LRGAN: 一种运算轻量化图像修复网络

廖年鸿^{1,2}, 张效娟^{1,2*}, 彭春燕^{1,2}, 范虹³

1. 青海师范大学计算机学院, 西宁 810016;

2. 青海师范大学省部共建藏语智能信息处理及应用国家重点实验室, 西宁 810016;

3. 陕西师范大学计算机科学学院, 西安 710119

* E-mail: zhxj@qhnu.edu.cn

收稿日期: 2021-10-09; 接受日期: 2022-01-17; 网络版发表日期: 2022-03-01

青海省重点研发与成果转化项目(编号: 2021-GX-111)和国家重点研发计划重点专项(编号: 2020YFC1523305)资助

摘要 近来的图像修复工作大多聚焦于如何提升修复质量, 但并没有考虑在计算资源不足的平台如何运行. 本文提出了一种结合组卷积与注意力机制的模块(LABGC)用于替换传统卷积, 在此模块中利用深度可分离卷积实现了多层次图像修复工作, 同时针对传统卷积处理过程中通道间信息流动性的问题提出了轮循通道注意力分配机制(C2AM), 并将其应用于构建一个轻量级的图像修复网络(LRGAN). 在整体网络的设计阶段使用了并行判别器结构保证图像生成的全局与局部的一致性. 实验结果表明, 在保证修复质量的情况下, 本文提出的图像修复网络的推理时间与资源占用均远远小于同类方法.

关键词 人工神经网络, 图像修复, 轻量化网络, 图像合成, 深度学习, 计算机视觉

1 引言

图像修复(image inpainting)是一种对图像的可用区域进行综合分析并且最终填充缺损区域的图像处理任务. 传统的图像修复方法是通过从可用区域中选择物理位置上的近邻区域或者数学意义上距离的近邻区域填充到缺损区域或者将可用区域进行扩散来得到修复结果的. 不同于传统的图像修复方式, 基于深度学习的图像修复模型应用不同尺度的卷积神经网络(convolution neural network, CNN)特征生成整体更为合理的结果, 这在很大程度上促进了图像修复任务的发展. 近年来, 随着视觉Transformer(visual transformer, ViT)

在图像处理任务上的应用, 图像修复的质量有望进一步提升.

然而, 有两个主要问题制约了图像修复的进一步发展: (1) 图像修复的质量受限于图像修复方法本身; (2) 随着图片本身分辨率的增加, 需要修复的破损区域会随之增加.

针对第一个问题, 已有的图像修复方法可以分为4类. 第一类方法^[1~4]通过从图像的完好区域选取相关像素块或者通过对完好区域的扩散来填补破损区域, 这一类方法很难处理具有大量信息的自然图像. 大多数改进都是通过增加搜索区域或者将多个相关像素进行混合以确保修复质量, 但这也大大增加了运算复杂度.

引用格式: 廖年鸿, 张效娟, 彭春燕, 等. LRGAN: 一种运算轻量化图像修复网络. 中国科学: 技术科学, 2022, 52: 447–457

Liao N H, Zhang X J, Peng C Y, et al. LRGAN: A computational lightweight image inpainting neural network (in Chinese). Sci Sin Tech, 2022, 52: 447–457, doi: 10.1360/SST-2021-0462

第二类方法^[5,6]是通过数据驱动训练大量的数据集来对破损区域进行推理. 然而, 由于缺损区域信息的大量缺失, 对于该区域的端到端推理是十分困难的. 此外, 数据驱动的方法依赖于大量高质量的图像数据集, 这在训练过程中将花费大量的代价. 上面两种方法均只聚焦于像素或语义方面的修复, 都没有完全解决图像修复中的问题.

第三类方法^[7~14]尝试综合像素修复和语义层级修复的优点, 克服CNN在长距离信息表达中的不足, 相比于第一与第二类方法, CNN克服了类马尔科夫过程带来的无法获取图像全局信息以及不能并行的问题. 但相似地, 基于CNN的图像修复方法也有一些共性的问题. 首先, 不论是低维的像素层级还是高维的特征图层级都是由可用和不可用区域组成的, 在高维层级中, 甚至存在可用与不可用信息的叠加. 简单地将这些区域在高维与低维层级中使用相同的方法进行处理会导致修复模型忽略关键信息, 造成修复质量不佳. 其次, 随着网络层级的增加, 数据的维度扩张导致的高维稀疏矩阵会使对网络的进一步优化难以进行. 这些因素启发和引导了改进卷积层来提升图像修复质量的工作. 例如, 文献^[7]将部分卷积(PCConv)方法引入CNN结构中, 通过对遮罩部分的步进式更新来改进修复质量. 而文献^[9]则更进一步, 将门控卷积(GatedConv)的方法应用到图像修复任务中, 为每个更新区域赋予0~1之间的一个值, 进一步增强修复质量. 同样地, 也有基于网络结构的修改多层次获得信息提升修复质量的设计, 例如文献^[10]在RetinaNet网络的基础上提出了PENet, 它通过上下文注意力多层次地获得图像的信息, 并且通过注意力迁移机制完成对图像的修复; 文献^[12]提出了CA方法, 它通过两个阶段的修复网络对图像进行修复, 并通过上下文注意力层次从图像的较远区域提取近似待修复区域的特征.

