[文章编号] 1007－7405(2023)06－0568－09 DOI:10.19715/j.jmuzr.2023.06.11

FMSWFormer：基于频率分离和自适应多尺度窗口的视觉Transformer

蔡岱立，谢维波

（华侨大学计算机科学与技术学院，福建 厦门 361021）

[摘要] 由于Vision Transformer具有二次方的patch复杂度和较差的局部归纳偏置，导致需要大量的数据和更专业的数据增强策略及更多的训练技巧来超越高效卷积网络。为了解决这些问题，从多尺度特征提取和图像频率的角度进行研究，提出具有轻量注意力机制的FMSWFormer。FMSWFormer采用卷积-自注意力机制混合模块构建起不同频率间的通信，通过窗口划分实现局部注意力机制，以此限制过高的计算成本；参考自适应尺度感知卷积的做法，并创新性地将多尺度算子引入到自注意力计算中，从而实现了多头自注意力机制的自适应尺度感知能力。在各种基准识别任务数据集上进行广泛的实验，结果表明了FMSWFormer的有效性，在多个数据集中都取得了优越的性能，且不增加时间成本。其中在CIFAR100数据集上，FMSWFormer比SepViT的性能高出4.2%，延迟降低了47.8%；在参数量比EfficientNetv2减少了22%的情况下，FMSWFormer的性能依然能高出3.94%。

[关键词] 窗口自注意力机制；多尺度特征提取；深度学习；卷积神经网络；图像高低频解耦

[中图分类号] TP 391.4

FMSWFormer: Visual Transformer with Frequency Separation and Adaptive Multi-Scale Window Attention

CAI Daili, XIE Weibo

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: Vision Transformer's quadratic patch complexity and weaker local inductive bias, require a large amount of data, more sophisticated data augmentation strategies, and additional training tricks to surpass efficient convolutional networks. To address these issues, this paper investigates multi-scale feature extraction and image frequency perspectives and introduces FMSWFormer. FMSWFormer employs convolution-self-attention hybrid modules to establish communication between different frequencies and implements local attention mechanisms through window partitioning to constrain excessive computational costs. One of the main contributions of the paper is the incorporation of multi-scale operators into self-attention computation. This technique is inspired by adaptive-scale perception convolution, and it serves to enhance the adaptive-scale perception capabilities of multi-head self-attention mechanisms. By doing so, FMSWFormer effectively combines the strong local inductive bias of CNNs with the dynamic long-range dependency modeling abilities of

[收稿日期] 2023－08－07

[基金项目] 国家自然科学基金项目（61271383）

[作者简介] 蔡岱立（1998—），男，硕士生，从事计算机视觉模型框架与算法的研究。通信作者：谢维波（1964—），男，博士，教授，从事现代信号处理与计算机应用技术方向研究。

transformers. Extensive experiments on various benchmark recognition task datasets demonstrate the effectiveness of FMSWFormer, which achieves superior performance on multiple datasets without increasing time costs. Notably, FMSWFormer outperforms SepViT by 4.2% on the CIFAR100 dataset while reducing latency by 47.8%. Even with 22% fewer parameters than EfficientNetv2, FMSWFormer's performance still surpasses it by 3.94%.

Keywords: window self-attention; multi-scale feature extraction; deep learning; convolutional neural networks; image high and low frequency decoupling

0　引言

在计算机视觉领域，图像理解是视觉下游任务的研究热点。卷积神经网络 CNNs（convolutional neural networks）由于具有平移不变性和局部感知能力，在各种视觉任务中表现出色。近年来，ViTs（vision transformer）通过全局自注意力机制，证明了图像理解任务中长距离上下文依赖的重要性，这启发研究者去探索卷积网络和 Transformer 相结合的混合模型，以兼具卷积的高效计算和 Transformer 的全局建模能力。

Transformer 的自注意力机制可以捕捉全局信息，但计算量随 patch 数量的平方增长，限制了其在高分辨率图像及硬件条件有限的场景中的应用。为降低计算量，Wang 等[1]采用特征金字塔结构减少 token 数量。Liu 等[2]将 CNN 的局部感受野概念引入 Transformer，采用滑动窗口限制自注意力感受野，使计算量大幅降低并同时保留了 CNN 的局部建模能力。Dong 等[3]在该方向上进一步提出使用交叉窗口实现更高效的局部建模。这些方法主要用于降低自注意力模块的计算开销。

另一些方法将卷积引入 Transformer 网络。Wu 等[4]使用卷积提取局部特征，再用自注意力建立全局依赖。Touvron 等[5]将卷积作为先验知识引入 Transformer，以更好地对局部信息建模。这些方法通过卷积和自注意力的有效融合来利用两者的优势。

但是，现有方法忽略了不同尺度特征的多样性关系，导致表达能力和泛化能力不足。本文提出一种新颖的混合模型，在自注意力模块内部引入多尺度特征融合机制，加强不同视图特征的多样性，以提高模型的表达能力和泛化能力。具体而言，本研究提出了一种频率分离的重建算法和一种拥有自适应尺度感知的自注意力机制，即 MSSA（multi-scale self-attention）。

