智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合前景特征增强与区域掩码自注意力的细粒度图像分类

刘万军, 赵思琪, 曲海成, 王宇萍

引用本文:
刘万军,赵思琪,曲海成,王宇萍. 结合前景特征增强与区域掩码自注意力的细粒度图像分类[J]. 智能系统学报, 2022, 17(6): 1134–1144.
LIU Wanjun,ZHAO Siqi,QU Haicheng,WANG Yuping. Combining foreground feature reinforcement and region mask self–attention for fine–grained image classification[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(6): 1134–1144.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202109029

DOI: 10.11992/tis.202109029
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20221008.0940.002.html

结合前景特征增强与区域掩码自注意力的细粒度图像分类

刘万军,赵思琪,曲海成,王宇萍
(辽宁工程技术大学 软件学院,辽宁 葫芦岛 125105)

摘 要:为解决细粒度图像分类中不相关背景信息干扰以及子类别差异特征难以提取等问题,提出了一种结合前景特征增强和区域掩码自注意力的细粒度图像分类方法。首先,利用 ResNet50 提取输入图片的全局特征;然后通过前景特征增强网络定位前景目标在输入图片中的位置,在消除背景信息干扰的同时对前景目标进行特征增强,有效突出前景物体;最后,将特征增强的前景目标通过区域掩码自注意力网络学习丰富、多样化且区别于其他子类的特征信息。在训练模型的整个过程,建立多分支损失函数约束特征学习。实验表明,该模型在细粒度图像数据集 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 的准确率分别达到了 88.0%、95.3% 和 93.6%,优于其他主流方法。

关键词:细粒度图像分类;目标定位;区域掩码;自注意力;多样化特征;特征增强;残差网络;深度学习
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1673−4785(2022)06−1134−11

中文引用格式:刘万军,赵思琪,曲海成,等. 结合前景特征增强与区域掩码自注意力的细粒度图像分类 [J]. 智能系统学报, 2022, 17(6): 1134–1144.
英文引用格式:LIU Wanjun, ZHAO Siqi, QU Haicheng, et al. Combining foreground feature reinforcement and region mask self-attention for fine-grained image classification[J]. CAAI transactions on intelligent systems, 2022, 17(6): 1134–1144.

Combining foreground feature reinforcement and region mask self-attention for fine-grained image classification

LIU Wanjun, ZHAO Siqi, QU Haicheng, WANG Yuping
(School of Software, Liaoning Technical University, Huludao 125105, China)

Abstract: This study presents a method of foreground feature reinforcement and region mask self-attention for fine-grained image classification due to the difficulty in extracting subtle features of subordinate classes that are difficult to distinguish irrelevant background noise interference. The ResNet50 is used first to extract global features of the input image, followed by the foreground feature reinforcement, which predicts the position coordinates of the foreground object in the input image. While eliminating background information interference, the features of foreground objects are enhanced to effectively highlight foreground objects. Finally, the region mask self-attention network is used to teach feature-enhanced foreground objects with rich and diverse fine-grained information that is different from other subclasses. The multi-branch loss function constrains the network's feature learning throughout the process. The comprehensive experiments show that our approach outperforms other mainstream methods on CUB-200-2011, Stanford Cars datasets, and FGVC-Aircraft, with 88.0%, 95.3%, and 93.6%, respectively.

Keywords: fine-grained image classification; object localization; region-based mask; self-attention; diverse feature; feature reinforcement; residual network; deep learning

如何让机器分辨出不同狗的种类?[1] 这是一个很常见的问题。由于狗的物理结构特征相似,因此要找到每只狗的局部细微特征才能正确区分不同狗的种类。在计算机视觉研究中的细粒度图像分类就是解决识别不同子类间的类别问题,如

收稿日期:2021−09−15. **网络出版日期**:2022−10−08.
基金项目:国家自然科学基金面上基金项目 (42071351);辽宁省教育厅基础研究项目(LJ2019JL010).
通信作者:刘万军. E-mail: liuwanjun@lntu.edu.cn.

分辨鸟的种类[2]和飞机的型号[3]。

细粒度图像分类的难点归纳为以下几点：1)类间差异小, 属于不同类别的对象除了一些细微的差别外, 可能非常相似; 2)类内差异大, 属于同一类别的对象通常呈现不同的姿态; 3)具有多尺度特征。由于拍摄距离、角度以及目标不同的姿态, 输入图像中的目标比例会变化很大。

当前针对细粒度图像分类方法包括强监督训练和弱监督训练。强监督细粒度图像分类算法是指在训练模型时, 不仅使用了类别标签, 而且还使用了额外的标注信息。早期一些研究工作就是利用额外的人工部件标注点和边界框直接定位对象的关键语义部分(如鸟类的头部、尾部和躯干等关键局部特征)。Zhang 等[4]提出的基于部件算法, 通过选择性搜索[5]产生关键部位(整体、头部和身体区域)的候选框, 利用 R-CNN(regions with CNN)[6]完成对这些部位的检测, 利用约束条件对提取的关键部位信息进行修正, 进行卷积特征提取, 将不同区域的特征进行连接, 最后通过支持向量机(support vector machine, SVM)分类器进行分类训练。Branson 等[7]提出了姿态归一化算法, 对图片先局部定位和特征提取, 根据局部定位结果剪裁图片, 分别提取多层次的卷积特征, 送入 SVM 分类器进行分类。但额外的手工标注信息费用昂贵, 且容易出现局部语义信息的判断错误, 导致分类性能下降, 限制了细粒度算法应用的可扩展性[8]。因此在仅利用类别标签的前提下, 用弱监督的方式训练模型成为近期研究热点[9]。

弱监督细粒度图像分类方法是学习输入图像和输出标签之间的映射。一部分学者利用单阶段方法直接学习对象的细粒度特征, Lin 等[10]提出了双线性模型, 使用两个独立的卷积神经网络计算成对的特征交互来捕获图像的局部差异, 但是双线性特征表示通常存在高维问题, 增加了计算量, 需要大量的训练样本进行拟合。还有一些学者关注如何定位物体的前景或语义部分, 提取可区分性区域的特征。Zheng 等[11]提出了递归注意力卷积神经网络算法(recurrent attention convolutional neural network, RA-CNN), 通过多分支的循环网络逐步定位关键性局部区域, 同时将该区域进行剪裁放大, 然后再在该放大的区域寻找判别性特征, 即在重点中寻找重点。但是 RA-CNN 只学习到单一的判别性特征, 忽略了其他语义特征。Zheng 等[12]提出了多注意力神经网络算法, 通过通道分组损失函数产生用于聚类的多个局部区域。但局部语义区域的数量有限, 存在细微且判别性特征丢失的问题。

