软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(9):3719−3738 [doi: 10.13328/j.cnki.jos.007534] [CSTR: 32375.14.jos.007534]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于难样本挖掘的灰度图像着色模型评估*

鄢杰斌[1,2], 彭振团[1,2], 祝文涛[1,2], 蔡 超[1,2], 毛阿敏[1,2], 方玉明[1,2]

[1](江西财经大学 计算机与人工智能学院, 江西 南昌 330032)

[2](多媒体智能处理江西省重点实验室 (江西财经大学), 江西 南昌 330013)

通信作者: 方玉明, E-mail: leo.fangyuming@foxmail.com

摘 要: 随着深度学习和计算机视觉的快速发展, 灰度图像着色研究已从传统手工特征设计转向数据驱动的深度神经网络范式. 然而, 现有的灰度图像着色模型评估体系面临双重挑战: 其一, 由于评价指标的局限性以及着色任务的高度病态性本质, 传统评价指标 (如 PSNR、SSIM 和 FID 等) 难以准确量化着色模型性能; 其二, 开展大规模主观实验进行定性分析耗时费力且可行性差. 针对上述问题, 提出了基于难样本挖掘的灰度图像着色模型评估方法. 该方法旨在通过多维度差异化 (包括图像质量、美学表现和颜色差异) 比较, 高效地挖掘用于比较着色模型的代表性样本; 随后开展可控小规模主观实验, 可靠地比较不同模型的性能, 并指出不同模型的优势和不足. 实验结果表明: 提出的方法能够高效、准确地找到模型的难样本, 在极大幅度地减小主观实验规模的同时, 揭示模型的优缺点, 为灰度图像着色模型评估提供了新范式, 并为模型优化指明方向.

关键词: 计算机视觉; 图像着色; 模型评估; 难样本挖掘; 模型优化

中图法分类号: TP391

中文引用格式: 鄢杰斌, 彭振团, 祝文涛, 蔡超, 毛阿敏, 方玉明. 基于难样本挖掘的灰度图像着色模型评估. 软件学报, 2026, 37(9): 3719–3738. http://www.jos.org.cn/1000-9825/7534.htm

英文引用格式: Yan JB, Peng ZT, Zhu WT, Cai C, Mao AM, Fang YM. Evaluation of Grayscale Image Colorization Models by Exposing Hard Examples. Ruan Jian Xue Bao/Journal of Software, 2026, 37(9): 3719–3738 (in Chinese). http://www.jos.org.cn/1000-9825/7534.htm

Evaluation of Grayscale Image Colorization Models by Exposing Hard Examples

YAN Jie-Bin[1,2], PENG Zhen-Tuan[1,2], ZHU Wen-Tao[1,2], CAI Chao[1,2], MAO A-Min[1,2], FANG Yu-Ming[1,2]

[1](School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330032, China)

[2](Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing (Jiangxi University of Finance and Economics), Nanchang 330013, China)

Abstract: With the rapid advancement of deep learning and computer vision, grayscale image colorization has evolved from traditional handcrafted feature-based methods to data-driven deep neural network paradigms. However, existing evaluation systems for grayscale image colorization models face the following two challenges: First, due to the limitations of evaluation metrics and the highly ill-posed nature of the colorization task, traditional quantitative metrics such as PSNR, SSIM, and FID cannot effectively quantify the performance of grayscale image colorization models. Second, it is time-consuming, laborious, and infeasible to conduct qualitative analyses through large-scale subjective experiments. To address these issues, a new evaluation method for grayscale image colorization models based on hard sample mining is proposed. The method aims to efficiently identify representative samples for model comparison through multi-dimensional evaluation (including image quality, aesthetics epression, and color difference), and then conduct a controlled small-scale subjective experiment to reliably compare different models. Subsequently, the advantages and shortcomings of the models are revealed. Experimental results show that the proposed method can efficiently and accurately find hard samples, and reveal the strengths and weaknesses of the models while drastically reducing the scale of subjective experiments, providing a new paradigm for grayscale image

* 基金项目: 国家自然科学基金 (U24A20220, 62461028); 江西省自然科学基金 (20243BCE51139, 20252BAC240227); 江西省教育厅科学技术研究项目 (GJJ210509); 江西财经大学研究课题 (2024112516274306O)
收稿时间: 2025-03-31; 修改时间: 2025-06-04, 2025-08-18; 采用时间: 2025-08-28; jos 在线出版时间: 2025-12-24
CNKI 网络首发时间: 2025-12-25

colorization model evaluation and indicating the direction for model optimization.

Key words: computer vision; image colorization; model evaluation; hard sample exposing; model optimization

灰度图像着色作为计算机视觉领域中的一项重要技术, 旨在将灰度图像转化为具有自然色彩的彩色图像. Zhong 等人[1]提出了一种基于灰度图像引导的红外图像着色方法, 通过迁移可见光域的结构信息实现伪彩色合成. 传统的灰度图像着色方法[2,3]往往依赖于手工设计的特征和先验知识, 通过规则或模型推测图像的颜色; 然而, 这些方法的着色效果并不可观, 且鲁棒性较差, 难以在多样化的场景中保持稳定表现. 近年来, 基于深度学习的灰度图像着色方法[4−6]逐渐成为主流, 通过训练卷积神经网络[7]或生成对抗网络[8]等模型, 能够学习丰富的图像特征和颜色分布规律, 从而生成更为精准且自然的色彩. 虽然这些基于深度学习的模型取得了显著的性能提升, 但它们在复杂场景下仍然存在色彩预测准确性不足、图像质量差以及美感低等问题, 在不同测试样本上的表现仍然存在较大差异. 总体而言, 现有的灰度图像着色模型评估方法存在以下缺陷. 一方面, 当前灰度图像着色模型往往依赖于定量指标, 如峰值信噪比 (peak signal-to-noise ratio, PSNR)[9]、结构相似性方法 (structural similarity index, SSIM)[10]、弗雷歇初始距离 (Fréchet inception distance, FID)[11]、学习感知图像块相似度 (learned perceptual image patch similarity, LPIPS)[12]和颜色鲜艳度分数 (colorfulness score, CF)[13]等传统指标, 使用这些指标评估灰度图像着色模型存在如下一些问题.

(1) PSNR 和 SSIM 方法不考虑颜色信息, 其度量结果无法反映着色任务中的色彩还原程度, 不适用于灰度图像着色模型的评估.

(2) 尽管 FID 和 LPIPS 等指标能综合度量生成图像与真实样本在特征空间的相似性, 但由于灰度图像着色任务固有的高度病态性 (即灰度图着色存在多种合理结果), 这些指标与着色性能的相关性较弱. 指标值高, 可提示潜在质量缺陷, 但无法直接判断绝对质量优劣. 因此, 可将此类指标定位成辅助性评估工具.

(3) CF 值过低表示着色模型性能不足, 而过高 CF 值往往对应着色过度饱和现象, 导致视觉真实性下降. 因此, 该指标与着色质量间的关联性较弱, 仅适用于参考性评估.

另一方面, 使用大规模主观实验对模型进行评价耗时费力且可行性差, 因此灰度图像着色模型[14−17]通常在验证集中随机抽取少量图像进行主观评估. 此做法存在显著局限性: 首先, 随机采样具有很大不确定性, 受样本选择的影响较大; 其次, 样本量过小不足以支撑评估结论. 因此, 评估结果可能不准确.

为解决上述问题, 我们提出基于难样本挖掘的灰度图像着色模型评估方法. 该方法通过引入组最大化差异竞争机制, 自动挖掘出难样本集. 其核心思想是通过衡量灰度图像着色模型在生成图像的图像质量、颜色差异和美学表现这 3 个维度上的综合性能差异, 自动找出能够揭示各个模型优缺点的潜在难样本, 通过在难样本集上进行小规模主观实验, 实现对灰度图像着色模型性能的全面准确评估.

本文的主要贡献包括如下.

(1) 提出一种灰度图像着色模型评估方法, 解决了由传统评价指标局限性以及着色任务本身的高度病态性本质而导致的评估结果不确定性问题.

(2) 构建了一个面向灰度图像着色模型评估的难样本集, 可用于模型的评估和改进.

(3) 提出的评估方法和难样本集的可扩展性非常好, 可补充或直接测试新的比较模型 (相关模型)、增加难样本的规模等.

本文第 1 节对模型评估和最大差异化竞争进行详尽概述, 第 2 节介绍基于组最大差异化竞争评估方法的框架设计, 第 3 节介绍实验结果和分析, 第 4 节介绍跨模型一致性验证, 最后进行总结.

1 相关工作

本节, 首先对模型评估的评估方法进行概述, 然后对最大差异化竞争方法及应用进行阐述.

1.1 模型评估

随着深度学习技术的快速发展, 评估方法也在不断演进, 涵盖了从传统的量化指标到新兴技术的多个方面. 在

图像质量评价 (image quality assessment, IQA) 模型的评估中, 常用皮尔逊相关系数 (Pearson linear correlation coefficient, PLCC) 和斯皮尔曼秩相关系数 (Spearman's rank correlation coefficient, SROCC) 来表示 IQA 模型[18−22]的性能. 在目标检测领域, 交并比 (intersection over union, IoU) 是目标检测中常用的一个指标, 取值范围在 0–1 之间, 值越大表示模型的性能越好. 除了各个领域的传统评价指标, 许多新的指标也在被提出, 以更准确地评估模型性能. 例如, Cordts 等人[23]使用新的量化指标用于评估实例级语义标注的准确性, 考虑了物体实例的大小, 更公平地评估了预测结果. Gu 等人[24]指出传统 IQA 方法 (如 PSNR、SSIM) 在评估基于 GAN 的图像修复算法时存在局限性, 因为它们无法有效区分 GAN 生成的纹理和噪声. 为了解决这个问题, 作者提出了空间翘曲差分网络 (space warping difference network, SWDN), 能够更有效地捕捉图像特征, 并提高对空间错位的鲁棒性. 然而, 仅依靠整体性能指标来评估模型存在局限性, 需要更深入地分析模型的行为和性能, 从不同的角度理解模型的优缺点. 例如, Nekrasov 等人[25]把语义分割模型的错误分为不同的类型, 根据错误类型和来源, 探索改进模型的方法, 例如, 针对小目标检测不准确的问题, 可以尝试使用生成对抗网络来学习大型目标的特征, 并将其迁移到小目标上. Zhang 等人[26]通过研究不同特征对模型性能的影响, 例如特征通道、滤波器的类型等, 可以帮助模型选择更合适的特征表示. Hoiem 等人[27]通过分析模型作用好坏的案例, 可以识别出模型不足的方面. Alwassel 等人[28]提出了用于诊断时间动作检测器错误的工具, 可以帮助研究人员了解模型的失败模式, 并针对性地改进算法. Li 等人[29]指出将图像去雾与其他计算机视觉高级任务相结合可以从另一个角度评估去雾模型的性能, 并为实际应用提供更可靠的解决方案. Borji[30]指出应该识别和减少模型在学习过程中引入的偏差, 提高模型的可靠性和可信度. Chen 等人[31]对不同类型的模型 (卷积神经网络和 Transformer) 进行了比较, 发现 Transformer 模型在损坏数据集上表现更优, 这提示我们需要根据不同的任务需求选择合适的模型类型. 李学龙等人[32]指出了深度学习方法的不可解释性问题, 并强调了知识的方向指导和数据的优化驱动; 在评估模型性能时, 需要关注模型的可解释性, 并尝试将专业知识融入模型中, 以提高模型的可靠性和可信赖性. Liu 等人[33]通过使用新的 4K 分辨率进行模型的评估, 为评估不同方法的性能提供了更好的基础. Li 等人[34]构建了一个多用途图像去雨基准数据集, 涵盖了雨迹、雨滴和雨雾等多种雨型, 并进行了语义标注, 为全面评估去雨算法提供了可靠的平台. Hendrycks 等人[35]提出了两个新的基准数据集, 关注模型对重建图像腐蚀和图像扰动的鲁棒性, 拓展了鲁棒性评估的维度, 更全面地评估模型的泛化能力. 侯淇彬等人[36]利用互联网图像训练语义分割模型, 大大降低了数据获取的成本, 并拓展了模型的适用范围.

