◎图形图像处理◎

结合Swin及多尺度特征融合的细粒度图像分类

项剑文,陈泯融,杨百冰
华南师范大学 计算机学院,广州 510631

摘 要:针对细粒度图像类间差异小、类内差异大等问题,提出了一种基于Swin及多尺度特征融合的模型(SwinFC)。基准骨干网络采用具有多阶段层级架构设计的Swin Transformer模型作为全新视觉特征提取器,从中获取局部和全局信息以及多尺度特征。然后在每个阶段的分支通道上嵌入融合外部依赖及跨空间注意力模块,以捕获数据样本之间的潜在相关性,同时捕捉不同空间方向上具有判别力的特征信息,进而强化网络每个阶段的信息表征。进一步地,引入特征融合模块将每个阶段提取的特征进行多尺度融合,促使网络学习更加全面、互补且多样化的特征信息。最后构建特征选择模块来筛选重要且具有辨别力的图像块,以此增大类间差异,减小类内差异,增强模型的判别力。实验结果表明,该方法在CUB-200-2011、NABirds和WebFG-496三个公开细粒度图像数据集上分别达到了92.5%、91.8%和85.84%的分类准确率,性能优于大部分主流模型方法,并且与基准模型Swin相比,分别提高了1.4、2.6和4.86个百分点的分类性能。

关键词:细粒度图像分类;Swin Transformer;注意力机制;多尺度特征融合;特征选择
文献标志码:A **中图分类号:**TP391.41 doi:10.3778/j.issn.1002-8331.2211-0456

Fine-Grained Image Classification Combining Swin and Multi-Scale Feature Fusion

XIANG Jianwen, CHEN Minrong, YANG Baibing
School of Computer Science, South China Normal University, Guangzhou 510631, China

Abstract: Challenged by high intra-class variances and low inter-class variances in fine-grained image classification, this paper proposes a fine-grained image classification model based on Swin and multi-scale feature fusion(SwinFC). The Swin Transformer model with multi-stage hierarchical design is used as a new visual backbone network to extract local and global information and multi-scale features. Then, a module integrating external-dependency attention and cross-space attention is embedded on the branches of each stage, which aims to capture potential correlations among data samples and discriminative feature information from different spatial directions, enhancing the information representation in each stage of the network. Further, a feature fusion module is introduced to perform multi-scale fusion of the features extracted at each stage, so that the network can learn more comprehensive, complementary and diverse feature information. Finally, in order to enlarge inter-class differences, narrow the intra-class differences, a feature selection module is adopted to select important and discriminative image patches, enhancing the discriminative power of the network. Experimental results show that the proposed method achieves classification accuracy of 92.5%, 91.8% and 85.84% on three public fine-grained image datasets, CUB-200-2011, NABirds and WebFG-496, respectively, outperforming most of the mainstream methods in classification performance. Moreover, compared with the benchmark model Swin, the classification performance is improved by 1.4, 2.6 and 4.86 percentage points, respectively.

Key words: fine-grained image classification; Swin Transformer; attention mechanism; multi-scale feature fusion; feature selection

基金项目:国家自然科学基金(61872153,61972288)。

作者简介:项剑文(1998—),男,硕士研究生,研究方向为细粒度图像分类、视觉Transformer、计算机视觉;陈泯融(1977—),通信作者,女,博士,教授,研究方向为计算智能与信息安全、计算机视觉,E-mail:chenminrong@scnu.edu.cn;杨百冰(1998—),女,硕士研究生,研究方向为图像风格迁移、生成对抗网络、计算机视觉。

收稿日期:2022-11-29 **修回日期:**2023-01-30 **文章编号:**1002-8331(2023)20-0147-11

近年来,细粒度图像分类逐渐成为计算机视觉、模式识别等领域一个热门的研究课题,其是对同属于一个基础类别下的图像进行更加细致的子类划分。细粒度图像分类重点在于区分具体对象的类别,例如鸟的种类、猫的品种、汽车的品牌等。以鸟类图像为例,同一种鸟类可以有数十种,甚至数百种不同的子类别。比如以海鸥来说,就有燕尾鸥、渔鸥、黑嘴鸥、红嘴鸥等数十种不同子类别的海鸥,这些海鸥之间的差异十分细微,因此具有很大的分类难度。与普通图像分类相比,细粒度图像的类间差异小而类内差异大,并且受到姿态、视角等诸多因素的影响,使得细粒度图像分类成为一项极具挑战性的任务。

为了避免繁琐的人工部位标注,目前大部分的研究主要集中在不需要额外标注信息且仅使用类别标签的弱监督细粒度图像分类任务上。细粒度图像分类的算法大致上可以分为三类,即基于特征编码的方法、基于区域定位的方法以及基于注意力的方法。基于特征编码的方法[1]主要通过丰富特征表示以获得更好的分类性能。与基于特征编码的方法相比,基于区域定位的方法可以精确地捕获不同子类之间的细微差异,并且具有更好的可解释性,通常可以取得更好的结果。早期基于区域定位的方法依靠部位标注来定位判别性区域,而近期的研究[2-3]主要采用区域提议网络(region proposal network,RPN)的方法在图像上提取具有判别性区域的边界框,进而筛选出目标对象可能存在的关键区域。如Ge等人[2]以弱监督的方式构建互补部位模型,以检索由卷积神经网络(convolution neural network,CNN)检测到的目标部位所抑制的信息。然而,基于区域定位的方法忽略了所选区域之间的关系,并且为了能够获得正确的分类结果,其往往会促使RPN提议更大的边界框以包含大部分前景对象。当这些所选的边界框不准确且覆盖了大量的背景信息时,目标对象的关键特征就很容易被混淆。此外,具有不同优化目标的RPN模块会使得骨干网络的训练难度加大,并且重用骨干网络也会使得整体算法流程复杂化。

基于注意力的方法通过自注意力机制自动检测图像中具有判别性的区域,这些方法摆脱了对人工标注判别性区域的依赖,并取得了令人鼓舞的效果。如Zheng等人[4]提出了一种渐进式注意力方法,以在多个尺度上逐步检测具有判别性的部位。最近,Dosovitskiy等人[5]成功将纯Transformer模型引入到计算机视觉领域中,提出vision Transformer(ViT)模型,其是一种完全基于自注意力机制来动态建模元素间关系的新兴视觉特征提取器。在大规模数据集上,无需依赖于CNN,ViT模型即可在各种各样的视觉任务中展现优异的性能。随后,Liu等人[6]构建一种多尺度层级Transformer架构,即Swin Transformer,并通过设计移动窗口将自注意力计算限制在不重叠的局部窗口上,以有效地建模局部信息和全局信息,从而提高模型的性能和效率。ViT模型在视觉任务上的巨大成功表明,纯Transformer架构固有的自注意力机制可以自动检测图像中有助于视觉识别的关键部位。然而,目前很少有研究探索基于视觉Transformer的细粒度图像分类。TransFG[7]网络作为首次在细粒度图像分类任务上研究视觉Transformer的工作,提出将ViT模型中所有原始注意力权值集成到一个注意力图中,以引导网络有效地选择具有判别性的图像块。然后将这些筛选出来的图像块输入到最后一层的Transformer模块中进行融合,最后实现了良好的分类性能。然而,ViT模型更多关注的是全局信息,而对局部信息和低级特征关注较少,由于局部信息在细粒度图像分类中起着极为重要作用,这可能会限制模型对局部关键信息的提取。此外,ViT模型遵循原始Transformer的单级柱状架构设计,并且在不同层之间,特征图始终维持固定的尺度,这不利于模型捕获更多细节信息以及多尺度细粒度的识别特征,进而限制了模型对特征信息的表达。