弱监督细粒度图像分类方法已经取得很大进展, 但是仍存在以下问题有待解决:

1)输入图片中的物体尺度变化。当物体只占据输入图片的一小块区域时, 在特征学习过程中, 目标特征很容易被背景抹去, 降低分类准确率。

2)细粒度图像具有类内差异大、类间差异小的特点, 因此只能借助于微小的局部差异才能区分出不同的子类别。

为此, 本文提出了一种结合前景特征增强和区域掩码自注意力的细粒度图像分类方法。本文的主要工作:

1)提出前景特征增强模块, 消除背景噪音干扰和前景目标多尺度变化对前景特征提取的影响, 实现前景目标增强, 有效突出前景;

2)提出区域掩码自注意力模块, 利用掩码机制的特性, 遮挡激活映射图的高响应特征, 从而使网络关注局部细微特征, 充分学习到不同子类间的局部差异性特征, 挖掘出更多有用信息;

3)多分支损失函数的协同作用共同约束网络的特征学习。

1　相关研究

1.1　目标定位

图像中目标检测和目标定位是计算机视觉中重要而又具有挑战性的任务。目标定位大致归纳为 3 类: 强监督方法、弱监督方法和无监督方法。本文不介绍强监督定位方法。

弱监督物体定位用类别标签实现目标定位, 早期是用类激活映射(class activation mapping, CAM)[13]实现目标定位, CAM 通过全局平均池化生成针对每一张图片的激活映射图, 这个激活映射图能反应出物体的大致位置。但是用交叉熵损失函数训练模型时, 通常会使模型关注高响应的局部区域而非整个目标, 同时没有充分利用浅层特征信息。Wei 等[14]对浅层和深层特征图进行元素相乘生成 CAM, 可以滤出背景噪音并同时生成更清晰的边界。Sohn 等[15]提出了一种新颖的残差细粒度注意方法, 此方法通过利用分布在通道和特征图中的位置信息, 结合残差操作, 自动激活对象的较少激活区域, 生成目标边界框。Pan 等[16]则充分利用卷积特性中包含的结构信息, 利用高阶自相关提取模型中保留的固有结构信息, 聚合多个模型的高阶自相关点实现精确定位。Zhang 等[17]提出了通过训练两个对抗互补分类器

来发现整个目标的互补学习方法，该分类器可以定位不同目标的部分，发现属于同一目标的互补区域。

无监督目标定位更具有挑战性，因为它只需要一张图片实现前景目标定位，不需要任何辅助信息。一些研究表明，卷积激活图能够同时表示空间信息和语义信息，并具有显著的定位能力。Wei 等[18]提出了一种选择性卷积描述子聚合方法（selective convolutional descriptor aggregation, SCDA），融合多层卷积特征实现目标定位。然后，采用阈值策略定位细粒度图像中的目标。但是定位效果不理想，一些重要信息丢失。因此前景特征增强模块是在 SCDA 算法上引入上下文注意力增强像素的空间相关性，提高前景目标的定位能力。

1.2　掩码注意力

注意力机制是有选择的关注重点数据，而不是平等对待全部数据。注意力机制是通过神经网络自动地学习特征的权重分布，并以“动态加权”的方式施加在特征之上进一步强调感兴趣的区域，并同时抑制不相关背景区域。

掩码注意力机制不同于普通的注意力机制，它是对被选择的重点数据进行遮掩，使网络关注其他数据信息。通常来讲，首先通过通道池化生成注意力图，然后对注意力图进行归一化操作，接近 1 的像素点是判别性特征，反之是细微特征。掩码注意力的应用范围很广，在目标定位、行人重识别、图像分类、3D 点云上都有所应用。

Qiao 等[19]使用掩码注意力生成注意权重，分配给文本实例。它允许一个图像中的不同文本实例被分配到不同的特征映射通道上，这些特征映射进一步被分组为一批实例特征。Wang 等[20]通过预训练好的语义分割模型产生辅助监督信号，即掩码注意力，实现了判别表示学习。这个掩码注意力让分类器过滤掉图像中不重要的部分。Sun 等[21]利用掩码注意力弱化高响应区域的特征值，使模型可以挖掘出图像中更有价值的区域。Li 等[22]通过变形卷积和掩码注意力，将稀疏的特征映射到目标区域，同时用掩码注意力突出复杂背景下的目标像素值。Xie 等[23]通过一个可见区域边界框信息生成一个空间掩码，同时调节由 RoI（region of interest）层生成的多通道特征。这些掩码有效地强化可见区域特征，隐藏模糊区域特征。Choe 等[24]通过对高响应特征进行掩码遮挡，强化低响应特征，有助于网络关注输出目标的轮廓信息。

1.3　残差网络

残差网络解决了网络层数增加引起的梯度弥散或梯度爆炸问题。残差网络由一系列残差块组成，残差块有两种，分别是恒等映射块和卷积块。其中恒等映射块不改变维度，卷积块改变特征维度。本文模型使用的是 ResNet50。其中，conv2、conv3、conv4、conv5 分别有 3、4、6、3 个残差块，每个残差块有 2 个 1×1 卷积和 1 个 3×3 卷积，conv2 到 conv5 有 48 个卷积层，再加上 conv1 层的 7×7 卷积层和 3×3 最大池化层，共有 50 个卷积层。

2　实验方法

为了解决网络在提取特征时，会掺杂背景信息以及无法提取多样化的局部区域特征等问题，本文提出了结合前景特征增强和区域掩码自注意力网络模型（foreground feature reinforcement and region mask attention, FFRMA）。FFRMA 整体框架如图 1 所示。

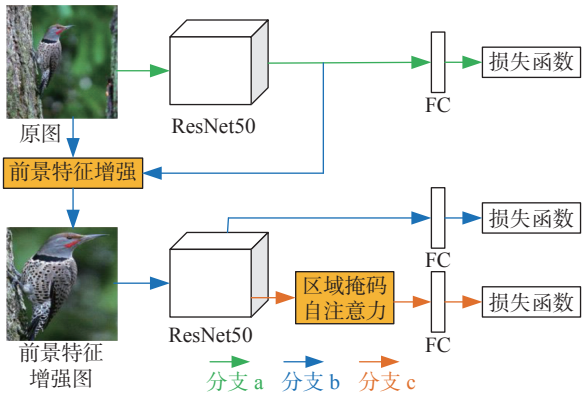


图 1　FFRMA 整体框架

Fig. 1　Overall framework of method FFRMA

由图 1 可以看出，FFRMA 框架主要由 3 部分组成：1) 前景特征增强模块（foreground feature reinforcement, FFR），定位输入图片的前景目标，然后从原图中剪裁前景目标实现特征增强；2) 区域掩码自注意力模块（region mask attention, RMA），用来提取更多的多样化局部细微且重要的特征；3) 多分支损失函数约束网络学习特征的能力。

