工程科学学报, 第 47 卷, 第 x 期: 1−10, 2025 年 x 月
Chinese Journal of Engineering, Vol. 47, No. x: 1−10, x 2025
https://doi.org/10.13374/j.issn2095-9389.2024.09.12.006; http://cje.ustb.edu.cn

GLIHamba: 基于 Mamba 的整体–局部上下文图像和谐化

孙金胜[1)], 潘　姣[1)], 郭　宇[2,3)], 姚　超[4,5)]✉

1) 北京科技大学智能科学与技术学院, 北京 100083　2) 北京科技大学北京材料基因工程高精尖创新中心, 北京 100083　3) 北京科技大学高精尖学院, 北京 100083　4) 北京科技大学计算机与通信工程学院, 北京 100083　5) 骨骼健康医疗大数据创新应用体系北京市重点实验室, 北京 100083

✉通信作者, E-mail: yaochao@ustb.edu.cn

摘　要　近年来, 包含 Transformer 组件的深度学习模型已经推动了包括图像和谐化在内的图像编辑任务的快速发展. 与使用静态局部滤波器的卷积神经网络 (CNN) 相反, Transformer 使用自注意力机制允许自适应非局部滤波来敏感地捕获远程上下文. 现有基于 CNN 和 Transformer 等方法图像和谐化方法, 未能很好的兼顾局部内容和整体风格的一致性, 导致前景与背景的视觉一致性不足. 本文提出了一种用于图像和谐化的新型网络模型, 基于 Mamba 的整体–局部上下文图像和谐化(Global-local context image harmonization based on Mamba, GLIHamba), 将全局特征和局部特征引入到 Mamba 模型, 建立具有整体–局部上下文感知能力的图像和谐化模型. 具体来说, 介绍了一种新的基于学习的图像和谐化模型 GLIHamba, 其核心组件包括局部特征序列提取器 (LFSE) 和全局特征序列提取器 (GFSE). LFSE 维护图像高维特征中相邻特征的局部一致性, 显式地确保空间上邻近的特征沿着通道保持一致性, 从而保证和谐化结果的局部内容完整一致. 另一方面, GFSE 在所有空间维度上建立全局序列, 保持图像的整体风格一致性. 研究结果表明, GLIHamba 提供了优于最先进的基于 CNN 和 Transform 的方法的性能.

关键词　图像和谐化; 状态选择空间模型; 图像编辑; Mamba 模型; 全局–局部特征提取

分类号　TG142.71

GLIHamba: global–local context image harmonization based on Mamba

SUN Jinsheng[1)], *PAN Jiao*[1)], *GUO Yu*[2,3)], *YAO Chao*[4,5)]✉

1) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China

2) Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China

3) School of Advanced Materials Innovation, University of Science and Technology Beijing, Beijing 100083, China

4) School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

5) Beijing Key Laboratory of Big Data Innovation and Application for Skeletal Health Medical Care, Beijing 100083, China

✉Corresponding author, E-mail: yaochao@ustb.edu.cn

ABSTRACT　Image harmonization is a technique that ensures the consistency and coordination of appearance features, such as lighting and color, between the background and foreground of a composite image. Image harmonization has emerged as a significant research area in the field of image processing. With the rapid development of image processing technologies in recent years, it has gradually become a focal point of attention in both academia and industry. The primary challenge in this research area is the development of image harmonization methods that achieve local content integrity and global style consistency. Traditional image harmonization methods rely primarily on matching low-level features, such as gradients and color histograms, to maintain good color coherence. However, these methods lack semantic awareness of the contextual relationship between the foreground and background, which leads to a lack of realism owing to inconsistencies between content and style. In recent years, harmonization methods based on deep learning have achieved

收稿日期: 2024−09−12

基金项目: 国家自然科学基金资助项目(62332017, 62306032, U22A2022); 广东省自然科学基金资助项目(2023A1515030177, 2022A1515110350)

significant progress. Pixel-wise matching methods utilize convolutional encoder-decoder models to learn the transformations from background to foreground pixel features. However, because of the limited receptive fields of convolutional neural networks (CNNs), these methods primarily use local regional features as references, which makes it difficult to incorporate the overall background information into the foreground. In contrast, region-based matching methods treat the foreground and background regions as two different styles or domains. Although these methods achieve global consistency in harmonization results, they often overlook the spatial differences between the two regions. Breakthroughs in state-space models (SSMs), particularly the Mamba model based on the selective state-space model, have brought about significant advancements. The Mamba model utilizes a selective scanning mechanism to achieve linear complexity in capturing global relationships and demonstrated excellent performance in a series of computer vision tasks. However, the Mamba model cannot maintain spatial local dependencies between adjacent features and thus lacks local consistency. In this study, we draw inspiration from the operational methods of CNNs and transformer models as well as introduce global and local features into the Mamba model to establish an image harmonization model with global-local context awareness. Specifically, we propose a novel learning-based image harmonization model called GLIHamba (Global-local context image harmonization based on Mamba). The core components of GLIHamba include a local feature sequence extractor (LFSE) and global feature sequence extractor (GFSE). The LFSE preserves the locality of adjacent features in high-dimensional arrays to explicitly ensure consistency among spatially neighboring features along the channels and thereby guarantee the local content integrity and consistency of the harmonization results. In contrast, GFSE compresses features across all spatial dimensions to maintain the overall style consistency of the image. Our experimental results demonstrate that the proposed GLIHamba model outperforms previous methods based on CNNs and transformers in image harmonization tasks. On the iHarmony4 dataset, our model achieved a PSNR value of 39.76 and exhibited excellent performance on real scene data. In summary, the proposed GLIHamba model provides a novel solution to the challenges of image harmonization by integrating global and local context awareness and thus achieves superior performance compared with existing methods.

