用于图像修复的金字塔上下文编码器网络

钟伟东
21821073
Zhejiang University
21821073@zju.edu.cn

摘要（Abstract）

高质量图像修复需要在受损图像中缺失部分填入合适的内容。现有的研究要么通过复制图像补丁patch，要么从区域背景生成小补丁来填充区域，而忽略了对视觉和语义可信性都有很高要求的事实。该文章提出了一种基于深度生成模型的金字塔上下文图像编码网络(PEN-Net)。PEN-Net是在U-Net结构的基础上构建的，它通过对全分辨率输入的上下文语义进行编码，将学习到的语义特征解码回图像，从而实现对图像的恢复。具体地说，它通过注意从高级语义特征图中逐步学习区域亲和力，并将学习到的注意转移到先前的低级特征图中。由于缺失的内容可以通过金字塔式的从深到浅的注意力转移来填补，从而保证了填充中图像的视觉和语义的一致性。进一步提出了一种具有深度监督金字塔损失和对抗性损失的多尺度解码器。这样的设计不仅能在训练中快速收敛，而且在测试中也能得到更真实的结果。对各种数据集的大量实验表明，该网络具有较好的性能。

1. 介绍

图像修复是指利用图像中已知区域的内容来学习，填充图像中受损缺失的区域，例如老化，污渍等。同样也可以利用图像修复来去除照片中不感兴趣的部分，例如拍摄的照片中出现了路人挡住了背景，可以将这一区域去除，作为缺失区域，利用图像修复技术来填充背景。

目前图像修复领域，依旧存在较大的挑战。对于修复结果，需要有两个方面的要求（1）符合图像的整体结构，（2）细致的图像纹理。传统图像修复算法存在部分局限，在下个小节会具体介绍到。

为了保证视觉和语义的一致性，建议同时在图像和特征层填充区域。首先，文章采用U-Net[4]结构作为主干，它可以将上下文从低级像素编码为高级语义特征，并将这些特征解码回图像中。具体来说，文章提出了一个金字塔上下文编码器网络(PEN-Net)与三个定制的关键组件，即金字塔上下文编码器、多尺度解码器和对抗性损失，用于提高U-Net在图像修复方面的能力。其次，一旦压缩后的潜在特征被编码到图像中，金字塔上下文编码器在解码之前会在金字塔路径中填充从高级语义特征到低级特征(具有更丰富的细节)的区域。为此，提出了一种注意转移网络(ATN)来学习高层次特征图中缺失区域内外的补丁之间的相似性，然后进行转移（即通过相似性加权参数），使得相关特征图补丁从外部进入到内部，从而生成更高分辨率的图。第三，提出的多尺度解码器以ATNs通过跳接连接的特征和潜在特征作为输入，进行最终解码。PEN-Net通过最小化深度监督金字塔L1 损失和对抗性损失进行优化。

