基于协同注意力的图文多任务美学分析研究

苗好田[1] 梁嘉欣[1]
MIAO Haotian LIANG Jiaxin

摘 要 随着网络技术的进步和移动智能终端的普及，社交图像已成为重要的信息载体，人们对于图像内容趣味度等美学属性也有了进一步的追求。在此背景下，社交图像的美学评价研究备受瞩目。过往相关研究大多仅将图像自身的视觉信息映射到人工标注，而未充分挖掘利用其他维度的相关内容。因此，利用模态数据或相关属性等来帮助突破图像美学研究的“瓶颈”极具意义。文章提出了一个基于多模态协同注意力及评分和风格多任务学习的图文美学分析模型。首先，提取图像和文本的局部特征组成向量矩阵，作为模型的输入。然后，基于 Transformer 编码结构，构建自注意力模块和协同注意力模块，并在此基础上设计了一个多模态协同注意力网络。通过网络中堆叠的多模态注意力层对不同模态特征反复的协同学习来实现信息互补。之后将重新编码后的特征进行拼接，用于美学评价任务。此外，将图像风格作为额外的训练任务，结合多任务学习的策略，对模型进一步优化。最后，爬取图像美学数据集对应的文本评论实现数据的扩展，通过在得到的多模态数据集 AVA-m 上进行实验，验证了所提模型的有效性。

关键词 美学评价；多模态；协同注意力；多任务学习；Transformer

doi：10.3969/j.issn.1672-9528.2025.05.035

0 引言

随着科技的飞速发展，成千上万张图像可轻松存储在尺寸较小的芯片中，使数码摄影在全球范围内普及。同时，社交网络的兴起也为信息传播提供了有力支持，图像分享逐渐成为一种广受欢迎的社交方式。

图像美学分析旨在构建能够模拟人类审美感知的自动评价算法，这一过程本身极具挑战性。现有研究多依赖对单一模态的视觉特征进行建模，难以充分应对审美的主观性与复杂性，因此在预测效果上提升空间有限。为此，探索更多模态或维度的信息以辅助美学分析显得尤为必要。图像与文本作为两种核心的信息载体，各有优势：图像直观、生动、富有艺术表现力，但往往较为抽象，难以直接获取高层次语义；相比之下，文本语义清晰且信息密度高，有助于揭示图像背后的含义。

在诸如 Flickr、DPChallenge 等图像分享平台上，用户除提供评分外，往往还会附带评论，这些评论内容体现了其主观观点，能够解释其评分依据，因而与图像的视觉特征形成了高度互补关系。若能有效融合这两种模态的信息，有望进一步提升图像美学评分的预测能力。此外，尽管美学判断带有较强主观性，但大众审美中仍存在一定共性，而摄影风格正是基于这种共性发展出的表达手段。恰当运用风格，能够显著增强图像的表现力和吸引力。即便拍摄者尚未熟练掌握风格运用技巧，了解这些规律本身也能帮助其从更加客观、专业的角度对图像美学质量做出判断。在 DPChallenge 等平台中，这些摄影风格也被明确标注，便于用户参考与学习。

1 相关工作

图像美学评价的核心目标在于使计算机具备对图像主观美感的判断能力，从而模拟人类的审美感知并做出相应决策，在图像检索、图像增强以及图像情感分析等领域[1]展现出重要的应用价值。其中，特征提取在整个美学评价过程中扮演着至关重要的角色。Murray 等人[2]在早期研究中不仅构建了对美学分析产生深远影响的大规模图像数据集 AVA，还使用 SIFT、Fisher Vector 等传统视觉特征，结合线性 SVM 模型对图像的美学质量进行了预测。近年来，随着深度学习，尤其是卷积神经网络（CNN）的广泛应用，越来越多研究者开始将其引入到美学分析中。考虑到美学评价较为依赖图像的细节特征，Lu 等人[3]提出了多列式 CNN 架构，分别从原图和随机裁剪得到的图像补丁中提取全局和局部特征，并在网络中进行融合。Hosu 等人[4]基于迁移学习思想，从预训练 CNN 的多个卷积模块中提取多层空间池化特征，拼接后输入浅层网络进行训练。Liu 等人[5]则从图结构的角度出发，将