KEY WORDS　image harmonization; selective state-space modeling; image editing; mamba model; global–local feature extraction

图像和谐化是使合成图像背景与前景的外观特征(如光照、颜色等)一致协调化的技术,作为图像处理领域的重要研究方向,近年来伴随着图像处理技术的迅猛发展,逐渐成为了学界和业界的一个关注焦点[1−4]. 随着生成对抗网络(Generative adversarial nets, GANs)[5−6]和扩散模型(Diffusion models)[7−8]等深度学习技术的涌现,显著提高了图像生成与编辑的品质与效率,同时对合成图像的和谐化技术也提出了更为严苛的要求[9]. 这些先进的技术能够生成出既逼真又富有创造性的图像内容,但与此同时也突出了合成图像的前景与背景在内容、风格等方面不协调的问题. 图像和谐化的内涵已远远超出了色彩、光影调整的范畴,更加关注图像局部元素之间的逻辑连贯性以及整体的视觉真实性[10]. 在这一背景下,实现局部内容完整,整体风格一致的图像和谐化方法,成为了当前研究领域的主要挑战.

传统的图像和谐化方法主要基于底层特征(如梯度、颜色直方图等)匹配等方法[11−12],能够保持较好的颜色连贯性,但缺少对前后内容语义感知,因而内容与风格不协调导致缺乏真实性. 近年来,基于深度学习的和谐化方法取得了显著的进展. 目前的方法主要有两种思路:基于像素间匹配方法和基于区域间匹配的方法. 其中,前者使用卷积编码器–解码器模型来学习背景像素特征到前景像素特征的转换[13−15]. 但是因为 CNN 感受野的限制,这些方法主要基于局部区域特征作为参考,背景的整体信息无法附加到前景当中. 后者方法将前景和背景区域分为两种风格或两种域,通过风格特征匹配和域鉴别器等技术将和谐化问题转化为风格(或域)迁移任务来处理[16−19]. 这些方法的和谐化结果具有整体一致性,但忽略了两个区域的空间差异.

状态空间模型(State space model, SSM)的突破,特别是以选择性状态空间模型(Selective state space model, S6)为基础的 Mamba 模型,其利用选择性扫描机制实现线性复杂性的全局关系[20−21]. Mamba 已成功应用于一系列计算机视觉任务中,显示出卓越的构建全局关系的性能[22−25]. 但是,Mamba 状态参数容量的有限性以及序列顺序建模的必要性,其很难维持邻近和远处令牌之间的依赖关系,同时无法保持空间相邻特征之间的相对关系,因此不具备局部一致性. 本文中,借鉴 CNN 和 Transformer 模型[26]的操作方法,将全局特征和局部特征引入到 Mamba 模型,建立具有整体–局部上下文感知能力的图像和谐化模型 GLIHamba (Global-local context image harmonization based on Mamba). 具体来说,GLIHamba 利用 Mamba 模型的

视觉状态空间(Visual state space, VSS),构建的核心组件包括局部特征序列提取器 (LFSE) 和全局特征序列提取器 (GFSE). LFSE 维护图像特征的高维数组中相邻特征的空间性,显式地确保空间上邻近的特征沿着通道保持一致性,从而保证和谐化结果的局部内容完整一致. 另一方面,GFSE 在所有空间维度上压缩特征,保持图像的整体风格一致性.

本文的主要创新点如下:

(1) 提出一种基于 Mamba 的融合整体–局部上下感知能力的图像和谐化模型 GLIHamba. 通过将全局特征和局部特征引入到 Mamba 模型,获得了具有局部内容和整体风格一致的和谐化结果.

(2) 在选择性状态空间模型基础上,提出了两种特征序列提取器,LFSE 和 GFSE,增强 SSM 模块提取合成图像中局部和整体的特征的能力,建立前景与背景的语义特征与风格特征关联关系.

(3) 实验结果表明本文提出的 GLIHamba 模型在图像和谐化任务中超过了之前基于 CNN 和 Transformer 的方法,在 iharmony4 数据集,本文的模型的峰值信噪比(PSNR)值达到 39.76,并在真实场景的数据中具有优异的表现.

1 相关工作

1.1 图像和谐化

图像和谐化的目标是调整合成图像的前景图像的外观和视觉风格,使其与背景图像和谐一致[10]. 图像和谐化技术包括传统方法和基于深度学习的方法. 传统的图像和谐化方法采用低层级的外观统计数据匹配的方式,如匹配局部和全局的颜色分布、调整局部像素和全局梯度[11] 以及多尺度的颜色直方图均衡化[12] 等方法. 近年来,随着 RealHM[13], iHarmony4[14] 等图像和谐化的数据集公开,基于深度学习的图像和谐化成为了研究的热点. 一些工作基于图像转换和颜色迁移的方法实现前景与背景一致性效果. Tsai 等[13] 使用编码器–解码器的卷积模型来感知前景背景之间像素到像素的风格转换. Cong 等[14] 把域迁移思路引入到和谐化中,采用了一种验证鉴别器来区分前景域和背景域. Hang 等[16] 在风格迁移块中加入了背景注意计算,并引入了对比学习的思想. 此外,Guo 等[27] 基于 Retinex 理论,利用自编码器将图像分解为反射率和照度进行分离和谐化. 这些方法将当前像素与局部参考进行和谐化,但未能有效捕捉全局上下文的外观信息,导致前景像素无法附加背景参考. 为了学习前景与背景的全局依赖关系,Ling 等[17] 将图像和谐化任务重构为从背景到前景的风格转移问题,通过渲染前景图像以保持背景图像的相似视觉风格. 其在编码–解码的 U 型卷积架构基础上,采用自适应实例正则化模块使前景的颜色特征与背景的整体颜色信息保持一致. Guo 等[18–19] 基于 Transformer 模型[28–29] 调整前景光使其与背景光兼容,同时保持结构和语义不变. 这些方法的结果具有较好的色彩光照一致性,但是当背景前景差异性较大时,容易因为过度和谐化导致前景视觉真实性不足. Chen 等[15] 提出图像和谐化任务在考虑整体一致性的前提下,应该关注背景不同区域对前景的影响,利用局部动态掩模感知方法,分层地适应模型的参数和特征. 上述方法忽略了空间邻近先验,不能很好地模拟远距离参考. 基于此,本研究着重探索了基于 Mamba 模型建立全局指导策略的图像和谐化方法.

