中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-16

论文引用格式:Ren Yuehe, Liu Bo, Chi Jing, Zhang Yunfeng, Xing Yinan. Multi-dimensional expert routing transformer for fine-grained image classification[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260565.(任玥赫, 刘博, 迟静, 张云峰, 邢怡楠. 多维专家路由Transformer细粒度图像分类[J/OL]. 中国图象图形学报,XXXX:1-16. DOI: 10.11834/jig.260565.)[DOI:10.11834/jig.260565]

多维专家路由Transformer细粒度图像分类

任玥赫,刘博,迟静,张云峰,邢怡楠

山东财经大学计算机与人工智能学院,济南　250014

摘　要:**目的**　细粒度图像分类旨在区分同一大类下视觉相似的不同子类别,其关键是在类内变化显著、类间差异细微的条件下提取局部判别特征。现有方法面临两项瓶颈:一是采用固定方式融合不同层的特征,忽略了不同图像对浅层细节和深层语义的需求差异;二是未能有效区分不同注意力头对判别特征建模的贡献差异,导致在特征聚合过程中引入冗余信息和背景干扰。针对上述问题,提出了一种多维专家路由Transformer(multi-dimensional expert routing transformer,MERT),以提高模型对关键局部判别特征的表达能力。**方法**　设计层级专家路由器(layer expert router,LER),利用全局语义对不同层表征进行差异化加权,形成跨层聚合特征;提出注意力头专家路由器(head expert router,HER),对注意力头响应进行稀疏重加权,抑制冗余与背景干扰;构建判别性Token选择(discriminative token selection,DTS)模块,融合层级与注意力头路由信息筛选关键Token,突出局部判别线索;构建跨层聚合与局部精炼双分支,并引入头部负载平衡约束,以增强特征判别性并促进不同注意力头的均衡利用。**结果**　本方法在Stanford Dogs、NABirds和Oxford 102 Flowers上的Top-1准确率分别达到92.97%、90.94%和99.65%,其中在Stanford Dogs和Oxford 102 Flowers上取得最优结果。相较ViT-B/16基线,参数量和FLOPs仅小幅增加,在三个数据集上的准确率分别提高2.34、1.25和0.24个百分点。消融实验评估了LER、HER与DTS模块的作用,完整模型在所比较的配置中取得最佳分类性能。**结论**　MERT能够减少冗余信息与背景噪声对关键局部特征的干扰,在仅增加少量参数量和计算量的情况下提高细粒度图像分类精度。

关键词:细粒度图像分类;视觉Transformer;多维专家路由;跨层特征聚合;负载均衡

Multi-dimensional expert routing transformer for fine-grained image classification

Ren Yuehe, Liu Bo, Chi Jing, Zhang Yunfeng, Xing Yinan

School of Computing and Artificial Intelligence, Shandong University of Finance and Economics, Jinan, 250014, China

Abstract: **Objective** Fine-grained image classification aims to distinguish visually similar subcategories within the same superordinate category. Its key challenge lies in identifying local discriminative features under large intra-class variations and subtle inter-class differences. Methods based on the vision transformer (ViT) produce hierarchical representations across layers and diverse attention responses across heads. However, layerwise representations are commonly combined using fixed fusion schemes, although the relative importance of low-level texture, part structure, and high-level semantics varies across images. Second, attention-head responses are often treated uniformly despite their different contributions to

收稿日期:2026-09-10;**修回日期**:2026-11-18

基金项目:国家自然科学基金区域创新发展联合基金(基金号:U25A20438)、山东省自然科学基金(基金号:ZR2025QC1629)

Supported by: National Natural Science Foundation of China (NSFC) Joint Fund for Regional Innovation and Development (Grant No. U25A20438); Natural Science Foundation of Shandong Province, China (Grant No. ZR2025QC1629)

© 中国图象图形学报版权所有

individual images. Consequently, redundant responses and background interference may be retained during feature aggregation, diluting the contribution of local discriminative features. To address these issues, we propose a multi-dimensional expert routing transformer (MERT). MERT adaptively organizes the layerwise representations and multi-head attention responses generated by the ViT backbone according to each input image, thereby improving the utilization of local discriminative information. **Method** MERT adopts ViT-B/16 pretrained on ImageNet-21K as its backbone. For each input image, the ViT backbone first generates token representations and multi-head self-attention responses from all Transformer layers. The class tokens from these layers are then stacked to form a shared routing signal. A layer expert router (LER) estimates image-dependent routing weights from the shared signal to differentiate the contributions of different Transformer layers. The resulting weights aggregate token representations across layers to construct the cross-layer representation. Using the same routing signal, a head expert router (HER) estimates the relative contributions of attention heads and performs sparse Top-k selection within each layer. A straight-through estimator preserves sparse selection during forward propagation while passing gradients to the head-scoring network. A head load-balancing loss is further introduced to mitigate excessive concentration of the routing distribution on a small subset of attention heads. Based on the two types of routing weights, a discriminative token selection (DTS) module combines the layer and head routing weights with the class-token-to-image-token attention scores to estimate the importance of each spatial token. The selected spatial positions are shared across layers, and tokens at the corresponding locations are aggregated using the layer routing weights to integrate local details with higher-level semantics while preserving spatial correspondence. The resulting local tokens are combined with the final-layer class token and refined by reusing the last Transformer layer. MERT therefore produces a cross-layer aggregated representation and a locally refined representation, whose predictions are fused for final classification. During training, an auxiliary classification loss is applied to the locally refined prediction, while a head load-balancing loss is jointly optimized to encourage a more balanced routing distribution. **Result** MERT was evaluated on three fine-grained image classification benchmarks: Stanford Dogs, NABirds, and Oxford 102 Flowers. MERT achieved Top-1 accuracies of 92.97%, 90.94%, and 99.65% on Stanford Dogs, NABirds, and Oxford 102 Flowers, respectively. MERT achieved the best Top-1 accuracy among the compared methods on Stanford Dogs and Oxford 102 Flowers. Compared with the matched ViT-B/16 baseline, MERT improved accuracy by 2.34, 1.25, and 0.24 percentage points on the three datasets, respectively. These gains were achieved with only 0.5 M additional parameters and 0.8 G additional floating-point operations (FLOPs). Ablation experiments evaluate the contributions of the LER, HER, and DTS modules, with the complete model achieving the best classification performance among the evaluated configurations. Visualizations obtained using gradient-weighted class activation mapping (Grad-CAM) showed more concentrated responses on category-relevant local regions with reduced background activation in the displayed samples. Embeddings obtained using t-distributed stochastic neighbor embedding (t-SNE) further exhibited more compact intra-class distributions and reduced overlap among visually similar categories. **Conclusion** By assigning image-dependent contributions to layerwise representations and attention-head responses, MERT improves the organization and utilization of local discriminative information while reducing interference from redundant and background responses. These results indicate that image-dependent organization of multi-layer representations and multi-head attention responses can improve the utilization of fine-grained discriminative features with only limited additional computational cost.

Key words: fine-grained visual classification; vision transformer; multi-dimensional expert routing; cross-layer feature aggregation; load balancing

论文引用格式:DOI:[10.11834/jig.260565]

0 引　言

细粒度视觉识别(fine-grained visual categorization,FGVC)致力于区分隶属于同一大类的不同子类别,在零售产品识别、智能交通和生物多样性保护等领域具有不可替代的应用价值(Angelova 和 Zhu,2013)。相较于通常依赖全局特征区分大类的通用图像分类(He 等,2016;Dosovitskiy 等,2021),FGVC 面临两方面的技术挑战:1)类内差异显著:同一子类样本因拍摄姿态、视角、光照或物体状态变化,视觉表现差异极大;2)类间差异细微且局部化:不同子类的判别特征多集中于局部区域,整体视觉相似度高,

© 中国图象图形学报版权所有

全局特征难以区分。上述特点使模型容易遗漏局部特征而混淆相似子类，从而限制分类精度。

针对上述挑战，早期的技术方案主要以卷积神经网络（convolutional neural networks，CNN）为基础，并逐渐演化形成部位定位与特征编码两条解决路径。前者利用边界框、部件标注、目标分割或弱监督注意力引导模型关注关键区域（Angelova 和 Zhu，2013；Yang 等，2018）；后者则通过双线性池化和高阶统计建模增强特征表达能力（Lin 等，2015；Li 等，2017；阳治民和宋威，2023）。以上两种方案均能提高模型对局部细节的感知能力，但受卷积操作局部感受野的限制，难以在单层特征中直接建立远距离判别区域间的联系。当判别特征分散于多个局部区域或目标受到复杂背景干扰时，模型难以汇集多个关键部位的信息进而削弱类别表示的判别性。

为弥补CNN在远距离特征交互方面的不足，视觉Transformer（vision transformer，ViT）通过多头自注意力实现不同图像Token之间的信息交互，使空间距离较远的局部判别特征能够共同参与图像表征。现有基于ViT的细粒度分类方法主要通过注意力引导的Token选择和多层特征聚合挖掘局部判别信息。例如，TransFG沿不同Transformer层传播注意力并选取高响应Token（He 等，2022）；FFVT利用多层注意力信息进行Token筛选与特征融合（Wang 等，2021）；IELT则结合注意力头投票和跨层Token分配机制逐步更新局部特征（Xu 等，2023）。上述研究表明，不同层级的视觉表征和不同注意力头产生的响应能够为细粒度分类提供互补信息。

