中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2025)08-2743-15

论文引用格式:Yu B, Xiang X, Fan Z H, Huang D J and Ding Y D. 2025. Image colorization via color query and feature enhancement. Journal of Image and Graphics, 30(8):2743-2757(于冰, 相雪, 范正辉, 黄东晋, 丁友东. 2025. 融合颜色查询与特征增强的图像上色方法. 中国图象图形学报, 30(8):2743-2757)[DOI:10.11834/jig.240506]

融合颜色查询与特征增强的图像上色方法

于冰[1,2*],相雪[1],范正辉[1],黄东晋[1,2],丁友东[1,2]

1. 上海大学上海电影学院,上海　200072; 2. 上海大学上海电影特效工程技术研究中心,上海　200072

摘　要:目的　图像上色在老照片修复和黑白电影增强等方面具有重要应用。现有方法在颜色预测过程中由于无法保证色彩的一致性,缺乏对局部细节的精细处理,导致某些色彩区域的上色效果不佳。**方法**　采用编码器—解码器结构,编码器用于提取灰度图像特征,解码器用于恢复空间分辨率。颜色预测网络通过丰富的视觉特征来细化颜色查询,并通过像素增强模块学习空间上的关注度来增强特定区域的像素。进一步,所提方法通过特征增强模块优化原图和生成图之间的颜色匹配关系,全面地捕捉图像的特征,实现细节保持的图像上色,减少颜色溢出。**结果**　实验在数据集ImageNet(val5k)、ImageNet(val50k)、COCO-Stuff、ADE20K和CelebA-HQ上与5种灰度图像自动上色方法进行比较,在上色结果的客观质量对比中,与性能第2的模型相比,所提方法在评价指标Frechet初始距离(Frechet inception distance, FID)上降低0.2,峰值信噪比(peak signal-to-noise ratio, PSNR)提升0.13 dB;在色彩指标的对比中,所提方法在色彩丰富度方面取得最高分;在主观评价和用户调查中,所提方法的上色效果与人类的审美感受较为一致,得到了最为优异的评价。此外,消融实验结果进一步证明了所提方法采用的模型结构在提升上色性能方面的有效性。**结论**　所提出的上色模型,更能捕捉并再现图像中的细节与色彩关系,实现了高质量的上色效果。

关键词:图像上色;灰度图像;空间注意力;颜色查询;特征增强

Image colorization via color query and feature enhancement

Yu Bing[1,2*], Xiang Xue[1], Fan Zhenghui[1], Huang Dongjin[1,2], Ding Youdong[1,2]

1. Shanghai Film Academy, Shanghai University, Shanghai　200072, China;

2. Shanghai Film Special Effects Engineering Technology Research Center, Shanghai University, Shanghai　200072, China

Abstract: Objective Image colorization refers to the process of predicting plausible colors for each pixel in a grayscale image, with the goal not necessarily being an exact restoration of the original color. Given that the same object can often be assigned different colors, colorization inherently exhibits multimodal characteristics. This complexity presents challenges for researchers and sustained interest in the field. As an important direction in computer vision, image colorization has gained considerable attention, particularly with the advancements in deep learning. Traditional colorization methods often rely on user input, such as scribble-based or reference image-based techniques, which, while effective, are hindered by limitations such as the inability to handle batch processing and extended processing times. To address these issues and reduce manual intervention, deep learning techniques have propelled the development of automatic colorization. Fully auto-

收稿日期:2024-08-27;**修回日期:**2025-01-06;**预印本日期:**2025-01-13

*通信作者:于冰　yubing@shu.edu.cn

基金项目:教育部产学合作协同育人项目(230801284291638,231104276162735,220504276214223);上海市人才发展资金项目(2021016)

Supported by: Collaborative Education Project of Industry-University Cooperation, Ministry of Education(230801284291638,231104276162735,220504276214223);The Development Fund for Shanghai Talents(2021016)

matic methods, which eliminate the need for user interaction, can efficiently colorize images; however, challenges including the accuracy of color restoration and the preservation of image details remain. Deep learning approaches, particularly those based on convolutional neural networks and generative adversarial networks, have demonstrated improvements in colorization performance but continue to face issues such as insufficient semantic understanding and blurred details. Recently, Transformer models have shown promise in image colorization tasks by leveraging their ability to capture long-range dependencies, further enhancing results. However, existing methods still struggle with challenges such as color bleeding and loss of detail clarity, especially in highly detailed images. Achieving fully automated, natural, and plausible colorization remains an ongoing research challenge. **Method** In this study, we propose an end-to-end grayscale image colorization method utilizing encoder-decoder architecture for fully automatic colorization. Given an input grayscale image, the network predicts the chrominance channels in the CIELAB color space to generate a colorized image. The encoder employs ConvNeXt to leverage its multiscale semantic representation capabilities, effectively extracting high-level semantic features from the grayscale image. Multiscale feature maps are passed from the encoder to the decoder through convolutional connection layers, progressively restoring the image's spatial resolution. These feature maps are then fed into the color prediction network, where a pixel enhancement block (PEB) refines the color predictions. The PEB is designed to focus on and enhance specific regions of the image, improving color matching accuracy by utilizing convolutional layers and pooling operations to generate spatial attention weights. These weights are element-wise multiplied with the original image, enabling spatial enhancement and better capturing important regions for improved color precision. The color query block in the color prediction network employs a Transformer-based approach, incorporating learnable color embedding memories that store sequences of color representations. Through cross-attention and self-attention mechanisms, color embeddings are progressively correlated with image features, reducing dependence on manual priors and improving sensitivity to semantic information. This approach mitigates issues such as color bleeding. Furthermore, to enhance the learning of color information and latent features from the grayscale image, the feature enhancement block generates attention maps using convolutional operations with varying kernel sizes. These maps are fused with the original image through convolutional layers to produce the final output tensor. This methodology ensures effective reconstruction of color and structural information, thereby enhancing the overall performance of the image colorization process. **Result** In the experiments, the proposed colorization model was trained on the large-scale ImageNet dataset and extensively evaluated across multiple benchmark datasets, including ImageNet (val5k), ImageNet (val50k), COCO-Stuff, ADE20K, and CelebA-HQ. The evaluation metrics included Frechet inception distance (FID), colorfulness score (CF), and peak signal-to-noise ratio (PSNR), which assess the realism, color quality, and reconstruction accuracy of the generated images. The model utilized a pretrained ConvNeXt-L as the encoder, paired with a multiscale decoder, and was optimized using the AdamW optimizer. All experiments were performed on four Tesla A100 GPUs. Comparative results showed that the proposed method remarkably outperformed existing approaches such as DeOldify, Wu et al., BigColor, ColorFormer, and DDColor, particularly in terms of color richness and realism. Quantitative comparisons across five test datasets further demonstrated that the proposed model consistently achieved the lowest FID scores, indicating superior image quality and strong generalization. Compared with the second-best model, the FID is reduced by 0.2, while the PSNR is improved by 0.13 dB. While previous methods achieved higher CF scores, a higher colorfulness score does not always correlate with better visual quality. Thus, the metric $\triangle$CF was introduced to measure the difference in colorfulness between generated and real images. The proposed method achieved the lowest $\triangle$CF scores across all datasets, reflecting its ability to generate more natural and realistic colorizations while preserving image diversity. Given the subjective nature of image colorization, a user study was also conducted, showing that over 30% of users preferred the colorization results produced by the proposed method. Additionally, ablation studies confirmed the effectiveness of the model architecture in enhancing colorization performance. **Conclusion** The proposed colorization model is good at capturing and reproducing the details and color relationships in the image, achieving high-quality colorization results.

