软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(11):4379−4395 [doi: 10.13328/j.cnki.jos.006325]

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

多尺度拼图重构网络的食品图像识别*

刘宇昕[1,2], 闵巍庆[1,2], 蒋树强[1,2], 芮 勇[3]

[1](中国科学院 智能信息处理重点实验室 (中国科学院 计算技术研究所), 北京 100190)

[2](中国科学院大学, 北京 100049)

[3](联想集团, 北京 100085)

通信作者: 蒋树强, sqjiang@ict.ac.cn

摘 要: 近年来, 食品图像识别由于在健康饮食管理、无人餐厅等领域的广泛应用而受到了越来越多的关注. 不同于其他物体识别任务, 食品图像属于细粒度图像, 具有较高的类内差异性和类间相似性, 而且食品图像没有固定的语义模式和空间布局, 这些特点使得食品图像识别更具挑战性. 为此, 提出了一种用于食品图像识别的多尺度拼图重构网络(multi-scale jigsaw and reconstruction network, MJR-Net). MJR-Net 由拼图重构模块、特征金字塔模块和通道注意力模块这 3 部分组成. 拼图重构模块使用破坏重构学习方法将原始图像进行破坏和重构, 以提取局部的判别性细节特征; 特征金字塔模块可以融合不同尺寸的中层特征, 以捕获多尺度的局部判别性特征; 通道注意力模块对不同特征通道的重要程度进行建模, 以增强判别性的视觉模式, 减弱噪声干扰. 此外, 还使用 A-softmax 和 Focal 损失, 分别从增大类间差异和修正分类样本的角度优化网络. MJR-Net 在 ETH Food-101, Vireo Food-172 和 ISIA Food-500 这 3 个食品数据集上进行实验, 分别取得了 90.82%, 91.37%和 64.95%的识别准确率. 实验结果表明, 与其他食品图像识别方法相比, MJR-Net 表现出较大的竞争力, 并在 Vireo Food-172 和 ISIA Food-500 上取得了最优识别性能. 全面的消融实验和可视化分析证明了该方法的有效性.

关键词: 食品图像识别; 深度学习; 拼图重构; 特征金字塔; 注意力机制

中图法分类号: TP393

中文引用格式: 刘宇昕, 闵巍庆, 蒋树强, 芮勇. 多尺度拼图重构网络的食品图像识别. 软件学报, 2022, 33(11): 4379−4395. http://www.jos.org.cn/1000-9825/6325.htm

英文引用格式: Liu YX, Min WQ, Jiang SQ, Rui Y. Food Image Recognition via Multi-scale Jigsaw and Reconstruction Network. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4379−4395 (in Chinese). http://www.jos.org.cn/1000-9825/6325.htm

Food Image Recognition via Multi-scale Jigsaw and Reconstruction Network

LIU Yu-Xin[1,2], MIN Wei-Qing[1,2], JIANG Shu-Qiang[1,2], RUI Yong[3]

[1](Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (Institute of Computing Technology, Chinese Academy of Sciences), Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

[3](Lenovo Group, Beijing 100085, China)

Abstract: Recently, food image recognition has received more and more attention for its wide applications in healthy diet management, smart restaurant, and so on. Unlike other object recognition tasks, food images belong to fine-grained ones with high intra-class variability and inter-class similarity. Furthermore, food images do not have fixed semantic patterns and specific spatial layout. These make food recognition more challenging. This study proposes a multi-scale jigsaw and reconstruction network (MJR-Net) for food recognition. MJR-Net is composed of three parts. The jigsaw and reconstruction module uses a method called destruction and reconstruction learning

* 基金项目: 国家自然科学基金(61972378, U1936203, U19B2040)

收稿时间: 2020-09-23; 修改时间: 2020-12-01, 2021-01-11; 采用时间: 2021-02-13

to destroy and reconstruct the original image to extract local discriminative details. Feature pyramid module can fuse mid-level features of different sizes to capture multi-scale local discriminative features. Channel-wise attention module can model the importance of different feature channels to enhance the discriminative visual patterns and weaken the noise patterns. The study also uses both A-softmax loss and Focal loss to optimize the network by increasing the inter-class variability and reweighting samples respectively. MJR-Net is evaluated on three food datasets (ETH Food-101, Vireo Food-172, and ISIA Food-500). The proposed method achieves 90.82%, 91.37%, and 64.95% accuracy, respectively. Experimental results show that, compared with other food recognition methods, MJR-Net shows greater competitiveness and especially achieves the state-of-the-art recognition performance on Vireo Food-172 and ISIA Food-500. Comprehensive ablation studies and visual analysis also prove the effectiveness of the proposed method.

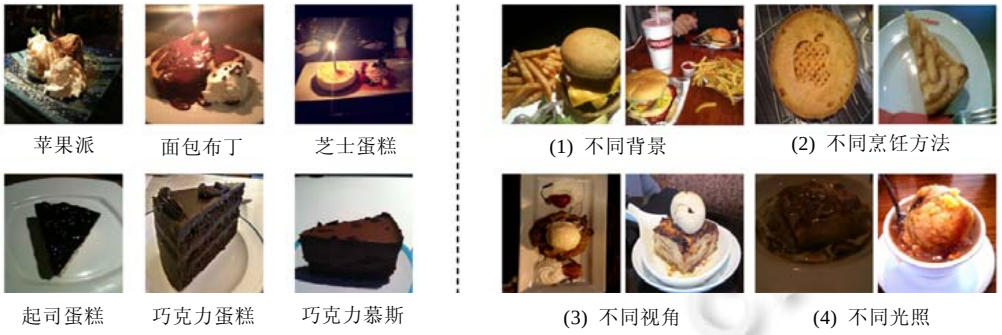
Key words: food image recognition; deep learning; jigsaw and reconstruction; feature pyramid; attention mechanism

食品对人类的生存、健康和福祉有重要意义. 当下, 越来越多的人由于不合理的饮食习惯饱受由肥胖引起的各种疾病(如心血管性疾病、糖尿病、各种癌症等)的困扰, 合理膳食显得尤为重要. 此外, 食品在文化传承上具有重要意义, 在定义我们的身份、社会地位、宗教意义和文化方面起着重要作用[1]. 得益于深刻的现实意义和广泛的应用前景, 近年来, 越来越多的研究者开始从事食品计算[2]的研究. 食品图像识别(food image recognition)是食品计算的核心任务之一, 指的是在自然的拍摄环境下, 给定一张食品图像, 预测该食品所属的细粒度类别(如番茄意面、巧克力蛋糕、鱼香肉丝等), 是一类特殊的细粒度图像识别任务. 食品图像识别是各种与食品相关的应用的基本和首要步骤, 在许多领域中都有着广泛的应用. 在健康饮食管理系统[3]中, 食品图像识别通过自动识别食品类别, 可进一步实现后续的营养分析和卡路里估算. 在智能厨具[4−6]中, 食品图像识别在智能冰箱的食品种类、新鲜度识别和智能厨房的食物追踪、食材识别中发挥着重要的作用. 在智能餐厅[7]中, 食品图像识别可以实现个性化餐饮推荐及自动结账, 并能根据顾客所选择的菜品及食用的食物量来衡量其饮食偏好. 在餐饮推荐[8]中, 食品图像识别对理解用户需求及改进个性化推荐结果有重要影响. 食品图像识别在许多领域中都有着广泛的应用.

现有的食品识别方法[9−11]大都直接使用 CNN 提取的视觉特征来进行食品图像识别, 而没有考虑食品图像自身的特点. 不同于其他的物体识别任务, 食品图像识别有其独特的挑战.

