多尺度特征融合的细粒度图像分类

李思瑶，刘宇红，张荣芬*
贵州大学大数据与信息工程学院，贵州 贵阳 550002

摘要 提出了一种基于多尺度特征融合的细粒度图像分类方法。通过运用特征金字塔结构对不同层次的特征进行尺度变换，再进行信息融合；之后筛选其中包含细节特征最多的前三个区域图，将其与图像的全局特征共同作用以判断图片所属的子类类别。在公开的细粒度数据集 CUB-200-2011、Stanford Dogs 上进行了实验，得到的分类精度分别为 85.7%和 83.5%。实验结果表明该方法对于精细化物体分类具有一定的优越性。

关键词 图像处理；细粒度图像分类；多尺度特征；特征金字塔；卷积神经网络
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.121002

Fine-Grained Image Classification Based on Multi-Scale Feature Fusion

Li Siyao, Liu Yuhong, Zhang Rongfen*
College of Big Data and Information Engineering, Guizhou University, Guiyang, Guizhou 550002, *China*

Abstract A fine-grained image classification method based on multiscale feature fusion is proposed. By using feature pyramid structure, the scales of different levels of features are transformed, and the information fusion is then carried out. After that, the first three regions with the most detailed features are screened out, combining with the global feature of the image to determine the subclass category of the image. The experiments are conducted on the open fine-grained data sets CUB-200-2011 and Stanford Dogs, and the classification accuracy is 85.7% and 83.5%, respectively. Experimental results show that the method has certain advantages for fine object classification.

Key words image processing; fine-grained image classification; multi-scale feature; feature pyramid; convolutional neural network
OCIS codes 100.2000; 100.3008; 100.4996

1 引 言

深度学习的概念源于人工神经网络的研究，近年来，深度学习在目标检测、表情识别、目标跟踪等诸多领域有了巨大的研究进展，特别是在图像分类方面有许多突破[1]。不过，目前大部分的图像分类工作都集中在通用分类，用以区分图像是否属于特定大类类别[2]。因此，子类别图像的区分，即细粒度图像分类成为这几年计算机视觉领域的研究热点，其研究目标从不同类别转换为同一类别不同子类[3]，例如汽车车型、动物品种以及食物菜肴等。这是一项极具挑战的研究任务，因为一个大类下的不同子类的全局几何构造及其外观可以非常相似。同一子类的物体在颜色、姿态、背景等方面具有一定差异性，这些因素会影响其相关类别的判定。如何识别关键部位的微妙差异至关重要[4]，例如对于不同犬类，需要关注它们的耳朵、尾巴等具有显著性区别的区域。

细粒度图像分类的重点工作在于如何有效地自动定位关键局部区域并整合到分类模型中。此前的一些研究利用了人工手动的注释[5-6]（如鸟类数据集对不同部位的标注，斯坦福犬类数据集对边界框的划定等），虽然取得了不错的成果，但人工标注成本昂贵、有一定的出错率且在实际中难以适用；另外一

收稿日期：2019-09-26；**修回日期**：2019-11-18；**录用日期**：2019-12-11
基金项目：贵州省科技计划（黔科合基础[2019]1099）
* **E-mail**：rfzhang@gzu.edu.cn

些方法[7-10]采取了弱监督机制来定位局部区域，例如文献[7]利用神经网络的深度滤波响应作为局部描述符；文献[9]利用注意力建议网络得到注意力区域信息；文献[10]利用检测的前景对象，选取其中重要局部候选作为聚类簇，将其编码后作为全局信息。这些方法均取消了人工注释的使用，但在一定程度上缺乏相应的机制确保网络关注的是正确区域，对关键区域的敏感性不强导致分类精度较低。在学习了之前各种研究的经验后，考虑将小目标多尺度检测的思路引入细粒度图像分类中，使得子类区分可以关注更加细节的区域，同时还要解决如何在目标定位出的大量区域中选取适合区分该特定子类的关键性区域的问题。