同时，不同图像对这种互补信息的需求存在差异，如图1（a）所示，部分图像的分类主要依赖纹理等低层细节，部分图像需要关注目标部件的语义特征，而背景复杂的图像则需要排除环境干扰，定位具有判别性的局部区域。这要求模型能够适应不同图像的判别需求，并合理调节各层级与各区域特征的贡献。现有研究通过动态选择与剪枝减少Transformer中的冗余计算。例如，DynamicViT依据Token重要性分数，在后续Transformer层执行前逐步剪除冗余Token（Rao 等，2021）；AdaViT根据输入图像动态决定图像块、注意力头和Transformer层是否参与后续计算（Meng 等，2022）。针对注意力头的冗余性，Michel 等（2019）发现，移除部分注意力头通常不会显著影响模型性能；Voita 等（2019）进一步分析了不同注意力头的功能差异，并通过稀疏门控实现冗余头剪枝。此外，稀疏门控混合专家模型（mixture of experts，MoE）通过路由器选择具有独立参数的专家网络，以条件计算方式扩展模型容量（Shazeer 等，2017）；V-MoE将这一机制引入ViT，通过动态分配图像Token实现专家选择（Riquelme 等，2021）。动态选择与剪枝能够减少冗余计算，但细粒度图像分类需要综合不同阶段形成的局部细节与高层语义，部分局部特征的判别价值往往需要结合深层信息才能充分体现。若在主干网络尚未完成特征提取时提前剪除Token、关闭注意力头或跳过Transformer层，可能使部分具有潜在判别价值的信息无法参与后续特征交互，从而影响最终的类别表示。此外，如图1（b）所示，现有部分FGVC方法在利用多层特征时采用预定义的跨层融合规则或统一处理注意力头响应，难以根据不同图像的判别需求调整不同层级表征与注意力头响应的贡献，可能引入冗余信息参与特征聚合，削弱关键局部特征的表达能力。



图1　MERT的研究动机与基本思路

Fig. 1　Motivation and overview of MERT

© 中国图象图形学报版权所有

针对上述问题，本文提出一种多维专家路由Transformer（multi-dimensional expert routing transformer，MERT），其基本思路如图1（c）所示。MERT保留ViT主干的完整特征提取过程，将不同Transformer层的视觉表征和各层注意力头响应作为路由对象，根据输入图像的判别需求动态调整其对特征聚合的贡献。

针对不同图像对层级信息需求不一的问题，本文设计层级专家路由器（layer expert router，LER），以多层全局语义作为路由依据，评估不同层级表征的相对贡献，实现样本相关的跨层特征聚合。为减少注意力头等权处理引入的冗余响应，本文提出注意力头专家路由器（head expert router，HER），对各层已经生成的注意力头响应进行排序与稀疏重加权，使路由评分较高的注意力头参与后续Token重要性估计。在此基础上，构建判别性Token选择模块（discriminative token selection，DTS），联合层级与注意力头路由权重确定关键空间位置，并聚合不同层中相同位置的Token进行特征精炼，从而加强局部判别信息与全局语义之间的联系。此外，构建跨层聚合与局部精炼双分支，并引入注意力头负载均衡约束进行联合训练，以增强两类表征的判别性并缓解注意力头选择过度集中（Fedus等，2022）。

与FFVT基于注意力信息进行多层Token筛选与融合、IELT结合注意力头投票和跨层Token分配构建局部表征的方式不同，MERT在ViT主干完成特征提取后，利用共享路由信号联合建模不同层级表征和注意力头响应的相对贡献。其中，LER生成的层级权重同时用于跨层特征聚合与Token重要性计算，HER生成的稀疏头部权重则用于调整不同注意力响应对Token评分的贡献。两类路由权重共同确定关键Token的位置，模型根据统一的位置索引汇集不同层的对应特征，使层级信息利用、注意力头选择与局部特征精炼形成相互关联的处理过程。

本文主要贡献如下：

1）提出面向细粒度图像分类的多维专家路由Transformer（MERT），对已经生成的多层表征和注意力响应进行样本相关的双维路由，实现不同层级与头部判别信息的差异化利用；

2）构建由层级专家路由器（LER）和注意力头专家路由器（HER）组成的多维专家路由机制，分别对不同层级表征进行加权聚合，并对注意力头响应进行稀疏重加权，有效抑制冗余响应和背景噪声干扰；

3）设计双维路由引导的判别性Token选择模块（DTS），融合层级与注意力头路由权重定位关键空间位置，强化关键局部特征与高层语义之间的联系；

4）在Stanford Dogs、NABirds和Oxford 102 Flowers三个基准数据集上的实验表明，MERT在Stanford Dogs和Oxford 102 Flowers上取得最优结果，并以较小的额外参数量和计算量有效提升ViT-B/16的分类精度，验证了所提方法的分类性能与计算效率。

1 本文方法

1.1 整体框架

针对现有细粒度图像分类方法难以根据输入图像调整多层表征和注意力头响应贡献的问题，提出MERT，其整体框架如图2所示。MERT以ViT各层已经生成的视觉表征为处理对象，利用共享路由信号评估层级与注意力头的相对贡献，并以路由结果引导跨层特征聚合和关键局部特征提取，进而形成跨层聚合与局部精炼两条分类路径。前者整合不同深度的互补特征，后者对筛选出的关键局部Token进行进一步精炼，最终融合两路预测得到分类结果。

MERT首先利用ViT主干提取多层视觉特征。给定输入图像$\boldsymbol{X}$，模型首先将其划分为N个固定大小的图像块，经线性投影后映射至C维特征空间。在图像Token序列中加入类别Token和位置编码后，依次输入L个Transformer层。第l层输出Token序列$\boldsymbol{Z}_l$及对应的多头自注意力响应$\boldsymbol{A}_l$。其中，$\boldsymbol{Z}_l$由类别Token$\boldsymbol{z}_{l,\mathrm{c}}$和$N$个图像Token组成，$\boldsymbol{A}_l$包含该层$E$个注意力头的响应。上述过程获得不同深度下包含局部细节、部件结构和类别语义的视觉表征，为后续评估不同层级信息的相对贡献提供基础。

为同时向层级路由和注意力头路由提供当前图像的全局语义信息，将各Transformer层的类别Token沿层级维度堆叠，构成共享路由信号$\boldsymbol{M}$：

$$\boldsymbol{M}=\left[\boldsymbol{z}_{1,\mathrm{c}}^{\mathrm{T}};\boldsymbol{z}_{2,\mathrm{c}}^{\mathrm{T}};\cdots;\boldsymbol{z}_{L,\mathrm{c}}^{\mathrm{T}}\right]\in\mathbf{R}^{L\quad C}\tag{1}$$

式中，$\boldsymbol{M}$为由各层类别Token组成的共享路由信号，$\boldsymbol{z}_{l,\mathrm{c}}$为第$l$层类别Token。LER根据$\boldsymbol{M}$评估各层表征的相对贡献，生成层级权重$\boldsymbol{W}^{\mathrm{L}}=[w_1,w_2,\cdots,w_L]^{\mathrm{T}}$，其中$w_l$表示第$l$层的权重。HER根据同一共享路由

© 中国图象图形学报版权所有

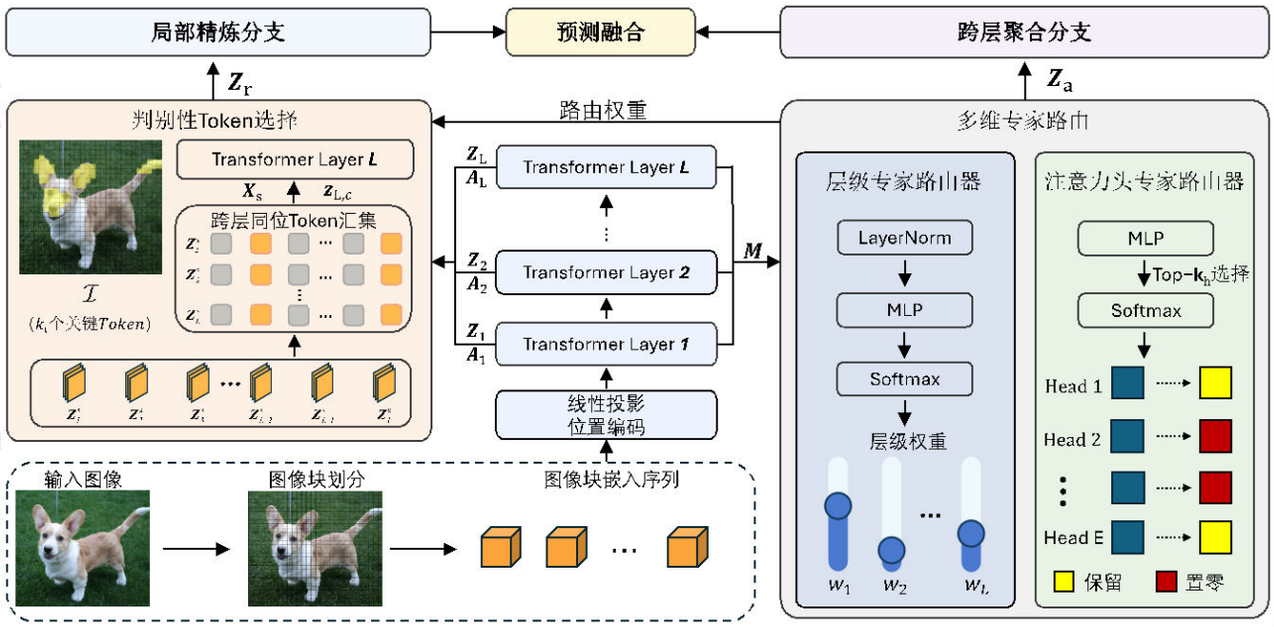

图2　MERT整体网络结构

Fig. 2　Overall architecture of MERT

信号评估各层内不同注意力头的相对贡献，并通过 $Top-k_h$ 选择得到稀疏头部权重 $\boldsymbol{w}_l^h$，其中 k_h 为每层保留的注意力头数量。两类路由权重分别刻画不同层级表征与注意力头对当前图像的相对重要性，并用于控制后续跨层特征聚合和Token重要性估计。