鉴于上述分析,本文提出了一种基于Swin及多尺度特征融合的细粒度图像分类模型(fine-grained image classification model based on Swin Transformer and multi-scale feature fusion,SwinFC),如图1所示,基准骨干网络采用具有多阶段层级架构设计的Swin Transformer模型作为全新视觉特征提取器,以完成图像特征的级联提取,在此基础上,进一步构建融合外部依赖及跨空间注意力模块、特征融合模块以及特征选择模块,以促进模型学习更加全面、细微以及多样化的特征信息,进而增强模型的判别能力和表征能力。在仅使用类型标签的前提下,本文模型能够有效捕获目标关键部位并实现较为理想的分类性能。主要贡献如下:

(1)利用Swin Transformer网络作为全新视觉特征提取器,从中获取局部和全局信息,建模多尺度特征;提出融合外部依赖及跨空间注意力模块(external-dependency attention and cross-space attention module,EACA),以捕获数据样本间的潜在相关性以及不同空间方向上具有判别力的特征信息,从而强化网络每个阶段的信息表征。

(2)引入特征融合模块[8](feature fusion module,FFM),以完成多尺度特征的成对融合;构建特征选择模块(feature selection module,FSM),筛选具有辨别力的图像块,以此增大类间差异,减小类内差异,增强模型判别力。

(3)在三个公开的细粒度图像数据集上进行一系列的对比实验,结果表明,本文模型的分类性能均高于大部分主流模型。

1　相关理论基础

1.1　Swin Transformer模型概述

Swin Transformer[6]是一种基于多尺度层级设计的特征金字塔网络架构，采用移动窗口的设计模式将自注意力的计算限制在不重叠的局部窗口上，并允许跨窗口连接。Swin Transformer的网络架构如图1的上半部分所示。与ViT模型类似，为了将输入的RGB图像（大小为 $H\times W\times 3$ ）转化为Transformer结构能够处理的序列数据，Swin Transformer首先通过块分割模块（patch partition）将原始二维图像转化为互不重叠的 4×4 图像块（patch tokens）序列，其特征被设置为原始像素RGB值的拼接，特征维度为 $48(4\times 4\times 3)$ ，再利用线性嵌入层（linear embedding）将特征维度投影到任意大小（记为 C ）。随后，将图像块序列输入到堆叠的Swin Transformer模块中以建模特征间的相互关系。特别地，块合并层（patch merging）用于对视觉特征进行降采样和增维操作，以构建多阶段的层级架构，进而可以学习不同空间尺度和维度的特征表示。如图1上半部分所示，第一个块合并层以 $4(2\times 2)$ 的倍数减少图像块的数量（即分辨率为 $\frac{H}{8}\times\frac{W}{8}$ ），输出维度设置为 $2C$ ，紧接着输入到Swin Transformer模块中进行特征交换，此过程为模型的第二阶段。类似地，重复该操作两次，分别得到分辨率为 $\frac{H}{16}\times\frac{W}{16}$ 和 $\frac{H}{32}\times\frac{W}{32}$ 的第三阶段和第四阶段。最后，将输出的图像块序列进行平均池化，并将平均池化结果输入到分类层中以完成模型最终的分类预测。

1.2　多尺度特征融合

具有多阶段层级架构设计的网络往往能够建模不同尺度的特征，这些不同尺度的特征图所提取到的信息重点是不同的，低层特征能够捕获更多细节信息，关注更多关键区域，如边缘纹理、形状颜色等，高层特征具有更加丰富的语义信息，从整体上关注目标区域。因此，有效地将不同尺度的特征进行融合，能够增强模型的特征表示能力，提高模型的识别性能。例如，FPN[9]和SSD[10]尝试利用卷积固有的特征金字塔网络架构，将不同尺度的特征进行融合，从而在目标检测任务中展现出很好的性能。SG-Net[11]利用非局部操作融合不同层的特征图，以高效地提取不同的潜在特征，提高模型的特征表示能力。受此启发，本文在基准骨干网络Swin Transformer的基础上，首先通过注意力模块来强化每个阶段的信息表征，然后利用模型的多阶段层级构架将不同尺度的特征进行融合，从而促使模型学习更加丰富的特征表示，增强模型的判别力。

2　本文方法

2.1　SwinFC模型整体架构

基于视觉Transformer的细粒度图像分类的初步探索表明，作为视觉领域的新兴特征提取器，视觉Transformer能够有效地建模有利于细粒度分类的视觉特征。然而，原生的ViT模型完全采用全局注意力机制建模特征间关系，并且遵循单级柱状架构设计，这不利

