

Image Inpainting

图像修复这一概念非常好理解，该任务可用于多种应用，其目的是多样的。可以是为了防止图像质量进一步恶化（例如，照片中的裂缝或胶片中的划痕和灰尘斑点）；可以是为了添加或删除元素（例如，从照片中删除加盖日期和红眼），也可以用于图像编辑：移除不需要的图像内容，用合理的图像内容填补移除后的空缺。



Fig. 1. Masked images and corresponding inpainted results using our partial-convolution based network.

第一篇 Context Encoders: Feature Learning by Inpainting

由 Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros 发表在 CVPR 2016 上的一篇 paper，核心思想是利用卷积神经网络来学习图像中的 high-level feature，利用这些这些 feature 来指导图像缺失部分的生成。文章提出的网络结构如下，包括 3 个部分：Encoder, Channel-wise fully-connected layer, Decoder。

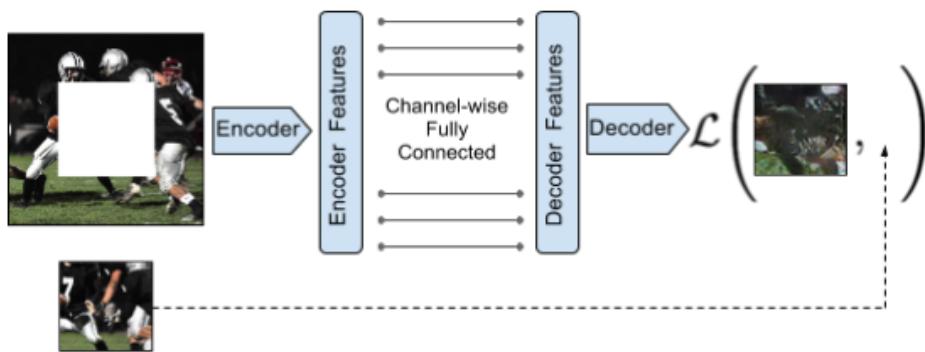


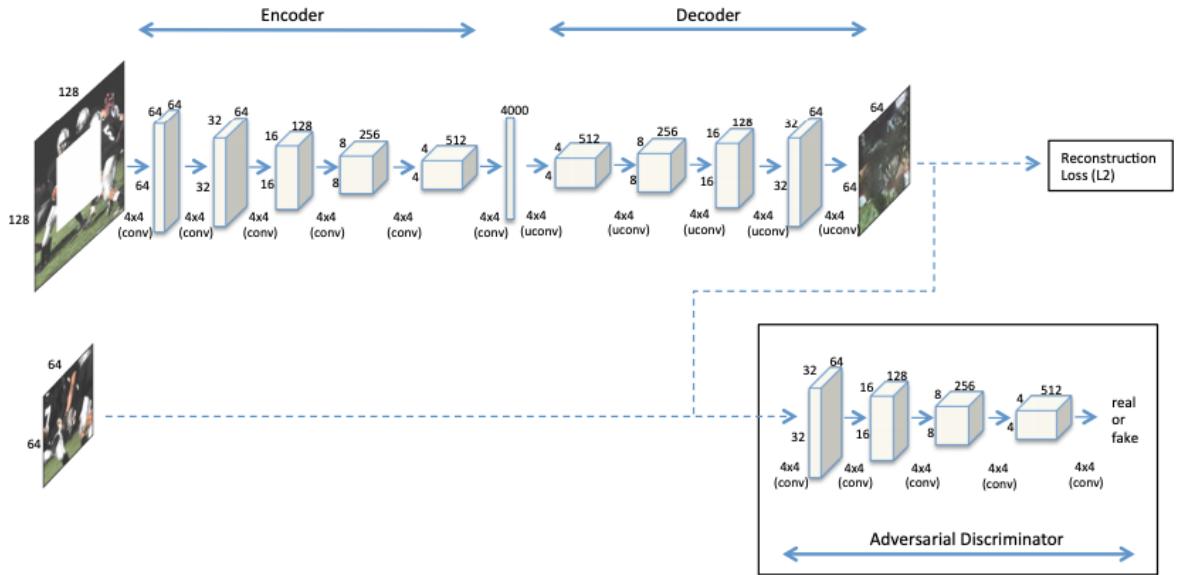
Figure 2: Context Encoder. The context image is passed through the encoder to obtain features which are connected to the decoder using channel-wise fully-connected layer as described in Section 3.1. The decoder then produces the missing regions in the image.

