DOI:10.19651/j.cnki.emt.2517840

基于改进的 CVT 细粒度图像识别算法研究*

冀得魁[1,2] 李冰锋[1,2] 杨 艺[1,2]
(1.河南理工大学电气工程与自动化学院 焦作 454000;2.河南省煤矿装备智能检测与控制重点实验室 焦作 454003)

摘 要:针对细粒度图像背景信息对目标区域干扰以及目标最具判别区域特征难以辨识的问题,本文提出了一种基于改进的 CVT 细粒度图像识别算法研究。首先,在 CVT 模型中引入目标区域定位模块,该模块通过多层次的特征聚合方法提取目标区域的特征,并通过阈值判定方式进行目标区域的确定,之后对原始图像进行等比例裁剪,以减少背景信息的干扰。其次,提出了 MDCSAIA 机制,采用维度转换的方法,促进通道位置相邻的空间信息和空间位置相邻的通道信息间的有效交互,从而增强网络对目标局部细节区域的感知能力。实验结果表明,与基线算法相比,该方法在 CUB-200-2011、Stanford Cars 和 Stanford Dogs 三个数据集上的识别准确率分别提高了 2.1%、1.7%和 1.5%。此结果验证了所提方法的有效性。

关键词:细粒度图像识别;Transformer;目标区域定位;MDCSAIA 机制;热力图
中图分类号:TP391.4;TN791 **文献标识码**:A **国家标准学科分类代码**:520.60

Investigation into an enhanced CVT-based algorithm for fine-grained image recognition

Ji Dekui[1,2] Li Bingfeng[1,2] Yang Yi[1,2]
(1. School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China;
2. Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo 454003, China)

Abstract: In response to the issues of background interference in fine-grained images and the challenge of identifying the most discriminative features in the target region, this paper proposes an improved CVT-based fine-grained image recognition algorithm. First, a target region localization module is introduced into the CVT model. This module extracts features of the target region using a multi-level feature aggregation method and determines the target region via threshold-based decision-making. The original image is then cropped proportionally to reduce the interference of background information. Furthermore, a mechanism called MDCSAIA (Multi-Dimensional Channel Spatial-Aware Interaction) is proposed. This mechanism employs dimensional transformation to facilitate effective interaction between spatial information of adjacent channels and channel information of adjacent spatial positions, thereby enhancing the network's ability to perceive the local details of the target region. Experimental results show that, compared to baseline algorithms, the proposed method improves recognition accuracy by 2.1%, 1.7%, and 1.5% on the CUB-200-2011, Stanford Cars, and Stanford Dogs datasets, respectively. These results validate the effectiveness of the proposed approach.

Keywords: fine-grained image classification; Transformer; object region localization; MDCSAIA mechanism; heatmap

0 引 言

细粒度图像识别[1]近年来成为计算机视觉和模式识别领域的热门研究课题,其主要任务是对同一类别中的不同子类进行更细致的区分。在日常生活中,细粒度图像识别有广泛的应用,例如植物和动物识别、车辆识别、医疗诊断等。相比传统的图像识别,细粒度图像识别面临以下挑战:首先,不同子类之间的物种差异微小,难以区分;其次,相同子类在形态、姿势、背景等因素的影响下存在显著差异。因此,如何减少背景区域信息的干扰、并有效提取目标区域的最具判别力的特征,成为当前细粒度图像识别领域的关键难题。

收稿日期:2025-01-07
*基金项目:河南省科技攻关项目(222102210230)、河南理工大学博士基金(B2018-33)项目资助

近年来,卷积神经网络[2-5](convolution neural network, CNN)凭借其在提取图像局部特征方面的优势性能,推动了基于卷积的分类算法[6-7]在细粒度图像识别研究领域中的广泛应用,并逐渐成为该领域的主要研究方法。细粒度图像识别方法主要分为强监督和弱监督两大类。强监督方法依赖于人工标注关键区域用来训练目标检测网络,进而利用特征提取网络进行目标分类,从而提高网络模型对目标关键区域的识别精度。然而,在处理大规模数据集或复杂分类任务时,获取精确标注信息的成本通常较高。相比之下,弱监督图像识别方法仅依赖于图像的类别标签进行模型训练,减少人工标注的介入,因此已成为细粒度图像识别的主要研究方向。

例如,王培森等[8]提出一种多通道视觉注意力机制的方法,通过生成多个视觉注意力图来描述物体的不同区域对应关系,并进一步提取高阶统计特征,以获取相应的视觉表示。然而,该方法不能准确识别目标区域位置,容易受背景区域信息的干扰,导致目标区域的准确识别受到影响。基于此,刘万军等[9]提出了一种结合前景特征增强与区域掩码自注意力的分类方法,该方法通过前景特征增强网络精确定位前景目标所在位置,在消除背景信息的同时增强对目标区域的关注,最后通过区域掩码自注意力网络捕获目标区域中区别于其他子类的特征信息,从而提升分类性能。然而,该方法仅关注单张图像显著局部区域提取,未充分考虑多张图像间的局部异构语义关系,导致所学习的特征缺乏足够的区分性。随后,陈权等[10]提出了一种融合目标定位与异构局部交互学习的分类方法,该方法不仅针对单张图像进行语义目标的集成定位,还设计了一个多图像异构局部交互图模块,以提取每张图像的显著局部区域特征。这些方法均有效提升了细粒度图像的识别精度,展现了优异的识别性能。

