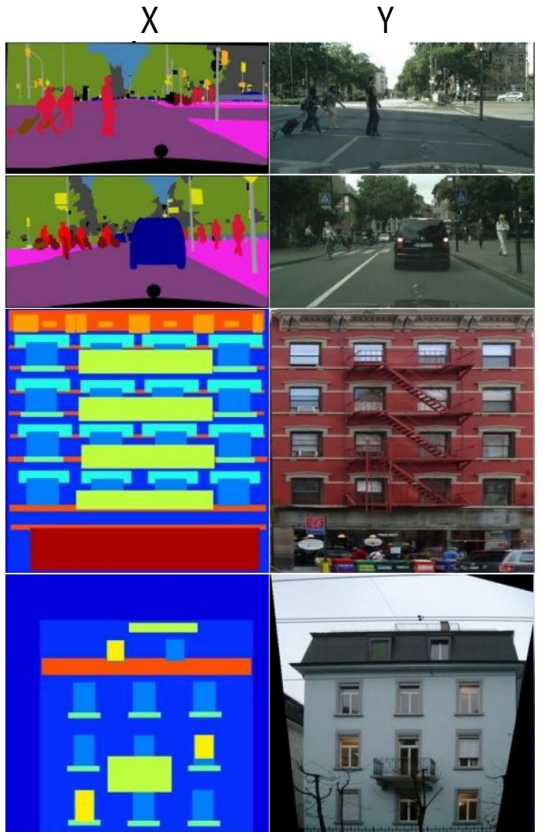
Image-to-image Translation

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

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

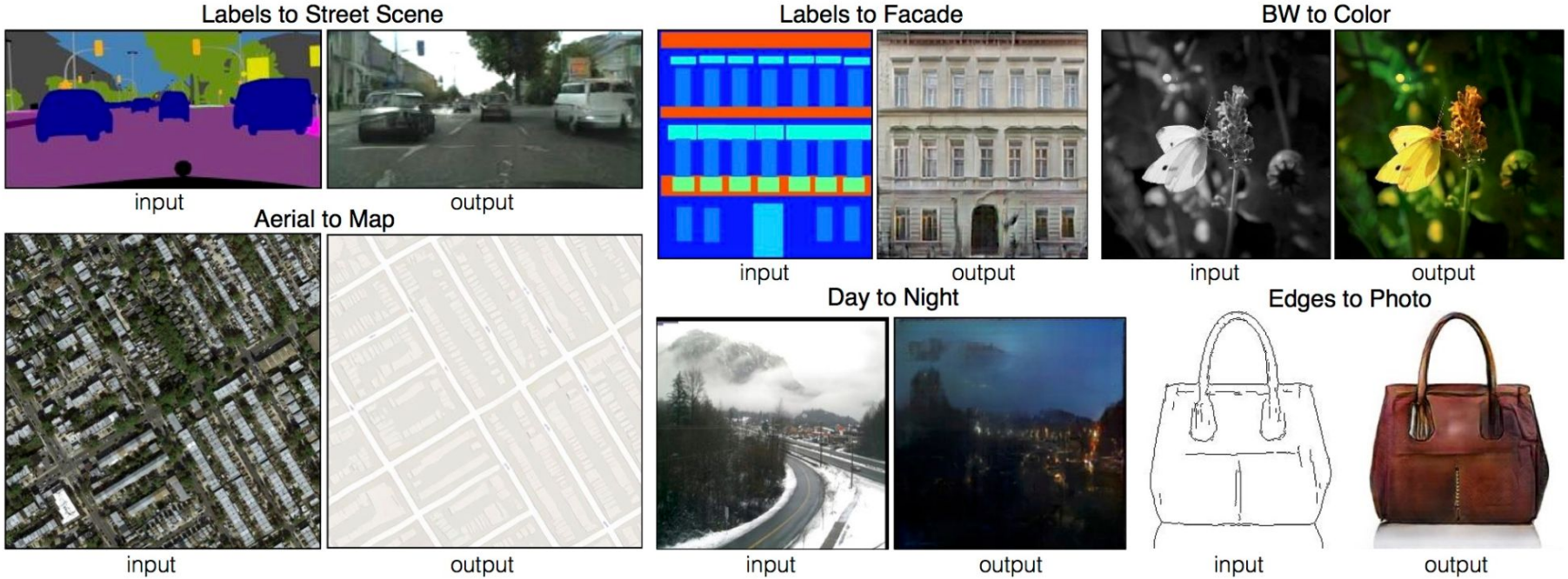
Paired image-to-image translation

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



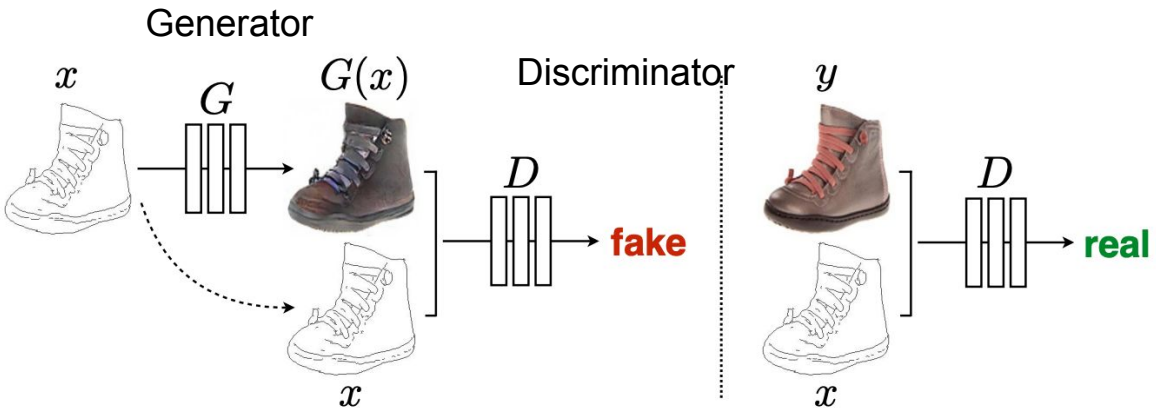
Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." CVPR'2017

Image-to-image Translation



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

Image-to-image Translation: Pix2Pix



Combined GAN-loss and reconstruction loss:

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

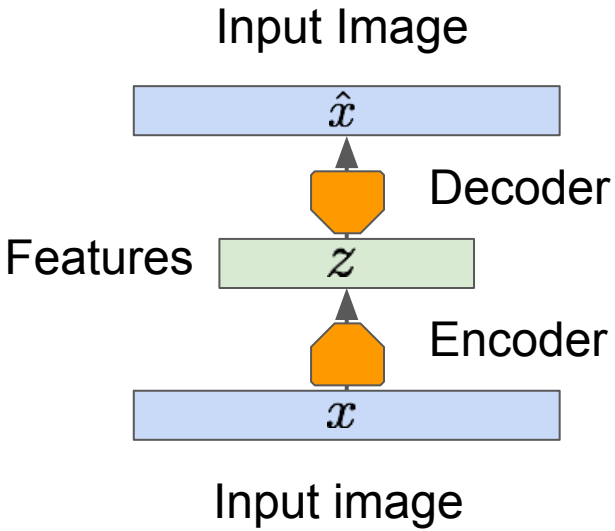
$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1]$$

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

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

Encoder-Decoder

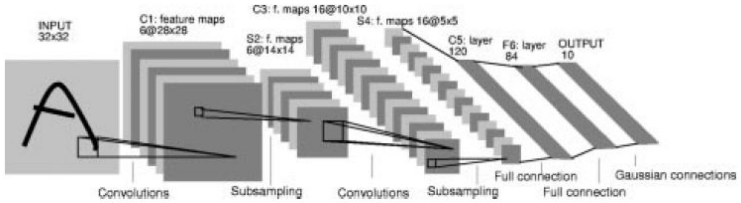
What is the architecture of G

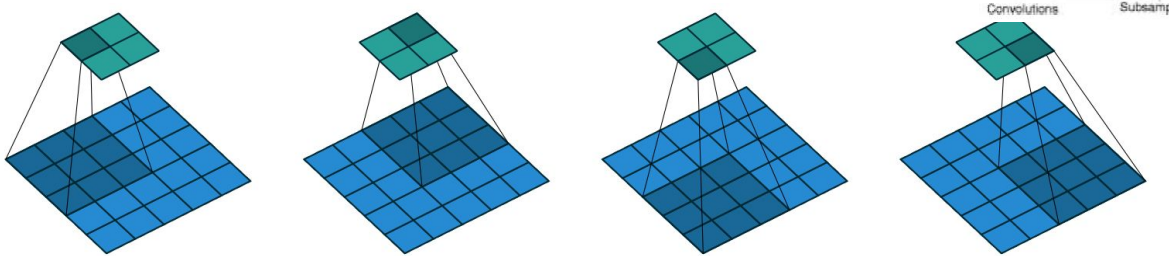


“De-convolution”

Architecture:

- Encoder: ConvNet architecture

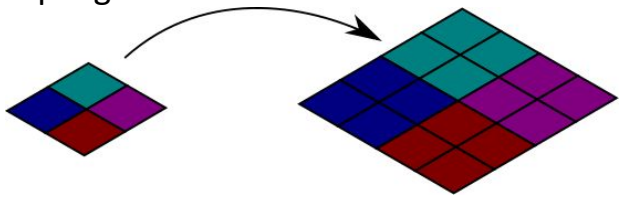




Convolving a 3 × 3 kernel over a 5 × 5 input using 2 × 2 strides

- Decoder: How to go from a vector to an image? How to “deconvolve”?

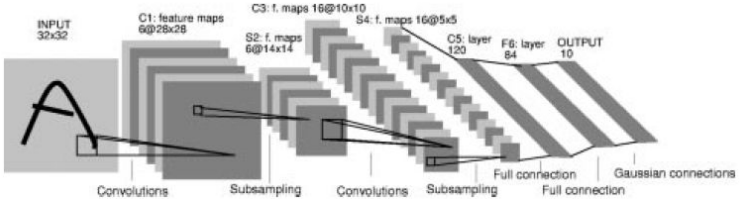
Up-sampling

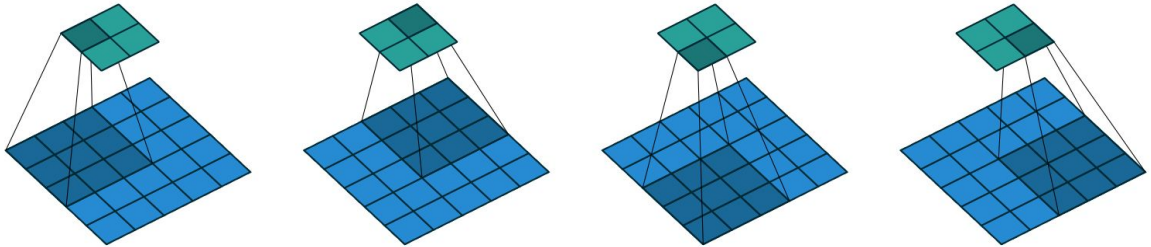


“De-convolution”

Architecture:

- Encoder: ConvNet architecture

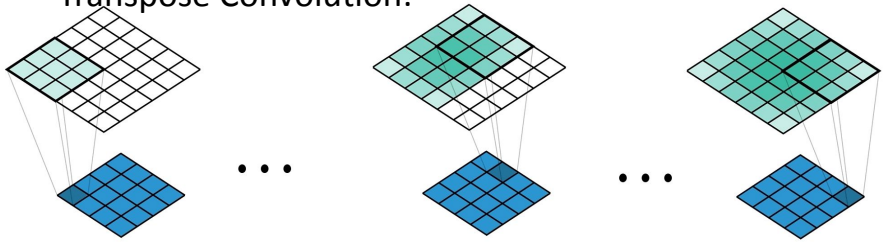




Convolving a 3 × 3 kernel over a 5 × 5 input using 2 × 2 strides

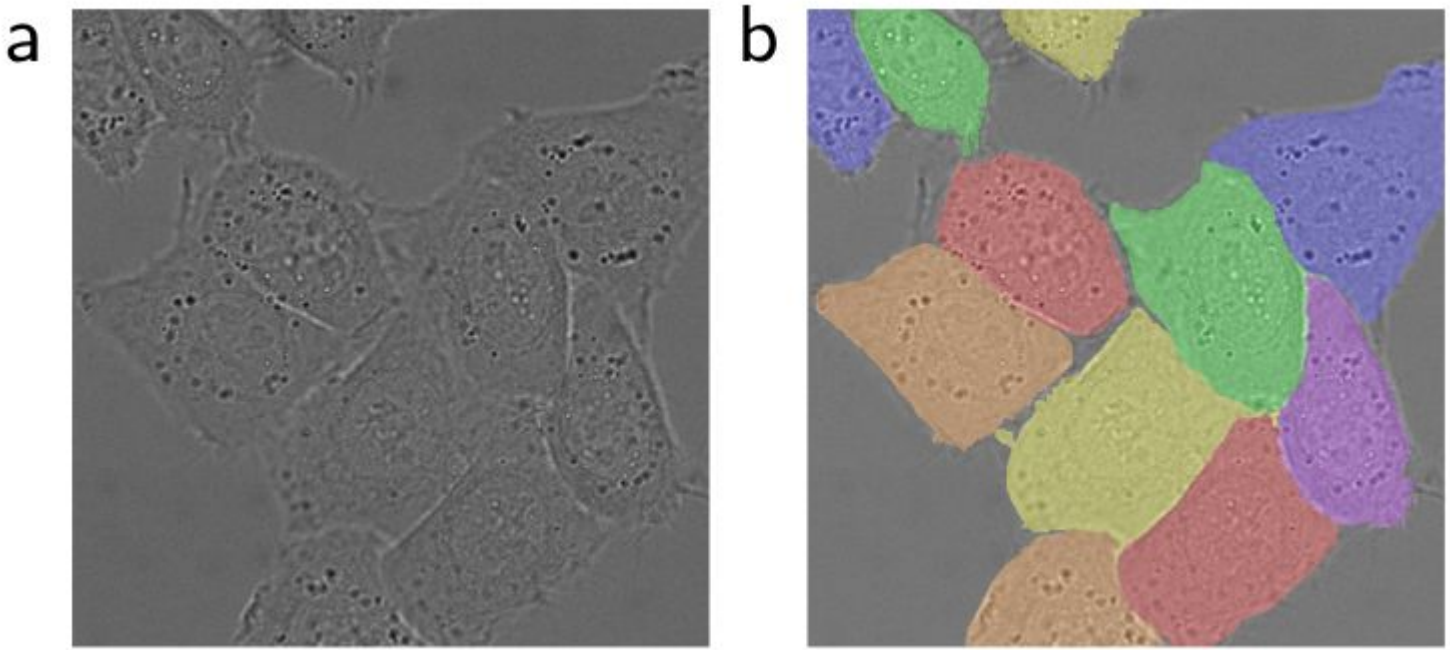
- Decoder: How to go from a vector to an image? How to “deconvolve”?

Transpose Convolution:



A guide to convolution arithmetic for deep learning ,Vincent Dumoulin, Francesco Visin

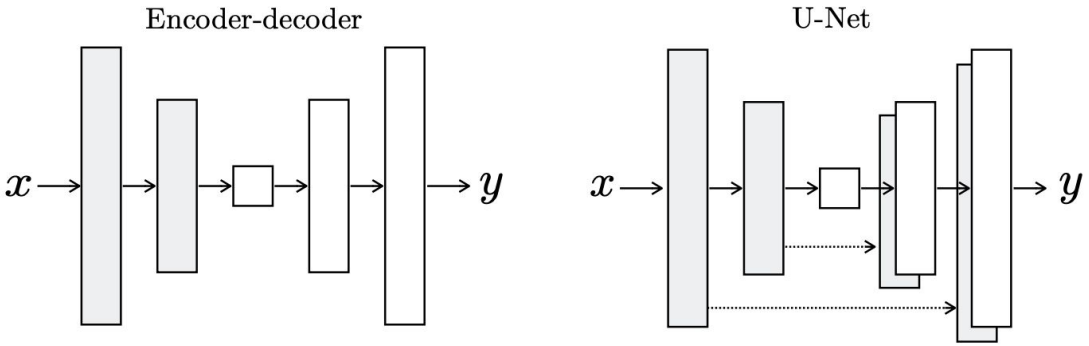
Image-to-image Translation: Pix2Pix



U-Net: Convolutional Networks for Biomedical Image Segmentation Olaf Ronneberger, Philipp Fischer, Thomas Brox MICCAI 2015

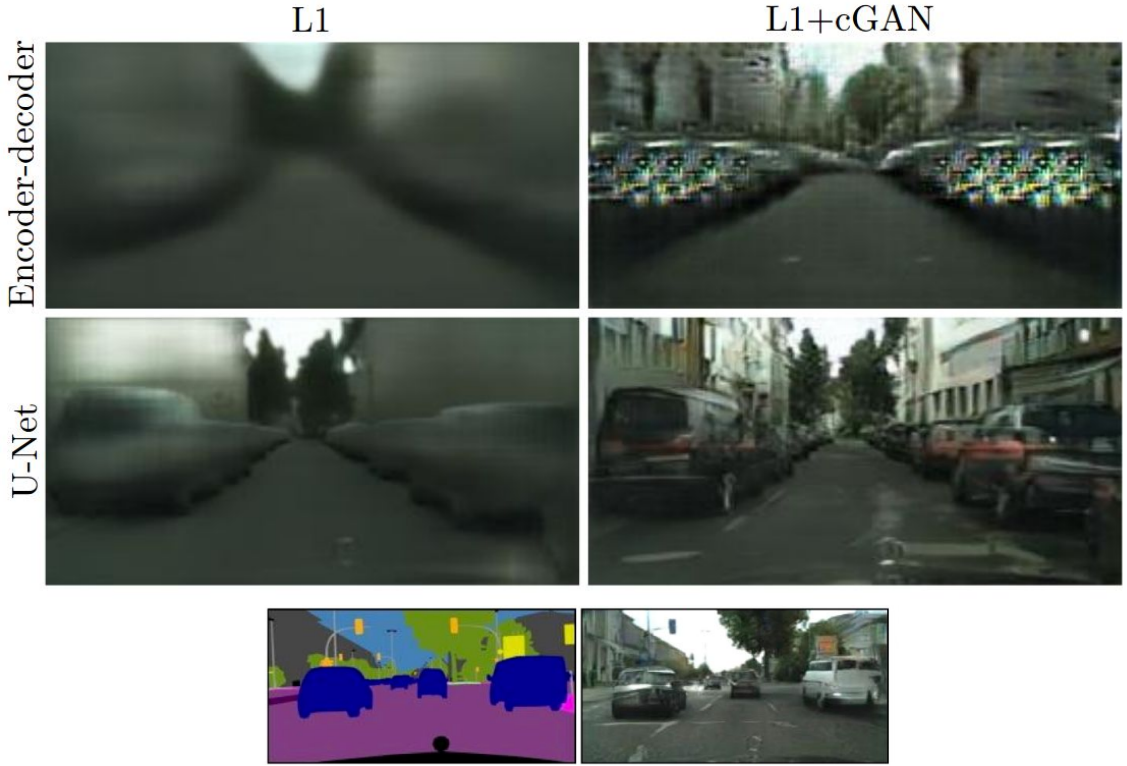
Image-to-image Translation: Pix2Pix

Skip connections in generator



U-Net: Convolutional Networks for Biomedical Image Segmentation Olaf Ronneberger, Philipp Fischer, Thomas Brox MICCAI 2015

Image-to-image Translation: Pix2Pix

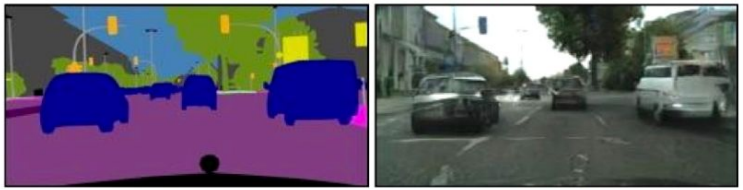


U-Net: Convolutional Networks for Biomedical Image Segmentation Olaf Ronneberger, Philipp Fischer, Thomas Brox MICCAI 2015

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.)



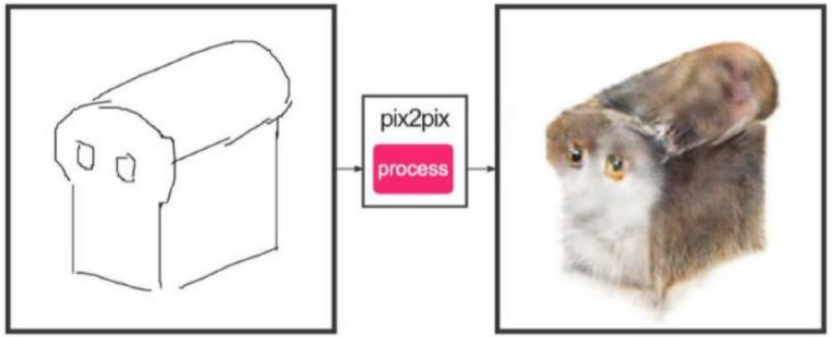
Pix2Pix: Ablations

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.