2. 相关工作

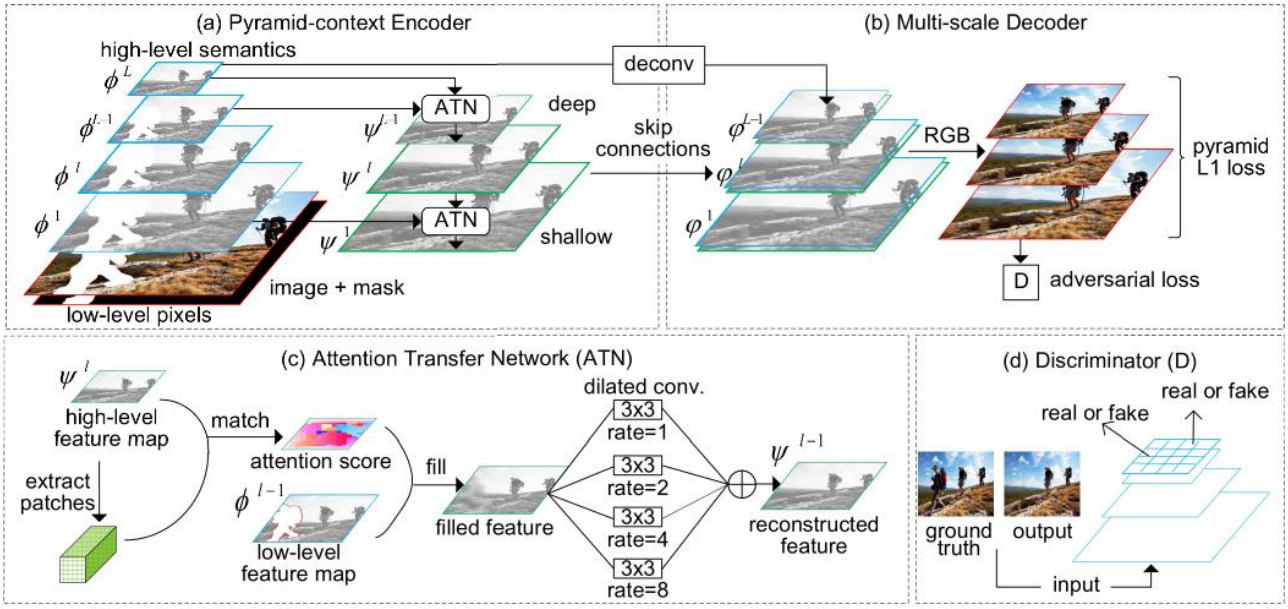
2.1. 基于偏微分方程的图像修复

这类方法是早期使用传统方法。其中具有代表性的有BSCB模型[1]，TV模型[2]，CDD模型[3]。2001年，BSCB模型由四位学者共同提出，其主要思想是将图像修复过程类比热扩散运动，模拟手工修复的过程，通过延伸等照度线把完整区域的信息各向异性向待修复区域扩散。该算法计算复杂度较高，且仅适用于较小待修复区域的图像，对于大的待修复区域的修复效果较差。TV(total variation)模型的基本思想是通过求解最小能量泛函完成对图像待修复区域的修复。其算法实现简单，效率高，收敛速度快。该方法同时能够达到去除图像噪声的目的，能较好的保持图像边缘，但是在图像特征信息分布均匀的区域，容易引入“阶梯效应”。

为了解决TV模型存在的问题，Tony.Chan等人提出了基于曲率驱动扩散（CDD）的图像修复模型。“阶梯效应”的产生，是由于TV模型中，扩散系数只与梯度值有关。CDD模型将图像特征信息的曲率引入到扩散强度中，使得满足了“连接整体性准则”。CDD模型图像修复能力要比TV模型强，能修复较大区域，但其模型收敛性较差，需要较大的计算量，修复效率低。

2.2. 基于补丁的图像修复

基于patch的纹理合成方法首先在[6,7]被提出了。然后[5]将它们应用于图像修复中，以填充图像级别的缺失



区域。他们通常根据补丁之间的距离度量(例如欧式距离,SIFT 距离等)，从数据库或未损坏的环境中采样并粘贴类似的补丁到缺失的区域。Bertalmio等人提出结合补丁纹理合成技术与基于扩散传播图像分解[8]的方法。许多方法试图通过提供更好的填充顺序或最佳补丁来提高性能[10,11]。PatchMatch被提出用于快速查找图像补丁之间的相似匹配。基于补丁的图像绘制方法能够生成与上下文相似的清晰结果。然而，由于缺乏对图像的高水平理解，基于补丁的方法很难产生语义上合理的结果。

2.3. 基于深度生成模型的图像修复

图像的深度生成模型通常将图像编码为一个潜在的特征，在特征级填充缺失的区域，然后将特征解码回图像。近年来，基于深度生成模型的研究取得了可喜的成果。基于深度特征学习和对抗性训练的上下文编码(Context Encoder)是最早的深度生成模型之一，能够给出语义合理的填充结果[11]。引入引导损失，使解码器生成的特征图与编码器生成的ground-truth特征图尽可能接近。Iizuka等人引入扩张卷积[12]来增加网络的感受域。设计了PConv和ShCNN等特殊卷积运算，消除了图像掩码区域占位符值的影响。提出了一种基于上下文注意层和补丁交换层的高层特征图，利用未受损区域的相似补丁填充缺失像素。受图像风格化的启发，MNPS提出了在推断时使用一个预先训练的分类网络来优化纹理细节。Isola等人尝试用一个通用的图像转化框架来解决图像填充中的问题。利用高级语义特征学习，深度生成模型能够为缺失区域生成语义一致的结果。然而，从一个紧凑的潜在特性生成视觉上真实的结果仍然具有挑战性。

3. 金字塔上下文编码网络PEN-Net

金字塔上下文编码器网络(PEN-Net)由三部分组成(如图所示)，即金字塔上下文编码器，多尺度解码器，判别器。PEN-Net是建立在U-Net结构上的，U-Net可以将带掩码的受损图像编码为紧凑的潜在特征，并将特征解码回图像。

