控制与决策
Control and Decision

随机选择全局多样化细粒度图像分类

刘光辉, 占华, 孟月波

引用本文:
刘光辉,占华,孟月波. 随机选择全局多样化细粒度图像分类[J]. 控制与决策, 2023, 38(9): 2622−2631.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1258

随机选择全局多样化细粒度图像分类

刘光辉†, 占 华, 孟月波

(西安建筑科技大学 信息与控制工程学院，西安 710055)

摘 要: 针对细粒度图像分类任务中潜在的可区分特征太过细微难以捕捉、忽视不同特征间的关系等问题,提出一种随机选择全局多样化分类网络模型. 首先,尝试以ConvNeXt作为主干来提升分类性能,并设计随机消除增强选择策略(REBS),通过特征消除分支和特征增强分支相互作用,促进网络学习更多相关信息,捕获潜在的可区分特征;然后,提出全局多样化模块(GDM),对不同层次的特征图进行交互建模,提高网络对比线索的能力;最后,建立内标压印数据集,将细粒度算法应用于真伪鉴定工作,实现细粒度图像分类任务在自然场景下的实际应用. 所提出方法在CUB-200-2011、Stanford Cars和FGVC-Aircraft三个公开数据集上分别达到了91.9%、93.8%和93.5%的准确率,相比其他先进对比方法性能有较大幅度提升. 在自建的内标压印数据集上达到了96.8%的准确率,能够实现真伪图像的准确分类.

关键词: 细粒度分类；可区分特征；随机消除增强选择策略；全局多样化；真伪鉴定

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1258

引用格式: 刘光辉,占华,孟月波. 随机选择全局多样化细粒度图像分类[J]. 控制与决策, 2023, 38(9): 2622-2631.

Random selection global diversification fine-grained image classification

LIU Guang-hui†, ZHAN Hua, MENG Yue-bo

(College of Information and Control Engineering，Xi'an University of Architecture and Technology，Xi'an 710055，China)

Abstract: A random selection global diversified classification network model is presented to deal with the difficulty of capturing the potential distinguishable features in fine-grained image classification and the tendency to ignore the relationship between different features. Firstly, the ConvNeXt is taken as the backbone to improve classification performance, and a random elimination boosting selection (REBS) strategy is designed to promote network learning more image information and capture potential distinguishable features through the enhancement of the interaction between the feature elimination and feature boosting branches. After that, the global diversification module (GDM) is proposed focusing on modelling feature maps of different levels interactively to enhance the comparison ability of network. Meanwhile, the dataset of logo imprint image is established, and the fine-grained algorithm is applied to conduct authenticity identification, which realizes the practical application of fine-grained image classification task in natural scenes. This network achieves 91.9%, 93.8% and 93.5% accuracy on three open datasets, CUB-200-2011, Stanford Cars and FGVC-Aircraft, respectively. Compared with other advanced comparison methods, the presented method greatly upgrades the comparison performance. The accuracy of 96.8% is achieved on the self-built dataset, which indicates the capacity of the network in accurate classification of true and fake images.

Keywords: fine-grained image classification；distinguishable feature；random elimination boosting selection strategy；global diversification；authenticity identification

0 引 言

细粒度图像分类,又称作子类别图像分类,其目的是对粗粒度的大类别进行更加细致的子类划分,但是,由于子类别间细微的类间差异和较大的类内差异,较之普通的图像分类任务,细粒度图像分类难度更大[1]. 虽然细粒度图像识别存在诸多难点,但是,其

收稿日期: 2022-07-14；**录用日期:** 2022-12-30.
基金项目: 国家自然科学基金项目(52278125)；陕西省重点研发计划项目(2021SF-429).
责任编委: 胡清华.
†通讯作者. E-mail: guanghuil@163.com.

对于工业界和学术界均有重大研究意义以及广泛的应用场景[2].

细粒度图像识别算法可分为强监督和弱监督两个方向: 强监督是指在模型中除使用类别标签, 还需要标注框, 部位标注点信息等; 弱监督是指仅仅使用类别标签完成模型的训练. 早期, 一些方法多采用两阶段的强监督方式来解决. 文献[3]提出了基于局部的R-CNN (part-based R-CNNs)分类算法, 借助细粒度图像中的边界框和零部件注释信息进行训练得到对象级和部件级的图像块特征, 将得到的特征级联后进行分类操作. 文献[4]提出了一个结合部位定位、对齐以及分类的细粒度分类网络, 设计了阀门连接函数(valve linkage function, VLF)来优化定位以及分类子网络间的连接, 协调分类结果与定位结果. 强监督方式十分依赖标注的边界框和部件注释信息等额外的人工标注信息, 而这些标注信息获取又十分昂贵, 使得这类强监督算法实用性受到限制, 因此, 近年来, 仅依赖类别标签完成分类的弱监督方式成为细粒度图像研究的一大趋势.

得益于深度学习的发展以及相关研究的深入, 不借助额外的人工标注信息也能够达到良好的分类性能. 文献[5]利用跨层特征间的相互作用, 在多个层次间建立空间依赖关系来学习更多的可区分特征. 文献[6]提出了区分性面向特征的高斯混合模型(discriminative feature-oriented Gaussian mixture model, DF-GMM)以解决区分性区域扩散问题, 并找到更好的细粒度细节. 文献[7]设计了一种双线性卷积网络模型(bilinear CNN, BCNN), 通过向量外积相乘方式组合两个网络特征, 实现更具区分性的细粒度特征表示. 注意力机制的应用, 使得网络关注有区分性的部位, 提高细粒度分类任务的准确率. 文献[8]提出了一种对象部位注意力模型(object-part attention model, OPAM), 通过对象级和部位级注意力机制来增强图像特征的表示. 文献[9]采用了多尺度注意力模型, 层次化表征注意力信息, 最后在输出层融合得到的注意力图特征图. 文献[10]提出了三线性注意力抽样网络, 包括注意力模块、注意力采样器以及特征蒸馏器. 文献[11]设计了一种注意力卷积二叉神经树结构, 沿树结构的边缘引入卷积计算, 并使用每个节点中的路由函数来确定树中的根到叶计算路径, 实现由粗至细的层次特征学习.