Pix2Pix: Applications

#edges2cats *by Christopher Hesse*



sketch by Ivy Tsai

Background removal

by Kaihu Chen

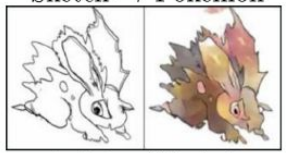
Palette generation

by Jack Qiao

Sketch→ Portrait

by Mario Klingemann

Sketch → Pokemon

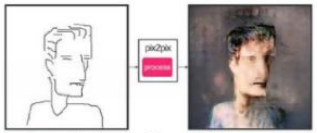


by Bertrand Gondouin

"Do as I do"

by Brannon Dorsey

#fotogenerator



sketch by Yann LeCun

Unpaired Image-to-image Translation

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

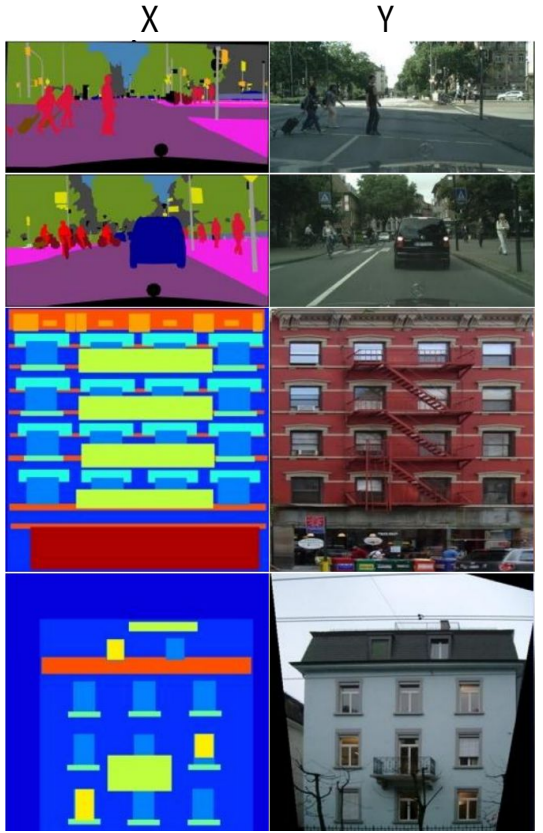
$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{y})$$
$$\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$$

Paired image-to-image translation

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

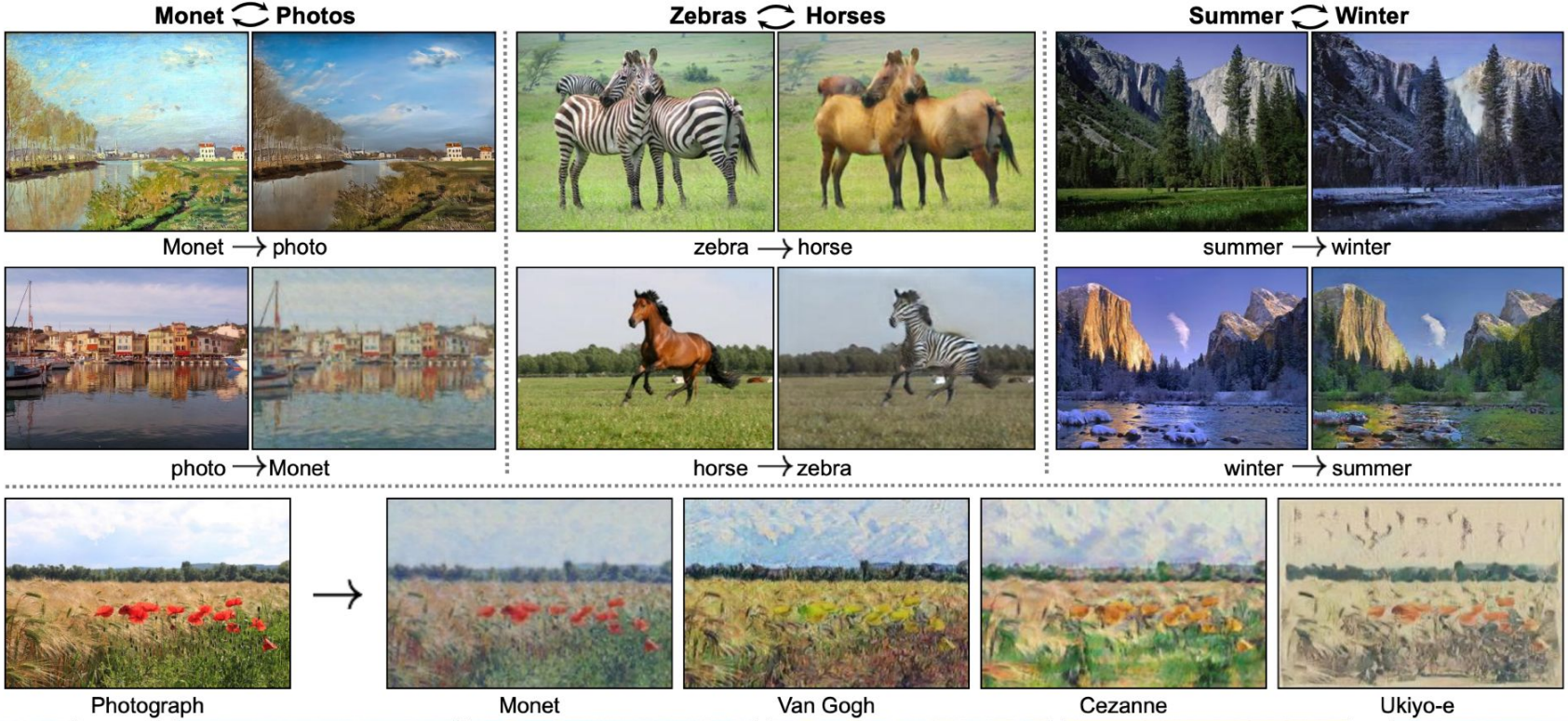
Unpaired

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



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

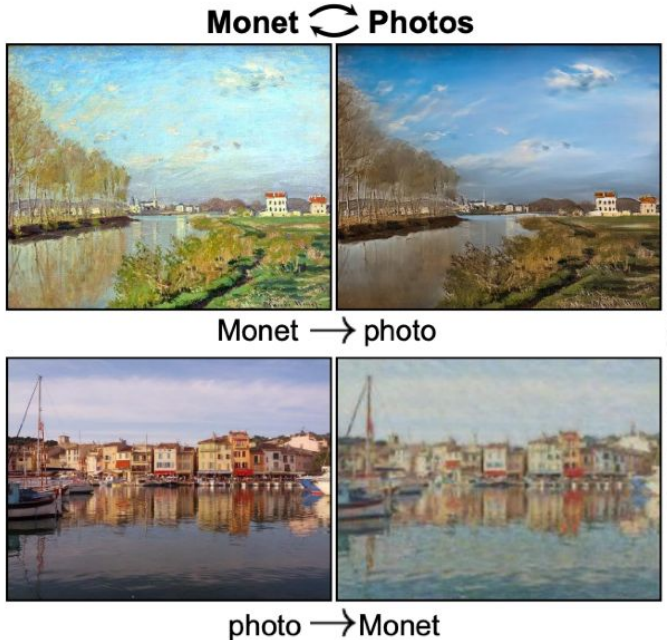
Unpaired Image-to-image Translation



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

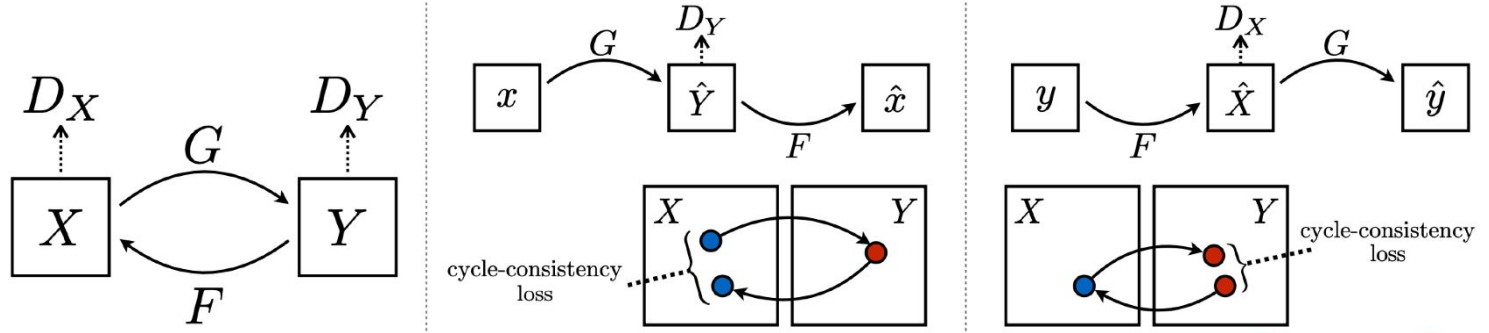
Unpaired Image-to-image Translation

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.

CycleGAN



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:

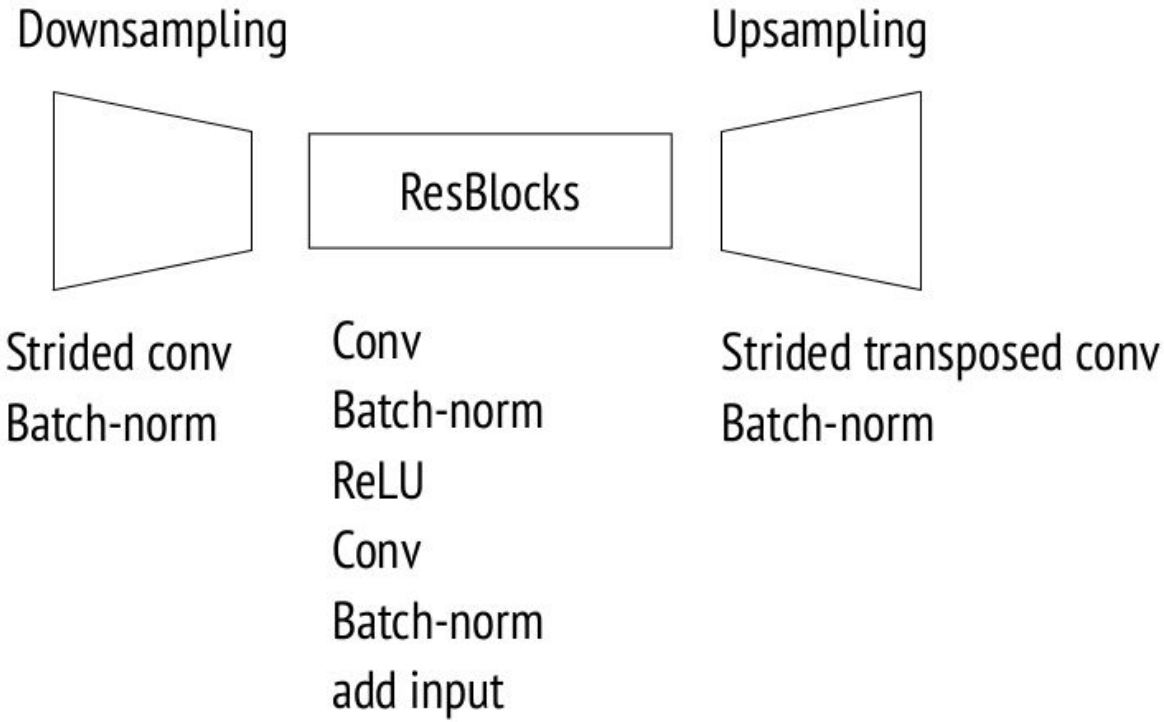
$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$$

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

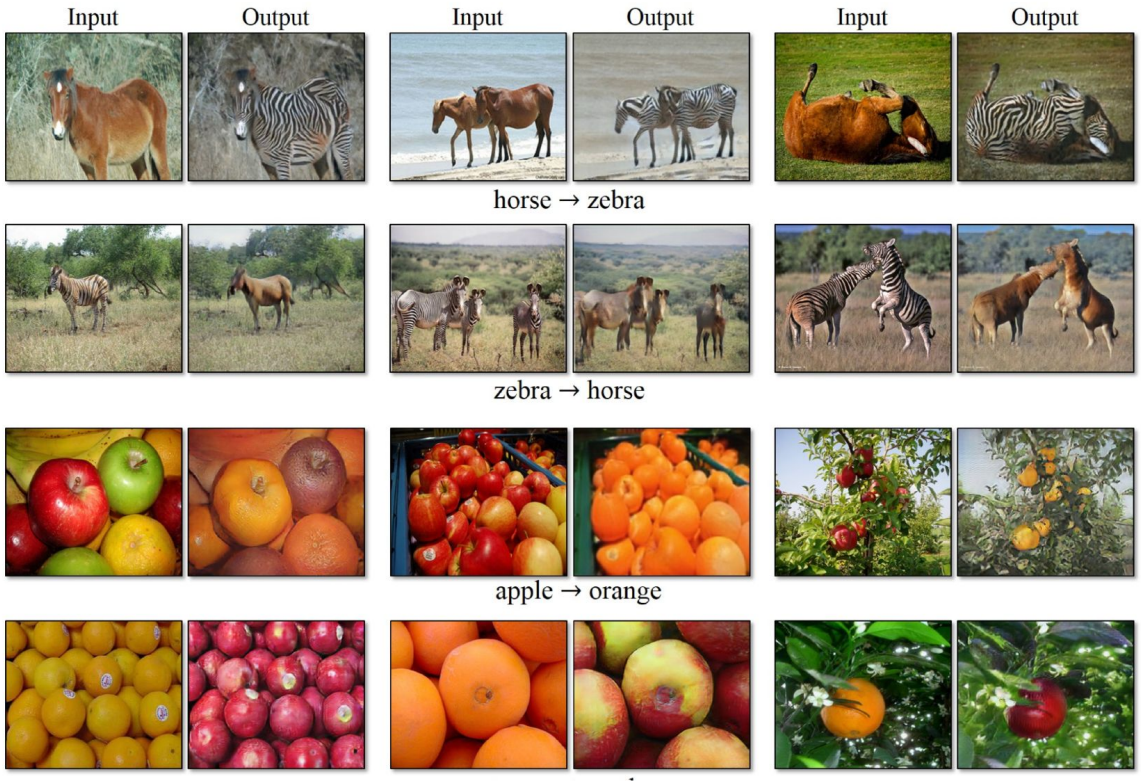


Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

CycleGAN

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

CycleGAN

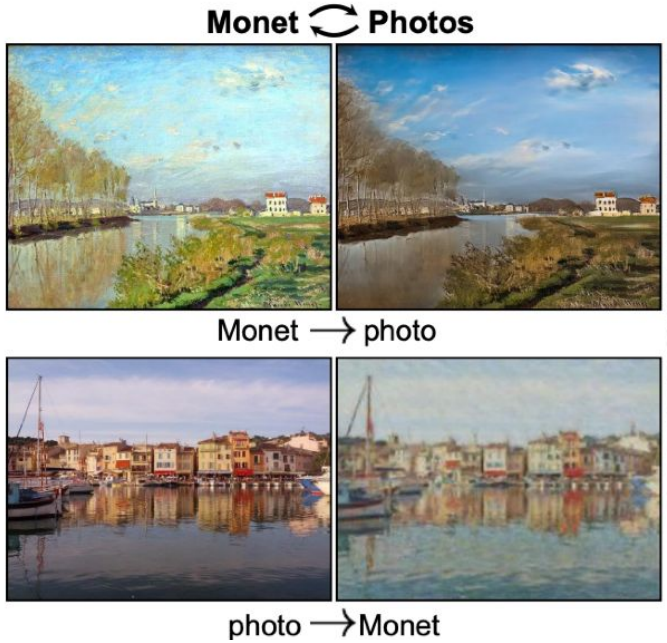


Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

Unpaired Image-to-image Translation

To solve it, constraints are necessary:

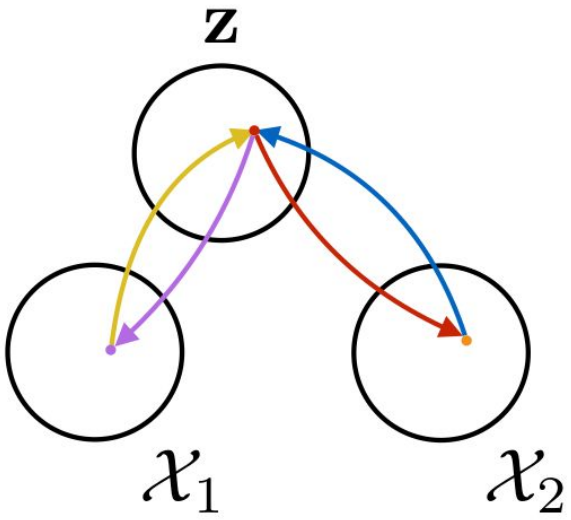
- Cycle-consistency constraint
- **Weight-sharing constraint**
- Geometry-consistency constraint

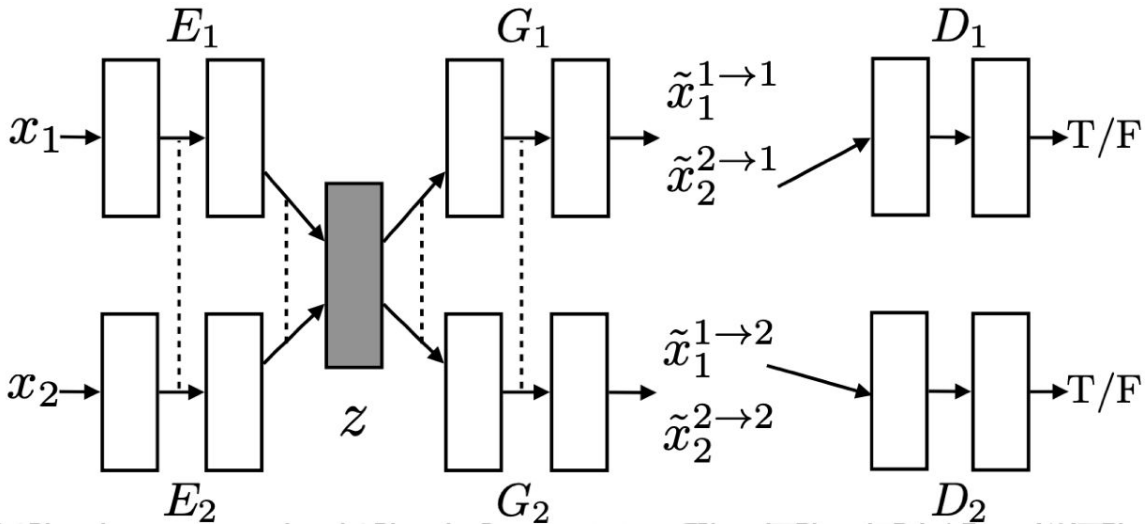


Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

Unpaired Translation with weight sharing

Train encoders with shared weights to go from image space to shared latent space

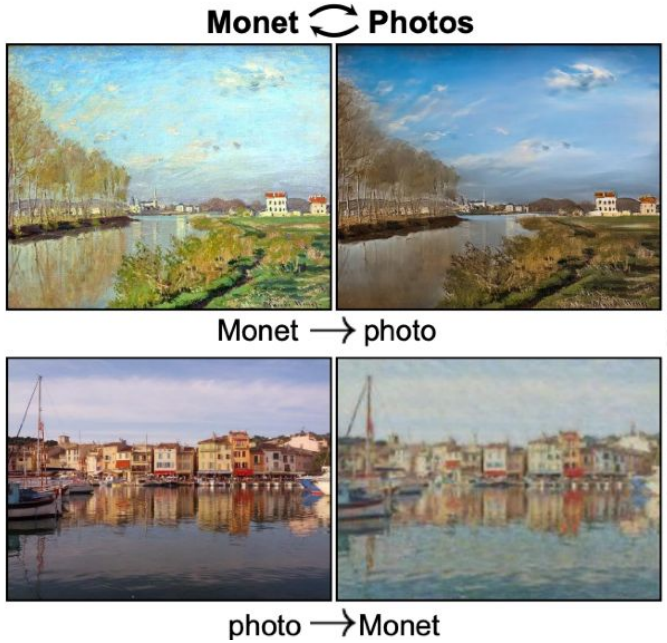




Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

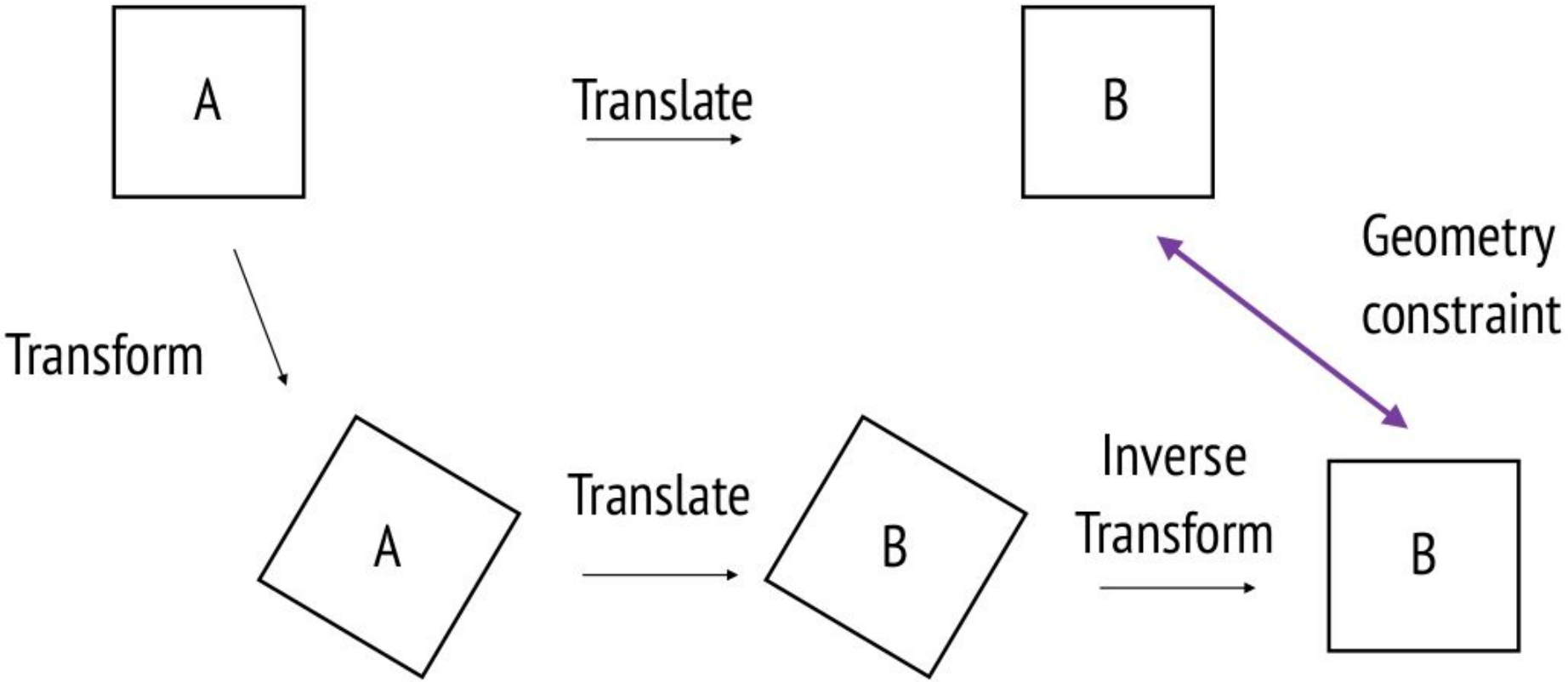
Unpaired Image-to-image Translation

To solve it, constraints are necessary:
- Cycle-consistency constraint
- Weight-sharing constraint
- **Geometry-consistency constraint**