Encoder 的结构直接借鉴了 AlexNet 前 5 层的卷积层结构。

Channel-wise fully-connected layer 是对普通 fc 层的一种改进。之所以加入 fc 层是为了使 feature map 每一层的信息可以在内部交流。但传统的 fc 层参数太多，因此作者提出可以在 fc 中去掉 feature map 层间的信息交流，从而减少参数规模。在 fc 之后会接一个 stride 为 1 的卷积层，来实现层间的信息交流。

Decoder 的目的是将压缩的 feature map 一步步放大，恢复到原始图片丢失部分的尺寸。文章提出采用 5 个 up-convolutional 层，每层后接一个 RELU。

整体结构如下：



(a) Context encoder trained with joint reconstruction and adversarial loss for semantic inpainting. This illustration is shown for *center region dropout*. Similar architecture holds for arbitrary region dropout as well. See Section 3.2.

接下来是 Loss 的设计。作者提出了 2 个 Loss: Reconstruction Loss 和 Adversarial Loss。

Reconstruction Loss 如下，就是直接计算生成部分和原始图像的 pixel-wise error:

Reconstruction Loss We use a normalized masked L_2 distance as our reconstruction loss function, \mathcal{L}_{rec} ,

$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1 - \hat{M}) \odot x))\|_2^2, \quad (1)$$

Adversarial Loss 如下，这个 Loss 借鉴了 GAN 的 Loss。不过一个主要的区别是，这里只固定 Generator，试图通过极大化 Loss 来训练更强的 Discriminator。

$$\begin{aligned}\mathcal{L}_{adv} = \max_D \quad & \mathbb{E}_{x \in \mathcal{X}} [\log(D(x)) \\ & + \log(1 - D(F((1 - \hat{M}) \odot x)))],\end{aligned}$$

最终的 Loss 为如下的组合，Reconstruction Loss 是为了提高补全部分和周围 context 的相关性；而 Adversarial Loss 则是为了提高补全部分的真实性。通过保持二者的平衡，可以得到如下尽可能好的补全效果。

Joint Loss We define the overall loss function as

$$\mathcal{L} = \lambda_{rec} \mathcal{L}_{rec} + \lambda_{adv} \mathcal{L}_{adv}.$$

