基于文本引导的人脸图像修复

廉 敬[1],张继保[1],刘冀钊[2],张家骏[1],董子龙[1]

(1.兰州交通大学 电子与信息工程学院,兰州 730030;2.兰州大学 信息科学与工程学院,兰州 730030)

摘 要:针对目前人脸修复方法存在结构扭曲、纹理模糊以及不可控等问题,提出了一种文本引导的人脸图像修复方法。该方法通过融合图像特征和相应的文本特征来重建图像中的缺失区域。在网络训练中,设计了视觉-文本模态融合模块,用于关联图像和文本特征,使重建人脸缺失区域不仅以图像中可见的视觉语义为基础,还以具有丰富的文本语义为指导。在编码和解码特征之间添加了一个注意力感知层,以提高可见区域和生成区域外观的一致性。在CelebA-HQ人脸数据集上的实验结果表明:本文方法能够得到在纹理和结构上更自然且符合文本语义的修复结果,其视觉效果和评价指标均优于对比算法。

关键词:图像修复;文本指导;跨模态融合;深度学习

中图分类号:TP391 **文献标志码**:A **文章编号**:1671-5497(2025)08-2732-09

DOI:10.13229/j.cnki.jdxbgxb.20240051

Text-based guided face image inpainting

LIAN Jing[1],ZHANG Ji-bao[1],LIU Ji-zhao[2],ZHANG Jia-jun[1],DONG Zi-long[1]

(1. *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730030, *China*; 2. *School of Information Science and Engineering, Lanzhou University, Lanzhou* 730030, *China*)

Abstract: A text-guided face image inpainting method is proposed to address the problems of structural distortion, texture blurring, and uncontrollability in current face inpainting methods. The method reconstructs the missing regions in an image by fusing image features and corresponding text features. In the network training, a visual-textual modal fusion module is designed for associating image and textual features so that the reconstruction of missing regions of the face is not only based on the visual semantics visible in the image, but also guided by textual semantics with rich text. An attention-aware layer is added between the encoded and decoded features to improve the consistency of the appearance of the visible and generated regions. Experimental results on the CelebA-HQ face dataset show that the method in this paper is able to obtain restoration results that are more natural and consistent with the textual semantics in terms of texture and structure, and its visual effect and evaluation metrics are better than those of the comparison algorithms.

Key words: image inpainting; textual guidance; cross-modal fusion; deep learning

---

收稿日期:2024-01-15.

基金项目:国家自然科学基金项目(62061023, 82260364);甘肃省杰出青年基金项目(21JR7RA345);甘肃省自然科学基金项目(23JRRA1485, 22JR5RA543, 22JRJ5RA166);兰州市青年科技人才创新项目(2023-QN-109).

作者简介:廉敬(1983-),男,教授,博士.研究方向:图像处理与模式识别.E-mail:lian322scc@163.com

0　引　言

随着人脸识别技术的迅速发展，人脸图像修复技术已经成为当前研究的热点问题[1,2]。然而，由于获取的人脸图像通常受到自然和人为因素的影响，增加了人脸识别的技术难度。人脸图像修复技术升级对于提高人脸识别准确率至关重要。

近年来，深度学习的发展推动了基于卷积神经网络和生成式对抗网络的广泛应用，许多先进适用的图像修复方法被提出。Pathak 等[3]提出了上下文编码器方法，先行探索将编解码架构应用于图像修复。然而，这种方法无法应用于任意的修复掩膜。同时，由于自动编码器网络的生成能力欠缺，这一方法会产生不可信的结果。为了克服上述问题，Iizuka 等[4]提出使用全局和局部判别器进行图像修复。该方法引入了全局和局部一致性的约束，在修复图像中大尺度缺失区域时，能够确保修复结果与原始图像在整体结构和细节上保持一致。然而，这是一种空间支持模型，无法填充大的孔洞。为了进一步细化全局结构和突出局部纹理细节，Yan 等[5]提出了一种卷积神经网络优点和传统复制粘贴结合的方法。然而，该方法对每一个掩膜输入仅产生一个修复结果。为此，Liu 等[6]提出了基于概率多样性的生成对抗网络方法，旨在解决图像修复中的多样性和不确定性问题，以生成具有多样化和真实感的修复图像；Wan 等[7]提出了一种多样化的图像修复方法，以增加修复结果的多样性和可选性。然而，当关键的结构或细节被破坏，导致全局和局部特征不一致时，该方法不能有效应用。此外，Li 等[8]利用 Transformer 中的多头注意力建模特征之间的非局部关系在全局上完成图像重建；Huang 等[9]提出了一种基于 U-Net 的新型 Transformer 网络以降低多头注意力的计算成本；Lian 等[10]从纹理和结构研究入手重建图像缺失区域。上述方法在人脸图像修复中，能够提取到图像深层语义信息，均为图像重建的有效方法，但在复杂背景或缺失区域较大的情况下，所修复出的图像尚存在伪影、结构扭曲和纹理模糊等问题。

基于此，本文提出了一种基于文本引导人脸图像修复的方法。该方法使用预训练的语言表征（Bidirectional encoder representation from transformers，BERT）模型[11]对文本描述进行编码，然后在解码过程中将文本语义特征转换成图像特征。在人脸图像重建过程中，设计了视觉-文本模态融合（Visual-textual modalities fusion，V-TMF）模块，利用丰富的文本特征来指导重建孔洞区域。