一些工作[8-9]发现，图像的低频特征主要反映了全局结构和颜色信息，而高频特征主要反映了物体的细微细节，如尖锐的边缘和识别度高的纹理特征等。因此，基于频率信息处理的神经网络引起了广泛关注。具体来说，Fritsche 等[10]提出了一种图像频率特征分离的训练方式，引导模型实现高频分量的域迁移，最终应用于图像超分辨率任务并取得优异的效果。Zhou 等[11]利用图像低频中的颜色信息提出了一种颜色引导的域映射网络，解决了域变换过程中的颜色偏移问题。Xiao 等[12]提出一种采用小波变换提取出的频率矩阵，它同时包含高频信息和低频信息，让低频信息协助高频建模过程中的上下文理解，让高频信息补充低频建模过程中的结构细节和线条边缘。

根据 Zhuang 等[6]的研究发现，自注意力机制专注于长距离 patches 之间的低频信息建模，而忽略了边缘、条纹等高频信息。相反地，卷积对高频信息更加敏感，能够充当一个高频信息提取器。但频率的分解过程往往也伴随着信息的丢失。为了解决这个问题，本研究提出了一种交叉形式的残差频率特征图，使用高频原信息和低频特征图的分步特征，和低频原信息对高频的距离计算高低频之间的分步特征，以重构高低频特征图。

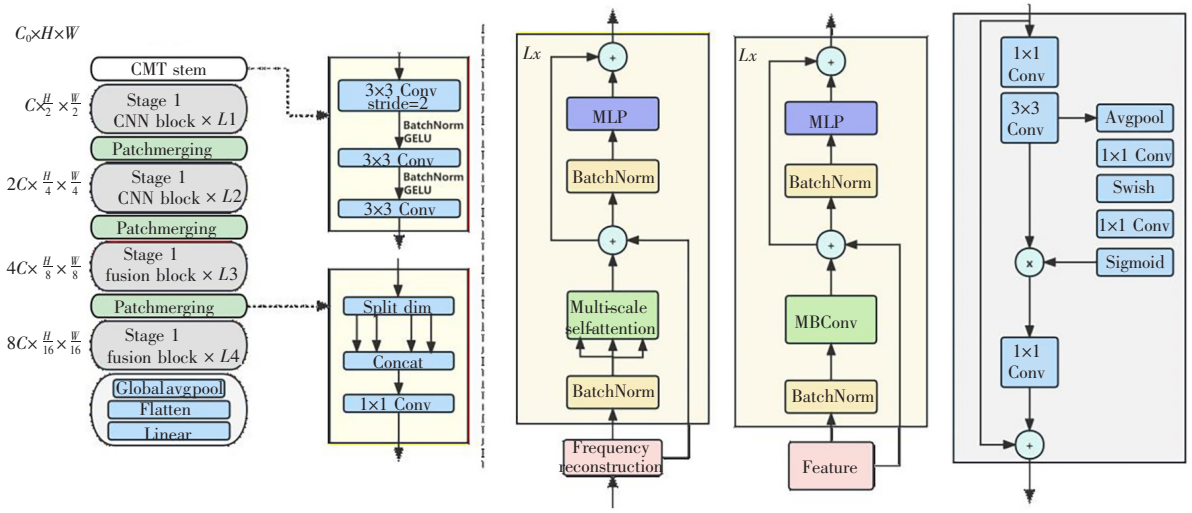
1　FMSWFormer

1.1　主要框架

本研究提出了一种新颖的视觉主干模型 FMSWFormer，它结合了卷积和自注意力机制，并考虑了

图像的多尺度特征和频率特征。FMSWFormer 遵循了 MetaFormer[7] 这一通用架构，即 LayerNorm-token mixer-LayerNorm-channel mixer，在不指定 token mixer 模块的情况下，使用高效卷积模块和 MSSA 模块进行特征提取。

如图 1 所示，FMSWFormer 由多个阶段（Stem、CNN Block、PatchMerging、Fusion Block 和输出层）组成，每个阶段包含若干个混合模块。其中：Stem 由 3 个核大小为 3 的卷积层组成；Fusion Block 和 CNN Block 具有 MetaFormer 结构；“ + ”表示残差连接；“L”表示模块的数量；“ × ”表示按元素相乘的操作；C、H、W 分别表示特征图的通道大小、空间维度的高度和宽度。在 Fusion Block 中，特征图先经过频率重构模块，这个模块将图像的高低频信息先进行解耦，再对解耦后的高低频特征图进行重建，分别输出高频率特征图和低频率特征图；接着，将其输入到混合模块中，分别用高效卷积模块和 MSSA 模块对高频率特征图和低频率特征图进行处理；然后，通过通道分组和交叉残差连接将两部分特征图重新组合起来。这样做可以让不同频率的特征图输入给不同的模块，并保留原始频率信息。在每个阶段结束后，使用空间下采样模块降低特征图的尺寸，并使用条件位置编码增强位置信息。最后，采用一个全局池化层和线性层进行特征分类。



a)整体网络框架 The overall framework　b)混合模块 Fusion block　c)CNN 模块 CNN block　d)MBConv

图 1　FMSWFormer 架构

Fig.1　FMSWFormer architecture

1.2　基于频率的混合模块

特征图高低频率信息的解耦分析有助于提高模型的泛化能力。Zhuang 等[6] 提出一种将 multi-head self-attention 中的 head 分解成两组来解耦自注意力层中的高低频率模式，其中一组的每个局部窗口内独立使用 self-attention 进行高频建模，另一组通过平均池化层提取每个窗口内的低频信息后输入进 self-attention 进行全局低频建模。

