多级注意力传播驱动的生成式图像修复方法

曹承瑞 [1] 刘微容 [1] 史长宏 [1] 张浩琛 [1]

摘 要 现有图像修复方案普遍存在着结构错乱和细节纹理模糊的问题, 这主要是因为在图像破损区域的重建过程中, 修复网络难以充分利用非破损区域内的信息来准确地推断破损区域内容. 为此, 本文提出了一种由多级注意力传播驱动的图像修复网络. 该网络通过将全分辨率图像中提取的高级特征压缩为多尺度紧凑特征, 进而依据尺度大小顺序驱动紧凑特征进行多级注意力特征传播, 以期达到包括结构和细节在内的高级特征在网络中充分传播的目标. 为进一步实现细粒度图像修复重建, 本文还同时提出了一种复合粒度判别器, 以期实现对图像修复过程进行全局语义约束与非特定局部密集约束. 大量实验表明, 本文提出的方法可以产生更高质量的修复结果.

关键词 注意力传播, 特征压缩, 复合粒度判别器, 图像修复

引用格式 曹承瑞, 刘微容, 史长宏, 张浩琛. 多级注意力传播驱动的生成式图像修复方法. 自动化学报, 2022, **48**(5): 1343−1352

DOI 10.16383/j.aas.c200485

Generative Image Inpainting With Attention Propagation

CAO Cheng-Rui[1] LIU Wei-Rong[1] SHI Chang-Hong[1] ZHANG Hao-Chen[1]

Abstract There are disordered structures and blurred detail textures in most existing image inpainting methods, because the inpainting network cannot make full use of the information in the non-damaged regions to infer the contents of the damaged regions during reconstructing process. To address the issues, an image inpainting network driven by multi-scale attention propagation is proposed in this paper. Firstly, the high-level features extracted from the full-resolution images are compressed into multi-scale compact features, and then the compact features are driven to perform multi-scale attention feature propagation in order of scale size. As a result, the high-level features including structures and details are fully propagated in the network. In order to realize fine-grained image inpainting, a compound granularity discriminator is proposed to constrain image inpainting process with global semantic constraints and non-specific local dense constraints. A large number of experimental results show that the proposed method can restore higher quality inpainting results.

Key words Attention propagation, feature compression, compound granularity discriminator, image inpainting

Citation Cao Cheng-Rui, Liu Wei-Rong, Shi Chang-Hong, Zhang Hao-Chen. Generative image inpainting with attention propagation. *Acta Automatica Sinica*, 2022, **48**(5): 1343−1352

图像修复是指对图像中缺失或损坏区域进行修复重建的过程, 它是计算机视觉技术领域的重点研究内容之一, 其在图像编辑、图像渲染等诸多领域具有重要实用价值[1−8]. 如何在图像破损区域合成与现有上下文区域结构语义一致、内容准确、细节丰富的局部图像信息, 是图像修复方法需要解决的难点问题.

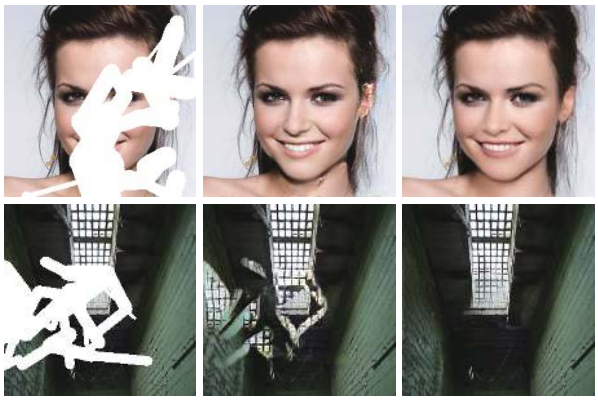
根据所利用特征级别的不同, 现有图像修复方法可分为两大类: 1) 利用低级非语义特征的方法; 2) 利用高级语义特征的方法. 其中, 利用低级非语义特征的图像修复方法为传统的图像修复方法, 通常基于扩散或图像块匹配机制将非破损区域的低级特征 “粘贴”到破损区域. 此类方法对特定的图像缺损类型有着优秀的修复效果. 例如基于扩散的方法将图像信息从破损区域边界往内部进行传播, 可以有效地修复 “抓痕”这样的细小破损. 基于图像块匹配的方法在背景修复方面性能强大, 并广泛应用于商用软件中. 然而, 此类利用低级非语义特征的图像修复方案无法对破损区域的上下文进行深入理解, 即无法获取图像的高级语义特征, 使得此类方法对高度模式化的图像 (比如人脸) 无法实现很好的修复效果.

利用高级语义特征的方法, 从大规模数据中学习高级语义特征, 大大提升了修复性能. 其中, 基于

收稿日期 2020-07-01 录用日期 2020-12-14
Manuscript received July 1, 2020; accepted December 14, 2020
国家自然科学基金 (61461028, 61861027) 资助
Supported by National Natural Science Foundation of China (61461028, 61861027)
本文责任编委 张向荣
Recommended by Associate Editor ZHANG Xiang-Rong
1. 兰州理工大学电气工程与信息工程学院 兰州 730050
1. College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050

生成式对抗网络 GANs[9] (Generative adversarial nets) 的方法已成为图像修复领域的主流. 基于 GANs 的方法将图像修复问题转化为基于条件生成对抗网络[10] 的条件生成问题. 此类方法通常以破损图像与标定破损区域的掩码作为条件输入, 采用自动编码器网络作为生成器来重建缺损区域的内容, 并结合判别器网络以对抗方式训练, 最终得到完整的图像输出. 为有效地综合利用图像上下文区域的特征, GL[11] (Globally and locally consistent image completion) 引入级联扩张卷积, 并将其集成到自动编码器网络的 "瓶颈区". 虽然扩张卷积可以在一定程度上将远距离特征纳入其感受野中, 以达到综合利用远距离特征的目标; 但是扩张卷积有较大的空穴区域, 以规则对称的网格方式采样图像特征, 从而造成远距离重点区域特征被忽略. MC[1] (Multi-column convolutional), CA[2] (Contextual attention) 以及 CI[12] (Contextual-based inpainting) 等方案采用单级上下文注意力方案, 计算图像上下文的语义相似度, 显式地从破损图像的未破损区域中借取有意义的图像表达, 缓解了远距离特征无法有效利用的问题.

