Survey of image to image transform

Jingqiu Zhang
2017/9/21

未来媒体研究中心
CENTER FOR FUTURE MEDIA
电子科技大学
University of Electronic Science and Technology of China

Outline

- Introduction
- Applications
- Classification
  - Supervised/Unsupervised
  - Content/Style
- Method
  - Architecture
  - Experiment
  - Evaluation
- Conclusion

1 Introduction

Definition of “image to image transform”：

- Learn the mapping between an input image and output image,thus giving an input image, we can get the relevant output image

Motivation：

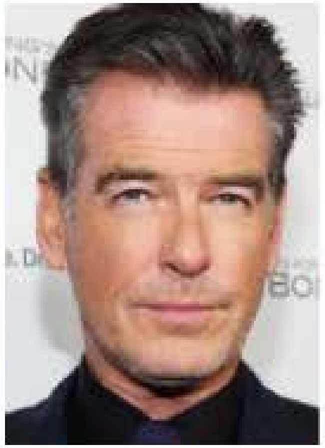
- Create “new” image.

The reason for this research：

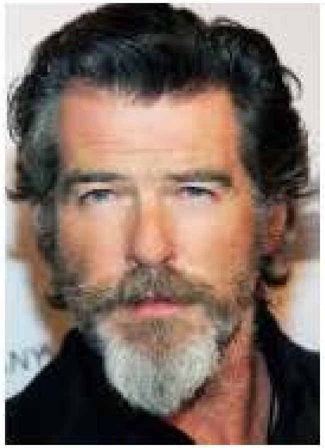
- Multi-domain images are views of an object with different attributes.

Non-smiling

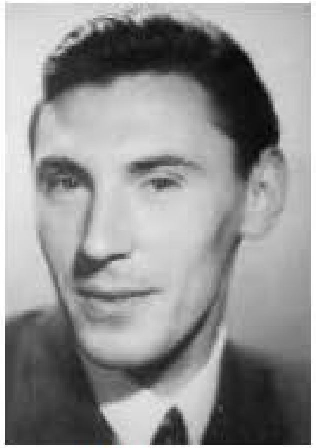
Smiling



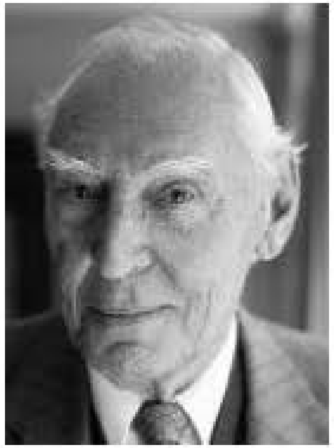
Non-beard



Beard



Young



Senior

恭發 恭發
喜財 喜財

Font#1 Font#2

images

Hand-drawings

summer

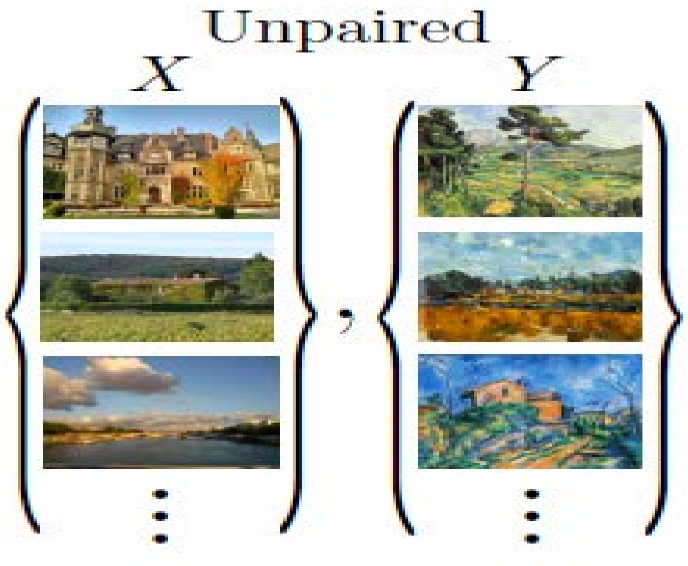
winter

2 Applications

- Object transfer
  (color/texture)
- Enhance image quality
  - Blurry to sharp
  - Low-res to high-res
  - Noisy to clean
- Style transfer
  - Season transfer/day to night
  - Painting style transfer
  - Photo generation from paintings
- Edge to photo
  - sketch to photo
  - Aerial photo to map
- Face feature change
  - Gender transfer
  - Age transfer
  - Hair color/sunglasses/expression

3 Classification

- Supervised (paired dataset)
  --One to one.
- Unsupervised (unpaired dataset)
  --Transform between two domains.



未来媒体研究中心 CENTER FOR FUTURE MEDIA
电子科技大学 University of Electronic Science and Technology of China

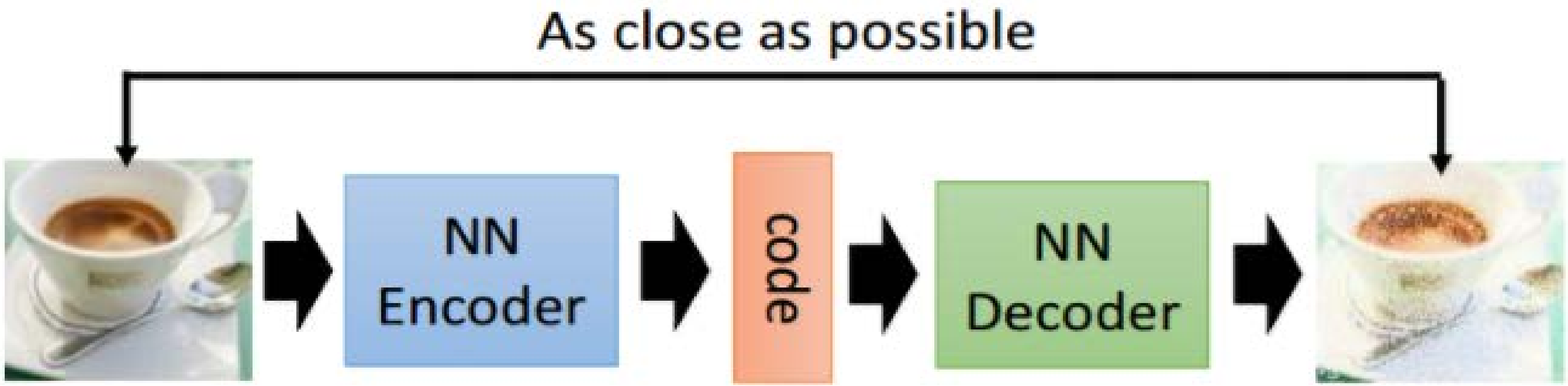
3 Classification

- transfer particular content
  - Encode input image into high-feature latent , then change the particular latent. (face feature change)

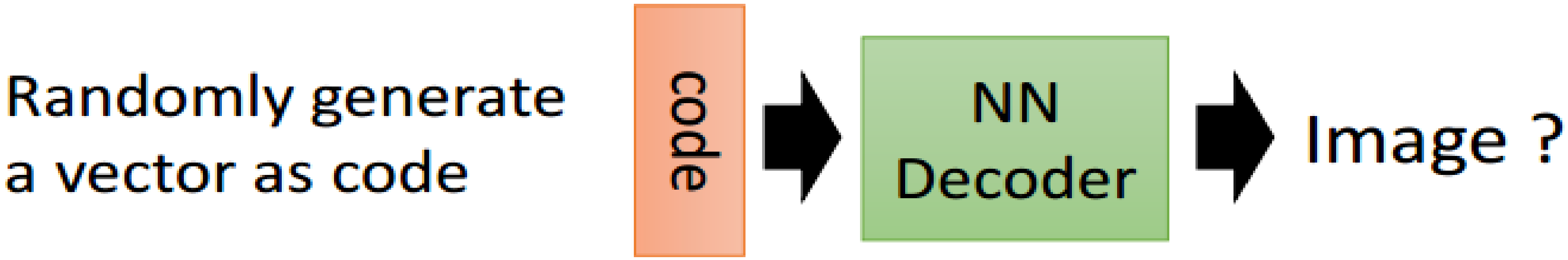
- transfer the basic style
  - cross domain (style transfer\face to emoji\photo to sketch)