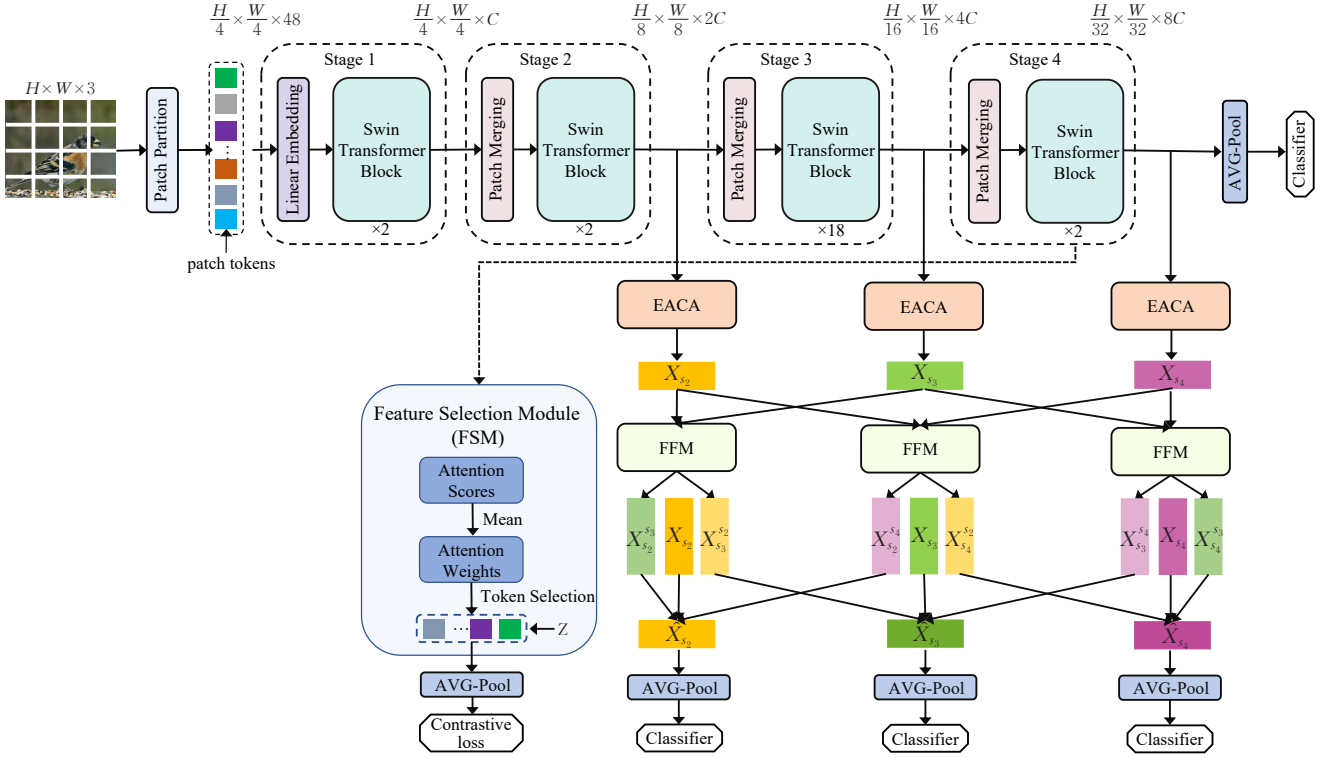

图1　SwinFC网络整体结构图

Fig.1　Overview structure of SwinFC network

于模型捕获更加细微以及多尺度细粒度的识别特征,从而限制了模型对特征信息的表达。为此,本文提出了一种基于Swin及多尺度特征融合的细粒度图像分类模型(SwinFC)。基准骨干网络采用Swin Transformer模型[6]作为输入图像的特征提取器。在骨干网络的基础上,进一步构建融合外部依赖及跨空间注意力模块、特征融合模块以及特征选择模块,以促进模型学习更加全面、细微以及多样化的特征信息,进而增强模型的判别能力和表征能力。

本文提出的SwinFC整体结构如图1所示。具体而言,采用具有层级结构的Swin Transformer骨干网络作为细粒度图像分类的全新特征提取器,以完成对视觉特征由浅入深的级联提取。然后在骨干网络每个阶段的末端增加多尺度特征融合分支(第一个阶段除外),并在每个分支的通道上嵌入融合外部依赖及跨空间注意力模块(EACA)以及特征融合模块(FFM)。将每个阶段的输出特征图并行输入到骨干网络及其分支通道上。在每个阶段的分支通道上,特征图首先被输入到EACA模块中,以挖掘特征样本间的潜在关系,同时捕捉不同空间方向上具有判别力的特征信息,进而强化网络每个阶段的信息表征。随后采用FFM模块对不同阶段的特征图进行多尺度的特征融合操作,使得高分辨率的底层特征与低分辨的高层特征能够被同时利用,从而促进网络学习更加全面、互补且多样化的特征信息。此外,重用骨干网络最后一个阶段的多头自注意力机制来构建特征选择模块(FSM),以筛选重要且具有辨别力的图像块,并对所选图像块进行平均池化操作,接着对池化结果计算对比损失,以此增大类间特征差异的同时减小类内特征差异。最后,用于分类预测的总损失函数由骨干网络的交叉熵损失、不同阶段的交叉熵损失以及对比损失融合而成,从而使得模型学习到更加全面的视觉表征知识,提高模型的性能收益。

2.2　融合外部依赖及跨空间注意力模块

细粒度图像往往因其数据样本类间差异小、类内差异大而导致模型预测类别信息易混淆。如果网络能够挖掘样本间潜在的相关性,并能够有效定位到对图像分类影响较大的部位,则可以提升网络的分类性能[12]。基于此,本文提出了融合外部依赖及跨空间注意力模块(EACA),并将其作用于每个阶段的输出特征图,以强化网络每个阶段的信息表征。

具体而言,EACA模块由两个注意力子模块并行组成:外部依赖注意力子模块(external-dependent attention,EA)以及跨空间注意力子模块(cross-spatial attention,CA)。将骨干网络每个阶段输出的特征图序列分别输入到EACA模块的两个子模块中,特别地,对于跨空间注意力子模块,由于其是对空间结构的建模,因此需将输出的特征图序列重塑回二维图像形式。在外部依赖注意力子模块中,利用外部依赖注意力来挖掘数据样本之间的潜在关系,使相同类别下的特征更具关联性,从而得到更具鲁棒性的特征;在跨空间注意力子模块中,聚合两个不同空间方向上的注意力,以感知空间位置信息,增强特征关注的丰富性,促进模型更加准确地定位判别性的局部区域。最后,将两个子模块的输出特征图进行相加,以得到EACA模块的输出特征图。如图2所示。

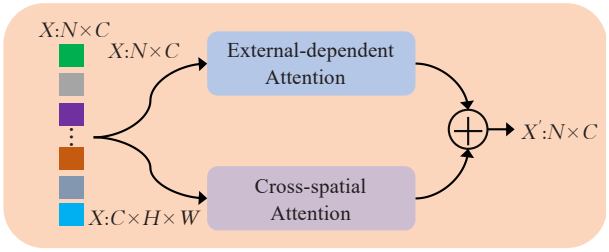

图2　融合外部依赖及跨空间注意力模块

Fig.2　EACA module

图中,$X\in\mathbb{R}^{N\times C}$ 和 $X'\in\mathbb{R}^{N\times C}$ 分别表示每个阶段输出的特征图以及EACA模块输出的特征图,N 表示特征图序列的长度,C 表示特征图的通道数,H 和 W 分别表示特征图的高度和宽度(其中 $N=H\times W$)。

2.2.1　外部依赖注意力子模块

属于同一类别但分布在不同样本中的特征应该被一致地对待,从而捕获同类样本间的内在关联性,减少其他不同类别样本的干扰[12]。受此启发,构建外部依赖注意力子模块,通过引入额外的外部可学习参数来捕获样本内和样本间的相关性,促使网络学习同类样本的潜在关联性,强化模型的学习能力。外部依赖注意力子模块如图3所示。

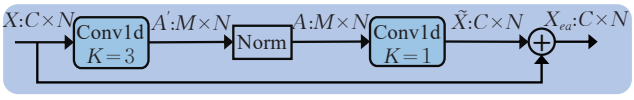

图3　外部依赖注意力子模块

Fig.3　External-dependent attention submodule

首先,将输入特征图 $X\in\mathbb{R}^{C\times N}$($N$ 为序列长度,C 为通道数)输入到一维卷积中以生成中间注意力图 $A'\in\mathbb{R}^{M\times N}$,其中,一维卷积的卷积核大小设置为3(即 $K=3$),紧接着 A' 经过正则化处理得到注意力图 $A\in\mathbb{R}^{M\times N}$,然后将注意力图 A 输入到卷积核大小为1的一维卷积中,以计算得到更为精细的特征图 $\tilde{X}\in\mathbb{R}^{C\times N}$。事实上,两个一维卷积的卷积权重 $W_1\in\mathbb{R}^{C\times M\times 3}$ 和 $W_2\in\mathbb{R}^{M\times C\times 1}$ 都是可学习的外部记忆矩阵,共享于整个数据样本。因此,由输入特征图 X 与外部记忆矩阵 $\boldsymbol{W}_1$ 乘积并正则化而来的注意力图 A 可视为独立于单个输入样本的外部依赖注意力,注意力图 A 与外部记忆组件 $\boldsymbol{W}_2$ 联合计算得到的特征图 $\tilde{X}$ 则蕴含着数据样本间的潜在相关性。最后,将特征图 X 与 $\tilde{X}$ 进行残差操

