计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):124−132 [doi: 10.15888/j.cnki.csa.009794] [CSTR: 32024.14.csa.009794]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向人脸图像超分辨率重建的 CNN-Mamba 属性引导网络①

刘晓亚, 韦姿煜, 周　迪, 宋廷强, 孙媛媛

(青岛科技大学 数据科学学院, 青岛 266061)
通信作者: 孙媛媛, E-mail: yysun@qust.edu.cn

摘　要: 针对现有的方法通常面临全局感受野和高效计算之间难以有效平衡以及重建图像细节不清晰的问题, 提出了基于 CNN-Mamba 的属性引导网络 (CMANet). 首先, 模型在进行重建时, 引入了属性信息并且考虑了这些属性之间的相互关系, 帮助模型提高整个重建过程的可靠性和精确度. 其次, 引入了沙漏状态空间模块, 发掘人脸图像的关键特征, 并保持了在长距离依赖建模方面具有线性复杂度的优势. 最后, 引入了自适应 Mamba 融合模块, 在图像特征学习多个方向长距离依赖关系时, 将属性针对不同方向进行自适应补充, 并将不同方向补充后的特征进行自适应融合, 使得模型在处理多样化的图像时更加灵活和高效. 大量的实验证明了所提方法的优越性.

关键词: 人脸图像; 属性; 超分辨率重建; 状态空间模块; 图注意力网络; 自注意力机制

引用格式: 刘晓亚,韦姿煜,周迪,宋廷强,孙媛媛.面向人脸图像超分辨率重建的 CNN-Mamba 属性引导网络.计算机系统应用,2025,34(3):124–132. http://www.c-s-a.org.cn/1003-3254/9794.html

Attribute-guided Network Based on CNN-Mamba for Super-resolution Reconstruction of Face Images

LIU Xiao-Ya, WEI Zi-Yu, ZHOU Di, SONG Ting-Qiang, SUN Yuan-Yuan

(School of Data Science, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Aiming at the difficult balance between the global receptive field and efficient computation and unclear details of image reconstruction, an attribute guided network based on CNN-Mamba (CMANet) is proposed. Firstly, when the model is reconstructed, attribute information is introduced and interrelationships among these attributes are considered, which helps the model to improve the reliability and accuracy of the whole reconstruction process. Secondly, the hourglass state space module is introduced to explore the key features of face images and maintain the advantage of linear complexity in long-distance dependency modeling. Finally, an adaptive Mamba fusion module is introduced. When image features learn long-distance dependencies in multiple directions, attributes are adaptively supplemented in different directions, and features supplemented in different directions are adaptively fused, making the model more flexible and efficient in processing diverse images. A large number of experiments prove the superiority of the proposed method.

Key words: face image; attribute; super-resolution reconstruction; state space module; graph attention network; self-attention mechanism

人脸图像超分辨率重建, 也称为面部幻觉, 是一种将给定的低分辨率人脸图像恢复为高分辨率人脸图像的技术. 高分辨率人脸图像可以捕捉更多面部细节, 能够提供更多的特征点, 在人脸识别与身份验证、医疗

① 基金项目: 国家自然科学基金 (32301702); 山东省自然科学基金 (ZR2021QC120)
收稿时间: 2024-09-04; 修改时间: 2024-09-30; 采用时间: 2024-10-14; csa 在线出版时间: 2025-01-21
CNKI 网络首发时间: 2025-01-22

与健康以及广告等多个方面有着广泛的应用.

许多流行的方法采用 CNN 来学习 LR 到 HR 之间的映射[1–3]. Song 等[4]提出使用两个阶段生成人脸图像. 首先, 使用卷积神经网络生成输入图像的人脸组成部分, 然后从高分辨率训练图像中合成细粒度面部结构, 并将这些细节转移到面部组成部分中以进行增强. 但人脸图像通常具有全局的几何结构, 使用卷积核处理人脸图像虽可以较好地获取图像的局部特征, 但无法有效捕获浅层特征的长距离依赖关系, 因此, 很难获得好的重建图像. Zeng 等[5]提出了一个三阶段的自注意力学习网络, 探索低级和高级空间的相互依赖关系, 补充人脸图像重建过程中缺失的信息. Xie 等[6]使用 EM 距离和 RMSProp 来改进基于 GAN 的模型, 以提升重建人脸图像视觉效果. Kim 等[7]利用同一个人的高分辨率图像作为参考图像, 帮助当前低分辨率人脸图像进行重建. 得益于 Transformer 能够捕捉序列中远程依赖关系的能力, Gao 等[8]提出了局部-全局特征协同模块, 将 CNN 与 Transformer 结合, 以同时捕获局部面部纹理细节和全局面部结构. 然而 Transformer 中自注意力机制在计算时, 每个元素都需要与序列中的每个其他元素进行计算. 因此, 随着输入序列长度增加, 计算量会呈二次方增长, 导致计算负担. 最近出现的 Mamba 其核心组件状态空间模块[9–12], 具有输入自适应和全局信息建模的能力, 同时保持线性复杂性, 其变体在自然语言处理、图像分类以及医学图像分割等多个方面都获得了良好效果[13–15]. 我们将其引入到人脸图像超分辨率重建任务中, 以提高远程像素的利用与多模态融合的效率.

此外, 人脸图像具有丰富的面部先验信息如面部解析图、关键点热图以及属性信息等, 可以利用这些先验信息来帮助人脸图像重建. Teng 等[16]提出了一个多先验协作网络, 结合了生成先验和面部特定几何先验的优势. Li 等[17]提出了局部和全局交互式混合网络, 利用面部语义和几何先验获得更具辨别性的结果. 然而, 对于低分辨率人脸图像来说, 通常丢失了大量的细节信息, 面部关键点或解析图可能变得模糊, 且人脸的姿态变化和光照条件的变化等也会影响其准确性. 相比之下, 人脸属性通常表现出相对的稳定性. 但目前使用人脸属性信息进行人脸图像超分辨率重建, 并没有充分考虑属性之间的关系, 且仅通过简单的嵌入操作, 很难充分利用属性信息.