Key words: image colorization; grayscale image; spatial attention; color inquiry; feature enhancement

0 引 言

图像上色是计算机视觉和图像处理领域的一个重要研究方向。随着深度学习技术的迅猛发展，图像自动上色已成为一个备受关注的课题。具体而言，图像上色需要解决色彩还原的准确性问题，并同时保持图像的细节和质感，这一过程耗时耗力。因此如何快速且准确地对灰度图像进行自动化上色仍然是一个具有挑战性的研究问题。

图像上色方法主要分为两类：传统上色方法和深度学习上色方法。传统上色方法需要用户辅助或参考彩色图像以减少不确定性，这些方法在计算机图形学领域已有广泛应用。Qu 等人（2006）提出一种基于涂鸦操作的快速上色方法，通过自动颜色扩散分析亮度联系，提升了视觉质量，但在色彩丰富度和细节表现方面表现不足。为此，Liu 等人（2008）提出基于参考图像的上色技术，通过网络搜索参考图像，将颜色转移到目标图像，并结合照明分量生成最终结果。为减少对参考图像的依赖，Chia 等人（2011）引入语义文本标签和分割提示，通过用户界面选择最优效果，大幅减少交互，提升了自然度和上色质量。然而，这些传统的上色方法速度慢、特征提取依赖人工设计并且缺乏自适应性，导致上色效果不理想且不自然。

随着深度神经网络技术的快速发展，自动上色研究取得了巨大进展。与传统手动干预着色不同，早期的一些基于深度学习方法（Cheng 等，2015；Zhang 等，2016；Antic，2024；Su 等，2020）使用卷积神经网络（convolutional neural network，CNN）预测每个像素的颜色分布，但是基于 CNN 的方法由于缺乏有效的语义理解导致了不合理的语义颜色。后来，研究者发现利用生成对抗网络（generative adversarial network，GAN）丰富的表示作为着色的生成先验，能够学习真实的颜色分布的特点。然而，基于 GAN 的方法（Wu 等，2021；Kim 等，2022；陈缘 等，2024）在生成细节丰富的图像时可能会出现细节模糊或不真实的问题，特别是在处理复杂场景和纹理时。

随着自然语言处理的巨大成功，Transformer 已经扩展到许多计算机视觉任务中。通过其卓越的长距离依赖关系建模能力和并行处理特性，显著提升了图像上色的效果。Kumar 等人（2021）首次将此架构应用于高分辨率图像上色任务，利用条件 Transformer 层和轴向自注意力机制，能够有效地建模全局像素关系。此外，还引入了辅助并行预测模型和多阶段生成策略。Ji 等人（2022）通过引入混合注意力机制，同时利用颜色记忆模块存储和检索常见颜色模式。其提出的整个 ColorFormer 框架采用端到端的 Transformer 架构，并通过联合优化策略实现着色任务，但在单尺度图像特征上执行颜色注意会导致颜色溢出。Xu 等人（2023）提出一种基于参考的端到端学习框架，通过融合多尺度颜色直方图，能够同时修复和上色退化的老照片。Kang 等人（2023）以端到端方式学习自适应语义感知的颜色嵌入，在此基础上，提出一种新的色彩损失函数，以提高生成结果的色彩丰富度。华夏等人（2024）提出将大窗口 Transformer 划分为不重叠小块，并对每块采样最大值点参与自注意力运算，以此提取局部特征，并在块内耦合局部与非局部信息，能够有效捕捉丰富特征。Liang 等人（2024）提出基于预训练稳定扩散（stable diffusion，SD）大模型（Rombach 等，2022）的可控上色方法，通过自注意力引导和学习内容引导的可变形自动编码器解决颜色溢出问题。但在复杂的图像或特殊场景下仍然存在一定的局限性，难以确保整个语义间有一致的颜色，需要进一步增强，且难以生成多样化的逼真图像。以上方法在颜色预测过程中由于无法保证色彩的一致性，缺乏对局部细节的精细处理，导致复杂背景和小物体的上色效果不佳。因此，实现完全自动化且合理自然的图像上色任务，仍需要进一步的研究和改进。

本文提出基于颜色查询和特征增强的图像上色方法，充分利用 Transformer 在处理颜色和细节方面的优势，旨在实现语义合理且视觉生动的自动着色。该方法采用编码器—解码器结构，其中 ConvNeXt（Liu 等，2022）编码器用于提取原始灰度图像的特征，解码器用于恢复空间分辨率。特别地，颜色预测网络包含颜色查询模块（color query block，CQB）和多尺度的像素增强模块（pixel enhancement block，PEB），以强调特定区域和颜色的匹配关系。在特征增强模块（feature enhancement block，FEB）的作用下，最终融合生成自然逼真的彩色图像。在公共基准数据集 ImageNet（val5k）（Deng 等，2009）、ImageNet（val50k）、COCO-Stuff（Caesar 等，2018）测试集、

ADE20K(Zhou等,2017)测试集、CelebA-HQ(Karras等,2018)测试集和真实老照片上验证了本文模型的性能。本文方法在语义一致性和色彩丰富度等方面取得了显著提升,在复杂背景和小物体的上色效果上尤为明显。

本文贡献概括如下:1)提出基于Transformer的图像上色网络,通过空间注意力引导,实现语义一致、色彩丰富的图像上色;2)提出特征增强模块,学习图像中不同感受野的特征,优化模型中颜色与原灰度图之间隐藏的映射关系,增强原图与生成图像之间的关联性,更加适用于处理环境复杂的图像;3)提出基于颜色查询和像素增强的颜色预测网络。颜色查询模块提供多尺度语义表示来指导颜色查询的优化,有效减少颜色溢出。像素增强模块旨在通过学习空间上的关注度,强调特定区域的像素,实现对图像像素的增强效果。

1 本文方法

1.1 网络结构

本文以端到端的方式对灰度图进行上色,图1是整个上色网络的架构图。首先,输入一幅灰度图像$\boldsymbol{x}^l \in \mathbf{R}^{H\times W\times 1}$,上色网络预测了两个消失的颜色通道$\boldsymbol{x}^{ab} \in \mathbf{R}^{H\times W\times 2}$,其中$l$和$ab$分别代表CIELAB色彩空间里的亮度和色度。本文构造了一个自动着色的编码器—解码器网络。

本文采用ConvNeXt作为编码器,充分利用其强大的特征提取能力和多尺度语义信息表达能力来提取灰度图像的高级语义信息。具体而言,输入灰度图$\boldsymbol{x}^l$,经过4个阶段输出4幅不同分辨率的中间特征图。每个阶段逐步减少特征图的空间尺寸并增加特征通道数,最终输出多尺度的特征图集合。前3幅特征图通过快捷连接传递到解码器,最后1幅特征图直接作为解码器的输入。解码器由4个上采样级组成,每个上采样级都与编码器的相应级别通过快捷连接相连,以此恢复图像特征的空间分辨率。

本文设计的颜色预测网络由一系列多尺度的颜色查询模块和像素增强模块构成,颜色查询模块利用不同尺度的多个图像特征,逐步改进语义感知的颜色向量。同时,像素增强模块通过空间注意力机制为特定区域提供了重要的权重,使得模型能够自动关注图像中的重要区域。因此,区域上的增强也可以理解为像素级的增强。融合模块是一种轻量级模块,将解码器的输出$\boldsymbol{E}_i \in \mathbf{R}^{C\times H\times W}$和颜色预测网络的输出$\boldsymbol{E}_c \in \mathbf{R}^{K\times C}$通过简单的点积,输出$\boldsymbol{y}^f \in \mathbf{R}^{C\times H\times W}$,其中$C$和$K$分别表示嵌入维数和颜色查询次数。具体过程为

$$\hat{\boldsymbol{F}} = \boldsymbol{E}_i \times \boldsymbol{E}_c \tag{1}$$

$$\boldsymbol{y}^f = conv(\hat{\boldsymbol{F}}) \tag{2}$$

$$\hat{\boldsymbol{y}}^{ab} = \boldsymbol{y}^l + \boldsymbol{y}^f \tag{3}$$

最后,通过特征增强模块将通道输出和原灰度

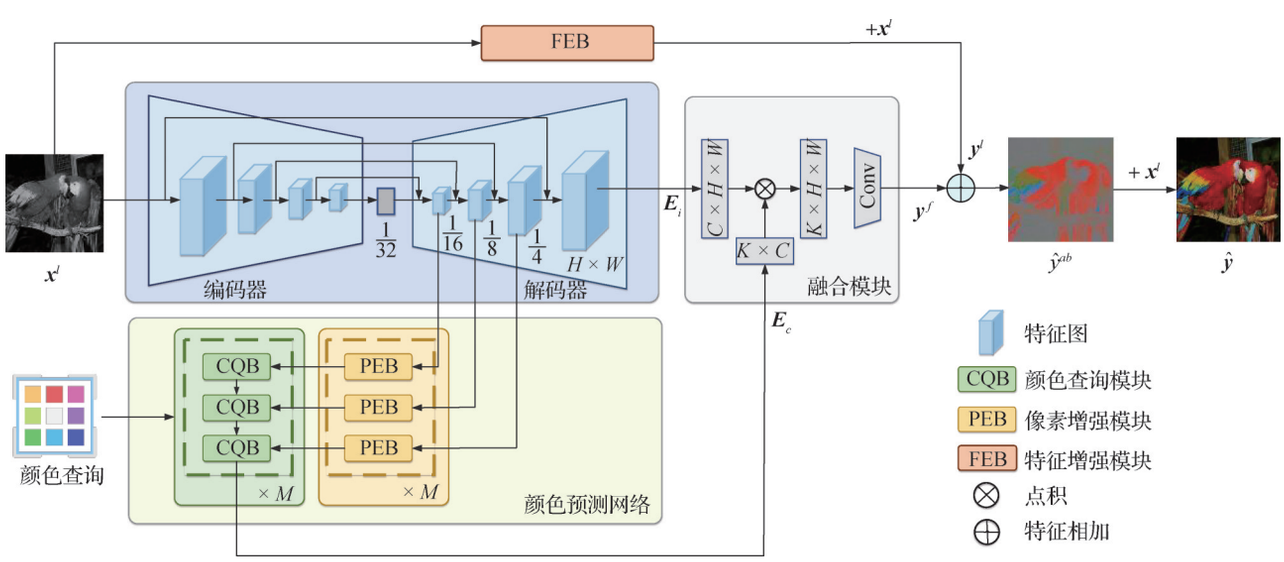


图1 融合颜色查询与特征增强的图像上色方法

Fig. 1 Image colorization via color query and feature enhancement

图连接起来输出 $\boldsymbol{y}^l$，与融合模块的输出 $\boldsymbol{y}^f$ 相加得到最终的颜色通道 $\hat{\boldsymbol{y}}^{ab}$。接下来详细描述这些模块。

