本科生《计算机视觉》
基于深度学习的视觉理解与生成

黄雷
人工智能研究院
huangleiAI@buaa.edu.cn

2023年10月26日

主要内容

- 深度学习基础
  - 神经网络及反向传播算法
  - 卷积神经网络中的视觉表示思想
- 视觉理解任务
  - 目标检测
  - 分割
- 视觉生成
  - 深度生成模型
  - 图像翻译任务详解
- 深度神经网络训练技巧

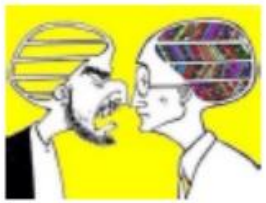
Outline

- Neural Style Transfer
- Image-to-Image Translation

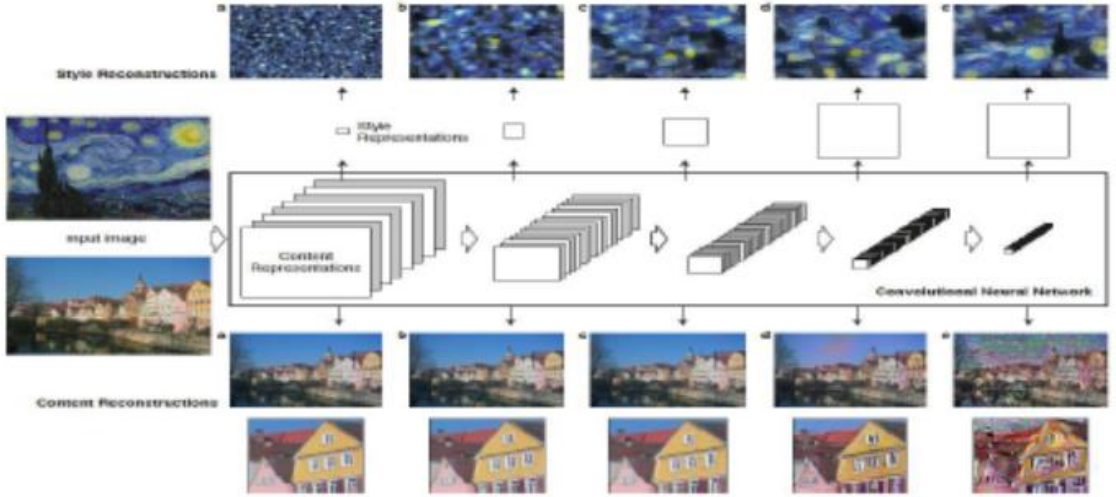
Image Style Transfer

- Content: Global structure. Style: Colours; local structures
  Like naturalistic, photographic, abstract, symbolistic



- Use CNNs to capture style from one image and content from another image.

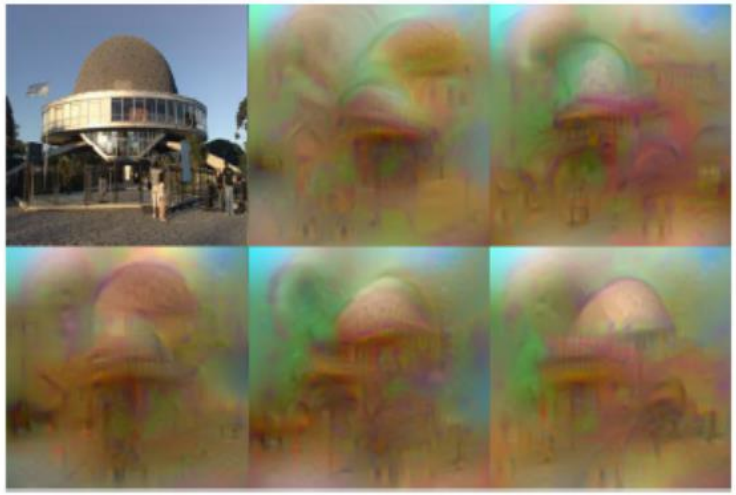


- Feature representations → Content
  Filter correlations → Style

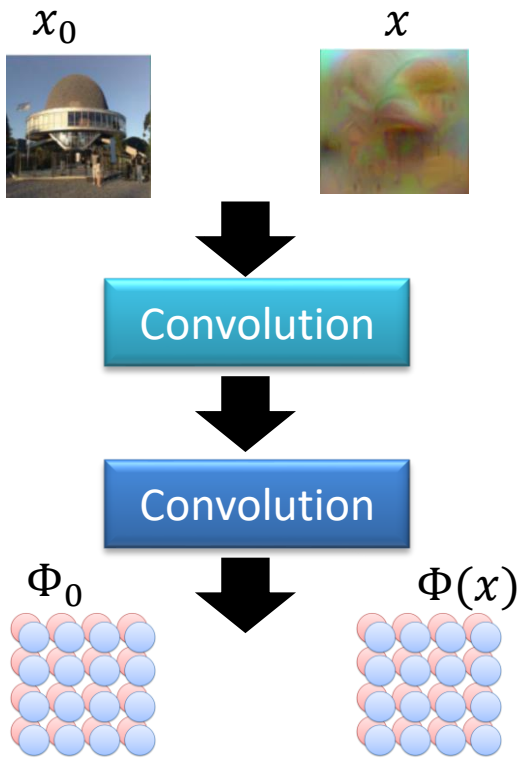
Reconstructing an image from a convolutional layer

- Representation function: $\Phi : \Re^{H \times W \times C} \rightarrow \Re^{d}$ (image space to feature space)
- Target Representation: $\Phi_0 = \Phi(x_0)$ (x0 is the original image)
- We need to find: $x \in \Re^{H \times W \times C}$ by minimizing:

$$x^* = \arg\min_{x \in \Re^{H \times W \times C}} l(\Phi(x), \Phi_0) + \lambda R(x)$$



"Understanding Deep Image Representations by Inverting Them", by Aravindh Mahendran and Andrea Vedaldi.

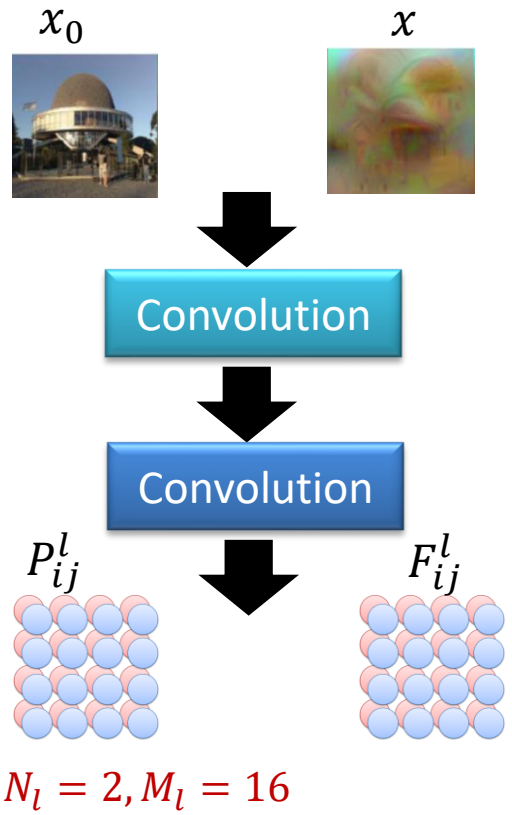


Content Loss Function

- Filters (Depths) at layer l: N_l
- The height times the width of the feature map at layer l: M_l
- Response at layer l: $F_l \in \Re^{N_l \times M_l}$