基于这个方法，本研究进一步提出了频率混合模块，如图 2 所示。混合模块先对特征图进行频率分解，将其分为高低频两部分。接着，分别对两部分特征图进行处理：高频特征图经过卷积层提取高频纹理和条纹等细节特征；低频特征图经过 MSSA 模块提取其局部和全局关系特征。然后，将两部分特征图通过通道聚合的形式重新组合起来，得到一个新的特征图，它同时包含了高频细节和低频内容的特征。最后，通过一个线性层对新的特征图进行特征映射，得到混合模块的输出，图 2 中“ + ”表示元素加。频率混合模块的公式表示为 $[X^{\text{high}}, X^{\text{low}}] = \text{FreqSep}(X)$，其中：high 表示高频特征图；low 表示低频特征图；FreqSep 是特征频率分解模块。高低频分量之后被送到相应的频率模块中以获

得高低频特征图，公式为 $X^{high'}=f_{high}(X^{high})$，$X^{low'}=f_{low}(X^{low})$，其中 $f_{high}()$ 和 $f_{low}()$ 分别表示高频特征提取模块和低频特征提取模块。

1.3　频率分解

本研究采用基于频率分解预处理的方法。如图 3 所示，将输入特征图的深度分为高低频两部分，并分别用不同的模块进行处理，即卷积模块处理图像的高频部分，而使用平均池化来提取特征图的低频信息。卷积模块具有对纹理特征敏感的特点，因此适合作为高频特征提取器；平均池化层具有对感受野范围内的空间信息取平均的能力，因此适合作为低频特征提取器。图 3 中“－”号表示残差连接。输入特征图 $X=[X^{high},X^{low}]$，由 $X^{high'}=\mathrm{Conv}_{3\times3}(X^{high})$，$X^{low'}=\mathrm{AvgPool}(X^{low})$，$\hat{X}^{low}=X^{low'}+X^{low}$，$\hat{X}^{high}=X^{high'}+X^{high}$（其中 $\mathrm{Conv}_{3\times3}()$ 表示 3x3 卷积层，AvgPool () 表示平均池化），得到频率重建的高低频特征图 $\hat{X}^{high}$ 和 $\hat{X}^{low}$。

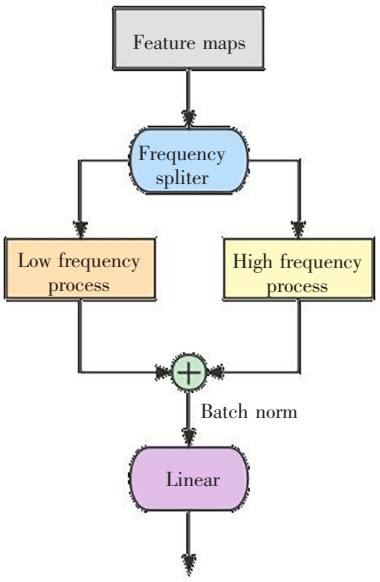


图 2　Fusion Layer 的整体框架

Fig.2　The overall framework of the Fusion Layer

1.4　多尺度自注意力机制（MSSA）

本研究的主要贡献是提出了一种新颖的多头自注意力机制 MSSA 模块，如图 4 所示，它具有自适应尺度感知的能力。MSSA 模块借鉴了 Swin Transformer[2] 和 SepViT[13] 的思想，它们在整体模块中采用了局部和全局交互运算而非串联进行。

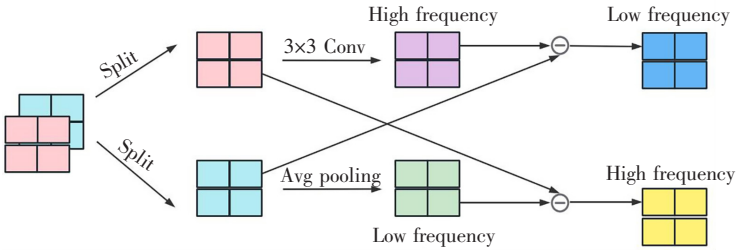


图 3　频率分解器的结构

Fig.3　The structure of the frequency decomposer

具体来说，MSSA 模块首先将特征图划分为多个窗口，并在每个窗口内添加一个额外的 window-token。接着，对每个窗口内的特征采用全局自注意力机制，并将 window-token 与其他特征分离。然后，对所有 window-token 都采用全局自注意力机制，并将其与原始窗口内特征重新组合。最后，将所有窗口内特征进行聚合，得到 MSSA 模块的输出。这样做可以有效地降低传统 MHSA 的计算和内存开销，并同时捕捉不同尺度下特征之间的关系。