然而, 以上这些方法通常无法为场景复杂图像的缺损区域生成结构合理、细节丰富的内容. 如图 1(b) 所示, 修复结果图像中明显存在整体性或局部性结构错乱, 此外生成图像还存在语义特征重建不够细致的问题, 即对图像语义 (比如人脸图像的眼睛、鼻子等部分) 重建比较模糊.



(a) 输入 (a) Inputs (b) MC[1]/CA[2] 结果 (b) MC[1]/CA[2] results (c) 本文结果 (c) Our results

图 1 当前图像修复方法所存在的结构和细节问题展示

Fig. 1 The structure and detail issues encountered in current image inpainting method

如图 2 所示为当前主流图像修复方案通常采用的自动编码器生成网络. 缺损图像经过编码器编码得到浅层特征, 将浅层特征送入 "瓶颈区"进行特征提取, 然后再由解码器解码为完整图像. 我们通过研究发现此类自动编码器结构存在非常严重的特征传递受阻问题, 其 "瓶颈区"高级特征的截面过大 (一般为 64×64 像素大小). 大截面特征使得扩张卷积与单级注意力特征匹配等方案[2, 11−12] 无法充分获取结构与细节特征, 同时阻碍了结构和细节特征在网络中传播, 从而导致了修复结果中出现结构错乱和语义对象模糊等现象.

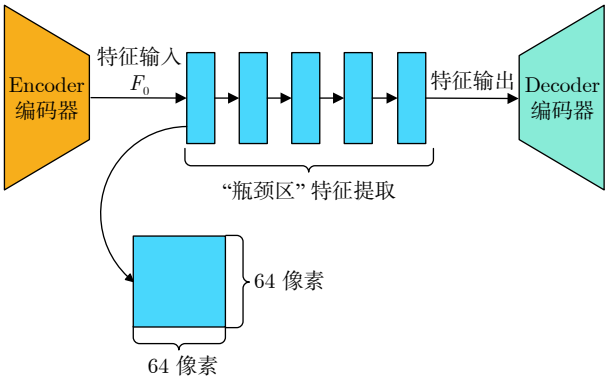


图 2 常规自动编码器

Fig. 2 Conventional autoencoder

如图 3 所示, 针对特征传递受阻问题, 我们对自动编码器结构中的 "瓶颈区"网络部分进行以下两步改进: 第 1 步, 多级特征压缩. 将编码器与解码器之间的 "瓶颈网络"中大小为 $h\times w\times c$ 像素的高级特征分别按照 0、2、4、8 压缩率进行缩放, 构建多级压缩特征, 即 F_0、F_{c2}、F_{c4} 和 F_{c8}. 越高压缩率的特征, 其尺度越小. 若按照特征尺度大小对多级压缩特征进行排列, 其结果为 $F_0>F_{c2}>F_{c4}>F_{c8}$. 多级压缩特征在特征表达方面是互补的, 越小尺度的特征中有着越小的结构特征空间, 网络更容易从中搜索出有意义的结构表达, 但是越小尺度特征越缺乏细节信息; 与之相反, 越大尺度特征中虽然在结构表达能力上更弱, 却有着越丰富的细节特征, 网络更容易从中搜索出有意义的细节表达. 因此, 大小尺度特征之间的这种互补性为第 2 步, 即多级注意力传播, 提供了巨大潜力. 多级注意力传播可以充分利用不同压缩特征对不同特征 (结构/细节) 表达方面的优势. 具体来说, 我们分别对各级压缩特征 F_{c8}、F_{c4}、F_{c2} 和 F_0 依次执行注意力匹配与替换, 得到注意力特征; 并依据从小尺度到大尺度的顺序对注意力特征进行分级传播. 如图 3 所示注意力特征 A_8 与压缩特征 F_{c4} 结合, 将小尺度注意力特征传播至更高尺度. 其后注意力特征 A_4 再以相同的过程传播至 A_2 和 A_0. 由于前一级注意力特征匹配替换

的结果总比后一级有更准确的结构表达; 后一级紧凑的压缩特征总比前一级有更多的细节特征. 因此, 多级注意力的传播方案可以促使网络在多个尺度下既保持图像结构准确, 又不断地丰富细节. 相比当前基于单级注意力的图像修复方案[1−2, 12], 我们的多级方案可以得到更加丰富的深度特征.

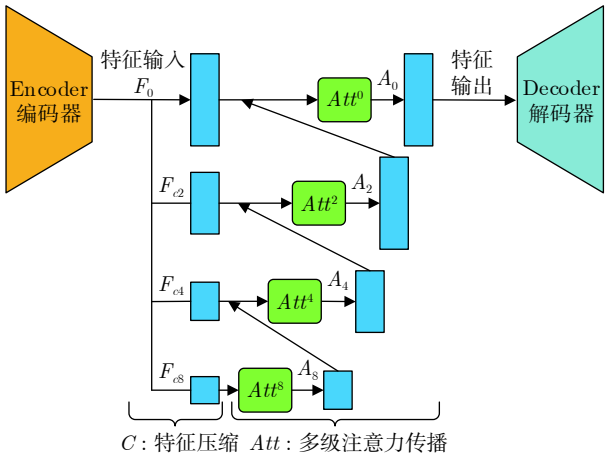


图 3　多级注意力特征传播自动编码器

Fig. 3　Multi-scale attention propagation driven autoencoder

同时, 与当前主流方法中由 “粗” 到 “细” 的多阶段方案不同, 我们期望在一个阶段内实现细粒度图像重建. 为此, 我们还提出了一种复合粒度判别器网络对图像修复过程进行全局语义约束与非特定局部密集约束. 其中, 全局语义约束由全局判别器实现, 该判别器的输出为一个评价图像整体真实度得分的值; 非特定局部密集约束由局部密集判别器实现, “非特定局部” 与 “密集” 体现在我们的局部密集判别器所执行的是对图像内多个相互重叠的局部区域进行密集地判别. 因此, 这种密集局部判别方式非常适合处理不规则破损情况下的修复任务.