4 Method

- Prior： AE 、 VAE



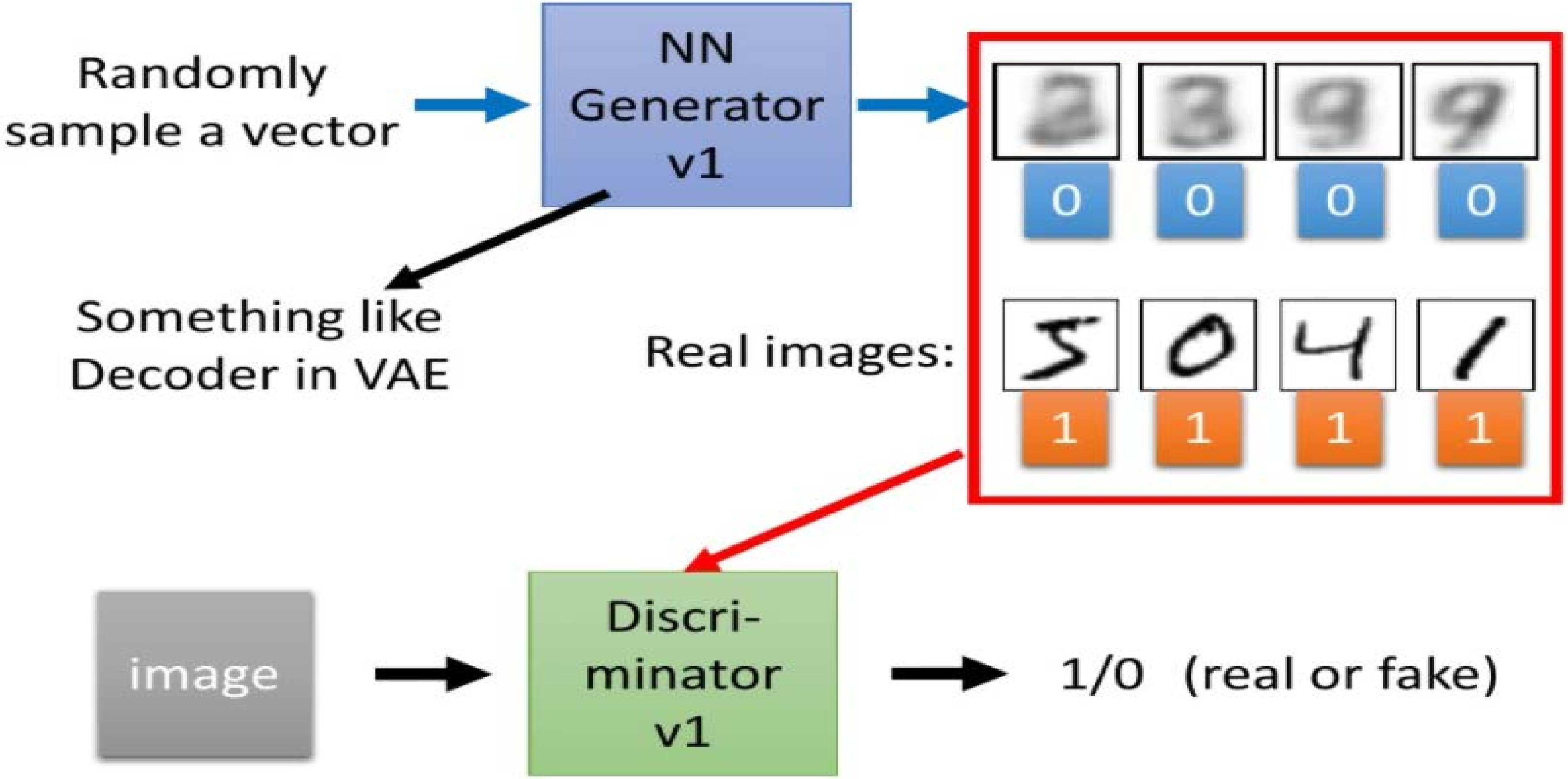
Auto Encoder： no-label dataset， unsupervised， learn features



Variational Auto Encoder： latent vector has prior distribution， sampling， decode

4 Method

- Prior： GAN



4 Method

- Basic-1： VAE + GAN

《Auto-encoding beyond pixels using a learned similarity metric 》ICML,2016

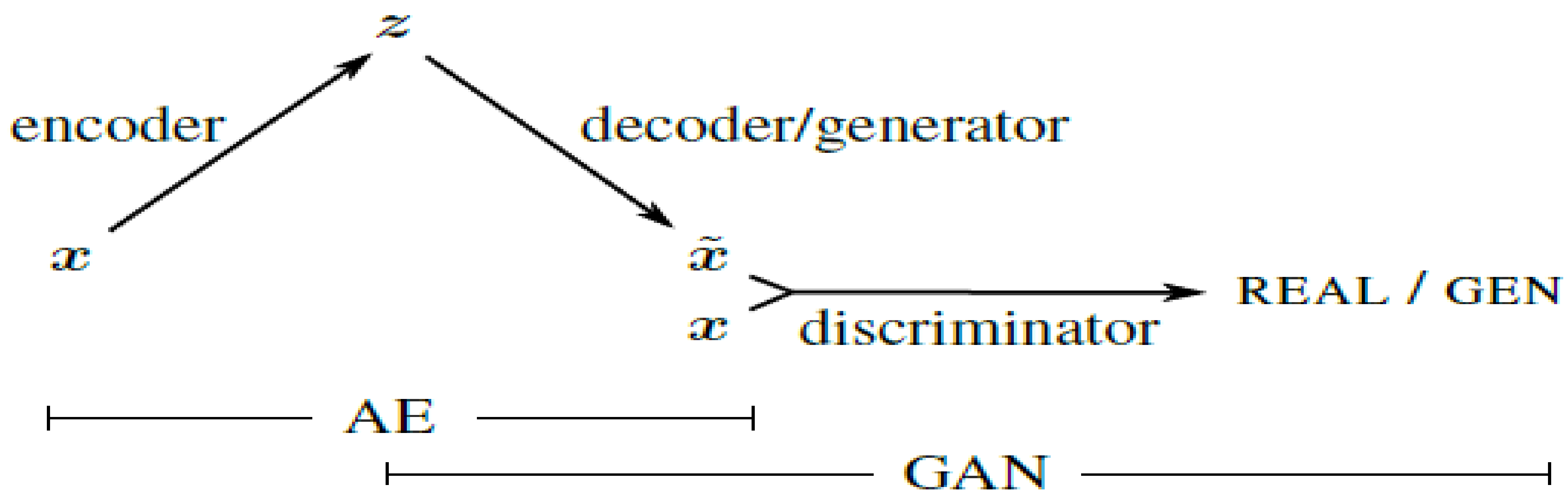
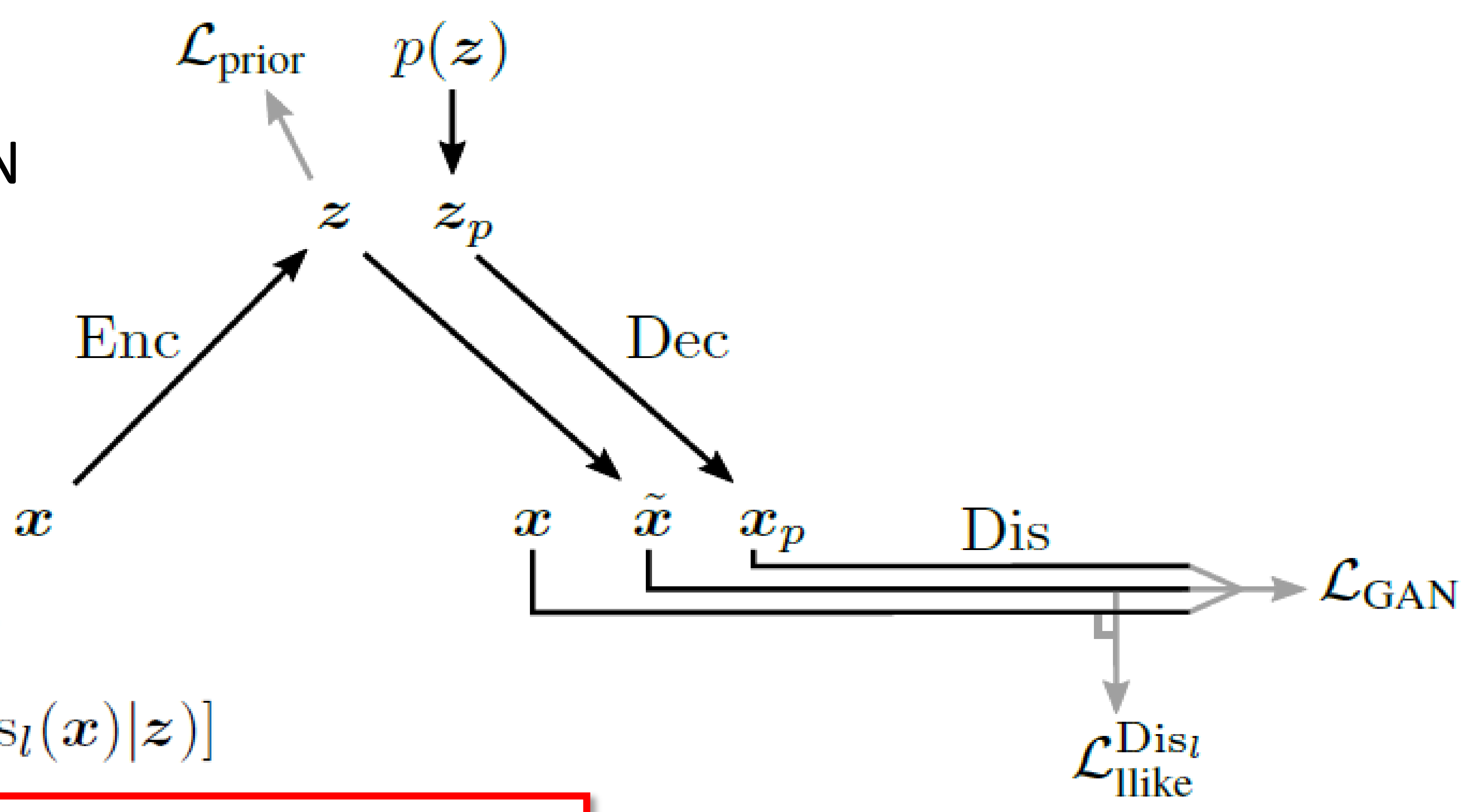


Figure 1. Overview of our network. We combine a VAE with a GAN by collapsing the decoder and the generator into one.

