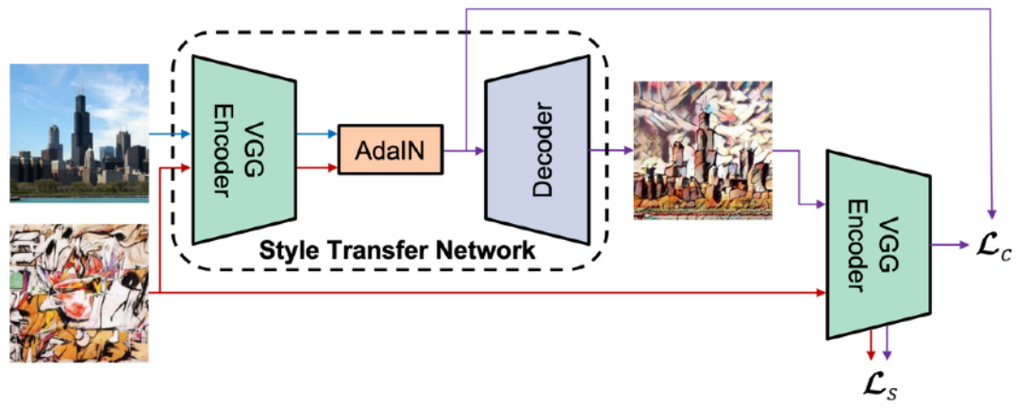
Adaptive Instance Normalization

Idea: If IN normalizes the input to a single style specified by the affine parameters, is it possible to adapt it to arbitrarily given styles by using adaptive affine transformations?

Layer:

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

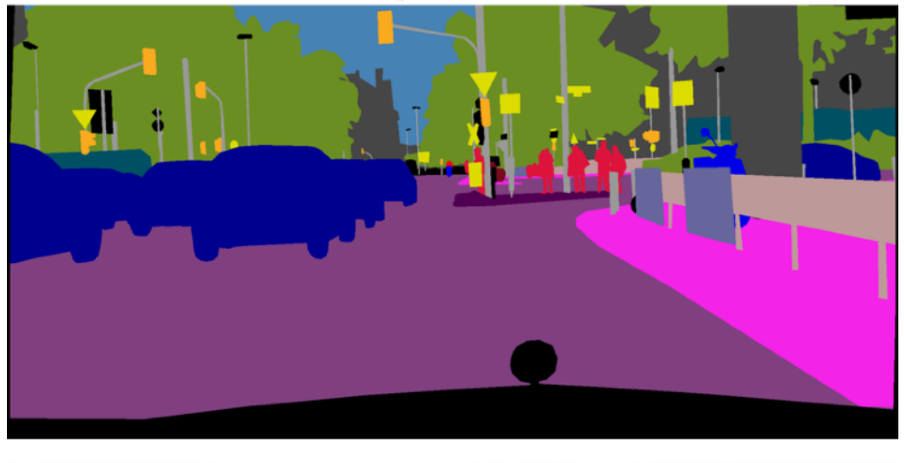
Style transfer framework:



Huang, Xun, and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization." ICCV'2017.