为了解决上述问题, 我们提出了一个基于 CNN-Mamba 的属性引导网络: CMANet. 该网络在重建图像时, 引入了属性信息, 并通过自注意力机制和图注意力网络充分考虑了属性之间的关系, 使得模型能够更好地生成高质量的重建图像. 具体来说, 网络采用类似 U-Net 结构, 通过残差主模块来提取不同尺寸下的特征信息. 在残差主模块中, 设计了沙漏状态空间模块, 通过不同方向的特征捕捉人脸图像中的不同细节和模式, 并在模块中引入沙漏结构以提取和利用多尺度特征, 确保重建结果在保持全局结构和细节精确性方面的良好表现. 获得当前尺度下的特征后, 将其与属性信息一起输入到自适应 Mamba 融合模块. 该模块对不同方向展开的特征信息赋予不同属性信息的补充, 并通过可学习因子将增强后的特征进行自适应融合, 提升了模型的特征表达能力并节省了计算资源.

1 本文方法

1.1 网络概述

我们使用类似 U-Net 网络进行人脸图像超分辨率重建, 在本文中低分辨率图像、高分辨率图像以及超分辨率图像分别表示为: I_{LR}、I_{HR}、I_{SR}. 整体结构如图 1 所示. 网络实现主要包含 3 个阶段: 浅层特征提取阶段、深层特征提取阶段和重建阶段. 首先, 使用一个卷积层获取图像浅层特征 F_{pre}. 然后, 将对应的属性信息进行处理, 通过自注意力机制获取属性间的相关矩阵, 将该矩阵送入到图注意力网络中, 利用图网络的灵活性进一步分析属性之间的关系, 构建属性关系图, 得到属性信息 F_{attr}. 然后, 将 F_{pre} 和 F_{attr} 一起输入 CMANet 中, 通过自适应 Mamba 融合模块将属性信息与图像特征进行融合, 融合后的特征输入到残差主模块中用于进行特征提取和不同尺寸下的特征生成. 由于 HAT[18] 中的 HAB 可以通过激活更多的输入像素提高重建质量, 因此, 在 CMANet 的编解码器中间, 我们引用其对编码特征进行特征的细化增强. 得到输出特征 F_{out} 后, 通过重建模块中的卷积操作得到重建结果 I_{SR}.

1.2 残差主模块

先前的研究表明[3–8], 使用全局和局部信息可以有效帮助图像重建. 然而, 卷积运算由于其处理像素的局部性, 削弱了网络捕获长距离依赖关系的能力. 因此, 我们设计了残差主模块 (residual main module, RMM), 来探索不同尺度下的特征表示. 采用与标准 Swin Trans-

former 模块类似的结构, 不同的是由于自注意力的二次复杂性, 使用改进的 Mamba 模块来获取像素之间的远距离信息. 为了更多地关注关键面部信息, 添加沙漏结构[5−16]来改进 Mamba 模块, 以捕捉多个尺度的面部特征标志. 具体来说, 模块输入融合信息之后的增强特征, 经过 LayerNorm 层处理输入到沙漏状态空间模块, 在该模块进行特征图尺寸处理, 并进行主要的全面特征与关键特征提取操作. 处理后的特征送入 LayerNorm 与 MLP 中, 并伴随局部残差连接减轻梯度消失问题.

