第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

面向细粒度图像识别的通道注意力多分支网络

王彬州，肖志勇*

江南大学人工智能与计算机学院，江苏 无锡 214122

摘要 细粒度图像识别研究的内容是大类下的子类别识别问题，其关键是找到图像中的关键区域并从中提取有效特征。针对现有方法在定位关键区域时无法兼顾准确性和计算量的问题，提出了一种引入高效通道注意力模块的多分支网络。首先，在递归注意力卷积神经网络的基础上引入通道注意力定位图像中目标的位置。然后，用深度超参数化卷积替换传统卷积操作，增加了网络可学习的参数。最后，用改进的注意力部件模块切割出多个图像关键区域部件，以捕捉丰富的局部信息。实验结果表明，本方法在弱监督情况下的识别效果较好，在两个常用细粒度数据集 Stanford Cars、Food-101 上的识别准确率分别为 95.4%和 90.6%。

关键词 图像处理；细粒度图像识别；通道注意力；深度超参数化卷积；卷积神经网络

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2210008

Channel Attention Multi-Branch Network for Fine-Grained Image Recognition

Wang Binzhou, Xiao Zhiyong*

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract The content of fine-grained image recognition research is the problem of sub-category recognition under broad categories. The key is to find the key regions in the image and extract effective features from them. Aiming at the problem that the existing methods cannot balance the accuracy and the amount of calculation when locating key areas, a multi-branch network that introduces an efficient channel attention module is proposed in this paper. First, the channel attention is introduced on the basis of the recurrent attention convolutional neural network to locate the target position in the image. Then, the traditional convolution operation is replaced with depthwise over-parameterized convolution, which increases the parameters that the network can learn. Finally, the advanced attention part module is used to cut out multiple image key area components to capture rich local information. Experimental results show that the method has a better recognition effect in weakly supervised situations, and the recognition accuracy rates on the two commonly used fine-grained datasets Stanford Cars and Food-101 are 95.4% and 90.6%, respectively.

Key words image processing; fine-grained image recognition; channel attention; depthwise over-parameterized convolution; convolutional neural network

OCIS codes 100.4996 100.5010; 100.3008

1 引 言

近年来对象识别方法在通用图像分类中取得了稳步进展，对特定大类别的图像分类技术已经比较成熟[1]，但在大类别物体识别中进行精细的子类别识别，即对细粒度图像的识别仍存在一系列问题。细粒度图像不同子类别之间的相似性和同一子类别中的差异性，使细粒度图像识别比普通图像识别具

收稿日期：2021-01-04；**修回日期**：2021-01-12；**录用日期**：2021-01-27

基金项目：江苏省优秀青年基金(BK20190079)

通信作者：*zhiyong.xiao@jiangnan.edu.cn

有更大的挑战性。

细粒度图像识别的关键是提取有区分度的视觉特征[2-4]。相关研究表明，细粒度视觉分类任务的关键在于使用有效的方法准确识别图像中的关键信息区域[5-8]。Wei 等[2]通过人工标注的边框信息实现目标检测和物体识别，Zhang 等[5]用人工标注的语义信息，联合语义检测与语义提取实现图像识别。但获得密集的人工标注信息需要大量人工劳动，限制了该方法实际应用的可扩展性和使用场景。因此，在仅有类别信息的情况下用弱监督定位关键区域并提取关键区域的特征变得尤为重要。Fu 等[6]提出的递归注意力卷积神经网络(RA-CNN)通过递归方法优化判别区域，仅能关注到单一的局部信息。Zhang 等[7]提出的多分支多尺度学习网络(MMAL-Net)不需要训练就能定位图像的关键区域，提高了识别速度，但识别效果不够精准。Zoph 等[9]提出的神经结构搜索网络通过叠加复杂的卷积核组合学习选择最优的网络结构，从而实现比较精确的识别，但需要耗费大量计算资源。部分研究通过双线性网络进行优化，以提取传统单线性网络未能提取的特征[10-11]，但只对网络结构进行优化，识别效果还存在一定的提高空间。生成对抗网络利用生成器和判别器进行对抗训练，也能得到较好的识别效果[12-13]。基于弱监督定位关键区域和提取关键区域特征的方法中，Yang 等[8]提出的 Navigator 模块通过类似目标检测的手段进行子网络学习，并加入了注意力机制，可以很好地定位关键区域。Zhang 等[7,14]在网络分支中不进行子网络学习，在运算量较小的情况下就能定位部件区域。

本文在 RA-CNN 的基础上，提出了一种引入高效通道注意力(ECA)模块[15]的多分支网络，以实现端到端的弱监督图像分类。首先，引入带有 ECA 模块的子网络，通过 ECA 子网络的通道注意力机制学习获取目标物体更准确的坐标信息。然后，用深度过参数化卷积代替传统卷积，在通道中使用多样的二维(2D)卷积核，以提高卷积性能和网络的学习能力。最后，引入不需要进行训练的改进注意力部件模块(AAPM)，用改进的滑动窗口代表激活值和距离交并比(DIoU)损失函数[16]，在不增加网络复杂度的情况下得到多个局部信息更准确的图像部件。