注意力机制能够引导模型关注有区分性的部位, 但是其通常只聚焦于最显著的部分, 而忽略了潜在的细微特征, 未从全局角度全面探索潜在的有辨别力的部分, 且孤立地对待各显著特征. 本文认为有效挖掘潜在的显著信息, 并提升得到信息的丰富性, 是提高网络提取细粒度特征能力的有效手段.

主干网络的选择也是细粒度任务的关键环节, 近期开发的transformer架构在细粒度任务上取得了非凡的效果. 文献[12]提出了选择性注意收集模块(selective attention collection module, SACM), 利用ViT[13](vision transformer)中的注意力权重, 按照输入图像块的相对重要性自适应地过滤它们, 以此来弥补ViT在细粒度任务中的不足. 文献[14]提出了相互注意权重选择模块(mutual attention weight selection, MAWS), 在不引入额外参数的情况下有效地选择有区别的图像块, 并以此来聚集每个transformer层的重要标记, 以补偿低级和中级信息. 文献[15]在ViT上提出了区域选择模块, 指导网络选择有鉴别力的图像块来获取最有判别力的图像区域. 文献[16]提出了递归注意多尺度transformer (recurrent attention multi-scale transformer, RAMS-Trans), 利用transformer的自注意力以多尺度的方式递归学习辨别性区域注意. 新的transformer架构结合注意力机制的思想是解决细粒度任务的一个有效思路, 但是对于transformer, 其架构与CNN存在一定探索和争论, 文献[17]提出了ConvMixer, 整个网络结构通过传统的卷积来实现, 验证了像ViT这种架构强大的性能并非全部来自于transformer, 至少部分是来自patch作为输入表示实现的. 文献[18]更是以纯卷积主干网络搭建ConvNeXt, 在多个任务性能上超越了Swin transformer[19], 这令人们重新思考卷积在计算机视觉中的重要性.

另外, 细粒度任务的挑战性还存在于应用方面, 目前主流研究所采用的细粒度图像数据库包括狗(stanford dogs)[20]、飞机(FGVC aircraft)[21]、汽车(stanford cars)[22]和鸟类(CUB200-2011)[23], 尽管这些数据集有一定类别数量和标注质量, 但是, 上述数据并非实际生活场景中获得的图像, 如手机拍摄的照片与数据集中图像存在较大差异, 这造成所取得的识别技术也局限于现有数据集上, 影响了细粒度分类任务的现实应用. 因此, 如何将细粒度分类算法结合实际应用, 并密切服务于现实生活, 是该领域值得展开的工作.

根据上述分析, 本文提出一种弱监督的随机选择全局多样化分类方法, 使用ConvNeXt作为主干

网络,在网络训练阶段提出随机消除增强选择策略(random elimination boosting selection, REBS),通过抑制最显著信息和奖励最具辨别力部分这两种方式,进行全局的潜在可辨别特征探索. 进一步地,设计全局多样化模块(global diversification module, GDM),建立各特征的共性关系,提高其特征丰富性. 同时,根据手提包真伪鉴定过程,搜集并建立手提包中内标压印部位的真伪图像数据集,通过该数据集结合所提出算法构建一个能够精确分类的模型,可大规模筛查假冒产品,辅助鉴定师进行高效鉴定.

1 随机选择全局多样化分类网络

随机选择全局多样化分类网络框架具体如图1所示,主要包括主干网络ConvNeXt、随机消除增强选择策略REBS和全局多样化模块GDM.