部分结果如下：

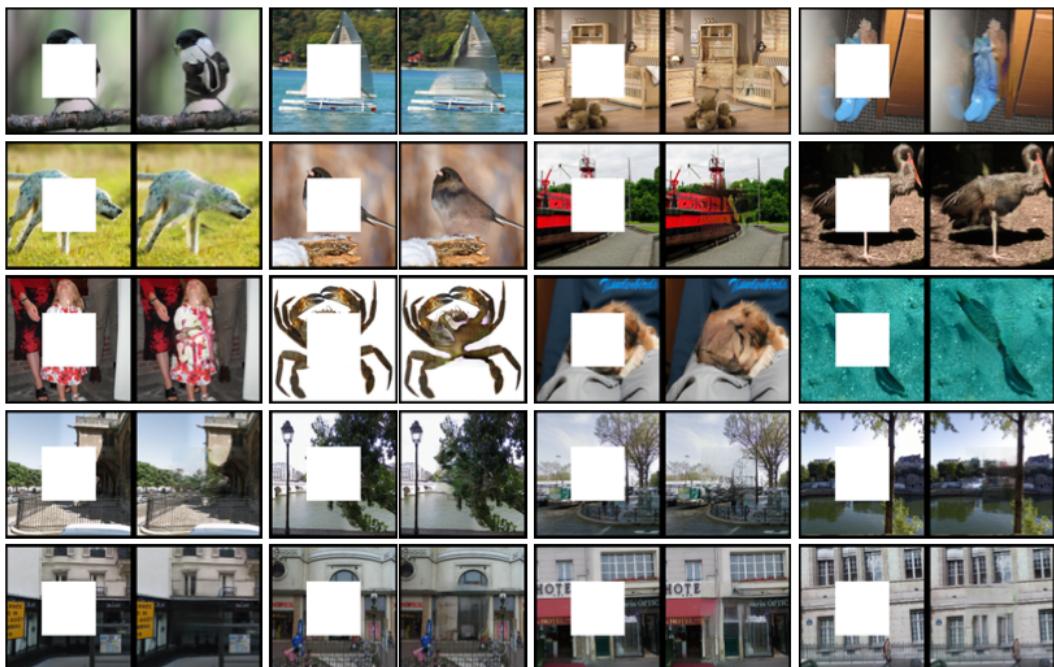


Figure 4: Semantic Inpainting results on *held-out* images for context encoder trained using reconstruction and adversarial loss. First three rows are examples from ImageNet, and bottom two rows are from Paris StreetView Dataset. See more results on author's project website.

参考：<https://arxiv.org/abs/1604.07379>

第二篇 Generative Image Inpainting with Contextual Attention

由 Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang 发表在 CVPR 2018 上的一篇 paper，这篇文章也被称作 **deepfill v1**，作者的后续工作 "[Free-Form Image Inpainting with Gated Convolution](#)" 也被称为 **deepfill v2**。两者最主要的区别是，v2 支持任意形状的 mask（标记图像的待修复区域），且支持标记黑线来指定修复的大致形状。文章主要是针对之前的基于 CNN 的方法的问题：在修复区域的边界生成扭曲的结构和模糊的纹理。作者发现是因为卷积神经网络无法很好地提取远距离的图像内容（**distant contextual information**）和不规则区域的图像内容（**hole regions**）。举例来说，一个像素点的内容被 64 个像素点以外的内容影响，那么它至少要使用 6 层 3×3 的卷积核才能够有这么大的感受野（**receptive field**）。而且由于这个感受野的形状是非常标准且对称的矩形（**regular and symmetric grid**），所以在不规则的一些图像内容上，无法很好地给对应特征分配正确的权值。

文章首先通过复制和改进近期最新的修复模型 **Globally and locally consistent image completion** 来构建其基础的生成图像修复网络，作者引入了粗略到细化的网络结构，其中第一个网络进行初始粗略预测，第二个网络将粗略预测作为输入并预测精确结果。粗略网络的训练损失为 **reconstruction loss**，细化网络的损失为 **reconstruction loss** 和 **GAN losses**。直观地说，精细网络比缺少区域的原始图像的粗略网络拥有更完整的场景输入，所以它的编码器可以比粗略网络学习更好的特征表示。

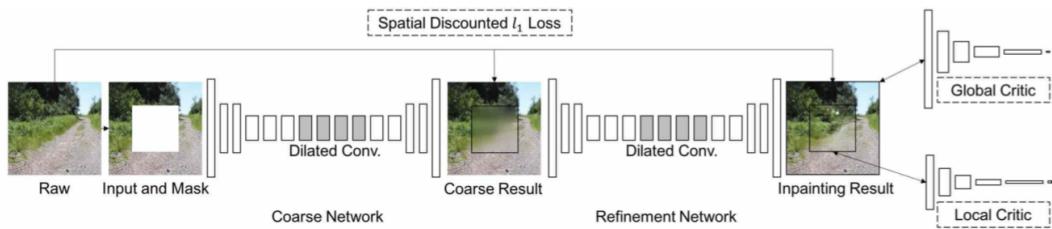


Figure 2: Overview of our improved generative inpainting framework. The coarse network is trained with reconstruction loss explicitly, while the refinement network is trained with reconstruction loss, global and local WGAN-GP adversarial loss.