4 Method

- Basic-1： VAE + GAN



$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}(q(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))$$

$$\mathcal{L}_{\text{llike}}^{\text{Dis}_l} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\text{Dis}_l(\boldsymbol{x})|\boldsymbol{z})\right]$$

$$\boxed{p(\text{Dis}_l(\boldsymbol{x})|\boldsymbol{z}) = \mathcal{N}(\text{Dis}_l(\boldsymbol{x})|\,\text{Dis}_l(\tilde{\boldsymbol{x}}), \mathbf{I})}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \quad & \log(\text{Dis}(\boldsymbol{x})) + \log(1 - \text{Dis}(\text{Dec}(\boldsymbol{z}))) \\ & + \log(1 - \text{Dis}(\text{Dec}(\text{Enc}(\boldsymbol{x}))))\end{aligned}$$

4 Method

- Basic-2： Condition GAN

《Conditional Generative Adversarial Nets》NIPS,2014

Why to add Condition

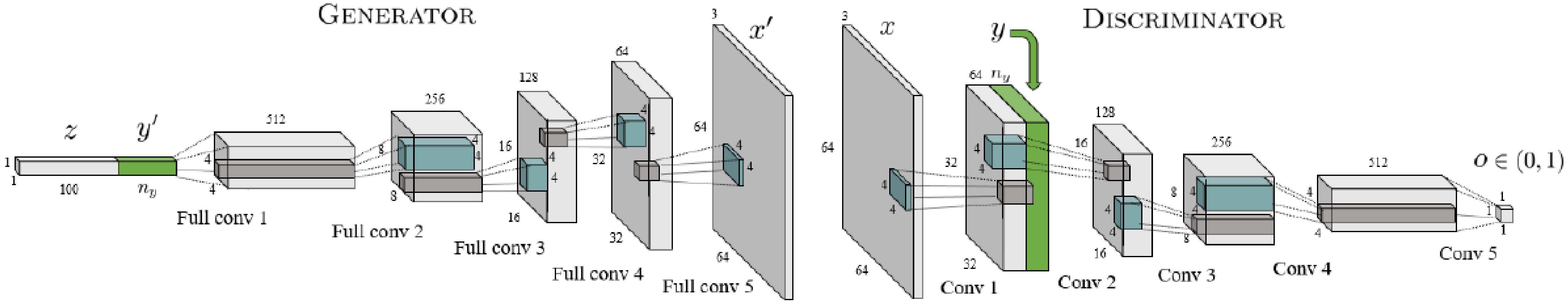
- stable model
- multimodal feature

How to add Condition

- Add in G
- Add in D

4 Method

- Basic-2： Condition GAN

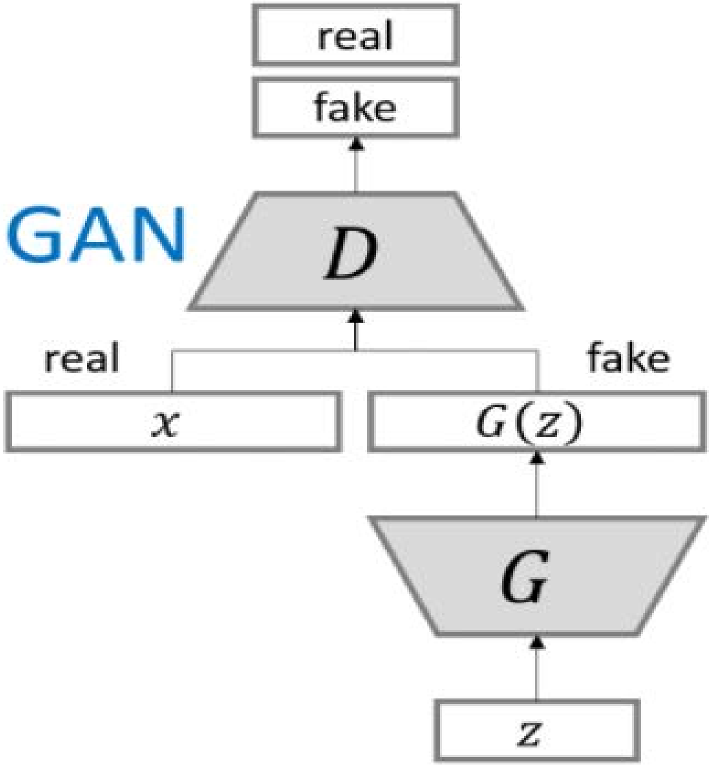


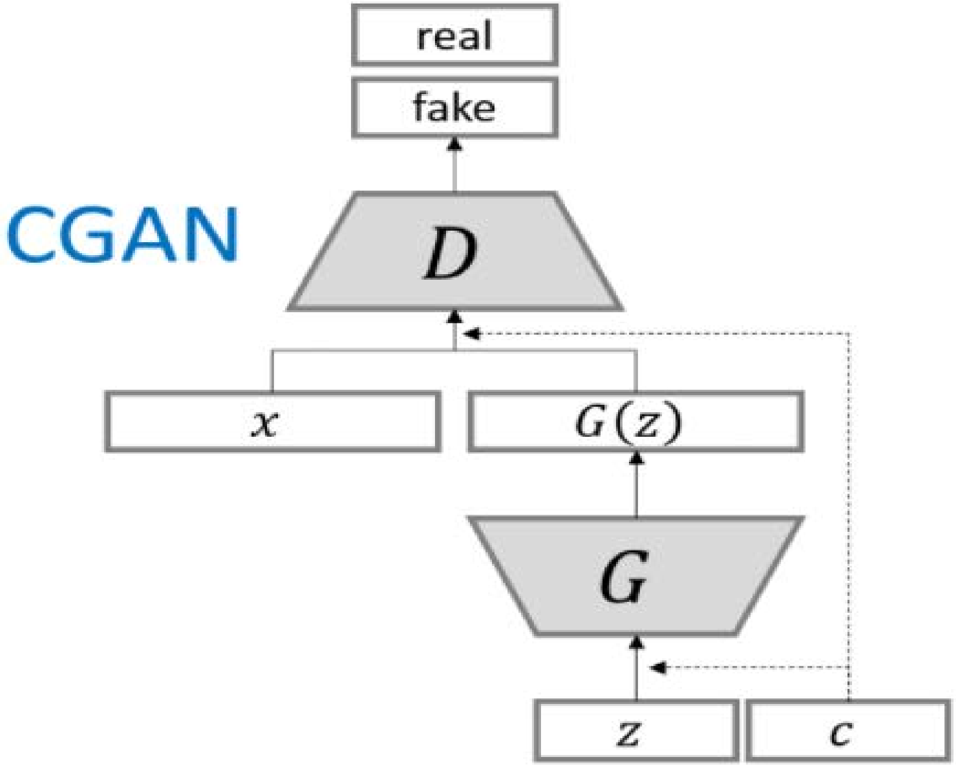
z【1x1x100】+y【1x1x10】-- 【1x1x110】

【1x1x10】spatially replicate→【32x32x10】---- 【32x32x(10+64)】

4 Method

- Basic-2： Condition GAN





$$L_D^{GAN} = E[\log(D(x))] + E[\log(1 - D(G(z)))]$$

$$L_G^{GAN} = E[\log(D(G(z)))]$$

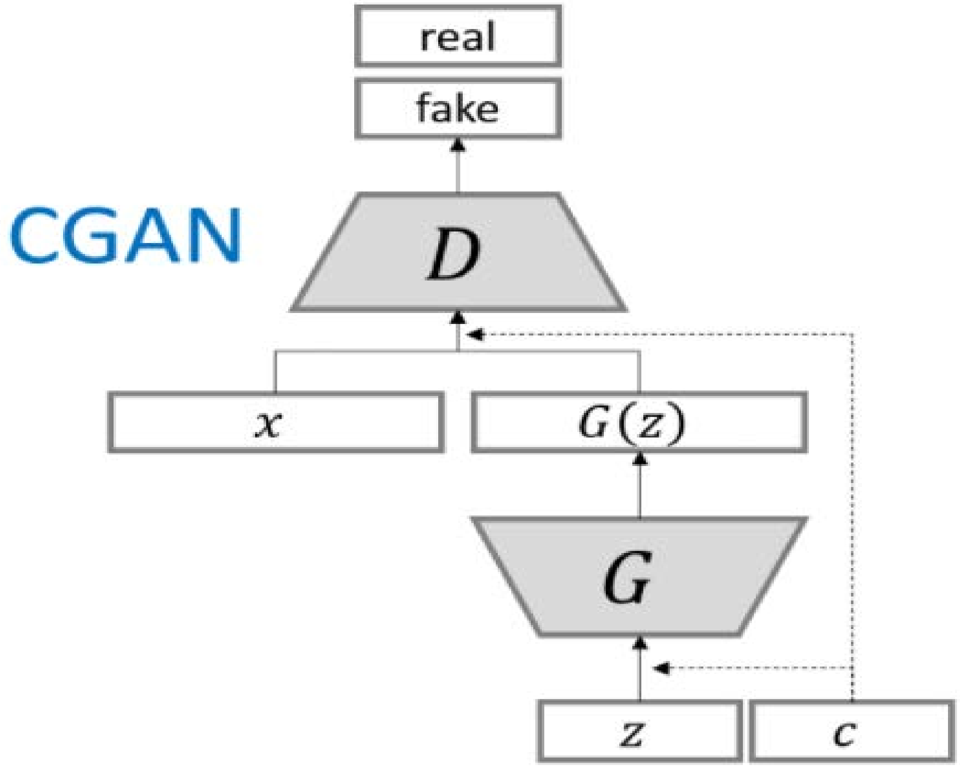
$$L_D^{CGAN} = E[\log(D(x,c))] + E[\log(1 - D(G(z),c))]$$