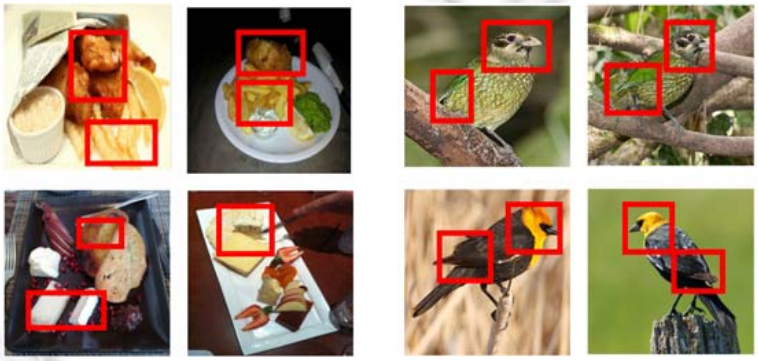
- 首先, 食品图像属于细粒度图像. 一方面, 由于不同种类的食品在原料或烹饪方法上具有一定的重合性, 使得食品图像呈现出较高的类间相似性(如图 1(a)所示, 其中, 每一行指示具有相似视觉外观的不同食品); 另一方面, 受背景、烹饪方法、视角、光照等因素的影响, 食品图像还呈现出较高的类内差异性(如图 1(b)所示). 因此, 能否准确地识别具有高度混淆性的食品类别, 依赖于对相似食品图像局部判别性特征的提取和学习. 近年来, 拼图机制在提取图像的局部细节特征方面表现出优异的性能, 其原理为: 将原始图像分割为均匀的、尺寸较小的子图像并随机打乱, 以强迫网络更关注局部细节特征而非全局结构, 从而增强网络对局部细节特征的学习能力. 受其启发, 本文使用一种带有对抗损失和重构损失的拼图机制来提取食品图像的局部判别性特征, 将拼图机制应用到食品图像识别任务中.
- 其次, 不同于常规的细粒度识别任务, 食品图像没有固定的语义模式, 其判别性细节特征呈现出多尺度、不规则的特点. 现有的细粒度识别方法将局部细节特征(如鸟类的头部和飞机的机翼等)看作固定语义, 通过提取这些语义特征进行识别. 但是食品图像不具有固定的语义模式, 很难在不同的食品之间提取公共的语义特征. 如图 2 所示, 很难像定义鸟类的头部、翅膀(图 2(b)中方框部分)一样定义食品的固定语义部分(图 2(a)中方框部分), 不同种类食品图像的局部细节特征在尺度和形状上具有较大的差异性. 因此, 食品图像识别还依赖于充分提取和学习多尺度和不规则的细节特征. 受限于固定的分割尺寸, 拼图机制只适合于学习固定大小和形状的局部特征, 而对尺度和形状变化较大的判别性细节不敏感. 基于此, 本文将特征金字塔引入到食品图像识别任务中, 并使用辅助训练的方式对主干网络进行优化, 旨在克服拼图机制无法充分提取多尺度特征的问题, 更好地适应食品图像判别性细节特征多尺度、不规则的特点.
- 最后, 食品图像是非刚性的, 缺少统一的空间布局且具有较多的噪声. 这些噪声会严重损害网络的特征学习能力, 进一步加剧相似类别之间的混淆性. 食品图像不同的视觉模式存在于特征图的不同通

道之中, 有的通道蕴含判别性的视觉模式, 有的通道则表示噪声. 因此, 本文还考虑将通道注意力机制引入食品识别框架中, 通过对不同的特征图进行重加权以增强判别性的视觉模式, 降低噪声干扰.



(a) 食品图像的类间相似性 (b) 食品图像的类内差异性

图 1 食品图像的类间相似性和类内差异性特性示例



(a) ETH Food-101 数据集 (b) CUB200-2011[12]鸟类数据集

图 2 ETH Food-101 数据集与 CUB200-2011[12]鸟类数据集部分图像对比

基于上述考虑, 本文提出了一种多尺度拼图重构网络 MJR-Net, 用于食品图像识别. 该网络能够提取食品图像中多尺度的局部判别性特征, 并增强对分类有利的视觉模式, 降低噪声干扰. 具体地, MJR-Net 由拼图重构模块、特征金字塔模块和注意力模块这 3 个部分组成. 考虑到食品图像的细粒度特点, 本文将破坏重构学习(destruction and construction learning, DCL)方法作为拼图重构器来挖掘食品图像的局部判别性特征. DCL 使用拼图机制将原始图像划分为大小相同的子图像并随机打乱, 以强迫网络关注局部区域而非全局结构, 进而增强主干网络的局部判别性特征学习能力. 考虑到食品图像的局部细节特征呈现出不规则和多尺度的特点, 本文提出使用特征金字塔模块(feature pyramid module, FPM)来融合多尺度的特征表示. 主干网络中, 不同深度的中层特征对应不同的感受野: 浅层特征的分辨率较高, 感受野较小, 含有较多的小尺度特征; 而深层特征分辨率较小, 感受野较大, 含有较多的大尺度特征. FPM 融合这些中层特征以生成多尺度的特征表示, 从而使得网络能够学习不同尺度的细粒度视觉特征. 最后, 考虑到食品图像非刚性和多噪声的特点, 本文提出了一种通道注意力模块(channel-wise attention module)用于建模不同特征的重要程度, 以突出判别性视觉模式, 减弱噪声干扰. 最后, 本文还使用 A-softmax 和 Focal 损失分别从增加类间差异性和修正分类样本的角度对网络进行训练.

本文的贡献主要总结为以下两个方面.

(1) 本文提出了一种用于食品图像识别的多尺度拼图重构网络 MJR-Net. 该网络能够提取多尺度的细粒度判别性特征, 并增强有判别性的视觉模式, 减弱噪声干扰. MJR-Net 能够很好地适应食品图像的

特点, 在食品识别任务上具有良好的表现.

(2) 本文在 3 个食品数据集 ETH Food-101[13], Vireo Food-172[14], ISIA Food-500[15]上评估了所提出的方法. 在 Vireo Food-172 和 ISIA Food-500 上取得了最优的识别性能.

本文第 1 节介绍食品图像识别和细粒度图像识别的相关工作. 第 2 节介绍本文提出的 MJR-Net. 第 3 节为实验验证和结果分析. 第 4 节为本文的总结展望.

1 相关工作

1.1 食品图像识别

按照特征的类型划分, 食品图像识别方法可以分为基于手工特征的识别和基于深度特征的识别两种类型. 早期的研究采用手工特征来进行食品图像识别. 例如, Chen 等人[16]使用颜色直方图和 SIFT[17]特征词袋模型进行食品图像识别. 随着计算机视觉技术的发展, 基于 CNN 的食品图像识别方法逐渐成为主流. Kagaya 等人[10]使用 AlexNet 网络[18]提取图像特征来进行食品检测和识别. Ming 等人[11]则使用 ResNet50[19]来进行食品图像识别. 这些方法都直接使用现有的 CNN 网络来直接提取图像特征, 而没有考虑食品图像自身的特点, 因此识别性能并没有达到最优. 近年来, 一些研究开始着手设计针对食品图像的专用深度网络. Martinel 等人[20]提出了 WISeR 的网络, 该网络由广度残差网络(wide residual networks, WRN)[21]分支和切片卷积网络分支两部分组成, WRN 分支用于提取食品图像一般性的视觉特征, 而切片卷积分支则提取食物图像的垂直结构(如汉堡、比萨等), 两个分支的特征级联以得到最终的特征表示. 得益于融合食品图像的全局特征和垂直结构特征, WISeR 在当时的多个数据集上取得了最优的识别性能. 但该方法更适合带有垂直结构的食品类别, 比如西餐, 没有总结出食品图像更一般性的特点. Jiang 等人[9]提出了 MSMVFA 网络, 该网络是一种两级特征融合网络, 首先将从图像中获得的高级语义特征、深度视觉特征和从食材中获得的中级属性特征进行融合, 然后进行多尺度特征融合来进行食品图像识别. 该方法利用食材作为监督信息来提高分类性能, 但由于食品原材料和烹饪方法的多样性, 很难为所有的食品标注明确的食材信息, 同类的食品也可能由不同的食材组成, 因此方法的通用性不高. Liang 等人[22]提出了一种多级卷积特征金字塔的细粒度食品图像识别方法, 该方法由级联的三级食品特征提取网络组成, 每级之间使用注意力定位网络将图像按照全局-食品对象-局部特征进行裁剪, 特征金字塔用于定位高分辨率的细粒度区域. 该方法使用三级主干网络来进行特征提取, 但其计算复杂度和空间成本较高, 不利于方法的拓展和应用. 本文的方法受其启发, 但有以下不同: 考虑到金字塔特征的通道数较小, 不足以完整表示食品图像复杂的视觉模式, 本文使用特征金字塔对主干网络进行辅助训练, 而不直接作为特征表示, 并且添加了光滑函数以防止特征的过拟合.

