第 60 卷 第 2 期/2023 年 1 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

基于改进Transformer的细粒度图像分类模型

田战胜，刘立波*

宁夏大学信息工程学院，宁夏 银川 750021

摘要 细粒度图像具有不同子类间差异小、相同子类内差异大的特点。现有网络模型在处理过程中存在特征提取能力不足、特征表示冗余和归纳偏置能力弱等问题，因此提出一种改进的Transformer图像分类模型。首先，利用外部注意力取代原Transformer模型中的自注意力，通过捕获样本间相关性提升模型的特征提取能力；其次，引入特征选择模块筛选区分性特征，去除冗余信息，加强特征表示能力；最后，引入融合的多元损失，增强模型归纳偏置和区分不同子类、归并相同子类的能力。实验结果表明，所提方法在CUB-200-2011、Stanford Dogs和Stanford Cars三个细粒度图像数据集上的分类精度分别达89.8%、90.2%和94.7%，优于多个主流的细粒度图像分类方法，分类结果较好。

关键词 细粒度图像分类；Transformer；外部注意力；特征选择；多元损失

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP220453

Fine-Grained Image Classification Model Based on Improved Transformer

Tian Zhansheng, Liu Libo*

School of Information Engineering, Ningxia University, Yinchuan 750021, *Ningxia, China*

Abstract For the characteristics of subtle differences between various subclasses and large differences between same subclasses in a fine-grained image, the existing neural network models have some challenges in processing, including insufficient feature extraction ability, redundant feature representation, and weak inductive bias ability; therefore, an enhanced Transformer image classification model is proposed in this study. First, an external attention is employed to replace the self-attention in the original Transformer model, and the model's feature extraction ability is enhanced by capturing the correlation between samples. Second, the feature selection module is introduced to filter differentiating features and eliminate redundant information to improve feature representation capability. Finally, the multivariate loss is added to improve the model's ability to induce bias, differentiate various subclasses, and fuse the same subclasses. The experimental findings demonstrate that the proposed method's classification accuracy on three fine-grained image datasets of CUB-200-2011, Stanford Dogs, and Stanford Cars reaches 89.8%, 90.2%, and 94.7%, respectively; it is better than that of numerous mainstream fine-grained image classification approaches.

Key words fine-grained image classification; Transformer; external attention; feature selection; multivariate loss

1 引 言

细粒度图像分类作为区分同一父类下不同子类的研究任务，通常用于识别不同种类的鸟、狗、汽车等。细粒度图像具有类间差异小和类内差异大的特点，区分性特征通常存在于局部区域，难以提取和捕获，极具挑战性[1-2]。

针对以上问题，研究人员提出了基于强监督的分类方法，其中极具代表性的有Part R-CNN[3]，该类方法虽然分类精度较高，但过分依赖人工标注信息，缺乏实用性。目前，以B-CNN[4]、RA-CNN[5]、DVAN[6]等为代表的弱监督方法成为主要研究趋势，通过改进卷积神经网络(CNN)模型来提高分类精度。张志刚等[7]改进ResNeXt50[8]，改进后的方法在野生菌分类任务中取得较好的分类结果。王彬州等[9]通过在RA-CNN中引入基于多重注意力机制的方法增强模型的特征提取能力，但该类模型存在因感受野较小无法捕获长距离依赖关系的问题，导致模型的特征提取能力受到限制[10]。Vision Transformer(ViT)模型由Dosovitskiy等[11]于2020年提出，通过自注意力(SA)模块捕获长

收稿日期：2022-01-05；**修回日期**：2022-02-28；**录用日期**：2022-03-14；**网络首发日期**：2022-03-24

基金项目：宁夏自然科学基金(2020AAC03031)、国家自然科学基金(61862050)

通信作者：*liulib@163.com

距离依赖关系，提取图像全局特征，分类准确率明显得到提高。但ViT模型只能捕获单个图像样本内像素间的相关性，导致输出特征提取能力不足，且参数量较大[12]。此外，ViT模型使用末层Transformer输出的class patch作为最终特征表示，存在大量冗余，导致区分性特征表示能力不佳[13]。虽然ViT克服了CNN无法捕获长距离依赖的缺点，但其归纳偏置的能力较弱[14]。

综上，本文以ViT为基础，提出基于改进Transformer的细粒度图像分类模型（TransFC）。主要贡献：采用外部注意力（EA）模块[12]代替自注意力模块，同时捕获单个样本内的长距离依赖关系和样本之间的潜在相关性，增强特征表示能力的同时降低原模型参数量；在模型当中引入特征选择（FS）模块[13]，在提取并融合区分性区域特征的同时去除冗余特征；引入一种融合的多元损失[13-14]，以扩大不同子类差异，缩小相同子类差异，并使模型具有归纳偏置的能力。在3个公用细粒度数据集上通过与原模型及主流弱监督分类方法进行对比实验，结果表明所提方法具有较好的分类结果。

2　Vision Transformer概述

Vision Transformer采用多层Transformer架构完成特征提取过程，每层内部均使用自注意力作为特征函数，并利用后层Tansformer对前层特征函数的输出进行特征细化，逐渐捕获到图像全局特征。该模型架构如图1所示。

