小样本细粒度图像分类的Mamba-小波多尺度建模方法

仝傲[1]，任劼[1]，孟宗阳[1]，鲁磊[2]
（1. 西安工程大学电子信息学院，陕西 西安 710600；
2. 西安交通大学信息与通信工程学院，陕西 西安 710049）

摘　要：小样本细粒度图像分类旨在在有限标注样本条件下识别类别间细微差异，广泛应用于智能识别、生态监测及自动驾驶等领域。现有卷积结构受限于固定感受野和局部建模方式，对多尺度特征的关联描述不足，注意力或频域方法虽提升了细粒度特征的判别性，但在跨尺度依赖建模与特征融合方面仍存在局限。为提升多尺度细粒特征的表达能力，提出了一种小样本细粒度图像分类的Mamba-小波多尺度建模方法，该方法构建了Mamba状态空间建模的多尺度特征关系网络（MSFRNet）。该网络包含两大核心创新模块：小波引导动态Mamba多尺度特征提取（WDMFE）模块与交叉尺度注意力融合（CAF）模块。其中，WDMFE模块通过小波引导的动态自适应Mamba结构强化不同尺度下的频率感知与上下文建模，CAF模块采用通道与空间注意力机制整合多尺度特征以实现跨尺度补充。实验结果在CUB-200-2011、Stanford-Dogs和Stanford-Cars等基准数据集上获得了较高分类准确率，并呈现出稳定的性能提升。结果表明，该网络能够有效增强细粒度特征表达与跨任务泛化能力，并为小样本细粒度识别模型的多尺度建模提供可拓展框架。

关键词：小样本细粒度图像分类；Mamba状态空间模型；多尺度特征建模；小波引导特征提取；注意力机制；特征融合
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202606

Mamba-wavelet-based multi-scale modeling method for few-shot fine-grained image classification

Tong Ao[1], Ren Jie[1], Meng Zongyang[1], Lu Lei[2]
1. School of Electronic Information, Xi'an Polytechnic University, Xi'an 710600, China
2. School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Abstract: Fine-grained few-shot image classification aims to recognize subtle inter-class differences under limited annotated samples and has been widely applied in intelligent recognition, ecological monitoring, and autonomous driving. However, existing convolutional architectures are constrained by fixed receptive fields and local modeling schemes, resulting in insufficient characterization of multi-scale feature relationships. Although attention-based or frequency-domain methods have improved the discriminability of fine-grained features, limitations still exist in modeling cross-scale dependencies and feature fusion. To address these issues, a Mamba-wavelet-based multi-scale modeling method for few-shot fine-grained image classification was proposed. Specifically, a multi-scale feature relation network (MSFRNet) based on Mamba state space modeling was constructed. The proposed network consisted of two core modules, namely a wavelet-guided dynamic Mamba multi-scale feature extraction (WDMFE) module and a cross-scale attention fusion (CAF) mod-

收稿日期：2025−12−05；修回日期：2026−03−02
通信作者：任劼，renjie@xpu.edu.cn
基金项目：陕西省自然科学基础研究计划（No.2025JC-YBMS-765）；陕西省教育厅重点项目（No.23JY029）
Foundation Items: The Shaanxi Province Basic Research Program in Natural Sciences (No.2025JC-YBMS-765), The Key Project of Shaanxi Provincial Department of Education (No.23JY029)

ule. In the WDMFE module, a wavelet-guided dynamic adaptive Mamba structure was introduced to enhance frequency perception and contextual modeling across different scales. In the CAF module, multi-scale features were integrated through channel and spatial attention mechanisms to achieve cross-scale feature complementation. Experimental results on benchmark datasets, including CUB-200-2011, Stanford Dogs, and Stanford Cars, demonstrated that higher classification accuracy was achieved and stable performance improvements were obtained. It is concluded that the proposed network effectively enhances fine-grained feature representation and cross-task generalization ability, and provides a scalable framework for multi-scale modeling in few-shot fine-grained classification.

Key words: fine-grained few-shot classification, Mamba state space model, multi-scale feature modeling, wavelet-guided feature extraction, attention mechanism, feature fusion

0　引言

小样本细粒度图像分类（fine-grained few-shot image classification，FGFSIC）[1]任务是指在每类仅有少量标注样本的条件下，对鸟类、花卉或车辆等全局外观高度相似，但在局部纹理或结构特征上存在细微差异的细粒度类别进行识别。该任务在生态监测、智能制造与医疗诊断等领域具有重要应用价值，能够在样本稀缺的情况下实现新类别的快速识别与迁移学习，为实际场景中的智能视觉感知提供关键支撑。传统深度学习方法在大规模数据下能够有效提取全局特征，但在样本数量不足的小样本场景下，模型容易出现过拟合全局特征而忽略局部细节，从而限制了分类性能的提升。因此，如何在有限样本下捕捉细粒特征差异，是FGFSIC任务中的核心挑战。

当前FGFSIC研究虽取得一定进展，但仍面临特征建模与多尺度信息融合的双重瓶颈。现有方法多依赖单尺度或全局特征聚合，卷积神经网络（convolutional neural networks，CNN）固定感受野限制全局依赖建模，Transformer则因高复杂度在小样本下易过拟合、出现特征漂移[2]；与此同时，仅依赖单一空间域特征难以充分刻画细粒度目标的复杂结构。已有研究表明，在多种视觉任务中引入多尺度特征建模能够通过融合局部与全局信息或跨尺度交互提升特征判别能力[3-5]，并在小样本学习（few-shot learning，FSL）任务中进一步增强模型对细粒度差异的感知能力[6]。然而，现有方法大多仍局限于空间域特征建模，难以同时兼顾不同尺度与不同频率层次的信息表达，从而限制了细粒度特征的判别能力。Mamba状态空间模型（state space model，SSM）以线性复杂度实现全局依赖建模，能够在保持轻量化的同时建模长程序列关系，为小样本场景下的结构化特征学习提供了新的思路。

此外，双向特征重构网络（bi-directional feature reconstruction network，BiFRN）[7]通过构建双向特征重建机制对嵌入特征进行约束，使同类别样本在特征空间中更加紧致，同时扩大不同类别之间的间隔，从而在小样本分类任务中取得较好的性能。受该思想启发，本文在BiFRN框架基础上进行改进，提出一种小样本细粒度分类的Mamba-小波多尺度建模方法，并构建多尺度特征关系网络（multi-scale feature relation network，MSFRNet）作为实现框架。

