基于内容语义感知多模态融合的图像增强方法

祝汉城[1,2],刘新宇[1,2],姚 睿[1,2*],邵志文[1,2],周 勇[1,2],李雷达[3]
(1. 中国矿业大学计算机科学与技术学院/人工智能学院,江苏徐州 221116;2. 矿山数字化教育部工程研究中心,江苏徐州 221116;
3. 西安电子科技大学人工智能学院,陕西西安 710126)

摘 要: 在图像增强方法中,基于曲线映射的修饰策略因其能够很好地保留图像的原始内容信息而成为研究的热点. 现有的基于曲线映射方法通常只关注修饰前后图像色彩空间的映射关系,而忽略了图像内容对修饰结果的影响,导致具有相似色彩的不同图像内容修饰得不够精细和自然. 针对上述问题,本文提出了一种基于内容语义感知多模态融合的图像增强方法,旨在通过引入描述图像内容语义感知信息的文本特征作为图像特征的补充,将图像和文本两个模态的特征进行融合得到内容语义感知的多模态特征,从而实现对图像不同内容的精细化修饰. 首先,本文利用多模态大语言模型生成描述图像内容的文本信息,并将文本信息对图像的内容进行多模态提示学习,该方法能够使模型学习在内容文本信息的提示下对图像进行辅助增强;随后,提出了一种注意力机制将文本特征与图像特征进行充分交互融合生成多模态特征;最后,利用多模态特征建立修饰图像的曲线映射关系,从而可以有效地根据图像的内容进行针对性的修饰与增强. 实验结果表明,本文提出方法在多个公开的基准数据集上取得了最优的性能表现,充分证明了融入内容语义感知信息在图像修饰任务上的有效性和优越性.

关键词: 图像增强;文本生成;内容感知;多模态融合;曲线映射
基金项目: 国家自然科学基金(No.62101555,No.62172417,No.62472424,No.62272461,No.62106268)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)07-2252-14
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241088

Image Enhancement via Content Semantic-Aware Multimodal Fusion

ZHU Han-cheng[1,2], LIU Xin-yu[1,2], YAO Rui[1,2*], SHAO Zhi-wen[1,2], ZHOU Yong[1,2], LI Lei-da[3]
(1. *School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*; 2. *Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, Jiangsu* 221116, *China*;
3. *School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi* 710126, *China*)

Abstract: Among image enhancement techniques, curve mapping-based retouching strategies have attracted significant research interest due to their ability to effectively retain the original content information of images. However, current curve-mapping methods primarily focus on the changes in color space before and after enhancement, often neglecting the influence of image content on the enhancement results. This limitation leads to suboptimal adjustments for images with similar colors but different content, resulting in less refined and natural enhancements. To address this issue, this paper proposes an image enhancement method based on content-aware multimodal fusion, which supplements image features by incorporating text features that describe the semantic perception of image content. By fusing features from both image and text modalities, the proposed approach captures multimodal content-aware semantics, enabling fine-grained adjustments tailored to different image content. Firstly, a multimodal large language model is employed to extract textual descriptions of image content, which are then used for multimodal prompt learning to guide the understanding of the image content. This method enables the model to leverage content-based text prompts for auxiliary image enhancement. Then, an attention mechanism is then applied to effectively integrate and fuse the textual and image features into a unified multimodal representation. Finally, this representation is used to construct a curve-mapping function, enabling content-specific image adjustments and enhancements. Experimental results on multiple public benchmark datasets demonstrate that the proposed method achieves state-of-the-art performance, highlighting the effectiveness and advantages of incorporating content-aware semantic information into image enhancement tasks.

收稿日期:2024-12-03;录用日期:2025-04-17;责任编辑:覃怀银
*通讯作者:姚睿

Key words: image enhancement; text generation; content-aware; multimodal fusion; curve mapping
Foundation Item(s): National Natural Science Foundation of China (No.62101555, No.62172417, No.62472424, No.62272461, No.62106268)

1 引言

随着智能手机和数码相机等设备的普及,摄影已成为日常生活中不可或缺的一部分. 然而,由于拍摄环境的复杂性、用户摄影水平的参差不齐以及设备本身的限制,导致拍摄出的图像视觉效果欠缺,难以满足用户对图像美感和质量的要求. 在此背景下,图像增强技术应运而生,旨在通过对图像进行后期处理,改善图像的视觉效果,提升图像的整体美感. 传统的图像增强方法通常依赖于人工设计的规则,如直方图均衡化、伽玛校正和滤波器等,但这些方法难以适应复杂的图像内容和多样化的用户需求. 随着深度学习技术的迅猛发展,图像增强领域迎来了新的突破[1]. 与传统的增强方法不同,深度学习能够通过大规模的数据训练,自动学习图像特征并生成增强结果.

目前,主流的深度学习图像增强方法主要分为两大类. 第一类是基于像素重建的增强方法[2~7],其通过深度神经网络直接预测增强后的图像. 尽管该类方法具有较强的图像生成能力,但其局限性也较为明显. 由于需要对大量像素进行逐一重建,训练和推理过程的计算开销较高;此外,其生成的图像中常常会出现伪影[8],影响最终的视觉效果. 第二类方法则是基于曲线映射的修饰方法[8~18]. 该类方法将颜色变换从模型中分离出来,通过估计一组中间参数,用于像素值的映射. 这类方法具有以下优势:首先,通过像素映射的方式进行增强,计算效率高,并且能够很好地保留图像原始细节和相邻像素间的内在关系,避免了图像重建方法中常见的伪影问题;其次,曲线映射过程简单明了,具有较强的可解释性,这使得该类方法在实际应用中更具吸引力. 因此,本文主要研究基于曲线映射的图像修饰方法.

当前基于曲线映射的修饰方法主要侧重于对图像色彩空间的变换[17,18]和映射方式的改进[9~11,19],忽略了图像内容对增强结果的影响. 因此,在色彩相似的情况下,不同图像内容的增强效果差异不显著,无法实现精细的内容感知修饰. 例如,CURL[18]和 NamedCurves[17]等方法主要专注于图像色彩的调整,但未能结合图像内容的语义信息,难以针对不同的图像内容进行差异化处理. Zero-DCE[9]、FlexiCurve[10]、3D LUT[11]和 AdaInt[19]等方法专注于改进像素值的映射方式,但仍然主要依赖于图像的低层次特征,忽略了图像中各个对象在语义上的差异,导致在处理复杂场景时,修饰效果不够自然. SpliNet[14]和 StarEnhancer[15]主要专注于对特定风格的建模学习,同样缺乏对图像内容的深入理解,难以实现真正的内容自适应增强与修饰.

上述方法的主要问题在于过度关注色彩信息和映射方式,而忽视了图像的内容语义信息. 在本文的研究中,内容语义指的是图像中视觉对象的实际含义. 与传统语义方法(如语义分割)主要依赖预定义的类别标签不同,本文的内容语义更侧重于对象的具体特征、属性及其在场景中的关系.

在图像修饰过程中,专业摄影师通常会遵循“先理解,后调整”的原则,即先分析图像中不同对象的语义信息,再根据其特性进行针对性调整. 如图 1 所示,尽管云层与地面在原始图像中的颜色相近,摄影师在对图像进行修饰时会基于不同内容的语义对其进行差异化处理,最终修饰成不同颜色. 受此启发,本文提出了一种基于内容语义感知多模态融合的图像增强方法. 通过引入内容语义信息,模拟人们修图过程中对图像不同内容的差异化处理,实现更精细化的图像增强效果.

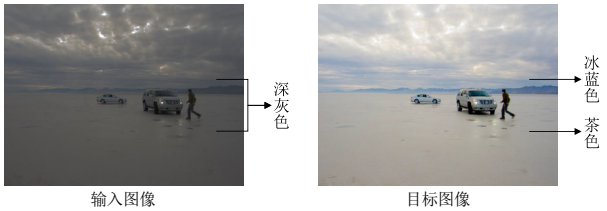


图 1　输入图像和目标图像对比图

为了实现上述目的,需要首先获取对图像内容的准确描述. 本文设计了特定的提示词,以引导多模态大语言模型生成描述图像内容的文本信息. 为了将这些文本描述转化为可用于图像增强的特征,本文采用了基于 CLIP(Contrastive Language-Image Pre-training)模型[20]的方法. CLIP 模型的优势在于其通过大规模的图文对数据预训练,能够同时理解图像和文本的语义关联. 这使得 CLIP 能够有效地捕捉图像中的内容信息,将其与相应的文本描述进行匹配,从而为后续的多模态融合提供高质量的文本和图像特征. 为了实现文本特征与图像特征的有效融合,本文设计了一个特征融合模块. 该模块采用基于注意力机制的多模态特征交互网络,通过将图像特征和文本特征相互结合,生成包含语义信息的多模态融合特征. 最终,将多模态融合特征转换为映射曲线,并对输入图像进行映射,以生成增强后的图像.