1.2 颜色预测网络

编码器提取的多尺度特征图通过卷积连接层传递到解码器，并整合编码器对应级的特征，逐步恢复图像的空间分辨率。然后，这些多尺度特征进一步作为颜色预测网络的输入，经过像素增强模块的处理后，优化指导颜色查询。

1.2.1 像素增强模块

为了更好地强调特定区域与颜色的匹配，本文设计了像素增强模块(PEB)。由于现有方法通过使用预训练得到的块作为编码器，并直接采用卷积层(convolutional layer，Conv)作为解码器，模型在特征提取方面存在局限性。整个模型是端到端训练的，因此在原始灰度图与最终输出之间进行的是有监督的损失计算，即每一层网络隐式地学习颜色之间的映射关系。该模块在学习过程中重点关注图像的特定区域，为这些特定区域提供了重要的权重，有效提高了网络在颜色匹配方面的精度和效果。

如图2所示，像素增强模块从解码器得到的特征图是一个维度为 $H \times W \times C$ 的三维图像。首先，通过一个 1×1 的卷积层将其降维为 $H \times W$ 大小、1通道的图像。接着，经过两个卷积层进行映射处理，通过池化将图像按原比例缩小。然后，再通过两层卷积映射，将缩小的图像放大至原尺寸，经过反卷积操作后逐元素与原始图像相加，从而实现对像素的增强效果，实现图像细节的增强和恢复，达到提升图像质量的效果。由于池化的步长是可调节的，初始感受野也可以控制，这相当于在每个空间位置上设置一个固定大小的感受野，使网络能够学习到图像中核心的区域。通过强调关键部分的像素值，并将其加回到原始图像中，从而增强特定像素的表现力。在获得增强后的图像后，通过卷积层进行映射，并使用 sigmoid 激活函数生成一个取值范围在0～1之间的权重。最后，将该图像与原始图像逐元素相乘，实现空间上的增强。

在单尺度图像特征映射上执行颜色注意，并不能充分捕获低级语义线索，在处理复杂上下文时可能导致颜色溢出。本文选择了3种不同尺度的图像特征，在像素增强模块中，以不同倍率的下采样率生成的中间视觉特征将数据块分为每组3个PEB，在每组中，将多尺度特征按顺序馈送到PEB，以循环方式重复 M 次。本文像素增强模块共由 $3M$ 个PEB组成。

1.2.2 颜色查询模块

对于图像的自动上色，引入不同的颜色是结果多样性的关键。许多现有的上色方法依赖于额外的先验来获得生动的结果，这些方法需要大量的前期

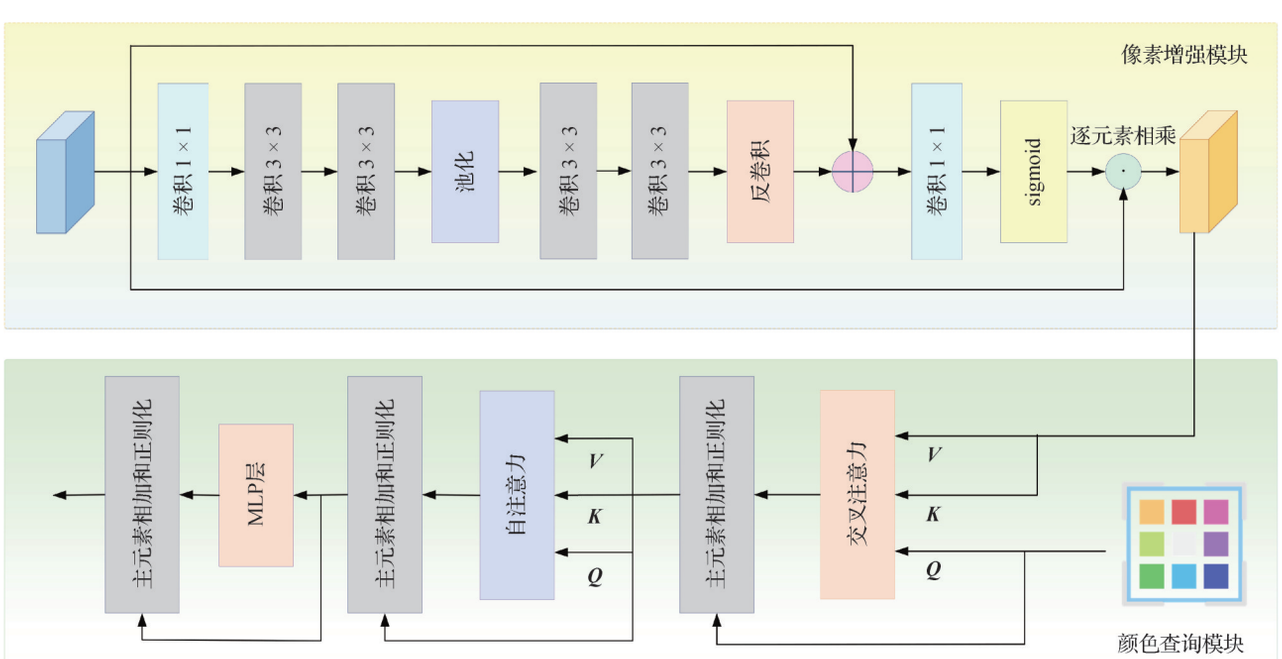


图2 颜色预测网络

Fig. 2 Color prediction network

投入工作，且在不同场景中的适用性可能有限。为了减少对人工设计先验的依赖，本文设计了一种基于查询的Transformer颜色查询模块（CQB），利用注意力机制来捕获全局相关性，这些特性可以提高着色性能。

为了学习一组基于视觉语义信息的自适应颜色查询，本文创建了可学习的颜色嵌入来存储颜色表示序列：$\boldsymbol{Z}_0 = \left[\boldsymbol{Z}_0^1, \boldsymbol{Z}_0^2, \cdots, \boldsymbol{Z}_0^K\right] \in \mathbf{R}^{K \times C}$。这些颜色嵌入在训练阶段初始化为零，并在第1个CQB中用做颜色查询。首先通过交叉注意层建立语义表示与颜色嵌入之间的相关性，具体为

$$\boldsymbol{Z}_l' = f_{\text{softmax}}(\boldsymbol{Q}_l \boldsymbol{K}_l^{\mathrm{T}})\boldsymbol{V}_l + \boldsymbol{Z}_{l-1} \tag{4}$$

式中，l为层索引，$\boldsymbol{Z}_l \in \mathbf{R}^{K \times C}$表示第$l$层的$K$个$C$维颜色嵌入。$\boldsymbol{Q}_l$、$\boldsymbol{K}_l$和$\boldsymbol{V}_l$分别表示径$f_\theta(\cdot)$、$f_K(\cdot)$和$f_V(\cdot)$线性变换后的图像特征，表示为

$$\begin{cases} \boldsymbol{Q}_l = f_Q(\boldsymbol{Z}_{l-1}) \in \mathbf{R}^{K \times C} \\ \boldsymbol{K}_l = f_K(\boldsymbol{Z}_{l-1}) \in \mathbf{R}^{H_l \times W_l \times C} \\ \boldsymbol{V}_l = f_V(\boldsymbol{Z}_{l-1}) \in \mathbf{R}^{H_l \times W_l \times C} \end{cases} \tag{5}$$