MSFRNet整体由3个核心组成部分构成：小波引导动态Mamba多尺度特征提取（wavelet-guided dynamic Mamba multi-scale feature extraction，WDMFE）模块、交叉尺度注意力融合（cross-scale attention fusion，CAF）模块和特征重构与分类模块。其中，WDMFE与CAF模块为本文核心创新模块，二者协同作用，实现了多尺度空间结构与多频率细节特征的联合建模，从而全面捕获细粒度目标的显著特征。

具体而言，WDMFE模块采用动态自适应Mamba卷积网络进行多尺度特征提取，在传统卷积结构中引入动态状态空间建模与小波分解机制，每个卷积阶段均嵌入小波引导动态自适应Mamba模块（WDA-Mamba），通过离散小波变换（DWT）与动态自适应状态空间混合（DASSM）实现频率感知的局部上下文建模。在此基础上，CAF模块通过通道与空间注意力机制自适应融合不同尺度特征，生成可解释的尺度权重，实现跨尺度的信息互补与增强。经过处理后的融合特征被输入特征重构与分类模块，结合其特征关系建模机制完成支持集与查询集间的特征匹配，最终实现小样本分类性能的有效提升。

1 相关工作

1.1 基于CNN、Transformer与混合架构的小样本细粒度方法

FGFSIC的核心挑战在于如何在有限样本下建模局部细节与全局结构之间的差异性。为解决这一问题，研究者围绕特征提取主干的不同架构形式进行了大量探索，主要可分为基于CNN、基于视觉Transformer（ViT）以及网络架构融合的混合结构。

早期的小样本细粒度学习方法多以CNN为特征提取主干，通过局部建模增强判别性区域的响应，以缓解小样本条件下全局特征的过拟合问题，如Li等[8]提出局部描述符的图像到类别度量，Wertheimer等[9]进一步将局部特征建模转化为特征图重建问题。然而，这些方法依赖单尺度特征，难以捕捉跨尺度的细粒度差异。为解决上述问题，后续研究尝试引入注意力机制以强化局部特征的筛选与融合，如Li等[10]利用空间和通道注意力机制来定位判别性区域。尽管在一定程度上改善了CNN的表达能力，但其注意力建模仍局限于局部范围，而且多尺度融合策略多为静态加权，特征响应难以根据样本复杂性动态调整。

随着Transformer架构的兴起，研究者进一步利用自注意力机制实现全局依赖建模，以克服CNN局部建模的限制[11-12]。He等[13]提出的方法在细粒度识别任务中引入多头注意力机制，通过标记交互实现跨区域特征对齐，从而增强模型对微小类别差异的感知能力。ViT结构使模型能够同时捕获全局与上下文信息，但其计算复杂度随特征维度平方增长，导致在小样本场景中训练不稳定且易出现过拟合。此外，Transformer缺乏CNN的归纳偏置，对局部纹理和形状的敏感度不足，使得在细粒度任务中的性能提升仍受限。

为兼顾CNN的局部细节捕捉能力与Transformer的全局建模优势，近期研究趋势转向混合或融合架构。例如，Lin等[14]提出的卷积型Transformer（ConvFormer）与Dai等[15]设计的卷积注意力网络（CoAtNet）在前层采用卷积结构提取低层纹理特征，在高层引入自注意力实现全局语义聚合，显著提升了模型的判别性与泛化能力。这类CNN-Transformer融合模型证明了多层次表征对于细粒度特征建模的重要性，但其计算复杂度与结构设计仍相对较高，难以在资源受限的小样本场景中高效部署。

近年来，受状态空间模型启发的Mamba架构被引入视觉任务中，以线性复杂度实现全局依赖建模。与注意力机制不同，Mamba通过动态状态更新捕获长程依赖关系，在保持轻量化的同时兼具CNN的局部建模效率与Transformer的全局建模能力。已有研究表明，Mamba框架凭借其线性复杂度与动态状态更新机制，在多类视觉任务中展现出优越的特征建模能力与泛化性能。例如，Zhu等[16]将Mamba状态空间模型引入视觉表征学习，通过双向选择性扫描机制实现对二维图像特征的高效建模；Wang等[17]在医学图像分割中引入Mamba，实现了在保持轻量化结构的同时显著提升边界分割精度；Lu等[18]则在高光谱图像分类任务中利用Mamba的序列建模特性实现了光谱−空间特征的高效融合。这些成果充分验证了Mamba在复杂视觉场景下的建模潜力与计算优势。

然而，在细粒度小样本图像分类领域，Mamba的潜力尚未得到充分挖掘。现有研究多集中于全局序列建模，忽视了对多尺度细粒特征与频域信息的联合建模，因此，如何将Mamba的动态状态建模机制与多尺度特征表达及频域相结合，成为进一步提升FGFSIC性能的关键方向。

1.2 基于多尺度建模的小样本细粒度方法

多尺度特征融合是提升细粒特征覆盖度的关键思路，但现有方法多缺乏对尺度相关性的自适应建模。原型网络ProtoNet[19]作为经典度量学习方法，通过全局平均池化（global average pooling，GAP）生成类别原型，完全丢失尺度信息。特征嵌入自适应网络（feature embedding adaptation transfer，FEAT）[20]引入Transformer实现多尺度特征交互，却因二次计算复杂度难以处理高分辨率细粒特征。

近年来，研究者们提出了多种改进方案。Ding等[21]通过显式建模不同尺度特征间的互补关系，但其尺度权重分配策略相对固定。Zhang等[22]通过最优传输理论匹配多尺度特征、保留空间结构，却存在计算成本高的问题。与此相比，Chen等[23]在ImageNet数据集预训练基础上，通过全局池化后的特征进行余弦相似度度量，虽然大幅简化了模型流程，但由于缺乏对多尺度结构的显式建模，其在细粒度特征对齐与判别性方面仍存在不足。

为在效率与多尺度建模能力间取得平衡，研究

者开始尝试引入SSM以实现低复杂度的全局建模。如视觉Mamba模型（VMamba）[24]提出了二维选择性扫描（SS2D）机制，以线性复杂度处理2D特征，但其扫描策略（固定垂直/水平方向）无法根据细粒度特征分布动态调整尺度关注区域。类似地，高效视觉状态空间模型（EVSSM）[25]通过设计视觉扫描块和几何变换，在图像恢复任务中实现了高效的空间域长程依赖建模，但主要关注固定空间结构的扫描，尚未探索多尺度特征建模。这些方法在效率和长程依赖捕获上具有潜力，为SSM框架与更精细的特征分析工具（如频域分解）相结合提供了研究空间。