作，以得到最终的输出结果 X_{ea}。公式化计算过程如式(1)~(3)所示。

$$A = Norm(X\boldsymbol{W}_1) \tag{1}$$

$$\tilde{X} = A\boldsymbol{W}_2 \tag{2}$$

$$X_{ea} = X + \tilde{X} \tag{3}$$

式中，Norm为正则化操作。$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为可学习的外部记忆矩阵。

2.2.2　跨空间注意力子模块

跨空间注意力子模块利用两个不同空间方向上的全局平均池化操作分别将输入特征图聚合为两个并行的方向感知特征图，然后将两个嵌入特定方向的特征图分别编码为两个并行的注意力图，每个注意力图能够捕获输入特征图沿着一个空间方向上更加细粒度的依赖关系，进而学习到更具区分性的局部细节特征。跨空间注意力子模块如图4所示。

首先分别使用尺寸为 $(H, 1)$ 和 $(1, W)$ 的池化核沿水平方向（W Avg-Pool）和垂直方向（H Avg-Pool）对输入特征图进行全局平均池化操作，以得到两个并行的方向感知特征图 $X_H \in \mathbb{R}^{C\times H\times 1}$ 和 $X_W \in \mathbb{R}^{C\times 1\times W}$，两个特征图分别聚合了一个空间方向的全局依赖性，并保留沿另一个空间方向上的位置信息。随后，分别采用卷积核大小为 3×1 和 1×3 的二维卷积对 X_H 和 X_W 进行卷积运算，以得到更具信息的特征图 $X_H^{'} \in \mathbb{R}^{C\times H\times 1}$ 和 $X_W^{'} \in \mathbb{R}^{C\times 1\times W}$。再使用Softmax函数分别为两个方向上不同的空间位置分配软性注意力概率值，以得到注意力图 $A_H \in \mathbb{R}^{C\times H\times 1}$ 和 $A_W \in \mathbb{R}^{C\times 1\times W}$。接着通过整合两个方向上的注意力权重，得到跨空间注意力图 $A \in \mathbb{R}^{C\times H\times W}$。最后，将注意力图 A 与输入特征图 X 进行元素相乘，实现跨空间注意力的加权融合，得到最终的输出特征图 $X_{ca} \in \mathbb{R}^{C\times H\times W}$。公式化该实现过程如式(4)~(6)所示。

$$A_H = \sigma(F_H(X_H)) \tag{4}$$

$$A_W = \sigma(F_W(X_W)) \tag{5}$$

$$X_{ca} = X \otimes (A_H + A_W) \tag{6}$$

式中，σ 表示Softmax函数，F_H 和 F_W 分别表示卷积操作，$\otimes$ 表示元素相乘。

2.3　特征融合模块

在骨干网络的不同阶段中，特征图具有不同的尺度，所包含的视觉信息重点不同。为了能够提取更加全面且互补的特征信息，本文采用了Song等人[8]提出的特征融合方法来构建特征融合模块(feature fusion module，FFM)，以将每个阶段提取的特征进行成对融合，从而增强每个阶段的视觉表征。特征融合模块的详细结构如图5所示。

给定任意两个阶段的特征图 $X_{s_i} \in \mathbb{R}^{C\times H_i\times W_i}$ 和 $X_{s_j} \in \mathbb{R}^{C\times H_j\times W_j}$，首先将两个特征矩阵相乘以得到相似矩阵 $\boldsymbol{M} \in \mathbb{R}^{H_iW_i\times H_jW_j}$，如式(7)所示：

$$\boldsymbol{M} = F(X_{s_i}, X_{s_j}),\ F(X, Y) = X^{\mathrm{T}}Y \tag{7}$$

其中，相似度越低，表明具有更多的互补信息，对相似矩阵取反以得到互补相关矩阵（即 $-\boldsymbol{M}$）。随后对互补相关矩阵进行Softmax归一化操作，接着将其分别与两个阶段的特征图进行矩阵相乘，以得到具有互补的输出特征图 $X_{s_j}^{s_i}$ 和 $X_{s_i}^{s_j}$。公式化该计算过程如式(8)~(11)所示：

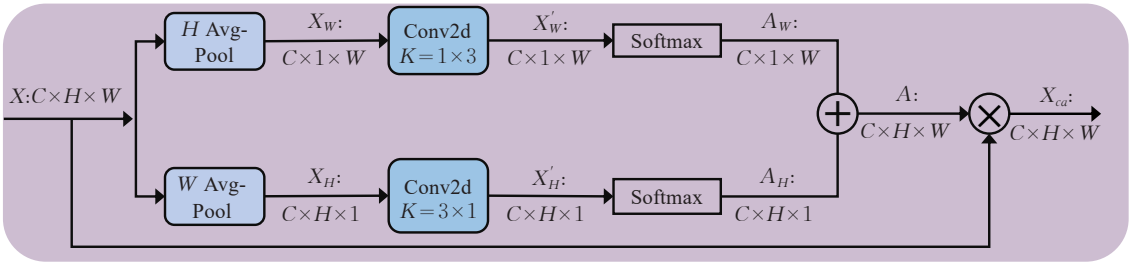

图4　跨空间注意力子模块

Fig.4　Cross-spatial attention submodule

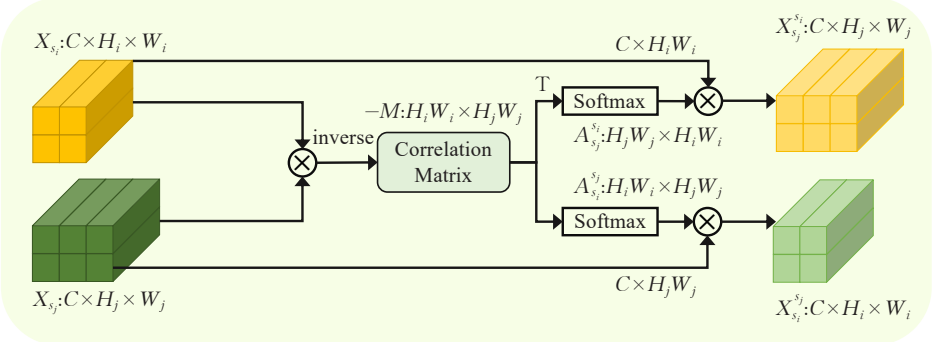

图5　特征融合模块

Fig.5　Feature fusion module

$$A_{s_j}^{s_i}=Softmax(-\boldsymbol{M}^{\mathrm{T}})\in[0,1] \tag{8}$$

$$A_{s_i}^{s_j}=Softmax(-\boldsymbol{M})\in[0,1] \tag{9}$$

$$X_{s_j}^{s_i}=X_{s_i}A_{s_j}^{s_i}\in\mathbb{R}^{C\times H_j\times W_j} \tag{10}$$

$$X_{s_i}^{s_j}=X_{s_j}A_{s_i}^{s_j}\in\mathbb{R}^{C\times H_i\times W_i} \tag{11}$$

式中，$A_{s_j}^{s_i}\in\mathbb{R}^{H_jW_j\times H_iW_i}$，$A_{s_i}^{s_j}\in\mathbb{R}^{H_iW_i\times H_jW_j}$，对于 $X_{s_j}^{s_i}$，其是 X_{s_j} 相对于 X_{s_i} 的互补信息输出特征图，对于 $X_{s_i}^{s_j}$ 同理可得。