除此之外，在下一部分，作者引入了 **attention mechanism**：

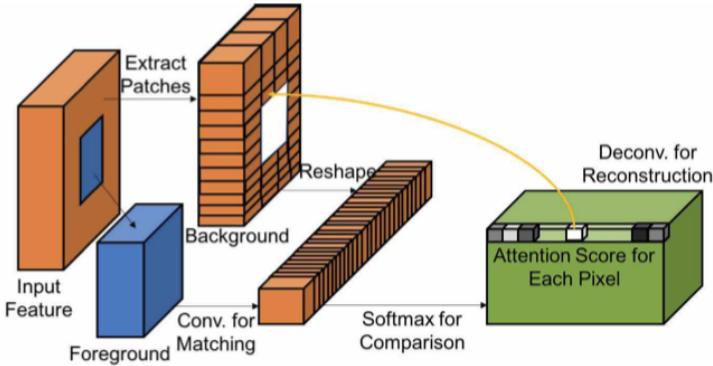


Figure 3: Illustration of the contextual attention layer. Firstly we use convolution to compute matching score of foreground patches with background patches (as convolutional filters). Then we apply softmax to compare and get attention score for each pixel. Finally we reconstruct foreground patches with background patches by performing deconvolution on attention score. The contextual attention layer is differentiable and fully-convolutional.

卷积神经网络逐层地处理具有局部卷积核的图像特征，因此对于从远处空间位置获取特征没有效果。为了克服这个局限性，作者考虑了 **attention mechanism**，并在深度生成网络中引入了一个新的 **contextual attention layer**。其核心思想是：使用已知图像 **patch** 的特征作为卷积核来加工粗略网络生成出来的 **patch**，来精细化这个模糊的修复结果。具体实现：作者首先在背景区域提取 3×3 的 **patch**，并作为卷积核。为了匹配前景（即待修复区域）**patch**，使用标准化内积（余弦相似度）来测量，然后用 **softmax** 来为每个背景 **patch** 计算权值，最后选取一个最好的 **patch**，并反卷积出前景区域。对于反卷积过程中的重叠区域（**overlapped pixels**）取平均值。

为了让网络能“想象”（**hallucinate**）出新的图像内容，还有另一条卷积通路（**convolutional pathway**），这条通路和内容感知卷积通路是平行的。这两个通路最终聚合并送入一个解码器来产生最后的输出。第二阶段的网络通过两个损失值来训练（重建损失值 **reconstruction losses** 和两个 **WGAN-GP** 损失（**Wasserstein GAN losses**），其中一个 **WGAN** 来观察全局图像，另一个 **WGAN** 来观察局部生成出的图像。

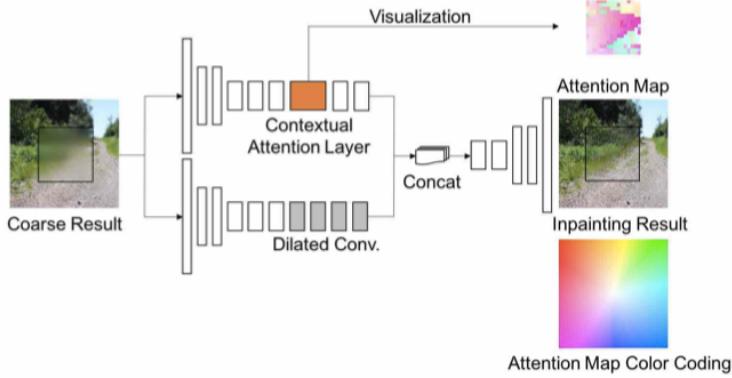


Figure 4: Based on coarse result from the first encoder-decoder network, two parallel encoders are introduced and then merged to single decoder to get inpainting result. For visualization of attention map, color indicates relative location of the most interested background patch for each pixel in foreground. For examples, white (center of color coding map) means the pixel attends on itself, pink on bottom-left, green means on top-right.

文章使用包括 Places2，CelebA 人脸，CelebAHQ 人脸，DTD 纹理和 ImageNet 在内的四个数据集进行了模型的评估。对比的模型为 Globally and locally consistent image completion 所提出的模型。

Method	ℓ_1 loss	ℓ_2 loss	PSNR	TV loss
PatchMatch [3]	16.1%	3.9%	16.62	25.0%
Baseline model	9.4%	2.4%	18.15	25.7%
Our method	8.6%	2.1%	18.91	25.3%

Table 1: Results of mean ℓ_1 error, mean ℓ_2 error, PSNR and TV loss on validation set on Places2 for reference.