1.3　基于频率域建模的特征增强方法

频率域分析为细粒特征建模提供了新视角。早期的研究表明，图像在空间域中表现出的复杂结构往往可以通过频率域变换获得更具判别性的表征。Baaziz等[26]率先在纹理特征提取中系统研究了空间–频率域的联合建模方法，通过引入离散小波变换（discrete wavelet transform，DWT）等多尺度分解，实现了纹理的细粒度表达与特征增强。在此基础上，研究者开始将频域分解方法引入更复杂的视觉任务中，以进一步增强特征判别性。Ahmad等[27]将小波变换与Mamba结合，通过哈尔小波（Haar）分解高光谱数据的空间–光谱特征。然而，该方法仅适用于高光谱数据，难以推广至自然场景细粒度图像。

近年来，部分工作开始探索在细粒度小样本学习中融合频率与空间特征，以实现更稳健的特征表征。例如，Zhu等[28]设计多频邻域模块提取多频结构表示，使模型在复杂背景下保持稳定判别能力。Guo等[29]进一步提出空间–频率特征融合网络，在小数据集细粒度分类中通过可学习权重动态融合空间特征与频域特征，增强模型的细节敏感性与抗噪性。此外，Sun等[30]将小波卷积引入小样本缺陷检测任务，实现频率卷积与原型学习的联合优化，展示了小波卷积在小样本学习中的潜力。

总体而言，尽管上述方法在不同任务中展示了频率域融合的潜力，但现有细粒度小样本分类方法仍难以同时兼顾局部判别性、尺度鲁棒性与频率敏感性。未来的研究可进一步探索如何将Mamba的动态状态建模机制与频域特征分解相结合，以实现跨尺度、跨频率的协同增强。

2　方法框架

2.1　问题定义

给定一个包含N个类别的数据集，将其划分为3个互不重叠的子集：训练集D_{train}、验证集D_{val}和测试集D_{test}，满足$D_{\text{train}}\cap D_{\text{val}}\cap D_{\text{test}}=\varnothing$，$D_{\text{train}}\cup D_{\text{val}}\cup D_{\text{test}}=D$。

在FSL任务中，每次实验由支持集S和查询集Q组成。两者共享相同的标签空间，并且每个类别都包含k个标注样本。支持集S中包含c个不同的图像类别（$c\in C$），用于提供少量标注样本以支持模型的学习；查询集Q则用于评估模型在这些类别上的分类能力。

小样本学习的目标是训练一个模型，使其能够将查询样本$X_q\in Q$正确分类到对应的c个类别中。按照小样本任务的设定方式，该任务被称为“c-way k-shot”任务，其中“way”表示类别数，“shot”表示每个类别的样本数。

2.2　基于Mamba的多尺度特征关系网络总体架构

针对上述不足，本文构建了MSFRNet，通过将小波引导的动态自适应Mamba结构融入特征提取过程，并结合多尺度注意力融合机制，实现频域与空间域的协同建模。MSFRNet的整体框架如图1所示。

如图1所示，支持集与查询集图像首先通过WDMFE生成多尺度特征表示。该模块以小波变换为引导，采用多尺度并行分支提取不同频率与层级的特征：低尺度分支捕获局部纹理，中尺度分支提取区域语义，高尺度分支结合动态小波与Mamba状态空间结构建模全局上下文。在多尺度特征提取后，各尺度分支内部的WDA-Mamba进一步对特征进行频域增强与动态建模。

随后，来自不同尺度分支的特征被送入CAF进行统一融合。该模块融合通道注意力与空间注意力机制，自适应地强调关键通道与显著空间区域，实现跨尺度语义的高效整合。

最终，融合后的特征被输入特征重构与分类模块完成类别判别。该模块由特征空间重构模块（feature space reconstruction module，FSRM）和特征映射重构模块（feature mapping reconstruction module，FMRM）及欧氏距离计算单元组成。其中，FSRM通过重建支持样本特征完成类别原型建模，FMRM进一步建立查询特征与类别表示之间的映射关系，

