GAN-based Image Inpainting

xyfJASON

CONTENTS

1 INTRODUCTION

INTRODUCTION

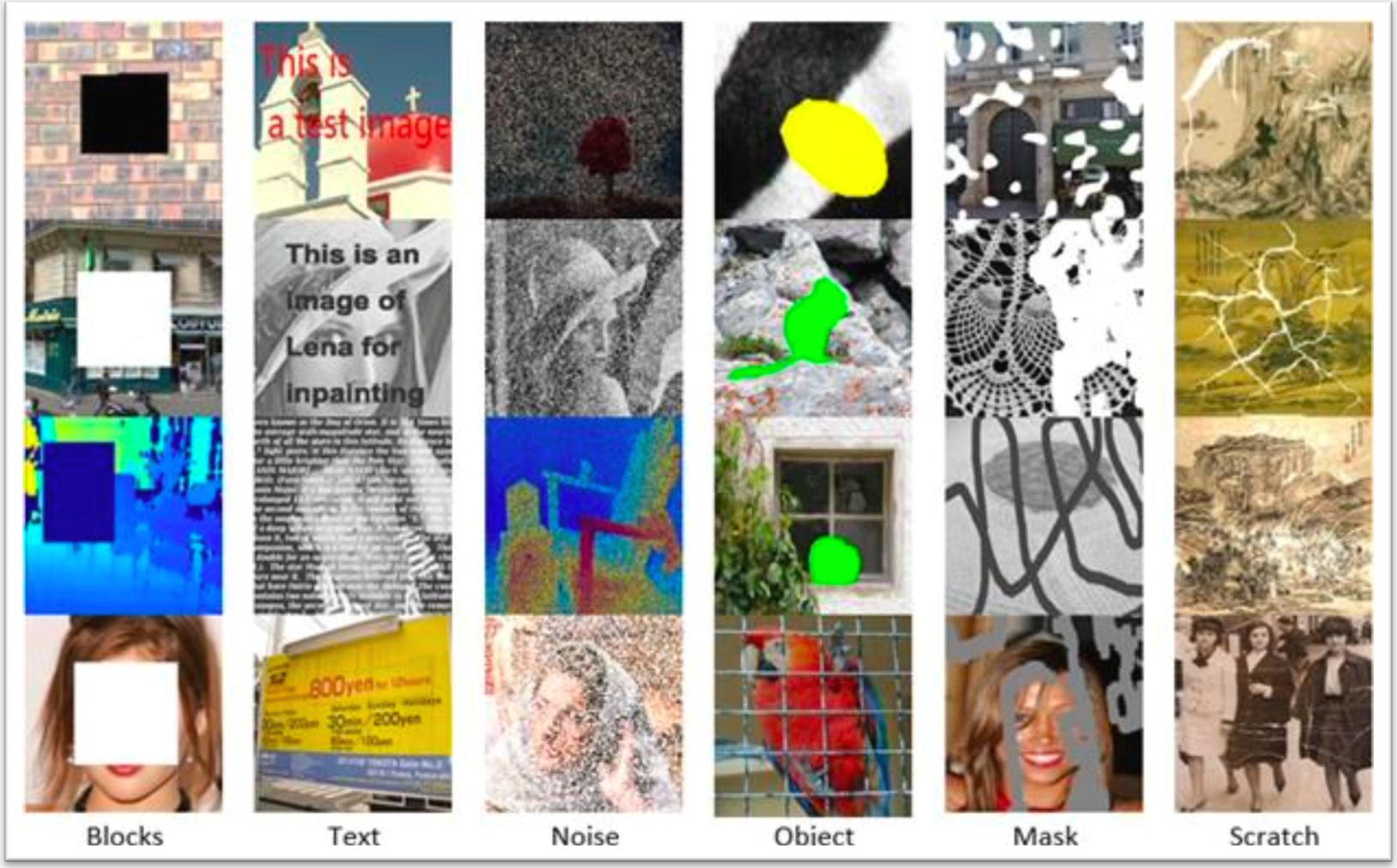


Image source: Elharrouss, Omar, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. "Image inpainting: A review." *Neural Processing Letters* 51, no. 2 (2020): 2007-2028.

INTRODUCTION

Keys

- 如何得到**全局一致性**
  - High-level features, semantic
- 如何得到**局部真实性**
  - Low-level features, textures / shapes, pixel level
- 如何获取远处信息
  - 与卷积操作的局部性相违背
  - Network structure
  - Loss function

Copy-and-paste

DL methods
(Context Encoder)

2 EVALUATION METRICS

EVALUATION METRICS

- Quantitative
  - MSE (Mean Squared Error)
  - PSNR (Peak Signal to Noise Ratio)
  - SSIM (Structural Similarity)
  - Image inpainting 是一个 ill-posed problem，即存在多种合理的填充方式
- Qualitative
- User study

3 MODELS

MODELS – Common things

Architecture

- Generator:
  - Encoder-decoder
  - U-Net
- Discriminator:
  - Global discriminator
  - Local discriminator / PatchGAN

Loss

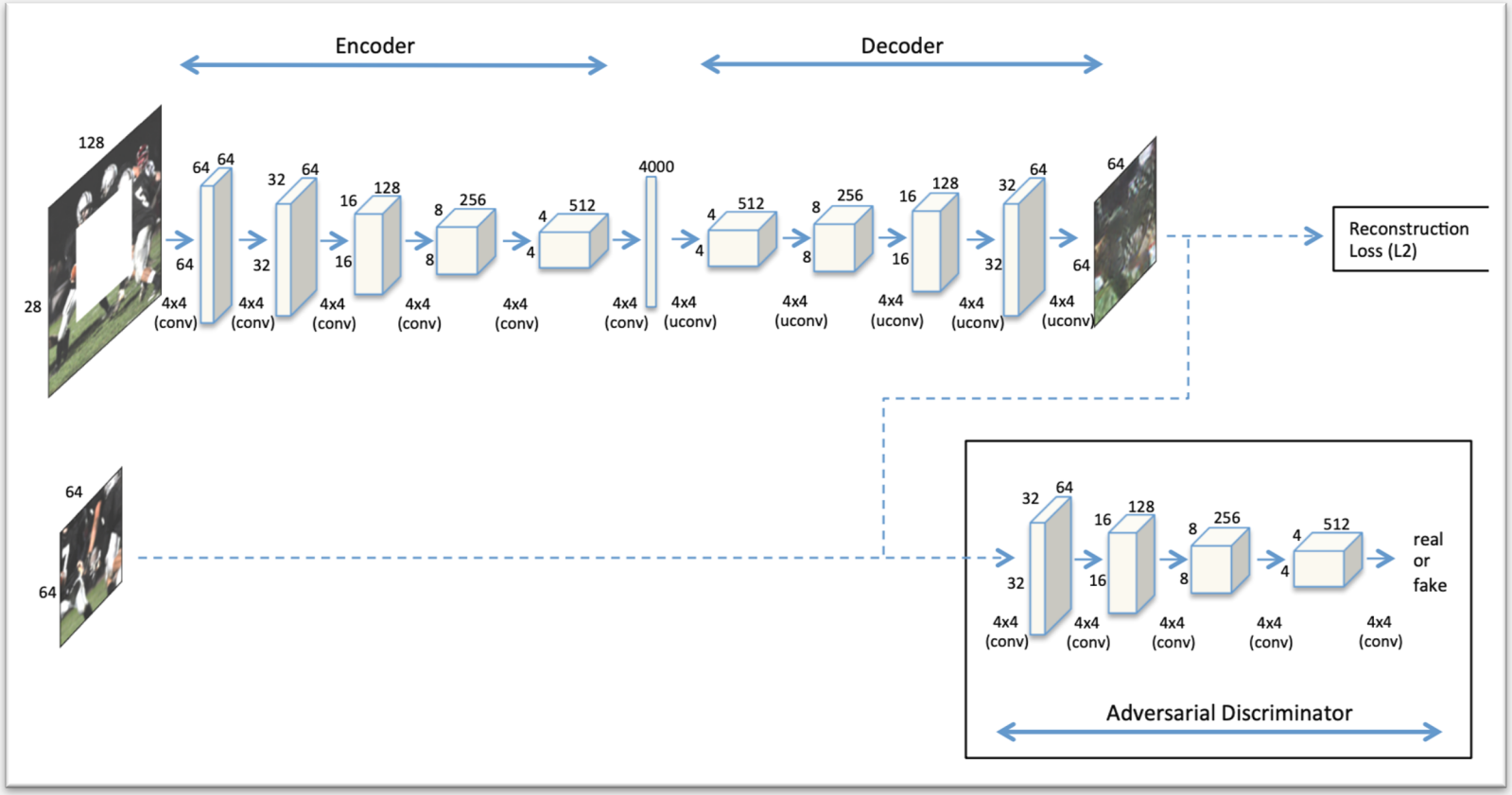
- Reconstruction loss
  - L1 loss → weighted
  - L2 loss
- Adversarial loss (GAN loss)
- Perceptual loss: VGG
- Style loss: Gram matrix
- ID-MRF loss
- TV loss