路由结果沿跨层聚合与局部精炼两条路径发挥作用。在跨层聚合分支中，模型根据层级权重 $\boldsymbol{W}^L$ 对各层Token序列进行加权求和，得到跨层聚合序列，并取该序列中的类别Token作为跨层聚合表示 $\boldsymbol{Z}_a \in \mathbf{R}^C$，使不同深度的视觉信息能够根据当前图像的判别需求进行融合。在局部精炼分支中，DTS模块联合层级权重、头部权重以及类别Token指向图像Token的注意力响应，计算各空间位置的重要性分数。经过二维空间平滑后，模型选取得分较高的 k_t 个位置，其中 k_t 为关键图像Token数量，不包含类别Token。随后依据同一组位置索引从各Transformer层提取对应Token，并利用层级权重聚合各层同位Token，得到关键局部表示 $\boldsymbol{X}_s$。

模型将 $\boldsymbol{X}_s$ 与最后一层类别Token $\boldsymbol{z}_{L,c}$ 组成新的Token序列，并再次输入最后一个Transformer层，使不同关键Token相互传递信息，同时建立局部特征与类别语义之间的联系，得到局部精炼表示 $\boldsymbol{Z}_r$。跨层聚合表示 $\boldsymbol{Z}_a$ 和局部精炼表示 $\boldsymbol{Z}_r$ 分别产生分类预测，并通过预测融合得到最终输出。训练阶段进一步引入注意力头负载平衡约束，缓解HER长期集中选择少数注意力头。

1.2　多维专家路由模块

ViT不同层的表征具有不同的特征侧重，同一层内的不同注意力头对图像区域的响应分布也存在差异。而不同细粒度图像所依赖的判别线索存在差异，因此对不同层级表征和注意力头响应的利用需求并不一致。采用固定规则融合不同层级表征或等权处理注意力头响应，既无法根据当前图像调节各层信息的贡献，也难以区分不同注意力头对关键区域定位的作用。为此，本文将不同Transformer层的视觉表征和各层注意力头响应分别作为层级与头级路由对象，构建由LER和HER组成的多维专家路由模块，分别从层级和注意力头两个维度生成样本相关的路由权重。

LER和HER均以式(1)得到的共享路由信号 $\boldsymbol{M}$ 为输入，其中第 l 行对应第 l 层的类别Token $\boldsymbol{z}_{l,c}$。由于不同Transformer层的特征分布存在差异，LER首先对类别Token进行层归一化(layer normalization，LN)，以减小这种差异对路由评分的影响，再利用参数共享的两层全连接网络将其映射为标量，得到相应的原始路由分数。第 l 层的路由分数计算为：

$$s_l = \boldsymbol{U}_2 ReLU\left(\boldsymbol{U}_1 LN\left(\boldsymbol{z}_{l,c}\right) + \boldsymbol{b}_1\right) + b_2 \tag{2}$$

式中，s_l 为第 l 层的原始路由分数；$\boldsymbol{U}_1$ 和 $\boldsymbol{U}_2$ 为两层线

© 中国图象图形学报版权所有

性映射矩阵，$\boldsymbol{b}_1$和b_2为相应的偏置项，*ReLU*为修正线性单元激活函数。得到全部层级的路由分数后，在层级维度上进行Softmax归一化，第l层的权重为

$$w_l = \frac{exp(s_l)}{\sum_{j=1}^{L} exp\ (s_j)}, \qquad l = 1,2,\cdots,L \tag{3}$$

式中，j表示层编号，取值为$1,\cdots,L$。分母对所有层的原始路由分数取指数后求和，使各层权重之和为1。所得层级权重用于刻画不同深度表征对当前图像分类的相对贡献。HER同样利用共享路由信号计算注意力头的路由分数。对于第l层，HER通过独立的头部评分网络得到评分向量$\boldsymbol{r}_l = [r_{l,1}, r_{l,2}, \cdots, r_{l,E}]^{\mathrm{T}}$，其中$r_{l,e}$表示第$e$个注意力头的原始路由分数。随后，根据$\boldsymbol{r}_l$选取得分最高的$k_{\mathrm{h}}$个注意力头。以$m_{l,e}$表示对应的选择状态，被选中时$m_{l,e}$的值为1，否则$m_{l,e}$等于0。仅在被选中的注意力头之间进行Softmax归一化，得到稀疏头部权重：

$$\hat{w}_{l,e}^{\mathrm{h}} = \frac{m_{l,e} exp(r_{l,e})}{\sum_{j=1}^{E} m_{l,j} exp(r_{l,j})}, \qquad e = 1,2,\cdots,E \tag{4}$$

式中，$\hat{w}_{l,e}^{\mathrm{h}}$为第$l$层第$e$个注意力头的稀疏归一化权重，$j$为注意力头编号，取值为$1,\cdots,E$。由于$Top-k_{\mathrm{h}}$产生的离散选择状态不可导，分类损失无法通过选择掩码直接优化头部排序过程。为在保持前向稀疏选择的同时完成反向传播，本文采用直通估计器（straight-through estimator，STE）。令完整头部评分对应的连续概率为$\boldsymbol{p}_l = Softmax(\boldsymbol{r}_l)$，最终的注意力头权重为

$$\boldsymbol{w}_l^{\mathrm{h}} = \boldsymbol{p}_l + sg\left(\hat{\boldsymbol{w}}_l^{\mathrm{h}} - \boldsymbol{p}_l\right) \tag{5}$$

式中，$\boldsymbol{w}_l^{\mathrm{h}}$为采用直通估计得到的第$l$层注意力头权重向量，$\hat{\boldsymbol{w}}_l^{\mathrm{h}}$为式(4)得到的稀疏权重向量，$sg(\cdot)$表示停止梯度操作。前向传播时，$\boldsymbol{w}_l^{\mathrm{h}}$等于稀疏权重$\hat{\boldsymbol{w}}_l^{\mathrm{h}}$；反向传播时，停止梯度项不参与梯度计算，分类损失的梯度通过连续概率$\boldsymbol{p}_l$传递至头部评分网络，从而利用连续概率对头部评分网络进行近似梯度优化。

LER输出的层级权重用于跨层特征聚合，同时与HER输出的注意力头权重共同参与图像Token重要性计算。训练阶段还对HER的头部选择分布施加负载平衡约束，以避免路由结果长期集中于少数注意力头。

式中，$\overline{\boldsymbol{M}}$为层归一化后的共享路由信号，$f$和$g$分别为层级路由与注意力头路由的评分映射，

算法1：多维专家路由算法

输入：共享路由信号$\boldsymbol{M}$，Transformer层数L，注意力头数E，保留头数k_{h}

输出：层级权重$\boldsymbol{W}^{\mathrm{L}}$，注意力头权重$\boldsymbol{W}^{\mathrm{H}}$

```
路由归一化 M̄ ← LN(M)
层级路由 W^L ← Softmax(f(M̄))
头部评分 T ← g(M̄) = [r_1, r_2, ..., r_L]^T
for r_l in {r_1, r_2, ..., r_L} do
    m_l ← TopK(r_l, k_h);   #保留高分注意力头
    ŵ_l^h ← Softmax(r_l, m_l);   #稀疏重加权
    p_l ← Softmax(r_l)
    w_l^h ← p_l + sg(ŵ_l^h − p_l)   #直通估计
end for
W^H ← Stack(w_1^h, ..., w_L^h)   #权重汇集
return W^L, W^H
```

$\boldsymbol{T} \in \mathbf{R}^{L \times E}$为全部层的注意力头评分矩阵，$\boldsymbol{W}^{\mathrm{H}} \in \mathbf{R}^{L \times E}$为各层注意力头权重组成的矩阵，*Stack*表示将各层权重向量按行堆叠。

1.3 判别性Token选择模块

多维专家路由模块输出层级路由权重和注意力头路由权重后，本文进一步设计DTS模块。该模块利用两类路由权重对自注意力响应进行加权，计算不同空间位置的Token重要性，随后根据重要性分数选取关键局部Token，并汇集不同层中相同位置的特征，为局部特征精炼提供输入。

设第l个Transformer层中，第e个注意力头的自注意力矩阵为$\boldsymbol{A}_{l,e}$。DTS利用注意力头路由权重对同一层内的多头分数进行加权汇集，再利用层级路由权重融合不同Transformer层的评分。经过上述两级加权，第i个图像Token的重要性分数表示为

$$q_i = \sum_{l=1}^{L} w_l \sum_{e=1}^{E} w_{l,e}^{\mathrm{h}} \boldsymbol{A}_{l,e}(0,i), \qquad i = 1,2,\cdots,N \tag{6}$$