FFRMA 是多分支网络结构，使用同一个 ResNet50 作为特征提取网络，共享 ResNet50 网络的全部参数信息（图 1 中画了两个 ResNet50 网络，但其实用的是同一个 ResNet50）。首先将原图送入预训练的 ResNet50 提取全局特征，输出卷积结果（图 1 分支 a）；然后将卷积结果和原图送入前景特征增强模块定位前景目标，将前景目标放大到原图尺寸，生成前景特征增强图，有效避免背景信息的干扰，接下来把前景特征增强图送入 ResNet50

进行特征提取, 输出卷积结果(图 1 分支 b); 最后将卷积结果通过区域掩码自注意力模块学习到更多的微小但重要的细粒度特征(图 1 分支 c)。整个过程以多分支损失函数约束特征学习。

2.1　前景特征增强

在细粒度图像分类任务中, 使用边界框裁剪目标对象可以减少背景噪音, 使目标对象特征加强从而提高分类准确率, 但是人工标注的边界框信息代价过于昂贵。前景特征增强模块解决的就是在仅有输入图片的前提下, 利用卷积特征的分布响应定位输入图片的前景目标, 消除背景噪声的干扰并同时对前景特征进行增强。FFR 模型结构如图 2 所示。

FFR 模块在 SCDA 算法上引入了上下文注意力(context attention, CA)。因为卷积运算会导致局部感受野, 前景目标的一些局部细微特征存在一些差异, 这种差异会造成类内不一致性, 即属于前景目标的局部细微特征被归为背景区域, 直接影响前景目标的定位性能。因此 CA 主要是在特征间建立全局空间的上下文关联, 使用注意力机制捕获特征图任意两个位置之间的空间依赖关系, 减少类内不一致性。CA 结构如图 3 所示。

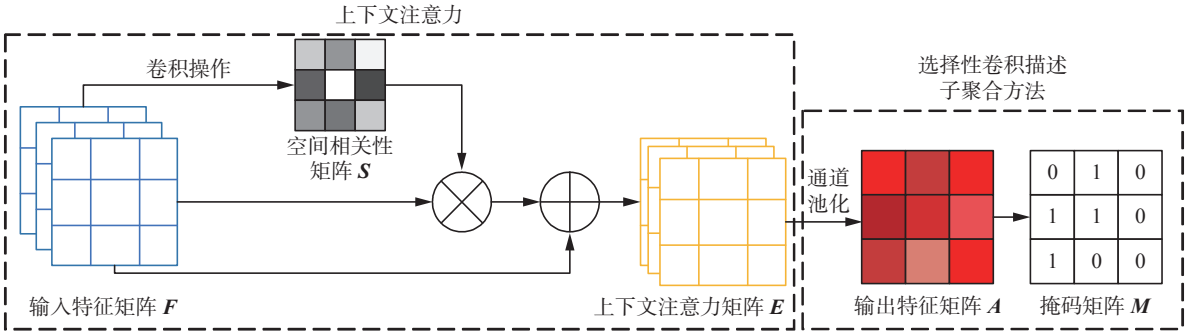


图 2　前景特征增强模型结构

Fig. 2　Architecture of FFR

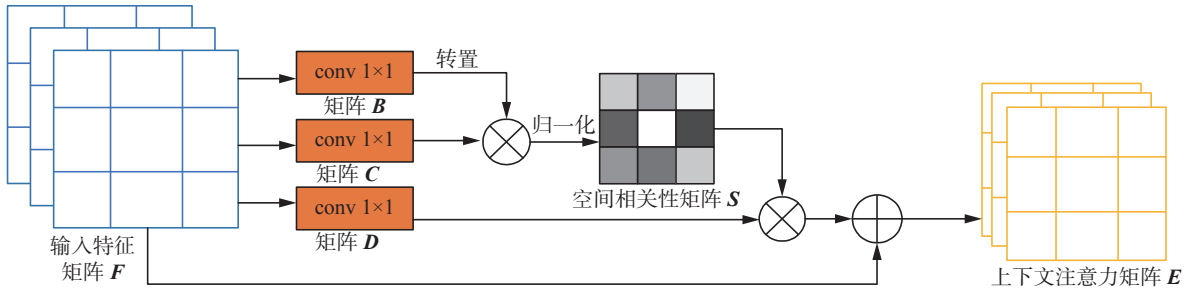


图 3　上下文注意力

Fig. 3　Context attention

假设$\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$表示输入图像的最后一个卷积特征图, 其中$C$表示通道数, 空间大小为$H\times W$。然后分别通过 3 个 1×1 卷积层得到 3 个特征图, 对 3 个特征图进行 reshape 操作得到$\{\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\}\in\mathbf{R}^{C\times N}$, 其中$N=H\times W$, 将$\boldsymbol{B}^{\mathrm{T}}$与$\boldsymbol{C}$进行乘积运算, 利用 Softmax 函数对结果进行归一化得到空间相关性矩阵$\boldsymbol{S}\in\mathbf{R}^{N\times N}$, 表示为

$$\boldsymbol{S}_{ji}=\frac{\exp(\boldsymbol{B}_i\cdot\boldsymbol{C}_j)}{\sum_{i=1}^{N}\exp(\boldsymbol{B}_i\cdot\boldsymbol{C}_j)} \tag{1}$$

式中: $\boldsymbol{S}_{ji}$表示i位置对j位置的影响, 两个位置的特征越相似, 它们之间的相关性就越强, 即属于同一个目标。

$\boldsymbol{D}$与空间相关性矩阵$\boldsymbol{S}$进行乘积运算得到中间矩阵, 将中间矩阵的空间形状恢复到$H\times W$, 再与特征图$\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$相加得到上下文注意力矩阵$\boldsymbol{E}\in\mathbf{R}^{C\times H\times W}$。矩阵$\boldsymbol{E}$表示了长距离上下文的特征信息, 增强了像素点间的空间相关性。$\boldsymbol{E}$按通道方向进行求和, 得到激活图$\boldsymbol{A}$, 具体为

$$\boldsymbol{A}=\sum_{i=0}^{C-1}f_i \tag{2}$$

式中f_i表示第i个通道的激活图。在不同通道维度上, 局部特征具有不同激活响应值。每一个通道对应的激活图学习到的细粒度局部区域特征都有所不同, 对应的最高响应的区域也不同。因此, 通过对激活图$\boldsymbol{A}$沿通道方向进行聚合, 将目标出现的位置进行响应累加确定目标的整体轮廓。为了准确地定位物体, 设置一个阈值$\bar{a}$ ($\boldsymbol{A}$的均值),