1.2 Mamba 模型

状态空间模型(SSM),特别是结构化状态空间模型(S4)在序列分析中显示出巨大的潜力[20], 它们能够以线性计算复杂度进行长距离序列建模. 通过将选择机制引入到中,Mamba[21] 进一步优化了其上下文压缩能力,并且性能优于 Transformers. 其允许模型根据输入的相关性动态地调整其关注点,使得 Mamba 模型在处理长序列时更加高效和准确.

考虑到其在长序列数据处理中的卓越性能,许多研究探索了 Mamba 在计算机视觉中的潜力,并取得了有希望的发展. 其中,视觉状态空间模型(VMamba)[22] 和视觉 Mamba(Vision Mamba)[23] 由于其在建模长距离依赖方面的效率,最近已成为各种计算机视觉任务的有力工具. 基于此,一系列视觉状态空间模型被提出,应用于医学图像分析[24] 和图像编辑[30–32] 等领域. U-Mamba 通过提出混合 CNN–SSM 块,有效地扩展了 Mamba 用于生物医学图像分割的能力[33]. UVM–Net 针对图像去雾任务,建立了特征与非通道域的长依赖关系[25]. Chen[34] 将 Mamba 模型与 Transformer 模型相结合,实现了模型在图像在像素(Pixel)和图像块(Patch)两个层面的交互学习,提升了图像补全的上下文一致性. 受这些工作的启发,本文将 Mamba 模型应用于图像和谐化,有效建立背景与前景风格之间的长距离依赖关系,提升复杂场景下图像和谐化的鲁棒性.

2 Mamba 模型核心概念

Mamba 是一种新的序列架构,集成了结构化

状态空间序列模型(S4)来管理数据序列. Gu 和 Dao[21] 在 S4 的基础上引入了选择性扫描机制, 提出了选择性状态空间模型(S6). 下面, 将简要阐述 Mamba 相关的核心概念.

2.1　结构化状态空间序列模型

对于线性不变系统的连续信号的 SSM 模型, 用 $\boldsymbol{h}(t)$ 表示时间 t 时刻模型的隐藏状态, 用 $\boldsymbol{x}(t)$ 和 $\boldsymbol{y}(t)$ 分别表示模型的输入和输出, 可以用如下公式描述:

$$\begin{aligned}&\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t),\\&\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t)\end{aligned}\tag{1}$$

其中: $\boldsymbol{h}'(t)$ 表示隐藏状态 $\boldsymbol{h}(t)$ 关于时间 t 的导数; $\boldsymbol{A}$、$\boldsymbol{B}$ 和 $\boldsymbol{C}$ 为可学习的模型参数矩阵, 并且独立于模型输入和时间.

为了处理离散输入信号, S4 引入零阶保持(Zero-order hold, ZOH)技术, 即每当接收到一个离散信号时, 就保持该信号值不变, 直到下一个离散信号出现. 保持信号值的时间由步长参数 $\boldsymbol{\Delta}$ 表示. 应用 ZOH 技术可以将连续参数 $\boldsymbol{A}$、$\boldsymbol{B}$ 转化为离散化参数 $\overline{\boldsymbol{A}}$、$\overline{\boldsymbol{B}}$, 公式如下:

$$\begin{aligned}&\overline{\boldsymbol{A}}=\boldsymbol{\Delta}\boldsymbol{A},\\&\overline{\boldsymbol{B}}=(\boldsymbol{\Delta}\boldsymbol{A}(\mathrm{e}^{\boldsymbol{\Delta}\boldsymbol{A}}-\mathbf{I})\cdot\boldsymbol{\Delta}\boldsymbol{B}\end{aligned}\tag{2}$$

其中, $\mathbf{I}$ 表示单位矩阵.

这样, 将 SSM 模型由连续函数 $\boldsymbol{x}(t)$ 到函数 $\boldsymbol{y}(t)$ 的映射, 变为离散序列 $\boldsymbol{x}_k$ 到序列 $\boldsymbol{y}_k$ 的映射, k 表示离散化的时间步长. 离散化后的 SSM 模型可以表示为:

$$\begin{aligned}&\boldsymbol{h}_k=\overline{\boldsymbol{A}}\boldsymbol{h}_{k-1}+\overline{\boldsymbol{B}}\boldsymbol{x}_k,\\&\boldsymbol{y}_k=\boldsymbol{C}\boldsymbol{h}_k\end{aligned}\tag{3}$$

为了有效地并行化训练, 这个递归过程被表示为卷积化的形式:

$$\begin{aligned}&\overline{\boldsymbol{K}}=(\boldsymbol{C}\overline{\boldsymbol{B}},\boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}},\cdots,\boldsymbol{C}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}),\\&\boldsymbol{y}=\boldsymbol{x}*\overline{\boldsymbol{K}}\end{aligned}\tag{4}$$

其中, $\boldsymbol{x}$、$\boldsymbol{y}$ 分别表示输入、输出序列, L 表示输入序列的长度, $*$ 表示卷积运算, $\overline{\boldsymbol{K}}$ 表示卷积核.

S4 结合了循环模型、卷积模型和连续时间模型的优点, 能够有效且高效地仿真长周期的依赖关系. 这使得它能够处理不规则采样的数据, 具有无限的上下文, 并在整个训练和测试过程中保持计算效率.

2.2　选择性状态空间模型

上述的 SSM 模型中, 模型参数相对于输入和时间动态保持不变. Gu 和 Dao[21] 认为这种时不变特性不利于上下文推理, 因此提出了选择性状态空间模型. 通过将 SSM 的参数设置为输入的函数来实现选择机制, 模型根据输入的相关性动态地调整关注点. 具体来说, 模型的参数依赖输入向量, 其关系为:

$$\begin{aligned}&\boldsymbol{B}=s_{\boldsymbol{B}}(\boldsymbol{x}),\\&\boldsymbol{C}=s_{\boldsymbol{C}}(\boldsymbol{x}),\\&\boldsymbol{\Delta}=\tau_{\boldsymbol{\Delta}}(s_{\boldsymbol{\Delta}}(\boldsymbol{x}))\end{aligned}\tag{5}$$

其中, $s_{\boldsymbol{B}}$、$s_{\boldsymbol{C}}$ 和 $s_{\boldsymbol{\Delta}}$ 均为 $\boldsymbol{x}$ 映射函数, $\tau_{\boldsymbol{\Delta}}$ 为 $softplus$ 函数. Mamba 采用了一种硬件感知算法来有效地计算选择性状态空间模型, 使其处理长序列更加高效和准确.