Pix2PixHD: Results

Input labels



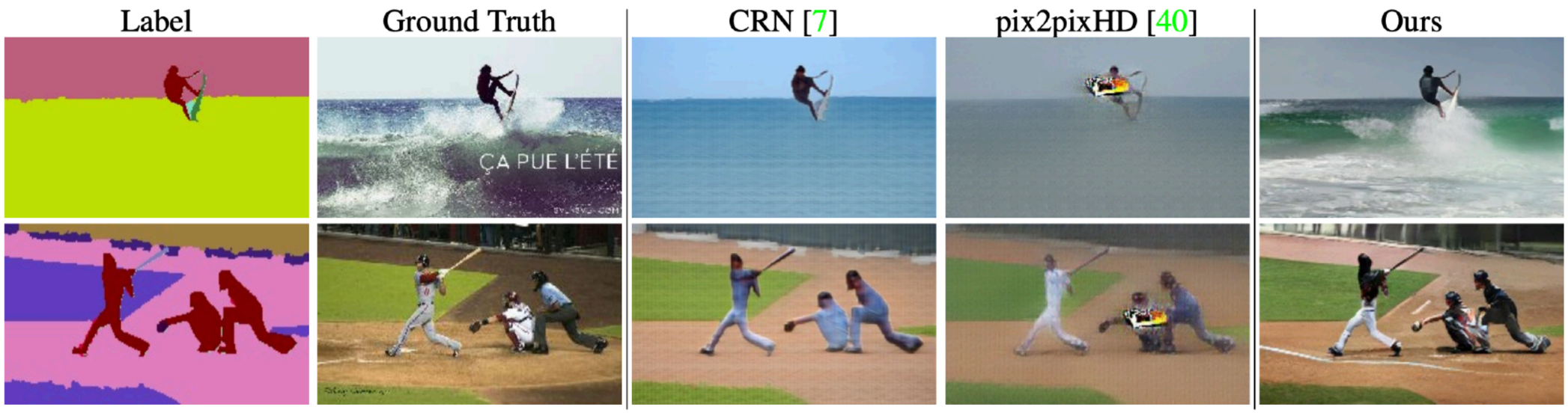
Synthesized image

There are many diverse and complex segments

Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Pix2PixHD: Results

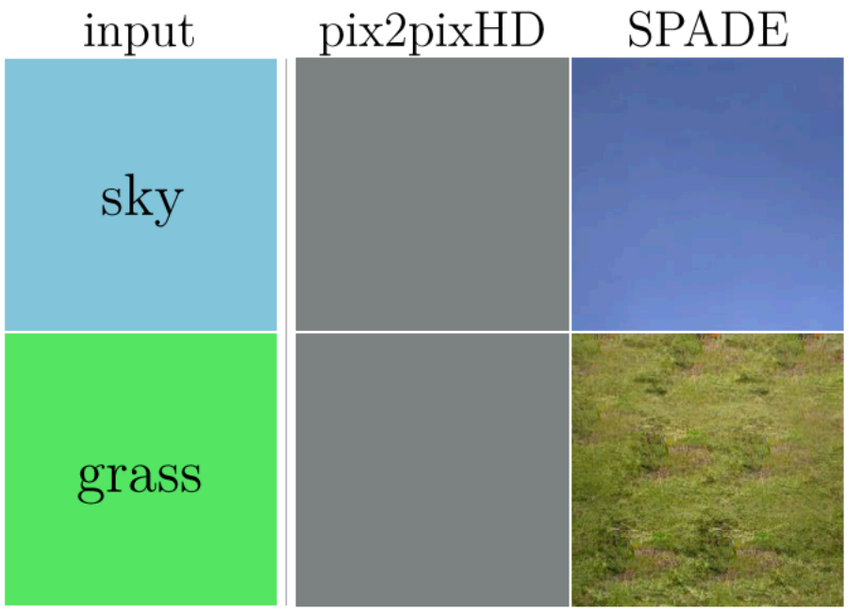
Problem: when segments are large normalization hurts performance



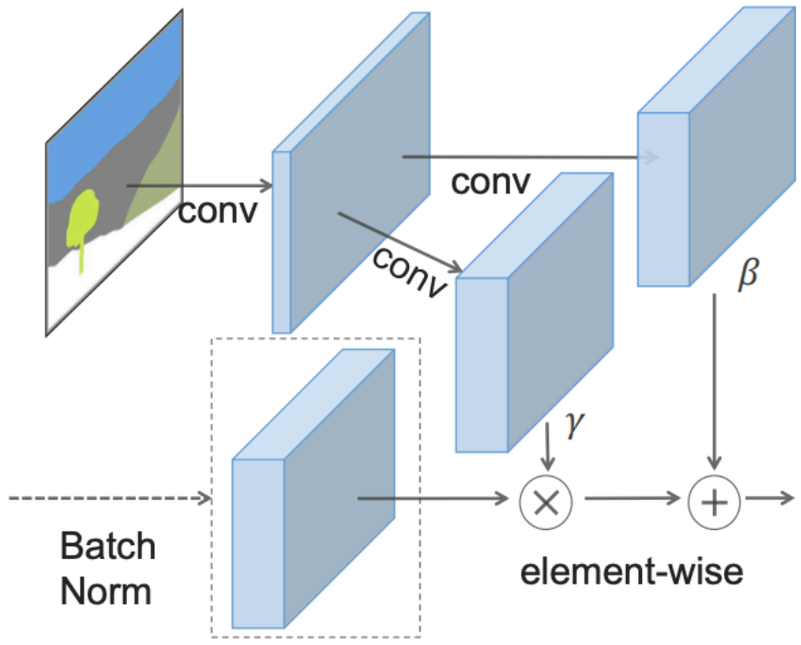
Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

SPatially Adaptive DEnormalization (SPADE)

Problem: when segments are large normalization hurts performance



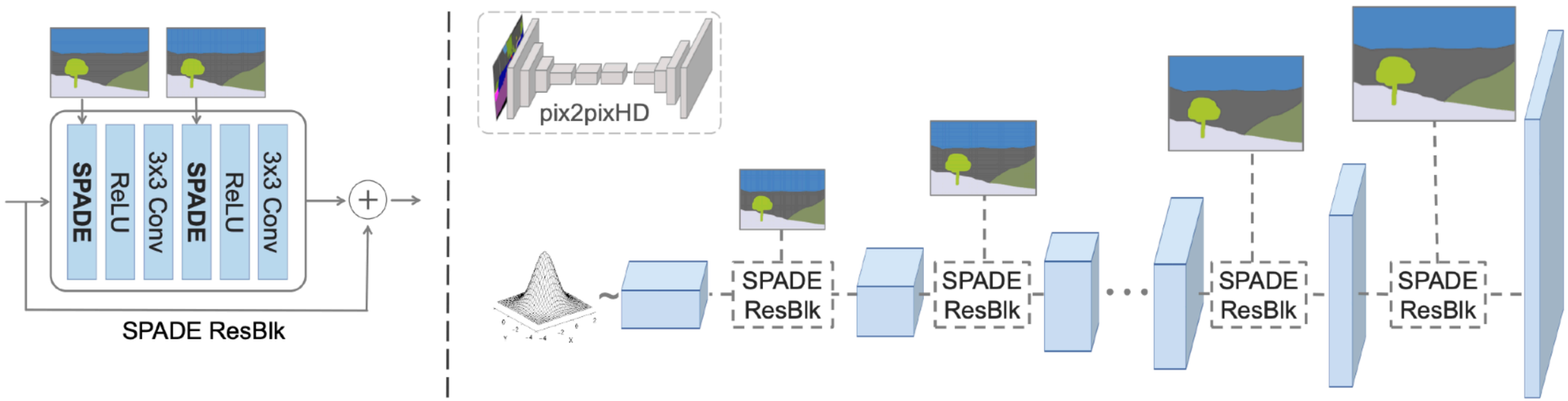
SPADE Block:



Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

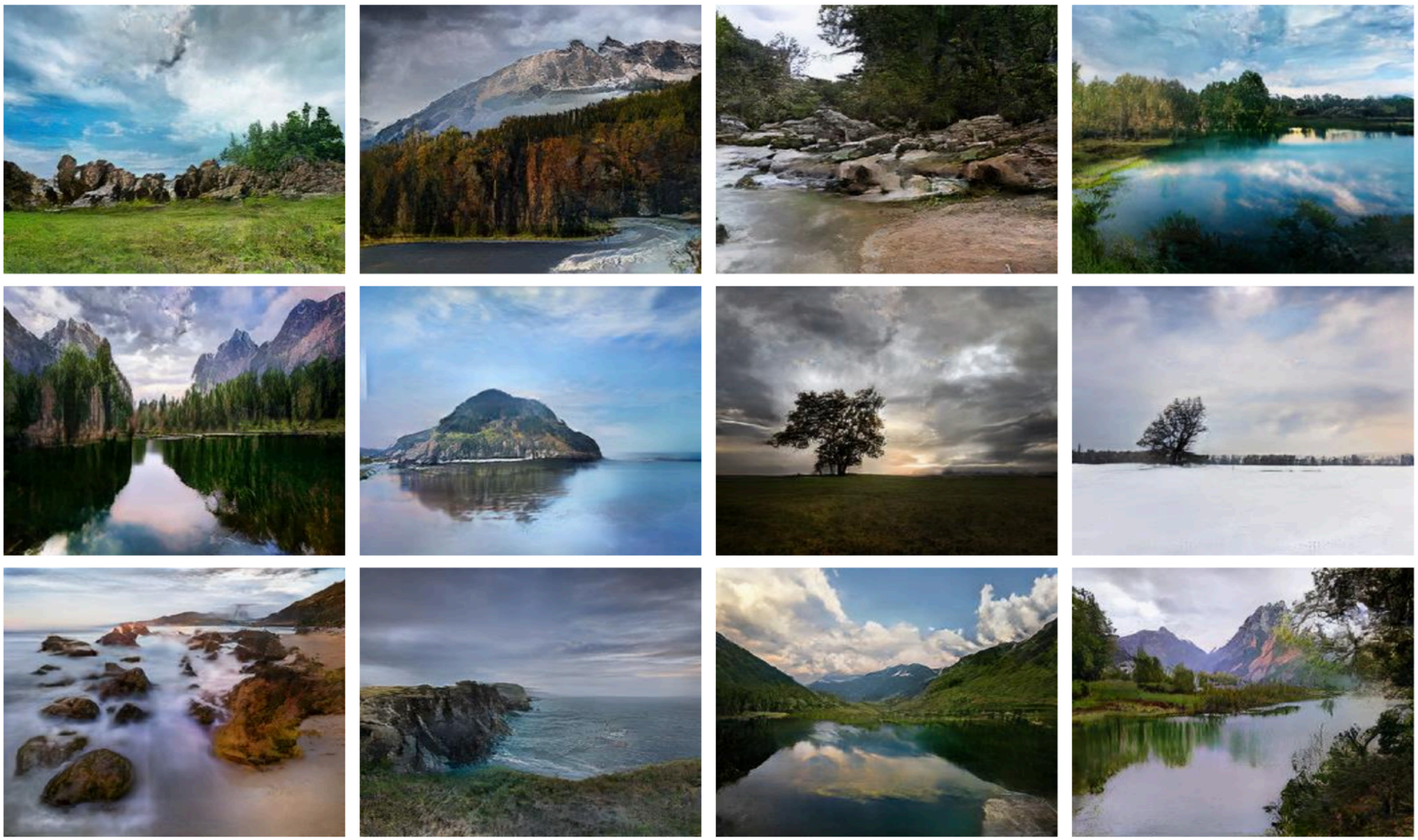
SPADE Generator

No Encoder is required



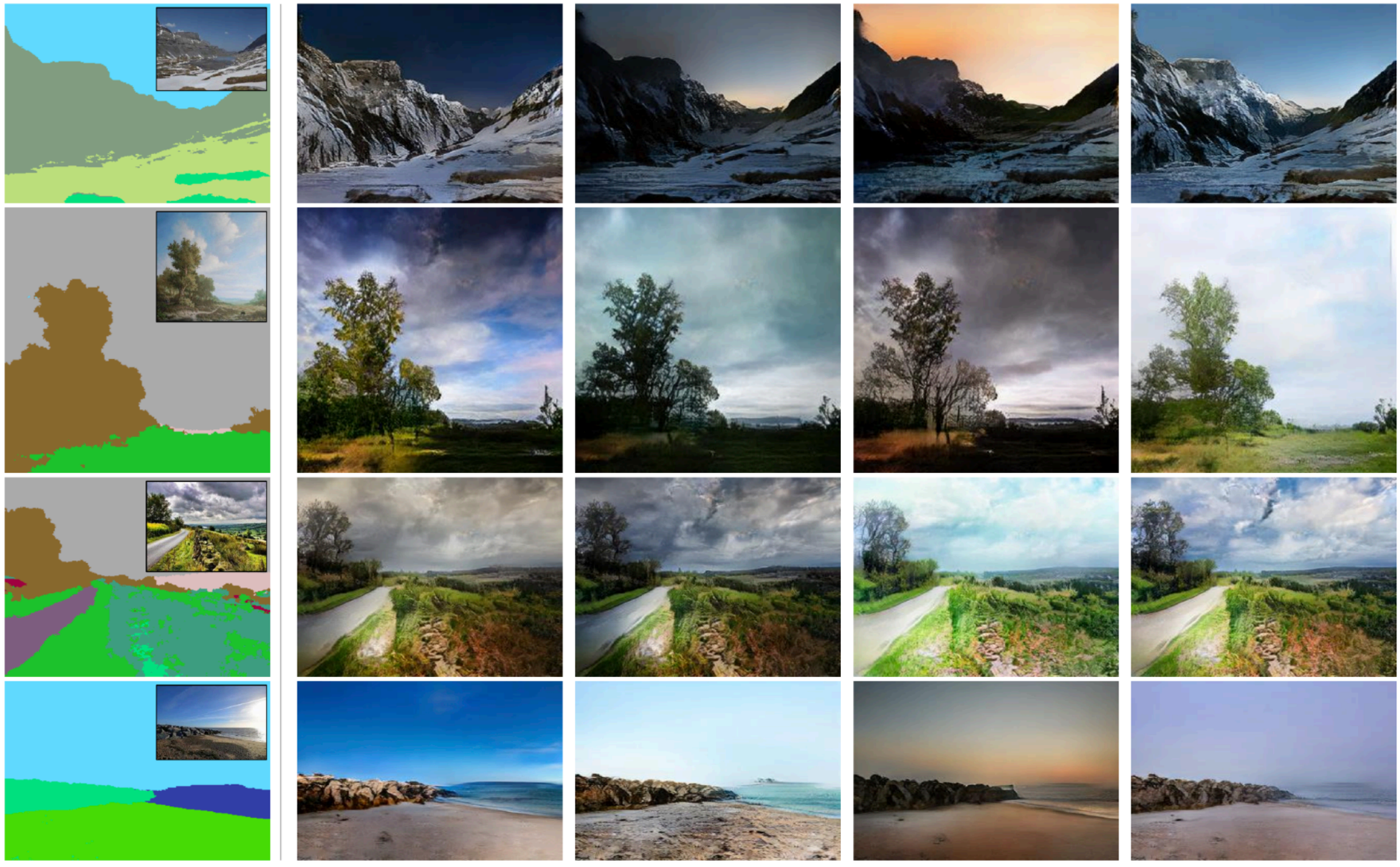
Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

SPADE Results



Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