综上所述,本文主要贡献如下:

(1)本文提出了一种基于内容语义感知多模态融合的图像增强方法,通过在曲线映射修饰中融入图像内容语义感知信息,有效地解决了现有曲线映射修饰方法仅考虑图像色彩信息的问题.

(2)本文提出的方法可以为图像生成描述内容语义的文本信息,并通过多模态提示和注意力机制将文本信息对图像信息进行引导学习,还可以通过生成文本信息实现基于内容语义感知多模态特征融合的图像增强方法.

(3)本文提出的方法在多个公开的基准数据集中取得了最佳的实验结果,验证了本文方法的有效性和优越性.

2 相关工作

目前,图像增强方法可根据输入信息模式的不同分为两大类:基于单模态的图像增强方法与基于多模态的图像增强方法. 单模态方法只依赖于图像自身信息进行增强;多模态方法则结合了其他模态的信息,如文本提示信息等.

2.1 基于单模态的图像增强方法

基于单模态的增强方法可以主要分为两大类:基于像素重建的增强方法和基于曲线映射的修饰方法.

2.1.1 基于像素重建的增强方法

基于像素重建的增强方法直接预测增强后的图像,利用深度神经网络对每个像素进行处理. 该类方法主要分为两类:全卷积网络(Fully Convolutional Network,FCN)[21]和视觉Transformer(Vision Transformer,ViT)[5].

在全卷积网络方面,Chen等人[3]率先将FCN用于图像增强任务,并证明了该方法在不同数据集和分辨率上的泛化能力. 随后,Ronneberger等人[2]提出的U-Net,作为一种经典的FCN结构,也被广泛应用于图像增强领域. Chen等人[4]提出的DPE采用生成对抗网络(Generative Adversarial Network,GAN)框架[22],以U-Net作为生成器,提出了一种无配对学习的方法,通过增强U-Net的全局特征处理能力,提高了图像增强的效果. 在视觉Transformer[5]方面,Zamir等人[6]提出的Restormer将ViT架构应用于图像增强,设计了一种高效的Transformer模型,使其能够捕捉长距离的像素交互. Wang等人[7]提出的LLFormer则针对低光照图像增强任务,设计了基于轴的多头自注意力和跨层注意力融合模块,降低了Transformer模型的线性复杂度,使其能够高效地处理高分辨率图像.

尽管基于像素重建的方法取得了不错的结果,但这类方法也存在一些固有的缺点. 首先,由于需要对每个像素进行预测,训练过程需要大量的计算资源和时间成本,模型参数量大. 其次,这类方法的推理速度较慢,尤其是在处理高分辨率图像时,难以满足实时增强的需求. 最后,伪影问题也常见于这类方法的生成图像中. 模型在逐像素生成图像时,可能会忽略全局上下文或局部像素之间的细微关联,导致增强后的图像出现不自然的纹理或结构,影响了增强后图像的视觉效果. 因此,基于曲线映射的图像修饰方法可以很好地保留原始图像的内容信息,成为图像增强领域的研究热点.

2.1.2 基于曲线映射的修饰方法

基于曲线映射的修饰方法通常估计一组中间参数,例如曲线或查找表,用于像素值的映射,从而避免了直接进行像素重建带来的计算负担. 其中,曲线映射方法因其高效和可解释性而受到广泛关注.

Guo等人[9]提出的Zero-DCE是首个将曲线映射应用于低光照图像增强的方法. 该方法将低光照图像增强视为曲线估计问题,通过训练轻量级深度网络DCE-Net来估计高阶曲线,实现图像动态范围的调整. 随后,Song等人[23]提出的MCT进一步优化了曲线映射的计算效率. 该方法不仅预测输入像素的翻译结果,还能预测邻近像素的翻译结果. 这使得MCT可以通过处理降采样图像来完成高分辨率图像的增强,大幅降低了计算成本. Li等人[10]提出的FlexiCurve则设计了一种分段映射曲线,并考虑了非线性调整和可微性,以更好地适应真实图像的复杂特性. 另一类曲线映射方法则关注于色彩信息的调整. Moran等人[18]提出了CURL,该方法在多种颜色空间中估计分段线性曲线,以实现对图像色彩的精细控制. Serrano-Lozano等人[17]提出的NamedCurves将曲线映射与颜色命名[24]结合,将图像分解为一小组命名颜色,并通过调色曲线对每个特定命名颜色进行全局调整,从而实现更直观的图像增强. 此外,曲线映射还被用于模拟专家的修图风格. Bianco等人[14]提出的SpliNet通过训练网络输出曲线的控制点,并将用户的配置文件注入网络,从而做到了使用单个神经网络建模多种修饰风格. Song等人[15]提出的StarEnhancer进一步扩展了这一思路,设计了额外的网络生成风格向量,并引入风格迁移模块[25],实现了单一模型在多种风格之间的灵活映射.

查找表(Look-Up Table,LUT)是另一种广泛应用于图像增强的技术,通常用于相机成像流程和图像编辑工具中[26]. 近年来,随着深度学习的发展,查找表与深度学习的结合进一步提升了图像增强的效果. Zeng等人[11]提出的3D LUT是首个将查找表与深度学习相结合的工作,通过成对或无配对数据集学习图像自适应的三维查找表,实现了图像增强过程的自动化. 为了进一步提升3D LUT的灵活性,Yang等人[19]提出的AdaInt引入了自适应间隔学习机制,通过在三维色彩空间中进

行非均匀采样,实现了灵活的采样点分布. 在高度非线性变换的颜色区域内,进行密集采样;在接近线性变换的区域内,采用稀疏采样,从而显著提升查找表的颜色映射能力. Yang等人[16]还提出SpeLUT,将单一的颜色变换分解为两个子变换,分别使用串联的1D LUT和3D LUT来实现变换,既保留了1D LUT的高效性,又提升了3D LUT的表现能力. 此外,Liu等人[13]提出的4D LUT将查找表架构扩展到四维空间,通过增加上下文信息来对不同内容进行更精细的颜色调整.

尽管上述方法已经取得了显著进展,但它们仍然主要关注图像的色彩信息的特征表示或像素数值之间的映射关系,忽略了图像内容对修饰效果的影响. 这导致在处理具有复杂语义信息的图像时,修饰结果可能不够精细和自然. 因此,在进行图像修饰时也有必要提取图像的内容信息,实现同时根据图像内容与色彩进行自适应的增强.

2.2　基于多模态的图像增强方法

目前,多模态图像处理技术已在计算机视觉领域得到广泛应用和发展,通过充分融合不同来源的数据来有效提升算法模型的表现[27,28]. 基于此,本文主要研究图像和文本两个模态的数据融合,利用Radford等人[20]提出的CLIP模型对图像进行增强. CLIP通过在大规模图文对数据上进行预训练,具备了同时理解图像和文本语义的强大能力,为多模态图像增强提供了丰富的视觉语言先验. Liang等人[29]提出了一种名为CLIP-LIT的无监督逆光图像增强方法. 该方法利用CLIP的先验来区分逆光图像和光照良好的图像,并感知不同亮度区域,从而指导增强模型的优化. Kosugi关注于可解释性图像增强,提出了名为IA-NILUT的滤波器架构[30]. 该方法通过采用CLIP实现了提示引导损失,使得每个滤波器具有可解释的名称. Chen等人[31]提出了一种名为CLIP-LUT的图像增强方法. 该方法结合了LUT和CLIP引导的提示学习. Lee等人[32]提出的CLIP-Tone通过将文本特征直接应用于无监督的图像色彩信息调整,实现了与文本信息描述一致的图像修饰.