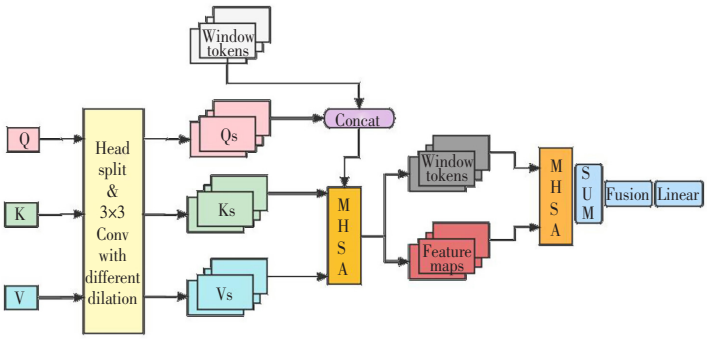


图 4　MSSA 的结构框架图

Fig.4　The structural framework of MSSA

为了让 MSSA 模块能够捕捉不同尺度下特征之间的关系，本研究提出了一种基于注意力头部的尺度划分策略。具体地，对每个注意力头部使用不同空洞率的卷积来生成不同尺度下的视图模式。这样做可以使得每个注意力头部关注不同范围内的特征，并且可以简单地插入到现有的 MHSA 模块中。

首先，设输入特征图生成的 Query 矩阵为 $\boldsymbol{Q}\in\mathbf{R}^{N\times H\times W\times C}$，其中 N 表示自注意力机制的头部数量，H 和 W 分别表示特征图的长与宽，C 表示特征图维度的大小。按 N 分为 3 组，得到小尺度组 $\boldsymbol{Q}_{small}$、中尺度组 $\boldsymbol{Q}_{medium}$ 和大尺度组 $\boldsymbol{Q}_{large}$。流程如下：$\boldsymbol{Q}=[\boldsymbol{Q}_1,\boldsymbol{Q}_2,\cdots,A_N]$，$\boldsymbol{Q}_{small}=\mathrm{Conv}_{3\times3,r=1}(\boldsymbol{Q}_{small})$，$\boldsymbol{Q}_{medium}=$

$\mathrm{Conv}_{3\times3,r=2}(\boldsymbol{Q}_{\mathrm{medium}}),\boldsymbol{Q}_{\mathrm{large}}=\mathrm{Conv}_{3\times3,r=3}(\boldsymbol{Q}_{\mathrm{large}})$，其中 r 表示空洞率。这样可分别提取出不同感受野的特征图。接着，将额外的 token 添加进每个窗口中，并执行窗口自注意力机制，以获得局部和全局交互的特征图。公式为 $X^s=W_1,W_2,\cdots,W_n,W_i\in\mathbf{R}^{m\times C}$，表示将特征图划分为 n 个窗口，即 $W_1,W_2,\cdots,W_n$，每个窗口包含 m 个元素。然后，进行基于自注意力机制的局部与全局的交互：$\widetilde{W}=[W_i,z_i],z_i\in\mathbf{R}^C,Y^s=\mathrm{MSA}(\widetilde{W}^s),Y^s=[Y^s_{\mathrm{local}};Y^s_{\mathrm{global}}],Y^{s'}_{\mathrm{global}}=\mathrm{MSA}(Y^s_{\mathrm{global}}),\widetilde{Y}^s=[Y^s_{\mathrm{global}};Y^{s'}_{\mathrm{global}}]$。其中：$z^s\in\mathbf{R}^{1\times1\times C}$为额外的 token，表示每个窗口的全局信息；MSA 表示乘性自注意力。这样就可以得到具有局部和全局交互的特征图。每个尺度都进行此类流程，获得不同尺度的全局和局部的交互特征图，而不仅局限于单个尺度中。

1.5　多尺度特征融合

本研究提出了一种多尺度特征混合结构，以有效地聚合不同尺度下的特征信息。如图 5 所示，该结构首先将每个注意力头部输出的特征图按照不同尺度进行划分，并将相同尺度下的特征图进行元素加，以增加特征图的信息量。

使用核大小为 1 * 1 的卷积进行线性投影，将不同尺度下的特征图进行融合。本研究采用将大尺度和小尺度融合进中尺度的特征图中。通过局部 CNN 模块，表征聚合了大小尺度的中尺度特征图，自适应提取其尺度之间的差异化结构信息。再利用残差连接重新注入到大尺度和小尺度特征图中，从而让大尺度特征图拥有中尺度和小尺度特征图所具有的高细粒度的高层语义。同样地，让小尺度特征图具有中尺度和大尺度特征图所具有的具体对象的特征信息。