第四类方法是近来从自然语言处理领域中移植并发展而来的视觉Transformer注意力^[15~17]. 这种基于多头注意力机制的方法优于卷积方法的地方在于, 卷积需要通过网络深度的增加使感受野逐步覆盖到整个图像, 而视觉Transformer注意力能够一次性获得整个图像的信息. 在这一点上视觉Transformer要远远优于传统的卷积结构, 一次性获得的图像信息中包含图像的纹理与结构, 这有助于神经网络理解图像的整体结构与语义信息, 将会大大促进图像修复相关方面的工作.

但视觉Transformer有三个需要解决的问题, 首先, 由于视觉Transformer技术是由自然语言处理领域中的Transformer技术移植而来的, 而在自然语言领域中, Transformer面向的输入是序列化的语句, 其本质上是一种一维张量, 直接将其应用于图像处理任务中时, 会将一个二维的图片拆分为多个一维张量进行处理, 这在执行效率上将会大打折扣, 同样也不利于获取图像的全局信息. 其次, 由于所处理任务的不同, 直接将NLP任务中的网络结构搬运到计算机视觉领域中会出现许多的层次冗余现象, 进一步导致执行效率的降低. 最后, 从实验结果来看, 已有的几个基于视觉Transformer构建的图像分类任务均存在计算量大、参数量高的问题, 这对训练设备的要求进行了进一步的提升, 大多数研究者难以承担这样的负担.

总而言之, 现有的图像修复方法主要依赖于大量的数据集以及在其他计算机视觉任务上网络结构的迁移进行修改并且应用, 获取图像的全局特征依然是图像处理任务的共性问题. 注意力机制, 尤其是视觉Transformer的出现给图像修复任务提供了新的解决方案, 但是由于其消耗运算资源的问题, 相较于工业界, 在学术界上较难有较大的运用.

针对第二个问题, 大多数有关图像研究的重心都聚焦于所面向任务质量的提升而非修复效率和运算资源的节约. 仅有一部分研究者对此进行了探索, MobileNet^[18]和GoogleNet^[19]将深度可分离卷积作为朴素卷积的一种有效替代. 然而, 训练过程带来的过拟合问题和高维空间的稀疏是分组卷积亟待解决的问题. 此外, 通道之间的信息流动性低下也限制了模型的进一步优化.

现行的主流图像修复网络为双阶段修复网络, 双阶段图像修复网络分为两个部分: 通常在首个阶段进行图像的大致修复, 得到图像的大致修复结果, 在第二个阶段依据第一阶段对修复结果进行精细化处理, 或者依据约束条件使之更为贴合原图像. 这一方式提升了修复结果的鲁棒性与准确性, 但这样做有几个缺陷. 其一, 图像修复网络需要构建得十分深, 不利于图像修复模型应用于边缘计算设备. 其二, 由于两个阶段之间的线性约束不一致, 比较难在两个阶段之间取得数学意义上的统一性, 而两者的约束函数规约一致时, 又失去了模型的可泛化性.

为了解决上述的问题, 本文结合了组卷积与注意

力机制各自的优势, 提出了轻量化的注意力组卷积模块, 同时结合了注意力机制能够获取图像全局信息优势与组卷积能够减少运算量与参数量的优点, 在有限的计算资源下取得了良好的图像修复结果. 在这一设计的基础上, 本文提出了一个单阶段图像修复网络, 这一网络利用改进的残差连接进行了编码层信息的复用, 进一步提升了图像的修复质量.