金字塔上下文编码器可以通过填充从潜在特征到低层特征(具有更高的分辨率和更丰富的细节)的缺失区域，进一步提高编码效率。在解码前，根据编码器的深度，通过重复使用所提出的注意转移网络(ATN) (c)多次来填充缺失区域。具体来说，ATN从高级语义特征中学习缺失区域内外补丁之间的相似性，并将学习到的注意力转移到补丁区域（通过相似性从上下文中加权复制)，使得具有更高分辨率。在ATN中，通过四组不同比率的膨胀卷积进一步聚合多尺度信息，细化填充特征。最后，多尺度解码器将ATNs通过跳跃连接的特征和潜在的特征作为输入进行解码。对抗性损失与金字塔L1 损失被用来逐步细化解码器在所有尺度上的预测输出。

3.1. 金字塔上下文编码器

为了提高编码器的效率，提出了一种金字塔上下文编码器，用于在解码前填充缺失区域。一旦学习了一个紧凑潜在特征，金字塔上下文编码器就会以金字塔的方式重复使用ATN，从而填充从高级语义特征到低级特征(具有更高的分辨率)的区域。在假设语义相似的像素具有相似的细节的前提下，每一层都使用一个ATN从高级语义特征中学习区域相似性，从而可以进一步指导相邻层中缺失区域内外的特征转移，具有更高的分辨率。

如图给出了L层的金字塔上下文编码器，其特征图从深到浅标记为$\varphi^{L}, \varphi^{L-1}, ... \varphi^{1}$。由深到浅每层由ATNs构造的特征记为:

$$\begin{aligned} \psi^{L-1} &= f(\phi^{L-1}, \phi^{L}), \\ \psi^{L-2} &= f(\phi^{L-2}, \psi^{L-1}), \\ &\cdots, \\ \psi^{1} &= f(\phi^{1}, \psi^{2}) = f(\phi^{1}, f(\phi^{2}, ...f(\phi^{L-1}, \phi^{L}))), \end{aligned} \tag{1}$$

其中，将ATN的操作表示为f。通过这种跨层的注意力转移和金字塔填充机制，可以保证缺失区域的视觉和语义一致性。f(即ATN)介绍如下

3.2. 注意力转移网络

注意力机制通常通过缺失区域内外的补丁之间的区域(通常为 3×3)相似度来获得，因此可以将缺失区域外的相关特征进行转移到内部区域。图所示(c), ATN首先从高层特征图ψ^{l}学习相似性映射。它从ψ^{l}中提取补丁，计算内部外部补丁之间的余弦相似性计算:

$$s_{i,j}^{l} = \langle \frac{p_i^l}{\|p_i^l\|_2}, \frac{p_j^l}{\|p_j^l\|_2} \rangle, \tag{2}$$

其中p_i^l在ψ^{l}掩模外部的第i个补丁，p_j^l是在ψ^{l}掩模内部的第j个补丁。然后用softmax计算相似性，得到每个补丁的注意值。

$$\alpha_{j,i}^{l} = \frac{\exp(s_{i,j}^{l})}{\sum_{i=1}^{N} \exp(s_{i,j}^{l})}. \tag{3}$$

从高层特征图中获取注意分后，在相邻低层特征图上的补丁可以用注意分值加权得到相应的填充。

$$\alpha_{j,i}^{l} = \frac{\exp(s_{i,j}^{l})}{\sum_{i=1}^{N} \exp(s_{i,j}^{l})}. \tag{3}$$

计算所有补丁之后,可以获得一个由注意力转移的填充。值得说明的是，所有这些操作都可以被表示为卷积操作，用于端到端训练。

$$p_j^{l-1} = \sum_{i=1}^{N} \alpha_{j,i}^{l} \; p_i^{l-1}, \tag{4}$$

进一步细化ATN中的特性如图中的(c)所示。具体来说，多尺度上下文信息可以通过四组不同速率的膨胀卷积进行聚合。这样的设计保证了结构在最终特征的一致性，从而提高了测试中的填充效果。

3.3. 多尺度解码器

提出的多尺度译码器将ATNs通过跳跃连接与编码器的潜在特征作为输入。多尺度解码器生成的特征图从深到浅标记为是$\varphi^{L}, \varphi^{L-1}, ... \varphi^{1}$。公式如下

$$\begin{aligned} \varphi^{L-1} &= g(\psi^{L-1} \oplus g(\phi^{L})), \\ \varphi^{L-2} &= g(\psi^{L-2} \oplus \varphi^{L-1}), \\ &\cdots, \\ \varphi^{1} &= g(\psi^{1} \oplus \varphi^{2}), \end{aligned} \tag{5}$$

其中g表示反卷积操作，表示特征图级联，ψ^{l}表示第L层ATN重建的特征。

一方面，ATN生成的特征图为缺失区域编码了更低层次的信息。这样的设计使解码器能够生成具有细粒度细节的视觉上真实的结果。另一方面，通过卷积得到的潜在特征能够在缺失区域合成新的目标，即使在缺失区域之外找不到目标。通过这两种特征，解码器能够结合图像的上下文，在语义和纹理上合成具有高度一致性的新填充。例如，所提出的解码器能够在双眼被蒙住的情况下合成人脸图像中的眼睛。