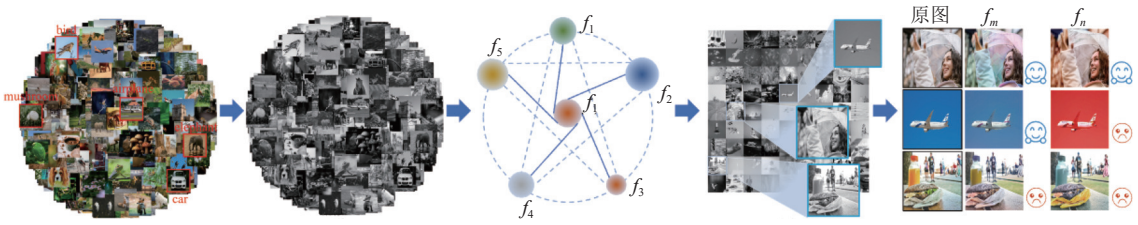
1.2 最大差异化竞争

传统的模型评估方法通常依赖大规模的数据集, 通过计算模型在这些数据集上的量化指标来表示模型的性能优劣. 这些传统方法通常也需要大量的主观测试, 但是, 收集足够的主观数据来覆盖整个空间几乎是不可能的, 即使进行了大量的主观测试, 也无法充分验证模型的准确性. 为了解决该问题, Wang 等人[37]提出了最大差异化竞争 (maximum differentiation competition, MAD). 传统的图像分类模型比较方法使用小规模、固定的测试集来比较模型, 这可能导致模型过拟合和泛化能力差, 针对该问题, Wang 等人[38]通过使用 MAD 方法, 从大规模未标注图像集中自适应地采样测试集, 发现 VGG 网络比残差网络更具泛化能力, 也可以更有效地发现模型的弱点, 并识别改进模型的方法. Wang 等人[39]通过运用 MAD 来比较模型在开放世界图像上的性能, 证明了他们的模型在开放世界图像上取得了更好的性能. MAD 方法虽然能有效比较两个模型, 但存在一些局限性, 比如需要梯度信息、难以扩展到多个模型和难以约束生成的样本落在特定领域等. 因此, Ma 等人[40]提出了组最大差异化竞争 (group maximum differentiation competition, gMAD), 能够同时比较多个模型, 通过攻击和防御的博弈来最大化模型的差异化, 从而在少量样本下有效比较模型性能. 具体而言, 该方法的核心思想是: 首先, 每个模型轮流作为"防御者", 其他模型作为"攻击者"; 然后, 攻击者在防御者预测值相近的样本区间内, 寻找预测差异最大的样本对; 最后, 基于实测数据计算模型的攻击性 (攻击其他模型的能力) 和抵抗性 (抵御攻击的能力), 最终通过聚合统计实现全局排序. Wang 等人[41]使用 gMAD 思想寻找对模型改进最有帮助的图像对, 并进行人工标注, 然后利用这些标注数据进行微调, 通过 gMAD 比较基线模型与一组全参考 IQA 方法, 寻找能够解释基线模型弱点的强反例, 将人工标注的数据与现有数据库进行合并, 从而对基线模型进行迭代微调, 从而不断提高模型的泛化能力.

在选择灰度图像着色模型的评估方法时, 需考虑图像着色任务本身的高度病态特性, 其解空间充满高度的不确定性和多样性, 导致传统的量化指标往往难以全面反映模型的性能优劣. 因此, 结合主观视觉感知的方法成为衡量着色结果质量的重要手段. 然而, 面对大规模数据集, 完全依赖人工进行主观评估不仅耗时耗力, 而且成本高昂. 为了解决这一问题, 我们引入了组最大差异化竞争的方法. 该方法通过比较多个灰度图像着色模型的输出结果, 筛选出模型间差异显著的潜在难样本从而大幅缩小需要主观评估的图像数据. 在此基础上, 我们对筛选出的难样本集进行主观实验评估, 既保证了评估结果的代表性, 又显著降低了评估成本, 为图像着色模型的性能评估提供了一种高效且可靠的解决方案. 该方法通过 3 个维度挖掘出能够区分模型的边界案例, 从而减少样本数量, 同时大幅度地减小了主观实验评估的规模, 在提高模型比较效率的同时, 提高了图像着色模型评估的准确性.

2 基于组最大差异化竞争评估方法框架设计

本节首先对使用组最大差异化竞争方法进行模型评估要解决的问题进行问题表述, 然后对基于组最大差异化竞争评估方法的主要思想进行详细介绍, 最后对多个模型的比较进行阐述, 本文提出的框架如图 1 所示.



(a) 彩色图像集 (b) 灰度图像集 (c) 着色模型两两对比 (d) 潜在难样本灰度图像集 (e) 着色结果

图 1 基于组最大差异化竞争评估方法的框架

图 1(a) 为包含 154 个类别的彩色图像集, 涵盖了多样化的场景和语义内容, 红色字体为红框中的图像类别; (b) 为由 (a) 转化来的灰度图像集, 作为模型评估的输入数据; (c) 表示参与评估的 L 个图像着色模型需要两两进行对比; (d) 为通过组最大差异化竞争方法挑选出的潜在难样本灰度图像集; (e) 展示了两个模型对潜在难样本图像集中灰度图像的着色结果的 3 种情况: 第 1 列为 (a) 中的原始彩色图; 第 1 行展示了两个着色模型均生成高视觉质量的着色结果示例; 第 2 行展示了其中一个着色模型结果显著优于另一着色模型示例; 第 3 行则展示了两个着色模型均生成低视觉质量的着色结果示例.

2.1 问题表述

图像着色任务是典型的不确定性病态问题, 其核心挑战源于解空间的不确定性, 即存在多种视觉合理的着色方案. 由于着色结果具有多样性, 并且着色评价具有主观性, 同一张灰度图像可能会生成多种不同的彩色版本, 这些结果可能都是合理的, 也可能都无法完全满足预期, 或者可能在特定方面表现优异而在其他方面存在不足. 传统量化评估指标, 如 PSNR、SSIM、FID、LPIPS 和 CF 等, 在模型性能比较中被广泛使用, 但由于评价指标的局限性以及着色任务的高度病态性本质, 传统评价指标难以准确比较灰度图像着色模型的性能, 也无法全面捕捉着色结果在图像质量、颜色差异和美学表现等方面的综合表现. 因此, 我们以主观感知作为评判着色模型优劣的依据. 对于一个灰度图像数据集 $G=\{g^{(n)}\}_{n=1}^{N}$, 包含 $N(100, 185)$ 张灰度图, 这些图像涵盖了现实世界中的各类场景, 若通过人工方式对每个着色模型生成的彩色图像进行逐一评估, 在实际操作中几乎不可行; 而从验证集中随机选取少量样本又不具合理性, 样本少且缺乏代表性. 为解决这一难题, 我们引入组最大差异化竞争思想来优化评估流程.

2.2 主要思想

本文评估方法的核心思想是通过对比灰度图像着色模型 (简称为着色模型) 的输出结果, 筛选出能够显著暴露模型差异的图像子集, 从而在保证评估效果的同时大幅降低人工评估的工作量. 我们从图像质量、颜色差异以及美学表现这 3 个维度综合比较着色模型 $F=\{f_l\}_{l=1}^{L}$, 其中包含 L 个着色模型, 旨在找出在各维度分布边缘且结果差异比较大所对应的灰度图像集 O. 这些图像往往能够更直观地反映各模型的优缺点, 同时大概率是对模型具有

挑战性的难样本, 通过这种方式, 可以高效地识别出最具代表性的图像子集, 进而集中资源对这些图像进行人工主观评估. 这种方法不仅能够显著减少评估工作量, 还能更精准地揭示模型在实际应用中的性能差异, 为模型优化提供更有价值的参考依据.

我们以两个着色模型 f_1 和 f_2 为例, 阐述模型比较的方法. 通过模型 f_1 和 f_2 对灰度图像 $g\in G$ 进行着色, 我们挑选出 top-K 对综合差异 (图像质量、颜色差异以及美学表现) 最大的彩色图像所对应的原始灰度图, 得到两个模型的潜在难样本集 O^*. 用数学公式表示如下:

$$O_k^* = \arg\max_{O_k\in G}(D(f_1(g),f_2(g)) + |Q(f_1(g)) - Q(f_2(g))| + \lambda\times|A(f_1(g)) - A(f_2(g))|) \tag{1}$$

其中, $k\in\{1,2,3,\ldots,K\}$, O_k^* 表示潜在难样本集 O^* 中使得着色差异第 k 大的灰度图, $D(\cdot)$ 为度量图像色差的模型, 计算两个着色模型对一张灰度图处理后的两张彩色图之间的色差; $Q(\cdot)$ 为一种无参考的 IQA 模型, 计算图像的质量得分; $A(\cdot)$ 表示美学评估模型, 度量图像的整体美感; λ 表示美学差异的权重, 经实验验证, 我们将 λ 设为 20, 作为实验有效值. 通过公式 (1), 我们可得到 K 对差异最大化的彩色图像, 这些图像对存在如下 3 种情况.