1. 沈阳理工大学 辽宁沈阳 110159

图像建模为由多个局部区域组成的图形，通过图卷积方法探索图像各区域之间的视觉联系，从而增强美学属性的建模能力。

近年来，融合多模态信息的图像美学预测方法取得了良好效果。Zhou 等人[6]提出了一种多模态深度信念网络（DBM），同时对图像和文本信息进行编码以支持美学预测；Hii 等人[7]将图像和文本的特征进行拼接，从而构建多模态输入；Wang 等人[8]采用多任务学习框架，以图像评分与用户评论作为监督信号，使模型不仅能输出美学得分，还能生成文本描述。与此同时，也有研究引入语义信息或图像风格属性作为辅助特征，以提升预测精度。例如，Kong 等人[9]提出结合图像内容与风格属性的自适应网络，并通过成对输入的排序损失进行训练；Kao 等人[10]在目标函数中建模美学与语义之间的关系，提出基于多任务关系学习的 CNN 模型；Pan 等人[11]则构建了一个基于 GAN 的多任务网络，联合学习图像的评分与风格特征。此外，Vaswani 等人[12]提出的 Transformer 模型基于注意力机制，完全摒弃 RNN 和 CNN 架构，通过将序列输入映射为连续表示并进行解码，在多种任务中取得优异效果，成为 Google 云推荐的基础架构之一，并被广泛应用于各类上下文建模任务[13]。

尽管已有方法在图像美学分析中取得了显著进展，但基于 CNN 的模型通常通过滑动窗口卷积获取图像的全局语义特征，难以有效捕捉图像中不同区域之间的长距离依赖，而这种信息对于构图分析尤为关键。同时，现有多模态模型多数侧重于图像与文本的整体特征融合，尚未深入建模图像区域与文本词语之间的细粒度关系。为解决这些问题，本文提出一种多模态协同注意力网络（Co-Transformer），通过图像和文本之间局部语义信息的深度交互，提升模态间的协同表达能力。同时，模型采用多任务学习机制，将美学评分与图像风格识别联合建模，从而进一步提升整体预测性能。

2 基于协同注意力的图文多任务美学分析

针对图像美学评价任务，本文提出了一种基于多模态协同学习的图文美学多任务分析模型 MCN-MTL。该模型以美学评分网站提供的图像及其配套评论文本为研究对象，首先分别提取图像区域特征与文本语义特征，并将两者作为 MCN-MTL 模型的输入。在此基础上，引入了 Co-Attention 机制以实现跨模态特征之间的深层协同学习。具体而言，模型采用了在 Attention 领域中广泛应用的深度自注意力架构 Transformer，并在其结构上设计了一个多模态注意力层（multimodal transformer layer, MTRL），通过图文特征的多轮交互学习，有效建构图像与文本间的语义联系，挖掘其所包含的美学信息。此外，为进一步提升模型性能，MCN-MTL 模型还引入了对多种潜在的美学相关属性的分析，例如互补色搭配、三分法构图、动态模糊效果等图像风格特征，以防止模型忽略可能具有显著影响的细节因素，从而增强整体的美学识别与判断能力。

图 1 展示了所提出的 MCN-MTL 模型的整体结构。整个训练过程主要由三部分组成：（1）特征提取阶段，分别从图像与文本中提取对应的向量表示，并将其输入至 MCN 网络；（2）多模态协同注意力机制，通过跨模态的迭代交互过程，实现图文信息的深度融合；（3）美学评分与风格识别的联合建模，在不同的美学维度上进行多任务训练，通过最小化局部损失函数优化各自的全连接层，同时进一步优化全局损失以训练 MCN 网络参数，从而获得最终的最优模型。在测试阶段，训练好的 MCN-MTL 模型可分别对图像进行美学打分与风格预测，二者互不依赖，因此在对测试集进行评分时，无需提供风格标签。