在包括人脸、建筑立面和自然图像在内的多个数据集上进行的大量实验表明, 本文所提出的多级注意力传播驱动的生成式图像修复方法所生成的图像修复结果比现有方法拥有更高的图像质量.

综上所述, 本文的贡献如下: 1) 提出了一种端到端的图像修复模型, 该模型通过对全分辨率的图像上下文进行编码, 将提取的高级特征压缩为多尺度紧凑特征, 并依据尺度大小顺序驱动紧凑特征进行多级注意力特征传播, 实现了包括结构和细节在内的高级特征在网络中的充分传播. 2) 提出了一种复合粒度判别器, 对图像进行全局语义约束与非特定局部密集约束, 使得图像修复在单个前向过程中同时实现高质量的细粒度重建.

1　相关工作概述

1.1　传统图像修复方法

利用图像级低级非语义特征的传统图像修复方法[7, 13−18] 可分为两类: 基于扩散的方法和基于图像块的方法. 基于扩散的方法利用距离场等机制将图像信息从相邻像素传播到目标区域, 对于图像的小面积或类抓痕的窄缺损区域有着非常有效的修复效果. 当缺损区域面积过大或纹理变化很大时, 它们通常会生成明显的视觉伪影. 基于图像块的方法首先用于纹理合成, 然后扩展到图像修复. 与基于扩散的方法相比, 基于图像块的方法能够修复场景更复杂的图像. 通常, 基于图像块的方法采用迭代方式, 从同一图像的非缺损区域或外部图像库中采样相似的信息来填补缺损区域. 由于必须计算每个目标-源对的相似度分数, 因此此类方法需要大量的计算和内存开销. PatchMatch[3] 是一种典型的基于图像块的方法, 它通过快速最近邻域算法解决了这个问题, 极大地加快了传统算法的速度, 取得了较高质量的修复效果. 基于图像块的方法假设修复区域的纹理可以在图像的其他区域找到, 然而这种假设未必时时成立, 因此限制了该方法的应用范围; 此外, 由于缺乏对图像的高层语义理解, 基于图像块的方法无法为人脸等高度模式化破损图像重建出语义合理的结果. 因此, 无论基于扩散还是基于图像块的传统修复方法, 均不具备感知图像高级语义的能力.

1.2　基于深度学习的图像修复方法

近年来, 基于深度学习的图像修复方法从大规模数据中学习高级语义表示, 大大提高了修复效果. Context Encoder[19] 是最早用于语义图像修复的深度学习方法之一. 它采用自动编码器结构, 通过最大限度地降低像素级重建损失和对抗损失, 实现了对 128×128 图像中心区域存在的 64×64 矩形缺损区域的修复. 编码器将带有破损区域的图像映射到高级特征空间, 该特征空间用于解码器重构完整的输出图像. 然而, 由于通道维全连通层的信息瓶颈以及对图像局部区域缺乏约束, 该方法输出图像的重建区域往往出现明显的视觉伪影. Iizuka 等[11] 通过减少下行采样层的数量, 用一系列膨胀卷积层代替通道全连接层, 在一定程度上解决了上下文编码器的信息瓶颈问题. 同时, Iizuka 等[11] 还引入了一种局部判别器来提高图像的质量. 然而, 这种方法需要复杂的后处理步骤, 如泊松混合, 以增强孔边界附近的颜色一致性. Yang 等[12] 和 Yu 等[2] 将粗到细的卷积网络配置方案引入到了图像修复中. 该

方案在第 1 步使用深度卷积神经网络实现对破损区域的粗略估计. 进而, 在第 2 步的深度卷积网络中, 利用注意力机制或特征块交换操作, 搜索图像上下文中最相似的特征块并替换缺失区域内的特征块, 从而得到细化的输出结果. 然而, 这两种方案在不规则破损区域修复上并没有很好的泛化能力. Wang 等[1] 提出了一种用于图像修复的多列生成网络, 设计了置信值驱动的重建损失, 并采用了隐式多样马尔科夫随机场 (Implicit diversified Markov random field, ID-MRF) 正则化方案来增强局部细节. 它在矩形和不规则掩码上都取得了很好的效果. Liu 等[20] 在图像修复中引入部分卷积, 对卷积进行了掩盖和重新归一化, 仅利用非破损区域的有效像素, 有效地解决了基于卷积所带来的色差、模糊等伪影问题.

2 多级注意力传播网络

如图 4 所示, 我们提出的多级注意力传播网络由两部分组成: (a) 多级注意力传播生成器 G, (b) 复合判别器 D. 多级注意力传播网络生成器是针对图像修复任务改进的自动编码器, 通过编码过程、多级注意力传播过程与解码过程重建图像的破损区域. 复合判别器网络 D 通过将 G 生成的图像判别为 “假”来惩罚 G, 从而促进 G 生成真实图. 我们将从破损图像到完整图像的学习过程描述为一个映射函数, 该映射函数将破损图像流形 z 映射到完整图像流形 x. 为了简化符号, 我们还将使用这些符号来表示它们各自网络的功能映射.

2.1 多级注意力传播网络生成器

如图 4 所示, 我们的多级注意力传播生成器 G 主要由特征提取网络、多级注意力传播网络、上采样网络等 3 个子网络构成. 设 $I_{\text{input}} = z$ 和 $I_{\text{output}} = G(z)$ 为多级注意力传播网络生成器的输入和输出. 在浅层特征提取阶段, 提取浅层特征 F_{-1}:

$$F_{-1} = Enc(I_{\text{input}}) \tag{1}$$

其中 $Enc(\cdot)$ 为编码器网络. 该网络的编码器首先进行平坦卷积, 然后采用下采样与卷积操作对受损图像进行压缩编码.