MODELS: Context Encoder

- Encoder-decoder
- Channel-wise fc
- 探讨 L2 loss 导致模糊的原因：L2 loss 倾向于给出较为平均化的结果，这样平均误差小并且较“安全”

■ 固定 128x128 中的 64x64 缺失区域，不够灵活

■ Channel-wise fc



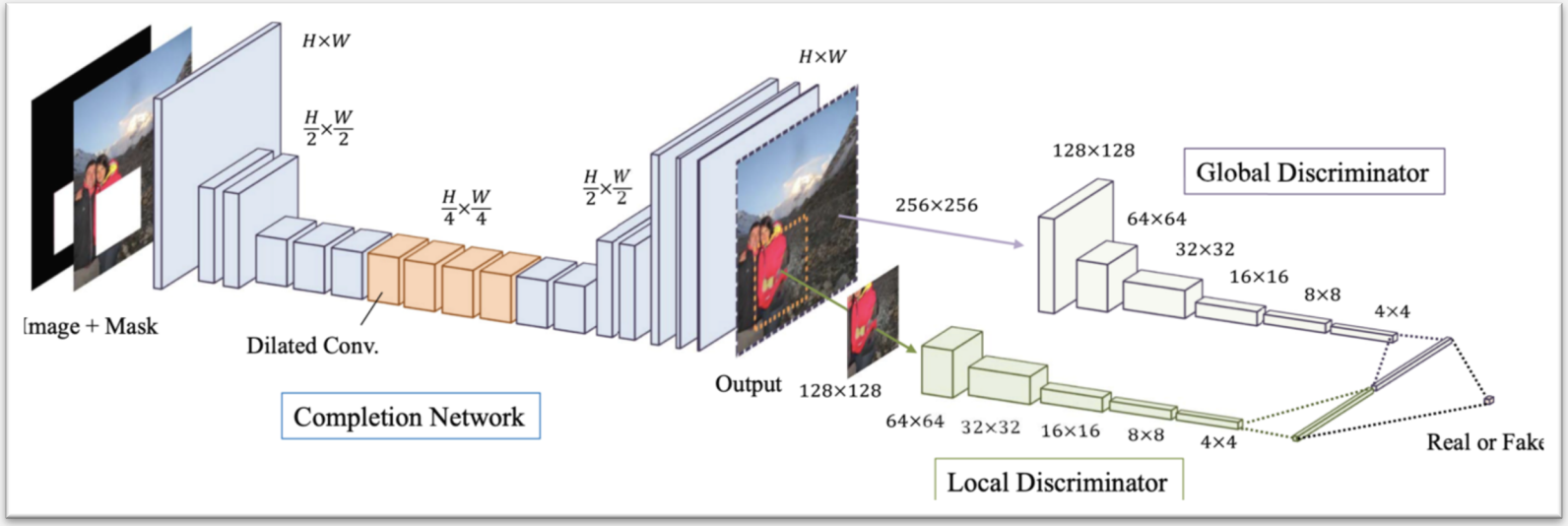
Pathak, Deepak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. "Context encoders: Feature learning by inpainting." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536-2544. 2016.

MODELS: GLCIC (GLGAN)

- FCN with dilated conv
- Global discriminator

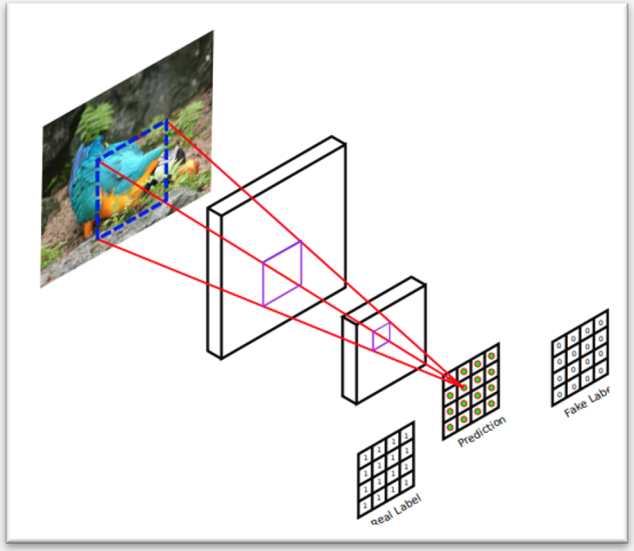
  Local discriminator

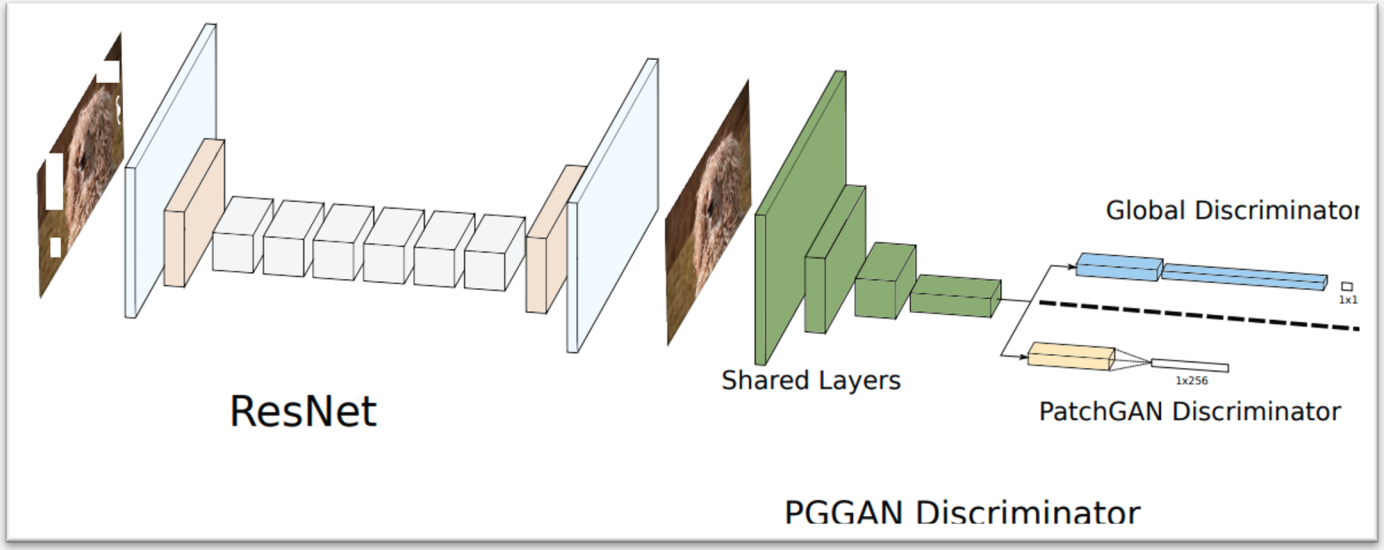
- ■ Local discriminator 只能接受一个缺失区域
- ■ 训练与测试的不一致



Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. "Globally and locally consistent image completion." ACM Transactions on Graphics (ToG) 36, no. 4 (2017): 1-14.

MODELS: PGGAN

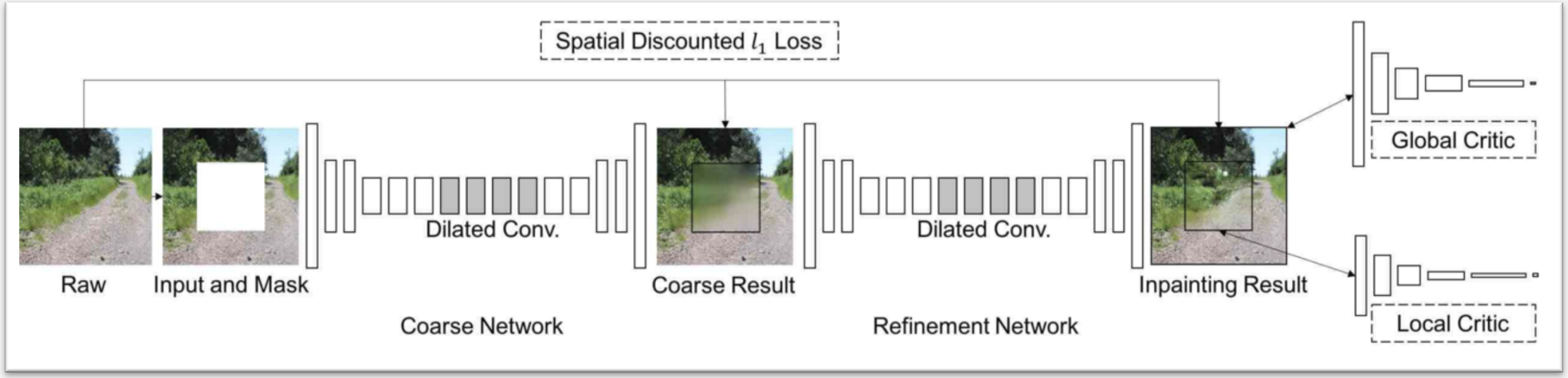




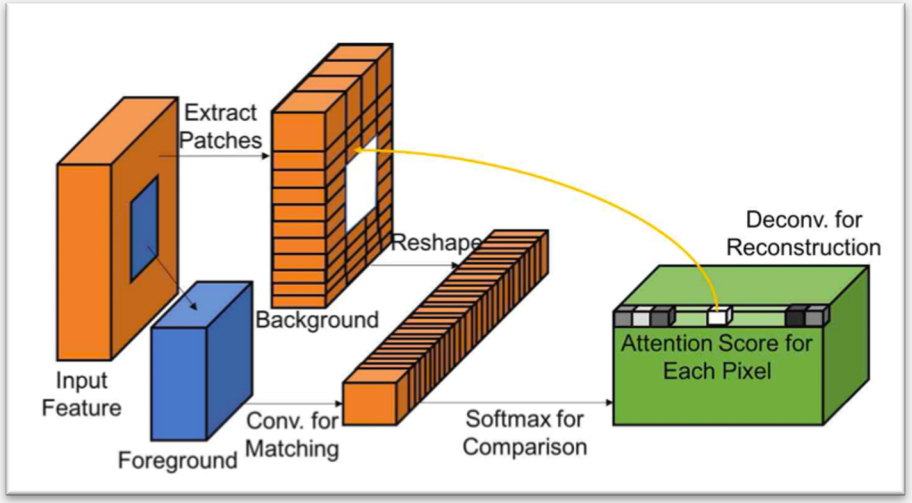
- PatchGAN & GlobalGAN
- Dilated residual block
- Local discriminator → PatchGAN discriminator

Demir, Ugur, and Gozde Unal. "Patch-based image inpainting with generative adversarial networks." arXiv preprint arXiv:1803.07422 (2018).

MODELS: Contextual Attention (DeepFill v1)

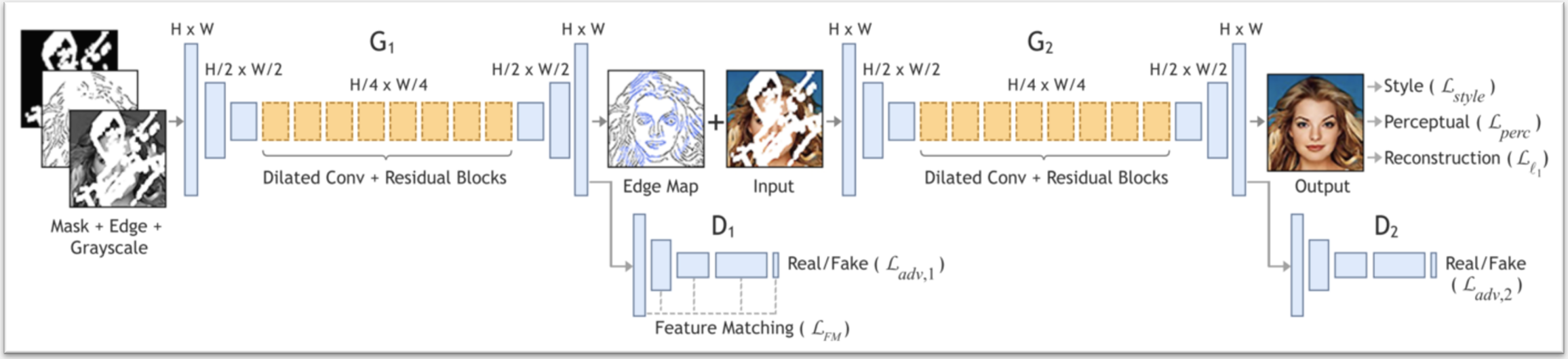


- Contextual Attention
- Spatially discounted L1 loss
- Coase-to-fine two-stage



Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. "Generative image inpainting with contextual attention." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505-5514. 2018.

MODELS: EdgeConnect



- Lines first, color next
- Edge map 是一种先验信息

Nazeri, Kamyar, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. "Edgeconnect: Generative image inpainting with adversarial edge learning." *arXiv preprint arXiv:1901.00212* (2019).