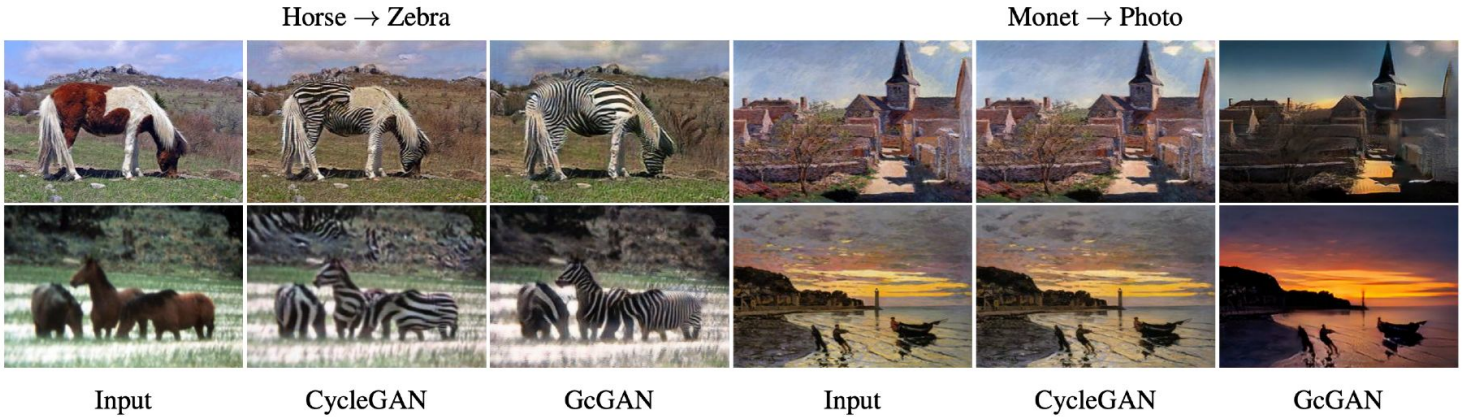


Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

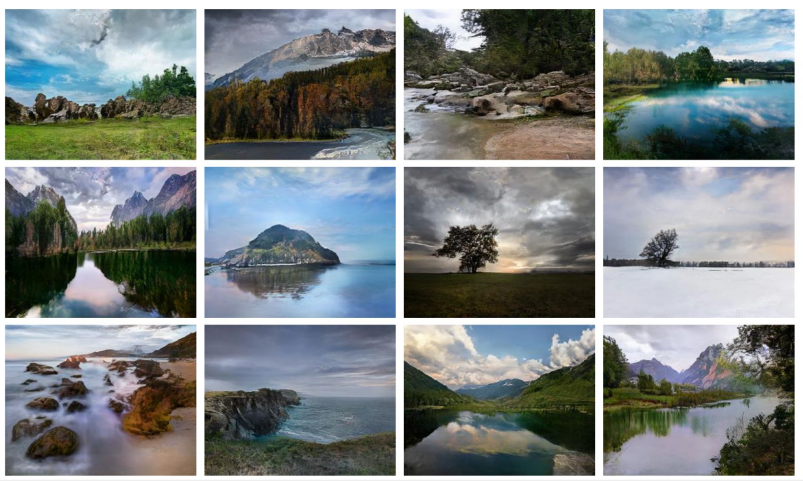
Geometry-consistency Constraint



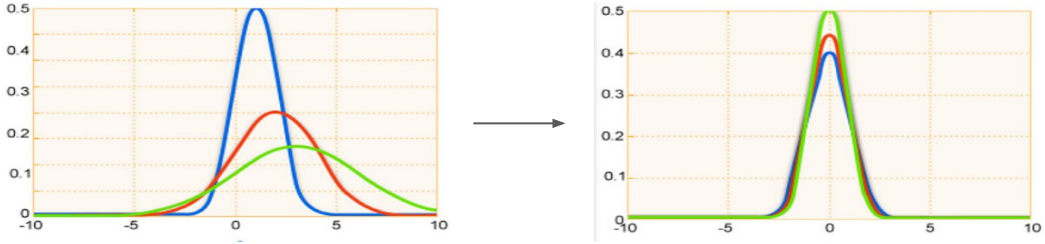
Geometry-consistency Constraint



Advanced methods



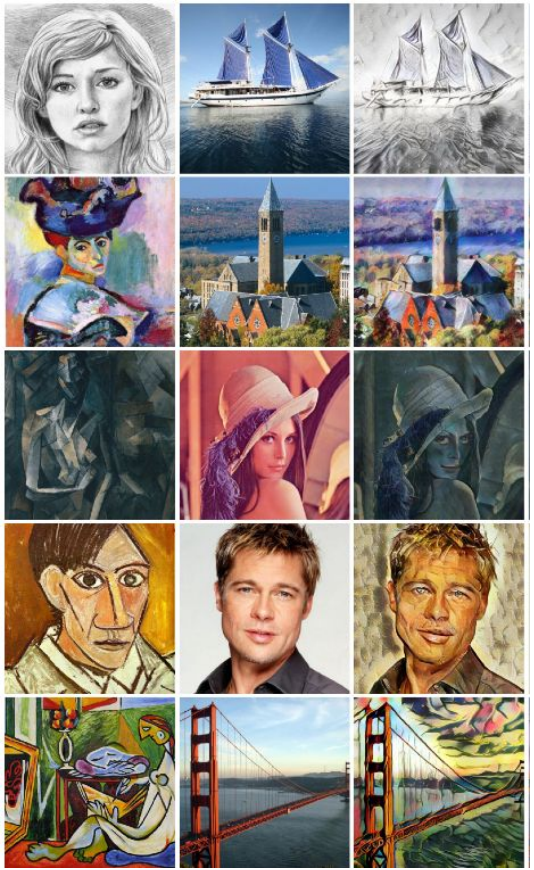
Style transfer



(a) Content image.

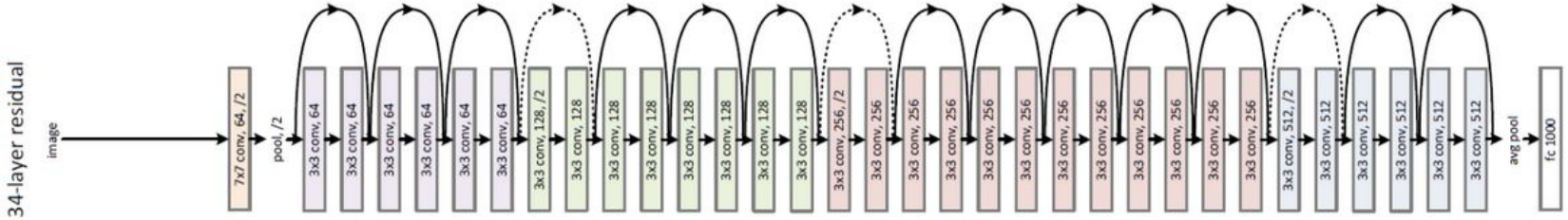
(c) Low contrast content image.

The two images are identical after instance normalization.
We obtain style invariant representations.



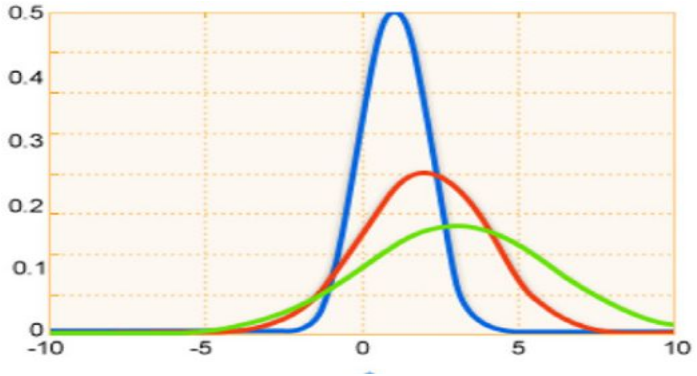
Instance Normalization: The Missing Ingredient for Fast Stylization,Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky

Layers: Batch-normalization

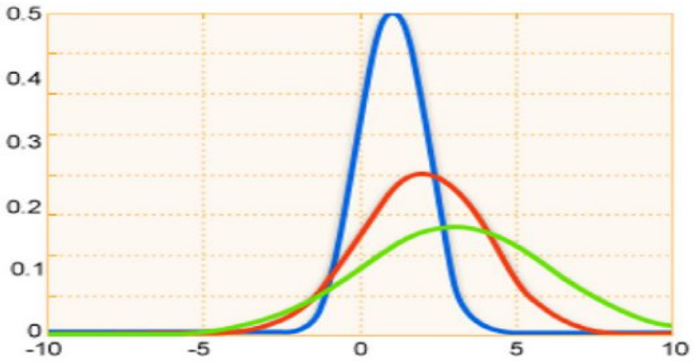


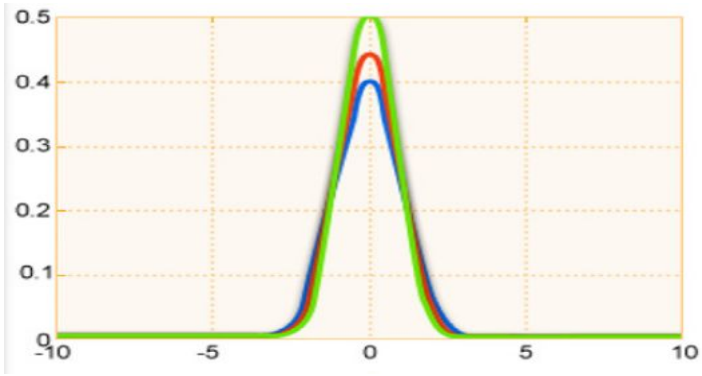
Parameter update:

$$\theta_{n+1} = \theta_n - \gamma \nabla L_b(\theta_n)$$



Layers: Batch-normalization





Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
Sergey Ioffe, Christian Szegedy

Layers: Batch-normalization

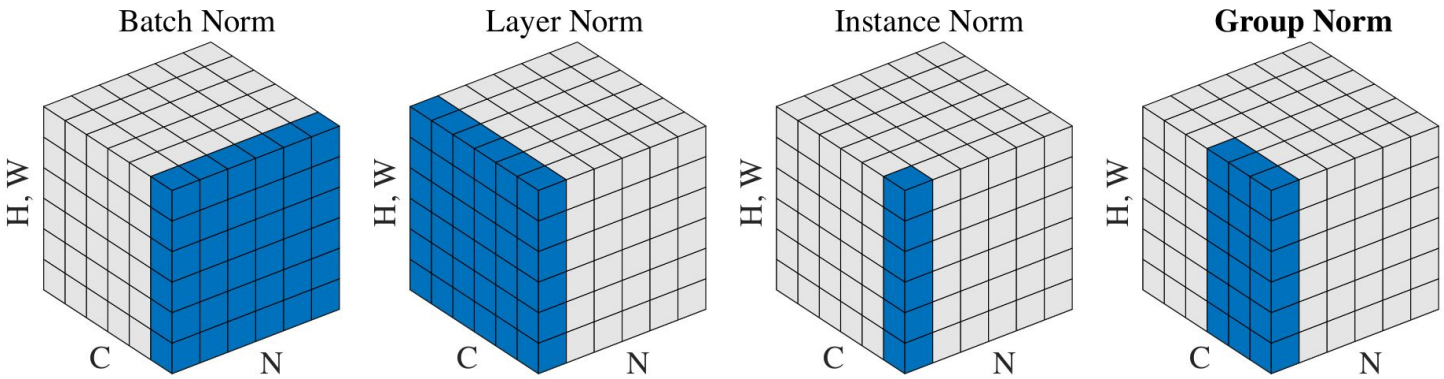
Advantages:

- Faster training
- Better optimization (lower final loss)
- Better generalization
- Specific usages in:
  - few-shot learning
  - domain adaptation
  - generation

Layers: Other normalizations

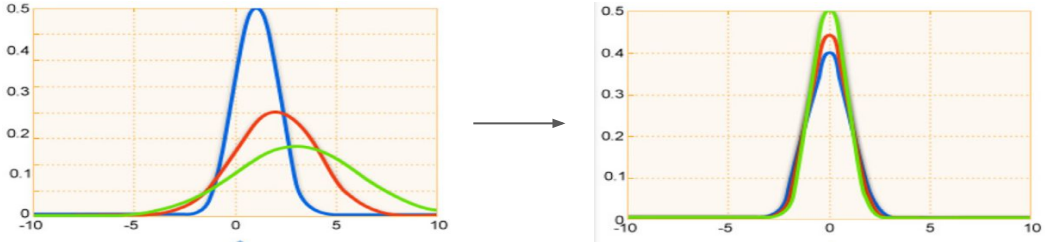
$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$



Group Normalization,Yuxin Wu, Kaiming He, ECCV 2018

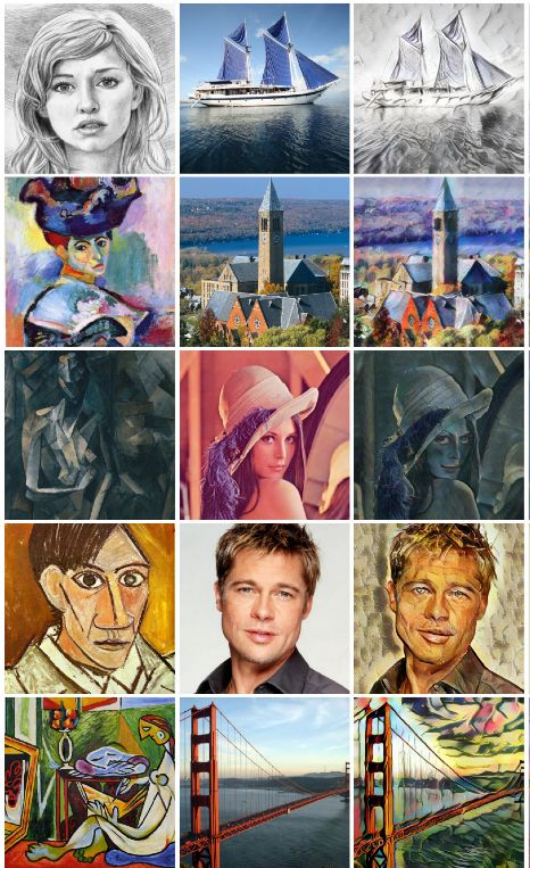
Style transfer



(a) Content image.

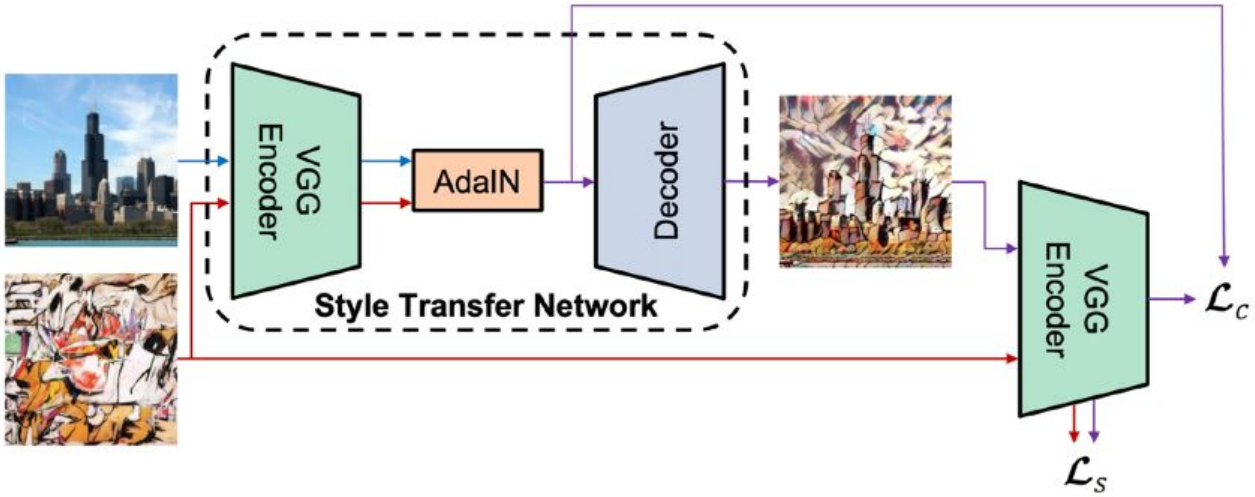
(c) Low contrast content image.

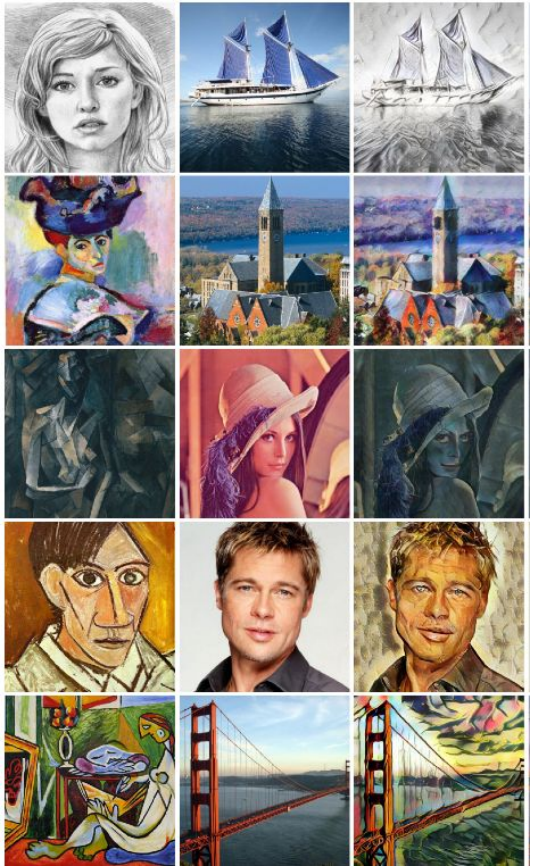
The two images are identical after instance normalization.
We obtain style invariant representations.



