Computer Science and Application 计算机科学与应用, 2022, 12(1), 83-94
Published Online January 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.121010

Hans汉斯

基于掩码Transformer的图像修复网络

康延亭，王直杰

东华大学，上海

收稿日期：2021年12月21日；录用日期：2022年1月18日；发布日期：2022年1月25日

摘　要

现有的基于深度学习的图像修复网络通常采用注意力机制以相似匹配的方式将完好区域信息填充到待修复区域来提升待修复区域的纹理细节。然而，现有的注意力机制的度量方式仅考虑特征纹理而缺少对语义的理解以至于会利用到一些语义不相似区域的信息。为了解决这一问题，本文提出一种基于掩码transformer的图像修复网络，该掩码transformer模块相较于基本的transformer层的区别主要包括两部分：1) 通过掩码将特征图分为有效区域和无效区域并提出掩码注意力机制有效的建模待修复区域和完好区域的相似性；2) 提出用查询集和相似度矩阵加权融合的方法为待修复区域精确填充信息。与传统的注意力机制相比，基于transformer的方法能够较为有效的提升修复的纹理效果。

关键词

掩码，注意力机制，Transformer，查询集，相似度矩阵

An Image in Painting Model Based on Mask Transformer

Yanting Kang, Zhijie Wang

Donghua University, Shanghai

Received: Dec. 21st, 2021; accepted: Jan. 18th, 2022; published: Jan. 25th, 2022

Abstract

Existing deep learning-based image repair networks usually use an attention mechanism to fill intact area information into the area to be repaired in a similar matching manner to improve the texture details of the area to be repaired. However, the existing measurement method of attention mechanism only considers the feature texture and lacks the understanding of semantics, so that it will use the information of some semantically dissimilar regions. In order to solve this problem, this paper proposes an image restoration network based on mask transformer. The difference

文章引用: 康延亭, 王直杰. 基于掩码 Transformer 的图像修复网络[J]. 计算机科学与应用, 2022, 12(1): 83-94.
DOI: 10.12677/csa.2022.121010

between the masked transformer module and the basic transformer layer mainly includes two parts: 1) The feature map is divided into valid regions and invalid regions by mask, and the mask attention mechanism is proposed to effectively model the similarity between the regions to be repaired and the intact regions; 2) A method of weighted fusion of query set and similarity matrix is proposed to accurately fill in information for the region to be repaired. Compared with the traditional attention mechanism, the transformer-based method can effectively improve the texture effect of repair.

Keywords

Mask, Attention, Transformer, Query Set, Similarity Matrix

1. 引言

图像修复是计算机视觉领域的一个长期而关键的问题，该方法的目的是通过利用完好区域和语义的内容来填补图像中缺失的部分[1]。高质量的图像修复可以广泛应用于许多领域，例如物体去除，照片恢复，图像伪造和图像重构等。

传统的图像修复算法主要依据图像像素间等相关性和内容相似性来进行推理修复[1]-[11]。但这些方法仅适合简单的纹理图像，当图像内容较为复杂时，很难通过这种方法生成合理的语义结果。近年来，基于深度学习的方法[12]-[30]开始主导这一领域。[14] [17]等模型通过编码器和生成对抗网络能够生成合理的修复图像。但在缺失区域卷积运算时难以利用完好区域信息，这使得修复图像常存在视觉模糊的情况。[19] [24] [28]等通过注意力机制的方法提高了网络的纹理细节修复效果。注意力机制借鉴传统方法中的块匹配思想，通过度量缺失区域特征块与完好区域特征块的相似性并将高相似度的完好区域匹配到缺失区域的方式有效提升了修复图像的纹理细节。尽管如此，注意力机制的度量方式在计算时仅考虑特征纹理而忽略了语义是否相似，这使得网络较容易地引入语义不相似区域的信息。这种语义不相似区域的信息会使得填充后的待修复区域存在伪影、模糊的现象。

Transformer，作为语言任务中的结构，在许多计算机任务中也正在兴起。与卷积神经网络(CNN)相比，该结构摒弃了固有的归纳偏置，通过多头注意力模块[31]进行长期的交互。[32]的一些初步工作也证明了该结构在自然图像合成建模结构关系方面的能力(全局建模感知能力)。

受 transformer 对全局建模感知能力的启发，本文提出一种基于掩码 transformer 的图像修复网络。其中掩码 transformer 部分通过替代常规的注意力机制以达到精准利用全局语义相似区域信息。本文提出的掩码 transformer 首先将特征图切分成若干同尺寸小块，并且根据掩码划分为完好块和待修复块。其中完好块按顺序生成一个查询集，完好块和待修复块都通过线性投影转换成令牌(Token)并嵌入位置信息。令牌通过本文设计的掩码注意力机制学习待修复块与完好块间的相似性信息并通过全连接层和 softmax 激活层生成相似度矩阵，最后将相似度矩阵与查询集加权融合并填充到对应待修复块处。其中本文设计的掩码注意力机制与传统 transformer 中的自注意力方式不同，该机制的查询(Q)为待修复块，而键(K)和值(V)都使用完好块。这种方式使得待修复区域仅关注完好区域中相似信息的位置，从而达到对完好区域全局建模的作用。

