高技术通讯 2024 年 第 34 卷 第 12 期:1266-1278

doi:10.3772/j.issn.1002-0470.2024.12.003

多视角解耦增强整合的细粒度分类算法①

孟月波② 王 博 刘光辉

(西安建筑科技大学信息与控制工程学院 西安 710055)

(西安市建筑制造智能化技术重点实验室 西安 710055)

摘 要 针对细粒度图像分类中由于背景环境、光照条件、样本姿态和拍摄角度等外部因素导致类内差异显著增加的问题,本文提出了多视角解耦增强整合的细粒度分类算法。首先,为了降低图像中外部因素的干扰,设计多视角注意力(MPA)模块,此模块通过将模型分解为数个视角,迫使每个视角关注不同尺度,实现干扰因素的解耦,并通过对特征进行自注意力建模,引导各个视角进一步挖掘关键特征。其次,提出递进式动态加权融合(PDWF)策略,旨在有效整合解耦后的多个视角信息,该策略通过获取不同视角下通道和空间关系动态调整融合系数,实现多尺度信息的高阶融合。最后,采用递进式训练方法促进视角交互,进一步捕获和整合多尺度特征的互补语义信息。在 CUB-200-2011、Stanford-Cars、FGVC-Aircraft 公开数据集上进行实验,实验结果表明所提方法分类准确率分别达到 90.5%、95.5% 和 94.2%,优于当前细粒度图像分类任务主流方法。

关键词 细粒度;多视角注意力(MPA);递进式动态加权融合(PDWF);图像分类

与传统的大类别图像分类相比,细粒度图像分类旨在进行更细致的子类别划分,例如狗的品种或车的型号等。这一领域的研究不仅在学术界具有重要的科研意义,且在实际场景中有着广泛的应用前景。细粒度图像分类的重要性不言而喻,然而它面临着独特的挑战,其数据分布的特性使得类间样本差异较小,而类内差异相对较大,对模型分类性能提出了较高要求。

细粒度图像分类算法通过训练标签是否包含额外标注可分为强监督和弱监督 2 个方向,强监督采用额外监督信息(如标注框、锚框等)辅助模型进行训练,弱监督则仅采用类别标签进行训练。早期受技术发展的限制,多采用强监督算法。Zhang 等人[1]采用了在图像目标检测(objection detection)[2-3]领域中广泛应用的区域卷积神经网络(region-based convolutional neural network,R-CNN),提出了 Part R-CNN。根据事先的标注信息,通过对图像进行检测,算法在分类精度上取得了不错的效果。但由于强监督的算法依赖手工标注、训练成本等问题导致其在后续应用中难以扩展和普及。

随着神经网络的发展,弱监督算法逐渐成为主流的研究方法。该类算法主要集中在提取和定位最具区别性的部分、与其他特征互补的部分以及各种尺度的特征部分。在这些研究中,注意力机制有着非常重要的地位。Xiao 等人[4]提出两级注意力算法,其分别关注对象级和局部级 2 个不同层次的特征,通过强大的网络结构得到良好的分类性能。Peng 等人[5]提出了一种对象部位注意力模型,通过对象级和部件级注意力来使得模型获取到强大的特征表示。Zheng 等人[6]利用基于注意力的非均匀采样器来突出感兴趣的部分。Hu 等人[7]利用注意力机制将特征图中最为显著的部分进行裁剪和丢弃,并将处理之后的图像进行二次分类。Zheng 等人[8]通过多注意力同时定位几个身体部件,并通过聚类

① 国家自然科学基金(52278125)资助项目。

② 女,1976 年生,博士;研究方向:计算机视觉理解,建筑环境智能感知与调控,智动化技术与系统;联系人,E-mail:mengyuebo@163.com。

(收稿日期:2024-04-16)

生成多个部分。这类研究利用注意力机制引导模型聚焦于图像中的显著区域,但缺乏对多尺度特征特别是小尺度特征的深入挖掘。

Lin 等人[9]通过设计双线模型对同一目标进行特征提取,能够端到端地实现区域检测和特征提取。有几种方法[10-11]通过挖掘多尺度下部件的互补信息来提取强大的特征表示。Chao 等人[12]在双线性池化的基础上提出了分层双线性池化,隐式地进行不同特征层的互补信息融合建模。Ye 等人[13]通过在多个层次之间建立空间依赖关系,对不同的特征层显示地进行融合建模,以学习更多关键性特征。文献[14]指出关键特征隐藏在各个尺度下,该算法迫使网络学习不同尺度特征的方式为本文提供了一定的思路。这类研究致力于挖掘多尺度特征,从而强化各尺度间的互补性。然而,由于缺乏有效的注意力引导机制,模型在捕捉关键信息方面存在明显的不足。

基于上述分析,本文借鉴分解、融合的思想,提出了多视角解耦增强整合的细粒度分类算法。首先,将模型的感受野通过模块固定在较小的尺度上,可以显著减少外部因素(如光照、姿态等)对分类结果的干扰。然而,固定感受野会限制模型对多尺度特征的感知能力,为此,本文设计了多视角注意力(multi-perspective attention,MPA)模块,确保了模型提取特征尺度的多样性以及对关键信息的聚焦能力。最后,为解决视角之间存在的信息割裂问题,即不同视角提取的特征可能无法有效地相互补充和增强,进一步提出了递进式动态加权融合(progressive dynamic weight fusion,PDWF)策略以整合不同视角信息。通过以上改进,所提出的方法能够在减少外部因素干扰的同时,充分挖掘和融合多尺度特征,综合提高模型的分类性能。

1 多视角解耦增强整合网络

多视角解耦增强整合网络的整体结构如图 1 所示,采用 ResNet-50 作为主干网络,将其中后 3 个阶段(stage_3,stage_4,stage_5)的输出视为算法的