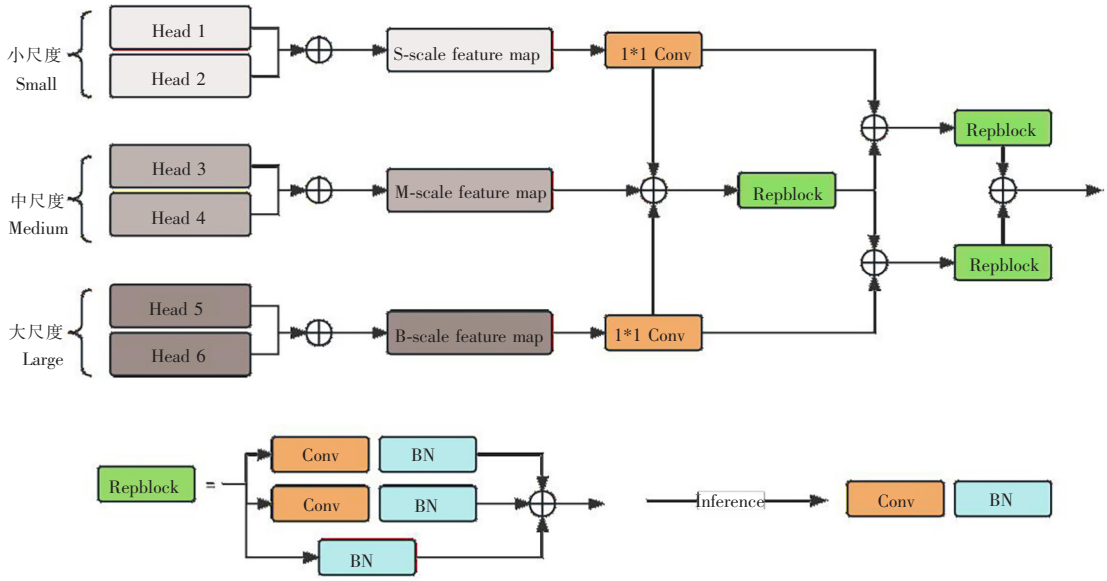


图 5　聚合 MSSA 模块中的多尺度自注意力头的算法流程

Fig.5　The algorithmic flow of the multi-scale self-attention heads in the MSSA module

这样做可以利用“特征复用”的思想来建立多尺度之间的特征关联，避免直接拼接操作可能导致尺度之间的通信关系不充分的问题。Raghu 等[14]证明在 Transformer 中，残差连接比 CNN 能提取更抽象的特征，而这对于以低频内容为导向的自注意力机制是非常有益的。本研究的多尺度特征混合结构也考虑了不同注意力头部之间的尺度特征模式，通过“特征复用”的方式，将不同尺度信息进行结合，并将学习更好尺度特征信息的任务交给了自注意力机制而不是随后的卷积层。本研究使用多个点卷积对富含多尺度信息的特征图进行特征映射，使得卷积算子能够同时学习到不同尺度之间的信息，而不是仅仅对单尺度信息进行处理。

设三个尺度组输出特征图为 $X^s\in\mathbf{R}^{H\times W\times C}$，$s=1$，2，3 分别表示小尺度、中尺度和大尺度特征，$N$

等于 C，为通道维度的大小。整体流程可表示为：

$$X^{\text{small}} = \sum X^{s}, s = 1, \cdots, N_{\text{small}};\quad X^{\text{middle}} = \sum X^{s}, s = N_{\text{small}} + 1, \cdots, N_{\text{middle}};$$

$$X^{\text{large}} = \sum X^{s}, s = N_{\text{middle}} + 1, \cdots, N;\quad X^{\text{small}'} = \text{Conv}_{1\times1}(X^{\text{small}});$$

$$X^{\text{large}'} = \text{Conv}_{1\times1}(X^{\text{large}});\quad X^{\text{fuse}} = X^{\text{small}'} + X^{\text{middle}} + X^{\text{large}'};$$

$$X^{\text{fuse}'} = \text{RepBlock}(X^{\text{fuse}});\quad X^{\text{residual}} = X^{\text{small}'} + X^{\text{large}'} + X^{\text{fuse}'};$$

$$X^{\text{small}''} = \text{RepBlock}(X^{\text{residual}});\quad X^{\text{large}''} = \text{RepBlock}(X^{\text{residual}});$$

$$X^{\text{output}} = X^{\text{small}''} + X^{\text{large}''}。$$

其中，small、middle 和 large 分别表示小尺度、中尺度和大尺度。另外，就如图 5 下部分所示，RepBlock 是采用重参数化模块聚合方法的 CNN 模块。具体来说，在训练人工神经网络时，使用普通的 CNN 模块进行表征，这里采用 3 条分支分别提取输入特征图的三种不同的特征信息。当训练结束时，即模型的推理阶段，可以把多分支转换成单分支的结构，以减少模型的计算开销。