论文主要贡献包括：

1) 提出掩码注意力机制，通过该机制有效地建立了待修复区域与完好区域的全局相似性信息。

2) 提出使用查询集和相似度矩阵的方式以精确填充待修复区域信息。

2. 相关工作

2.1. 传统图像修复方法

传统的图像修复方法分为基于填充的方法[1] [2] [3] [4]和基于扩散的方法[3]-[8]。基于填充的方法通常利用不同的度量方式(例如，欧几里得距离，SIFT 距离[10]等)寻找相似的完好区域并填充。基于扩散的方法通常利用不同的变分算法将完好区域信息传播到缺失区域。例如 Bertalmio 等人[2]提出使用三阶偏微分方程来模拟平滑传输过程的 BSCB (Bertalmio-Sapiro-Caselles-Ballester)模型。全变分模型[4]通过将待修复区域周围的边缘信息向待修复区域扩散的方法修复图像。PatchMatch [2]通过利用在完好区域寻找待修复区域的相似块并填充的方法修复图像。但是这些传统的图像修复方法有效的前提是背景区域中含有待修复区域的色块且结构较为简单。由于这些方法不能捕获高级语义信息，所以在非重复性结构、背景未包含待修复区域的情况下未能修复出较为合理的区域。

2.2. 基于深度学习的图像修复方法

用于图像修复的深度学习模型通常将图像编码为低维特征，在特征级别填充缺失区域，最后将特征解码回图像。上下文编码器[14]通过生成对抗网络能够为语义填充提供合理的结果。特殊的卷积运算(如部分卷积[18]，门控卷积[21])旨在通过设定合理的掩码以屏蔽缺失区域对图像修复效果的影响。能够建模长距离信息的注意力模块(例如上下文注意力模块[19]，金字塔注意力模块[26])通过匹配填充与缺失区域相似度较高的完好区域信息以达到丰富纹理细节的作用。这些注意力模块通常采用余弦相似度求相似性，这种度量方式存在特征语义表示失真和语义歧义的缺点，这使得建模长距离信息仍存在一定的挑战性。

3. 基于掩码 transformer 的图像修复网络

该网络由生成器和鉴别器组成。其中生成器为两段式设计，本文提出的掩码 transformer 用于第二阶段中。后续 3.1 节中介绍掩码 transformer 的详细结构，3.2 节介绍整体网络结构，3.3 节介绍损失函数。

3.1. 掩码 Transformer

现有的注意力模块通过相似匹配填充的方式以丰富待修复区域内容信息。具体来说，将待修复区域和完好区域分为大小相同的参考块和背景块，然后将两种块通过度量方式确定块间的相似度，最后将相似度高的背景块填充到参考块的位置。该方式虽然能够较好的将完好区域的信息填充到待修复区域，但现有的度量方式在全局相似度匹配时仅考虑局部块间的纹理相似度，并不考虑块的语义相似问题，这导致匹配的背景块存在不同程度的语义差异性。针对上述问题，本文借鉴 Chen 等人[33]将图像转化为一维序列的方式设计了一种掩码 transformer 模块。如图 1 所示，该模块包括查询集、掩码 transformer 结构、全连接层、softmax 激活函数和相似度矩阵。其中查询集通过将输入切分成小块并展平得到。经过掩码 transformer 结构学习得到待修复块的全局相似信息，然后通过全连接层将全局相似信息进一步增强，最后通过 softmax 得到相似度矩阵，通过该矩阵与查询集加权得到最终修复的结果。

具体来说，该模块首先对输入进行预处理：给定输入特征，将其切分为小尺寸块并根据掩码分为完好块和待修复块，其中完好块组成查询集 Q。然后将完好块和待修复块通过线性投影变成完好块令牌和待修复块令牌并嵌入位置信息。

