普通卷积层和残差块均采用步幅 2 来减少图像大小并双倍增加输出的通道数 (第一层卷积输出通道数为 64), 最后两层通道数不变.

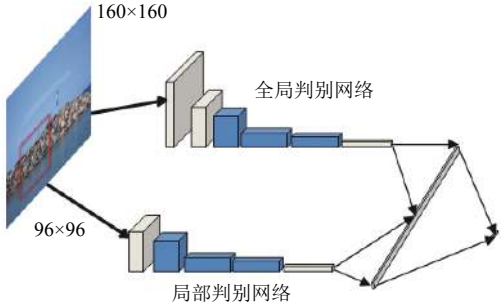

图 2 判别网络结构图

局部上下文判别器与全局类似, 局部网络的输入为 96×96 大小的图像, 此图像包括了缺失区域, 但不全是缺失区域, 还有一部分是未缺失图像. 由于输入只是全局网络输入的缺失部分, 该文去掉全局网络中的第一层作为局部网络. 输出同样为一个 1024 维的向量, 表示局部上下文信息.

最后全局和局部上下文判别网络的输出被拼接成一个 2048 维的向量, 然后送入一个全连接层并输出一个实数, 最后经过一个 Sigmoid 函数将其映射到 [0, 1] 范围代表图像真假的概率.

判别器的损失如下:

$$L_{\text{dis}} = -E_{x\sim p_r}[\log(D(x)) + \log(1 - D(G(x)))] \tag{5}$$

2.4 上下文感知损失

对于生成网络, 我们希望它具有能够约束局部特征的相似性, 使图像修复得更加逼真. 从近期的图像修复所使用的损失来看, 普遍使用的是像素级的损失即要求修复后的图像与完整图像对齐[10–12], 这些方法都不利于生成一个完整而又清晰的修复结果. 因此该文提出了联合上下文感知损失网络训练生成网络, 将图像视为特征的集合, 忽略特征的空间位置, 不要求图像完全对齐, 允许局部形变, 并根据特征之间的相似性度量图像之间的相似性. 让生成的图像和原始图像经过 VGG16 特征提取器, 得到图像的特征图, 通过特征图计算相似度作为损失来对生成网络进行训练.

上下文感知损失网络接收两张 160×160 大小的图像, 经过已经训练好的 VGG16 后输出感知损失. 假设 x为输入图像, 则表示生成网络, 表示 VGG16 网络, 表示计算相似度的函数, 那么上下文感知网络的损失可以表示为:

$$L_{CX} = -\log[CX(\Phi(x), \Phi(G(M \odot x)))] \tag{6}$$

更进一步的, 对于两个输入的图像x, y经过 VGG16 的特征提取后的x_i, y_j, 其中 CX 函数计算两张图像的相似性函数如下, 对于每个特征y_j, 找到与它最相似的特征x_i, y_j, 然后对所有y_j求和相应的特征相似值:

$$CX(x,y) = CX(X,Y) = \frac{1}{N}\sum_j \max_i CX_{ij} \tag{7}$$

其中,

$$CX_{ij} = w_{ij} \Big/ \sum_k w_{ik} \tag{8}$$

其中, w_{ij}表示特征x_i与y_j的相似性, 通过下式计算得到.

$$w_{ij} = \exp\left(\frac{1 - d_{\text{similar}}}{h}\right) \tag{9}$$

上式通过求幂从距离转换到相似性, 距离由下式计算得到.

$$d_{\text{similar}} = \frac{d_{ij}}{\min_k d_{ik} + \varepsilon} \tag{10}$$

对d_{ij}归一化, 其中d_{ij}为x_i与y_j的余弦距离. 上述d_{ij}计算公式为:

$$d_{ij} = \left(1 - \frac{(x_i - \mu_y)\cdot(y_j - \mu_y)}{\|x_i - \mu_y\|_2 \|y_j - \mu_y\|_2}\right) \tag{11}$$

其中,

$$\mu_y = \frac{1}{N}\sum_j y_j \tag{12}$$

在训练过程中, 通过不断的减小此损失 (L_{CX}) 来优化生成网络, 生成网络因此而具有约束局部特征的相似性的功能.