其次, 将提取的有用局部特征 F_{-1} 进行特征细化:

$$F_0 = Bot(F_{-1}) \tag{2}$$

其中 $Bot(\cdot)$ 为由 4 层扩张卷积级联组成的 “瓶颈区”网络, 卷积核尺寸为 3×3, 膨胀率分别为 2、4、8、16.

接下来, 进行多级注意力传播. 注意力多级传播的第一步是将细化后的高级特征缩放为多级压缩特征:

$$F_{c8} = C^8(F_0) \tag{3}$$

$$F_{c4} = C^4(F_0) \tag{4}$$

$$F_{c2} = C^2(F_0) \tag{5}$$

其中 $C^n(\cdot)$ 为特征缩放操作, n 为缩放率, 表示特征尺寸缩放为原来的 $1/n$.

随后, 对压缩特征进行基于注意力的多级特征匹配与传播, 以小尺度结果引导后续处理:

$$A_0 = Att^0(A_2 \oplus F_0) \tag{6}$$

$$A_2 = Att^2(A_4 \oplus F_{c2}) \tag{7}$$

$$A_4 = Att^4(A_8 \oplus F_{c4}) \tag{8}$$

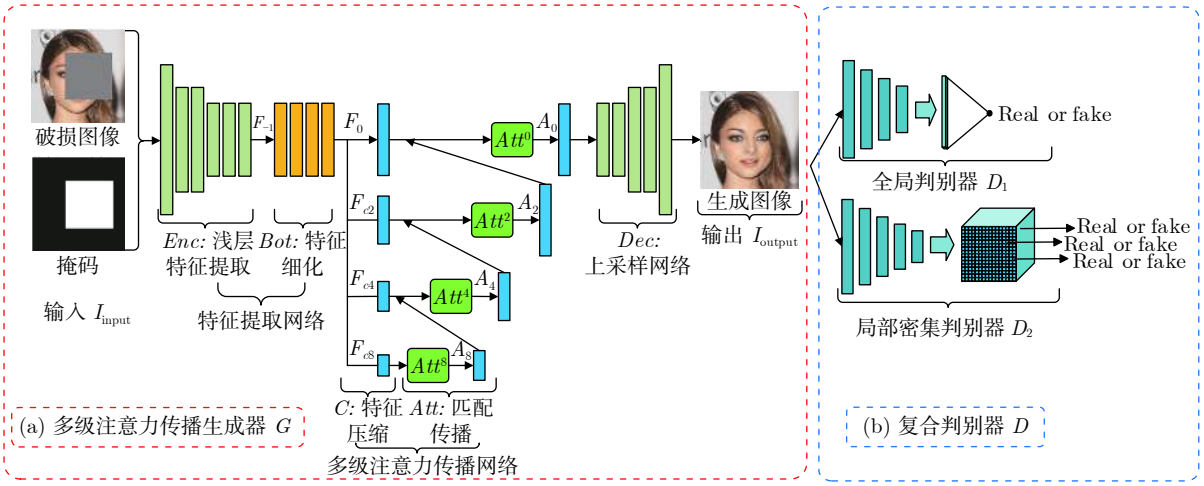


图 4 多级注意力传播网络整体框架

Fig. 4 The framework of multistage attention propagation network

$$A_8 = Att^8(F_{c8}) \tag{9}$$

其中 $\oplus$ 表示通道维叠加, $Att^l(\cdot)$ 为在压缩率为 l 的特征上进行的匹配替换与传播操作, 更多细节将在第 3.2 节中给出.

最终, 经过多级注意力特征配替换与传播后, 采用上采样网络将高级特征映射转化为完整的输出图像:

$$I_{\text{output}} = Dec(A_0) \tag{10}$$

其中 $Dec(\cdot)$ 为解码器网络, 对特征 A_0 进行两次上采样得到完整的重建图像.

2.2　基于注意力的特征匹配与传播

我们采用当前最先进的注意力特征匹配方案[2, 12, 21]. 注意力通常是通过计算缺失区域内外的图像块或特征块之间的相似度来获得的. 因此可以将缺失区域外的相关特征进行转移, 即通过相似度关系将图像上下文的图像块/特征块加权复制到缺失区域内部. 图 5 所示, $Att^l(\cdot)$ 首先从压缩特征 F_c 中学习区域亲和力, 即从 F_c 中提取特征块并计算破损区域内部特征块和外部特征块之间的余弦相似性:

$$s_{i,j}^l = \left\langle \frac{p_i^l}{\|p_i^l\|_2}, \frac{p_j^l}{\|p_j^l\|_2} \right\rangle \tag{11}$$

其中 p_i^l 是提取自 F_c 破损区域之外第 i 个特征块, p_j^l 为从 F_c 破损区域内提取的特征块. 然后用 softmax 对相似性进行处理, 得到每个图像块的注意分值:

$$a_{j,i}^l = \frac{\exp(s_{i,j}^l)}{\sum_{i=1}^{N} \exp(s_{i,j}^l)} \tag{12}$$

从高级特征图中获取注意分值后, 采用基于注意分值加权的上下文填充相似特征块中的破损区域:

$$p_j^l = \sum_{i=1}^{N} a_{j,i}^l p_i^l \tag{13}$$

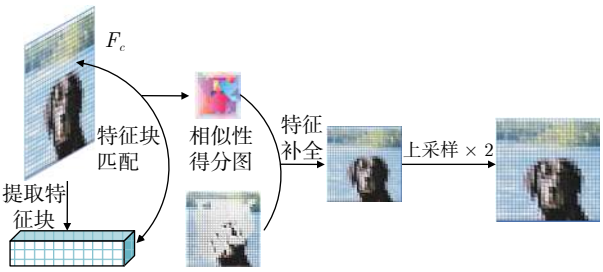


图 5　注意力特征匹配与传播

Fig. 5　Flowchart of attention feature matching and propagation

其中 p_i^l 为从 F_c 破损区域外提取的第 i 个特征块, p_j^l 为填充缺失区域的第 j 个特征块. 所有这些运算都可以表示为卷积运算, 用于端到端训练[2]. 我们将每一级 $Att^l(\cdot)$ 得到的特征进行上采样, 以引导下一层的注意力的传播. 这样的设计在保证图像结构在多个尺度上一致性的同时, 并逐级丰富图像细节. 值得注意的是, 在我们的方案中最紧凑的压缩特征的大小只有 $8\times8\times c$, 因此在注意力匹配的过程中无需额外的扩张卷积进行远距离特征借取.