1.2 细粒度图像识别

细粒度图像识别主要研究如何识别一个大类下的不同子类, 例如识别不同物种的鸟类、不同类型的汽车等. 本文所研究的食品识别也属于细粒度识别. 早期的细粒度方法多为强监督的方法, 需要对数据集进行边界框和大量关键点的标注. 例如, Huang 等人[23]使用堆栈式的 CNN 来进行细粒度图像识别. Zhang 等人[24]通过训练两个检测器, 即物体检测器和局部细节检测器来进行细粒度的图像识别. 随着研究的深入, 使用弱监督的方法来进行细粒度识别成为主流. Zheng 等人[25]提出了 MA-CNN 网络, 旨在通过相互促进局部细节区域的定位和特征学习两个过程来进行细粒度图像识别. Yang 等人[26]提出了 NTS-Net, 旨在基于强化学习自监督地定位判别性的局部区域来进行细粒度识别. Chen 等人[27]提出了 DCL, 通过将原图像进行破坏并重构来提取局部判别性的视觉特征来进行细粒度图像识别. 本文的方法受 DCL 启发, 但有以下几点不同.

(1) 食品图像没有固定的语义模式, 局部判别性特征呈现多尺度和不规则的特点. 受限于固定的破坏尺寸, 原始的 DCL 仅能提取部分细粒度特征, 它们不足以正确区分混淆的食品类别. 因此, 本文使用特征金字塔网络来捕获多尺度的局部判别性特征.

(2) 食品图像没有统一的空间布局, 其视觉外观相比于常规的细粒度数据集(如 CUB200-2011 等)更为复

杂, 且含有较多噪声. 因此, 本文使用通道注意力来对不同特征图的重要程度进行建模, 以增强判别性的视觉模式, 降低噪声干扰.

(3) 本文还使用 A-softmax 和 Focal 损失分别从增大类间差异和修正分类样本的角度优化网络, 以进一步提升识别性能.

2 方法模型

本文提出的 MJR-Net 主要由破坏重构学习(包括区域混淆模块、对抗学习网络和区域对齐网络)、特征金字塔模块、通道注意力模块这 3 个部分组成. 网络结构如图 3 所示, 对于输入的食品图像, 首先使用区域混淆模块将其打乱, 以生成破坏图像, 如图 3(a)所示. 原图和破坏图像然后被输入到主干网络中以实现特征提取. 特征金字塔模块从主干网络的不同层中获取不同尺寸的中层网络特征并进行融合, 得到的金字塔特征用于对主干网络进行辅助训练, 如图 3(b)所示. 通道注意力模块用于建模特征图中各个通道的重要程度, 并对主干网络的特征图进行重新加权, 如图 3(c)所示. 主干网络可以选择任何一种适合于计算机视觉任务的流行 CNN 网络, 如 ResNet[19], DenseNet[28]和 InceptionV3[29]等. 对抗学习网络(如图 3(d)所示)通过对抗损失拉近原始图像和破坏图像的特征空间, 从而降低随机混淆模块在破坏图像时产生的噪声影响. 为了恢复局部区域的原始空间分布, 区域对齐网络(如图 3(e)所示)用于对局部细节区域之间的语义相关性进行建模. 最后, 分类器(如图 3(g)所示)将根据增强后的特征预测食品的类别标签.

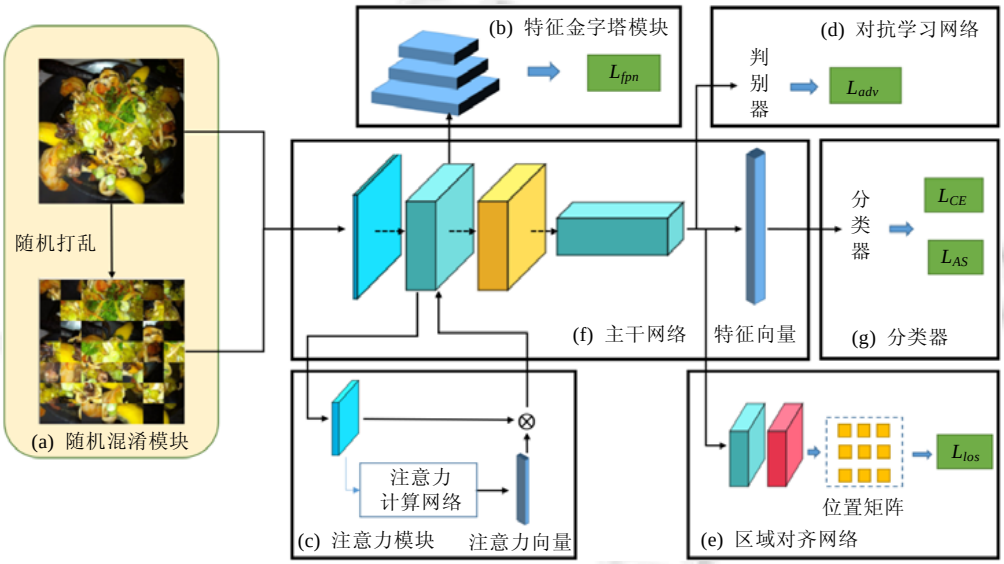


图 3 MJR-Net 网络结构

2.1 破坏重构学习

本文使用破坏重构学习 DCL[27]作为拼图重构器. DCL 由随机混淆模块(region confusion mechanism, RCM)、对抗学习网络和区域对齐网络组成. 这 3 个部分协同工作, 以学习图像的局部判别性特征. 详细的实现细节请参见原始文献.

RCM 通过将原始图像进行局部打乱, 旨在使神经网络不再关注全局结构, 而更多关注局部的判别性特征. 具体地, RCM 首先将图像拆分成大小相同的 $N\times N$ 个子图像, 然后将这些子图像进行随机打乱以生成破坏图像. 图 4 展示了 RCM 模块的功能示意(其中, N=7).

图 4 RCM 功能示意图，第 1 行是原始图像，第 2 行是经过 RCM 后的破坏图像

RCM 生成的破坏图像中存在干扰性的视觉噪声，它们会损害网络学习真正局部细节特征的能力，进而影响分类性能. 对抗学习网络旨在降低由 RCM 引入的噪声对分类的干扰. 该网络包含一个判别器 D，能够对输入的特征向量执行二分类操作，以区分主干网络提取的特征来自原始图像还是破坏图像. 与原始的 DCL 不同，本文使用两层的全连接网络作为对抗学习网络中的判别器，以提高判别器的性能. 对抗损失 L_{adv} 的计算公式如下:

$$L_{adv}=-\sum[\log D[F(X)]+\log D[F(X^{RCM})]] \tag{1}$$

其中，X 为原始图像，X^{RCM} 为破坏图像，$F(\cdot)$为主干网络的计算函数，$D(\cdot)$为判别器的计算函数. 通过最小化 L_{adv}，使得判别器能够分辨当前特征属于原始图像还是破坏图像. 分类损失和判别损失以一种对抗式的方式联合地进行优化，以迫使主干网络拟合两个特征空间的公共部分，以降低噪声的干扰.

区域对齐网络用于恢复局部区域的原始空间布局. 具体地，对于原始图像和破坏图像的输出特征 f 和 f_{RCM}，网络预测其在 RCM 划分后的每个子图像的横纵坐标 $M(X)$和 $M(X^{RCM})$，并通过最小化区域对齐损失 L_{loc} 使主干网络能够对结构信息进行建模. 对齐损失 L_{loc} 计算如下:

$$L_{loc}=\sum_{X\in\Gamma}\sum_{i=1}^{N}\sum_{j=1}^{N}|M(X)_{i,j}-[i,j]^T|+|M(X^{RCM})_{i,j}-[i,j]^T| \tag{2}$$

其中，Γ为输入图像集合，$[i,j]^T$ 为每个子图像的真实位置坐标.

2.2 特征金字塔模块

食品图像缺乏固定的语义模式，其局部细节特征呈现出多尺度的特性. 受限于固定的破坏尺寸和感受野，DCL 无法提取多尺度的判别性特征，因此，本文使用特征金字塔[30]来增强网络的多尺度特征学习能力.

特征金字塔的核心思想为: 融合主干网络中不同深度的中层特征，这些特征蕴含着不同尺度的局部细粒度视觉模式. 使用融合后的金字塔特征对网络模型进行辅助训练，能够增强主干网络多尺度判别性特征学习能力.

特征金字塔由自下而上的路径、桥接路径和自上而下的路径这 3 个部分组成. 自下而上的路径表示主干网络的前馈计算过程. 设主干网络共有 k 个阶段，我们将每个阶段最后一个卷积层输出的特征定义为一个中层特征，于是可以得到特征集合 $C=\{c_1,c_2,\ldots,c_{k-1},c_k\}$.