Instance Normalization: The Missing Ingredient for Fast Stylization,Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky

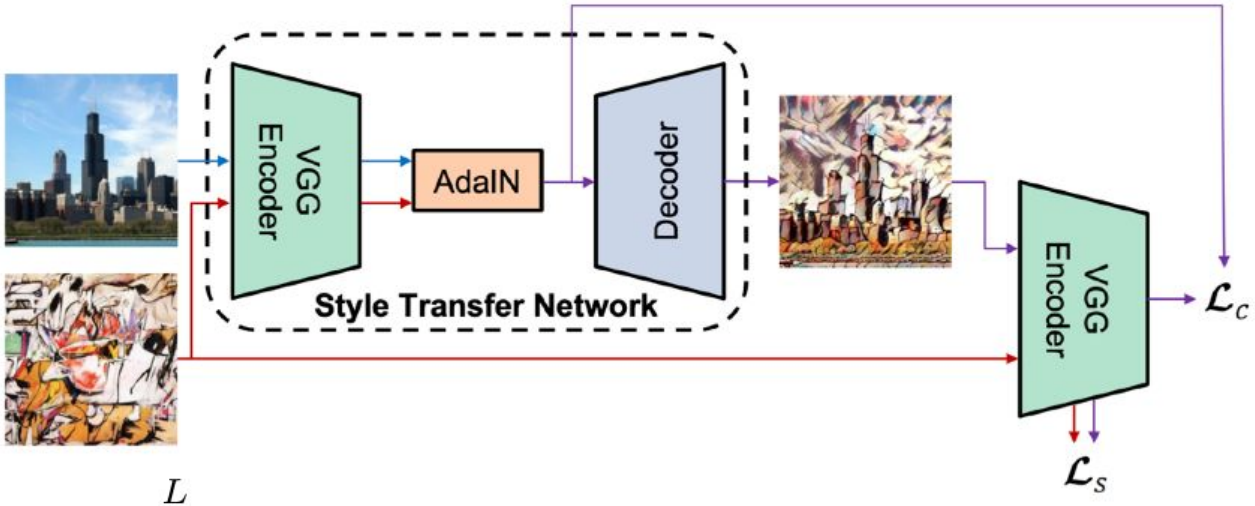
Style transfer





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

Style transfer



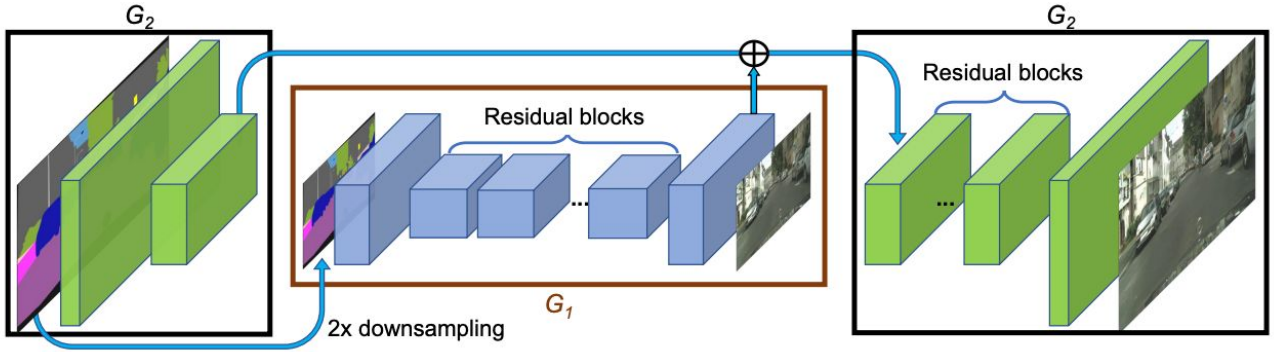
$$\mathcal{L}_c = \|f(g(t)) - t\|_2$$

$$\mathcal{L}_s = \sum_{i=1}^{L} \|\mu(\phi_i(g(t))) - \mu(\phi_i(s))\|_2 \quad + \quad \sum_{i=1}^{L} \|\sigma(\phi_i(g(t))) - \sigma(\phi_i(s))\|_2$$

where each φ_i denotes a layer in VGG-19

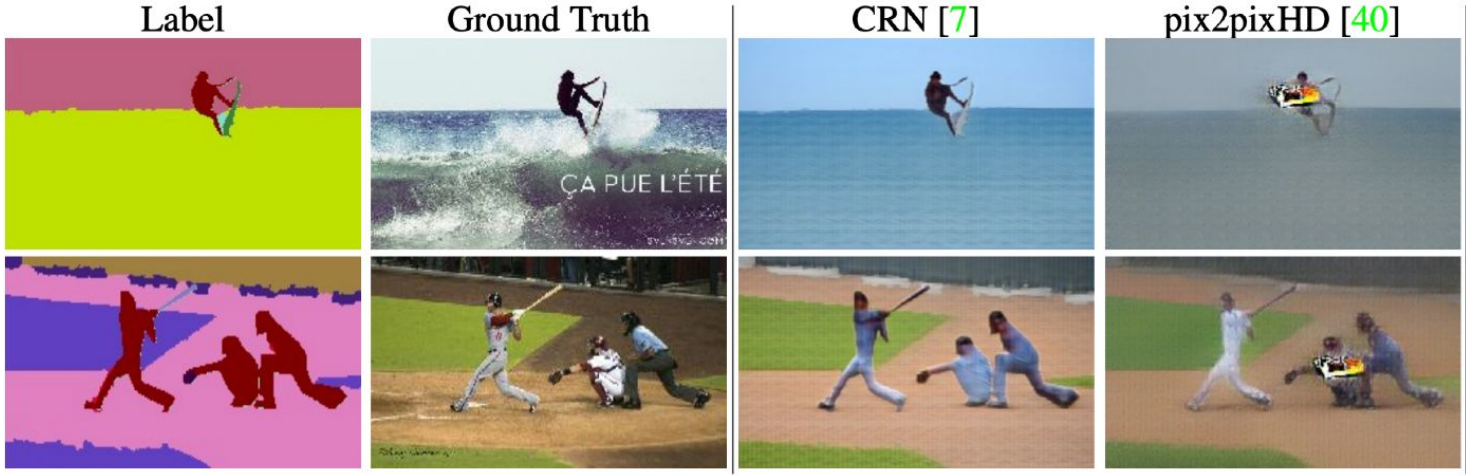
Pix2PixHD

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

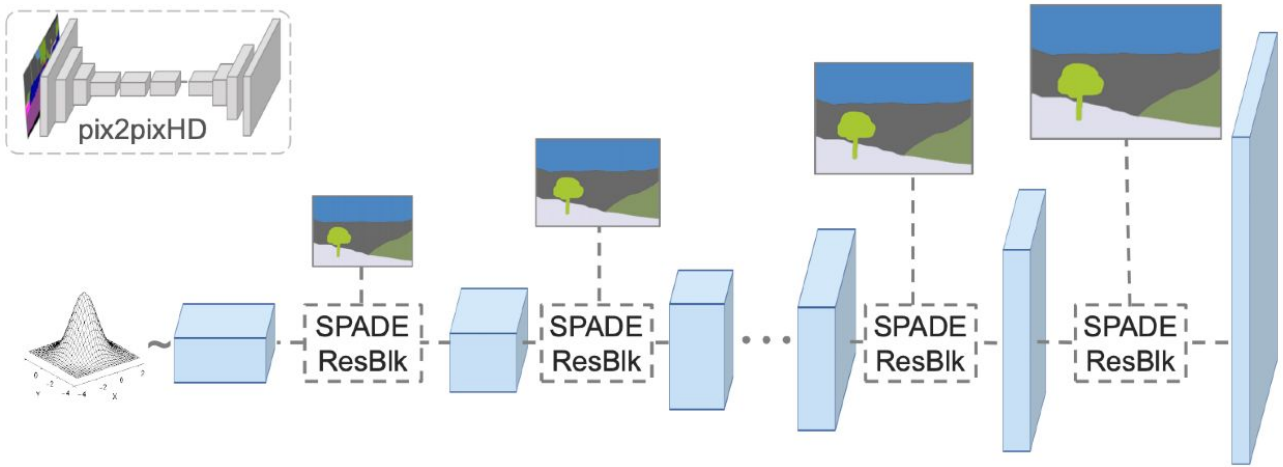


Input labels

Pix2PixHD

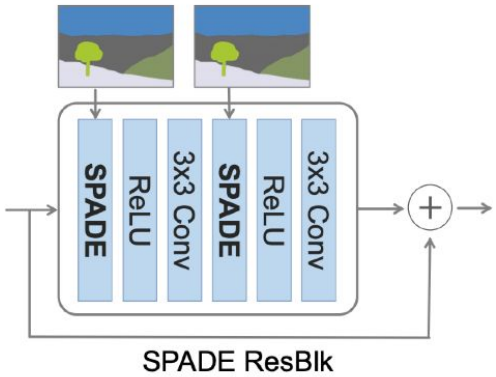


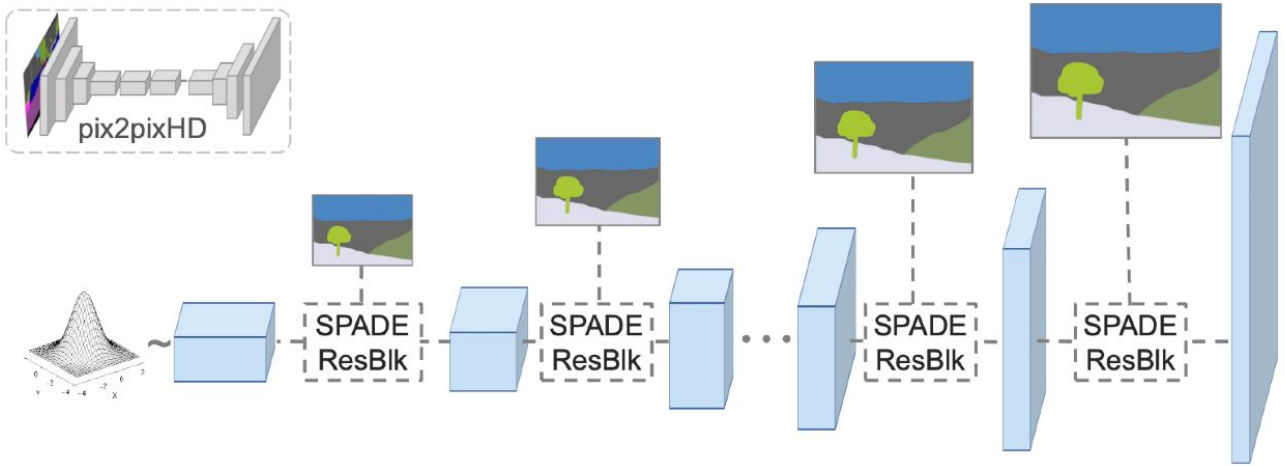
Spade

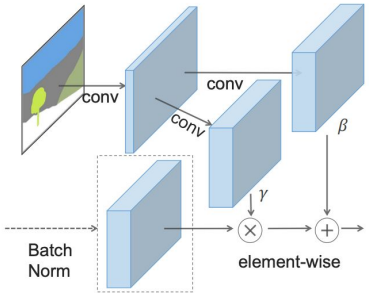


Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

Spade

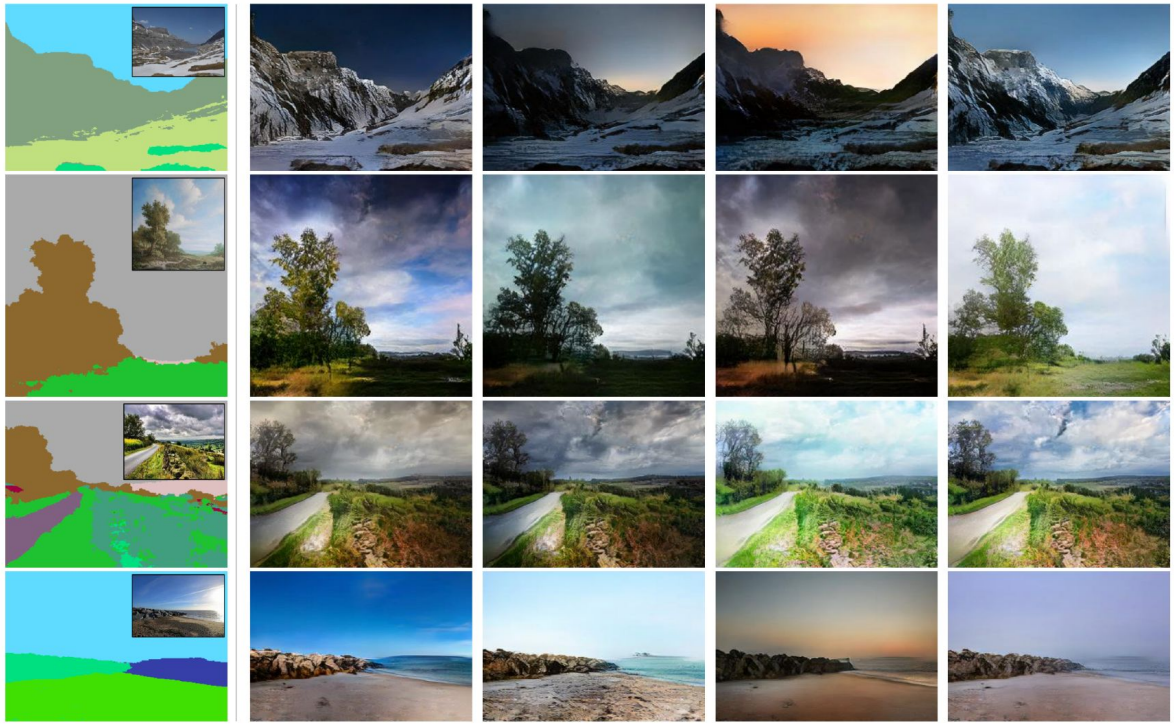






Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." CVPR'2019

Spade



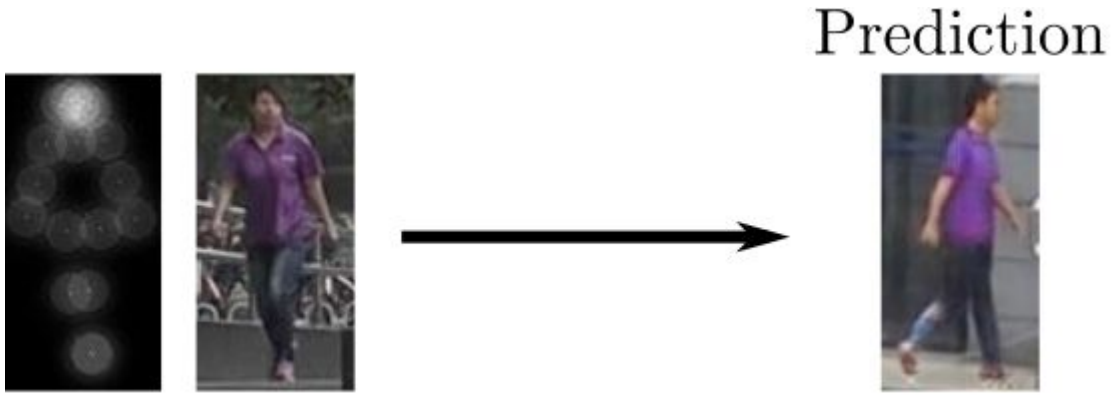
Live demo

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

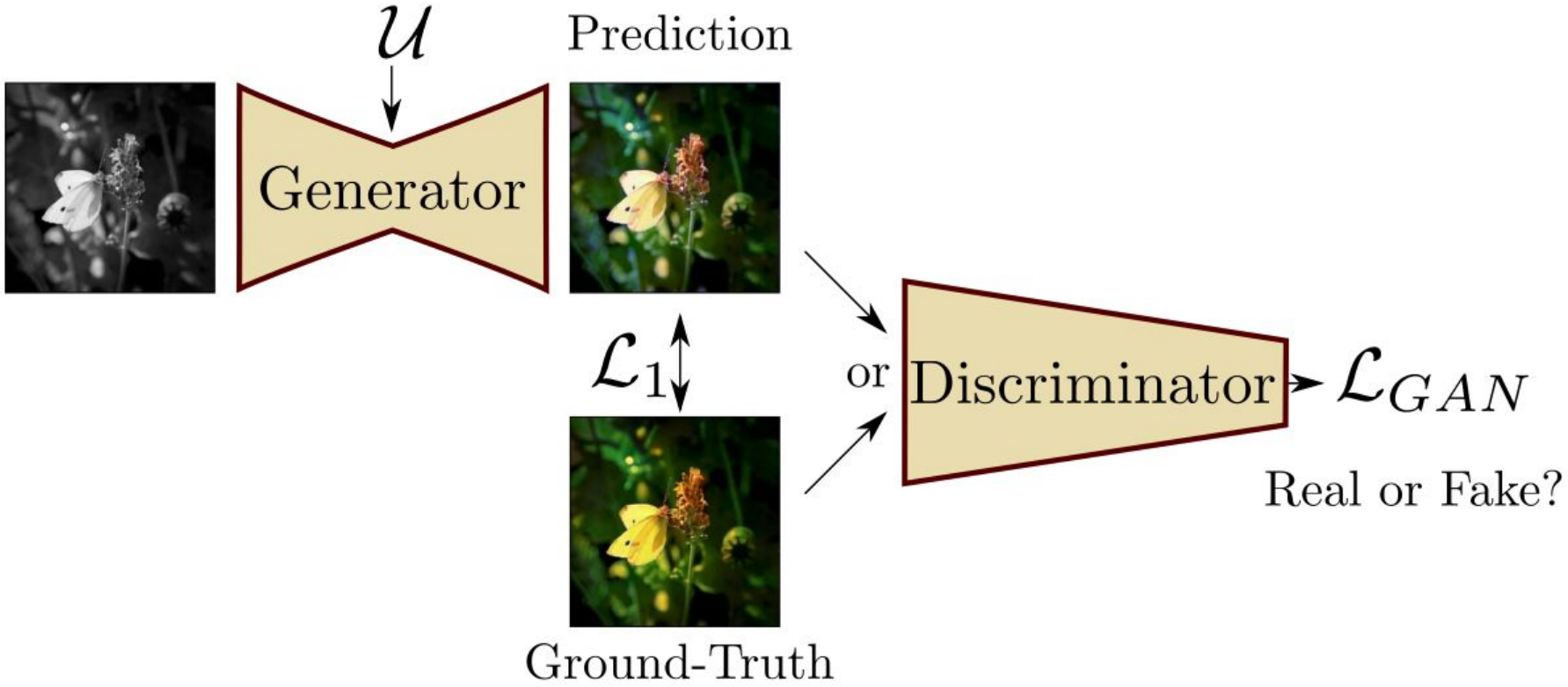
Pose-guided Image Generation

Stéphane Lathuilière

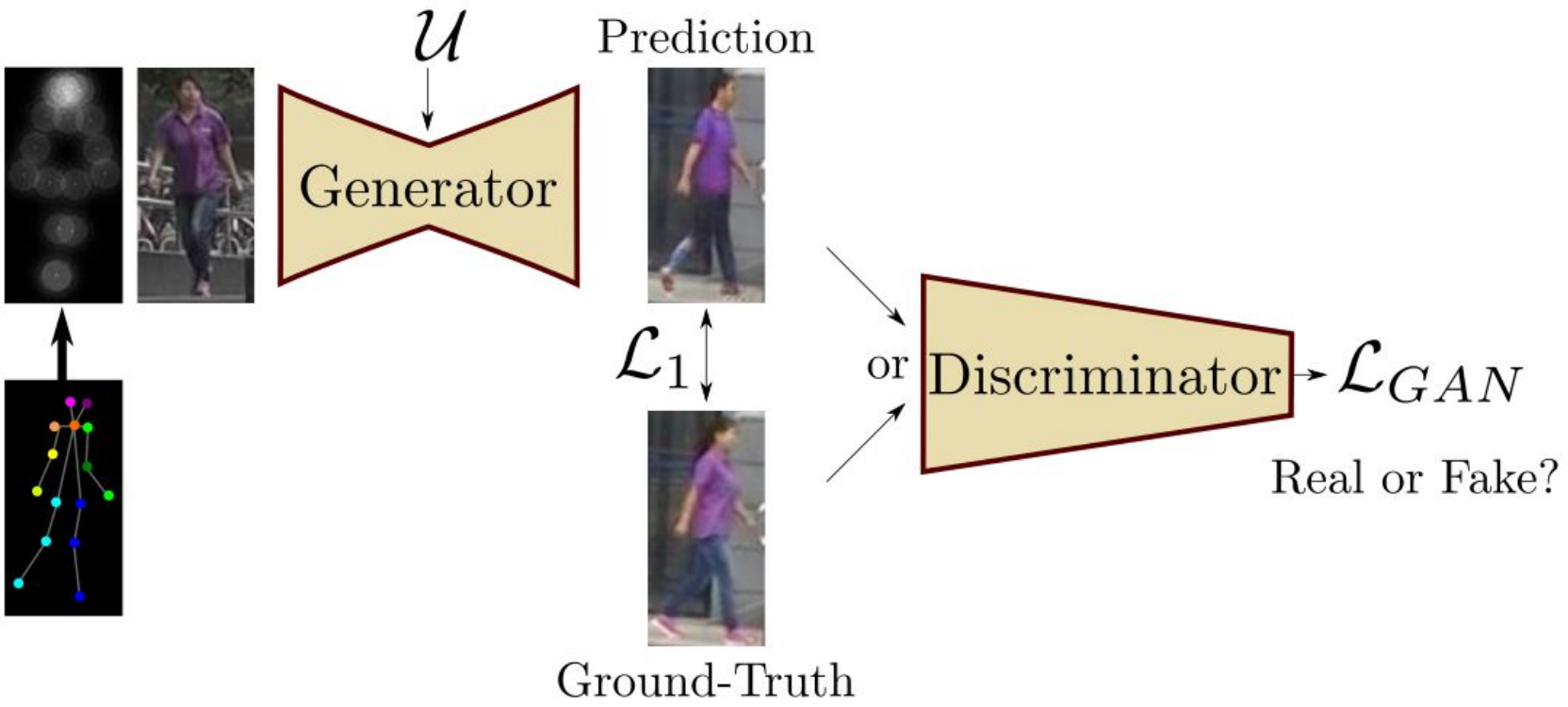
Introduction



Pix2Pix


$\mathcal{U}$
Prediction
Generator
$\mathcal{L}_1$
or
Discriminator
$\mathcal{L}_{GAN}$
Real or Fake?
Ground-Truth