2.3　复合判别器网络

作为生成网络的补充, 复合判别器网络 D 用于判断 G 生成的图像是否足够真实. 在图像修复中, 高质量的图像不仅取决于图像的整体特征, 还取决于图像局部对象的特征. 不同于全局与局部判别器来分别约束全局与局部破损区域, 我们设计了复合判别器来实现全局语义约束与非特定局部密集约束.

如图 4(b) 所示, 全局语义约束与非特定局部密集约束分别由全局判别器 D_1 与非特定局部密集判别器 D_2 来实现. 全局判别器由卷积层与全连接层构成, 输出为一个评价图像整体真实度得分的值. 非特定局部密集判别器类似 Patch-GAN[22] 结构, 由 5 个的步长卷积 (内核大小为 5, 步长为 2) 进行叠加构成. 输入由图像和掩模通道组成, 输出为形状为 $R^{h\times w\times c}$ 的三维特征图, 其中 h、w、c 分别表示通道的高度、宽度和数量. 然后, 我们将判别器的损失直接应用到判别器最后一层特征图的每个元素上, 形成针对输入图像局部不同位置的数量为 $h\times w\times c$ 的生成对抗网络. 复合判别器网络中全局判别器与非特定局部密集判别器在功能方面为相互补充的. 全局判别器针对全局的约束, 促使生成的图像破损区域与非破损区域在全局层面实现自然过渡; 而非特定局部密集判别器对图像内多个局部区域进行密集的有重叠的判别, 使得图像局部拥有丰富的细节纹理.

3　损失函数

损失函数由三部分组成: 1) 对抗损失 L_{adv}; 2) 特征匹配损失 L_{match}; 3) 重构损失 L_{rec}. 整体的目标函数可以表示为:

$$L = L_{\text{adv}} + \omega_1 L_{\text{match}} + \omega_2 L_{\text{rec}} \tag{14}$$

其中损失项的平衡参数 $\omega_1 = 1$、$\omega_2 = 1\,000$.

3.1　生成对抗损失 L_{adv}

我们方法采用改进的 Wasserstein GAN[23], 对

抗损失同时应用于网络 G 和网络 D, 最终影响生成网络 G 对破损图像的重构过程. 复合判别器网络 D 的输出值代表生成网络 G 的输出图像与真实图像的相似程度, 被用来惩罚并促使生成网络 G 生成更真实图像. 我们的复合判别器网络 D 由 D_1 和 D_2 组成. 对抗性损失可以表示为:

$$\begin{aligned} L_{adv} = & E_{x\sim pdata}\left[\log D_1(x)\right] + \\ & E_{x\sim pdata}\left[\log D_2(x)\right] + \\ & E_{z\sim pz}\left[\log(1-D_1(G(z)))\right] + \\ & E_{z\sim pz}\left[\log(1-D_2(G(z)))\right] \end{aligned} \tag{15}$$

3.2 特征匹配损失 L_{match}

特征匹配损失 L_{match} 用来比较判别器中间层的激活映射, 迫使生成器生成与真实图像相似的特征表示, 从而稳定训练过程, 这类似于感知损失[24−26]. 不同于感知损失比较从预先训练的 VGG 网络获取到来自真值图像与输出图像的激活映射, 特征匹配损失比较的是判别器中间层激活映射. 我们定义特征匹配损失 L_{match} 为:

$$\begin{aligned} L_{\text{match}} = & \sum_{i=1}^{L}\frac{1}{N_i}\left\|{D_1}^{(i)}(x)-{D_1}^{(i)}(G(z))\right\|_1 + \\ & \sum_{i=1}^{L}\frac{1}{N_i}\left\|{D_2}^{(i)}(x)-{D_2}^{(i)}(G(z))\right\|_1 \end{aligned} \tag{16}$$

其中 L 为判别器的最终卷积层, N_i 为第 i 个激活层的元素个数, ${D_1}^{(i)}$ 为判别器 D_1 第 i 层的激活映射, ${D_2}^{(i)}$ 为判别器 D_2 第 i 层的激活映射.

3.3 重建损失 L_{rec}

图像修复不仅要保证修复好的图像具有语义真实感, 而且要对图像进行像素级精确重建. 因此, 对于像素级重建过程, 我们定义了 $L1$ 重建损失:

$$L_{\text{rec}} = \|x - G(z)\|_1 \tag{17}$$

4 实验

4.1 数据集

我们使用 3 个面向于图像修复任务的国际公认通用图像数据集来验证我们的模型 (数据集分割如表 1 所示).

–Places2[27] 数据集: MIT 发布的数据集, 包含超过 800 万张来自 365 个场景的图像.

–CELEBA-HQ[28] 数据集: 来自 CelebA 的高质量人脸数据集.

–Facade[29] 数据集: 世界各地不同城市建筑立面集合.

表 1 3 个数据集的训练和测试分割

Table 1 Training and test splits on three datasets

数据集	训练	测试	总数
Facade	506	100	606
CelebA-HQ	28 000	2 000	30 000
Places2	8 026 628	328 500	8 355 128

4.2 实验设置

在 Windows 10 系统上使用 Python 开发编译了本文所提出方法的程序代码. 编译测试所用的深度学习平台软件配置为 TensorFlow v1.8、CUDNN v7.0 和 CUDA v9.0; 核心硬件配置为 Intel 8700 3.20 GHz 的 CPU, 12G NVIDIA TITAN XP 的 GPU. 我们使用 Adam 优化器对批量大小为 6 的模型进行训练, beta1 与 beta2 分别设定为 0 和 0.9. 在模型训练初始阶段的学习率设置为 1×10^{-4}, 随后再使用 1×10^{-5} 学习率对模型进行微调. 在模型训练过程中, 训练集中的全部图像均被缩放至 256×256 大小. 训练好的模型可在 CPU 及 GPU 上运行, 不论缺损面积大小, 修复过程在 Intel(R) Core(R) CPU 上平均运行时间为 1.5 秒, 在 NVIDIA(R) TITAN XP GPU 上平均运行时间为 0.2 秒. 本文中全部实验结果都是从训练好的模型中直接输出的, 未进行任何后期处理.