桥接路径主要用于压缩特征图维数以便于特征融合，每个连接是一个卷积核为 1×1、步长为 1 的卷积层. 经过桥接路径，C 中的每个特征被压缩为相同的维数，我们将压缩之后的特征定义为 $S=\{s_1,s_2,\ldots,s_{k-1},s_k\}$.

自上而下的路径表示特征金字塔的特征融合过程. 设 $P=\{p_2,\ldots,p_k\}$是融合后的一组金字塔特征，则 p_i 的计算公式如下:

$$\begin{cases}p_i=S_i(U(p_{i+1})+s_i)\\ p_k=S_k(s_k)\end{cases} \tag{3}$$

其中, $U(\cdot)$是上采样函数, 用于统一来自不同网络层特征的尺寸. S 是卷积核大小为 3×3 的卷积层, 用于对融合特征进行光滑处理以防止冗余. 反复迭代这一过程, 可以得到最终的金字塔特征 p_{fin}(一般取 $p_{fin}=p_2$).

与常规的特征金字塔的使用方式不同, 本文使用特征金字塔对主干网络进行辅助训练. 主要有以下 3 个原因.

(1) 任务类型的差异性. 在目标检测任务中, 特征金字塔通常被直接用于边界框的回归和对象的分类, 不同尺度的物体可以在对应分辨率的金字塔特征中独立地进行检测. 而食品图像识别属于细粒度的图像识别任务, 需要融合多尺度的局部细节以增强图像的特征表示, 因此不适合将某层的金字塔特征直接作为最终的特征表示.

(2) 使用目的的差异性. 目标检测任务中, 特征金字塔主要用于检测因感受野较大而无法精确定位的小目标, 而在 MJR-Net 中, 特征金字塔主要用于增强主干网络的多尺度特征学习能力, 而非直接作为特征表示, 因此使用辅助训练的方式.

(3) 特征维度的差异. 受限于分辨率, 特征金字塔的维度一般都较小(一般为 256), 这对局部目标的分类和回归是足够的, 但不足以表达含有复杂视觉模式的食品图像.

在实现上, 本文通过对最终的金字塔特征 p_{fin} 引入一个分类损失函数来优化主干网络. 定义分类器 C_1, 其参数为 $W_1\in R^{c\times d}$, 其中, c 为类别数目, d 为 p_{fin} 的通道数目. 于是, 特征金字塔损失函数 L_{fpn} 定义如下:

$$L_{fpn}=-l\times\log\left[softmax(C_1(p_{fin}))\right]=-l\times\log\left[softmax(W_1\times p_{fin})\right] \tag{4}$$

其中, l 是图像的真实类别标签.

因为对不同深度的中间特征进行融合, 特征金字塔能够捕获多尺度的局部判别性特征, 从而使网络关注多样化的细粒度语义, 增强主干网络的特征学习能力. 特别地, 破坏重构学习模块和特征金字塔模块都只用于在训练阶段中增强网络的特征学习能力, 而在测试阶段不参与计算, 因此不会显著增加模型复杂度从而影响计算效率. 关于模型复杂度更详细的分析请参见第 3.5.6 节.

2.3 通道注意力模块

食品图像是一种非刚性图像, 没有固定的空间布局, 其视觉外观千差万别, 并且含有较多噪声. 针对此特性, 本文提出了通道注意力模块, 通过建模不同特征图的重要程度来增强判别性的视觉特征、降低噪声干扰.

通道注意力的具体计算过程如下: 设网络从输入图像中学到的特征为 $W_1\in R^{H\times W\times C}$, 首先使用全局平均池化方法获得特征图的全局上下文特征$\Theta\in R^C$; 接下来, 定义全连接层 FC_1, 其参数为 $W_1\in R^{C/r\times C}$, 用于将特征Θ压缩为 C/r 维以减少参数量, 其中, r 为缩放系数; 然后使用 ReLU 激活函数进行激活以拟合非线性特征. 得到的特征通过参数为 $W_2\in R^{C\times C/r}$ 的全连接层 FC_2 恢复为 C 维特征, Sigmoid 激活函数用于将特征中的元素映射到区间(0,1)中, 以得到每个特征图的权重. 整个计算过程如下:

$$s=\sigma(W_2\times\delta(W_1\times AVG(f))) \tag{5}$$

其中, $AVG(\cdot)$为全局平均池化, $\delta(\cdot)$为 ReLU 函数, $\sigma(\cdot)$为 Sigmoid 函数. 最终的特征为

$$f_c=s\odot f \tag{6}$$

其中, $\odot$为逐元素相乘. 在实现层面, 本文将通道注意力添加到每个网络层的最后.

此外, 本文在 MJR-Net 上还探索了其他多种不同的注意力机制对食品识别性能的影响, 详见第 3.5.3 节.

2.4 损失函数设计

考虑到食品图像属于细粒度图像, 不同类别的食品图像之间有极其相似的视觉特征, 类间差异性较低. 本文选用 A-softmax 损失[31]对网络进行辅助训练. A-softmax 能够将特征从欧式空间映射到角空间, 并引入余量以增大类间差异性. 其计算公式如下:

$$L_{AS}=\frac{1}{N}\sum_i\frac{-\log\left[\exp(\|x_i\|\cos(m\theta_{y_i,i}))\right]}{\exp(\|x_i\|\cos(m\theta_{y_i,i}))+\sum_{j\neq i}\exp(\|x_i\|\cos(\theta_{j,i}))} \tag{7}$$

其中, x_i 为第 i 个输入图像经过主干网络的最后一个卷积层输出的深度特征, m 为余量系数, $\theta_{y_i,i}$ 表示第 i 个图像的对应的特征向量与第 y_i 类的特征向量之间的夹角, $\|\cdot\|$为均方根函数.

同时, 考虑到食品图像具有较高的类内差异性, 本文还使用Focal损失[32]代替交叉熵损失. Focal损失可以为不同的样本赋予不同的权重, 减少易分类样本的权重, 使得模型在分类中更侧重于难分类的样本. Focal 损失的计算公式为

$$L_{CE}=\sum_{i=1}^{N}-\alpha\times[1-f(x_i)]^{\gamma}\times\log(f(x_i)) \tag{8}$$

其中, α和γ为超参数, $f(x_i)$为网络输出的评分.

整个 MJR-Net 的损失函数由 5 个部分组成: 衡量分类损失的 L_{CE}、用于优化分类的 A-softmax 损失 L_{AS}、对抗学习损失 L_{adv}、特征金字塔损失 L_{fpn}、区域对齐损失 L_{los}. 最终的损失 L 如下:

$$L=\alpha\times L_{CE}+\beta\times L_{adv}+\gamma\times L_{los}+\mu\times L_{fpn}+\nu\times L_{AS} \tag{9}$$

其中, α, β, γ, μ, ν为每一种损失的权重.

3 实验结果与分析

3.1 数据集

- ETH Food-101[13]是一个主要由西餐组成的食品图像数据集(https://vision.ee.ethz.ch/datasets/food-101/), 该数据集共有 101 个类, 每个类含有 1 000 张图像, 整个数据集共有 101 000 张食品图像. 训练集和测试集按照 3:1 的比例进行随机划分, 无验证集.
- Vireo Food-172[14]是一个由中国菜肴组成的食品图像数据集(http://vireo.cs.cityu.edu.hk/VireoFood172/), 该数据集共有 172 个类, 共 110 241 张食品图像. 训练集、验证集和测试集按照 6:1:3 的比例随机划分.
- ISIA Food-500[15]由中国菜和西方菜共同组成, 该数据集共有 500 个类(http://123.57.42.89/FoodComputing-Dataset/ISIA-Food500.html), 每个类包含的图像数目都大于 500, 共 399 726 张食品图像. 和 Vireo Food-172 的划分比例一致, 训练集、验证集和测试集按照 6:1:3 的比例随机划分.

表 1 展示了 3 个数据集的统计信息, 图 5 展示了 3 个数据集中的部分图像.

表 1 3 个数据集的统计信息

数据集	类别数	图片数	训练集	验证集	测试集	菜系
ETH Food-101	101	101 000	75 750	–	25 250	西方菜
Vireo Food-172	172	110 241	60 071	11 016	33 154	中国菜
ISIA Food-500	500	399 724	239 378	40 204	120 142	中西方菜

图 5 3 个食品数据集中的图像样例

3.2 实验设置和细节实现

本文选择基于动量的随机梯度下降方法作为优化器进行训练, 动量设置为 0.9, 权值衰减系数设置为 0.000 5, 训练周期设置为 100, 初始学习率设置为 0.001, 每隔 2 个周期下降为原来的 0.9. 本文使用 PyTorch 训练所有的模型, 每个主干网络都在 ImageNet 数据集上进行预训练. 在训练阶段中, 本文使用随机裁剪、随机角度旋转、颜色饱和度和色调调整、随机垂直翻转等数据增强方法, 在测试阶段中, 本文使用 10-crop 方法.