虽然上述方法取得了不错的性能表现,但是现有基于多模态的图像增强方法主要还是通过文本信息来约束图像增强的方向,如指定期望的增强风格或修饰效果等. 但是这些方法往往忽略了图像本身不同内容的实际语义信息,无法根据图像中的不同对象进行有针对性的修饰. 针对这一问题,本文方法借助于文本信息来描述图像的具体内容,通过多模态融合方式为算法模型提供描述图像内容的语义特征,从而实现基于内容语义感知多模态融合的图像增强方法.

3　基于内容语义感知多模态融合的图像增强

如图2所示,本文提出的模型由特征提取、特征融合和曲线映射三个模块组成. 为了获取多模态特征,首先将提示信息和图像输入到多模态大语言模型生成描述图像内容的文本信息,然后采用预训练的CLIP编码器的特征提取模块,得到输入图像和对应文本信息的多模态特征. 为了实现输入图像和文本信息的相互补充,本文设计了特征融合模块,采用并行交互式注意力机制,将图像特征和文本特征进行有效融合. 基于该融合特征,曲线映射模块对输入图像进行像素级映射,以生成残差图像,并最终将其与输入图像融合,从而得到增强后的图像.

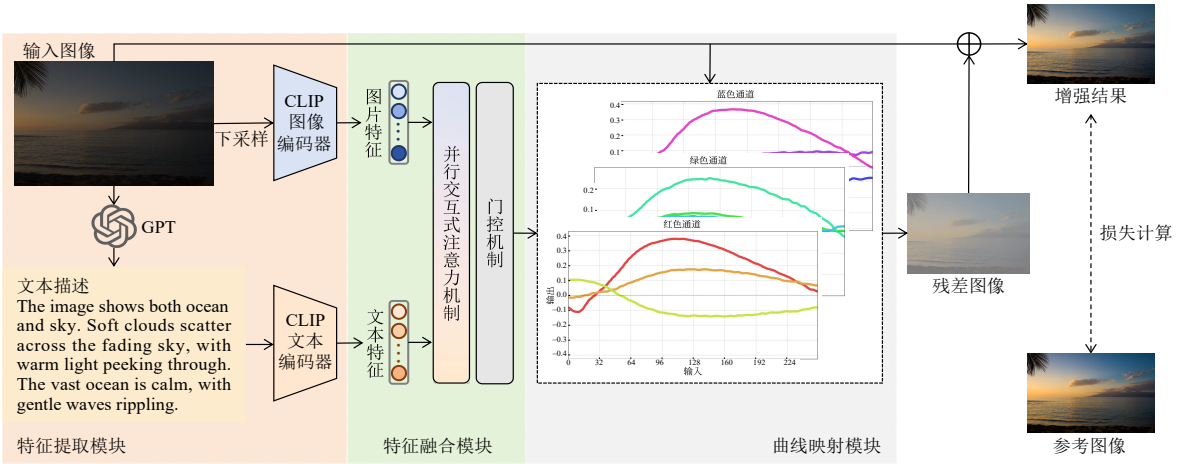


图2　本文提出方法的模型架构图

3.1 特征提取模块

3.1.1 文本数据集构建

在特征提取之前，首先需要为图像数据集$\{x_i\}(i=1,2,\cdots,N)$生成对应的文本描述数据集$\{t_i\}(i=1,2,\cdots,N)$. 其中，x_i和t_i代表训练集中第i幅图像及其对应的文本描述，数据集中图像和文本的数量用N来表示. 本文采用了最新的多模态生成式预训练模型(Generative Pre-trained Transformer, GPT)[33]生成文本描述，以确保文本能够准确捕捉图像中的内容信息，为后续的图像增强过程提供有效保障. 为实现增强过程中的内容语义感知，本文希望生成的文本描述主要关注图像中不变的核心信息，如图像的主要对象，而非那些在增强过程中可能会调整的属性(如颜色、亮度和对比度等).

因此，本文设计了特定的提示词，引导GPT在生成文本时专注于描述图像的核心内容，具体的提示词为："Describe the image in 40 words. Use short phrases. Follow this order: 1. Main subject: appearance, shape, structure, texture, patterns. 2. Secondary elements: key features. 3. Spatial composition and perspective. Avoid mentioning colors, lighting, or image quality". 该提示词要求生成模型按照图像的主要对象、次要元素及空间构图的顺序进行描述，确保生成的文本能够捕捉图像的核心语义特征，同时避免提及颜色、光照等可能在增强过程中发生变化的因素. 通过这种方式，生成的文本描述为后续的增强任务提供了精确的语义支持. 具体而言，给定输入图像x_i和提示词p，GPT生成对应的文本描述t_i. 该过程可表示为

$$t_i=\mathrm{GPT}(x_i;\ p) \tag{1}$$

3.1.2 多模态特征提取

为有效地提取多模态信息，本文选择CLIP预训练模型[20]作为图像和文本的编码器. CLIP模型能够将图像与文本映射到同一语义空间中，从而实现跨模态的特征对齐. 具体地，图像特征通过CLIP图像编码器提取得到. 对于输入图像x_i，CLIP的图像编码器E_x将其转换为d维的特征向量$\boldsymbol{F}_{x_i}$. 文本特征通过CLIP文本编码器提取得到. 对于文本描述t_i，CLIP的文本编码器E_t将其转换为d维的特征向量$\boldsymbol{F}_{t_i}$. 因此，对于输入图像x_i和对应的文本描述t_i，特征提取过程可表示为

$$\boldsymbol{F}_{x_i}=E_x(x_i),\boldsymbol{F}_{t_i}=E_t(t_i) \tag{2}$$

其中，$\boldsymbol{F}_{x_i}\in\mathbb{R}^d$和$\boldsymbol{F}_{t_i}\in\mathbb{R}^d$分别代表第$i$幅图像的特征和对应的文本特征，$d$为特征维度.

3.2 特征融合模块

为了充分利用图像特征和文本特征，本文提出了一种特征融合模块，其结构如图3所示. 该模块包含两条并行的特征处理路径，分别用于处理图像特征与文本特征. 为了实现更紧密的跨模态关联，本文引入多头注意力机制(Multi-Head Attention, MHA)[5]进行特征交互，并设计了门控机制[7]，通过动态调整不同模态的贡献度，得到表征能力更优越的多模态融合特征.

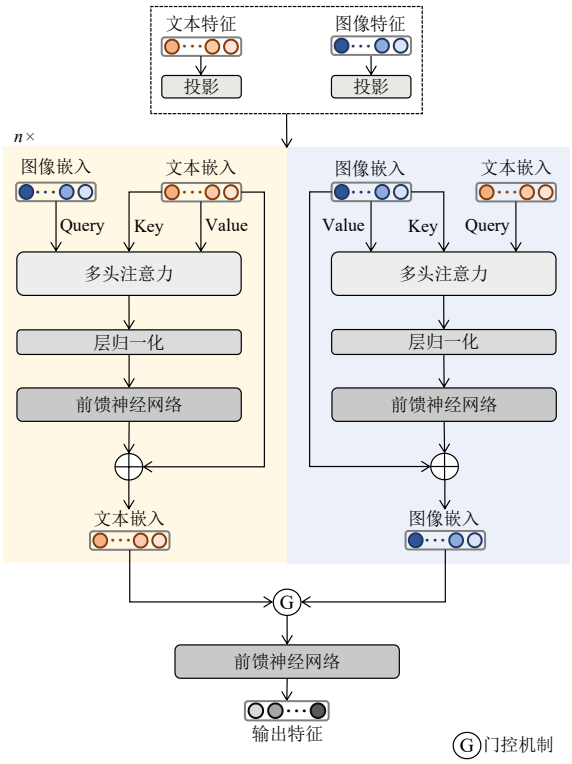


图3 特征融合模块

3.2.1 多头注意力机制

在多模态特征融合阶段，本文基于多头注意力机制，分别从文本指导图像和图像指导文本两个方向进行特征的交互融合.