4.3 对比模型

我们将与以下经典主流方案进行比较:

–PatchMatch (PM)[3]: 一种典型的基于图像块的方法, 从周围环境复制类似的图像块.

–CA[2]: 一个两阶段的图像修复模型, 利用了高层次的上下文注意特征.

–MC[1]: 为图像修复模型设计了一个置信值驱动的重建损失, 并采用了隐式多样马尔可夫随机场正则化来增强局部细节.

5 结果与验证

5.1 实验结果

我们将本文方法与第 4.3 节中当前经典主流方案分别进行了定性和定量分析, 以证明本文方法的优越性.

定性比较. 图 6、图 7 和图 8 分别展示了我们的方法在 Places2、Facade 和 CelebA-HQ 数据集上和对比方法之间的对比结果. 在大多数情况下, 我们的图像修复结果比对比方法在结构重建方面表现得

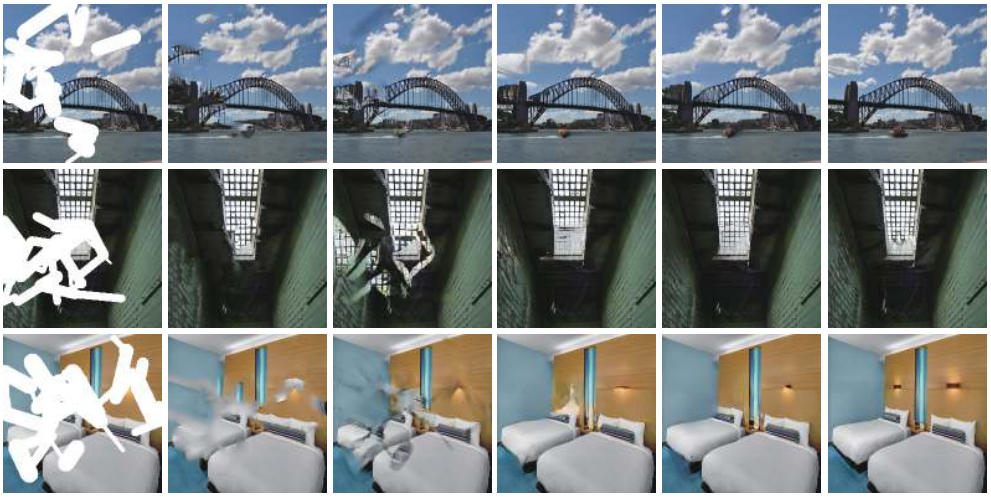


图 6　Places2 数据集上的结果比较

Fig. 6　Comparisons on the test images from Places2 dataset

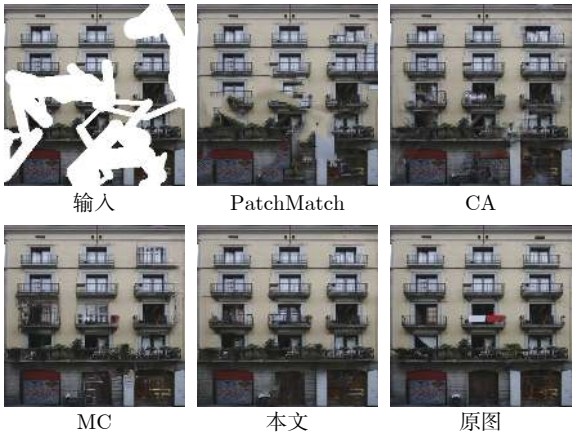


图 7　Facade 数据集上的结果比较

Fig. 7　Comparisons on the test image from Facade dataset

图 8　CelebA-HQ 数据集上的结果比较

Fig. 8　Comparisons on the test image from CelebA-HQ dataset

更准确合理. 与其他方法相比, 我们提出的方法在细节纹理重建上表现得更加细致.

定量比较. 我们使用 PSNR、SSIM 和平均 $L1$ 损失等指标来客观衡量修复结果的质量. 其中, PSNR 和 SSIM 可以大致反映模型重构原始图像内容的能力, 为人类的视觉感知提供了良好的近似. 平均 $L1$ 损失直接测量重建图像与真值图像之间的 $L1$ 距离, 是一个非常实用的图像质量评估指标. 如表 2 所示, 我们的方法在 Places2、CelebA-HQ 和 Facade 数据集中取得了最优的结果, 其中 SSIM、PSNR 是最高的, 平均 $L1$ 损失是最低的.

5.2　方案有效性分析

我们在建筑立面数据集上分别进行了两个分解实验来验证我们所提出方案的有效性. 为了更清楚地展示实验结果, 所有的实验均为矩形中心掩码情况下的图像修复结果.

1) 多级注意力传播的有效性

图 9(a) 为输入图像, 图 9(b) 为有注意力传播时的图像修复结果, 图 9(c) 为无注意力传播时的图像修复结果, 图 9(d) 为原图. 具体来说, 这次试验参与对比的分别为本文提出方案的结果与本文方案除去多级注意力传播时的结果. 可以看出在多级注意力传播的帮助下本文所提出的方案有着更准确的结构重建能力.