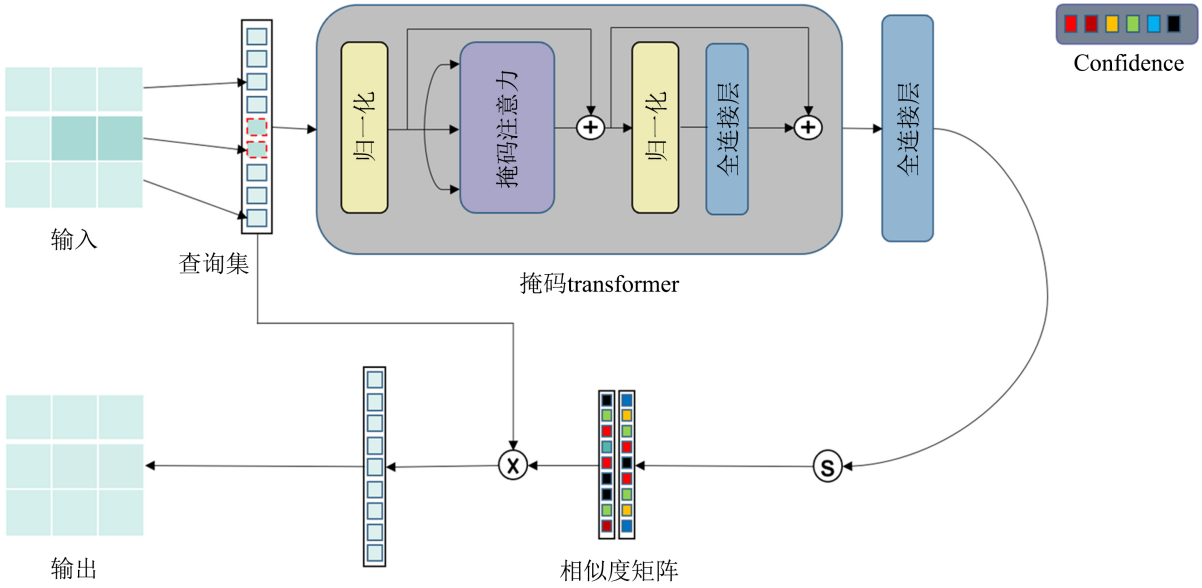

Figure 1. The structure of masked transformer module
图 1. 掩码 transformer 模块框架图

然后通过掩码 transformer 结构学习相似度矩阵。给定完好块令牌 R_b 和待修复块令牌 R_m，待修复块令牌 R_m 与完好块令牌 R_b 进行掩码注意力操作获取缺失块相似度可表达为：

$$I_R = LN\left(RMA\left(R_m, R_b\right)\right) + R_m \tag{1}$$

其中 LN， RMA 分别表示层归一化[34]，掩码多头注意力[31]。I_R 代表待修复块相似度信息。其中掩码多头注意力借鉴自注意力结构。不同的是，掩码多头注意力的查询符号(Q)为经过线性投影的待修复块令牌 R_m，键(K)，值(V)为经过线性投影的完好块令牌 R_b。

掩码注意力模块如图 2 所示，该模块将待修复块经过线性投影得到 Q，完好块经过线性投影得到 K 和 V，然后通过点乘和缩放操作获取待修复块投影 Q 与完好块投影 K 的相似性信息，然后经过 softmax 表示出相似度较高的信息，最后通过与完好块投影 V 进行点乘进一步获取更精准的相似信息。

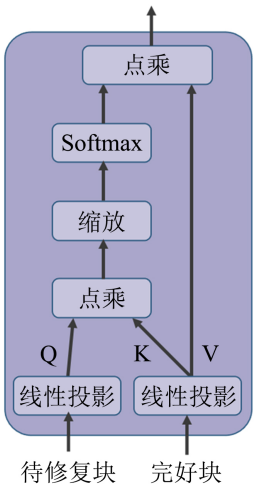

Figure 2. Mask attention module
图 2. 掩码注意力模块

针对区域 a，给定该区域的待修复块令牌 R_m^a，完好块令牌 R_b^i，其中 i 代表第 i 个区域块。掩码多头注意力可表达为：

$$head_j = soft\max\left(\frac{R_m^a W_Q^j \left(R_b^i W_K^j\right)^T}{\sqrt{d}}\right)\left(R_b^i W_V^j\right) \tag{2}$$

$$RMA = \left[head_1, \cdots, head_n\right] W_o \tag{3}$$

其中 h 为头的个数，W_Q^j，W_K^j，W_V^j 为三个可学习的线性层，$1 \le j \le h$。W_O 也是一个可学习的线性层，W_O 可以融合多头的输出以获取更精准的多头信息。获取多头信息后通过归一化和全连接层对信息进一步增强。

最后输出的信息通过全连接层进行维度变换以获取与相似度矩阵相同的维度，然后通过 softmax 激活函数将该相似度信息约束到 0~1 之间生成相似度矩阵。对输出的相似度信息 I_R 通过上述操作获取相似度矩阵 C_R 可表示为：

$$C_R = \sigma\left(MLP\left(I_R\right)\right) \tag{4}$$

其中 σ 代表 softmax 函数，MLP 代表全连接层，相似度矩阵 C_R 的维度为特征图中局部块的个数。

得到相似度矩阵之后通过加权融合查询集 Q 以得到待修复块的填充特征信息。对于输入的第 l 个待修复块的相似度矩阵 C^l，得到填充后的待修复块 O^l 可表达为：

$$O^l = \sum_{m=1}^{M} Q_m \cdot C_m^l \tag{5}$$

其中 M 为全局完好局部块的个数，Q_m 代表第 m 个完好块的信息，C_m^l 代表第 l 个待修复块对应的与第 m 个完好块的相似度。

每个待修复块通过查询集和相似度矩阵加权融合后能够有效的丰富待修复区域的纹理信息。本文的掩码 transformer 结构通过掩码将特征分为完好区域和待修复区域并用完好区域建立查询集，然后通过 transformer 结构学习相似度矩阵并加权融合查询集以填充待修复区域信息。这种通过将特征块线性投影成令牌再进行相似性学习的方式缓解了常规注意力操作中只考虑局部纹理相似而忽略语义相似的缺点。并且通过建立查询集的方式以将完好区域信息精确的填充到待修复区域。

3.2. 网络结构

在生成器阶段，我们的网络全部采用门控卷积。如图 3 所示，粗修复生成器包括编码器、四组级联扩张门控卷积和解码器，其中扩张卷积的扩张率为 2，4，8，16。解码器生成的图像含有粗略的纹理轮廓，精细化的细节修复在细阶段生成器。

细修复生成器包括编码器，一个局部分支、一个全局分支和解码器。其中局部分支采用四组级联扩张门控卷积，扩张卷积的扩张率同上。全局分支采用掩码 transformer 模块。具体来说，输入的特征图切分成 4*4 的小块并通过掩码确定完好块和待修复块。其中完好块信息组成一个查询集，并且将完好块和待修复块线性投影成完好块令牌和待修复块令牌。将完好块令牌和待修复块令牌通过掩码注意力模块建模相似性信息并输出为相似度矩阵。该相似度矩阵与查询集加权融合得到待修复块的填充信息。最后将局部分支与全局分支的信息组合并输出到解码器中以生成精细化图像。