  $F^l{}_{ij}$ represents the ith filter at position j in layer l
- Original image: $\vec{p}$
- We generate image: $\vec{x}$ (randomly initialized)
- Squared-error loss:

$$L_{content} = \frac{1}{2}\sum_{i,j}(F^l{}_{ij} - P^l{}_{ij})^2$$



Style Loss Function

- Filter correlations are given by the Gram matrix:

$$G^l \in \Re^{N_l \times N_l}$$

- $G^l{}_{ij}$ is the inner product between the filters i and j in layer l:

$$G^l{}_{ij} = \sum_k F^l{}_{ik} F^l{}_{jk}$$

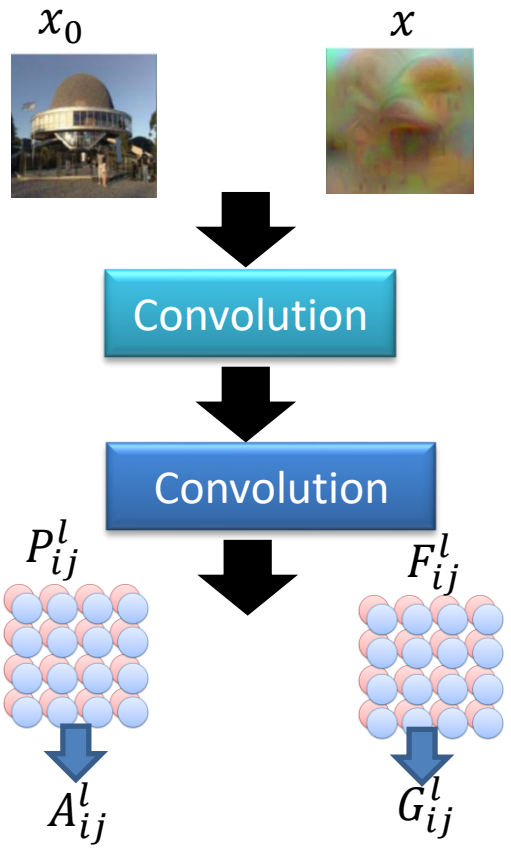
- The loss at layer l:

$$E_l = \frac{1}{4N^2{}_l M^2{}_l} \sum_{i,j} (G^l{}_{ij} - A^l{}_{ij})^2$$

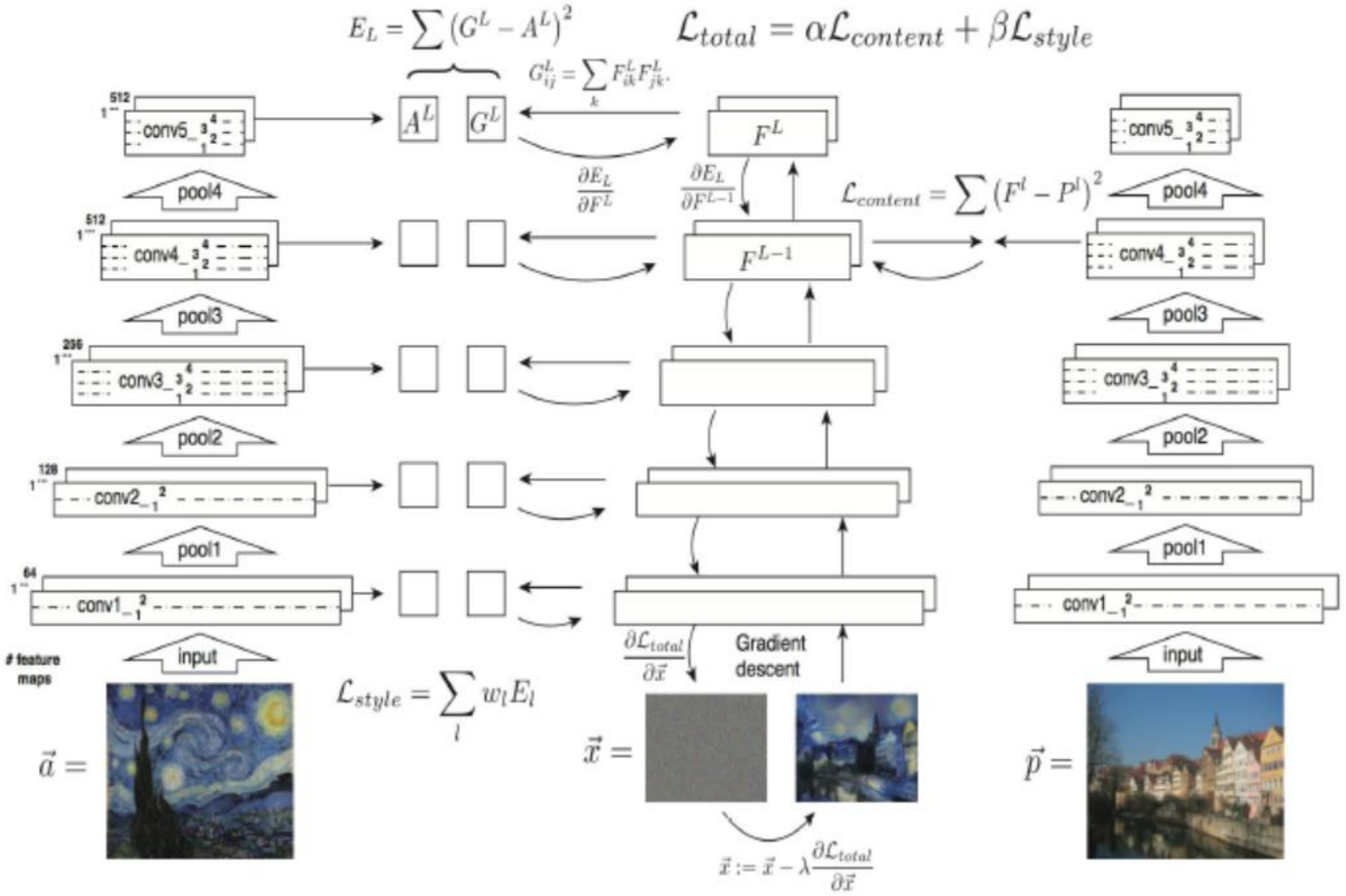
A <-> original image
G <-> generated image

- The total style loss:

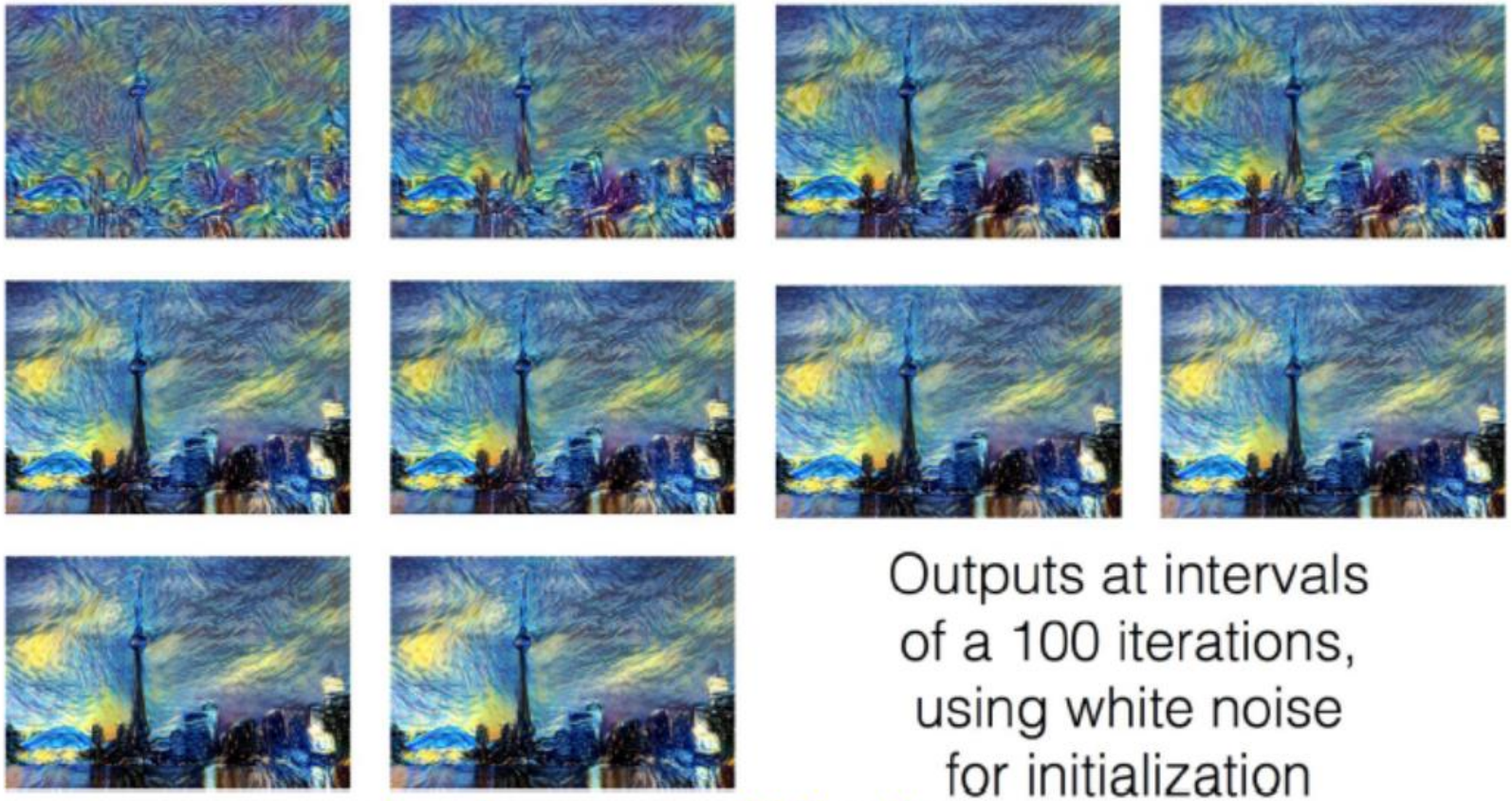
$$L_{style} = \sum_{l=0}^{L} w_l E_l$$



The Total Loss Function

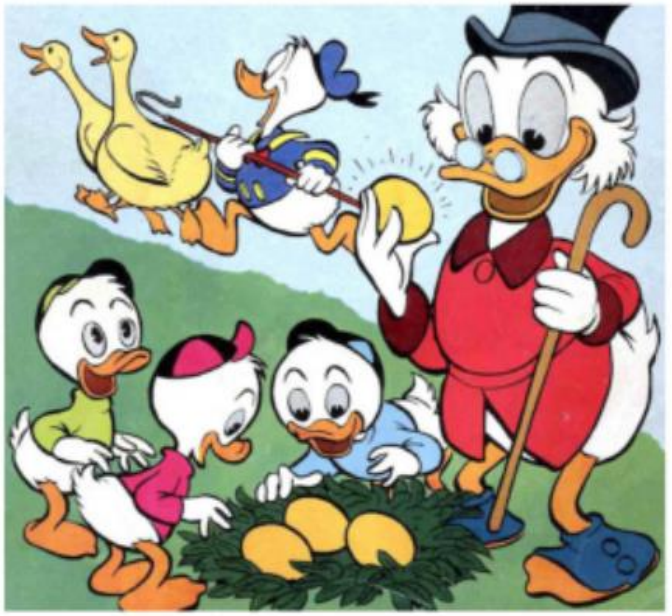


Results



Outputs at intervals of a 100 iterations, using white noise for initialization

show image every 10 iterations



The optimization-based methods

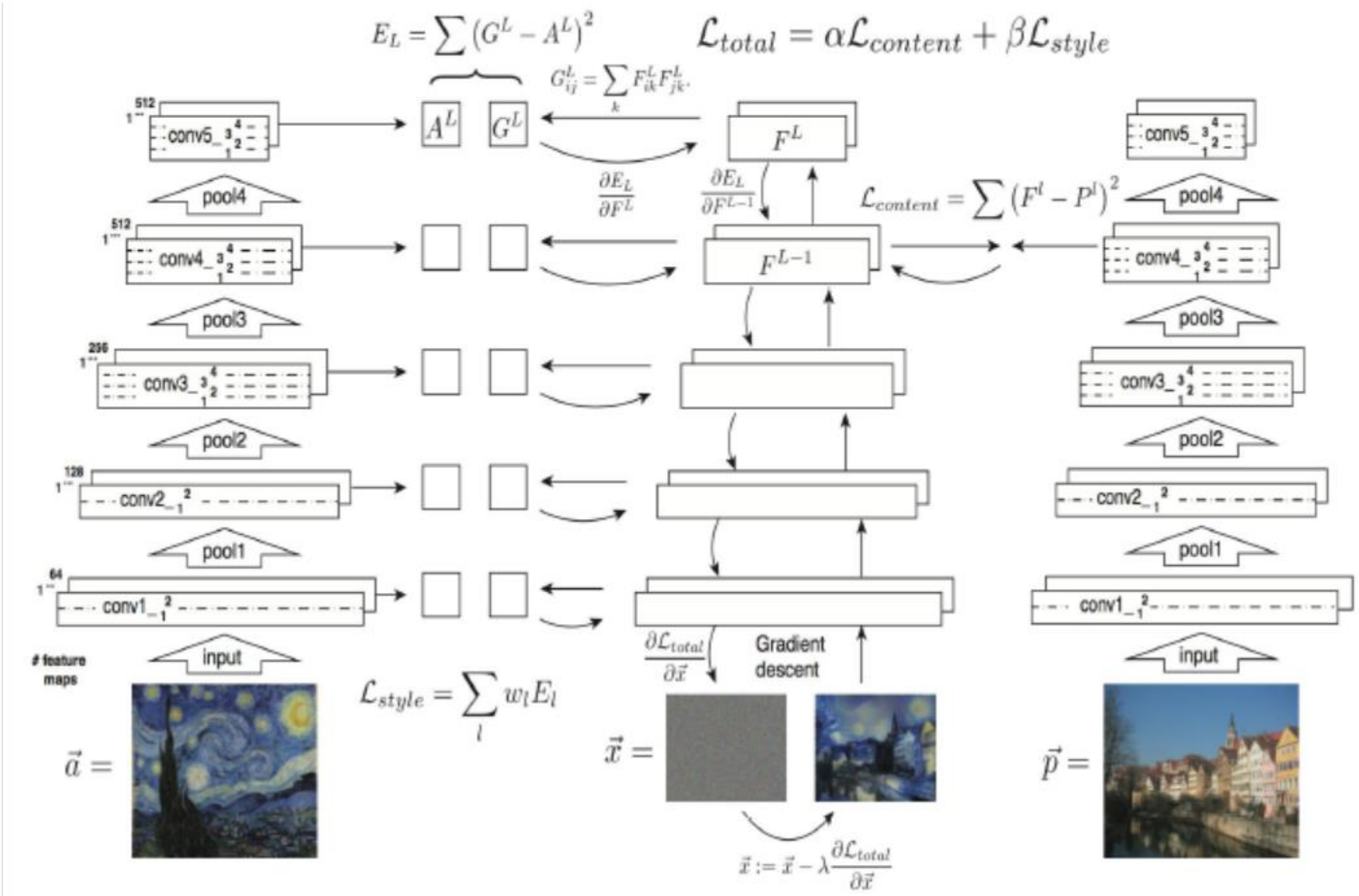
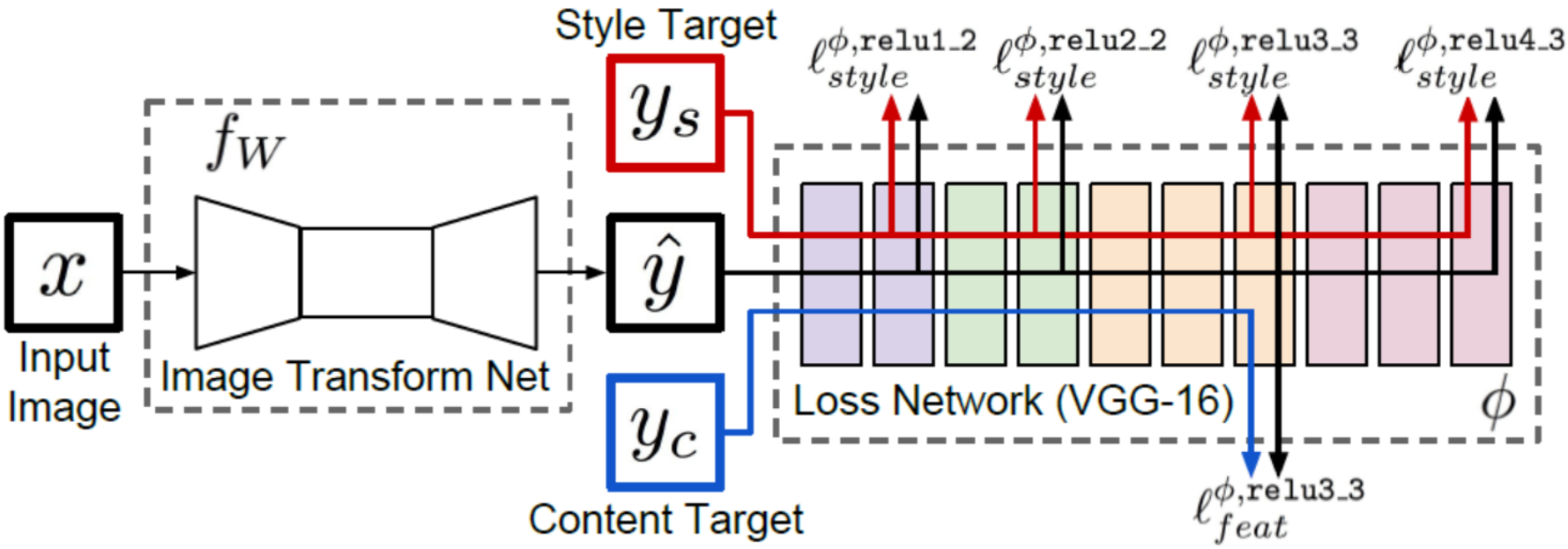