定义为

$$\bar{a}=\frac{\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}A(x,y)}{H\times W} \tag{3}$$

阈值$\bar{a}$是用来确定激活图$\boldsymbol{A}$中某个位置是否是目标的一部分。将小于阈值$\bar{a}$的响应标记为 0，大于阈值$\bar{a}$的位置标记为目标所在，$M_{(x,y)}$定义为

$$\boldsymbol{M}_{(x,y)}=\begin{cases}1, & \boldsymbol{A}_{(x,y)}>\bar{a}\\ 0, & \boldsymbol{A}_{(x,y)}\leqslant\bar{a}\end{cases} \tag{4}$$

不同的卷积特征映射对目标具有不同的激活响应，因此 FFR 将融合不同卷积层的特征值实现精确定位。假设 layer4_2、layer4_3 分别是 ResNet50 网络中的 conv5 层的特征图，通过式(4)获得相应的$M_{(4,2)}$、$M_{(4,3)}$。然后对$M_{(4,2)}$、$M_{(4,3)}$进行点乘获得准确的$\boldsymbol{M}$。

细粒度物体通常处于$\boldsymbol{M}$中最大的联通分量中，因此使用包含最大联通区域的最小边界框作为定位对象的结果，再将结果调整为输入图像$\boldsymbol{X}$的大小。

2.2　区域掩码自注意力

不同于通用图像分类，细粒度图像分类不仅需要关注全局语义特征还要关注区别于其他子类的局部细微特征，因此设计了区域掩码自注意力模块(region mask attention, RMA)。该模块的整体结构如图 4 所示。

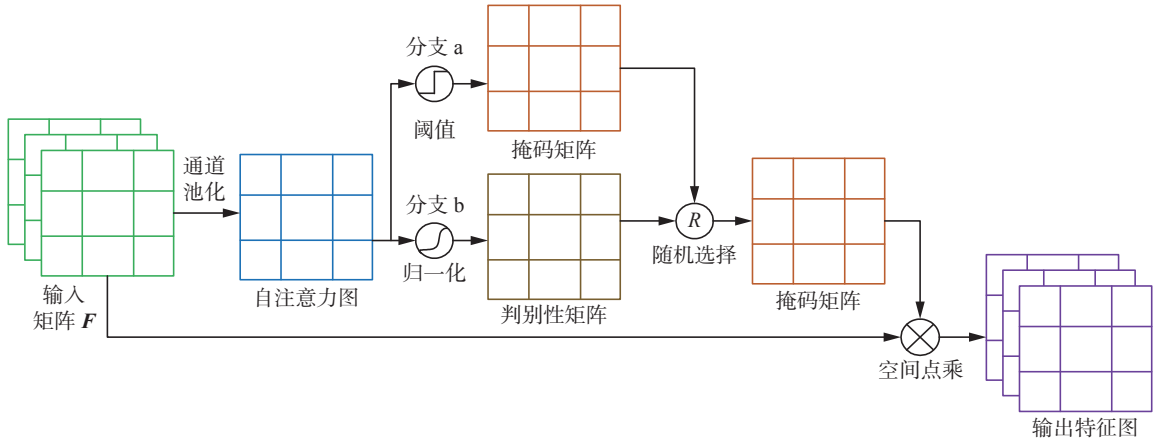


图 4　区域掩码自注意力模型结构

Fig. 4　Structural diagram of RMA

RMA 通过自注意力图产生两个作用不同的特征矩阵，即掩码矩阵$\boldsymbol{M}_{\text{mask}}$和判别性矩阵$\boldsymbol{M}_{\text{dis}}$，$\boldsymbol{M}_{\text{dis}}$学习高响应区域的特征，$\boldsymbol{M}_{\text{mask}}$通过设置阈值抑制高响应区域，使网络关注局部细微特征，然后以均匀分布概率 p 随机选择特征矩阵，$\boldsymbol{M}_{\text{mask}}$和$\boldsymbol{M}_{\text{dis}}$的相互合作，可以让网络学习到不同的特征。

RMA 模块的设计理念是为了提取所有对细粒度分类有用的特征信息，而不是只提取一些识别能力强的高响应信息。此外，RMA 并没有引入额外的参数开销。

具体来讲，RMA 首先通过通道平均池化，将输入特征图$\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$映射到自注意力图$\boldsymbol{M}_{\text{att}}$，自注意力图中每个像素的大小可以有效近似出目标中判别性特征的空间分布。随后，从自注意力图中产生两个分支，图 4 中分支 a 通过阈值γ抑制自注意力图中的高响应区域，得到掩码特征矩阵$\boldsymbol{M}_{\text{mask}}$，使网络聚焦在除高响应区域之外的局部细微特征。图 4 中分支 b 用 Sigmoid 函数对自注意力图进行归一化得到判别特征矩阵$\boldsymbol{M}_{\text{dis}}$，$\boldsymbol{M}_{\text{dis}}$中接近 1 的像素点代表判别性特征，通过对分支 b 的训练可以让网络关注高响应区域，学习判别性特征。分支 a 和分支 b 的协同合作，使得输入图像的所有有用特征都可以被模型学习。在每次迭代过程中，以均匀概率 p 随机选择分支 a 或分支 b，大于 0.5 的概率选择分支 a，小于 0.5 的概率选择分支 b。最后将随机选择的特征矩阵与输入特征图$\boldsymbol{F}$进行点乘得到输出特征图。图 4 中随机选择的是掩码矩阵$\boldsymbol{M}_{\text{mask}}$。

RMA 中有一个超参数是阈值γ。自注意力图中超过阈值γ的区域视为高响应区域。$\boldsymbol{M}_{\text{mask}}$中像素值为 0 代表高响应区域，像素值为 1 代表其他区域信息，$\boldsymbol{M}_{\text{mask}}$表示为

$$\boldsymbol{M}_{\text{mask}}=\begin{cases}0, & \boldsymbol{M}_{\text{att}}>\gamma\\ 1, & \boldsymbol{M}_{\text{att}}\leqslant\gamma\end{cases} \tag{5}$$