本文还采用主流的 Patch-GAN 作为鉴别器，其中鉴别器堆叠了六个内核大小为 5，步长为 2 的卷积以捕获马尔可夫块的特征统计信息。并且通过该方法堆叠的卷积得到的输出图中每个神经元的感受野可以覆盖整个输入图像。本文还在在鉴别器中采用谱归一化[25]来进一步稳定生成对抗网络的训练。

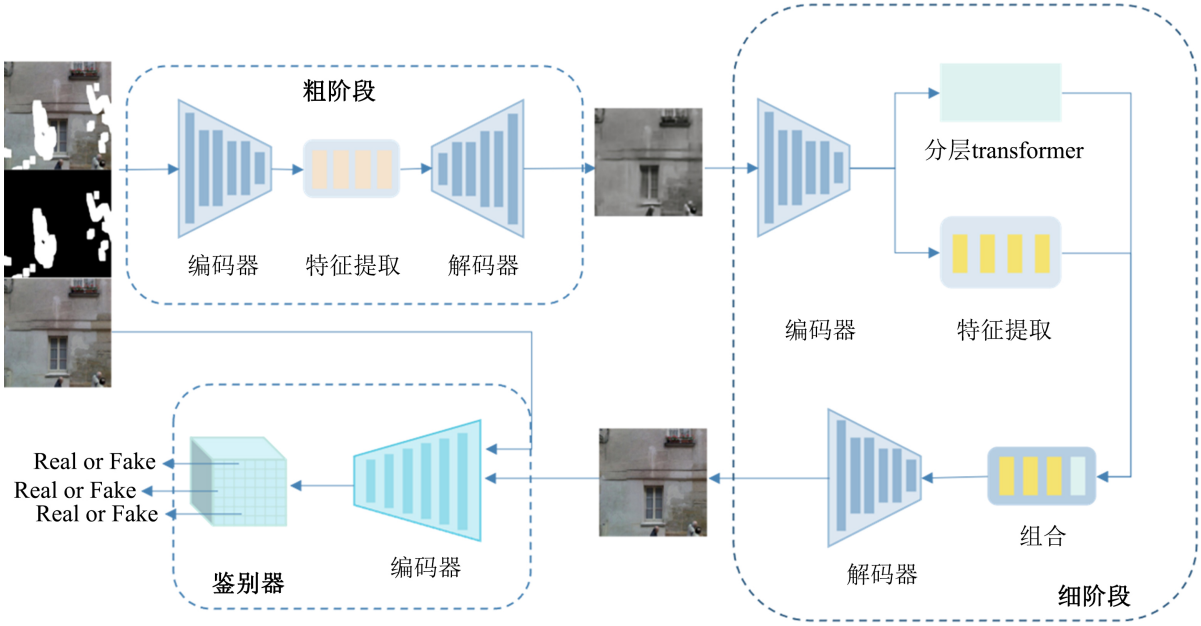

Figure 3. Image in painting network framework diagram based on mask transformer
图 3. 基于掩码 Transformer 的图像修复网络框架图

3.3. 损失函数

本文的损失函数包括对抗损失、感知损失和重构损失，总的损失函数可以表示为：

$$L = \lambda_1 L_{adv} + \lambda_2 L_{per} + \lambda_3 L_{rec} \tag{6}$$

其中 λ_1， λ_2， λ_3 三个超参数分别为 10，5，10。

对抗损失 由于图像修复对于缺失区域具有多种可能的结果，所以本文使用对抗训练来选择最真实的结果。对抗网络包括一个生成器(G)和一个鉴别器(D)，其中生成器 G 包括两部分，第一部分是粗阶段生成器 G_c，第二部分是细阶段的生成器 G_r。其目的是实现纳什均衡，使得鉴别器无法区分生成器生成的图像和真实图像。首先定义来自生成器的最终预测结果：

$$Z_c = G_c\left(I \odot (1-M) + M\right) \odot M + I \odot (1-M) \tag{7}$$

$$Z_r = G_r\left(Z_c \odot (1-M) + M\right) \odot M + I \odot (1-M) \tag{8}$$

其中 I 是真实图像，$\odot$ 是点乘运算，M 是掩码，掩码中 1 表示缺失区域，0 表示完好区域。鉴别器损失定义为：

$$L_{dis} = E_{I\sim P_{data}(I)}\left[\max\left(0,1-D(I)\right)\right] + E_{Z_r\sim p_{Z_r}(Z_r)}\left[\max\left(0,1+D(Z_r)\right)\right] \tag{9}$$