3 实验

3.1 数据集

本文使用来自香港中文大学的开放数据集 CelebA[14] 和 LFW[15]. CelebA 是一个大型的人脸属性数据集, 包含 1 万多个名人身份的 20 多万张图片. LFW 数据集是一个无约束自然场景人脸识别数据集, 该数据集由 13 000 多张全世界知名人士互联网自然场景环境人脸图片组成.

3.2 训练过程

在实验中, 将 CelebA 的 12 万张图片作为训练集,

剩下的图片作为测试集, 将式 (4) 中λ_1设置为 0.0004, λ_2设置为 0.004, 将 batch size 设置为 12, 图片被裁剪成 160×160 像素大小然后被送进网络训练. 首先单独对生成网络训练 90000 次, 然后对判别网络训练 10000 次, 最后联合生成网络, 判别网络训练 400 000 次. 对于 LFW 数据集我们将其中的 1 万多张图片作为训练集剩下的图片作为测试集, 将式 (4) 中λ_1设置为 0.0004, λ_2设置为 0.004, 将 batch size 设置为 16, 图片被裁剪成 160×160 像素大小, 训练步骤与上述相同.

3.3 SE-ResNet 的效果分析

在 3.1 节该文从理论上分析了基于 SE-ResNet 残差块的生成网络和判别网络的优势, 为了进一步从实验上证明, 该文使用了 CelebA 数据集对添加 SE-ResNet 残差块进行了定性分析. 如图 3. 第 1 列为原始图片, 第 2 列为缺失图片, 第 3 列为不使用 SE-ResNet 残差块的方法修复后图片, 第 4 列为使用 SE-ResNet 残差块的修复后图片. 可以看出, SE-ResNet 的使用对修复效果的影响显著, 这也验证了 SE-ResNet 残差块的有效性.

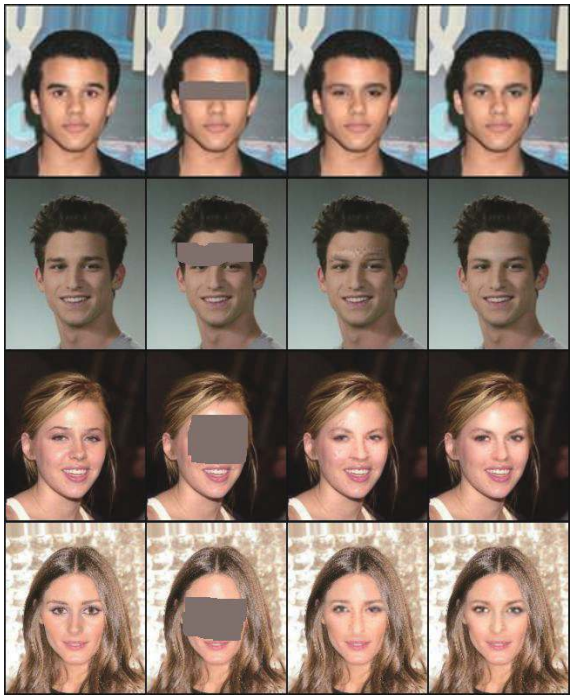

图 3 添加 SE-ResNet 残差块与否对比图

3.4 上下文感知损失的效果分析

在 3.4 节该文从理论上分析了联合上下文感知损失网络训练生成网络的有效性, 为了验证上下文感知损失网络在修复效果上的重要性, 分别在训练时采用上下文感知损失网络和不采用上下文感知损失网络对网络进行训练, 结果如图 4 所示. 第 1 列为原始图片, 第 2 列为缺失图片, 第 3 列为不采用上下文感知损失网络修复后图片, 第 4 列为采用上下文感知损失网络修复后图片. 从图 4 的第 4 行可以看出, 对缺失嘴部的修复中, 不采用上下文感知损失网络修复后的图片存在较为严重的修复痕迹, 该文的方法能够有效的减少修复痕迹, 与原始图像相似度更高.

