基于全局与局部特征融合的图像中干扰物移除方法

石昌鑫[1] 冯继凡[1] 杨 清[1]
SHI Changxin FENG Jifan YANG Qing

摘 要 如今受摄影者审美水平以及拍摄环境的限制，图片中存在一些干扰物，使得图片背景混乱，层次复杂，导致图片美感降低。虽然可以通过图像编辑软件的平滑褪色等工具进行去除，但需要掌握相关专业软件的操作知识，并且复杂的操作流程会给人们带来极大的不便。因此，合理并准确地识别并移除图片中影响美感的干扰物，可以极大地提升日常拍摄图片的美感。针对上述问题，提出了一种基于全局与局部特征融合的干扰物移除算法，构建了一个数据集，完成了图片中包含语义信息物体的标注，分为图片主体与干扰物。首先对原始图片进行目标检测，确定图片中包含语义信息的感兴趣区域；接着提取原始图片的全局特征与感兴趣区域的特征并完成融合分类；最后将分类出的干扰物区域输入图片修复模型，完成干扰物的移除。实验结果证明，所提出的算法可以完成对图片中主体与干扰物的分类，在实际应用中取得较好表现。

关键词 图像美学；干扰物移除；卷积神经网络；美学特征；特征融合；图像修复

doi：10.3969/j.issn.1672-9528.2024.05.015

0 引言

审美是人们与生俱来的能力，利用人工智能技术让计算机感知“美”、发现“美”和生成“美”，是一项有意义的研究。近年来，随着人民物质生活水平的提高以及微信、抖音等社交平台的广泛普及，人们随手拍摄照片记录美好瞬间并分享在社交平台的行为越来越普遍[1-3]。然而，大多数情况下，随手拍摄的图片不一定美观，需要一定程度上的美学编辑才能更好地满足人们分享的需求。在绝大多数场景下，图像干扰物会使所拍图片背景复杂混乱，并分散欣赏者对图片主体的注意力，严重影响图片的观赏性。此时，自动检测图像干扰物并完成移除的算法，可以帮助人们避免使用复杂的图像编辑软件，快速完成干扰物移除。人们可以方便快捷地将自己所拍摄的图片提高美感后分享在社交平台上。

Ohad 等人[4]（2015）首次提出了一种新的计算机视觉任务，称为“干扰物预测”，但现有有关干扰物预测的方法，还有着较大的提升空间，存在着较多的错误预测的情况。Ohad 等人的方法对图片硬阈值的分割还会将完整的物体进行分割，其中部分预测为干扰物。为了避免这种情况，本文提出的算法将干扰物重新描述为包含语义信息的、将注意力从图片主体转移并影响图片美学的区域。设计的融合全局与局部特征的算法，充分利用全局信息与感兴趣区域的局部信息，使网络能够学习到图片中主体与干扰物的细粒度的差别。基于全局与局部特征融合的图像中干扰物移除算法可以检测到图片中的干扰物并完成移除，经过处理后的图片背景简单明了，通过背景烘托使图片主体清晰明确，美学体验显著提高。

1 算法框架

本文所提出的基于全局与局部特征融合的图像中干扰物移除方法，预测框架如图 1 所示。

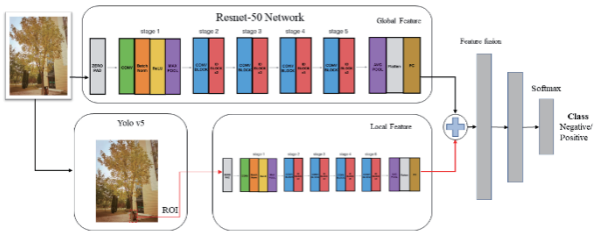

图 1 图片主体与干扰物预测框图

本算法主要分为三个模块：感兴趣区域检测模块、特征提取与分类模块、图像修复模块。具体来讲，首先通过感兴趣区域检测模块即目标检测将图片中包含语义信息的物体检测出来，此区域为感兴趣区域；接着通过深度学习网络提取原图的全局特征以及感兴趣区域的局部特征，同时采取普通拼接特征与基于注意力融合特征两种方式进行特征的融合，并将感兴趣区域分类出图片主体与干扰物两种；最后将分类出的干扰物区域通过图像修复达到干扰物移除的目标。移除

1. 中国航空工业集团公司西安航空计算技术研究所
陕西西安 710065

干扰物的图片，相较原图背景更加简洁，层次更为清晰，主体更为突出，整体美感大幅提升。

2 数据集的构建

针对本文提出的基于全局与局部特征融合的图像中干扰物移除方法，需构建一个数据集。具体来讲，通过网络爬虫技术 Selenium 在 Pexels[5] 网站爬取大量图片。该网站是一个免费的高质量无版权图片素材网站，每张图片均包含详细的信息，如拍摄的相机型号、光圈、聚焦、IOS、像素等。此网站的图片均为专业摄像师经过筛选的质量较高的图片，因此其图片中的人物大部分属于本文分类的正例。为了使本数据集中的正反例均衡分布，还收集了一部分非专业摄影师的日常生活或旅行时的图片扩充本数据集。