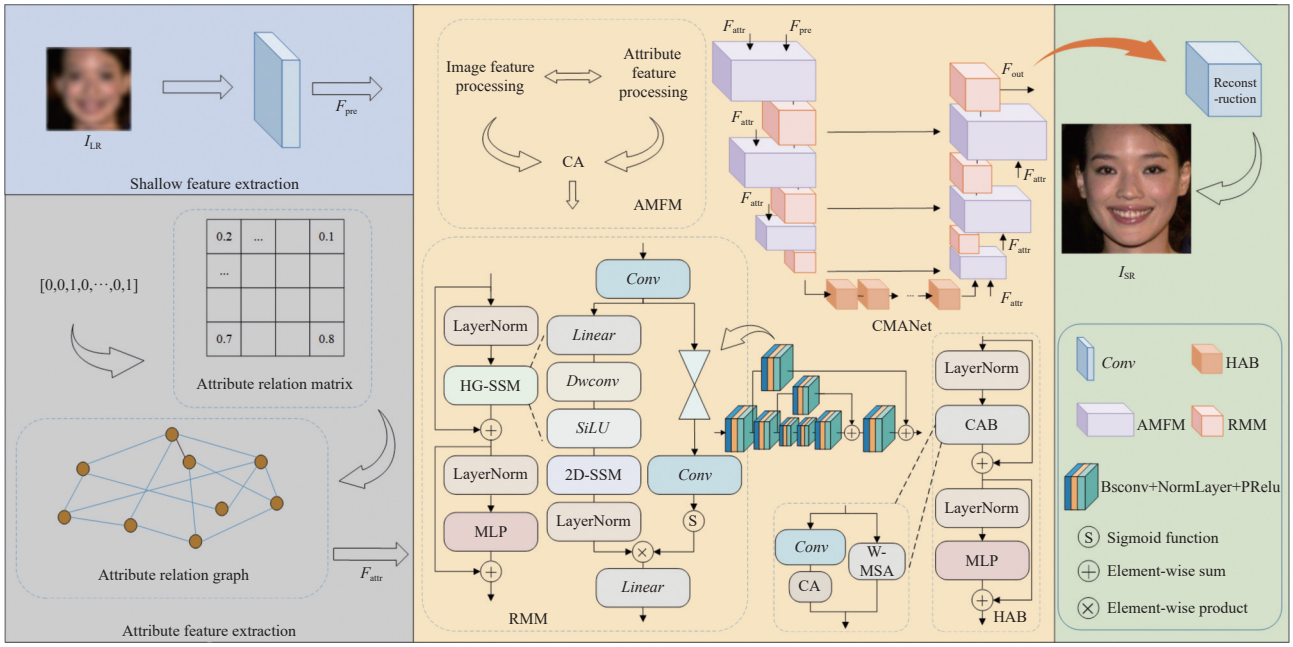

图 1 基于 CNN-Mamba 的属性引导网络架构

1.2.1 沙漏状态空间模块

作为残差主模块的核心模块, 沙漏状态空间模块 (hourglass state space module, HG-SSM) 将获取当前尺度特征, 并对特征进行进一步提取. 首先使用卷积操作对 LayerNorm 层处理后的特征 X_{pre} 进行特征提取, 获取目标尺寸的特征 X_{in}. 然后通过两个路径分别对特征进行处理. 在第 1 个路径中, 输入特征经过 *Linear*、*Dwconv* 以及 *SiLU* 激活后, 利用多个选择性扫描函数即 *2D-SSM*[13], 获取不同方向特征处理信息 X_1, 以保证特征提取的全面性. 在第 2 个路径中, 我们引入了轻量化的沙漏模块, 使用蓝图可分离卷积代替普通卷积进行轻量化处理, 并使用该模块来获取多尺度特征. 处理后的多尺度特征, 由于上下采样操作的存在, 可能会引入一些特征的混叠现象. 我们在输入该特征前, 加入了卷积层和 *Sigmoid* 函数, 这可以有效地消除这些混叠效应, 能够更好地捕捉局部细节信息, 使得最终的超分辨率结果更加干净和平滑. 经过第 2 个路径, 得到优化的多尺度关键特征 X_2. 最后, 我们通过相乘和 *Linear* 层将 X_1 与 X_2 进行融合, 得到最终输出特征 X_{out}. 其过程如下所示:

$$X_{\text{in}} = Conv(X_{\text{pre}}) \tag{1}$$

$$X_1 = LN(2D\text{-}SSM(SiLU(Dwconv(Linear(X_{\text{in}}))))) \tag{2}$$

$$X_2 = Sigmoid(Conv(BS\text{-}HourGlass(X_2))) \tag{3}$$

$$X_{\text{out}} = Linear(X_1 \times X_2) \tag{4}$$

其中, *Dwconv* 表示深度方向的卷积, *LN* 表示 LayerNorm.

1.2.2 属性处理

之前的一些方法[19,20]在利用属性信息时, 直接将原始属性信息添加到图像特征中, 没有考虑不同属性之间可能存在复杂的关系, 导致模型无法正确理解属性之间的影响. 基于此, 我们在进行属性信息的融入时, 先使用自注意力机制对属性信息进行全局性的关系建模, 强化重要区域. 通过计算每个属性的自相关性, 得到反映不同属性之间依赖关系的矩阵. 然后, 对该矩阵使用 *GAT* 以进一步挖掘属性之间更深层次的依赖关

系, 并通过分析局部属性间的关联来加强特征表示的细节和上下文信息. 两者结合可以使模型在处理人脸图像时, 既能关注整体的结构性特征, 也能保留局部的细节信息. 具体来说, 我们首先采用了归一化和嵌入层相结合的方法对属性信息进行编码. 将原本的值 −1 和 1 转换为 0 和 1, 通过 Embedding 层将 0 和 1 映射到更高维的特征空间, 增加表示能力, 使模型能够捕捉到更多的细微差别和复杂模式, 得到属性特征 X_{attr}. 然后, 通过自注意力机制处理, 得到 *Aout* 与 *Aweight*. 其中 *Aout* 是加权后的特征表示, 反映了不同部分的特征对整体人脸图像重建的重要性. *Aweight* 则显示了各个特征之间的相关性和重要性分布, 从而使得模型能够动态地关注图像中不同的区域. 然后, 我们利用 *Aout* 与 *Aweight* 创建图结构数据, 将其输入到图注意力网络中捕捉到更复杂的属性特征关系, 使得模型不仅能关注单一特征的重要性, 还能考虑不同属性特征之间的交互关系, 从而产生更自然、更高质量的超分辨率图像. 最后得到融合属性之间关系的属性特征 F_{attr}. 其过程如下.

$$Aout, Aweight = SA(X_{\text{attr}}) \tag{5}$$

$$F_{\text{attr}} = GAT(Aout, Aweight) \tag{6}$$

其中, *SA* 表示自注意力机制, *GAT* 表示图注意力网络.