SPADE Results



Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

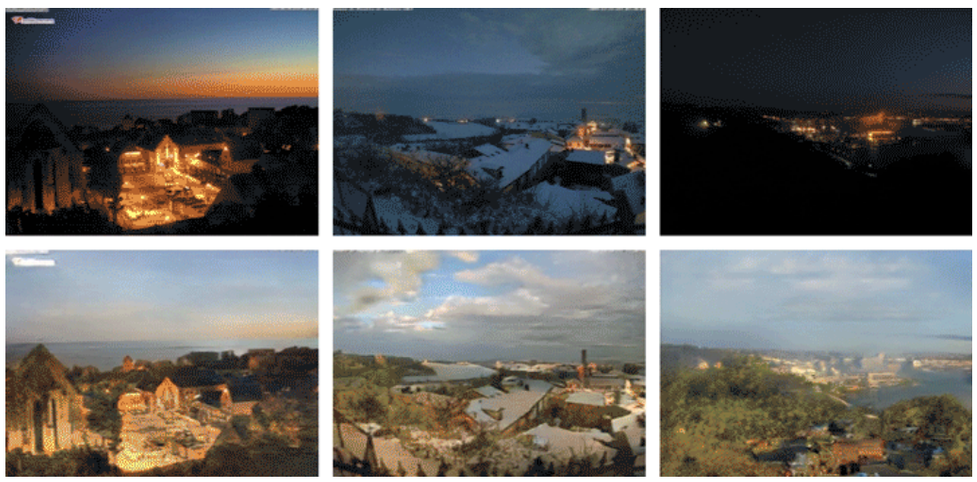
Multimodal Paired Image-to-image Translation

There are many ways of translating an image into another domain

$$\mathbf{x_z} \sim p(\mathbf{x}|\mathbf{y}, \mathbf{z}), \ \ \mathbf{z} \sim p(\mathbf{z})$$