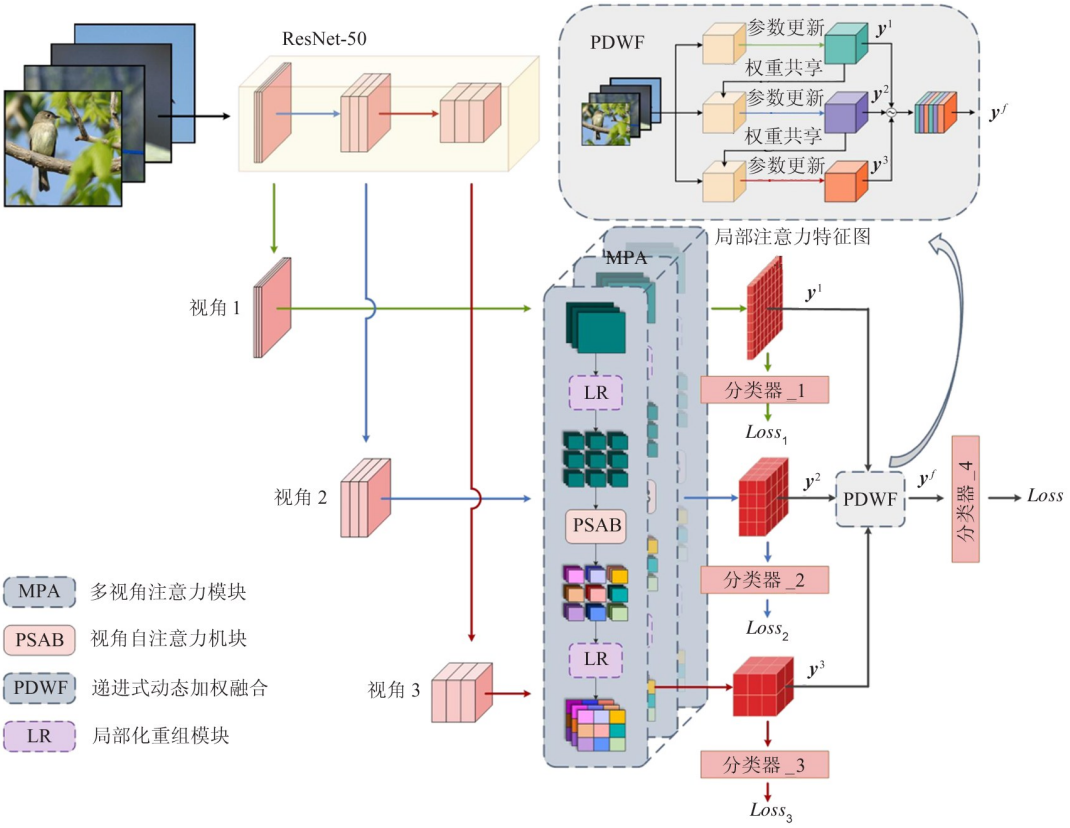

图 1 网络整体结构框图

3 个视角,每个视角都有独立的输出($i=1,2,3$),并依次进行训练,视角之间参数共享。为提高每个视角对于不同尺度的兴趣度,设计了 MPA 模块,并分别引出 3 条分支 MPA-i 作用于每个视角,每个 MPA-i 中都包含有局部化重组模块(localize and restore module,LR)和视角自注意力块(perspective self-attention block,PSAB)2 个子模块。MPA 作用在每个视角,使其专注于捕获特定尺度,并进行相应粒度的局部注意力建模。为了加强不同尺度特征间的互补关系,通过 PDWF 整合所有视角,促进视角之间的信息交互。

1.1　重新审视渐进式训练策略

文献[14]提出渐进式训练策略,通过逐步增加网络复杂度和引入新的模块来提高网络性能。该策略的每个阶段或模块具有特定的任务和网络结构。随着阶段的增加,模型逐渐变得更强大,从初始阶段的简单模型到后续阶段的深层网络结构。此外,该文献通过结合渐进式训练策略提出了随机拼图的方法,以限制模型的视野在不同尺度下进行学习。

但是该策略并未有效解决模型注意力引导的问题。由于缺乏明确的注意力机制,模型在处理复杂任务时可能无法准确聚焦于关键信息,进而影响训练效率和模型性能。模型中不同阶段之间的联系相对较弱,导致模型在各个阶段学习到的特征缺乏有效的信息交互。这种割裂式学习会使得模型难以形成对整体任务的全面理解,进而影响其在复杂场景下的泛化能力和适应性。各个阶段所学特征之间缺乏整合,可能会导致模型在处理实际问题时表现出片面或不准确的行为。

基于此,本文设计了多视角解耦增强整合网络。该方法不仅迫使模型挖掘多尺度的特征信息,并且引入注意力机制来指导模型精准挖掘和聚焦关键信息。并对多尺度特征采用高阶融合,强化不同尺度特征之间的互补与协同,从而提升模型的整体性能和泛化能力。

1.2　多视角注意力

图 2 为 MPA-i 的结构图。首先,LR 模块对每个视角的输入特征图进行局部化分块,以限制当前视角下模型的感受野;其次,局部化分块之后的特征会被送至 PSAB 模块中,该模块通过对特征图中任意两两像素之间的坐标关系进行自注意力建模,从而捕获丰富的空间信息;随后,LR 模块将特征重组至原始大小;最终,得到经过局部注意力建模的特征图。公式如式(1)所示。

$$y^k = R(\mathrm{PSAB}(\prod_{i=0}^{m}\mathrm{Conv}(L(x^k),\theta_k))) \tag{1}$$

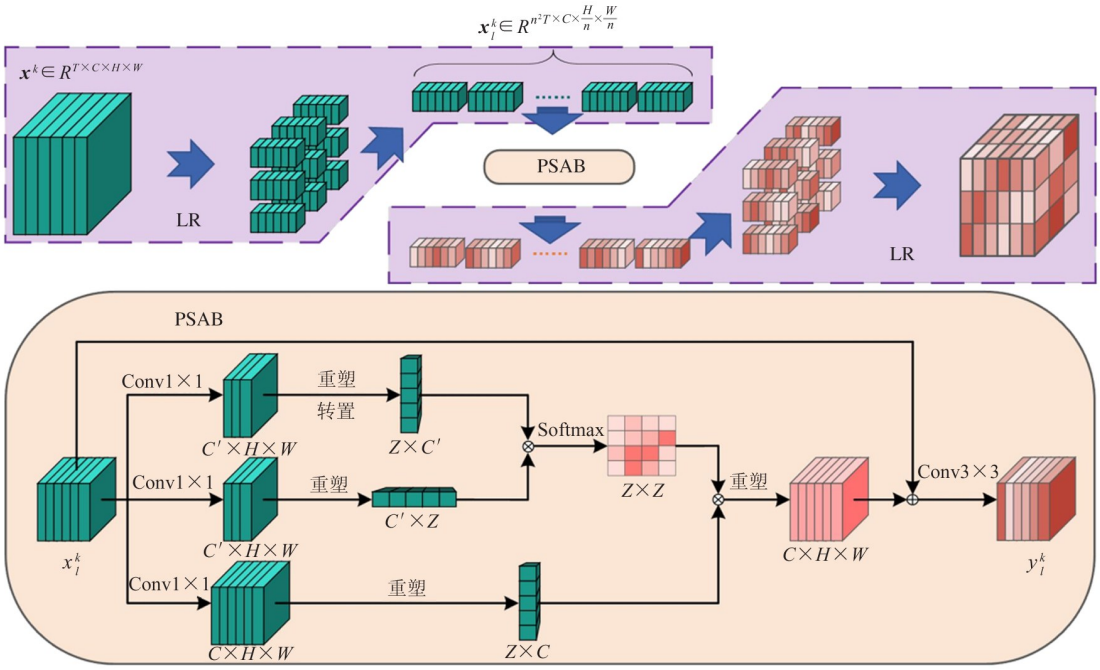

图 2　多视角注意力(MPA-i)结构图

其中,x^k 为视角 $k \in \{1,2,3\}$ 的输入;y^k 表示该视角下的输出;$\prod_{i=0}^{m}\text{Conv}$ 表示当前视角下包含的卷积层;θ_k 为卷积层的参数,其中 m 的取值与视角 k 相关;PSAB(·)表示视角自注意力块;$L(\cdot)$ 和 $R(\cdot)$ 分别表示 LR 模块的不同功能。接下来将详细介绍 LR 和 PSAB 的工作原理。