总的来看, 本文的贡献主要如下所述: (1) 通过结合组卷积与注意力机制设计了一种修复效果好并且更节约计算资源的轻量化注意力组卷积(lightweight attention-based group convolution, LABGC)模块, 这一模块执行高效修复的同时能够保证优秀的修复质量, 可以降低同尺寸朴素卷积的运算量与参数量, 并且该模块是可即插即用的, 可以用于替换其他神经网络中的朴素卷积层; (2) 本文设计了一个轻量级的图像修复模型——轻量化图像修复对抗生成网络(lightweight restoration generative adversarial network, LRGAN), 本次模型的训练以及测试均在单张NVIDIA Tesla P100 GPU上完成; (3) 本文所设计的网络结构, 采用结构相似度(structural similarity index, SSIM)和峰值信噪比(peak signal-to-noise ratio, PSNR)评价体系, 对LRGAN进行了定量、定性评估, 并且评价结果优于其他对比方法.

2 生成对抗网络的介绍

图像修复网络结构中, 一种重要的设计是對抗生成网络(generative adversarial network, GAN)结构, 它主要通过生成器(generator)以及判别器(discriminator)的互相作用逐步提升图像的修复质量. 其公式表示如下:

$$\begin{aligned} D &= \operatorname{argmin}_D \mathbb{E}_{x \sim p(x), z \sim q(z)} [f(D(x)) + g(D(G(z)))], \\ G &= \operatorname{argmin}_G \mathbb{E}_{z \sim q(z)} [h(D(G(z)))], \end{aligned} \quad (1)$$

其中, $x \in \mathbb{R}^{n_x}$, $z \in \mathbb{R}^{n_z}$, $p(x)$ 是原始数据集的分布, 而 $q(z)$ 是噪声的分布, 其中的 D 指判别器, 而 G 指生成器.

或者, 可以将式(1)中的 D 与 G 融合表示, 写为更简单的式(2)形式, 式(2)即为原始GAN的数学表达式:

$$\min_G \max_D \mathbb{E}_{x \sim p(x)} [D(x)] + \mathbb{E}_{z \sim q(z)} [1 - D(G(z))]. \quad (2)$$

GAN自被发现并应用以来就以其简单却深刻的思想与简明的结构大量应用于各类任务中, 在图像修复任务中也不例外, 包括PixelRNN, PGGAN等基于生成模型的图像修复网络都是以GAN为基础网络的. 相比于基于自编码器的图像修复方法, 基于生成模型的方法由于其输入噪声 z 的不同可以实现具有多种结果的修复, 满足不同场景下的需要. 但由于传统GAN是基于JS散度基础之上的优化结果, 在同时训练判别器与生成器时, 容易出现判别器梯度消失的问题, 使网络陷入局部最优或者造成生成器的梯度爆炸, 最终导致训练失败.

3 LRGAN神经网络的设计

3.1 网络整体架构设计

本文的模型在设计上沿用了對抗生成网络的主要设计, 通过生成器-判别器架构对图像进行修复, 构建了LRGAN网络. 与普通的對抗生成网络相区别的是, 本文在生成器部分使用一个编码-解码器结构替换了原本的生成器结构, 本文在判别器的设计部分使用了修改的PatchGAN判别器与Wasserstein判别器的并行结合, 通过两个判别器的权重共享对生成图像的质量与稳定性做进一步的约束. 本文设计的對抗生成网络在编码阶段使用五层连续的轻量化注意力组卷积模块, 在解码阶段使用五层连续的反卷积层最终得到生成图像. 其可视化表示如图1所示.

定义源图像为 I_{gt} , 定义遮罩为 I_m , 则遮罩的反码为 $1 - I_m$, 被遮罩的图像可以表示为 $I_M = I_{gt} \odot I_m$, 其中 $\odot$ 表示逐元素乘法运算符.

编码层的LABGC之间使用LeakyReLU激活函数, 在本文中LeakyReLU激活函数的 $\alpha=0.2$. 在编码器层级之间应用最大池化和谱归一化方法以降低运算复杂度以及参数量, 同时使生成的图像保留最为重要的部分(如, 边缘等).

在解码器部分, 本文使用了层级递增的方式, 通过反复对生成图像进行上采样来减少生成图像的棋盘效应, 并且与编码器中生成的特征实施了跳跃连接. 将解码器中的层次与其编码器中对应尺度的层次进行跳跃连接以进一步保证生成图像的质量与稳定性.

在判别器部分, 本文使用了PatchGAN判别器与

图 1 本文图像修复模型的设计可视化示意图

Figure 1 A visual display of proposed image completion network.