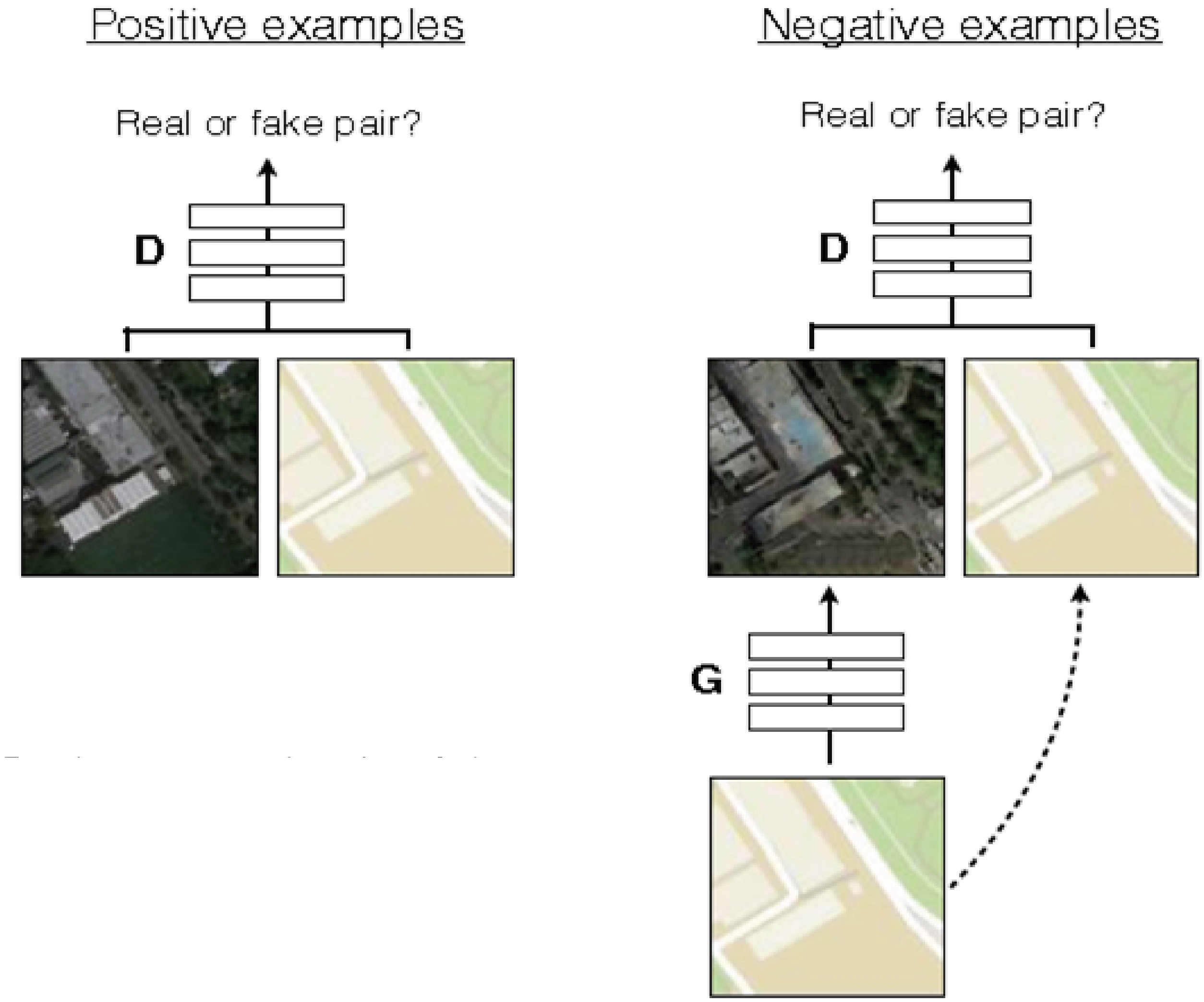
$$L_G^{CGAN} = E[\log(D(G(z),c))]$$

未来媒体研究中心 CENTER FOR FUTURE MEDIA | 电子科技大学 University of Electronic Science and Technology of China

4 Method

- Basic-2： Condition GAN





《Image-to-Image Translation with Conditional Adversarial Networks》 CVPR,2017[BAIR]

Novelty:

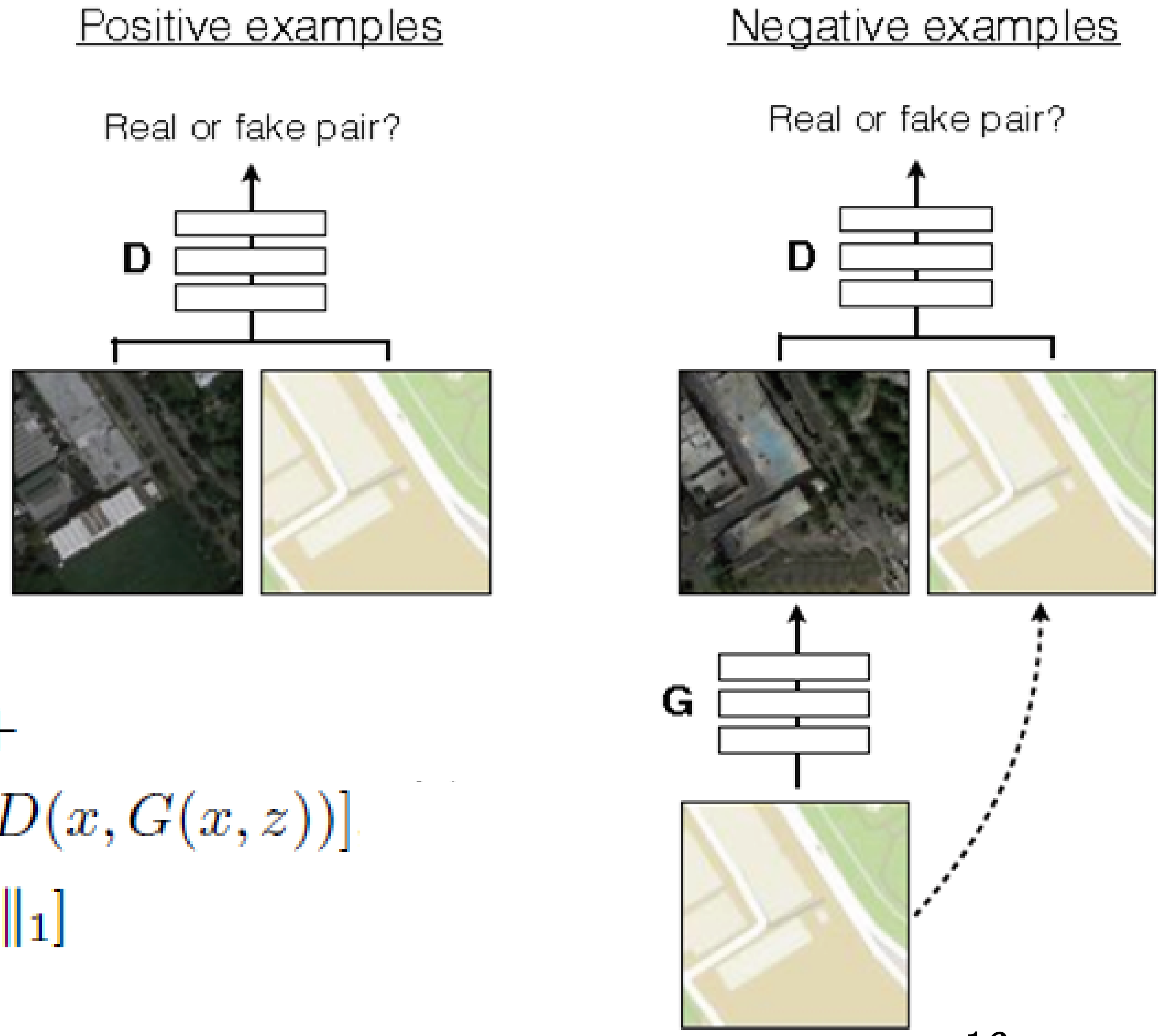
- Pix2pix-map pixels to pixels
- Build simple loss functions based on GAN

Loss function:

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

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) =& \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)] + \\ & \mathbb{E}_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z))]\end{aligned}$$

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



4 Method

《Image-to-Image Translation with Conditional Adversarial Networks》 CVPR,2017

	GAN	CGAN	pix2pix
Input of G	z	z+c	(z)+x
Output of G	G(z)	G(z,c)	G(z,x)
Input of D	G(z) / real image	G(z,c)+c / real image+c	G(z,x)+x / real image+x
Output of D	(0,1)	(0,1)	(0,1)