本文提出了一种基于多尺度特征融合的细粒度图像识别方法。基础网络使用已在大规模数据集 ImageNet 上预训练过的深度神经网络 Resnet101，提取输入图像全局特征；低层特征分辨率高，包含很多位置、细节信息，高层特征经过多层卷积操作后具有更强的语义性。利用深度卷积网络生成一个自底向上的具有不同尺度和维度的特征金字塔结构，并通过另一条自顶向下的通道增强特征信息的表达，通过横向连接进行同尺度的特征融合进而构成特征金字塔网络，保证细微的目标区域的高层特征中包含较强的语义信息；随后，通过计算多个显著性区域的负置信度滤除具有干扰性的背景区域，略微调整目标边界框，进一步精确确定坐标；接着，选取三个与图片真实类别约束力最强的关注区域，提取其相应位置上的特征信息与全局特征共同送入网络学习，用以预测子类类别。该方法的创新点是将多尺度特征的小目标检测用于细粒度图像分类领域，同时设计了一个筛选机制选取其中最具判别性的区域，有效提高了分类精度，具有潜在的应用价值。

2 相关理论

2.1 细粒度图像分类

对于细粒度图像分类，由于类间差异性是微妙的，需要更加关注显著性特征的学习以及物体部件级别的定位。当前，细粒度图像分类的研究方向可大致分为强监督学习与弱监督学习。强监督学习在图像识别过程中除了利用图像类别信息，还需要借助标注框以及特定位置的局部坐标等额外人工标注，排除背景噪声的干扰，完成对前景对象及局部区域的检测[11-12]。虽然这种方法在汽车、花卉等种类识别中取得了不错的成绩，但其需要耗费巨大的人力成本且在实际生活中并没有很高的实用性，因此近几年的研究热点都趋向于第二种监督方法。

弱监督学习在训练过程中仅依赖图像的类别标签，随着深度学习的快速发展，近些年来出现了许多先进的方法。最早尝试不使用额外人工标注信息的两级关注模型[13]，采用视觉注意力进行图像分类，图像分类分为物体级别与部件级别单独分类，将得到的分数相加生成最后结果；Lin 等[14]提出双线性网络，利用网络 1 对物体进行定位，利用网络 2 对上一步的检测区域进行特征提取，两个网络相互协调、共同作用。文献[15]提出通过通道关注和空间关注的递归混合关注网络等。本文方法在参考了这些技术特点的基础上，进行了进一步的改进，尽可能充分获取有用特征信息，为子类的区分学习到更加显著的区别。

2.2 目标检测

目标检测与定位是许多图像处理任务的基础，它为识别精细类别提供了可靠的信息。早期的目标检测方法采用尺度不变特征变换(SIFT)或者方向梯度直方图(HOG)等特征进行提取，最近的检测方法大多基于深度学习网络。具体来说，为了解决小目标检测的问题，之前的研究采用的是图像金字塔方法，即使用简单全卷积网络，将原图按不同比例缩放后送入网络以达到检测不同尺度目标的目的，这种方法虽然取得了不错的效果，但所需的计算时间与量级都是巨大的，不具有实用性；RCNN(region-based convolution neural networks)[16]算法将任一大小图片输入更深层的全卷积网络，在最后得到的特征图上提取信息，针对单尺度特征的检测速度明显变快，但该算法重点关注的是顶层语义信息，容易忽略具有判别性的小目标；文献[17]对上述方法进行了改进，利用卷积层固有的类金字塔结构，即不同层得到的特征图具有不同的尺度，对这些特征图分别进行预测，这种方法的计算量虽然不会增加，但对不同尺度的单独检测使得底部特征表达能力不足，不能充分运用具有细节信息的高分辨率图。

综合前面所述的方法，期望在不增加计算量的同时，能够准确、有效地定位出具有判别性的小目标区域，如图 1 所示。通过在多尺度特征图中定位到关键局部特征并与全局特征共同作用，进一步提高子类分类的精度。受文献[18]的启发，本文设计思路是通过自底向上的卷积、自顶向下的过程以及特征的横向连接生成一个在所有尺度上均具有强语义信息的特征金字塔，使网络可以充分学习到值得关

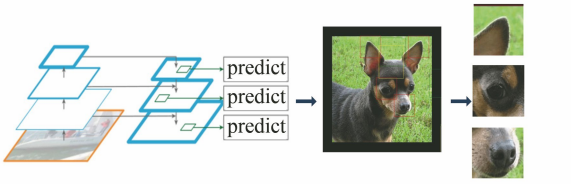

图 1 小目标区域检测效果

Fig. 1 Detection effect for small object regions

注的区域；计算这些区域的负置信度，设定相应阈值滤除掉部分负锚，即一些干扰性的无用区域，并且略微调整剩余目标的边界框；接着，利用本文设定的约束力函数 L_C 进行计算，由小到大选取前三个与图片真实类别最为相近的关注区域，提取相应的特征信息，将其与全局特征共同送入网络学习，用以预测子类类别。