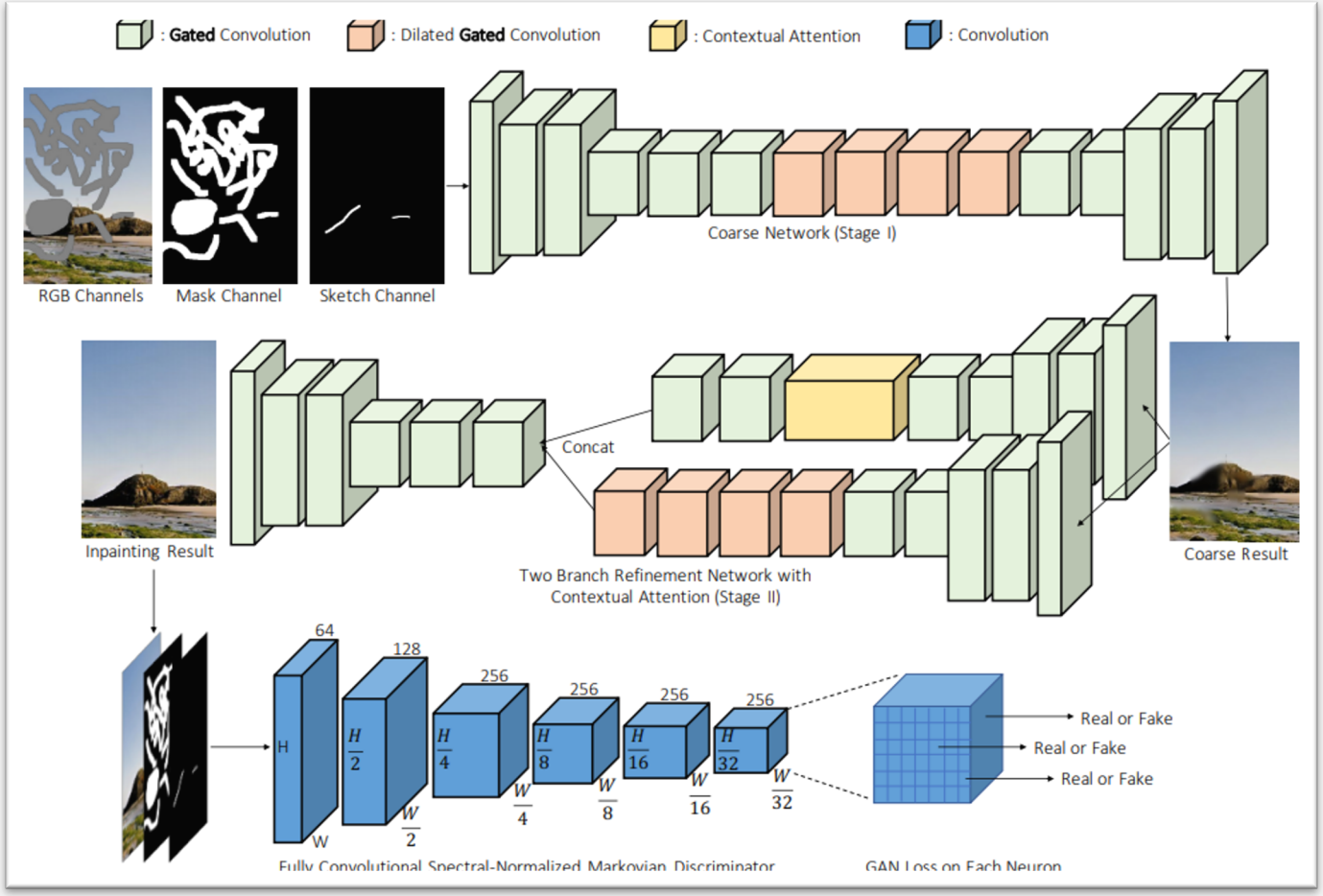
MODELS: DeepFill v2

- Gated convolution
- DeepFill v1 +

  Partial conv +

  EdgeConnect +

  PatchGAN



Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. "Free-form image inpainting with gated convolution."
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471-4480. 2019.

MODELS: DeepFill v2

- Gated convolution
- DeepFill v1 +

  Partial conv +

  EdgeConnect +

  PatchGAN

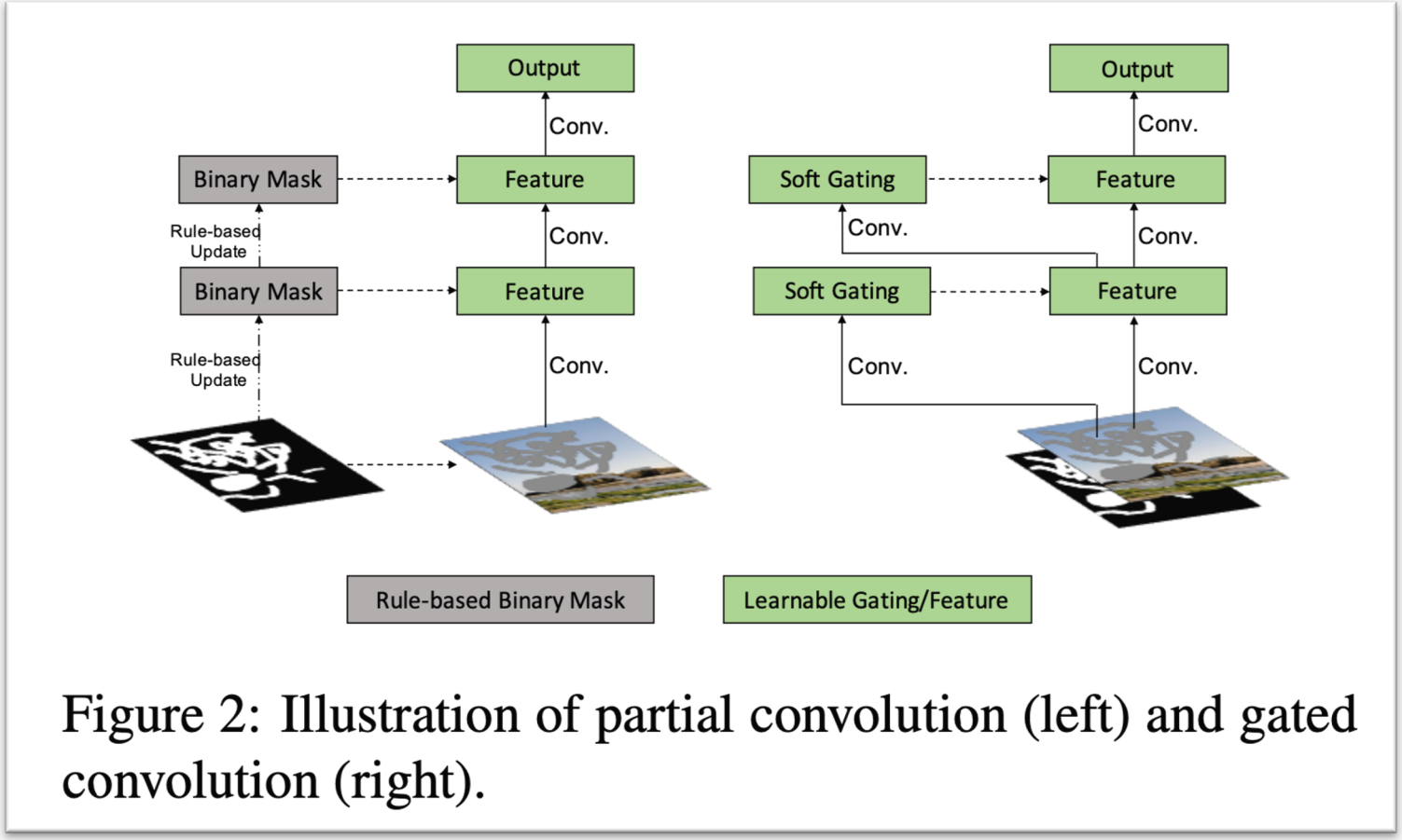


Figure 2: Illustration of partial convolution (left) and gated convolution (right).

Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. "Free-form image inpainting with gated convolution." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471-4480. 2019.

4 BLIND - VCNet

BLIND IMAGE INPAINTING

- Image inpainting 要求输入精确的缺失区域 mask，这导致它的实际应用场景局限在 object removal
- Blind image inpainting 在 **inference** 阶段不要求提供 mask 信息
  - 网络的输入没有 mask

BLIND: VCNet

Introduction

- Highly ill-posed:
  - The unknown degraded regions need to be located based on their **differences** from the intact ones instead of their known characteristics.
- Training data collection:
  - Damage should be **diverse and complicated** enough so that the high-level of structure inconsistency instead of the pattern in damage can be extracted.
  - **Natural images** are adopted as the filling content with random strokes.
- Network structure:
  - Two stages: mask prediction & robust inpainting
  - Biggest issue: neutralize the generation degradation brought by inevitable mask estimation errors in the first stage
  - Probabilistic context normalization (PCN)

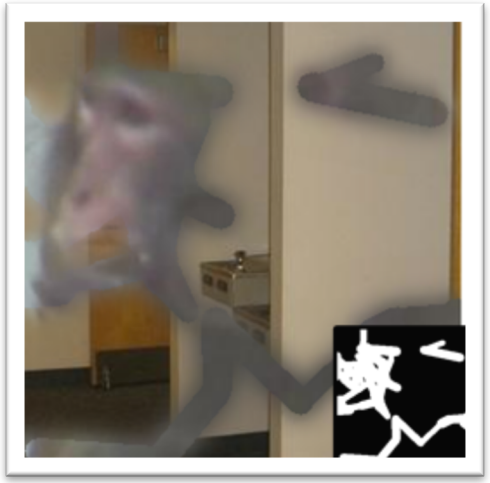
BLIND: VCNet

Training data generation

- $\langle I, O, M, N \rangle$: degraded image, ground-truth image, mask, noise

$$I = O \odot (1 - M) + N \odot M$$

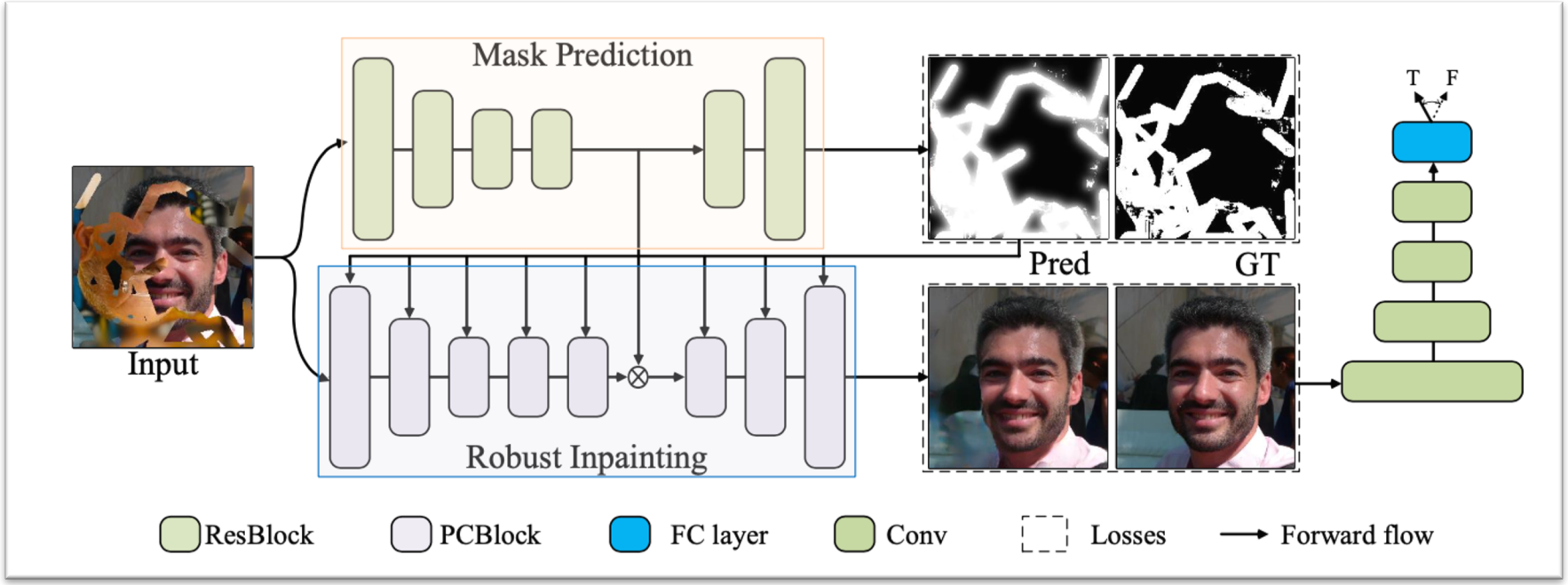
- Make N indistinguishable as much as possible with I on image pattern → real-world image patches
- M: free-form strokes
- Direct blending image O and N lead to noticeable edges → Guassian smoothing & alpha blending



BLIND: VCNet

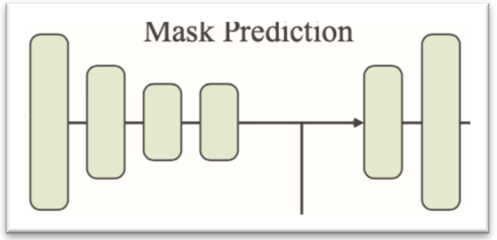
Method

- Visual Consistent Network (VCNet)
- Mask Prediction Network (MPN) + Robust Inpainting Network (RIN)



BLIND: VCNet

Method - MPN



- Aim: learn a mapping F where $F(I) \rightarrow M$.
- Encoder-decoder structure with residual blocks.
- Loss: binary cross-entropy

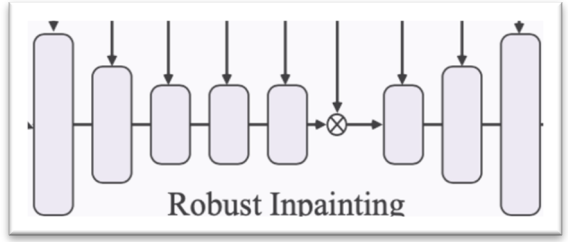
$$L_m(\widehat{M}, M) = -\tau \sum_p M_p \cdot \log(\widehat{M}_p) - (1-\tau) \sum_q (1-M_q) \cdot \log(1-\widehat{M}_q)$$

where $p \in \{p \mid M_p = 1\}$, $q \in \{q \mid M_q = 0\}$, $\tau = \frac{|\{q \mid M_q=0\}|}{h \times w}$