Paired image-to-image translation

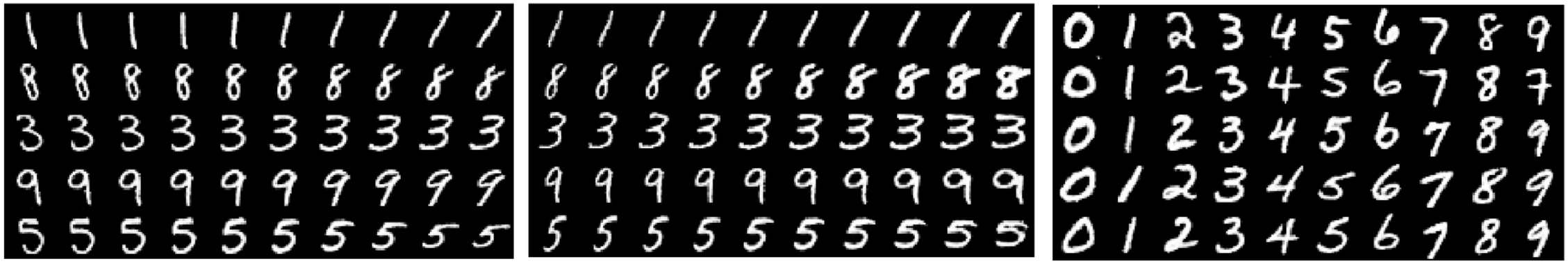
$$\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$$



Zhu, Jun-Yan, et al. "Toward multimodal image-to-image translation." NIPS'2017

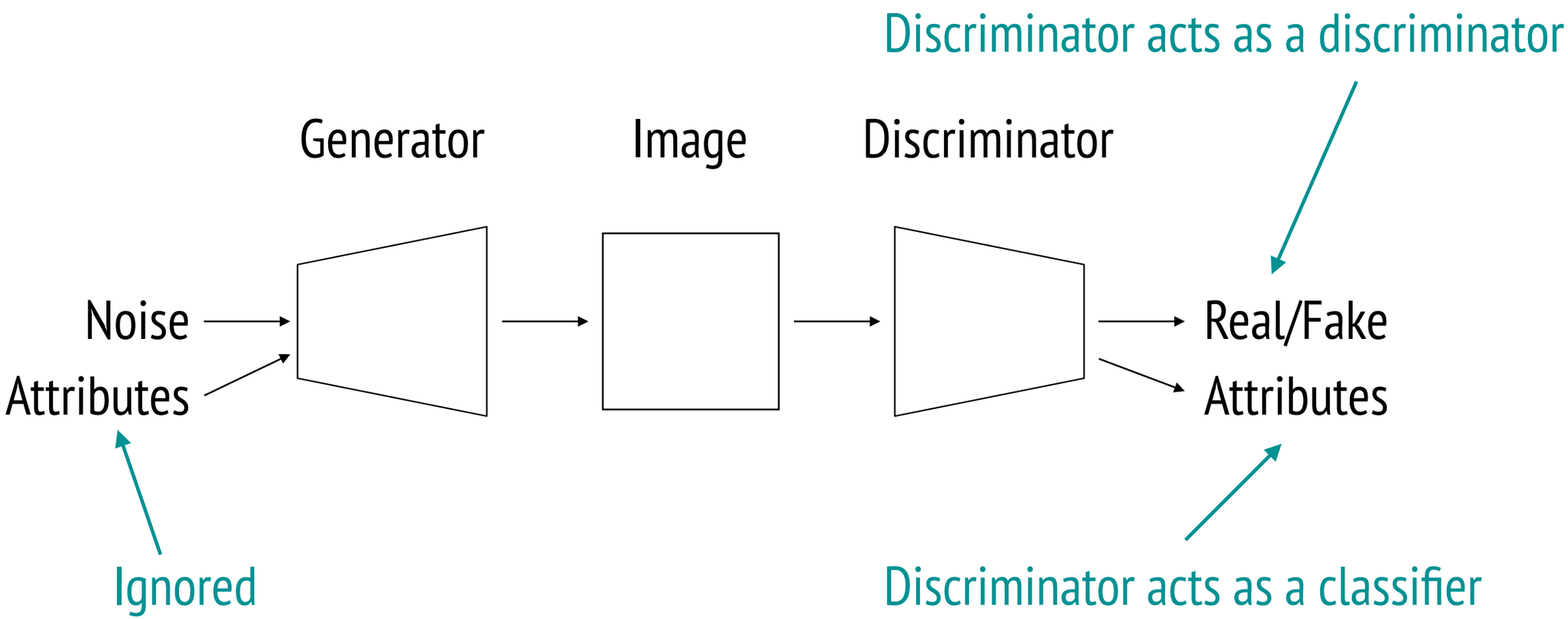
InfoGAN: Disentangling Modes of Variation

Idea: Can we condition the generative process of GANs to control particular attributes of interest in the generated images?



Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." NIPS 2016.