式中，q_i为第i个图像Token的重要性分数，$\boldsymbol{A}_{l,e}(0,i)$表示第$l$层第$e$个注意力头中，类别Token对第$i$个图像Token的注意力分数。最终得到的$\boldsymbol{q} = [q_1, q_2, \cdots, q_N]^{\mathrm{T}}$综合了不同层级和不同注意力头提供的空间信息，用于后续关键Token选择。得到Token重要性分数后，还需考虑相邻图像块之间的空间联系。单个图像块的高分可能来自局部噪声，而目标的眼睛、耳朵或花瓣边缘等判别部位通常由多个相邻图像块共同表征。为减小孤立高分对选择结果的影响，本文借鉴IELT的二维空间平滑方法（Xu等，2023），将$\boldsymbol{q}$恢复为二维图像块网格，并采用

© 中国图象图形学报版权所有

归一化的3 × 3平滑核$\boldsymbol{K}$进行卷积。平滑结果重新展平后表示为：

$$\bar{\boldsymbol{q}} = Flatten\left[\boldsymbol{K}*Reshape(\boldsymbol{q})\right] \tag{7}$$

式中，$\bar{\boldsymbol{q}}$为二维平滑后的Token重要性分数向量，*Flatten*表示将二维评分网格展平为向量，*表示二维卷积，*Reshape*表示按照图像块的空间排列将$\boldsymbol{q}$恢复为二维评分网格。二维平滑综合相邻图像块的评分，使空间上连续的高分区域在Token选择中得到更稳定的保留。随后，根据$\bar{\boldsymbol{q}}$选取得分最高的k_{t}个位置：

$$\mathcal{I} = TopK\left(\bar{\boldsymbol{q}},k_{\mathrm{t}}\right) \tag{8}$$

式中，$\mathcal{I}$为对应的位置索引，*TopK*表示返回分数最高的k_{t}个位置索引。该索引由融合全部层级和注意力头信息的Token分数统一确定，获得位置索引后，DTS依据$\mathcal{I}$从各Transformer层提取相同空间位置的Token，并利用层级路由权重进行跨层聚合。设$\boldsymbol{Z}_l^{\mathrm{s}}$为第$l$层输出的图像Token序列，则聚合后的关键Token表示为

$$\boldsymbol{X}_{\mathrm{s}} = LN\left[\sum_{l=1}^{L} w_l \boldsymbol{Z}_l^{\mathrm{s}}[\mathcal{I}]\right] \tag{9}$$

式中，$\boldsymbol{X}_{\mathrm{s}} \in \mathbf{R}^{k_{\mathrm{t}} \times C}$为跨层聚合后的关键Token表示，$\boldsymbol{Z}_l^{\mathrm{s}}[\mathbf{I}]$表示按照索引序列$\mathcal{I}$提取对应位置的Token。每个Token均融合了不同Transformer层在同一空间位置产生的特征。由此既保持了Token在各层之间的位置对应关系，又能根据层级路由权重综合局部细节、部件结构和类别语义。

最后，将关键Token表示$\boldsymbol{X}_{\mathrm{s}}$与最后一层类别Token$\boldsymbol{z}_{L,\mathrm{c}}$拼接，并再次输入最后一个Transformer层进行特征精炼。将精炼后的类别Token记为局部精炼表示$\boldsymbol{Z}_{\mathrm{r}}$，用于该分支的分类预测。该过程使选中的局部特征能够相互传递信息，并与类别Token中的全局语义进一步融合。

1.4 预测融合与损失函数

经过跨层聚合与局部特征精炼，MERT得到跨层聚合表示和局部精炼表示。前者利用层级权重对各层Token序列进行加权聚合，并取聚合序列中的类别Token作为$\boldsymbol{Z}_{\mathrm{a}}$；后者由DTS选出的关键Token经过最后一个Transformer层再次精炼得到，记为$\boldsymbol{Z}_{\mathrm{r}}$。模型通过权重矩阵为$\boldsymbol{W}_{\mathrm{c}} \in \mathbf{R}^{N_{\mathrm{c}} \times C}$、偏置向量为$\boldsymbol{b}_{\mathrm{c}} \in \mathbf{R}^{N_{\mathrm{c}}}$的共享分类器，将均为$C$维类别Token特征的跨层聚合表示$\boldsymbol{Z}_{\mathrm{a}}$和局部精炼表示$\boldsymbol{Z}_{\mathrm{r}}$映射至包含$N_{\mathrm{c}}$个类别的类别空间。考虑到两类信息具有互补性，为平衡跨层整体表征与局部判别线索对最终分类的贡献，对两分支预测进行融合，具体过程为：

$$\boldsymbol{o}_{\mathrm{a}} = \boldsymbol{W}_{\mathrm{c}}\boldsymbol{Z}_{\mathrm{a}} + \boldsymbol{b}_{\mathrm{c}} \tag{10}$$

$$\boldsymbol{o}_{\mathrm{r}} = \boldsymbol{W}_{\mathrm{c}}\boldsymbol{Z}_{\mathrm{r}} + \boldsymbol{b}_{\mathrm{c}} \tag{11}$$

$$\boldsymbol{o} = \boldsymbol{o}_{\mathrm{a}} + Softmax\left(\boldsymbol{o}_{\mathrm{r}}\right)\odot\left(\boldsymbol{W}_{\mathrm{c}}\boldsymbol{1}_C\right) \tag{12}$$

式中，$\boldsymbol{o}_{\mathrm{a}},\boldsymbol{o}_{\mathrm{r}} \in \mathbf{R}^{\mathbf{N}_{\mathrm{c}}}$分别为跨层聚合分支与局部精炼分支的分类得分，$\boldsymbol{o} \in \mathbf{R}^{\mathbf{N}_{\mathrm{c}}}$为最终分类结果。$\boldsymbol{1}_C$为$C$维全1向量，⊙表示逐元素乘法。局部精炼预测经Softmax转换为类别概率后，与分类器各类别权重的通道和相乘，形成对基础预测的类别相关补充。该方式避免直接拼接两类高维特征，使跨层信息和局部判别信息在类别空间中完成融合。

训练阶段，最终融合预测$\boldsymbol{o}$承担整体分类监督，同时对局部精炼预测施加辅助监督。前者直接优化最终分类结果，后者约束DTS筛选并精炼得到的局部表示保持类别判别能力。同时，HER采用*Top* − k_{h}稀疏选择，若路由长期集中于少数注意力头，容易造成头部使用失衡，并限制不同注意力模式的充分利用。为此，本文同时统计各注意力头的连续路由概率和实际选择频率，并构建负载平衡损失：

$$\bar{p}_e = \frac{1}{BL}\sum_{b=1}^{B}\sum_{l=1}^{L} p_{b,l,e} \tag{13}$$

$$\bar{m}_e = \frac{1}{BLk_{\mathrm{h}}}\sum_{b=1}^{B}\sum_{l=1}^{L} m_{b,l,e} \tag{14}$$

$$\mathcal{L}_{\mathrm{b}} = E\sum_{e=1}^{E}\bar{p}_e\bar{m}_e \tag{15}$$

式中，$\bar{p}_e$为第e个注意力头在当前批次全部样本和全部层上的平均连续路由概率，B为批大小，b为批次内的样本编号，$p_{b,l,e}$为第b个样本在第l层对第e个注意力头分配的连续路由概率。$\bar{m}_e$为第e个注意力头的归一化选择频率，$m_{b,l,e}$表示该头是否被选中，选中时取1，否则取0。$\mathcal{L}_{\mathrm{b}}$为注意力头负载平衡损失。连续路由概率反映HER对各注意力头的评分倾向，选择频率反映*Top* − k_{h}路由的实际使用情况。二者共同约束路由分布，以同时平衡连续评分倾向与实际头部使用频率，缓解路由集中现象。

综合最终融合预测的分类监督、局部精炼分支的辅助分类监督和注意力头负载平衡约束，模型的总损失为：

$$\mathcal{L} = (1-\alpha)CE(\boldsymbol{o},y) + \alpha CE(\boldsymbol{o}_{\mathrm{r}},y) + \lambda_{\mathrm{b}}\mathcal{L}_{\mathrm{b}} \tag{16}$$

式中，$\mathcal{L}$为训练总损失，$\alpha \in [0,1]$控制局部精炼分支

© 中国图象图形学报版权所有

辅助分类损失的权重，最终融合预测的分类损失权重相应为1 − α，CE表示交叉熵损失，y为真实类别标签，λ_b为负载平衡损失的权重。推理阶段保留跨层聚合与局部精炼两条分类路径，并输出融合后的类别预测$\boldsymbol{o}$。

2 实验结果与分析

2.1 数据集和实验设置

1）数据集：本文在Stanford Dogs、NABirds和Oxford 102 Flowers共3个公开细粒度图像分类数据集上进行实验。图3以NABirds数据集中的典型样本为例，展示了同一类别内部的外观变化以及不同相似类别之间的细微视觉差异。Stanford Dogs包含120个犬种，共20,580幅图像，不同犬种之间具有较高的外观相似性，部分图像同时包含复杂背景（Khosla等，2011）。NABirds包含555个北美鸟类类别，类别间差异通常集中于局部形态和纹理特征，是检验模型局部判别特征提取能力的常用基准数据集（Van Horn等，2015）。Oxford 102 Flowers包含102个花卉类别，不同图像在拍摄视角、目标尺度及背景环境等方面存在较大差异（Nilsback和Zisserman，2008）。各数据集的详细统计信息见表1。监督训练过程中仅使用类别标签，未引入边界框或部件标注等额外信息。