2.3　损失函数

为了使 FFRMA 模型能够充分有效地学习通过 FFR 和 RMA 获得的图像特征，设计了多分支损失函数。在训练阶段，FFRMA 是一个由 3 个分

支组成的网络结构, 不同分支学习到的特征不一样, 图 1 中分支 a 关注输入图片的整体特征; 分支 b 借助分支 a 中原始图像的特征映射获取前景目标的边界框信息, 裁剪边界框并放大到输入图片的大小, 实现特征增强。前景特征增强既包括目标的结构特征, 又包括细粒度特征; 分支 c 抑制前景特征图的判别性特征, 使网络充分学习到不同的局部细微多样性特征。3 个分支使用交叉熵函数作为分类损失, 分别表示为

$$L_{\text{raw}} = -\ln(P_r(c)) \tag{6}$$

$$L_{\text{object}} = -\ln(P_o(c)) \tag{7}$$

$$L_{\text{ram}} = -\ln(P_d(c)) \tag{8}$$

式中: c表示输入图片的类别标签; P_r、P_o、P_d分别代表 3 个分支中最后一个 Softmax 层输出的类别概率。总损失函数表示为

$$L_{\text{total}} = L_{\text{raw}} + L_{\text{object}} + L_{\text{ram}} \tag{9}$$

总损失函数是 3 个分支损失函数之和, 用以优化模型在反向传播时的性能。3 个分支损失函数的协同合作可以加快网络模型的收敛速度和提高模型的特征学习能力。分支 a 输出原始图片的粗粒度分类概率; 分支 b 获得前景目标的边界框, 从原始图片裁剪相应的区域并放大到原图尺寸, 将其送入网络获得细粒度预测概率。最终分类结果是对粗粒度分类结果和细粒度分类概率取平均值。

3　实验与结果分析

3.1　实验细节

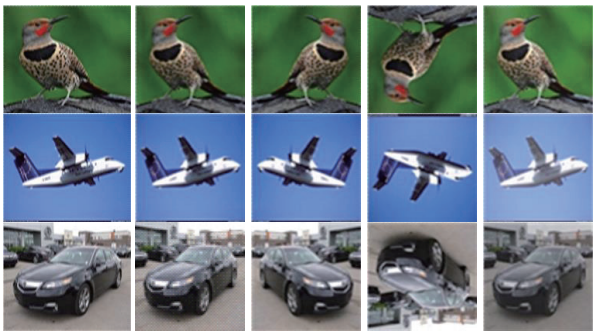
3.1.1　细粒度图像数据集与预处理

为了证明本文方法的有效性, FFRMA 在 3 个细粒度图像数据集 CUB-200-2011、FGVC-Aircraft 和 Stanford Cars[25] 上进行评估。3 个数据集的详细信息见表 1。由于细粒度数据集中每一类样本数量过少, 可能在模型训练的时候出现欠拟合现象。因此, 在训练 FFRMA 模型之前对数据集进行数据增强。具体数据增强方式如图 5 所示, 图 (a) 为原始图片、图 (b) 为缩放、图 (c) 为水平翻转、图 (d) 为垂直翻转、(e) 图为增强图片颜色和对比度。其中, 图片的水平翻转和垂直翻转都是以概率 0.5 进行数据扩充。

表 1　数据集信息

Table 1　Datasets information

数据集	对象	类	训练集	测试集
CUB-200-2011	bird	200	5994	5794
FGVC-Aircraft	aircraft	100	6667	3333
Stanford Cars	car	196	8144	8041



(a) 原图像　(b) 缩放　(c) 水平翻转　(d) 垂直翻转　(e) 亮度

图 5　3 个数据集中训练样本的数据增强

Fig. 5　Data augmentation of training samples in three datasets

3.1.2　实验环境及参数设置

FFRMA 模型使用预训练的 ResNet50 作为基础网络。所有的输入图片通过双线性插值调整到 448×448 大小, 在训练阶段对图片进行水平和垂直翻转, 测试阶段对图片仅做归一化操作。在训练时通过图 1 中 3 分支完成提取图像不同的局部特征和定位输入图片的前景物体, 测试阶段仅由图 1 中分支 a 和分支 b 完成细粒度图像分类。在训练阶段使用 3 个分支提高模型的鲁棒性, 3 个分支的作用各不相同, 彼此缺一不可, 共同完成图像局部细微且判别性特征的学习。而在测试阶段, 分支 c 主要用于获得局部细微特征, 但是经过训练阶段, 网络能够很好地学习局部细微特征, 因此去除了分支 c。分支 a 输出原始图片的粗粒度分类概率, 分支 b 获得前景目标的边界框, 从原始图片裁剪相应的区域并放大到原图尺寸, 将其输入网络获得细粒度预测概率。最终分类结果是对粗粒度分类预测概率和细粒度分类预测概率取平均值。

参数设置: FFRMA 模型采用随机梯度下降法 (stochastic gradient descent, SGD) 优化模型, 动量为 0.9, 权重衰减为 0.000 1, epoch 为 200, batch 为 6, 初始学习率为 0.001, 每经过 60 次 epoch 学习率乘上 0.1[26]。

实验设备: 实验环境为 Ubuntu 18.04.5, GeForce RTX 2080 Ti, 运行内存为 128 GB, 使用 1 个显卡进行训练。模型训练平台为基于开源深度学习框架 PyTorch, 版本为 PyTorch 1.2.0, Python 版本为 Python 3.7。

3.1.3　评价指标

在细粒度图像分类领域中, 研究者采用准确率作为评价指标[27]。因此本文方法采用分类准确率作为评估标准, 分类准确率为

$$\text{Accuracy} = \frac{R_a}{R} \tag{10}$$

式中：R为测试集的图片数量，R_a是测试实验中正确分类的样本数量。

3.2　实验结果及对比

3.2.1　消融实验

为了证明本文模型的有效性，以 CUB-200-2011 数据集为例，验证 FFR 和 RMA 的有效性。CAM、SCDA 和 FFR 的基础网络都为在 ImageNet 数据集上预训练的 ResNet50，且参数设置保持一致，数据集中的每一张图片采用中心裁减，将大小调整为 448×448。

为了验证 FFR 的定位性能，以目标边界框定位的准确性（percentage of correctly localized object, PCO）作为评价指标。PCO 是指预测物体边界框与真值框的交并比值。

FFR 模块预测的是检测框与真值框之间的交并比值，将前景目标定位的准确率转化为检测框与真值框的交并比大小，更加准确地验证前景特征放大模块的定位效果。接下来，从定量的角度分析前景特征放大算法的正确性，在基准网络 ResNet50 中，以数据集 CUB-200-2011 为例，分别与 CAM、SCDA 和 FFR 进行消融对比实验，以 PCO>0.5 作为 FFR 模块的定量分析指标，实验结果见表 2。

表 2　目标定位准确率

Table 2　Accuracy of object localization

方法	基础网络	PCO /%
CAM	ResNet50-GAP	65.7
SCDA(layer4_3)	ResNet50	76.8
FFR (layer4_3)	ResNet50	81.2
FFR (layer4_2&layer4_3)	ResNet50	**85.3**