Wasserstein判别器的并行联合分支进行图像生成的判别, 其中PatchGAN用于把图像分为 $n \times n$ 个块, 对这 n^2 个块分别进行生成质量的评价, 而Wasserstein判别器在继承传统的GAN判别器的同时加入了正则化运算参数, 很大程度上避免了梯度消失问题的产生。

通过这种参数设置与网络设计, 本文生成的图像的像素点保证了其与周边的可识别区域的强关联性, 使生成的图像更为自然、清晰。

3.2 LABGC模块的设计

3.2.1 组卷积部分的设计

大量的卷积运算是深度学习模型中消耗时间与运算资源的主要原因。在现有的工作中, 许多方法通过削减网络层次达到轻量化, 然而这种方式同时削弱了模型的泛化能力和图像修复的效果。有一部分方法是使用分组卷积执行的, 但分组卷积主要是通过不同尺度特征的卷积方式汇聚不同尺度的特征, 而没有考虑通道的信息流动性。

在两个连续层之间存在一个可进行泛化的子网络能够提高CNN的特征提取能力。在LABGC中, 本文通过连续的 1×1 卷积来实现多层感知机的作用, 并在层次间使用LeakyReLU激活函数来促进特征的稀疏性。 1×1 卷积的连接促进了参数化的跨通道信息交换, 并使网络能够学习矩阵的特征。

多路径分支卷积最早是由GoogleNet的Inception模块实现的。基于Inception的设计, 本文设计了一个具有 3×3 卷积与 1×1 卷积并行分支, 在此基础上辅以全局

注意力确定本轮修复的更新区域。在这种情况下, 初始 1×1 层通过将输入特征卷投影到较小的维度来作为嵌入层, 减少了执行 3×3 卷积的运算量。通过将 3×3 分支卷积与深度可分离的核分解, 本文进一步减少了它的运算数量。与最开始的 1×1 卷积一样, 最后的 1×1 卷积表示由周围和全局特征点组成的特征图上的每个点, 确保局部和全局特征的一致性。其图例表示如图2所示。在本文的LABGC的设计中, 运用到注意力机制的部分分为两部分, 在整体的LABGC架构中所运用的主要是全局注意力, 3×3 分支卷积的作用是用于更新和填充区域, 这部分填充区域的指导是通过全局的注意力进行的, 通过全局自注意力方式加大了图像的贴合度, 使训练的迭代与收敛得到了加快, 减少了计算成本。通过这种方式替换传统的 3×3 卷积或者 5×5 卷积, 使参数量与运算量得到了节省。这一结论可以通过下述的公式计算。

其与原始 3×3 卷积的运算量与参数量之比可以记为

$$r_{\text{flops}}^{3 \times 3} = \frac{\frac{13}{4}C_1^2 + 2C_1C_2 + 9C_1 + C_2^2}{9C_1C_2}, \quad (3)$$

$$r_{\text{params}}^{3 \times 3} = \frac{2C_1 + 3C_2 + 10 + \frac{1}{4}C_1}{9C_2}, \quad (4)$$

其中, C_1 为当前层次的输入通道数量, C_2 为当前层次的输出通道数量。基于上述计算, 可以发现相比于传统 3×3 卷积与 5×5 卷积, 本文提出的模块在参数量与运算量上都具有相当的优势。本文将设计的模块应用于编

图 2 LABGC卷积架构

Figure 2 A visual display of lightweight attention-based group convolution.

码层, 在此条件下均有 $C_2=2C_1$, 在替换原始 3×3 卷积时最大可以节省原始卷积37.5%的运算量与54.1%参数量, 在替换原始 5×5 卷积时最大可以节省原始卷积77.5%的运算量与83.5%参数量。

3.2.2 通道注意力部分的设计

深度可分离卷积的提出将静态卷积的参数和运算复杂度降低了许多。然而, 分组卷积并不能解决通道之间的流动性, 从而导致神经网络模型参数的增加。随着网络深度的加深, 层间的卷积操作引入了一个高维的隐藏空间, 稀疏矩阵抑制了卷积核的学习。LABGC解决了卷积操作中参数和浮点操作的问题, 但并没有解决通道之间的信息流动性问题, 而且也带来了新的卷积核参数量的增加。为了使通道之间的信息更具有关联性, 并且使各个分组之间的信息都能产生联系, 本文设计了轮循式通道注意力机制。

如图3所示, 本文在LABGC中设计了轮循式通道注意力机制(cycle channel attention mechanism, C2AM)。在分组维度方面, AlexNet^[20]除了可以看作是第一个深度CNN之外, 也是第一个使用分组思想对图