2　递归注意力卷积神经网络的原理

RA-CNN 的思想是以递归方式堆叠更大规模的网络，在图像识别中从粗糙到精细(如从身体到头部再到鸟喙)逐渐进入图像中区分度最高的单一区域进行图像识别。该模型可分为三个子网络，每个子网络的结构相同，但参数不同，如图 1 所示。每个子网络中都包含分类子网络和注意力建议子网络(APN)两个不同类型的网络。RA-CNN 的工作流程：首先，将原始尺寸图像送入第一个子网络中，直接通过包含全连接(FC)层的卷积神经网络(CNN)进行分类训练；然后，APN 基于分类网络提取的特征进行训练，并用带有注意力机制的网络训练得到

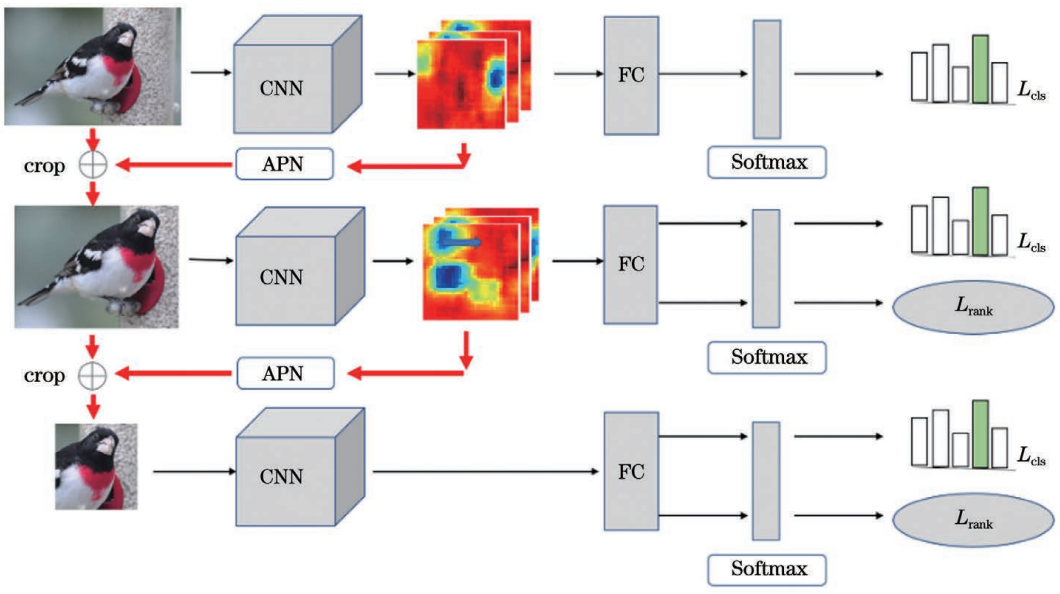

图 1　RA-CNN 的结构

Fig. 1　Structure of the RA-CNN

尺寸缩小的优化区域；最后，对图像进行裁剪，将尺寸更小的图像作为第二个子网络的输入。重复进行三次迭代后，将三个子网络的输出进行融合，得到最终的分类结果。

在每个子网络的训练过程中，首先将输入图像 $\boldsymbol{X}$ 送入预训练好的卷积层中提取基于区域的深度特征 $\boldsymbol{W}_c * \boldsymbol{X}$，其中，$*$ 运算符包括卷积、池化和激活操作，$\boldsymbol{W}_c$ 为预训练卷积层的总体参数。子网络输出细粒度类别的概率分布 p 可表示为

$$p(\boldsymbol{X}) = f(\boldsymbol{W}_c * \boldsymbol{X}), \tag{1}$$

式中，函数 f 可通过 Softmax 函数将特征向量转化为概率。APN 的本质是最后一层为 FC 层的 CNN，可将参与区域近似为正方形，可用 3 个参数表示为

$$[t_x, t_y, t_l] = g(\boldsymbol{W}_c * \boldsymbol{X}), \tag{2}$$

式中，t_x、t_y 分别为当前参与区域 x、y 轴的中心坐标，t_l 为边长的 1/2，函数 g 表示最后一层具有三个输出的 FC 网络。APN 的学习以弱监督方式进行训练，由于裁剪前后图像的尺寸不同，为了确保 APN 在训练中的反向传播迭代优化，用 Boxcar2D 函数作为 Attention Mask 近似裁剪操作，Attention Mask 在前向传播中可以选择最重要的区域。连续函数的特性可使整个网络在反向传播中对自身进行优化。APN 输出框的左上角坐标 $(t_x^{\mathrm{tl}}, t_y^{\mathrm{tl}})$ 和右下角坐标 $(t_x^{\mathrm{br}}, t_y^{\mathrm{br}})$ 可表示为

$$t_x^{\mathrm{tl}} = t_x - t_l,\ t_y^{\mathrm{tl}} = t_y - t_l,\ t_x^{\mathrm{br}} = t_x + t_l,\ t_y^{\mathrm{br}} = t_y + t_l, \tag{3}$$

连续的 Attention Mask 函数可表示为

$$M(\cdot) = [h(x - t_x^{\mathrm{tl}}) - h(x - t_x^{\mathrm{br}})] \cdot [h(y - t_y^{\mathrm{tl}}) - h(y - t_y^{\mathrm{br}})], \tag{4}$$