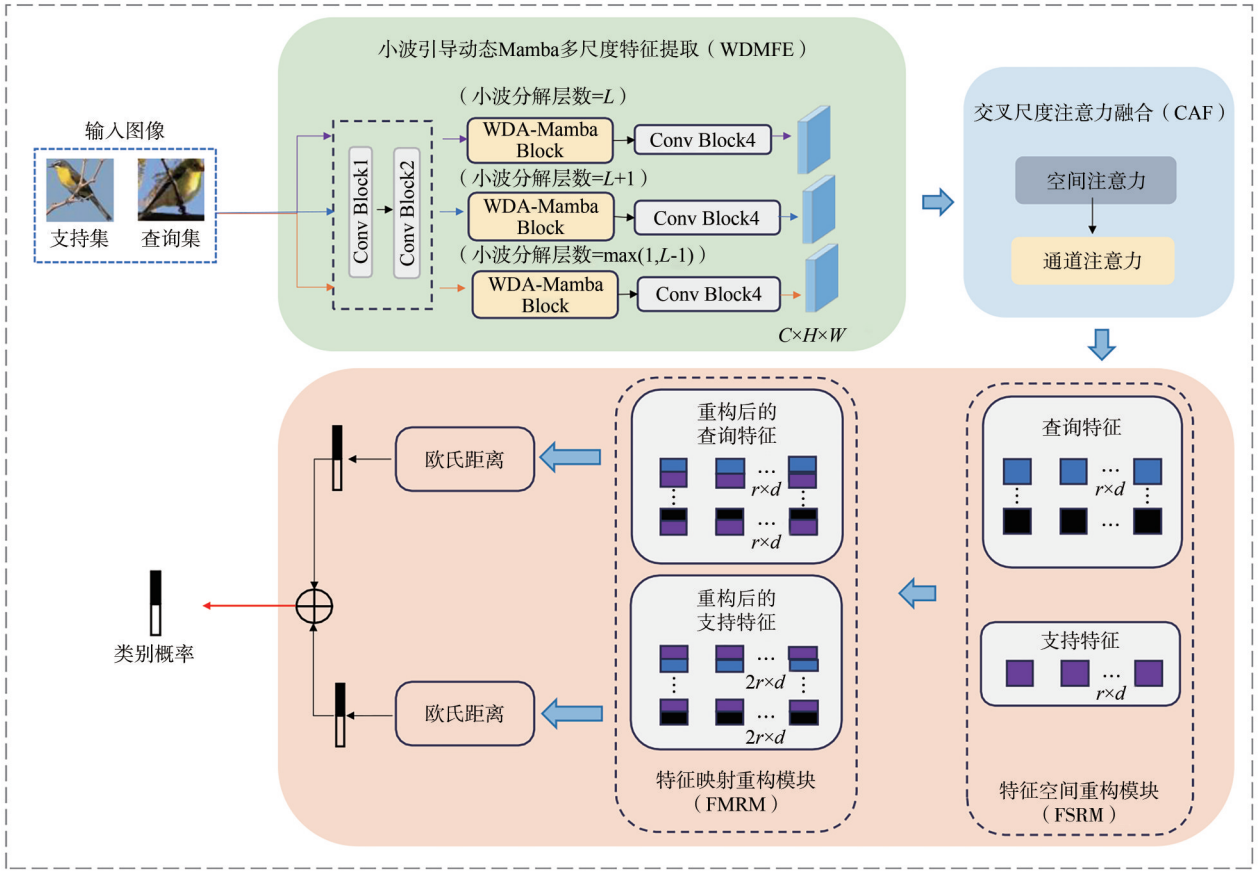

图 1　MSFRNet 的整体框架

挖掘支持集与查询集之间的语义对应关系，最终通过欧氏距离计算查询样本与各类别原型之间的相似度并输出类别概率。

2.3　小波引导动态 Mamba 多尺度特征提取模块

WDMFE 通过将小波引导的动态自适应 Mamba 结构融入特征提取过程来进行输入图像的多尺度特征提取，其核心组件是小波引导动态 Mamba 块（WDA-Mamba Block）。模块整体由 3 个并行分支组成，分别对应高分辨率、中分辨率与低分辨率路径，以实现从局部细节到全局语义的多层级建模（图 1 中标注为“WDMFE”）。

给定输入图像矩阵 $\boldsymbol{I}\in R^{3\times H\times \mathrm{W}}$（其中 H、W 分别表示图像的高度和宽度），WDMFE 将输入特征分配到 3 条独立路径中，每条路径均包含完整的卷积层级与 WDA-Mamba Block，但采用不同的小波分解深度，从而分别捕获细粒度纹理、中层语义与全局上下文信息。

$$\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3=\mathrm{Backbone}(\boldsymbol{I}) \tag{1}$$

其中，$\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$ 分别代表高、中、低尺度特征。三者互补，$\boldsymbol{F}_1$ 侧重捕获细粒度的局部纹理与边缘信息，$\boldsymbol{F}_2$ 关注于区域级的语义结构，$\boldsymbol{F}_3$ 通过更大感受野表征全局上下文信息。

每个尺度分支内部均嵌入一个 WDA-Mamba Block，如图 2 所示，输入特征首先经过归一化与卷积预处理后，通过小波分解提取低频与高频子带特征，之后将特征送入动态自适应状态空间混合与状态空间模型进行长距离依赖建模；建模后经逆小波重构与残差连接输出增强特征，实现频域感知的细粒度特征提取，具体过程如下。

对于输入特征 $\boldsymbol{X}\in R^{C\times H\times W}$（其中 C 表示特征图通道数），首先执行层归一化（LayerNorm）与残差深度可分离卷积（ResDWC）进行局部结构强化与特征平衡。

$$\boldsymbol{X}'=\mathrm{ResDWC}\left(\mathrm{LayerNorm}\left(\boldsymbol{X}\right)\right) \tag{2}$$

其中，$\boldsymbol{X}'$ 表示经过卷积与归一化后的增强特征图，该步骤在保持空间分辨率的同时，通过通道独立的卷积操作捕获局部上下文信息，并利用残差路径防止梯度消失。

为了引入多分辨率频率感知能力，对增强特征图 $\boldsymbol{X}'$ 进行二维离散小波变换（2D-DWT）。

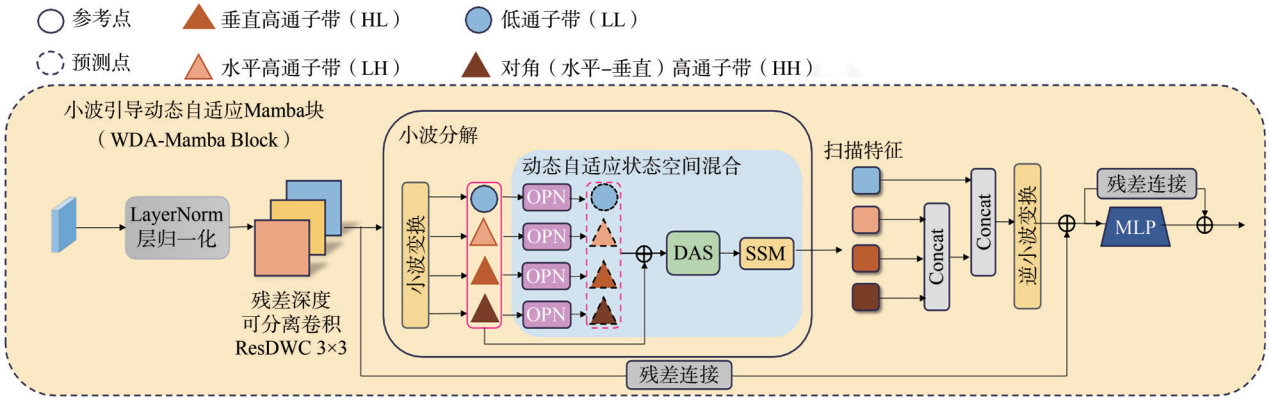

图2　WDA-Mamba Block 结构

$$\left\{\boldsymbol{X}_{\mathrm{LL}},\boldsymbol{X}_{\mathrm{LH}},\boldsymbol{X}_{\mathrm{HL}},\boldsymbol{X}_{\mathrm{HH}}\right\}=\mathrm{DWT}\left(\boldsymbol{X}'\right) \tag{3}$$

其中，$\boldsymbol{X}_{\mathrm{LL}}$ 表示低频结构信息，$\boldsymbol{X}_{\mathrm{LH}}$、$\boldsymbol{X}_{\mathrm{HL}}$、$\boldsymbol{X}_{\mathrm{HH}}$ 表示不同方向的高频边缘与纹理。通过小波分解，特征空间被映射到多频域表征，使网络能够同时关注图像的全局语义与局部细节。

之后，使用 DASSM 对小波分解得到的多频特征进行联合建模与加权融合得到 $\tilde{x}$。

$$\tilde{x}=\mathrm{DASSM}\left\{\boldsymbol{X}_{\mathrm{LL}},\boldsymbol{X}_{\mathrm{LH}},\boldsymbol{X}_{\mathrm{HL}},\boldsymbol{X}_{\mathrm{HH}}\right\} \tag{4}$$