像通道信息进行处理。在AlexNet提出信息分组的设计之初, 这一种操作只是为了将不同通道上的信息分别分配给不同的GPU进行训练, 并且最终再将信息通过Dense层进行聚合。GoogleNet^[19]对这一设计进行了强化, 通过大小不同的卷积核实现了对图像不同程度的感知。而ShuffleNet V2^[21]在上述基础上通过实验得到了几条重要的结论: 对卷积分组而言, 当分组的大小相同时, 内存访问代价(Memory Access Cost, MAC)最小; 当分支过多时, 会因为网络的碎片化而降低模型的并行度。C2AM遵循了上述两条原则, 在组卷积分支的设计中仅使用了 1×1 与 3×3 的分支, 并且将两个分支分流的通道大小都设为一致的, 即输入通道数量的50%。

在权重维度方面, C2AM参考了SENet^[22]中的设计。SENet面向的是图像分类任务, 它通过SE模块动态地学习每个特征通道的重要程度, 通过得到的通道重要程度提升对于识别物体具有帮助的特征, 而忽略那些对于识别帮助不大的特征。将此思想运用到图像修复任务上时, 显而易见的想法是, 当所有通道中的信息都尽可能与其周边特征结合良好, 并且通道之间具

图3 C2AM机制的设计

Figure 3 A visual display of cycle channel attention mechanism.

有强关联性时, 呈现的图像修复结果才会更好. 在此思想的指导下, C2AM的设计将通道注意力与全局注意力进行结合, 将计算得出的注意力分数低的通道与注意力分数高的通道通过两个分支进行分别处理, 以确保图像的局部与全局一致性.

C2AM机制将通道的内部与外部信息相比较, 并且将最终的比较结果进行排序. 注意力分数是根据区域对网络中高维特征的亲和力(affinity)来计算的. 本文对ACA机制中的注意力分数由余弦相似度定义. 设第 l 通道的破损区域为 R_i^l , 完好区域为 R_o^l , 两者之间的余弦相似度 d_l 可以表示为如下的形式:

$$d_l = \left\langle \frac{R_i^l}{\|R_i^l\|}, \frac{R_o^l}{\|R_o^l\|} \right\rangle, \quad (5)$$

$$e_L = e^M \sim 1 + M, \quad (6)$$

$$M = \frac{e_L^{d_l}}{\sum_{p=1}^N e_L^{d_p}}. \quad (7)$$

如式(6)所示, 对于任意 $M > -1$ 的情况, a_l 的非负性会得到保证. 通过将Softmax转为线性形式, 计算负担大大减少. 最终, 本文将得到的所有通道的注意力分数进行排序, 注意力分数最小的 n 个通道分流至 3×3 通道, 反之则分流至 1×1 通道. 在本文的实验中, n 设置为0.5倍当前通道数量, 并向上取整. 通过这种通道注意力方式与全局注意力方式的结合, 不仅保证了在整体上图

像呈现出一种好的修复效果, 也避免了由于某个通道的信息与外围完好的图像信息结合度不佳导致训练失效的问题.

3.3 损失函数的设计

本文使用多种损失函数的联合设计来确保生成图像的可靠性与一致性, 测量细化网络在原图和输出之间的语义和像素差异, 最终的优化目标如式(8)所示.

$$G = \underset{G}{\operatorname{argmin}} \underset{D}{\operatorname{max}} L_{\text{total}}, \quad (8)$$

其中, L_{total} 为总体损失函数的表示, 具体而言, L_{total} 由三部分损失函数组成, 分别包括L1损失、WGAN-GP损失以及PatchGAN损失, 其表示如式(9)所示.

$$L_{\text{total}} = \lambda_1 L_l + \lambda_2 L_p + \lambda_3 L_w. \quad (9)$$

由于基于GAN基础设计的网络容易出现梯度消失的问题, 而此问题主要源于判别器的训练过程中达到了局部最优的情况, 通过L1损失直接将生成图像与原始训练图像进行比较使得生成图像在过拟合前始终保持梯度, 一定程度上避免了GAN的梯度消失问题.