3　基于整体–局部视觉状态空间的图像和谐化模型

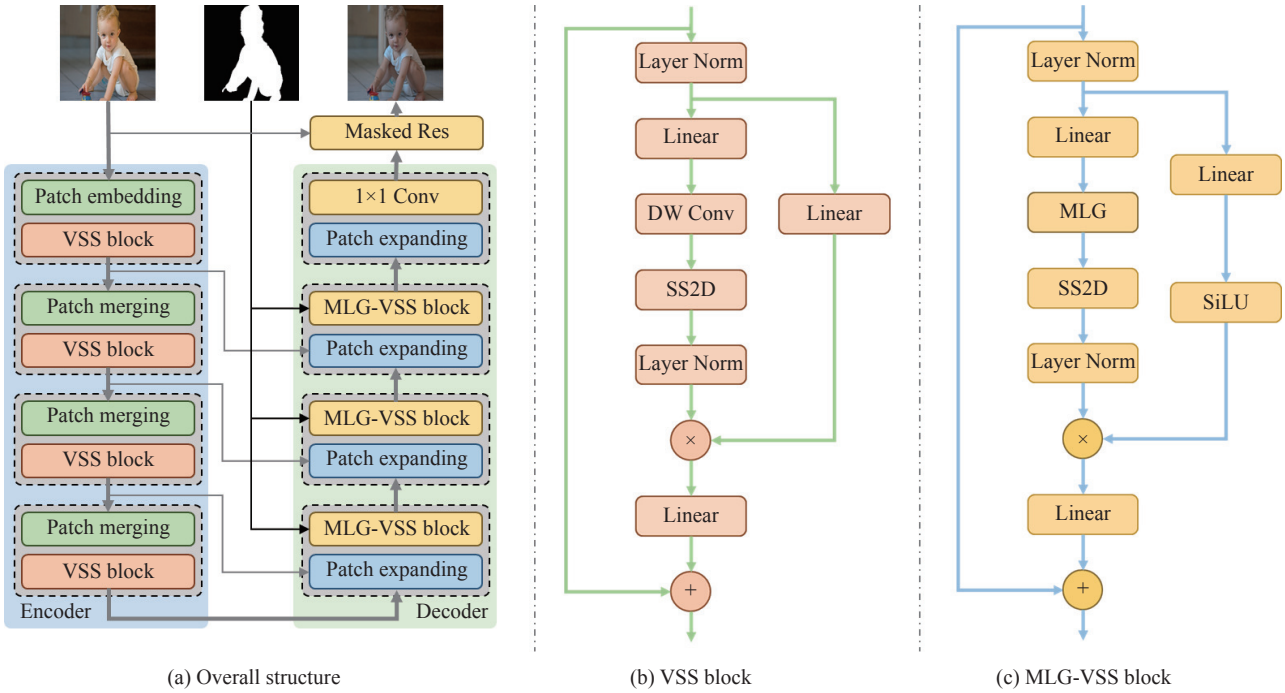
在本节, 首先描述用于图像和谐化任务的基于整体–局部视觉状态空间的图像和谐化模型的总体流程和模型结构, 然后介绍本文提出的用于建立前景与背景风格关系的整体–局部视觉状态空间的具体细节.

3.1　总体流程与模型结构

本文提出的基于整体–局部视觉状态空间模型 GLIHamba 的总体流程和结构如图 1(a)所示. 参考在视觉领域取得优异表现的 Swin–Transformer[29], 以及 Swin–UMamba[24] 等模型结构, 本文 GLIHamba 模型采用了 U 型层级网络模型, 包括编码器(Encoder)、解码器(Decoder)和它们之间的跳跃连接. 具体来说, 给定一个合成图像 $\boldsymbol{P}\in\boldsymbol{R}^{3\times H\times W}$, 其长宽分别为 H 和 W, 首先用一个带有 ReLU 的 3×3 卷积的块编码层(Patch embedding, PE), 提取图像特征 $\boldsymbol{X}_0\in\boldsymbol{R}^{N\times H\times W}$, 其中 N、H、W 分别表示通道数、长度和宽度. 进一步将特征图进行 M 次编码操作, 每次操作包括匹配特征形状(Patch merging)块, 以及基于选择性状态空间的视觉状态空间块, 用于建立合成图像特征的语义和风格的全局上下文关系. VSS 块的结构如图 1(b)所示, 参考 Swin–UMamba 的处理方式, 本文中 VSS 采用 2D 选择性扫描 (2D-selective-scan, SS2D) 实现 2D 图像数据的扫描方式. 给定第 m 阶段的输入特征 $\boldsymbol{X}_m$(其中, $m\in\{0,\cdots M-1\}$), VSS 过程可以表示为:

$$\begin{aligned}&\hat{\boldsymbol{X}}_m=\mathrm{LayerNorm}(\boldsymbol{X}_m),\\&\hat{\boldsymbol{X}}_m=\mathrm{Linear}(\hat{\boldsymbol{X}}_m)\otimes\mathrm{LayerNorm}\\&\qquad(\mathrm{SS2D}(\mathrm{DWConv}(\mathrm{Linear}(\hat{\boldsymbol{X}}_m)))),\\&\boldsymbol{X}_{m+1}=\boldsymbol{X}_m+\mathrm{Linear}(\hat{\boldsymbol{X}}_m)\end{aligned}\tag{6}$$

其中, $\hat{\boldsymbol{X}}_m$ 和 $\boldsymbol{X}_{m+1}$ 分别表示第 m 阶段的中间特征以及输出的特征(下一阶段的输入特征), $\otimes$ 表示元素



(a) Overall structure (b) VSS block (c) MLG-VSS block

图 1 GLIHamba 模型结构图. (a) 模型总体流程图; (b) 视觉状态空间模块结构图; (c) 掩膜化的局部–全局视觉状态空间模块结构图

Fig.1 Overview of the GLIHamba structure: (a) overall pipeline of the model; (b)VSS block; (c) MLG–VSS block

相乘. LayerNorm、Linear 和 DWConv 分别表示层规范化函数、线性函数和深度可分离卷积.