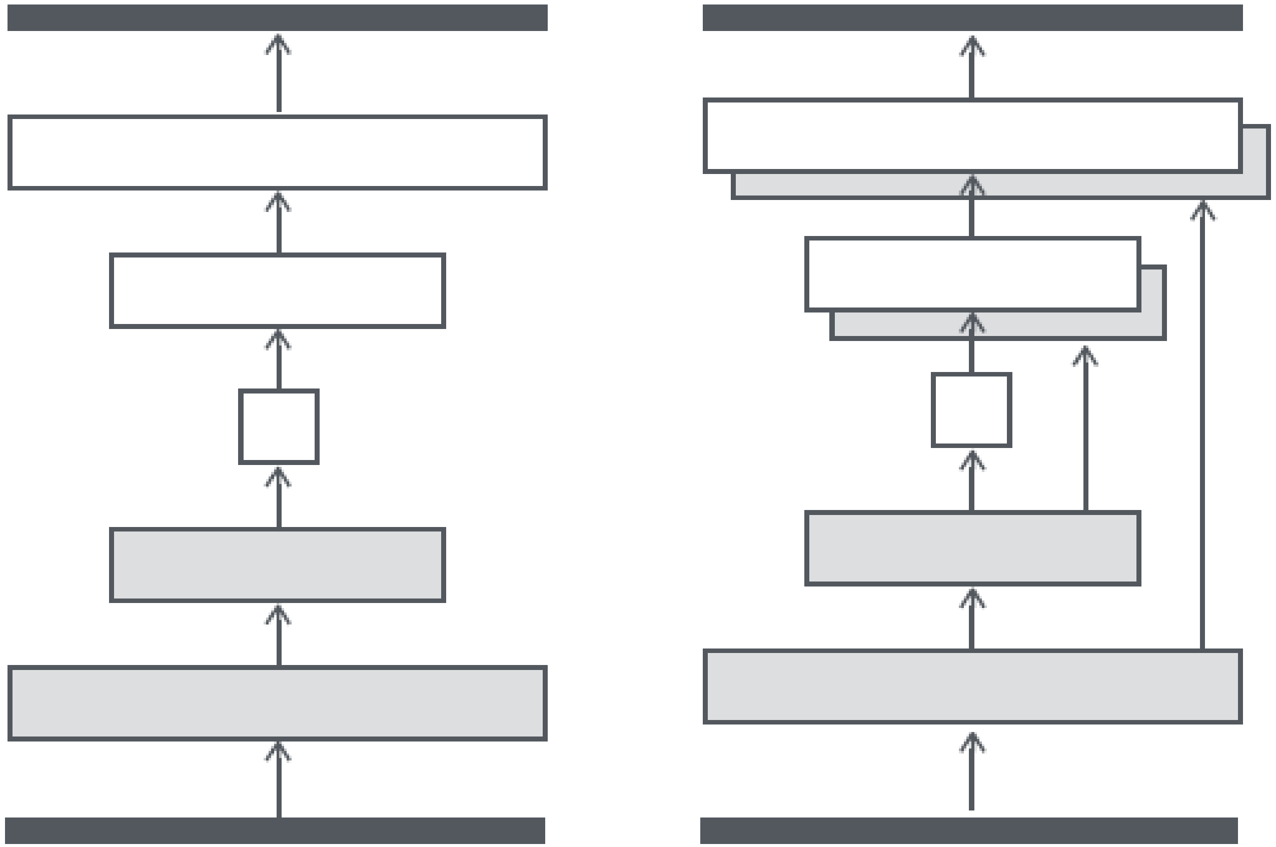
Loss function: GAN_loss + L1_loss

Architecture: PatchGAN for D -- N×N batch as input, high level feature

U_net for G -- Encoder+Decoder with skip connection →

4 Method

《U-net: Convolutional networks for biomedical image segmentation》MICCAI,2015



Encoder-decoder　　U-Net

Skip connection：

how -- Concatenate layers

where -- ResNet/DenseNet

why -- Share feature

4 Method--Evaluation

- Error metrics

1. Amazon Mechanical Turk (AMT) perceptual loss:
Each participant will be shown pairs of images and asked to click on the image they thought is correct. The rate at which the algorithm fools the participants is recorded.
2. Fully-Connected Network (FCN) score:
First apply FCN to the image and then compute the semantic segmentation metrics by comparing with the ground-truth label.
3. Semantic segmentation metrics:
Three metrics are used: per-pixel accuracy, mean class IoU and per-class accuracy.

$$\text{Per-pixel accuracy:} \quad \frac{\text{The number of corrected labled pixels}}{\text{The total number of pixels}}$$

Per-class accuracy: The per-pixel accuracy for each class.

$$\text{Mean class IoU:} \quad \frac{\text{The Intersection of Pixel with the same label}}{\text{The Union of Pixel with the same label}}$$

未来媒体研究中心 CENTER FOR FUTURE MEDIA | 电子科技大学 University of Electronic Science and Technology of China

4 Method--Evaluation

Loss	Photo → Map % Turkers labeled *real*	Map → Photo % Turkers labeled *real*
L1	2.8% ± 1.0%	0.8% ± 0.3%
L1+cGAN	6.1% ± 1.3%	**18.9% ± 2.5%**

Table 1:
AMT "real vs fake" test on maps↔aerial photos.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.44	0.14	0.10
GAN	0.22	0.05	0.01
cGAN	0.61	**0.21**	**0.16**
L1+GAN	**0.64**	0.19	0.15
L1+cGAN	0.63	**0.21**	**0.16**
Ground truth	0.80	0.26	0.21

Table 2:
FCN-scores for different losses, evaluated on Cityscapes Labels→photos.