2　实验结果与分析

2.1　环境设置

为了验证 FMSWFormer 在图像分类任务上的有效性和优越性，本研究在 CIFAR10、CIFAR100 和 Mini-ImageNet 三个公开数据集上进行了实验。这些数据集涵盖了不同类别数目、不同图像尺寸和不同难度等方面。还将 FMSWFormer 与当前流行的轻量视觉模型进行了对比，并参考了最近提出的 ViT-base[2] 和 CNN-base[15] 作为基准设置。具体来说，在一张 RTX 3090 的 GPU 上训练 300 个 epoch，总批次大小为 512，输入图像大小为 224 px × 224 px。使用 AdamW 作为优化器，并设置权重衰减系数为 0.01；使用余弦策略调整学习率，并将初始学习率设为 0.001；使用线性预热学习率的策略，并在前 5 个 epoch 进行预热学习；使用 DropPath 的技术，并将概率设为 0.3；在模型模块上，采用轻量化设计，如减少模型的深度和宽度，降低浮点计算量，以展示模型在极端轻量化情况下的表现能力。

2.2　CIFAR10/100 数据集测试

表 1 展示了 FMSWFormer 与其他主流模型在 CIFAR10 数据集上的性能和模型复杂度比较。由表 1 可见，ResNet 作为经典的卷积神经网络，参数量较小但精度仅达到 76.32%。轻量级卷积神经网络如 MobileNetv3 和 ShuffleNetv2 进一步降低了参数量，但精度也下降到 70.84% 左右。EfficientNet 通过神经架构搜索获得了很好的精度（84.46%），但参数量和计算量较大。基于 Transformer 的视觉模型如 LiT 和 PVT 也取得了不错的精度，并通过一定设计降低了参数量，但计算量仍较大。SepViT 和 Swin Transformer 通过引入一定形式的局部归纳偏置的操作来提升性能并降低计算量。CageViT 和 RelCol 等大模型虽然精度很高但参数和计算量都非常大。EMO 和 SwiftFormer 通过模块化和轻量级设计也取得了良好平衡。本文提出的 FMSWFormer 在维持较高精度（88.00%）的同时，其参数量（6.9 M）和计算量（38.3 M）都是较低的，显示了该模型在精度与效率之间取得了很好的平衡。这主要归功于本文提出的特征融合模块和滑动窗口自注意力机制的设计。

为了进一步验证 FMSWFormer 在更加复杂的数据集上的泛化能力，将数据集换成了 CIFAR100。从表 1 中可以看出：虽然 ResNet 比 FMSWFormer 更加轻量，但同时精度也下降了约 42.63%；与 EfficientNetV2 相比，FMSWFormer 的参数量减少了 20.45%，而精度提高了 3.94%；LiTv2 将 Transformer 模型进行多分支轻量化处理，但 FMSWFormer 与它相比，精度提高了 13.76%，参数量增加了 9%；PVTv2 作为另一个轻量级设计的 Transformer，FMSWFormer 比它参数量减少了 20%，精度提高了 8.3%。这证明了 FMSWFormer 的计算效率更高。CageViT 和 RelCoL 都是高参数量的模型，FMSWFormer 分别比它们的参数量减少了 50% 和 77%，计算量减少了 48% 和 74%，而精度还分别提高了 0.6% 和 3.67%。这说明 FMSWFormer 实现了更好的平衡。

表 1 各模型在 CIFAR10 和 CIFAR100 数据集上的精度、参数和计算量的大小

Tab. 1 The precision, parameters magnitude and computational complexity of various models on the CIFAR10 and CIFAR100 datasets

模型 Model	CIFAR10			CIFAR100		
	Top-1 准确率 Precision/%	参数量 Parameters/M	浮点计算量 Computation/M	Top-1 准确率 Precision/%	参数量 Parameters/M	浮点计算量 Computation/M
ResNet[15]	76.32	3.2	17.68	35.71	3.7	17.69
MobileNetv3[16]	72.36	10.2	23.83	50.70	10.7	24.00
ShuffleNetv2[17]	69.33	3.4	24.83	37.54	3.8	14.88
EfficientNetv2[18]	84.46	8.2	82.18	58.31	8.8	82.35
LiTv2[6]	77.15	4.3	21.16	48.49	6.4	18.18
PVTv2[1]	78.26	8.7	25.98	53.95	8.8	15.79
SepViT[13]	85.89	3.7	62.78	58.05	6.7	62.79
Swin[2]	80.21	7	34.80	50.15	7.1	37.84
CageViT[19]	87.34	14	74.31	61.65	14.0	74.31
RelCol[20]	84.29	30	149.28	58.58	30.0	149.28
EMO[21]	86.48	10.3	39.09	60.28	10.3	39.09
SwiftFormer[22]	85.46	9.4	40.63	60.37	9.4	40.63
FMSWFormer	88.00	6.9	38.30	62.25	7.0	38.30

2.3 Mini-ImageNet 数据集测试

为了进一步验证 FMSWFormer 模型的泛化能力，在更复杂的 Mini-ImageNet 数据集上进行了测试。表 2 展示了各模型在 Mini-ImageNet 数据集上的精度、参数量和计算量。

表 2 各模型在 Mini-ImageNet 数据集上的精度、参数和计算量的大小

Tab. 2 The precision, parameters magnitude and computational complexity of various models on the Mini-ImageNet datasets