MJR-Net 中的超参数设置如下: 在 RCM 模块中, 混淆子图像数目 N 设置为 4, 表示将原始图像拆分为 4×4 个子图像, 然后进行随机打乱. 在通道注意力模块中, 缩放系数 r 设置为 16, 即每个残差块输出的特征压缩为原来的 1/16. 主干网络使用 ResNet50, 判别网络使用两层的全连接网络, 隐藏层的维度设置为 1 024, 并使用 ReLU 函数进行激活. 分类网络使用单层全连接网络. 特征金字塔网络使用 ResNet50 后 3 个阶段输出的特征作为中层特征, 每个特征被压缩为 256 维. 在 A-softmax 损失中, 余量参数 m 设置为 4, 在 Focal 损失中, α设置为 1, γ设置为 0.2. 总的损失函数中各个损失的权重均设置为 1, 即$\alpha=\beta=\gamma=\mu=\nu=1$.

3.3 评价指标

本文选择 Top-1 准确率和 Top-5 准确率作为评测 MJR-Net 的评价指标. Top-1 准确率指预测正确的图像占所有测试集图像的百分比. Top-5 准确率指的是预测结果前五的类别中含有正确类别的图像占所有测试集图像的百分比.

3.4 方法比较

本节主要评估 MJR-Net 的识别性能, 主干网络选择 VGG16 和 ResNet50, 与常用的细粒度图像识别方法保持一致, 输入分辨率设置为 448×448.

3.4.1 在 ETH Food-101 上的方法评估

表 2 展示了 MJR-Net 与其他食品识别方法在 Food-101 上的性能比较结果.

表 2 本文提出的 MJR-Net 与其他方法在 ETH Food-101 上的性能(%)

方法	主干网络	Top-1 准确率	Top-5 准确率
AlexNet[18]	–	56.40	–
GoogLeNet[33]	–	78.11	–
WARN[34]	–	85.50	–
ResNet50[19]	–	87.40	97.40
Inception V3[29]	–	88.28	96.88
SENet154[35]	–	88.62	97.57
WRN[21]	–	88.72	97.92
NTS[26]	–	89.40	97.80
WS-DAN[36]	ResNet50	90.13	98.23
WISeR[20]	WRN	90.27	**98.71**
DCL[27]	ResNet50	88.90	97.82
PAR-Net[37]	ResNet50	90.40	–
IG-CMAN[38]	VGG16+LSTM	90.40	98.42
MSMVFA[9]	VGG16	87.68	97.45
MSMVFA[9]	Densenet161	90.59	98.25
SGLANet[15]	SENet154	90.33	98.20
多级卷积特征金字塔[22]	VGG16	**91.40**	–
MJR-Net	VGG16	88.42	97.81
MJR-Net	ResNet50	90.82	98.32

从表中可以得出以下结论.

(1) WRN 和 SENet-154 的性能要好于其他单一的 CNN.

(2) 与主干网络 ResNet50 相比, MJR-Net 在识别性能上有较大的提升, 为 3.42%.

(3) 本文提出的 MJR-Net 在 Food-101 上取得了 90.82%的识别准确率, 领先于现有的大多数食品图像识别方法, 仅次于多级特征金字塔. 特别地, 本文提出的 MJR-Net 的参数量仅占主干网络的 22.0%左右, 而多级特征金字塔使用 3 个主干网络作为三级特征提取网络, 其参数量约为 3 个主干网络参数

量之和, 后者的参数量和计算成本都要远高于 MJR-Net.

(4) 考虑到 MSMVFA 和 IG-CMAN 使用了额外的食材信息作为监督信息, 而本文仅使用图像的类别标签进行训练, 且其他的食品识别方法大都使用性能较强的主干网络(如 Densenet, SENet 等). 综合上述考量, MJR-Net 在 Food-101 上具有较好的识别性能.

3.4.2 在 Vireo Food-172 上的方法评估

表 3 展示了 MJR-Net 与其他食品识别方法在 Food-172 上的准确率比较. 由表 3 可知, (1) SENet154 的性能要好于其他的单一 CNN; (2) 本文提出的 MJR-Net 在 Food-172 上取得了 91.37%的识别准确率, 大幅度领先于现有的食品识别方法, 比目前识别性能最好的方法 IG-CMAN 高出 0.74%, 取得了最优的识别性能; (3) MJR- Net 在 Food-172 上较大幅度地领先于在 Food-101 上取得最优性能的多级特征金字塔(1.07%), 在参数量和复杂度更少的情况下取得了更好的识别性能; (4) MJR-Net 仅使用图像的类别标签进行训练, 不使用额外的监督信息和性能更强的主干网络, 综合上述考量, MJR-Net 在 Food-172 上具有较好的识别性能.

表 3 本文提出的 MJR-Net 与其他方法在 Vireo Food-172 上的性能(%)

方法	主干网络	Top-1 准确率	Top-5 准确率
AlexNet[18]	–	64.91	85.32
VGG16[39]	–	80.41	94.59
DenseNet161[28]	–	86.93	97.17
MTDCNN[14]	VGG16	82.06	95.88
MTDCNN[14]	DenseNet16	87.21	97.29
SENet154[35]	–	88.71	97.74
PAR-Net[37]	ResNet50	90.20	–
IG-CMAN[38]	VGG16+LSTM	90.63	98.40
MSMVFA[9]	Densenet161	90.61	98.31
SGLANet[15]	SENet154	90.30	98.03
多级卷积特征金字塔[22]	VGG16	90.30	–
MJR-Net	VGG16	88.66	97.89
MJR-Net	ResNet50	**91.37**	**98.60**

3.4.3 在 ISIA Food-500 上的方法评估

为了验证方法的普适性, 本文还在另一个大规模食品图像数据集 ISIA Food-500 上进行实验. 表 4 展示了 MJR-Net 与其他方法在 ISIA Food-500 上的准确率比较. 由表 4 可知, MJR-Net 在 Food-500 上取得了 64.95%的识别准确率, 领先于其他食品识别方法, 比识别性能最好的 SGLANet 高出 0.21%. 实验结果进一步验证了 MJR- Net 的有效性, 能够应用于不同风格的食品类型, 在各种食品数据集上均能实现较好的性能.

表 4 MJR-Net 与其他方法在 ISIA Food-500 上的性能(%)

Method	主干网络	Top-1 准确率	Top-5 准确率
VGG16[39]	–	55.22	82.77
GoogLeNet[33]	–	56.03	83.42
ResNet152[19]	–	57.03	83.80
WRN50[21]	–	60.08	85.98
Densenet161[28]	–	60.05	86.09
SENet154[35]	–	63.83	88.61
SE-ResNeXt101[35]	–	61.95	87.54
NAS-NET[40]	–	60.66	86.38
NTS[26]	ResNet50	63.66	88.48
WS-DAN[36]	ResNet50	60.67	86.48
DCL[27]	ResNet50	64.10	88.77
SGLANet[15]	SENet154	64.74	89.12
MJR-Net	ResNet50	**64.95**	**89.29**

3.5 消融实验

本节主要评估 MJR-Net 各个组件之间的有效性. 所有的实验均在 ETH Food-101 数据集上进行.

3.5.1 破坏重构学习的有效性分析

本节主要验证破坏重构学习 DCL 的有效性. 实验设置方面, 本节在 3 种不同的主干网络 ResNet50,

ResNet152 和 InceptionV3 上进行实验, 输入图像的分辨率统一设置为 448×448. 表 5 展示了 DCL 与原始主干网络的准确率比较, 从表 5 中可以看出, 与单一的主干网络相比, DCL 能够显著提升食品图像识别性能. 当主干网络选择 ResNet50 时, Top-1 准确率有 1.50%的提升; 当主干网络选择 InceptionV3 时, Top-1 准确率有 1.68%的提升; 而当主干网络选择大规模的 ResNet152 时, Top-1 准确率仍有 1.68%的提升. 破坏重构学习能够有效地提取局部的判别性特征, 进而提高识别性能.