为了验证 RMA 模块中掩码矩阵$\boldsymbol{M}_{\text{mask}}$和判别性矩阵$\boldsymbol{M}_{\text{dis}}$对于网络学习特征能力的影响，在 FFR 的基础上，对每一个特征矩阵进行实验。实验结果见表 3。

表 3　RMA 模块不同组件的消融实验对比

Table 3　Ablation experiment of different components of the RMA model

方法	基础网络	精度/%
FFR	ResNet50	87.20
FFR+$\boldsymbol{M}_{\text{mask}}$	ResNet50	57.35
FFR+$\boldsymbol{M}_{\text{dis}}$	ResNet50	87.45
FFR+RMA	ResNet50	**88.0**

以结合 FFR 模块的 ResNet50 作为基准网络，分别测试掩码矩阵和判别性矩阵的有效性。通过分析表 3 中数据可以得知，结合了判别性矩阵$\boldsymbol{M}_{\text{dis}}$的 FFR 模块的准确率比单独使用 FFR 模块提高了 0.25%，但是结合了掩码矩阵$\boldsymbol{M}_{\text{mask}}$的 FFR 模块的准确率只有 57.35%，因为掩码矩阵掩盖掉大多数的高响应判别性特征（如鸟头、鸟喙等判别性特征），此时网络学习到的特征是不具有类针对性，虽然学习到了一些细微特征（如鸟喙的颜色），但是这些细微特征还是要辅以高响应特征才能发挥作用，因此单独使用掩码矩阵会导致模型判别能力不足。FFR 和 RMA 的结合，精度提高了 0.82%，说明 RMA 模块中两个矩阵对分类精度都做出了贡献。

3.2.2　RMA 参数敏感性分析

为了确定阈值γ对 RMA 的作用，分别在 CUB-200-2011、FGVC-Aircraft 和 Stanford Cars 数据集上进行实验，实验结果如图 6 所示。

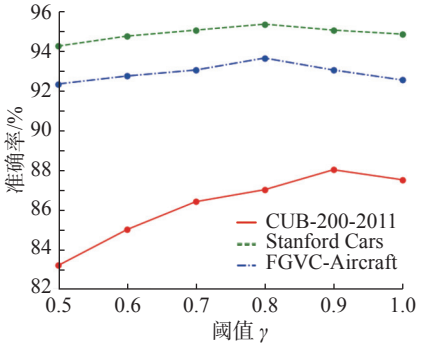


图 6　参数分析

Fig. 6　Parameter analysis

CUB-200-2011 数据集在阈值γ=0.9 时取得最大精度，88.0%；FGVC-Aircraft 数据集在阈值γ=0.8 时取得最大精度，93.6%；Stanford Cars 数据集在阈值γ=0.8 时取得最大精度，95.3%。观察图 6 中 3 条曲线的波动情况，可以得出，阈值γ对 CUB-200-2011 数据集的影响较大，曲线波动较为明显；阈值γ对数据集 FGVC-Aircraft 和 Stanford Cars 的影响较小，曲线波动不明显。

CUB-200-2011 对γ的敏感性比较大，原因是 CUB-200-2011 中鸟的种类繁多，每种鸟类都有丰富且多样性的局部特征，通过控制γ的取值可以有效学习区别于其他子类的局部细微且判别性特征。另外，当γ=1 时 3 个数据集的精度都有些下降，因为γ=1 表示仅抑制特征图中的峰值，峰值是通用图像的分类依据（区分大类）。对于细粒度分类而言，只关注峰值是不够的，也需要关注其他特征区域，这些区域包括更多的子类间差异特征。

3.2.3　对比实验

为了更进一步分析本文方法的分类性能, 将本文方法与其他方法进行比较, 实验结果见表 4。其中, FFRMA (Coarse) 是原始图像经过图 1 中分支 a 得到的分类精度, FFRMA (Finer) 是原始图像经过图 1 中分支 b 得到的分类精度, FFRMA (Ours) 是对分支 a 和分支 b 分类精度的均值。

表 4　不同弱监督细粒度图像分类方法实验对比

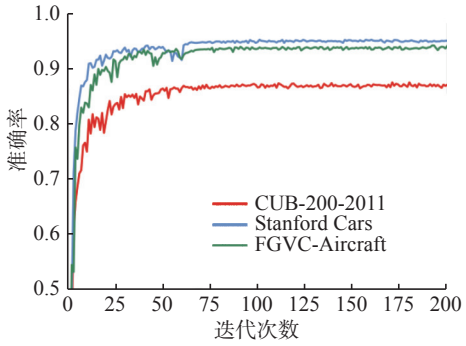
Table 4　Experimental comparison of different weakly supervised fine-grained image classification methods　%

方法	基础网络	CUB-200-2011	FGVC-Aircraft	Stanford Cars
Cross-X[28]	ResNet-50	87.7	92.6	94.6
DCL[29]	ResNet-50	87.8	93.0	94.5
ELoPE[30]	ResNet-50	87.4	93.4	94.5
MBPOL[31]	ResNet-50	87.75	91.1	93.84
NTS-Net[32]	ResNet-50	87.5	91.4	93.9
LAFE[33]	ResNet-50	87.6	93.6	94.8
FFRMA(Coarse)	ResNet-50	86.5	92.6	93.3
FFRMA(Finer)	ResNet-50	87.8	93.2	94.8
FFRMA(Ours)	ResNet-50	88.0	93.6	95.3

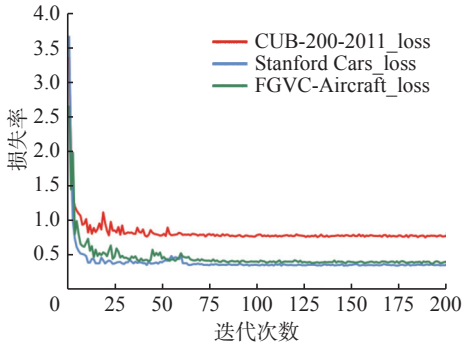
表 4 的实验数据说明了在与弱监督细粒度分类的主流方法进行对比时, FFRMA 模型在 3 个数据集上均取得不错的分类成绩。FFRMA 在 Cars 数据集的精度最高为 95.3%, 与性能最好的 LAFE 模型相比, 提高了 0.5%; 其次, 在 Aircraft 数据集取得精度为 93.6%, 与 LAFE 在 Aircraft 数据集的准确率相同, 但是 FFRMA 模型在另两个数据集上的精度均超过了 LAFE; FFRMA 在 CUB 数据集的精度为 88.0%。ELoPE、MBPOL 和本文方法都是基于定位–分类的细粒度图像分类, 但是 FFRMA 在 3 个数据集的准确率均优于 ELoPE、MBPOL。