式中，H_l和W_l是图像特征的空间分辨率。通过上述交叉注意操作，颜色嵌入表示被图像特征所丰富。然后使用Transformer结构对颜色嵌入进行变换，具体为

$$\boldsymbol{Z}_l'' = MSA(LN(\boldsymbol{Z}_l')) + \boldsymbol{Z}_l' \tag{6}$$

$$\boldsymbol{Z}_l''' = MLP(LN(\boldsymbol{Z}_l'')) + \boldsymbol{Z}_l'' \tag{7}$$

$$\boldsymbol{Z}_l = LN(\boldsymbol{Z}_l''') \tag{8}$$

式中，$MSA(\cdot)$表示多头自注意（Vaswani等，2017），$MLP(\cdot)$表示多层感知器，$LN(\cdot)$表示层归一化（Ba等，2016）。值得关注的是，在颜色查询模块中，交叉注意先于自注意进行操作。这是因为在应用第1个自注意层之前，颜色查询是初始化为零的，且语义独立。同样，颜色查询块也是拓展到多尺度图像特征，使颜色嵌入对语义信息更加敏感，从而更准确地识别语义边界，减少颜色溢出。通过从多个尺度的特征中提取信息，模型能够捕捉到大尺度语义结构等图像全局信息，还能够捕捉边界区域和纹理等局部细节。这种多尺度的特征融合有助于精细化颜色边界的生成，减少颜色溢出。

本文将颜色预测网络表示为

$$\boldsymbol{E}_c = f_{\text{ColorQuery}}(\boldsymbol{Z}_0, \boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3) \tag{9}$$

式中，$\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3$是3个不同尺度的视觉特征。

1.3 特征增强模块

在图像上色任务中，许多方法仅仅将生成的颜色信息与原始灰度图进行简单结合，这些方法虽然在一定程度上能够生成彩色图像，但往往存在细节缺失和生成效果不自然等问题。针对以上缺陷，本文设计了特征增强模块（FEB），不仅能够学习颜色信息的映射关系，还能学习原始灰度图的隐空间，从而更全面地捕捉图像的特征。

在不同尺度的关注过程中，首先使用最原始的特征，通过两个不同大小的卷积核来学习特征。较小的卷积核感受野相对较小，更加局部。而较大的卷积核感受野更大，覆盖更广泛的区域。如图3所示，具体来说，通过不同的卷积核对输入张量进行处理，生成中间特征图$\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbf{R}^{C \times H \times W}$。通过不同感受野的卷积核，模型可以同时关注到图像中的细节信息和全局结构。

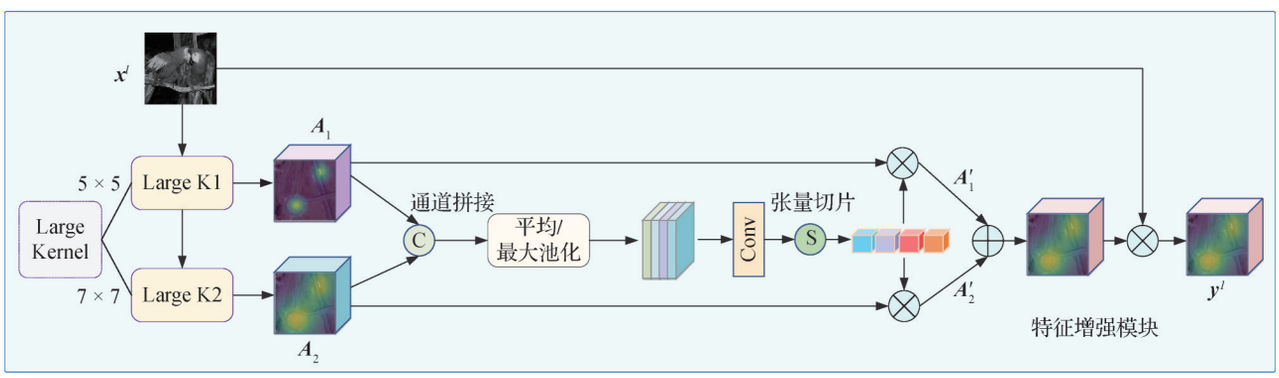


图3 特征增强模块

Fig. 3 Feature enhancement block

通过1 × 1卷积核对这些特征图进行通道压缩，得到特征图，通过平均池化和最大池化操作分别生成注意力图。这些操作分别提取了特征图的全局和局部信息，使模型能够生成更全面的特征表示。将这些注意力图拼接成聚合特征图，并通过一个7 × 7卷积层和sigmoid激活函数生成注意力权重，在这里

权重张量用Sig表示。

最后，将这些注意力权重分别与压缩后的特征图相乘，得到加权特征图。具体为

$$\boldsymbol{A}_1' = Conv_1(\boldsymbol{A}_1) \times Sig[:,0,:,:].unsqueeze(1) \tag{10}$$

$$\boldsymbol{A}_2' = Conv_2(\boldsymbol{A}_2) \times Sig[:,1,:,:].unsqueeze(1) \tag{11}$$

式中，$Conv_1(\cdot)$和$Conv_2(\cdot)$表示1×1的卷积层，$Sig[:,0,:,:].unsqueeze(1)$表示对张量$Sig$进行切片，并在指定维度上扩展维度。通过这种方式，注意力权重强调了特征图中的重要信息。接着，将这些加权特征图相加并通过卷积层恢复通道，得到整体的关注图，具体为

$$\boldsymbol{Attn} = Conv(\boldsymbol{A}_1' + \boldsymbol{A}_2') \tag{12}$$

这一过程确保了不同尺度的信息融合在一起，增强了特征图的表达能力。最终，将输入张量和关注图逐元素相乘，生成输出张量。

$$\boldsymbol{y}^l = \boldsymbol{x}^l \cdot \boldsymbol{Attn} \tag{13}$$

这种设计的优势在于可以有效地提升图像上色的效果，使得模型能够更好地理解和重构图像中的颜色和结构信息。

1.4 损失函数

在训练过程中，本文采用4种损失：像素损失提供了像素级监督，语义损失对齐语义特征，对抗损失保证了图像的真实性，颜色损失确保了颜色的多样性。

像素损失L_{pix}是彩色图像$\hat{\boldsymbol{y}}$与真实图像$\boldsymbol{y}$之间的L_1距离，它提供像素级监督，减少绝对差异来推动生成器生成与真实图像相似的输出，像素损失表示为

$$L_{pix} = \left\| \boldsymbol{y} - \hat{\boldsymbol{y}} \right\|_1 \tag{14}$$

为了衡量生成图像和真实图像在高层次特征上的差异，本文使用预训练的VGG16（Visual Geometry Group 16）（Simonyan和Zisserman，2014）网络来提取$\hat{\boldsymbol{y}}$和$\boldsymbol{y}$的深度特征，并计算它们之间的L_1距离，得出语义损失L_{per}，具体为

$$L_{per} = \sum_{l=1}^{5} w_l \left\| \Phi_l(\hat{\boldsymbol{y}}) - \Phi_l(\boldsymbol{y}) \right\|_1 \tag{15}$$

式中，$\Phi_l(\cdot)$为VGG16网络的层，w_l为对应层的权值。本文使用PatchGAN（Isola等，2017）鉴别器来区分预测结果和真实图像。具体来说，生成器对抗损失L_{adv}驱动生成器产生判断器难以区分的图像，判别器损失L_d则训练判别器准确区分真伪，二者可分别表示为

$$L_{adv} = \left\| 1 - D(\hat{\boldsymbol{y}}) \right\|_1 \tag{16}$$

$$L_d = \left\| 1 - D(\boldsymbol{y}) \right\|_1 + \left\| D(\hat{\boldsymbol{y}}) \right\|_1 \tag{17}$$

式中，D表示判别器，它的输出是输入数据为真实数据的概率。判别器损失L_d仅用于PatchGAN判别器自身的参数更新，不参与总损失的计算。

为了生成颜色更加饱满、对比更为鲜明的图像，本文使用的颜色损失L_{col}（Kang等，2023）为

$$L_{col} = \frac{1 - \left[\sigma_{rgyb}(\hat{\boldsymbol{y}}) + 0.3 \times \mu_{rgyb}(\hat{\boldsymbol{y}}) \right]}{100} \tag{18}$$

式中，$\sigma_{rgyb}(\cdot)$和$\mu_{rgyb}(\cdot)$分别为彩色平面像素云的标准差和平均值。最后，将上述所有损失函数组合成一个整体损失函数，在此损失函数下训练上色网络，表示为

$$L_\theta = \lambda_{pix} L_{pix} + \lambda_{per} L_{per} + \lambda_{adv} L_{adv} + \lambda_{col} L_{col} \tag{19}$$