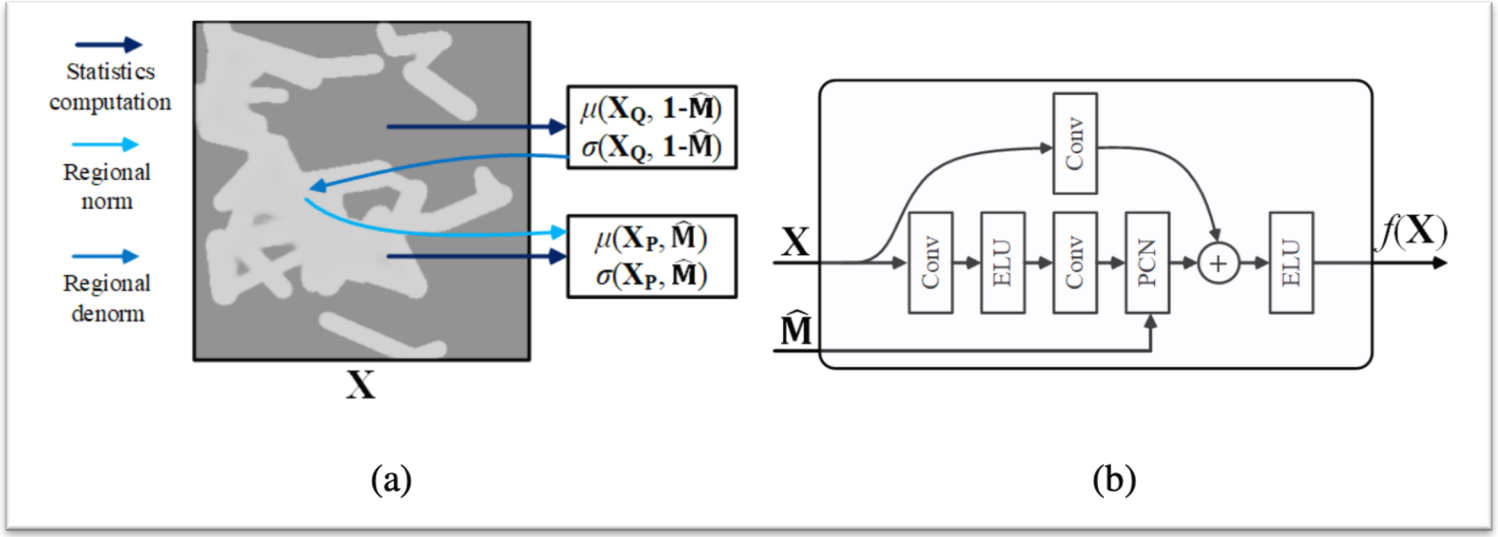
- Mask prediction is a per-pixel binary classification task → a segmentation task

BLIND: VCNet

Method - RIN



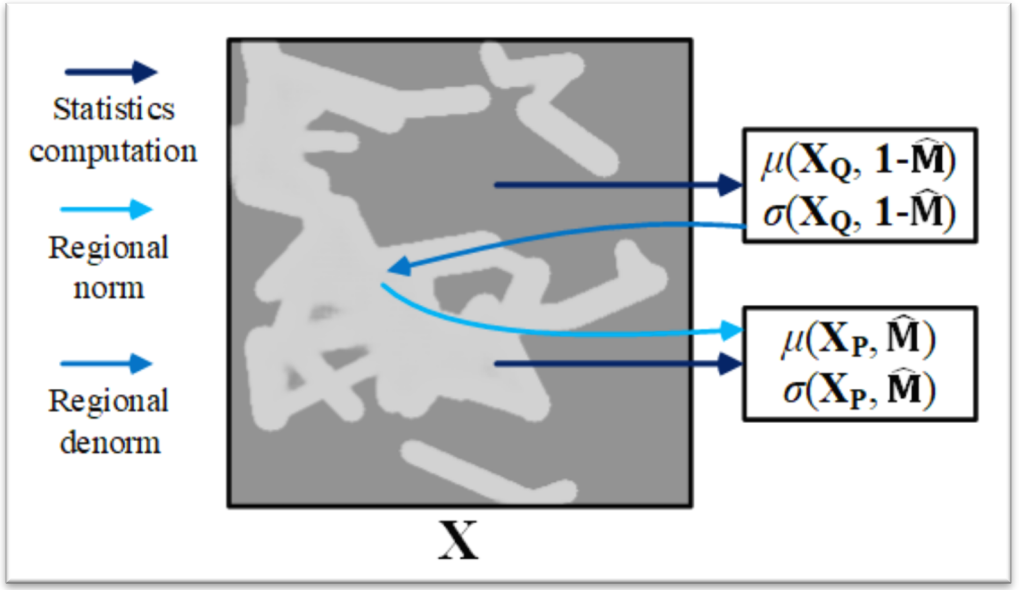
- Aim: learn a mapping G where $\mathrm{G}(I \mid \widehat{M}) \rightarrow O$.
- Encoder-decoder structure with probabilistic contextual blocks (PCB, a residual block variant with a new normalization)
- Probabilistic context normalization (PCN)



BLIND: VCNet

Method - RIN

- $PCN(X,H) = [\beta \cdot \mathrm{T}(X,H) + (1-\beta)X \odot H] \odot H + X \odot \bar{H}$
- $\mathrm{T}(X,H) = \frac{X_P - \mu(X_P,H)}{\sigma(X_P,H)} \cdot \sigma(X_Q,\bar{H}) + \mu(X_Q,\bar{H})$
- $\beta = f(\bar{x})$ squeeze-and-excitation module
- Literature shows feature mean is related to its global semantic information and variance is highly correlated to local patterns like texture.



Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural net- works. In: CVPR. pp. 2414–2423 (2016)
Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711 (2016)

BLIND: VCNet

Method - RIN

- Learning objective (reconstruction + semantic consistency + texture consistency + adversarial):

$$L_g(\hat{O}, O) = \lambda_r \left\| \hat{O} - O \right\|_1 + \lambda_s \left\| V_{\hat{O}}^l - V_O^l \right\|_1 + \lambda_f L_{mrf}(\hat{O}, O) + \lambda_a L_{adv}(\hat{O}, O)$$

where $\hat{O} = G(I \mid \widehat{M})$.

V: VGG19

L_{mrf}: ID-MRF loss

$(\lambda_r, \lambda_s, \lambda_f, \lambda_a)$: $(1.4, 10^{-4}, 10^{-3}, 10^{-3})$

Training procedure

- MPN and RIN are separately trained at first. After both networks are converged, we jointly opimize

$$\min_{\theta_F, \theta_G} \lambda_m L_m(F(I), M) + L_g(G(I|F(I)), O)$$

BLIND: VCNet

Experimental results

- Contextual Attention, Multi-column, PartialConv, GatedConv (DeepFill v2)
- Equipped with MPN in front of their inputs, trained from scratch.

- Quantitative results

Table 1: Quantitative results on the testing sets from different methods.