其中，函数 h 为 Sigmoid 函数，可表示为

$$h(x) = \frac{1}{1 + \exp(-kx)}, \tag{5}$$

式中，k 为常系数，表示逻辑增长率。当 k 足够大时，Sigmoid 函数可认为是阶跃函数，其特性与 Boxcar2D 函数相同，从而很好地进行近似裁剪操作。裁剪图像操作在 RA-CNN 中可以实现反向传播，RA-CNN 通过分类网络和 APN 组成的子网络进行三次递归，学习图像不同尺寸的特征，最终仅用标签信息实现对图像的端到端分类训练。

3　本方法的原理

首先，介绍了通道注意力多分支网络的整体结构和流程。然后，分别介绍了 ECA 模块的结构以及用深度过参数化卷积层代替普通卷积层的细节。最后，分析了 AAPM 的结构。

3.1　联合通道注意力的多分支网络结构

联合通道注意力机制的网络是一个多分支网络，主干部分包括整体分支、目标分支和部件分支，网络结构如图 2 所示。原始图像在整体分支中通过

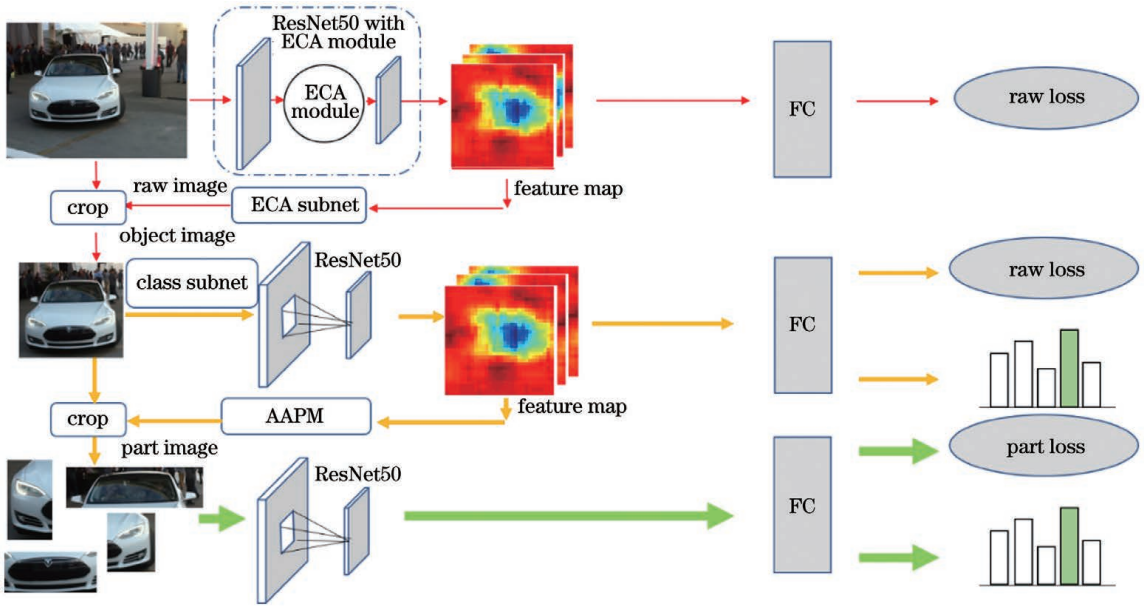

图 2　本网络的结构

Fig. 2　Structure of our network

CNN 训练后进入 FC 层，以捕获图像的整体特征。目标分支包括 ECA 子网络和分类子网络，ECA 子网络与整体分支共享卷积层，通过添加 ECA 模块的 CNN 训练得到对象的边界框信息，从而利用边界框信息对图像进行裁剪。分类子网络将 ECA 子网络裁剪的图像送入带有 FC 层的 CNN 进行分类训练，以提取图像中的目标特征。在部件分支中，先通过 AAPM 处理分类子网络中裁剪的图像，得到多个区分度最高、冗余度最小的部分区域，进而得到多个局部信息更准确的图像部件。AAPM 不需要进行训练，可在不增加网络学习参数的情况下进行图像切块。切块完成后部件分支将部件图像送入 CNN 中进行训练，得到丰富的局部特征。三个分支使用的网络都以残差网络（ResNet50）为基础进行训练。同时将 ResNet50 中的卷积操作全部替换为深度过参数化卷积操作。在训练阶段，用所有分支共同训练本网络。在测试阶段时，仅使用整体分支和目标分支联合得到图像的分类结果。

3.2　高效通道注意力子网络

人类具有快速关注图像中各个部分，而不是处理整个场景的能力，这种选择性注意力机制也被称为视觉注意力预测或视觉显著性检测，被广泛应用于计算机视觉和神经科学领域中。对物体进行定位时，常用通道注意力机制定位模块（CALM）[17] 和 SENet（Squeeze-and-excitation network）[18] 等模块提取图像特征，进而用于细粒度图像中物体位置信息的提取。ECA 模块是 SENet 中通道注意力模块的改进方法。ECA 模块通过避免降维操作和适当的跨通道交互保持网络的性能，同时可以显著降低模型的复杂性，其结构如图 3 所示。相比经典的 SENet 模块，ECA 模块可在不降低维度的情况下有效捕捉通道与通道间的关系；且 ECA 模块作为一种无需降维的局部跨通道交互策略，能在不明显增加计算量的情况下具有更优的性能。