由于 CNN 在特征提取时受限于感受野,其处理范围仅局限于图像的局部区域,且通过逐层结构提取特征,导致对图像全局特征信息的关注不足。为了解决这一问题,Dosovitskiy 等[11]提出了 Vision Transformer (ViT)架构,为图像分类任务提供了一种新颖的方法。ViT 通过自注意力机制处理全局特征,建立图像块之间的长距离依赖关系,使得网络能够有效地捕获图像中的全局上下文信息,从而显著提升细粒度图像的分类准确率。然而,ViT 架构中的自注意力机制需要计算各图像块之间的相似性,导致计算复杂度过高,对硬件性能要求较高。为解决此问题,Wu 等[12]提出了卷积视觉 Transformer(convolutional vision transformer,CVT),采用层次化架构,逐层缩小特征层的空间维度并提取多层次特征,显著降低了特征图的分辨率和数量。此外,CVT 引入了卷积层代替传统的线性投影,以更高效的方式提取局部特征。该方法显著降低了网络的计算复杂度和运算量,同时进一步提升了模型的性能。

然而,CVT 模型仍存在一些局限性。首先,尽管引入了卷积层以代替线性投影降低了计算复杂度,但卷积层固定感受野的局限性限制其在提取不规则目标区域特征时的表现能力,无法准确捕获复杂结构的细节信息。此外,该模型采用的全局自注意力机制仅考虑了空间维度的信息,忽略了其他维度的特征,从而限制了模型在目标区域细粒度特征提取方面的能力,影响其对多维度特征的全面感知和学习。

针对此问题,本文基于 CVT 网络架构,提出一种基于改进的 CVT 细粒度图像识别的研究方法(research method on improved CVT for fine-grained image recognition, IMCVT)。该方法引入目标区域定位(target region localization, TRL)模块,采用基于阈值判别的多层次特征聚合技术进行目标区域的准确定位,突出目标区域的细节信息,并有效抑制输入样本中非目标背景区域的干扰,增强网络模型的目标定位和识别性能。此外,为了更高效地提取目标最具判别力区域的细粒度特征,提出一种多维跨空间自适应交互注意力(multi-dimensional cross-space adaptive interaction attention mechanism, MDCSAIA)机制。该注意力机制不仅捕捉每个通道空间相邻位置之间的依赖关系,还实现每个空间位置相邻通道之间的信息交互,从而增强模型对目标局部特征的感知能力。此外,本文引入平均池化和标准差池化自适应组合的方法,以更有效地聚合跨维度特征响应,进一步提升网络在目标区域细粒度特征提取中的表现。最终,通过在 3 个公共细粒度数据集上的实验验证,结果表明,本文提出的方法具有良好的分类性能。

1 本文方法

1.1 IMCVT 模型整体框架

IMCVT 的整体网络架构如图 1 所示。其主干网络由三阶段金字塔结构的子网络组成,每个子网络包括卷积令牌嵌入、令牌序列处理和卷积 Transformer 模块 3 个主要部分。

首先,输入的样本图像经过卷积令牌嵌入层(convolutional token embedding,CTE)进行处理,以提取相同尺寸的局部特征图。随后,将此特征图展平为令牌序列,以便与 Transformer 编码器兼容,从而利用自注意力机制关注图像全局特征信息。接着,将展平后的令牌序列重塑为二维令牌图,并输入到卷积 Transformer 模块中,以进一步提取图像的空间全局特征。该卷积 Transformer 模块包括卷积映射层、多头自注意力模块和多层感知机(MLP),其内部结构如图 2 所示。

卷积映射层采用深度可分离卷积替代传统的位置编码方式,从而有效降低了多头自注意力操作的计算复杂度。多头自注意力子模块用于捕捉令牌之间的长距离依赖关系,增强模型对全局特征的学习能力。此外,在网络的第三阶段,将分类令牌 X_class 添加到令牌序列首部,以整合全