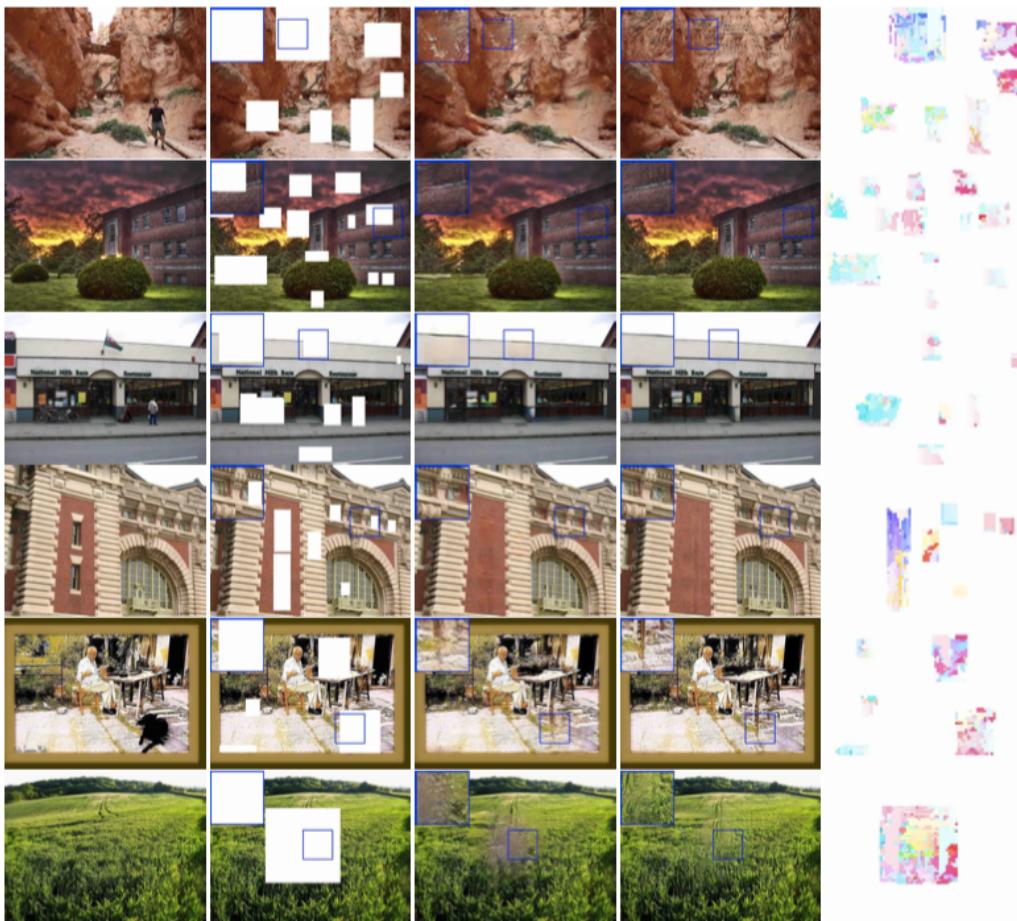


Figure 6: Qualitative results and comparisons to the baseline model. We show from left to right the original image, input image, result of our baseline model, result and attention map (upscaled 4×) of our full model. Best viewed with zoom-in.