SS2D 沿着四个方向展开图像块, 形成四个不同的序列. 随后, 每个特征序列将通过 SSM 进行处理. 最后, 对输出的特征进行合并, 形成完整的二维特征. 给定输入特征 z, SS2D 的输出特征 $\bar{z}$ 可表示为:

$$\begin{aligned}\bar{z}_v &= FS6(\text{Expand}(z, v)),\\ \bar{z} &= \text{Merge}(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4)\end{aligned} \tag{7}$$

其中, $v \in \{1,2,3,4\}$表示四个扫描方向, Expand(·)和 Merge(·)分别表示扫描展开操作和扫描合并操作, $\boldsymbol{F}_{S6}$即表示选择性状态空间模型函数. 选择性状态空间模型是 VSS 块的核心操作, 通过状态参数, 它能使特征序列中的每个元素与先前扫描的样本交互, 从而有效地建立合成图像中背景与前景的语义和风格特征之间的关系.

在图像特征解码阶段, 和编码阶段相同, 也包括 M 个解码阶段. 每个阶段包括块扩展(Patch expanding)块和用于根据背景与前景的语义关系实现风格和谐化的掩码化的整体-局部视觉状态空间(Masked local-global visual state space, MLG–VSS)块, 一层卷积融合(1×1Conv)块和基于掩码的残差连接(Masked Res)层. MLG–VSS 利用整体–局部特征提取器, 打破了 VSS 在视觉特征中无法关注局部信息的限制, 有效建立合成图像特征的局部和整体的关联关系. 和编码阶段的 VSS 块类似, 对于第 m 阶段的输入特征 $\boldsymbol{X}'_m$, MLG–VSS 过程可以表示为:

$$\begin{aligned}\hat{\boldsymbol{X}}'_m &= \text{LayerNorm}(\boldsymbol{X}'_m),\\ \hat{\boldsymbol{X}}'_m &= \text{SiLU}(\text{Linear}(\hat{\boldsymbol{X}}'_m)) \otimes \text{LayerNorm}\\ &\quad (\text{SS2D}(\text{MLG}(\text{Linear}(\hat{\text{X}}'_{\text{m}})))),\\ \boldsymbol{X}'_{m+1} &= \boldsymbol{X}'_m + \text{Linear}(\hat{\boldsymbol{X}}'_m)\end{aligned} \tag{8}$$

其中, SiLU 为激活函数, MLG 表示掩码化的整体-局部特征提取器, 用于整合合成图像的局部和整体的信息, 这个模块将在下面一节详细介绍.

在解码的最后阶段, 特征被扩展到 2D 后经过一个 1×1 卷积层, 进而与输入的合成图像合并, 组成和谐化的图像结果. 其中合并操作描述为:

$$\tilde{\boldsymbol{P}} = \boldsymbol{X}'_M \otimes \text{Mask} + \boldsymbol{P} \otimes (1 - \text{Mask}) \tag{9}$$

其中, Mask为图像掩码, $\tilde{\boldsymbol{P}}$为合成图像 $\boldsymbol{P}$ 和谐化的图像结果. 在编码与解码之间, 采用了和 Swin–UMamba 一样的跳跃连接, 提升模型的稳定性.

3.2 掩码化的整体-局部视觉状态空间

为了改进 Mamba 模型在图像和谐化任务中, 保持局部一致性的同时建立整体相关性, 本文将图像特征抽象为局部特征序列和全局特征序列. 其中, 局部特征序列由带有掩码的局部特征序列提取器(Masked local feature sequence extractor, Masked LFSE)模块建立, 如图 2(a) 所示. 具体来说, 为了分别学习前景和背景信息, 将掩码信息和图像特征

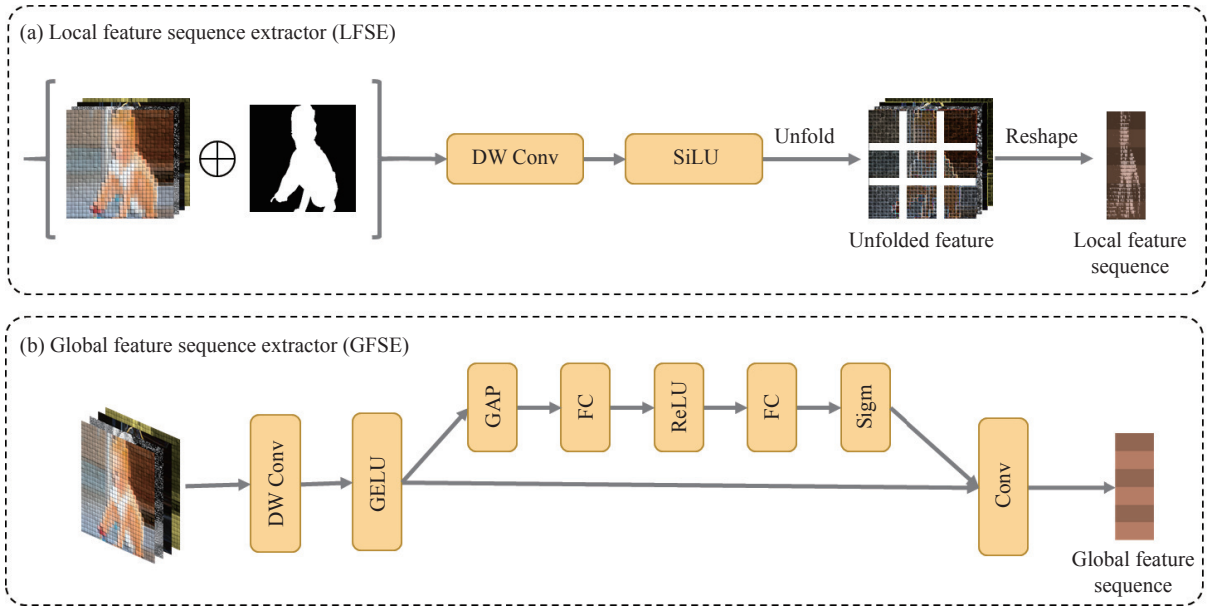


图 2　局部–全局特征提取器结构图. (a)本地特征序列提取器; (b)全局特征序列提取器

Fig.2　Local-global feature sequence extractor: (a) local feature sequence extractor; (b) global feature sequence extractor