Pix2Pix for pose guided



Pix2Pix for pose guided

Pose guided

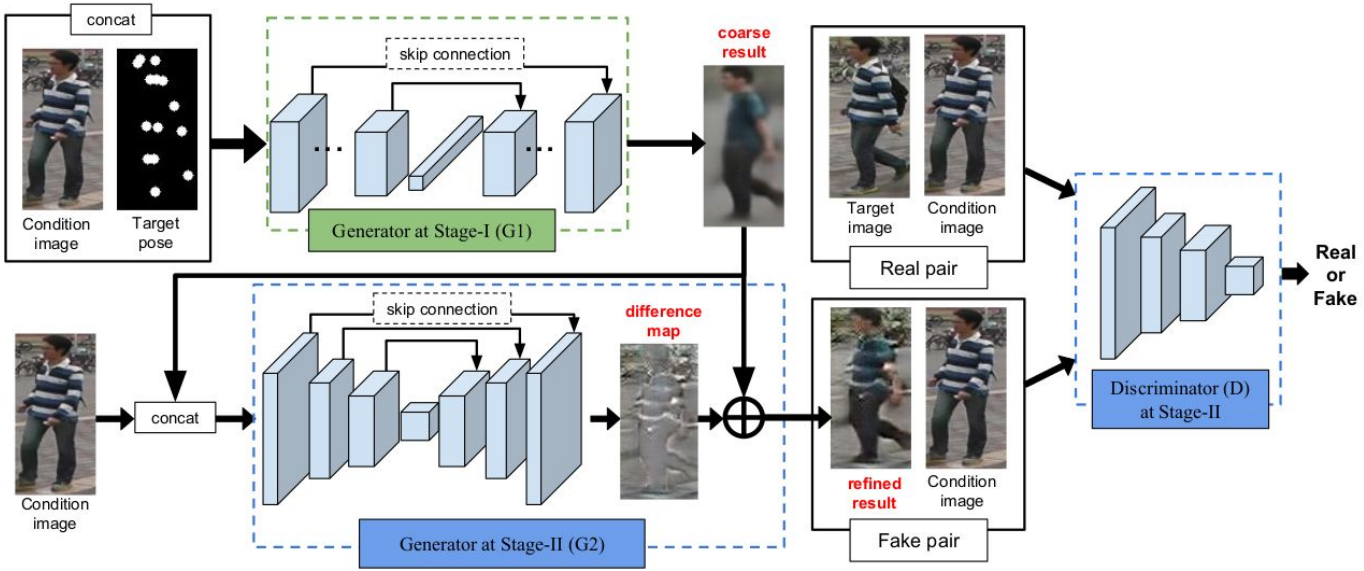
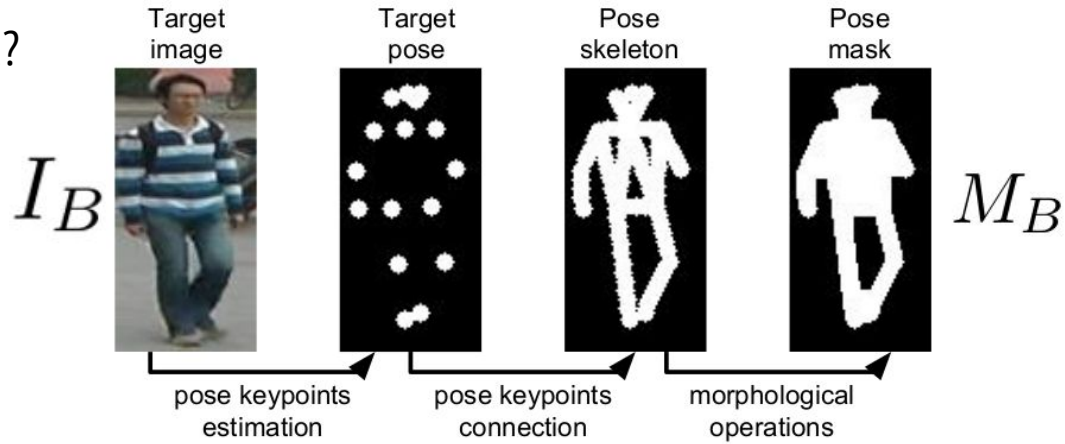


Figure 2: The overall framework of our Pose Guided Person Generation Network (PG^2). It contains two stages. Stage-I focuses on pose integration and generates an initial result that captures the global structure of the human. Stage-II focuses on refining the initial result via adversarial training and generates sharper images.

Pose Guided Person Image Generation, Ma et. al. Nips 2017

Pose guided

What about the background?

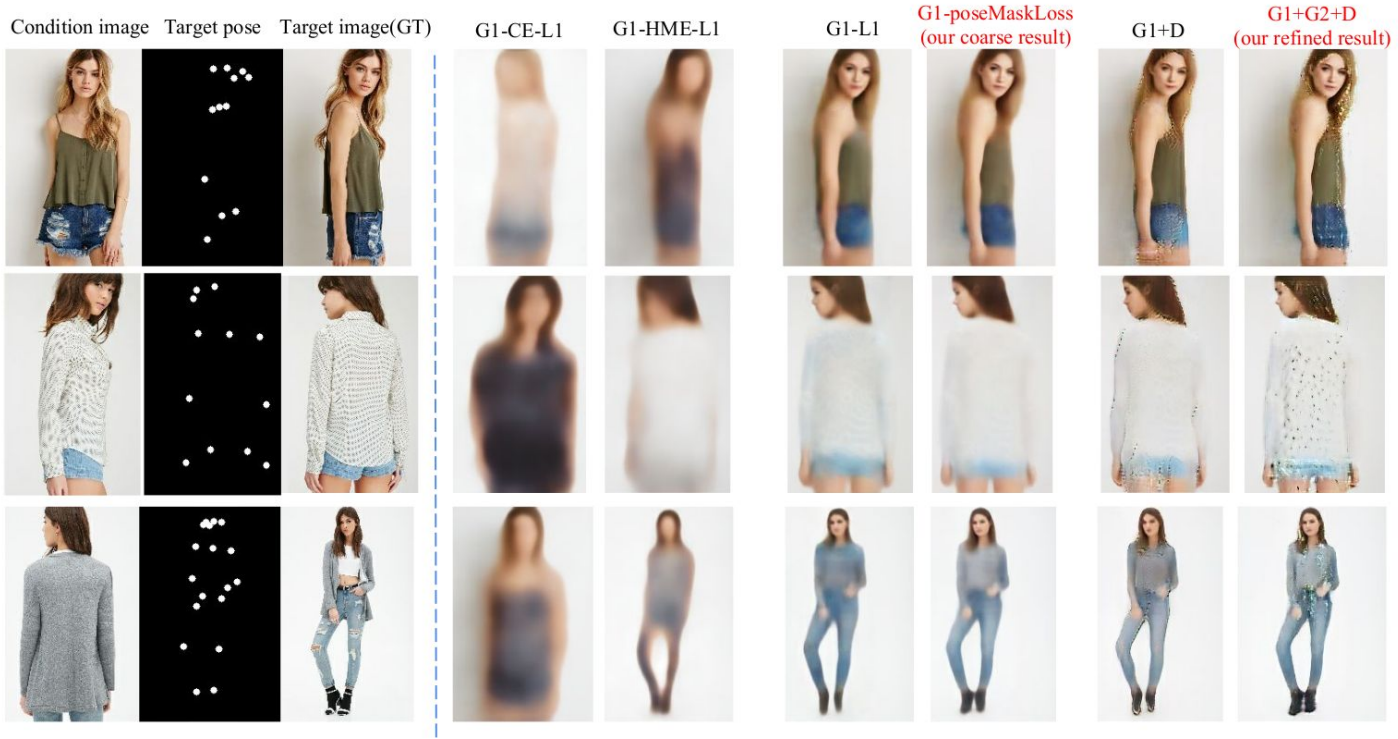


They give more weight to the foreground:

$$\mathcal{L}_{\mathrm{G1}} = \|(\mathrm{G1}(I_A, P_B) - I_B) \odot (1 + M_B)\|_1$$

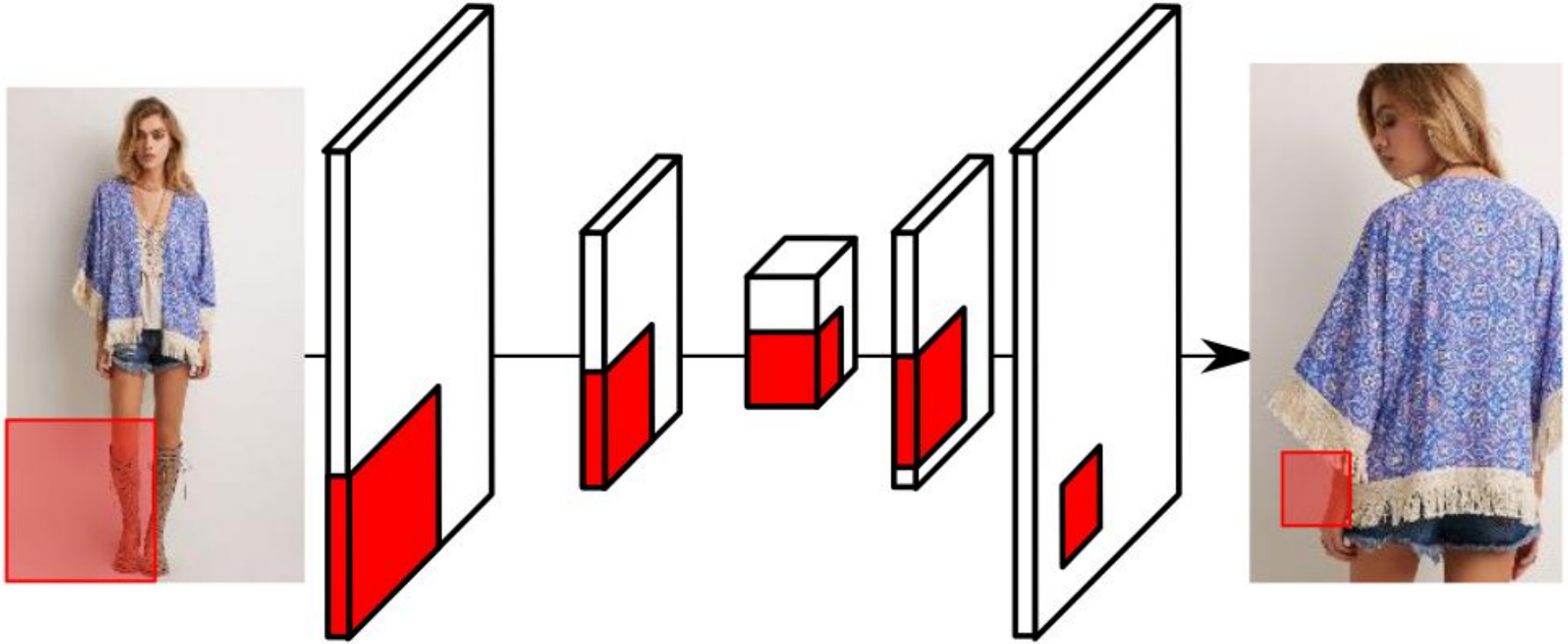
Pose Guided Person Image Generation, Ma et. al. Nips 2017

Pose guided



Pose Guided Person Image Generation, Ma et. al. Nips 2017

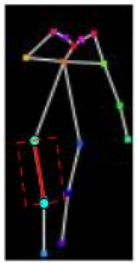
Pix2Pix for pose guided: problem



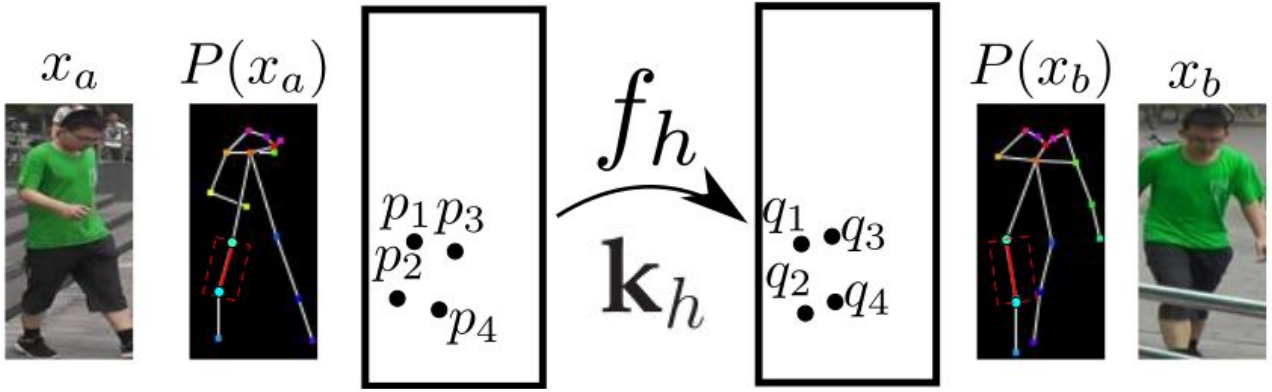
Pose guided: deformable GAN

x_a $P(x_a)$

$P(x_b)$ x_b

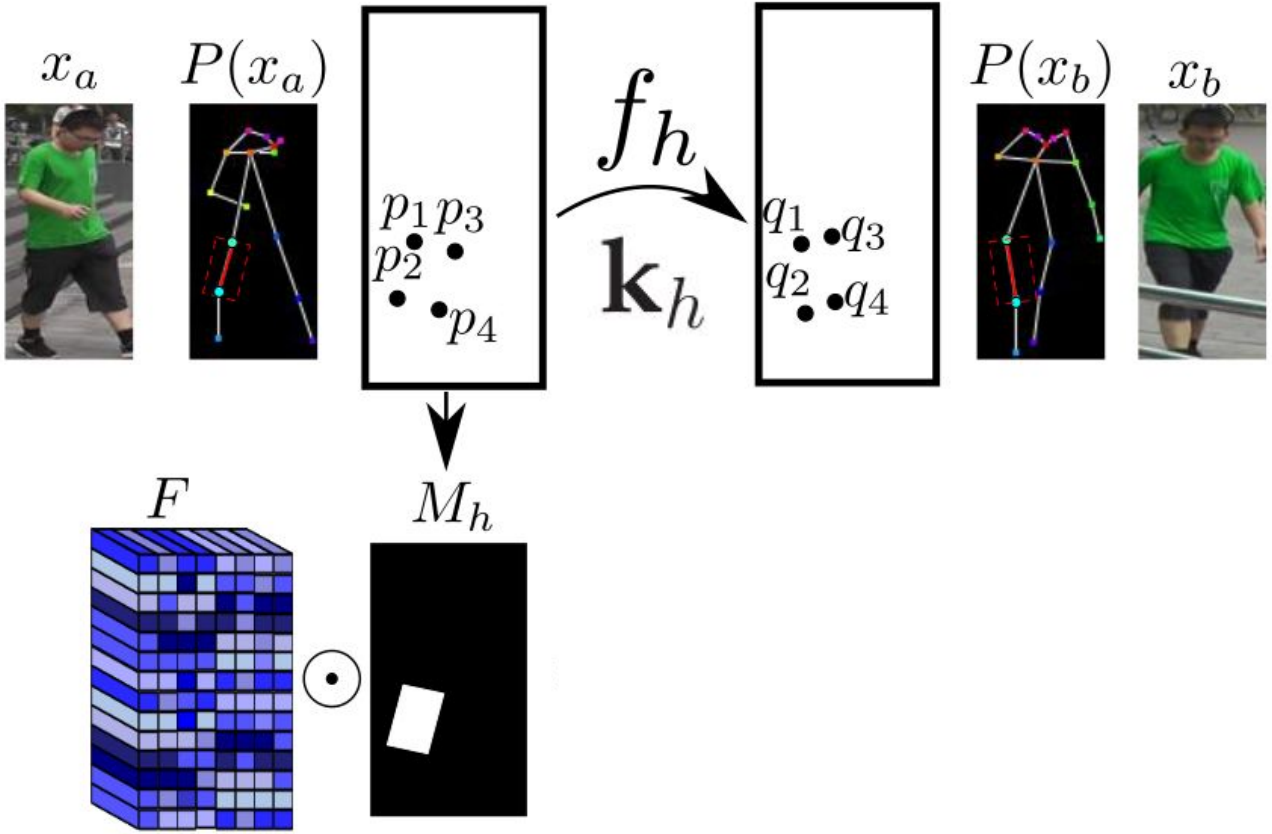


Pose guided: deformable GAN

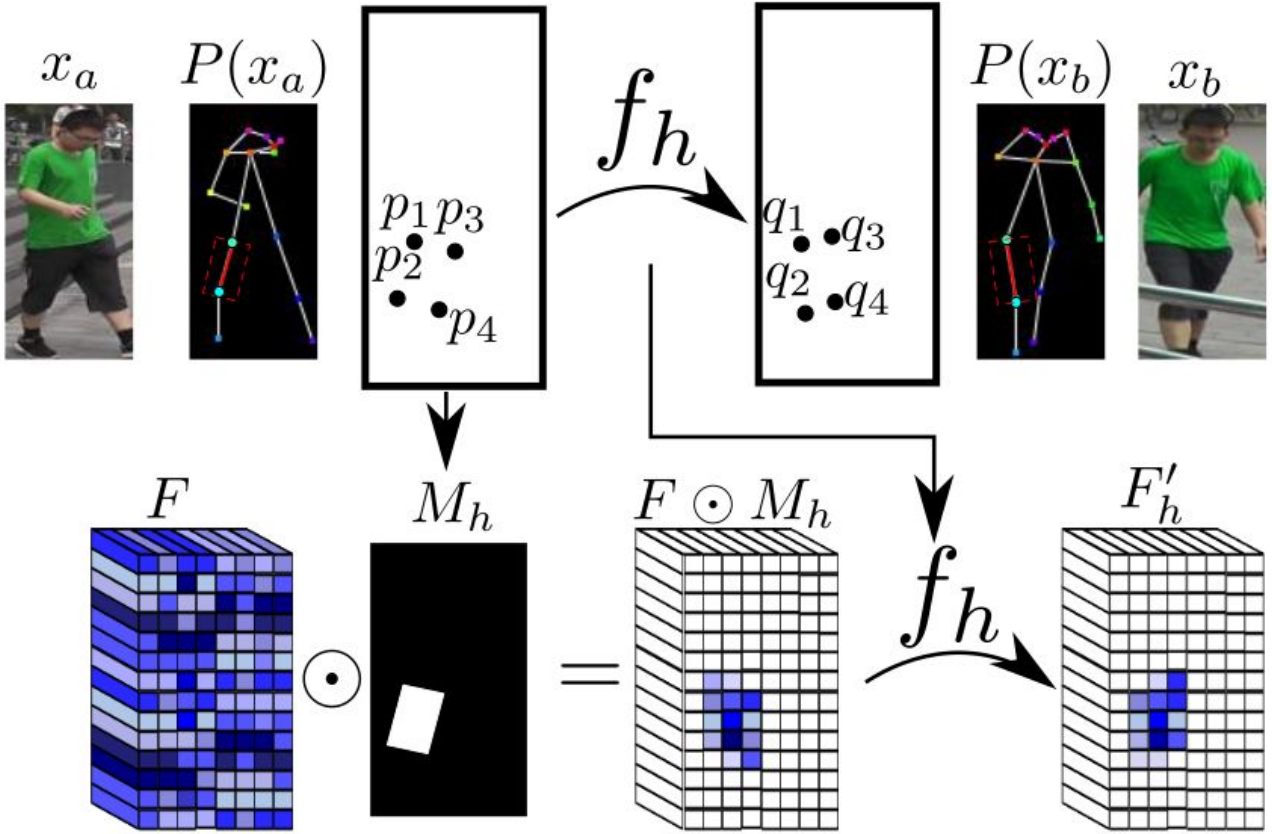


$$\min_{\mathbf{k}_h} \sum_{\mathbf{p}_j \in R_h^a, \mathbf{q}_j \in R_h^b} ||\mathbf{q}_j - f_h(\mathbf{p}_j; \mathbf{k}_h)||_2^2.$$