1　本文算法

1.1　网络框架与流程

本文算法的框架如图 1 所示，主要分为文本语义学习网络和人脸重建网络两个部分，图中 $\mathcal{L}_{rec}$、$\mathcal{L}_{style}$、$\mathcal{L}_{perc}$ 和 $\mathcal{L}_{adv}$ 分别为像素重建损失、风格损失、感知损失和对抗损失 4 种损失函数。

首先，文本语义学习网络通过 BERT 对真实图像 I_{gt} 的文本描述编码，在解码阶段，通过卷积神经网络将文本特征转换成图像特征。其次，人脸重建网络对孔洞图像 I_m 和掩码图像 I_{mask} 编码。在解码过程中，通过视觉-文本模态融合（V-TMF）

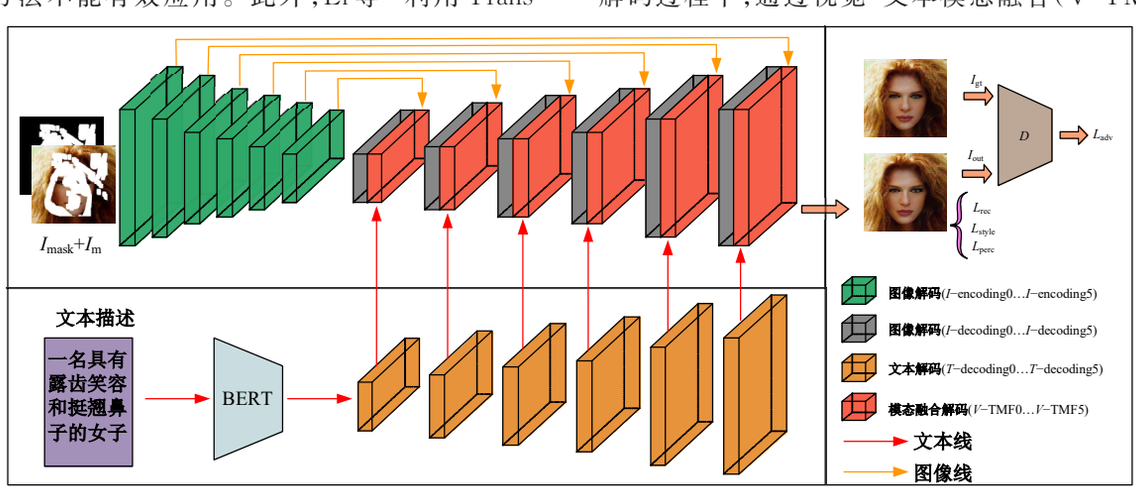

图 1　本文方法的整体架构

Fig. 1　Overall architecture of proposed method

模块交互图像和文本特征完成图像重建。网络的详细结构如表1所示。

1.2 视觉-文本模态融合模块

文本描述作为一种重要参考,可以为图像修复提供丰富的信息。通过将文本特征与视觉特征融合,修复模型能更全面地理解图像内容。这确保了在修复过程中重建区域与掩码周围像素的一致性,同时也保证了填充部分在整个图像中的语义合理性。本文设计的视觉-文本模态融合模块如图2所示。

表1 网络结构

Table 1 Network structure

输入	网络层	输出尺寸	通道数	卷积核	步长	填充	归一化	激活函数
文本描述	BERT	4	512	None	None	None	None	None
T	T-decoding0	8	512	7	2	3	BN	ReLU
T_Dec0	T-decoding1	16	512	5	2	2	BN	ReLU
T_Dec1	T-decoding2	32	128	3	2	1	BN	ReLU
T_Dec2	T-decoding3	64	128	3	2	1	BN	ReLU
T_Dec3	T-decoding4	128	64	3	2	1	BN	ReLU
T_Dec4	T-decoding5	256	3	3	2	1	BN	ReLU
$I_{_mask}$, I_m	I-encoding0	128	64	3	2	1	BN	ReLU
I_Enc0	I-encoding1	64	128	3	2	1	BN	ReLU
I_Enc1	I-encoding2	32	128	3	2	1	BN	ReLU
I_Enc2	I-encoding3	16	256	3	2	1	BN	ReLU
I_Enc3	I-encoding4	8	512	3	2	1	BN	ReLU
I_Enc4	I-encoding5	4	512	3	2	1	BN	ReLU
I_Enc5	I-decoding0	8	512	3	2	1	BN	ReLU
I_Enc5, I_Dec0, T_Dec0	V-TMF0	8	512	None	None	None	BN	ReLU
I_V-TMF0	I-decoding1	16	512	3	2	1	BN	ReLU
I_Enc4, I_Dec1, T_Dec1	V-TMF1	16	512	None	None	None	BN	ReLU
I_V-TMF1	I-decoding2	32	128	3	2	1	BN	ReLU
I_Enc3, I_Dec2, T_Dec2	V-TMF2	32	128	None	None	None	BN	ReLU
I_V-TMF2	I-decoding3	64	128	3	2	1	BN	ReLU
I_Enc2, I_Dec3, T_Dec3	V-TMF3	64	128	None	None	None	BN	ReLU
I_V-TMF3	I-decoding4	128	64	3	2	1	BN	ReLU
I_Enc1, I_Dec4, T_Dec4	V-TMF4	128	64	None	None	None	BN	ReLU
I_V-TMF4	I-decoding5	256	3	3	2	1	BN	ReLU
I_Enc0, I_Dec5, T_Dec5	V-TMF5	256	3	None	None	None	None	None