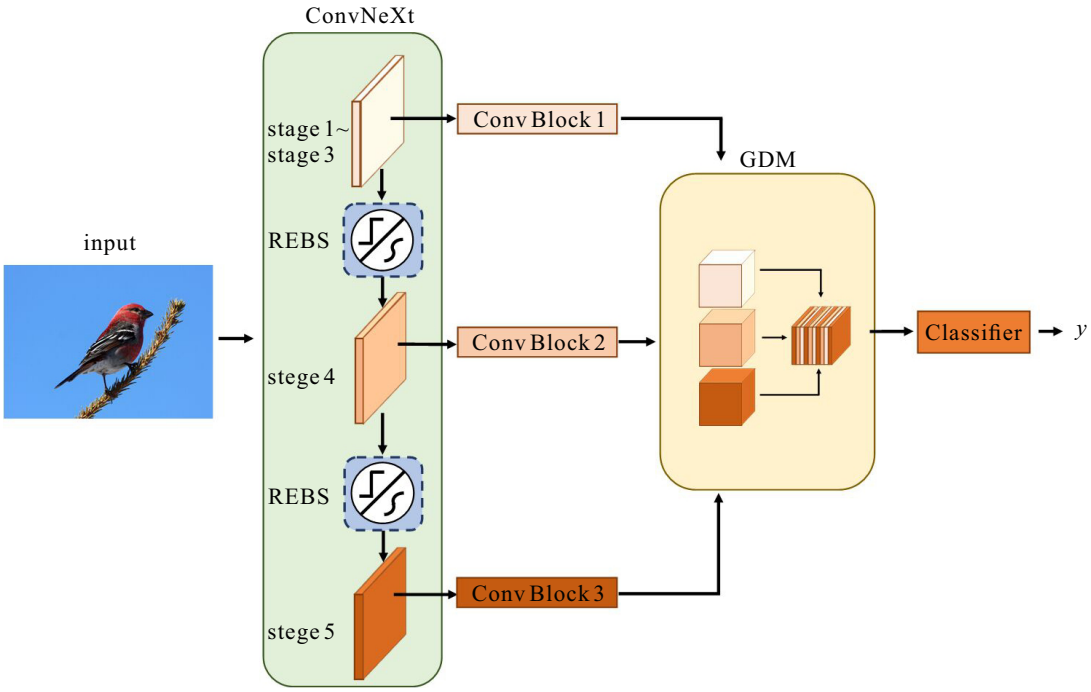

图 1 随机选择全局多样化分类网络结构

1.1 主干网络

对于细粒度任务,若想构造强大的特征表示,则主干网络的选择十分重要,其决定了细粒度特征的提取能力. 如后文表5中的Mask-CNN[24]在使用了相同标注信息以及相同算法后分别在Alex-Net、VGG以及Resnet 50[25]3个主干网络中进行对比实验,实验结果表明,当主干网络的分类能力足够好时,下游任务也会随之取得好的性能,文献[12, 14-16]更是以ViT作为主干获得了强大的性能. 根据上述分析,如何选择优异的主干网络,构建强大的特征表示,是细粒度任务的关键之一.

ConvNeXt是由Facebook AI研究院于2022年提出的,其完全由标准卷积模块构建,将Swin transformer以及ViT中的特殊设计集于一身,依次从宏观设计、深度可分离卷积[26]、逆瓶颈层[27]、大卷积核以及其他细节出发,升级了ResNet架构,拥有比Swin transformer更快的推理速度和更高的准确率,因此,本文选择ConvNeXt作为主干网络. 如图1所示,该网络包含5个stage. 其中: stage 1结构简单,可视为对输入图像的预处理; stage 2 ~ stage 4由ConvNeXt Block堆叠组成,结构相似. 网络深度随着stage的增加而加深,所包含信息也随之愈加丰富. 当输入图像经过不同stage后可得到不同尺度下的特征图 $X \in F^{C\times W\times H}$,其中$C$、$W$和$H$分别为特征图的通道数、宽度和高度.

1.2 随机消除增强选择策略

在细粒度任务上网络通常只关注最显著的部分而忽视了其他潜在的可辨别部分. 为了避免网络只聚焦于最显著的局部特征,而忽视整体的全局特征,本文设想在训练过程中将特征图切片后,通过抑制每个切片中最显著的部分来迫使网络学习更多相关性的信息,促进网络关注全局信息. 然而,在整个训练过程若均采用抑制操作,则将会造成网络完全忽视最显著的特征,从而导致精度降低. 因此,还需要一个增强操作奖励最具辨别力的部分来提高模型的预测能力. 基于上述分析,本文提出了随机消除增强选择策略,通过对特征图进行均匀切片,并在每一份切片中随机执行上述两种操作,迫使网络学习更加全面的有

效特征.

REBS 的具体结构如图 2 所示, 该策略输入可以是主干网络任意一层的输出特征图 F, 本文将以 stage 3 和 stage 4 的输出作为 REBS 的输入. 首先, 沿着宽度维度对特征图 $F \in F^{C\times W\times H}$ 执行 n 份均匀切片操作得到 $F_{(k)} \in R^{C\times (W/n)\times H}, k \in [1, n]$; 然后, 对每个切片 $F_{(k)}$ 随机执行特征消除操作或特征增强操作, 即该策略提供了两条候选分支, 分别为特征消除分支和特征增强分支, 每个切片各有 50 % 概率执行特征消除操作或特征增强操作. REBS 采用 0 和 1 表征 2 个分支, 以随机抽取方式决定分支的选择, 从而实现切片 $F_{(k)}$ 两类操作的随机化执行.

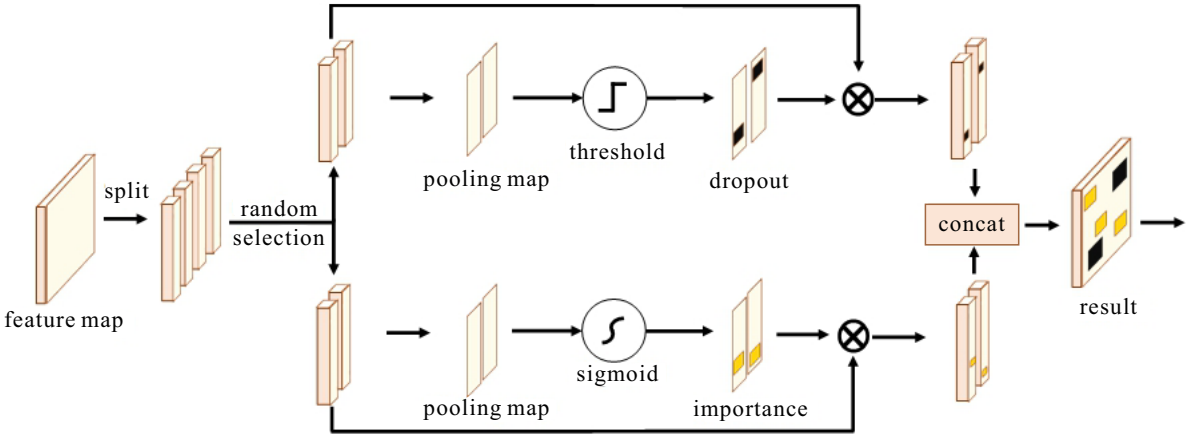

图 2　随机消除增强选择策略结构

对于特征消除分支, 采用下式对输入特征图执行通道平均池化操作, 得到 $F_{P(k)} \in R^{(W/n)\times H}$:

$$F_{P(k)} = \mathrm{CAP}(F_{(k)}) \in R^{(W/n)\times H}, \tag{1}$$

其中 CAP 为通道平均池化 (channel-wise average pooling).