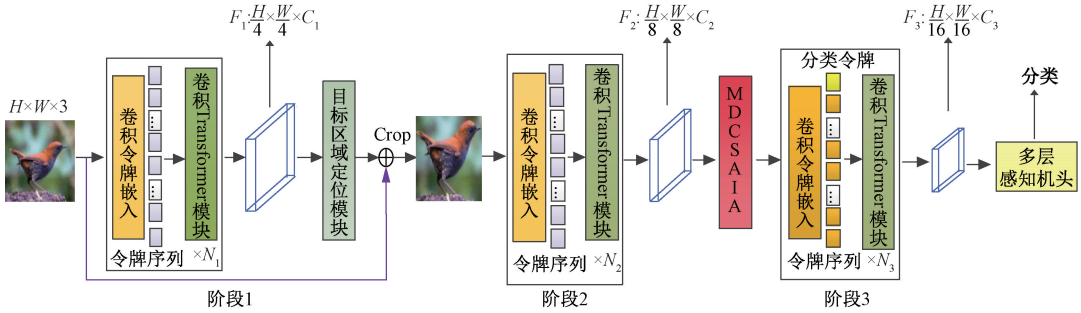

图 1　IMCVT 网络架构图

Fig. 1　IMCVT network architecture diagram

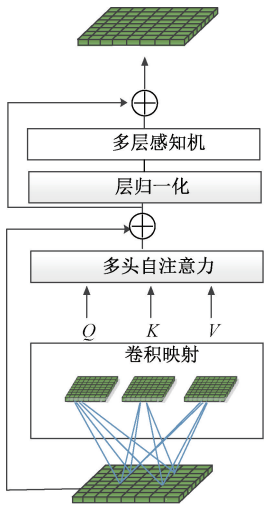

图 2　多层感知机结构图

Fig. 2　Multi-layer perceptron structure diagram

局特征进行最终的分类任务，而在第一、二阶段不添加分类令牌。最终，通过 MLP 层对网络输出的特征进行处理，生成最终的分类标签。

本文的网络模型基于 CVT 架构，在网络的第一阶段引入了目标区域定位模块，通过对目标特征区域进行提取，并将其与原始样本图像进行裁剪，从而有效减少背景信息对目标区域的干扰。此外，为进一步增强网络对目标区域内局部细节的关注，并弥补 Transformer 自注意力机制在通道特征信息交互方面的不足，在第二阶段后引入了多维跨空间自适应交互注意力（multi-dimensional cross-space adaptive interaction attention，MDCSAIA）机制，该机制通过捕捉通道空间相邻位置之间的依赖关系以及空间位置相邻通道之间的信息交互，从而提升网络对细粒度图像中最具判别力区域的识别性能，达到更好的分类效果。

1.2　目标区域定位模块

在细粒度图像分类任务中，复杂背景信息往往会干扰网络对图像目标区域的识别，进而影响分类性能。因此，准确识别并定位样本图像中的目标区域对细粒度分类至关重要。为减少复杂背景信息对细粒度分类任务的影响，本文采用基于文献[13]所提出的多层次特征聚合的策略获取相应样本图像的目标区域特征信息，再利用阈值判断的方法来判定特征图中目标区域，最后将判定目标区域与样本图像做等比例裁剪处理，减少背景噪声对分类任务的干扰，从而提升网络模型对细粒度图像的分类精度。该目标区域定位模块的网络结构如图 3 所示。

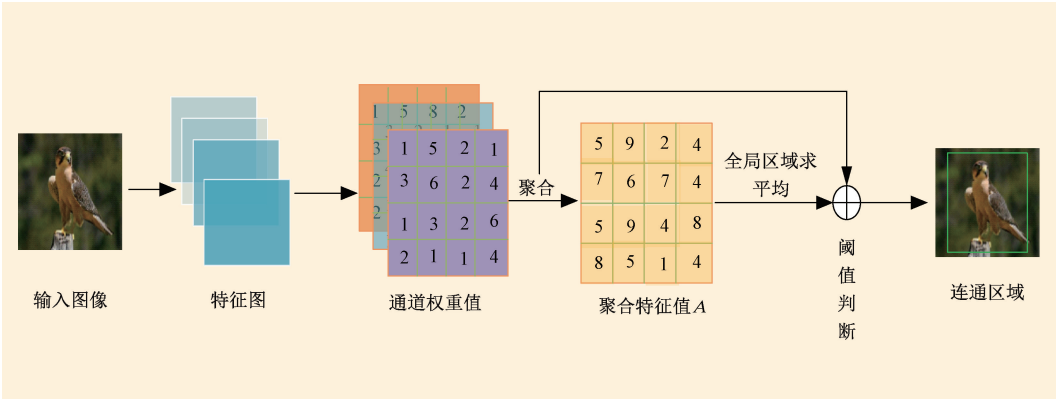

图 3　目标区域定位结构图

Fig. 3　Target region localization structure diagram

1）如图 3 所示，首先，输入图像通过 CVT 第一阶段的子网络模型进行提取特征，该阶段模型通过卷积神经网络

与视觉 Transformer 相结合,提取图像的高层次语义,并将最后 Block 块中的输出特征向量 $\boldsymbol{X}_1$,作为目标区域定位模块的输入特征。