按照通道合并, 然后利用深度可分离卷积(DW Conv)将输入图像特征沿着通道方向按比例因子ε压缩, 将压缩后的特征依次使用 SiLU 激活函数和3×3卷积操作. 最后, 将压缩后的特征按照空间维度展开, 构建局部特征序列(Local feature sequence), 该序列保留了局部特征的空间关系, 从而建立背景与前景语义和风格的局部一致性.

为了生成全局令牌, 进一步提出了全局特征序列提取器 (Global feature sequence extractor, GFSE) 模块, 如图 2(b)所示. 具体而言, 给定输入特征, 使用深度可分离卷积 (DW Conv) 对其进行空间压缩, 在通过 GELU 后, 使用压缩激励(Squeeze and excitation, SE)操作以建立全局相关性, 最后使用1×1卷积输出全局令牌, 并保持器维度与局部令牌维度一致, 从而在随后的步骤中将 GFSE 的输出与局部特征序列合并.

最后, 将提取的整体–局部特征提取器整合起来, 以创建具有整体–局部上下文感知的 GLIHamba 模型. 该组合利用了全局特征提取器的全局视野以及局部特征提取器的局部依赖关系, 结合 Mamba 模型的全局关系构建能力, 能够实现图像和谐化的局部内容和整体风格的一致性. 同时, 为了学习背景与前景的不同风格信息, 将掩码信息与图像特征信息合并后提供给特征提取器, 从而实现背景信息到前景的迁移. 具体来说, 对于带有掩码信息的给定的图像特征$z \in \boldsymbol{R}^{H\times W\times(N_C+1)}$(其中, N_C+1表示图像特征通道数基础上增加 1 层掩码通道数), 该过程可以用如下公式表示:

$$\begin{aligned} z^{\mathrm{L}} &= \mathrm{LFSE}(z), \\ z^{\mathrm{G}} &= \mathrm{GFSE}(z), \\ z^{\mathrm{GL}} &= \mathrm{concat}(z^{\mathrm{L}}, z^{\mathrm{G}}) \end{aligned} \tag{10}$$

其中, z^{L}、z^{G}分别表示局部特征信息和全局特征信息, z^{GL}则表示含有全局特征信息和局部特征信息, 进而结合 SS2D 操作(如图 1(c)所示), 实现整体–局部上下文的整合.

4　实验

4.1　数据集

为了评估本文方法在图像和谐化方面的性能, 本文在广泛使用的 iHarmony4[14] 公开数据集上进行了实验. 该数据集由 4 个子数据集组成, 包括 HCOCO、HAdobe5k、HFlickr 和 Hday2night, 共包含 73147 组前景掩码图像、合成图像和对应的真实图像. 在这项工作中, 本文遵循与 DoveNet 相同数量的训练集测试集划分, 65742 组用于训练, 7404 组用于测试. 同时, 为了验证本文方法在真实场景中的有效性, 本文使用 100 张真实场景合成图片数据集进行测试. 该数据集由 CDTNet[35] 公开.

4.2　实验设置

本文以 Adam 优化器来训练模型, 参数为 β_1=0.9, β_2=0.999, 总共 200 个 epoch. 设置初始学习率为 10^{-3}, 衰减系数 γ=0.5. 本文的训练样本通过水平翻转和随机大小裁剪来增强模型的泛化能力, 并将输入图像尺寸调整到 256×256. 此外, 本文的实验环境: 并行计算平台模型 CUDA 版本为 11.2, 深

度学习开发环境 PyTorch 版本为 3.8.2, 在 4 张 Nvidia GTX 3090Ti GPU 上进行训练.

4.3 对比方法与评价指标

本文对比方法包括: DoveNet[14], RainNet[17], D-HT[18], HDNet[15], GKNet[36]. 其中, D-HT 是基于 Transformer 的方法, 其余模型均基于 CNN 方法. 评估指标采用均方误差 (Mean square error, MSE)、前景均方误差 (fore-ground MSE, fMSE) 和峰值信噪比 (Peak-signal-to-noise-ratio, PSNR) 来评估本文模型性能. MSE 本质上测量了整个数据集中所有像素的平均误差, 而 fMSE 仅计算前景区域 MSE, 更加适合在背景区域像素不变的情况下对图像和谐化的结果真实性进行衡量. 其中, MSE 和 fMSE 值越小, PSNR 值越大, 代表模型性能越好.

4.4 对比实验结果

4.4.1 定量分析

表 1 展示了本文方法与图像和谐化领域前沿方法的定量对比结果. 表中的加粗数字表示最优结果. 可以看到, 本文的方法在 HCOCO、HAdobe5k、HFlickr 以及数据集总体结果上超过了表中其他的方法结果, 其中在整体数据集中实现了 PSNR 指标 0.2 dB, MSE 指标 1.32 数据提升.

表 1 iHarmony4 数据集对比结果

Table 1 Quantitative comparison across four sub-datasets of iHarmony4

Models	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR
Composite	69.37	33.94	345.54	28.16	264.35	28.32	109.65	34.01	172.47	31.63
DoveNet	36.72	35.83	52.32	34.34	133.14	30.21	54.05	35.18	52.36	34.75
RainNet	29.52	37.08	43.35	36.22	110.59	31.64	57.4	34.83	40.29	36.12
D-HT	16.89	38.76	38.53	36.88	74.51	33.13	53.01	37.10	30.30	37.55
HDNet	15.59	39.49	22.67	38.56	63.85	33.96	**35.92**	38.11	23.42	38.58
GKNet	12.95	40.32	17.84	39.97	57.58	34.45	42.76	**38.47**	19.90	39.53
Ours	**12.05**	**40.35**	**16.53**	**40.23**	**56.29**	**34.56**	43.27	38.05	**17.58**	**39.76**

Note: **bold** and underline indicate the optimal and suboptimal results, respectively.