$F_{P(k)}$ 每个像素点的取值范围与输入特征图相同, 表征分类模型获取到的关键特征表达. 由于所提出的随机选择全局多样化分类网络是为分类任务而训练, $F_{P(k)}$ 可近似反映最具辨别力部分的空间分布, 其元素取值越高, 辨别力越强. 即对于分类任务而言, $F_{P(k)}$ 中每个像素点的强度代表了其辨别的能力. 为了消除最具辨别力的部分, 根据 $F_{P(k)}$ 中最大强度的像素值设定阈值率 δ 生成消除掩码 $P_{(k)\mathrm{drop}}$, 将大于阈值的部分像素设置为 0, 反之, 将小于阈值的部分像素设置为 1, 具体如下式所示:

$$P_{(k)\mathrm{drop}} = \begin{cases} 0, & F_{P(k)}(i,j) > \delta \times \max(F_{P(k)}); \\ 1, & \text{otherwise}. \end{cases} \tag{2}$$

对于特征增强分支, 采用下式对 $F_{P(k)}$ 使用 sigmoid 激活函数生成增强掩码 $P_{(k)\mathrm{important}}$:

$$P_{(k)\mathrm{important}} = \mathrm{sigmoid}(F_{P(k)}) \in [0,1]^{(W/n)\times H}. \tag{3}$$

当切片 $F_{(k)}$ 选择特征消除分支时, 则通过下式得到消除特征图 $F_{(k)\mathrm{drop}}$:

$$F_{(k)\mathrm{drop}} = F_{(k)} \times P_{(k)\mathrm{drop}}. \tag{4}$$

对应地, 若选择特征增强分支, 则通过下式得到增强特征图 $F_{(k)\mathrm{important}}$:

$$F_{(k)\mathrm{important}} = F_{(k)} \times P_{(k)\mathrm{important}}. \tag{5}$$

最后, 将 $F_{(k)\mathrm{drop}}$ 和 $F_{(k)\mathrm{important}}$ 按宽度维度进行拼接得到 $F_{\mathrm{result}} \in R^{C\times W\times H}$, 有

$$F_{\mathrm{result}} = \mathrm{concat}(F_{(k)\mathrm{drop}}, F_{(k)\mathrm{important}}), \tag{6}$$

其中 concat 为将被切分处理后的每份特征图在宽度维度进行拼接.

1.3　全局多样化模块

通过上述 REBS 得到潜在的全局特征, 若直接输出则限制了模型从不同特征中进行线索对比的能力, 本文认为不应孤立地对待这些不同的全局特征, 更合理的做法是对不同层次的特征图进行交互建模, 迫使网络不同层共享挖掘到的信息, 增强语义互补信息.

GDM 的过程如图 3 所示. 首先, 将主干网络 stage 3 ~ stage 5 经过 ConvBlock 1 ~ ConvBlock 3 后的特征图作为一个图像对 $(F_1, F_2, F_3) \in (R^{C\times W_1\times H_1}, R^{C\times W_2\times H_2}, R^{C\times W_3\times H_3})$, 将这 3 个特征图的宽度和高度由 $W_1\times H_1$、$W_2\times H_2$、$W_3\times H_3$ 压缩为 L_1、L_2、L_3, 得到 $(F_1', F_2', F_3') \in (R^{C\times L_1}, R^{C\times L_2}, R^{C\times L_3})$; 然后, 通过 $F_1'^{\mathrm{T}}$ 和 F_2'、$F_2'^{\mathrm{T}}$ 和 F_3' 以及 $F_1'^{\mathrm{T}}$ 和 F_3' 进行内积运算获取相似度矩阵 M_1、M_2 和 M_3, 相似度矩阵中元素 $M_{i,j}$ 为不同特征图像素的相似度, 2 个像素值的相似度越低, 它们之间的互补性越强, 因此, 将 $-M_1$、$-M_2$ 和 $-M_3$ 作为交互矩阵并按照下式对其行列进行归一化

操作得到W_{12}、W_{23}、W_{13}:

$$W_{12}=\text{softmax}(-M_1^{\mathrm{T}})\in[0,1]^{L_1\times L_2},\ M_1=F_1'^{\mathrm{T}}F_2';\tag{7}$$

$$W_{23}=\text{softmax}(-M_2^{\mathrm{T}})\in[0,1]^{L_2\times L_3},\ M_2=F_2'^{\mathrm{T}}F_3';\tag{8}$$

$$W_{13}=\text{softmax}(-M_3^{\mathrm{T}})\in[0,1]^{L_1\times L_3},\ M_3=F_1'^{\mathrm{T}}F_3'.\tag{9}$$

接着,采用下式将上述归一化操作得到的交互特征图W_{12}、W_{23}、W_{13}加权至F_1'、F_2'、F_3'中得到W_{F_1}、W_{F_2}、W_{F_3},并将其尺寸L_1、L_2、L_3转换回$W_1\times H_1$、$W_2\times H_2$、$W_3\times H_3$,得到$(W_{F_1},W_{F_2},W_{F_3})\in(R^{C\times W_1\times H_1},R^{C\times W_2\times H_2},R^{C\times W_3\times H_3})$,有

$$W_{F_1}=F_2'\times W_{12}^{\mathrm{T}}+F_3'\times W_{13}^{\mathrm{T}},\tag{10}$$

$$W_{F_2}=F_1'\times W_{12}+F_3'\times W_{23}^{\mathrm{T}},\tag{11}$$

$$W_{F_3}=F_1'\times W_{13}+F_2'\times W_{13}.\tag{12}$$

最后,通过降采样融合W_{F_1}、W_{F_2}、W_{F_3},获得具有丰富语义信息的特征图.

在分类器部分,本文将融合后的特征图映射为一维特征向量,采用全连接层方式通过softmax逻辑回归实现图像最终的分类.