情况 1: 两个着色模型均生成高视觉质量的着色结果, 如图 1(e) 第 1 行所示. 这种情况在实际应用中较为常见, 充分体现了图像着色任务的不确定性本质. 由于着色问题本身是一个病态问题, 同一张灰度图像可能存在多种合理的颜色分布方案. 例如, 天空可以被渲染为深蓝色或浅蓝色, 树叶可以是翠绿色或黄绿色, 这些差异并不影响图像的整体自然度和视觉吸引力. 这种情况表明, 某些图像对颜色分布的容忍度较高, 模型在着色时具有较大的灵活性.

情况 2: 其中一个着色模型结果显著优于另一个着色模型, 如图 1(e) 第 2 行所示. 这种情况是我们希望重点关注的, 因为这类图像能够清晰揭示不同模型在着色能力上的差异. 例如, 一个模型可能在渲染自然景观时表现出色, 而另一个模型在处理人造物体时颜色更加准确. 通过分析这些差异显著的图像对, 我们可以识别出每个模型的优势和不足, 从而为模型评估提供更具针对性的样例. 这类图像不仅有助于理解模型的性能特点, 还能为后续的模型优化提供明确的方向.

情况 3: 两个着色模型均生成低视觉质量的着色结果, 如图 1(e) 第 3 行所示. 这种情况表明, 所选的灰度图像对这两个模型来说都是具有挑战性的难样本. 例如, 某些图像可能包含复杂的纹理、光照条件或罕见的物体, 导致模型难以准确预测颜色分布. 通过识别这些难样本, 我们可以深入分析模型的短板, 例如颜色偏差、细节丢失或纹理失真等问题. 这些发现为着色模型的改进提供了宝贵的参考, 例如可以通过增加训练数据的多样性或改进网络架构来提升模型在复杂场景下的表现.

通过将上述筛选出的图像用于主观实验来对着色模型的性能进行比较, 我们能够在保证评估效果的同时显著降低评估成本.

2.3 多个模型比较

基于上述两个模型的比较思想, 本文进一步将这一方法扩展到 L 个着色模型进行比较. 具体而言, 我们对这 L 个模型进行两两比较, 共比较 $J=C_L^2$ 次, 每对模型通过组最大差异化方法筛选出一组潜在难样本集 $O^j, j\in\{1,2,3,\ldots,J\}$, 这些潜在难样本集反映了每对模型在着色性能上的显著差异, 能够有效暴露模型在特定场景或复杂条件下的局限性. 通过对所有模型对进行比较, 初步得到 $O^T=\cup O^j, j\in\{1,2,3,\ldots,J\}$, 共 $J\times K$ 张灰度图像. 再对 O^T 进行去重处理, 最终得到一个潜在难样本集 $O=\{g^{(m)}\}_{m=1}^{M}(M\ll N\&M\leqslant J\times K)$. 我们总结了难样本挖掘的伪代码描述 (算法 1). 得到潜在难样本集 O 后, 我们可以通过主观实验对 L 个着色模型进行综合评估. 通过各模型给该样本集着色后, 邀请评估者对每组结果进行主观排序. 这种基于主观实验的评估方法不仅能够准确反映模型在实际应用中的表现, 还能捕捉传统量化指标无法衡量的着色合理性和美学表现. 例如, 评估者可以从颜色自然度、颜色真实性、整体协调性等多个维度对模型进行评分, 最终根据综合得分对模型进行排名.

算法 1. 难样本集的挖掘.

输入: 灰度图像数据集 G, 着色模型 $F=\{f_l\}_{l=1}^{L}$, 色差度量模型 $D(\cdot)$, 无参考 IQA 模型 $Q(\cdot)$, 美学评估模型 $A(\cdot)$;
输出: 难样本集 O.

1. $O^T \leftarrow \varnothing$, $j \leftarrow 1$
2. **for** $l \leftarrow 1$ **to** L **do**
3. 生成着色结果 $\{f_l(g)|g \in G\}$
4. **end**
5. **for** $l \leftarrow 1$ **to** L **do**
6. **for** $i \leftarrow l+1$ **to** L **do**
7. //计算模型着色结果的颜色差异
   $\{D(f_l(g), f_i(g))|g \in G\}$
8. //计算模型着色结果的质量评分
   $\{Q(f_l(g), Q(f_i(g))|g \in G\}$
9. //计算模型着色结果的美学评分
   $\{A(f_l(g), A(f_i(g))|g \in G\}$
10. 利用公式 (1) 筛选出着色综合差异最大的 K 张灰度图形成图像集 O^j
11. $O^T \leftarrow O^T \cup O^j$
12. $j \leftarrow j+1$
13. **end**
14. **end**
15. 对 O^T 进行去重处理得到难样本集 O

我们的模型评估方法的优势在于其高效性、全面性以及良好的可扩展性. 通过聚焦潜在难样本集, 我们能够在保证评估效果的同时, 显著节省人力、物力和财力资源. 例如, 相比于对全部数据集进行评估, 这种方法可以将评估工作量减少一个数量级, 同时确保评估结果的代表性和可靠性. 此外, 该方法能够更精准地揭示不同模型的优点和缺点, 为后续的着色模型设计和改进提供重要的参考依据. 例如, 针对模型在特定场景下的不足, 可以通过增加相关训练数据或优化网络结构来提升性能; 而对于表现优异的模型, 可以进一步分析其成功经验, 为其他模型的设计提供借鉴. 并且, 我们能够在不增加额外工作量的情况下, 轻松地将若干个需要评估的模型添加到已有的评估模型组中 (如算法 2 所示).

算法 2. 加入一个新的着色模型.

输入: 灰度图像数据集 G, 着色模型 $F = \{f_l\}_{l=1}^{L}$, 色差度量模型 $D(\cdot)$, 无参考 IQA 模型 $Q(\cdot)$, 美学评估模型 $A(\cdot)$, 新的着色模型 f_{L+1}, 更新前的难样本集 O;
输出: 更新后的难样本集 O'.

1. $j \leftarrow J+1$
2. 生成着色结果 $\{f_{L+1}(g)|g \in G\}$
3. **for** $l \leftarrow 1$ **to** L **do**
4. //计算模型着色结果的颜色差异
   $\{D(f_l(g), f_{L+1}(g))|g \in G\}$
5. //计算模型着色结果的质量评分
   $\{Q(f_l(g), Q(f_{L+1}(g))|g \in G\}$
6. //计算模型着色结果的美学评分
   $\{A(f_l(g), A(f_{L+1}(g))|g \in G\}$
7. 利用公式 (1) 筛选出着色综合差异最大的 K 张灰度图形成图像集 O^j

8. $O \leftarrow O \cup O^j$
9. $j \leftarrow j+1$
10. **end**
11. 对 O 进行去重处理得到更新后的难样本集 O'

3 实 验

本节基于第 2.3 节描述的框架, 对 8 个着色模型进行了全面的性能评估. 其中, 5 个模型为无需参考图像的自动着色模型, 另外 3 个模型为需要参考图像的引导着色模型. 评估所使用的数据集为一个包含 100 185 张图像的大规模数据集, 涵盖多样化的场景和内容, 以确保评估结果的广泛性和代表性. 最后进行消融实验, 包括: 对依赖于参考图像的模型进行探究; 对参考图像的选择进行分析; 对难样本集的特征进行分析; 对难样本挖掘选择的指标以及指标权重系数进行分析.

3.1 实验设置

3.1.1 图像数据集 G 的构建

我们基于 50 万张图像的数据集来对图像数据集 G 进行构建, 用于着色模型的评估. 首先, 我们对原始 50 万张图像进行了类别筛选. 由于部分类别 (如“fly”类) 包含语义范围过于广泛 (例如同时包含飞机、鸟类等不同语义的图像) 或样式过于繁杂的图像, 这些类别不利于后续的模型评估. 因此, 我们从中筛选出 154 个语义明确且样式相对统一的类别, 作为后续处理的基础. 接下来, 对这 154 个类别中的每一类进行随机筛选. 通过随机删除大部分图像, 我们初步筛选出约 30 万张图像. 随后, 对这 30 万张图像进行了更严格的筛选, 剔除了大量不符合要求的图像, 包括但不限于以下情况; (1) 图像内容与所属类别不符; (2) 图像为卡通或非真实场景; (3) 图像质量较差 (如模糊、噪声过多); (4) 图像中目标物体占比较小, 难以清晰识别.

经过严格筛选, 得到了 100 185 张较高质量的彩色图像. 为获得着色任务中具有参考价值的图像, 我们主观筛选出各类别中语义关联性最强且内容最具代表性的图像作为参考图像. 具体选择标准为: (1) 目标类别物体在图像中的视觉占比不低于 50%; (2) 排除卡通等非真实风格图像; (3) 最大程度覆盖数据集中该类别的典型视觉特征. 最后将这 100 185 张彩色图像转换为灰度图像, 构建最终的图像数据集 G. 该数据集的图像大小均为 512×512, 不仅涵盖多样化的场景和语义类别, 还确保了图像的质量和语义一致性, 为着色模型性能评估提供了可靠基础.

3.1.2 评估模型 F 的构建

我们对具有代表性的 8 个着色模型进行了系统性评估. 这些模型可分为两类: 一类是不依赖参考图像的自动着色模型, 另一类是需要指定参考图像的引导着色模型. 具体来说, 自动着色模型包括在 ImageNet 数据集上训练的 DDColor[14]、ColorFormer[15]、InstColor[17]、CT^2[16], 以及在 ImageNet ILSVRC 2012 数据集上训练的 UniColor[42]; 引导着色模型则包括在 ImageNet 数据集上训练的 PDNLA[43]、在 COCO 数据集上训练的 SAC[44], 以及在 ImageNet ILSVRC 2012 数据集上训练的 AC[45].

在第 3.1.1 节中, 我们提到参考图像的选取基于语义相关性, 但这种相关性并非绝对严格, 而是相对意义上的. 这是因为数据集 G 中的每个类别可能包含数千张图像, 这些图像在语义内容和结构上存在显著差异. 例如, 同一大类下的不同子类别可能涵盖多样化的场景或对象, 导致图像之间的语义关联性较弱. 此外, 即使在同一类别中, 图像也可能因拍摄角度、光照条件或背景复杂度的不同而呈现出较大差异. 因此, 尽管我们尽可能选择最具代表性的参考图像, 但仍无法保证这些图像在所有情况下都能为着色模型提供理想的指导.