1.3 自适应 Mamba 融合模块

原始的状态空间模块 (SSM) 仅支持单个输入, 为了整合图像特征与人脸属性特征, 我们对其进行了扩展, 并以此为基础得到了自适应 Mamba 融合模块 (adaptive Mamba fusion module, AMFM). 在 AMFM 中, 首先对图像特征和属性特征进行初步融合. 由于, 大多数基于 SSM 的融合方法并没有考虑到, 不同展开方向, 对另一种信息利用情况的不同, 导致图像感知质量下降. 因此, 我们在使用属性帮助人脸图像重建时, 对不同方向属性信息进行了调整, 使其更适合于当前展开方向属性信息的利用. 此外, 在将不同方向信息进行整合时, 大多数关于 SSM 的方法平等地融合来自不同方向生成的特征信息, 并没有考虑他们之间的相似性和差异性, 将处理后的特征平等地整合到生成的特征, 这将大大影响计算效率, 影响生成图像的质量. 为此, 我们使用可学习因子来控制每个方向生成的人脸图像特征. 由于使用 SSM 缺乏通道之间的交互, 我们在属性分支中, 将初步融合后的特征输入到了通道注意力机制中, 并将该特征与图像特征分支初步融合后的特征相加, 使图像特征可以在通道方面进一步融合. 具体来说, 对于输入图像特征 X^a 和属性特征 X^b. 首先, 经过 LayerNorm 进行归一化操作. 然后, 针对得到的两个特征分别进行信息补充. 对当前特征 X_i 进行补充时, 由于从 4 个方向进行扫描获取的远程依赖关系之间存在一定的差异, 因此, 对补充特征进行了自适应处理, 如算法 1 的第 16 行所示. 获取不同方向的图像属性融合特征后, 不同于直接将 4 个方向的特征进行整合, 在 AMFM 中, 考虑不同融合信息的相似性和多样性, 将各方向融合特征进行整合前, 首先通过可学习因子对最终融合特征的贡献程度进行了调节, 如算法 1 的第 18 行所示. 经过特征融合补充后, 我们得到图像特征 $Y^{a\prime}$ 与属性特征 $Y^{b\prime}$. 最后, 我们对 $Y^{b\prime}$ 应用通道注意力机制, 突出属性特征的关键信息. 由于当前属性特征经过了图像特征的补充, 因此将通道注意力机制处理后的关键属性特征与图像特征直接进行相加整合, 得到最终融合后的特征 *Y*. 其伪代码和网络结构分别如算法 1 及图 2 所示.

算法 1. Adaptive Mamba fusion module (AMFM)

输入: X^a, X^b: (B,L,D).

输出: Y: (B,L,D).

1. X^a: $(B,L,D) \leftarrow Norm_a(X^a)$
2. X^b: $(B,L,D) \leftarrow Norm_b(X^b)$
3. **for** $i \in \{a,b\}$ **do**
4. z^i: $(B,L,E) \leftarrow Linear_z(X^i)$
5. **for** $j \in \{1,2,3,4\}$ **do**
6. x_j^i: $(B,L,E) \leftarrow Linear_x(X^i)$
7. x_j^t: $(B,L,E) \leftarrow Linear_t(X^t)$ ▷ t is the opposite of i
8. $x_j^{i\prime}$: $(B,L,E) \leftarrow SiLU(Conv1d(x_j^i))$
9. A: $(D,N) \leftarrow Parameter_A$
10. ▷ A represents D sets of structured $N \times N$ matrices
11. B: $(B,L,D) \leftarrow Linear_B(x_j^t)$
12. C: $(B,L,D) \leftarrow Linear_C(x_j^t)$
13. Δ: $(B,L,D) \leftarrow \log(1+\exp(Linear(x_j^t + Parameter_\Delta)))$
14. $\overline{A}$: $(B,L,D,N) \leftarrow \exp(\Delta \odot A)$
15. $\overline{B}$: $(B,L,D,N) \leftarrow \Delta \odot B$
16. y_j^i: $(B,L,E) \leftarrow SSM(\overline{A},\overline{B},C) \odot (\alpha_j^i, x_j^{i\prime})$
17. **end for**
18. y^i: $(B,L,E) \leftarrow \sum_{k=1}^{4} \beta_i^k \odot y_k^i$
19. Y^i: $(B,L,E) \leftarrow y^i \odot SiLU(z^i)$
20. **end for**
21. $Y^{a\prime}$: $(B,L,D) \leftarrow Linear_a(Y^a) + Y^a$
22. $Y^{b\prime}$: $(B,L,D) \leftarrow Linear_b(Y^b) + Y^b$
23. Y: $(B,L,D) \leftarrow Y^{a\prime} \odot CA(Y^{b\prime})$ ▷ CA represents channel attention

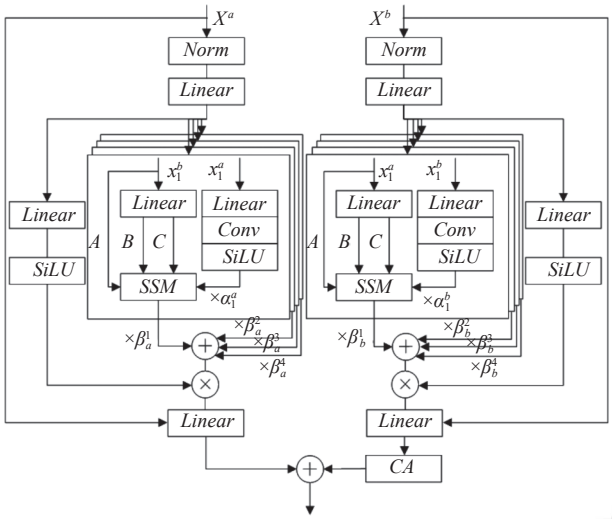

图 2 AMFM 网络结构图

1.4 损失函数

我们使用端到端的方式优化模型, 为了保证人脸重建图像在像素方面的准确性、感知质量以及语义方面的准确性. 我们使用了像素损失、感知损失和语义分割损失共同构建目标损失函数. 其中, 使用 L1 范数 (L1-norm) 作为 ISR 和 IHR 之间的像素级损失. 在感知损失中, 使用预训练好的 VGG19[21]来提取图像的人脸特征, 然后计算结果图像与 HR 图像之间的 L1 距离. 由于人脸图像相比于其他图像具有较强结构信息, 为了更好地利用这一优势, 我们引入了语义分割损失. 使用 BiSeNet[22]对图像进行语义分割生成人脸图像的解析图, 并对 SR 与 HR 解析图进行交叉熵损失的计算. 其公式如下所示:

$$\mathcal{L}_{\text{pix}}=\frac{1}{N}\sum_{i=1}^{N}\|G(I_{\text{LR}}^{i})-I_{\text{HR}}^{i}\|_1 \tag{7}$$

$$\mathcal{L}_{\text{pcp}}=\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\text{VGG}}}\frac{1}{M_{\text{VGG}}^{l}}\|f_{\text{VGG}}^{l}(I_{\text{SR}}^{i})-f_{\text{VGG}}^{l}(I_{\text{HR}}^{i})\|_1 \tag{8}$$

$$\mathcal{L}_{\text{semseg}}=\frac{1}{N}\sum_{i=1}^{N}\left\{-\sum_{j=1}^{C}Seg_{j}^{i}\times\log(Pseg_{j}^{i})\right\} \tag{9}$$

$$\mathcal{L}=\lambda_{\text{pix}}\mathcal{L}_{\text{pix}}+\lambda_{\text{pcp}}\mathcal{L}_{\text{pcp}}+\lambda_{\text{semseg}}\mathcal{L}_{\text{semseg}} \tag{10}$$

其中, N 表示模型训练中一个批次的数据量, G 表示提出的 CMANet, f_{VGG}^{l}表示 VGG 网络中第 l 层, C 表示 BiSeNet 中可以分割的种类数量, Seg 表示真实高分辨率图像的人脸语义分割图, $Pseg$ 表示重建图像的人脸语义分割图, λ_{pix}、λ_{pcp}、λ_{semseg} 分别表示像素损失、感知损失和语义分割损失的权重参数.

2 实验

2.1 数据集

我们使用 CelebA 数据集进行模型实验. 该数据集包含 202 599 张人脸图片, 且每一张图像都有相应的 40 个属性标记. 首先, 对原有的图片进行预处理操作, 使用 OpenCV 的人脸检测器识别人脸, 基于识别的信息裁剪 128×128 大小的人脸区域, 作为高分辨率图像, 并将其进行双三次下采样 8 倍处理, 得到相应的低分辨率图像. 取 168 854 张处理后的图像以及相应的属性对模型进行训练. 测试时, 在处理后的剩余图像中, 随机取 1 000 张图像将其与对应属性一起输入模型中进行模型总体性能的测试. 由于人脸图像的不同部位, 其重建的难度不同. 因此, 我们对特定属性人脸图像重建进行了评估测试. 选取“5_o_Clock_Shadow”“Arched_Eyebrows”“Bags_Under_Eyes”“Bushy_Eyebrows”“Pointy_Nose”“Smiling”属性, 分别选取 100、300、200、100、300、500 张图片进行特定属性重建性能评估.

2.2 实验细节

我们使用 PyTorch 框架来实现模型, 在 NVIDIA GeForce RTX 3090 上运行. 同时, 使用 Adam 并设置 β_1=0.9, β_2=0.99 来优化模型. 学习率固定为 2×10^{-4}, 批大小设置为 16, 训练周期设置为 20. 设置损失函数参数 λ_{pix}=1.0, λ_{pcp}=0.01, λ_{semseg}=0.1[23,24]. HAB 模块个数设置为 10.

为评估模型生成的超分辨率图像的质量, 使用了 3 个评价指标: 峰值信噪比 (peak signal to noise ratio, PSNR)、结构相似性 (structural similarity, SSIM)[25]以及学习感知图像块相似度 (learned perceptual image patch similarity, LPIPS)[26].

2.3 消融实验

2.3.1 各组件的有效性验证

我们进行了实验来证明所提出的 RMM 与 AMFM 的有效性, 将在网络框架中移除 RMM 与 AMFM 的剩余部分作为 Baseline. 如表 1 所示. 结果表明: (1) 在 Baseline 的基础上融入了 RMM, 在 PSNR/SSIM 中取得了 0.13 dB/0.002 2 的提升, 在 LPIPS 上降低了 0.023 6. 这是因为 RMM 可以获得比 Baseline 更全面的信息,

并突出人脸图像特征中对重建更重要的信息. (2) 在 Baseline 的基础上融入了 AMFM, 在 PSNR/SSIM 上提升了 0.14 dB/0.001 5, 在 LPIPS 上降低了 0.020 4. 相比于 Baseline 仅使用拼接方法进行属性信息的融合, 引入 AMFM 可以从多个方向进行自适应信息补充, 可以针对每个方向特定的局部特征进行细致的处理. 这种针对性的增强使得模型在重建细节时更为精确, 减少了重建误差, 因而提高了 PSNR 和 SSIM 值. (3) 在 Baseline 的基础上将提出的 RMM 与 AMFM 均进行了融合, 在 PSNR/SSIM 上提升了 0.24 dB/0.004 2, LPIPS 降低了 0.028 2. 通过比较表 1 中的数据, 可以发现虽然仅融入 AMFM 可以更加充分的利用属性信息, 但缺乏对融合特征的进一步特征提取. 如果仅融入 RMM, 虽然可以有效处理当前尺度特征, 但缺乏对属性信息的充分利用. 而将这两个模块均进行了融合, 可以结合各自优势并相互补充, 因此取得最佳结果.