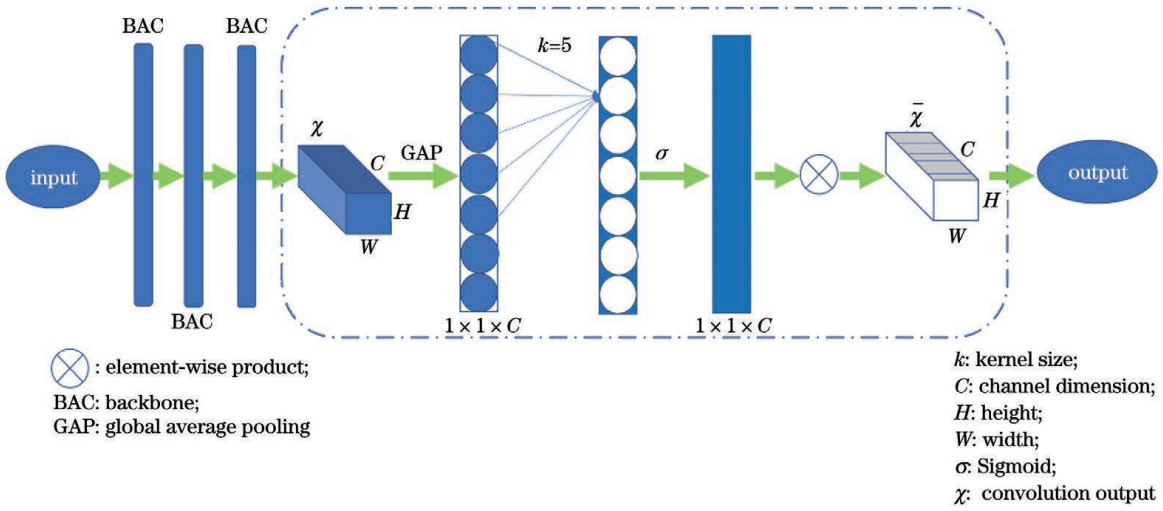

图 3　ECA 模块的结构

Fig. 3　Structure of the ECA module

ECA 模块的注意力可表示为

$$f_{(\boldsymbol{W}_1,\boldsymbol{W}_2)}(\boldsymbol{y})=\boldsymbol{W}_2R(\boldsymbol{W}_1\boldsymbol{y}),\tag{6}$$

式中，$\boldsymbol{W}_k$ 为学习到的通道注意力，$\boldsymbol{y}$ 为逐通道进行的全局平均池化结果，R 为线性修正单元（ReLU）激活函数。ECA 模块中，矩阵 $\boldsymbol{W}_k$ 可表示为[7]

$$\boldsymbol{W}_k=\begin{bmatrix} w^{(1,1)} & \cdots & w^{(1,k)} & 0 & 0 & \cdots & \cdots & 0\\ 0 & w^{(2,2)} & \vdots & w^{(2,k+1)} & 0 & \vdots & \vdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & 0 & 0 & \cdots & w^{(C,C-k+1)} & \cdots & w^{(C,C)} \end{bmatrix},\tag{7}$$

式中，$\boldsymbol{W}_k$ 包含 $k\times C$ 个参数，k 为通道序号，C 为通道维度。任意 y_i 的权重仅考虑 y_i 与其 k 个邻近元素之间的相互作用，可表示为

$$w_i=\sigma\left[\sum_{j=1}^{k}w_i^{(j)}y_i^{(j)}\right],y_i^{(j)}\in\boldsymbol{\Omega}_i^{(k)},\tag{8}$$

式中，$\boldsymbol{\Omega}_i^{(k)}$ 为 y_i 的 k 个相邻通道的集合，σ 为

Sigmoid 函数。为了使所有通道共享训练参数，简化得到

$$w_i=\sigma\left[\sum_{j=1}^{k}w^{(j)}y_i^{(j)}\right],y_i^{(j)}\in\boldsymbol{\Omega}_i^{(k)}。\tag{9}$$

通过卷积核大小为 k 的一维卷积实现通道间的信息交换，在 k 和通道维度 C 之间可能存在函数映射 ϕ

$$C=\phi(k)。\tag{10}$$

由于映射中通道维度 C 通常为 2 的幂次方，为了避免映射过于简单，令

$$C=\phi(k)=2\cdot\exp(\gamma\times k-b),\tag{11}$$

式中，参数 γ 和 b 分别设置为 2 和 0.8。在给定通道维度 C 的情况下，卷积核大小 k 可表示为

$$k=\psi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{\text{odd}},\tag{12}$$

式中，$|t|_{\text{odd}}$ 为最接近 t 的奇数，ψ 函数为 ϕ 函数的反函数。

ECA 子网络中与整体分支共享卷积过程后的输入向量为整体分支网络最后一个卷积层的输出向量。通过 FC 层可预测目标物体的长宽及其左下角起始的横纵坐标 l、d 和 t_x、t_y。本方法使用 RA-CNN 中的类 Boxcar2D 连续函数作为 Attention Mask 近似裁剪，以确保反向传播的顺利进行。