在实验设置方面, 我们使用灰度图集 G 中原始分辨率的图像 (512×512) 作为模型的输入, 且未对图像进行其他预处理. 这种做法旨在保留图像的原始细节和结构信息, 避免预处理操作引入额外的失真或信息损失, 从而更真实地反映模型在实际应用中的性能. 通过这种方式, 我们能够更准确地评估模型的着色能力和鲁棒性, 同时确保实验结果的可靠性和可重复性.

3.1.3 潜在难样本集 O 的构建

我们基于第 2.3 节所描述的方法构建差异最大化图像集 O, 利用公式 (1) 计算模型预测结果之间的差异, 并通过设定参数 K=10, 从图像集 G 中筛选出每对模型结果差异最大的前 10 张灰度图像. 这一设计旨在平衡主观实验的成本与图像集的规模, 确保评估效率的同时保留足够的代表性样本. 在模型比较过程中, 我们对 L=8 个着色模型进行两两比较, 共需进行 $J=C_8^2=28$ 轮比较. 每轮比较生成 10 张潜在难样本灰度图像, 因此初步得到 $J\times K=280$ 张灰度图像. 由于不同模型对之间可能存在部分重叠的难样本, 我们对这 280 张图像进行去重处理, 最终得到 211 张不同的灰度图像, 构成潜在难样本集 O. 该图像集的构建具有以下优势.

(1) 通过聚焦于差异显著的图像, 大幅减少了需要评估的图像数量, 降低了主观实验的成本.

(2) 所选的 211 张图像集中反映了各模型在着色任务中的性能差异, 能够有效暴露模型的优缺点.

(3) 覆盖了 8 个模型之间的所有两两比较, 确保评估结果的广泛性和可靠性.

接下来, 我们将基于潜在难样本集 O 对这 8 个着色模型进行深入评估. 这一方法不仅节省了资源, 还为模型性能的全面分析和优化提供了高质量的数据支持.

3.1.4 主观实验

我们基于潜在难样本集 O 开展主观实验, 以评估着色模型 F 的性能. 具体实验流程如下.

对于图像集 O 中的每一张灰度图像, 我们通过模型 F 生成对应的彩色结果图, 每组包含 8 张结果图 (分别由 8 个模型生成), 最终得到 211 组彩色图像. 为了确保实验的科学性和一致性, 我们邀请了 12 名计算机视觉方向的研究生参与评估, 包括 9 位男生和 3 位女生, 并且都进行过 5–12 次不等的 IQA 相关主观实验, 具有丰富的图像评估经验. 受试者通过我们在 PyCharm 平台上开发的图形用户界面 (如图 2 所示) 完成实验.

图 2 第 1 行最左侧显示原始彩色图像, 作为参考基准; 第 1 行第 2 幅图像为灰度图着色时使用的参考图; 第 1 行后 3 幅图像为需要参考图的模型生成的着色结果. 第 2 行图像则为无需参考图的模型生成的着色结果.

图 2 模型评估图形用户界面

我们将 12 位受试者分为 4 组, 每组受试者分工完成 211 组彩色图像的评估, 对每组中的 8 张结果图进行排名. 排名规则为: 第 1 名得 8 分, 第 2 名得 7 分, 以此类推, 第 8 名得 1 分. 完成所有图像的排名后, 每个模型对应图像的平均得分即为该组受试者在图像集 O 上的评估分数. 在实验开始前, 我们对受试者进行了简单的培训, 介绍了图像质量 (如颜色自然度、细节保留度)、颜色丰富度和图像美学 (如色彩协调性、视觉吸引力) 的评判标准, 以确保评分的一致性和客观性. 为了减少实验误差, 所有受试者均在同一显示设备 (经过颜色校正处理) 上进行评估,

以避免因设备差异导致的视觉偏差. 此外, 为了保证受试者在评估过程中保持专注, 减少因疲劳或分心导致的误判, 我们允许受试者在评估过程中根据需要适当休息.

在所有受试者完成评估后, 我们对每位受试者的评分结果进行汇总, 分别计算每个模型的得分. 这一分数反映了着色模型在颜色生成合理性、细节保留和美学表现等方面的综合性能排名. 通过这种系统化的评估方法, 我们能够更准确地揭示各模型的优势与不足, 为后续模型优化提供重要参考, 结果如表 1 所示. 为了进一步验证主观实验的可靠性和一致性, 我们计算了每组图像主观评分之间的 SROCC 和 PLCC. 结果如表 2 所示, SROCC 范围为 [0.87, 0.98], PLCC 范围为 [0.88, 0.98], 证明了该主观实验结果的可靠性.

表 1　主观实验模型分数和排名

受试者	DDColor[14]	ColorFormer[15]	InstColor[17]	CT^2[16]	UniColor[42]	PDNLA[43]	SAC[44]	AC[45]
第1组	6.18/1	6.07/2	4.32/4	5.38/3	3.96/5	2.94/8	3.63/6	3.54/7
第2组	5.62/2	5.99/1	4.93/4	4.99/3	4.06/5	3.34/8	3.59/6	3.49/7
第3组	5.75/2	6.05/1	4.99/3	4.18/4	3.73/7	3.27/8	4.08/5	3.96/6
第4组	5.48/2	5.75/1	4.76/4	4.74/3	3.90/6	3.65/7	4.09/5	3.63/8

注: /左边的值表示模型在该组的总得分, /右边的值表示模型在该组评分中的排名

表 2　各组分数间的皮尔逊相关系数和斯皮尔曼秩相关系数

受试者	第1组	第2组	第3组	第4组
第1组	—	0.96	0.87	0.95
第2组	0.98	—	0.91	0.98
第3组	0.88	0.90	—	0.95
第4组	0.90	0.93	0.93	—

注: —下方区域为皮尔逊相关系数, —上方区域为斯皮尔曼秩相关系数

接下来, 基于表 1 的主观实验结果以及各模型的着色结果进行分析.

针对基于参考图像的着色方法 (PDNLA[43], SAC[44]和 AC[45]), 其性能受到参考图像的严重制约. 此类模型需依赖用户提供的语义强关联参考图, 但在实际场景中, 找到与输入灰度图高度匹配的参考图极具挑战性. 如图 3 第 1 行所示, 参考图语义偏差导致生成结果色彩单一且丰富度显著低于其他模型; 而图 3 第 2 行进一步表明, 当参考图包含误导性内容 (大面积蓝色天空) 时, 模型倾向于引入与场景语义冲突的异常颜色. 此类局限性使得上述模型在主观实验中评分垫底, 验证了其对参考图的强依赖缺陷.

相比之下, 无需参考图像的自监督模型展现出显著优势. ColorFormer[15]通过全局-局部混合注意力机制增强长距离依赖建模能力, 结合颜色记忆模块提供多样化颜色先验, 有效减少语义错误并提升色彩生动性. 其训练策略融合像素级重建损失 (L1)、语义感知损失 (使用 VGG16 抽取特征) 与对抗损失, 从多层级约束生成质量, 生成符合人类视觉感知的真实图像着色结果. DDColor[14]则采用双解码器架构, 像素解码器专注于细节恢复, 而颜色解码器通过交叉注意力和自注意力机制从多尺度特征中提取语义感知的颜色查询, 增强语义边界的识别能力, 显著减少颜色渗色现象, 辅以颜色丰富度损失函数优化色彩饱和度与审美性. 如图 3 第 3、4 行所示 (黄色方框标出的为渗色区域), 老虎的皮毛变成与周围绿叶相同的颜色, 人脸也被周围环境所影响变成了绿色, 道路的颜色也并不一致. 这两个模型在色彩合理性、语义一致性及颜色丰富度这 3 个方面均显著优于其他模型, 主观评分都在 5 分以上, 其中 ColorFormer 接近 6 分, 也验证了其设计的有效性.

其余模型在特定场景下存在明显缺陷: UniColor[42]为多模态模型, 为了保证评估的公平性, 我们评估该模型的无条件模式, 在无条件模式下因缺乏控制信号而产生语义失配问题. 此外, 其训练集 (ImageNet ILSVRC 2012) 规模相比前几个模型的训练集 (ImageNet) 较小, 导致色彩预测偏差, 如图 3 第 5–7 行中出现的反直觉颜色, 如图 3 第 5 行中异常的背景颜色, 图 3 第 6 行中绿色的蜜蜂, 图 3 第 7 行奇怪的天空与云朵颜色等, 严重偏离实际. CT^2[16]将着色建模为分类任务, 通过 313 个颜色令牌约束输出空间, 虽能提升饱和度, 但离散化颜色空间导致细节丢失.

如图 4, 图像相对模糊, 其亮度选择模块虽抑制了异常颜色, 却无法覆盖自然场景的连续色彩分布, 在颜色的自然性方面也不如人意. InstColor[17]虽通过目标检测实现实例级着色以减少背景干扰, 但其性能高度依赖检测精度. 漏检或误检将导致模型退化为全局着色, 引发颜色失真或特征融合失效 (如局部颜色突变), 图 5 展示了该模型语义检测失败时的普遍失败样例, 普遍出现了与周围颜色不相融的区域性异常颜色.

(a) 灰度图 (b) 参考图像 (c) 原始图 (d) ColorFormer (e) DDColor (f) CT^2 (g) InstColor (h) UniColor (i) PDNLA (j) SAC (k) AC