图3 细粒度图像分类的主要挑战

Fig. 3 Main challenges of fine-grained image classification

2）实验设置：所有实验在NVIDIA GeForce RTX 4090 GPU上使用PyTorch框架进行训练和测试，GPU显存为24 GB，系统内存为1 TB。Python版本为3.12.13，PyTorch版本为2.11.0+cu130，CUDA版本为13.0，cuDNN版本为9.19.0。

采用在ImageNet-21K预训练的ViT-B/16作为主干网络，输入图像尺寸设置为448×448，图像块大小为16×16，对应784个图像Token。三个数据集的保留注意力头数量均设置为9；Stanford Dogs、NABirds和Oxford 102 Flowers的关键Token数量分别设置为96、128和512。训练阶段采用随机裁剪、水平翻转和颜色抖动进行数据增强，测试阶段采用中心裁剪。模型使用随机梯度下降优化器，动量设置为0.9，批大小设置为16，共训练50轮。本方法的主要训练参数参照IELT（Xu等，2023）的实验设置，具体地，在Stanford Dogs数据集的初始学习率设置为0.002，NABirds和Oxford 102 Flowers数据集设置为0.02，并采用余弦退火策略调整学习率。实验评价指标为测试集的Top-1分类准确率。

表1 实验数据集统计信息

Table 1 Statistics of the experimental datasets

数据集	类别数	训练集	测试集
Oxford 102 Flowers	102	2040	6149
Stanford Dogs	120	12000	8580
NABirds	555	23929	24633

2.2 对比实验

为验证MERT的分类性能，本文在Stanford Dogs、NABirds和Oxford 102 Flowers数据集上，将其与代表性的细粒度图像分类方法进行比较。对比方

© 中国图象图形学报版权所有

法包括传统CNN细粒度方法和近年来基于ViT的代表性方法，具体结果见表2。MERT在不同随机种子下独立运行5次，表2报告其Top-1准确率的均值和标准差，其他方法的实验结果均引自相应文献。

1）Stanford Dogs实验结果：MERT取得了92.97%±0.04%的Top-1准确率，相较ViT-B/16基线（Dosovitskiy等，2021）提高了2.34个百分点。FFVT（Wang等，2021）利用多层注意力信息进行Token筛选与特征融合，IELT（Xu等，2023）则结合注意力头投票和跨层Token分配机制提取局部判别特征。EViT（Liang等，2022）通过保留重要Token并融合低重要性Token，减少后续Transformer层的计算量。MERT利用样本相关的层级权重和注意力头权重确定关键Token，并聚合不同层中相同位置的特征，其准确率相较FFVT，IELT和EViT分别提高1.45，1.11和2.55个百分点。与DAH-Trans（Yang等，2026）相比，MERT的准确率提高0.06个百分点，而参数量和FLOPs分别减少31.4M和227.9G，在保持较低计算开销的同时取得了更高的分类准确率。

2）NABirds实验结果：MERT在NABirds数据集上的Top-1准确率为90.94%±0.08%，相较ViT-B/16基线提高了1.25个百分点。NABirds中的类别差异往往集中于鸟类头部、翅膀及羽毛纹理等局部区域，对关键Token的选择提出了较高要求。其准确率在数值上较TransFG、IELT和DAH-Trans分别提高0.11、0.10和0.08个百分点。PCT-ViT（Gao等，2026）结合注意力信息与反事实Token选择增强判别性特征，在该数据集上取得91.47%的准确率，高于MERT的90.94%；ACC-ViT（Zhang等，2024）同样取得了91.38%的准确率。然而，MERT的参数量和FLOPs分别为86.9M和78.6G，较PCT-ViT分别减少4.3M和29.0G，较ACC-ViT分别减少7.1M和92.6G。因此，MERT在NABirds上虽然未取得最高分类准确率，但在参数量和计算开销较低的情况下保持了具有竞争力的分类性能。

3）Oxford 102 Flowers实验结果：MERT取得了99.65%±0.03%的Top-1准确率，在所有对比方法中排名第1。相较ViT-B/16基线，MERT提高了0.24个百分点；相较IELT和PCT-ViT分别提高了0.02和0.04个百分点。Oxford 102 Flowers中的花卉图像存在尺度、视角及背景差异，关键判别区域可能涉及花瓣纹理、花蕊结构等不同层次的信息。R^2-Trans（Ye等，2024）通过动态掩码减少输入图像中的冗余信息，并利用信息瓶颈约束压缩特征表示，在该数据集上取得99.51%的准确率。相比之下，MERT根据输入图像动态调整不同层级表征和注意力头响应的贡献，并利用联合路由结果引导关键Token选择与跨层特征聚合，在参数量同为86.9M的情况下，准确率提高0.14个百分点，FLOPs减少22.3G。此外，相较ViT-B/16基线，MERT仅增加0.5M参数量和0.8G计算量，表明其能够以较小的额外计算开销提高细粒度图像分类精度。

综合3个数据集的结果，MERT相较ViT-B/16基线的准确率分别提高2.34、1.25和0.24个百分点，而参数量和FLOPs仅增加0.5M和0.8G。在Stanford Dogs和Oxford 102 Flowers上取得最优结果，在NABirds上则以较低计算开销获得具有竞争力的分类性能。上述结果表明，通过联合利用不同层级表征与注意力头响应引导判别性Token选择和跨层特征聚合，MERT能够在仅增加少量额外开销的情况下改善细粒度图像分类性能。

2.3 消融实验

为分析MERT各组成模块对分类性能的影响，在Stanford Dogs、NABirds和Oxford 102 Flowers三个数据集上开展组件消融实验。所有变体均采用相同的ViT-B/16主干网络、固定随机种子和训练配置，仅调整LER、HER、DTS及负载平衡损失的启用状态。未启用LER时，跨层聚合分支使用最后一层表征，DTS中的层级权重设置为均匀权重，使不同层Token等权参与重要性计算和跨层聚合；未启用HER时，DTS对全部注意力头的响应取平均。由于HER生成的路由权重仅用于DTS中的Token重要性计算，因此在不启用DTS时，单独加入HER不会改变后续特征构建过程，其结果与ViT-B/16基线相同。实验结果如表3所示。

从单模块结果来看，LER在Stanford Dogs和NABirds上分别将准确率提高1.82和0.88个百分点，增幅高于单独使用DTS时的0.69和0.56个百分点。该结果表明，不同Transformer层提供的视觉表征具有一定互补性，利用LER调整各层贡献能够提升分类性能。DTS通过筛选关键空间位置并构建局部精炼表示，在3个数据集上均带来一定性能增益。

单独启用HER时，模型性能与ViT-B/16基线一致，这是因为HER本身不直接生成分类特征，其路

© 中国图象图形学报版权所有

表2　不同细粒度图像分类方法的分类性能与计算开销比较

Table 2　Comparison of classification performance and computational cost among fine-grained image classification methods

方法	来源	主干网络	Top-1准确率/%			参数量	FLOPs
			Dogs	NABirds	Flowers	(M)	(G)
基于CNN的方法							
API-Net(Zhuang等,2020)	AAAI′20	DenseNet-161	90.37	88.12	99.31	**29.0**	30.9
Cross-X(Luo等,2019)	ICCV′19	ResNet-50	88.95	86.28	99.14	48.3	32.2
Grafit(Touvron等,2021)	ICCV′21	ResNet-50	88.49	88.75	99.10	39.0	**8.0**
基于ViT的方法							
ViT(Dosovitskiy等,2021)	ICLR′21	ViT-B/16	90.63	89.69	99.41	86.4	77.8
TPSKG(Liu等,2022)	Neuro′22	ViT-B/16	92.54	90.17	99.55	86.6	134.4
TransFG(He等,2022)	AAAI′22	ViT-B/16	90.68	90.83	99.43	86.9	129.4
FFVT(Wang等,2021)	BMVC′21	ViT-B/16	91.52	90.17	99.45	86.4	256.0
EViT(Liang等,2022)	ICLR′22	ViT-B/16	90.42	88.65	99.21	86.4	50.16
IELT(Xu等,2023)	TMM′23	ViT-B/16	91.86	90.84	99.63	93.5	72.6
MPSA(Wang等,2024)	TIP′24	ViT-B/16	91.46	89.97	99.50	98.5	67.9
ACC-ViT(Zhang等,2024)	PR′24	ViT-B/16	92.71	91.38	99.58	94.0	171.2
R^2-Trans(Ye等,2024)	IVC′24	ViT-B/16	92.45	90.26	99.51	86.9	100.9
FAL-ViT(Huang等,2025)	TCSVT′25	ViT-B/16	91.13	91.19	99.52	88.4	175.5
KR-MHV(Khairunnisa和Wibowo,2025)	JOIV′25	ViT-B/16	92.03	90.49	99.53	93.5	72.6
PCT-ViT(Gao等,2026)	DSP′26	ViT-B/16	92.62	**91.47**	99.61	91.2	107.6
DAH-Trans(Yang等,2026)	ESWA′26	ViT-B/16	92.91	90.86	99.59	118.3	306.5
MERT(本文)	—	ViT-B/16	**92.97** **0.04**	90.94±0.08	**99.65** **0.03**	86.9	78.6