表 5 破坏重构学习(DCL)与主干网络在 EHT Food-101 上的性能(%)比较

方法	主干网络	Top-1 准确率	Top-5 准确率
CNN	ResNet50	87.40	97.40
	ResNet152	**88.51**	**97.72**
	InceptionV3	84.32	–
DCL	ResNet50	88.90	97.80
	ResNet152	**90.19**	**98.10**
	InceptionV3	86.00	–

3.5.2 金字塔模块和 Focal 损失、A-softmax 损失有效性分析

本节主要验证特征金字塔模块和 Focal 损失、A-softmax 损失的有效性. 实验设置方面, 本节选择 ResNet50 作为主干网络, 在两种分辨率 224×224 和 448×448 上分别进行验证, 以全面证明模块的有效性. 表 6 展示了消融实验的结果, 其中, FPM 表示特征金字塔模块, FA 表示 Focal 损失和 A-softmax 损失函数. 由表 6 可知, 在 224×224 的输入分辨率下, 当使用原始的 ResNet50 时, 使用特征金字塔能提升 0.75%的 Top-1 准确率. 引入 DCL 后, 特征金字塔能够提升 0.63%的 Top-1 准确率, 使用两种损失函数训练网络能提升 0.63%的准确率, 特征金字塔和两个损失联合训练则能提升 1.28%的准确率. 当输入分辨率调整为 448×448 时, 两者联合使用能提升 1.0%的准确率. 特征金字塔模块能够有效地提高主干网络定位和学习多尺度特征的能力, 进而提高识别性能, 关于特征金字塔模块的定性分析可参见第 3.6 节. 而 A-softmax 损失和 Focal 损失则分别从增大类间差异和修正分类样本的角度优化网络训练. 在两个分辨率上的实验结果证明了两者的有效性.

表 6 特征金字塔模块和 Focal 损失、A-softmax 损失在 EHT Food-101 上的消融实验结果(%)

方法	分辨率	Top-1 准确率	Top-5 准确率
ResNet50	224×224	84.55	96.88
+FPM	224×224	85.30	96.80
+DCL	224×224	85.68	96.71
+DCL+FPM	224×224	86.31	97.05
+DCL+FA	224×224	86.31	**97.21**
+DCL+FA+FPM	224×224	**86.96**	97.19
ResNet50	448×448	87.40	97.39
+DCL	448×448	88.90	97.80
+DCL+FA+FPM	448×448	**89.90**	**98.00**

进一步地, 本文还探究了特征金字塔不同使用方式的实验性能比较. 实验设置方面, 基准网络选择 ResNet50, 输入分辨率设置为 224×224. 表 7 展示了消融实验的结果, 其中, FPM_{p_2} 表示使用最终的金字塔特征 p_{fpn}(这里为 p_2)进行分类, FPM_{p_2,p_3,p_4} 表示使用多级金字塔特征集合 $P=\{p_2,p_3,p_4\}$的连接进行分类, FPM_{AUX} 表示使用特征金字塔辅助训练以进行分类.

表 7 特征金字塔使用方式在 EHT Food-101 上的消融实验结果(%)

方法	特征维数	Top-1 准确率	Top-5 准确率
ResNet50	2 048	84.55	96.88
+ FPM_{p_2}	256	83.10	95.54
+ FPM_{p_2,p_3,p_4}	768	83.64	95.90
+FPM_{AUX}	2 048	**85.30**	**96.80**

表 7 第 2 列展示了特征金字塔不同使用方式下, 最终用于分类特征的维度. 实验结果与我们之前的论述分析相一致. 由表 7 可知, 直接使用金字塔特征进行分类的准确率低于使用主干网络的原始特征. 其原因为:

金字塔特征的特征维数过低(一般为 256 维或 512 维), 无法充分的表示食品图像中复杂的视觉模式, 因此不适合单独地作为特征表示. 而使用特征金字塔进行辅助训练能够使网络更好的关注不同尺寸的局部判别性特征, 进而增强主干网络的细粒度特征学习能力.

3.5.3 注意力模块有效性分析

通道注意力模块用于对深层特征通道之间的重要程度进行建模, 通过对不同特征通道进行重新加权以突出判别性的视觉模式. 本文选择 ResNet50 作为主干网络, 并在 DCL 上探究通道注意力模块对性能的影响. 特别地, 本文还探究了其他 3 种注意力: 通道和空间注意力 CBAM[41]、自注意力[42]和互补通道注意力[43]对于分类性能的影响. 表 8 展示了各类注意力机制的识别准确率比较, 其中, Channels, Self-attention, Channels&Spatial, SCI 分别表示通道注意力、自注意力、通道和空间注意力和互补通道注意力. 由表 8 可知, 当使用小分辨率(224×224)的图像时, 采用本文提出的通道注意力模块能够明显提升分类性能(0.82%), 说明对通道重要性进行建模以挖掘判别性的视觉模式能够增强食品图像的识别性能, 减少噪声干扰. 在大分辨率的输入图像上,通道注意力仍有较大的性能提升(0.70%). 使用互补通道注意力对识别性能也有显著的提升(1.03%), 但随着分辨率的增大, 性能提升幅度较小(0.20%). 当使用通道和空间注意力时, 性能提升(0.59%)低于单一通道注意力(0.82%), 说明食品图像中不存在明显的结构信息, 很难利用空间上的注意力机制来增强特征表示. 使用自注意力对分类性能提升不明显(0.02%), 说明食品图像中不存在明显的长距离依赖关系.

表 8 各类注意力机制在 EHT Food-101 上的性能(%)

方法	输入分辨率	Top-1 准确率	Top-5 准确率
DCL	224×224	85.68	96.71
+Channels	224×224	86.50	**97.30**
+Self-attention	224×224	85.70	96.87
+Channels&Spatial	224×224	86.27	97.12
+SCI	224×224	**86.71**	96.92
DCL	448×448	88.90	97.82
+Channels	448×448	**89.60**	**98.15**
+Self-attention	448×448	89.30	97.90
+Channels&Spatial	448×448	89.56	98.10
+SCI	448×448	89.10	97.80

3.5.4 不同输入分辨率对性能的影响

本文还探究了两种输入分辨率, 224×224 和 448×448, 对于食品图像识别性能的影响. 本节在 DCL 网络上进行实验. 表 9 展示了两种输入分辨率的识别准确率比较. 由表 9 可知, 当主干网络使用 ResNet50 时, 相比于小分辨率的输入(224×224), 使用大分辨率(448×448)能够提升 3.22%的分类准确率; 而当主干网络使用 DenseNet161 时, 大的输入分辨率同样能够带来 1.80%的提升. 实验结果表明, 使用大分辨率的图像能够有效地提高食品识别的性能.

表 9 224×224 和 448×448 两种输入分辨率在 EHT Food-101 的性能(%)比较

输入分辨率	主干网络	Top-1 准确率	Top-5 准确率
224×224	ResNet50	85.68	96.71
224×224	ResNet152	87.00	97.44
224×224	DenseNet161	**88.50**	**97.80**
448×448	ResNet50	88.90	97.80
448×448	ResNet152	90.19	98.10
448×448	DenseNet161	**90.30**	**98.11**

3.5.5 不同混淆子图像数目 N 对于性能的影响

进一步地, 本文还探究了不同的混淆子图像数目 N 对于性能的影响. N 决定了 DCL 中每个子图像的尺寸, N 越小, 表明每个子图像的尺寸越大. 由表 10 可知, 在 224×224 的分辨率下, N 设置为 4 时有最优的识别准确率; 在 448×448 分辨率下, N 设置为 4 时有最优的识别准确率. 实验结果表明, 食品图像中的局部细粒度特征的尺度较大, 识别性能依赖于特征的完整性, 较小的 N 能够极大程度地保留完整的局部视觉特征.