图 3 模型着色结果视觉比较

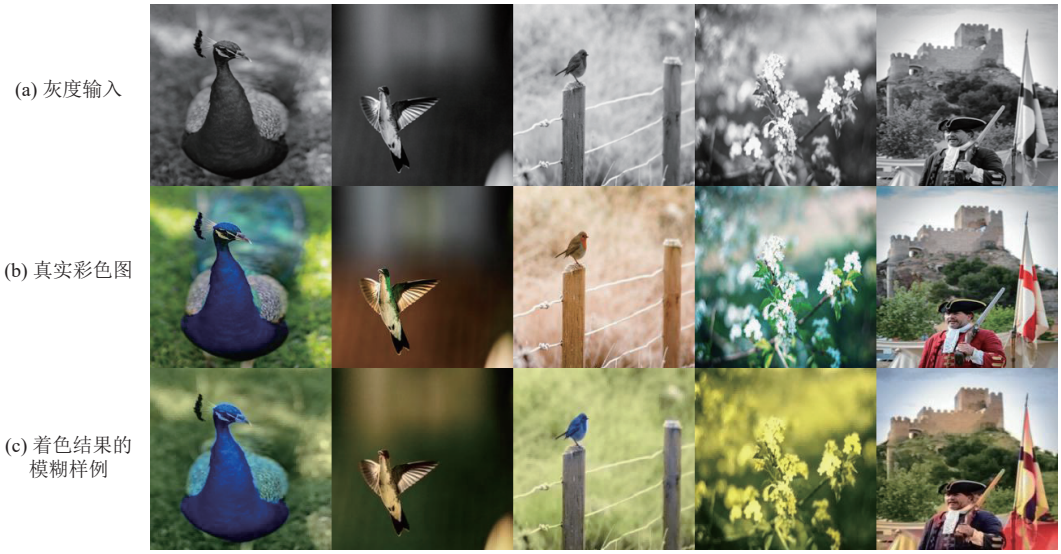


图 4 CT^2 相对模糊样例

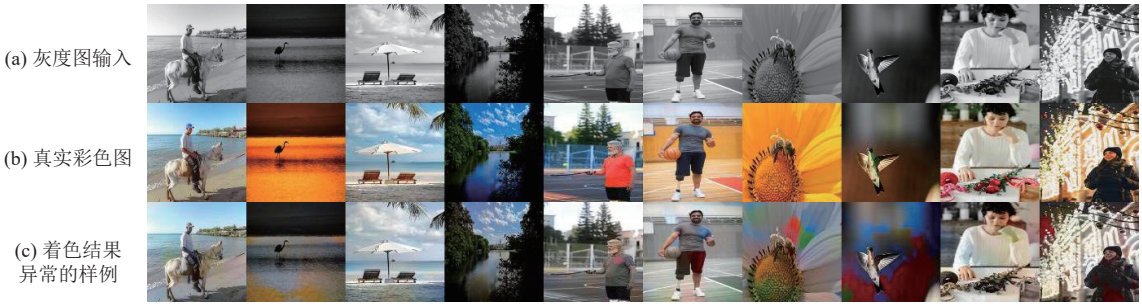


图 5 InstColor 失败样例

3.1.5 量化分析

我们基于潜在难样本集 O 进行进一步的量化分析, 辅助评估着色模型 F 的性能. 对于图像集 O 中的每一张灰度图像, 通过模型 F 生成对应的彩色结果图; 对于需要参考图像的模型, 使用在构建图像数据集 G 时挑选出来的语义相关的参考图像, 然后利用无参考图像质量评价模型 (Hyper-IQA[21]) 和美学评估模型[46]对生成结果进行综合评估. 具体而言, Hyper-IQA 模型为每张结果图预测其质量评分, 而美学评估模型则预测其美学评分. 我们对每个模型生成的所有结果图的质量评分和美学评分分别取平均值, 作为该模型在图像质量和美学表现上的最终分值. 此外, 还借助原始彩色图像计算每个模型着色结果的 FID 值, 以量化生成图像与真实图像之间的分布差异, 评估着色结果的相对真实性; 还测量 LPIPS 以评估生成图像与原始彩色图像在人类感知上的相似性; 然后, 通过 CF 量化模型着色结果的色彩鲜艳程度, 利用ΔCF 衡量模型着色结果与真实图像在颜色丰富度上的一致程度, 反映了模型生成逼真、自然颜色分布的能力, 结果如表 3 所示.

表 3 各模型在数据集 O 上的量化结果

模型	质量分数↑	美学分数↑	FID↓	LPIPS↓	CF↑	ΔCF↓
DDColor[14]	59.62	70.59	**73.61**	<u>0.251</u>	52.56	**20.82**
ColorFormer[15]	<u>60.24</u>	70.51	<u>78.67</u>	**0.240**	42.17	<u>22.12</u>
InstColor[17]	**60.61**	<u>71.74</u>	101.90	0.313	31.69	27.73
CT^2[16]	58.50	69.58	114.37	0.266	**60.49**	22.79
UniColor[42]	58.17	70.90	96.56	0.339	<u>53.35</u>	31.28
PDNLA[43]	59.77	71.46	101.74	0.283	29.27	30.50
SAC[44]	58.68	**72.10**	99.90	0.287	24.85	32.52
AC[45]	40.22	69.79	102.69	0.337	30.32	29.91

注: 加粗表示最优数值, 加下划线表示次优数值, ↑ (↓)表示数值越高 (越低)相对越好

从评估结果可以看出, 所有模型在潜在难样本集上的表现均欠佳, 图像质量得分和美学得分普遍不高, 表明这些图像对模型具有较高的挑战性. 在 FID 方面, 除了 DDColor 和 ColorFormer 两个模型的 FID 值低于 80, 其余模型的 FID 值均接近或超过 100, 其中 CT^2 模型的 FID 值达到最高的 114.37. 而对于 LPIPS, 达到最好性能的 ColorFormer 也仅为 0.240, 这一结果揭示模型生成图像与真实彩色图像分布之间存在显著偏差. 对于ΔCF 指标, 模型的性能更不如意, 所有模型的ΔCF 均超过 20, 最高达 32.52, 进一步验证模型在颜色还原能力上的不足. 这些结果验证了潜在难样本集 O 的有效性, 能够有效暴露模型在复杂场景下的性能局限性.

总体来看, ColorFormer 和 DDColor 的 FID、LPIPS 和ΔCF 都大幅度领先于其他模型, 质量分数和美学分数与最高值也相差无几, 而 CF 虽然与最高的 60.49 存在差距, 但这两个模型不因追求颜色丰富度而降低颜色真实性, 如图 6 所示, 其中黄色数值表示该图像的颜色丰富度, 即使 CT^2 和 UniColor 着色结果的颜色丰富度更高, ColorFormer 和 DDColor 的结果也更符合主观视觉感知.

(a) 灰度图 (b) 原始图 (c) ColorFormer (d) DDColor (e) UniColor (f) CT²

图 6 4 个模型着色结果视觉比较

3.2 消融实验

3.2.1 基于参考图像模型的局限性

为了深入探究基于参考图像的着色方法的局限性, 我们采用了先前研究的 3 个基于参考图像的着色模型 (PDNLA[43]、SAC[44]和 AC[45]), 并在图像集 O 上进行了重新着色实验. 与之前的着色模型不同的是, 我们在选择参考图像时, 特意选取了与目标图像在语义上完全不相关的图像. 此外, 我们还采用了两种颜色迁移模型, 分别基于语义相关和语义不相关的图像进行颜色迁移.

为了客观辅助性评估这些模型的着色效果, 我们使用与第 3.1.5 节相同的量化指标进行测试, 结果如表 4 所示. 可以看出, 使用语义完全不相关的参考图像对模型的着色性能产生了显著影响. 相较于使用语义相关的参考图像, 模型的各项指标普遍下降; 而对于颜色迁移模型, 输入为两张彩色图 (其中一张为目标图像, 另一张为参考图像), 即使模型已知初始图像的颜色分布特性, 进行颜色迁移后得到结果的 LPIPS 指标远差于两个最优的基于参考图像的着色模型. 这一结果充分揭示了参考图像与目标图像在语义相关性上的重要性, 对于基于参考图像的着色方法而言, 选择合适的参考图像是至关重要的. 这验证了基于参考图像的模型其使用场景是具有局限性的.

表 4 基于参考图像模型在数据集 O 上的量化结果

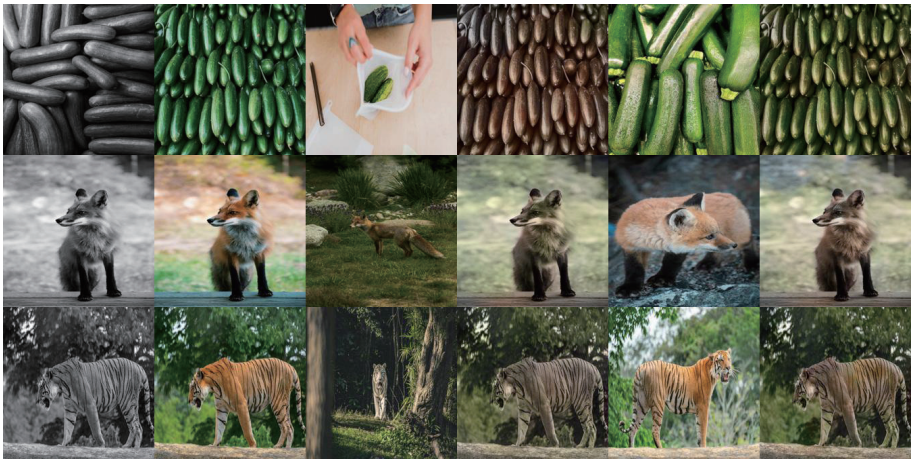
模型	质量分数↑	美学分数↑	LPIPS↓	FID↓	CF↑	ΔCF↓
PDNLA[43]	59.00 (59.77)	71.46 (71.46)	**0.291** (**0.283**)	105.42 (101.74)	22.61 (29.27)	36.66 (30.35)
SAC[44]	58.42 (58.68)	72.01 (72.10)	0.300 (0.287)	104.98 (99.90)	24.28 (24.85)	33.54 (32.52)
AC[45]	39.91 (40.22)	69.82 (69.79)	0.360 (0.337)	110.71 (102.69)	27.02 (30.32)	31.89 (29.91)
DCT[47]	56.27 (56.22)	71.21 (70.98)	0.355 (0.317)	**71.03** (**61.78**)	**32.00** (**37.62**)	**31.27** (**26.73**)
SCTNet[48]	**60.53** (**60.59**)	**72.05** (**72.28**)	0.414 (0.376)	122.73 (114.67)	24.01 (31.26)	34.52 (29.73)

注: 加粗表示最优数值, 加下划线表示次优数值, 括号中的数值表示基于语义相关参考图的测试性能, ↑ (↓)表示数值越高 (越低)相对越好

3.2.2 参考图像的选择分析

在基于参考图像的着色任务中, 我们主观筛选出各类别中语义关联性最强且内容最具代表性的图像作为参考图像. 具体选择标准如下: (1) 目标类别物体在图像中的视觉占比不低于 50%; (2) 排除卡通等非真实风格图像; (3) 最大程度覆盖数据集中该类别的典型视觉特征. 为验证该参考图像选择策略的有效性, 我们控制输入灰度图像不变, 使用着色模型 SAC[44], 分别采用符合与不符合标准的参考图像进行着色引导实验. 着色结果如图 7 所示, 目标物体像素占比过低时 (参考图像 1), 着色结果 (结果 1) 出现显著语义失真, 如生成的黄瓜呈现非自然的褐色, 老虎和狐狸的皮毛也被错误的染成绿色. 相比之下, 根据标准挑选出的参考图像 2 引导的着色结果 (结果 2) 虽与真实

原始图存在视觉差异, 但色彩分布符合自然先验. 对比实验表明: 有效 (符合选择标准) 的参考图像引导生成的着色结果, 在视觉保真度上显著优于无效 (不满足标准) 参考图像引导生成的结果, 后者直接导致引导型模型着色失败, 根据该策略选出的参考图像 2 展现出稳定的引导性, 证实了: (1) 参考图像的选择对图像着色质量具有关键影响; (2) 该选择策略能有效提升模型的着色效果.