具体来说，图像特征$\boldsymbol{F}_{x_i}$和文本特征$\boldsymbol{F}_{t_i}$分别作为注意力机制中的查询(Query)、键(Key)和值(Value). 在图像指导文本的路径中，图像特征$\boldsymbol{F}_{x_i}$作为查询，文本特征$\boldsymbol{F}_{t_i}$作为键和值，模型通过计算图像特征与文本特征的匹配程度，筛选出与图像内容最相关的文本信息，并赋予更高的权重. 该过程生成新的文本特征：

$$\boldsymbol{F}'_{t_i}=\mathrm{MHA}(\boldsymbol{F}_{x_i},\boldsymbol{F}_{t_i},\boldsymbol{F}_{t_i}) \tag{3}$$

其中，$\boldsymbol{F}'_{t_i}$为更新后的文本特征；MHA为多头注意力机制.

同样地，文本指导图像的路径中，文本特征$\boldsymbol{F}_{t_i}$作为查询；图像特征$\boldsymbol{F}_{x_i}$作为键和值，生成新的图像特征$\boldsymbol{F}'_{x_i}$：

$$\boldsymbol{F}'_{x_i}=\mathrm{MHA}(\boldsymbol{F}_{t_i},\boldsymbol{F}_{x_i},\boldsymbol{F}_{x_i}) \tag{4}$$

在注意力操作后，首先对图像特征$\boldsymbol{F}'_{x_i}$和文本特征

$\boldsymbol{F}'_{t_i}$进行层归一化(Layer Normalization,LN),来规范不同模态的数据分布,减少图像与文本特征在数值尺度上的差异. 随后,通过前馈神经网络(Feed-Forward Network,FFN)对特征进行非线性变化,以提升特征表示能力. 最后,采用残差连接用于缓解梯度消失问题,从而得到最新的图像特征$\hat{\boldsymbol{F}}_{x_i}$和文本特征$\hat{\boldsymbol{F}}_{t_i}$:

$$\hat{\boldsymbol{F}}_{x_i},\hat{\boldsymbol{F}}_{t_i}=\mathrm{FFN}(\mathrm{LN}(\boldsymbol{F}'_{x_i},\boldsymbol{F}'_{t_i}))+(\boldsymbol{F}_{x_i},\boldsymbol{F}_{t_i}) \tag{5}$$

3.2.2　门控机制

尽管图像特征和文本特征在注意力机制的作用下已经建立了跨模态关联,但它们仍然保持一定的独立性. 为进一步融合两个模态信息,使模型能够自适应地调整不同模态的贡献度,本文设计了一种门控机制. 该机制能够动态学习图像与文本的重要性权重,实现更有效的特征融合.

具体而言,将图像特征和文本特征进行拼接,通过全连接(Fully Connected,FC)层并应用Sigmoid激活函数,得到门控权重,用于衡量图像和文本在融合过程中各自的贡献程度. 最终,利用门控权重对图像特征和文本特征进行加权求和,生成包含两种模态信息的融合特征:

$$\boldsymbol{g}_i=\sigma\left(\mathrm{FC_g}\left([\hat{\boldsymbol{F}}_{x_i},\hat{\boldsymbol{F}}_{t_i}]\right)\right) \tag{6}$$

$$\boldsymbol{F}_{f_i}=\boldsymbol{g}_i\cdot\hat{\boldsymbol{F}}_{x_i}+(1-\boldsymbol{g}_i)\cdot\hat{\boldsymbol{F}}_{t_i} \tag{7}$$

其中,σ为Sigmoid激活函数;$\mathrm{FC_g}$为门控模块中的全连接层;$[\cdot,\cdot]$表示特征拼接操作;$\boldsymbol{g}_i$和$\boldsymbol{F}_{f_i}$分别是得到的门控权重和融合特征.

通过上述步骤,本文最终生成了包含文本语义信息和图像视觉信息的多模态融合特征$\boldsymbol{F}_{f_i}$,为后续的图像增强提供了丰富的多模态信息.

3.3　曲线映射模块

在图像增强过程中,曲线映射模块的主要作用是通过生成映射曲线对输入图像的像素进行非线性调整,从而实现图像的增强. 该模块分为两个步骤:生成映射曲线和像素映射.

在生成映射曲线的过程中,首先通过FC层对多模态融合特征$\boldsymbol{F}_{f_i}\in\mathbb{R}^d$进行特征维度的变化,生成曲线参数:

$$\boldsymbol{C}_{p_i}=\mathrm{FC_c}(\boldsymbol{F}_{f_i}) \tag{8}$$

其中,$\boldsymbol{C}_{p_i}\in\mathbb{R}^{c_u\times c_n}$,$c_u$表示曲线条数,$c_n$表示曲线节点数,$\mathrm{FC_c}$表示曲线映射模块中的全连接层.

由于曲线节点的数量c_n通常小于8 bit色深图像对应的灰度级,本文使用双三次插值(Bicubic Interpolation,BI),对曲线参数进行插值调整,公式如下:

$$\boldsymbol{C}_i=\mathrm{BI}(\boldsymbol{C}_{p_i},256) \tag{9}$$

其中,插值后得到映射曲线$\boldsymbol{C}_i\in\mathbb{R}^{c_u\times 256}$.

得到曲线之后,对输入图像进行像素值的映射. 对于RGB图像x_i的三个颜色通道$s\in\{r,g,b\}$的映射可以表示为

$$\begin{aligned}\hat{y}_i^s&=T(x_i^s;\ \boldsymbol{C}_i^s)+x_i^s\\&=\hat{y}_{\mathrm{res}_i}^s+x_i^s\end{aligned} \tag{10}$$

其中,T代表曲线映射函数,也就是利用曲线$\boldsymbol{C}_i^s$在s通道上对训练集上第i幅输入图像$x_i^s(x_i^s\in x_i)$进行映射,$\hat{y}_{\mathrm{res}_i}^s$代表在映射过程中生成的残差图像,$\hat{y}_i=[\hat{y}_i^r,\hat{y}_i^g,\hat{y}_i^b]$表示模型输出的增强图像.

通过上述过程,曲线映射模块实现了对图像的逐像素非线性调整,利用残差图像的方式进行增强,从而在保留原始图像细节的同时,依据内容信息实现自适应的修饰与增强.

3.4　损失函数

在图像增强任务中,合适的损失函数能够有效指导模型生成高质量的图像[34]. 具体而言,均方误差(Mean Squared Error,MSE)损失常用于衡量增强图像与目标图像之间像素值的差异,确保增强后的图像在像素层面接近目标图像. 结构相似度(Structural Similarity Index Measure,SSIM)损失则关注图像的结构信息,衡量两幅图像结构的相似性. 此外,ΔE损失用于评估图像颜色差异,确保增强后的图像在颜色空间中逼近目标图像.

为了使得增强图像像素、结构和颜色三个维度上都能接近目标图像,本文设计了联合特征损失函数. 该函数由MSE损失、SSIM损失和ΔE损失三部分组成. 下面对该损失函数进行详细介绍,对于训练集$\{x_i\}(i=1,2,\cdots,N)$,y_i和$\hat{y}_i$分别表示训练数据集中第i幅目标图像和本文模型增强后的图像.

MSE损失计算增强图像与目标图像之间像素值差平方的均值,其公式如下:

$$L_{\mathrm{MSE}}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\left(\hat{y}_i(k)-y_i(k)\right)^2 \tag{11}$$

其中,K为图像中像素的总数,$\hat{y}_i(k)$和$y_i(k)$分别表示增强图像和目标图像第k个像素的值.

SSIM损失用于衡量两幅图像结构的相似性. 与MSE损失不同,SSIM损失考虑了图像的亮度、对比度和结构信息,更符合人类的视觉感知. SSIM值的范围为0~1,值越大表示两幅图像越相似,因此损失函数可以表示为

$$L_{\mathrm{SSIM}}=\frac{1}{N}\sum_{i=1}^{N}\left(1-\mathrm{SSIM}(\hat{y}_i,y_i)\right) \tag{12}$$

其中,$\mathrm{SSIM}(\hat{y}_i,y_i)$表示增强图像和目标图像之间的结构相似度,可以表示为

$$\mathrm{SSIM}(\hat{y}_i,y_i)=\frac{(2\mu_{\hat{y}_i}\mu_{y_i}+C_1)(2\sigma_{\hat{y}_i y_i}+C_2)}{(\mu_{\hat{y}_i}^2+\mu_{y_i}^2+C_1)(\sigma_{\hat{y}_i}^2+\sigma_{y_i}^2+C_2)} \tag{13}$$

其中，$\mu_{\hat{y}_i}$和μ_{y_i}分别代表增强图像和目标图像的均值；$\sigma_{\hat{y}_i}^2$和$\sigma_{y_i}^2$分别表示它们的方差；$\sigma_{\hat{y}_i y_i}$表示它们的协方差；C_1和C_2是用于避免除零操作的常数.