注:**加粗**字体和下划线分别表示各列最优值和次优值;各方法输入图像尺寸均为448×448。

由结果需要通过DTS作用于Token选择过程。将HER与DTS结合后,三个数据集的准确率均高于仅使用DTS的结果,其中Stanford Dogs上的提升达到1.24个百分点。可见,不同注意力头对关键区域定位的作用存在差异。HER对注意力头响应进行选择与重加权后,DTS的Token重要性计算能够减少低贡献头部响应的影响。

不同模块组合的结果进一步反映了LER、HER与DTS之间的作用关系。相较单独使用LER,LER+DTS在Stanford Dogs、NABirds和Oxford 102 Flowers上的准确率分别仅提高0.06、0.05和0.01个百分点。LER与DTS均涉及跨层聚合,判别信息存在重叠,联合收益受限。同时,在未启用HER的情况下,DTS采用各注意力头的平均响应计算Token重要性,部分冗余响应仍可能影响关键区域的定位。从不同组合的表现来看,HER+DTS在Stanford Dogs上略优于LER+DTS,而LER+DTS在NABirds上取得更高的准确率,两种组合在Oxford 102 Flowers上的表现较为接近,说明层级信息与注意力头信息在不同数据集中的作用存在差异。

同时启用LER、HER和DTS后,模型在三个数据集上的准确率分别达到92.81%、90.85%和99.51%,均优于相应的双模块组合。说明层级信息与头信息从不同方面支持DTS,联合改善跨层组织与局部Token选择。在此基础上加入注意力头负载平衡损失,三个数据集上的准确率进一步提高0.16、0.09和0.14个百分点,表明负载平衡约束能够缓解注意力头使用过度集中的现象,并进一步改善分类性能。

完整模型相较ViT-B/16基线,在Stanford Dogs、NABirds和Oxford 102 Flowers上的准确率分别提高2.34、1.25和0.24个百分点。综合各项结果可以看

© 中国图象图形学报版权所有

表3　MERT组成模块的消融实验结果

Table 3　Component ablation results of MERT

模型配置	Top-1准确率/%		
	Stanford Dogs	NABirds	Oxford 102 Flowers
ViT-B/16	90.63	89.69	99.41
+LER	92.45	90.57	99.47
+HER	90.63	89.69	99.41
+DTS	91.32	90.25	99.43
+HER+DTS	92.56	90.45	99.49
+LER+DTS	92.51	90.62	99.48
+LER+HER+DTS	92.81	90.85	99.51
完整模型(+L_b)	**92.97**	**90.94**	**99.65**

注：**加粗**字体和下划线分别表示各列最优值和次优值。

出，跨层特征聚合构成了性能提升的主要基础，注意力头路由进一步提高了关键Token定位的有效性，负载平衡约束则增强了稀疏路由训练的稳定性。

2.4　超参数分析

为考察关键超参数对MERT分类性能的影响，本文在Stanford Dogs、NABirds和Oxford 102 Flowers数据集上分别分析了保留注意力头数量k_h和关键Token数量k_t。实验中保持其他训练设置不变，参数k_h用于控制参与Token重要性计算的头部注意力图数量，k_t用于控制进入局部特征精炼过程的Token数量。

2.4.1　保留注意力头数量k_h的影响

HER从各层已经生成的12个注意力头响应中选取k_h个参与后续Token重要性计算。不同k_h取值下的实验结果如表4所示。

表4　不同保留注意力头数量下的Top-1准确率

Table 4　Top-1 accuracy with different numbers of retained attention heads

k_h	Dogs	NABirds	Flowers
2	92.797	90.178	99.480
4	92.692	90.083	99.409
6	92.762	90.199	99.545
8	92.797	90.260	99.464
9	**92.972**	**90.941**	**99.653**
10	92.774	90.876	99.545
12	92.774	90.766	99.464

注：**加粗**字体和下划线分别表示各列最优值和次优值。

由表4可知，当保留注意力头数量k_h=9时，MERT在Stanford Dogs、NABirds和Oxford 102 Flowers上的Top-1准确率分别达到92.972%、90.941%和99.653%，均为相应数据集上的最高结果。在Stanford Dogs上，仅保留2个注意力头时，模型仍能达到92.797%的准确率，较k_h=9时仅下降0.175个百分点，表明少量高贡献注意力头能够提供一定的局部判别信息。而在NABirds上，两种配置的准确率相差0.763个百分点，反映出不同数据集对多头注意力信息的利用需求存在差异，适当增加保留头数有助于综合不同注意力模式提供的互补信息。当k_h增加至10或12时，准确率均出现不同程度下降，与保留全部12个注意力头相比，k_h=9时三个数据集的准确率分别提高0.198、0.175和0.189个百分点，说明适度筛除注意力头响应有助于减少冗余信息对Token重要性计算的干扰。综合来看，较小的k_h可能限制多头注意力信息的利用，而较大的k_h又会降低头部筛选作用，因此本文后续实验将k_h设置为9，以兼顾判别信息的充分利用与冗余响应的抑制。

2.4.2　关键Token数量k_t的影响

由表5可知，在Stanford Dogs上，k_t=96和k_t=320时均取得最高准确率92.972%，考虑到较少的Token能够降低后续局部特征精炼的计算开销，最终选择96；NABirds在k_t=128时达到最高准确率90.941%；Oxford 102 Flowers则在k_t=512时取得最高准确率99.653%。

© 中国图象图形学报版权所有

表5　不同关键Token数量下的Top-1准确率

Table 5　Top-1 accuracy with different numbers of selected tokens

k_t	Dogs	NABirds	Flowers
32	92.762	90.716	99.528
64	92.751	90.766	99.512
96	**92.972**	90.790	99.545
128	92.786	**90.941**	99.540
160	<u>92.890</u>	90.814	99.431
192	92.879	90.810	99.512
256	92.786	90.770	<u>99.593</u>
320	**92.972**	<u>90.866</u>	99.528
384	92.879	90.659	99.496
448	92.832	90.768	99.496
512	92.844	90.778	**99.653**
640	92.840	90.659	99.528
784	92.786	90.649	99.528

注：**加粗**字体和<u>下划线</u>分别表示各列最优值和次优值。

根据上述实验结果可知，不同数据集对应的最优Token数量存在差异。考虑到各数据集的输入图像均被调整为相同尺寸，上述差异可能与目标在图像中的相对尺度，以及判别区域的空间覆盖范围有关。Stanford Dogs的类别差异通常体现在耳朵形态、面部结构和毛发纹理等局部特征上，NABirds的判别线索主要涉及鸟喙、头部及羽毛纹理等区域，较小的Token预算可能已能覆盖主要判别信息。相比之下，Oxford 102 Flowers的类别识别不仅依赖花蕊等局部细节，还涉及花瓣形态、纹理及整体结构，判别信息可能分布于更广的空间区域，因此需要保留更多Token以充分表征相关特征。

同时，由表5可知，分类准确率并未随Token数量增加而持续提高，说明过多保留低判别性Token可能引入冗余信息，影响局部特征精炼的效果。因此，Token预算的选择需要综合考虑不同数据集的判别特征及其空间分布。

2.4.3　分类损失权重α的影响

分类损失权重α用于平衡最终融合预测与局部精炼分支辅助预测的分类监督。为分析不同监督权重对模型分类性能的影响，本文在Stanford Dogs、NABirds和Oxford 102 Flowers三个数据集上考察不同α取值下的分类性能，相关实验结果如表6所示。

由表6可知，当分类损失权重α从0逐步增加至0.6时，Stanford Dogs、NABirds和Oxford 102 Flowers上的Top-1准确率分别由92.662%、90.623%和99.463%提高至92.972%、90.941%和99.653%，均达到所测试配置中的最高值。当α进一步增加至1.0时，三个数据集的准确率分别下降至92.755%、90.777%和99.491%。上述结果表明，三个数据集的分类准确率均随α的增加呈现先上升后下降的趋势，适当平衡最终融合预测与局部精炼分支的分类监督有助于提高模型性能。这一现象可能与两类分类监督的作用差异有关。当α较小时，局部精炼分支受到的直接分类监督较弱，可能限制关键Token特征的判别能力；而当α较大时，最终融合预测对应的分类损失权重降低，可能削弱对整体分类结果的直接优化。

表6　不同分类损失权重下的Top-1准确率

Table 6　Top-1 accuracy with different classification loss weights

α	Dogs	NABirds	Flowers
0.0	92.662	90.623	99.463
0.2	92.744	90.656	99.480
0.4	92.786	90.711	99.496
0.6	**92.972**	**90.941**	**99.653**
0.8	<u>92.890</u>	<u>90.809</u>	<u>99.534</u>
1.0	92.755	90.777	99.491

注：加粗字体和下划线分别表示各列最优值和次优值。

因此，本文综合三个数据集的实验结果，将α设置为0.6，以兼顾最终融合预测的分类监督与局部精炼分支的判别特征学习。

2.5　分类性能与计算开销分析

为分析MERT在分类性能与计算开销之间的权衡，图4给出了不同ViT细粒度分类方法在Stanford Dogs数据集上的Top-1准确率和浮点运算量（FLOPs）。图中位置越靠近左上方，表示方法在较低计算开销下取得了较高的分类准确率。