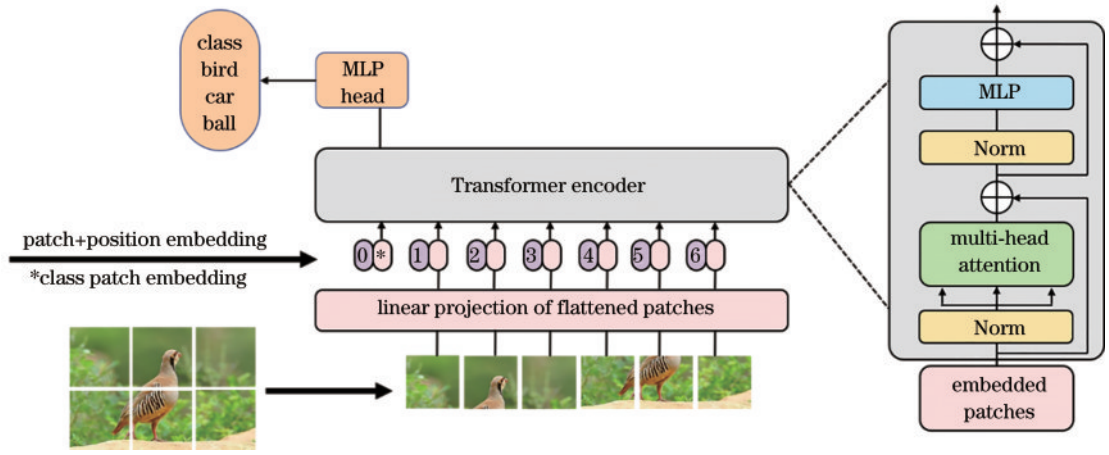

图1　Vision Transformer架构

Fig. 1　Framework of Vision Transformer

首先，采用不重叠方式将图像$\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$划分为patch序列$\boldsymbol{x}_{\mathrm{p}}\in\mathbb{R}^{N\times(P\times P\times C)}$，$H\times W$为图像的分辨率，$C$为通道数，$P\times P$和$N=HW/P^2$分别为patch的分辨率和数量；之后，利用可学习线性映射向量$\boldsymbol{E}\in\mathbb{R}^{(P\times P\times C)\times D}$将每个patch映射到$D$维空间，再将分类向量$\boldsymbol{x}_{\mathrm{class}}\in\mathbb{R}^{1\times D}$添加到patch序列首部，用于集成全局特征；最后，利用$\boldsymbol{E}_{\mathrm{pos}}\in\mathbb{R}^{(N+1)\times D}$为每个patch赋予位置特征后得到首层Transformer的输入。首层Transformer的输入为

$$\boldsymbol{Z}_0=\left[\boldsymbol{x}_{\mathrm{class}};\ \boldsymbol{x}_{\mathrm{p}}^1\boldsymbol{E};\ \boldsymbol{x}_{\mathrm{p}}^2\boldsymbol{E};\ \cdots;\ \boldsymbol{x}_{\mathrm{p}}^N\boldsymbol{E}\right]+\boldsymbol{E}_{\mathrm{pos}}。\tag{1}$$

$\boldsymbol{Z}_0$输入第一层Transformer后，分别经具有残差结构[15]的多头自注意力（MSA）模块和多层感知机（MLP）模块进行特征提取，数据在输入这两个模块之前均利用Layernorm（LN）进行标准化。Transformer的内部结构如图1右侧所示。

为提取更为精细有效的特征，ViT采用多层Transformer架构细化前层的输出特征，即

$$\boldsymbol{Z}'_\ell=\mathrm{MSA}\left[\mathrm{LN}\left(\boldsymbol{Z}_{\ell-1}\right)\right]+\boldsymbol{Z}_{\ell-1},\ \ell=1,\cdots,L,\tag{2}$$

$$\boldsymbol{Z}_\ell=\mathrm{MLP}\left[\mathrm{LN}\left(\boldsymbol{Z}'_\ell\right)\right]+\boldsymbol{Z}'_\ell,\ \ell=1,\cdots,L,\tag{3}$$

式中：ℓ为层数；$\boldsymbol{Z}'_\ell$和$\boldsymbol{Z}_\ell$分别为数据经第ℓ层MSA和MLP模块后的结果。通过在多层Transformer中进行流动处理，输入图像的全局特征被逐渐精细化并聚合到class patch中。因此，对末层输出的class patch（$\boldsymbol{Z}_L^0$）进行LN处理，即$\boldsymbol{y}=\mathrm{LN}\left(\boldsymbol{Z}_L^0\right)$，得到最终的全局特征$\boldsymbol{y}$，将$\boldsymbol{y}$输入分类器中进行分类预测、损失计算、反向传播，最终完成模型的训练。

3　所提方法

3.1　TransFC模型整体结构

使用滑动窗口取代ViT中的不重叠方式，生成图像patch序列，以解决分割边缘特征后难以提取的问题。设图像分辨率为$H\times W$，patch数量N计算方式为

$$N=\frac{H-P+S}{S}\times\frac{W-P+S}{S}。\tag{4}$$

TransFC的整体架构如图2所示。首先在patch序列末尾添加一个$\boldsymbol{x}_{\mathrm{dis}}\in\mathbb{R}^{1\times D}$用于计算多元损失；然后将每个Transformer层内部的自注意力替换为外部注意力；再采用特征选择模块对末层Transformer的输入进行筛选，去除冗余特征；最后利用末层Transformer的输出从多方面计算损失并融合。