(a) 灰度图 (b) 原始图 (c) 参考图像1 (d) 结果1 (e) 参考图像2 (f) 结果2

图 7 参考图对着色结果的影响

3.2.3 难样本集 O 特征分析

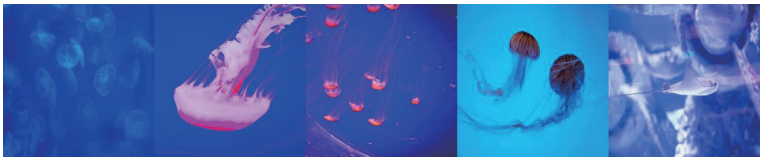
遵循第 2.2 节的描述, 我们从图像质量、颜色差异以及美学表现这 3 个维度共同挖掘难样本, 旨在系统挖掘在各维度分布边缘且结果差异显著的灰度图像集. 为了深入剖析着色模型难样本的特性, 我们对构建的难样本集进行可视化分析. 后文图 8 中均为难样本灰度图的真实彩色图像, 后文图 9 后 4 列为本文评估模型 F 中的其中 4 个模型所生成的结果. 通过观察与分析, 我们发现所挖掘出的难样本集呈现以下显著共性: (1) 如图 8(a) 所示, 难样本集中包含了大量花朵的图像 (为了展示更多图像信息, 我们用难样本集的原始彩色图进行展示), 这种图像挖掘可能归功于颜色差异维度上的评估, 因为花朵种类繁多、万紫千红, 若模型所使用的训练集并不包含所要着色的花朵的种类, 模型可能倾向于错误映射到常见色域, 这可能导致不同的模型预测出颜色差异极大的花朵, 如图 9 第 1 行所示. (2) 如图 8(b) 所示, 难样本集中还涵盖众多处于水下环境的透明或半透明生物的图像, 如水母等, 这些难样本的挖掘归功于图像质量维度下的差异评估, 因为许多模型在处理透明或半透明图像时存在薄弱环节, 容易生成带有伪影的结果, 从而导致图像质量显著下降, 如图 9 第 2 行所示, 这和 DDColor[14]的失败样例相一致. (3) 如图 8(c) 所示, 难样本集中, 显著比例图像包含的人物处于颜色单一的背景下, 例如密集植被覆盖区域, 此类图像在着色过程中, 模型容易发生溢色现象, 如图 9 第 3 行所示, 进而影响图像的美感, 因此这一类型难样本的挖掘可能主要归因于图像质量和美学变现维度上的差异评估; (4) 如图 8(d) 所示, 图像集中有许多复杂光照条件下的天空, 生成的图像往往难以准确对应真实图片中的太阳光或其他光源下的天空及水面, 如图 9 第 4 行所示, 真实图像可能是黄昏时分的昏暗天空, 但生成结果可能呈现出正午时分强烈太阳光下的天空, 不同模型的着色结果在图像质量、颜色以及美学上都差异极大, 此类难样本的挖掘归功于这 3 个维度上的综合差异评估.

3.2.4 基于多维差异联合指标的难样本挖掘策略分析

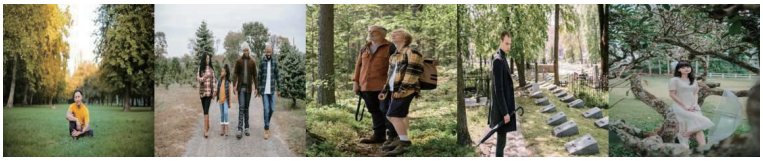
本文提出并采用一种融合了颜色差异、图像质量差异和美学差异的三维联合差异度量指标进行难样本的挖掘. 为了系统性评估该多维联合指标的必要性与有效性, 本节设计了消融实验, 探究不同差异指标组合在样本挖掘中的作用. 具体而言, 分析了以下 7 种策略: (1) 颜色差异、图像质量差异和美学差异的三维差异联合指标; (2) 颜色差异和图像质量差异的二维差异联合指标; (3) 颜色差异和美学差异的二维差异联合指标; (4) 图像质量和美学差异的二维差异联合指标; (5) 单一颜色差异指标; (6) 单一图像质量差异指标; (7) 单一美学差异指标. 实验保持参数 K=10, 针对 ColorFormer[15]和 DDColor[14]两个着色模型, 分别应用上述 7 种策略, 各自挖掘出包含 10 幅图像的

难样本集. 随后, 我们对两个模型在这些难样本集上的着色性能进行定量评估, 结果如表 5 所示. 综合表 5 中数据, 单一色彩差异指标筛选出的样本集对两个模型展现出最高的挑战性, 尤其是 ColorFormer 在各性能指标上均呈现显著劣势. 此外, 包含颜色差异的组合策略 (如三维差异联合指标、颜色差异和图像质量差异的二维差异联合指标) 所挖掘的样本, 相较于不含颜色差异的策略, 普遍导致模型性能下降更为明显.

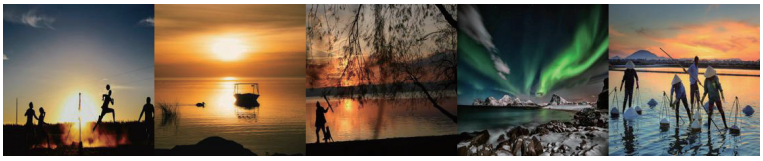
(a) 颜色多样性的花朵



(b) 水下环境的透明或半透明生物



(c) 密集植被覆盖背景中的人



(d) 复杂光照条件下的天空和水面

图 8　难样本特性可视化

(a) 输入　(b) 真实图像　(c) ColorFormer　(d) DDColor　(e) UniColor　(f) InstColor

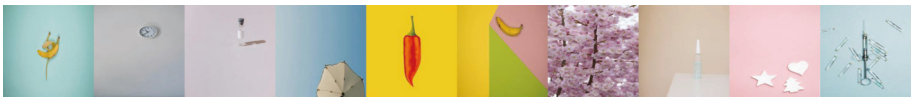
图 9　难样本代表性失败样例

表 5 差异指标选择的消融分析

差异指标	ColorFormer[15]				DDColor[14]			
	LPIPS↓	FID↓	CF↑	ΔCF↓	LPIPS↓	FID↓	CF↑	ΔCF↓
CD+IQ+IA	0.197	**134.66**	30.02	16.94	0.315	**142.32**	56.36	28.38
CD+IQ	0.211	97.84	35.22	9.46	0.316	96.80	65.30	31.01
CD+IA	0.252	124.82	34.39	21.57	**0.465**	113.05	89.33	**62.33**
IQ+IA	0.158	98.24	33.48	12.58	0.174	92.87	37.65	15.82
CD	**0.271**	118.82	**25.93**	**25.58**	0.452	120.64	90.96	57.55
IQ	0.185	98.63	34.44	17.29	0.201	92.93	37.03	16.84
IA	0.184	116.51	35.30	23.65	0.194	131.75	**36.64**	29.82

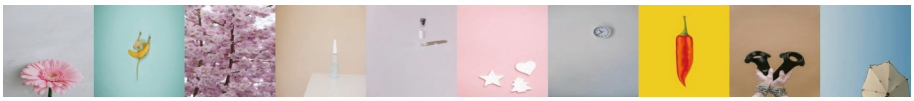
注: 加粗表示最差数值, 加下划线表示次差数值, ↑(↓)表示数值越高 (越低)相对越好, CD、IQ和IA分别表示颜色差异、图像质量差异和美学差异

然而, 视觉对比分析揭示了不同指标策略在样本筛选效果上的显著差异: 单一颜色差异指标筛选的样本集呈现高度类型单一化的特征, 主要表现为单一元素置于素色背景的极简主义风格图像, 如图 10(a). 单一美学差异指标挖掘出的难样本集未观察到上述类型的单一性问题, 如图 10(b), 表 5 的定量结果进一步辅助证实了该指标在识别有效难样本方面的重要贡献. 而引入美学差异构建的二维差异联合指标 (颜色差异+美学差异) 未能有效缓解该问题, 其挖掘的样本视觉特征与单一颜色差异指标的结果相似, 如图 10(c). 同时, 表 5 数据显示, 图像质量差异与颜色差异的二维差异联合指标所挖掘的样本, 其构成的挑战性相对不足. 相比之下, 三维差异联合指标 (颜色差异+图像质量差异+美学差异) 有效克服了上述局限性. 该策略挖掘的样本集不仅对模型着色性能构成了显著挑战, 同时样本多样性显著提升, 涵盖了更广泛的图像类型与内容类别, 如图 10(d). 图 10 中均为难样本灰度图的真实彩色图像, 综合定量与定性分析结果, 可以看出, 结合美学差异、图像质量差异与颜色差异的三维差异联合指标在构建高挑战性与丰富多样性兼备的难样本集方面具有显著优势.

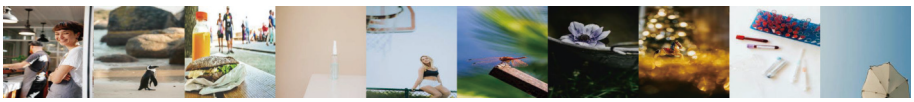


(a) 单一颜色差异指标挖掘的难样本

(b) 单一美学差异指标挖掘的难样本



(c) 颜色差异和美学差异的二维联合指标挖掘的难样本



(d) 三维差异联合指标挖掘的难样本

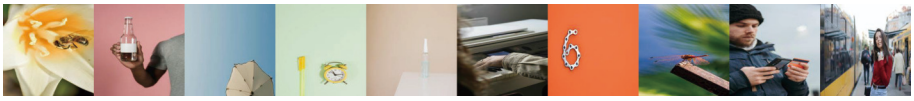
图 10 各指标组合挖掘的难样本视觉分析

3.2.5 美学差异指标的权重系数分析

从第 3.2.4 节的分析可以看出, 美学差异指标在难样本挖掘识别与样本集多样性提升方面均展现出显著价值. 因此, 在确定多指标联合权重时, 本文倾向于赋予美学差异指标更高权重系数 (λ). 本节系统研究美学差异权重 λ 对样本选择的影响. 具体包括 5 组对比策略: (1) 各指标归一化后等权相加; (2) 非归一化, 固定 λ=1; (3) 非归一化, 固定 λ=10; (4) 非归一化, 固定 λ=20; (5) 非归一化, 固定 λ=30. 实验保持参数 K=10, 针对 ColorFormer[15]和

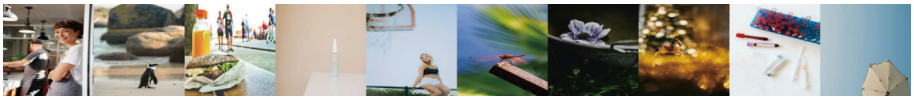
DDColor[14]两个着色模型, 分别应用上述 5 种策略, 各自挖掘出包含 10 幅图像的难样本集. 图 11 中也均为难样本灰度图的真实彩色图像, 视觉对比分析表明: 归一化策略挖掘的样本集存在显著样本同质化问题, 主要表现为极简风格主义, 如图 11(a). 非归一化时, 将 λ 设为 1 时, 部分缓解样本同质化问题, 如图 11(b). 我们发现, 通过增大 λ, 同质化问题能够得到有效解决, 如图 11(c)–(e). 在解决同质化问题的同时, 为避免美学差异指标权重过高导致失衡, 本研究选取 λ=20 作为最优权衡点.