DASSM 模块以频域融合为核心，内部集成了输入投影与深度可分离卷积（OPN）、动态自适应扫描（dynamic adaptive scan，DAS）以及 SSM 3 种机制，实现了频域、空间域与时序依赖的统一建模。具体而言，首先通过 OPN 对输入特征进行空间增强。

$$\boldsymbol{X}_{\mathrm{opn}}=\mathrm{SiLU}\left(\mathrm{Conv}_{\mathrm{dw}}\left(\mathrm{Conv}_{1\times1}\left(\boldsymbol{X}\right)\right)\right) \tag{5}$$

其中，$\mathrm{Conv}_{\mathrm{dw}}(\cdot)$ 为深度可分离卷积操作，$\mathrm{SiLU}(\cdot)$ 为激活函数，$\boldsymbol{X}_{\mathrm{opn}}$ 为空间增强后的输出特征，上述操作用于捕获局部上下文特征。随后，DAS 机制在空间维度上自适应地捕获局部与全局依赖关系。

$$\boldsymbol{X}_{\mathrm{das}}=\mathrm{DAS}\left(\boldsymbol{X}_{\mathrm{opn}},\boldsymbol{X}\right) \tag{6}$$

对于缺乏序列一致性和远程依赖的问题，引入基于 Mamba 结构的状态空间 SSM 在通道维度建模长程依赖。

$$\boldsymbol{Y}=\mathrm{SSM}\left(\boldsymbol{X}_{\mathrm{das}}\right)+\boldsymbol{X}' \tag{7}$$

SSM 通过选择性扫描机制在通道维度内建模长程依赖，同时保持线性复杂度。

融合后的特征 $\boldsymbol{Y}$ 经逆小波变换（IDWT）将频域特征重新投影至空间域，实现局部纹理的还原与全局结构的重构。

$$\hat{\boldsymbol{X}}=\mathrm{IDWT}\left(\boldsymbol{Y}\right) \tag{8}$$

最后通过多层感知机（MLP）与残差连接增强表达能力。

$$\mathbf{Output}_i=\hat{\boldsymbol{X}}+\mathrm{MLP}\left(\hat{\boldsymbol{X}}\right) \tag{9}$$

2.4　交叉尺度注意力融合模块

低尺度特征包含丰富的局部纹理信息，而高尺度特征则提供更强的语义表达。为实现多尺度特征的自适应融合，CAF 引入了通道注意力与空间注意力两级机制，有效平衡局部与全局特征。通道注意力机制通过全局平均池化与轻量 MLP 生成权重向量，学习不同尺度的通道重要性，经通道加权后的特征在尺度间进行逐元素求和，实现多尺度语义的自适应融合。随后引入空间注意机制，通过 1×1 卷积与 Softmax 生成空间注意图，对融合特征进行位置再标定，突出关键目标区域并抑制背景干扰。最终输出的融合特征兼具全局语义一致性与局部细节辨识力，为后续的特征重构模块提供高质量输入，CAF 结构如图 3 所示。

具体而言，给定来自 3 个尺度的特征：

$$\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3\in\mathbb{R}^{3\times H\times W} \tag{10}$$

在第一阶段，模块通过 GAP 提取每个尺度特征的统计描述。

$$\boldsymbol{Z}_i=\mathrm{GAP}\left(\boldsymbol{F}_i\right)\in\mathbb{R}^{C},i\in\{1,2,3\} \tag{11}$$

描述向量 $\boldsymbol{Z}_i$ 表征了特征图在每个通道维度上的平均响应，反映了该尺度的整体语义贡献，i 为尺度索引，GAP 表示全局平均池化操作。随后，将其输入一个两层的 MLP，并通过 Sigmoid 激活函数生成通道注意力权重。

$$\boldsymbol{\alpha}_i=\sigma\left(\mathrm{MLP}\left(\boldsymbol{Z}_i\right)\right) \tag{12}$$

其中，$\boldsymbol{\alpha}_i\in R^{C}$ 表示第 i 个尺度特征在各通道上的权

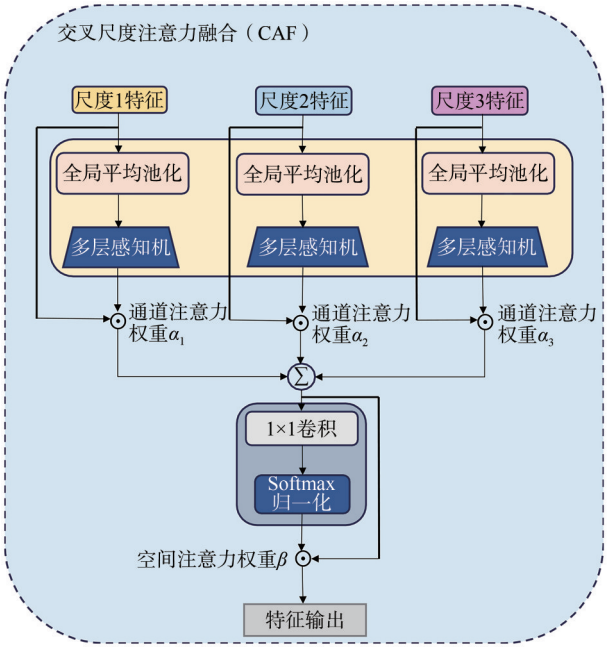

图3 CAF结构

重，σ表示Sigmoid激活函数，该权重可自适应调节不同尺度特征在通道维度的贡献程度。通道注意力权重随后作用于原始特征图，进行逐通道加权。

$$\tilde{\boldsymbol{F}}_i=\boldsymbol{\alpha}_i\odot\boldsymbol{F}_i \tag{13}$$

经过通道加权后，各尺度特征通过逐元素求和方式进行融合。

$$\boldsymbol{F}_c=\sum_{i=1}^{3}\tilde{\boldsymbol{F}}_i \tag{14}$$

该加法操作可理解为一种学习到的跨尺度残差组合，其中每个尺度的贡献由其通道权重$\boldsymbol{\alpha}_i$动态控制。与传统的拼接或直接求和不同，这种融合策略能够根据任务需求自动平衡深层与浅层特征的重要性，当高层语义更具判别力时，模型倾向于提高深层特征的权重；而在需要细节辨识时，则增强浅层特征的贡献。

尽管通道注意机制提升了语义层面的区分能力，但仍未显式考虑空间维度上的特征差异。为此，CAF在融合后引入了空间注意力机制，以进一步突出关键区域并抑制背景噪声，具体步骤如下。

首先，通过1×1卷积和Softmax归一化生成空间注意图。

$$\boldsymbol{\beta}=\mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}\left(\boldsymbol{F}_c\right)\right) \tag{15}$$