为了使提取到的特征表示更加多样化，将每个阶段分支通道上的FFM模块输入特征图与输出互补特征图进行融合。具体而言，如图1右下半部分所示，给定当前阶段 S_i，则当前阶段所在分支最终的输出特征图 $\hat{X}_{s_i}$ 可由式(12)计算得到：

$$\hat{X}_{s_i}=X_{s_i}+\sum_{j=2}^{P}X_{s_i}^{s_j},j\neq i \tag{12}$$

式中，P 表示模型的阶段数量。

2.4　特征选择模块

为了定位细粒度图像分类中子类之间具有区别性的区域和细微差异，本文充分利用最后一个阶段(即Stage4)中的多头注意力来筛选更具区别性的图像块，并以此构建特征选择模块(FSM)。特征选择模块详细结构如图1左下部分所示。

具体而言，给定当前层的单头注意力矩阵 $\boldsymbol{A}\in\mathbb{R}^{N\times N}$，$N$ 表示图像块序列长度，通过对矩阵中的每列取平均以得到平均注意力向量 $\boldsymbol{A}_{\mathrm{avg}}$（长度为 N），平均注意力向量中的每个元素表示对应图像块对模型的响应，权值越大，表明对模型分类发挥更重要的作用。公式化该过程如式(13)、(14)所示：

$$a_j=\frac{1}{N}\sum_{i=1}^{N}A_{(i,j)} \tag{13}$$

$$\boldsymbol{A}_{\mathrm{avg}}=[a_0,a_1,\cdots,a_j,\cdots,a_N] \tag{14}$$

式中，$A_{(i,j)}$ 表示注意力矩阵 $\boldsymbol{A}$ 中第 i 行第 j 列的注意力权重，a_j 表示第 j 个图像块对模型的重要性得分。接下来根据平均注意力向量 $\boldsymbol{A}_{\mathrm{avg}}$ 来筛选出权值最大所对应的图像块，并以此作为候选图像块。由于每一层具有多头注意力矩阵，分别对多头注意力矩阵执行上述操作，可得到 K 个候选图像块，其中 K 为每一层的注意力头数。特别地，对于模型最后一个阶段，其仅有两层Swin Transformer模块，并且两层具有相同注意力头数，因此，可得到 $2K$ 个候选图像块。将 $2K$ 个候选图像块组成的特征集合记为 Z，并对其进行全局平均池化以得到该集合的全局表示 $\hat{Z}$，随后，将 $\hat{Z}$ 输入到对比损失函数(contrastive loss)[7]中，以增大类间特征差异，减小类内特征差异，捕获更具判别性图像块。对比损失计算如式(15)所示：

$$L_{\mathrm{con}}=\frac{1}{N_B^2}\sum_{i}^{N_B}\Big[\sum_{j:y_i=y_j}^{N_B}(1-\cos(\hat{Z}_i,\hat{Z}_j))+\sum_{j:y_i\neq y_j}^{N_B}\max(\cos(\hat{Z}_i,\hat{Z}_j)-\alpha,0)\Big] \tag{15}$$

式中，N_B 表示批处理的大小，y_i 表示第 i 个图像的真实标签，$\hat{Z}_i$ 表示第 i 个图像经过特征选择模块后得到的特征表示，cos(,)表示两个特征图的余弦相似度，其大于超参数 α 时才会在对比损失函数中发挥作用。L_{con} 表示对比损失，其经过反向传播可以扩大不同子类别间的特征表示，缩小相同子类别的特征表示，促使模型筛选更具判别性的图像块。

2.5　损失函数

综合上述分析，本文提出的模型最终损失函数如式(16)所示：

$$L_{\mathrm{total}}=L_{\mathrm{swin}}+\sum_{i}^{P}\beta_iL_{\mathrm{stage}}^{i}+L_{\mathrm{con}} \tag{16}$$

式中，L_{total} 表示总损失函数，L_{swin} 表示骨干网络的交叉熵损失函数，L_{stage}^{i} 表示模型的第 i 个阶段交叉熵损失函数，L_{con} 表示对比损失函数，β_i 为第 i 阶段的超参数，通过对模型第 i 阶段的损失函数进行加权来控制模型第 i 阶段对模型性能的影响程度，P 表示模型的阶段数量。

3　实验结果与分析

3.1　数据集

为了评估本文方法的分类性能，在CUB-200-2011[13]、NABirds[14]以及WebFG-496[15]三个公共的细粒度图像数据集上进行实验分析。特别地，WebFG-496是网络监督细粒度图像数据集(webly supervised fine-grained datasets)，其由三个子数据集组成，总共有53 339幅网络训练图片，包含200种鸟类(web-bird)、100种飞机(web-aircraft)以及196种汽车模型(web-car)。网络监督数据集除了有细粒度图像常见的特性以外，还存在较大的数据偏差以及较多的噪声数据，因此具有更大的挑战性[15-16]。本文实验中数据集的详细信息如表1所示。

表1　细粒度图像数据集详细信息

Table 1　Details of fine-grained image datasets

Datasets		Training	Testing	Category
CUB-200-2011		5 994	5 794	200
NABirds		23 929	24 633	555
WebFG-496	Web-Bird	18 388	5 794	200
	Web-Aircraft	13 503	3 333	100
	Web-Car	21 448	8 041	196

3.2　实验设置与评价指标

3.2.1　实验设置

实验环境为Ubuntu 18.04.3 LTS系统，使用四个

RTX 2080 TI GPU并行训练。模型训练平台采用基于Python编程语言的PyTorch深度学习框架。实验中,所有图像的尺寸首先统一调整为512×512,然后再裁剪为384×384,同时采用常见的数据增强策略来扩充数据,如随机水平翻转、随机旋转等。本文采用官方[6]公布的Swin-B模型作为骨干网络和特征提取网络,并使用官方[6]发布的预训练权重对骨干网络初始化,对新增模块采用随机初始化。所有模型使用随机梯度下降[17](stochastic gradient descent,SGD)优化器进行训练,并设置动量为0.9。批处理大小设置为32,余弦退火(cosine annealing)调整学习率。对比损失中的超参数α设置为0.4,损失函数中最后三个阶段的超参数$\{\beta_2,\beta_3,\beta_4\}$设置为{0.4,0.6,0.8}。针对不同的数据集,本文对SwinFC模型采用不同的学习率进行训练:对于CUB-200-2011数据集,骨干网络学习率为2E−3,新增模块学习率为5E−3;NABirds数据集和WebFG-496数据集,骨干网络和新增模块学习率为3E−2。

3.2.2 评价指标

本文使用测试集的分类准确度(Accuracy)作为模型的评价指标,最终结果取多次实验的平均值,以更加客观地反映模型的分类性能,计算公式如式(17)所示:

$$Accuracy=\frac{\text{分类预测正确的样本数量}}{\text{总的测试样本数量}} \tag{17}$$

3.3 消融实验

为了验证SwinFC模型以及提出的各个模块的有效性,本文首先设计了不同方法的消融实验。除非有必要的说明,本文所有消融实验都是在CUB-200-2011数据集下展开。