生成器的对抗损失可表示为：

$$L_{adv} = -E_{Z_r\sim p_{Z_r}(Z_r)}\left[D(Z_r)\right] \tag{10}$$

感知损失 在神经网络中随着网络层数的加深，深度特征能够获取更高层次的语义信息。网络通过惩罚生成图和原图深度特征差异的监督信号，使得生成图可以保留原图中较高层次的语义信息。在本文的感知损失中，生成图和真实图经过预训练 VGG 网络[35]得到深度特征。损失函数可表示为：

$$L_{per}=\frac{1}{C_jH_jW_j}\left\|\Phi_j(I)-\Phi(Z_r)\right\|_2^2 \tag{11}$$

其中 Φ 代表预先训练的 VGG16 网络，我们采取最后一个卷积层的特征图作为图像语义结果，即 j = 15。

重构损失 网络修复不仅要保证语义的一致，还要对像素进行精确的重建，因此，我们使用 L1 重构损失约束重构过程。具体来说，我们将生成的修复图像与原图在像素级上计算一阶偏差作为重构损失进行优化，以期望达到修复图像与真实图像一致的结果。

$$L_{rec}=\left\|I-Z_c\right\|_1+\left\|I-Z_r\right\|_1 \tag{12}$$

4. 实验设置

4.1. 数据集

本文在 3 个图像修复任务的通用数据集上评估本文的修复模型。

Place2 数据集[36]：MIT 发布的数据集，包含超过 180 万张来自 365 个场景的图像，我们随机挑选 1000 张作为测试图像，其余作为训练图像。

CelebA-HQ [37]数据集：来自名人的高质量人脸数据集，包含 3 万张高分辨率人脸图像，我们随机挑选 1000 张作为测试图像，其余作为训练图像。

Paris StreetView [38]数据集：一个主要集中在城市建筑的巴黎街景数据集，包含了 14,900 张训练图像和 100 张测试图像。

在训练过程中采用随机生成掩码的方式进行训练，测试的时候采用 NVIDIA 的不规则掩码数据集[18]进行测试。

4.2. 对比实验

将本文的模型与以下几种主流方法进行对比，其中包括 DeepFill v2 (GC)、EdgeConnect (EC)、Pluralistic-inpainting (PI)。

DeepFillv2 (GC) [21]：一种两段的图像修复网络模型，采用门控卷积的方式更加合理的利用待修复区域的信息。

EdgeConnect (EC) [22]：设计了一种先生成边界再利用边界补全颜色的修复模型。

Pluralistic-inpainting (PI) [20]：设计了一种平行双分支的概率学习模型，重建分支获取缺失区域的先验，生成分支将条件先验耦合重建分支的分布。

4.3. 实验环境

在 Ubuntu18.04 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 PyTorchv1.6.0、CUDNN v7.2、CUDA v10.2；核心硬件配置为 Inteli7-9700K 3.60 GHz 的 CPU，12GNVIDIA 2080Ti 的 GPU。我们使用 Adam 优化器对批量大小为 4 的模型进行训练，beta1 和 beta2 分别是 0.5 和 0.999。在模型训练时，我们的学习率使用分段常数衰减的策略进行训练。训练时图像全部缩放为 256 × 256 大小。

5. 结果与验证

所有模型均在 CelebA、Place2、ParisStreet 三个通用数据集上进行实验验证。为了公平对比，所有测试结果均采用相同的掩码方案。与其他模型对比实验在 5.1 节详细介绍，并在 5.2 节展示一些对掩码 transformer 的可视化结果以验证该模块的有效性。

5.1. 实验结果

本文在上述三个通用数据集上与其他先进的模型进行定量和定性的比较以证明本文方法的优越性。在测试时，我们固定读取掩码和测试图像的顺序以保证测试结果的公平性。

5.1.1. 定量比较

本文使用了 FID、PSNR、SSIM、平均 L1 损失这四种常用的评价指标来客观衡量修复结果的质量。其中，FID 可以大致反映修复图像和原始图像在特征层面的相似度。PSNR 和 SSIM 可以大致反映模型修复图像和原始图像结构信息相似度的能力。平均 L1 损失可以大致反映修复图像重构原始图像内容的相似度。这几种评价指标能够客观的从不同方面评价模型的好坏。

表 1 中给出本文模型与其他主流模型的定量对比结果。通过表格可以看出，本文提出的模型在三个通用数据集上均有一定程度的提升。相较于基础网络(baseline) Deepfillv2，FID 指标在人脸数据集上提升 25%左右，在巴黎街景数据集上提升 40%左右，在风景数据集上仅提升 18%左右。PSNR 指标在人脸数据集上提升 4%左右，在巴黎街景数据集上提升 6%左右，在风景数据集上提升 2%左右。SSIM 指标在人脸数据集上提升 2.4%左右，在巴黎街景数据集上提升 3.6%左右，在风景数据集上提升 1.5%左右。而与其他主流模型对比，本文模型在各项指标上均有所提升。相较于不同数据集来说，巴黎街景数据集中含有较多相似信息，而这个数据集中本文的模型提升效果最好。这也验证了本文模型使用的掩码 transformer 结构比传统的注意力机制能够更加有效的利用全局相似信息。

Table 1. Quantitative comparisons on CelebA, Paris Street View and Place2
表 1. CelebA、Paris Street View 和 Place2 数据集上的定量对比