2) 复合判别器网络的有效性

如图 10(a) 为输入图像, 图 10(b) 为有复合判别器时的图像修复结果, 图 10(c) 为无复合判别器

表 2 CelebA-HQ、Facade 和 Places2 数据集上的定量对比

Table 2 Quantitative comparisons on CelebA-HQ, Facade and Places2

数据集	掩码率	PSNR			SSIM			Mean $L1$ loss		
		CA	MC	Ours	CA	MC	Ours	CA	MC	Ours
CelebA-HQ	10% ~ 20%	26.16	29.62	**31.35**	0.901	0.933	**0.945**	0.038	0.022	**0.018**
	20% ~ 30%	23.03	26.53	**28.38**	0.835	0.888	**0.908**	0.066	0.038	**0.031**
	30% ~ 40%	21.62	24.94	**26.93**	0.787	0.855	**0.882**	0.087	0.051	**0.040**
	40% ~ 50%	20.18	23.07	**25.46**	0.727	0.809	**0.849**	0.115	0.069	**0.052**
Facade	10% ~ 20%	25.93	27.05	**28.28**	0.897	0.912	**0.926**	0.039	0.032	**0.028**
	20% ~ 30%	25.30	24.49	**25.36**	0.870	0.857	**0.871**	0.064	0.052	**0.047**
	30% ~ 40%	22.00	23.21	**24.53**	0.780	0.815	**0.841**	0.084	0.068	**0.059**
	40% ~ 50%	20.84	21.92	**23.32**	0.729	0.770	**0.803**	0.106	0.086	**0.074**
Places2	10% ~ 20%	22.49	27.34	**27.68**	0.867	0.910	**0.912**	0.059	0.031	**0.029**
	20% ~ 30%	19.95	24.58	**25.05**	0.786	0.854	**0.857**	0.097	0.051	**0.048**
	30% ~ 40%	18.49	22.72	**23.41**	0.714	0.800	**0.805**	0.131	0.071	**0.066**
	40% ~ 50%	17.54	21.42	**22.29**	0.658	0.755	**0.765**	0.159	0.089	**0.081**

(a) (b) (c) (d)

图 9 有/无注意力传播时的图像修复结果

Fig. 9 Results with/without attention propagation

(a) (b) (c) (d)

图 10 有/无复合判别器时的图像修复结果

Fig. 10 Results with/without compound discriminator

时的图像修复结果, 图 10(d) 为原图. 可以看出在复合粒度判别器的帮助下本文所提出的方案有着更细腻的细节重建能力.

5.3 组件研究

为验证多级注意力机制以及复合粒度判别器网络的有效性, 我们以平均 $L1$ 损失为性能参考 (平均 $L1$ 损失越小性能越好), 进行了对比定量研究, 结果如表 3 所示. 其中, $Att0$ 至 $Att8$ 为注意力组件, Single-D 为单全局判别器, Cg-D 为本文所提出的复合粒度判别器.

表 3 组件有效性研究

Table 3 Effectiveness study on each component

$Att8$	无	有	有	有	有	有
$Att4$	无	无	有	有	有	有
$Att2$	无	无	无	有	有	有
$Att0$	无	无	无	无	有	有
Single-D	有	有	有	有	有	无
Cg-D	无	无	无	无	无	有
Mean $L1$ loss	0.091	0.089	0.086	0.081	0.078	0.074

从表 3 中我们可以看出, 多级注意力传播可以在很大程度上提升网络性能, 同时由于复合粒度判别器对全局语义与非特定局部的密集约束, 网络性能得到了进一步提升.

5.4 泛化应用研究

为进一步验证我们方法的泛化能力, 我们还通过对所提出模型进行对象移除实际应用研究.

如图 11 所示, 在示例 (a) 中, 我们尝试删除人脸图像中的眼镜. 我们可以看到本文方法都成功地删除了眼镜, 并在眼镜区域重建出了清晰自然的人眼. 在示例 (b) 中, 我们的模型将面部大面积区域移除, 并重建出合理的结果. 值得注意的是, 示例 (a) 与示例 (b) 人脸图像均不是正视前方, 而在训练过程中, 整个训练集中的非正视前方图像只占据少数, 这从侧面说明了本文方法具有良好的泛化能力.

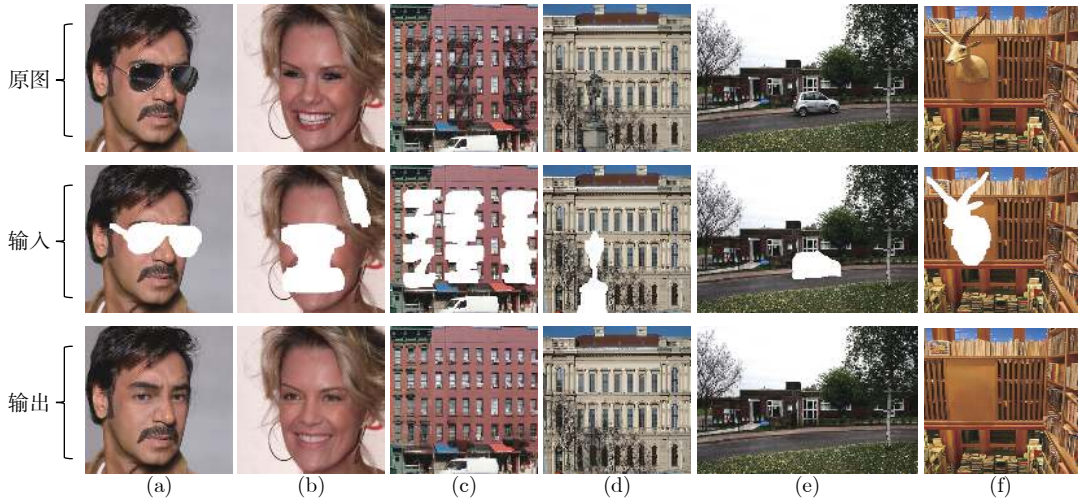


图 11　在 Facade、CelebA-HQ 和 Places2 数据集上的实例研究结果

Fig. 11　Case study on Facade, CelebA-HQ and Places2

更多的成功移除特定对象, 重建出高品质的结果见示例 (c)、(d)、(e)、(f).

6　总结

本文提出了一种基于层级注意力传播的图像修复网络. 为解决图像修复结果中的结构错乱与语义对象模糊问题, 我们提出将编码器编码的高级语义特征进行多尺度压缩和多层级注意力特征传播, 以实现包括结构和细节在内的高级特征的充分利用. 同时, 为实现在一个阶段内完成粗粒度与细粒度图像的同步重建, 我们提出了一种复合粒度判别器网络对图像修复过程进行全局语义约束与非特定局部密集约束. 大量实验表明, 与经典主流方法相比, 我们提出的方法可以产生更高质量的修复结果.