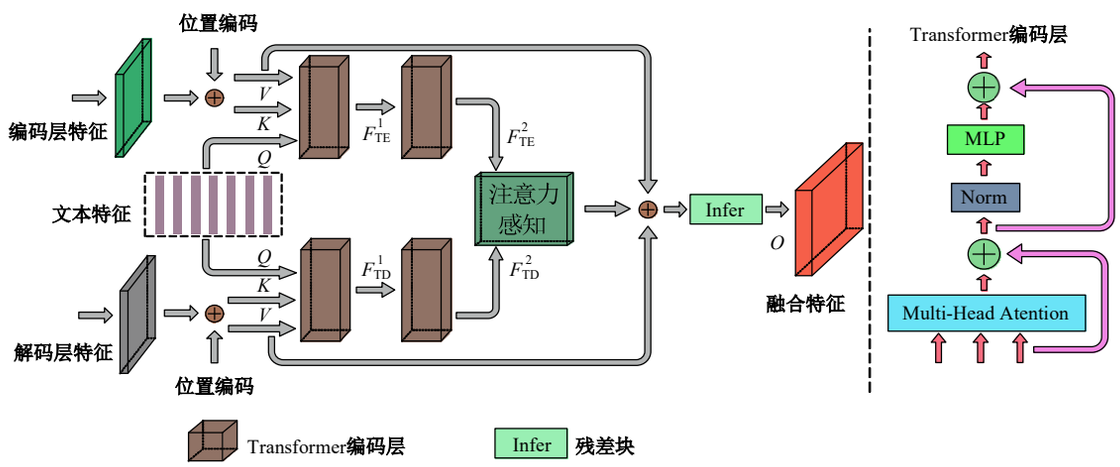

图2 视觉-文本模态融合模块

Fig. 2 Visual-textual modal fusion module

该模块采用了Transformer编码器，用于交互图像和文本信息。通过利用注意力机制，其能够捕捉输入序列中的全局依赖关系。在处理图像和文本特征时，它能够聚焦于彼此之间的重要信息，并通过注意力权重进行跨模态的信息传递和交互。这种方式使得模型能够更有效整合图像和文本信息，从而提高修复结果的准确性和一致性。

为了减少不同模态之间信息融合的损失，使用双向Transformer编码层，每个分支使用两个独立的编码器处理不同模态的输入数据，通过这种架构使文本和图像特征充分融合。为了捕捉不同模态特征的语义信息，引入了多头注意力(Multi-head attention，MHA)。这允许模型更好地关注不同特征，并学习到不同位置和特征之间的关联信息。多头注意力的结构如图3所示。

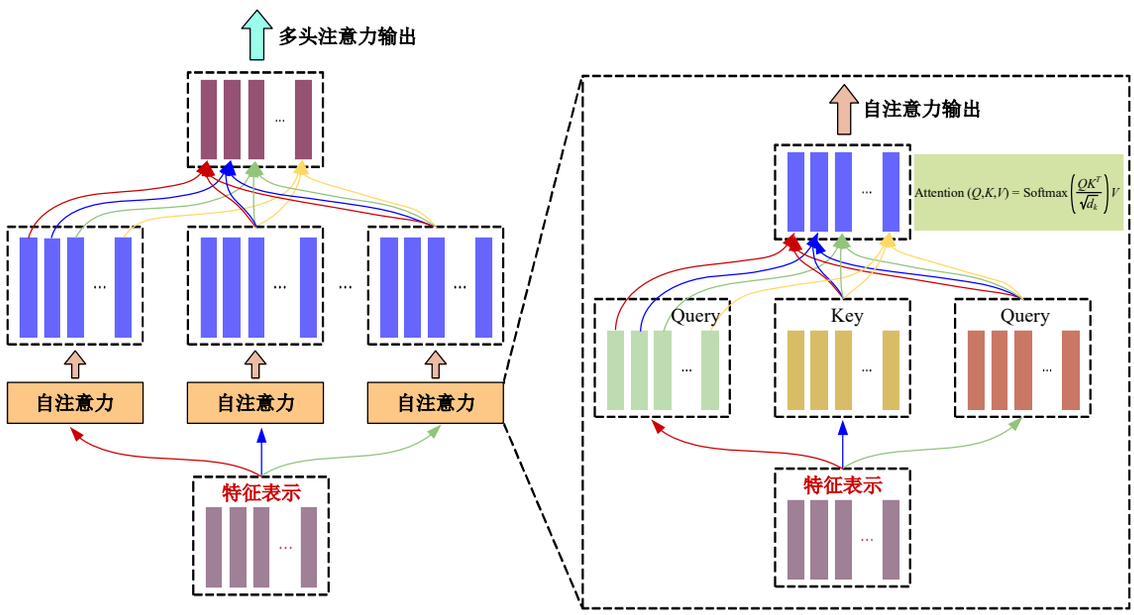

图 3　多头注意力模型结构

Fig. 3　Structure of the multi-attention model

在MHA下，将每个头独立计算的结果拼接在一起：

$$\mathrm{MHA}=\mathrm{Concat}(head_1,\cdots,head_i)L \tag{1}$$

$$head_i=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{2}$$

$$A=F\boldsymbol{W}^{Q},\boldsymbol{K}=F\boldsymbol{W}^{K},\boldsymbol{V}=F\boldsymbol{W}^{V} \tag{3}$$