数据集		CelebA			Paris Street View			Place2		
掩码率		10%~20%	30%~40%	50%~60%	10%~20%	30%~40%	50%~60%	10%~20%	30%~40%	50%~60%
FID	EdgeConnect	17.6	33.37	65.38	32.54	66.61	117.53	55.53	83.23	130.9
	Deepfillv2	19.61	41.26	74.64	39.2	85.53	145.97	54.54	89.22	136.28
	PI	32.83	65.88	99.52	41.43	96.24	143.32	67.11	128.86	181.27
	Our	**14.58**	**29.63**	**56.39**	**21.24**	**51.75**	**101.48**	**45.7**	**73.59**	**109.54**
PSNR	EdgeConnect	31.35	25.94	20.97	29.83	25.46	21.54	29.86	26.44	22.76
	Deepfillv2	30.85	25.41	20.67	29.45	24.94	21.28	29.49	26.82	22.51
	PI	28.13	22.81	18.55	26.11	21.75	18.11	27.69	23.13	18.27
	Our	**31.89**	**26.55**	**21.82**	**31.53**	**26.59**	**22.14**	**30.48**	**27.3**	**22.97**
SSIM	EdgeConnect	0.883	0.744	0.566	0.8658	0.7163	0.5362	0.8445	0.7003	0.5398
	Deepfillv2	0.8859	0.7532	0.5732	0.8757	0.7242	0.5435	0.8423	0.714	0.5391
	PI	0.8624	0.7034	0.5264	0.8414	0.6608	0.4738	0.8307	0.6447	0.463
	Our	**0.905**	**0.776**	**0.6121**	**0.8965**	**0.758**	**0.5654**	**0.8605**	**0.7213**	**0.5491**
Mean l1	EdgeConnect	0.0088	0.0213	0.0449	0.0114	0.0251	0.0468	0.0112	0.0211	0.038
	Deepfillv2	0.009	0.0216	0.0459	0.011	0.0257	0.0475	0.0103	0.0206	0.0362
	PI	0.0124	0.033	0.0661	0.0164	0.0388	0.0734	0.0134	0.038	0.0814
	Our	**0.007**	**0.018**	**0.0376**	**0.0082**	**0.0202**	**0.0416**	**0.0079**	**0.018**	**0.0432**

5.1.2. 定性比较

下图 4 比较了 EC、GC、PI 和本文的模型生成的图像。在三个数据集中，本文的修复效果均好于其他结果。其中 EC 网络在修复边界语义较为复杂时难以生成有效的语义边界，导致生成的修复图像内部模糊。例如第一行中的眼睛位置难以生成合理的眼球区域。而对于语义较为简单时则能够通过边界生成较为良好的修复效果。例如第三行中两条柱子位置相较于其他模型更真实。而对于 GC 模型，由于注意力机制存在特征语义失真现象，所以全局相似填充的区域并不准确，导致待修复区域出现伪影和模糊现象。而本文模型相较于其他模型在语义上更加清晰、连贯，也没有伪影现象。考虑相较于 GC 网络本文仅改进 transformer 结构来替代注意力模块，但 GC 网络存在的视觉伪影和模糊现象在本文模型则改善了很多。所以我们认为改进的掩码 transformer 结构能够有效的解决注意力机制带来的弊端。并且从整体上来看，本文模型相较于其他模型在纹理细节上更加丰富，在修复质量上也更具有优势。

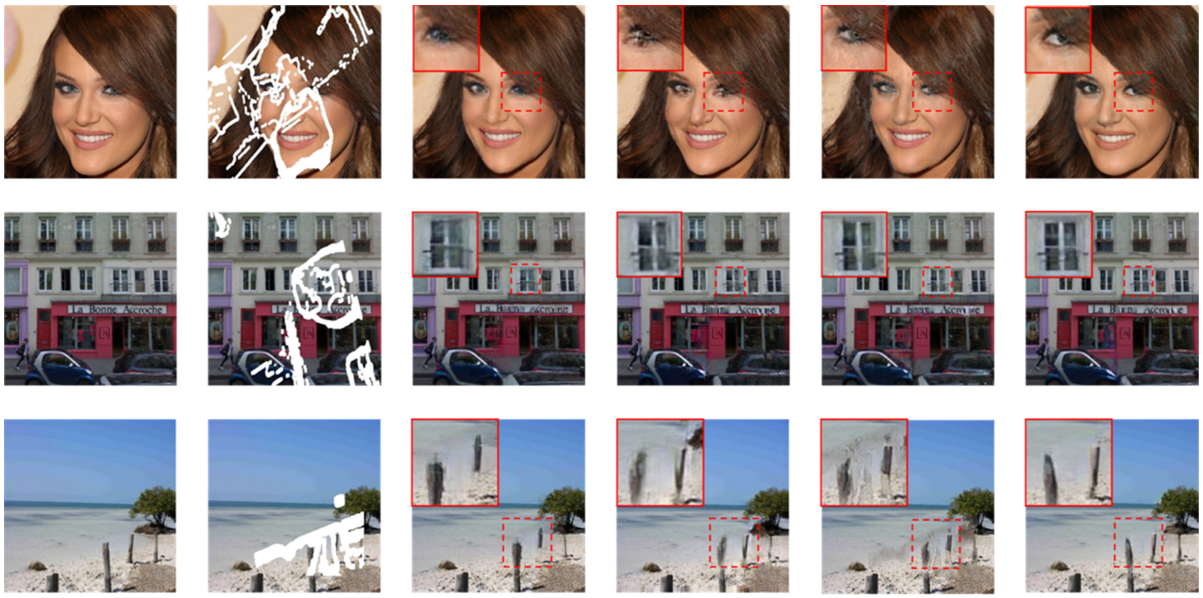

Figure 4. Results on different data sets
图 4. 不同数据集上的结果

5.2. 可视化实验

本文在巴黎街景数据集和人脸数据集上都做了可视化实验来验证本文所提出方案的有效性。本文选取了主流的上下文注意力模块[21]来代表传统注意力模块的结果，然后与本文的掩码 transformer 进行对比。其中本节实验均在同一掩码下对同一位置求全局相似度并用热力图的方式呈现。