3.3　深度过参数化卷积层

卷积层是 CNN 最重要的组成部分之一，但实际中一般不使用仅增加 FC 层的方法提高网络性能，原因是该方法可能造成过参数化问题。最新研究结果表明，过参数化能起到加速神经网络训练并提高聚合模型性能的作用，因此，Cao 等[19]利用超参数化特性构建一个深度卷积操作，并将引入这种操作的卷积层称为深度过参数化卷积层（DO-Conv）。本方法采用的深度过参数化卷积层用一个附加的深度方向卷积增加卷积层，其中，每个输入通道用不同的 2D 卷积核进行卷积。两个卷积组合形成一次超参数化，从而增加可学习的参数，且得到的线性运算可用一个卷积层表示。通过增加一个额外过参数化的分量对卷积层进行过参数化运算，其本质是对每个输入通道分别进行卷积运算。即用深度过参数化卷积层代替分类网络中的卷积层，生成的线性运算可由单个卷积层表示，增大了网络可学习的参数，提高了网络的性能。

DO-Conv 可看作一个深度卷积和常规卷积的组合，用 D_{mul} 表示卷积核的深度，M 和 N 表示深度卷积每个通道特征的大小且 $D_{\text{mul}}\geqslant M\times N$，$C_{\text{in}}$ 和 C_{out} 为卷积操作前后的通道数目。其中，深度卷积中可训练的卷积核 $\boldsymbol{D}\in\boldsymbol{R}^{(M\times N)\times D_{\text{mul}}\times C_{\text{in}}}$，而常规卷积的卷积核 $\boldsymbol{W}\in\boldsymbol{R}^{C_{\text{out}}\times D_{\text{mul}}\times C_{\text{in}}}$。给定一个输入的块（Patch）$\boldsymbol{P}\in\boldsymbol{R}^{(M\times N)\times C_{\text{in}}}$，则 DO-Conv 的输出维度与卷积层相同，可表示为

$$\boldsymbol{O}=(\boldsymbol{D},\boldsymbol{W})\circledast\boldsymbol{P},\tag{13}$$

深度过参数化卷积操作可表示为

$$\boldsymbol{O}=(\boldsymbol{D},\boldsymbol{W})\circledast\boldsymbol{P}=\boldsymbol{W}*(\boldsymbol{D}\circ\boldsymbol{P})=(\boldsymbol{D}^{\mathrm{T}}\circ\boldsymbol{W})*\boldsymbol{P},\tag{14}$$

式中，$\circ$ 为深度卷积运算符号，$*$ 为普通卷积，$\circledast$ 为 DO-Conv 卷积操作，$\boldsymbol{D}^{\mathrm{T}}$ 为 $\boldsymbol{D}$ 在前两个坐标轴上的转置。实验选择 Kernel composition 进行训练，以提高训练效率。Kernel composition 可表示为

$$\begin{cases}\boldsymbol{W}'=\boldsymbol{D}^{\mathrm{T}}\circ\boldsymbol{W}\\ \boldsymbol{O}=\boldsymbol{W}'*\boldsymbol{P}\end{cases},\tag{15}$$

式中，H 和 W 为特征图的高度和宽度，$\boldsymbol{W}'$在整个特征图中仅计算一次。

3.4　AAPM

AAPM 是 RA-CNN 中 APN 和 MMAL 中 APM 模块的改进，可在不进行网络训练，不添加学习参数的情况下得到多个目标部件的局部信息。通过激活图中激活值的数值特性设定各窗口的代表激活值，用 DIoU 和非极大值抑制（DIoU-NMS）方法选择无重叠区域的固定数量窗口。图像通过 CNN 生成激活图时，参考全卷积网络中滑动窗口的思想，AAPM 确定多个含语义信息的窗口作为部件图像。具体实现过程中，首先通过神经网络得到图像的激活值 $\boldsymbol{A}_w$，并用平方平均数作为每个窗口的代表激活值。由图像的特征图分布特点可知，激活图中激活值较高的区域通常是语义信息集中的区域，而激活值较低的区域语义信息较少。在二维及更高维情况下，平方平均数对于较大激活值比算术平均数更敏感，即对于语义信息集中的区域更敏感。各窗口的代表激活值可表示为

$$\overline{\boldsymbol{A}_w}=\sqrt{\frac{\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}[\boldsymbol{A}_w(x,y)]^2}{H_wW_w}},\tag{16}$$

式中，H_w、W_w 为窗口要素图的高度和宽度。按照各窗口代表激活值$\overline{\boldsymbol{A}_w}$ 的大小对各窗口进行排序，由于各窗口之间会发生重叠，采用 DIoU-NMS 选择固定数量的窗口作为不同大小的部件分块图像，最终得到图像中的目标部件部分。DIoU-NMS 的主要思想：当前框的中心点越靠近当前最大得分框的

中心点，则当前框更有可能是冗余框。在 IoU 大小相同的情况下，不同的中心点位置对于冗余与否的判断起到重要作用。因此，用 DIoU 替代常见的 IoU 作为 NMS 的判断标准，IoU 可表示为

$$s_i=\begin{cases}0, T_{\mathrm{DIoU}}(M,B_i)\geqslant T_{\mathrm{thresh}}\\ s_i, T_{\mathrm{DIoU}}(M,B_i)<T_{\mathrm{thresh}}\end{cases},\tag{17}$$

式中，s_i 为当前框的得分，T_{DIoU} 为 DIoU 方法得到的比例，M 为当前最大得分框，B_i 为当前框，T_{thresh} 为设定的阈值，实验中设定为 0.43。DIoU 通过引入参数 β 控制 d^2/c^2 的惩罚范围，可表示为

$$T_{\mathrm{DIoU}}=T_{\mathrm{IoU}}-(d^2/c^2)^{\beta},\tag{18}$$

式中，T_{IoU} 为重合度量值，d 为当前框和当前最大框中心点的距离，c 为一个能覆盖两框的最小框对角线长度，β 为 DIoU-NMS 的超参数。可以发现，当 β 趋近于∞时，DIoU 退化为普通 IoU；当 β 趋近于 0 时，几乎所有中心点不与当前最大得分框中心重合的框都被保留，这表明将中心点距离纳入考虑有助于缓解遮挡情况的发生。实验中设置的 β 为 3。