ΔE损失用于评估图像颜色差异. 由于Lab颜色空间更接近人类的视觉感知，本文采用CIE Lab颜色空间计算ΔE损失. 其公式如下：

$$L_{\Delta E}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\left\|\mathrm{Lab}(\hat{y}_i(k))-\mathrm{Lab}(y_i(k))\right\|_2 \tag{14}$$

其中，Lab表示将RGB图像转换到Lab颜色空间的操作，采用欧几里得距离衡量两幅图像在Lab颜色空间中的差异.

综上所述，本文使用的联合损失函数由MSE损失、SSIM损失和ΔE损失加权组合而成，其公式如下：

$$L_{\mathrm{total}}=L_{\mathrm{MSE}}+\alpha L_{\mathrm{SSIM}}+\beta L_{\Delta E} \tag{15}$$

其中，α和β分别是SSIM损失和ΔE损失的权重系数.

在训练阶段，本文采用成对的数据进行监督学习. 模型通过最小化联合损失函数L_{total}来优化所有参数，确保增强结果在多个维度上接近目标图像. 为了更好地适应不同模块的特性，本文对不同部分的参数采用了不同的学习率. 具体地，对于CLIP编码器的参数，使用较低的学习率进行微调；而对于特征融合模块和曲线映射模块，则使用较高的学习率训练. 模型使用Adam优化器进行参数更新，并采用余弦退火策略动态调整学习率，以在训练过程中获得更好的收敛效果.

在测试阶段，给定一幅待增强的图像. 首先使用GPT模型生成文本描述，随后将图像和文本输入训练好的模型. 模型依次通过特征提取、多模态特征融合和曲线映射三个核心模块，最终输出增强后的图像. 整个推理过程是端到端的，能够自适应地根据内容进行修饰与增强.

4 实验

4.1 图像数据集

本文使用MIT-Adobe-5K数据集[1]进行实验，以验证提出方法的有效性. MIT-Adobe-5K数据集包含5 000幅由数码单反相机拍摄的RAW格式图像，并由五位专业摄影师(A/B/C/D/E)使用Adobe Lightroom进行后期修饰. 尽管数据集中所有图像的内容保持一致，但由于图像预处理和数据集划分方式的不同，形成了几个版本. 为了公平地与现有先进方法进行对比，本文选用了三个常用版本的数据集：DPE[4]、UPE[35]和3D LUT[11].

DPE数据集[4]采用前2 250幅图像作为训练集，中间2 250幅作为验证集，最后500幅作为测试集. 输入图像采用相机原始白平衡设置，没有经过额外调整；与图像增强领域主流方法相同，目标图像选用专家C修饰的结果. 图像分辨率统一调整为短边512像素，采用PNG格式存储.

UPE数据集[35]使用前4 500幅图像构建训练集，最后500幅作为测试集. 输入图像在专家C白平衡设置的基础上，减少了1.5的白平衡值；目标图像依然选用专家C修饰的结果. 图像分辨率统一调整为短边512像素，采用PNG格式存储.

3D LUT数据集则是Zeng等人[11]挑选了4 500幅图像用于训练，剩余500幅用于测试. 输入图像采用专家C的白平衡设置；目标图像同样选用专家C修饰的结果. 图像分辨率统一调整为短边480像素，采用JPG格式存储.

通过使用以上三个版本的数据集进行算法模型训练和测试实验，本文能够全面评估所提出方法在不同数据集设定下的性能表现，从而验证其在多种场景中的泛化能力，确保实验结果的公正可靠.

4.2 实验细节

本文的实验在配备了一张NVIDIA RTX 3090显卡的服务器中进行，使用深度学习框架PyTorch实现模型的训练和评估. 本文采用预训练的CLIP模型作为骨干网络.

模型训练采用Adam优化器[36]进行参数更新. 考虑到模型不同模块的特性，对CLIP编码器参数采用较小的学习率(1×10^{-5})，以保持其预训练特征；对特征融合和曲线映射等其他模块采用了1×10^{-4}的学习率，以加快模型收敛. 为了使模型训练更加稳定，本文引入余弦退火学习率调度器[37]来动态调整学习率，并将最大训练轮次设定为80轮，批次大小为10.

为提升模型的泛化能力，本文采用了随机水平垂直翻转、随机裁剪和随机转置三种数据增强策略. 模型训练的损失函数由MSE损失、SSIM损失和ΔE损失组成，α和β分别设置为0.4和0.1，以确保增强后的图像在像素、结构和颜色三个方面都有较好的视觉效果.

4.3 结果对比

为了测试提出方法的性能，本文采用峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、SSIM和ΔE_{ab}三种定量评价指标对方法进行客观评价. PSNR和SSIM值越高，表示增强后图像质量越好[38]；ΔE_{ab}值越低，表示颜色差异越小. 本文在MIT-Adobe-5K[1]数据集的三个版本上分别对提出的方法进行了训练和测试，并将实验结果与每个数据集中主流的图像增强方法进行了对比. 实验结果如表1所示，其中，最优结果以加粗标记，次优结果以下划线标记.

在DPE数据集上，本文方法的PSNR和SSIM分别

表1　与现有主流图像增强方法的比较结果

数据集	方法	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
DPE[4]	DPE[4]	23.80	0.900	—
	CURL[18]	24.04	0.900	—
	DeepLPF[39]	23.93	0.903	—
	FlexiCurve[10]	24.03	0.910	—
	4D LUT[13]	24.61	0.918	9.74
	NamedCurves[17]	24.91	0.927	**7.82**
	本文方法	**25.14**	**0.932**	8.30
UPE[35]	UPE[35]	23.04	0.893	—
	CURL[18]	24.20	0.880	—
	DeepLPF[39]	24.48	0.887	—
	NamedCurves[17]	25.20	0.906	7.58
	StarEnhancer[15]	25.51	0.890	7.22
	本文方法	**26.07**	**0.912**	**7.03**
3D LUT[11]	HDRNet[8]	24.66	0.915	8.06
	CSRNet[40]	25.17	0.924	7.75
	3D LUT[11]	25.21	0.922	7.61
	SepLUT[16]	25.47	0.921	7.54
	AdaInt[19]	25.49	0.926	7.47
	本文方法	**25.52**	**0.934**	**7.41**

达到了25.14 dB和0.932,均优于现有的图像增强方法.与其他最佳的NamedCurves[17]方法相比,PSNR提升了0.23 dB,SSIM提升了0.5%. 这表明,本文方法在图像质量上具有显著优势. 尽管在色彩差异指标ΔE_{ab}上,本文方法略低于NamedCurves,这一差异主要源于NamedCurves采用了颜色命名机制,能够更有效地调整特定颜色区域,从而在色彩差异上获得更优越的表现.

在UPE数据集上,本文方法展现出更为显著的性能优势. 与之前最优方法相比,本文方法的PSNR提升了0.56 dB,达到了26.07 dB;SSIM提升了0.6%,达到了0.912;ΔE_{ab}则降低了0.19,达到7.03. 充分证明了本文方法在该数据集上的全面优势.

在3D LUT数据集上,本文方法同样取得了最优的结果,PSNR为25.52 dB,SSIM为0.934,均超过了现有最优方法AdaInt[19]的25.49 dB和0.926. 特别是在SSIM指标上,本文方法的优越性更加显著,表明其在保持图像结构和细节方面具有明显的优势. 此外,本文方法在ΔE_{ab}指标上以7.41取得了最优结果,进一步验证了本文方法在色彩还原能力上的有效性.

为进一步检验本文方法的计算效率,本文方法与代表性轻量级LUT方法AdaInt[19]在计算复杂度和推理速度方面进行比较. 在相同硬件环境(NVIDIA RTX 3090显卡)下处理单张1080p图像时,AdaInt的平均推理时间为1.52 ms,参数量为619.7 K;本文方法的平均推理时间为29 ms,参数量为457 M. 主要原因在于:本文方法需执行CLIP编码器的前向传播及特征融合模块的计算,而AdaInt只需通过一个小型网络生成LUT参数. 虽然本文方法在计算复杂度和推理速度方面存在劣势,但是这一计算开销带来了显著的性能提升. 正如表1所示,本文方法在各项指标上均取得最优或接近最优的结果. 同时,29 ms的推理时间,即以34帧/s的速度,能满足实时视频处理等下游任务的需求.