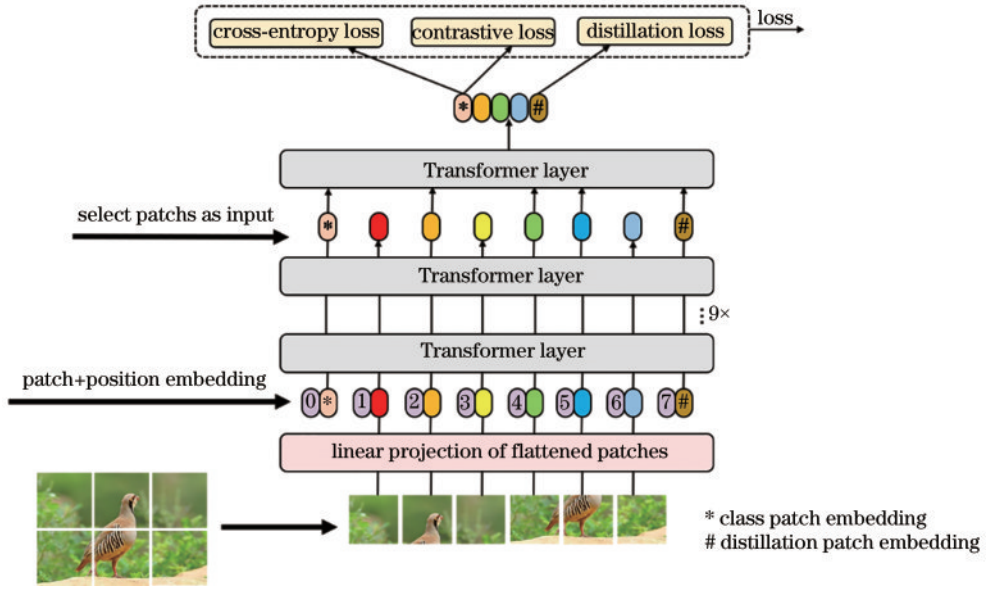

图 2　TransFC 的整体架构

Fig. 2　Framework of TransFC

3.2　引入外部注意力

自注意力机制作为 ViT 中的主要特征提取方法，详细架构如图 3 所示。首先，将输入特征图 $\boldsymbol{F}\in\mathbb{R}^{N'\times d}$ 线性映射为 $\boldsymbol{Q}_{\text{query}}\in\mathbb{R}^{N'\times d'}$、$\boldsymbol{K}_{\text{key}}\in\mathbb{R}^{N'\times d'}$ 和 $\boldsymbol{V}_{\text{value}}\in\mathbb{R}^{N'\times d}$，其中 N' 为像素数量，d 为特征图维度，并利用 $\boldsymbol{Q}_{\text{query}}$ 和 $\boldsymbol{K}_{\text{key}}$ 计算得到注意力权重矩阵，具体计算过程为

$$\boldsymbol{A}=(\alpha)_{i,j}=\text{Softmax}\left(\boldsymbol{Q}_{\text{query}}\otimes\boldsymbol{K}_{\text{key}}{}^{\text{T}}\right),\tag{5}$$

式中：$\boldsymbol{A}\in\mathbb{R}^{N'\times N'}$ 为注意力权重图；$(\alpha)_{i,j}$ 为第 i 个像素和第 j 个像素之间的相似度。对 $\boldsymbol{A}$ 和 $\boldsymbol{V}_{\text{value}}$ 进行矩阵相乘，再和 $\boldsymbol{F}$ 进行残差连接后即得到最终的输出特征，表达式为

$$\boldsymbol{F}_{\text{out}}=\boldsymbol{F}\oplus\boldsymbol{A}\otimes\boldsymbol{V}_{\text{value}},\ \boldsymbol{F}_{\text{out}}\in\mathbb{R}^{N'\times d}。\tag{6}$$

以上计算过程中，式(5)利用枚举的方式计算同一样本内像素点之间的相关性，忽略了样本间的潜在相关性，导致模型特征提取能力不足[12]；单个样本内大多数像素点只和其他少数像素点之间有相关性，枚举的计算方式造成大量冗余计算，导致模型参数量较大[16]。

为解决自注意力存在的问题，本文引入 Guo 等提出的具有线性结构的外部注意力[12]，通过两个可学习的外部记忆单元使模型可以同时捕获样本内和样本间相关性，增强模型特征提取能力，同时减少模型参数量，详细结构如图 4 所示。首先，将输入特征图 $\boldsymbol{F}\in\mathbb{R}^{N'\times d}$ 映射为向量 $\boldsymbol{Q}_{\text{E}}\in\mathbb{R}^{N'\times d'}$，之后利用一个可学