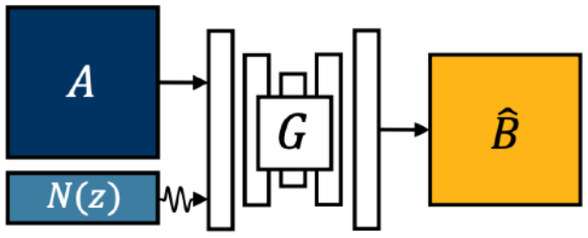
InfoGAN: Disentangling Modes of Variation



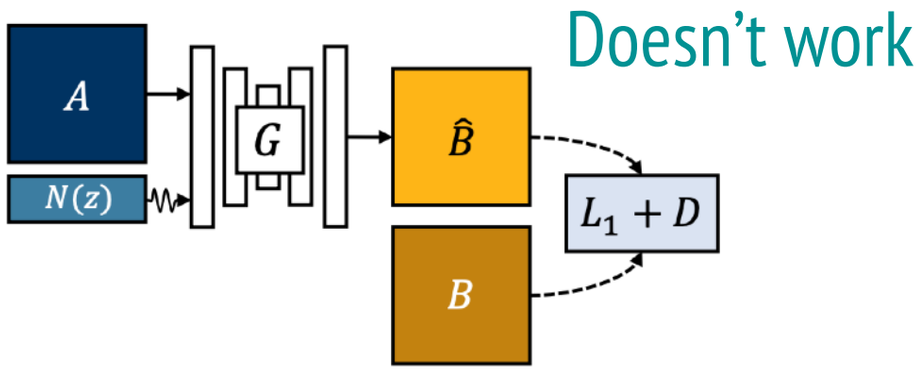
InfoGAN: To enforce the use of the required attributes one has to penalize the Generator if the images do not have this attribute

Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." NIPS 2016.