图片通过目标检测模型，确定出图片中包含语义信息的感兴趣区域进行补充标注，注释分为图片主体与干扰物两类。通过不同场景、不同风格的筛选，数据集图片数量为 2168 张，标注数据为 4398 组用于实验，即该数据集中平均每个图片约包含 2 个注释，数据类型分布如图 2 所示。其中，正例即图片主体为 1704 组，反例即干扰物为 2694 组，正反例约为 1:2。图片中干扰物的存在形式、位置以及大小具有多样性。从图像美学的角度来讲，干扰物颜色、构图以及场景等特征也与图片主体有较大差异。因此，本文在构建数据集的过程中充分考虑了各类干扰物的多样性。

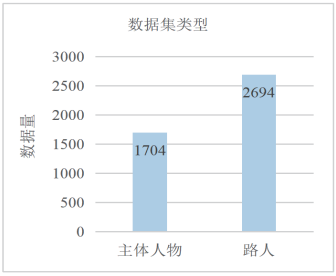

图 2 数据集类型分布

本文创建的标注任务也具有较大的主观性，因此每张图片依然会得到本课题组五名同学的标注，得到最后的注释。本文为数据集的标注者提供了 labelimg 标注工具 [6]，标注者将每张图片中的物体进行框选，随后将该感兴趣区域分类，如图 3 所示。

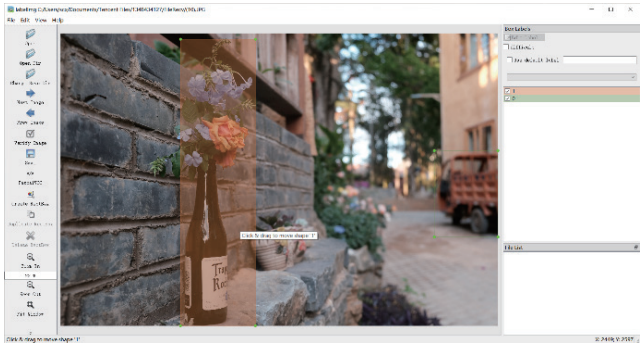

图 3 标注示意图

3 算法模块分解

3.1 感兴趣区域的提取

本模块的作用为确定输入图片中包含语义信息物体的区域，需对图片整体进行目标检测，故采用目前最快最准确的目标检测模型 YOLO-v5[7]。其网络结构包括输入端、backbone、neck 和 prediction 四部分。其中，backbone 是卷积神经网络，用于在不同图像细粒度上聚合得到图像特征。neck 是一系列网络层，用于混合和组合图像特征，并将其传递到预测层。prediction 层用于对图像特征进行预测，生成边界框和预测类别。

3.2 特征提取模块

在本章提出的方法中，通过基于卷积网络的深度学习方法，使用相同的卷积架构分别提取图片的全局特征以及感兴趣区域的局部特征，这有助于算法对不同的感兴趣区域进行特征差异化的计算。

3.2.1 特征的提取

具体来说，本节使用 ResNet[8] 策略来加深网络层数，提取全图更加细粒度的特征 Global_feature 以及 Local_feature，采取 ResNet50[9] 作为主干网络来进行特征的提取。

若输入输出的特征图数量一致时，残差块分成直接映射部分和残差部分两部分，x_l 是直接映射，$F(x_l, W_l)$ 是残差部分，一个残差块公式为：

$$x_{l+1} = x_l + F(x_l, W_l) \tag{1}$$

当输入输出的特征图数量不一致时，为了能够匹配它们的通道数，需要进行 1*1 卷积操作来升维或降维。这样做的目的是确保在网络层之间进行有效的信息传递，并且避免由于通道数量不一致而导致的错误或信息丢失，一个残差块公式为：

$$x_{l+1} = h(x_l) + F(x_l, W_l) \tag{2}$$

式中：$h(x_l)=W_l'x$，W_l' 是 1*1 卷积操作，$F(x_l, W_l)$ 是残差部分，其通常包含三个不同的卷积操作，分别为 1*1、3*3、1*1 三个卷积核。每一个卷积核都被放置在一个批量归一化（BN）层和一个 RELU 激活函数之间，其作用是在保留有效特征的同时进一步防止过拟合。这种卷积核组合提供了一种提取特征的有效方法。