其中，$\boldsymbol{\beta}\in R^{1\times H\times W}$表示每个空间位置的注意分布。随后，空间注意图通过逐像素乘法作用于融合特征。

$$\boldsymbol{F}_{\mathrm{out}}=\boldsymbol{\beta}\odot\boldsymbol{F}_c \tag{16}$$

该操作可增强目标区域的激活响应，削弱无关背景区域，并在空间维度上实现特征的显著性增强。

最终输出的融合特征表示为：

$$\boldsymbol{F}_{\mathrm{CAF}}=\boldsymbol{F}_{\mathrm{out}}\in\mathbb{R}^{C\times H\times W} \tag{17}$$

3 实验与结果分析

3.1 实验数据集

本文的实验在3个常用于小样本细粒度图像分类的公开数据集上进行，包括CUB-200-2011[31]、Stanford-Dogs[32]和Stanford-Cars[33]。

CUB-200-2011[31]：该数据集包含200种鸟类，共11 788张图像，每个类别之间差异细微，主要体现在形状、羽毛纹理和色彩上，是细粒度分类领域的经典难题。

Stanford-Dogs[32]：该数据集包含120种犬类，共20 580张图像，来源于ImageNet，图像中存在较大的姿态、尺度及背景变化，能够有效检验模型在复杂场景下的泛化能力。

Stanford-Cars[33]：该数据集包含196种汽车型号，共16 185张图像，该数据集中不同车型间的差异非常细微，对模型的特征判别能力提出较高要求。

对于每个数据集，本文按照常规的小样本学习设置，将其划分为3个互不重叠的子集：训练集D_{train}、验证集D_{val}和测试集D_{test}。数据集的划分情况见表1，其中，N_{train}、N_{val}和N_{test}分别为训练集、验证集和测试集的类数，N_{total}为总类数。

表1 数据集的划分情况

数据集	N_{train}	N_{val}	N_{test}	N_{total}
CUB-200-2011	130	20	50	200
Stanford-Dogs	70	30	20	120
Stanford-Cars	130	17	49	196

3.2 实验细节

本文在标准的“c-way k-shot”小样本学习框架下评估模型性能。在每一次任务单元（episode）中，从测试集中随机采样C个类别，每个类别选取K张标注样本组成支持集S，并从相同类别中额外

选取Q张未标注样本构成查询集Q，模型需根据支持集预测查询样本的类别标签。本文默认采用5-way 1-shot与5-way 5-shot设置，每个训练任务的迭代周期（epoch）包含600个episode，测试阶段在10 000个随机episode上评估模型性能，并报告平均准确率及95%置信区间。

在实现上，本文的网络框架分别采用ResNet-12与Conv-4作为骨干网络进行实验比较。每个卷积块由卷积层、批归一化层和ReLU激活组成。ResNet-12包含4个残差块，通过全局平均池化生成维度为640的特征表示。所有输入图像均调整为84×84并归一化至[0,1]。优化器采用随机梯度下降，动量设为0.9，权重衰减系数为5×10^{-4}。初始学习率为0.1，并采用余弦退火周期重启策略动态调整学习率，每个训练时的小批量单元（batch）包含16个episode。训练阶段使用随机裁剪、水平翻转与颜色扰动等数据增强策略以提升模型泛化能力，测试阶段仅采用中心裁剪与归一化操作。本文提出的模块在端到端框架下联合优化，所有实验均在PyTorch 2.0平台上实现，并在NVIDIA RTX 4090 GPU上运行。

3.3　实验结果

表2与表3给出了本文提出的MSFRNet框架与多种主流小样本学习方法在CUB-200-2011[31]、Stanford-Dogs[32]和Stanford-Cars[33]3个细粒度数据集上的定量比较结果，其中表2的骨干网络是Conv-4，表3的骨干网络是ResNet-12。所有方法均在5-way 1-shot和5-way 5-shot设置下进行评估，实验协议与3.2节一致，表格结果以平均分类准确率及其95%置信区间的形式展现。

在Conv-4网络中，与基准模型BiFRN相比，本文模型在CUB-200-2011、Stanford-Dogs、Stanford-Cars 3个数据集的5-way 1-shot任务分别提升3.51%、2.85%、3.07%，5-way 5-shot任务分别提升1.71%、2.40%、1.36%；在ResNet-12网络中，该模型在上述3个数据集的5-way 1-shot任务分别提升2.37%、

表2　使用Conv-4网络在3个细粒度数据集上的分类准确率

模型	CUB-200-2011		Stanford-Dogs		Stanford-Cars	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
文献[7]	79.08%±0.21%	91.89%±0.12%	65.27%±0.41%	81.71%±0.24%	76.94%±0.10%	91.62%±0.20%
文献[8]	57.45%±0.89%	84.41%±0.58%	39.08%±0.76%	69.81%±0.69%	34.12%±0.68%	87.47%±0.47%
文献[10]	62.84%±0.95%	85.39%±0.56%	43.42%±0.86%	71.90%±0.68%	40.89%±0.77%	86.88%±0.50%
文献[19]	64.82%±0.23%	85.74%±0.14%	46.66%±0.21%	70.77%±0.16%	50.88%±0.23%	74.89%±0.18%
文献[22]	64.08%±0.50%	80.55%±0.11%	46.73%±0.49%	65.74%±0.63%	61.63%±0.27%	72.95%±0.38%
文献[28]	75.27%±0.61%	88.48%±0.37%	64.74%±0.69%	79.23%±0.46%	77.31%±0.58%	89.47%±0.32%
文献[34]	72.61%±0.21%	86.23%±0.14%	57.86%±0.21%	73.59%±0.16%	66.35%±0.21%	82.25%±0.14%
文献[35]	63.94%±0.92%	77.87%±0.64%	47.35%±0.88%	66.20%±0.74%	46.04%±0.91%	68.52%±0.78%
文献[36]	74.43%±0.95%	83.11%±0.67%	55.86%±0.97%	68.06%±0.72%	66.01%±0.94%	73.74%±0.70%
文献[37]	65.35%±0.65%	78.47%±0.41%	45.46%±0.36%	59.65%±0.51%	61.07%±0.47%	88.73%±0.49%
文献[38]	76.55%±0.21%	90.33%±0.58%	62.68%±0.22%	79.59%±0.15%	71.16%±0.21%	89.21%±0.10%
本文	82.59%±0.51%	93.60%±0.21%	68.12%±0.61%	84.11%±0.32%	80.01%±0.32%	92.98%±0.43%