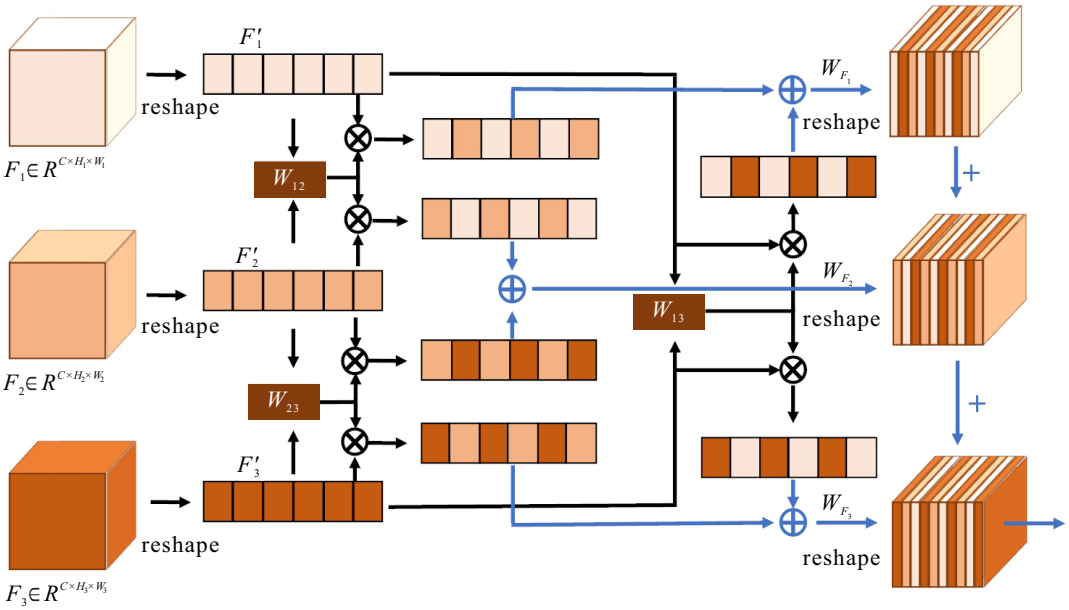

图 3 全局多样化模块结构

2 实验结果分析

2.1 实验数据集

本节将在CUB-200-2011、Stanford Cars、FGVC-Aircraft这3个常用数据集以及自建的手提包内标压印部位真伪图像数据集上进行对比实验来表明所提出算法的有效性. 上述4个数据集的具体信息如表1所示.

表 1 4种数据集参数

数据集	CUB-200-2011	Stanford Cars	FGVC-Aircraft	内标压印
训练样本量	5 994	8 144	6 667	5 794
测试样本量	5 794	8 041	3 333	2 483
类别数	200	196	100	2

2.2 实验参数

实验的环境配置如下:操作系统为Ubuntu 16.04, GPU 型号为RTX 2 080Ti, 采用 PyTorch 深度学习框架. 输入图像的大小为448×448. 使用带动量的批处理随机梯度下降法(SGD),动量设置为0.9,权重衰减为0.000 01,每批次内包含20张图像,以0.02的初始学习率训练200个epoch,数据增广方式为图像的随机水平翻转,同时利用余弦退火调整学习速率. 细粒度分类多采用准确率作为评价指标,为了便于多算法对比,本文亦如此.

2.3 消融实验

CUB-200-2011数据集拥有非常详细和准确的标注信息,每张图像除图像级标签(label)还有对象级的位置边界标注框(BBox)和15个部位标注点信息(parts),其包含了200种鸟类的图像,每一类别统一包含了30张左右的训练图像和测试图像,拥有较丰富的种类以及均衡的数据,是细粒度图像分类任务中最常用的数据集. 鉴于此,本节选择在该数据集上进行消融实验.

2.3.1 设置超参数δ

本节将探索REBS中阈值δ对于网络的影响,式(1)中的切片份数n将根据经验设置为7. 网络中间

层所包含特征较为丰富,蕴含上下文信息的表达,鉴于此,将以stage 3和stage 4的输出作为REBS的输入. 在REBS的特征消除分支中,由于消除的区域会随着δ的减小而增大,反之亦然. 若δ取值太小,则消除区域太多,会对结果造成负面影响;若δ取值太大,则使得消除区域过小,达不到预期的理想效果. 由表2可见,当δ取0.85时,REBS实现了最佳性能.

表2　阈值δ在CUB-200-2011数据集上的实验结果

阈值δ	acc/%
0.95	91.4
0.9	91.5
0.85	**91.6**
0.8	91.5
0.75	91.5
0.65	91.4
0.6	91.4
0.55	91.4
0.5	91.3

2.3.2　REBS插入位置的选择

在网络训练过程中应具体在哪个阶段采用REBS来优化网络能够达到最佳效果,对于此问题,在上述设置超参数δ的实验中,本文按照经验设想REBS应在网络的中间层使用将有利于网络性能的提升,为了验证这一设想,本文分别在主干网络的不同stage下采用REBS进行了实验,在此过程中阈值δ设置为0.85.

首先,设置单独只在stage 2～stage 5的输出后采用REBS,由于stage 1只是对输入图像的预处理,其包含有效信息较少,本文不考虑在stage 1后采用REBS. 由表3可见,在stage 4后采用REBS准确率为91.4%,高于应用在其他stage后的性能;然后,设置在多个不同的stage输出后采用REBS,由表3中结果可发现,将REBS应用于stage 3和stage 4后取得了最理想的效果.

表3　不同stage结合REBS在CUB-200-2011数据集上的实验结果

REBS	acc/%
stage 2	91.2
stage 3	91.3
stage 4	91.4
stage 5	91.2
stage 2 + stage 3	91.3
stage 3 + stage 4	**91.6**
stage 2 + stage 3 + stage 4	91.4
stage 3 + stage 4 + stage 5	91.4
stage 2 + stage 3 + stage 4 + stage 5	91.2

2.3.3　不同模块的性能分析

为了验证REBS和GDM的有效性,表4给出了主干网络结合不同模块对于最终分类性能以及模型复杂度的影响.