3.2.2 注意力机制模块

本模块是在提取全局特征时增加的，可以使计算机更加准确地注意到图片中感兴趣区域的特征。通过使用注意力机制，本文的模型可以加权地给予更多的关注和重要性给图片中感兴趣区域的特征。这样，即使在图片中存在干扰或者不相关的特征，本文的模型仍然可以确定感兴趣区域的语义信息。最终，通过深度融合的特征作为输入，本文的模型将实现更加准确的感兴趣区域是否为干扰物的分类。

CABM[10] 是一个有效且轻量级的注意模块，由于该模块

输入端与输出端的特征维度是一样的，因此可以集成在任一个卷积神经网络架构中。具体而言，CBAM 模块分别在中间特征图的两个独立维度上逐步生成注意力图，然后将该注意力图与输入特征图相乘，从而实现自适应的特征优化。从结构上来说，CBAM 模块结合了空间注意模块和通道注意模块。

3.3.3 局部与全局特征融合模块

本节将会介绍如何使用注意力机制来融合从 ResNet 提取的图像的全局特征以及局部特征。注意力机制可以在特征融合过程中更好地捕捉不同特征的重要性，并将不同特征的贡献度加以区分。使用这种方法，可以提高图像特征的表征能力，并进一步提升分类和检测任务的性能。

具体来说，注意力融合方法是基于计算每个特征分量相对于总体特征的重要性，并使用 Softmax 函数来实现注意力系数的计算。在计算每个分量的注意力系数时，将所有分量组合成一个向量，然后计算每个分量与总体向量之间的相似度。这样，就可以得到一个注意力向量，其中每个分量表示对应分量的重要性。最后，将注意力向量与每个特征分量相乘，即可得到融合后的特征向量。这种注意力机制的融合方法能够充分利用图像的全局和局部信息。

$$F_{att}=\sum_{l=1}^{L\_global+L\_region} v_i^l F^{\{global,region\}} \tag{3}$$

$$F^{\{global,region\}}=contact(H^{global},H^{region}) \tag{4}$$

$$v_i^l=softmax(a_i^l)=\frac{exp((a_i^l)^T H)}{\sum_{j=1}^{L\_global+L\_region}\exp((a_i^j)^T H)} \tag{5}$$

$$a_i^j=\tanh(W_{att}f_i^j+b_{att}) \tag{6}$$

式中：F_{att} 是融合后的特征向量；v_i^l 表示第 l 级融合层表示的归一化注意力权重；f_i^j 是 l 层的图像表示特征；W_{att} 和 b_{att} 分别是可学习的权重矩阵和偏差项；H 是局部特征的特征向量，用于指导注意力系数的计算；*contact*() 表示拼接操作。

3.4 移除区域修复

对于前一阶段已经分类出的干扰物，为了达到移除的目标，本模块采取 Yi 等人 [11] 提出的图像修复算法。该算法提出了一种上下文残差聚合（CRA）机制，能够在资源有效的情况下生成超高分辨率的修复结果。由于内存限制，基于数据驱动的图像修复方法通常只能处理低于 1 K 分辨率（1920×1080）的图片。而现在使用移动设备拍摄的照片分辨率已经提高到了 8 K，如果单纯对低分辨率图片进行修复再进行上采样，只会产生较大且模糊的结果。但是，将高频残差图像添加到较大的模糊图像上，可以产生清晰的结果，增加细节和纹理。这种方法可以通过对上下文补丁中的残差进行加权聚合来生成丢失内容的高频残差，因此只需要网络的低分辨率预测即可。神经网络的卷积层只需要在低分辨率的输入和输出上进行操作，因此它可以有效地减少内存和计算能力的成本，并减轻对高分辨率训练数据集的需求。本文提出的模型在分辨率为 512×512 的小图像上进行，并在高分辨率图像上进行推理，从而达到令人信服的修复质量。该模型可以修补具有相当大孔尺寸的 8 K 图像，这对于以前的基于学习的方法来说是很难解决的。

4 实验结果与分析

4.1 实验设置

本文提出的算法将在所构建的数据集中进行测试，数据集中 88% 数据作为训练，12% 数据作为测试集进行实验。通过 ResNet50 主干网络分别对全图与感兴趣区域提取特征，对全局特征的提取增加注意力机制，使计算机加权地给予更多的关注和重要性给图片中感兴趣区域的特征。对于全局与局部特征的融合，本文也分别对比了普通拼接特征与使用注意力机制融合特征是否为干扰物的分类结果，通过精确率（precision）、召回率（recall）、准确率（accuracy）等评估指标来测试算法的表现。最后展示本算法进行干扰物移除后的结果实例。

4.2 实验结果

该部分将评估本文提出的基于全局与局部特征融合的图像中干扰物分类模型的性能。表 1 和表 2 分别给出了两种融合模型对图像中分类主体与干扰物的测试结果，其内容为评价图片中主体与干扰物分类的混淆矩阵。矩阵纵列的“主体”与“干扰物”代表预测分类结果，横行的“主体”与“干扰物”代表实际分类结果。