图 7 给出了 FFRMA 模型在 CUB-200-2011、FGVC-Aircraft 和 Stanford Cars 数据集上准确率和损失率的变化趋势。随着迭代次数的增多, 分类准确率逐渐上升, 在 125 次迭代后, 模型完全收敛, 分类准确率也达到最大。网络刚进行训练时, 测试集随着训练次数的增加不断减少, 网络模型参数在不断更新优化, 在 75 次迭代后, 模型基本收敛, 在 125 次迭代后, 优化算法找到极值点, 损失值几乎不变, FFRMA 模型的参数基本稳定。

对 SCDA 和 FFR 的定位效果进行可视化, 如图 8 所示。其中图 8(a)是卷积算子增强方法 (SCDA) 可视化的结果, 图 8(b)是前景特征增强方法可视化的效果。红色矩形代表真值框, 绿色代表网络学习到的边界框。通过图 8 的可视化结果, FFR 模块的定位性能优于 SCDA 方法, 边界框更加贴合真值框, 消除了更多的背景噪音, 获得了更纯粹的前景目标特征, 进一步提高了模型的表征能力。



(a) 精度变化趋势

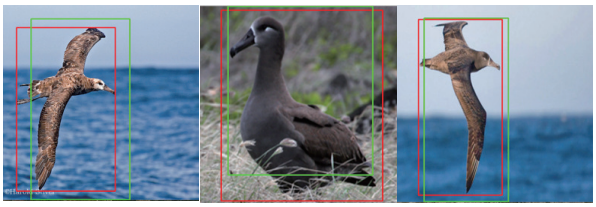


(b) 损失率变化趋势

图 7　FFRMA 准确率和损失率的变化趋势

Fig. 7　Accuracy and loss trend of FFRMA

(a) 卷积算子增强方法



(b) 前景特征增强方法 (本文方法)

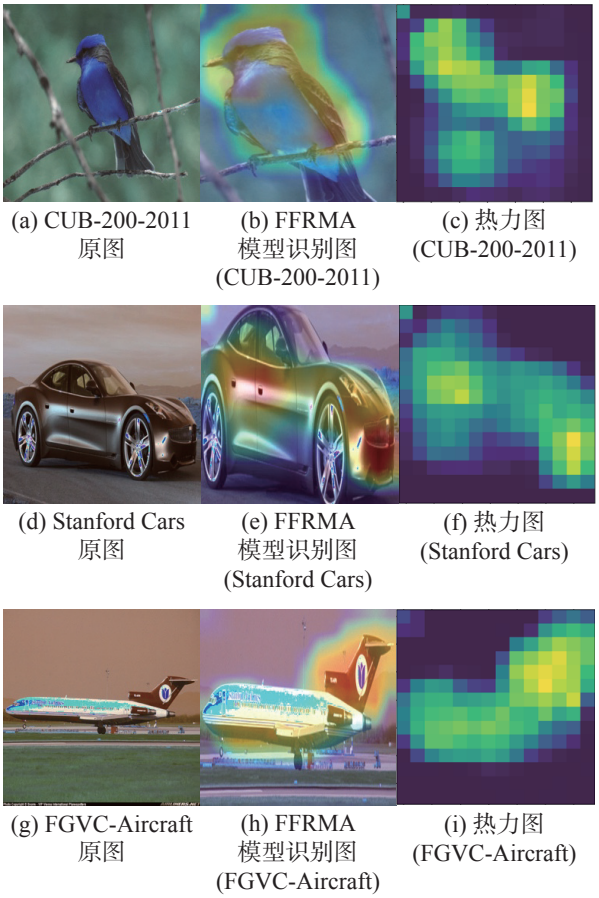
图 8　目标定位可视化效果对比图

Fig. 8　Comparison of object localization visual effect

图 9 给出了 3 个数据集经过 FFRMA 模型处理后的可视化图。

对比图 9(a)和图 9(b), 图 9(d)和图 9(e), 图 9(g)和图 9(h), 图 9(b)、(e)、(h)实现了目标增强, 消除了背景噪音干扰的同时关注到更多的

微小且重要的局部特征信息。



(a) CUB-200-2011 原图　(b) FFRMA 模型识别图 (CUB-200-2011)　(c) 热力图 (CUB-200-2011)

(d) Stanford Cars 原图　(e) FFRMA 模型识别图 (Stanford Cars)　(f) 热力图 (Stanford Cars)

(g) FGVC-Aircraft 原图　(h) FFRMA 模型识别图 (FGVC-Aircraft)　(i) 热力图 (FGVC-Aircraft)

图 9　3 种数据集的可视化图

Fig. 9　Visualization of three datasets

从鸟的可视化图中可以看出，热力图深红色区域主要集中在鸟头、鸟身子和鸟脚等关键局部区域，说明鸟头、鸟身子和鸟脚对最后分类结果贡献最大；汽车的热力图聚焦在车牌、车标等显著特征；从飞机的热力图看出，网络关注的特征是方向舵，飞机图标和机翼。通过以上的可视化效果，FFRMA 模型可以有效地定位前景目标和学习丰富的局部细微特征。

4　结束语

本文提出了将前景特征增强和区域掩码自注意力相结合的细粒度图像分类方法，能够减少背景噪音干扰，提取丰富多样化的局部判别性特征信息。首先，前景特征增强模块可以准确定位前景目标，在消除背景噪音干扰的情况下对前景目标进行特征加强；然后，区域掩码自注意力模块在前景目标增强的前提下，捕获更多丰富的且区别于其他子类的局部细微特征。两个模块的协同合作使得分类精度明显提高；最后，在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 三个数据集的实验结果表明，本文模型均取得了不错的分类精度，分别为 88.0%、95.3% 和 93.6%，性能优于其他模型。在未来工作中，将致力于针对一些局部小区域的特征提取考虑使用跨层特征融合；将目标检测算法用在细粒度图像分类上，提高模型的定位性能；将语义与空间信息联系起来提取更加丰富且多样化的特征信息，进一步提高分类性能。

参考文献：

[1] ZHANG Fan, LI Meng, ZHAI Guisheng, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization[M]. MultiMedia Modeling. Cham: Springer International Publishing, 2021: 136−147.

[2] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset[EB/OL]. (2011−04−12)[2021−09−15]. http://www.vision.caltech.edu/datasets/CUB_200_2011.

[3] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. (2013−06−21)[2021−09−15].https://arxiv.org/abs/1306.5151.

[4] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2014: 834−849.