图 5 为上下文注意力模块和掩码 transformer 模块的可视化显示，其中前两行是在巴黎街景数据集(Paris Street View)中的可视化结果，后两行是名人数据集(CelebA)中的可视化结果。第一列为数据的真实图像、第二列为模型的输入、第三列和第五列分别是含上下文注意力模块和掩码 transformer 的输出结果、第四列为对掩码区域内黑色小块全局相似度的热力图、第六列为对掩码区域内白色小块全局相似度的热力图。巴黎街景数据集(Paris Street View)相对于其他数据集来说大部分区域为建筑物，这导致该数据集含有大量纹理信息和重复区域。对于图 5 第一行，修复交界处的位置信息时，上下文注意力仅关注纹理相似信息导致输出结果存在交界模糊的情况。而掩码 transformer 在该位置时关注更多的则是像这种交界处

的纹理相似信息。对于图 5 第二行，待修复区域与部分位置存在语义相似处。上下文注意力在该热力图上不仅关注该横向区域的语义相似处，还考虑周围的含有条纹纹理信息的纹理相似处。而掩码 transformer 更关注于横向的语义相似处。对于人脸数据集来说，由于人脸含有确定的对称先验信息，于是本文设计在眼睛和牙齿等具有对称结构区域进行可视化实验。图 5 第三行对左眼进行遮盖，查看经注意力模块的眼角区域的全局相似度。其中上下文注意力模块关注点不仅在右眼，还有掩码外左眼和嘴角一部分区域。而掩码 transformer 模块仅关注右眼区域信息。图 5 中对右半区域的牙齿进行遮盖，查看经注意力模块的右嘴角区域的全局相似度。上下文注意力模块关注头发，左嘴角和右嘴角三部分区域信息，而掩码 transformer 模块则仅关注左嘴角区域信息。对于上述四种情况的概述，验证了掩码 transformer 模块相对于传统的注意力模块来说在这种纹理相似信息较多的情况下能够更好地关注语义相似处信息，而对于语义不同但纹理相似区域能够较好的屏蔽以达到摒除噪声信息影响的效果。

Figure 5. Visualization of global similarity between context attention module and mask transformer module
图 5. 上下文注意力模块和掩码 transformer 模块全局相似性可视化

6. 总结

本文提出了一种基于掩码 transformer 的图像修复网络。针对图像修复领域中注意力机制的度量方式存在仅考虑特征纹理而忽略特征语义的问题，提出了掩码 transformer 结构。该结构首先将输入完好区域分块组成查询集，并通过对输入特征块线性投影以获取高级特征语义，然后通过掩码注意力模块获取待修复区域与完好区域的相似性信息并通过全连接层和 softmax 激活层获取相似度矩阵。最后加权融合查询集和相似度矩阵以获取待修复区域的填充信息。后续通过定量和定性实验分析，该方法相较于传统的注意力机制能够更精确的获取完好区域相似的信息以生成具有丰富纹理细节的修复图像。并且与其他先进模型相比，本文提出的方法也能够产生更好的效果。

参考文献

[1] Ballester, C., Bertalmio, M., Caselles, V., *et al*. (2001) Filling-In by Joint Interpolation of Vector Fields and Gray Levels. *IEEE Transactions on Image Processing*, **10**, 1200-1211. https://doi.org/10.1109/83.935036

[2] Bertalmio, M., Sapiro, G., Caselles, V., *et al*. (2000) Image Inpainting. *SIGGRAPH Conference*, New Orleans, 23-28 July 2000, 417-424. https://doi.org/10.1145/344779.344972

[3] Bertalmio, M., Vese, L., Sapiro, G., *et al*. (2003) Simultaneous Structure and Texture Image Inpainting. *IEEE Transactions on Image Processing*, **12**, 882-889. https://doi.org/10.1109/TIP.2003.815261

[4] Shen, J. and Chen, T. (2003) Euler's Elastica and Curvature-Based Inpainting. *SIAM Journal on Applied Mathematics*, **63**, 564-592. https://doi.org/10.1137/S0036139901390088

[5] Barnes, C., Shechtman, E., Finkelstein, A., *et al*. (2009) Patchmatch: A Randomized Correspondence Algorithm for Structural Image Editing. *Proceedings of ACM SIGGRAPH*, Vol. 28, 1-11. https://doi.org/10.1145/1531326.1531330

[6] Drori, I., Cohen-Or, D. and Yeshurun, H. (2003) Fragment-Based Image Completion. *ACM Transactions on Graphics*, **22**, 303-312. https://doi.org/10.1145/882262.882267

[7] Esedoglu, S. and Shen, J. (2003) Digital Inpainting Based on the Mumford-Shah-Euler Image Model. *European Journal of Applied Mathematics*, **13**, 353-370. https://doi.org/10.1017/S0956792502004904

[8] Xu, Z. and Sun, J. (2010) Image Inpainting by Patch Propagation Using Patch Sparsity. *IEEE Transactions on Image Processing*: *A Publication of the IEEE Signal Processing Society*, **19**, 1153-1165. https://doi.org/10.1109/TIP.2010.2042098

[9] Wang, Z., Bovik, A., Sheikh, H.R., *et al*. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/TIP.2003.819861