1.2.1 局部化重组模块

如图 2 上半部分所示,LR 模块拥有局部化和重组 2 个功能(便于区分,这里用 LR-L 表示局部化功能,LR-R 表示重组功能),局部化功能会将特征进行局部化分块,重组功能则能使局部特征块恢复至原始大小。当输入 x^k 被送至 LR 模块中,该模块首先通过 LR-L 对输入 x^k 进行局部化特征处理,然后通过在空间维度上进行分割将输入特征划分为多个局部特征块,并将特征块在批次维度上进行拼接。具体而言,对于每个视角的输入 $x^k \in R^{T\times C\times H\times W}$(其中 T 表示训练的批量大小,C 表示特征的通道,H、W 分别表示特征的高度和宽度),在高宽维度上通过局部化超参数 n,使得对应的特征图被分离成 n^2 个局部特征块,这些局部的特征块会独立作为一个训练批次参与卷积等运算,并在一系列计算之后通过 LR-R 恢复至原始大小。LR 模块中局部化功能和重组功能的计算公式如下所示。

$$x_l^k \in R^{n^2T\times C\times \frac{H}{n}\times\frac{W}{n}} = L(x_l^k \in R^{T\times C\times H\times W}) \tag{2}$$

$$y^k \in R^{T\times C\times H\times W} = R(x_l^k \in R^{n^2T\times C\times \frac{H}{n}\times\frac{W}{n}}) \tag{3}$$

其中,x_l^k 表示输入 x^k 经过 LR 模块局部化处理之后的输出;$L(\cdot)$ 函数表示 LR 模块的局部化功能,该函数通过调整超参数来控制局部特征块的大小;$R(\cdot)$ 表示 LR 模块的重组功能,可将局部特征图重新恢复为整体的特征图。

如式(2)所示,经过 LR-L 处理后的所有局部特征会进行卷积运算和注意力建模。由于局部特征块的大小固定,模型在卷积运算过程中无法获得全局信息,因此模型在训练过程中会被迫关注特定大小的局部区域。这种做法增强了模型对局部特征的专注度,提高了其挖掘小尺度特征的能力。

1.2.2 视角自注意力块

MPA 会为每个视角分配特定的视角自注意力块,并通过该自注意力块在局部特征图中捕捉空间信息。受位置注意力模块[15]和自注意力机制[16]的启发,本文设计了视角自注意力块 PSAB,用于捕获局部特征图上任意 2 个像素之间的空间相关性,从而进行空间注意力建模。

PSAB 的框架如图 2 下半部分所示,对于任意给定的局部特征图 $x_l^k \in R^{T\times C\times H\times W}$(其中 n^2T 为批量大小,由于在 PSAB 中的运算并不会改变批量大小,因此在后文中会省略批量维度的表示),通过 1×1 卷积获取任意位置的特征 $K_A \in R^{C'\times\frac{H}{n}\times\frac{W}{n}}$、$K_B \in R^{C'\times\frac{H}{n}\times\frac{W}{n}}$ 以及 $K_C \in R^{C\times\frac{H}{n}\times\frac{W}{n}}$($C'$ 表示位置特征嵌入通道数,本文取 $C' = C/16$),分别将 K_A、K_B 重塑为 $R^{Z\times C'}$ 和 $R^{C'\times Z}$,在 K_A 和 K_B 之间使用矩阵乘法运算,其中 Z 为局部特征图的像素总数。通过函数将原始特征转化为概率分布,得到空间相关性特征图 $K_P \in R^{Z\times Z}$,公式如下:

$$p_{ij} = \frac{\exp(K_{A_i}K_{B_i})}{\sum_{i=1}^{Z}\exp(K_{A_i}K_{B_i})} \tag{4}$$

其中,p_{ij} 表示特征图中第 i 个位置对第 j 个位置的相关性概率分布。之后将位置特征 $K_C \in R^{C\times\frac{H}{n}\times\frac{W}{n}}$ 重塑为 $K_C \in R^{Z\times C}$,与空间相关性特征图 K_P 执行矩阵乘法,并将输出特征矩阵 $\boldsymbol{K}_{O'}$ 重塑回 $R^{C\times\frac{H}{n}\times\frac{W}{n}}$,通过残差连接特征 $x_l^k \in R^{C\times\frac{H}{n}\times\frac{W}{n}}$。最后将结果输入到 3×3 卷积层中,得到最终的输出特征 $y^k \in R^{C\times\frac{H}{n}\times\frac{W}{n}}$,融合公式为

$$y_l^k = \gamma\sum_{i=1}^{Z}(K_{P_{ij}}K_{C_i}) + x_l^k \tag{5}$$

其中,γ 为残差系数,该参数初始值为 0,随着训练的推进,模型具备了一定的特征分辨能力,位置相关性特征图能产生正确的引导,此时 γ 会被分配更多的权重,从而使模型获取到更多全局信息。

每个位置上的最终输出 y_l^k 都为所有位置上的特征图和输入特征图的加权和,因此 y_l^k 具有丰富的空间信息,能够使得局部特征输入在聚合上下文信息时进一步获取到关键特征,从而提高类内相关性和语义一致性。