Image Style Transfer

- Fast inference based methods



Johnson et al, ECCV 2016

Image Style Transfer

- Condition Instance Normalization (Dumoulin et al, ICLR 2017)

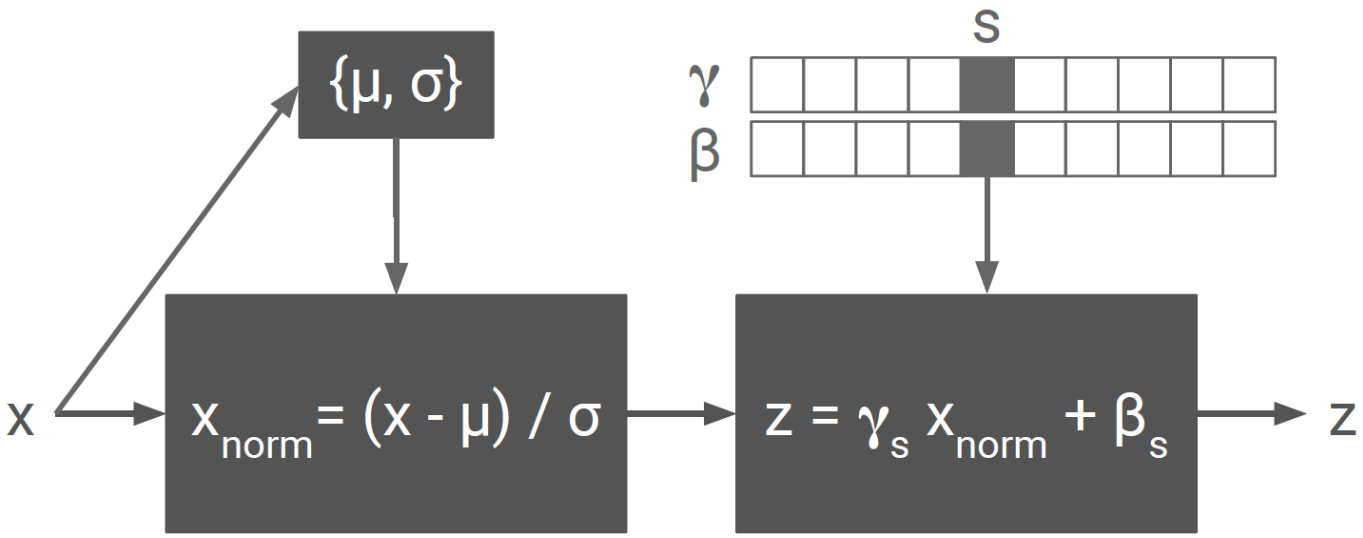
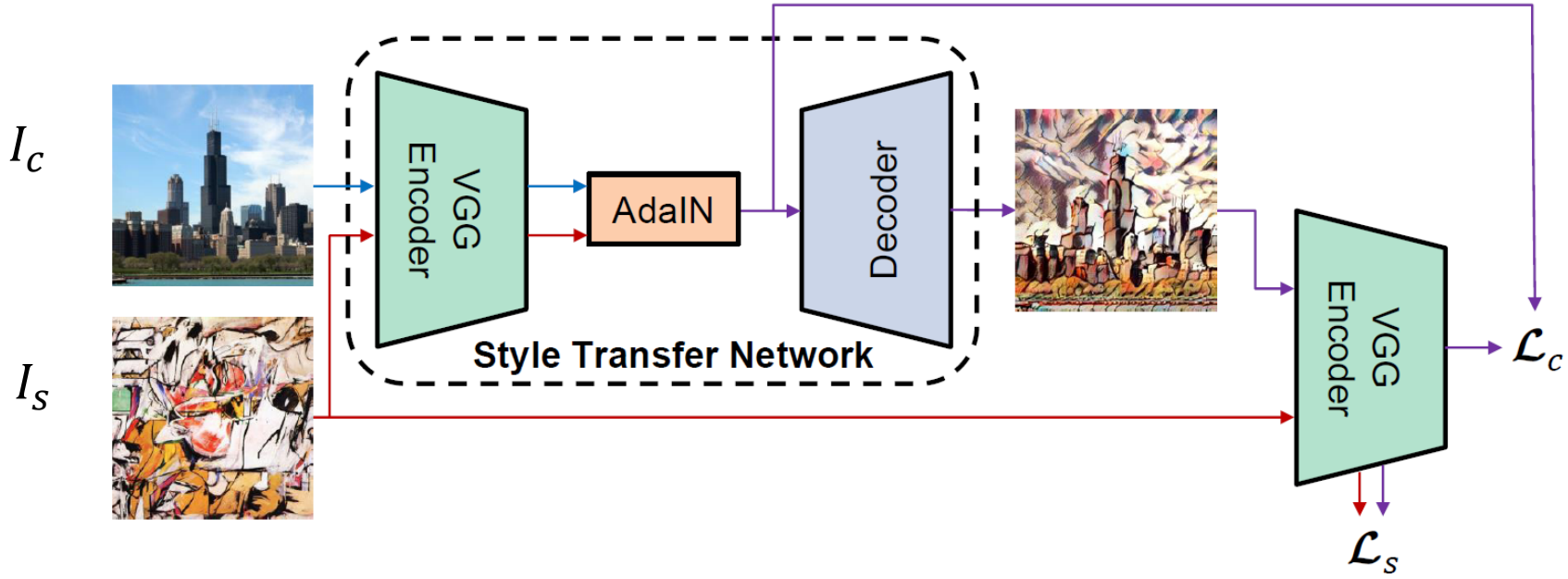