3 本文方法

3.1 基本思想

细粒度图像分类的难点在于精确定位具有判别性的关键区域并且从这些区域中提取有效特征来进行分类。为了实现上面提到的关键思想，本文提出了如图 2 所示的网络架构以找到更多的有用信息，从而进一步提高识别准确率。

首先，将输入图像送入特征提取器以得到图像的全局特征信息。使用没有完全连接层的完全卷积网络作为特征提取器，具体来说，选择预训练的 Resnet101 网络结构作为 CNN 特征提取器。然后，将得到的特征送入多尺度目标检测网络用来进一步寻找显著性小的目标区域，通过设定的边界框筛选规则，即定义一个损失函数 L_C 用来计算小目标区域与其所属子类之间的关联性强弱，得到的结果越小，关联性越强。通过所有结果的排序寻找其中与子类类别的约束力最大，即最接近、最有利于子类真实类别判定的前三个判别区域。接着，自动截取这三个区域的图片信息，调整其大小为 224×224，再次送入特征提取器以提取其特征，得到具有强语义的细节信息。与全局信息进行特征融合，增加了特征信息的维度，使得网络可以学习到多个区域的图像特征表示，特征融合后能加强子类分类的效果。其中，图像经过卷积处理后，被送入完全连接层及损失层，得到的损失函数 L_G 的表达式为

$$L_G=-\ln S(X), \tag{1}$$

式中：X 为输入图像；$S(X)$ 为归一化指数函数。

将全局特征与小目标特征进行融合后，得到的损失函数 L_S 可定义为

$$L_S=-\ln S(X,R_1,R_2,R_3), \tag{2}$$

式中：$R_i(i=1,2,3)$ 为选取的三个小目标区域。

在设定的区域筛选规则中，定义该区域与真实类别的约束力强度的损失函数 L_C 为

$$L_C=-\ln C(R_i)-\ln C(X), \tag{3}$$

式中：$C(X)$ 是一个置信度函数，映射了输入图像与真实类别之间的关联性。首先，需要筛选出具有判别性的小区域，此时(3)式中的第一项是每个小目标区域的交叉熵损失函数，第二项是原始图像的交叉熵的值。将不同区域得到的结果由小到大进行排序，选取前三个更易于判别出真实类别的小目标区域，之后，将这三个区域的交叉熵函数总和代入(3)式中的第一项与原始图像的交叉熵相减，得到最后的 L_C 取值。

最后用于分类判断的总的损失函数是由这三部分相加而成：

$$L=L_G+L_S+L_C。 \tag{4}$$

总的损失函数度量的是预测值与真实值之间的差异。上述三个损失函数使用不同输入值分别计算模型预测属于每个品种的概率损失，通过三项结果共同衡量不同预测类别与真实类别之间的损失大小，选取损失最小值所对应的图像类别作为最后的预测结果。

由于每种子类训练数量较少，选择能在相关方向上加速或减速的带动量的 SGD 优化器，以有效抑制振荡、加快收敛。训练过程中，梯度下降时，每个批次包含 16 个样本，迭代轮数设为 500 轮。

3.2 多尺度小目标检测

在尺度空间中，不同尺度的图像模糊程度与物体由近到远的成像过程类似。在不确定待检测目标尺寸的前提下，考虑利用图片的不同特征尺度检测相同的感兴趣区域以进行匹配。总的来说，多尺度检测在提取全局整体信息的同时又可提取到局部详细信息，这样可以得到更全面的信息。

实验所用的多尺度小目标检测方法如图 3 所示。参考特征金字塔的设计，本文使用了由卷积网络前向传播构成的一条自底向上的通道与采用上采样构成的一条自顶向下的通道，通过横向连接将相同尺寸的特征图进行融合，使得高分辨率的底层特征与低分辨率的顶层特征能够同时被使用。对融合结果使用 3×3 的卷积核进行卷积，消除上采样的混叠效应。由于不同尺度的特征图所包含的信息重点是不同的，低层特征包含更多的细节信息，可以关注到更多关键性目标，顶层特征具有更强的语义信息，