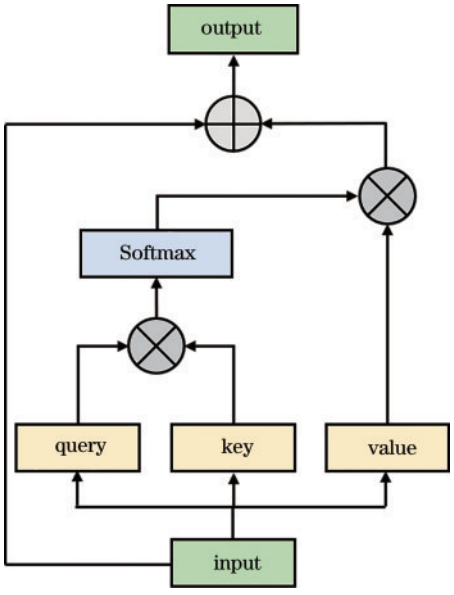

图 3　自注意力模块

Fig. 3　Self-attention module

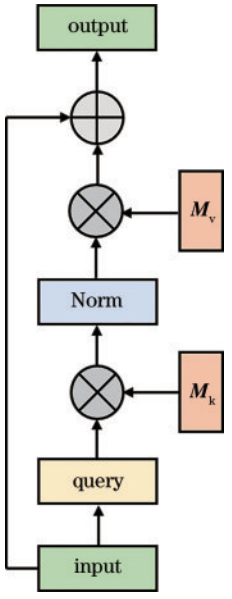

图 4　外部注意力模块

Fig. 4　External-attention module

习的外部记忆单元 $\boldsymbol{M}_{\mathrm{k}}\in\mathbb{R}^{S\times d'}$ 与 $\boldsymbol{Q}_{\mathrm{E}}$ 相乘，对结果进行正则化处理后得到注意力权重图 $\boldsymbol{A}_{\mathrm{E}}$，表达式为

$$\boldsymbol{A}_{\mathrm{E}}=\mathrm{Norm}\left(\boldsymbol{Q}_{\mathrm{E}}\otimes\boldsymbol{M}_{\mathrm{k}}^{\mathrm{T}}\right),\tag{7}$$

然后，使用 $\boldsymbol{A}_{\mathrm{E}}$ 与另一个外部记忆组件 $\boldsymbol{M}_{\mathrm{v}}\in\mathbb{R}^{S\times d}$ 联合计算出一个更为精细的特征图，再与输入特征进行残差操作，得到最终的输出结果 $\boldsymbol{F}_{\mathrm{out}}$，表达式为

$$\boldsymbol{F}_{\mathrm{out}}=\boldsymbol{F}\oplus\boldsymbol{A}_{\mathrm{E}}\otimes\boldsymbol{M}_{\mathrm{v}}\ 。\tag{8}$$

3.3 引入特征选择模块

在 TransFC 采用滑动窗口方式生成的 patch 序列中，有些 patch 只包含背景信息或少部分前景对象，当滑动窗口步长减小时，类似 patch 还会增多。由于细粒度图像区分性特征一般存在于具有细微差异的局部区域，而大量 patch 内缺乏有效信息会导致输出特征存在冗余，因此，为去除冗余，引入特征选择模块来提取区分性区域特征。

对于原始 patch 序列，在没有任何注意力权重信息的情况下无法选取区分性 patch。特征选择模块根据前 $L-1$ 层的注意力权重筛选出第 $L-1$ 层输出的区分性特征，并将其作为第 L 层的输入进一步细化特征。设第 $L-1$ 层 Transformer 输出为

$$\boldsymbol{Z}_{L-1}=\left[\boldsymbol{Z}_{L-1}^{0};\boldsymbol{Z}_{L-1}^{1},\boldsymbol{Z}_{L-1}^{2},\cdots,\boldsymbol{Z}_{L-1}^{N};\boldsymbol{Z}_{L-1}^{N+1}\right],\tag{9}$$

前 $L-1$ 层的注意力权重可表示为

$$\boldsymbol{a}_{\ell}=\left[\boldsymbol{a}_{\ell}^{0},\boldsymbol{a}_{\ell}^{1},\boldsymbol{a}_{\ell}^{2},\cdots,\boldsymbol{a}_{\ell}^{K}\right],\ \ell\in1,2,\cdots,L-1,\tag{10}$$

$$\boldsymbol{a}_{\ell}^{k}=\left[\boldsymbol{a}_{\ell}^{k_0};\boldsymbol{a}_{\ell}^{k_1},\cdots,\boldsymbol{a}_{\ell}^{k_N};\boldsymbol{a}_{\ell}^{k_{N+1}}\right],k\in0,1,\cdots,K,\tag{11}$$

式中：K 为注意力头的数量。特征选择模块使用连乘操作整合前 $L-1$ 层的注意力权重，即

$$\boldsymbol{a}_{\mathrm{final}}=\prod_{\ell=0}^{L-1}\boldsymbol{a}_{\ell},\tag{12}$$

式中：$\boldsymbol{a}_{\mathrm{final}}$ 记录了注意力权重由第 1 层到第 $L-1$ 层的传递过程。之后从中分别选取 k 个注意力头的最大值对应的索引 $\left[A_1,A_2,\cdots,A_K\right]$，并根据索引选择输入到第 L 层的特征，被选中的序列可表示为

$$\boldsymbol{Z}_{\mathrm{select}}=\left[\boldsymbol{Z}_{L-1}^{0};\boldsymbol{Z}_{L-1}^{A_1},\boldsymbol{Z}_{L-1}^{A_2},\cdots,\boldsymbol{Z}_{L-1}^{A_K};\boldsymbol{Z}_{L-1}^{N+1}\right]。\tag{13}$$