Method	FFHQ-2K			Places2-4K			ImageNet-4K		
	BCE↓	PSNR↑	SSIM↑	BCE↓	PSNR↑	SSIM↑	BCE↑	PSNR↑	SSIM↑
CA [37]	1.297	16.56	0.5509	0.574	18.12	0.6018	0.450	17.68	0.5285
GMC [33]	0.766	20.06	0.6675	0.312	20.38	0.6956	0.312	19.56	**0.6467**
PC [21]	**0.400**	20.19	0.6795	0.273	19.73	0.6682	0.229	19.53	0.6277
GC [38]	0.660	17.16	0.5915	0.504	18.42	0.6423	0.410	18.35	0.6416
Our VCN	**0.400**	**20.94**	**0.6999**	**0.253**	**20.54**	**0.6988**	**0.226**	**19.58**	0.6339

BLIND: VCNet

Experimental results

- Qualitative results

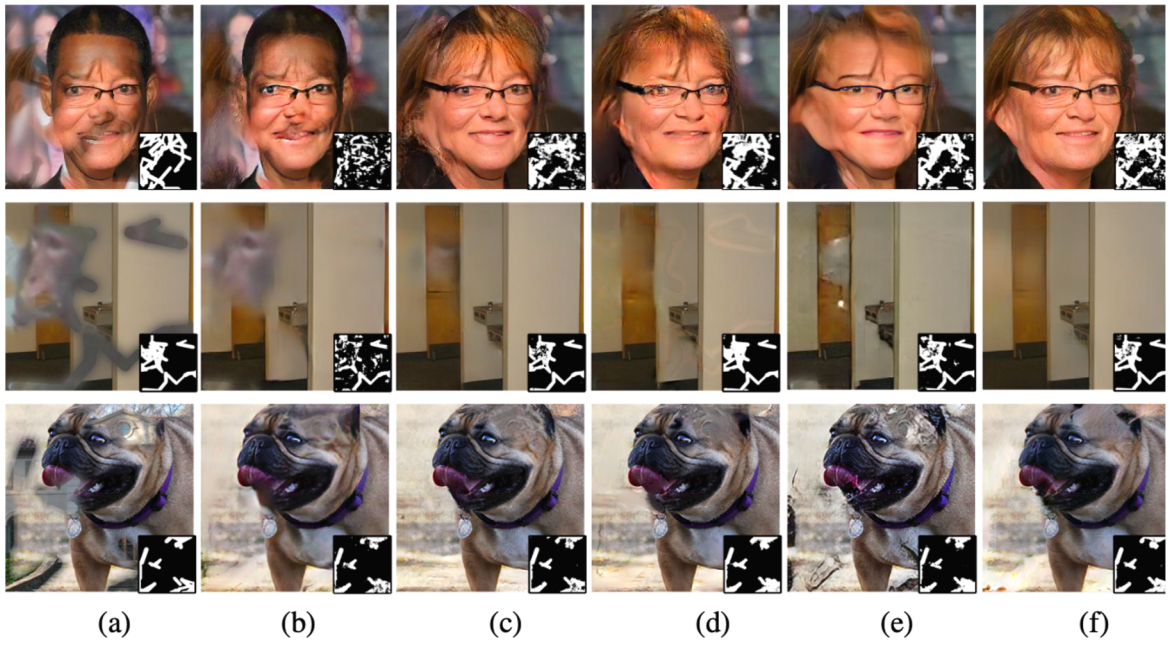
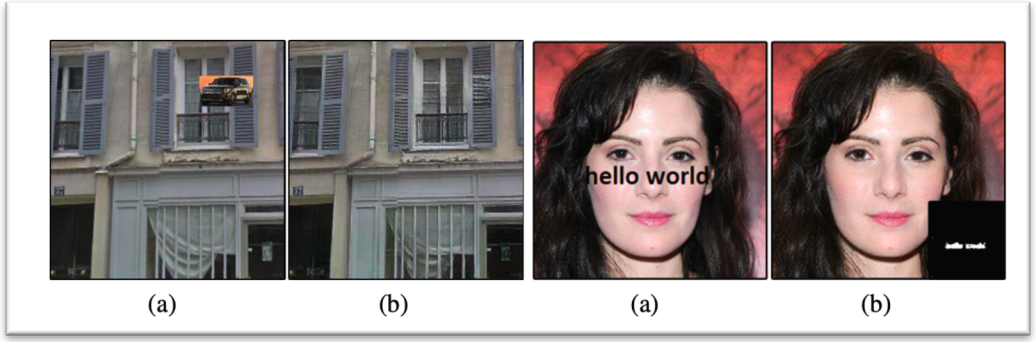
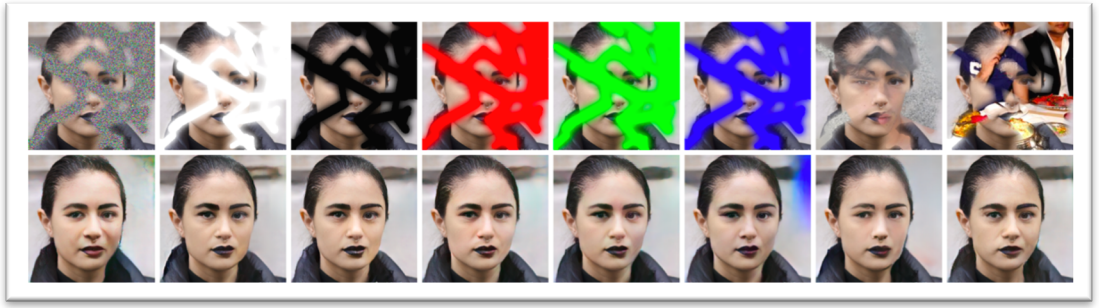


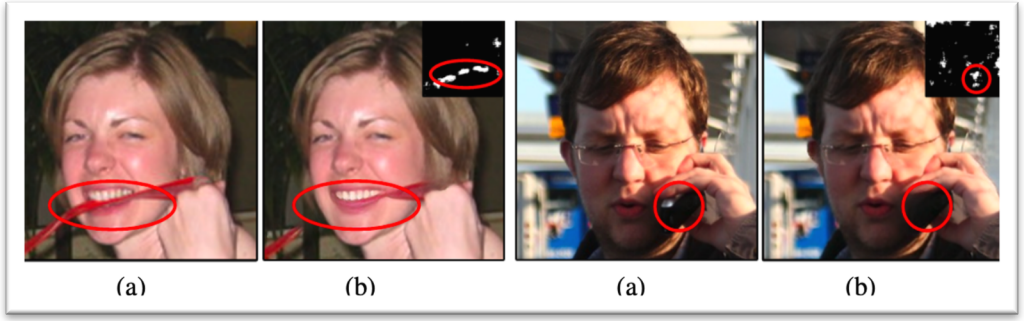
Fig. 5: Visual comparison on synthetic data (random stroke masks) from FFHQ (top), Places2 (middle), and ImageNet (bottom). (a) Input image. (b) CA [37]. (c) GMC [33]. (d) PC [21]. (e) GC [38]. (f) Our results. The ground truth masks (shown in the first column) and the estimated ones (in binary form) are shown on the bottom right corner of each image. More comparison is given in the supplementary file.

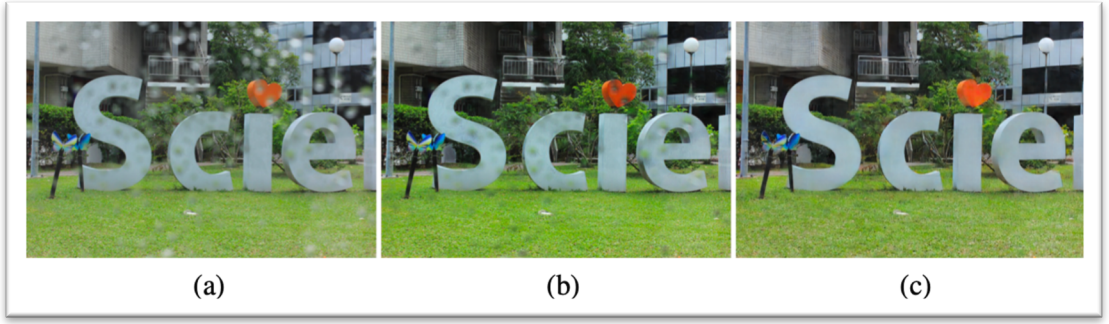
BLIND: VCNet

Experimental results









BLIND: VCNet

Experimental results

- Ablation

Table 3: Quantitative results of component ablation of VCN on FFHQ dataset (ED: Encoder-decoder; fusion: the bottleneck connection between MPN and RIN; -RM: removing the estimated contamination as $G(\mathbf{I} \odot (\mathbf{1} - \hat{\mathbf{M}})|\hat{\mathbf{M}})$; SC: semantic consistency term).