4 Method--Experiment

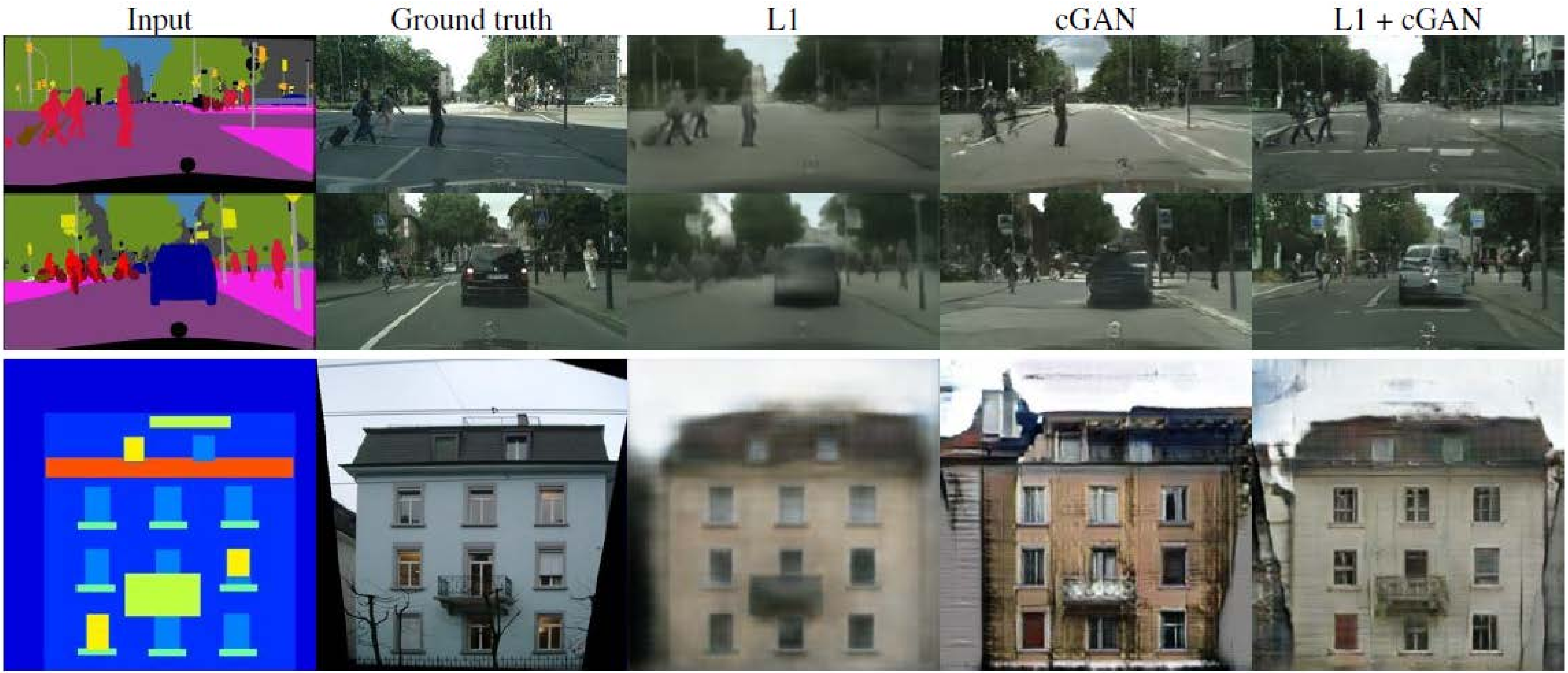


Figure1： Example results of our method on Cityscapes labels → photo

4 Method

- Supervised： VAE +Condition GAN

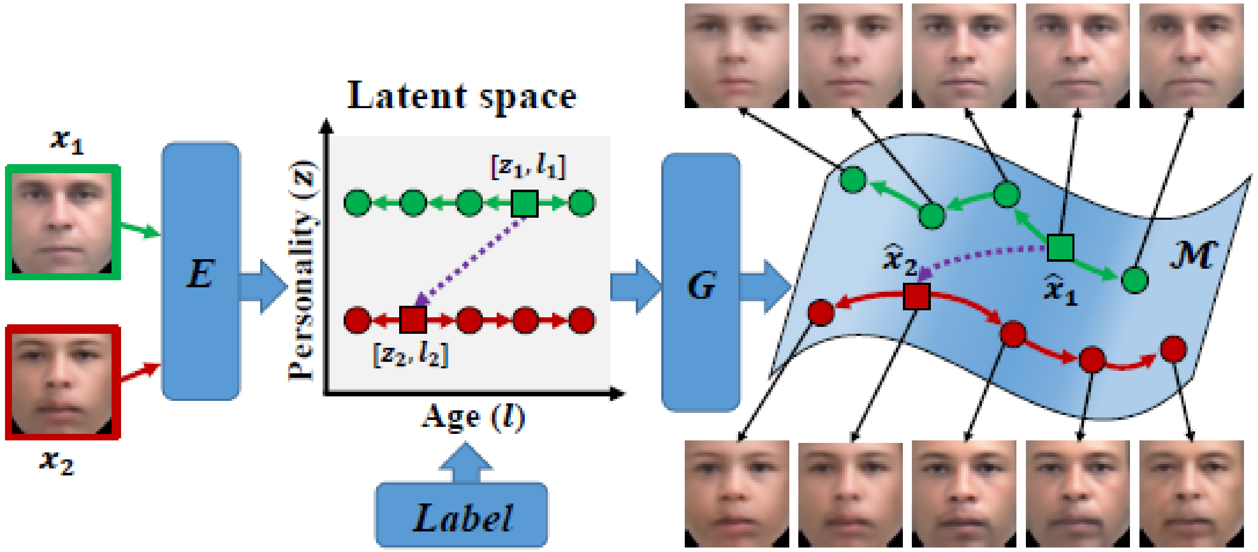
《Age Progression/Regression by Conditional Adversarial Auto-encoder》 CVPR,2017

Traditional methods：

- physical model-based： too complex
- prototype-based： age group-based

Novelty：

Progression/Regression



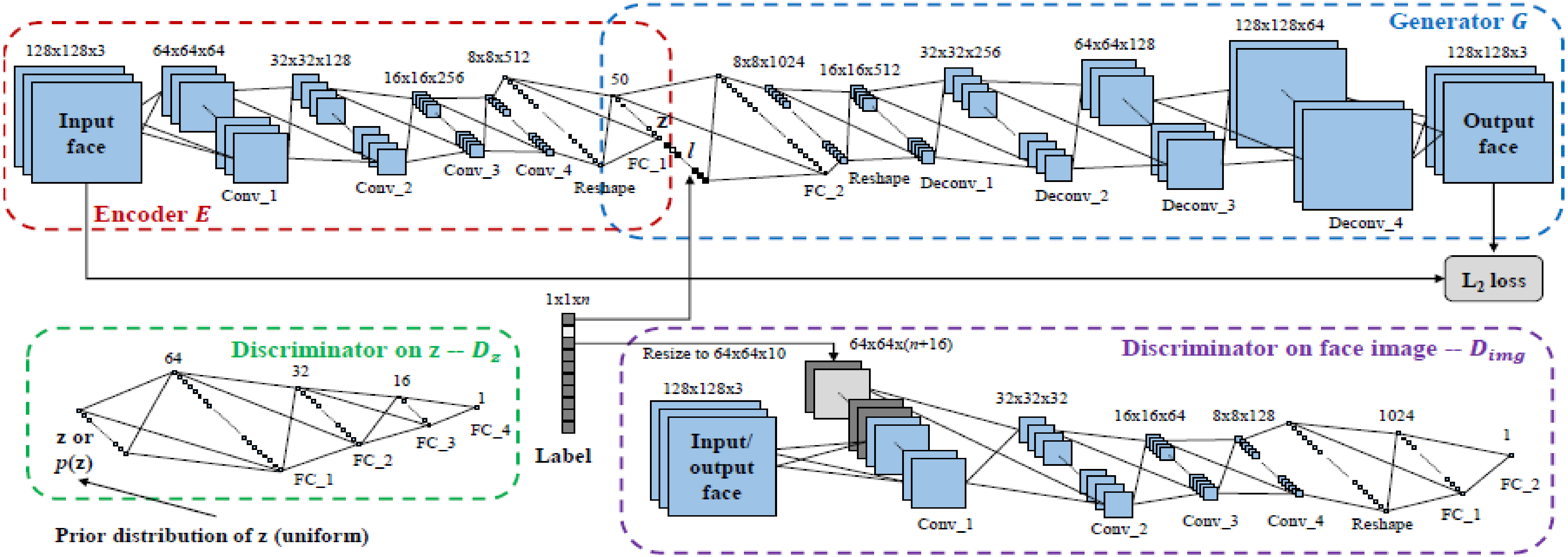
X1→E→z1[personality]+Age label[L1]=latent vector

[z1,L1]→G=x1’

未来媒体研究中心 CENTER FOR FUTURE MEDIA | 电子科技大学 University of Electronic Science and Technology of China

《Age Progression/Regression by Conditional Adversarial Auto-encoder》

CVPR,2017 [TN, USA]



$$\min_{E,G} \max_{D_z,D_{img}} \lambda\mathcal{L}(x, G(E(x),l)) + \gamma TV(G(E(x),l))$$

$$+\mathbb{E}_{z^*\sim p(\mathbf{z})}[\log D_z(z^*)] \qquad +\mathbb{E}_{x,l\sim p_{data}(\mathbf{x},\mathbf{l})}[\log D_{img}(x,l)]$$

$$+\mathbb{E}_{x\sim p_{data}(\mathbf{x})}[\log(1-D_z(E(x)))] \qquad +\mathbb{E}_{x,l\sim p_{data}(\mathbf{x},\mathbf{l})}[\log(1-D_{img}(G(E(x),l)))]\ ,$$

4 Method--Experiment

Figure1：

- *Comparison to prior works of face aging.*
- *The first column shows input faces, and second column are the best aged faces cited from prior works.*
- *The rest columns are our results from both age progression and regression.*
- *The red boxes indicate the comparable results to the prior works.*

未来媒体研究中心
CENTER FOR FUTURE MEDIA
电子科技大学

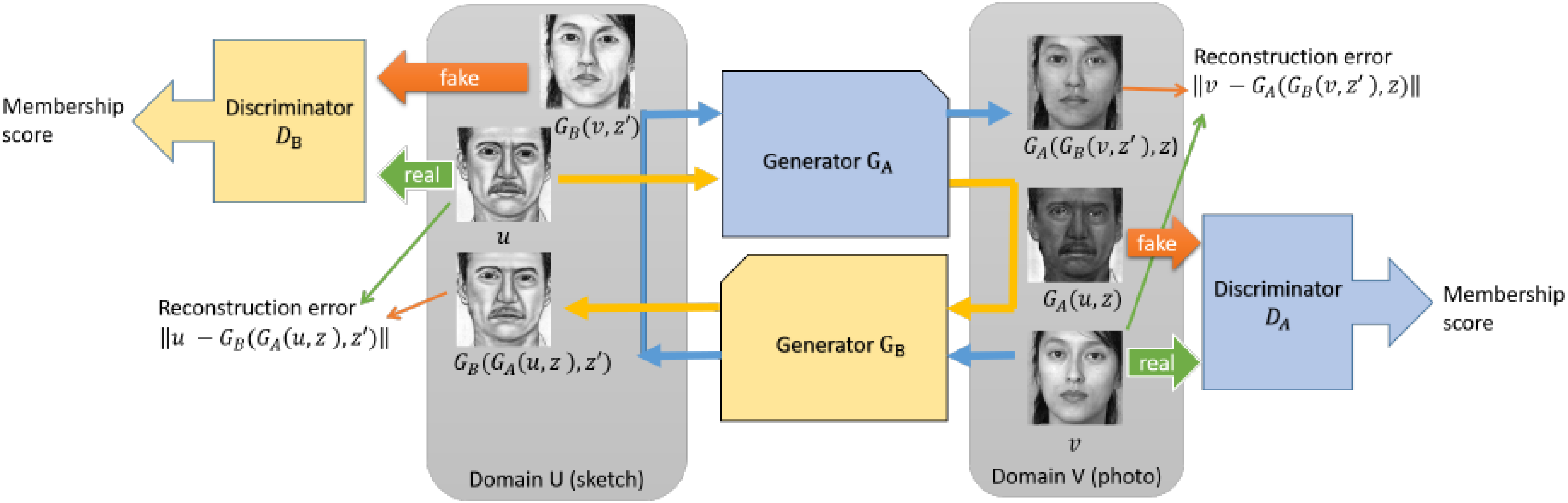
4 Method

- Unsupervised：

Dual model(Cross domain)

- Dual GAN -- ICCV,2017.4.30
- Cycle GAN -- ICCV,2017.3.30
- Disco GAN -- ICML,2017.3.15