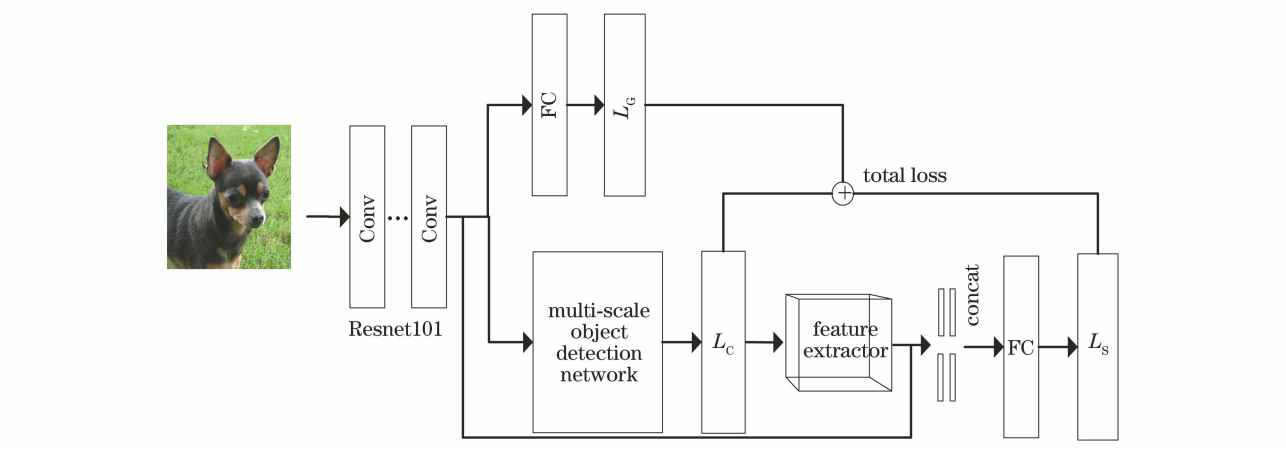

图 2 网络架构设计

Fig. 2 Design of network architecture

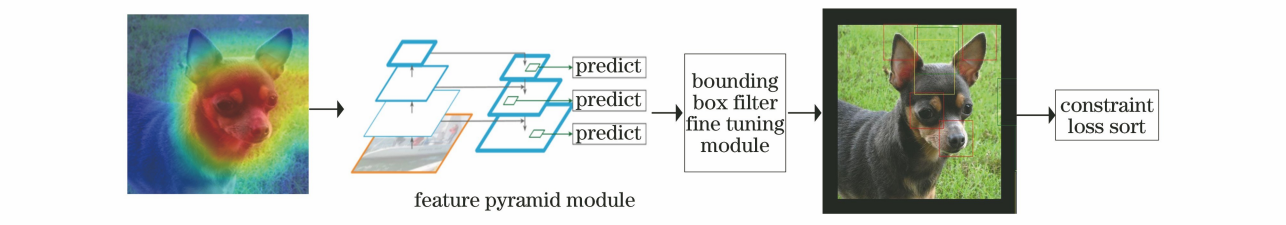

图 3 多尺度小目标检测过程

Fig. 3 Process of multi-scale small object detection

对整体的目标区域更为关注。因此设计在不同尺度的层中定义不同大小的边界框，且边界框大小由底向上不断递增。对于一张大小为 448×448 的输入图像，设置划定区域的大小分别为 48×48，96×96，128×128，另外每个区域大小都有三个长宽对比度(1∶1，1∶2，2∶1)，这样整个特征金字塔会产生多个不同尺度的矩形框。对每个框定区域进行负打分，若其得分大于设定的阈值 $\theta=0.99$，说明这部分信息可能不具备判别性，是无效的背景信息，则将其剔除，对剩余区域的位置进行微调整，过滤和位置精修的操作使得小目标区域的定位更加精确。考虑到定位的区域经常有重叠产生，使用非极大抑制(NMS)算法迭代建议列表，消除部分重复定位的边界框。最后，分别计算剩余区域的约束力损失函数，以找到与真实类别最为相关的前三个目标子区域。将这三个判别性区域截取出来，调整到统一大小，送入预训练网络中提取特征，其与整体图像特征的共同作用有利于精细检测子类类别。

4 实验验证与分析

4.1 数据集

本文所使用的细粒度数据集包括斯坦福大学搜集的 120 种犬类数据 Stanford Dogs 以及加利福尼亚理工学院的 200 类鸟类数据集 CUB-200-2011。在实验中没有使用任何边界框或部件标注，仅依靠图像类别的标签实现弱监督分类。