表 10 MJR-Net 中不同混淆子图像数目 N 在 EHT Food-101 上的性能(%)

混淆子图像数目 N	分辨率	Top-1 准确率	Top-5 准确率
4×4	224×224	**87.52**	97.34
7×7	224×224	87.47	**97.40**
14×14	224×224	86.87	97.15
4×4	448×448	**90.82**	98.28
7×7	448×448	90.64	98.32
14×14	448×448	90.40	**98.35**

3.5.6 模型复杂度分析

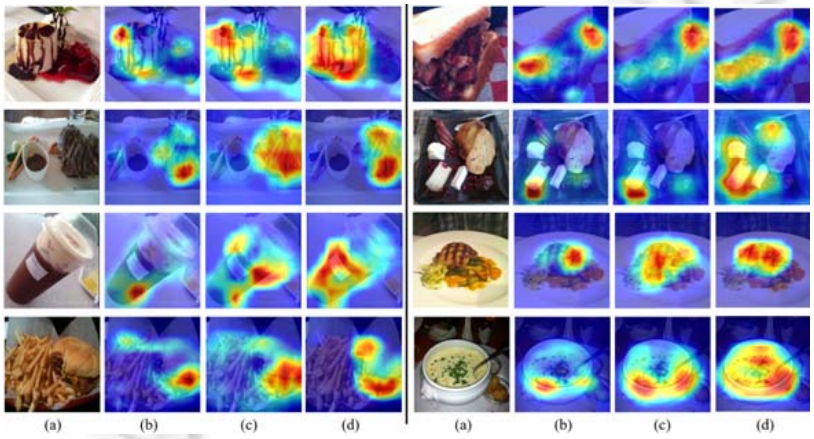
本文进一步分析了 MJR-Net 的模型复杂度, 主干网络为 ResNet50. 表 11 展示了 MJR-Net 的复杂度分析结果, 括号内为引入对应模块后相对主干网络增加的参数量百分比. 由表 11 可知, 引入 DCL 模块带来的复杂度与主干网络相比可以近似忽略, 引入特征金字塔模块增加的参数量仅为 2.7 M, 仅占主干网络参数量的 11.5%, 且引入 DCL 和特征金字塔模块几乎不会增加模型的计算量. 整个 MJR-Net 引入的参数量约为 5.3 M, 仅占主干网络参数量的 22.0%. 在预测阶段, DCL 和特征金字塔模块均不会参与计算, 因此, MJR-Net 增加的参数量仅为 2.5 M (约占 10.4%), 增加的计算量仅为 0.011 G. 因此, MJR-Net 的模型复杂度较低, 在拥有较高识别性能的同时, 还拥有较高的计算效率.

表 11 MJR-Net 的模型复杂度分析

模型	Params (M)	GFlops (G)
训练阶段		
ResNet50	23.948	4.132
+DCL	23.954	4.132
+FPM	26.70 2(+11.5%)	4.132
MJR	29.223 (+22.0%)	4.143
测试阶段		
ResNet50	23.948	4.132
MJR	26.443 (+10.4%)	4.143

3.6 可视化分析

为了进一步验证方法的有效性, 本节对 MJR-Net 进行定性分析和可视化. 图 6 是 MJR-Net 在一些食品图像的可视化分析结果, 每栏中每一行表示一幅图像在不同模型下从主干网络的最后一个卷积层中得到的热力图. 其中, 图 6(a)为原始图像, 图 6(b)为从 DCL 中学到的特征热力图, 图 6(c)为从使用 FPM 辅助训练的 DCL 中学到的特征热力图, 图 6(d)为从本文提出的 MJR-Net 中学到的热力图, 由蓝色到红色表示网络对这部分区域的关注程度逐渐增加.



(a) 原始图像 (b) DCL 得到的热力图 (c) FPM 和 DCL 得到的热力图 (d) MJR-Net 得到的热力图

图 6 MJR-Net 的可视化分析结果

可以看出，基准细粒度识别方法 DCL 倾向于挖掘最明显地细粒度特征，使得网络能够关注最具判别性的局部区域(如图 6(b)所示)，但由于食品图像不具备固定的语义模式，且受破坏尺寸的限制，DCL 学到的视觉特征不足以区分具有较强混淆性的食品类别．引入特征金字塔后，主干网络可以关注不同粒度的细粒度视觉特征，从而获得丰富的多尺度判别性特征表示．从图 6 中可以看出，相比于从 DCL 中得到的热力图(如图 6(b)所示)，使用特征金字塔辅助训练后得到的热力图(如图 6(c)所示)能够突出各种尺度的局部细节部分．这表明在特征金字塔的辅助训练下，主干网络能够更好地学习多尺度的细粒度特征．MJR-Net 在特征金字塔的基础上引入注意力机制对特征进行增强，通过对通道之间的重要程度进行建模以强化判别性视觉模式，弱化噪声干扰，从而使得网络能够精确定位判别性区域，而不会拟合无关的背景等信息(如图 6(d)所示，与图 6(c)相比，模型对判别性特征有更高的关注度，而对噪声的关注度会降低)．以图 6 左侧第 1 行的意式奶冻图像为例，DCL 能够定位局部的细粒度区域，学到意式奶冻的部分特征，但这些特征不足以将其与提拉米苏(DCL 的预测结果)区分开．特征金字塔模块可以提取到多尺度的食品特征，但由于过度拟合了一些无关噪声和背景，使得网络错误地将图像预测为起司蛋糕．而本文提出的 MJR-Net 通过使用通道注意力对特征进行增强，使得网络可以精确的定位到食品图像的判别性区域，并减少对背景等无关噪声的拟合(如图所示，热力图精准覆盖在食品对象上)，因此图像被正确地预测为意式奶冻．

接下来，本文还尝试对被错误分类的图像进行分析．图 7 展示了 MJR-Net 在 3 个数据集上的混淆矩阵，图 8 进一步展示了一些被错误分类的食品图像．可以看出，这些食品在视觉上都具有极其相似的外观，很难单纯借助视觉信息将其区分开．后续可以考虑使用一些额外的监督信息(如食材原料信息)来辅助食品图像识别，以从其他角度分辨视觉特征相似的食品．

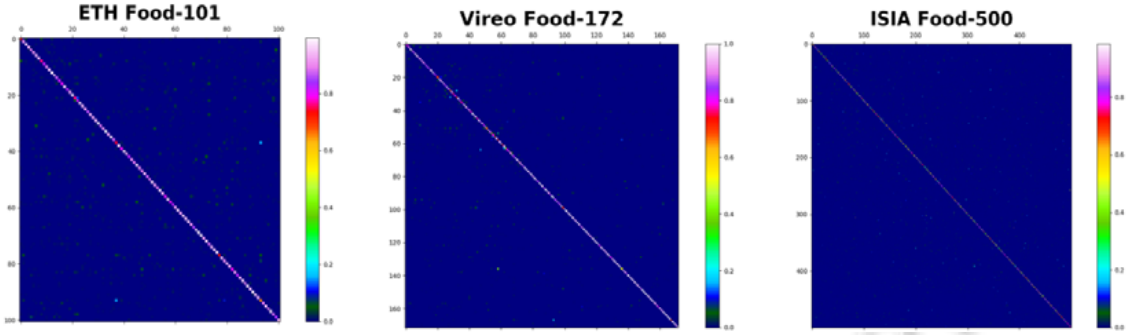


图 7　MJR-Net 在 3 个数据集上的混淆矩阵

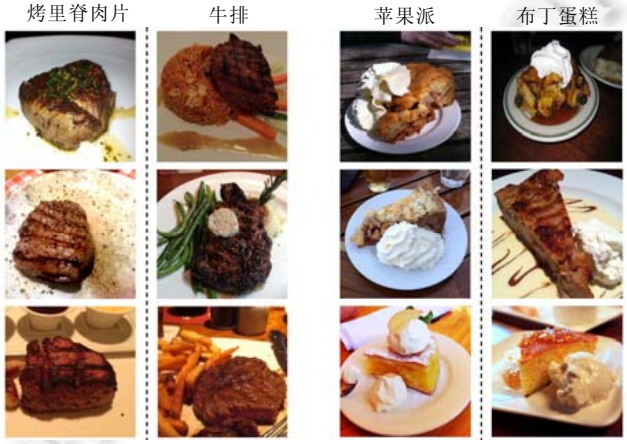


图 8　MJR-Net 的错分图像举例

4 结 论

本文提出了一种 MJR-Net 的食品图像识别方法. 为了适合食品图像细粒度的特点, 本文使用了破坏重构学习模块, 通过将原始图像按区域破坏和重建, 以强迫网络关注局部的判别性细节特征. 为了适应食品图像没有固定的局部语义的特性, 本文还提出使用特征金字塔模块来挖掘多尺度的非固定语义特征. 最后, 为了适应食品图像非刚性和多噪声的问题, 本文还提出使用注意力机制对图像特征进行增强, 通过对不同特征图的重要程度进行建模以增强判别性的视觉模式, 降低噪声干扰. 实验结果表明, 本文提出的MJR-Net在3个食品数据集上均有较好的识别性能. 在今后的工作中, 我们将主要研究如何使用食品图像的额外信息(如食材信息, 深度信息等)[44,45]提高食品图像识别方法的性能.