References

1　Wang Y, Tao X, Qi X, Shen X, Jia J. Image inpainting via generative multi-column convolutional neural networks. In: Proceedings of the 32nd Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2018. 331−340

2　Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T S. Generative image inpainting with contextual attention. In: Proceedings of the 2018 IEEE Conference on Computer Vision And Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 5505−5514

3　Barnes C, Shechtman E, Finkelstein A, Goldman D B. PatchMatch: A randomized correspondence algorithm for structural image editing. In: Proceedings of the ACM SIGGRAPH Conference. New Orleans, LA, USA: ACM, 2009. 1−11

4　Levin A, Zomet A, Peleg S, Weiss Y. Seamless image stitching in the gradient domain. In: Proceedings of the 8th European Conference on Computer Vision. Prague, Czech Republic: Springer, 2004. 377−389

5　Voronin V V, Sizyakin R A, Marchuk V I, Cen, Y, Galustov G G, Egiazarian K O. Video inpainting of complex scenes based on local statistical model. *Electronic Imaging*, 2016, **111**(2): 681−690

6　Park E, Yang J, Yumer E, Berg A C. Transformation-grounded image generation network for novel 3D view synthesis, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 702−711

7　Simakov D, Caspi Y, Shechtman E, Irani M. Summarizing visual data using bidirectional similarity. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008. 1−8

8　Yeh R, Chen C, Lim T Y, Hasegawa J M, Do N M. Semantic image inpainting with perceptual and contextual losses. arXiv: 1607.07539v2, 2016.

9　Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde F D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 2014 Advances in Neural Information Processing Systems, arXiv: 1406.2661v1, 2014.

10　Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 14111784, 2014.

11　Iizuka Satoshi, Edgar. S-S, Hiroshi I. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 2017, **36**(4): 107:1−107:14

12　Song Y, Yang C, Lin Z, Liu X, Huang Q, Li H, et al. Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 3−18

13　Ballester C, Bertalmio M, Caselles V, Sapiro G, Verdera J, et al. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 2001, **10**(8): 1200−1211

14　Bertalmio M, Sapiro G, Caselles V, Ballester C, et al. Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA: SIGGRAPH, 2000. 417−424

15　Efros A A, Freeman W T. Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA: SIGGRAPH, 2001. 341−346

16　Zhu Wei, Li Guo-Hui. Image completion based on automatic structure propagation. *Acta Automatica Sinica*, 2009, **35**(8): 1041−1047
(朱为, 李国辉. 基于自动结构延伸的图像修补方法. 自动化学报, 2009, **35**(8): 1041−1047)

17　Wang Zhi-Ming, Zhang Li. Local-structure-adapted image diffusion. *Acta Automatica Sinica*, 2009, **35**(3): 244−250
(王志明, 张丽. 局部结构自适应的图像扩散. 自动化学报, 2009, **35**(3): 244−250)

18 Meng Xiang-Lin, Wang Zheng-Zhi. Image diffusion based on visual masking effect. *Acta Automatica Sinica*, 2011, **37**(1): 21−27
(孟祥林, 王正志. 基于视觉掩蔽效应的图像扩散. 自动化学报, 2011, **37**(1): 21−27)

19 Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros A A. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2016. 2536−2544

20 Liu G, Reda F A, Shih K J, Wang T, Tao A, Catanzaro b. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 89−105

21 Zeng Y, Fu J, Chao H, Guo B. Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA: IEEE, 2019. 1486−1494

22 Hao D, Neekhara P, Chao W, Guo Y. Unsupervised image-to-image translation with generative adversarial networks. arXiv: 1701.02676, 2017.

23 Gulrajani I, Ahmed F, Arjovsky M, Vincent D, Aaron C. Improved training of wasserstein gans. In: Proceedings of the 2017 Advances in Neural Information Processing Systems, Long Beach, CA, USA: Curran Associates, Inc., 2017. 5767−5777

24 Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the 2016 European Conference on Computer Vision, Amsterdam, the Netherlands: Springer, 2016. 694−711

25 Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2016. 2414−2423

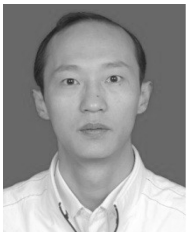
26 Gatys L, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems, Montreal, Quebec, Canada: Curran Associates, Inc., 2015. 262−270

27 Zhou B, Lapedriza A, Khosla A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(6): 1452−1464

28 Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv: 1710.10196, 2017.

29 Tyleček R, Šára R. Spatial pattern templates for recognition of objects with regular structure. In: Proceedings of the 2013 German Conference on Pattern Recognition, Münster, Germany: Springer, 2013. 364−374

曹承瑞 兰州理工大学硕士研究生. 主要研究方向为深度学习和图像处理.
E-mail: xiaocao1239@outlook.com
(**CAO Cheng-Rui** Master student at Lanzhou University of Technology. His research interest covers deep learning and image processing.)



刘微容 兰州理工大学教授. 主要研究方向为机器视觉与人工智能, 复杂系统先进控制理论与应用. 本文通信作者.
E-mail: liu_weirong@163.com
(**LIU Wei-Rong** Professor at Lanzhou University of Technology. His research interest covers machine vision and artificial intelligence, advanced control theory and application. Corresponding author of this paper.)

史长宏 兰州理工大学博士研究生. 主要研究方向为深度学习和图像处理.
E-mail: changhong_shi@126.com
(**SHI Chang-Hong** Ph. D. candidate at Lanzhou University of Technology. Her research interest covers deep learning and image processing.)

张浩琛 兰州理工大学电气工程与信息工程学院讲师. 主要研究方向为机器人传感与控制.
E-mail: zhanghc@lut.edu.cn
(**ZHANG Hao-Chen** Lecturer in the Department of Electrical and Information Engineering at Lanzhou University of Technology. His research interest covers sensing and control of robots.)