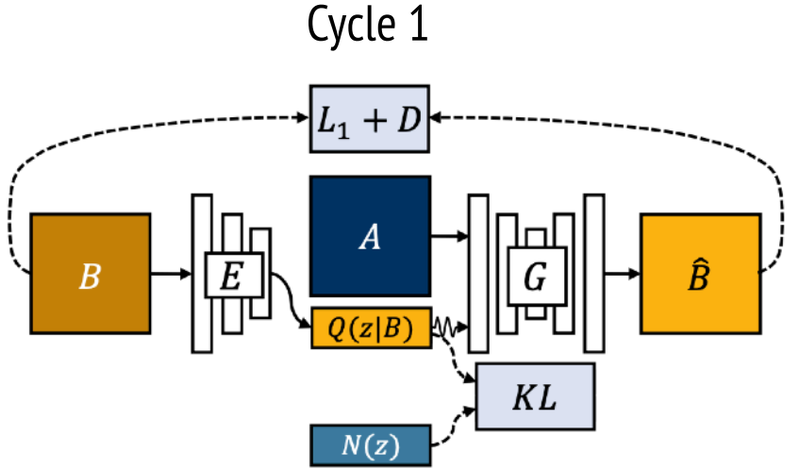
BiCycleGAN: Multimodal Image-to-image

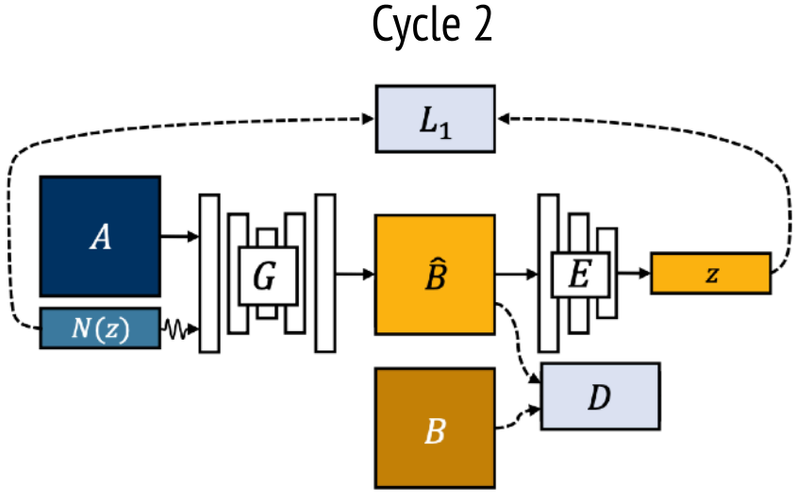


Inference



Augmented pix2pix

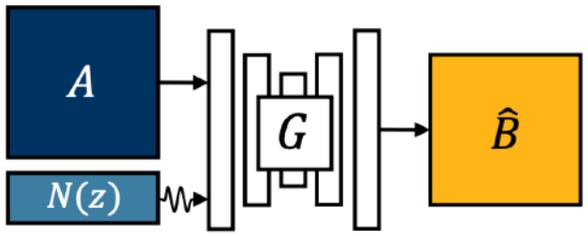




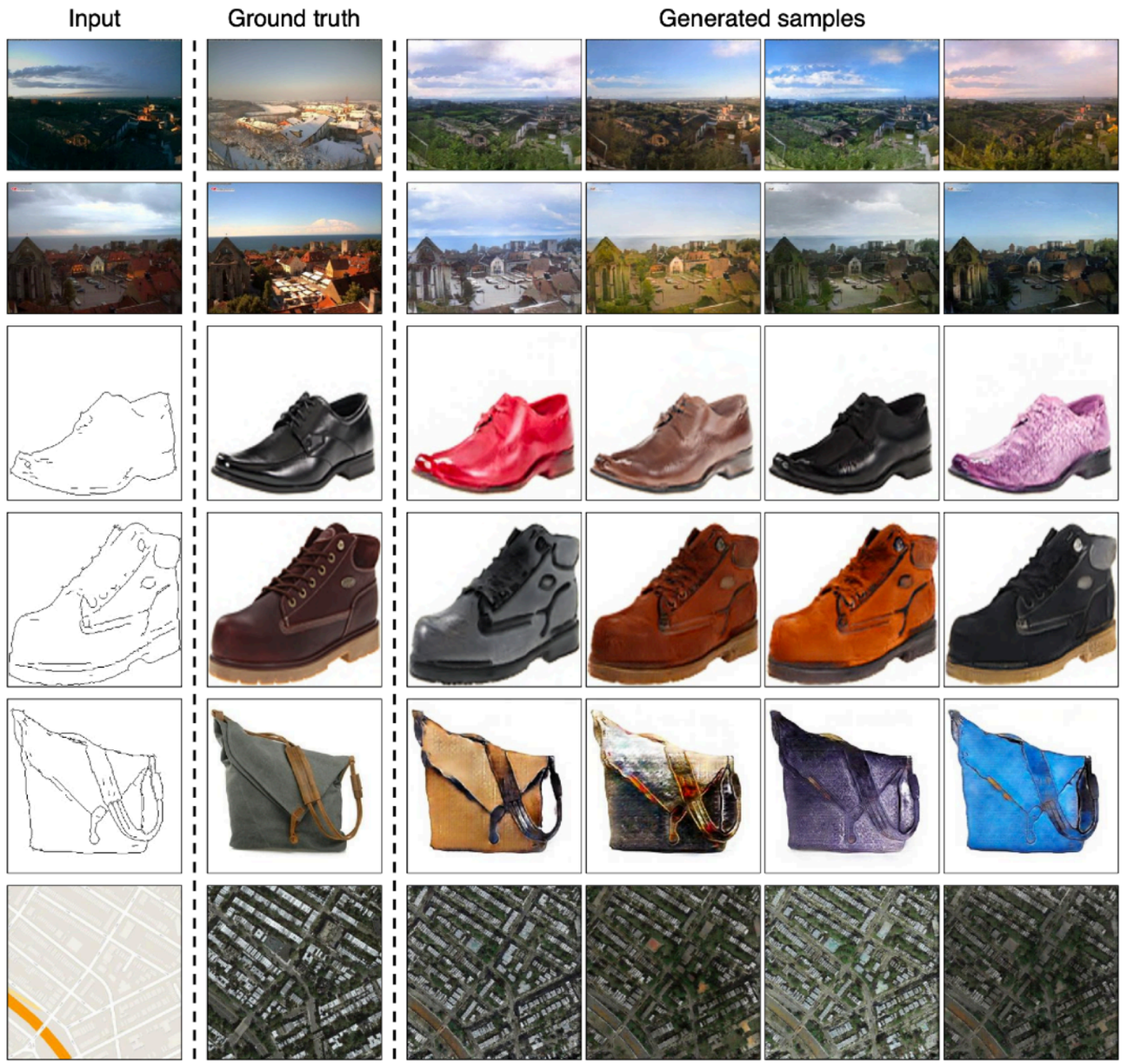
Training BicycleGAN

Zhu, Jun-Yan, et al. "Toward multimodal image-to-image translation." NIPS'2017

BiCycleGAN: Multimodal Image-to-image



Inference



Zhu, Jun-Yan, et al. "Toward multimodal image-to-image translation." NIPS'2017

Multimodal Unpaired Image-to-image Translation

There are many ways of translating an image into another domain

$$\mathbf{x_z} \sim p(\mathbf{x}|\mathbf{y}, \mathbf{z}),\ \mathbf{z} \sim p(\mathbf{z})$$