1.3 递进式动态加权融合策略

MPA 通过注意力和局部化使每个视角专注于不同尺度特征的学习。然而,各视角输出的特征缺

乏有效高阶交互，导致模型所学习到的特征缺乏互补性。本小节设计了递进式动态加权融合策略 PDWF，模块结构如图 3 所示。

如图 3 上半部分所示，第 1 列部分代表每个视角的参数，当前视角训练完毕后所学习到的权重参数将共享至下个视角的训练，并在所有视角都训练

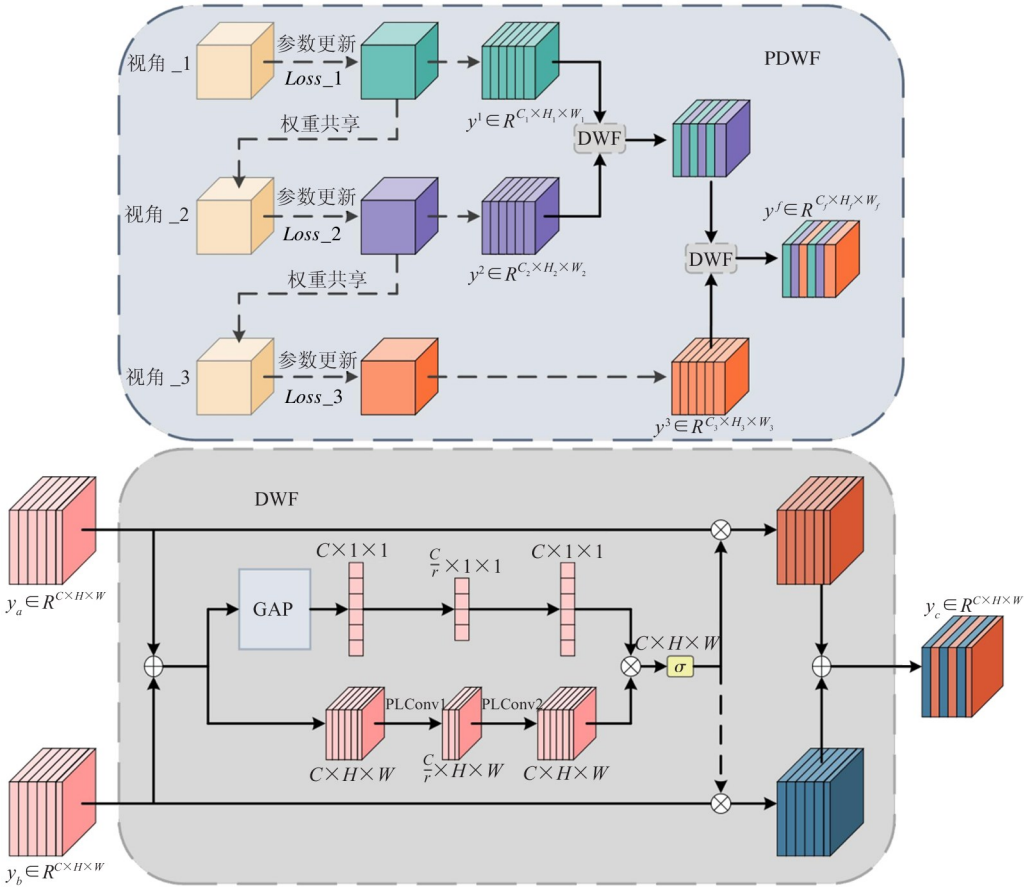

图 3　动态加权特征融合示意图

完毕后通过动态加权融合模块（dynamic weight fusion，DWF）整合所有视角的特征信息。此模块迫使网络共享各视角所挖掘到的特征信息，增进视角之间的信息交互。该策略遵循了从细节到全局的视野层次，有效提升了不同尺度之间特征的互补关系。DWF 对特征进行融合过程如图 3 下半部分所示，对于不同视角特征输出 $y^a \in R^{C\times H\times W}$，$y^b \in R^{C\times H\times W}$（其中 $a,b \in \{1,2,3\}$）。首先将 2 个特征进行加权和；然后将加权之后的特征分别输入到 2 个分支，第 1 个分支对加权特征进行空间注意力建模，捕获位置相关性。本文采用像素级卷积（pixel-level convolution，PLConv）通过对特征图中的空间位置逐像素进行交互，从而聚合局部空间语义信息。第 2 个分支通过压缩激励块[17]捕获通道相关性，将加权特征进行全局平均池化操作之后，接入全连接层得到输出。将 2 条分支得到的结果进行广播加法，并使用激活函数得到大小在 0 ~ 1 之间的特征加权比例，计算公式如下：

$$\boldsymbol{P}_{a+b} = \mathrm{PLConv}_2(\mathrm{ReLU}(\mathrm{PLConv}_1(y^a + y^b))) \tag{6}$$

$$\boldsymbol{C}_{a+b} = fc(\mathrm{GAP}(y^a + y^b)) \tag{7}$$

$$W_{a+b} = \mathrm{sigmoid}(\boldsymbol{P}_{a+b} \oplus \boldsymbol{C}_{a+b}) \tag{8}$$

其中，$\boldsymbol{P}_{a+b} \in R^{T\times C\times H\times W}$ 为位置相关性矩阵，$\boldsymbol{C}_{a+b} \in R^{T\times C\times 1\times 1}$ 为通道相关性矩阵，GAP 为全局平均池化操作，$W_{a+b} \in R^{T\times C\times H\times W}$ 则为最终的加权特征的权重，$\oplus$ 为矩阵广播加法，fc 表示全连接层，PLConv 表示像素级卷积。这些权重 W_{a+b} 为特征的融合系数，具有丰富的空间相关性和通道相关性。具体而言，该模块将融合系数作为可学习的参数。通过对加权特征

的激活响应进行预测,从而获取简单融合后的特征。根据该响应,得到特征的融合比例,特征响应经过激活函数处理,因此融合比例 W_{a+b} 取值范围在 0 ~ 1 之间。通过自适应调整融合程度,融合后的权重能够根据相关性的变化进行动态调整。具体公式如下所示。

$$y_{a\odot b} = W_{a+b} \otimes y^a + (1 - W_{a+b}) \otimes y^b \tag{9}$$

其中,$y_{a\odot b} \in R^{T\times C\times H\times W}$ 表示自适应融合特征;⊗表示逐元素相乘;通过注意力模块得到第 1 个特征 y^a 的融合比例 W_{a+b};相应地,$1 - W_{a+b}$ 表示另一个特征 y^b 的融合比例。因此,网络能够在特征 y^a 和 y^b 之间进行软注意力权重选择,并根据注意力的预测调整参数,实现自适应融合。各个不同阶段的特征输出经自适应特征融合模块整合为最终的输出特征,并将其输入至分类得到的预测输出。当所有视角均训练完毕后,会在最后阶段通过 DWF 模块对所有视角特征进行高阶融合,结合图 3 上半部分和式(3) ~ (9),可最终融合特征输出 y^f 的具体公式为

$$y^f = (y^1 \odot y^2) \odot y^3 \tag{10}$$

1.4 损失函数

模型通过损失函数进行模型参数的更新,而损失函数的选择是网络提升性能的关键因素之一。交叉熵(cross entropy,CE)损失函数是细粒度图像分类中较为常用的损失函数,其通过量化预测概率分布与真实概率分布之间的“距离”,从而有效反映出细粒度数据中的类内相似度和类间差异性。本小节采用损失函数的数学描述如式(11)、(12)所示。

$$L_{\mathrm{CE}}^k(\tilde{y}^k, y) = -\sum_{i=1}^{T} y_i^k \times \log(y) \tag{11}$$

$$L_{\mathrm{CE}}^f(\tilde{y}^f, y) = -\sum_{i=1}^{T} y_i^f \times \log(y) \tag{12}$$