参考: <https://arxiv.org/abs/1801.07892>

https://github.com/JiahuiYu/generative_inpainting

第三篇 EdgeConnect:Generative image inpainting with Adversarial Edge learning

由 Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, Mehran Ebrahimi
发表在 ICCV 2019 上的一篇 paper, 核心思想是结合边缘信息先验的图像修复方法, 可以更好地再现显示精细节的填充区域。生成效果如下所示, 补全模型会先生成中间所示的完整边缘信息, 然后结合失真信息一起生成最终的修复图像。

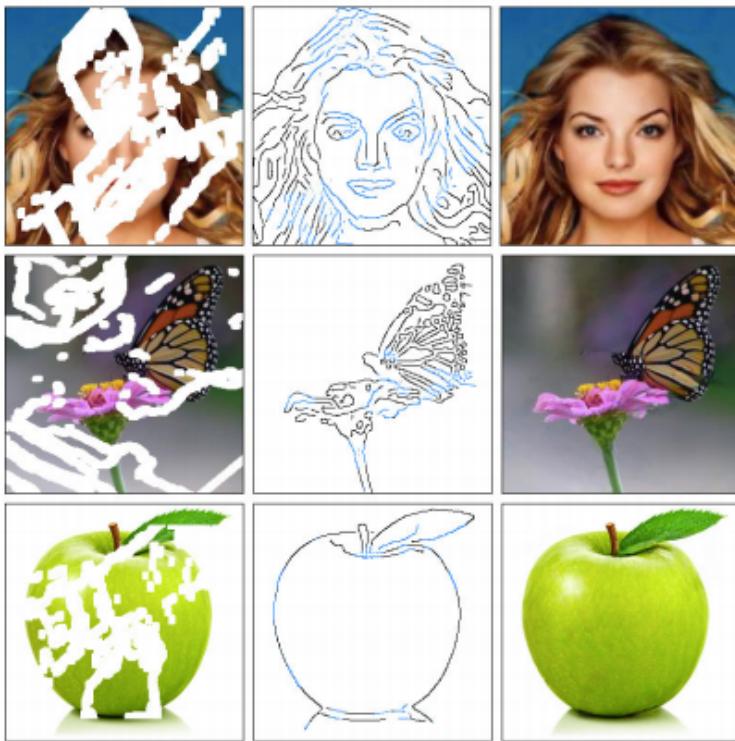


Figure 1: (Left) Input images with missing regions. The missing regions are depicted in white. (Center) Computed edge masks. Edges drawn in black are computed (for the available regions) using Canny edge detector; whereas edges shown in blue are hallucinated (for the missing regions) by the edge generator network. (Right) Image inpainting results of the proposed approach.

包括两个阶段：1) edge generator 和 2) image completion network。首先利用利用启发式的生成模型得到了缺失部分的边缘信息，随后将边缘信息作为图像缺失的先验部分和图像一起送入修复网络进行图像重建。每个 stage 包括一个对抗模型（生成器和判别器），生成器的结构类似解决 style transfer, super-resolution, and image-to-image translation 问题中的生成器，包括两次下采样，8 个 residual blocks，使用的是空洞卷积（dilated convolution），最后一个 block 包括大小为 205 的感知域。而判别器使用的是 70×70 PatchGAN。

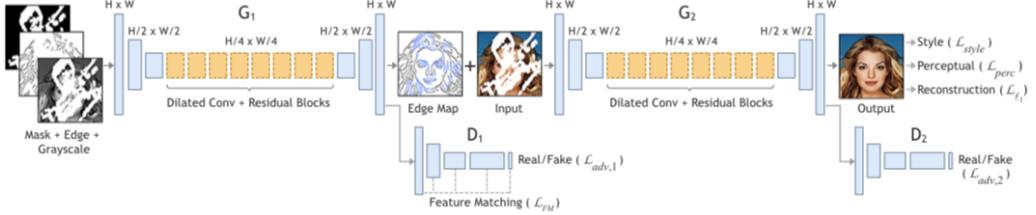


Figure 2: Summary of our proposed method. Incomplete grayscale image and edge map, and mask are the inputs of G_1 to predict the full edge map. Predicted edge map and incomplete color image are passed to G_2 to perform the inpainting task.

edge generator 阶段生成器的输入是三个单通道图像的和：一个是 mask 后的 gt 灰度图，一个是 mask 后的 gt edge 图，一个是 mask 图。这里在训练的时候 edge 是使用的 Canny 算子进行边缘提取。

$$\mathbf{C}_{pred} = G_1 \left(\tilde{\mathbf{I}}_{gray}, \mathbf{C}_{gt}, \mathbf{M} \right).$$