2)然后,将该输入特征图 X_1 在通道维度进行加权求和,得到聚合特征图 Q。根据 Q 各通道特征值计算不同样本图像区域对应的聚合特征值 A。如式(1)所示,其中 f_i 表示对应通道的第 i 个特征图。

$$Q = \sum_{i=0}^{C-1} f_i \tag{1}$$

基于训练数据的统计特性,对聚合特征图 Q 在空间维度上进行全局平均池化操作,以获取该空间区域内平均值。将全局空间区域的平均值设定为阈值,表示为 $\overline{a}$,将其与聚合特征值 $Q(x, y)$进行比较,若 $Q(x, y)>\overline{a}$,将其视为样本图像的目标区域,反之则视为背景区域。式(2)中 $\overline{a}$ 表示计算所得的空间池化平均值,式(3)确定空间目标区域位置所对应二值化区域,其具体操作公式如下:

$$\overline{a} = \frac{\sum_{x=0}^{W-1}\sum_{y=0}^{H-1} Q(x,y)}{H \times W} \tag{2}$$

$$\tilde{M}(x,y) = \begin{cases} 1, & Q(x,y) > \overline{a} \\ 0, & \text{其他} \end{cases} \tag{3}$$

此外,在获取目标区域后,使用最大连通域分析对该区域进行处理,以获得该 Block 块下输出特征图的最大连通区域,最后依据最大连通区域裁剪原始样本图像,减少背景噪声对分类任务的干扰,从而提高分类准确性。

TRL 模块通过聚合不同层级的特征信息来提取样本目标区域特征,并通过空间区域阈值判断来定位目标区域。这种方法在一定程度上减少了与分类任务无关的背景干扰,增强了网络模型对样本图像目标区域的感知能力,从而提升了模型对细粒度图像的识别精度。

1.3 注意力机制

1)视觉注意力分析

由于细粒度图像各子类别间的特征差异极为细微,导致其难以区分,从而影响了网络模型的分类精度。而视觉注意力机制是一种有效的解决方法,视觉注意力机制按类型通常分为通道和空间注意力两大类。通道注意力主要通过对特征图的通道施加不同的权重,关注权重值更高的通道,减少对无关通道的关注。例如,SE-Net[14] 通过全局平均池化有效聚合特征图的二维全局信息,并通过压缩和扩展操作建立通道之间的相互依赖关系,但未能实现空间特征信息的交互。随着注意力机制的发展,CBAM[15] 通过自适应学习特征图中通道和空间注意力权重,不仅构建了通道依赖关系,而且引入空间位置编码信息,有效实现了空间信息和通道信息的交互。但处理需要跨通道或空间维度交互的特征时仍然会有所局限。基于此问题,Wah 等[16] 提出了多维协作注意力机制,该机制强调不同维度特征之间的协同作用,整合各维度信息,获得更全面的特征表示。尽管该方法有效提升了网络对细微目标特征的表达能力,但仍未显著解决空间相邻通道信息的交互问题,并且缺乏通道相邻空间位置信息间的依赖关系。

2)MDCSAIA 机制

为了进一步增强网络模型对目标区域最具判别力特征的识别能力,本文提出了一种多维跨空间自适应交互注意力机制,该机制采用多分支网络架构:左侧分支负责捕捉空间位置相邻通道之间的依赖关系,实现跨空间的通道信息融合。右侧分支则关注样本图像中通道相邻空间位置之间的依赖关系,实现跨通道的空间信息融合,从而提高网络对局部区域细节的感知能力。此外,中间通道维度分支采用自适应压缩变换机制,有效利用平均池化和标准差池化的方法,进一步提取特征,并通过 Sigmoid 激活函数获得通道权重信息,最后与原始输入特征图相乘以实现通道加权。最终,将各分支输出结果融合,提升网络模型的权重分配性能和目标最具判别力区域的识别能力。图 4 展示了该注意力机制的整体网络架构,其详细原理解释如下:

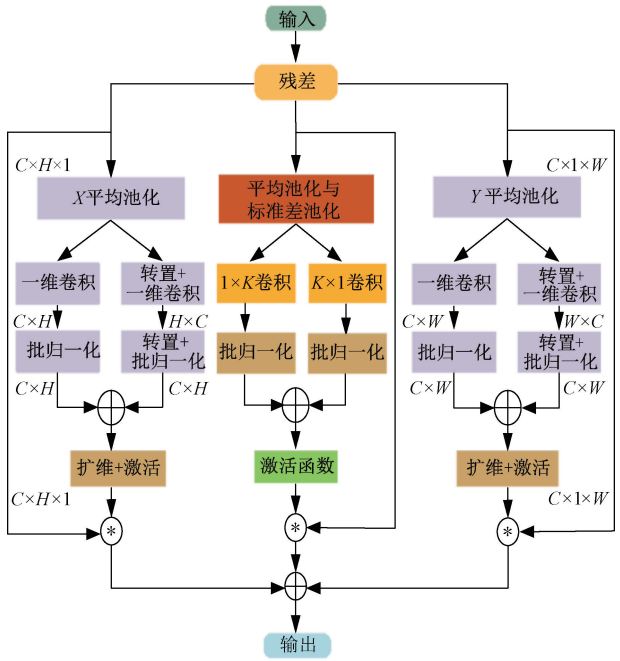

图 4 MDCSAIA 机制网络架构图

Fig. 4 MDCSAIA mechanism network architecture diagram

(1)首先,该注意力机制对输入特征图 $X \in R^{C\times H\times W}$ 沿着空间维度的高度方向进行平均池化,这将减小输入特征图的高度,但不改变其宽度。第 C 个通道在高度方向第 h 位置的特征表达如式(4)所示。

$$Z_c^h(h) = \frac{1}{W}\sum_{i=1}^{W} x_c(h,i) \tag{4}$$

同样地,对输入特征图沿空间维度的宽度方向进行平

均池化，第 C 个通道在宽度方向第 w 位置的特征表达如式(5)所示：

$$Z_c^w(w)=\frac{1}{H}\sum_{j=1}^{H}x_c(j,w) \tag{5}$$

在图像特征提取过程中，通过式(4)和(5)计算生成两个具有方向感知能力的特征向量，将其表示为 $Z(h)\in R^{C\times H}$、$Z(w)\in R^{C\times W}$。然而，空间邻域信息在单个通道内部具有较高的相关性，而将不同通道的空间邻域信息进行融合可能会引入噪声，从而影响空间邻域依赖关系的构建。

为减少不同通道间空间邻域信息的交互，本文采用了一种基于分组一维卷积的方法，对特征向量 $Z(h)\in R^{C\times H}$ 和 $Z(w)\in R^{C\times W}$ 进行通道维度的特征提取，并将分组数设置为与通道数相匹配。具体而言，使用一维分组卷积和批归一化(BatchNorm)处理特征向量，详细计算过程如式(6)所示。

$$\begin{cases}F_S(h)=f_s^h(Z(h))\\F_S(w)=f_s^w(Z(w))\end{cases} \tag{6}$$

其中，f_s^h 和 f_s^w 分别表示在空间高度和宽度方向上的一维分组卷积和批归一化操作，$F_S(h)$ 和 $F_S(w)$ 表示其对应运算结果，这些操作描述了通道相邻空间信息的交互。

(2)此外，为了捕捉空间区域内相邻通道之间的特征依赖关系，本文对特征向量 $Z(h)\in R^{C\times H}$ 和 $Z(w)\in R^{C\times W}$ 进行维度转置操作，并沿着高度方向和宽度方向维度分别执行一维分组卷积。分组卷积特征提取的运算过程如式(7)所示。

$$\begin{cases}F_C(h)=f_c^h(Z(h)^{\mathrm{T}})\\F_C(w)=f_c^w(Z(w)^{\mathrm{T}})\end{cases} \tag{7}$$

其中，$\boldsymbol{Z}(h)^{\mathrm{T}}$ 和 $\boldsymbol{Z}(w)^{\mathrm{T}}$ 分别表示对特征向量 $Z(h)$ 和 $Z(w)$ 的转置操作；f_ch 和 f_cw 表示对转置后的特征向量在通道方向上进行的一维分组卷积和批归一化运算。卷积运算的结果分别表示为 $F_C(h)$ 和 $F_C(w)$，描述了空间相邻通道信息的交互。随后，将空间邻域信息的交互结果与通道邻域信息的交互结果进行融合，并使用 Sigmoid 函数激活，之后在空间方向进行扩展，最后与初始特征进行点乘操作，其计算过程如式(8)所示。

$$\begin{cases}P^H=\sigma(F_S(h)\oplus F_c(h))\otimes X\\P^W=\sigma(F_S(w)\oplus F_c(w))\otimes X\end{cases} \tag{8}$$

(3)在该注意力模块中间分支部分，结合了平均池化和标准差池化，以同时获取局部区域的平均值和变化状况，从而提供更全面的特征描述。随后，通过自适应组合机制生成按通道划分的特征向量，其过程如式(9)所示。

$$F_c=\frac{1}{2}\otimes(F_{avgc}\oplus F_{stdc})\oplus\alpha\otimes F_{avgc}\oplus\beta\otimes F_{stdc} \tag{9}$$