Image Style Transfer

- Adaptive instance Normalization (Huang et al, ICCV 2017)



$$AdaIN(I_c, I_s) = \sigma(I_s) * \frac{I_c - \mu(I_c)}{\sigma(I_c)} + \mu(I_s)$$

Outline

- Neural Style Transfer
- Image-to-Image Translation

Image-to-Image Translation

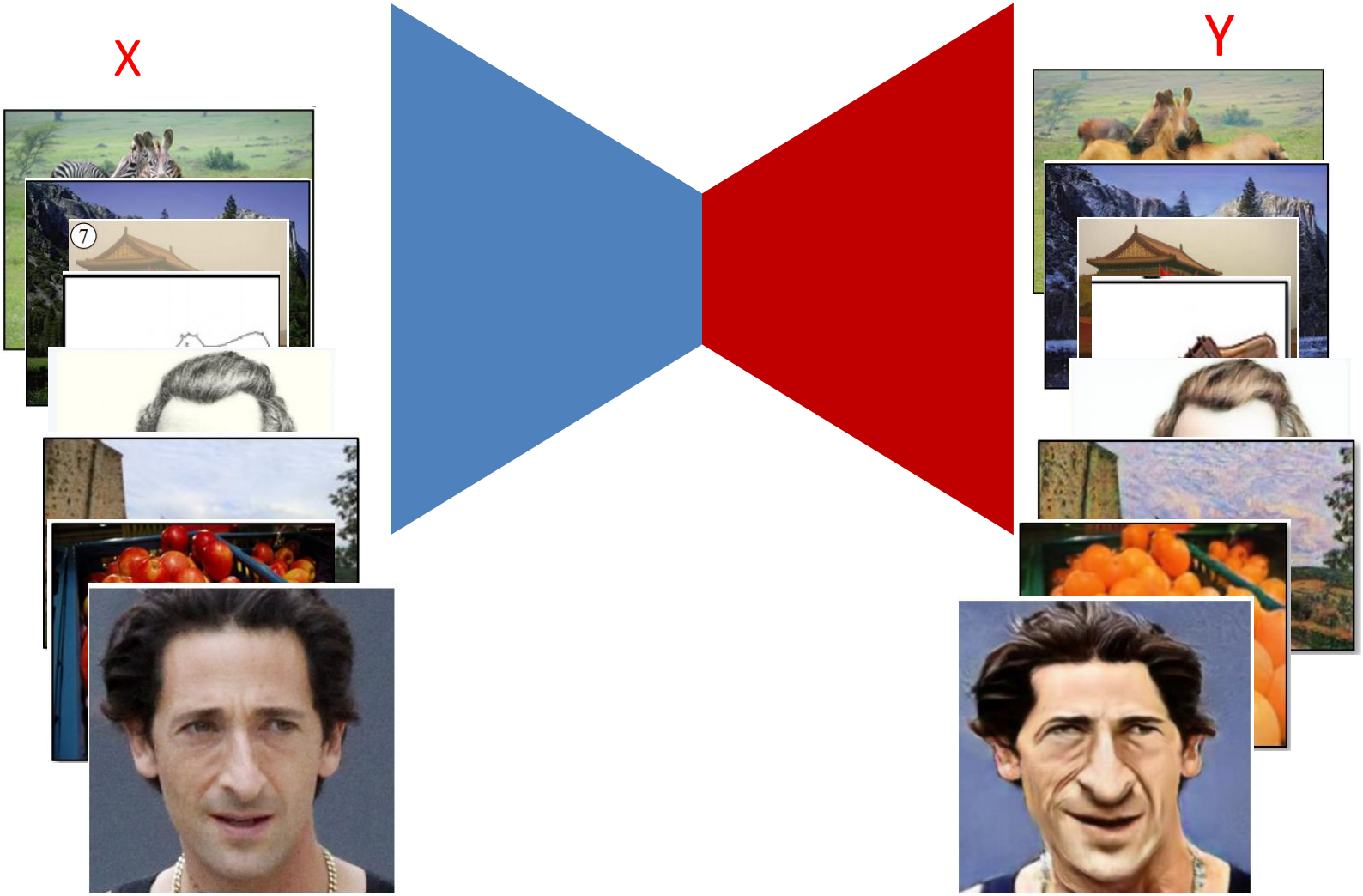


Image-to-Image Translation

Problem Definition

- Supervised/Paired image-to-image translation
- Unsupervised/Unpaired image-to-image translation

Image Reconstruction/Super Resolution

- Supervised image super resolution

Better feature reconstruction

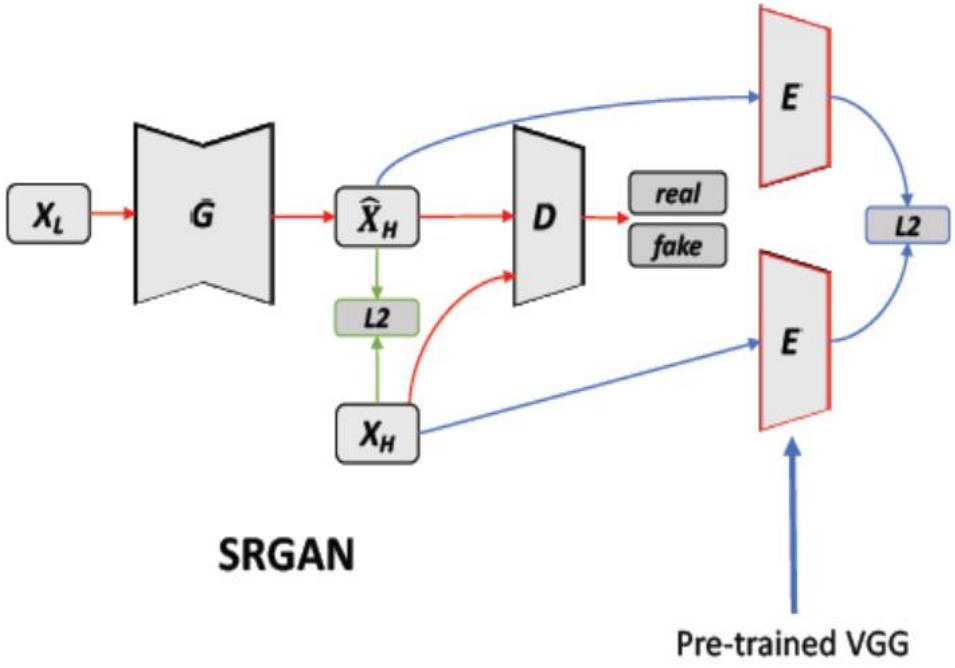


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *C. Ledig, L. Theis et al. CVPR 2017.*