由图4可知，MERT以78.6GFLOPs取得92.97%的Top-1准确率，位于散点图的左上区域。与ViT-B/16基线相比，MERT的计算量仅增加0.8G，增幅约为1.0%，而分类准确率提高2.34个百分点。

© 中国图象图形学报版权所有

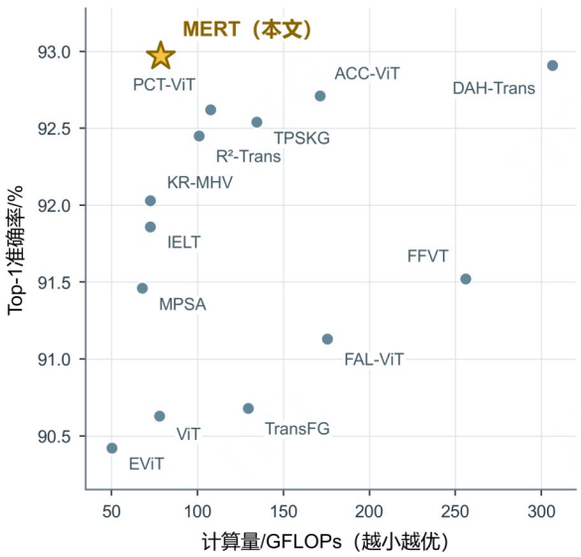

图4　Stanford Dogs数据集上不同方法的分类准确率与计算量比较

Fig. 4　Comparison of classification accuracy and computational cost of different methods on the Stanford Dogs dataset

这表明，层级与头部路由以及关键Token精炼模块在不改变ViT主干完整特征提取过程的情况下，仅引入了较小的额外计算开销。

在准确率相近的方法中，DAH-Trans以306.5GFLOPs取得92.91%的准确率，ACC-ViT以171.2GFLOPs取得92.71%的准确率；MERT在获得更高准确率的同时，计算量分别减少227.9G和92.6G，相较这两种方法的计算量分别降低74.4%和54.1%。与计算量较低的IELT、KR-MHV和MPSA相比，MERT虽然增加了少量计算，但获得了更高的分类准确率。上述结果说明，MERT的性能提升并非依赖大幅增加计算量。结合前述消融结果可知，MERT的性能提升主要得益于跨层路由、注意力头路由及其引导的局部特征精炼，在仅增加少量计算开销的情况下提高了分类准确率。

2.6　可视化分析

为考察不同方法对局部判别特征的关注情况及其分类特征分布，本节采用梯度加权类激活映射（gradient-weighted class activation mapping，Grad-CAM）和t分布随机邻域嵌入（t-distributed stochastic neighbor embedding，t-SNE）进行可视化分析。Grad-CAM用于观察模型分类过程中具有较高响应的空间区域，t-SNE则用于比较不同方法所提取特征的类内紧凑性和类间可分性。

图5展示了ViT、IELT和MERT在Stanford Dogs、Oxford 102 Flowers及NABirds数据集上的Grad-CAM结果。每组从左至右依次为原始图像以及ViT、IELT和MERT的响应图，亮色区域表示对类别预测贡献较大的位置。

从图5所示样本可以看出ViT能够覆盖目标主体，但高响应区域较为分散，部分样本在草地、枝叶等背景位置仍有明显响应，说明分类特征容易受到场景信息干扰。IELT通过Token选择增强了前景响应，但部分样本的激活较弱或较为零散。例如，犬类和鸟类图像中的响应多集中于有限区域，对躯干、翅膀等互补结构的覆盖不足，显示其在部分样本中对多个互补判别部位的覆盖仍相对有限。

相比之下，MERT的响应范围更完整，关注重点

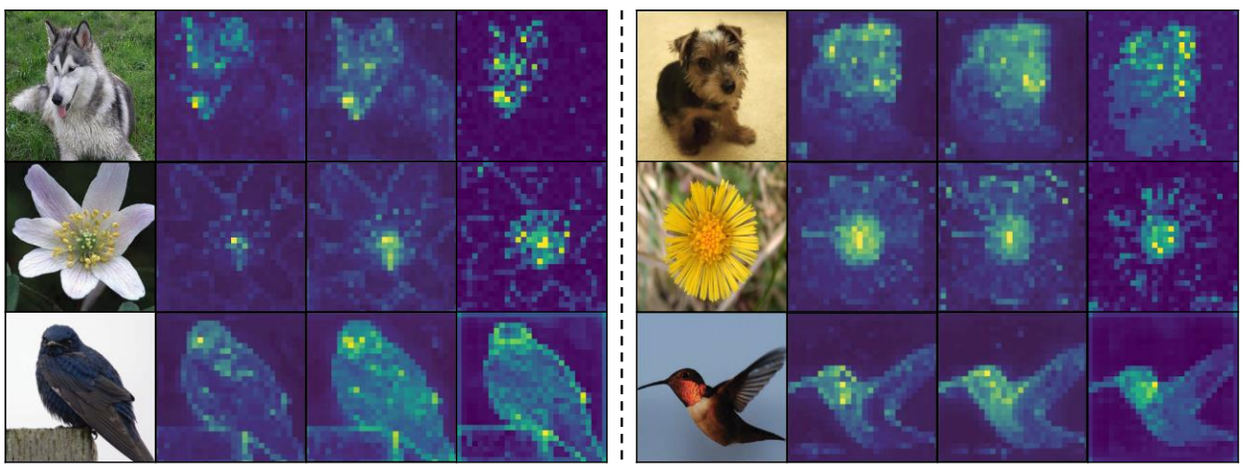

图5　不同方法的Grad-CAM激活图比较

Fig. 5　Comparison of Grad-CAM activation maps from different methods

© 中国图象图形学报版权所有

也更明确。在犬类图像中,高响应区域覆盖面部、耳部、头部轮廓及部分躯干;在花卉图像中,主要集中于花蕊及邻近花瓣;在鸟类图像中,则分布于头部、喙部、胸腹部和翅膀纹理。与上述方法相比,MERT的响应模式有所不同,其并非简单缩小响应区域,而是借助层级与注意力头路由筛选关键Token,使响应更集中于判别部位,避免因过度裁剪而丢失互补线索,在减少部分背景响应的同时保留多个与类别判别相关的局部结构。

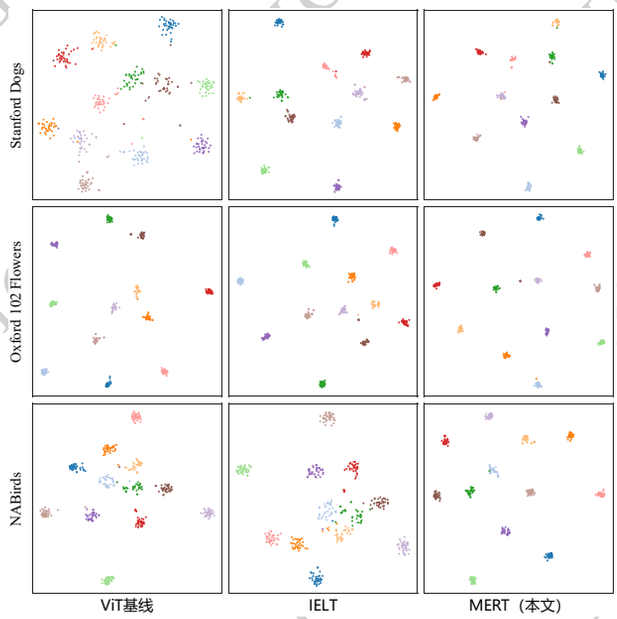

图6　不同方法分类特征的t-SNE可视化结果

Fig. 6　t-SNE visualization of classification features learned by different methods

在Stanford Dogs上,ViT已经形成基本的类别簇,但部分簇的分布范围较大,并存在少量离散样本。IELT改善了部分类别的聚集程度,而MERT形成的类别簇整体更加紧凑,相邻类别之间的混叠有所减少,类别边界也更加清晰。

Oxford 102 Flowers中的3种方法均表现出较好的类别分离,说明颜色、花瓣形态和花蕊结构已经能够提供较稳定的分类信息。MERT的变化主要体现在减少类内离散点和保持类别簇紧凑,而不是显著改变原有分布结构。这一变化与该数据集较高的基线准确率及相对有限的性能增幅相一致。

NABirds包含大量外观相近的鸟类亚种,ViT和IELT仍有多个类别在局部区域相互接近,并伴随一定程度的类内离散。MERT得到的多数类别簇边界更加清楚,分布也更为紧凑,说明经过层级路由、注意力头路由和关键Token选择后,头部、羽毛及翅膀纹理等局部特征能够形成更稳定的类别表示,并减弱相近类别间的特征混叠。

综合Grad-CAM与t-SNE结果,可以观察到MERT在所示样本中具有更集中的局部判别响应和更紧凑的类别特征分布。结合前述消融实验,这些现象进一步支持多维专家路由及其引导的Token选择有助于改善细粒度判别特征的组织与利用。