CUB-200-2011 是一个鸟类数据集，包含 200 种不同鸟类，共计 11788 张图片。由于该数据集每个子类只有 30 张左右的图片用于训练，通常被认为是最具竞争力的数据集之一。在实验中，划分的训练集有 5994 张图片，测试集有 5794 张图片。

Stanford Dogs 是一个犬类数据集，包含 120 种不同犬种，共计 20580 张图片，实验中自行划分的训练集与测试集分别包含 11632 和 8948 张图片。图像中的犬类取自许多角度和场景，同类之间的背景与毛色差异具有一定迷惑性。

4.2 实验环境及参数设置

本实验采用的 GPU 显卡为 GTX 1 080 Ti，CPU 处理器型号为 intel Core I7-7800X，内存为 64 GB。在 Linux 系统下，采用 Pytorch 实现算法设计。在所有实验过程中，输入的预处理图片大小为 448×448，使用预先训练过的全卷积网络 Resnet101 作为特征提取器，通过批量标准化(Batch Normalization)处理调节梯度。初始学习率为 0.001，

每迭代 100 次学习率调整为之前的 10%。权重衰减量为 0.0001，NMS 阈值设定为 0.3。该实验可以完全进行端到端的训练，不需要预先训练好的检测模型，进行细粒度图像分类的平均计算时间为 120.3 ms。

4.3 实验结果可视化

图像处理的可视化过程如图 4 所示。其中，第一列为从两个数据集中随机选取的 4 幅原始图像；第二列为原始图像经过预处理后被送入特征提取器所得到的最后一层卷积的特征图映射，即图像的全局特征；第三列显示了将上述特征信息用于构造特征金字塔并利用矩形框表示的具有多尺度的小目标区域；对这些区域进行筛选和微调，选择其中最接近真实类别的三个判别区域，提取其中相应的特征信息，如最后三列所示，即图像关键部位的局部特征。将得到的这些特征进行融合，增大了特征维度的同时丰富了特征信息，使得网络学习到的分类信息更加详细。

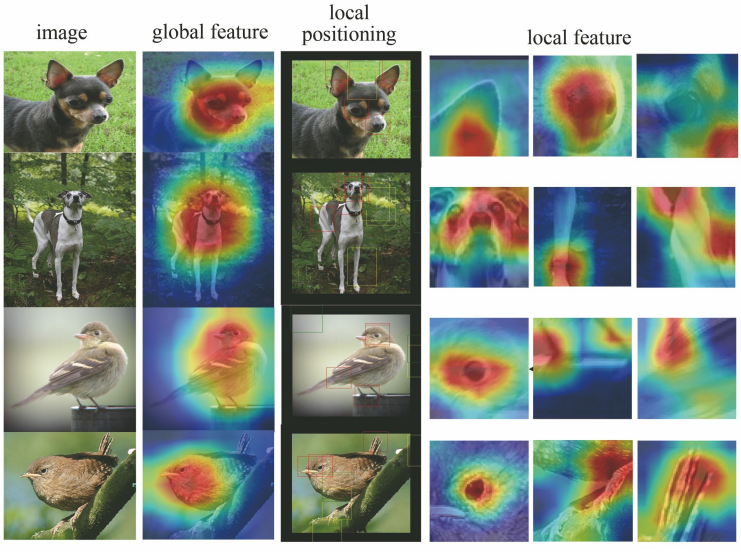

图 4 图像处理可视化

Fig. 4 Visualization of image processing

4.4 实验验证与分析

1）网络模型对比

为了分析模型中不同组件的影响，设计在两个数据集中不同模块组件的运行，分别使用 Resnet101 基础网络、Resnet101 的特征金字塔网络、加入边界框过滤微调的 Resnet101 特征金字塔网络进行对比实验，实验结果如表 1 所示。Resnet101 的基础网络使用了传统图像分类的方法，通过最后得到的全局特征进行分类。这种方法中输入图片的大小是固定的，得到的特征图尺寸也是单一不变的，并且分类精度与数据集的数量、网络的复杂度有着密切的关系。由表 1 中结果可知，只使用基础网络对于细粒度图像的分类效果提升并不明显，需要进一步探索具有判别性特征的小目标区域；Resnet101＋特征金字塔的网络直接将得到的多个建议区域进行排序选择，其分类精度与加入边界框过滤微调的 Resnet101 特征金字塔网络相比在鸟类和犬类数据集中得到的最佳结果分别减少了 2.3 和 2.2 个百分点。推测得到这样的结果是由于没有对多个建议区域进行后续的筛选和调整，许多随机区域干扰了分类。