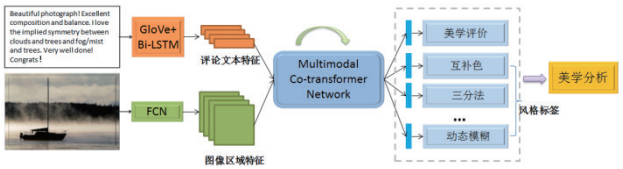

图 1 美学评价模型 MCN-MTL 的整体框架

图像和文本经过特征提取后所形成的向量矩阵被作为输入送入 MCN 网络，进一步用于多模态特征的学习与融合。如图 2 所示，MCN 网络的整体框架由两个关键注意力模块构成：（1）自注意力模块（self-transformer module, STRM）；（2）协同注意力模块（co-transformer module, CTRM）。对于输入的图像和文本特征，首先分别通过 STRM 模块进行重新编码，使模型能够自主识别本模态中更具代表性的特征，形成新的向量表达；随后，通过由 CTRM 与 STRM 组成的多模态注意力层，实现图像与文本之间的深度协同学习，进一步挖掘跨模态中对当前模态具有补充意义的信息。通过多层 MTRL 的堆叠使用，可以显著增强模态间的交互与整合。最终，将图像与文本模态的输出通过 Add & Norm 层完成残差连接和归一化处理，拼接后形成融合的多模态表示，用于后续的美学分析任务。

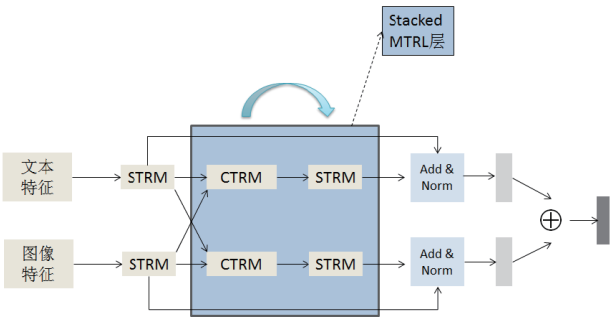

图 2 多模态协同注意力网络 MCN 的框架图

本文将图像风格作为辅助信息引入，将单一的美学评分任务拓展为多任务学习（multi-task learning, MTL）问题。美学评价与风格识别任务共用同一个网络结构，即 MCN 网络。在训练阶段，不同任务间相互作用，共同影响模型参数的更新。当所有任务完成收敛后，整体性能得到提升，且优于传统的单任务方法。

MCN 网络的主干部分参数共享，不同任务的全连接（FC）层各自独立。分类任务采用二元交叉熵损失，回归任务则使用均方误差作为损失函数。在训练过程中，美学评分与风格识别均为有监督任务，各自的 FC 层通过优化本任务的损失独立更新；MCN 主网络参数则由多个任务的总损失联合驱动优化。任务损失的权重采用动态加权策略，结合任务 t 在上一个 epoch 的损失值与下降速率，调整其权重。对于收敛较慢的任务，会赋予更大权重以加快其优化进度。

由于各任务的 FC 层独立，在测试阶段无需提供风格标签，模型即可对图像的美学得分进行单独预测与评估。

3 实验与分析

在视觉美学分析研究中，AVA 数据集由 Murray 等人提出，是一个具有代表性的大规模图像数据集，为后续研究提供了坚实的基础。多数相关工作均围绕该数据集展开。AVA 数据集包含超过 25 万张附带美学评分的图像，这些图像来源于数码摄影比赛平台 dpchallenge.com [14]，并由专业摄影师、图像从业者及摄影爱好者进行标注。每张图像获得的评分数量在 78~549 之间，平均约为 210 个，评分范围为 1~10 分，分数越高表示图像质量越高。此外，AVA 数据集中的部分图像还带有风格信息，包括互补色、三分法、长曝光等共 14 类风格标签。表 1 为数据集中各类风格标签的分布情况。

表 1 AVA 数据集中的风格分布

风格标签	互补色	双色调	动态渲染	纹理图	亮白
数据数量	949	1 301	396	840	1 199
风格标签	长曝光	微距	动态模糊	负片	三分法
数据数量	845	1 698	609	959	1 031
风格标签	浅景深	剪影	软焦距	消失点	
数据数量	710	1 389	1 479	674	