4 结果与分析

4.1 数据集与评价方法

本文应用了三个数据集作为训练与测试的数据

集, 分别为CelebA-HQ^[23]数据集、Places2^[24]数据集以及唐卡图像数据集。CelebA-HQ数据集是由香港科技大学收集的大量高清人脸数据集, 每张人脸图像的大小均为 1024×1024 ; Places2数据集是一个有着400多个分类, 100多万张自然场景图像的数据集, 每张图片的大小均为 512×512 ; 唐卡图像数据集为一个主要包括了青海唐卡数字图像的数据集, 共有30多类1000余张数字图像, 分辨率经过裁剪之后统一为 512×512 大小。

在CelebA-HQ数据集中, 本文随机选用了20000张人脸图像作为训练集, 并额外选择与训练集不重复的4000张图像作为测试集, 在可视化展示中选取两张有代表性的人脸图像作为测试集。

在Places2数据集中, 本文随机选用了40000张不同分类的场景图像作为训练集, 并且额外选择与训练集同类但是不重复的8000张图像作为测试集, 在可视化展示中选取两张有代表性的场景图像作为测试结果的展示。

在青海唐卡数字图像中, 选取了400张图片作为训练集, 100张图片作为测试集, 在可视化展示中选取一张有代表性的场景图像作为测试结果的展示。

公开数据集与私有数据集各有其特点, 其中, 公开数据集面向的任务是多样本、多分类的, 且由于自然图像固有的复杂性与弱边缘性质, 需要重构其图像特征所需的是网络对图像整体信息的理解与利用; 而唐卡数字图像表现出的特点是少分类、少样本的, 且由于绘画图像的强边缘性质, 对绘画图像的边缘与结构的理解则更为重要。

实验采用了PSNR与SSIM两种标准评价修复的图像。本文的实验均在Intel Xeon E5 CPU, Nvidia Tesla P100 GPU, 16GB运行内存的硬件环境下完成, 使用的Pytorch版本号为1.8.0, CUDA版本为10.2。

矩阵中心掩码虽然是图像修复任务中最为常见的图像破坏形式, 但在实际的图像修复场景中, 不规则的破坏则更为常见。因此在训练期间, 本文与6个对照组所使用的遮罩均为随机生成的不规则遮罩, 参数设定为随机范围, 覆盖全图像的25%~50%面积。

4.2 实验结果

本文先参考原始WGAN的参数设置, 然后加入LABGC和C2AM后对学习率、BatchSize等超参数进

行调整, 进行多次实验, 把内容损失和对抗损失结合的组合损失函数达到最小值的实验参数作为最佳参数。为了更好地比较各类方法修复数字图像的细微差别, 本文选择对二维数字图像放大特定区域以便观察重构细节。在本文最终的确定版本中, 损失函数部分的超参数分别设定为 $\lambda_1=\lambda_2=1$, $\lambda_3=0.5$ 。学习率设定为 10^{-4} , BatchSize在进行 512×512 分辨率训练时设定为10, 在进行 1024×1024 分辨率时设定为5。

在对照组的选取方面, 本文选取了近年来具有代表性的几类先进算法进行比较, 包括在网络结构上做出改进的PEN-Net^[12]与CA^[10], 在卷积层本身进行改进的部分卷积方法^[7]与门控卷积方法^[9], 以及同样基于GAN框架的PICNet^[25]以及基于编码-解码架构的LISK^[26]方法。

为了全面测试对照组与本文方法在不同分辨率下的表现, 本文选取了在Places2中的 512×512 分辨率和CelebA-HQ中的 1024×1024 分辨率的两类数据集, 在两类数据下分别比较对照组之间的表现。在对照组提供的原始文献中, LISK用于训练与测试的原始数据为 256×256 分辨率的, 本文将其通过输入与输出的修改分别重新训练 512×512 以及 1024×1024 分辨率下的表现。

PConv方法通过对遮罩部分的步进式更新来改进修复质量, 主要通过将图像的可用区域与不可用区域基于0-1掩码进行区分, 这一方法不利于对可用区域与不可用区域之间的交接区进行区分, 从实验结果中也可以看出基于此方法的修复结果并不理想。Gate-Conv方法在部分卷积方法的基础上进行了改进, 考虑到了交界区域的计算, 因此在 512×512 分辨率上的结果优于本文提出的方法, 但在高分辨率修复状态下不如本文提出的方法。PEN-Net方法基于RetinaNet架构, 在人脸数据修复的场景中修复的效果较好, 但在街景的测试结果中, 去除人及影子时均出现了伪影。CA方法通过引入内容感知层与全局-局部判别器方法提升了修复区域与原始图像的结合度。PICNet方法结合了VAE和GAN架构, 采用一种概率策略来生成多元化的修复结果, 但由于其约束相对其他的同架构网络较弱, 因此修复的效果不如本文所提出的方法。LISK方法采用了多任务架构, 应用了共享生成器和并行判别器方法以同时修复图像的内容和结构, 其修复效果好于本文提出的方法, 但其修复时间较长。