2 实验结果及分析

2.1 实验设置

为证明本文算法的有效性和通用性，本文在ImageNet上训练上色模型，ImageNet已被大多数现有的着色方法广泛使用，本文采用的ImageNet训练集包含130万幅图像。本文的测试集包括5类，分别是由5 000幅图像构成的ImageNet（5 k）数据集，由50 000幅图像构成的ImageNet（50 k）数据集。另外，COCO-Stuff测试集由10 000幅图像构成，包含了多种复杂场景和丰富语义信息。ADE20K测试集是由以场景为中心的2 000幅图像组成，如室内、室外、城市和乡村等，涵盖了丰富的场景类型。CelebA-HQ测试集是CelebA的高质量人脸图像，共由30 000幅图像组成，分辨率可达1 024像素。

在上色模型评估中，本文使用3种关键的评价指标：Frechet初始距离（Frechet inception distance，FID）（Heusel等，2017）、色彩丰富度（colorfulness score，CF）（Hasler和Suesstrunk，2003）和峰值信噪比（peak signal-to-noise ratio，PSNR）（Huynh-Thu和Ghanbari，2008）。FID通过比较生成图像和真实图像在预训练InceptionV3（Szegedy等，2016）网络中的特征分布，衡量生成图像的逼真度和多样性，数值越低表示生成图像与真实图像越接近。CF评估生成图像的色彩饱和度和多样性，数值越高表示图像的色彩更加丰富。PSNR通过比较生成图像和真实图

像的像素差异，衡量图像的重建质量，数值越高表示图像质量越好。这些指标共同作用，全面评估图像上色模型在生成图像的逼真度、色彩质量和重建精度方面的性能。

本文使用预训练的ConvNext-L（Liu等，2022）作为编码器，并对该模型进行了微调，以更好地适应图像上色的需求。此外，配置多尺度解码器，输出2个通道。优化器使用AdamW（Loshchilov和Hutter，2019）训练网络，初始学习率为0.000 1，动量参数β_1和β_2分别设置为0.9和0.99，并通过多步学习率进行调整，权重衰减为0.01。对于颜色预测网络，本文设置循环重复次数M为3，查询次数K为100。整个网络以端到端自监督的方式进行20万次迭代训练，在第8万次迭代时，学习率减半，并在随后的每4万次迭代后再次减半。每次迭代的批量大小为4，数据加载使用8个线程。每1万次迭代进行一次验证，并在验证时评估FID、CF和PSNR等指标。对于损失函数，本文设置各个权重$\lambda_{\text{pix}}=0.1$，$\lambda_{\text{per}}=5.0$，$\lambda_{\text{adv}}=1.0$和$\lambda_{\text{col}}=0.5$。所有实验均在配备有4块Tesla A100 GPU的服务器上进行。

2.2 对比实验

为了说明本文方法的有效性，本文对比了5种方法，分别是DeOldify（Antic，2024）、Wu等人（2021）、BigColor（Kim等，2022）、ColorFormer（Ji等，2022）和DDColor（dual decoders color）（Kang等，2023）。

在ImageNet测试集（5k、50k）、COCO-Stuff测试集、ADE20K测试集和CelebA-HQ测试集上的可视化效果对比图表明，本文方法的效果更加生动自然。图4是在ImageNet测试集上的可视化结果对比，来自各测试集的综合测试结果如图5所示。具体来说，DeOldify模型在处理图像时，容易生成暗淡且饱和度不足的图像，例如图4的鹦鹉的颜色显得暗淡，图5中，卡车的色彩也缺乏生气。相比之下，Wu等人（2021）提出的方法和BigColor模型虽然基于GAN先验，但在实际应用中常常会出现色彩不统一甚至不合理的现象。例如，在图5中，这些方法在对水果进行上色时，生成了绿色的橘子和草莓，在男士的脖颈处也生成了不合理的鲜艳色彩，导致视觉效果失真。ColorFormer和DDColor模型在某些情况下也会出现不正确的着色结果，尤其是在具有复杂背景的场景中容易出现颜色溢出现象，例如在图4中的树枝红果、草坪灯以及图5中的卡车车尾，甚至在男士的眼白处也出现了蓝色的迹象。

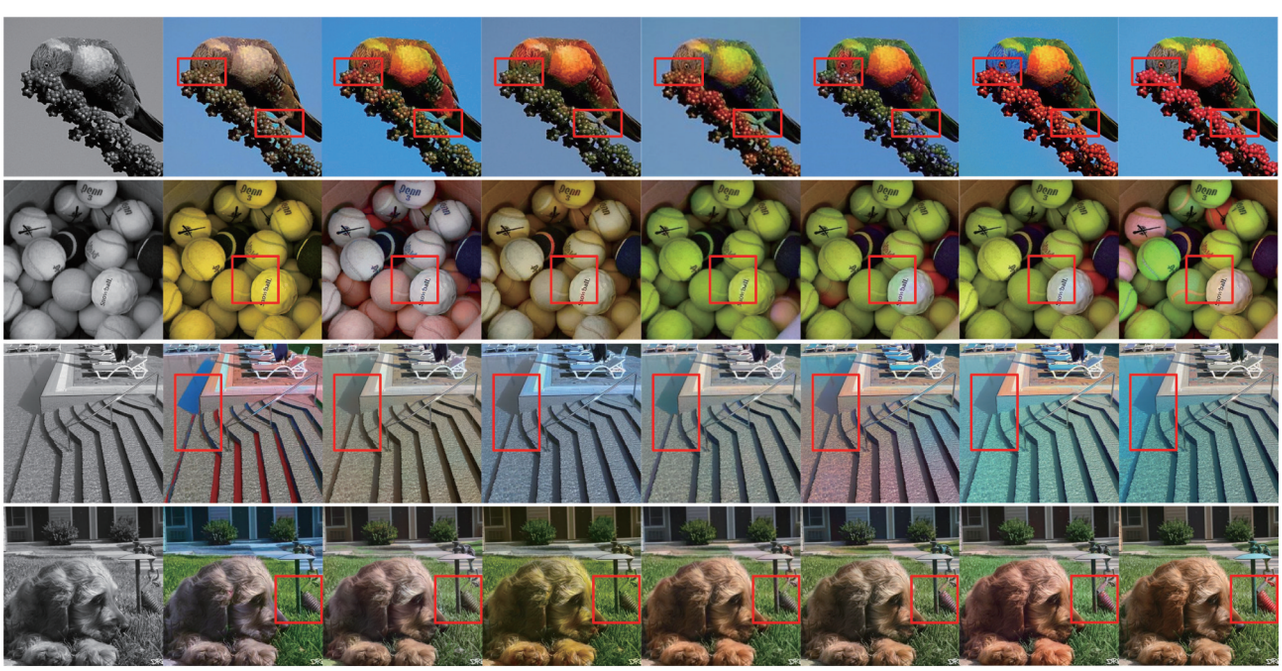
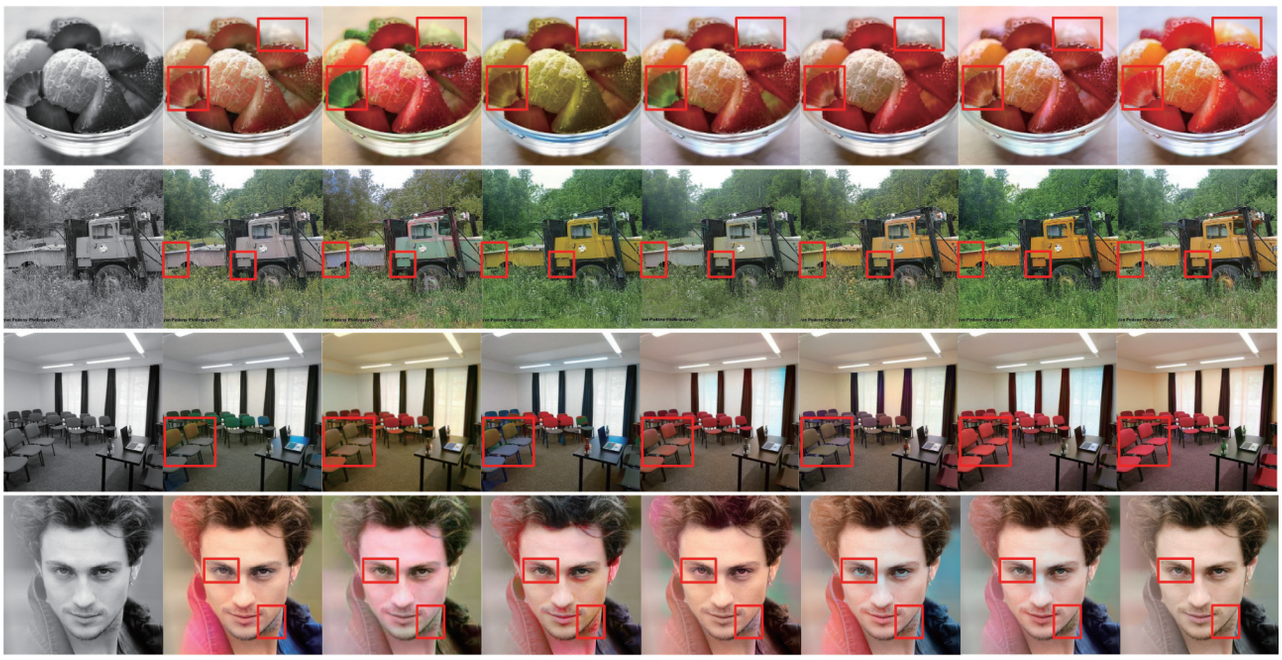