相比算术平均数，平方平均数作为窗口的代表激活值更具有代表性。相比传统的 NMS 方法，DIoU-NMS 不仅能提高算法的识别准确性，其性能与传统 NMS 方法也基本一致。AAPM 的工作流程及可视化如图 4 所示，可以发现，该模块不经额外训练就能得到比原始 APN 子网络局部信息更多的图像部件。

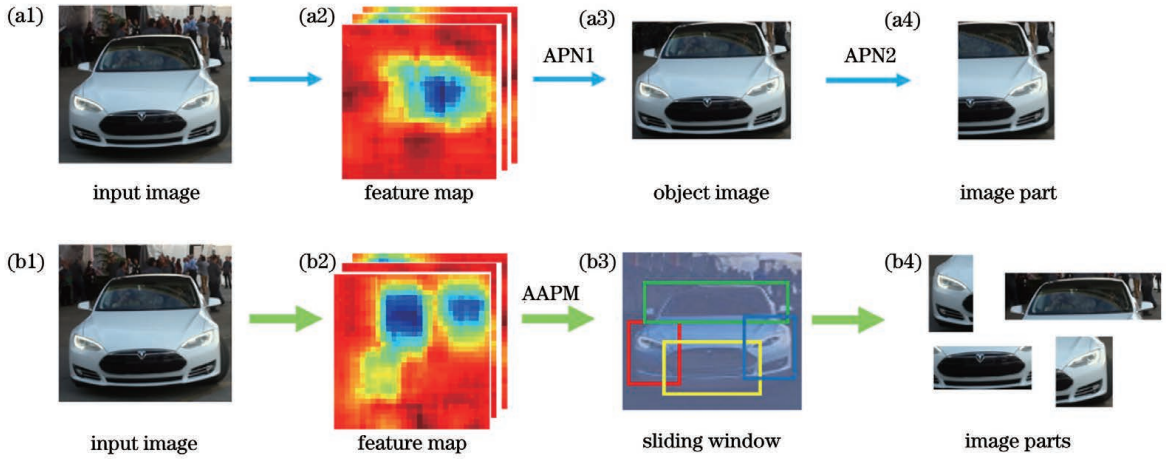

图 4　AAPM 的工作流程及可视化结果。(a)APN；(b)AAPM

Fig. 4　Workflow and visualization results of the AAPM. (a) APN; (b) AAPM

4　实验结果与分析

4.1　数据集

实验使用的细粒度数据集包括 Stanford Cars 和 Food-101 数据集，这两个数据集常被用作细粒度图像分类的基准，示例图像如图 5 所示。Food-101 数据集包含来自 101 类食物的 101000 张图像，训练图像共 75750 张，测试图像共 25250 张。Stanford

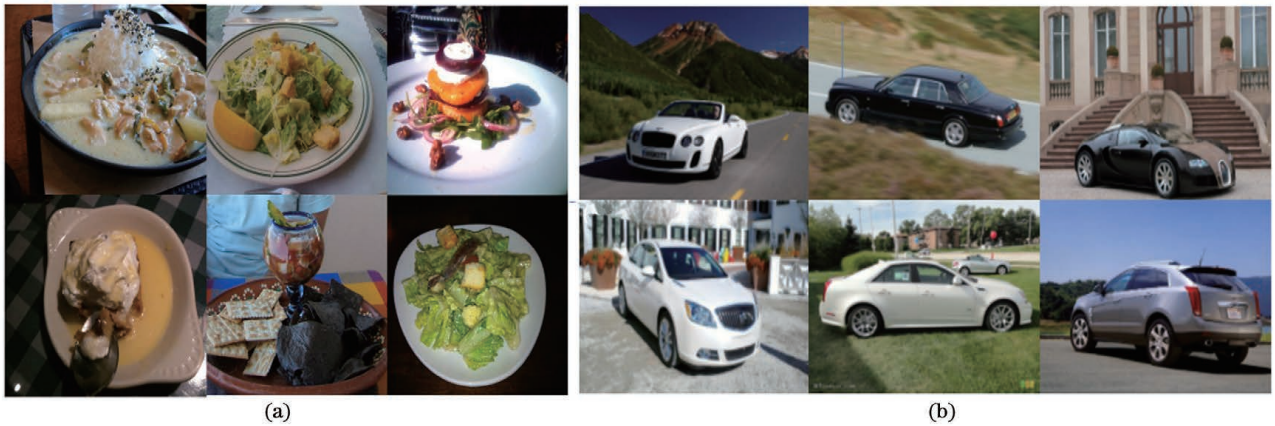

图 5　数据集示例图像。(a)Food-101 数据集；(b)Stanford Cars 数据集

Fig. 5　Example image of the datasets. (a) Food-101 dataset; (b) Stanford Cars dataset

Cars 数据集包含 196 类汽车的 16185 张图像。将所有数据分为 8144 张训练图像和 8041 张测试图像，且每个类别的训练集和测试集数量大致相等。实验中未采用边界框(Bounding box)和分割坐标等除类别外的额外标注信息。