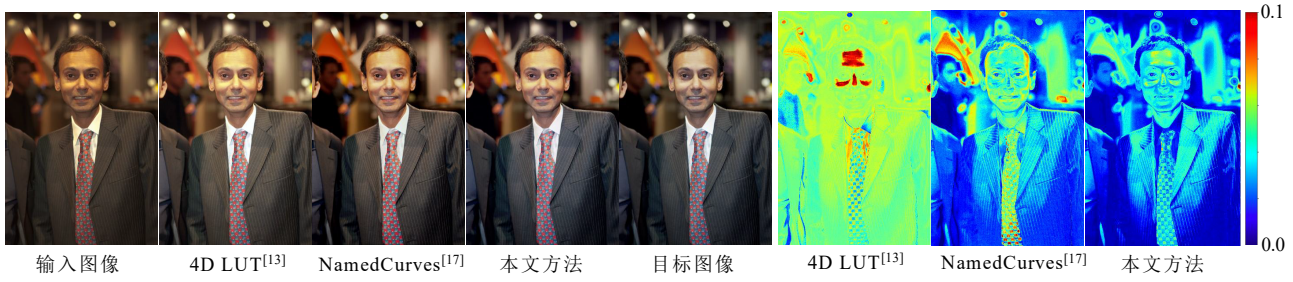
此外,为了更直观地展示本文方法的有效性,本文在MIT-Adobe-5K数据集中随机选取若干图像进行可视化实验,并对比不同方法的定性增强效果. 如图4所示,本文不仅展示了不同方法的增强结果,还展示了增强图像与目标图像的差分图. 差分图中色彩的变化用于直观反映增强结果与目标图像之间的差异:颜色越接近蓝色,差异越小;反之,颜色越偏向红色,则表示差异越大. 从图中可以观察到本文方法相对于其他主流方法在不同场景下均得到优秀的增强效果.

具体来说,在图4(*a*)的室内人像场景中,4D LUT和NamedCurves方法虽然提高了整体亮度,但未能准确还原西服和领带的颜色,导致色彩失真. 相比之下,本文方法正确地将西服增强为蓝灰色,使增强后的图像更接近目标图像,同时保持丰富的细节. 差分图整体呈现深蓝色,表明本方法与目标图像的误差较小.

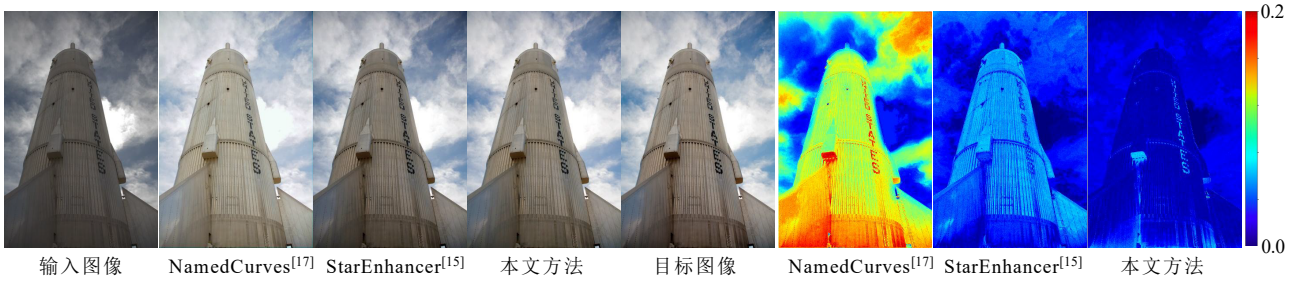
图4(*b*)展示了一个典型的逆光场景. 逆光场景亮度的调整难度较大,容易出现过曝或过暗的情况. NamedCurves方法调整幅度过大,导致火箭及云层区域出现过曝现象. 在差分图中,火箭主体和天空云层区域呈现明显的黄红色. 相比之下,StarEnhancer方法未能充分提升亮度,使画面整体偏暗. 本文方法在亮度调整上实现了更好的平衡,使火箭与天空之间的明暗对比更加自然. 其最终增强结果更接近目标图像,差分图中呈现更为一致的深蓝色,进一步验证了本方法在亮度增强方面的有效性.

在图4(*c*)的室内运动场景中,台球桌的蓝色和背景的颜色存在较大差异,这对图像增强方法提出了更高的挑战. 3D LUT和AdaInt方法对整体色调的调整不够准确,导致图像偏色严重,这些方法的差分图中台球桌区域和人物区域均显示出较大的偏差. 相比之下,本文方法能够准确理解场景内容,对台球桌的蓝色进行精准还原,同时保持人物与背景的自然过渡,使增强结果更接近目标图像. 在差分图中也可以直观证明本文方法的增强效果,特别是在台球桌和人物区域,其误差明显小于其他方法.

为了更直观地展现文本描述信息的有效性,本文随机选取DPE数据集中的一幅图像,并与该数据集上其他性能最好的两种方法进行对比,实验结果如图5所示. 可以观察到,4D LUT[13]和NamedCurves[17]两种方法在增强该图像时未能有效区分前景人物和背景墙,导



(*a*) DPE数据集中图像增强效果对比



(*b*) UPE数据集中图像增强效果对比

(*c*) 3D LUT数据集中图像增强效果对比

图4 可视化实验的对比结果

致整体增强效果较为平均,缺乏层次感,最终增强结果显得平淡且不够自然. 相比之下,本文方法在人物和背景墙的处理上表现出显著的优势. 具体而言,本文方法有效提升了人物的亮度,使人物的面部和手部细节更加清晰. 与此同时,墙体的颜色得到了更准确的还原,墙面的纹理也得到了很好的保护. 这表明,基于内容语义感知的图像增强策略能够根据图像内容进行针对性处理,灵活调整不同区域的增强效果,从而使结果更加贴近目标图像.

为进一步验证本文方法在内容语义感知方面的有效性,本文随机选取DPE数据集中三幅图像,并在图6中展示了增强前后图像的视觉对比结果. 从图中可以看出,输入图像中存在大范围色彩相似但语义不同的内容,而本文方法能够更精准地根据不同内容的语义信息进行针对性的增强. 例如,在第一行的输入图像中,天空和海洋均呈现灰蓝色,在采用本文方法增强的图像中,天空被调整为更明亮的天蓝色,而水面则呈现出更具层次感的水绿色. 在第三行的输入图像中,天空和雪地都呈现深灰色,导致画面缺乏层次感. 采用本文方法增强后,天空被优化为更富有层次的蓝灰色,而雪地则被调整为更自然的浅灰色. 从上述对比结果可以看出,本文方法能够有效利用文本感知图像内容的语义信息,实现对图像中不同内容的精准修饰,使得增强图像的效果更加自然.

4.4 消融实验

为了进一步验证引入多模态框架对图像增强效果的影响,本文设计了针对模型编码器权重的消融实验. 本次实验选用的预训练权重分为两大类:单模态预训练权重和多模态预训练权重. 单模态预训练权重来源于PyTorch提供的ResNet[41]和ViT[5]模型权重,这些权重仅在图像数据集上训练. 在单模态权重的实验设计中,所使用的模型直接采用图像编码器生成映射曲线,不包含文本编码器和特征融合模块. 作为对比,多模态预训练权重由CLIP官方提供[20],这些权重基于大规模

图5　带有文本描述的可视化对比结果

图6　增强前后图像的视觉对比结果

文本-图像配对数据集进行训练,具备处理图像和文本的能力. 在多模态权重的实验设计中,模型架构与图2所示一致,包含完整的特征提取模块和特征融合模块.

