Computer Science and Application 计算机科学与应用, 2025, 15(7), 17-26
Published Online July 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.157176

Hans汉斯

基于ViT-X在小型数据集上的图像分类

钟士辉

西南民族大学电气工程学院，四川 成都

收稿日期：2025年6月8日；录用日期：2025年7月7日；发布日期：2025年7月14日

摘　要

近年来，在图像分类等计算机视觉任务中，Vision Transformer (ViT)展现出了卓越的进展，但ViT网络在建模图像中的局部依赖关系方面常显不足，尤其是在小规模数据集上训练时，可能导致归纳偏置不足的问题。针对该问题，本文提出了一种改进的ViT模型。该模型通过引入功能更强的交叉协方差注意力机制(XCA)，增强对多尺度上下文全局依赖关系的建模能力，同时在保持性能优势的情况下减少参数数量。在此基础上，本文还提出一种新颖的模块(Septh-Wise Convolution，简称SWConv)，进一步增强局部特征提取能力。实验结果表明，本文提出的ViT-X模型在CIFAR10等经典数据集中取得了优异的性能，该模型识别准确率达到95.6%，较原始ViT模型提升了1.8%，显著提高了模型的识别性能。

关键词

图像分类，Vision Transformer网络，归纳偏置，局部特征

Image Classification Based on ViT-X on Small-Scale Datasets

Shihui Zhong

College of Electrical Engineering, Southwest Minzu University, Chengdu Sichuan

Received: Jun. 8th, 2025; accepted: Jul. 7th, 2025; published: Jul. 14th, 2025

Abstract

In recent years, Vision Transformers (ViT) have demonstrated remarkable progress in computer vision tasks such as image classification. However, ViT networks often struggle to model local dependencies within images, especially when trained on small-scale datasets, which can lead to insufficient inductive bias. To address this issue, this paper proposes an improved ViT model. The proposed model introduces a more powerful cross-covariance attention mechanism (XCA) to enhance the modeling of multi-scale contextual global dependencies while reducing the number of parameters without compromising performance. Furthermore, a novel module (Septh-Wise Convolution, SWConv) is proposed to further

文章引用: 钟士辉. 基于 ViT-X 在小型数据集上的图像分类[J]. 计算机科学与应用, 2025, 15(7): 17-26.
DOI: 10.12677/csa.2025.157176

strengthen local feature extraction capabilities. Experimental results show that the proposed ViT-X model achieves outstanding performance on benchmark datasets such as CIFAR10, reaching an accuracy of 95.6%, which is 1.8% higher than the original ViT model, significantly improving the recognition performance of the model.

Keywords

Image Classification, Vision Transformer Network, Inductive Bias, Local Features

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

Transformer 模型在多个方向表现出卓越的性能，尤其是在自然语言处理(NLP)领域中，原因是通过注意力机制，Transformer 能够捕捉长距离依赖关系。Vision Transformer (ViT) [1]将这个特点成功引入视觉领域，先将图像划分为若干个固定尺寸的图像块(patches)，再将每个 patch 嵌入为一个 token，类似于 NLP 中的单词处理方式。因此，ViT 被广泛应用于图像分类[2]-[4]、目标检测[5] [6]和语义分割[7]等任务。然而，直接将 Transformer 应用于视觉任务仍面临若干挑战：1) 图像的固有局部结构和自注意力机制的二次复杂度往往限制 ViT 的性能，尤其是在处理小型数据集或资源受限环境时，更为严重。原因在于其对自注意力机制的依赖，自注意力机制带来了较高的计算成本，其复杂度随着输入序列长度的增加呈二次增长。这就导致较长的训练时间和较大的数据需求。尽管线性注意力机制[8]被探索为一种计算效率更高的替代方案，但由于其简化假设，通常相比于传统的 softmax 注意力性能更低。2) 图像的固有局部结构尤其是在处理小型数据集或资源受限环境时，更为严重。ViT 缺乏 CNN 中固有的归纳偏置[7]，对局部特征信息的提取能力和多尺度信息处理方面表现较弱。