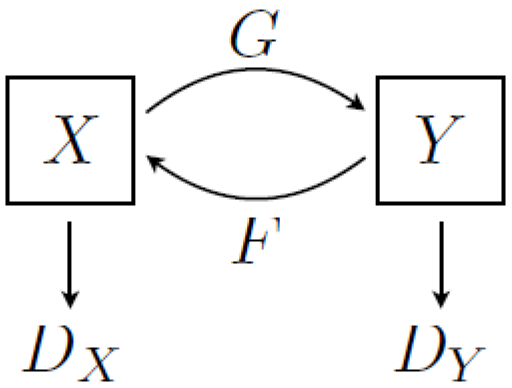
《DualGAN: Unsupervised Dual Learning for Image-to-Image Translation》
ICCV,2017

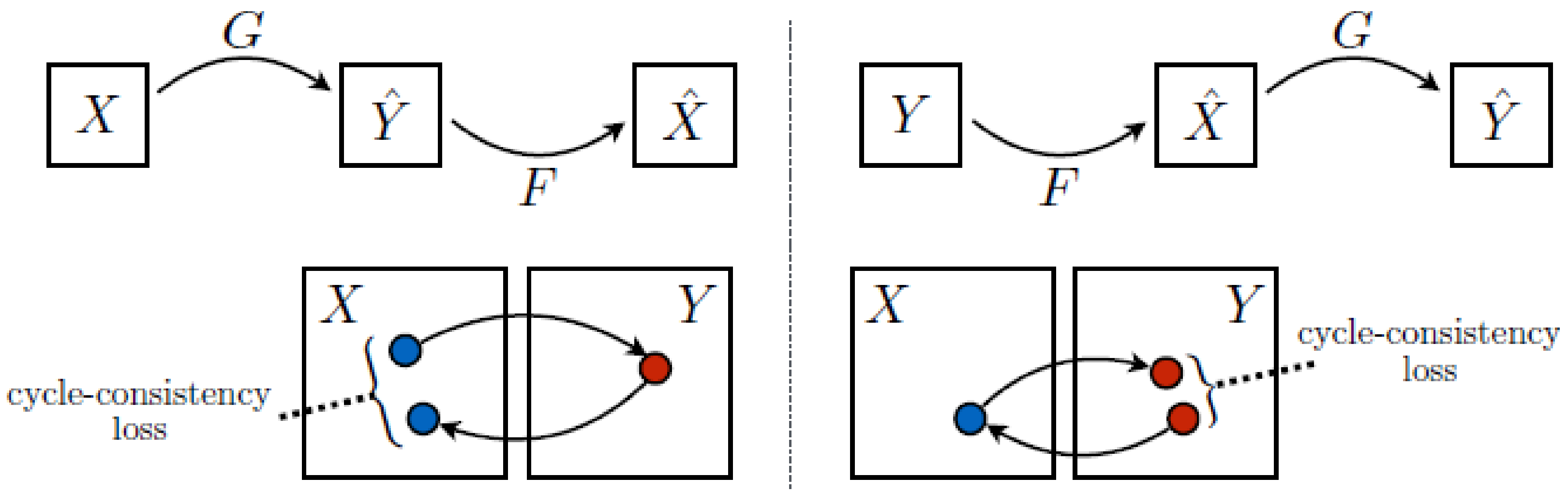


$$l_A^d(u,v) = D_A(G_A(u,z)) - D_A(v),$$
$$l_B^d(u,v) = D_B(G_B(v,z')) - D_B(u)$$

$$l^g(u,v) = \lambda_U \|u - G_B(G_A(u,z),z')\| + \lambda_V \|v - G_A(G_B(v,z'),z)\| - D_A(G_B(v,z')) - D_B(G_A(u,z)),$$

《Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks》 ICCV,2017



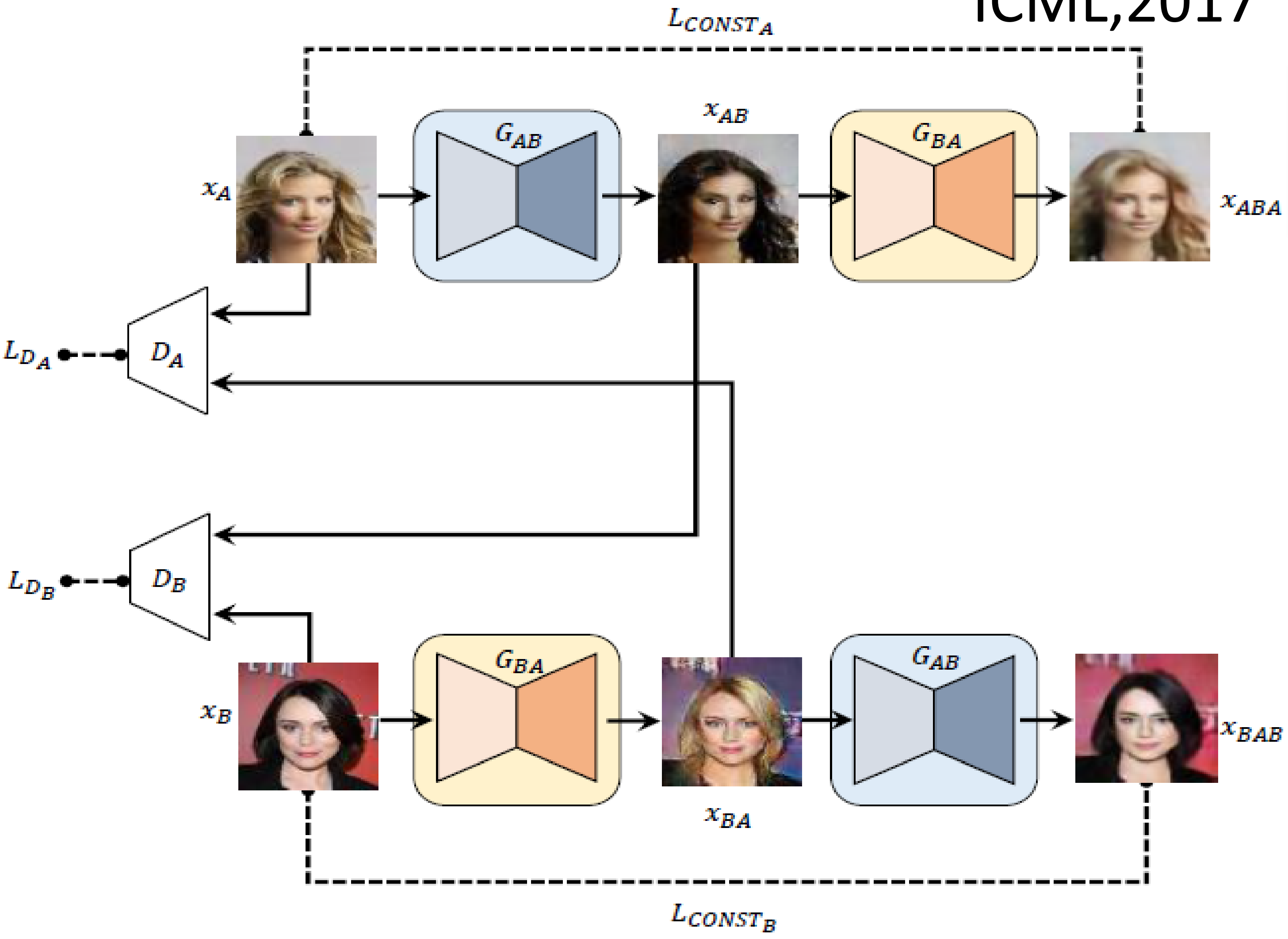


$$\boxed{\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) =& \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ &+ \lambda \mathcal{L}_{\text{cyc}}(G, F),\end{aligned}}$$

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

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

《Learning to Discover Cross-Domain Relations》 ICML,2017



$$\begin{aligned} L_G &= L_{G_{AB}} + L_{G_{BA}} \\ &= L_{GAN_B} + L_{CONST_A} + L_{GAN_A} + L_{CONST_B} \end{aligned}$$

$$L_D = L_{D_A} + L_{D_B}$$

$$\begin{aligned} L_{CONST_A} &= d(\mathbf{G}_{BA} \circ \mathbf{G}_{AB}(x_A), x_A) \\ L_{GAN_B} &= -\mathbb{E}_{x_A \sim P_A}[\log \mathbf{D}_B(\mathbf{G}_{AB}(x_A))] \\ L_{D_B} &= -\mathbb{E}_{x_B \sim P_B}[\log \mathbf{D}_B(x_B)] \\ &\quad - \mathbb{E}_{x_A \sim P_A}[\log(1 - \mathbf{D}_B(\mathbf{G}_{AB}(x_A)))] \end{aligned}$$

4 Method--Evaluation

- Error metrics

1. Amazon Mechanical Turk (AMT) perceptual loss:
Each participant will be shown pairs of images and asked to click on the image they thought is correct. The rate at which the algorithm fools the participants is recorded.
2. Fully-Connected Network (FCN) score:
First apply FCN to the image and then compute the semantic segmentation metrics by comparing with the ground-truth label.
3. Semantic segmentation metrics:
Three metrics are used: per-pixel accuracy, mean class IoU and per-class accuracy.

$$\text{Per-pixel accuracy:} \quad \frac{\text{The number of corrected labled pixels}}{\text{The total number of pixels}}$$

Per-class accuracy: The per-pixel accuracy for each class.

$$\text{Mean class IoU:} \quad \frac{\text{The Intersection of Pixel with the same label}}{\text{The Union of Pixel with the same label}}$$

4 Method--Evaluation

Compare our approach against recent methods for unpaired image-to-image translation on paired datasets.