3.3.1 融合外部依赖及跨空间注意力模块的实验分析

为了验证融合外部依赖及跨空间注意力模块(EACA)及其子模块(EA子模块、CA子模块)的有效性,本小节的消融实验分别在骨干网络的最后三个阶段依次引入EACA模块的每个子组件,并单独进行实验训练。实验结果如表2所示,在未引入任何模块的情况下,基准骨干网络Swin可以实现91.13%的分类准确率,在此基础上,进一步引入EACA模块并将两个子模块进行并行训练,模型的分类准确率可以达到91.80%,实现了0.67个百分点的性能提升。特别地,在仅使用EA子模块时,模型可以实现0.39个百分点的性能提升;在仅使用CA子模块时,模块可以实现0.5个百分点的性能提升;通过实验结果显示,将EA子模块与CA子模块并行组合可以进一步带来性能上的收益,其能够联合捕获样本间的相关性以及更具判别性的区域,进而提高模型的分类性能。

表2 EACA模块中不同组件模块消融实验分析

Table 2 Ablation experiment analysis of different components in EACA module

Method	EACA module		Accuracy/%
	EA submodule	CA submodule	
Swin(baseline)	—	—	91.13
SwinFC(EA)	√	—	91.52
SwinFC(CA)	—	√	91.63
SwinFC(EACA)	√	√	91.80

3.3.2 特征融合模块与特征选择模块的实验分析

在上一小节实验的基础上,进一步引入特征融合模块(FFM)以整合不同阶段的特征。实验结果如表3所示,当使用FFM模块来融合不同阶段特征时,模型的性能进一步得到了提升,实现了92.12%的分类准确率,并且与原始Swin模型相比,模型分类准确率提升了0.99个百分点。此外,如表3最后一行所示,当引入特征选择模块时,模型的性能进一步提高了0.41个百分点,并且与原模型Swin相比,模型整体的分类准确率到达了92.53%,实现了1.4个百分点的性能提升。实验结果表明,本文提出的各组件模块能够有效捕获有利于细粒度图像分类的视觉特征,进而提高模型整体的分类性能。

表3 FFM模块与FSM模块的消融实验分析

Table 3 Ablation experiment analysis of FFM module and FSM module

Method	EACA	FFM	FSM	Accuracy/%
Swin(baseline)	—	—	—	91.13
SwinFC(EACA)	√	—	—	91.80
SwinFC(EACA+FFM)	√	√	—	92.12
SwinFC(EACA+FFM+FSM)	√	√	√	92.53

3.3.3 不同阶段中超参数β设置的实验分析

本小节以网格搜索的方式对损失函数中最后三个阶段的超参数$\{\beta_2,\beta_3,\beta_4\}$的设置进行消融实验分析,其中搜索范围为(0,1]。实验结果如表4所示,表中第一列SwinFC后所注明的序号为实验组号,第二列为每组实验所对应的超参数$\{\beta_2,\beta_3,\beta_4\}$的设置。从表中可知,当对不同阶段的超参数设置不同的权值时,模型的分类准确度都高于对不同阶段设置相同的权值,其原因是:模型级联式提取不同层次的特征,所包含的视觉信息重点不同,底层更多关注位置、边缘和低层次的细节信息,经过多层特征提取操作后,高层特征往往具有更强的语义信息,更有利于模型的分类,因此,通过对不同阶段的超

表4 不同阶段中超参数β设置的消融实验分析

Table 4 Ablation experiment analysis of hyperparameter β settings in different stages

Method	$\{\beta_2,\beta_3,\beta_4\}$	Accuracy/%
Swin(baseline)	—	91.13
SwinFC-1	{1,1,1}	92.18
SwinFC-2	{0.4,0.6,0.8}	92.53
SwinFC-3	{0.5,0.5,0.5}	92.29
SwinFC-4	{0.2,0.3,0.5}	92.37

参数设置不同权值来控制模型不同阶段的作用程度,进而有效促使模型学习更加全面且多样的特征信息。特别地,当最后三个阶段的超参数 $\{\beta_2,\beta_3,\beta_4\}$ 设置为{0.4,0.6,0.8}时,模型取得了最优的分类准确率,为此,本文将其作为默认的参数设置。

3.3.4　不同优化器的实验分析

图6展示了本文模型在SGD和Adam两种优化器下的损失函数收敛曲线以及准确率收敛曲线。特别地,模型训练步长(train steps)设置为15 000,每隔100 Steps获取对应的损失值和准确率。从图中可知,相比于Adam优化器,SGD优化器能够更好地优化本文模型,使得模型收敛于更小的损失值,从而实现更高的分类准确率。为此,本文采用SGD作为模型默认的优化器。

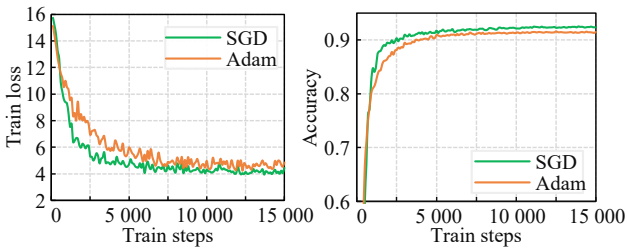

图6　不同优化器的实验分析

Fig.6　Experimental analysis of different optimizers

3.4　不同细粒度图像分类方法的比较

表5展示了本文模型SwinFC在CUB-200-2011数据集上与其他模型的实验对比结果。从表5可看出:本文方法明显高出了所有基于CNN的方法和基于视觉Transformer方法,展现了最先进的性能,例如,与性能最优的CNN模型API-Net相比,分类准确率提高了2.5个百分点,与性能最优的Transformer模型TransFG相比,提高了0.8个百分点的准确率;与基准骨干网络Swin相比,提升了1.4个百分点的分类性能。其次,表6展示了在NABirds数据集上的实验对比结果,特别地,相对于CUB-200-2011数据集,NABirds是一个更大的鸟类数据集,有555种类别,因此具有更大的挑战性。从表6可看出,本文方法高于大部分的主流方法,具有明显的性能优势,实现了91.8%的分类准确率,相比较于最优模型CAP,高出了0.8个百分点,并且与基准模型Swin相比,提高了2.6个百分点。实验结果表明,本文模型能够有效学习到有利于细粒度图像分类的关键特征,捕获更具多样且丰富的特征信息,从而提高了模型的分类性能和泛化能力。

表7展示了在WebFG-496数据集上的实验对比结果,从表中可知,本文模型SwinFC在Web-496数据集的三个子数据集上均获得了高于所有主流方法的分类准确率。例如,相比于CMW-Net-SL模型,本文方法在Web-Bird、Web-Aircraft以及Web-Car上分别高出了9.51、7.95以及6.48个百分点;与基准模型Swin相比,在

表5　不同分类算法在CUB-200-2011上的准确率对比

Table 5　Comparison of accuracy of different classification methods on CUB-200-2011

Method	Backbone	Accuracy/%
ResNet-50[18]	ResNet-50	84.5
MaxEnt[19]	DesenNet-161	86.6
Cross-X[20]	ResNet-50	87.7
BARAN[21]	B-CNN+ResneXt29	87.9
DBTNet[1]	ResNet-101	88.1
Ding[22]	Xception	89.3
PMG[23]	ResNet-50	89.6
API-Net[24]	DesenNet-161	90.0
ViT[5]	ViT-B_16	90.3
TransFG[7]	ViT-B_16	91.7
FFVT[25]	ViT-B_16	91.6
TransAA[26]	ViT-B_16	91.4
Swin[6]	Swin-B	91.1
SwinFC(our)	Swin-B	**92.5**