Image Reconstruction/Super Resolution

- Supervised image super resolution

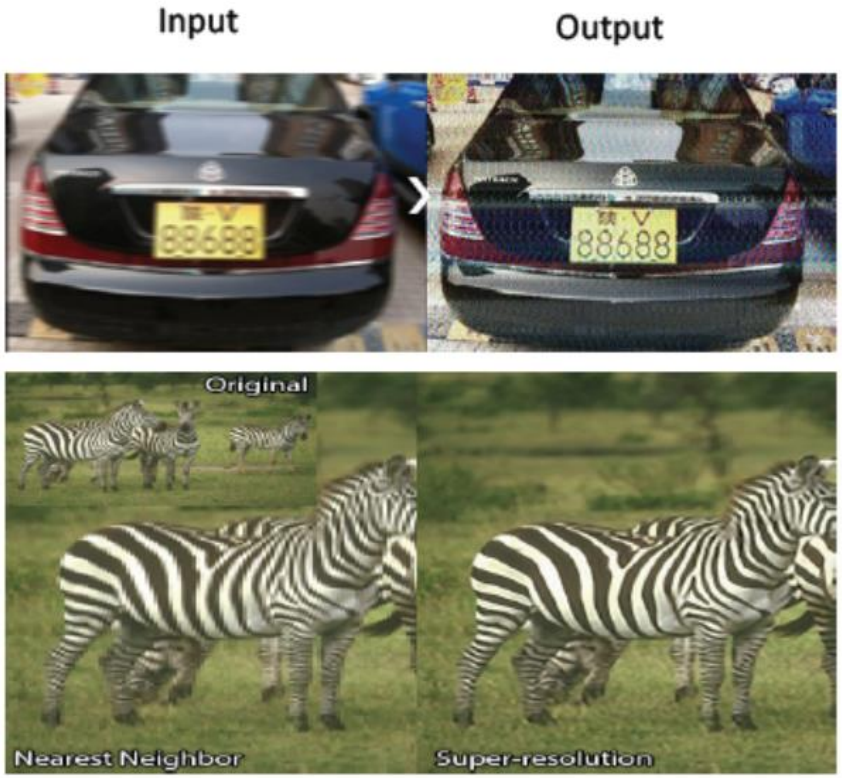
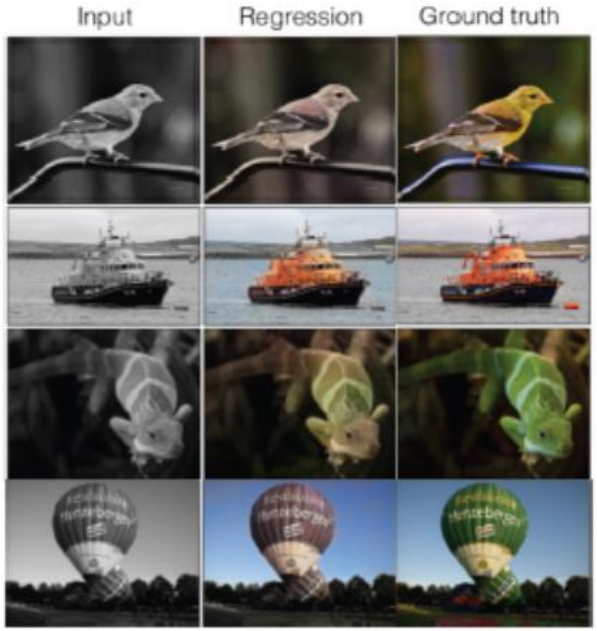
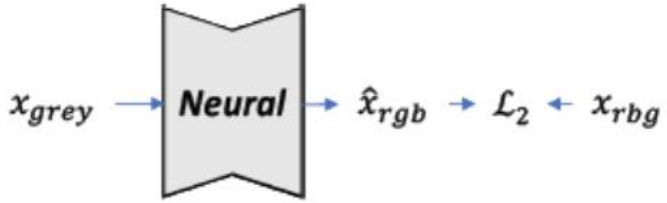


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *C. Ledig, L. Theis et al. CVPR 2017.*

Pix2Pix: paired data

- Pix2Pix: Supervised Image-to-Image Translation
  - Beyond MLE: Adversarial Learning





Different colors will have conflicts,

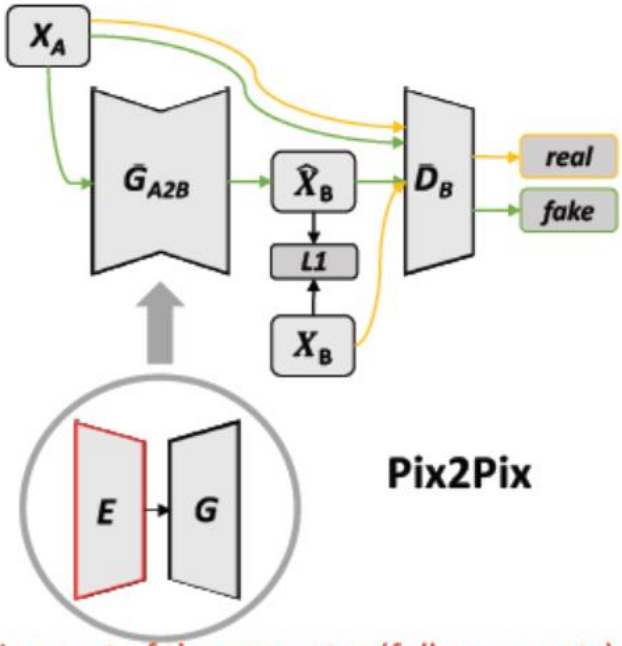
(some want red, some want blue, …)

resulting "grey" outputs

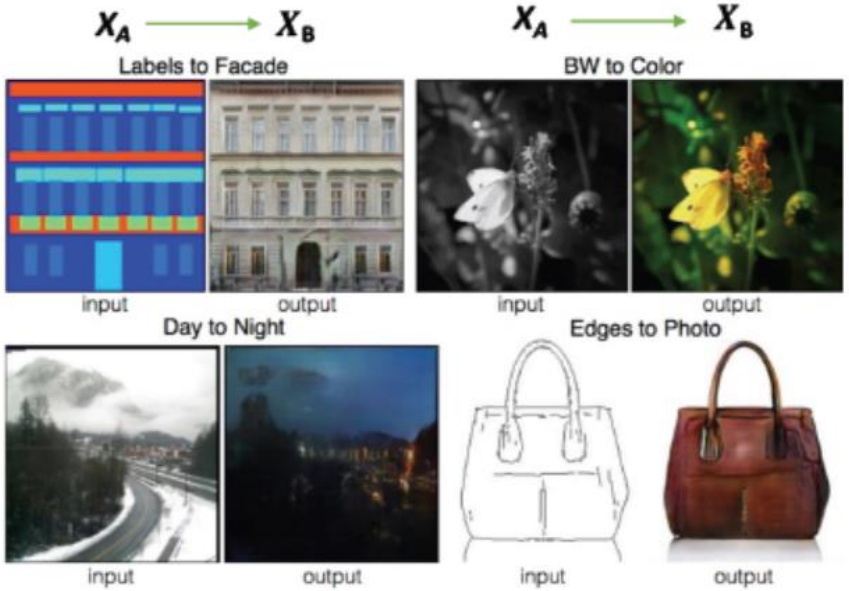
Colorful Image Colorization. *R. Zhang, P. Isola, A.A. Efros. ECCV. 2016.*
Image-to-Image Translation with Conditional Adversarial Networks. *P. Isola, J. Zhu et al. CVPR 2017.*