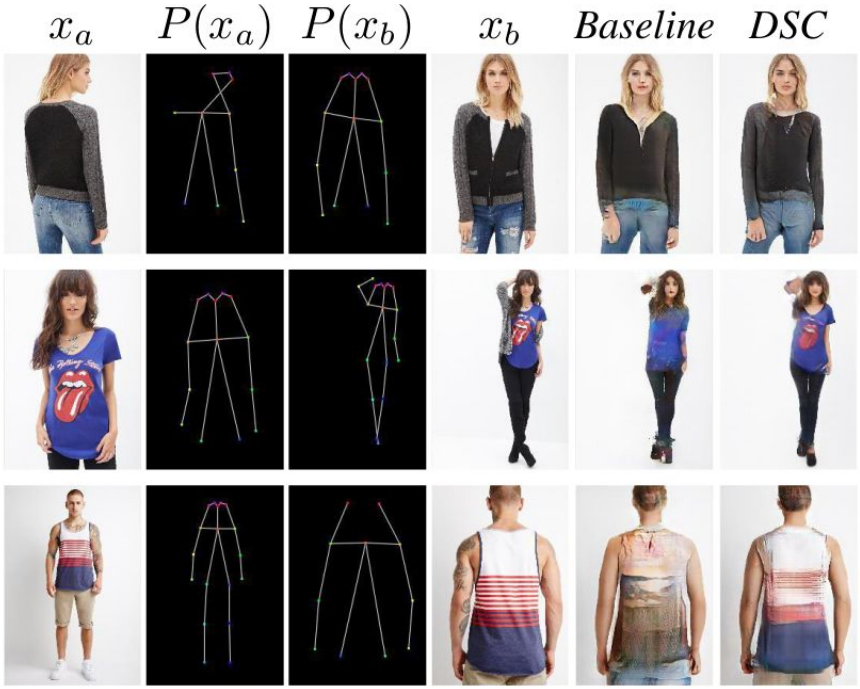
Pose guided: deformable GAN



Pose guided: deformable GAN



Pose guided: deformable GAN

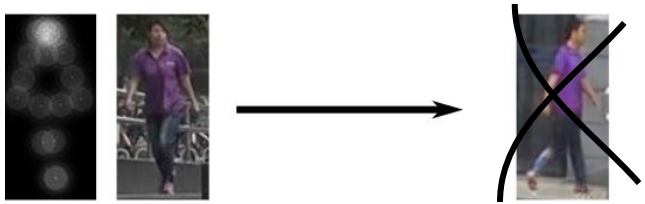


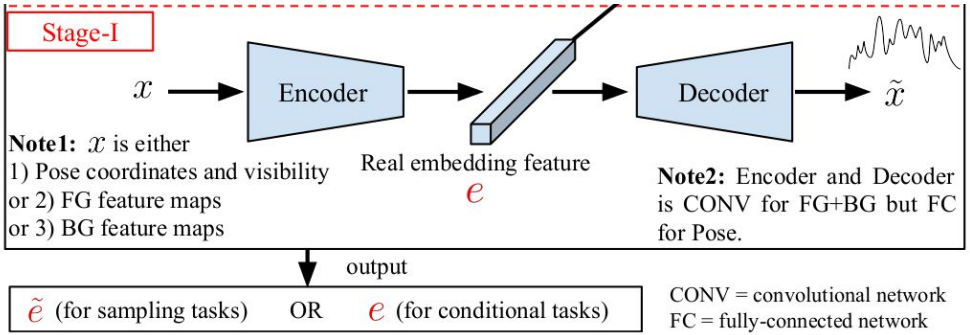
Pose guided: reconstruction vs GAN

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

	Market-1501				
λ	*SSIM*	*IS*	*mask-SSIM*	*mask-IS*	*DS*
0.1	**0.292**	2.621	**0.808**	3.168	0.697
0.01	0.290	3.185	0.805	3.502	**0.720**
0.001	0.245	**3.566**	0.779	**3.634**	0.609

Pose guided: Disentangled Generation





Disentangled Person Image Generation, Ma et. al., CVPR 2018

Pose guided: Disentangled Generation

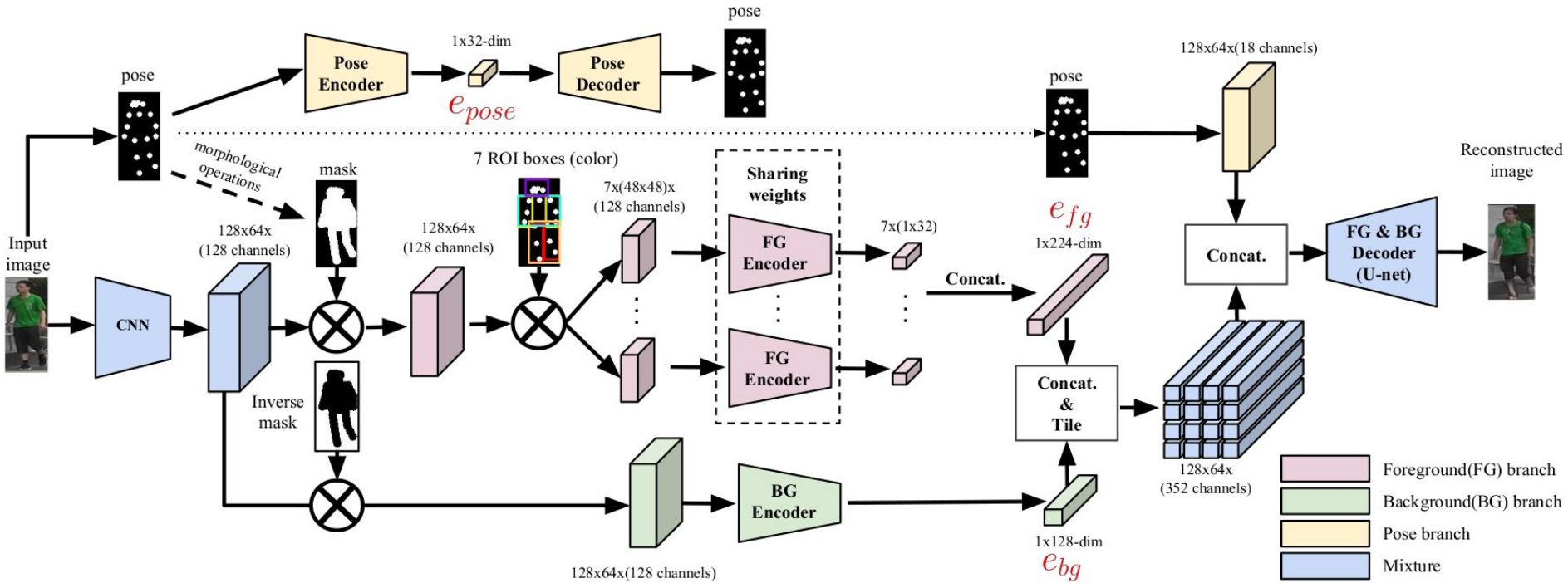


Figure 3: Stage-I: disentangled image reconstruction. This framework is composed of three branches: foreground, background and pose. Note that we use a fully-connected auto-encoder network to reconstruct the pose (incl. keypoint coordinates and visibility), so that we can decode the embedded pose features to obtain the heatmaps at the sampling phase.

Pose guided: Disentangled Generation

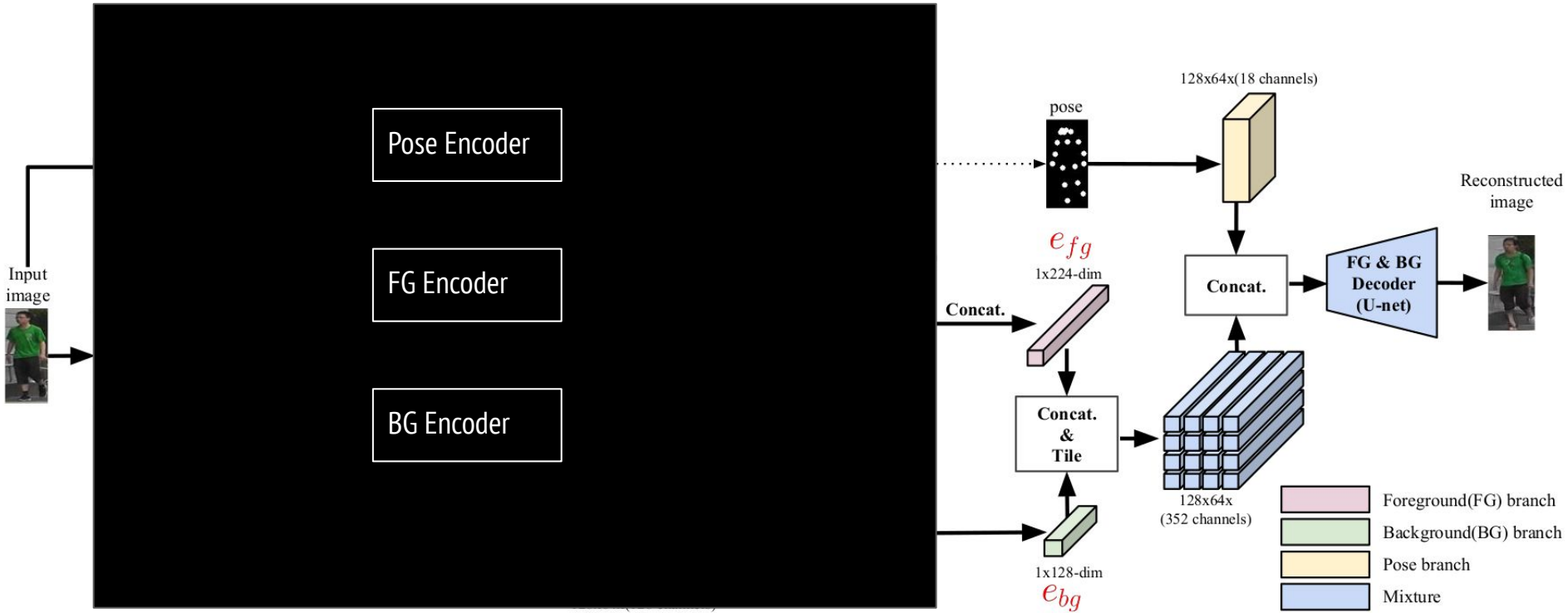
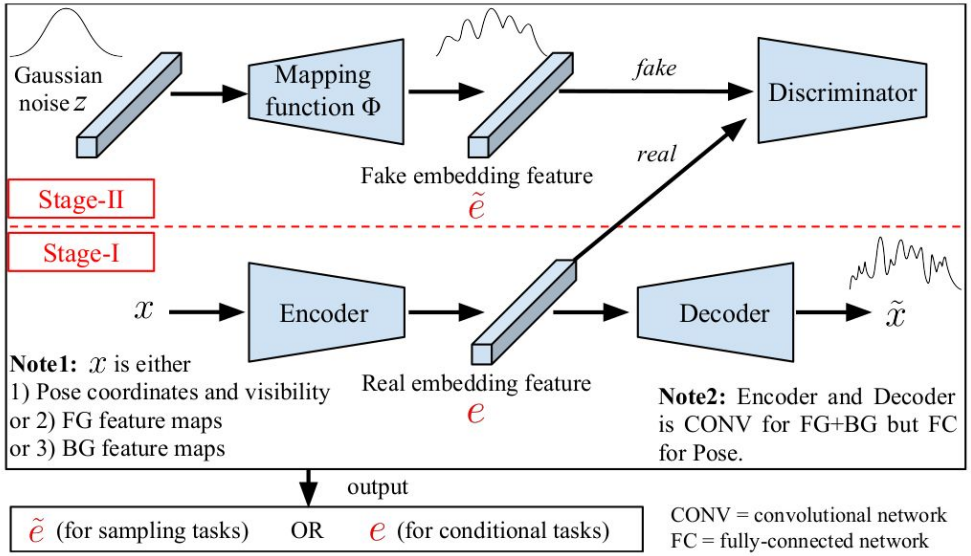


Figure 3: Stage-I: disentangled image reconstruction. This framework is composed of three branches: foreground, background and pose. Note that we use a fully-connected auto-encoder network to reconstruct the pose (incl. keypoint coordinates and visibility), so that we can decode the embedded pose features to obtain the heatmaps at the sampling phase.

Pose guided: Disentangled Generation



Disentangled Person Image Generation, Ma et. al., CVPR 2018

Pose guided: Disentangled Generation

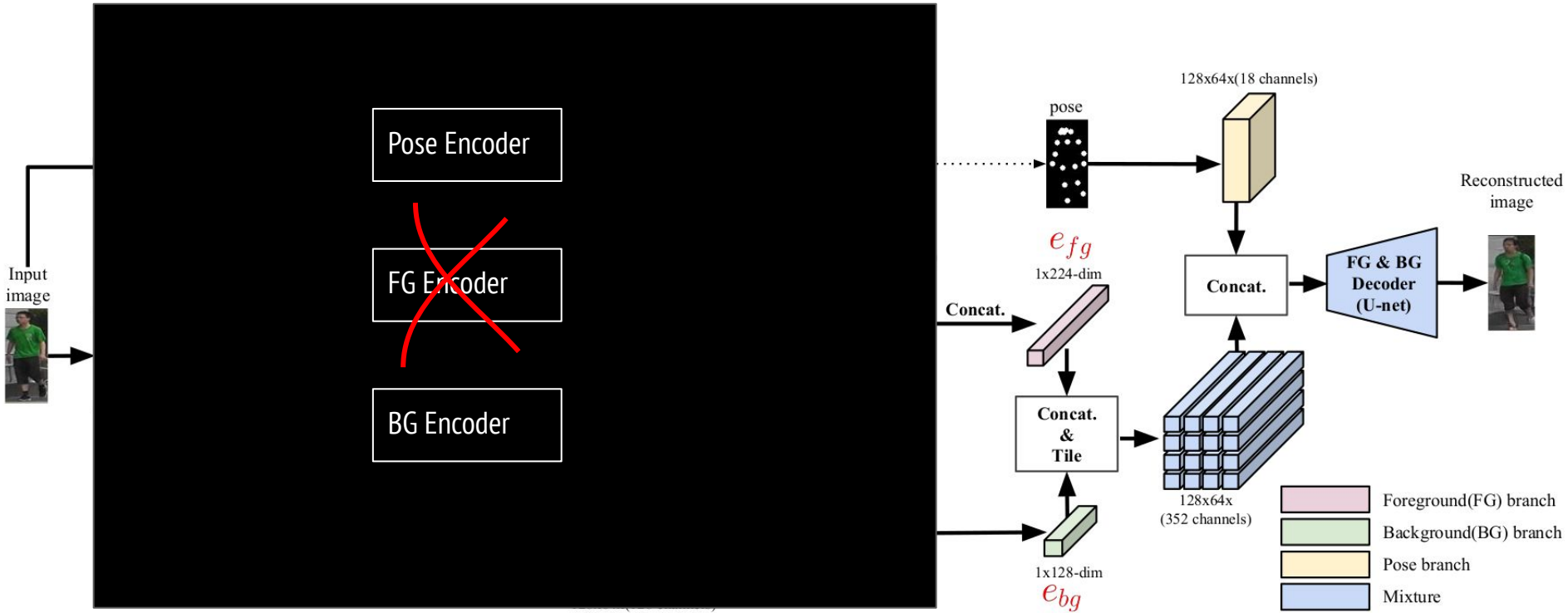


Figure 3: Stage-I: disentangled image reconstruction. This framework is composed of three branches: foreground, background and pose. Note that we use a fully-connected auto-encoder network to reconstruct the pose (incl. keypoint coordinates and visibility), so that we can decode the embedded pose features to obtain the heatmaps at the sampling phase.

Pose guided: Disentangled Generation

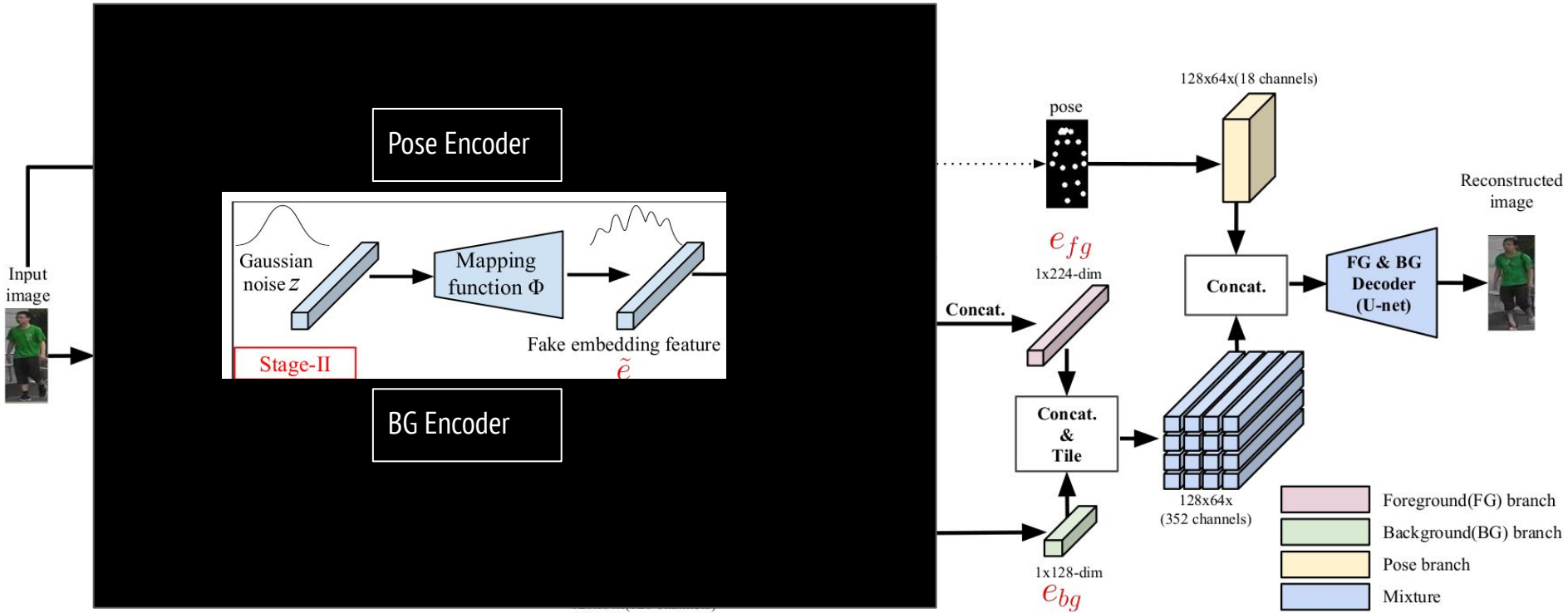


Figure 3: Stage-I: disentangled image reconstruction. This framework is composed of three branches: foreground, background and pose. Note that we use a fully-connected auto-encoder network to reconstruct the pose (incl. keypoint coordinates and visibility), so that we can decode the embedded pose features to obtain the heatmaps at the sampling phase.