References:

[1] Khanna SK. Food and Culture: A Reader. 2nd ed., Carole Counihan and Penny Van Esterik, 2009. 157−159.

[2] Min WQ, Jiang SQ, Liu LH, Rui Y, Jain RC. A survey on food computing. ACM Computing Surveys, 2019, 52(5): Article No.92.

[3] Mezgec S, Eftimov T, Bucher T, Seljak BK. Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment. Public Health Nutrition, 2019, 22(7): 1193−1202.

[4] Zhang WS, Zhang YJ, Zhai J, Zhao DH, Xu L, Zhou JH, Li ZW, Yang S. Multi-source data fusion using deep learning for smart refrigerators. Computers in Industry, 2018, 95: 15−21.

[5] Zhu YS, Zhao X, Zhao CY, Wang JQ, Lu HQ. Food det: Detecting foods in refrigerator with supervised transformer network. Neurocomputing, 2020, 379: 162−171.

[6] Mohammad I, Mazumder MSI, Saha EK, Razzaque ST, Chowdhury S. A deep learning approach to smart refrigerator system with the assistance of IOT. In: Proc. of the Int'l Conf. on Computing Advancements. 2020. 1−7.

[7] Aguilar E, Remeseiro B, Bolaños M, Radeva P. Grab, pay, and eat: Semantic food detection for smart restaurants. IEEE Trans. on Multimedia, 2018, 20(12): 3266−3275.

[8] Min WQ, Jiang SQ, Jain RC. Food recommendation: Framework, existing solutions, and challenges. IEEE Trans. on Multimedia, 2020, 22(10): 2659−2671.

[9] Jiang SQ, Min WQ, Liu LH, Luo ZD. Multi-scale multi-view deep feature aggregation for food recognition. IEEE Trans. on Image Processing, 2020, 29: 265−276.

[10] Kagaya H, Aizawa K, Ogawa M. Fooddetection and recognition using convolutional neural network. In: Proc. of the ACM Int'l Conf. on Multimedia. Orlando, 2014. 1085−1088.

[11] Ming ZY, Chen JJ, Cao Y, Forde C, Ngo CW, Chua TS. Food photo recognition for dietary tracking: System and experiment. In: Proc. of the Int'l Conf. on Multimedia Modeling. Osaka, 2018. 129−141.

[12] Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD birds-200-2011 dataset. Technical Report, CNS-TR-2011-001, 2011.

[13] Bossard L, Guillaumin M, Van Gool L. Food-101-mining discriminative components with random forests. In: Proc. of the European Conf. on Computer Vision. Zurich, 2014. 446−461.

[14] Chen JJ, Ngo CW. Deep-based ingredient recognition for cooking recipe retrieval. In: Proc. of the ACM Int'l Conf. on Multimedia. Amsterdam, 2016. 32−41.

[15] Min WQ, Liu LH, Wang ZL, Luo ZD, Wei XM, Wei XL, Jiang SQ. ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network. In: Proc. of the ACM Int'l Conf. on Multimedia. Seattle, 2020. 393−401.

[16] Chen M, Dhingra K, Wu W, Yang L, Sukthankar R, Yang J. PFID: Pittsburgh fast-food image dataset. In: Proc. of the Int'l Conf. on Image Processing. Cairo, 2009. 289−292.

[17] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004, 60(2): 91−110.

[18] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification withdeep convolutional neural networks. In: Proc. of the Annual Conf. on Neural Information Processing Systems. Lake Tahoe, 2012. 1106−1114.

[19] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas, 2016. 770−778.

[20] Martinel N, Foresti GL, Micheloni C. Wide-slice residual networks for food recognition. In: Proc. of the IEEE Workshop on Applications of Computer Vision. 2018. 567−576.

[21] Zagoruyko S, Komodakis N. Wide residual networks. In: Proc. of the British Machine Vision Conf. York, 2016. Article No.87.

[22] Liang HG, Wen XQ, Liang DD, Li HD, Ru F. Fine-grained food image recognition of a multilevel convolution feature pyramid. Journal of Image and Graphics, 2019, 24(6): 870−881 (in Chinese with English abstract).

[23] Huang SL, Xu Z, Tao DC, Zhang Y. Part-stacked CNN for fine-grained visual categorization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas, 2016. 1173−1182.

[24] Zhang N, Jeff D, Girshick RB, Darrell T. Part-based R-CNNs for fine-grained category detection. In: Proc. of the European Conf. on Computer Vision. Zurich, 2014. 834−849.

[25] Zheng HL, Fu JL, Mei T, Luo JB. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proc. of the Int'l Conf. on Computer Vision. Venice, 2017. 5219−5227.

[26] Yang Z, Luo TG, Wang D, Hu ZQ, Gao J, Wang LW. Learning to navigate for fine-grained classification. In: Proc. of the European Conf. on Computer Vision. Munich, 2018. 438−454.

[27] Chen Y, Bai YL, Zhang W, Mei T. Destruction and construction learning for fine-grained image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach, 2019. 5157−5166.

[28] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Hawaii, 2017. 2261−2269.

[29] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas, 2016. 2818−2826.

[30] Lin TY, Piotr D, Ross BG, He KM, Bharath H, Belongie SJ. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Hawaii, 2017. 936−944.

[31] Liu WY, Wen YD,Yu ZD, Li M, Raj B, Song L. SphereFace: Deep hypersphere embedding for face recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Hawaii, 2017. 6738−6746.

[32] Lin TY, Priya G, Ross BG, He KM, Dollár P. Focal loss for dense object detection. In: Proc. of the Int'l Conf. on Computer Vision. Venice, 2017. 2999−3007.

[33] Szegedy C, Liu W, Jia YQ, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston, 2015. 1−9.

[34] Rodríguez P, Dorta DV, Cucurull G, Gonfaus JM, Roca FX, Gonzàlez J. Pay attention to the activations: A modular attention mechanism for fine-grained image recognition. IEEE Trans. on Multimedia, 2020, 22(2): 502−514.

[35] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Utah, 2018. 7132−7141.

[36] Hu T, Qi H, Huang Q, *et al*. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv:1901.09891, 2019.

[37] Qiu JN, Po F, Luo W, Sun YN, Wang SY, Lo B. Mining discriminative food regions for accurate food recognition. In: Proc. of the British Machine Vision Conf. Cardiff, 2019. Article No.165.

[38] Min WQ, Liu LH, Luo ZD, Jiang SQ. Ingredient-guided cascaded multi-attention network for food recognition. In: Proc. of the ACM Int'l Conf. on Multimedia. Nice, 2019. 1331−1339.

[39] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[40] Zoph B, Vasudevan V, Shlens J, Le Q. Learning transferable architectures for scalable image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Utah, 2018. 8697−8710.

[41] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the European Conf. on Computer Vision. Munich, 2018. 3−19.

[42] Cao Y, Xu JR, Stephen L, Wei FY, Hu H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In: Proc. of the Int'l Conf. on Computer Vision Workshops. Seoul, 2019. 1971−1980.

[43] Gao Y, Han XT, Wang X, Huang WL, Scott MR. Channel interaction networks for fine-grained image categorization. In: Proc. of the AAAI Conf. on Artificial Intelligence. New York, 2020. 10818−10825.

[44] Min WQ, Bao BK, Mei SH, Zhu YH, Rui Y, Jiang SQ. You are what you eat: Exploring rich recipe information for cross-region food analysis. IEEE Trans. on Multimedia, 2018, 20(4): 950−964.

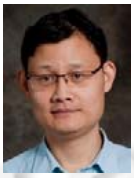
[45] Min WQ, Jiang SQ, Wang SH, Sang JT, Mei SH. A delicious recipe analysis framework for exploring multi-modal recipes with various attributes. In: Proc. of the ACM Int'l Conf. on Multimedia. California, 2017. 402−410.

附中文参考文献:

[22] 梁华刚, 温晓倩, 梁丹丹, 李怀德, 茹锋. 多级卷积特征金字塔的细粒度食物图片识别. 中国图像图形学报, 2019, 24(6): 870−881.

刘宇昕(1998－), 男, 博士生, 主要研究领域为多媒体分析, 计算机视觉.

闵巍庆(1985－), 男, 博士, 副研究员, CCF 高级会员, 主要研究领域为多媒体内容分析和理解, 食品计算.



蒋树强(1977－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为多媒体内容分析, 多模态智能, 食品计算.

芮勇(1970－), 男, 博士, 研究员, 博士生导师, CCF 会士, 主要研究领域为多媒体检索, 知识挖掘.