通过比较图4中的实验结果, 可以发现基于

图4 结果可视化展示. 左起: 遮罩原图, 原图, CA, PConv, GatedConv, LISK, PICNet, PEN-Net, LRGAN(本文方法). 第一行与第二行为Places2数据集中的数据, 第三行与第四行为CelebA-HQ数据集中的人脸数据, 最后两行为唐卡数据集数据

Figure 4 A visual display of result. From left to right: Masked GT, GT, CA, Pconv, GateConv, LISK, PICNet, PEN-Net, LRGAN. Lines 1-2 from Places2, lines 3-4 from CelebA-HQ, lines 5-6 from proposed Thangka dataset, zoom in for best view.

LRGAN网络的图像修复算法在细节上具有更清晰的纹理结构和更准确的色彩还原. 表1显示了通过不同的算法构建的PSNR和SSIM指标的平均值, 以及每个算法的参数数量与运算时间. 可以看出, 无论是在低分辨率下还是较高的高分辨率图像数据集上, 本文所提出的算法都能以较低的运算时间得到近似于其他算法的结果.

究其原因, 相比于其他修复方法, 本文通过层次内通道注意力机制对各个通道的修复质量都做了评价, 又通过全局注意力的特征图更新了每一轮的修复区域, 因此, 无论是通道之间还是特征之间的关联度都得到了了一定程度的提升.

与通过一系列卷积运算提供的渐进式传播相反, 自注意力机制计算任意两个位置之间的直接相互作用, 而与距离无关. 自注意力机制允许网络感知那些分散在不同位置, 却有结构关联的区域, 学习图像的上下文关系, 使生成的图像更具有全局一致性, 从而提升重建图像细节, 提高图像的修复质量.

此外, 为了验证本文提出的两种机制的有效性, 设计了消融实验, 分别在不使用深度可分离卷积、仅使用深度可分离卷积、同时使用深度可分离卷积与C2AM机制三种情况下进行实验, 以验证本文模型的有效性 with 可靠性. 该消融实验均在训练集图像分辨率为 512×512 条件下进行, 除编码层部分应用层次不同之外, 其余参数均相同. 从实验结果中可以看出, 深度可分离卷积在网络的设计中起到了大幅度减少参数量与运算时间的作用, 但相应地, 作为代价, 修复图像的质量出现了明显的下降. 通道注意力一定程度上解决了这一问题, 通过运算时间与参数量的少量增加使修复质量得到了一定的提升. 全局注意力则进一步提升了图像的修复质量.

5 结论

本文提出了一种结合组卷积与注意力机制的模块(LABGC)用于替换传统卷积, 在此模块中结合深度可

表 1 图像修复的定量参数比较^{a)}

Table 1 A quantitative comparison of image completion

使用方法	512×512			1024×1024			参数量 (M)
	SSIM	PSNR	时间 (ms)	SSIM	PSNR	时间 (ms)	
PEN-Net ^[10]	0.8283	26.10	35	0.8131	25.66	289	14.7
PConv ^[7]	0.8401	26.53	41	0.8119	24.98	143	51.6
CA ^[12]	0.8436	25.58	62	0.8298	25.32	610	9.0
GatedConv ^[9]	0.8659	27.18	78	0.8612	25.51	696	37.4
LISK ^[26]	0.8721	26.53	61	0.8362	25.82	255	15.1
PICNet ^[25]	0.8501	25.12	58	0.8335	24.37	178	9.1
LRGAN (ours)	0.8571	26.63	34	0.8411	26.12	111	7.3

a) 加粗字体表示最优结果

表 2 图像修复的消融实验比较^{a)}

Table 2 An ablation study of image completion

实验编号	全局注意力	深度可分卷积	通道注意力	传统卷积	SSIM	PSNR	时间 (ms)	参数量 (M)
5×5VConv				√	0.8487	26.45	485	10.4
3×3VConv				√	0.8291	24.17	268	7.6
DWC		√			0.8224	23.76	76	3.7
DWC+C2AM		√	√		0.8386	25.96	97	4.3
DWC+C2AM+NLA (ours)	√	√	√		0.8411	26.12	111	7.3

a) 加粗字体表示最优结果

分离卷积实现了多层次图像修复工作, 同时针对传统卷积处理过程中通道间信息流动性的问题提出了轮循通道注意力分配机制(C2AM), 并在C2AM模块中应用线性注意力方法进一步降低了运算的复杂度, 并将LABGC应用于构建一个轻量级的图像修复网络(LRGAN). 实验证明了在资源受限的平台上显示了本文的模块在设计计算效率高但视觉上准确的模型方面

的有效性. 此外, 基于组卷积的方法被广泛用于超分辨率、图像去噪的领域中, LABGC模块具有能够提取上一层浅层特征的特性, 并且通过前一层次的全局注意力提取与本层次中的深度可分卷积结果进行结合, 未来可以尝试将LABGC模块进行再次修改, 例如针对超分辨率功能, 可以通过对LABGC模块的输入输出尺寸进行修改, 以使得其可以适应不同的图像处理任务.