原始 AVA 数据集仅包含视觉图像，未涵盖用户在网络平台上的评论信息，而这些评论有助于理解评分背后的主观依据，为美学分析提供更多语义支持。为此，本文在 AVA 数据集基础上，利用图像 ID 从 dpchallenge.com 网站爬取对应的用户评论，并进行清洗处理，去除引号、HTML 标签及链接等无效内容，构建了包含文本评论的 AVA-Comments 数据集。本文将原始 AVA 扩展为多模态美学分析数据集 AVA-m，规模约为 220 000 条图文样本。

表 2 展示了本文所提出的图文多模态美学评价模型 MCN-MTL 与多种对比模型在 AVA-m 数据集上的实验结果。与同样引入辅助信息进行美学分析的模型，如 Reg+Rank+Att+Cont、MT-CNN 和 Att-GAN 相比，MCN-MTL 在 ACC 和平均分上至少分别提升了 7.07% 和 0.152 2。这一优势源于本文不仅融合图像风格信息，还通过协同注意力机制深入挖掘了文本模态的语义。与 Multimodal DBM 和 MULTGAP 这类多模态分类模型相比，MCN-MTL 在 ACC 上分别提升了 6.75% 和 3.36%，进一步说明所提出的 MCN 网络在模态间特征学习与融合方面能力更强。值得注意的是，近年来的方法如 MLSP、MPA 和 RGNet 等多采用图像局部内容建模，在某些指标上甚至超过了多模态模型 MULTGAP。相比之下，MCN-MTL 在学习图像区域特征的同时，还引入了文本词序列信息，实现了不同模态局部特征的交互融合，因此在 ACC 和平均分上至少提升了 2.04% 和 0.027 5。此外，去除多任务学习策略的 MCN+FC 模型，也在 ACC 和平均分上超越现有方法，分别提升了 0.85% 和 0.019 3，进一步验证了 MCN 结构本身的有效性。综上，MCN-MTL 模型通过局部交互、整体融合及风格引入，在图文多模态美学评价任务中表现出色，具备较强的实用性和推广价值。

表 2 不同美学评价模型间的效果比较

模型	ACC/%	ρ
Murray et al.	68.00	—
Reg+Rank+Att+Cont	77.33	0.558 1
MT_CNN	78.56	—
Att_GAN	—	0.631 3
Multimodal DBM	78.88	—
MULTIGAP	82.27	—
NIMA	81.51	0.612
MLSP	81.72	0.756
A-Lamp	82.0	—
MPA	83.03	—
RGNet	83.59	—
MCN	84.44	0.775 3
MCN-MTL	85.63	0.783 5

图 3 展示了部分美学预测的实验结果，图像风格的分类结果均与真实标签一致，() 外代表图像的预测值，() 内代表图像的实际值。如图 3 所示，在美学与风格的分类任务中，模型表现较为理想，同时在美学评分的拟合方面也展现出良好能力。此外，图中还反映出用户在评论时较为关注图像风格的运用；同时，不同风格及其组合会对美学评价产生影响。某些摄影风格有助于提升图像美感，而如双色调或动态模糊等风格较难掌控，容易引发负面观感。