4.2　实验环境及参数设置

实验采用的 GPU 显卡为 GTX2080 Ti，CPU 处理器为 Intel Xeon CPU E5-2650，内存为 32 GB。在 Ubuntu 系统下用开源深度学习框架 Pytorch 作为平台，实验过程中预处理输入的图像尺寸为 448 pixel×448 pixel。用在 ImageNet 预训练的 ResNet50 提取图像特征，初始学习率设为 0.001，每隔 50 次迭代后衰减至 0.1 倍，权重衰减量为 0.0001。模型使用 Adam 优化器，并用小批次梯度下降方法，批大小为 6，最大迭代次数 200。DIoU-NMS 中的超参数 β 设为 3，阈值设为 0.43。

4.3　可视化

图 6 为通过 ECA 子网络得到的类激活图和 AAPM的定位结果，可以发现，ECA子网络得到的特征图可以忽略环境信息，准确提取整体目标的位置，而 AAPM 能很好地定位目标物体的多个不同部件。如在车辆的定位结果中，AAPM 很好地定位了车的车灯、前车脸、车挡风玻璃和车门等区域，在食物识别过程中，AAPM 则准确定位了食物的表面特性、边缘形状和层次特征等。

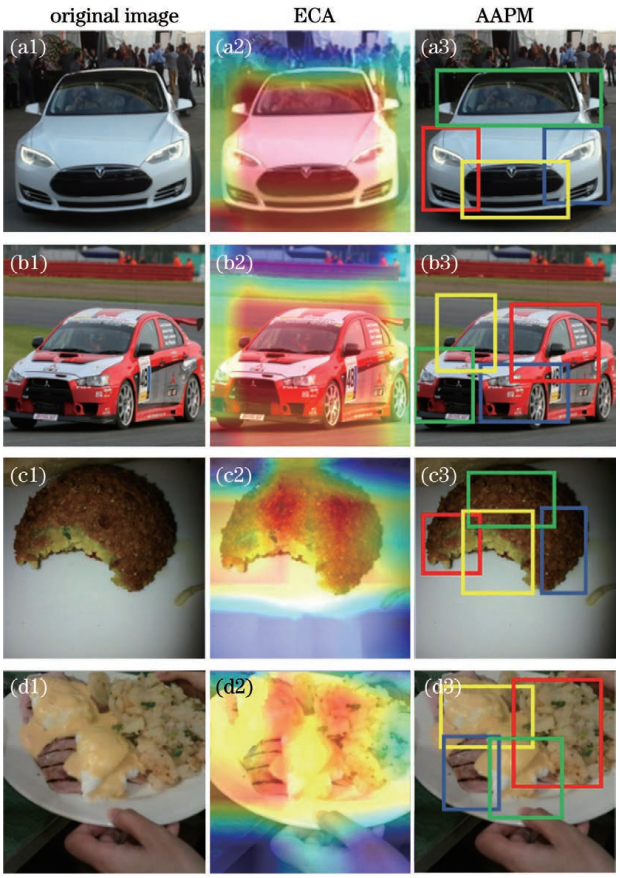

图 6　处理结果的可视化。(a)汽车 1；(b)汽车 2；(c)食物 1；(d)食物 2

Fig. 6　Visualization of processing results. (a) Car 1; (b) car 2; (c) food 1; (d) food 2

4.4　消融实验

消融实验中，主要对比了四个多分支网络结构：1)未加入 ECA 的网络；2)未用深度过参数化卷积层代替常规卷积的网络；3)未加入 AAPM 及部件分支的网络；4)ECA 多分支网络。不同多分支网络在两个数据集上的实验结果如表 1 所示，可以发现，相比未加入 ECA 的网络、未用深度过参数化卷积层代替常规卷积以及未提取部件信息的网络，同时添加所有模块的网络在 Stanford Cars 数据集上的分类准确率分别提高了 0.6、0.2 和 2.9 个百分点，在 Food-101 数据集上的分类准确率分别提高了 2.1、0.8 和 5.2 个百分点。

表 1　本方法在不同数据集上的消融实验结果

Table 1　Ablation experimental results of our method on different datasets　　unit: %

Method	Stanford Cars	Food-101
Without ECA sub-network	94.8	88.5
Without DO-Conv	95.2	89.8
Without AAPM	92.5	85.4
Ours	**95.4**	**90.6**

4.5　对比实验

将本方法与目前主流的弱监督和部分强监督分类方法进行对比。在 Stanford Cars 数据集中选取双线性 CNN(Bilinear-CNN)、RA-CNN、启发式后继网络(HS-Net)[20]、属性感知注意力网络(AAA Model)[21]和 MMAL 的结果进行对比，在 Food-101 数据集中选取 Bilinear-CNN、宽块残差网络(WISeR)[22]、食物部件 CNN(FPCNN)[23]、全卷积网络(FCA)的结果进行对比。不同方法在两个数据集上的实验结果如表 2 和表 3 所示。可以发现，相比其他方法，本方法在 Stanford Cars 数据集上的分类精度最大提升了 4.1 个百分点，在 Food-101 数据集上最大提升了 5.9 个百分点，相比其他弱监督分类方法均有所提升；且本方法在不同细粒度图像数据集上均取得了较优的分类效果，验证了该网络结构的泛化性。相比使用非学习模块进行定位的 MMAL 方法，本方法的定位精度更高，这表明注意