ViT 中的自注意力机制能够在全局范围内捕捉图像特征信息之间的关联，有效处理图像数据中的长距离依赖关系，弥补了 CNN 在全局信息提取方面的不足。例如，Li 等[9]提出的(PMVT)模型，该模型将 MobileNet 与 ViT 相结合，使用倒置残差结构替换模型中的卷积块，并将卷积注意力模块(Convolutional Block Attention Module, CBAM)集成到模型中，还将固有局部结构和自注意力机制相结合，但是准确率较低。He 等[10]提出了另外一种改进 ViT 模型 ECA-ViT，在 ViT 网络中添加 ECA 模块，弥补了 ViT 缺少对图像局部特征信息提取的能力。Wu 等[11]提出了一种基于 ViT 多粒度特征提取的模型，通过学习不同尺度的图像特征信息，使模型能够更精细地识别相似图片之间的细微差异。Sharma 等[12]使用迁移学习的 ViT 模型，强化了全局特征提取能力。综合分析，已有的模型对图像局部特征的提取能力和准确率仍有提升空间。

针对以上问题，本文提出了一种改进的 ViT 图像分类算法，在 CIFAR10 等经典的数据集上进行训练。将 CNN 的归纳偏置(局部性和尺度不变性)集合成 SWConv 模块并引入到 ViT 模型中，并引入交叉协方差注意力机制 XCA (Cross-Covariance Attention)，提高模型在图像分类任务中的准确性，并减小模型参数量。

2. 方法

2.1. 改进的 Vision Transformer 算法

本文在 ViT 的基础上进行了两方面的结构改进。首先，采用 XCA 替换 Transformer [13]中的多头注意力机制，引入 CNN 的固有归纳偏置，即局部性和尺度不变性[14]，集合成 SWConv 模块。改进前后的

ViT 结构如图 1 和图 2 所示。视觉任务中的 ViT 处理流程[15]：将图像分割为若干固定长度小块(patches)，再将这些小块转化为输入 token 即 x。为了保留 patches 的位置信息，在每个 token 中引入可学习的位置嵌入，具体操作如公式(1)所示，其中 x_c 表示类别 token，x_p 表示位置嵌入。

$$x^n = \left(x_1, x_2, \cdots, x_l; x_c\right) + x_p \tag{1}$$

$$x^{n+1} = x'^n + \mathrm{FF}\left(\mathrm{LN}\left(x'^n\right)\right) \tag{2}$$

$$x'^n = x^n + \mathrm{XCA}\left(\mathrm{LN}\left(x'^n\right)\right) \tag{3}$$

式(2)和(3)分别描述了交叉协方差注意力(XCA)层和前馈(FF)层，均结合了残差连接和预层归一化(Pre-LayerNorm)。若将 XCA 与 FFN 层相结合，可以构建优化后的 Transformer 块，如图 3 所示。因此，Transformer 模型就是通过堆叠多个这样的块构建而成，从而实现高效的特征提取和信息整合。此外，类别 tokens 被用于图像分类和生成最终输出，从而提高模型的分类性能和鲁棒性。

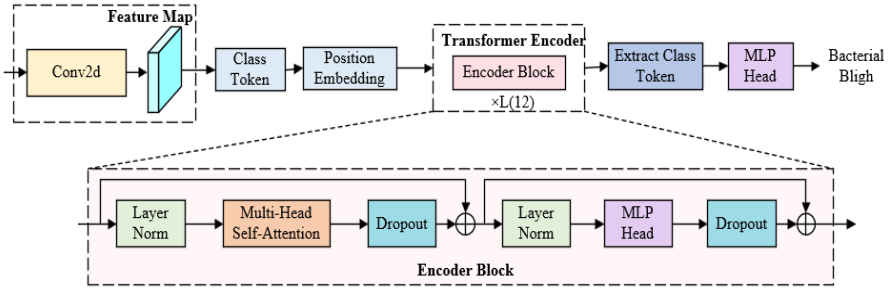

Figure 1. Original Vision Transformer architecture
图 1. 原始 Vision Transformer 结构

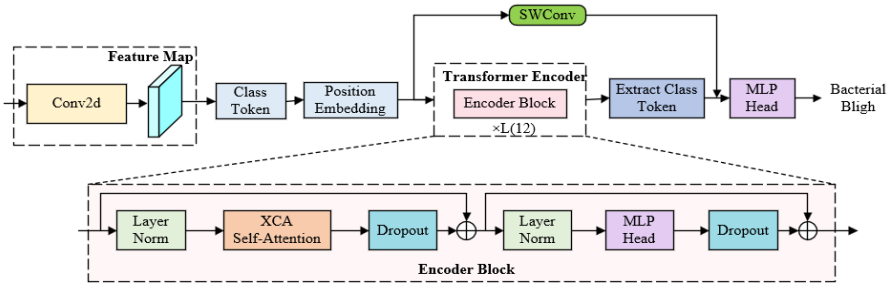

Figure 2. Improved Vision Transformer architecture used in this paper
图 2. 本文使用的改进 Vision Transformer 结构