式中：$\boldsymbol{Q}\in\mathbb{R}^{N\times D}$、$\boldsymbol{K}\in\mathbb{R}^{N\times D}$和$\boldsymbol{V}\in\mathbb{R}^{N\times D}$由输入特征$F\in\mathbb{R}^{N\times C}$得到，分别为查询向量(Query)、键向量(Key)和值向量(Value)；$N=H\times W$为输入特征的长度，H和W分别为输入特征的高和宽；C为特征通道，D为变换后特征的维度；$\boldsymbol{W}_i^{Q}$、$\boldsymbol{W}_i^{K}$和$\boldsymbol{W}_i^{V}\in\mathbb{R}^{C\times D}$分别为第$i$个头$Q$、$K$、$V$使用的线性变阵矩阵；$i$为MHA中头的数量，在本文中$i$设置为8，$\mathcal{L}\in\mathbb{R}^{(i\times D)\times Z}$为线性变换矩阵；$Z$为MHA最后输出特征的维度；Concat为通道拼接；Attention(•)为注意力计算函数，其表达式如下所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

式中：d_k为$\boldsymbol{Q}$、$\boldsymbol{K}$矩阵的向量维度；通过Softmax函数计算特征之间的相似度权重；然后将权重系数与V相乘得到注意力输出。

第一层编码器将文本特征与图像编码层和解码层特征融合，使图像重建过程中不仅基于可见部分的视觉信息，还利用丰富的文本语义，其表达式如下：

$$F_{\mathrm{TE}}^{1}=F_{\mathrm{T}}+\mathrm{MHA}(F_{\mathrm{T}},F_{\mathrm{E}},F_{\mathrm{E}}) \tag{5}$$

$$F_{\mathrm{TD}}^{1}=F_{\mathrm{T}}+\mathrm{MHA}(F_{\mathrm{T}},F_{\mathrm{D}},F_{\mathrm{D}}) \tag{6}$$

式中：F_{T}为文本特征；F_{E}为图像编码层特征；F_{D}为图像解码层特征；F_{TE}^{1}和F_{TD}^{1}分别为融合后的特征。

第二层编码器将融合后的特征进一步细化，表达式如下：

$$F_{\mathrm{TE}}^{2}=F_{\mathrm{T}}+\mathrm{MHA}(F_{\mathrm{TE}}^{1},F_{\mathrm{TE}}^{1},F_{\mathrm{TE}}^{1}) \tag{7}$$

$$F_{\mathrm{TD}}^{2}=F_{\mathrm{T}}+\mathrm{MHA}(F_{\mathrm{TD}}^{1},F_{\mathrm{TD}}^{1},F_{\mathrm{TD}}^{1}) \tag{8}$$

为提高可见区域和生成区域外观的一致性，本文在编码和解码特征之间提出了一个新颖的注意力感知层(Attention-aware layer，AAL)如图4所示。该设计旨在更好地利用远距离相关的高频

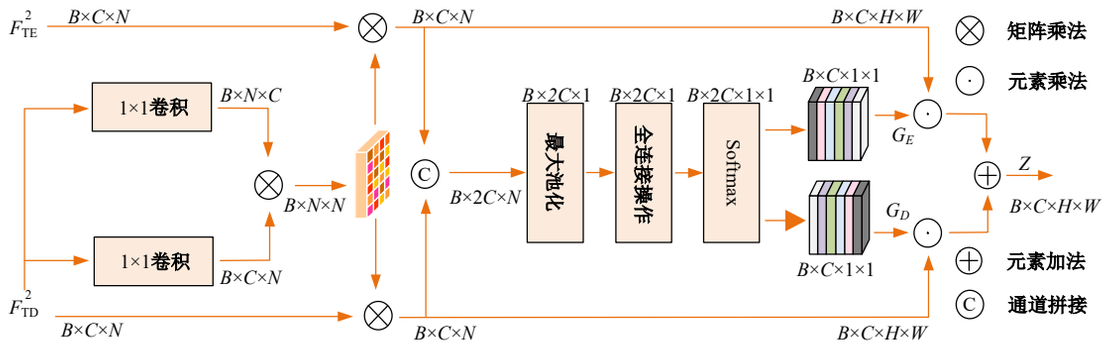

图 4 注意力感知层

Fig. 4 Attention-Aware Layer

特征,自适应地平衡对可见内容和生成内容的注意力,从而实现语义上更优越的特征转换。

具体来说,本文通过式(9)计算解码层特征 F_{TD}^2 的注意力分数:

$$A=\theta(F_{TD}^2)\Phi(F_{TD}^2) \tag{9}$$

式中:A 为特征的相似性;θ、Φ 为 1×1 卷积。

接下来,通过以下方式获得长期依赖关系:

$$F=\mathrm{Concat}\left[(A\odot F_{TE}^2),(A\odot F_{TD}^2)\right] \tag{10}$$

式中:F 为包含编码层中的高频细节信息和解码层的自注意力特征;Concat 为通道拼接。

最后,通过学习动态权重来组合 F_{TD}^2 和 F_{TE}^2:

$$G_E,G_D=\mathrm{Chunk}\left[\mathrm{Softmax}(\mathrm{Mlp})(\mathrm{Maxp}(F))\right] \tag{11}$$

$$Z=(G_E F_{TE}^2)\oplus(G_D F_{TD}^2) \tag{12}$$

式中:Z 为注意力感知层的输出;G_E、G_D 为特征的门控;Maxp 为最大池化;Mlp 为全连接层;Chunk 为特征通道维度分割操作。

视觉-文本模态融合模块的输出 O 表达式如下:

$$O=\mathrm{Infer}(F_{TE}^2\oplus Z\oplus F_{DE}^2) \tag{13}$$

式中:⊕ 为逐元素相加操作;Infer 为残差卷积操作。

2 损失函数

在网络训练过程中,采用了像素重建损失、感知损失、风格损失和对抗损失使模型重建结果尽可能地接近真实图像。

重建损失采用 L_1 范数进行计算,用于计算生成图像和真实图像之间的差异,损失函数表达式如下:

$$\mathcal{L}_{\mathrm{hole}}=\sum_i\left\|(I_{\mathrm{out}(i)}-I_{\mathrm{gt}(i)})\odot I_{\mathrm{mask}}\right\|_1 \tag{14}$$

$$\mathcal{L}_{\mathrm{valid}}=\sum_i\left\|(I_{\mathrm{out}(i)}-I_{\mathrm{gt}(i)})\odot(1-I_{\mathrm{mask}})\right\|_1 \tag{15}$$

$$\mathcal{L}_{\mathrm{rec}}=\lambda_{\mathrm{rec}}(\mathcal{L}_{\mathrm{hole}}+\mathcal{L}_{\mathrm{valid}}) \tag{16}$$

式中:I_{out} 为图像的修复结果;I_{gt} 为图像的真实值;I_{mask} 为掩码图像(mask=1);$\mathcal{L}_{\mathrm{hole}}$ 为孔洞区域的重建损失;$\mathcal{L}_{\mathrm{valid}}$ 为非孔洞区域的重建损失;i 分别对应图像 64×64、128×128 和 256×256 这 3 个尺度;λ_{rec} 为平衡因子,由经验设置为 20。

感知损失[12]不局限于像素级的差异,可以捕捉到图像的结构信息且可以弥补重建损失难以捕获的高级语义。本文中使用在 ImageNet[13]上预训练的 VGG-16[14]网络作为特征提取器,提取生成网络中的特征表示,通过计算这些特征之间的差异来度量生成图像与目标图像之间的感知损失。感知损失定义如下:

$$\mathcal{L}_{\mathrm{pere}}=\sum_i\left\|\phi_i(I_{\mathrm{out}})-\phi_i(I_{\mathrm{gt}})\right\|_1 \tag{17}$$

式中:$\phi_i(\cdot)$ 为图像在预训练的 VGG-16 网络中的第 i 层特征。

对抗损失用于确保重建图像与文本描述以及纹理和结构的一致性。本文中的对抗损失函数采用 LSGAN 方法[15]。相比于传统 GAN 网络的对抗损失设置,该方法可以使生成器产生更真实的结果,降低了梯度消失或梯度爆炸的风险。对抗损失定义如下:

$$\mathcal{L}_{\mathrm{D}}=\frac{1}{2}E_{I_{\mathrm{gt}}}[(D(I_{\mathrm{g}})-1)^2]+\frac{1}{2}E_{I_{\mathrm{out}}}[(D(I_{\mathrm{out}}))^2] \tag{18}$$

$$\mathcal{L}_{\mathrm{G}}=E_{I_{\mathrm{out}}}[(D(I_{\mathrm{out}})-1)^2] \tag{19}$$

风格损失用于使重建区域和图像可见部分保持一致。通过引入风格损失[16]可以使修复区域与周围环境在风格上融合,提高修复结果的自然度。风格损失定义如下:

$$\mathcal{L}_{\mathrm{style}}=\sum_i\left\|(\psi_i(I_{\mathrm{out}})-\psi_i(I_{\mathrm{gt}}))\right\|_1 \tag{20}$$

式中:$\psi_i(\cdot)$ 为 Gram 矩阵,用于量化图像特征表示

中的风格信息，通过比较Gram矩阵之间的差异来度量图像之间的风格差异。

综上所述，网络的总损失为：

$$\mathcal{L}_{joint}=\mathcal{L}_{rec}+\lambda_{rec}\mathcal{L}_{perc}+\lambda_{style}\mathcal{L}_{style}+\lambda_{g}\mathcal{L}_{G}+\lambda_{d}\mathcal{L}_{D} \quad (21)$$