表4　本文算法在CUB-200-2011数据集上消融实验结果

方法	acc/%	param/M	FLOPs/G
ConvNeXt	91.2	87.5	61.4
ConvNeXt + REBS	91.6 (+0.4)	87.5 (+0)	62.8 (+1.4)
ConvNeXt + DM	91.3 (+0.1)	92.5 (+5.0)	65.2(+3.8)
ConvNeXt + REBS + DM	**91.9 (+0.7)**	92.5 (+5.0)	66.6 (+5.2)

由表4可见:仅采用ConvNeXt主干网络,算法准确率为91.2%;根据上述实验设定阈值并在stage 3和stage 4后采用REBS准确率达到91.6%,提升了0.4%;为了观察GDM模块的作用,移除REBS后单独在主干网络中采用GDM,由结果可知,准确率为91.3%,相较于基线只提升了0.1%;当REBS结合GDM共同采用时,准确率达到了91.9%,相比基线提升了0.7%. 由此可见,单独采用GDM对于网络性能提升十分有限,合理的做法应是在REBS捕捉到潜在可区分特征的基础上,利用GDM建立各特征的共性关系,提高其特征丰富性来提升网络整体的分类性能.

模型复杂度包括计算复杂度和空间复杂度,可分别采用浮点运算次数(floating point operations, FLOPs)、参数量(parameters, Param)进行度量. REBS主要功能是对输入特征图进行消除操作或增强操作,策略不包含可训练参数,因此,不会给模型带来额外的参数量负担;但是计算量增加不可避免,引入REBS后,FLOPs升高了1.4 G. 单独引入GDM模块,参数量增加了5.0 M,FLOPs升高了3.8 G. REBS、GDM均采用时,参数量增加了5.0 M,FLOPs升高了5.2 G.

2.4　与现有方法的比较

2.4.1　CUB-200-2011数据集实验与分析

图4为所提出方法在CUB-200-2011数据集下的网络损失收敛和准确率变化情况. 由图4可见,当训练损失train_loss和验证损失test_loss趋向于稳定时,训练准确率train_acc和验证准确率test_acc收敛,test_acc达到91.9%.

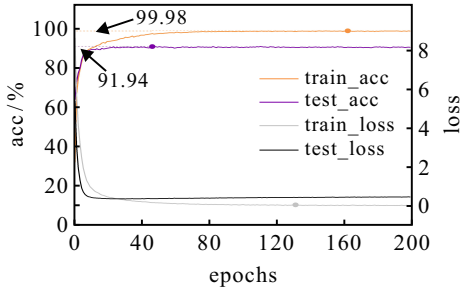

图4　网络损失收敛和准确率变化曲线

表5为多算法对比实验. 由表5可见,所提出方法在该数据集上达到了准确率91.9%的卓越性能,远高于强监督方式的PoseNorm、KERL、Mask-CNN,且能够只采用图像级标签(label)进行弱监督的端到端训练. 表5中PMG和FBSD以89.6%和89.8%的准确率领先于其他基于CNN架构的弱监督方式,但是均落后于所提出方法. 基于transformer架构的方法,如TPSKG、AFTrans、FFVT、TransFG借助ViT作为主干网络取得了较为优秀的准确率,而所提出方法以ConvNeXt作为主干网络,取得了更加优异的性能.

表5 不同算法在CUB-200-2011数据集上的准确率以及模型复杂度实验结果

方法	主干网络	使用标签	acc/%	param/M	FLOPs/G
PoseNorm[28]	VGG	label + BBox + parts	75.7	–	–
KERL[29]	VGG	label + knowledge	87.0	–	–
Mask-CNN[24]	Alex-Net	label + parts	78.6	–	–
	VGG	label + parts	85.7	–	–
	Resnet50	label + parts	87.3	–	–
BCNN[7]	D-Net + M-net	label + BBox	85.1	–	–
	D-Net + M-net	label	84.1	–	–
MAMC[30]	Resnet 50	label	86.2	–	–
LIO[31]	Resnet 50	label	88.0	–	–
FDL[32]	Resnet 50	label	88.6	–	–
NTS[33]	Resnet 50	label	87.5	29.0	83.6
PMG[34]	Resnet 50	label	89.6	45.1	37.4
CIN[35]	Resnet 101	label	88.1	–	–
API-Net[36]	Resnet 101	label	88.6	46.1	63.0
DTB-Net[37]	Resnet 101	label	88.1	–	–
FBSD[38]	Densenet 161	label	89.8	46.8	53.1
ViT[13]	ViT-B_16	label	90.2	86.1	62.0
TPSKG[39]	ViT-B_16	label	91.3	–	–
AFTrans[12]	ViT-B_16	label	91.5	–	–
FFVT[14]	ViT-B_16	label	91.6	86.2	62.0
TransFG[15]	ViT-B_16	label	91.7	86.4	62.0
ours	ConvNeXt	label	**91.9**	92.5	66.6

注:表中“–”表示原文献源码未给出,无法对其复杂度进行精准衡量.

同时,所提出方法模型总参数量达到92.5 M,FLOPs为66.6 G,在追求更高分类准确率的过程中,牺牲了一定的模型复杂度性能. 模型复杂度表征(FLOPs、param)主要与模型本身有关,因此,其他数据集实验将以准确率作为主要性能指标进行实验结果说明.

2.4.2 Stanford Cars数据集实验与分析

Stanford Cars数据集包含196类汽车,共16 185张图像,其中关键特征包括车辆制造商、汽车品牌、车型等,该数据集除图像级标签(label)只提供了标注框信息(BBox). 由表6可见,所提出方法在该数据集上取得了93.8%准确率的性能,相较于VGG为主干的RA-CNN与MA-CNN,分别高出了1.3%和1%,相比于Resnet 50为主干的MAMC、ISQRT-COV、NTS分别高出了1.0%、1.0%、0.5%、0.2%,对比ViT高出了0.3%.