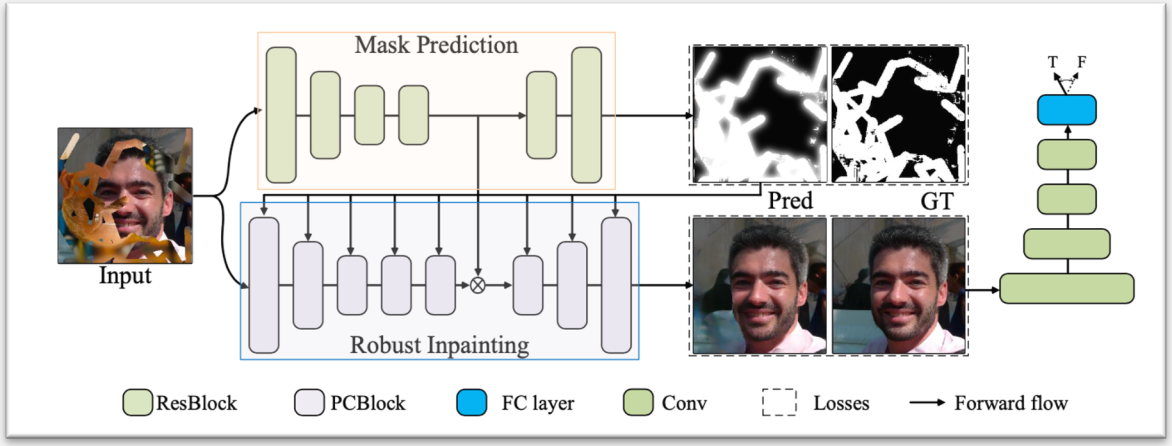
Model	ED	VCN w/o MPN	VCN w/o fusion	VCN w/o SC	VCN-RM	VCN full
PSNR↑	19.43	18.88	20.06	20.56	20.87	**20.94**
SSIM↑	0.6135	0.6222	0.6605	0.6836	**0.7045**	0.6999
BCE↓	-	-	0.560	0.653	0.462	**0.400**

(a) (b) (c) (d) (e) (f)

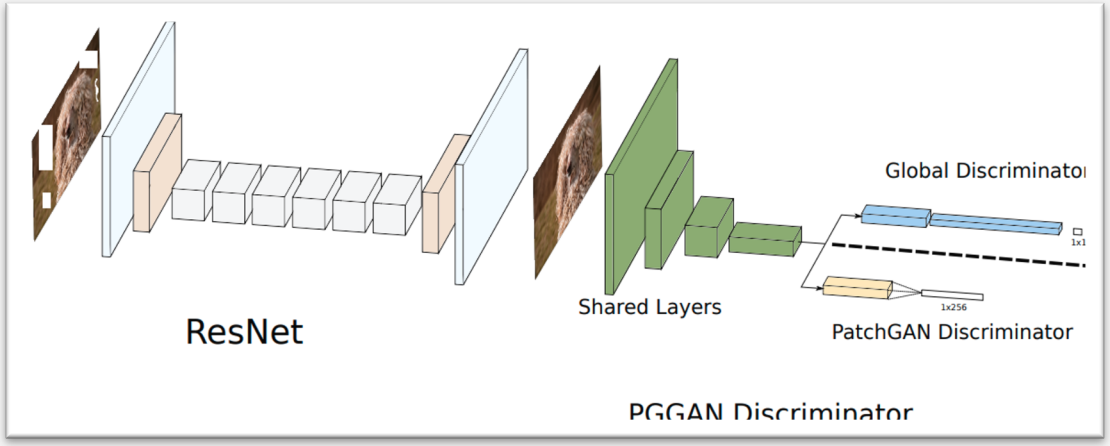
Fig. 10: Visual comparison on FFHQ using VCN variants. (a) Input image. (b) VCN w/o MPN. (c) VCN w/o skip. (d) VCN w/o semantics. (e) VCN-RM. (f) VCN full.

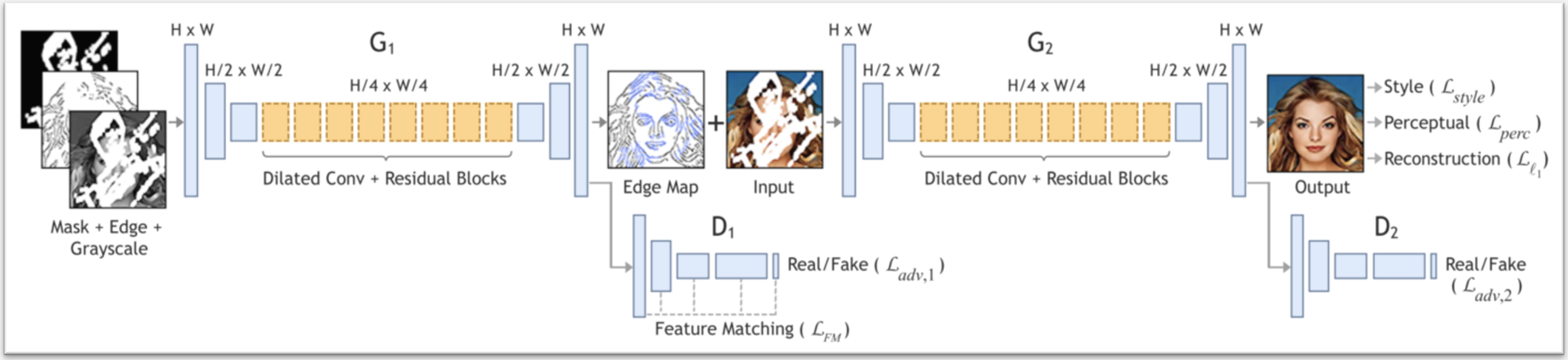
BLIND: VCNet

VCNet



PGGAN





EdgeConnect

5 THINKING

Thinking…

由 dataset 引发的思考

- Dataset
  - Paris StreetView / ImageNet / Places / CelebA + artificial distortion
- Applications
  - Object removal——mask 可以人为构造
  - 修复污染图像——人为构造的 mask 与污染形成的 mask 存在差异
- Unpaired image inpainting?
  - 训练时，也不提供 corrupted images 对应的 ground-truth images
  - Corrupted image domain → Uncorrupted image domain

Thinking...

由 reconstruction loss 引发的思考

- L1 loss / L2 loss
- Image inpainting 本身是一个 ill-posed problem，那么 reconstruction loss 的合理性值得质疑
- Reconstruction loss 会导致生成的图像模糊，即对局部真实性有害
- [Contextual Attention] Our conclusion is that the pixel-wise reconstruction loss, although tends to make the result blurry, is an essential ingredient for image inpainting. The reconstruction loss is helpful in capturing content structures and serves as a powerful regularization term for training GANs.

- 现有方法非常依赖于 reconstruction loss，能否找到只用 adversarial loss 的方式呢？

Thinking...

由 EdgeConnect 引发的思考

- Edge map 是一种先验
  - 还有什么先验可以帮助 inpainting？例如语义分割图？

由 ill-posed 引发的思考

- 能否让同一个模型输出不同的、但都合理的填充图像？

THE END