同时, 表 2 展示了不同比例(0 ~ 5%, 5% ~ 15%, 15% ~ 100%)的前景占比的指标, 可以看到本文的方法均达到了最优结果. 在三个比例组以及整体平均值中, MSE 分别提升了 0.61、1.43、6.41、5.84, fMSE 分别提升了 21.9、50.08、35.47、42.28. 说明本文的方法在 iharmony4 数据集上具有领先的优异表现, 并且在前景占比较大的时候, 具有更优异的表现, 说明本文的方法具有更好的鲁棒性.

表 2 各方法在 iHarmon4 数据集中不同前景占比的 MSE 和 fMSE 指标

Table 2 Quantitative comparison of foreground ratios of iHarmony4

Models	Ratio range 0–5%		Ratio range 5%–15%		Ratio range 15%–100%		Average	
	MSE	fMSE	MSE	fMSE	MSE	fMSE	MSE	fMSE
Composite	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.3
DoveNet[5]	14.03	591.88	44.9	504.42	152.07	505.82	52.36	549.96
RainNet[37]	11.66	550.38	32.05	378.69	117.41	389.8	40.29	469.6
HDNet[7]	5.95	230.75	20.32	265.31	68.95	318.15	23.42	258.8
Ours	**5.34**	**208.85**	**18.89**	**215.23**	**62.54**	**272.68**	**17.58**	**216.52**

Note: **bold** and underline indicate the optimal and suboptimal results, respectively.

4.4.2 定性分析

为了进一步说明本文提出方法的有效性, 分别展示了 iHarmony4 数据集和真实场景数据集的可视化结果, 如图 3 和图 4 所示. 其中, 在图 3 中, 前两列分别为合成图、相应掩码, 后面的 5 列分别展示了 DoveNet、RAINNet、HDNet、本文提出的方法, 以及真实图片. 可以看到, 本文的方法在整体外观具有更好的一致性. 在图 4 中, 前两列同样为

(a) Composite　(b) Masks　(c) DoveNet　(d) RainNet　(e) HDNet　(f) Ours　(g) Real

图 3　不同方法在 iHarmony4 数据集上的结果

Fig.3　Qualitative comparison of iHarmony4

(a) Composite　(b) Masks　(c) RainNet　(d) HDNet　(e) Ours

图 4　不同方法在真实场景数据集上的结果对比

Fig.4　Visualization of real-world images

合成图、相应掩码, 后面的 3 列分别为 RAINNet、HDNet 以及本文方法的结果. 如第二行的汽车、第三行的人物, 本文的方法明显具有更好的外观一致性, 说明本文提出的 GLIHamba 模型在图像和谐化任务应用于真实场景图片, 实现了领先的结果.

4.4.3　计算复杂性分析

为了验证所提出模型的优越性, 使用三个指标进行计算复杂度评估: 浮点运算次数 (FLOPs)、

模型的参数和内存占用. FLOPs 是衡量模型计算复杂度的重要指标,而模型参数和内存占用则分别评估了网络的规模和内存需求.

由表 3 可知, 相比于 RainNet 和 HDNet, 本文所提出的具有更多的模型参数, 这是由于引入了整体及局部特征提取器造成的. 但是由于 HDNet 引入了动态卷积, 因此计算量较高. 相比于基于 Transformer 的 D-HT 模型, 本文的参数量和计算复杂度均有更好地表现.

表 3 计算复杂性分析实验

Table 3 Complexity analysis experiments

Models	GFLOPs	Parameters/10^6	Inference time/ms
RainNet	3.80	54.75	12.06
HDNet	48.05	10.41	15.08
D-HT	120.10	89.42	20.45
Ours	**14.86**	**64.20**	**19.67**

4.5 消融实验结果

为了验证本文提出的 GLIHamba 模型的有效性, 本文在 iharmony4 数据集上构建了消融实验. 结果如表 4 所示. 其中第一行为数据集中合成图像与真实图像的差值; 基线模型(Baseline)采用使用卷积的 Unet 模型. 在此基础上, +VSS Block 表示将 mamba 的 VSS 块替换 Unet 中的卷积块; +MLG–VSS 则是在解码阶段使用了 MLG–VSS 块的完整的模型. 由表 4 可知, 当基线模型添加了 VSS Block 后, 整个数据集的 MSE 值降低了 12.2, PSNR 值提升了 2.07, 并且各个子数据集的 MSE 值均有降低, PSNR 值均有提升, 表明 VSS Block 带来的全局关系构建能力, 能够一定程度的提升和谐化的效果. 进一步, 添加了 MLG–VSS 后, 相比于基线模型, 整个数据集的 MSE 值降低了 34.78, PSNR 值提升了 5.01, 达到了 39.76. 表明本文提出的局部–整体上下文方法能够明显提升图像和谐化任务的效果.

表 4 消融实验结果

Table 4 Results on ablation study

Methods	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR
Composite	69.37	33.94	345.54	28.16	264.35	28.32	109.65	34.01	172.47	31.63
Baseline	36.72	35.83	52.32	34.34	133.14	30.21	54.05	35.18	52.36	34.75
+VSS	29.52	37.08	43.35	36.22	110.59	31.64	57.4	34.83	40.16	36.82
+MLG–VSS	**12.05**	**40.35**	**16.53**	**40.23**	**56.29**	**34.56**	43.27	38.05	**17.58**	**39.76**

Note: **bold** indicates the optimal value.

5 结论

在本文中, 本文设计了一种新的基于选择性状态空间的图像和谐化模型. 本文通过将全局特征序列和局部特征序列引入到 Mamba 模型, 增强 SSM 模块提取合成图像中整体和局部特征的能力, 建立前景与背景的语义特征与风格特征关联关系. 通过实验验证了本文方法有效性. 在未来工作中, 将进一步提升图像和谐化的真实感, 加快运行速度, 以便生成更高质量的和谐化效果, 提升在自动动化图像处理领域产生更多的贡献.