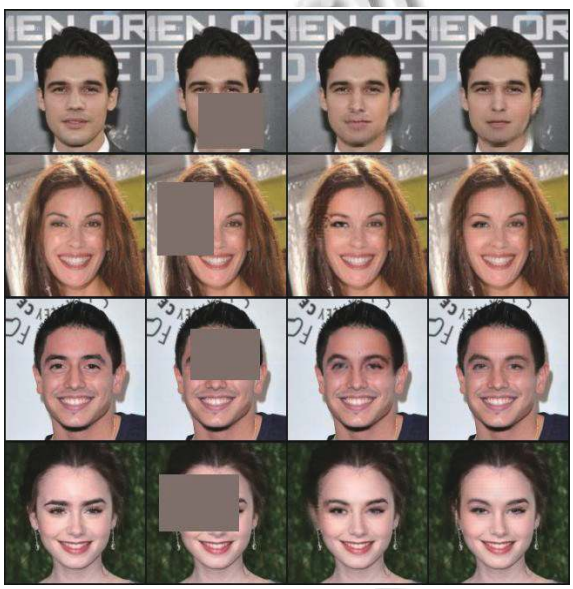

图 4 采用上下文感知损失与否对比图

3.5 与现有方法的比较

为了证明该文方法的优越性, 接下来将从各个方面与文献 [3] 的方法进行对比分析.

该文使用了两种评价标准来评测修复效果. 峰值信噪比 (Peak Signal to Noise Ratio, PSNR)[16] 是一种全参考的图像质量评价指标. 结构相似性 (Structural SIMilarity, SSIM)[17], 是一种全参考的图像质量评价指标, 它分别从亮度、对比度、结构 3 方面度量图像相似性. SSIM 取值范围 [0, 1], 值越大, 表示图像失真越小.

从 CelebA 的测试数据集中随机的选取 128 张图片, 分别计算中心缺失为 1/4 和中心缺失为 1/3 的原图, 用文献 [3] 方法修复的图像以及本文的方法修复的图像的 PSNR 以及 SSIM 值, 然后进行对比. 从表 1 可以看出, 本文的方法优于文献 [3] 的方法, 更加接近原图. 另外, 图 5 定性分析了对比效果. 前两行为中心缺失 1/2 的修复效果对比图, 后两行为中心缺失 1/4 的修复

效果对比图, 从左到右依次为原图, 缺失图片, 文献 [3] 的方法修复后图片, 本文方法修复后图片. 可以看出文献 [3] 的修复结果存在较多的瑕点, 该文的修复效果更加清晰.

表 1 与文献 [3] 方法的 PSNR 和 SSIM 指标对比

方法	PSNR (dB)	SSIM
文献[3]	30.18	0.80
本文方法	36.90	0.92

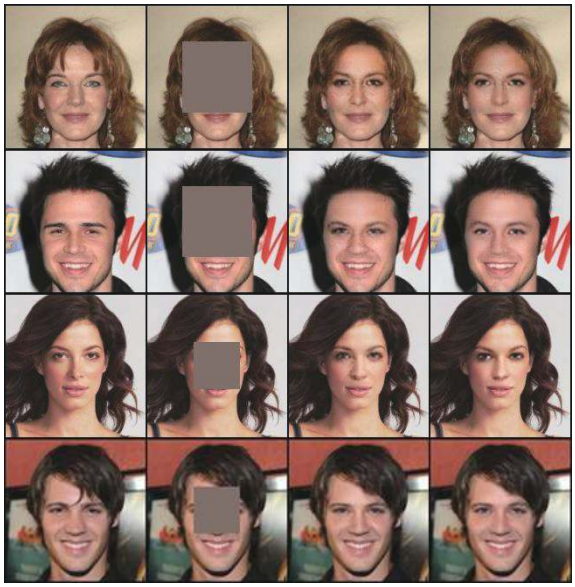

图 5 与文献 [3] 方法的中心缺失修复效果对比图

另外图 6 展示了在随机缺失的情况下文献 [3] 方法与本文方法的对比, 可以看出在图像的高频信息部分缺失时 (如眼睛、鼻子、嘴巴等), 文献 [3] 方法修复细节较差, 且存在较多的伪影, 而本文的方法不存在, 修复效果更好.