Pix2Pix: paired data

- Pix2Pix: Supervised Image-to-Image Translation
  - Beyond MLE: Adversarial Learning



Pix2Pix



$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{data}}[logD(x_A, x_B)] + \mathbb{E}_{x\sim p_{data}}[log(1 - D(x_A, G(x_A)))]$$

$$\mathcal{L}_G = \mathbb{E}_{x\sim p_{data}}[logD(x_A, G(x_A))]$$

Encoder is a part of the generator (fully conv nets)

Image-to-Image Translation with Conditional Adversarial Networks. *P. Isola, J. Zhu et al. CVPR 2017.*

Pix2Pix: paired data

- Pix2Pix: Supervised Image-to-Image Translation

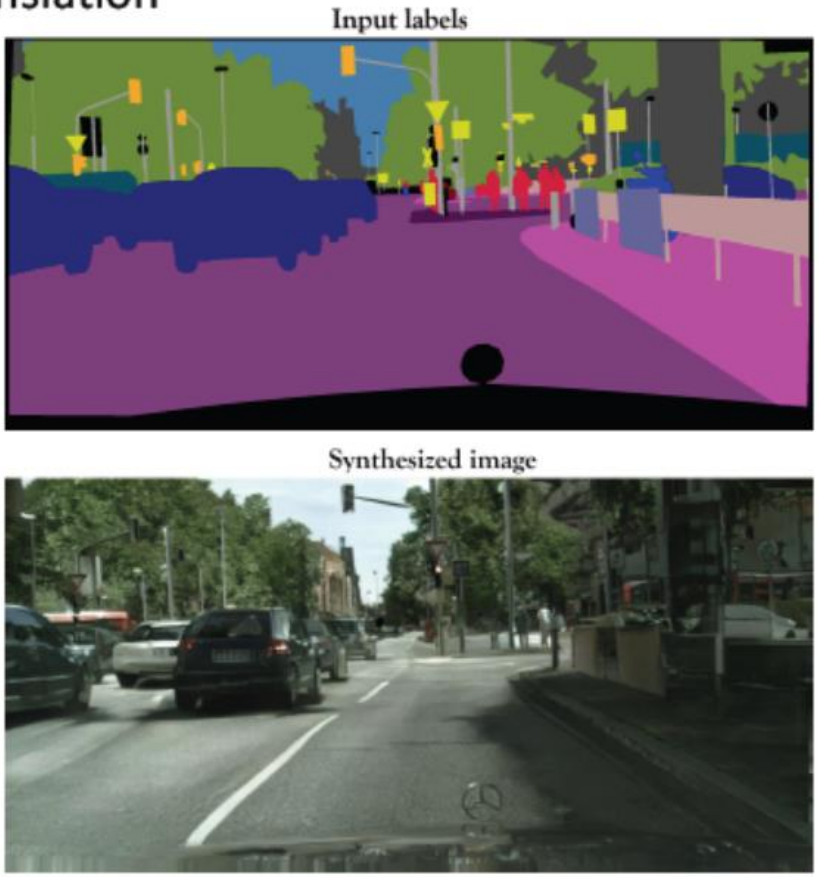
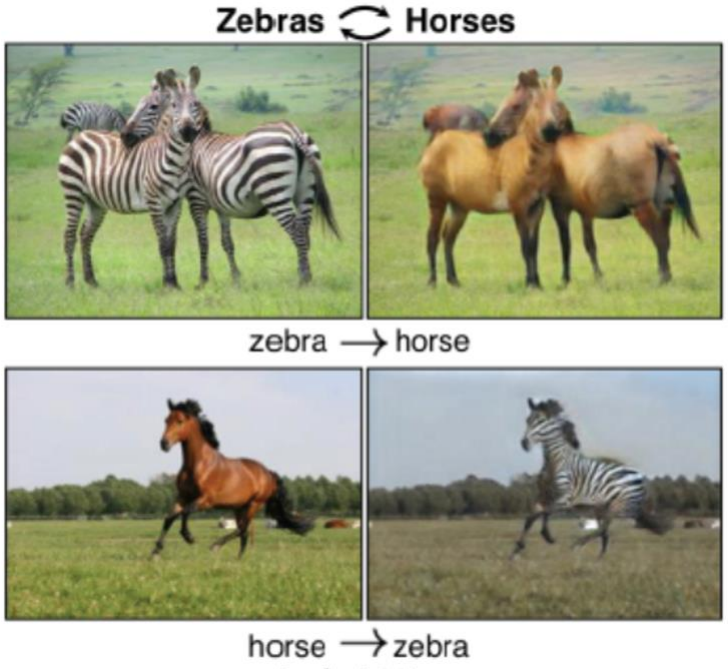
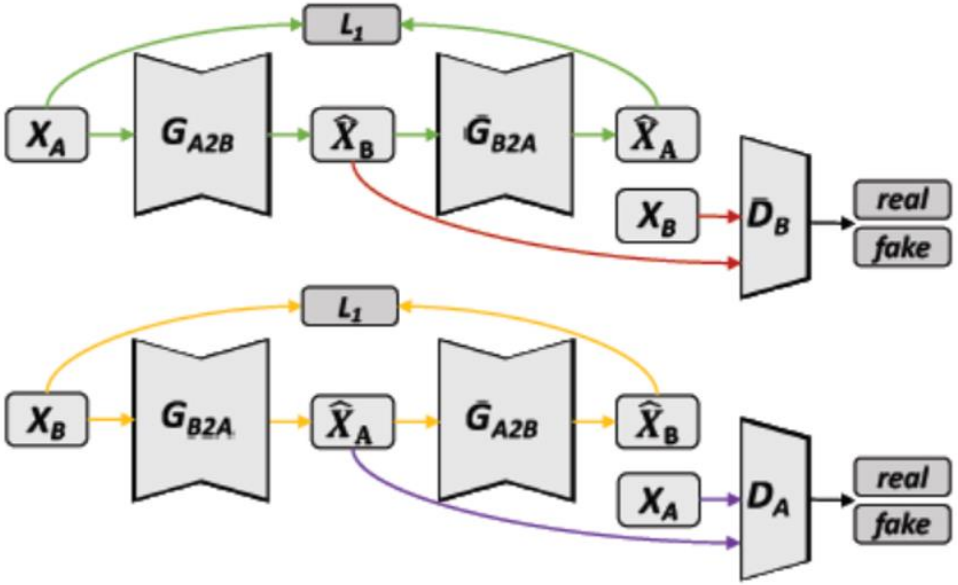


Image-to-Image Translation with Conditional Adversarial Networks. *P. Isola, J. Zhu et al. CVPR 2017.*