参 考 文 献

[1] Wang Z H. *Research and Application of Multi-modal Image Harmonization* [Dissertation]. Beijing: Beijing University of Posts and Telecommunications, 2024
(汪正徽. 多模态图像和谐化的研究与应用[学位论文]. 北京: 北京邮电大学, 2024)

[2] Hays S P, Remedios S W, Zuo L R, et al. Beyond MR image harmonization: Resolution matters too // *Simulation and Synthesis in Medical Imaging*. Marrakesh, 2024: 34

[3] Bao Z Y, Long C J, Fu G, et al. Deep image-based illumination harmonization // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 18521

[4] Li R B, Guo J C, Zhou Q H, et al. FreePIH: Training-free painterly image harmonization with diffusion model // *Proceedings of the* 32*nd ACM International Conference on Multimedia*. Melbourne, 2024: 7464

[5] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J/OL]. *arXiv preprint* (2014-06-10) [2024-09-12]. https://arxiv.org/abs/1406.2661v1

[6] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 12868

[7] Ho J, Jain A, Abbeel P, et al. Denoising diffusion probabilistic models // *Proceedings of the* 34*th International Conference on*

Neural Information Processing Systems. Vancouver, 2020: 6840

[8] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 10674

[9] Hu F L, Chen A A, Horng H, et al. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, 2023, 274: 120125

[10] Niu L, Cong W Y, Liu L, et al. Making images real again: A comprehensive survey on deep image composition [J/OL]. *arXiv preprint* (2021-06-28) [2024-09-12]. https://arxiv.org/abs/2106.14490v6

[11] Pérez P, Gangnet M, Blake A. Poisson image editing // *SIGGRAPH*03: *Special Interest Group on Computer Graphics and Interactive Techniques*. San Diego, 2003: 313

[12] Sunkavalli K, Johnson M K, Matusik W, et al. Multi-scale image harmonization. *ACM Trans Graph*, 2010, 29(4): 1

[13] Tsai Y H, Shen X H, Lin Z, et al. Deep image harmonization // 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu, 2017: 2799

[14] Cong W Y, Zhang J F, Niu L, et al. DoveNet: Deep image harmonization *via* domain verification // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 8391

[15] Chen H X, Gu Z X, Li Y H, et al. Hierarchical dynamic image harmonization // *Proceedings of the* 31*st ACM International Conference on Multimedia*. Ottawa, 2023: 1422

[16] Hang Y C, Xia B, Yang W M, et al. Scs-Co: Self-consistent style contrastive learning for image harmonization // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 19678

[17] Ling J, Xue H, Song L, et al. Region-aware adaptive instance normalization for image harmonization // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 9357

[18] Guo Z H, Guo D S, Zheng H Y, et al. Image harmonization with transformer // 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2021: 14850

[19] Guo Z H, Gu Z R, Zheng B, et al. Transformer for image harmonization and beyond. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(11): 12960

[20] Gu A, Gupta A, Goel K, et al. On the parameterization and initialization of diagonal state space models [J/OL]. *arXiv preprint* (2022-06-23) [2024-09-12]. https://arxiv.org/abs/2206.11893

[21] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces [J/OL]. *arXiv preprint* (2023-12-01) [2024-09-12]. https://arxiv.org/abs/2312.00752

[22] Liu Y, Tian Y, Zhao Y, et al. VMamba: Visual state space model [J/OL]. *arXiv preprint* (2024-01-18) [2024-09-12]. https://arxiv.org/abs/2401.10166

[23] Zhu L, Liao B, Zhang Q, et al. Vision Mamba: Efficient visual representation learning with bidirectional state space model [J/OL]. *arXiv preprint* (2024-01-17) [2024-09-12]. https://arxiv.org/abs/2401.09417

[24] Liu J, Yang H, Zhou H Y, et al. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining [J/OL]. *arXiv preprint* (2024-02-05) [2024-09-12]. https://arxiv.org/abs/2402.03302

[25] Zheng Z R, Wu C. U-shaped Vision Mamba for single image dehazing [J/OL]. *arXiv preprint* (2024-02-06) [2024-09-12]. https://arxiv.org/abs/2402.04139

[26] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J/OL]. *arXiv preprint* (2020-10-22) [2024-09-12]. https://arxiv.org/abs/2010.11929

[27] Guo Z H, Zheng H Y, Jiang Y F, et al. Intrinsic image harmonization // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 16362

[28] Waswani A, Shazeer N, Parmar N, et al. Attention is all you need [J/OL]. *arXiv preprint* (2017-06-12) [2024-09-12]. https://arxiv.org/abs/1706.03762

[29] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows // 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2021: 9992

[30] Wang Z, Liu Z S. StyleMamba: State space model for efficient text-driven image style transfer [J/OL]. *arXiv preprint* (2024-05-08) [2024-09-12]. https://arxiv.org/abs/2405.05027

[31] Peng S R, Zhu X Y, Deng H Y, et al. FusionMamba: efficient remote sensing image fusion with state space model [J/OL]. *arXiv preprint* (2024-05-10) [2024-09-12]. http://arxiv.org/abs/2404.07932

[32] Xie X, Cui Y, Ieong C I, et al. FusionMamba: Dynamic feature enhancement for multimodal image fusion with Mamba [J/OL]. *arXiv preprint* (2024-07-23) [2024-09-12]. https://arxiv.org/abs/2407.16126

[33] Ma J, Li F F, Wang B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation[J/OL]. *arXiv preprint* (2024-01-09) [2024-09-12]. https://arxiv.org/abs/2401.04722

[34] Chen S, Atapour-Abarghouei A, Zhang H Z, et al. MxT: Mamba x transformer for image inpainting [J/OL]. *arXiv preprint* (2024-07-23) [2024-09-12]. https://arxiv.org/abs/2407.16126

[35] Cong W Y, Tao X H, Niu L, et al. High-resolution image harmonization *via* collaborative dual transformations // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 18449

[36] Shen X T, Zhang J N, Chen J, et al. Learning global-aware kernel for image harmonization // 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Paris, 2023: 7501