表6 不同算法在Stanford Cars数据集上的准确率对比

方法	主干网络	使用标签	acc/%
BCNN[7]	D-Net + M-net	label	91.3
RA-CNN[40]	VGG	label	92.5
MA-CNN[41]	VGG	label	92.8
MAMC[30]	Resnet 50	label	92.8
ISQRT-COV[7]	Resnet 50	label	92.8
NTS[33]	Resnet 50	label	93.3
MGE-CNN[42]	Resnet 101	label	93.6
ViT[13]	ViT-B_16	label	93.5
ours	ConvNeXt	label	**93.8**

2.4.3 FGVC-Aircraft数据集实验与分析

FGVC-Aircraft数据集提供了10 200张飞机图片,只提供图像级标签(label)和标注框信息(BBox),按照variants进行划分,可分为100个类别,在细粒度任务中一般多采用这种划分方式. 由表7实验结果可

知,所提出算法在此数据集上取得了93.5%准确率的优秀性能,高于其他对比方法.

表 7　不同算法在FGVC-Aircraft数据集上的准确率对比

方法	主干网络	使用标签	acc/%
LIO[31]	Resnet 50	label	92.7
FDL[32]	Resnet 50	label	93.4
NTS[33]	Resnet 50	label	91.4
CIN[35]	Resnet 101	label	92.8
API-Net[36]	Resnet 101	label	93.4
DTB-Net[37]	Resnet 101	label	91.6
FBSD[38]	Densenet 161	label	93.2
ours	ConvNeXt	label	**93.5**

2.4.4　内标压印数据集实验与分析

对于如何建立基于手提包饰真伪图像的数据集,本文根据研究人工鉴定师的鉴定方法可知,鉴定师们凭借专业知识和经验对品牌指定区域进行详细观察,根据商品特定部位的物理特征如纹理、光泽、印压深度等特性判定其是否为假冒产品. 以路易威登(LOUIS VUITTON)高端手提包为例,专业鉴定师会根据内标压印上的文字和外部logo上文字的物理特征来对其进行真伪鉴定. 由图5可见,路易威登包袋上均会印有“Louis Vuitton”字样的内标压印,对于内标压印部位,赝品往往从纹路和深度上无法做到与真品一模一样,从这个部位的文字印章来鉴别真伪是鉴定师们常用的方法之一,也是较为可靠的鉴定手段之一. 同时对该部位进行真伪分类,十分符合细粒度分类任务中存在的类内差异大,类间差异小等特点. 基于此,本文建立了内标压印数据集,包含真伪两大类别,该数据集图像是用户通过手持相机拍摄后上传得来,为了确保数据集的准确性,其标签全部由专业鉴定师给出. 本文将细粒度算法应用于该数据集,通过完成真伪分类来开发一种非侵入式真伪识别方案,该方案可轻松区分制造商生产的原始产品与假冒的仿造产品.

对于内标压印数据集,本文搜集了8 277张该部位的真伪图像. 其中:4 139张为真品图像,4 138张为赝品图像. 在内标压印数据集上,首先对几个细粒度任务常用的主干网络进行了实验,由表8可见,Resnet 50、ViT、ConvNeXt在标压印数据集上准确率分别达到了90.4%、95.2%和95.6%,在内标压印数据集上ConvNeXt取得了最好的性能. 根据上述3个常用细粒度图像数据集上的实验,本文从中选取了几个性能优异的方法在该数据集上进行对比实验,具体算法包括FBSD、PMG、AFTrans以及TransFG. 所提出

(a) 真品　(b) 真品

(c) 真品　(d) 赝品

(e) 赝品　(f) 赝品

图 5　路易威登高端手提包内标压印部位真伪图像

方法在自建的内标压印数据集上达到了96.8%的准确率,优于其他算法. 同时通过在该数据集上的实验,验证了细粒度图像算法可有效应用于手提包的真伪鉴定工作中.

表 8　不同算法在内标压印数据集上的准确率对比

方法	主干网络	使用标签	acc/%
Resnet 50[25]	Resnet 50	label 1	90.4
ViT[13]	ViT-B_16	label 1	95.2
ConvNeXt[18]	ConvNeXt-B	label 1	95.6
FBSD[38]	Densenet 161	label 1	95.2
PMG[34]	Resnet 50	label 1	95.4
AFTrans[12]	ViT-B_16	label 1	96.4
TransFG[15]	ViT-B_16	label 1	96.1
ours	ConvNeXt	label 1	**96.8**

2.4.5　可视化分析

为了进一步验证所提出算法的有效性,本节将在CUB-200-2011数据集和自建的内标压印数据集上进行可视化实验分析. 如图6所示,对比原始基线模型,所提出方法除了能够聚焦到最显著区域,还能够关注到其他被忽视的关键点,这得益于REBS和GDM的共同作用,迫使网络全面挖掘多个不同的可辨别特征.

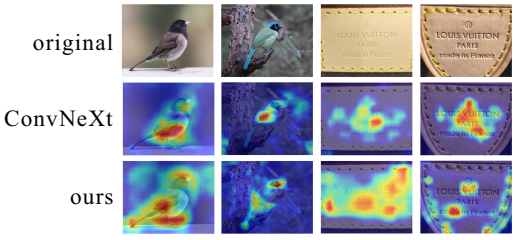

图 6　可视化效果图

3 结 论

本文提出了一种新的细粒度分类网络,在训练过程中提出了随机消除增强选择策略REBS,通过特征消除分支与特征增强分支相互作用,以此促进网络学习更多相关性的信息,同时为了进一步增强语义互补信息,提出了全局多样化模块GDM,使得网络不同层共享挖掘到的信息,从不同特征中进行对比来促进网络的分类性能. 所提出方法不需要边界框或部位标注信息,可进行弱监督的端到端训练. 实验结果表明,所提出网络在CUB-200-2011、Stanford Cars和FGVC-Aircraft三个公开数据集上分别达到了91.9%、93.8%和93.5%准确率的优异性能,优于其他先进方法,同时通过消融实验进一步表明了每个模块的有效性. 另一方面,本文还根据鉴定师的鉴定工作,提出了内标压印数据集,并将本文细粒度算法应用于商品真伪分类任务,为细粒度任务服务于实际应用提供了一个新思路.