[10] Lindeberg, T. (2012) Scale Invariant Feature Transform. *Scholarpedia*, **7**, 10491. https://doi.org/10.4249/scholarpedia.10491

[11] Efros, A.A. and Leung, T.K. (1999) Texture Synthesis by Non-Parametric Sampling. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, 20-27 September 1999, 1033-1038. https://doi.org/10.1109/ICCV.1999.790383

[12] Criminisi, A., Perez, P. and Toyama, K. (2004) Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Transactions on Image Processing*, **13**, 1200-1212. https://doi.org/10.1109/TIP.2004.833105

[13] Levin, A., Zomet, A., Peleg, S., *et al*. (2004) Seamless Image Stitching in the Gradient Domain. 8*th European Conference on Computer Vision*, Prague, 11-14 May 2004, 377-389. https://doi.org/10.1007/978-3-540-24673-2_31

[14] Pathak, D., Krahenbuhl, P., Donahue, J., *et al*. (2016) Context Encoders: Feature Learning by Inpainting. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2536-2544. https://doi.org/10.1109/CVPR.2016.278

[15] Lowe, D.G. (1999) Object Recognition from Local Scale-Invariant Features. *Proceedings of IEEE International Conference on Computer Vision*, Corfu, 20-25 September 1999, 1150-1157. https://doi.org/10.1109/ICCV.1999.790410

[16] Simakov, D., Caspi, Y., Shechtman, E., *et al*. (2008) Summarizing Visual Data Using Bidirectional Similarity. *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 24-26 June 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587842

[17] Satoshi, L., Edgar, S.-S. and Hiroshi, I. (2017) Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics*, **36**, 107:1-107:14. https://doi.org/10.1145/3072959.3073659

[18] Liu, G., Reda, F.A., Shih, K.J., *et al*. (2018) Image Inpainting for Irregular Holes Using Partial Convolutions. In: *European Conference on Computer Vision*, Springer, Cham, 85-100. https://doi.org/10.1007/978-3-030-01252-6_6

[19] Yu, J., Lin, Z., Yang, J., *et al*. (2018) Generative Image Inpainting with Contextual Attention. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 5505-5514. https://doi.org/10.1109/CVPR.2018.00577

[20] Zheng, C., Cham, T.J. and Cai, J. (2019) Pluralistic Image Completion. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 1438-1447. https://doi.org/10.1109/CVPR.2019.00153

[21] Yu, J., Lin, Z., Yang, J., *et al*. (2018) Free-Form Image Inpainting with Gated Convolution. 2019 *IEEE*/*CVF International Conference on Computer Vision Workshops*, Seoul, 27-28 October 2019, 4471-4480. https://doi.org/10.1109/ICCV.2019.00457

[22] Nazeri, K., Ng, E., Joseph, T., *et al*. (2019) EdgeConnect: Structure Guided Image Inpainting Using Edge Prediction. 2019 *IEEE*/*CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 1-8. https://doi.org/10.1109/ICCVW.2019.00408

[23] Li, J., Wang, N., Zhang, L., *et al*. (2020) Recurrent Feature Reasoning for Image Inpainting. *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 7760-7768. https://doi.org/10.1109/CVPR42600.2020.00778

[24] Xie, C., Liu, S., Li, C., *et al*. (2019) Image Inpainting with Learnable Bidirectional Attention Maps. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 8858-8867. https://doi.org/10.1109/ICCV.2019.00895

[25] Miyato, T., Kataoka, T., Koyama, M., *et al*. (2018) Spectral Normalization for Generative Adversarial Networks. 6*th International Conference on Learning Representations*, Vancouver, 30 April-3 May, 2018.

[26] Zeng, Y., Fu, J., Chao, H., *et al*. (2019) Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 1486-1494. https://doi.org/10.1109/CVPR.2019.00158

[27] Xiao, Q., Li, G. and Chen, Q. (2018) Deep Inception Generative Network for Cognitive Image Inpainting.

[28] Yang, C., Lu, X., Lin, Z., *et al*. (2017) High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6721-6729. https://doi.org/10.1109/CVPR.2017.434

[29] Song, Y., Yang, C., Lin, Z., *et al*. (2017) Contextual-Based Image Inpainting: Infer, Match, and Translate. 15*th European Conference on Computer Vision*, Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01216-8_1

[30] Sagong, M.C., Shin, Y.G., Kim, S.W., *et al*. (2020) PEPSI: Fast Image Inpainting with Parallel Decoding Network. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 11360-11368. https://doi.org/10.1109/CVPR.2019.01162

[31] Vaswani, A., Shazeer, N., Niki, P., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 5998-6008.

[32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[33] Chen, M., Radford, A., Child, R., *et al*. (2020) Generative Pretraining from Pixels. *International Conference on Machine Learning*, Vienna, 13-18 July 2020, 1691-1703.

[34] Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) Layer Normalization.

[35] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science.

[36] Zhou, B., Lapedriza, A., Khosla, A., *et al*. (2018) Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **40**, 1452-1464. https://doi.org/10.1109/TPAMI.2017.2723009

[37] Karras, T., Aila, T., Laine, S., *et al*. (2017) Progressive Growing of GANs for Improved Quality, Stability, and Variation.

[38] Efros, A.A., *et al*. (2015) What Makes Paris Look Like Paris? *ACM Transactions on Graphics*, **31**, 1-9. https://doi.org/10.1145/2185520.2185597