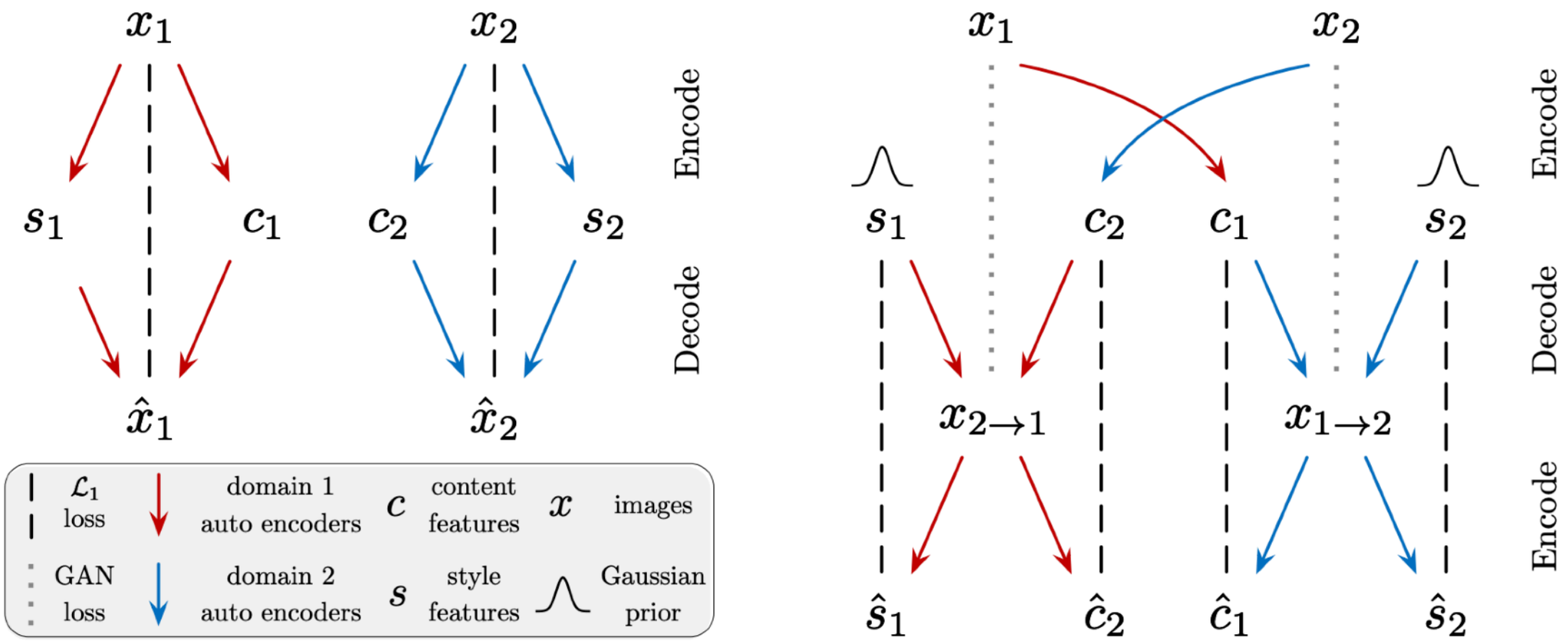
Unpaired

$$\mathbf{x} \sim p(\mathbf{x}),\ \mathbf{y} \sim p(\mathbf{y})$$

Huang, Xun, et al. "Multimodal unsupervised image-to-image translation." ECCV'2018.

Multimodal Unpaired Image-to-image Translation

Assumption: Images can be decomposed into style and content



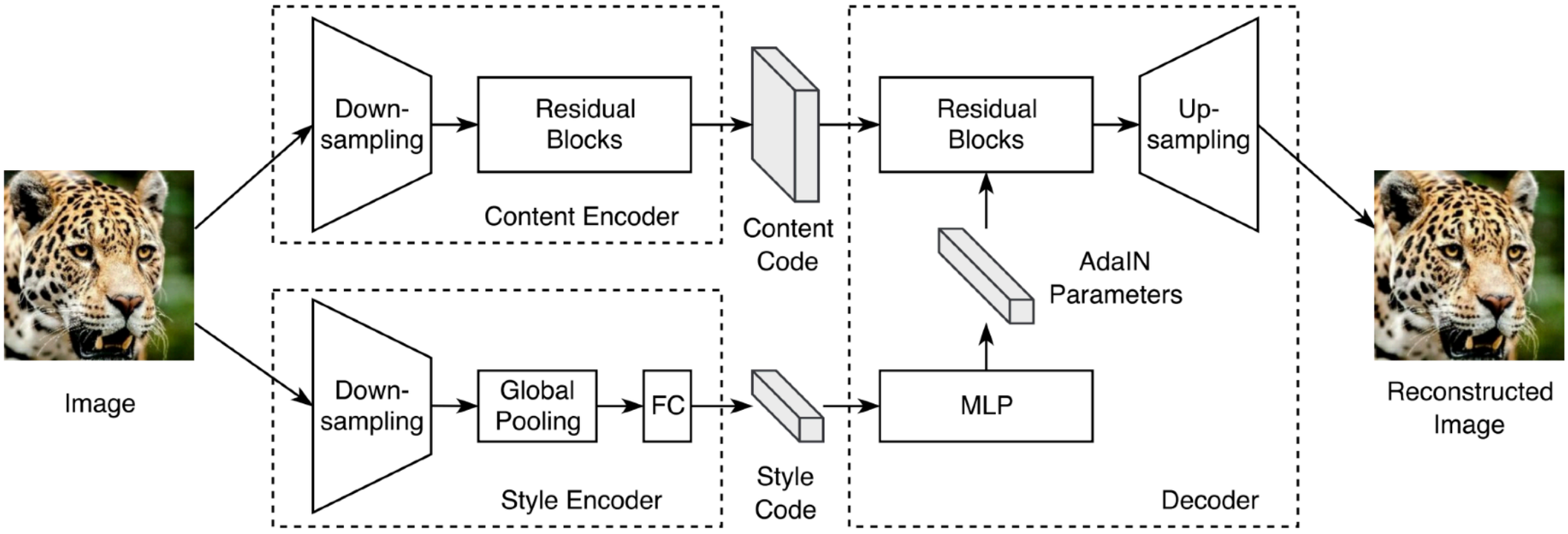
(a) Within-domain reconstruction

(b) Cross-domain translation

Huang, Xun, et al. "Multimodal unsupervised image-to-image translation." ECCV'2018.

Multimodal Unpaired Image-to-image Translation

Assumption: Images can be decomposed into style and content



Huang, Xun, et al. "Multimodal unsupervised image-to-image translation." ECCV'2018.

MUNIT: Results

Assumption: Images can be decomposed into style and content

Huang, Xun, et al. "Multimodal unsupervised image-to-image translation." ECCV'2018.