力机制能更好地识别目标区域。相比 RA-CNN 方法，本方法的效果更优，也证明了提取图像中多个含局部信息关键区域的重要性。此外，本方法使用添加 ECA 模块和 DO-Conv 及 AAPM 的多分支网络，能够有效分割细粒度图像的判别性部位，同时提取和整合细粒度图像的整体信息和局部特征，在多个数据集上取得了良好的效果。

表 2 不同方法在 Stanford Cars 数据集上的实验结果

Table 2 Experimental results of different methods on the Stanford Cars dataset unit: %

Method	Added training	Accuracy
Bilinear-CNN(2015)	×	91.3
RA-CNN(2017)	×	92.5
HS-Net(2017)	√	93.8
AAA Model (2019)	√	95.4
MMAL(2020)	×	95.0
Ours	×	**95.4**

表 3 不同方法在 Food-101 数据集上的实验结果

Table 3 Experimental results of different methods on the Food-101 dataset unit: %

Method	Added training	Accuracy
Bilinear-CNN(2015)	×	84.7
WISeR(2018)	×	90.3
FPCNN(2018)	×	87.9
FCA(2019)	√	86.3
Ours	×	**90.6**

5 结 论

提出了一种引入通道注意力模块的多分支网络细粒度图像识别方法，在原有 RA-CNN 中引入高效通道注意力模块子网络和深度过参数化卷积及 AAPM，提高了原始网络的识别精度。实验结果表明，本方法在 Stanford Cars 和 Food-101 数据集上的识别精度分别达到了 95.4%与 90.6%。原因是本方法通过获取更准确的目标位置信息和目标中多个不同部件的局部特征信息及提高卷积运算，学习到更多的参数，有效提高了网络的识别性能。相比其他方法，本方法在弱监督条件下能够达到较优的识别结果，甚至优于部分强监督模型的性能，为图像信息的特征提取提供了一种新思路。后续研究可着眼于对关键区域的定位和特征提取，以取得更好的识别结果。

参 考 文 献

[1] 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.

[2] 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.

[3] Li S Y, Liu Y H, Zhang R F. Fine-grained image classification based on multi-scale feature fusion[J]. Laser & Optoelectronics Progress, 2020, 57(12): 121002.
李思瑶, 刘宇红, 张荣芬. 多尺度特征融合的细粒度图像分类[J]. 激光与光电子学进展, 2020, 57(12): 121002.

[4] Lin X N, Qin F W, Peng Y, et al. Fine-grained pornographic image recognition with multiple feature fusion transfer learning[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(1): 73-86.

[5] Zhang H, Xu T, Elhoseiny M, et al. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1143-1152.

[6] 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, USA. New York: IEEE Press, 2017: 4476-4484.

[7] Zhang F, Li M, Zhai G S, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization[EB/OL]. (2020-07-21)[2021-01-03]. https://arxiv.org/abs/2003.09150.

[8] Yang Z, Luo T G, Wang D, et al. Learning to navigate for fine-grained classification[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 438-454.

[9] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition[C]//2018 IEEE/CVF Conference on

Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8697-8710.

[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), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1449-1457.

[11] Li Q N, Sun H X, Sun K J. Fine-grained classification of sleeper shoulder crack images based on improved B-CNN[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141013.
李启南, 孙海鑫, 孙可佳. 基于改进 B-CNN 的轨枕挡肩裂纹图像细粒度分类[J]. 激光与光电子学进展, 2020, 57(14): 141013.

[12] Liu K, Wang D, Rong M X. X-ray image classification algorithm based on semi-supervised generative adversarial networks[J]. Acta Optica Sinica, 2019, 39(8): 0810003.
刘坤, 王典, 荣梦学. 基于半监督生成对抗网络 X 光图像分类算法[J]. 光学学报, 2019, 39(8): 0810003.

[13] Xu Z J, Wang D. Multi-pose face recognition with two-cycle generative adversarial network[J]. Acta Optica Sinica, 2020, 40(19): 1910002.
徐志京, 王东. 基于双路循环生成对抗网络的多姿态人脸识别方法[J]. 光学学报, 2020, 40(19): 1910002.

[14] Sermanet P, Eigen D, Zhang X, et al. Overfeat: integrated recognition, localization and detection using convolutional networks[EB/OL]. (2014-02-24)[2021-01-03]. https://arxiv.org/abs/1312.6229.

[15] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[16] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[18] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[19] Cao J M, Li Y Y, Sun M C, et al. DO-Conv: depthwise over-parameterized convolutional layer[EB/OL]. (2020-06-22)[2021-01-03]. https://arxiv.org/abs/2006.12030.

[20] Lam M, Mahasseni B, Todorovic S. Fine-grained recognition as HSnet search for informative image parts[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6497-6506.

[21] Han K, Guo J Y, Zhang C, et al. Attribute-aware attention model for fine-grained representation learning[C]//Proceedings of the 26th ACM International Conference on Multimedia, October 22-26, 2018, Seoul Republic of Korea. New York, NY, USA: ACM, 2018: 2040-2048.

[22] Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 567-576.

[23] Zheng J N, Zou L, Wang Z J. Mid-level deep food part mining for food image recognition[J]. IET Computer Vision, 2018, 12(3): 298-304.