- Baselines

1. CoGAN:
The method learns two generators. The two generators share the weights for the first few layers so they can learn similar latent representation.

2. Pixel loss + GAN:
Requires the generated image to be similar to the input images by adding a pixel-wise identity loss.

3. Feature loss + GAN:
Similar to Pixel loss + GAN, but replaces the pixel-wise identity loss with a perceptual loss, which is the L1 distance in the deep learning feature space.

4. BiGAN:
Jointly learn two generator so that the joint distribution of the input domain and output domain can be similar. Please refer to the original paper for more detail.

5. pix2pix:
Directly trained with paired image. The results should be better than all the other methods listed above. Used as an upper bound.

4 Method--Evaluation

task	Avg. 'realness' score			
	DualGAN	cGAN[3]	GAN	ground-truth
sketch → photo	**1.78**	1.64	1.07	3.61
day → night	**2.37**	1.93	0.14	3.02
label → facades	1.90	**2.65**	1.40	3.34
maps → aerial photo	2.55	**2.91**	1.89	3.17

Table 1:
The average AMT score of outputs of various tasks.

	Per-pixel acc.	Per-class acc.	Class IOU
DualGAN	0.27	0.13	0.06
cGAN [3]	**0.54**	**0.33**	**0.19**
GAN	0.22	0.10	0.05

Table 2:
The segmentation accuracy for facades→architecture label task.

未来媒体研究中心 CENTER FOR FUTURE MEDIA | 电子科技大学 University of Electronic Science and Technology of China

4 Method--Evaluation

Loss	Map → Photo % Turkers labeled *real*	Photo → Map % Turkers labeled *real*
CoGAN [27]	0.6% ± 0.5%	0.9% ± 0.5%
BiGAN [6, 5]	2.1% ± 1.0%	1.9% ± 0.9%
Pixel loss + GAN [41]	0.7% ± 0.5%	2.6% ± 1.1%
Feature loss + GAN	1.2% ± 0.6%	0.3% ± 0.2%
CycleGAN (ours)	**26.8% ± 2.8%**	**23.2% ± 3.4%**

Table 3:
AMT "real vs fake" test on maps→aerial photos.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
CoGAN [27]	0.40	0.10	0.06
BiGAN [6, 5]	0.19	0.06	0.02
Pixel loss + GAN [41]	0.20	0.10	0.0
Feature loss + GAN	0.07	0.04	0.01
CycleGAN (ours)	**0.52**	**0.17**	**0.11**
pix2pix [18]	0.71	0.25	0.18

Table 4:
FCN-scores for different methods, evaluated on Cityscapes labels→photos.

4 Method--Experiment

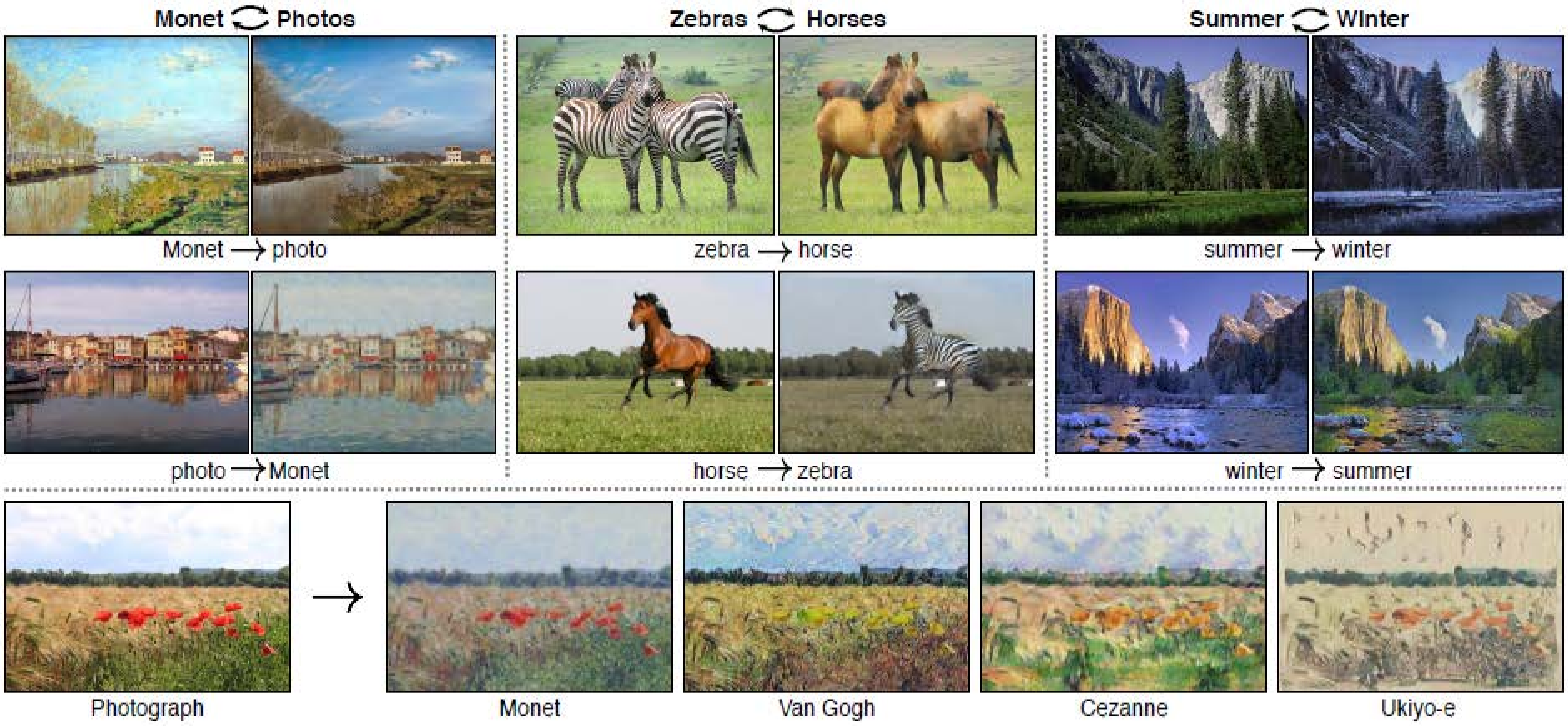


Figure1：Given any two unordered image collections X and Y , our algorithm learns to automatically "translate" an image from one into the other and vice versa.

4 Method--Experiment

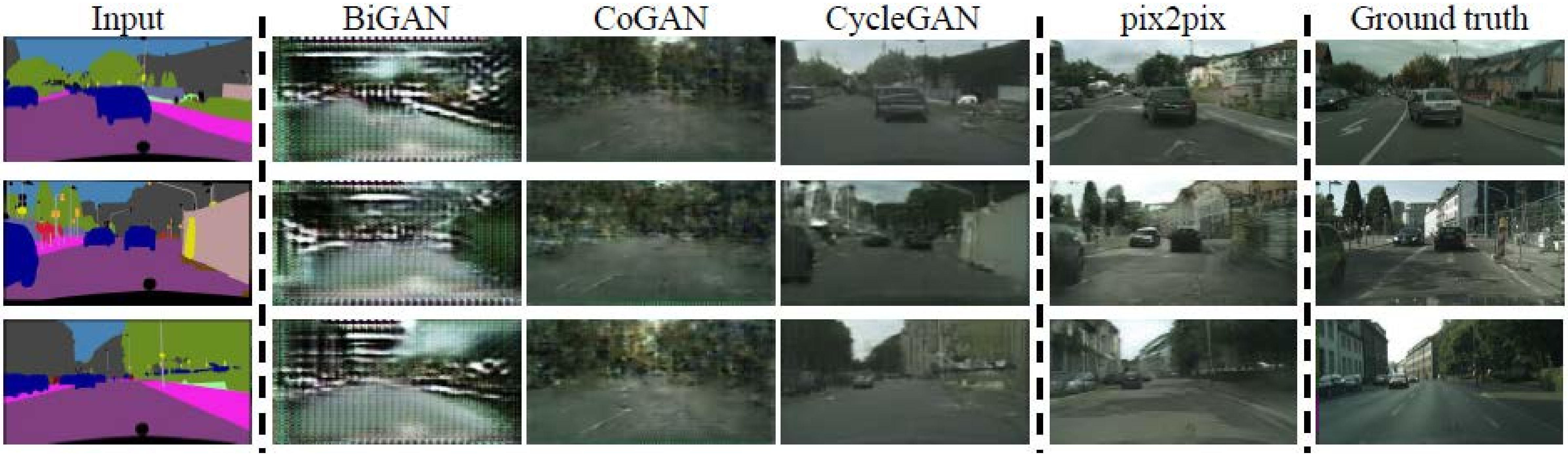


Figure2： Different methods for mapping labels→photos trained on cityscapes.

5 Conclusion

Dual – “supervised”、 “unsupervised”

Basic model：

pix2pix , conv+residual block+deconv , DCGAN

Q&A