模型 Model	Top-1 准确率 Precision/%	参数量 Parameter magntitude/M	浮点计算量 Computational complexity/M	模型 Model	Top-1 准确率 Precision/%	参数量 Parameter magntitude/M	浮点计算量 Computational complexity/M
ResNet[15]	53.30	5.7	216.61	Swin[2]	54.90	7.1	398.40
MobileNetv3[16]	65.30	10.7	278.10	CageViT[19]	65.21	17.6	285.63
ShuffleNetv2[17]	68.03	6.8	189.16	RelCol[20]	69.19	60.0	1104.96
EfficientNetv2[18]	62.10	8.3	1010.0	EMO[21]	63.48	15.3	603.48
LiTv2[6]	58.00	10.4	152.12	SwiftFormer[22]	67.29	12.1	447.30
PVTv2[1]	51.47	8.8	158.81	FMSWFormer	71.90	6.0	204.96
SepViT[13]	64.30	6.7	769.01				

从整体上看，FMSWFormer 在 Mini-ImageNet 这个更加复杂的数据集上仍然显示出出色的精度和效率。在精度方面，FMSWFormer 达到了 71.9% 的准确率，超过所有对比模型。具体来说，与 EfficientNetv2 相比，提升了 9.8%；与 SepViT 相比，提升了 7.6%；与 Swin Transformers 相比，提升了 17%；与 CageViT 相比，提升了 6.69%；与 RelCol 相比，提高了 2.71%；与 EMO 相比，提高了 8.42%；与 SwiftFormer 相比，提高了 4.61%。这说明 FMSWFormer 模型具有很强的泛化能力，能够在不同的数据集上保持状态的精度水平。

在参数量方面，FMSWFormer 只有 6 M 参数。与 EfficientNetv2 相比，减少了 15.9%；与 SepViT 相比，减少了 5.3%；与 CageViT 相比，减少了 66%；与 RelCol 相比，减少了 90%；与 EMO 相比，

减少了 61%；与 SwiftFormer 相比，减少了 50%。这说明 FMSWFormer 的参数利用效率很高，模型更加轻量和易部署。

在计算量方面，FMSWFormer 只有 204.96 M flops。与 EfficientNetv2 相比，减少了 79.7%；与 SepViT 相比，减少了 73.4%；与 CageViT 相比，减少了 28%；与 RelCol 相比，减少了 81%；与 EMO 相比，减少了 66%；与 SwiftFormer 相比，减少了 54%。这说明 FMSWFormer 模型非常高效，能够大幅减少计算资源的需求。

总之，FMSWFormer 在保持精度的同时，参数量和计算量都比其他模型显著减少，达到了最佳的复杂度和均衡的性能。这充分验证了 FMSWFormer 模型的泛化能力及对计算资源等实际约束的适应性。

2.4　消融实验

为了验证本研究提出的 FMSWFormer 中每个关键组件的重要性，包括频率分离、局部自注意力机制和多尺度方案，在 CIFAR-10 数据集上进行了一系列消融实验。表 3 展示了消融实验结果。

表 3　消融实验

Tab. 3　Ablation study

局部卷积阶段	频率重构模块	混合模块	多尺度聚合模块	精度 Accuracy
有 Yes	有 Yes	有 Yes	有 Yes	88.03%
有 Yes	有 Yes	有 Yes	否 No	87.62%
有 Yes	有 Yes	否 No	否 No	85.50%
有 Yes	否 No	否 No	否 No	81.88%
否 No	否 No	否 No	否 No	80.59%

从表 3 中可以看出，在去除所有高效组件设计的情况下，FMSWFormer 退化为普通的自注意力机制模型，达到了 80.59% 的测试精度。在模型前期加入卷积层，测试精度提升到了 81.88%。这是因为卷积层可以帮助模型更快地收敛，并提高局部特征的学习能力。在模型中引入频率分解模块，测试精度进一步提升到了 85.50%。这说明频率分解模块可以为后续的卷积层和自注意力层提供更适合的特征图频率，如卷积层对高频特征更敏感，而自注意力层对低频特征的全局建模更擅长。按照 BoTNet 所提出的模型阶段设计准则，即在前期阶段使用 CNN，在高层语义阶段使用 Transformer，测试精度再次提升，达到了 87.62%。这样的设计不仅使模型的容量和效率之间达到了最佳平衡，也提高了模型对于局部和全局特征的提取能力。最后，比较了多尺度特征之间的聚合方式，发现利用“特征复用”的思想来建立多尺度之间的特征关联效果更好，测试精度又提高了 0.38%。

3　结论

本文提出了一种基于 Transformer 的高效多尺度方案，称为 FMSWFormer。它包含三个关键组成部分。首先，将图片进行频域上的解耦，施行一种更加有效的残差连接方式进行信息修补；然后，添加额外的 global tokens 来学习窗口之间的相互关联，使得模块能够同时了解局部和全局信息；最后，基于 Transformer 的多尺度操作使得模型在不依赖卷积进行尺度学习的情况下，不仅能够了解特征的长距离依赖关系，也学习到了不同尺度之间的依赖关系。在分类识别任务的各个数据集实验表明，FMSWFormer 不仅速度更快，而且性能也是最好的，同时达到了更优的复杂度和均衡的性能。

[参 考 文 献]

[1] WANG W H, XIE E Z, LI X, et al. PVTv2: improved baselines with Pyramid Vision Transformer[J]. Computational Visual Media, 2022, 8: 415-424. DOI: 10.1007/s41095-022-0274-8.