该自适应机制引入了 α 和 β 两个可训练的浮动参数，其值范围为 0 至 1 之间，并通过随机梯度下降进行优化。其中 F_{avgc} 和 F_{stdc} 分别表示平均池化和标准差池化的输出结果。该机制基于输入特征的动态性，在不同的特征提取阶段灵活调整两者的权重分配，使网络可以优先关注关键特征区域，从而提升整体特征提取的准确性和有效性。随后，采用 $1\times K$ 和 $K\times 1$ 二维卷积对水平和垂直方向上的各区域位置进行特征提取，并将所得特征在通道维度上进行叠加。最后，执行归一化和激活函数处理，并与原输入特征进行点乘运算，增强模型的表达能力，其计算表达式如式(10)和(11)所示。

$$\begin{cases}Q_S(w)=(Conv_{1\times K}(X))\\Q_S(h)=(Conv_{K\times 1}(X))\end{cases} \tag{10}$$

$$Q^S=\sigma(BN(Q_S(h)\oplus Q_S(w)))\otimes X \tag{11}$$

$$Q^S=Q^S\oplus P^H\oplus P^W \tag{12}$$

其中，式(11)中 σ 表示 Sigmoid 激活函数，BN 代表批归一化处理，$Q_S(w)$ 代表水平方向的输出特征，而 $Q_S(h)$ 代表垂直方向的输出特征。最终，整合各分支所得输出特征，如式(12)所示，从而提升对局部判别区域特征提取的准确性和有效性。

2　实验设计与结果分析

2.1　实验数据集和评价指标

为了公平地验证本文方法的性能，所有实验均在以下 3 个公共细粒度图像数据集上进行：CUB-200-2011[17]、Stanford-Cars[18]、Standford-Dogs[19]。其中，Standford-Dogs 数据集包含 120 种不同狗类的 20 580 张图像，CUB-200-2011 数据集包含 200 种不同鸟类的 11 788 张图像，而 Stanford-Cars 数据集包含 196 种不同汽车类型的 16 185 张图像。这些数据集的类别及训练集、测试集的详细划分情况如表 1 所示。

表 1　细粒度图像数据集详细信息

Table 1　Detailed information of fine-grained image datasets

数据集	类别	训练集	测试集
CUB-200-2011	200	5 994	5 794
Stanford-Dogs	120	12 000	8 580
Stanford-Cars	196	8 144	8 041

同时，本文使用通用的准确率(accuracy，Acc)、召回率(recall，R)、精确率(precision，P)来评估所提方法的分类效果，其计算公式如下：

$$Acc=\frac{TP+TN}{TP+FP+TN+FN} \tag{13}$$

$$R=\frac{TP}{TP+FN} \tag{14}$$

$$P=\frac{TP}{TP+FP} \tag{15}$$

其中，TP、TN、FP、FN 分别表示真正例(true

positive)、真负例(true negative)、假正例(false positive)和假负例(false negative)的样本数量。

2.2 实验配置与数据预处理

本文的网络架构基于 Pytorch 实现,所有实验均在配置有 Ubuntu18.04.6LTS 操作系统的服务器上进行,该服务器配备了一块 12 GB 显存的 NVIDIA GeForce RTX 3080 显卡以及一款 Intel® Core™ i7-12700 的处理器。软件环境包括:Cuda12.0 版本,Cudnn8.8.0 版本,以及 Python3.9.13 版本。

IMCVT 采用 CVT 网络作为分类任务的骨干架构,以提取输入图像的特征信息。为了避免 IMCVT 模型在训练过程中由于计算量过大而导致难以收敛,所有样本图像均被调整为 224×224 的统一尺寸,并使用 ImageNet-1K 的预训练权重作为 IMCVT 网络模型训练的初始值,以便模型能够更好的完成训练。模型参数优化采用 Adam 方法,该方法具有解耦权重衰减,动量(Momentum)设置为 0.9,批量大小(Batchsize)设定为 32,以提高计算效率,迭代次数(epoch)设定为 100。训练阶段学习率初始值设定为 0.001,并使用余弦退火(Cosine Annealing)策略来调整学习率。此外,通过水平翻转、颜色变换等数据增强技术,增加了样本数据的多样性,以提升模型的泛化能力和鲁棒性,样本数据的展示如图 5 所示。

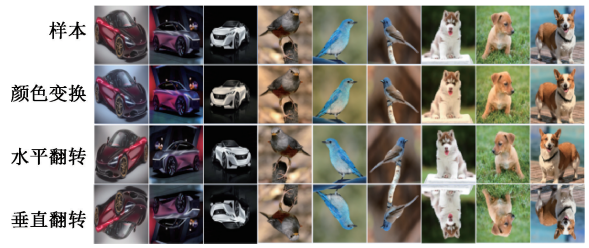

图 5 样本增强效果展示图

Fig. 5 Sample augmentation effect diagram

2.3 消融实验

为验证本文所提出的 TRL 模块和 MDCSAIA 机制的有效性,本文在 CUB-200-2011、Stanford-Cars、Standford-Dogs 数据集上进行了消融实验。实验结果如表 2 所示。