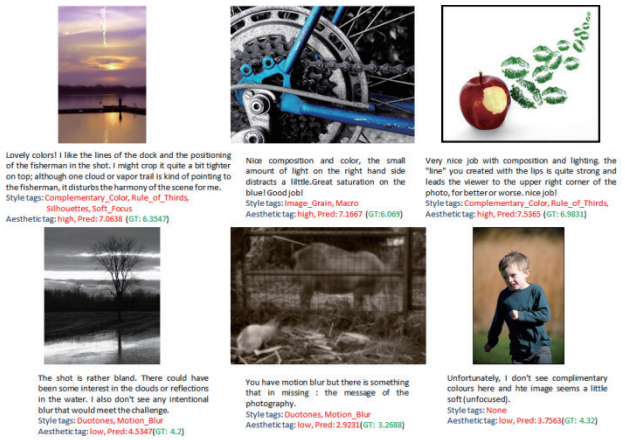

图 3 部分美学预测结果展示

4 结论

本文提出了一种新颖的图文美学多任务分析模型 MCN-MTL，基于多模态协同学习框架。首先，从多模态数据中分别提取图像和文本的局部特征，形成向量矩阵，作为协同注意力网络 MCN 的输入。在 MCN 中，基于 Transformer 结构构建了两个模块，自注意力模块 STRM 与协同注意力模块 CTRM，并据此设计了多模态注意力层 MTRL。通过堆叠 MTRL 层，实现图像与文本模态的深层交互学习。将 MCN 网络不同模态的输出拼接后，生成统一的多模态特征表示，用于后续美学评分预测。此外，模型采用多任务学习策略，将美学评价与多种图像风格预测任务联合建模，以进一步提升性能。在实验方面，基于大规模图像美学数据集扩展构建了多模态数据集 AVA-m，引入图像评论文本作为额外信息。对比实验显示，MCN-MTL 在各项指标上均优于现有方法。本文还通过对 STRM、CTRM 模块及 MTRL 层数的消融实验，进一步验证了模型结构设计的有效性与合理性。

参考文献：

[1] SAMII A, MECH R, LIN Z. Data-driven automatic cropping using semantic composition search [J]. Computer graphics forum, 2015, 34(1):141-151.

[2] MURRAY N, MARCHESOTTI L, PERRONNIN F. AVA: a large-scale database for aesthetic visual analysis [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012:2408-2415.

[3] LU X, LIN Z, SHEN X H, et al. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation [C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE,2015:990-998.

[4] HOSU V, GOLDLUCKE B, SAUPE D. Effective aesthetics prediction with multi-level spatially pooled features [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2019: 9367–9375.

[5] LIU D, PURI R, KAMATH N, et al. Composition-aware image aesthetics assessment [C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV).Piscataway: IEEE, 2019: 3558-3567.

[6] ZHOU Y, LU X, ZHANG J P, et al. Joint image and text representation for aesthetics analysis [C]//MM'16: Proceedings of the 24th ACM international conference on Multimedia. NewYork: ACM, 2016:262-266.

[7] HII Y L, SEE J, KAIRANBAY M, et al. Multigap: multi-pooled inception network with text augmentation for aesthetic prediction of photographs [C]// 2017 IEEE International Conference on Image Processing(ICIP). Piscataway: IEEE, 2017: 1722-1726.

[8] WANG W S, YANG S, ZHANG W S, et al. Neural aesthetic image reviewer [J]. IET computer vision, 2019, 13(8):749-758.

[9] KONG S, SHEN X H, LIN Z, et al. Photo aesthetics ranking network with attributes and content adaptation [C]//Computer Vision–ECCV 2016. Berlin: Springer, 2016:662-679.

[10] KAO Y Y, HE R, HUANG K Q. Deep aesthetic quality assessment with semantic information [J]. IEEE transactions on image processing, 2017, 26(3):1482-1495.

[11] PAN B W, WANG S F, JIANG Q S. Image aesthetic assessment assisted by attributes through adversarial learning [C]//AAAI'19/IAAI'19/EAAI'19: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI, 2019: 679-686.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. (2023-08-02)[2024-10-11].https://doi.org/10.48550/arXiv.1706.03762.

[13] TAN H, BANSAL M. LXMERT: learning crossmodality encoder representations from transformers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Brussels: ACL, 2019: 5100-5111.

[14] A digital photography contest.[DB/OL]. (2021.3.20)[2024-06-19].http://www.dpchallenge.com .

【作者简介】

苗好田（1990—），男，辽宁沈阳人，博士研究生，副教授、教师，研究方向：机器视觉，email:mht@sylu.edu.cn。

梁嘉欣（1999—），女，山西阳泉人，硕士研究生，研究方向：图像处理技术，email:liangjiaxin12_31@163.com。

（收稿日期：2025-04-07 修回日期：2025-05-07）