3　结　论

针对现有ViT细粒度分类方法在多层表征和多头注意力响应利用过程中缺乏样本适应性的问题,提出MERT。MERT利用共享路由信号对不同层级表征和注意力头响应进行样本相关加权与选择,并以联合路由结果引导关键Token筛选和跨层同位特征聚合,形成跨层聚合与局部精炼两条互补分类路径。对比实验结果表明,MERT在Stanford Dogs和Oxford 102 Flowers数据集上取得最优Top-1准确率,在NABirds数据集上也获得了具有竞争力的分类性能。相较ViT-B/16基线,MERT在3个数据集上的Top-1准确率分别提高2.34、1.25和0.24个百分点,而参数量和FLOPs仅增加0.5 M和0.8 G。消融实验进一步表明,层级路由能够改善跨层特征组织,注意力头路由通过DTS作用于关键Token选择,两类路由信息联合使用可获得更好的分类性能。上述结果说明,对ViT已经生成的多层表征和多头注意力响应进行样本相关组织,能够在保持较低额外计算开销的同时增强细粒度局部判别特征的利用。

本文提出的多维专家路由机制为Transformer多层表征与多头注意力响应的后置式样本自适应组织提供了一种通用范式。未来工作可将该机制拓展至医学图像、遥感图像及其他复杂视觉识别场景,以探索后置式多维路由在视觉表征自适应建模中的潜力。

参考文献(References)

Angelova A and Zhu S. 2013. Efficient object detection and segmentation for fine-grained recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland:

© 中国图象图形学报版权所有

IEEE：811-818［DOI：10.1109/CVPR.2013.110］

Dosovitskiy A，Beyer L，Kolesnikov A，Weissenborn D，Zhai X，Unterthiner T，et al. 2021. An image is worth 16×16 words：transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Virtual Conference：OpenReview.net

Fedus W，Zoph B and Shazeer N. 2022. Switch transformers：scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research，23(120)：1-39

Gao Z，Lu J K，Zhao L，Li Q，Chen N H and Li P. 2026. PCT-ViT：a vision transformer incorporating fine-grained perception enhancement and counterfactual token selection. Digital Signal Processing，170：105796［DOI：10.1016/j.dsp.2025.105796］

He J，Chen J N，Liu S，Kortylewski A，Yang C，Bai Y T，et al. 2022. TransFG：a transformer architecture for fine-grained recognition//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Conference：AAAI Press：852-860［DOI：10.1609/aaai.v36i1.19967］

He K M，Zhang X Y，Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE：770-778［DOI：10.1109/CVPR.2016.90］

Huang Y T，Hechen Z Z，Zhou M L，Li Z G and Kwong S. 2025. An attention-locating algorithm for eliminating background effects in fine-grained visual classification. IEEE Transactions on Circuits and Systems for Video Technology，35(6)：5993-6006［DOI：10.1109/TCSVT.2025.3535818］

Khairunnisa M and Wibowo S A. 2025. Multi-head voting based on kernel filtering for fine-grained visual classification. JOIV：International Journal on Informatics Visualization，9(2)：690-698［DOI：10.62527/joiv.9.2.2920］

Khosla A，Jayadevaprakash N，Yao B P and Li F F. 2011. Novel dataset for fine-grained image categorization：Stanford Dogs//Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization. Colorado Springs：IEEE

Li P H，Xie J T，Wang Q L and Zuo W M. 2017. Is second-order information helpful for large-scale visual recognition? //Proceedings of the IEEE International Conference on Computer Vision. Venice：IEEE：2089-2097［DOI：10.1109/ICCV.2017.228］

Liang Y W，Ge C J，Tong Z，Song Y B，Wang J and Xie P T. 2022. Not all patches are what you need：expediting vision transformers via token reorganizations//Proceedings of the 10th International Conference on Learning Representations. Virtual Conference：OpenReview.net

Lin T Y，RoyChowdhury A and Maji S. 2015. Bilinear CNN models for fine-grained visual recognition//Proceedings of the IEEE International Conference on Computer Vision. Santiago：IEEE：1449-1457［DOI：10.1109/ICCV.2015.170］

Liu X D，Wang L L and Han X G. 2022. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing，492：137-149［DOI：10.1016/j.neucom.2022.04.037］

Luo W，Yang X T，Mo X J，Lu Y H，Davis L S，Li J，et al. 2019. Cross-X learning for fine-grained visual categorization//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul：IEEE：8241-8250［DOI：10.1109/ICCV.2019.00833］

Meng L C，Li H D，Chen B C，Lan S Y，Wu Z X，Jiang Y G，et al. 2022. AdaViT：adaptive vision transformers for efficient image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE：12299-12308［DOI：10.1109/CVPR52688.2022.01199］

Michel P，Levy O and Neubig G. 2019. Are sixteen heads really better than one? //Advances in Neural Information Processing Systems 32. Vancouver：Curran Associates，Inc.：14014-14024

Nilsback M E and Zisserman A. 2008. Automated flower classification over a large number of classes//Proceedings of the 6th Indian Conference on Computer Vision，Graphics and Image Processing. Bhubaneswar：IEEE：722-729［DOI：10.1109/ICVGIP.2008.47］

Rao Y M，Zhao W L，Liu B L，Lu J W，Zhou J and Hsieh C J. 2021. DynamicViT：efficient vision transformers with dynamic token sparsification//Advances in Neural Information Processing Systems 34. Virtual Conference：Curran Associates，Inc.：13937-13949

Riquelme C，Puigcerver J，Mustafa B，Neumann M，Jenatton R，Pinto A S，et al. 2021. Scaling vision with sparse mixture of experts//Advances in Neural Information Processing Systems 34. Virtual Conference：Curran Associates，Inc.：8583-8595

Shazeer N，Mirhoseini A，Maziarz K，Davis A，Le Q V，Hinton G E，et al. 2017. Outrageously large neural networks：the sparsely-gated mixture-of-experts layer//Proceedings of the 5th International Conference on Learning Representations. Toulon：OpenReview.net

Touvron H，Sablayrolles A，Douze M，Cord M and Jégou H. 2021. Grafit：learning fine-grained image representations with coarse labels//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal：IEEE：854-864［DOI：10.1109/ICCV48922.2021.00091］

Van Horn G，Branson S，Farrell R，Haber S，Barry J，Ipeirotis P G，et al. 2015. Building a bird recognition app and large scale dataset with citizen scientists：the fine print in fine-grained dataset collection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE：595-604［DOI：10.1109/CVPR.2015.7298658］

Voita E，Talbot D，Moiseev F，Sennrich R and Titov I. 2019. Analyzing multi-head self-attention：specialized heads do the heavy lifting，the rest can be pruned//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence：Association for Computational Linguistics：5797-5808［DOI：10.18653/v1/P19-1580］

Wang J，Yu X H and Gao Y S. 2021. Feature fusion vision transformer

© 中国图象图形学报版权所有

for fine-grained visual categorization//Proceedings of the 32nd British Machine Vision Conference. Online: BMVA Press: article 170 [DOI: 10.5244/C.35.192]

Wang J H, Xu Q, Jiang B, Luo B and Tang J H. 2024. Multi-granularity part sampling attention for fine-grained visual classification. IEEE Transactions on Image Processing, 33: 4529-4542 [DOI: 10.1109/TIP.2024.3441813]

Xu Q, Wang J H, Jiang B and Luo B. 2023. Fine-grained visual classification via internal ensemble learning transformer. IEEE Transactions on Multimedia, 25: 9015-9028 [DOI: 10.1109/TMM.2023.3244340]

Yang H J, He H M, Li Y and Cao F Y. 2026. Revisiting fine-grained classification: a dual-branch method for noise-resilient and global-local discriminative learning. Expert Systems with Applications, 295: 128802 [DOI: 10.1016/j.eswa.2025.128802]

Yang Z, Luo T G, Wang D, Hu Z Q, Gao J and Wang L W. 2018. Learning to navigate for fine-grained classification//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer: 438-454 [DOI: 10.1007/978-3-030-01264-9_26]

Yang Z M and Song W. 2023. Selecting and fusing coarse-and-fine granularity features for fine-grained image recognition. Journal of Image and Graphics, 28(7): 2081-2092 (阳治民, 宋威. 2023. 选择并融合粗细粒度特征的细粒度图像识别. 中国图象图形学报, 28(7): 2081-2092) [DOI: 10.11834/jig.220052]

Ye S, Yu S J, Wang Y and You X G. 2024. R^2-Trans: fine-grained visual categorization with redundancy reduction. Image and Vision Computing, 143: 104923 [DOI: 10.1016/j.imavis.2024.104923]

Zhang Z C, Chen Z D, Wang Y X, Luo X and Xu X S. 2024. A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information. Pattern Recognition, 145: 109979 [DOI: 10.1016/j.patcog.2023.109979]

Zhuang P Q, Wang Y L and Qiao Y. 2020. Learning attentive pairwise interaction for fine-grained classification//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press: 13130-13137 [DOI: 10.1609/aaai.v34i07.7016]

作者简介

任玥赫,男,硕士研究生,主要研究方向为计算机视觉。E-mail:sduferyh@mail.sdufe.edu.cn

刘博,通信作者,男,讲师,主要研究方向为多源遥感图像处理。E-mail:liubo24@sdufe.edu.cn

迟静,女,教授,主要研究方向为计算机动画、图形图像智能处理及大数据建模。E-mail:chijing@sdufe.edu.cn

张云峰,男,教授,主要研究方向为计算机视觉与人工智能。E-mail:yfzhang@sdufe.edu.cn

邢怡楠,女,硕士研究生,研究方向为遥感图像处理。E-mail:xingyinan@mail.sdufe.edu.cn

© 中国图象图形学报版权所有