(a) 各指标归一化



(b) λ=1

(c) λ=10



(d) λ=20

(e) λ=30

图 11　各策略挖掘的难样本视觉分析

4　跨模型一致性验证

为了系统验证本文所构建的难样本集在其他模型框架中的泛化能力, 我们选取了近期的主流着色模型作为测试对象, 包括 DeOldify[49]、BigColor[5]以及 Liang 等人[50]提出的一种多模态着色框架 CtrlColor. CtrlColor 有许多参数可以自由选择, 支持两种典型的着色范式, 包括: (1) 无条件着色: 直接将输入的灰度图生成合理颜色分布的彩色图, 包含两种参数选择 (使用或不使用可变形自动编码器 (VAE)); (2) 条件着色: 基于自然语言提示 (prompt) 引导色彩生成, 同样包括两种参数选择 (使用或不使用可变形自动编码器). 我们使用 DeOldify[49]、BigColor[5]以及这 4 种条件组合下的模型在难样本集上进行跨模型的一致性测试. 定量分析如表 6 所示, 模型在一些关键指标上均呈现较低的性能. 图 12 为失败结果可视化样例, 其中黄色框内的区域为着色异常的区域. 具体而言, DeOldify[49]着色结果的 CF 值过低, 表现为颜色单调、饱和度低且层次感弱, 对于未使用可变形自动编码器的 CtrlColor 系列模型[50], 其 CF 偏高, 导致视觉混乱且缺乏自然感与真实性 (如图 12 第 1、3 行). 并且, 所有模型着色结果的 FID 值显著劣化, 且颜色保真度 (ΔCF) 也较高. 这种现象表明, 模型倾向于生成过饱和色彩以提升视觉吸引力, 但以牺牲颜色分布真实性为代价, 导致与真实场景的色彩统计特性产生显著分歧.

通过图 12 的典型失败案例可视化分析发现, CtrlColor 系列模型[50]在该样本集的难点主要有以下 3 类: (1) 人体部位出现反生物规律的色偏, 如紫色的眼睛、青蓝色的面部以及诡异颜色的肢体; (2) 生成结果与常识认知冲突, 例如呈现有毒理学暗示的青绿色饮料、非自然光照条件下的蓝色人影以及黑色鸭嘴等生物特征畸变; (3) 日照条件下的异常天空色温, 和主观印象中的光照特性相违背.

表 6 跨模型验证结果

模型	LPIPS↓	FID↓	CF↑	ΔCF↓
CtrlColor	0.365	96.76	62.01	30.37
CtrlColor-VAE	0.316	88.91	40.85	28.17
CtrlColor-prompt	0.351	91.85	**69.96**	29.59
CtrlColor-VAE&prompt	0.298	**81.27**	48.66	**23.88**
DeOldify[49]	**0.260**	82.32	22.19	34.14
BigColor[5]	0.265	85.85	44.34	24.01

注: 加粗表示最优数值, ↑ (↓)表示数值越高 (越低) 越好

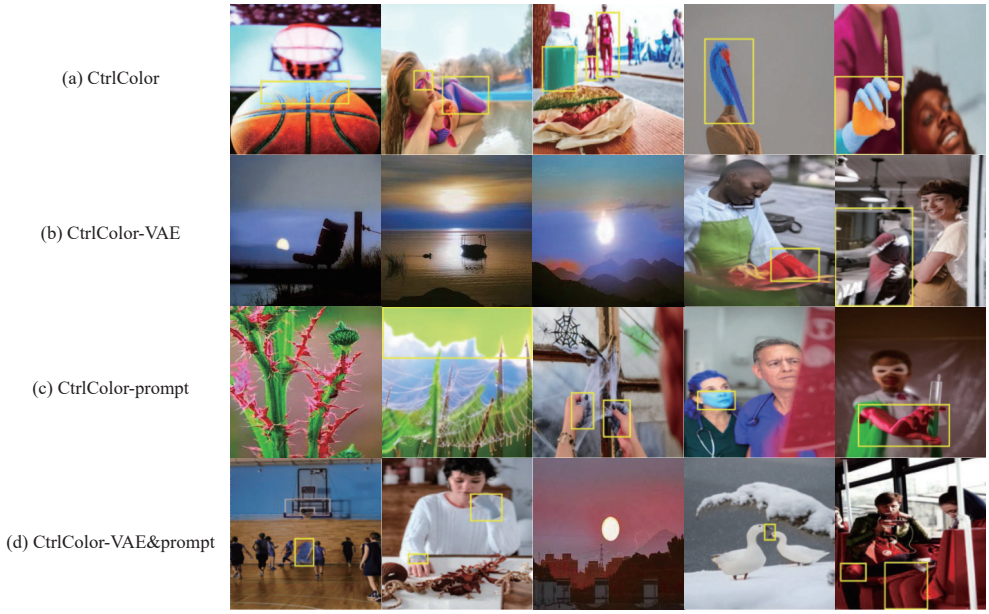


图 12 CtrlColor 系列模型失败结果可视化样例

如图 13 所示, BigColor[5]的着色失效模式与 CtrlColor 系列模型[50]呈现出较高相似性, 大多表现在语义色彩失真, 该现象表现为生成图像的色彩违背客观世界的物理规律或语义先验. 例如, 生成的汉堡及辣椒等物体呈现违背自然先验的异常色调, 人体皮肤区域也生成了异样的颜色. 其根本缺陷在于着色过程受环境干扰, 具体表现为: 人体肤色被同化为背景天空的灰蓝色; 处于密集植被环境下的人体肤色以及衣物的颜色也被误导; 石碑背景下手臂区域被错误渲染为石灰色. 表 6 量化分析及图 14 可视化结果表明, DeOldify[49]在难样本集上暴露出显著的局限性: 色彩表达能力贫乏, 着色结果的颜色丰富度 (CF 值) 显著低于其他对比模型.

图 13 BigColor 失败结果可视化样例

图 14 DeOldify 失败结果可视化样例

本研究所构建的难样本集能够系统性揭示着色模型的关键失效模式. 这些现象共同验证了本文所构建的难样本集是一个包含大量复杂场景的有效图像集, 同时也验证了该样本集跨模型的一致有效性.

5 总 结

本文提出一种着色模型评估方法来对主流的 8 种着色模型进行比较, 为着色模型的评估提供了新的视角和方法. 通过该评估框架, 我们挖掘出了具有广泛的适用性和挑战性的难样本集, 该难样本集不仅对本文所重点评估的模型 F 构成显著挑战, 同样也对其他近期提出的着色模型展现出难度. 这表明, 我们挖掘的难样本集具有普遍的难例特性. 此外, 本文所提出的是任务无关的评估框架, 展现了极高的灵活性和扩展性, 除图像着色模型外, 也同样适用于其他类型的模型评价当中, 例如, 语义分割、图像去噪和图像超分辨率重建等. 同时, 该框架不依赖于模型的训练数据分布, 模型可使用不同分布的数据进行训练, 并在挖掘出的难样本集上进行泛化性主观评估或量化分析. 而且, 该框架支持任意数量的模型参与评估, 并且能够根据需求寻找任意数量的难样本, 为模型的改进和优化提供了指导意见. 我们挖掘的难样本集也可以作为模型训练集的重要组成部分, 从而增加训练难度来提升模型的泛化能力和鲁棒性, 通过深入分析所挖掘的难样本集, 我们可以更有针对性地调整模型参数、优化模型结构, 从而提升模型的综合性能. 在未来的工作中, 我们还将对该框架进行进一步的优化. 目前挖掘出的难样本集可能包含部分非难样本, 因此, 可以通过增大 K 值以获取规模更大的候选难样本集, 随后通过某种方式对其进行进一步的筛选, 得到更具代表性和通用的难样本集, 从而更有效地暴露模型的痛点.

References

[1] Zhong X, Lu TY, Huang WX, Ye M, Jia XM, Lin CW. Grayscale enhancement colorization network for visible-infrared person re-identification. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(3): 1418–1430. [doi: 10.1109/TCSVT.2021.3072171]

[2] Liu SG, Zhang X, Wu JT, Sun JZ, Peng QS. Gray-scale image colorization based on the control of single-parameter. Journal of Image and Graphics, 2011, 16(7): 1297–1302 (in Chinese with English abstract). [doi: 10.11834/jig.20110722]

[3] Hu W, Qin KH. Fast multi-resolution colorization of high-resolution gray images. Chinese Journal of Computers, 2009, 32(5): 1062–1068 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.01062]

[4] Wu YZ, Wang XT, Li Y, Zhang HL, Zhao X, Shan Y. Towards vivid and diverse image colorization with generative color prior. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 14357–14366. [doi: 10.1109/ICCV48922.2021.01411]

[5] Kim G, Kang K, Kim S, Lee H, Kim S, Kim J, Baek SH, Cho S. BigColor: Colorization using a generative color prior for natural images. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 350–366. [doi: 10.1007/978-3-031-20071-7_21]

[6] Yun J, Lee S, Park M, Choo J. iColoriT: Towards propagating local hints to the right region in interactive colorization by leveraging vision Transformer. In: Proc. of the 2023 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2023. 1787–1796. [doi: 10.1109/WACV56688.2023.00183]

[7] Li HA, Zheng QX, Ma T, Zhang J, Li ZL, Kang BS. Multi-field features representation based colorization of grayscale images. Pattern Recognition and Artificial Intelligence, 2022, 35(7): 637–648 (in Chinese with English abstract). [doi: 10.16451/j.cnki.issn1003-6059.202207006]

[8] Li HA, Zheng QX, Zhang J, Du ZM, Li ZL, Kang BS. Pix2Pix-based grayscale image coloring method. Journal of Computer-aided Design & Computer Graphics, 2021, 33(6): 929–938 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2021.18596]

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

[10] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

[11] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. arXiv:1706.08500, 2018.