表 1 不同组件的分类精度

Table 1 Classification accuracy of different components

%

Dataset	Resnet101	Resnet101＋FPN(feature pyramid networks)	Resnet101＋FPN＋object filter
CUB-200-2011	81.4	83.4	85.7
Stanford Dogs	77.5	81.3	83.5

进一步分析判别性较小区域的个数对细粒度图像分类结果的影响，实验分别使用 2，3，4 个判断区域进行网络训练，分类结果如表 2 所示。由表中结果可看到，仅选择 2 个小目标区域对于细粒度的图像分类精度影响较大，在鸟类和犬类数据集上的分类结果与选择 3 个区域相比分别减少了 2.1 和 2.6 个百分点。选择 4 个显著小区域得到的分类精度略高于使用 3 个显著小区域时的结果，但是从计算效率上来看，使用 4 个显著小区域进行品种分类的平均计算时间为160.7 ms，远大于使用 3 个显著小区域时的计算时间(120.3 ms)。因此，从分类精度与

表 2 不同小目标区域数目下的分类精度

Table 2 Classification accuracy for different numbers of small target areas %

Dataset	Two areas	Three areas	Four areas
CUB-200-2011	83.6	85.7	86.0
Stanford Dogs	80.9	83.5	84.0

计算时间综合考虑，使用 3 个显著小区域进行分类更为合理。

2）各类算法结果对比

进一步在两个数据集上对多尺度特征融合算法和其他文献算法的分类精度进行对比，对比结果如表 3 和表 4 所示。从表 3 可以看出，在 Stanford Dogs 数据集上，本文算法的分类精度均高于 DVAN (diversified visual attention networks)与多尺度金字塔匹配算法，但与 RACNN (recurrent attention convolutional neural network)算法相比还有一定的差距。但是 RACNN 的结构包含三个尺度的子网络，每个子网络又包含了分类与 APN (attention proposal sub-network)两种类型的网络，使得相应区域的特征提取需要经过多步实现，计算量较大，并且 RACNN 一次只能定位一个关键区域，但在真实场景中人类视觉分辨时往往是多个关键点共同作用；分析表 4 结果可知，文献[11]使用了目标边界框、关键点标注等人工注释，其检测精度明显低于其他方法，这是由于人工标注的关键特征点是固定的，而不同网络在学习过程中对判别区域的关注位置可能会发生变化。文献[14]同时使用了两个卷积神经网络提取特征，其输出经过外积相乘，池化后获得结合特征，本文方法对于局部模型不需要使用额外神经网络且定位更加精确，其分类精度比 B-CNN 算法提高 1.6%。文献[9]的训练需要经过多次迭代，本文方法的训练只需一个阶段，且分类精度比文献[9]略高 0.4%。

总的来说，本文提出的方法在不借助任何边界框和部件注释的前提下，得到的分类结果仍优于大多数现有的方案，实现了单阶段的端到端的简易训练。本文所提模型的分类过程的计算时间为 120.3 ms，是一个性能优良、实用性强的基于弱监督细粒度图像分类的网络模型。

5 结 论

提出了一种基于多尺度特征融合的细粒度图像分类的通用方法。通过自底向上的卷积和自顶向下的上采样构成图像的特征金字塔，通过相同尺寸特征的横向连接生成多个带有强语义信息的多尺度关键区域；计算每个区域的负置信度，设定相应阈值滤除部分具有干扰性的无用区域并且再次微调目标的位置回归；接着，通过约束损失函数排序选取三个最接近图片真实类别的关注区域，提取相应的特征信息，将其与全局特征共同送入网络学习，以预测细粒度图像类别。在 CUB-200-2011 和 Stanford Dogs 数据集上进行了实验，得到的分类精度分别达到了 85.7%与 83.5%，本文方法优于大多数现有的方案，且不需要使用任何人工标记的边界框或关键点标注，实现了端到端的单阶段训练，具有较强的鲁棒性和较好的实用性。下一步将深入研究网络架构的细节设计，进一步提高分类精度，以便更好地将细粒度图像分类应用于各个领域。