表 2 IMCVT 消融实验分析

Table 2 Ablation experiment analysis of IMCVT

模型	准确率/%			参数量/M
	CUB	Dogs	Cars	
CVT	80.3	87.8	89.5	**20.21**
CVT+TRL	81.1	88.4	90.2	20.23
CVT+MDCSAIA	80.9	88.7	90.8	20.30
IMCVT	**82.4**	**89.3**	**91.2**	20.51

在表 2 中,原始 CVT 网络模型在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 数据集上的分类准确率分别为 80.3%、87.8%和 89.5%。在该基准网络模型中嵌入 TRL 模块后,减少了背景信息的干扰,增强了模型对样本图像空间目标区域的感知能力,使分类准确率在 3 个数据集上分别提升至 81.1%、88.4%和 90.2%,提高了 0.8%、0.6%和 0.7%。当仅加入 MDCSAIA 模块时,模型对目标局部关键区域的识别能力得以增强,使分类准确率在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 数据集上分别提升至 80.9%、88.7%和 90.8%,较基准模型分别提升了 0.6%、0.9%和 1.3%。本文提出的 IMCVT 网络实验结果表明,将 TRL 模块和 MDCSAIA 模块串行组合基准网络中,可进一步提升模型的性能,最终在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 数据集上的分类准确率分别达到 82.4%、89.3%和 91.2%,显著提高了模型的分类性能。

2.4 注意力对比实验

为了进一步验证本文所提注意力机制的优越性,本文将 SE-Net、CBAM-Net、EMA-Net[20]、MCA-Net 分别嵌入 CVT 模型,并在 Standford-Dogs 公共数据集上与本文所提出的 IMCVT 算法进行对比实验。实验结果如表 3 所示。

表 3 注意力模型实验结果对比

Table 3 Comparison of experimental results of attention models

模型	参数量/M	计算量/M	准确率/%	召回率/%	精确率/%
			Standford-Dogs		
CVT	20.21	16 220.8	87.8	84.6	82.5
CVT+TRL+SE-Net	20.25	16 221.5	88.3	85.3	83.0
CVT+TRL+EMA-Net	20.34	16 244.2	89.0	85.6	83.6
CVT+TRL+CBAM-Net	20.28	16 238.7	88.6	85.0	83.2
CVT+TRL+MCA-Net	20.47	16 326.1	89.1	86.1	83.7
IMCVT	20.51	16 367.3	**89.3**	**86.6**	**84.0**

表 3 展示了本文模型 IMCVT 在 Standford-Dogs 数据集上与其它算法的实验对比结果。实验结果表明,MDCSAIA 机制的分类准确率明显高于现有的大多数主流注意力机制,展示了最先进的性能。例如,本文引入的

MDCSAIA 机制在减少参数量的同时，显著提升了分类精度，相比于 MCA-Net 和 EMA-Net 有显著提升。相比之下，SE-Net 和 CBAM-Net 使用不同数量的卷积核来学习通道间的重要性，但无法组合空间区域相邻的通道信息，而本文提出的注意力机制可以实现空间位置相邻通道间的信息交互，并捕捉通道相邻空间位置之间的依赖关系。

此外，引入 MDCSAIA 机制和 TRL 模块后的 IMCVT 算法能有效学习样本图像中目标区域局部关键特征，相较于基准骨干网络，该算法在 Standford-Dogs 细粒度图像数据集上的分类精度提升了 1.5%，且参数量增加较小，显著提升了网络模型的分类性能。

2.5　损失函数对比

通过与原 CVT 模型在训练过程中损失曲线的对比，如图 6 所示，可以明显观察到，本文方法在损失下降的速度和模型收敛的效率上均优于原模型。这一结果进一步验证了本文提出模型的优越性。

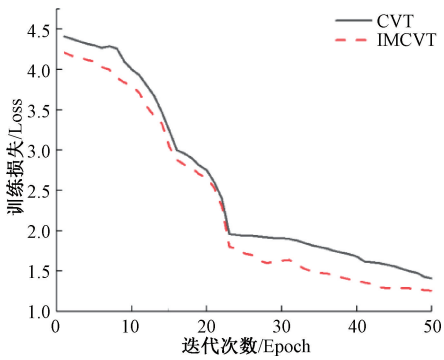

图 6　损失函数对比图

Fig. 6　Comparison of Loss Functions

2.6　IMCVT 可视化

为了验证本文方法对网络特征提取性能的影响，本文在三大公共数据集下的细粒度图像子类中对特征提取网络的输出层特征进行了可视化，并与基准网络 CVT 的可视化结果进行了对比。如图 7 所示，本文算法生成的热力图在显著性区域上更加全面，主要集中在具有更高辨别性的局部位置。例如鸟类的热力图在嘴部和头部包含较多的权重；狗类的关注点大多聚集在眼部和头部区域；而车辆的关注点则主要集中在车灯和车镜等部位。由此证明，该算法在学习细粒度图像的关键区域和捕获细节特征方面表现更佳。