表6　不同分类算法在NABirds上的准确率对比

Table 6　Comparison of accuracy of different classification methods on NABirds

Method	Backbone	Accuracy/%
MaxEnt[19]	DenseNet-161	83.0
Cross-X[20]	ResNet-50	86.4
API-Net[24]	DesenNet-161	88.1
Ding[22]	Xception	88.4
CS-Parts[27]	ResNet-50	88.5
MGE-CNN[28]	ResNet-50	88.6
FixSENet-154[29]	SENet-154	89.2
CAP[30]	Xception	91.0
DeepFVE[31]	Inception-V3	90.3
ViT[5]	ViT-B_16	89.9
TransFG[7]	ViT-B_16	90.8
TransAA[26]	ViT-B_16	90.4
Swin[6]	Swin-B	89.2
SwinFC(our)	Swin-B	**91.8**

表7　不同分类算法在WebFG-496上的准确率对比

Table 7　Comparison of accuracy of different classification methods on WebFG-496

单位:%

Method	Web-Bird	Web-Aircraft	Web-Car	Average
VGG-19[32]	67.69	70.99	67.21	68.63
ResNet-101[18]	66.74	63.46	65.51	65.24
GoogLeNet[33]	66.01	66.02	65.87	65.97
B-CNN[3]	66.56	64.33	67.42	66.10
Decouping[34]	70.56	75.97	75.00	73.84
Co-teaching[35]	73.85	72.76	73.10	73.24
Peer-learning[15]	76.48	74.38	78.52	76.46
MW-Net[36]	75.60	72.93	77.33	75.29
CMW-Net[37]	75.72	73.72	77.42	75.62
CMW-Net-SL[37]	77.41	76.48	79.70	77.86
ViT[5]	84.47	72.94	76.04	77.82
Swin[6]	83.03	77.98	81.92	80.98
SwinFC(our)	**86.92**	**84.43**	**86.18**	**85.84**

三个子数据集上分别提高了3.89、6.45以及4.26个百分点。此外，本文也是首次探索视觉Transformer在网络监督细粒度图像数据集上的应用，并且从实验结果可以看出，视觉Transformer作为基础视觉特征提取器，能够在网络监督细粒度图像分类中表现出较好的分类性能。

表8展示了本文模型分别在三个数据集上的平均准确率(Avg)、标准差(Std)以及方差(Var)，从表中可以看出，本文模型在三个数据集上的实验结果具有较小的方差，这表明本文模型具有较好的稳定性以及鲁棒性。

表8　SwinFC模型的平均准确率、标准差以及方差

Table 8　Average accuracy，standard deviation and variance of SwinFC model

Datasets		Avg/%	Std	Var
CUB-200-2011		92.53	0.055	0.003
NABirds		91.82	0.045	0.002
WebFG-496	Web-Bird	86.92	0.095	0.009
	Web-Aircraft	84.43	0.182	0.033
	Web-Car	86.18	0.055	0.003

3.5　模型复杂度分析

模型浮点计算量(floating-point operations，FLOPs)、推理时间(Inference time)以及吞吐量(Throughput)等是评价深度学习模型复杂度的重要指标。为此，本文在相同实验环境配置下，使用CUB-200-2011的测试集作为实验测试数据，分别对本文模型、基准模型Swin以及同样采用视觉Transformer为基准的模型进行模型复杂度实验对比分析。实验结果如表9所示，从表中可知，相比于基准模型，本文模型虽然在浮点计算量和推理时间上略有增加，吞吐量有下降，但更重要的是从表5可知，本文模型的分类准确率在很大程度上高于基准模型。此外，更值得一提的是，从表9可看出，与ViT和TransFG模型相比，本文模型不仅在分类准确率上有很大的提高，而且在模型浮点计算量、推理时间以及吞吐量上都具有明显的优势。

表9　模型复杂度分析

Table 9　Complexity analysis of model

Method	FLOPs/10^9	Inference time/s	Throughput/(image/s)
ViT[5]	67.1	115.7	50.1
TransFG[7]	107.6	524.7	11.0
Swin[6]	47.0	86.8	66.8
SwinFC(our)	59.0	103.3	56.1

3.6　可视化分析

为了进一步验证本文模型的有效性，采用类激活可视化(Grad-CAM)[38]的方法对模型的分类识别性能进行量化分析。本小节随机选取各个数据集中测试集的图像作为实验测试数据，并以可视化热度图的方式展示模型预测出的判别性区域位置。图7展示了本文模型SwinFC与原始模型Swin的可视化热度图结果，第一行为原始图像，第二行为基准模型Swin生成的热图，第三行为本文模型SwinFC生成热图，其中，热度图中的高亮区域(即红色)表示与模型预测类别相关的区域。从图7中可以看到：基准模型热度图中判别性区域显得更加分散、微小，并且关注了大量的背景信息，相反本文模型不仅能够聚焦于目标物体，而且能够有效定位到具有判别性的区域，如鸟的头部、羽毛等。这表明了本文方法能够有效学习到更加全面、细微且丰富的特征信息，增强了网络模型的判别力和表征能力，进而提高模型的分类性能。

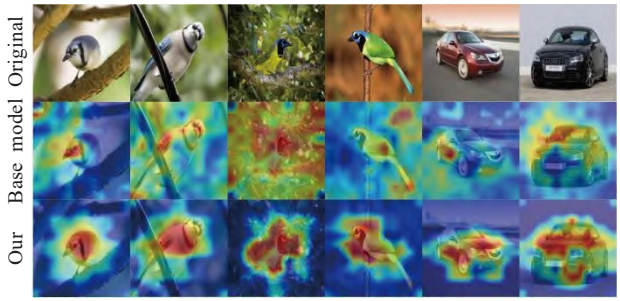

图7　基准模型与本文模型生成的热度图对比

Fig.7　Comparison of heat maps generated by baseline model and proposed model

4　结束语

本文提出了一种基于Swin及多尺度特征融合的细粒度图像分类模型(SwinFC)。采用具有多阶段层级架构设计的Swin Transformer模型作为全新视觉特征提取器。然后在骨干网络每个阶段的分支通道上嵌入融合外部依赖及跨空间注意力模块，以捕获数据样本之间的潜在相关性，同时捕捉不同空间方向上多样且具判别力的特征信息，强化网络每个阶段的信息表征。进一步地，引入特征融合模块以将每个阶段提取的特征进行多尺度融合，促使网络学习更加全面、互补且多样化的特征信息。最后构建特征选择模块来筛选重要且具有辨别力的图像块，以此增大类间差异，减小类内差异，增强模型的判别力。实验结果表明，本文提出的模型在多个细粒度数据集上均取得优异的性能，高于大部分的主流方法。下一步将深入研究视觉Transformer架构在细粒度图像分类中的内在特性以及模型自身过大问题，以探索出更加适用于细粒度图像分类的网络。

参考文献：

[1] ZHENG H L，FU J L，ZHA Z J，et al.Learning deep bilinear transformation for fine-grained image representation[C]// Proceedings of Annual Conference on Neural Information Processing Systems 2019，Dec 8-14，2019，Vancouver，BC，Canada，2019：4279-4288.