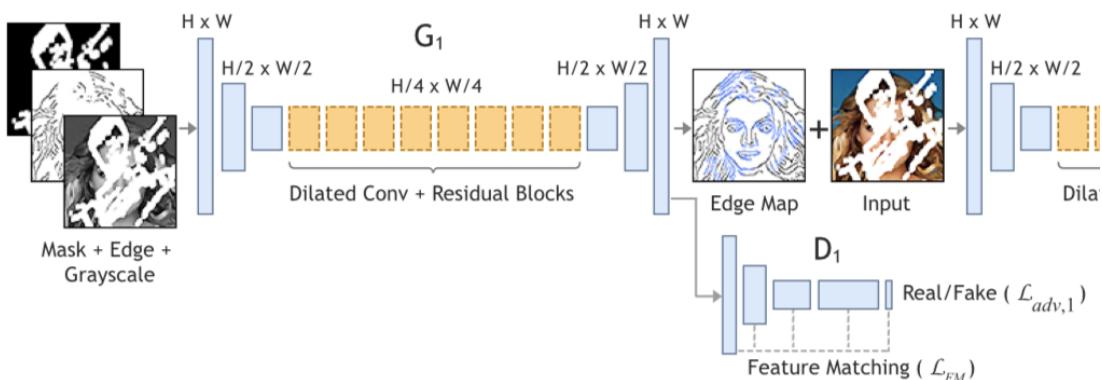
对于损失函数，包括两个，一个是对抗损失：

$$\begin{aligned} \mathcal{L}_{adv,1} &= \mathbb{E}_{(\mathbf{C}_{gt}, \mathbf{I}_{gray})} [\log D_1(\mathbf{C}_{gt}, \mathbf{I}_{gray})] \\ &\quad + \mathbb{E}_{\mathbf{I}_{gray}} \log [1 - D_1(\mathbf{C}_{pred}, \mathbf{I}_{gray})]. \end{aligned}$$

一个是 feature-matching loss：

$$\mathcal{L}_{FM} = \mathbb{E} \left[\sum_{i=1}^L \frac{1}{N_i} \left\| D_1^{(i)}(\mathbf{C}_{gt}) - D_1^{(i)}(\mathbf{C}_{pred}) \right\|_1 \right]$$

它比较 fake 和 real 图像在每个卷积层 activation map 的 element-wise error 并求和。



而在 Image Completion Network 阶段生成器的输入是 mask 后的 gt 图, 输入的条件是上一阶段生成的 edge 图(补充缺失的 edge 部分吗, 上图蓝色部分)和 mask 后的 edge 图的组合, 即一个完整的 edge。

rupted region from the previous stage, i.e. $\tilde{\mathbf{C}}_{comp} = \mathbf{C}_{gt} \odot (\mathbf{1} - \mathbf{M}) + \mathbf{C}_{pred} \odot \mathbf{M}$. The network returns a color im-

最后生成结果：

$$\mathbf{I}_{pred} = G_2 \left(\tilde{\mathbf{I}}_{gt}, \mathbf{C}_{comp} \right).$$

损失函数包括, l1 loss, adversarial loss, perceptual loss, and style loss。其中 perceptual loss :

$$\mathcal{L}_{perc} = \mathbb{E} \left[\sum_i \frac{1}{N_i} \|\phi_i(\mathbf{I}_{gt}) - \phi_i(\mathbf{I}_{pred})\|_1 \right]$$

即比较重构图片和 gt 图片在一个 pre-trained 的网络中每一个 activation map 的距离, 原文中是 relu1 1, relu2 1, relu3 1, relu4 1 and relu5 1 of the VGG-19 network pre-trained on the ImageNet dataset。

而 style loss 则通过从 activation map 构造 Gram matrix, 用以计算两个 activation map 的协方差, style loss 是为了防止反卷积造成的 “checkerboard” artifacts。