式中：λ_{perc}、λ_{style}、λ_{g}、λ_{d}为函数的平衡因子，由经验值设置。

3　实　验

3.1　实验介绍

为验证本文方法的有效性，在CelebA-HQ数据集上进行了定性分析和定量分析。除此之外，还设计了消融实验，并进行相应的定量分析，以验证引入的注意力感知层和视觉-文本模态融合模块在修复过程中的贡献。

3.2　实验设置

在实验中，使用了由Liu等[17]创建的大型不规则掩码数据集作为掩码图像。训练时，所有图像的尺寸被设置为256×256。根据掩码区域的占比率，将测试掩码分为6种，分别是10%、20%、30%、40%、50%和60%。模型的训练是在PyTorch框架下实现的，使用的GPU是NVIDIA 4090。Batch size被设置为8。在训练过程中，前100个epoch的初始学习率设置为固定值2×10^{-4}，后100个epoch的学习率线性衰减到0，并使用Adam优化器对网络进行优化。这种设置在训练初期使用较小的学习率有助于稳定模型的训练，随着训练的进行逐渐减小学习率以更细致地调整参数，并选择Adam优化器以提高优化效率。

3.3　定性分析

为验证本文方法的有效性，在CelebA-HQ数据集上进行了实验，并将实验结果与当前6种具有代表性的方法进行比较，结果如图5所示。

RFR方法[18]是一种基于递归特征推理的图像修复方法，然而，该方法的修复结果存在纹理模糊问题。从图5中RFR的修复结果可以看出，在人脸重建部分存在较多模糊情况。RePaint方法[19]是基于扩散模型的图像修复方法，能够重建出在视觉效果上逼真的内容。然而，重建区域与人脸可见部分存在不一致的情况。从图5中RePaint的修复结果的眼睛部分可以看出，重建结果显示明显的结构失真，使图像整体看起来不太自然。CTSDG方法[20]将图像修复分为两个子任务：纹理生成和结构重建。然而，从图5中CTSDG的修复结果可以看出，第一张人脸的右眼存在结果扭曲和第三张图像中眼镜失真的情况。MIGT方法[21]和TDANet方法[22]是基于文本描述引导图像修复，但修复图像存在掩码区域与文本语义不一致和失真的情况。如图5中MIGT方法的第二张和第三张图像左右眼睛颜色不一样。TDANet方法的第二张、第三张和第五张图像存