表 3 Stanford Dogs 数据集下不同算法分类精度对比

Table 3 Comparison of classification accuracy of different algorithms for Stanford Dogs dataset

Reference	Method	Accuracy /%
Ref. [10]	Pyramid matching	80.4
Ref. [8]	DVAN	81.5
Ref. [9]	RACNN	87.3
This papar	Multi-scale feature fusion	83.5

表 4 CUB-200-2011 数据集下不同算法分类精度对比

Table 4 Comparison of classification accuracy of different algorithms for CUB-200-2011 dataset

Reference	Method	Accuracy /%
Ref. [11]	Part-RCNN	81.6
Ref. [14]	B-CNN	84.1
Ref. [7]	PDFR (picking deep filter responses)	84.5
Ref. [9]	RACNN	85.3
This papar	Multi-scale feature fusion	85.7

参 考 文 献

[1] Tang C, Ling Y S, Yang H, et al. Decision-level fusion tracking for infrared and visible spectra based on deep learning [J]. Laser & Optoelectronics Progress, 2019, 56(7): 071502.
唐聪, 凌永顺, 杨华, 等. 基于深度学习的红外与可见光决策级融合跟踪[J]. 激光与光电子学进展, 2019, 56(7): 071502.

[2] Liu F, Lu L X, Huang G W, et al. Landform image classification based on discrete cosine transformation and deep network[J]. Acta Optica Sinica, 2018, 38(6): 0620001.
刘芳, 路丽霞, 黄光伟, 等. 基于离散余弦变换和深度网络的地貌图像分类[J]. 光学学报, 2018, 38

(6): 0620001.

[3] Zhao X. Research of fine-grained image recognition and classification algorithm based on deep convolutional neural network [D]. Hefei: Anhui University, 2018.
赵星. 基于深度卷积神经网络的细粒度图像识别与分类算法研究[D]. 合肥:安徽大学, 2018.

[4] Zhao Z Y, Cheng Y L, Shi X S, et al. Terrain classification of LiDAR point cloud based on multi-scale features and PointNet [J]. Laser & Optoelectronics Progress, 2019, 56(5): 052804.
赵中阳, 程英蕾, 释小松, 等. 基于多尺度特征和PointNet的LiDAR点云地物分类方法[J]. 激光与光电子学进展, 2019, 56(5): 052804.

[5] Xie L X, Tian Q, Hong R C, et al. Hierarchical part matching for fine-grained visual categorization[C]// 2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, Australia. IEEE, 2013: 1641-1648.

[6] Chai Y N, Lempitsky V, Zisserman A. Symbiotic segmentation and part localization for fine-grained categorization [C] // 2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, Australia. IEEE, 2013: 321-328.

[7] Zhang X P, Xiong H K, Zhou W G, et al. Picking deep filter responses for fine-grained image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 1134-1142.

[8] Zhao B, Wu X, Feng J S, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[9] Fu J L, Zheng H L, Mei T. 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 (CVPR), July 21-26, 2017, Honolulu, HI. IEEE, 2017: 440-455.

[10] Yuan Q D. The Research on fine-grained object classification algorithm based on depth learning[D]. Changsha: Changsha University of Science and Technology, 2017.
袁千丁. 基于多尺度卷积特征匹配的细粒度分类研究[D]. 长沙: 长沙理工大学, 2017.

[11] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection [M] // Computer Vision - ECCV 2014. Cham: Springer International Publishing, 2014: 834-849.

[12] Zhang N, Paluri M, Ranzato M, et al. PANDA: pose aligned networks for deep attribute modeling[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. IEEE, 2014: 1637-1644.

[13] Xiao T J, Xu Y C, Yang K Y, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. IEEE, 2015: 842-850.

[14] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015. Santiago, Chile. IEEE, 2015: 1449-1457.

[15] Gui J S, Ma C F, Bao X A, et al. Fine-grained image classification method for recurrent deep hybrid attention network[J]. Computer Engineering, 2019, 45(5): 205-209.
桂江生, 麻陈飞, 包晓安, 等. 递归深度混合关注网络的细粒度图像分类方法[J]. 计算机工程, 2019, 45(5): 205-209.

[16] 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, June 23-28, 2014, Columbus, OH, USA. IEEE, 2014: 580-587.

[17] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector [M] // Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[18] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. IEEE, 2017: 936-944.