图4 在ImageNet测试集上的可视化结果对比

Fig. 4 Comparison of visualization results on ImageNet testset ((a) input images; (b) DeOldify; (c) Wu et al., (2021); (d) BigColor; (e) ColorFormer; (f) DDColor; (g) ours; (h) ground truth)



(a) 输入图像　(b) DeOldify　(c) Wu等人 (2021)　(d) BigColor　(e) ColorFormer　(f) DDColor　(g) 本文　(h) 真实图像

图 5　在 COCO-Stuff、ADE20K 和 CelebA-HQ 测试集上的可视化结果对比

Fig. 5　Comparison of visualization results on COCO-Stuff, ADE20K and CelebA-HQ testset((a) input images; (b) DeOldify; (c) Wu et al. ,(2021); (d) BigColor; (e) ColorFormer; (f) DDColor; (g) ours; (h) ground truths)

相较于上述方法，本文方法展现出明显优势。在图像语义理解上更加合理，生成的图像不仅视觉效果生动，而且成功捕捉到图像中突出的物体细节。例如，在图 4 中，生成的网球颜色丰富、饱和度高，生成的泳池图像与真实图像的视觉效果高度相似；在图 5 中，生成的水果色彩鲜艳透亮，真实感极强，人脸等细微细节处也表现最好。这表明，本文方法在多种测试集上均能生成视觉上接近原始图像且生动自然的结果，展现了在图像上色领域的显著改进和优越性能。

此外，本文收集了真实的黑白照片来验证本文模型的泛化性。真实老照片测试集来自多个公开的摄影和图像库，共 30 幅，分辨率范围非常广泛，从较低分辨率 256 像素到高分辨率 1 024 像素均有包含。真实老照片测试集覆盖了广泛的场景与内容，如风景、人物以及城市景观等(本文自建真实老照片测试集链接：https://www.alipan.com/s/vra5w7mEH1y)。在真实黑白老照片上色对比实验中，本文还引入预训练大模型作为对比方法，该方法结合 SD1. 5(预训练模型下载链接：https://huggingface.co/digiplay/majicMIX_realistic_v7/tree/main/majicMIX_realistic_v7.safetensors)的生成能力与面向上色任务的 ControlNet(Zhang 等，2023)微调的生成控制模型(预训练模型下载链接：https://huggingface.co/lllyasviel/sd_control_collection/resolve/main/ioclab_sd15_recolor. safetensors)。该方法(SD1. 5 + ControlNet)通过输入黑白图像和相应的控制信息，使模型能够生成高质量、符合特定需求的彩色图像，已在艺术创作、历史照片上色等多个场景得到广泛应用。

真实照片上色结果如图 6 所示，DeOldify 模型产生的第 4 张测试老照片“Mother of seven children”色彩对比度不足，缺乏鲜艳的颜色。Wu 等人(2021)提出的方法在上色结果中出现颜色溢出现象严重，例如在第 3 张测试老照片“Abandoned Boy”的背景中出现了大面积的绿色溢出，在第 1 张测试老照片“Louis Armstrong”的西装胳膊处，BigColor 模型出现了很明显的颜色不协调。ColorFormer 在第 3 行的衣角处有不合理的颜色，且色彩的饱和度较低，如 DDColor 的第 3 张测试老照片整体色调单一。SD+ ControlNet 模型生成的颜色整体缺乏多样性和层次感，图像整体显得沉闷和缺乏生气。本文方法在复杂场景中的效果最优，如第 2 张测试老照片“vintage-photography”里的鸭头颜色各不相同，鲜艳生动。

为了进一步证明本文方法的优越性，将此模型

图6　在真实老照片测试集上的可视化结果对比

Fig. 6　Comparison of visualization results on real old photos testset((a) input images; (b) DeOldify; (c) Wu et al. ,(2021); (d) BigColor; (e) ColorFormer; (f) DDColor; (g) SD+ControlNet; (h) ours)

与之前的图像自动上色方法在ImageNet(val5k)、ImageNet(val50k)、COCO-Stuff、ADE20K和CelebA-HQ 5个测试集上进行了定量对比，关于各模型在评价指标FID、CF、△CF和PSNR上的表现如表1和表2所示。本文方法在5个数据集上均获得了最低的FID，表明本文方法能够生成高质量的彩色图像，且具有较好的泛化能力。虽然先前的模型在色彩分数指标上高于本文方法，但是高的色彩评分并不一定意味着良好的视觉质量，于是进一步设计了评价指标△CF用于表示生成的彩色图像与真实图像之间的色彩评分差异，可以发现本文方法在数据集ImageNet(val50k)、COCO-Stuff、ADE20K和CelebA-HQ上均达到了最低的评分差异，表明本文方法在确保生成图像多样性的同时，实现了更自然真实的着色。

图像上色是一项具有高度主观性的任务，需要进行详细的用户调研以确保效果符合预期。通过深入了解用户需求，可以更好地提升图像上色的质量和用户满意度。实验选取30名受试者，每位受试者从ImageNet 1K验证集中随机挑选了100幅输入图像及其相应的着色图像。受试者的任务是从每组5幅

表1　在ImageNet测试集上的定量结果对比

Table 1　Quantitative comparison of results on the ImageNet testset

方法	参数量/M	运行效率/(帧/s)	ImageNet (val5k)				ImageNet (val50k)			
			FID	CF	△CF	PSNR/dB	FID	CF	△CF	PSNR/dB
DeOldify	63.6	5	6.59	21.29	16.92	<u>24.11</u>	3.87	22.83	16.26	22.97
Wu等人(2021)	310.9	9	5.95	32.98	5.23	21.68	3.62	35.13	3.96	21.81
BigColor	105.2	31	5.36	**39.74**	1.53	21.24	1.24	**40.01**	0.92	21.24
ColorFormer	44.8	40	4.91	38.00	0.21	23.10	1.71	<u>39.76</u>	0.67	23.00
DDColor	227.9	19	<u>3.92</u>	38.26	**0.05**	23.85	<u>0.96</u>	38.65	<u>0.44</u>	<u>23.74</u>
本文	297.5	14	**3.72**	<u>38.28</u>	<u>0.06</u>	**24.24**	**0.74**	38.98	**0.30**	**23.84**

注：加粗、下划线字体分别表示各列最优、次优结果。

表 2　在 COCO-Stuff、ADE20K 和 CelebA-HQ 测试集上的定量结果对比

Table 2　Quantitative comparison of results on the COCO-Stuff, ADE20K and CelebA-HQ testset

方法	COCO-Stuff				ADE20K				CelebA-HQ			
	FID	CF	△CF	PSNR/(dB)	FID	CF	△CF	PSNR/(dB)	FID	CF	△CF	PSNR/(dB)
DeOldify	13.86	24.99	13.25	**24.19**	12.41	17.98	17.06	24.40	9.48	43.93	1.18	25.20
Wu 等人(2021)	13.97	28.41	9.83	24.03	13.27	27.57	7.47	22.03	9.05	45.15	0.84	24.51
BigColor	12.58	36.43	1.81	21.51	11.23	**35.85**	0.81	21.33	7.48	**45.92**	0.46	24.14
ColorFormer	8.68	36.34	1.90	23.91	8.83	32.27	2.77	23.97	7.54	42.43	0.32	25.62
DDColor	5.18	**38.48**	0.24	22.85	8.21	34.80	0.24	24.13	6.45	42.06	0.29	26.94
本文	**4.76**	38.12	**0.21**	23.92	**7.15**	35.06	**0.20**	**24.94**	**5.76**	45.45	**0.22**	**27.61**

注：加粗、下划线字体分别表示各列最优、次优结果。

彩色图像中挑选出他们认为效果最佳的图像。结果如图 7 的箱线图所示。超过 30% 的用户首选本文方法，优于所有其他竞争方法（DDColor 22. 9%、ColorFormer17. 4%、BigColor 14. 7%、DeOldify 13. 5%）。