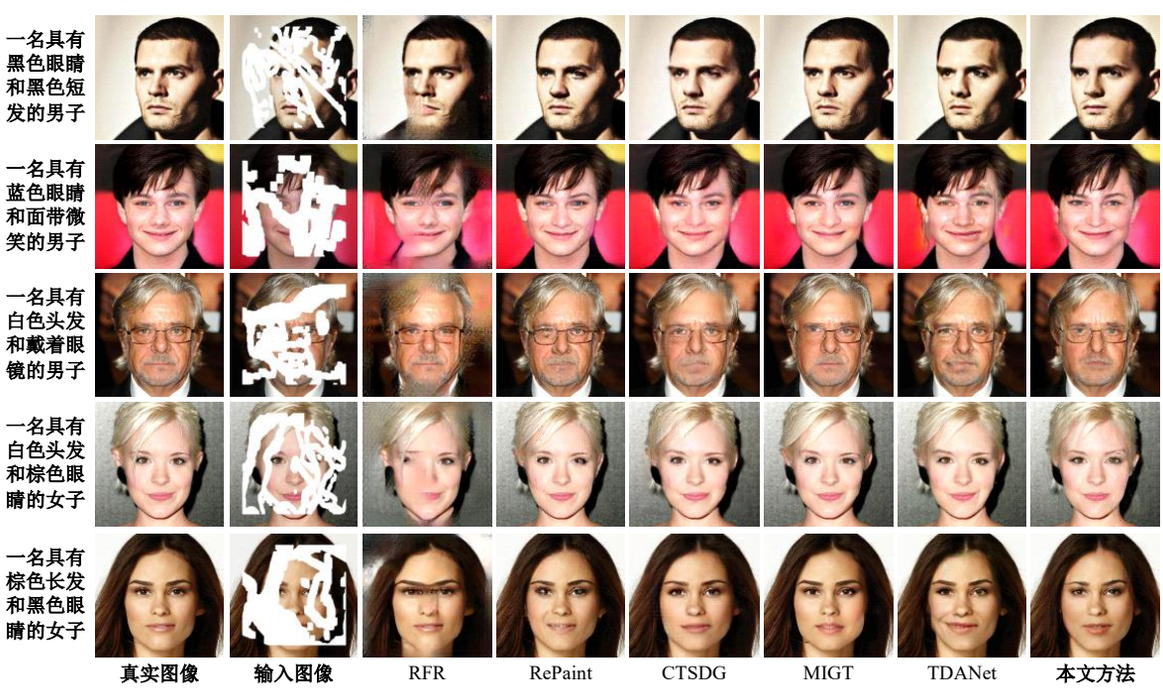

图5　不同图像修复方法测试结果对比

Fig. 5　Comparison of test results of different image inpainting methods

在结构扭曲和细节模糊的情况。综上所述,本文提出的方法能够重建出逼真的图像,并且与文本语义保持一致,没有明显修复痕迹,符合人眼的视觉特性。

3.4 定量分析

在6组不规则掩码(1%～10%、10%～20%、20%～30%、30%～40%、40%～50% 和 50%～60%)上,与当前先进算法进行了比较。本文方法使用了 L_1 误差、Learned perceptual image patch similarity (LPIPS)、Structural similarity index (SSIM)和Peak signal-to-noise ratio(PSNR)这4种主要的评价指标进行定量分析。这些评价指标分为像素级(L_1、SSIM、PSNR)和感知级(LPIPS)两类。像素级指标表现在重建图像和原图像之间的差异,图像修复任务是为缺失区域生成合理的纹理和连贯的结构而不是原始图像。为了客观评价图像重建的质量,感知评价指标可以模仿人类判断图像的视觉效果,从高级特征的信息出发而不局限于像素级信息。目前,感知和像素评价指标都已经被应用到图像修复工作中,以衡量图像的质量。表2展示了在CelebA-HQ数据集上的测试结果。从实验结果可以看出,虽然其他方法都能够重建出合理的内容,但在各项指标上,本文方法在整体上都优于其他对比方法。这表明本文方法修复的图像在视觉质量和与原图的相似度上都取得了较好的效果。

3.5 消融实验

为验证注意力感知层和视觉-文本模态融合模块在图像重建过程中的作用,进行了消融实验,以验证这两个模块的有效性。首先,进行一个基准实验,即使用完整的图像修复系统。在消融实验中,分别移除了注意力感知层和视觉-文本模态融合模块中的第二个编码器,然后使用同一组图像进行修复。对消融实验的测试结果进行定量比较分析,结果如表3所示,可以看出,本文模型在各项评价指标上都优于其他模型,验证了注意力

表2 不同图像修复方法、不同掩码区域比例的定量结果分析

Table 2 Analysis of quantitative results of different image inpainting methods and different mask area ratios

Masks		1%～10%	10%～20%	20%～30%	30%～40%	40%～50%	50%～60%
L_1 ↓	RFR	2.124	1.458	1.335	3.968	1.726	2.241
	RePaint	0.325	0.412	0.358	0.474	2.215	1.564
	CTSDG	0.094	0.143	0.422	0.335	0.726	1.102
	MIGT	0.096	0.122	0.422	0.417	0.624	1.934
	TDANet	0.104	0.139	0.387	0.433	0.643	1.024
	本文方法	**0.085**	**0.104**	**0.315**	**0.214**	**0.534**	**0.987**
PSNR ↑	RFR	14.13	13.82	17.52	13.21	16.74	22.23
	RePaint	29.54	29.14	25.17	25.44	23.86	22.64
	CTSDG	32.51	31.24	26.24	27.14	24.47	24.58
	MIGT	32.45	31.78	28.53	28.36	24.89	24.84
	TDANet	32.38	31.27	28.84	28.17	25.01	24.76
	本文方法	**33.21**	**32.94**	**29.04**	**29.75**	**25.61**	**25.10**
SSIM ↑	RFR	0.394	0.521	0.721	0.723	0.675	0.795
	RePaint	0.951	0.924	0.833	0.810	0.748	0.614
	CTSDG	0.967	0.951	0.838	0.854	0.825	0.787
	MIGT	0.847	0.825	0.813	0.800	0.821	0.794
	TDANet	0.935	0.871	0.811	0.794	0.857	0.824
	本文方法	**0.981**	**0.964**	**0.892**	**0.874**	**0.864**	**0.840**
LPIPS ↓	RFR	0.464	0.247	0.178	0.194	0.124	0.126
	RePaint	0.033	0.037	0.069	0.078	0.101	0.151
	CTSDG	0.021	0.029	0.048	0.051	0.087	0.132
	MIGT	0.025	0.028	0.057	0.063	0.087	0.121
	TDANet	0.028	0.033	0.047	0.054	0.092	0.146
	本文方法	**0.017**	**0.020**	**0.041**	**0.046**	**0.074**	**0.103**

注:↑表示越高越好,↓表示越低越好。

表 3　消融实验的定量结果分析

Table 3　Analysis of the quantitative results of the ablation experiments

a	b	L_1↓	PSNR↑	SSIM↑	LPIPS↓
√		0.724	28.337	0.868	0.047
	√	0.476	29.576	0.970	0.036
√	√	**0.287**	**32.214**	**1.316**	**0.017**

注:a 表示视觉-文本模态融合模块中第二个编码器;b 表示注意力感知层;↑表示越高越好,↓表示越低越好。

感知层和视觉-文本模态融合模块在图像修复中的重要作用。

3.6　人为指定文本特征引导图像重建实验

本文方法采用了人为指定的文本描述,用于引导图像重建实验。这种方法在一定程度上可以控制修复过程,使得修复结果符合特定的要求,如图 6 所示。其中,图 6(a)为掩膜图像,图 6(b)为重建图像,图 6(c)为原始图像。通过实验结果可以观察到,当图像中的目标被遮住时,该区域的上下文信息会受到破坏。然而,本文模型能够根据给定的文本描述对缺失区域进行合理的重建,使得重构后的图像看起来自然且与周围图像保持一致。这有助于改善图像的质量和视觉效果。通过引入人为指定的文本描述,使得图像修复更具可控性,能够满足用户特定需求。