表 1 选择不同组件的消融实验

Settings	PSNR (dB)	SSIM	LPIPS
Baseline	27.19	0.791 5	0.177 9
+ RMM	27.32	0.793 7	0.154 3
+ AMFM	27.33	0.793 0	0.157 5
+RMM +AMFM	**27.43**	**0.795 7**	**0.149 7**

2.3.2 总体性能

表 2 显示了我们提出的方法与其他最先进的超分辨率方法在人脸图像重建上的定量比较结果. 我们选择 Bicubic 的重建结果作为与其他方法进行比较的基准, 从表中数据可以看出, 由于 SwinIR、RCAN 以及 SAN 不是专门为人脸图像设计的, 重建的人脸图像性能不佳. 图 3 显示了不同方法之间的视觉比较结果, 可以注意到其无法很好地恢复面部特征. DICNet、CTCNet、SPARNet 和 SFMNet 是针对人脸图像的重建方法, 相比于之前的方法, 性能取得了一定的提升, 但恢复的人脸图像重要组件边界仍然模糊. 相反的, 我们提出的方法在 PSNR、SSIM 和 LPIPS 上均优于其他方法, 与直接通过 Bicubic 方法进行人脸图像重建相比, 在 PSNR/SSIM 上取得了 3.77 dB/0.157 9 的提升, 在 LPIPS 上降低了 0.386 0. 与最先进的 SFMNet 方法相比, 在 PSNR/SSIM 上提升了 0.13 dB/0.001 1, 在 LPIPS 上降低了 0.024 7. 这是因为我们提出的方法利用属性信息进行了自适应补充, 并通过从多个方向扫描图像特征, 更全面地捕捉到图像的细节信息. 从图 2 可以看出, 我们方法重建出的人脸图像其不仅可以有效恢复面部结构, 对于胡须、眼睛下面的卧蚕以及嘴角周围皱纹等面部纹理细节可以得到很好的恢复, 进一步证明了我们所提方法的有效性.

表 2 本文所提方法与其他超分辨率方法的定量比较

Method	PSNR (dB)	SSIM	LPIPS	Params (M)	Mul-Adds (G)
Bicubic	23.66	0.637 8	0.535 7	—	—
SwinIR[27]	25.74	0.720 5	0.303 3	3.6	0.92
RCAN[28]	26.32	0.755 5	0.238 5	15.7	4.7
SAN[29]	26.48	0.763 0	0.216 1	16	4.79
DICNet[30]	27.15	0.789 6	0.182 3	22.8	35.5
CTCNet[8]	26.84	0.778 0	0.206 5	22.41	47.17
SPARNet[23]	27.18	0.790 8	0.185 7	16.59	7.13
SFMNet[31]	27.30	0.794 6	0.174 4	8.6	30.6
Ours	**27.43**	**0.795 7**	**0.149 7**	19.63	3.34

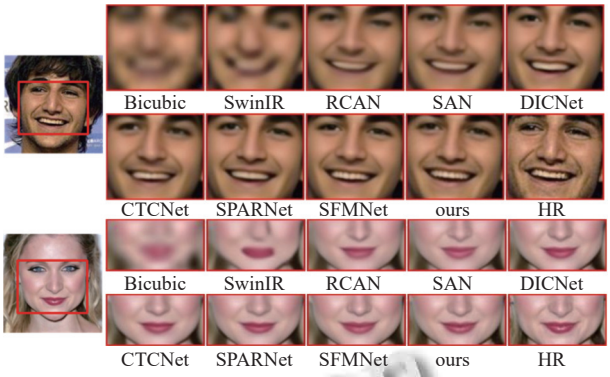

图 3 本文所提方法与其他超分辨率方法的视觉比较

2.3.3 对特定属性面部图像的评估

表 3 显示了不同方法在特定属性上的定量比较, 我们选取了 6 种属性作为本实验中要恢复的主要属性, 这些属性覆盖了眼睛、眉毛、鼻子以及嘴巴等区域, 恢复的难度之间存在差异. 根据表中数据可以看出, 对于上述属性“5_o_Clock_Shadow”“Arched_Eyebrows” “Bags_Under_Eye”“Bushy_Eyebrows”“Pointy_Nose” “Smiling”, 我们的方法与最先进的方法 SFMNet 相比在 PSNR 上分别提升了 0.18 dB、0.12 dB、0.17 dB、0.16 dB、0.12 dB、0.14 dB, 另外两个指标上也获得了最佳性能, 尤其对于“Bags_Under_Eye”属性, 其在 SSIM 得到了 0.002 2 的提升, 在 LPIPS 中下降了 0.022 7. 这源于我们的方法在获取不同方向的远距离依赖时, 可以更有效地捕捉找到这种特征的潜在模式, “5_o_Clock_Shadow”不同指标上的改进也证明了这一点. 综合上述属性数据, 可以看出所提方法在包含各种人

脸信息的情况下, 均可以得到好的重建结果. 图 4 展示了不同方法在这些属性上的视觉比较. 我们注意到, 在包含胡子和眉毛的属性中, 相比于其他方法, 我们的方法恢复的人脸图像毛发纹理以及边界轮廓更加清晰. 在“Pointy_Nose”属性中, 本文方法恢复的人脸图像, 鼻子更加挺拔, 和高分辨率图像更加接近. 在“Smiling”属性中, 虽然大多数方法总体可以恢复出嘴角微笑形状, 但微笑带动的嘴角附近面部扭曲以及鼻孔变化并不能清楚的表达, 而本文方法通过属性利用, 可以很好地恢复这些信息.

表 3 特定属性上不同方法之间的定量比较

Method	5_o_Clock_Shadow			Arched_Eyebrows			Bags_Under_Eyes		
	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS
Bicubic	23.59	0.6510	0.5318	23.80	0.6376	0.5320	23.86	0.6524	0.5311
SwinIR	26.01	0.7510	0.2891	25.73	0.7234	0.3036	26.10	0.7448	0.2947
RCAN	26.65	0.7816	0.2232	26.49	0.7599	0.2246	26.79	0.7743	0.2252
SAN	26.84	0.7898	0.2005	26.66	0.7677	0.2019	26.94	0.7815	0.2044
DICNet	27.68	0.8192	0.1631	27.35	0.7939	0.1730	27.75	0.8102	0.1673
CTCNet	27.33	0.8063	0.1904	27.01	0.7811	0.1964	27.40	0.7987	0.1930
SPARNet	27.72	0.8200	0.1682	27.37	0.7944	0.1754	27.78	0.8114	0.1700
SFMNet	27.84	0.8243	0.1537	27.47	0.7981	0.1643	27.90	0.8151	0.1589
Ours	**28.02**	**0.8264**	**0.1344**	**27.59**	**0.7984**	**0.1389**	**28.07**	**0.8173**	**0.1362**