式中,L_{CE}^k 代表视角 k 的损失函数(其中 $k \in \{1,2,3\}$);L_{CE}^f 表示最终阶段进行高阶融合的损失函数;y 表示样本的标签;$\tilde{y}^k$、$\tilde{y}^f$ 分别代表预测标签;T 代表批次所包含的图像数量。网络整体损失 L_{CE} 为各个阶段损失之和,公式如下所示。

$$L^{\mathrm{loss}} = L_{\mathrm{CE}}^f + \sum_{k=1}^{3} L_{\mathrm{CE}}^k \tag{13}$$

1.5 整体训练

为提升网络在挖掘不同粒度特征信息方面的效率,本文引入了递进式的训练策略。该策略为每个视角施加独立的分类损失,并基于这些损失更新网络参数。同时,当前视角更新后的参数将通过权重共享机制传递至下一个视角,以实现视角间信息的有效交互和融合。由式(1) ~ (3)可知,任意视角的图像输入 x^k 经过运算得到视角输出 y^k,随后会输入至该视角下的分类器中。分类器由 2 个批归一化(batch normalization, BN)层、1 个指数线性激活函数(exponential linear unit, ELU)和 2 个全连接层(fully connected layers, FC)组成,经过分类器得到最终预测输出 $\tilde{y}^k$,具体公式如下所示。

$$\tilde{y}^k = \mathrm{CLS}_k(y^k) \tag{14}$$

$$\mathrm{CLS}_f = \mathrm{CLS}_k = fc(\mathrm{ELU}(\mathrm{BN}(fc(\mathrm{BN}(\cdot))))) \tag{15}$$

其中,CLS_k 表示第 k 个视角的分类器,CLS_f 为最终阶段高阶融合的分类器。训练过程中,模型会在一个迭代(epoch)中经历 4 个迭代阶段,其中每个阶段专门负责训练一个特定的视角并更新网络模型参数。在所有视角完成训练后,通过 DWF 模块进行视角信息的高阶融合,整个训练过程使网络能够学习到丰富的多尺度特征和互补的语义信息,最终实现一次完整的训练。整体训练过程如算法 1 所示。

算法 1 模型整体训练流程

输入:训练数据集 D_{tr},样本批次大小 T,数据样本 x,样本标签 y,视角 k

输出:模型预测 $\tilde{y}^k$、$\tilde{y}^f$

for $epoch = 0$; $epoch < EPOCHS$; $epoch$ ++ **do**
  for $(x, y) \Leftarrow D_{tr}(T)$ **do**
    for $k = 1$; $epoch \leqslant 3$; k ++ **do**
      $y^k = (PSAB(\prod_{i=0}^{m} Conv(L(x^k), \theta_k)))$
      $\tilde{y}^k = CLS_k(y^k)$
      $L^k \Leftarrow L_{\mathrm{CE}}(\tilde{y}^k, y)$
      $L^k \Rightarrow backward$
    end for
    $\tilde{y}^f \Leftarrow CLS_f(y^f) = CLS_f((y^1 \odot y^2) \odot y^3)$
    $L^f \Leftarrow L_{\mathrm{CE}}(\tilde{y}^f, y)$
    $L^f \Rightarrow backward$
  end for
end for

2　实验对比分析

2.1　实验数据集

本文在实验部分采用 3 个标准的细粒度数据集 CUB-200-2011[19]、FGVC-Aircraft[20] 以及 Stanford Cars[21] 来进行算法的有效性评估。

表 1 展示了本文采用的 3 个数据集具体参数。CUB-200-2011 数据集包含约 11 788 张图像样本，涵盖了 200 种不同的鸟类；本文使用了 5 994 张图像作为训练集，剩下的 5 794 张图像作为测试集。Stanford Car 数据集共有约 16 185 张图像样本，包含了 196 种不同品牌和型号的汽车；其中的 8 144 张图像作为训练集，8 041 张图像作为测试集。FGVC-Aircraft 数据集是一个包含 10 000 张图像样本的大规模航空器分类数据集，包括 100 种不同的航空器类别；其中 6 667 张图像作为训练集，3 333 张图像作为测试集。这些数据集为本文的研究提供了广泛的样本和多样性，用于评估所提出的方法在细粒度图像分类任务中的性能和鲁棒性。本文在实验中仅使用数据集类别标签进行训练，均未采用标记框等额外信息。

表 1　数据集具体参数

数据集	训练集样本	测试集样本	类别数
CUB-200-2011	5 994	5 794	200
Stanford Cars	8 144	8 041	196
FGVC-Aircraft	6 667	3 333	100

细粒度图像分类多采用分类准确率（accuracy，Acc）作为评价指标，为方便进行算法的对比，本文采用准确率作为本文的评价指标，具体数学公式描述为

$$Acc = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i = y_i) \tag{16}$$

其中，n 表示数据集中包含的样本总数，$\tilde{y}_i$ 为模型对类别样本的预测值，y_i 为类别标签值。

2.2　实验环境和参数选择

本文算法基于开源 Pytorch 深度学习框架实现，采用的 GPU 型号为 RTX2080ti。将输入图像调整为 550×550 像素大小，使用随机裁剪得到大小为 448×448 的输入图片。采用动量批随机梯度下降方法作为优化器，初始学习率为 0.002 0，骨干网络初始学习率设置为 0.000 2，动量因子设置为 0.9，学习率采用余弦退火策略进行调整，设置权重衰退为 0.000 05，最大迭代次数为 200 轮，将所有数据集批大小设置为 12。本文采用水平翻转的图像增广方式，并且对于局部重组模块分离超参数的设置沿用了文献[14]对于随机拼图模块参数的设置（其中 $n \in \{8,4,2,1\}$）分别对应浅层到深层视角、以及最终高阶融合阶段的局部化超参数取值）。

模型参数结构信息如表 2 所示，其中 conv1 ~ conv5 代表主干网络 ResNet-50 的 5 个卷积阶段。为构建多视角特征，本模型选取了 conv3 _ x 的输出作为视角 1，conv4 _ x 的输出作为视角 2，以及 conv5 _ x 的输出作为视角 3。视角之间彼此共享主干网络参数。在每个视角上均部署了一个 MPA 模块的分支 MPA-i。MPA-i 首先通过一个 1×1 的卷积层将通道维度调整至 1 024，以适应后续操作。接着，该分支引入了一个 PSAB 模块，该模块包含 3 个 1×1 的卷积层，用于进行空间自注意力建模，从而捕捉