表3　使用ResNet-12网络在3个细粒度数据集上的分类准确率

模型	CUB-200-2011		Stanford-Dogs		Stanford-Cars	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
文献[7]	86.54%±0.18%	94.73%±0.25%	78.98%±0.21%	89.70%±0.12%	91.3%±0.15%	97.79%±0.05%
文献[19]	81.02%±0.20%	91.93%±0.11%	73.81%±0.21%	87.39%±0.12%	85.46%±0.19%	95.08%±0.08%
文献[22]	75.59%±0.30%	88.23%±0.18%	70.38%±0.30%	85.24%±0.18%	80.62%±0.26%	92.63%±0.13%
文献[28]	80.97%±0.57%	93.17%±0.32%	72.41%±0.64%	85.11%±0.37%	86.81%±0.47%	95.36%±0.22%
文献[34]	80.39%±0.20%	91.01%±0.11%	73.22%±0.22%	85.90%±0.13%	85.03%±0.19%	92.63%±0.11%
本文	88.91%±0.11%	96.25%±0.32%	81.33%±0.25%	91.86%±0.15%	92.65%±0.35%	98.52%±0.22%

2.35%、1.35%，5-way 5-shot 任务分别提升 1.52%、2.16%、0.73%。整体来看，模型在 Conv-4 网络的性能提升普遍高于 ResNet-12 网络，这是因为 ResNet-12 特征提取能力更强，BiFRN 性能基线高，优化空间小，而 Conv-4 基线低，模型改进机制更易发挥作用。

本文模型的有效性核心源于所引入的通道与空间注意力融合机制及小波引导的 Mamba 动态状态空间建模机制，二者协同实现了跨尺度特征的精准整合与频率感知的局部细节建模。该机制通过 CAF 构建跨尺度的通道与空间特征交互通道，自适应强化判别性通道与关键空间区域，实现了细粒度识别任务中关键微小类别差异（如 CUB-200-2011 数据集中鸟类的羽毛纹理差异、Stanford-Dogs 数据集中犬类的毛色细节差异）的精准捕捉；同时 WDMFE 结合小波分解的频域感知与 Mamba 的线性复杂度全局依赖建模，弥补了传统模型在多尺度频域特征捕捉与跨尺度关联上的不足。尤其在 Conv-4 这类浅层网络中，由于基线模型对细微特征的挖掘能力较弱，本文模型的频域-空间域协同建模以及跨尺度注意力融合优势更易凸显，进而带来更显著的性能提升；即便在 ResNet-12 这类深层网络中，面对高基线场景，该双重机制仍能通过优化多尺度特征关联效率与频域细节表征，实现稳定的性能增益，证明了其在不同网络架构下的适配性与有效性。

3.4 消融实验

为验证各模块对模型性能的贡献，本文在 CUB-200-2011 数据集上进行了消融实验，使用的骨干网络是 Conv-4，结果如表 4 所示。由表 4 可以看出，基础模型 BiFRN 在 5-way 1-shot 和 5-way 5-shot 任务中分别达到 79.08% 与 91.89% 的准确率。引入 DASSM 后，模型性能显著提升至 81.57% 和 93.27%，表明 DASSM 能够有效增强特征的动态建模能力。在此模型基础上进一步加入小波分解机制（Wavelet），性能提升至 81.84% 和 93.44%，说明频域建模有助于捕捉细粒局部纹理特征。最后，引入 CAF，通过通道与空间注意力实现跨尺度特征增强，使准确率进一步提升至 82.59% 和 93.60%。综合来看，各模块均对性能提升具有正向贡献，其中 DASSM 对性能增益最为显著，而小波引导与多尺度融合进一步提升了模型对细粒差异的表征能力。

4 结束语

本文提出了一种面向小样本细粒度图像分类的 Mamba-小波多尺度特征建模方法，该方法构建了 Mamba 状态空间建模的多尺度特征关系网络。该网络包含两大核心创新模块——WDMFE 与 CAF，通过频域与空间域的协同建模，实现了对细粒特征的高效捕获。其中，WDA-Mamba 嵌入 WDMFE 中，结合离散小波分解与动态状态空间混合机制，从频域感知的角度增强了模型对局部细节的建模能力；CAF 模块通过通道与空间注意力的联合作用，自适应整合多尺度特征，从而提升了跨尺度的判别能力与可解释性。在 CUB-200-2011 与 Stanford-Dogs 等细粒度数据集上的实验结果表明，MSFRNet 在不同小样本设置下均取得稳定的性能提升，具备较好的特征表达能力与泛化性能。消融实验进一步验证了各模块的协同贡献。

然而，本文方法也存在一定局限性。首先，WDMFE 中引入的小波分解与动态状态空间建模增加了网络复杂度，在实时性要求较高的应用中可能存在推理延迟问题。其次，当前方法主要针对自然图像中的细粒度分类任务，其在跨域场景（如医疗影像、遥感图像）中的泛化能力尚未充分验证。此外，模型对高频噪声的鲁棒性仍有提升空间，在背景复杂或标注质量较低的数据上可能出现性能波动。

未来的研究将探索 MSFRNet 在跨域与跨模态

表 4 模块消融实验分类准确率

模块	Wavelet	DASSM	多尺度	特征融合	5-way 1-shot	5-way 5-shot
BiFRN	×	×	×	—	79.08%	91.89%
+DASSM	×	√	×	—	81.57%	93.27%
+DASSM+Fusion	×	√	√	注意力	82.11%	93.38%
+DASSM+Wavelet	√	√	×	—	81.84%	93.44%
本文模型	√	√	√	注意力	82.59%	93.60%

小样本任务中的迁移能力，同时通过轻量化小波-状态空间结构的自适应优化实现更高的计算效率，并进一步研究其在噪声环境下的稳健性提升策略。

参考文献：

[1] Zhang Y B, Tang H, Jia K. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 241-256.

[2] Lu Z Y, Xie H T, Liu C B, et al. Bridging the gap between vision transformers and convolutional neural networks on small datasets[C]//Proceedings of the Advances in Neural Information Processing Systems. New York: ACM Press, 2022: 14663-14677.

[3] Xie T, Wang L, Wang K, et al. FARP-net: local-global feature aggregation and relation-aware proposals for 3D object detection[J]. IEEE Transactions on Multimedia, 2024, 26: 1027-1040.

[4] 黎拓新, 项凤涛, 陈君海, 等. 基于跨空间多尺度信息聚合和推理一致性的域泛化方法[J]. 智能科学与技术学报, 2025, 7(2): 200-210.
Li T X, Xiang F T, Chen J H, et al. Domain generalization method based on cross-space multi-scale information aggregation and inference consistency[J]. Chinese Journal of Intelligent Science and Technology, 2025, 7(2): 200-210.