[2] GE W F，LIN X R，YU Y Z.Weakly supervised complementary parts models for fine-grained image classifica-

tion from the bottom up[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition,2019:3034-3043.

[3] LIN T Y,ROYCHOWDHURY A,MAJI S.Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision,Santiago,Chile,December 7-13,2015:1449-1457.

[4] ZHENG S X,LU J C,ZHAO H S,et al.Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition,2021:6881-6890.

[5] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al.An image is worth 16×16 words:transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations,Austria,May 3-7,2021.

[6] LIU Z,LIN Y T,HU H,et al.Swin transformer:hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE International Conference on Computer Vision,Montreal,QC,Canada,Oct 10-17,2021:9992-10002.

[7] HE J,CHEN J N,LIU S,et al.TransFG:a transformer architecture for fine-grained recognition[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence,the 34th Conference on Innovative Application of Artificial Intelligence,the 12th Symposium on Educational Advances in Artificial Intelligence,Feb 22-March 1,2022:852-860.

[8] SONG J W,YANG R Y.Feature boosting,suppression,and diversification for fine-grained visual classification[C]//Proceedings of International Joint Conference on Neural Networks,Shenzhen,China,July 18-22,2021:1-8.

[9] 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,2017:936-944.

[10] LIU W,ANGUELOV D,ERHAN D,et al.SSD:single shot multibox detector[C]//Computer Vision ECCV 2016.Cham:Springer International Publishing,2016:21-37.

[11] CHEN X S,FU C M,ZHAO Y,et al.Salience-guided cascaded suppression network for person re-identification[C]//2020 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),June 13-19,2020,Seattle,WA,2020:3297-3307.

[12] GUO M H,LIU Z N,MU J T,et al.Beyond self-attention:external attention using two linear layers for visual tasks[J].arXiv:2105.02358,2021.

[13] WAH C,BRANSON S,WELINDER P,et al.The Caltech-UCSD Birds-200-2011 dataset[R].Pasadena:California Institute of Technology,2011.

[14] HORN V G,BRANSON S,HABER S,et al.Building a bird recognition app and large scale dataset with citizen scientists:the fine print in fine-grained dataset collection[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition,Boston,June 7-12,2015:595-604.

[15] SUN Z,YAO Y Z,WEI X S,et al.Webly supervised fine-grained recognition:benchmark datasets and an approach[C]//Proceedings of 2021 IEEE International Conference on Computer Vision,Montreal,QC,Canada,Oct 10-17,2021:10582-10591.

[16] 魏秀参,许玉燕,杨健.网络监督数据下的细粒度图像识别综述[J].中国图象图形学报,2022,27(7):2057-2077.
WEI X S,XU Y Y,YANG J.Review of webly-supervised fine-grained image recognition[J].Journal of Image and Graphics,2022,27(7):2057-2077.

[17] SUTSKEVER I,MARTENS J,DAHL G E,et al.On the importance of initialization and momentum in deep learning[C]//Proceedings of the 30th International Conference on Machine Learning,Atlanta,GA,June 16-21,2013:1139-1147.

[18] HE K,ZHANG X Y,REN S Q,et al.Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition,Las Vegas,June 27-30,2016:770-778.

[19] DUBEY A,GUPTA O,RASKAR R,et al.Maximum-entropy fine grained classification[C]//Proceedings of the Annual Conference on Neural Information Processing System,Montréal,Dec 3-8,2018:635-645.

[20] LUO W,YANG X T,MO X J,et al.Cross-x learning for fine-grained visual categorization[C]//Proceedings of the International Conference on Computer Vision,Oct 27-Nov 2,2019:8241-8250.

[21] 李宽宽,刘立波.双线性聚合残差注意力的细粒度图像分类模型[J].计算机科学与探索,2022,16(4):938-949.
LI K K,LIU L B.Fine-grained image classification model based on bilinear aggregate residual attention[J].Journal of Frontiers of Computer Science and Technology,2022,16(4):938-949.

[22] 丁文谦,余鹏飞,李海燕,等.基于Xception网络的弱监督细粒度图像分类[J].计算机工程与应用,2022,58(2):235-243.
DING W Q,YU P F,LI H Y,et al.Weakly supervised fine-grained image classification based on Xception network[J].Computer Engineering and Applications,2022,58(2):235-243.

[23] DU R Y,CHANG D L,BHUNIA A K,et al.Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]//Proceedings of the 16th European Conference on Computer Vision,Glasgow,Aug 23-28,2020:153-168.

[24] ZHUANG P Q,WANG Y L,QIAO Y.Learning attentive pairwise interaction for fine-grained classification[C]//

Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Conference on Innovative Application of Artificial Intelligence, the 10th Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020: 13130-13137.

[25] WANG J, YU X H, GAO Y S. Feature fusion vision transformer for fine-grained visual categorization[C]//Proceedings of 32nd British Machine Vision Conference, Nov 22-25, 2021: 170.

[26] CAI C L, ZHANG T K, WENG Z W, et al. A transformer architecture with adaptive attention for fine-grained visual classification[C]//Proceedings of the 7th International Conference on Computer and Communications, 2021: 863-867.

[27] KORSCH D, BODESHEIM P, DENALER J. Classification-specific parts for improving fine-grained visual categorization[C]//Proceedings of the 41th German Conference Pattern Recognition, Dortmund, Germany, Sep 10-13, 2019: 62-75.

[28] ZHANG L B, HUANG S L, TAO D H. Learning a mixture of granularity-specific experts for fine-grained categorization[C]//Proceedings of International Conference on Computer Vision, Seoul, Korea (South), Oct 27-Nov 2, 2019: 8330-8339.

[29] TOUVRON H, VEDALDI A, DOUZE M, et al. Fixing the train-test resolution discrepancy[C]//Proceedings of Annual Conference on Neural Information Processing Systems, Vancouver, Dec 8-14, 2019: 8250-8260.

[30] BEHERA A, WHARTON Z, HEWAGE P R P G, et al. Context-aware attentional pooling (cap) for fine-grained visual classification[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, the 33rd Conference on Innovative Application of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence, Feb 2-9, 2021: 929-937.

[31] KORSCH D, BODESHEIM P, DENZLER J. End-to-end learning of a fisher vector encoding for part features in fine-grained recognition[J]. arXiv: 2007.02080, 2020.

[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015.

[33] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of Conference on Computer Vision and Pattern Recognition, Boston, June 7-12, 2015: 1-9.

[34] MALACH E, SHALEV-SHWARTZ S. Decoupling "when to update" from" how to update"[C]//Proceedings of Annual Conference on Neural Information Processing Systems, Long Beach, Dec 4-9, 2017: 960-970.

[35] HAN B, YAO Q M, YU X R, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels[C]//Proceedings of Annual Conference on Neural Information Processing Systems, Montréal, Dec 3-8, 2018: 8536-8546.

[36] SHU J, XIE Q, YI L X, et al. Meta-Weight-Net: learning an explicit mapping for sample weighting[C]//Proceedings of Annual Conference on Neural Information Processing Systems, Vancouver, Dec 8-14, 2019: 1917-1928.

[37] SHU J, YUAN X, XU Z B. CMW-Net: learning a class-aware sample weighting mapping for robust deep learning[J]. arXiv: 2202.05613, 2022.

[38] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 618-626.