Pose guided: Disentangled Generation

Foreground (FG) sampling (fixed BG and Pose)

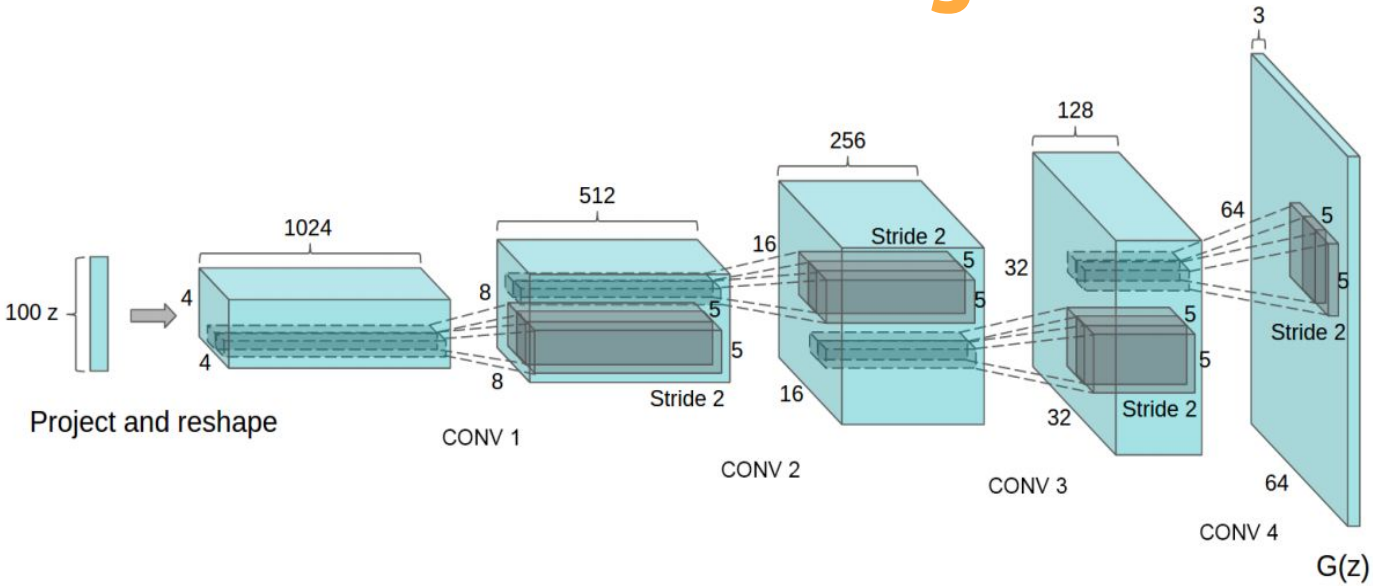
Background (BG) sampling (fixed FG and Pose)

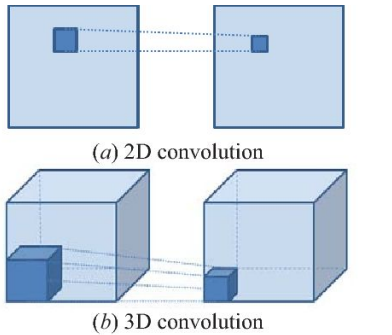
Pose sampling (fixed FG and BG)

FG, BG and Pose sampling

Video generation

From DCGAN to video generation

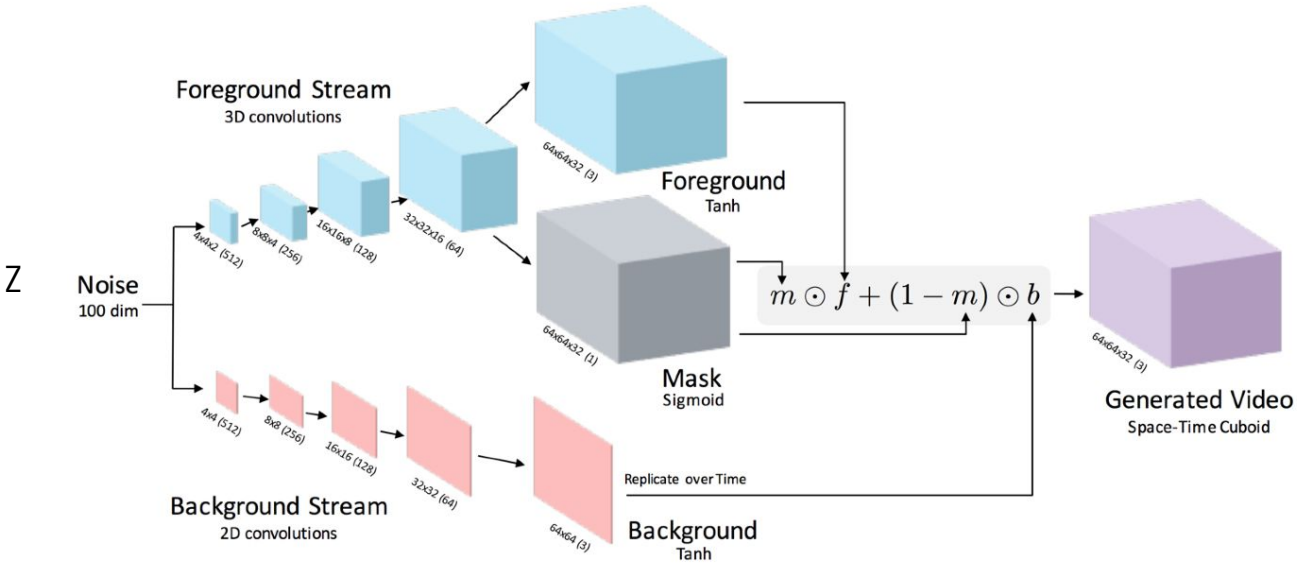




3. Comparison of (a) 2D convolution and (b)

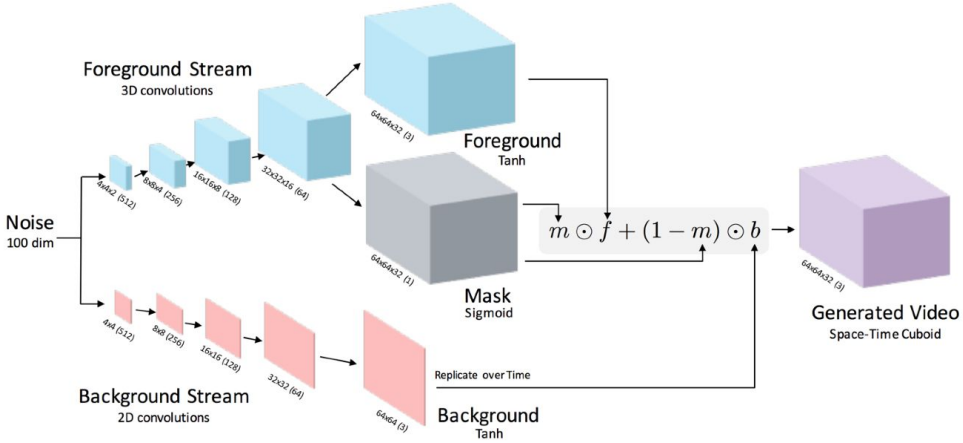
A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, ICLR 2016
Figure from Lung nodule detection based on 3D convolutional neural networks Lei Fan et al.

Video Generation



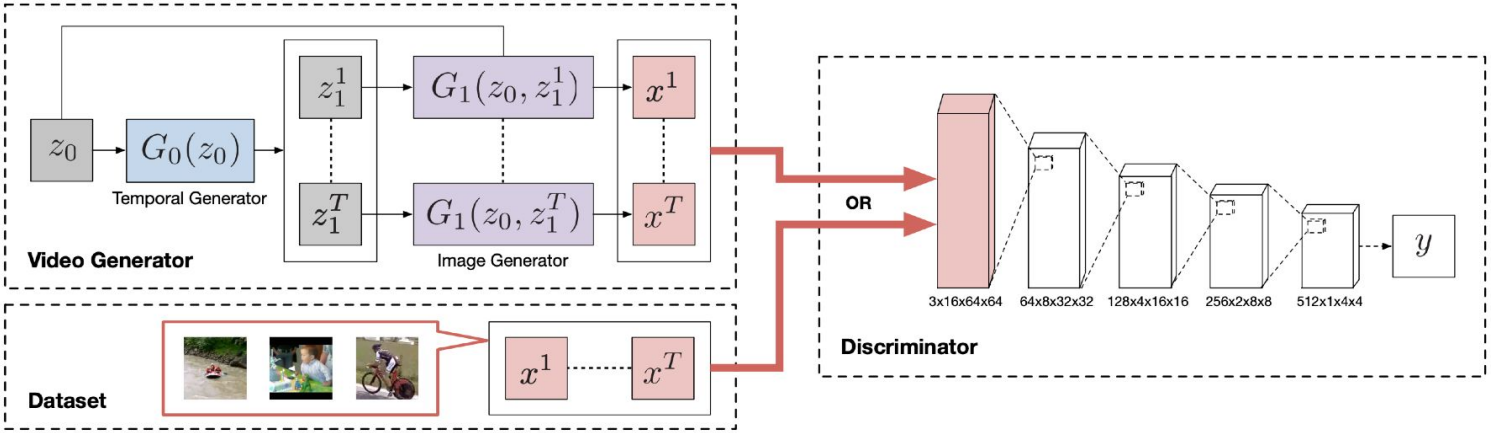
Saito, Matsumoto, and Saito. "Temporal generative adversarial nets with singular value clipping." ICCV'2017.

Video Generation



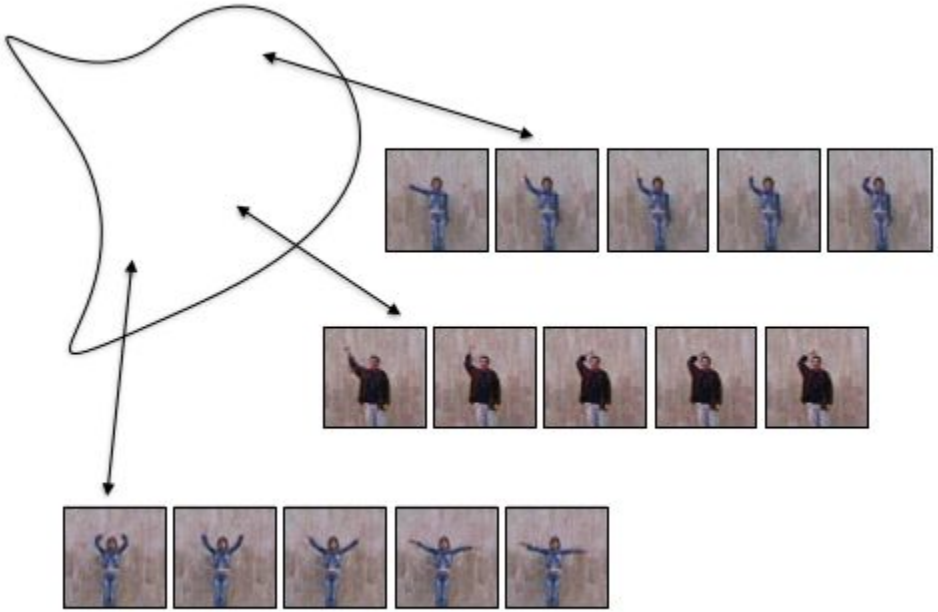
Saito, Matsumoto, and Saito. "Temporal generative adversarial nets with singular value clipping." ICCV'2017.

Video Generation



Saito, Matsumoto, and Saito. "Temporal generative adversarial nets with singular value clipping." ICCV'2017.

Video Generation



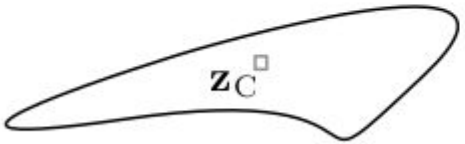
Limitations:

- Fixed length videos only
- No control over motion and content

Video Generation: MoCoGAN

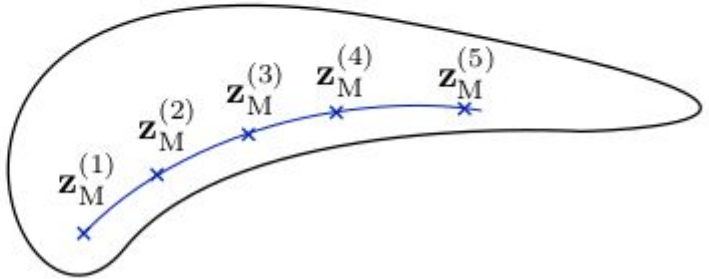
Sampled content

$$\mathbf{Z}_\mathrm{C} = [\mathbf{z}_\mathrm{C}, \mathbf{z}_\mathrm{C}, ..., \mathbf{z}_\mathrm{C}]$$



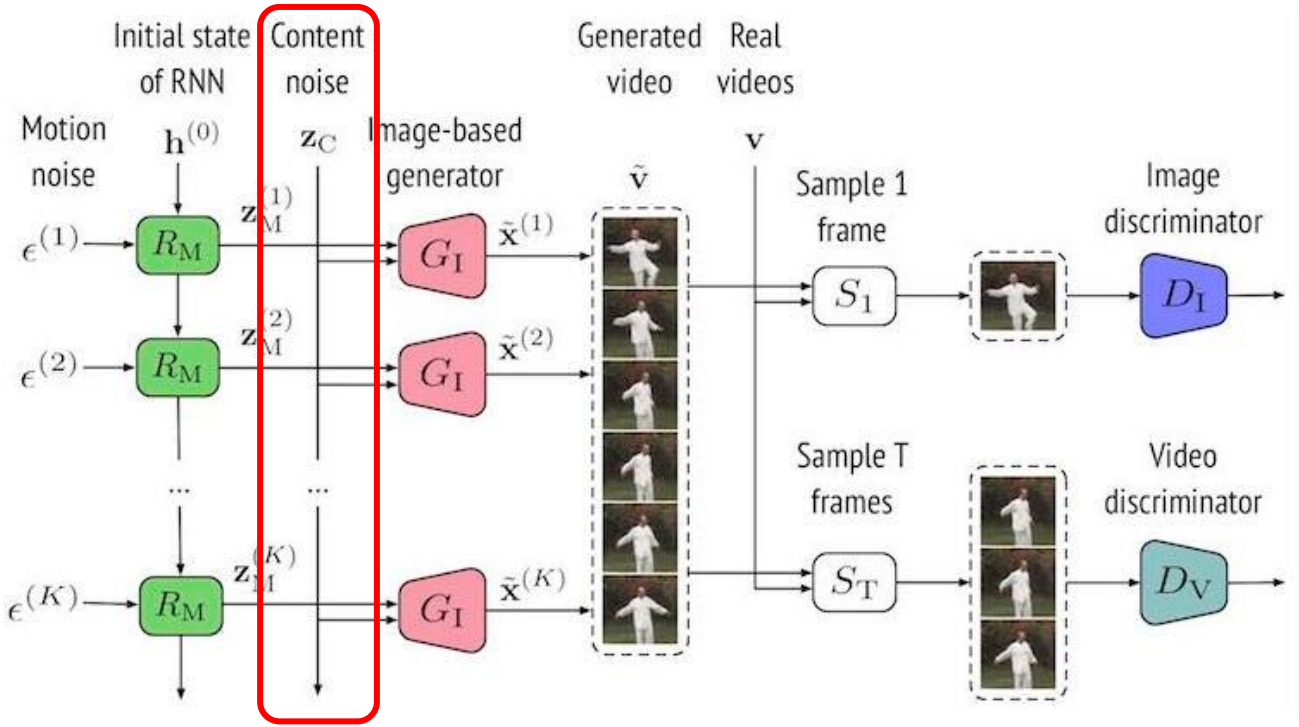
Motion trajectory

$$\mathbf{Z}_\mathrm{M} = [\mathbf{z}_\mathrm{M}^{(1)}, \mathbf{z}_\mathrm{M}^{(2)}, ..., \mathbf{z}_\mathrm{M}^{(K)}]$$



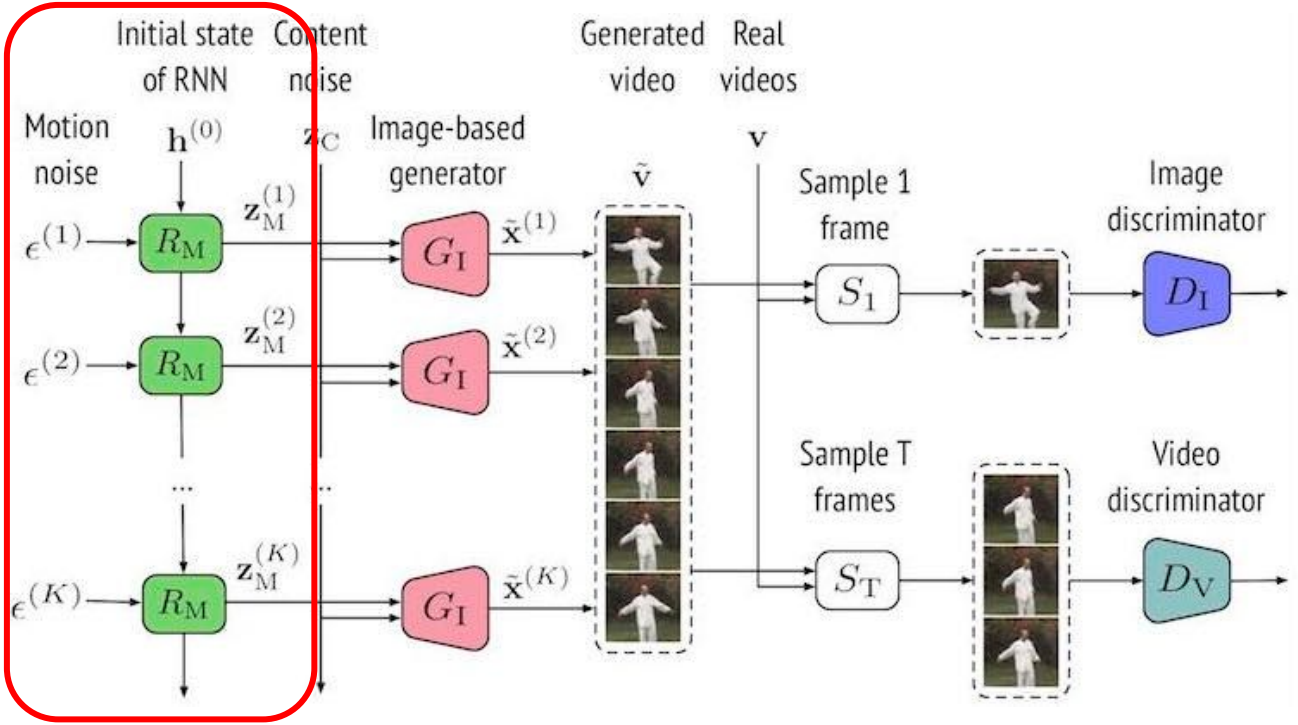
MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN



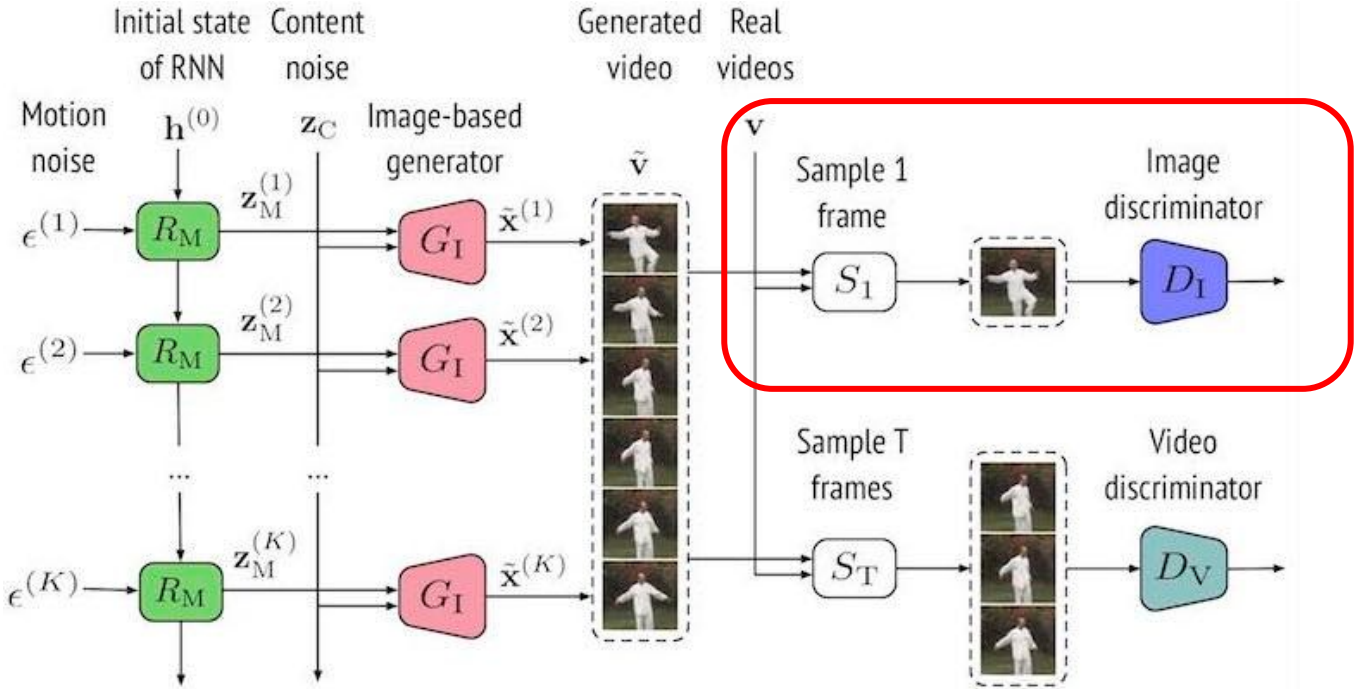
MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN



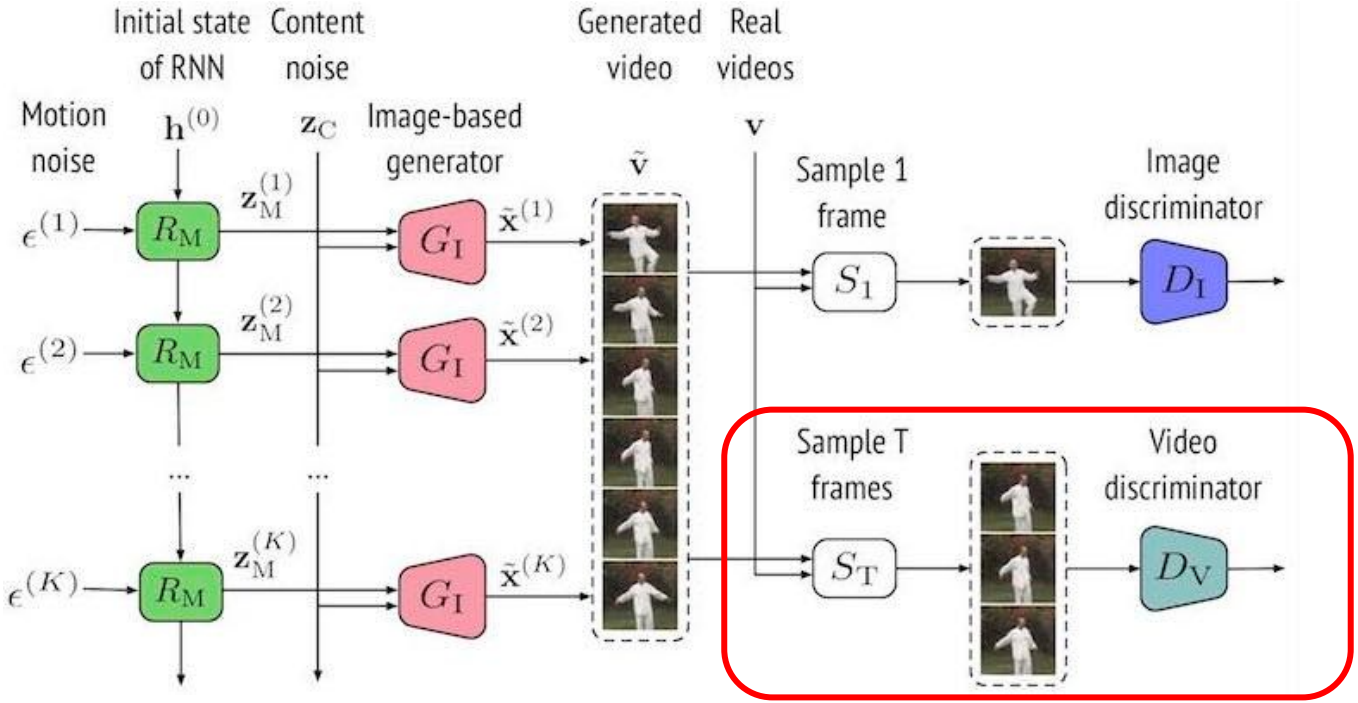
MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN



MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN



MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN

MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: MoCoGAN

We can change the motion without changing the content:

MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

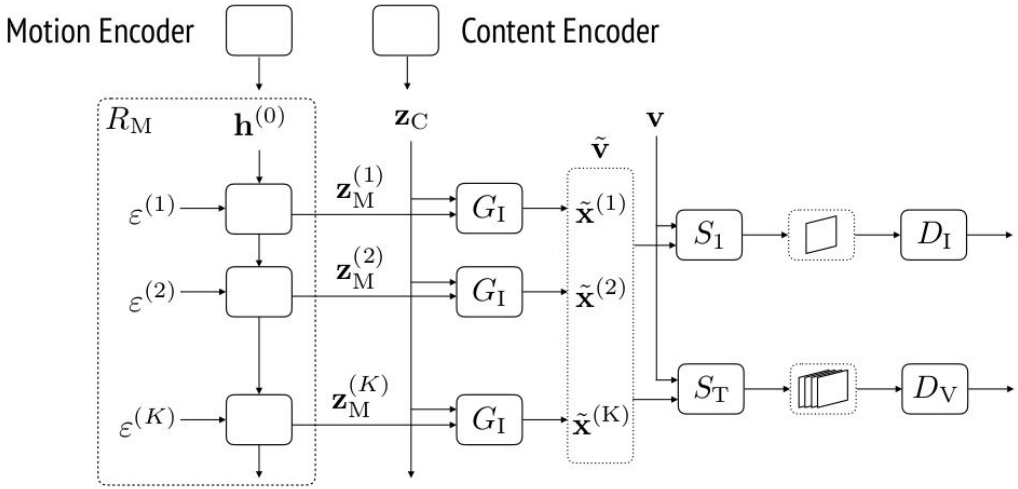
Video Generation: MoCoGAN

Video prediction:

MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

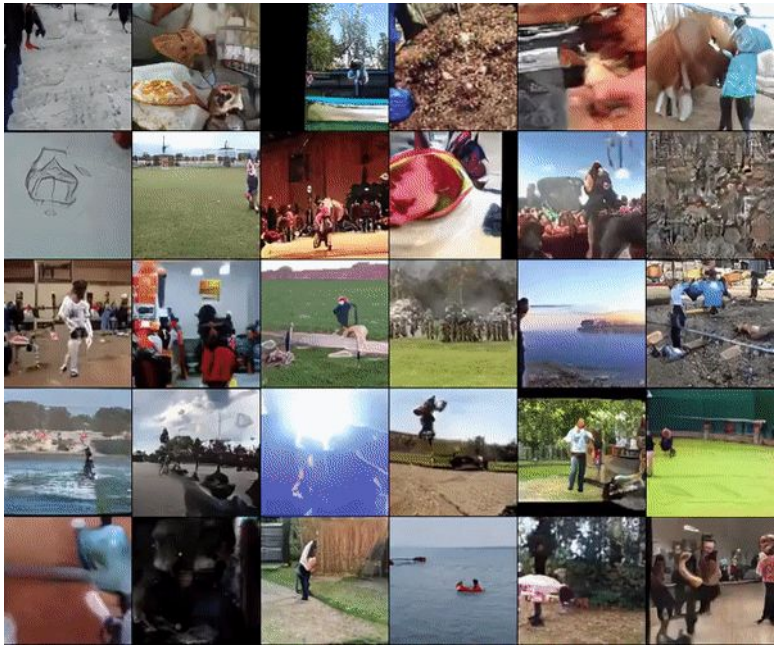
Video Generation: MoCoGAN

Video prediction:



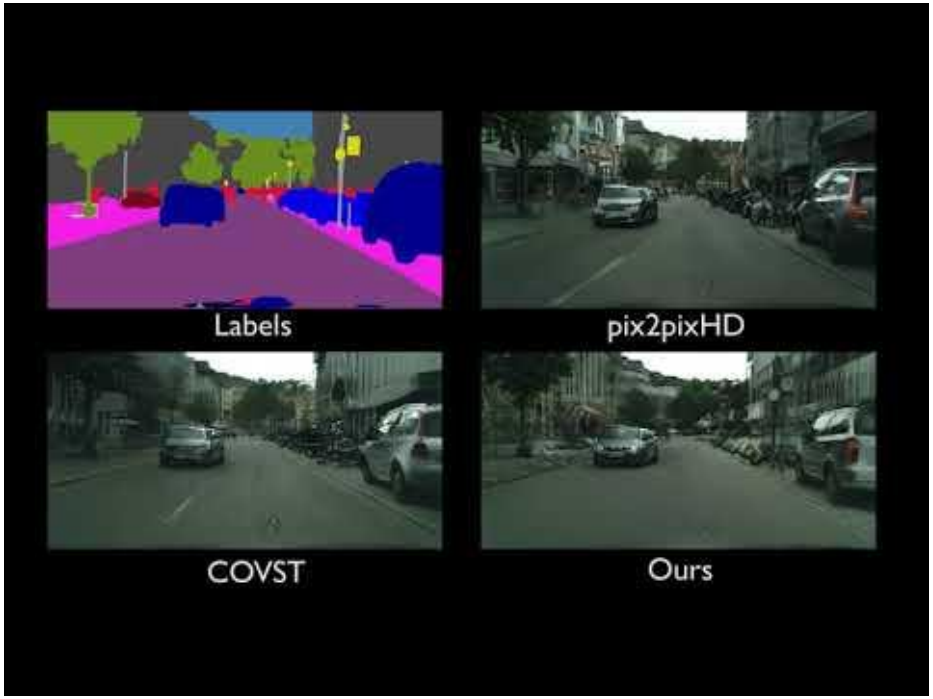
MoCoGAN: Decomposing Motion and Content for Video Generation Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR'2018.

Video Generation: DVDGAN



Clark, Aidan, Jeff Donahue, and Karen Simonyan. "Efficient video generation on complex datasets." ArXiv preprint

Video-to-video translation

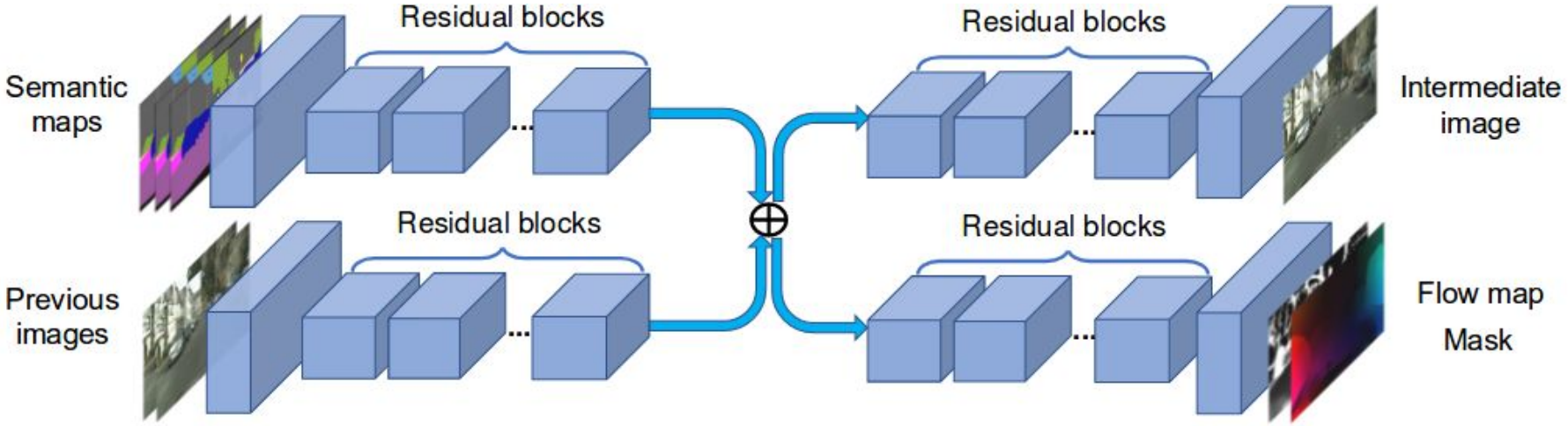


Wang, Ting-Chun, et al. "Video-to-video synthesis." NIPS'2018

Video-to-video translation

Wang, Ting-Chun, et al. "Video-to-video synthesis." NIPS'2018

Video-to-video translation



Previous generated frame

$$F(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}) = (\mathbf{1} - \tilde{\mathbf{m}}_t) \odot \tilde{\mathbf{w}}_{t-1}(\tilde{\mathbf{x}}_{t-1}) + \tilde{\mathbf{m}}_t \odot \tilde{\mathbf{h}}_t$$

Optical flow warping

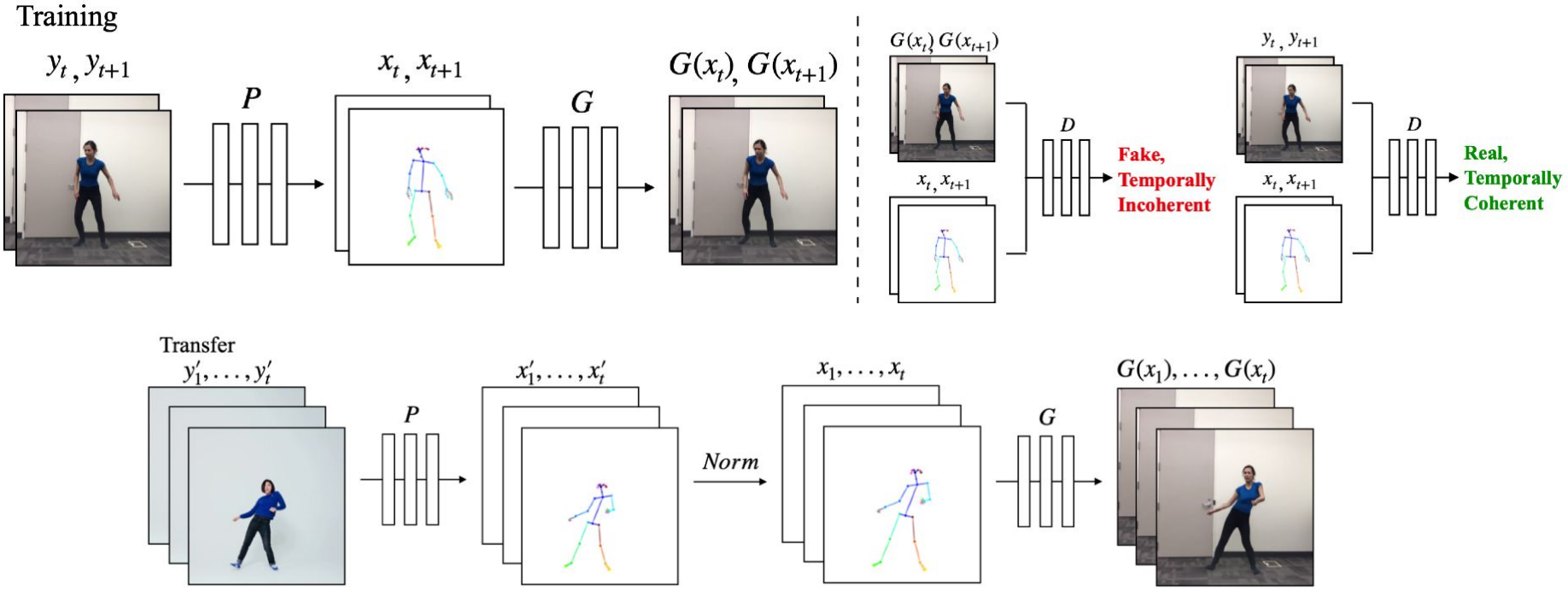
mask

Wang, Ting-Chun, et al. "Video-to-video synthesis." NIPS'2018

Animating Single Subject

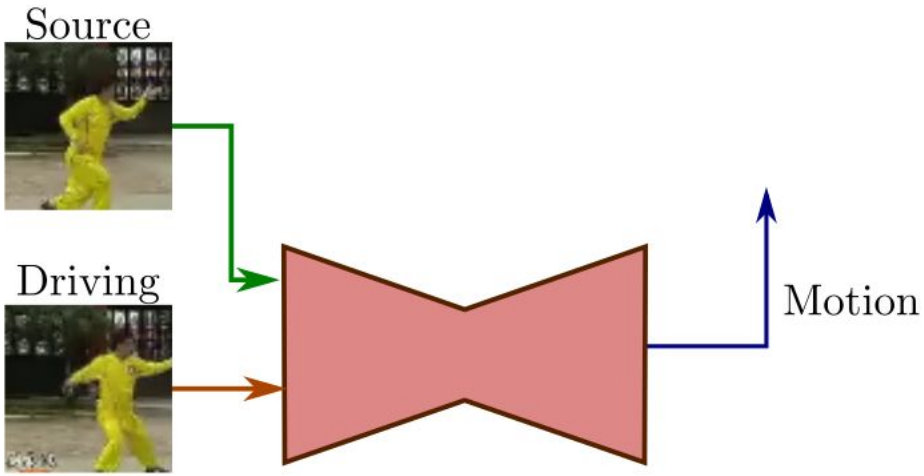
Chan, Caroline, et al. "Everybody dance now." ICCV'2019.

Animating Single Subject



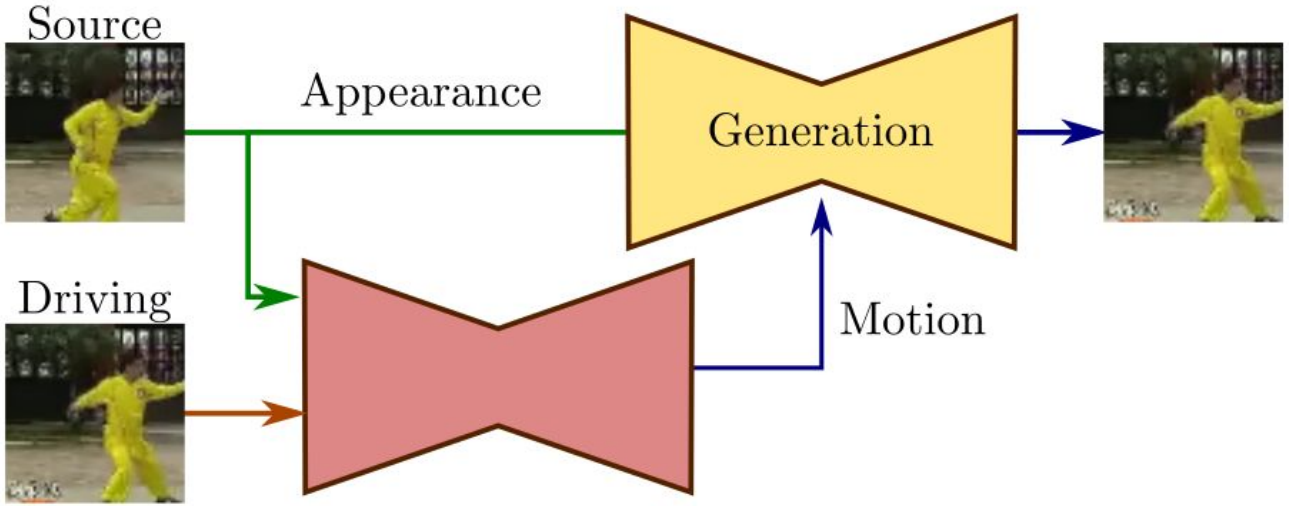
Chan, Caroline, et al. "Everybody dance now." ICCV'2019.

Retargeting



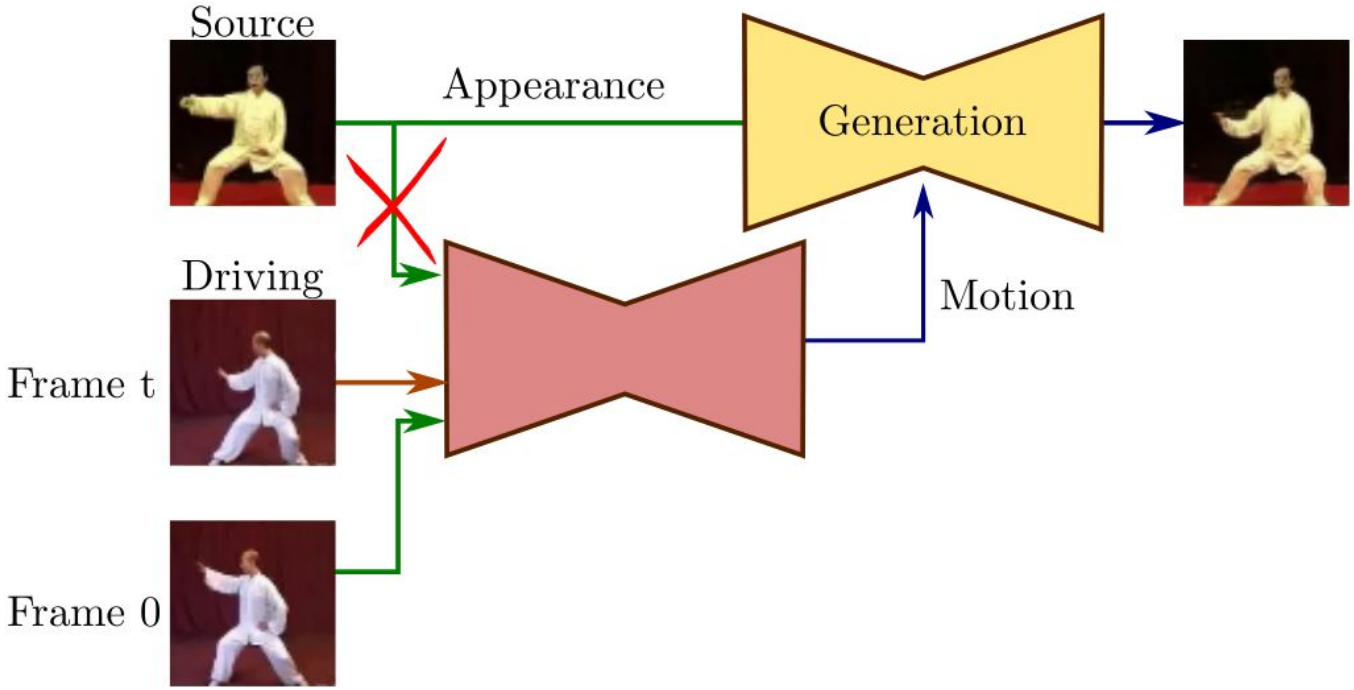
A. Siarohin, S. Lathuilière, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Retargeting



A. Siarohin, S. Lathuilière, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Retargeting



A. Siarohin, S. Lathuilière, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Retargeting

A. Siarohin, S. Lathuilière, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019