[5] 崔家豪, 江涛, 徐梦瑶. 基于同质多层图卷积的多尺度网络对齐模型[J]. 智能科学与技术学报, 2024, 6(4): 522-532.
Cui J H, Jiang T, Xu M Y. Multiscale network alignment model based on convolution of homogeneous multilayer graphs[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(4): 522-532.

[6] Han M Y, Wang R G, Yang J, et al. Multi-scale feature network for few-shot learning[J]. Multimedia Tools and Applications, 2020, 79(17/18): 11617-11637.

[7] Wu J J, Chang D L, Sain A, et al. Bi-directional ensemble feature reconstruction network for few-shot fine-grained classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6082-6096.

[8] Li W B, Wang L, Xu J L, et al. Revisiting local descriptor based image-to-class measure for few-shot learning[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 7253-7260.

[9] Wertheimer D, Tang L M, Hariharan B. Few-shot classification with feature map reconstruction networks[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 8008-8017.

[10] Li X X, Wu J J, Sun Z, et al. BSNet: bi-similarity network for few-shot fine-grained image classification[J]. IEEE Transactions on Image Processing, 2021, 30: 1318-1331.

[11] 张杨, 程智宇, 陈允降, 等. 注意力机制增强的输煤传送带异物检测[J]. 智能科学与技术学报, 2025, 7(2): 268-276.
Zhang Y, Cheng Z Y, Chen Y J, et al. Foreign object detection on coal conveyor belt enhanced by attention mechanism[J]. Chinese Journal of Intelligent Science and Technology, 2025, 7(2): 268-276.

[12] 姚云, 胡振虓, 邓涛, 等. 基于自适应池化注意力Transformer的唇语识别方法[J]. 智能科学与技术学报, 2025, 7(2): 211-220.
Yao Y, Hu Z X, Deng T, et al. A lip reading method based on adaptive pooling attention Transformer[J]. Chinese Journal of Intelligent Science and Technology, 2025, 7(2): 211-220.

[13] He J, Chen J N, Liu S, et al. TransFG: a transformer architecture for fine-grained recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 852-860.

[14] Lin X, Yan Z Q, Deng X B, et al. ConvFormer: plug-and-play CNN-style transformers forImproving medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Cham: Springer, 2023: 642-651.

[15] Dai Z, Liu H, Le Q V, et al. Coatnet: marrying convolution and attention for all data sizes[J]. Advances in neural information processing systems, 2021, 34: 3965-3977.

[16] Zhu L H, Liao B C, Zhang Q, et al. Vision mamba: efficient visual representation learning with bidirectional state space model[PP]. V3. (2024-11-14) [2025-12-05]. arXiv: arXiv.2401.09417.

[17] Wang Z Y, Zheng J Q, Zhang Y C, et al. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation[PP]. V2. (2024-03-30) [2025-12-05]. arXiv: arXiv.2402.05079.

[18] Lu S, Zhang M, Huo Y, et al. SSUM: spatial-spectral unified mamba for hyperspectral image classification[J]. Remote Sensing, 2024, 16(24): 4653.

[19] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing System. Massachusetts: MIT Press, 2017: 4080-4090.

[20] Ye H J, Hu H X, Zhan D C, et al. Few-shot learning via embedding adaptation with set-to-set functions[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8805-8814.

[21] Ding Y M, Tian X, Yin L R, et al. Multi-scale relation network for few-shot learning based on meta-learning[C]//International Conference on Computer Vision Systems. Berlin: Springer, 2019: 343-352.

[22] Zhang C, Cai Y J, Lin G S, et al. DeepEMD: differentiable earth mover's distance for few-shot learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5632-5648.

[23] Chen Y B, Liu Z, Xu H J, et al. Meta-baseline: exploring simple meta-learning for few-shot learning[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 9042-9051.

[24] Jiao J B, Liu Y, Liu Y F, et al. VMamba: visual state space model[C]// Proceedings of the Advances in Neural Information Processing Systems. New York: ACM Press, 2024: 103031-103063.

[25] Kong L S, Dong J X, Tang J H, et al. Efficient visual state space model for image deblurring[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2025: 12710-12719.

[26] Baaziz N, Abahmane O, Missaoui R. Texture feature extraction in the spatial-frequency domain for content-based image retrieval[PP]. V1. (2010-12-23) [2025-12-05]. arXiv: arXiv.1012.5208.

[27] Ahmad M, Usama M, Mazzara M, et al. WaveMamba: spatial-spectral wavelet mamba for hyperspectral image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 5500505.

[28] Zhu H G, Gao Z, Wang J Y, et al. Few-shot fine-grained image classification via multi-frequency neighborhood and double-cross modulation[J]. IEEE Transactions on Multimedia, 2024, 26: 10264-10278.

[29] Guo Y F, Li B, Zhang W Y, et al. Spatial-frequency feature fusion network for small dataset fine-grained image classification[J]. Scientific

Reports, 2025, 15: 9332.
[30] Sun Z H, Lin Y Y, Li Y, et al. Crossed wavelet convolution network for few-shot defect detection of industrial chips[J]. Sensors, 2025, 25(14): 4377.
[31] Welinder P, Branson S, Mita T, et al. Caltech-UCSD Birds 200-2011 dataset[R]. 2010.
[32] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]//Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC). Piscataway: IEEE Press, 2011: 1.
[33] Krause J, Stark M, Jia D, et al. 3D object representations for fine-grained categorization[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2013: 554-561.
[34] Doersch C, Gupta A, Zisserman A. Crosstransformers: spatially-aware few-shot transfer[J]. Advances in Neural Information Processing Systems, 2020, 33: 21981-21993.
[35] Sung F, Yang Y X, Zhang L, et al. Learning to compare: relation network for few-shot learning[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1199-1208.
[36] Wu Z Y, Li Y W, Guo L H, et al. PARN: position-aware relation networks for few-shot learning[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6658-6666.
[37] Hao F S, He F X, Cheng J, et al. Collect and select: semantic alignment metric learning for few-shot learning[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 8459-8468.
[38] Lee S, Moon W, Heo J P. Task discrepancy maximization for fine-grained few-shot classification[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 5321-5330.

[作者简介]

仝傲（2002− ），女，西安工程大学电子信息学院硕士生，主要研究方向为小样本细粒度图像分类、深度学习。

任劼（1984− ），女，西安工程大学电子信息学院副教授，主要研究方向为小样本细粒度图像分类、兴趣点检测、高光谱图像处理、深度学习。

孟宗阳（2005− ），男，西安工程大学电子信息学院本科生，主要研究方向为小样本细粒度图像分类。

鲁磊（1988− ），男，博士，西安交通大学信息与通信工程学院讲师，主要研究方向为计算机视觉、机器学习、图像处理。