参考文献

- Grossauer H. A combined PDE and texture synthesis approach to inpainting. In: Pajdla T, Matas J, eds. Computer Vision-ECCV 2004. ECCV 2004. Lecture Notes in Computer Science, vol 3022. Berlin, Heidelberg: Springer, 2004
- Rane S D, Sapiro G, Bertalmio M. Structure and texture filling-in of missing image blocks in wireless transmission and compression applications. *IEEE Trans Image Process*, 2003, 12: 296–303
- Criminisi A, Pérez P, Toyama K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process*, 2004, 13: 1200–1212
- Shen J, Chan T F. Mathematical models for local nontexture inpaintings. *SIAM J Appl Math*, 2002, 62: 1019–1043
- Hays J, Efros A A. Scene completion using millions of photographs. *ACM Trans Graph*, 2007, 26: 4

- 6 Whyte O, Sivic J, Zisserman A. Get out of my picture! Internet-based inpainting. In: British Machine Vision Conference (BMVC). London, 2009
- 7 Liu G, Reda F A, Shih K, et al. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018
- 8 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv: [1511.06434](#)
- 9 Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019
- 10 Zeng Y, Fu J, Chao H, et al. Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019
- 11 Pathak D, Krähenbühl P, Donahue J, et al. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016
- 12 Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018
- 13 Broad T, Grierson M. Light field completion using focal stack propagation. In: SIGGRAPH'16: Special Interest Group on Computer Graphics and Interactive Techniques Conference. Anaheim, 2016
- 14 Sagong M, Shin Y, Kim S, et al. Pepsi: Fast image inpainting with parallel decoding network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019
- 15 Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. arXiv: [2101.01169](#)
- 16 Lin T, Wang Y, Liu X, et al. A Survey of Transformers. arXiv: [2106.04554](#)
- 17 Chen Q, Zhao H, Li W, et al. Behavior sequence transformer for e-commerce recommendation in Alibaba. In: Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data. Anchorage, 2019
- 18 Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: [1704.04861](#)
- 19 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015
- 20 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. *Commun ACM*, 2017, 60: 84–90
- 21 Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical guidelines for efficient cnn architecture design. In: Ferrari V, Hebert M, Sminchisescu C, et al., eds. Computer Vision–ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11218. Cham: Springer, 2018
- 22 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 7132–7141
- 23 Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, 2015. 3730–3738
- 24 Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition. *IEEE Trans Pattern Anal Mach Intell*, 2017, 40: 1452–1464
- 25 Zheng C, Cham T J, Cai J. Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 1438–1447
- 26 Yang J, Qi Z, Shi Y. Learning to incorporate structure knowledge for image inpainting. *Proc AAAI Conf Artif Intell*, 2020, 34: 12605–12612

LRGAN: A computational lightweight image inpainting neural network

LIAO NianHong^{1,2}, ZHANG XiaoJuan^{1,2}, PENG ChunYan^{1,2} & FAN Hong³

¹ School of Computer, Qinghai Normal University, Xining 810016, China;

² State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining 810016, China;

³ School of Computer and Science, Shaanxi Normal University, Xi'an 710119, China

Currently, most image inpainting approaches focus on improving the quality of inpainted images while ignoring the requirement for a low-end platform. In this work, we propose a lightweight group convolution module in place of vanilla convolution. This module, named lightweight attention-based group convolution (LABGC), uses a depth-wise convolution to achieve multi-scale image inpainting. Meanwhile, to alleviate the information constraint between channels, we propose a cycle channel attention mechanism. Then in our proposed LRGAN inpainting network, we use LABGC. We use parallel discriminators in the overall network design to ensure local and non-local consistency. Experiments show that our model is as accurate as state-of-the-art approaches in terms of speed and FLOPs.

artificial neural network, image inpainting, lightweight network designation, image synthesis, deep learning, computer vision

doi: [10.1360/SST-2021-0462](https://doi.org/10.1360/SST-2021-0462)