[2] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2022: 9992-10002. DOI: 10.1109/iccv48922.2021.00986.

[3] DONG X Y, BAO J M, CHEN D D, et al. CSWin Transformer: a general vision Transformer backbone with cross-shaped windows[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA:

IEEE,2022:1181. DOI:10.1109/cvpr52688.2022.01181.

[4]WU H,XIAO B,CODELLA N,et al. CvT:introducing Convolutions to vision Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2022: 22-31. DOI: 10.1109/iccv48922.2021.00009.

[5]ASCOLI S D,TOUVRON H,LEAVITT M L,et al. ConViT:improving Vision Transformers with soft convolutional inductive biases[J]. Journal of Statistical Mechanics:Theory and Experiment,2022:114005. DOI:10.1088/1742-5468/ac9830.

[6]PAN Z Z,CAI J F,ZHUANG B H. Fast Vision Transformers with Hilo attention[J]. Advances in Neural Information Processing Systems,2022,35:14541-14554.

[7]YU W H,LUO M,ZHOU P,et al. MetaFormer is actually what you need for vision[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 10809-10819. DOI: 10.1109/cvpr52688.2022.01055.

[8]COOLEY J W,LEWIS P A W ,WELCH P D,The fast Fourier Transform and its applications[J]. IEEE Transactions on Education,1969,12(1):27-34. DOI:10.1109/TE.1969.4320436.

[9]DENG G,CAHILL L W. An adaptive Gaussian filter for noise reduction and edge detection[C]//1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference. San Francisco,CA,USA:IEEE,1993:1615-1619. DOI:10.1109/NSSMIC.1993.373563.

[10]FRITSCHE M,GU S H,TIMOFTE R. Frequency separation for real-world super-resolution[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul,Korea :IEEE,2020:3599-3608. DOI:10.1109/ICCVW.2019.00445.

[11]ZHOU Y B,DENG W,TONG T,et al. Guided frequency separation network for real-world super-resolution[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle,WA,USA:IEEE,2020:428-429. DOI:10.1109/CVPRW50498.2020.00222.

[12]XIAO M Q,ZHENG S X,LIU C,et al. Invertible image rescaling[C]//Computer Vision-ECCV 2020,16th European Conference. Glasgow:ACM,2020:126-144. DOI:10.1007/978-3-030-58452-8_8.

[13]LI W,WANG X,XIA X,et al. SepViT:separable Vision Transformer[EB/OL]. (2023-06-15)[2023-08-07]. https://doi.org/10.48550/arXiv.2203.15380.

[14]RAGHU M,UNTERTHINER T,KORNBLITH S,et al. Do Vision Transformers see like convolutional neural networks?[EB/OL]. (2022-03-03)[2023-08-07]. https://doi.org/10.48550/arXiv.2108.08810.

[15]HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas,NV,USA:IEEE,2016:770-778.

[16]HOWARD A,SANDLER M,CHEN B,et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul,Korea:IEEE,2019:1314-1324. DOI:10.1109/iccv.2019.00140.

[17]MA N N,ZHANG X Y,ZHENG H T,et al. ShuffleNet V2:practical guidelines for efficient CNN architecture design[C]//Computer Vision-ECCV 2018,Lecture Notes in Computer Science,2018:122-138. DOI:10.1007/978-3-030-01264-9_8.

[18]TAN MC,LE Q C. Efficientnetv2:smaller models and faster training[C]//International Conference on Machine Learning. NY:PMLR,2021:10096-10106.

[19]ZHENG H,WANG J,ZHEN X,et al. CageViT:convolutional activation guided efficient vision Transformer[EB/OL]. (2023-05-17)[2023-08-07]. https://doi.org/10.48550/arXiv.2305:09924.

[20]CAI Y X,ZHOU Y Z,HAN Q,et al. Reversible column networks[EB/OL]. (2020-12-22)[2023-08-07]. https://doi.org/10.48550/arXiv.2012.11696.

[21]ZHANG J N,LI X T,LI L,et al. Rethinking mobile block for efficient attention-based models[C]//2023 IEEE/CVF International Conference on Computer Vision(ICCV). Paris:IEEE,2023:1389-1400.

[22]SHAKER A,MAAZ M,RASHEED,et al. SwiftFormer:efficient additive attention for Transformer-based real-time Mobile Vision applications[EB/OL]. (2023-07-25)[2023-08-07]. https://doi.org/10.48550/arXiv.2303.15446.

(责任编辑 朱雪莲 英文审校 黄振坤)