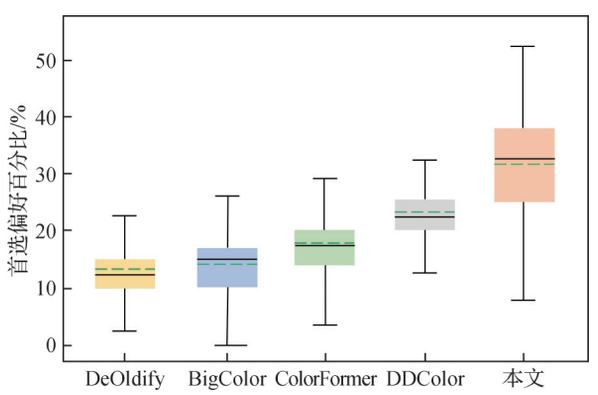


图 7　用户调研

Fig. 7　User research

2. 3　消融实验

为了进一步说明本文算法的有效性，本文从实验过程的角度说明网络为何能生成生动合理的图像。参考 DDColor（Kang 等，2023）的实验方式，进行了两组消融对比，一组是基于轻量级参数模型的实验（baseline-tiny），另一组是基于完整大规模参数模型的实验（baseline）。

每组消融实验均包括 3 种变体网络，第 1 种是不包括颜色预测网络和特征增强模块的模型变体，即整个网络结构仅包含主干和解码器。第 2 种是在变体基础上增强颜色查询模块的变体，第 3 种是包括颜色查询与特征增强模块的完整方法。

本文在 ImageNet 训练集上分别训练变体网络，表 3 和表 4 展示了不同变体网络在 ImageNet（val5k）测试集上的定量结果。两组不同参数规模的消融实验均表明，本文方法在上色任务上的有效性和优越性。由于从不同的语义特征信息中学习到颜色查询，本文所提出的颜色预测网络在 3 个指标上均有所提升。在像素增强模块的加持下，在最终的着色结果中起着重要作用。与基线相比，带颜色预测网络的方法可以实现更自然、语义更合理的图像上色。带特征增强模块的方法在指标 CF 和 △CF 上表现出众，更能帮助网络全面地捕捉和重构图像的颜色和结构信息。

为进一步讨论本文方法的有效性，本文对基于完整大规模参数模型的消融实验进行可视化展示，

表 3　消融实验（baseline-tiny）

Table 3　Ablation experiments（baseline-tiny）

方法	FID	CF	△CF	PSNR/dB
baseline-tiny	4.49	37.05	0.67	23.13
baseline-tiny + ColorQuery	4.43	37.74	0.53	23.51
baseline-tiny + ColorQuery + FEB	**4.32**	**38.28**	**0.49**	**23.84**

注：加粗字体表示各列最优结果。

表 4　消融实验（baseline）

Table 4　Ablation experiments（baseline）

方法	FID	CF	△CF	PSNR/dB
baseline	3.92	38.26	0.05	23.85
baseline + ColorQuery	3.88	38.27	**0.05**	24.11
baseline + ColorQuery + FEB	**3.72**	**38.28**	0.06	**24.24**

注：加粗字体表示各列最优结果。

如图8所示，baseline的效果往往在语义一致性方面效果不佳，例如消防栓、网球的整体颜色不一致，屋顶未能成功上色，屋顶上方出现黄色阴影。在baseline中结合颜色预测网络（+ ColorQuery），整体的颜色分布更加一致，例如消防栓的上方颜色显示更为准确，屋顶也成功上色为一致的红色，引入颜色预测网络更能帮助网络学习到最核心的区域，有助于提高最终结果的色彩，例如红色屋顶颜色更为突出，特别是矮层房屋以及后方房屋边缘则更加明显清晰，天空乌云更加栩栩如生。进一步，加入特征增强模块（+ ColorQuery + FEB），网络更能捕捉到细节信息和全局结构，使得图像的视觉吸引力更高。从特征可视化的效果热图中，可以更直观地看出模型在同一层次的特征空间中如何聚焦于图像的特定部分，特别是在物体边界、细节和语义区域的上色过程中。例如，网球的颜色分布更加均匀，边界更加清晰，消防栓的整体颜色也呈现出一致的鲜红色。虽对于边缘不明显的红色屋顶，增强效果不明显，但又进一步增强了矮层房屋以及后方房屋边缘和细节。由此可以得出，本文的完整模型捕获了更准确的语义边界识别特征，产生了更自然、更准确的着色结果。

为了验证循环重复次数M和查询次数K对上色模型性能的影响，本文分别进行了消融实验，以独立评估这些参数的选择对模型效果的贡献。实验结果如图9所示，随着M的增加，模型在一定程度上能够捕捉到更多的特征信息。然而，当M超过3时，性能提升有所下降。此外，性能在100次查询时达到峰值，当查询数量持续增加时，性能不再提升。本文最终模型使用了3次循环重复和100次查询，因为这种设置可以实现最佳性能。

从实验设定的损失函数的参数来看，语义损失的权重比较大，为了进一步分析语义损失权重的影响，本文设计了一个消融实验。在该实验中，固定像素损失、对抗损失和颜色损失的权重，逐渐调整语义损失的权重，观察其对最终上色效果的影响。实验结果如图10所示，较大的语义损失权重有助于提升图像的全局结构和内容的一致性，但当语义损失权重大幅增加时，局部细节和颜色过渡的自然性可能会受到影响。因此，本文设置语义损失的权重为5.0时，才能获得最佳的生成效果。

2.4 局限性

本文方法在大多数情况下表现优异，但仍存在

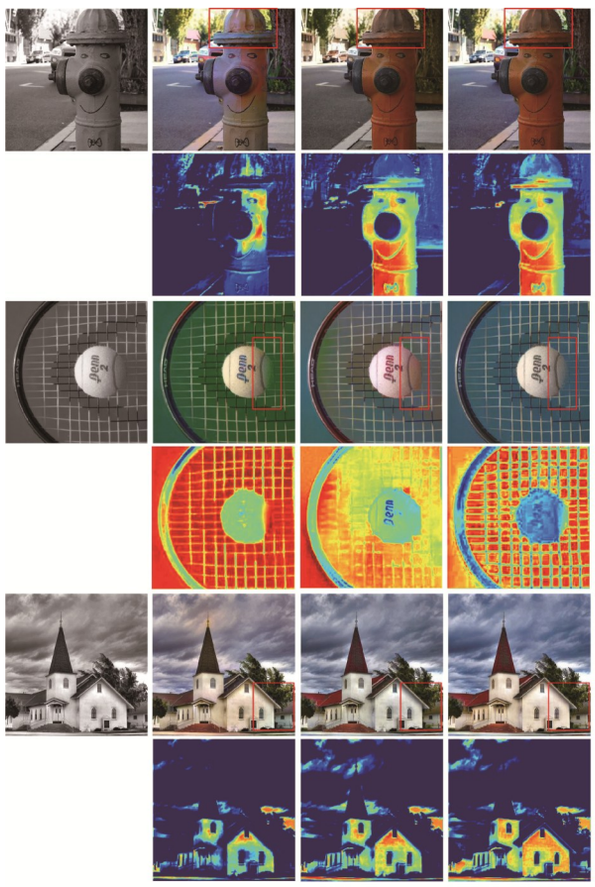


图8 消融实验可视化结果对比

Fig. 8 Comparison of visualization results of ablation experiment ((a) input images; (b) baseline; (c) baseline + ColorQuery; (d) baseline + ColorQuery + FEB)