$\boldsymbol{Z}_{\mathrm{select}}$ 作为末层输入，舍弃了大量从背景区域提取到的无效特征，避免了模型最终输出特征存在冗余的问题。

3.4 融合的多元损失函数

针对细粒度图像不同类间差异小、相同类内差异大的特点和 Transformer 模型偏置归纳能力弱的问题，从多角度对损失函数进行优化，提出融合的多元损失函数。

ViT 使用的交叉熵损失可捕获到比较明显的类间差异，但缺少对类间细微差异和类内差异的捕获能力；而对比损失（contrastive loss）可以在增大不同子类特征差异的同时减小相同子类特征差异。本文保留 ViT 利用末层 Transformer 输出的 class patch 计算的交叉熵损失，同时利用该 patch 计算对比损失，计算过程可表示为

$$\mathcal{L}_{\mathrm{con}}=\frac{1}{N_{\mathrm{B}}^2}\left\{\sum_{j:y_i=y_j}^{N_{\mathrm{B}}}\left[1-\cos\left(\boldsymbol{Z}_i,\boldsymbol{Z}_j\right)\right]+\sum_{j:y_i\neq y_j}^{N_{\mathrm{B}}}\max\left\{\left[\cos\left(\boldsymbol{Z}_i,\boldsymbol{Z}_j\right)-\alpha\right],0\right\}\right\},\tag{14}$$

式中：N_{B} 为 batch size 的大小；$\boldsymbol{Z}_i$ 为第 i 个图像经过 TransFC 后输出的 calss patch，也是最终的特征表示；$\cos\left(\boldsymbol{Z}_i,\boldsymbol{Z}_j\right)$ 表示 $\boldsymbol{Z}_i$ 和 $\boldsymbol{Z}_j$ 的余弦相似度，其大于超参数 α 时才会在对比损失中起作用。$\mathcal{L}_{\mathrm{con}}$ 经过反向传播可以扩大不同子类别之间的特征表示，缩小相同子类别内的特征表示，缓解了类间差异小和类内差异大造成的分类困难问题。

其次，由于归纳偏置能力是影响 Transformer 模型特征提取能力的关键因素，而 CNN 模型具有较强的偏置归纳能力，因此利用 CNN 引入蒸馏损失（distillation loss）[14]，使得 TransFC 能够从 CNN 中学习到归纳偏置能力，以进一步提升模型的特征提取能力。

Hinton 等[17]提出的知识蒸馏（knowledge distillation）是一种将知识从 teacher 模型转移到 student 模型的训练策略，其联合二者 Softmax 层输出的预测标签共同计算出蒸馏损失，实现知识迁移。本文引入蒸馏损失作为总损失的一部分，在输入的 patch 序列后增加一个 distillation patch，与 class patch 类似，distillation patch 在多个 Transformer 层内与其他 patch 相互作用，最终聚合图像的特征表示；但与 class patch 不同，distillation patch 的目标是再现 teacher 模型输出的预测标签，而不是真实标签。联合依据 distillation patch 计算得到的标签与 teacher 模型（CNN）的输出标签，通过计算二者之间 Kullback-Leibler（KL）散度的方式得到蒸馏损失，作为总损失的一部分，指导 student 模型（TransFC）进行反向传播，具体计算方法为

$$\mathcal{L}_{\mathrm{dis}}=\tau^2\mathrm{KL}\left[\psi\left(\frac{\boldsymbol{Z}_{\mathrm{s}}}{\tau}\right),\psi\left(\frac{\boldsymbol{Z}_{\mathrm{t}}}{\tau}\right)\right],\tag{15}$$

式中：$\boldsymbol{Z}_{\mathrm{s}}$ 为利用 distillation patch 进行分类时的 logist 函数输出；$\boldsymbol{Z}_{\mathrm{t}}$ 为 teacher 模型的 logist 函数输出；$\psi(\cdot)$ 表示 Softmax 函数；τ 表示蒸馏温度，使 Softmax 层的输出的概率分布更加接近。综上，从三个角度分别计算交叉熵损失、对比损失和蒸馏损失后进行融合，帮助模型区分不同子类差异，归并相同子类差异，并赋予模型归纳偏置能力，使输出特征更加精细化，更有区分性。因交叉熵损失和对比损失是利用同一个 class patch 计算得到的，所以融合过程中将二者之和视为总损失的一部分，蒸馏损失视为另一部分。具体的融合方式为

$$\mathcal{L}_{\mathrm{global}}=(1-\lambda)\left[\mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{y}',\boldsymbol{y}\right)+\mathcal{L}_{\mathrm{con}}(\boldsymbol{Z})\right]+\lambda\mathcal{L}_{\mathrm{dis}},\tag{16}$$

式中：$\mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{y}',\boldsymbol{y}\right)$ 为 class patch 的预测标签 $\boldsymbol{y}'$ 和真实标签 $\boldsymbol{y}$ 之间的交叉熵损失；λ 为超参数。

4 实验结果与分析

系统实验环境为 Ubuntu 16.04，软件环境为 cuda9.0 和 python3.6，硬件配置为 NVIDIA Quadro P500 GPU、16 GB 显存、Intel Xeon E5-2620 CPU。选用深度学习框架 Pytorch 搭建网络模型。

4.1 数据集选取与预处理

为充分验证所提方法的有效性，选取 CUB-200-2011[18]、Stanford Cars[19]和 Stanford Dogs[20]3 个公用细粒度图像数据集进行实验对比，三个数据集的子类数量和训练集、测试集的划分如表 1 所示。

表 1 细粒度图像数据集的详细信息

Table 1 Detailed information of fine-grained image datasets

Dataset	Number of subclasses	Number of samples on training set	Number of samples on test set
CUB-200-2011	200	5994	5794
Stanford Cars	120	8144	8041
Stanford Dogs	196	12000	8580