表2展示了在DPE数据集[4]上的消融实验结果,其中最优结果以加粗标记,次优结果以下划线标记. 从表中可以看出,基于ViT架构的模型在PSNR、SSIM和ΔE_{ab}三项指标上均优于基于ResNet架构的模型. 特别值得注意的是,多模态ViT权重与单模态ViT权重的对比. 使用CLIP提供的ViT权重,在PSNR上提升了0.70 dB,SSIM提升了0.7%,ΔE_{ab}降低了0.56. 这一显著的性能差异主要归功于CLIP模型的文本-图像多模态能力,证明了引入文本信息对图像增强任务的必要性.

表2　编码器权重消融实验

预训练模型权重		PSNR ↑	SSIM ↑	ΔE_{ab} ↓
单模态	ResNet	23.93	0.903	9.94
	ViT	<u>24.44</u>	<u>0.925</u>	<u>8.86</u>
多模态	ResNet	24.00	0.906	9.90
	ViT(本文方法)	**25.14**	**0.932**	**8.30**

为进一步分析CLIP编码器学习率对模型性能的影响,本文设计了CLIP编码器微调学习率的消融实验,在DPE数据集[4]上的结果如表3所示,表中对最优结果进行加粗标记,下划线标记次优结果. 实验过程中,除CLIP编码器外,其余模块(如特征融合模块与曲线映射模块)的学习率固定为1×10^{-4},仅对CLIP编码器的学习率进行调整. 从表3结果可见,当CLIP编码器的学习率设为1×10^{-5}时,模型在PSNR、SSIM和ΔE_{ab}三个指标上均取得最优性能,表明对预训练的CLIP编码器采用较低学习率进行微调是有效的. 这一现象可归因于CLIP模型在大规模图像和文本数据上预训练所获得的先验知识. 若采用较高的学习率(如5×10^{-5}或1×10^{-4}),则可能破坏这些有价值的特征表示,导致模型性能下降. 相反,若采用的学习率较低(如1×10^{-6}或5×10^{-6}),则编码器参数更新幅度过小,难以适应下游图像增强任务,同样限制了模型性能的提升.

表3　CLIP编码器学习率消融实验

学习率	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
1×10^{-6}	24.68	<u>0.928</u>	8.74
5×10^{-6}	<u>24.87</u>	0.927	<u>8.48</u>
1×10^{-5}(本文方法)	**25.14**	**0.932**	**8.30**
5×10^{-5}	24.66	0.926	8.60
1×10^{-4}	24.63	0.925	8.67

为了验证本文设计的基于并行交互式注意力机制的特征融合模块的有效性,本文进行了特征融合方式的消融实验. 实验将此方法与其他常见的融合方式进行对比,包括特征拼接(Concat)、加法融合(Add)和乘法融合(Multiply),在DPE数据集[4]上的实验结果如表4所示. 表中对最优结果进行加粗标记,下划线标记次优结果.

从表4可以看到,本文方法在所有指标上均取得了

表4　特征融合方式消融实验

特征融合方式	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
特征拼接(Concat)	24.73	<u>0.929</u>	8.58
加法融合(Add)	24.86	0.921	8.70
乘法融合(Mutiply)	<u>24.90</u>	0.926	<u>8.48</u>
本文方法	**25.14**	**0.932**	**8.30**

最佳性能. 具体而言,本文方法的PSNR达到了25.14 dB,SSIM为0.932,ΔE_{ab}为8.30,均高于其他多模态融合策略. 通过上述实验可以证明,基于并行交互式多头注意力机制的特征融合方式能够在图像增强任务中更好地捕捉图像与文本之间的语义关联. 相比于简单的特征拼接、加法或乘法,注意力机制能够在不同模态间自动选择重要特征进行融合,从而显著提升模型的性能.

为了验证本文提出的联合损失函数的有效性,本文设计了消融实验,分别分析了平均绝对误差(Mean Absolute Error,MAE)损失、MSE损失、SSIM损失和ΔE损失对模型性能的影响,在DPE数据集[4]上的实验结果如表5所示. 表中对最优结果进行加粗标记,下划线标记次优结果.

从表5可以看出,单独使用MAE损失或MSE损失性能较差. 在MSE损失的基础上增加SSIM损失后,模型性能获得了显著提升. PSNR和SSIM分别提高到25.06 dB和0.932,ΔE_{ab}降低到8.32. 这表明SSIM损失能够有效地保持图像的结构信息,使增强后的图像在视觉效果上更接近目标图像. 进一步加入ΔE损失使PSNR提升了0.08 dB,同时ΔE_{ab}也有所下降. 这说明ΔE损失能够指导模型在Lab颜色空间中进行优化,提高图像颜色还原的准确性.

表5 损失项消融实验

MAE	MSE	SSIM	ΔE	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
√				24.89	<u>0.929</u>	8.43
	√			24.95	0.928	8.38
	√	√		<u>25.06</u>	**0.932**	<u>8.32</u>
	√	√	√	**25.14**	**0.932**	**8.30**

上述消融实验结果充分证明了本文设计的联合损失函数的有效性. 三种损失函数的组合能够从像素值、结构信息和颜色差异三个维度共同约束模型的训练过程,使模型在各项评价指标上都达到较好的性能.

为进一步验证联合损失函数中各项权重设置的合理性,本文对SSIM损失权重α和ΔE损失权重β进行了系统的消融实验,在DPE数据集[4]上的结果如表6所示,表中对最优结果进行加粗标记,下划线标记次优结果. 在分析SSIM损失权重α的影响时,将β固定为0.1. 实验结果表明,随着α从0逐步增加至0.4,模型在SSIM指标上持续提升;当α达到0.4时,性能达到最优. 然而,进一步增大α至0.6或0.8时,PSNR数值明显降低,ΔE_{ab}数值明显上升,可能是由于结构信息约束过强,抑制了像素精度和颜色还原能力.

在分析ΔE损失权重β对模型的影响时,将α固定为0.4. 实验结果显示,随着β从0增加至0.1,PSNR由25.06 dB提升至25.14 dB,ΔE_{ab}达到最优值8.30,表明适度引入颜色差异损失有助于提升颜色还原效果. 然而,

表6 损失项权重消融实验

损失项权重		PSNR ↑	SSIM ↑	ΔE_{ab} ↓
α	0	<u>25.02</u>	0.926	<u>8.41</u>
	0.2	24.85	0.931	8.59
	0.4	**25.14**	**0.932**	**8.30**
	0.6	24.83	0.929	8.47
	0.8	24.89	<u>0.931</u>	8.52
β	0	25.06	**0.932**	<u>8.32</u>
	0.05	24.98	0.930	8.41
	0.1	<u>25.14</u>	**0.932**	**8.30**
	0.15	**25.15**	0.928	8.44
	0.2	24.95	<u>0.931</u>	8.42

当β进一步增大时,尽管PSNR变化不大,SSIM数值有所下降,ΔE_{ab}数值有所上升,说明过度强调Lab颜色空间中的颜色差异可能破坏图像结构一致性,甚至引发颜色失真.

综上,当α设为0.4、β设为0.1时,模型在PSNR、SSIM与ΔE_{ab}三项指标之间实现了较好的平衡.

为了验证文本描述对模型性能的影响,本文设计了相应消融实验. 在该实验中设置四种不同的文本输入,并分别训练模型以评估其性能:无文本,不使用任何文本信息;随机文本,使用随机生成的文本作为模型输入;简短文本,使用由GPT生成的简短概括性描述. 本文方法,使用精确描述图像视觉内容的文本,以充分发挥文本信息的辅助作用. 在UPE数据集[35]上的实验结果如表7所示. 表中对最优结果进行加粗标记,下划线标记次优结果.

表7 文本描述消融实验

文本描述	PSNR ↑	SSIM ↑	ΔE_{ab} ↓
无文本	25.92	0.900	7.08
随机文本	25.85	0.902	7.12
简短文本	<u>25.98</u>	<u>0.904</u>	<u>7.05</u>
本文方法	**26.07**	**0.912**	**7.03**

从表7中可以看出,在不使用文本描述时,与本文方法相比,模型的PSNR和SSIM分别下降至25.92 dB和0.900,ΔE_{ab}则增加至7.08. 此外,当输入为随机文本时,模型的性能甚至低于无文本的情况. 这表明,只有当文本描述能够准确反映图像内容时,文本信息才能有效指导图像增强,从而提升增强结果的质量. 值得关注的是,当采用仅包含概括性描述的简短文本时,PSNR和SSIM分别提升至25.98 dB和0.904,较无文本基准提升0.06 dB和0.4%. 而采用更具体文本描述的本文方法,PSNR和SSIM分别达到26.07 dB和0.912,较简短文本描述分别提升了0.09 dB和0.8%. 因此,上述消融实验结果表明,准确且详细地描述图像视觉内容的文本能够为本文模型提供有效的先验知识,从而实现内容感

知的多模态融合,提高图像增强的效果.