表 2　模型整体结构参数表

模型	参数
Conv1	7×7，64，步距 2
Conv2 _ x	$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$
Conv3 _ x	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times4$
Conv4 _ x	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times6$
Conv5 _ x	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$
MPA	$[1\times1,\ 1024]\times3$ $(PSAB)\begin{bmatrix}1\times1, & 64\\ 1\times1, & 64\\ 1\times1, & 1024\end{bmatrix}\times3$
PDWF	$\begin{bmatrix}1\times1, & 256\\ 1\times1, & 1024\end{bmatrix}\times2$

特征图中的空间依赖关系。最后,模型采用 PDWF 模块对多视角特征进行融合。PDWF 模块在建模通道关系时,利用 2 个全连接层对特征图的通道维度进行变换和整合,以捕捉不同通道之间的依赖关系。而在建模空间关系时,PDWF 模块采用 2 个 1×1 的像素级卷积层,在特征图的空间维度上进行,旨在提取和强化空间位置信息。

2.3 网络训练和测试

本文算法在 CUB-200-2011 鸟类数据集下网络的损失收敛情况和精度变化曲线如图 4 所示。通过观察发现,在训练前期模型的训练损失(*train _ Loss*)和验证损失(*test _ Loss*)波动较大,当训练达到 150 轮次时,模型损失趋于平衡,此时训练的分类精度(*train _ Acc*)和验证分类精度(*test _ Acc*)均收敛,其中 *test _ Acc* 达到 90.5%。

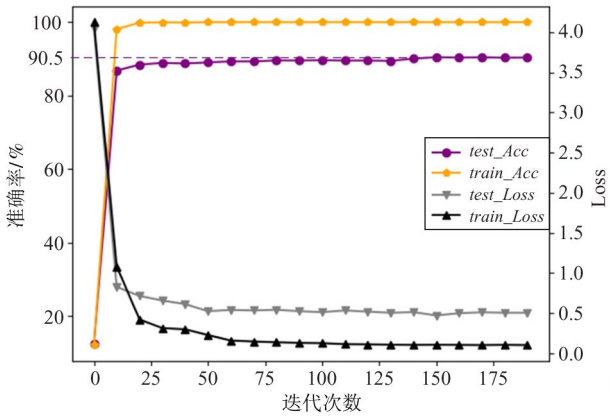

图 4 网络收敛和准确率变化情况

2.4 消融实验

本实验选用 CUB-200-2011 鸟类数据集作为消融实验数据集,并且使用 ResNet-50 作为主干网络探讨各模块的有效性。

2.4.1 LR 模块视角感受野对比实验

多视角注意力模块通过 LR 模块对特征进行局部化分块,以实现限制模型感受野的功能。在本小节中,将展示 LR 模块对于特征感受野的限制能力。由表 3 数据可以看出,在每个视角中模块都能将感受野限制在原本理论感受野大小的 1/2 左右,使模型能够挖掘更小的尺度特征。

2.4.2 LR 模块与 PSAB 组合方式分析讨论

本小节将讨论在多视角注意力 MPA 模块中,视角自注意力块 PSAB 与局部化重组 LR 模块的兼容性。

表 3 LR 在不同视角中对特征感受野的限制

视角	分块尺度	理论感受野	模块感受野
视角 1(浅层)	8	99×99	56×56
视角 2(中层)	4	291×291	112×112
视角 3(深层)	2	483×483	224×224

如表 4 所示,在分别采用了 PSAB、LR 之后,模型相较于基线网络分类准确率分别提高了 0.3%、1.1%,证明了子模块的有效性。表中第 2 部分为多视角注意力 MPA 模块中子模块组成方式,MPA 模块在每个视角中包含了对应视角的 PSAB 和 LR 模块。在实验过程中,发现不同方式组合的 MPA 对模型的提升效果差异显著。第 1 种线性串联的方式所得到的模型分类精度为 88.5%,相比于直接使用 LR 模块时精度降低了 0.5%。推测原因可能为通过线性串联的方式进行组合会使得两模块的缺点凸显,PSAB 无法充分挖掘除显著区域外的其他区域,LR 模块也会增加过多无关局部特征块,从而影响模型训练。而通过耦合连接的方式能够使 LR 模块通过局部化特征分块的方式,有效中和 PSAB 在训练过程中对显著特征而非关键特征的倾向性,使得模型能够更全面地挖掘和识别潜在的关键特征。同时,PSAB 能够引导模型过滤掉无关的局部特征,专注于学习并强化关键特征。实验证明该种方法所得到的分类精度为 90.0%,因此本文采用第 2 种组合方式构建 MPA 模块。

表 4 不同 MPA 设计在 CUB-200-2011 数据集上实验结果

算法	准确率/%
ResNet-50	84.1
ResNet-50 + PSAB	88.2
ResNet-50 + LR	89.0
ResNet-50 + MPA(PSAB + LR)	88.5
ResNet-50 + MPA(PSAB × LR)	90.0

注:表中“+”表示 2 模块线性串联,“×”表示 2 模块进行耦合连接

2.4.3 多视角注意力模块视角贡献度讨论

本节通过 3 组对照实验讨论不同视角对于模型

的贡献。通过以启用或禁用不同的视角,观察模型训练时的变化,分别验证 MPA 各个视角对于训练的有效性,实验结果如表 5 所示。

第 1 组实验通过单独启用其中一个视角所得出的分类效果较差,所训练的模型分类精度较低,其中仅启用视角 1(浅层视角)时效果最差。这是由于模型缺乏全局语义信息,而视角 1 进一步将特征分块缩小导致模型学习到大部分细节信息,这些细节信息会在特征空间中互相干扰,影响模型的训练。

第 2 组实验则是通过禁用其中一个视角观察模型分类精度以判断视角之间的重要程度,当禁用视角 1 时,模型分类精度在该组实验中最差,其次为视角 3(深层视角)。分析原因为区分不同类别的关键信息往往隐藏在小尺度当中,禁用视角 1 导致模型缺失了这些重要的细节信息,进而难以学习到具有判别性的特征。此外,视角 3 所捕捉的全局语义信息对于模型理解特征不同部位之间的互补关系也起到了关键作用,对于提高模型分类性能同样重要。

表 5　MPA 中不同视角的消融实验分析

视角_1	视角_2	视角_3	准确率/%
√	×	×	87.7
×	√	×	87.9
×	×	√	88.1
×	√	√	88.4
√	√	×	89.0
√	×	√	89.5
√	√	√	90.0

2.4.4　模块性能测试

在上一节本文探讨了 PSAB 对于递进式训练的兼容性,接下来将通过实验分别讨论 MPA 和 PDWF 模块对于算法整体性能和参数的提升情况,实验结果如表 6所示。