参考文献(References)

[1] 罗建豪, 吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述[J]. 自动化学报, 2017, 43(8): 1306-1318.
(Luo J H, Wu J X. A survey on fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica, 2017, 43(8): 1306-1318.)

[2] 王松, 纪鹏, 张云洲, 等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, 37(1): 119-126.
(Wang S, Ji P, Zhang Y Z, et al. Adaptive receptive network for person re-identification[J]. Control and Decision, 2022, 37(1): 119-126.)

[3] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]. European Conference on Computer Vision. Cham, 2014: 834-849.

[4] Lin D, Shen X Y, Lu C W, et al. Deep LAC: Deep localization, alignment and classification for fine-grained recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 1666-1674.

[5] Ye Z H, Hu F Y, Liu Y, et al. Associating multi-scale receptive fields for fine-grained recognition[C]. IEEE International Conference on Image Processing. Abu Dhabi, 2020: 1851-1855.

[6] Wang Z H, Wang S J, Yang S H, et al. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 9746-9755.

[7] Lin T Y, Roy C A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]. IEEE International Conference on Computer Vision. Santiago, 2016: 1449-1457.

[8] Peng Y X, He X T, Zhao J J. Object-part attention model for fine-grained image classification[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2018, 27(3): 1487-1500.

[9] Rodríguez P, Velazquez D, Cucurull G, et al. Pay attention to the activations: A modular attention mechanism for fine-grained image recognition[J]. IEEE Transactions on Multimedia, 2020, 22(2): 502-514.

[10] Zheng H L, Fu J L, Zha Z J, et al. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 5007-5016.

[11] Ji R Y, Wen L Y, Zhang L B, et al. Attention convolutional binary neural tree for fine-grained visual categorization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10465-10474.

[12] Zhang Y, Cao J, Zhang L, et al. A free lunch from ViT: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 3234-3238.

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J/OL]. 2020, arXiv: 2010.11929.

[14] Wang J, Yu X, Gao Y. Feature fusion vision transformer for fine-grained visual categorization[J/OL]. 2021, arXiv: 2107.02341.

[15] He J, Chen J N, Liu S, et al. TransFG: A transformer architecture for fine-grained recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 852-860.

[16] Hu Y Q, Jin X, Zhang Y, et al. RAMS-trans: Recurrent attention multi-scale transformer for fine-grained image recognition[C]. Proceedings of the 29th ACM International Conference on Multimedia. New York, 2021: 4239-4248.

[17] Trockman A, Kolter J Z. Patches are all you need?[J/OL]. 2022, arXiv: 2201.09792.

[18] Liu Z, Mao H Z, Wu C Y, et al. A ConvNet for the 2020s[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 11966-11976.

[19] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference

on Computer Vision. Montreal, 2022: 9992-10002.

[20] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]. Conference on Computer Vision and Pattern Recognition. Colorado Springs, 2011.

[21] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[J/OL]. 2013, arXiv: 1306.5151.

[22] Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization[C]. IEEE International Conference on Computer Vision Workshops. Sydney, 2014: 554-561.

[23] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[R]. Pasadena: Caltech, 2011.

[24] Wei X S, Xie C W, Wu J X, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76: 704-714.

[25] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[26] Xie S N, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5987-5995.

[27] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4510-4520.

[28] Branson S, van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J/OL]. 2014, arXiv: 1406.2952.

[29] Chen T, Lin L, Chen R, et al. Knowledge-embedded representation learning for fine-grained image recognition[J/OL]. 2018, arXiv: 1807.00505.

[30] Sun M, Yuan Y, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition[C]. Proceedings of the European Conference on Computer Vision. Munich, 2018: 805-821.

[31] Zhou M H, Bai Y L, Zhang W, et al. Look-into-object: Self-supervised structure modeling for object recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11771-11780.

[32] Liu C B, Xie H T, Zha Z J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11555-11562.

[33] Yang Z, Luo T, Wang D, et al. Learning to navigate for fine-grained classification[C]. Proceedings of the European Conference on Computer Vision. Munich, 2018: 420-435.

[34] Du R, Chang D, Bhunia A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]. European Conference on Computer Vision. Glasgow, 2020: 153-168.

[35] Gao Y, Han X T, Wang X, et al. Channel interaction networks for fine-grained image categorization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 10818-10825.

[36] Zhuang P Q, Wang Y L, Qiao Y. Learning attentive pairwise interaction for fine-grained classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13130-13137.

[37] Zheng H L, Fu J L, Zha Z J, et al. Learning deep bilinear transformation for fine-grained image representation[J/OL]. 2019, arXiv: 1911.03621.

[38] Song J W, Yang R Y. Feature boosting, suppression, and diversification for fine-grained visual classification[C]. International Joint Conference on Neural Networks. Shenzhen, 2021: 1-8.

[39] Liu X D, Wang L L, Han X G. transformer with peak suppression and knowledge guidance for fine-grained image recognition[J]. Neurocomputing, 2022, 492: 137-149.

[40] Fu J L, Zheng H L, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4476-4484.

[41] Zheng H L, Fu J L, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]. IEEE International Conference on Computer Vision. Venice, 2017: 5219-5227.

[42] Ding Y, Zhou Y Z, Zhu Y, et al. Selective sparse sampling for fine-grained image recognition[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 6598-6607.

作者简介

刘光辉 (1976–), 男, 副教授, 博士, 从事计算机视觉理解、建筑环境智能感知与调控、建筑智能化技术领域等研究, E-mail: guanghuil@163.com;

占华 (1996–), 男, 硕士生, 从事深度学习、计算机视觉等研究, E-mail: 1075585751@qq.com;

孟月波 (1979–), 女, 教授, 博士, 从事计算机视觉理解、建筑环境智能感知与调控、建筑智能化技术领域等研究, E-mail: mengyuebo@163.com.