表 1 普通融合特征模型的混淆矩阵

	主体	干扰物
主体	94	55
干扰物	119	274

表 2 基于注意力融合模型的混淆矩阵

	主体	干扰物
主体	99	50
干扰物	97	296

表 3 是二分类问题的常用评价指标，将两种融合模型进行对比。

表 3 两种融合方式的比较

	普通融合模型	基于注意力融合模型
精确率（precision）	0.441 3	0.505 1
召回率（recall）	0.630 8	0.664 4
准确率（accuracy）	0.678 9	0.728 7
F_1 值	0.519 3	0.573 9

具体来讲，精确率是指预测为正的样本中真正为正的比例，召回率是指所有正样本中被正确预测的比重，准确率是指正确预测的样本占所有样本的比重。F_1 值为精确率和召回率的兼顾指标，是精确率和召回率的调和平均数。此处，将

图片中的“主体”认为正类预测，“干扰物”认为负类预测。通过这四个指标来评估本文算法的性能。

两种分类模型的 ROC 曲线如图 4 所示，可看出采取注意力融合的模型具有更好的分类性能。

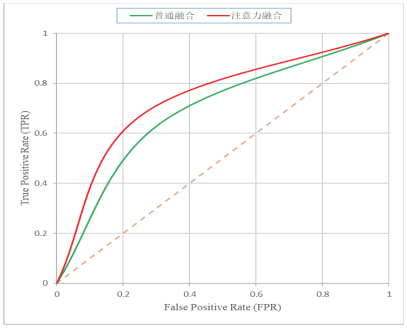

图 4 分类模型 ROC 图

4.3 结果实例

为了显示基于全局与局部特征融合的图像中干扰物移除方法的优越性，以及直观展示本算法的预测效果，本节给出了从构建数据集的测试数据中挑选出来的图片进行干扰物移除操作的结果，如图 5 所示。可以直观看出，该算法可以完成干扰物的识别判断，并完成了移除工作。将干扰物移除后的图片，主体清晰明确，背景简单明了，层次明显，美感体验显著提高。

（a）原图 1

（b）结果图 1

（c）原图 2

（d）结果图 2

图 5 干扰物移除结果实例

5 总结

本文提出的基于全局与局部特征融合的图像中干扰物移除算法，针对人们日常拍摄时面对的更普遍的影响美感的干扰物，这些干扰物会转移人们对主体的注意力，从而影响图片的层次感。将干扰物移除，可以使图片背景变得干净整洁，图片层次分明，提升美学体验。实验结果表明，本文提出的分类模型能够识别并分类出图片中的干扰物，通过结果实例可以直观看出所提出的方法可以有效进行图片中干扰物的移除。

参考文献：

[1] 李雪薇 . 基于美学的图像质量评价与提升算法研究 [D]. 北京 : 北京邮电大学 ,2021.

[2] 金鑫 , 周彬 , 邹冬青 , 等 . 图像美学质量评价技术发展趋势 [J]. 科技导报 ,2018,36(9):36-45.

[3] 祝汉城 , 周勇 , 李雷达 , 等 . 个性化图像美学评价的研究进展与趋势 [J]. 中国图象图形学报 ,2022,27(10):2937-2951.

[4]FRIED O, SHECHTMAN E, GOLDMAN D, et al. Finding distractors in images[C]//Proceedings of the IEEE Conference on Computer Vision and pattern Recognition. Piscataway:IEEE,2015:1703-1712.

[5]HEALEY C, ENNS J. Large datasets at a glance: combining textures and colors in scientific visualization[J]. IEEE transactions on visualization and computer graphics, 1999,5(2):145-167.

[6]AFIF M, AYACHI R, PISSALOUX E, et al. A novel dataset for intelligent indoor object detection systems[J]. Artificial intelligence advances,2019,1(1):52-58.

[7]BOCHKOVSKIY A, WANG C, MARK L, et al.YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2024-01-22].https://arxiv.org/abs/2004.10934.

[8]HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway:IEEE,2016:770-778.

[9]HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands.Berlin:Springer, 2016:630-645.

[10]WOO S, PARK J, LEE J, et al. Cbam: convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV).Berlin:Springer,2018:3-19.

[11]YI Z, TANG Q, AZIZI S, et al. Contextual residual aggregation for ultra high-resolution image inpainting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.Piscataway:IEEE,2020:7505-7514.

【作者简介】

石昌鑫（1998—），男，陕西渭南人，硕士，助理工程师，研究方向：计算机视觉、图像美学。

冯继凡（1999—），男，陕西咸阳人，硕士，助理工程师，研究方向：计算机视觉。

杨清（1998—），男，陕西榆林人，硕士，助理工程师，研究方向：计算机视觉。

（收稿日期：2024-02-26）