采用 ResNet-50 作为骨干网络时,算法准确率为 87.6%。引入 MPA 后,模型准确率提高至 90.0%,比基线网络提高了 2.0%。MPA 是 PMG 策略的改进算法,对比采用 PMG 策略,仅采用 MPA 时模型精度提高了 0.3%,验证了 MPA 的有效性。在考察 PDWF 模块有效性时,仅在骨干网络基础上实验,PDWF 提高了模型准确率至 88.0%,相比基线网络提升了 0.1%。同时采用 MPA 和 PDWF 后,模型准确率达到 90.5%,相较于基线网络提升了 2.4%。数据表明,将 MPA 和 PDWF 模块结合能够进一步增强各阶段特征交互、提升语义信息,从而提升模型分类性能。

模型复杂度可用计算复杂度和空间复杂度来衡量,分别通过浮点运算次数(floating point operations per second,FLOPs)和参数量(Param)来度量,表中采用 Speed 表示模型的运行速度。与基线模型对比,单独采用 MPA 时,模型参数量增加 14.8 M,计算量增加 40.6 G;单独采用 PDWF 模块时参数量增加 3.9 M,计算量增加 7.6 G。模型整体参数为 44.3 M,计算量为 64.7 G,速度为 2.78 $f \cdot s^{-1}$;相较于基线模型,参数量增加 18.7 M,计算量升高 48.2 G,速度降低 1.34 $f \cdot s^{-1}$。

就模块而言,MPA 较大幅度地增加了模型的计算量。这是由于 MPA 通过限制各视角感受野、注意力机制等方式来提高视角对不同尺度特征的专注度。尽管这种方法对模型参数量的增加并不显著,但却会大幅度提升训练过程中的计算需求。原因在于,卷积块必须独立处理每一个局部特征,特别是在浅层视角中,这种附加的计算量尤为突出。就模型

表 6　各模块对算法整体性能与参数提升情况

PMG[14]	MPA	PDWF	精度/%	增加量	参数量/M	增加量	计算量/G	增加量	速度/$f \cdot s^{-1}$	降低量
×	×	×	87.6	-	25.6	-	16.5	-	4.12	-
√	×	×	89.6	+2.0	45.1	19.5	40.4	23.9	2.91	1.21
×	√	×	90.0	+2.4	40.4	14.8	57.1	40.6	2.84	1.28
×	×	√	88.3	+0.7	29.5	3.9	24.1	7.6	3.47	0.65
×	√	√	90.5	+2.9	44.3	18.7	64.7	48.2	2.78	1.34

整体而言,本文算法在精度方面相比基线模型有较大提升,但也一定程度上牺牲了模型复杂度和实时性。

2.5 与现有方法进行比较

本小节将本文算法与细粒度图像分类任务中其他先进方法进行比较,表 7 列出了 3 个细粒度图像数据集的分类实验结果,并列出了一些当前主流的以 ResNet-50 为骨干网络的分类方法,结果表明,本文提出的方法要优于大部分主流细粒度图像分类方法。

表 7 本文算法与相关主流方法的分类准确率对比

出版年份	方法	主干网络	额外训练标签	CUB-200-2011/%	Stanford Cars/%	FGVC-Aircraft/%
DICTA 2018	PC-CNN[22]	VGG	√	85.1	—	—
PR 2018	Mask-CNN[23]	VGG	√	85.7	—	—
		ResNet-50		87.3	—	—
ICCV 2017	MA-CNN[7]	VGG	—	86.5	92.8	89.9
AAAI 2020	AENet[24]	DenseNet-161	—	90.0	94.5	94.0
IJCNN 2021	FBSD[25]	DenseNet-161	—	89.8	-	93.2
CVPR 2018	FT ResNet[26]	ResNet-50	—	84.1	91.7	88.5
CVPR 2019	Stacked LSTM[27]	ResNet-50	—	90.4	—	—
CVPR 2019	DCL[28]	ResNet-50	—	87.8	94.5	93.0
AAAI 2020	API-Net[29]	ResNet-50	—	87.7	94.8	93.0
ECCV 2020	PMG[30]	ResNet-50	—	89.6	95.1	93.4
AAAI 2021	DP-Net[31]	ResNet-50	—	89.3	94.8	93.9
IEEE TIP 2021	AP-CNN[32]	ResNet-50	—	88.4	95.4	94.1
PR 2022	SDNs[33]	ResNet-50	—	88.4	94.4	92.8
PR 2023	GDSMP-Net[31]	ResNet-50	—	89.9	95.3	94.4
Neural Netw 2023	FRe-Net[34]	ResNet-50	—	89.4	95.2	93.9
ICASSP 2024	HLS + PMG[35]	ResNet-50	—	89.9	—	93.5
2024	本文算法	ResNet-50	—	90.5	95.5	94.2

2.6 可视化分析

为更直观地进行展示,本小节分别对本文算法和 ResNet-50 骨干网络的最后 3 个阶段进行可视化实验。如图 5 所示,图左侧展示了本文算法中浅层、中层和深层 3 个视角的特征输出可视化,而图右侧对应于基线模型 ResNet-50 最后 3 个阶段的特征输出可视化。视角 1 对应浅层模型阶段,可以观察到该阶段的模型主要关注一些粒度较小的特征区域,如鸟的眼睛、鸟喙以及鸟羽毛上的图案或纹理。视角 3 中对应深层模型,可以发现该阶段关注到了鸟头、鸟身以及其他较大范围的判别性特征区域。同时,通过观察发现,在 3 个阶段中,本文模型对背景无关特征的关注明显降低,相较于基线模型,具有更强的目的性和准确性。对比图中结果可以看出,本文模型通过多视角注意力模块在早期阶段有效地屏蔽了大量无关特征信息,并挖掘出关键特征信息。可视化结果表明,本文算法能够在不同阶段准确地预测不同粒度大小的特征,为模型后续多粒度特征学习提供有力支持。

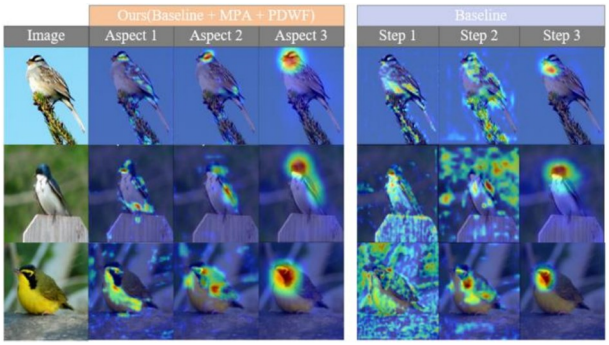

图 5 本文模型和基线模型在 CUB 上的特征激活效果图

3 结 论

本文提出了一种多视角解耦增强整合的细粒度分类算法。通过将模型分解为多个视角,鼓励网络对不同的粒度进行针对性地学习,尽可能多地捕捉关键性特征并屏蔽无关背景特征。本文设计了多视角注意力模块所分配的视角自注意力模块,它能在局部特征中获取重要关键信息并屏蔽无关信息的同时,尽量保留细微特征。为了更充分地利用网络学习到的粒度信息,本文引入了递进式动态加权融合模块,旨在整合不同阶段输出的不同尺度特征,以丰富特征的互补语义信息。