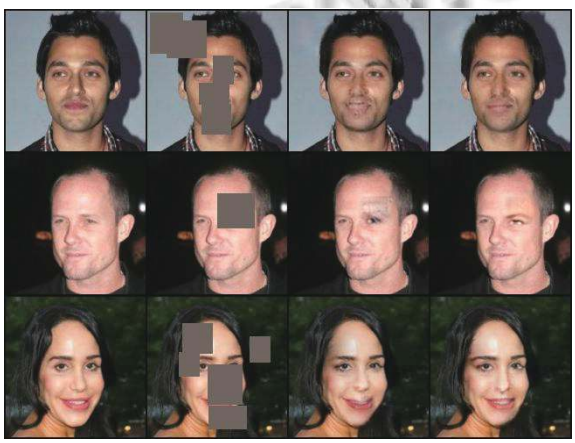

图 6 与文献 [3] 方法的随机缺失修复效果对比图

图像边缘检测图中标识了数字图像中亮度变化明显的点, 反映了图像中的重要的结构属性特征, 因此本文从修复前后边缘检测图的对比来判断图像的修复效果. 图 7 中第 1 列为原图, 第 2 列为原图的边缘检测图, 第 3 列为修复后图像的边缘检测图. 从图中可以看出, 修复后的边缘检测图与原图比较接近, 表明本文方法能从结构上理解并修复图像.

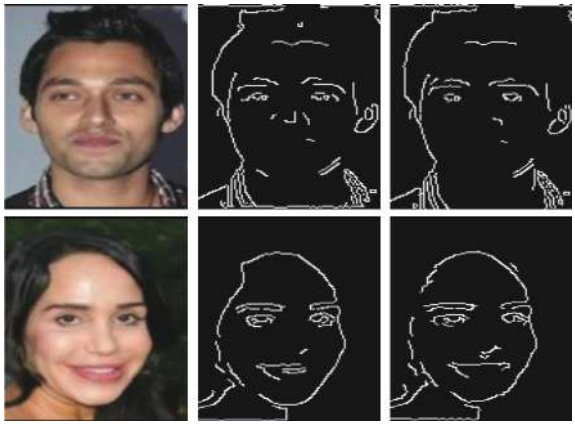

图 7 边缘检测图对比

4 结束语

本文提出了一种基于 SE-ResNet 的生成对抗网络联合上下文感知损失的方法来进行图像修复工作, 通过在生成网络和判别网络部分加入 SE-ResNet 模块, 提升网络特征利用, 使得修复图像更清晰. 通过联合 contextual loss 约束局部特征的相似性, 使得修复图像更加逼真. 多个实验证明, 该文的方法在图像修复上具有重要的作用. 然而当原始图像的分辨率较大或者缺失范围越大, 该方法修复结果还是会存在明显的模糊, 并且训练时间更长, 因此如何解决这个问题有待于进一步研究. 基于深度学习的语义分割技术已经较为成熟, 将语义分割技术与图像修复相结合是一项非常有意义的工作, 因此下一步的研究重点就是如何将语义分割应用于图像修复.

参考文献

1 Darabi S, Shechtman E, Barnes C, *et al*. Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on Graphics, 2012, 31(4): 82.

2 Pathak D, Krahenbühl P, Donahue J, *et al*. Context encoders: Feature learning by inpainting. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition.

Las Vegas, NV, USA. 2016. 2536–2544.
3 Iizuka S, Simo-Serra E, Ishikawa H, *et al*. Globally and locally consistent image completion. ACM Transactions on Graphics, 2017, 36(4): 107.
4 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 7132–7141.
5 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
6 Mechrez R, Talmi I, Zelnik-Manor L, *et al*. The contextual loss for image transformation with non-aligned data. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 800–815.
7 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
8 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA, USA. 2014. 2672–2680.
9 Song YH, Yang C, Shen YJ, *et al*. SPG-Net: Segmentation prediction and guidance network for image inpainting. arXiv: 1805.03356, 2018.
10 Yu JH, Lin Z, Yang JM, *et al*. Generative image inpainting with contextual attention. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 5505–5514.
11 Liu HY, Jiang B, Xiao Y, *et al*. Coherent semantic attention for image inpainting. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 4169–4178.
12 Yu JH, Lin Z, Yang JM, *et al*. Free-form image inpainting with gated convolution. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 4471–4480.
13 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv: 1511.07122, 2015.
14 Liu ZW, Luo P, Wang XG, *et al*. Large-scale celebfaces attributes (CelebA) dataset. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, 2018.
15 Huang GB, Ramesh M, Berg T, *et al*. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Amherst: University of Massachusetts, 2007.
16 佟雨兵, 张其善, 祁云平. 基于 PSNR 与 SSIM 联合的图像质量评价模型. 中国图象图形学报, 2006, 11(12): 1758–1763. [doi: 10.11834/jig.2006012307]
17 Horé A, Ziou D. Image quality metrics: PSNR vs. SSIM. Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey. 2010. 2366–2369.