Method	Bushy_Eyebrows			Pointy_Nose			Smiling		
	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS
Bicubic	24.04	0.6726	0.5057	23.83	0.6352	0.5302	24.23	0.6480	0.5224
SwinIR	26.66	0.7753	0.3034	25.78	0.7214	0.3016	26.20	0.7319	0.3034
RCAN	27.40	0.8069	0.1972	26.54	0.7576	0.2257	27.02	0.7687	0.2206
SAN	27.60	0.8143	0.1758	26.71	0.7655	0.2041	27.21	0.7773	0.1982
DICNet	28.60	0.8433	0.1412	27.43	0.7922	0.1738	27.94	0.8043	0.1671
CTCNet	28.11	0.8287	0.1674	27.05	0.7796	0.1973	27.59	0.7924	0.1898
SPARNet	28.58	0.8434	0.1443	27.42	0.7924	0.1769	27.99	0.8059	0.1689
SFMNet	28.80	0.8483	0.1298	27.54	0.7963	0.1650	28.08	0.8091	0.1573
Ours	**28.96**	**0.8490**	**0.1137**	**27.66**	**0.7964**	**0.1397**	**28.22**	**0.8102**	**0.1353**

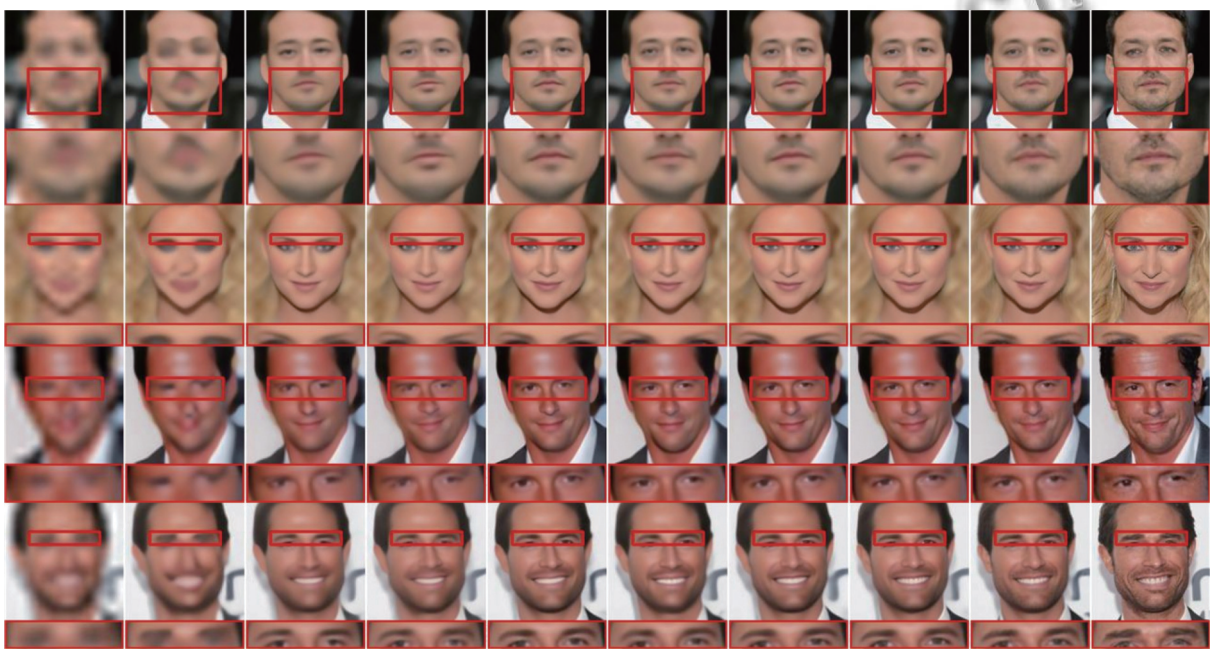

图 4 特定属性上不同方法之间的视觉比较

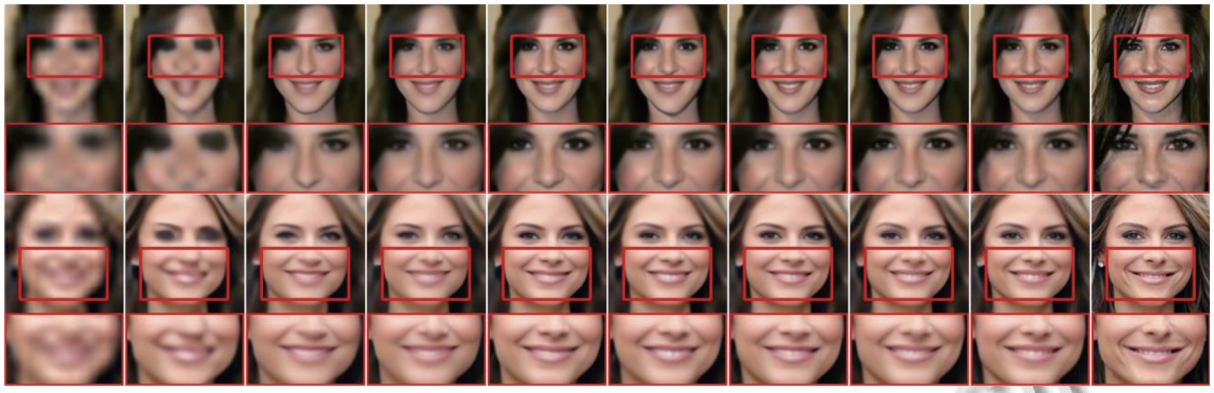

图 4 特定属性上不同方法之间的视觉比较 (续)