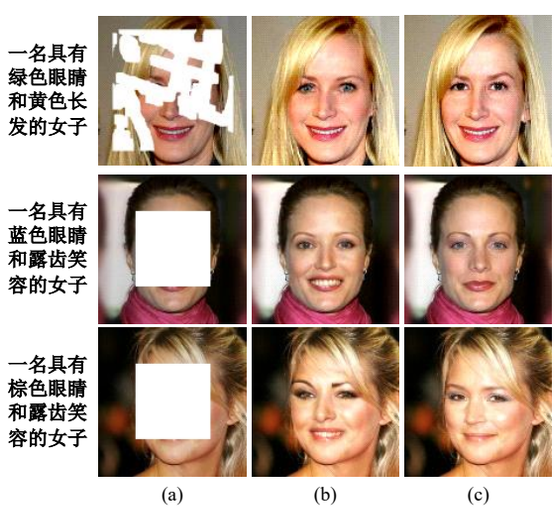

图 6　指定文本下的重建图像

Fig. 6　Reconstructing an image with specified text

4　结　论

(1) 提出了一种文本引导图像修复的方法,为图像重建提供了丰富的语义信息,解决了在遇到较大区域缺失时可利用的上下文信息不足的问题。

(2) 提出的视觉-文本模态融合模块和注意力感知层,通过融合图像和文本特征来指导图像修复。这使得修复网络能够重建出在整体结构上保持一致,并具有合理语义的内容。

(3) 在 CelebA-HQ 数据集上进行了定量和定性分析以及消融研究。实验结果证明了本文方法能够额外利用文本特征,指导生成更自然和上下文语义一致的修复结果。

(4) 本文方法在人脸识别、照片修复和医学影像处理等领域均具有参考和借鉴价值。除了针对 RGB 人脸图像的重建外,未来还将考虑其他类型的人脸图像重建任务,如红外图像、深度图像等。对于不同模态的图像修复,还需进一步深入探索。

参考文献:

[1] 周大可, 张超, 杨欣. 基于多尺度特征融合及双重注意力机制的自监督三维人脸重建[J]. 吉林大学学报: 工学版, 2022, 52(10): 2428-2437.
Zhou Da-ke, Zhang Chao, Yang Xin. Self-supervised 3D face reconstruction based on multi-scale feature fusion and dual attention mechanism[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(10): 2428-2437.

[2] 王小玉, 胡鑫豪, 韩昌林. 基于生成对抗网络的人脸铅笔画算法[J]. 吉林大学学报: 工学版, 2021, 51(1): 285-292.
Wang Xiao-yu, Hu Xin-hao, Han Chang-lin. Face pencil drawing algorithms based on generative adversarial network[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(1): 285-292.

[3] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: feature learning by inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2536-2544.

[4] Iizuka S, Simo S E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.

[5] Yan Z, Li X, Li M, et al. Shift-net: image inpainting via deep feature rearrangement[C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 1-17.

[6] Liu H, Wan Z, Huang W, et al. Pd-gan: probabilistic diverse gan for image inpainting[C]//Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 9371-9381.

[7] Wan Z, Zhang J, Chen D, et al. High-fidelity pluralistic image completion with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, USA, 2021: 4692-4701.

[8] Li W, Lin Z, Zhou K, et al. Mat: mask-aware transformer for large hole image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 10758-10768.

[9] Huang W, Deng Y, Hui S, et al. Sparse self-attention transformer for image inpainting[J]. Pattern Recognition, 2024, 145: 109897.

[10] Lian J, Zhang J, Liu J, et al. Guiding image inpainting via structure and texture features with dual encoder [J]. The Visual Computer, 2024, 40: 4303-4317.

[11] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J/OL]. [2023-12-16]. arXiv preprint arXiv: 1810.04805v2.

[12] Johnson J, Alahi A, Fei F L. Perceptual losses for real-time style transfer and super-resolution[C]//14th European Conference, Amsterdam, The Netherlands, 2016: 694-711.

[13] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115: 211-252.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. [2023-12-17]. arXiv preprint arXiv: 1409.1556v6.

[15] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2794-2802.

[16] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2414-2423.

[17] Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 85-100.

[18] Li J, Wang N, Zhang L, et al. Recurrent feature reasoning for image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 7760-7768.

[19] Lugmayr A, Danelljan M, Romero A, et al. Repaint: inpainting using denoising diffusion probabilistic models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 11461-11471.

[20] Chen L, Yuan C, Qin X, et al. Contrastive structure and texture fusion for image inpainting[J]. Neurocomputing, 2023, 536: 1-12.

[21] Li A, Zhao L, Zuo Z, et al. MIGT: multi-modal image inpainting guided with text[J]. Neurocomputing, 2023, 520: 376-385.

[22] Zhang L, Chen Q, Hu B, et al. Text-guided neural image inpainting[C]//Proceedings of the 28th ACM International Conference on Multimedia, New York, USA, 2020: 1302-1310.