3.4. 金字塔L1 损失

提出了深度监督金字塔L1 损失，以逐步完善预测的缺失区域在每个规模。具体而言，每个金字塔损失为特定尺度的预测与真实值之间的归一化L1 距离：

$$L_{pd} = \sum_{l=1}^{L-1} \left\| x^{l} - h(\varphi^{l}) \right\|_1, \tag{6}$$

3.5. 对抗性损失

图像填充是一个不确定性的问题，对于缺失的区域有许多可能的结果，使用对抗性训练来选择最真实的区域。对抗训练通常包括生成器(G)和鉴别器(D)，目的是实现纳什均衡，使生成器生成的虚假数据与鉴别器生成的真实数据无法区分。如图 (d)所示，金字塔上下文编码器和多尺度解码器构成一个生成器，采用PatchGAN[10]作为的鉴别器。使用频谱归一化的方法来稳定训练鉴别器[16]。定义生成器为：

$$z = G(x \odot (1 - M), M) \odot M + x \odot (1 - M), \tag{7}$$

其中x为真实值，$\odot$为element-wise乘积。M是掩模，其中 1 表示缺失区域，0 表示内部区域。判别器的对抗性损失表示为：

(a) Input (b) PatchMatch (c) GL (d) CA (e) PConv (f) Ours

$$L_D = \mathbb{E}_{x\sim p_{data}}\left[max(0, 1 - D(x))\right] + \mathbb{E}_{z\sim p_z}\left[max(0, 1 + D(z))\right], \tag{8}$$

其中是D(x)与D(z)是D的逻辑输出。生成器的对抗性损失为

$$L_G = -\mathbb{E}_{z\sim p_z}\left[D(z)\right]. \tag{9}$$

整个PEN-Net通过最小化对抗性损失与金字塔L1 损失来进行优化。定义整个目标函数为

$$L = \lambda_G L_G + \lambda_{pd} L_{pd}. \tag{10}$$

4. 实验与结论

4.1. 比对实验

在四个不同特征的数据集上进行实验

- Facade:世界各地不同城市的高结构化图片。
- DTD,一种收集了 47 种可描述的纹理。
- CELEBA-HQ,一个高质量的人类人脸数据集,来自CELEBA。
- Places2,一个包含来自自然世界的 365 个场景的图像的数据集。

将以下先进的方法作为基线与该篇文章方法进行比较

- PM:一种典型的基于拼合的方法,从周围的环境中复制类似的补丁。
- GL:一种生成模型,它利用了全局和局部鉴别器来完成图像的完成。
- CA:一种两阶段的图像修复模型,它利用了上下文的高层注意力特征。
- PConv:生成模型,它提出了一个特殊的卷积层来填充不规则的区域。

4.2. 结论

如图所示，典型的基于补丁的方法PatchMatch能够生成清晰的纹理，但是扭曲的结构与周围区域不一致，而GL、CA、PConv等深层生成模型在最终的结果中往往会生成模糊的纹理。在跨层注意力转移和金字塔填充机制的帮助下，该模型能够生成语义合理，视觉逼真，纹理清晰，结构与上下文一致的结果。

References

[1] Bertalmio, Marcelo, Sapiro, et al. Image inpainting 2000, 4(9):417--424.

[2] Tony F. Chan, Jianhong Shen. Mathematical Models for Local Nontexture Inpaintings. SIAM Journal on Applied Mathematics, 2001, 62(3):1019-1043.

[3] Chan T F, Shen J. Nontexture Inpainting by Curvature-Driven Diffusions Journal of Visual Communication & Image Representation, 2001, 12(4):436-449.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234–241, 2015.

[5] Alexandru Telea. An image inpainting technique based on the fast marching method. Journal of graphics tools, 9(1):23–34, 2004.

[6] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In SIGGRAPH, pages 341–346. ACM, 2001.

[7] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In ICCV, pages 1033–1038. IEEE, 1999.

[8] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. TIP, 12(8):882–889, 2003.

[9] Radim Tyleˇcek and Radim ˇS´ara. Spatial pattern templates for recognition of objects with regular structure. In GCPR, pages 364–374, 2013.

[10] Yonatan Wexler, Eli Shechtman, and Michal Irani. Spacetime completion of video. TPAMI, (3):463–476, 2007.

[11] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In CVPR, pages 2536–2544, 2016.

[12] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In CVPR, pages 1125–1134, 2017.

[14] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In ICLR, 2018.