所以总损失为：

$$\mathcal{L}_{G_2} = \lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_{adv,2} \mathcal{L}_{adv,2} + \lambda_p \mathcal{L}_{perc} + \lambda_s \mathcal{L}_{style}.$$

比较结果如下：

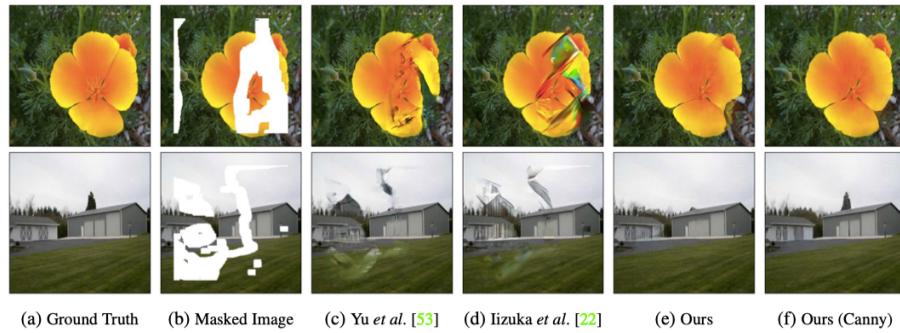


Figure 3: Comparison of qualitative results with existing models. (a) Ground Truth Image. (b) Ground Truth with Mask. (c) Yu et al. [53]. (d) Iizuka et al. [22]. (e) Ours (end-to-end). (f) Ours (G_2 only with Canny $\sigma = 2$).

提取不同程度的边缘信息作为条件：对于较大的 σ 值，可用的边缘太少，不能保证生成的图像质量。另一方面，当 σ 太小时，生成太多边缘，这对于所生成图像的质量也会产生不利影响。也就是说，存在合适最佳 σ 值，或者说我们只需要适当的边缘信息量。下图展示了修复图像的质量随 σ 的变化：

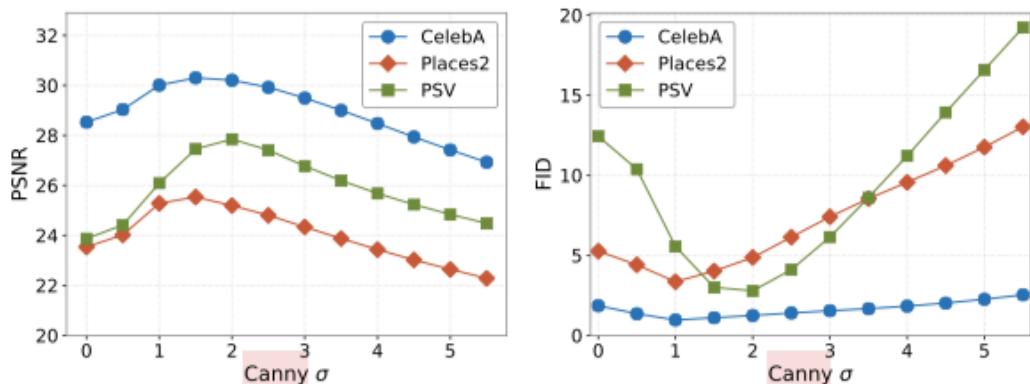


Figure 6: Effect of σ in Canny detector on PSNR and FID.

参考：<https://arxiv.org/abs/1901.00212>

总结

通过以上几个论文不难发现，使用深度学习解决图像修复的问题的基本方法是引入 GAN，一方面，GAN 的对抗损失可以让修复后的图像尽可能真实，另一方面，通过定义一些感知损失，如简单的基于 pixel 的 $L1, L2$ 重构损失，或者使用 CNN

提取一定特征后的损失，都可以保证修复后的图像与原图像尽可能保持一致。可以说，CNN 利用局部感知视野获得图像中缺失的部分周围的特征，即一些 context，从而作为一定的条件指导 GAN 中生成器的生成。对我论文要研究的课题提供了一定的启发和帮助，因为对于“sketch-image generation”来说，sketch 相当于缺失了部分信息的图片，如何对 sketch 和生成的图片进行特征的匹配是定义损失函数的关键。