[12] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: 10.1109/CVPR.2018.00068]

[13] Hasler D, Suesstrunk SE. Measuring colorfulness in natural images. In: Proc. Volume 5007, Human Vision and Electronic Imaging VIII. Santa Clara: SPIE, 2003. 87–95. [doi: 10.1117/12.477378]

[14] Kang XY, Yang T, Ouyang WQ, Ren PR, Li LZ, Xie XS. DDColor: Towards photo-realistic image colorization via dual decoders. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 328–338. [doi: 10.1109/ICCV51070.2023.00037]

[15] Ji XZ, Jiang BY, Luo DH, Tao GP, Chu WQ, Xie ZF, Wang CJ, Tai Y. ColorFormer: Image colorization via color memory assisted hybrid-attention Transformer. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 20–36. [doi: 10.1007/978-3-031-19787-1_2]

[16] Weng SC, Sun JM, Li Y, Li S, Shi BX. CT^2: Colorization Transformer via color tokens. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 1–16. [doi: 10.1007/978-3-031-20071-7_1]

[17] Su JW, Chu HK, Huang JB. Instance-aware image colorization. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7965–7974. [doi: 10.1109/CVPR42600.2020.00799]

[18] Zhang WX, Zhai GT, Wei Y, Yang XK, Ma KD. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14071–14081. [doi: 10.1109/CVPR52729.2023.01352]

[19] Ke JJ, Wang QF, Wang YL, Milanfar P, Yang F. MUSIQ: Multi-scale image quality Transformer. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5128–5137. [doi: 10.1109/ICCV48922.2021.00510]

[20] Pan ZQ, Yuan F, Lei JJ, Fang YM, Shao X, Kwong S. VCRNet: Visual compensation restoration network for no-reference image quality assessment. IEEE Trans. on Image Processing, 2022, 31: 1613–1627. [doi: 10.1109/TIP.2022.3144892]

[21] Su SL, Yan QS, Zhu Y, Zhang C, Ge X, Sun JQ, Zhang YN. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3664–3673. [doi: 10.1109/CVPR42600.2020.00372]

[22] Qin GY, Hu RZ, Liu YT, Zheng XW, Liu HT, Li X, Zhang Y. Data-efficient image quality assessment with attention-panel decoder. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 2091–2100. [doi: 10.1609/aaai.v37i2.25302]

[23] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3213–3223. [doi: 10.1109/CVPR.2016.350]

[24] Gu JJ, Cai HM, Chen HY, Ye XX, Ren J, Dong C. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. arXiv:2011.15002, 2020.

[25] Nekrasov V, Shen CH, Reid I. Diagnostics in semantic segmentation. arXiv:1809.10328, 2018.

[26] Zhang SS, Benenson R, Omran M, Hosang J, Schiele B. How far are we from solving pedestrian detection? In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1259–1267. [doi: 10.1109/CVPR.2016.141]

[27] Hoiem D, Chodpathumwan Y, Dai QY. Diagnosing error in object detectors. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 340–353. [doi: 10.1007/978-3-642-33712-3_25]

[28] Alwassel H, Heilbron FC, Escorcia V, Ghanem B. Diagnosing error in temporal action detectors. In: Proc. of the 2018 European Conf. on Computer Vision. Munich: Springer, 2018. 264–280. [doi: 10.1007/978-3-030-01219-9_16]

[29] Li BY, Ren WQ, Fu DP, Tao DC, Feng D, Zeng WJ, Wang ZY. Benchmarking single-image dehazing and beyond. IEEE Trans. on Image Processing, 2019, 28(1): 492–505. [doi: 10.1109/TIP.2018.2867951]

[30] Borji A. Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding, 2022, 215: 103329. [doi: 10.1016/j.cviu.2021.103329]

[31] Chen MH, Wang ZQ, Zheng F. Benchmarks for corruption invariant person re-identification. arXiv:2111.00880, 2022.

[32] Li XL, Zhao ZY. Pixel level semantic understanding: From classification to regression. Scientia Sinica Informationis, 2021, 51(4): 521–564 (in Chinese with English abstract). [doi: 10.1360/SSI-2020-0340]

[33] Liu JY, Liu D, Yang WH, Xia SF, Zhang XS, Dai YY. A comprehensive benchmark for single image compression artifact reduction. IEEE Trans. on Image Processing, 2020, 29: 7845–7860. [doi: 10.1109/TIP.2020.3007828]

[34] Li SY, Ren WQ, Wang F, Araujo IB, Tokuda EK, Junior RH, Cesar RM Jr, Wang ZY, Cao XC. A comprehensive benchmark analysis of

single image deraining: Current challenges and future perspectives. Int'l Journal of Computer Vision, 2021, 129(4): 1301–1322. [doi: 10.1007/s11263-020-01416-w]
[35] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. arXiv:1903.12261, 2019.
[36] Hou QB, Han LH, Liu JJ, Cheng MM. Autonomous learning of semantic segmentation from Internet images. Scientia Sinica Informationis, 2021, 51(7): 1084–1099 (in Chinese with English abstract). [doi: 10.1360/SSI-2020-0146]
[37] Wang Z, Simoncelli EP. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. Journal of Vision, 2008, 8(12): 8. [doi: 10.1167/8.12.8]
[38] Wang HT, Chen TL, Wang ZY, Ma KD. I am going MAD: Maximum discrepancy competition for comparing classifiers adaptively. arXiv:2002.10648, 2020.
[39] Wang HT, Chen TL, Wang ZY, Ma KD. Troubleshooting image segmentation models with human-in-the-loop. Machine Learning, 2023, 112(3): 1033–1051. [doi: 10.1007/s10994-021-06110-7]
[40] Ma KD, Duanmu ZF, Wang Z, Wu QB, Liu WT, Yong HW, Li HL, Zhang L. Group maximum differentiation competition: Model comparison with few samples. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(4): 851–864. [doi: 10.1109/TPAMI.2018.2889948]
[41] Wang ZH, Ma KD. Active fine-tuning from gMAD examples improves blind image quality assessment. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4577–4590. [doi: 10.1109/TPAMI.2021.3071759]
[42] Huang ZT, Zhao NX, Liao J. UniColor: A unified framework for multi-modal colorization with Transformer. ACM Trans. on Graphics, 2022, 41(6): 205. [doi: 10.1145/3550454.3555471]
[43] Wang HZ, Zhai DM, Liu XM, Jiang JJ, Gao W. Unsupervised deep exemplar colorization via pyramid dual non-local attention. IEEE Trans. on Image Processing, 2023, 32: 4114–4127. [doi: 10.1109/TIP.2023.3293777]
[44] Carrillo H, Clément M, Bugeau A. Super-attention for exemplar-based image colorization. In: Proc. of the 16th Asian Conf. on Computer Vision. Macao: Springer, 2023. 646–662. [doi: 10.1007/978-3-031-26284-5_39]
[45] Yin W, Lu P, Zhao ZR, Peng XJ. Yes, "Attention Is All You Need", for exemplar based colorization. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 2243–2251. [doi: 10.1145/3474085.3475385]
[46] Li LD, Huang YP, Wu JJ, Yang YZ, Li YQ, Guo YD, Shi GM. Theme-aware visual attribute reasoning for image aesthetics assessment. IEEE Trans. on Circuits and Systems for Video Technology, 2023, 33(9): 4798–4811. [doi: 10.1109/TCSVT.2023.3249185]
[47] Lee J, Son H, Lee G, Lee J, Cho S, Lee S. Deep color transfer using histogram analogy. The Visual Computer, 2020, 36(10): 2129–2143. [doi: 10.1007/s00371-020-01921-6]
[48] Fang YM, Yuan PW, Lv CL, Peng C, Yan JB, Lin WS. Saliency Guided deep neural network for color transfer with light optimization. IEEE Trans. on Image Processing, 2024, 33: 2880–2894. [doi: 10.1109/TIP.2024.3381833]
[49] Antic J. A deep learning based project for colorizing and restoring old images (and video!). 2019. https://github.com/jantic/DeOldify
[50] Liang ZX, Li ZC, Zhou SC, Li CY, Loy CC. Control color: Multimodal diffusion-based interactive image colorization. Int'l Journal of Computer Vision, 2025, 133: 7897–7923. [doi: 10.1007/s11263-025-02549-6]

附中文参考文献

[2] 刘世光, 张翔, 吴婧婷, 孙济洲, 彭群生. 单参数控制的灰度图像着色. 中国图象图形学报, 2011, 16(7): 1297–1302. [doi: 10.11834/jig.20110722]
[3] 胡伟, 秦开怀, 高分辨率灰度图像的快速多分辨率着色. 计算机学报, 2009, 32(5): 1062–1068. [doi: 10.3724/SP.J.1016.2009.01062]
[7] 李洪安, 郑峭雪, 马天, 张婧, 李占利, 康宝生. 多视野特征表示的灰度图像彩色化方法. 模式识别与人工智能, 2022, 35(7): 637–648. [doi: 10.16451/j.cnki.issn1003-6059.202207006]
[8] 李洪安, 郑峭雪, 张婧, 杜卓明, 李占利, 康宝生. 结合 Pix2Pix 生成对抗网络的灰度图像着色方法. 计算机辅助设计与图形学学报, 2021, 33(6): 929–938. [doi: 10.3724/SP.J.1089.2021.18596]
[32] 李学龙, 赵致远. 像素级语义理解: 从分类到回归. 中国科学: 信息科学, 2021, 51(4): 521–564. [doi: 10.1360/SSI-2020-0340]
[36] 侯淇彬, 韩凌昊, 刘姜江, 程明明. 互联网图像驱动的语义分割自主学习. 中国科学: 信息科学, 2021, 51(7): 1084–1099. [doi: 10.1360/SSI-2020-0146]

作者简介

鄢杰斌, 博士, 讲师, CCF 专业会员, 主要研究领域为人工智能, 深度学习, 多媒体技术.

彭振团, 硕士生, 主要研究领域为人工智能, 深度学习.

祝文涛, 硕士生, 主要研究领域为深度学习, 模型评估.

蔡超, 博士, 讲师, 主要研究领域为人工智能, 深度学习, 时间序列分析.

毛阿敏, 博士, 主要研究领域为人工智能, 显著性检测.

方玉明, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为视觉大数据, 计算机视觉, 多媒体技术处理, 智能信息处理.