[5] UIJLINGS J R R. VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154–171.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014 : 580−587.

[7] BRANSON S, VAN HORN G, BELONGIE S, et al. Bird species categorization using pose normalized deep convolutional nets[EB/OL]. (2014−06−11)[2021−09−15]. https://arxiv.org/abs/1406.2952.

[8] 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 Jianhao, WU Jianxin. A survey on fine-grained image categorization using deep convolutional features[J]. Acta automatica sinica, 2017, 43(8): 1306–1318.

[9] 陈立潮, 朝昕, 潘理虎, 等. 基于部件关注 DenseNet 的细粒度车型识别 [J]. 智能系统学报, 2022, 17(2): 402–410.
CHEN Lichao, CHAO Xin, PAN Lihu, et al. Fine-grained vehicle type identification based on part-focused DenseNet[J]. CAAI transactions on intelligent systems, 2022, 17(2): 402–410.

[10] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear

CNN models for fine-grained visual recognition[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 1449−1457.

[11] FU Jianlong, ZHENG Heliang, MEI Tao. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017 : 4476−4484.

[12] ZHENG Heliang, FU Jianlong, MEI Tao, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5219−5227.

[13] ZHOU Bolei, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016 : 2921−2929.

[14] WEI Jun, WANG Qin, LI Zhen, et al. Shallow feature matters for weakly supervised object localization[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021 : 5989−5997.

[15] SOHN J, JEON E, JUNG W, et al. Fine-grained attention for weakly supervised object localization[EB/OL]. (2021−04−11)[2021−09−15].https://arxiv.org/abs/2104.04952.

[16] PAN Xingjia, GAO Yingguo, LIN Zhiwen, et al. Unveiling the potential of structure preserving for weakly supervised object localization[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 11637−11646.

[17] ZHANG Xiaolin, WEI Yunchao, FENG Jiashi, et al. Adversarial complementary learning for weakly supervised object localization[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1325−1334.

[18] WEI Xiushen, LUO Jianhao, WU Jianxin, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE transactions on image processing, 2017, 26(6): 2868–2881.

[19] QIAO Liang, CHEN Ying, CHENG Zhanzhan, et al. MANGO: a mask attention guided one-stage scene text spotter[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(3): 2467–2476.

[20] WANG Jun, YU Xiaohan, GAO Yongsheng. Mask guided attention for fine-grained patchy image classification[C]//2021 IEEE International Conference on Image Processing. Anchorage: IEEE, 2021 : 1044−1048.

[21] SUN GUOLEI, CHOLAKKAL H, KHAN S, et al. Fine-grained recognition: accounting for subtle differences between similar classes[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(07): 12047–12054.

[22] LI J L, DAI H, SHAO L, et al. Anchor-free 3D Single Stage Detector with Mask-Guided Attention for Point Cloud[M]. New York, NY, USA: Association for Computing Machinery, 2021: 553−562.

[23] XIE Jin, PANG Yanwei, KHAN M H, et al. Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection[J]. IEEE transactions on image processing:a publication of the IEEE signal processing society, 2021, 30: 3872–3884.

[24] CHOE J, LEE S, SHIM H. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2214−2223.

[25] KRAUSE J, STARK M, JIA Deng, et al. 3D object representations for fine-grained categorization[C]//2013 IEEE International Conference on Computer Vision Workshops. Sydney: IEEE, 2013: 554−561.

[26] TAN Min, WANG Guijun, ZHOU Jian, et al. Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask[J]. IEEE access, 2019, 7: 117944–117953.

[27] WEI Xiu Sen, WU Jian Xin, CUI Quan. Deep Learning for Fine-Grained Image Analysis: A Survey[EB/OL]. (2019−07−06)[2021−09−15].https://arxiv.org/abs/1907.03069.

[28] LUO Wei, YANG Xitong, MO Xianjie, et al. Cross-X learning for fine-grained visual categorization[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 8241−8250.

[29] CHEN Yue, BAI Yalong, ZHANG Wei, et al. Destruction and construction learning for fine-grained image recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5152−5161.

[30] HANSELMANN H, NEY H. ELoPE: fine-grained visual classification with efficient localization, pooling and embedding[C]//2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020: 1236−1245.

[31] LI Ming, LEI Lin, SUN Hao, et al. Fine-grained visual classification *via* multilayer bilinear pooling with object localization[J]. The visual computer, 2022, 38(3): 811–820.

[32] YANG Ze, LUO Tiange, WANG Dong, et al. Learning to Navigate for Fine-Grained Classification[C]//European Conference on Computer Vision. Cham: Springer, 2018: 438−454.

[33] SHI Xiruo, XU Liutong, WANG Pengfei, et al. Beyond

the attention: distinguish the discriminative and confusable features for fine-grained image classification[C]// MM '20: Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 601−609.

作者简介：

刘万军，教授，博士生导师，主要研究方向为图像与智能信息处理。主持国家自然科学基金面上项目等各类科研项目 20 余项。发表学术论文 200 余篇。

赵思琪，硕士研究生，主要研究方向为图像与智能信息处理。

曲海成，副教授，博士，主要研究方向为图像与智能信息处理。主持省自然科学基金 1 项、省教育厅面上项目 2 项。发表学术论文 60 余篇。

2022 年杭州全球人工智能技术大会
The global artificial intelligence technology conference in Hangzhou (2022)

大会主题：交叉、融合、相生、共赢

2022 杭州全球人工智能技术大会（GAITC 2022）于 11 月 25—27 日在杭州未来科技城举办。作为我国智能科技界的标杆会议和国际化的交流平台，本届大会坚持“综合性、专业性、引领性”的办会宗旨，打通政企学研的跨界联动和跨域联合，推动智能科技的协同创新和融合发展。

本次大会力邀中外院士、技术精英、商业先锋等数百位知名专家同场交流，以多元化的内容和立体化的形式，全面呈现智能科技的最新热点和未来焦点，展现我国人工智能自主创新活力，赋能智能产业生态建设和人才教育培养，助力杭州打造国家人工智能创新应用先导区。

全球人工智能技术大会连续三届落地杭州，持续为当地带来了丰富资源和广泛关注。今年，恰逢《建设杭州国家人工智能创新应用先导区行动计划（2022—2024 年）》发布，产学研汇聚一堂，将为杭州“行动计划”的高效实施献良策、出实招，为我国人工智能发展绘路径、谋大势。

暮秋新冬，西子湖畔，期待与您携手向着数智未来扬帆远航。

主办单位：中国人工智能学会、杭州市人民政府

承办单位：杭州市委人才办、杭州市科技局、杭州市余杭区人民政府

执行单位：浙江杭州未来科技城（海创园）管委会