图 7　特征图可视化结果

Fig. 7　Feature map visualization results

3　结　　论

本文提出了一种基于改进的 CVT 细粒度图像识别的研究方法，该算法利用目标定位模块，通过集中聚合样本图像各特征层间目标区域特征信息，并根据聚合值与阈值比较来判定特征图中目标区域位置，再对原样本图像进行等比例裁剪。此外，本文提出了 MDCSAIA 机制，捕捉空间位置相邻通道之间的依赖关系，同时实现通道相邻空间位置之间的信息交互，从而进一步提升网络对目标关键区域细粒度特征的识别能力。实验结果表明，IMCVT 网络在 3 个细粒度图像数据集上均表现出优异的识别性能。

参考文献

[1]　ZHAO B, FENG J, WU X, et al. A survey on deep learning-based fine-grained object classification and semantic segmentation[J]. International Journal of Automation and Computing, 2017, 14(2): 119-135.

[2]　张志林，李玉鑑，刘兆英，等. 深度学习在细粒度图像识别中的应用综述[J]. 北京工业大学学报，2021，47(8)：942-953.
ZHANG ZH L, LI Y J, LIU ZH Y, et al. Deep learning for fine-grained image recognition: A survey[J]. Journal of Beijing University of Technology, 2021, 47(8): 942-953.

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

[4]　朱阳光，刘瑞敏，黄琼桃，基于深度神经网络的弱监督信息细粒度图像识别[J]. 电子测量与仪器学报，2020，34(2)：115-122.
ZHU Y G, LIU R M, HUANG Q T. Fine-grained image recognition of weak supervisory information based on deep neural network[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(2): 115-122.

[5]　齐爱玲，王宣淋. 融合通道与位置信息的 ResNet 细粒度图像识别[J]. 国外电子测量技术，2022，41(12)：103-111.
QI AI L, WANG X L. ResNet fine-grained image identification with fused channel and location information[J]. Foreign Electronic Measurement Technology, 2022, 41(12): 103-111.

[6]　朱丽，王新鹏，付海涛，等. 基于注意力机制的细粒度图像分类[J]. 吉林大学学报：理学版，2023，61(2)：371-376.
ZHU L, WANG X P, FU H T, et al. Fine-grained image classification based on attention mechanisms[J].

Journal of Jilin University: Science Edition, 2023, 61(2):371-376.

[7] 张文轩,吴秦.基于多分支注意力增强的细粒度图像分类[J].计算机科学,2022,49(5):105-112.
ZHANG W X, WU Q. Fine-grained image classification based on multi-branch attention augmentation[J]. Computer Science, 2022, 49(5): 105-112.

[8] 王培森,宋彦,戴礼荣.基于多通道视觉注意力的细粒度图像分类[J].数据采集与处理, 2019, 34(1): 157-166.
WANG P S, SONG Y, DAI L R. Fine-grained image classification with multi-channel visual attention[J]. Data Acquisition & Processing, 2019, 34(1): 157-166.

[9] 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.
LIU W J, ZHAO S Q, QU H CH, et al. Combining foreground feature reinforcement and region mask self-attention for fine-grained image classification[J]. Journal of Intelligent Systems, 2022, 17(6): 1134-1144.

[10] 陈权,陈飞,王衍根,等.融合目标定位与异构局部交互学习的细粒度图像分类[J].自动化学报, 2024, 50(11):2219-2230.
CHEN Q, CHEN F, WANG Y G, et al. Fine-grained Image Classification by Integrating Object Localization and Heterogeneous Local Interactive Learning[J]. Acta Automatica Sinica, 2024, 50(11):2219-2230.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv:2010.11929, 2020.

[12] WU H P, XIAO B, CODELLA N, et al. CVT: introducing convolutions to vision transformers[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 22-31.

[13] ZHANG F, LI M, ZHAI G, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization[C]. International Conference on Multimedia Modeling. Cham: Springer International Publishing, 2021: 136-147.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer vision(ECCV), 2018: 3-19.

[16] YU Y, ZHANG Y, CHENG Z, et al. MCA: multidimensional collaborative attention in deep convolutional neural networks for image recognition[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107079.

[17] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset[J]. California Institute of Technology, 2011, DOI:10.21227/fgbw-vh29.

[18] DATASET E. Novel datasets for fine-grained image categorization[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013: 554-561.

[19] 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 Workshops, 2013: 554-561.

[20] OUYANG D, HE S, ZHANG G, et al. Efficient multi-scale attention module with cross-spatial learning[C]. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2023:1-5.

作者简介

冀得魁,硕士研究生,主要研究方向为计算机视觉与目标检测。

E-mail:j1326553@163.com

李冰锋(通信作者),博士,讲师,主要研究方向为迁移学习、计算机视觉与目标检测。

E-mail:libingfeng@hpu.edu.cn

杨艺,博士,副教授,主要研究方向为深度学习与强化学习。

E-mail:1286535923@qq.com