GAN with Encoder—Unsupervised Image-to-Image Translation

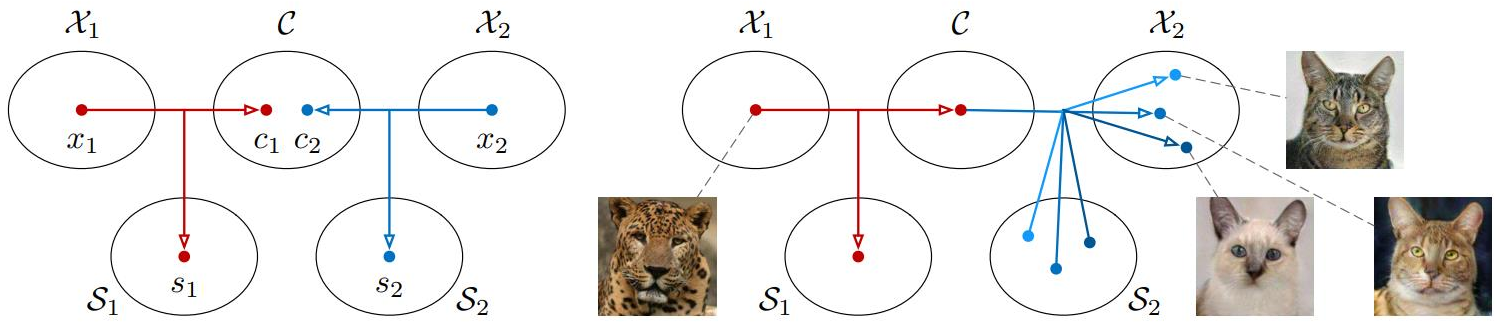
CycleGAN: Unpaired Image-to-Image Translation





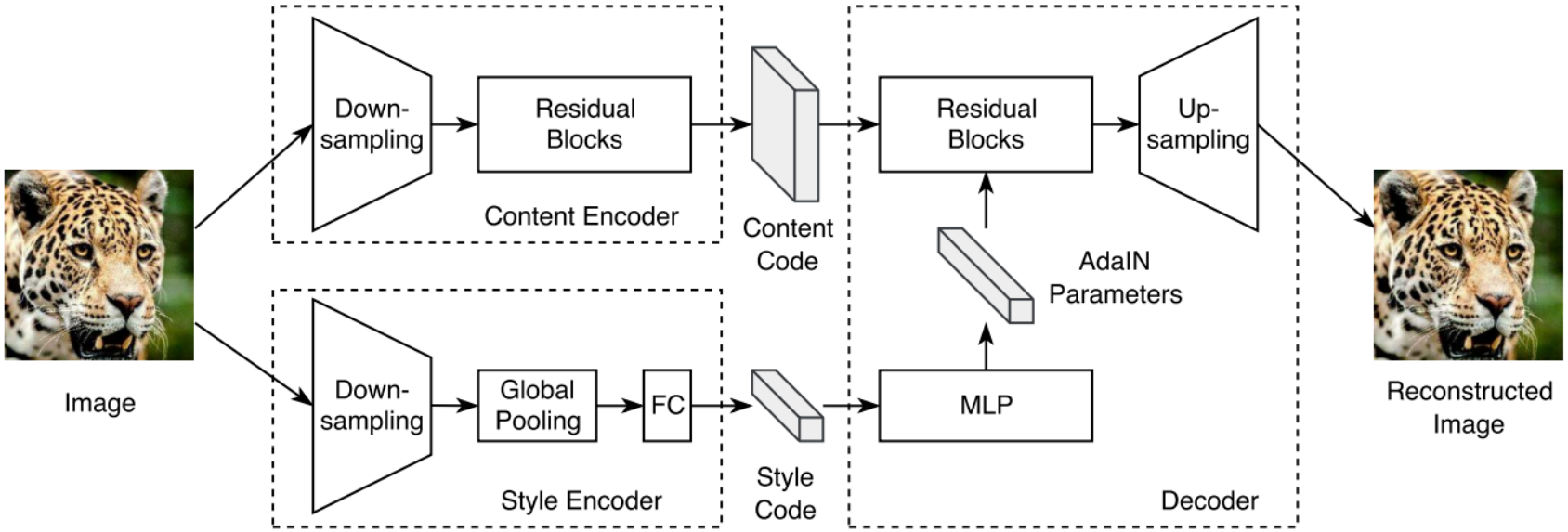
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *J. Zhu, T. Park et al. ICCV 2017.*

MUNIT : unpaired+multi-modal



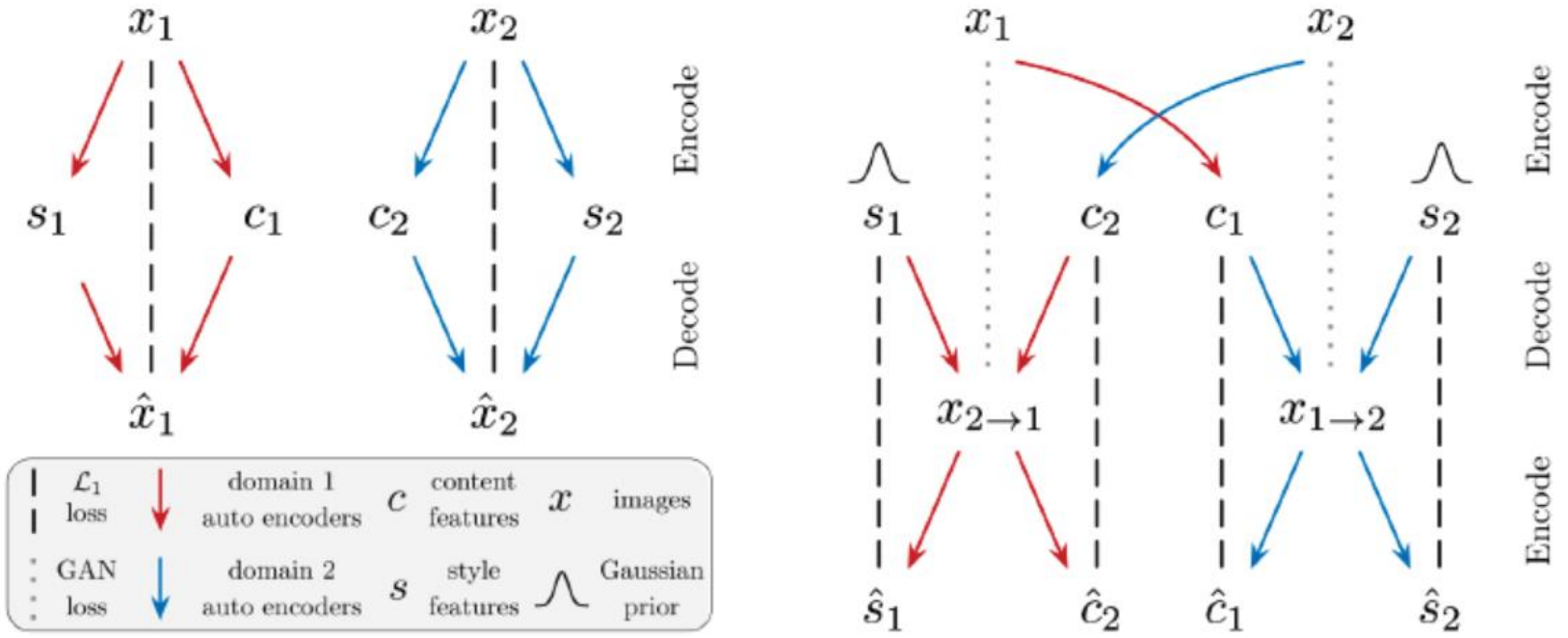
(a) Auto-encoding

(b) Translation



MUNIT : unpaired+multi-modal

- Latent reconstruction + Adversarial learning



(a) Within-domain reconstruction

(b) Cross-domain translation

MUNIT: Multimodal Unsupervised Image-to-Image Translation. ECCV 2018.

MUNIT : unpaired+multi-modal

- Goal: unpaired + multi-modal results

MUNIT: Multimodal Unsupervised Image-to-Image Translation. ECCV 2018.