在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 这 3 个常用公开数据集上进行了实验评估。结果显示,本文方法在这些数据集上分别取得了 90.5%、95.5% 和 94.2% 的分类准确率,优于基线方法以及其他先进方法的表现。此外,本文还进行了消融实验,验证了所设计的各个模块的合理性和有效性;论证了通过将图像的不同尺度分解为各个视角进行专项训练有利于降低姿态、光照等各类因素互相耦合带来的负面影响。

参考文献

[1] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[C]// The 13th European Conference on Computer Vision. Zurich, Switzerland:Springer, 2014:834-849.

[2] VIOLA P, JONES M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57:137-154.

[3] WU J, LIU N, GEYER C, et al. A real-time object detection framework[J]. IEEE Transactions on Image Processing, 2013,22(10):4096-4107.

[4] XIAO T, XU Y, YANG K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015:842-850.

[5] PENG Y, HE X, ZHAO J. Object-part attention model for fine-grained image classification[J]. IEEE Transactions on Image Processing, 2017,27(3):1487-1500.

[6] ZHENG H, FU J, ZHA Z J, et al. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 5012-5021.

[7] HU T, QI H, HUANG Q, et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification [J]. (2019-03-23) [2024-04-16]. https://arxiv.org/pdf/1901.09891.

[8] ZHENG H, FU J, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017:5209-5217.

[9] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. Boston, USA: IEEE, 2015:1449-1457.

[10] FU J, ZHENG H, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Venice, Italy: IEEE, 2017:4438-4446.

[11] ZHANG L, HUANG S, LIU W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA: IEEE, 2019:8331-8340.

[12] YU C, ZHAO X, ZHENG Q, et al. Hierarchical bilinear pooling for fine-grained visual recognition[C]// The 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018:574-589.

[13] YE Z, HU F, LIU Y, et al. Associating multi-scale receptive fields for fine-grained recognition [C] // 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, Arabs: IEEE, 2020:1851-1855.

[14] DU R, CHANG D, BHUNIA A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]// The 16th European Conference on Computer Vision. Online: Springer, 2020: 153-168.

[15] FU J, LIU J, TIAN H, et al. Dual attention network for

scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:3146-3154.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017,30:6000-6010.

[17] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Online: IEEE, 2021:3560-3569.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:7132-7141.

[19] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset[EB/OL]. [2024-04-16]. https://www.semanticscholar.org/paper/The-Caltech-UCSD-Birds-200-2011-Dataset-Wah-Branson/c069629a51f6c1c301eb20ed77bc6b586c24ce32?p2df.

[20] KRAUSE J, STARK M, DENG J, et al. 3d object representations for fine-grained categorization[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013:554-561.

[21] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. (2013-06-21) [2024-04-16]. https://arxiv.org/pdf/1306.5151.

[22] LIAO Q, HOLEWA H, XU M, et al. Fine-grained categorization by deep part-collaboration convolution net[C]//2018 Digital Image Computing: Techniques and Applications (DICTA). Canberra, Australia: IEEE, 2018:1-8.

[23] WEI X S, XIE C W, WU J, et al. Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018,76:704-714.

[24] HU Y, LIU X, ZHANG B, et al. Alignment enhancement network for fine-grained visual categorization[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021,17(1):1-20.

[25] SONG J, YANG R. Feature boosting, suppression, and diversification for fine-grained visual classification[C]//2021 International Joint Conference on Neural Networks. Shenzhen, China: IEEE, 2021:1-8.

[26] WANG Y, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:4148-4157.

[27] GE W, LIN X, YU Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:3034-3043.

[28] CHEN Y, BAI Y, ZHANG W, et al. Destruction and construction learning for fine-grained image recognition [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA: IEEE, 2019: 5157.

[29] ZHUANG P, WANG Y, QIAO Y. Learning attentive pairwise interaction for fine-grained classification[C]//AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020:13130-13137.

[30] WANG S, LI H, WANG Z, et al. Dynamic position-aware network for fine-grained image recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2021: 2791-2799.

[31] DING Y, MA Z, WEN S, et al. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification[J]. IEEE Transactions on Image Processing, 2021,30:2826-2836.

[32] ZHANG L B, HUANG S L, LIU W. Learning sequentially diversified representations for fine-grained categorization[J]. Pattern Recognition, 2022,121:1.

[33] KE X, CAI Y H, CHEN B T, et al. Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification[J]. Pattern Recognition, 2023,137:109305.

[34] ZHAO P, LI Y, TANG B, et al. Feature relocation network for fine-grained image classification[J]. Neural Networks, 2023,161:306-317.

[35] ZHANG S, ZHENG S, SHUI Z, et al. HLS-FGVC: hierarchical label semantics enhanced fine-grained visual classification[C]//The IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea: IEEE, 2024:7370-7374.

Multi-perspective decoupling enhancement and integration for fine-grained classification

MENG Yuebo, WANG Bo, LIU Guanghui
(College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055)
(Xi'an Key Labratory of Building Manufactaring Intelligent & Automation Technology, Xi'an 710055)

Abstract

To address the significant intra-class variation caused by external factors such as background environment, lighting conditions, sample posture, and shooting angle in fine-grained image classification, this paper proposes a fine-grained classification algorithm based on multi-perspective decoupling enhancement integration. Firstly, to reduce the interference of external factors in images, a multi-perspective attention (MPA) module is designed. This module decomposes the model into several perspectives, forcing each perspective to focus on different scales, thus decoupling the interference factors. By modeling features with self-attention, each perspective is guided to further mine key features. Secondly, a progressive dynamic weighted fusion (PDWF) strategy is proposed to effectively integrate the decoupled multi-perspective information. This strategy dynamically adjusts the fusion coefficient by acquiring channel and spatial relationships from different perspectives, achieving high-order fusion of multi-scale information. Lastly, a progressive training method is adopted to facilitate perspective interaction, further capturing and integrating complementary semantic information from multi-scale features. Experiments are conducted on three public datasets, CUB-200-2011, Stanford-Cars, and FGVC-Aircraft, and the results show that the proposed method achieves classification accuracy rates of 90.5%, 95.5%, and 94.2%, respectively, which outperforms current mainstream methods for fine-grained image classification tasks.

Key words: fine-grained, multi-perspective attention (MPA), progressive dynamic weighted fusion (PDWF), image classification