5 结论

本文提出了一种基于内容语义感知多模态融合的图像增强方法. 该方法通过引入描述图像内容语义感知信息的文本特征作为图像特征的补充,进而得到内容语义感知的多模态融合特征,从而实现对图像不同内容的精细化修饰,有效地增强了图像. 本文方法提出的特征提取模块能够有效地提取描述图像内容语义信息的多模态特征;提出的基于并行交互式注意力机制的多模态融合策略可以充分地将文本和图像特征进行交互融合;提出的曲线映射模块能够根据包含内容语义信息的多模态特征生成更有效的曲线映射关系,从而高效地完成对图像内容的修饰与增强. 通过大量定量实验验证、可视化分析以及算法模型的消融实验表明,本文方法在三个数据集上的性能表现均优于当前最先进的图像增强方法,充分验证了引入内容语义感知信息对提升图像增强效果的必要性.

参考文献

[1] BYCHKOVSKY V, PARIS S, CHAN E, et al. Learning photographic global tonal adjustment with a database of input/output image pairs[C]//CVPR 2011. Piscataway: IEEE, 2011: 97-104.

[2] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[3] CHEN Q F, XU J, KOLTUN V. Fast image processing with fully-convolutional networks[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2516-2525.

[4] CHEN Y S, WANG Y C, KAO M H, et al. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6306-6314.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22) [2024-11-24]. https://arxiv.org/pdf/2010.11929/1000.

[6] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5718-5729.

[7] WANG T, ZHANG K H, SHEN T R, et al. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 2654-2662.

[8] GHARBI M, CHEN J W, BARRON J T, et al. Deep bilateral learning for real-time image enhancement[J]. ACM Transactions on Graphics, 2017, 36(4): 1-12.

[9] GUO C L, LI C Y, GUO J C, et al. Zero-reference deep curve estimation for low-light image enhancement[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1777-1786.

[10] LI C Y, GUO C L, ZHOU S C, et al. FlexiCurve: Flexible piecewise curves estimation for photo retouching[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2023: 1092-1101.

[11] ZENG H, CAI J R, LI L D, et al. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2058-2073.

[12] WANG T, LI Y, PENG J Y, et al. Real-time image enhancer *via* learnable spatial-aware 3D lookup tables[C]// 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 2451-2460.

[13] LIU C X, YANG H, FU J L, et al. 4D LUT: Learnable context-aware 4D lookup table for image enhancement[J]. IEEE Transactions on Image Processing, 2023, 32: 4742-4756.

[14] BIANCO S, CUSANO C, PICCOLI F, et al. Personalized image enhancement using neural spline color transforms[J]. IEEE Transactions on Image Processing, 2020, 29: 6223-6236.

[15] SONG Y D, QIAN H, DU X. StarEnhancer: Learning real-time and style-aware image enhancement[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 4106-4115.

[16] YANG C Q, JIN M G, XU Y, et al. SepLUT: Separable image-adaptive lookup tables for real-time image enhancement[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 201-217.

[17] SERRANO-LOZANO D, HERRANZ L, BROWN M S, et al. NamedCurves: Learned image enhancement *via* color Naming[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 92-108.

[18] MORAN S A, MCDONAGH S, SLABAUGH G. CURL:

Neural curve layers for global image enhancement[C]// 2020 25th International Conference on Pattern Recognition. Piscataway: IEEE, 2021: 9796-9803.

[19] YANG C Q, JIN M G, JIA X, et al. AdaInt: Learning adaptive intervals for 3D lookup tables on real-time image enhancement[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 17501-17510.

[20] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//MEILA M, ZHANG T. Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.

[21] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3431-3440.

[22] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[23] SONG Y D, QIAN H, DU X. Multi-curve translator for high-resolution photorealistic image translation[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 126-143.

[24] BERLIN B, KAY P. Basic Color Terms: Their Universality and Evolution[M]. Berkeley: University of California Press, 1991.

[25] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1510-1519.

[26] KARAIMER H C, BROWN M S. A software platform for manipulating the camera imaging pipeline[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 429-444.

[27] LI C Y, ZHANG B, HONG D F, et al. CasFormer: Cascaded transformers for fusion-aware computational hyperspectral imaging[J]. Information Fusion, 2024, 108: 102408.

[28] 李晨玉, 洪丹枫, 张兵. 深度展开网络的高光谱异常探测[J]. 遥感学报, 2024, 28(1): 69-77.
LI C Y, HONG D F, ZHANG B. Deep unfolding network for hyperspectral anomaly detection[J]. National Remote Sensing Bulletin, 2024, 28(1): 69-77. (in Chinese)

[29] LIANG Z X, LI C Y, ZHOU S C, et al. Iterative prompt learning for unsupervised backlit image enhancement[C]// 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 8060-8069.

[30] KOSUGI S. Prompt-guided image-adaptive neural implicit lookup tables for interpretable image enhancement[C]// Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 6463-6471.

[31] CHEN W, KE Q, LI Z. CLIP guided image-perceptive prompt learning for image enhancement[EB/OL]. (2023-11-07)[2024-11-24]. http://arxiv.org/abs/2311.03943.

[32] LEE H, KANG K, OK J, et al. CLIPtone: Unsupervised learning for text-based image tone adjustment[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 2942-2951.

[33] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical re-port[EB/OL]. (2023-03-15) [2024-11-24]. http://arxiv.org/abs/2303.08774.

[34] ZHAO H, GALLO O, FROSIO I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on Computational Imaging, 2017, 3(1): 47-57.

[35] WANG R X, ZHANG Q, FU C W, et al. Underexposed photo enhancement using deep illumination estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6842-6850.

[36] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22)[2024-11-24]. http://arxiv.org/abs/1412.6980.

[37] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[EB/OL]. (2016-08-13) [2024-11-24]. https://openreview.net/pdf?id=Skq89Scxx.

[38] HORÉ A, ZIOU D. Image quality metrics: PSNR *vs*. SSIM[C]//2010 20th International Conference on Pattern Recognition. Piscataway: IEEE, 2010: 2366-2369.

[39] MORAN S A, MARZA P, MCDONAGH S, et al. DeepLPF: Deep local parametric filters for image enhancement[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 12823-12832.

[40] HE J W, LIU Y H, QIAO Y, et al. Conditional sequential modulation for efficient global image retouching[C]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 679-695.

[41] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

作者简介

祝汉城　男,1989年12月出生于江苏省徐州市. 现为中国矿业大学计算机科学与技术学院/人工智能学院副教授、硕士生导师. 主要研究方向为图像美学评价与增强、计算机视觉和人工智能.
E-mail: zhuhancheng@cumt.edu.cn

刘新宇　男,2000年10月出生于山东省滨州市. 现为中国矿业大学计算机科学与技术学院/人工智能学院硕士研究生. 主要研究方向为图像增强与计算机视觉.
E-mail: xinyu_liu_cs@163.com

姚　睿　男,1982年7月出生于河南省南阳市. 现为中国矿业大学计算机科学与技术学院/人工智能学院教授、博士生导师. 主要研究方向为计算机视觉和人工智能.
E-mail: ruiyao@cumt.edu.cn

邵志文　男,1994年12月出生于安徽省马鞍山市. 现为中国矿业大学计算机科学与技术学院/人工智能学院副教授、硕士生导师. 主要研究方向为情感计算、计算机视觉和人工智能.
E-mail: zhiwen_shao@cumt.edu.cn

周　勇　男,1974年9月出生于江苏省徐州市. 现为中国矿业大学计算机科学与技术学院/人工智能学院教授、博士生导师. 主要研究方向为计算机视觉、机器学习和人工智能.
E-mail: yzhou@cumt.edu.cn

李雷达　男,1982年10月出生于江苏省徐州市. 现为西安电子科技大学教授、博士生导师. 主要研究方向为计算机视觉、情感计算和人工智能.
E-mail: ldli@xidian.edu.cn