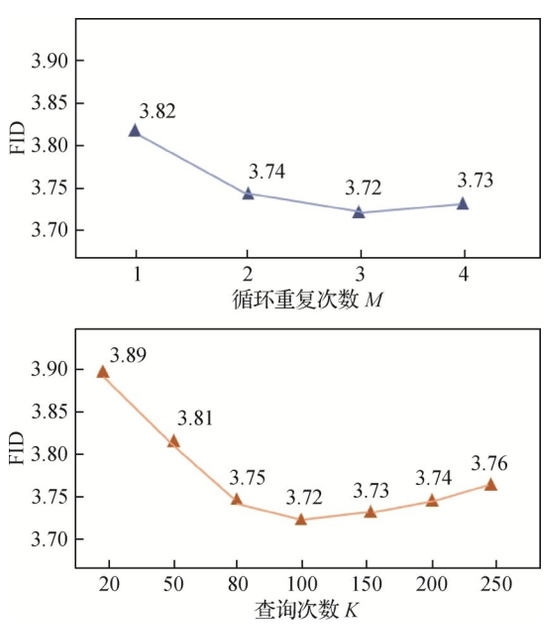


图9 不同循环次数和查询次数对模型性能的影响

Fig. 9 Impact of different cycle time and query time on model performance

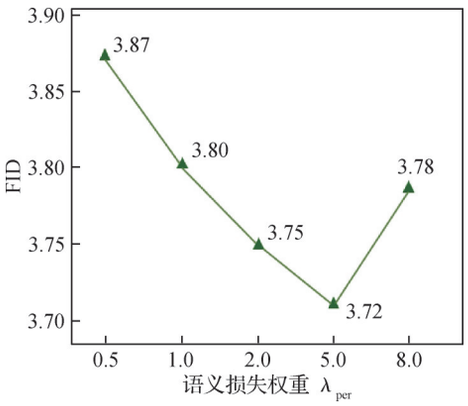


图 10　语义损失权重对模型性能的影响

Fig. 10　Impact of semantic loss weight on model performance

一些例外情况。如图 11 所示，本文方法并未能准确识别出杯子的把手和杯盖对应的语义信息，在处理包含透明物体的图像时，仍然存在一定的不足之处。为进一步改进，可能需要引入额外的语义监督，以帮助网络更好地理解此类场景。此外，与大多数自动着色方法相似，本文方法在生成的颜色控制或用户指导方面仍显不足。未来的研究可以在着色过程中引入更多用户交互方式，如文本提示、颜色涂鸦和参考图像等，以增强用户对最终结果的控制。

(a) 输入图像　(b) 本文方法　(c) 真实图像

图 11　失败案例

Fig. 11　Failure case

((a) input image; (b) ours; (c) ground truth)

3　结　论

现有的许多上色方法在处理复杂场景和细节丰富的图像时，面临着色彩还原不准确和细节丢失等问题，且方法较为烦琐且难以泛化。针对这一问题，本文提出一种端到端图像上色方法。主要贡献在于提出特征增强模块，优化颜色与原灰度图之间的映射关系，增强图像关联性，适用于复杂环境的图像；颜色预测网络通过学习空间关注度，重点关注特定区域，并提供多尺度语义表示，减少颜色溢出。同时，基于查询的颜色预测保证了色彩的多样性。综合实验表明，在测试集 ImageNet(val5k)、ImageNet(val50k)、COCO-Stuff、ADE20K 和 CelebA-HQ 中，与现有方法相比，本文方法在颜色丰富度和美观度方面达到了更好的性能。在本研究中，虽然已经取得了较好的上色效果，但仍有一些未能深入探讨的方面。例如，本文方法主要关注图像的颜色恢复，未来可以进一步研究如何结合深度学习中的风格迁移技术，使上色结果不仅在颜色上准确，还能在风格上保持多样性与一致性。此外，如何提升模型的计算效率，尤其是在高分辨率图像上的应用，也将是未来的研究重点。

参考文献(References)

Antic J. 2024. DeOldify: a deep learning based project for colorizing and restoring old images (and video!) [EB/OL]. [2024-08-17]. https://github.com/jantic/DeOldify

Ba J L, Kiros J R and Hinton G E. 2016. Layer normalization [EB/OL]. [2024-08-17]. https://arxiv.org/pdf/1607.06450.pdf

Caesar H, Uijlings J and Ferrari V. 2018. COCO-Stuff: thing and stuff classes in context//Proceedings of 2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1209-1218 [DOI: 10.1109/CVPR.2018.00132]

Chen Y, Zhao Y, Zhang X J and Liu X P. 2024. Sketch colorization with finite color space prior. Journal of Image and Graphics, 29(4): 978-988 (陈缘, 赵洋, 张效娟, 刘晓平. 2024. 有限色彩空间下的线稿上色. 中国图象图形学报, 29(4): 978-988) [DOI: 10.11834/jig.230189]

Cheng Z Z, Yang Q X and Sheng B. 2015. Deep colorization//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 415-423 [DOI: 10.1109/ICCV.2015.55]

Chia A Y S, Zhuo S J, Gupta R K, Tai Y W, Cho S Y, Tan P and Lin S. 2011. Semantic colorization with internet images. ACM Transactions on Graphics (TOG), 30(6): 1-8 [DOI: 10.1145/2070781.2024190]

Deng J, Dong W, Socher R, Li L J, Kai L and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Hasler D and Suesstrunk S E. 2003. Measuring colorfulness in natural images. Proceedings of SPIE 5007, Human Vision and Electronic Imaging VIII. Santa Clara, USA: SPIE: 87-95 [DOI: 10.1117/12.477378]

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a

local Nash equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6629-6640 [DOI: 10.5555/3295222.3295408]

Hua X, Shu T, Li M X, Shi Y and Hong H Y. 2024. Nonlocal feature representation-embedded blurred image restoration. Journal of Image and Graphics, 29(10): 3033-3046 (华夏, 舒婷, 李明欣, 时愈, 洪汉玉. 2024. 融合非局部特征表示的模糊图像复原. 中国图象图形学报, 29(10): 3033-3046) [DOI: 10.11834/jig.230735]

Huynh-Thu Q and Ghanbari M. 2008. Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 44(13): 800-801 [DOI: 10.1049/el:20080522]

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Ji X Z, Jiang B Y, Luo D H, Tao G P, Chu W Q, Xie Z F, Wang C J and Tai Y. 2022. ColorFormer: image colorization via color memory assisted hybrid-attention transformer//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 20-36 [DOI: 10.1007/978-3-031-19787-1_2]

Kang X Y, Yang T, Ouyang W Q, Ren P R, Li L Z and Xie X S. 2023. DDColor: towards photo-realistic image colorization via dual decoders//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 328-338 [DOI: 10.1109/ICCV51070.2023.00037]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada

Kim G, Kang K, Kim S, Lee H, Kim S, Kim J, Baek S H and Cho S. 2022. BigColor: colorization using a generative color prior for natural images//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 350-366 [DOI: 10.1007/978-3-031-20071-7_21]

Kumar M, Weissenborn D and Kalchbrenner N. 2021. Colorization transformer//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Liang Z X, Li Z C, Zhou S C, Li C Y and Loy C C. 2024. Control color: multimodal diffusion-based interactive image colorization [EB/OL]. [2024-08-17]. https://arxiv.org/pdf/2402.10855.pdf

Liu X P, Wan L, Qu Y G, Wong T T, Lin S, Leung C S and Heng P A. 2008. Intrinsic colorization. ACM Transactions on Graphics (TOG), 27(5): #152 [DOI: 10.1145/1409060.1409105]

Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T and Xie S N. 2022. A ConvNet for the 2020s//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11966-11976 [DOI: 10.1109/CVPR52688.2022.01167]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization//Proceedings of the 9th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Qu Y G, Wong T T and Heng P A. 2006. Manga colorization. ACM Transactions on Graphics (TOG), 25(3): 1214-1220 [DOI: 10.1145/1141911.1142017]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Su J W, Chu H K and Huang J B. 2020. Instance-aware image colorization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7965-7974 [DOI: 10.1109/CVPR42600.2020.00799]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc: 6000-6010. [DOI: 10.5555/3295222.3295349]

Wu Y Z, Wang X T, Li Y, Zhang H L, Zhao X and Shan Y. 2021. Towards vivid and diverse image colorization with generative color prior//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14357-14366 [DOI: 10.1109/ICCV48922.2021.01411]

Xu R S, Tu Z Z, Du Y Q, Dong X Y, Li J L, Meng Z B, Ma J Q, Bovik A and Yu H K. 2023. Pik-fix: restoring and colorizing old photos//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1724-1734 [DOI: 10.1109/WACV56688.2023.00177]

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhang R, Isola P and Efros A A. 2016. Colorful image colorization//Proceedings of the 14th European Conference on Computer Vision ——

ECCV 2016. Amsterdam, the Netherlands: Springer: 649-666 [DOI: 10.1007/978-3-319-46487-9_40]

Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A and Torralba A. 2017. Scene parsing through ADE20K dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5122-5130 [DOI: 10.1109/CVPR.2017.544]

作者简介

于冰，男，讲师，主要研究方向为图形图像处理与深度学习。E-mail: yubing@shu.edu.cn

相雪，女，硕士研究生，主要研究方向为图像修复与深度学习。E-mail: 22723302@shu.edu.cn

范正辉，男，硕士研究生，主要研究方向为图像增强与深度学习。E-mail: zhenghui_fan@shu.edu.cn

黄东晋，男，副教授，主要研究方向为虚拟现实与计算机图形学。E-mail: djhuang@shu.edu.cn

丁友东，男，教授，主要研究方向为计算机图形学与数字影视技术。E-mail: ydding@shu.edu.cn