为避免模型因数据量过少而出现过拟合的现象，采用水平翻转、垂直翻转和 AutoAugment[21]的方式对数据进行扩充。

4.2 评价指标与实验设置

选用准确率(accuracy)作为评价指标，表达式为

$$A=\frac{I_{\mathrm{ac}}}{I_{\mathrm{total}}}, \tag{17}$$

式中：I_{ac} 为正确分类的图像数量；I_{total} 为测试集图像总数量。

Dosovitskiy 等[11]在提出 ViT 模型时通过大量实验证明，当 Transformer 模型层数达 12 时，继续增加层数并不能明显提升模型分类准确率，却大大增加了模型参数量，因此本文采用 12 层 Transformer 架构。patch 的数量也是影响模型参数量的一个重要因素，其数量与 patch 大小成反比，与输入图像的分辨率成正比。为保证 TransFC 模型具有充足的数据输入量，并避免模型因参数量较大在训练阶段不易收敛，使用 448×448 分辨率的输入图像，在训练阶段采用随机裁剪，测试阶段采用中心裁剪，保留原 ViT 模型 16×16 的 patch 大小，同时将滑动窗口的步长设置为 12。模型训练阶段，均采用加载 ImageNet 预训练参数的方式进行微调，将对比损失中的超参数 α 设置为 0.4，使用随机梯度下降法(SGD)作为优化方法，动量(momentum)设置为 0.9，batch size 为 32。考虑到 Stanford Dogs 数据集的训练集相较于其他两个数据集较多，在 Stanford Dogs 数据集上训练时将学习率初始化为 0.003，而在其他两个数据集上将学习率初始化为 0.03，采用余弦退火(cosine annealing)控制学习率的下降幅度。

4.3 消融实验

为验证各模块的有效性，采用向 ViT 中逐步融入各改进模块的方式在 CUB-200-2011 数据集上展开消融实验分析。

4.3.1 外部注意力和特征选择模块的实验分析

采用交叉熵损失函数计算损失，对比 4 种 Transformer 网络结构：1)用 ViT(baseline)表示原 Transformer 模型；2)将 1)中自注意力替换为外部注意力模块，表示为 ViT(EA)；3)在 1)中添加特征选择模块，表示为 ViT(FS)；4)将 1)中自注意力替换为外部注意力，并同时添加特征选择模块，表示为 ViT(EA&FS)。具体实验结果如表 2 所示。

表 2 外部注意力和特征选择模块的消融实验分析

Table 2 Ablation experiment analysis of MEA and FS module

No.	Method	Model composition	Accuracy /%
1)	ViT(baseline)	SA	85.8
2)	ViT(EA)	EA	86.9
3)	ViT(FS)	SA+FS	86.6
4)	ViT(EA&FS)	EA+FS	87.9

从表 2 的结果可以看出：选用交叉熵损失函数计算损失，只替换注意力模块的网络模型、只引入特征选择模块的网络模型、替换注意力模块的同时添加特征选择模块的网络模型相比原 Transformer 模型的分类准确率分别提升 1.1 个百分点、0.8 个百分点、2.1 个百分点，既替换注意力模块又添加特征选择模块后的模型的分类准确率达到最高。

4.3.2 融入多元损失的实验分析

在利用交叉熵损失的基础上，为了进一步验证对比损失在模型中的有效性，将对比损失分别融入 1) ViT(baseline)和 4) ViT(EA&FS)，得到 5) ViT(SA&L_CON)和 6) ViT(EA&FS&L_CON)。利用交叉熵损失函数和对比损失函数联合计算出模型总损失，具体实验结果如表 3 所示。

表 3 对比损失的消融实验分析

Table 3 Ablation experiment analysis of contrastive loss

No.	Method	Model composition	Accuracy /%
1)	ViT(baseline)	SA	85.8
5)	ViT(SA&L_CON)	SA+L_CON	86.2
4)	ViT(EA&FS)	EA+FS	87.9
6)	ViT(EA&FS&L_CON)	EA+FS+L_CON	88.2

从表 3 可以发现：5）ViT（SA&L_CON）和 6）ViT（EA&FS&L_CON）模型的准确率相较对比损失融入前分别提升了 0.4 个百分点和 0.3 个百分点，表明对比损失可使模型增强区分不同子类、归并相同子类的能力。

为进一步验证引入蒸馏损失后模型的有效性，在 6）ViT（EA&FS&L_CON）的基础上，分别选用 VGG-16[22]、ResNet-50[15]、ResNet-101[15]和 DenseNet-121[23] 4 个网络层数逐渐加深的经典 CNN 骨干网络作为 teacher 模型，指导 6）完成训练，此时的模型总损失由交叉熵损失函数、对比损失函数和蒸馏损失函数共同计算得出。具体实验结果如表 4 所示。至此，注意力模块的替换、特征选择模块的引入和多元损失的融合已全部完成。

表 4 蒸馏损失的消融实验分析

Table 4 Ablation experiment analysis of distillation loss

Method	Model composition	Teacher	Accuracy /%
ViT(EA&FS&L_CON)	EA+FS+L_CON		88.2
TransFC(EA&FS&L_CON&L_DIS1)	EA+FS+L_CON+L_DIS	VGG-16	88.9
TransFC(EA&FS&L_CON&L_DIS2)	EA+FS+L_CON+L_DIS	ResNet-50	89.1
TransFC(EA&FS&L_CON&L_DIS3)	EA+FS+L_CON+L_DIS	ResNet-101	89.5
TransFC(EA&FS&L_CON&L_DIS4)(ours)	EA+FS+L_CON+L_DIS	DenseNet-121	89.8