3 结论与展望

我们提出了一种基于 CNN-Mamba 的属性引导网络用于人脸图像超分辨率重建. 我们的模型融合了属性信息, 并分析属性之间的关系来帮助人脸图像重建过程. 我们提出了残差主模块, 该模块融合了沙漏结构和状态空间模块的优势, 以突出人脸图像特征提取过程中的全面性和关键特征. 此外, 提出了自适应 Mamba 融合模块, 将图像特征与属性信息在不同依赖关系中进行自适应融合. 最后, 在多个属性上对所提出的方法进行了评估, 其表现均优于其他方法. 在未来的研究工作中, 可以考虑融合语音描述以及人脸语义分割图等信息, 推动该领域的进一步发展.

参考文献

1 Lu T, Wang YZ, Zhang YD, *et al*. Face hallucination via split-attention in split-attention network. Proceedings of the 29th ACM International Conference on Multimedia. ACM, 2021. 5501–5509.

2 Chen YT, Phonevilay V, Tao JJ, *et al*. The face image super-resolution algorithm based on combined representation learning. Multimedia Tools and Applications, 2021, 80(20): 30839–30861. [doi: 10.1007/s11042-020-09969-1]

3 Hou H, Xu J, Hou YK, *et al*. Semi-cycled generative adversarial networks for real-world face super-resolution. IEEE Transactions on Image Processing, 2023, 32: 1184–1199. [doi: 10.1109/TIP.2023.3240845]

4 Song YB, Zhang JW, He SF, *et al*. Learning to hallucinate face images via component generation and enhancement. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: ijcai.org, 2017. 4537–4543.

5 Zeng KL, Wang ZY, Lu T, *et al*. Self-attention learning network for face super-resolution. Neural Networks, 2023, 160: 164–174. [doi: 10.1016/j.neunet.2023.01.006]

6 Xie ZM, Zhang N. Improved algorithm for GAN super-resolution face image reconstruction. Proceedings of the 5th IEEE International Conference on Civil Aviation Safety and Information Technology. Dali: IEEE, 2023. 285–289.

7 Kim JS, Ko K, Kim H, *et al*. RPF: Reference-based progressive face super-resolution without losing details and identity. IEEE Access, 2023, 11: 46707–46718. [doi: 10.1109/ACCESS.2023.3274841]

8 Gao GW, Xu ZX, Li JC, *et al*. CTCNet: A CNN-Transformer cooperation network for face image super-resolution. IEEE Transactions on Image Processing, 2023, 32: 1978–1991. [doi: 10.1109/TIP.2023.3261747]

9 Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

10 Smith JTH, Warrington A, Linderman SW. Simplified state space layers for sequence modeling. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.

11 Fu DY, Dao T, Saab KK, *et al*. Hungry hungry hippos: Towards language modeling with state space models. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.

12 Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752, 2024.

13 Guo H, Li JM, Dai T, *et al*. MambaIR: A simple baseline for image restoration with state-space model. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 222–241.

14 Ma J, Li FF, Wang B. U-Mamba: Enhancing long-range

dependency for biomedical image segmentation. arXiv:2401.04722, 2024.
15 Yue YB, Li ZZ. MedMamba: Vision mamba for medical image classification. arXiv:2403.03849, 2024.
16 Teng Z, Yu XS, Wu CD. Blind face restoration via multi-prior collaboration and adaptive feature fusion. Frontiers in Neurorobotics, 2022, 16: 797231. [doi: 10.3389/fnbot.2022.797231]
17 Li L, Zhang Y, Yuan L, *et al*. PLGNet: Prior-guided local and global interactive hybrid network for face super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 10166–10181. [doi: 10.1109/TCSVT.2024.3403713]
18 Chen XY, Wang XT, Zhang WL, *et al*. HAT: Hybrid attention Transformer for image restoration. arXiv:2309.05239, 2023.
19 Li MY, Zhang ZY, Yu J, *et al*. Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. IEEE Transactions on Multimedia, 2021, 23: 468–483. [doi: 10.1109/TMM.2020.2984092]
20 Srivastava A, Chanda S, Pal U. AGA-GAN: Attribute guided attention generative adversarial network with U-Net for face hallucination. Image and Vision Computing, 2022, 126: 104534. [doi: 10.1016/j.imavis.2022.104534]
21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.
22 Yu CQ, Wang JB, Peng C, *et al*. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 334–349.
23 Chen CF, Gong DH, Wang H, *et al*. Learning spatial attention for face super-resolution. IEEE Transactions on Image Processing, 2021, 30: 1219–1231. [doi: 10.1109/TIP.2020.3043093]
24 Zhao S, Sun JF, Ou HS, *et al*. A novel multi-task face super-resolution framework embedding degraded augmented GAN networks. Journal of Physics: Conference Series, 2022, 2303: 012061. [doi: 10.1088/1742-6596/2303/1/012061]
25 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]
26 Zhang R, Isola P, Efros AA, *et al*. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595.
27 Liang JY, Cao JZ, Sun GL, *et al*. SwinIR: Image restoration using swin Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1833–1844.
28 Zhang YL, Li KP, Li K, *et al*. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 294–310.
29 Dai T, Cai JR, Zhang YB, *et al*. Second-order attention network for single image super-resolution. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11057–11066.
30 Ma C, Jiang ZY, Rao YM, *et al*. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5568–5577.
31 Wang CY, Jiang JJ, Zhong ZW, *et al*. Spatial-frequency mutual learning for face super-resolution. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22356–22366.

(校对责编: 张重毅)