从表 4 可以发现：在 6）ViT（EA&FS&L_CON）的基础上引入蒸馏损失后，随着 teacher 网络层数不断加深，TransFC 模型的分类准确率也逐渐提高，表明利用 CNN 指导 TransFC 模型训练可以增强模型偏置归纳能力。当选择 DenseNet-121 作为 teacher 网络时，相较其他 3 个 CNN 模型，效果最好，分类精度达 89.8%，比融入蒸馏损失前提升 1.6 个百分点。因此选用 TransFC（EA&FS&L_CON&L_DIS4）为最终模型，以下简称 TransFC（ours）。

4.4 对比实验

为进一步验证所提方法的优越性，选取 DB[24]、SEF[25]、B-CNN[4]、WS-DAN[26]、ACNet（VGG-16）[27]、ACNet（ResNet-50）[27]、MAMC[28]、DVAN[6]、RA-CNN[5]、MC Loss[29]和 ViT[11]等主流弱监督方法在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 三个公用数据集上进行实验，并与所提方法进行对比，结果如表 5 所示。

由表 5 的实验结果可知：所提方法在 CUB-200-2011 和 Stanford Dogs 两个数据集上的分类精度均优于主流方法；在 Stanford Cars 上的分类精度达 94.7%，优于大多数主流方法，与 DB[21]、WS-DAN[23]方法持平，相较于原 ViT 模型有 2.1 个百分点的提升。结果表明，引入外部注意力、特征选择模块、对比损失和蒸馏损失后的 Transformer 模型具有较强的局部区分性特征提取能力，并在多个细粒度图像数据集上取得较好的分类结果。

4.5 TransFC 可视化

从 3 个细粒度数据集中分别选取 2 幅样例图进行 TransFC 可视化：CUB-200-2011（bird1，bird2）、Stanford Dogs（dog1，dog2）、Stanford Cars（car1，car2）。

表 5 不同弱监督细粒度图像分类方法的实验对比

Table 5 Experiment comparison of different weakly supervised fined-grained image classification methods

Method	Base model	Accuracy /%		
		CUB-200-2011	Stanford Dogs	Stanford Cars
DB[24]	ResNet-50	88.6	87.7	94.9
SEF[25]	ResNet-50	87.3	88.8	94.0
B-CNN[4]	VGG-16	84.1		91.3
WS-DAN[26]	Inception V3	89.4	90.0	94.1
ACNet[27]	VGG-16	87.8		94.3
ACNet[27]	ResNet-50	88.1		94.6
MAMC[28]	ResNet-101	86.5	85.2	93.0
DVAN[6]	VGG-16	79.0	81.5	87.1
RA-CNN[5]	VGG-19	85.5	87.3	92.5
MC Loss[29]	ResNet-50	87.3		93.7
ViT[11]	SA	85.8	87.2	92.6
TransFC(ours)	EA	89.8	90.2	94.7

结果如图 5 所示，第一行为待识别的原图像，第二行为 TransFC 模型生成的热图。从图 5 可以看出：模型的注意力权重主要集中在图像的前景对象，避免了图像背景造成输出特征冗余的问题；权重最大的部分主要集中在具有区分性的局部位置，如鸟类的头部、眼睛和腹部，狗类的鼻子、眼睛、耳朵和腿部，汽车的车轮、车灯和车标，表明模型能聚焦区分性区域并且捕获特征的能力较强。

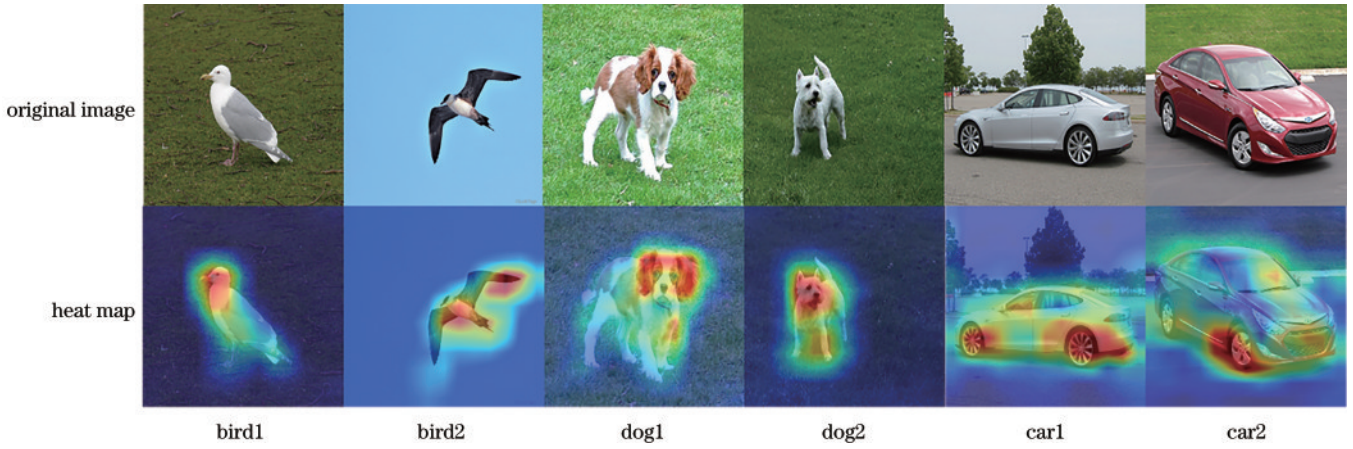

图 5　TransFC 模型在 3 个数据集上的可视化效果

Fig. 5　Visualization of TransFC model on three datasets

5　结　　论

在 ViT 的基础上，针对细粒度图像分类的特点和 Transformer 网络特征提取能力不足、特征表示冗余、归纳偏置能力弱等问题，提出基于改进 Transformer 的细粒度图像分类模型。采用外部注意力替换原有的特征提取方法，以捕获样本间和样本内的相关性，进而提升模型特征提取能力，同时降低模型的参数量；引入特征选择模块来去除冗余特征，使最终的特征表示更加精细；引入多元损失加强模型的偏置归纳能力，并增强模型区分不同子类、归并相同子类的能力，使模型更适用于细粒度图像分类任务。实验结果表明，所提方法在多个细粒度数据集上均具有较高的分类精度，优于多个主流的细粒度分类方法。通过蒸馏学习的方式，利用 CNN 指导 Transformer 模型训练的方式比较繁琐，在未来工作中，将 CNN 直接融入 Transformer 模型是后续工作的方向。

参　考　文　献

[1] 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 on fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica，2017，43(8)：1306-1318.

[2] Wei X S，Wu J X，Cui Q. Deep learning for fine-grained image analysis：a survey[EB/OL]. (2019-06-06)[2021-08-06]. https://arxiv.org/abs/1907.03069.

[3] Zhang N，Donahue J，Girshick R，et al. Part-based R-CNNs for fine-grained category detection[M]//Fleet D，Pajdla T，Schiele B，et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham：Springer，2014：934-849.

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

[5] 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，July 21-26，2017，Honolulu，HI，USA. New York：IEEE Press，2017：4476-4484.

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

[7] 张志刚，余鹏飞，李海燕，等. 基于多尺度特征引导的细粒度野生菌图像识别[J]. 激光与光电子学进展，2022，59(12)：1210016.
Zhang Z G，Yu P F，Li H Y，et al. Fine-grained image recognition of wild mushroom based on multiscale feature guide[J]. Laser & Optoelectronics Progress，2022，59(12)：1210016.

[8] Xie S N，Girshick R，Dollár P，et al. Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition，July 21-26，2017，Honolulu，HI，USA. New York：IEEE Press，2017：5987-5995.

[9] 王彬州，肖志勇. 面向细粒度图像识别的通道注意力多分支网络[J]. 激光与光电子学进展，2021，58(22)：2210008.
Wang B Z，Xiao Z Y. Channel attention multi-branch network for fine-grained image recognition[J]. Laser & Optoelectronics Progress，2021，58(22)：2210008.

[10] 王嘉楠，高越，史骏，等. 基于视觉转换器和图卷积网络的光学遥感场景分类[J]. 光子学报，2021，50(11)：1128002.
Wang J N，Gao Y，Shi J，et al. Scene classification of optical high-resolution remote sensing images using vision transformer and graph convolutional network[J]. Acta Photonica Sinica，2021，50(11)：1128002.

[11] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16×16 words：transformers for image recognition at scale[EB/OL]. (2020-10-22)[2021-05-08]. https://arxiv.org/abs/2010.11929.

[12] Guo M H，Liu Z N，Mu T J，et al. Beyond self-

attention: external attention using two linear layers for visual tasks[EB/OL]. (2021-05-05)[2021-08-06]. https://arxiv.org/abs/2105.02358.

[13] He J, Chen J N, Liu S, et al. TransFG: a transformer architecture for fine-grained recognition[EB/OL]. (2021-03-14)[2021-08-09]. https://arxiv.org/abs/2103.07976.

[14] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[EB/OL]. (2020-12-23) [2021-09-08]. https://arxiv.org/abs/2012.12877.

[15] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[16] Cao Y, Xu J R, Lin S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1971-1980.

[17] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2021-05-08]. https://arxiv.org/abs/1503.02531.

[18] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset[R]. Pasadena: California Institute of Technology, 2011.

[19] Krause J, Stark M, Jia D, et al. 3D object representations for fine-grained categorization[C]//2013 IEEE International Conference on Computer Vision Workshops, December 2-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 554-561.

[20] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[EB/OL]. [2022-01-04]. https://people.csail.mit.edu/khosla/papers/fgvc2011.pdf.

[21] Cubuk E D, Zoph B, Mane D, et al. AutoAugment: learning augmentation policies from data[EB/OL]. (2018-05-24)[2021-05-06]. https://arxiv.org/abs/1805.09501.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-08-09]. https://arxiv.org/abs/1409.1556.

[23] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[24] Sun G L, Cholakkal H, Khan S, et al. Fine-grained recognition: accounting for subtle differences between similar classes[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12047-12054.

[25] Luo W, Zhang H M, Li J, et al. Learning semantically enhanced feature for fine-grained image classification[J]. IEEE Signal Processing Letters, 2020, 27: 1545-1549.

[26] Hu T, Qi H. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification[EB/OL]. (2019-01-26)[2021-05-08]. https://arxiv.org/abs/1901.09891.

[27] Ji R Y, Wen L Y, Zhang L B, et al. Attention convolutional binary neural tree for fine-grained visual categorization[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10465-10474.

[28] Sun M, Yuan Y C, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11220: 805-821.

[29] Chang D, Ding Y, Xie J, et al. The devil is in the channels: mutual-channel loss for fine-grained image classification[J]. IEEE Transactions on Image Processing, 2020, 29: 4683-4695.