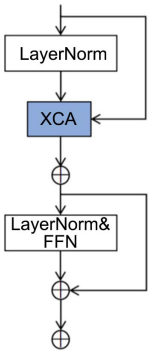

Figure 3. Modified Transformer
图 3. Transformer 改进

2.2. XCA 模块

在 ViT 中，多头自注意力机制(Multi-Head Self-Attention, MHSA)是建模全局依赖关系的关键组件。MHSA 通过在空间维度上计算不同位置之间的相关性，赋予模型捕获长距离依赖的能力。然而，传统 MHSA 也存在一定的局限性。

首先，MHSA 主要专注于空间位置间的信息交互，对于特征通道(Feature Channels)之间的相关性建模较为薄弱。在处理复杂图像时，通道间的特征协同关系同样重要，仅关注空间关系可能导致特征表达能力受限。其次，MHSA 在计算注意力矩阵时需要进行大规模的矩阵乘法，尤其在处理高分辨率特征图时，计算与存储开销显著增加，这会导致模型的参数量增加。

针对上述问题，本文引入了交叉协方差注意力机制(Cross-Covariance Attention, XCA)，代替 MHSA。XCA 通过在通道维度上计算特征的协方差矩阵，能有效捕捉不同通道间的交互关系。相比传统的空间自注意力机制，XCA 在建模特征通道相关性方面具有天然优势，能够进一步提升特征的判别能力。同时，XCA 在计算过程中摒弃了高维空间上的大规模矩阵运算，仅需处理通道数目相关的矩阵，因此在一定程度上降低了计算复杂度，提高了模型的推理效率。基于传统注意力机制对查询(Q)、键(K)和值(V)的定义，交叉协方差注意力机制的公式如式(4)和式(5)所示：

$$A_{\mathrm{XCA}}\left(K,Q\right)=\mathrm{Softmax}\left(K^{\mathrm{T}}Q/\tau\right) \tag{4}$$

$$\mathrm{XCA\text{-}Attention}\left(Q,K,V\right)=V\cdot A_{\mathrm{XCA}}\left(K,Q\right) \tag{5}$$

每个输出 token 嵌入是其在 V 中对应的 d 维特征的凸组合。注意力权重 A 是基于交叉协方差矩阵计算的。XCA 可以被视为一种动态的 1 × 1 卷积，其中所有 token 都与相同的数据依赖权重矩阵相乘。

ViT 模型采用自注意力机制主要计算每对图像小块(patch token)之间的相似性，为其分配对应的权重，以示其重要性。然而，ViT 模型会忽略图像中固有的归纳偏置，尤其是相邻像素或小块之间的强空间相关性，可能会导致训练速度变慢。因此，需要更多训练轮次来学习 patches 之间的关系，就需要更大的数据集。相比之下，卷积神经网络(CNN)通过局部感受也自然地融入了这种归纳偏置，可捕获局部模式和空间层次结构，这对于图像任务至关重要。这种特性使 CNN 在较小的数据集上表现出色。为此，本文提出了一种 ViT 与 CNN 相结合的混合架构，既可以保证整体信息的融合，又能捕捉局部模式和空间层次信息，且能降低对数据集规模的依赖度。

2.3. SWConv 模块

为进一步增强局部特征建模能力，本文设计了 SWConv 模块，优化了局部特征提取过程。SWConv 选择了空间卷积来处理局部细节。空间卷积作为捷径绕过了整个 Transformer 块。由于 patchtokens 被展平为一维，它们需要重新构建为二维特征图。所提出模型的架构如图 4 所示，其中 Conv 模块作为补充组件，集成在所有 Transformer 块中。从公式(6)得到的 1D token x^n 被重塑为 2D 特征图。这些重塑后的特征图经过 GELU 激活函数[16]和批量归一化[17]处理后，输入到空间卷积(SWConv)层。用于空间卷积的卷积核大小为 3 × 3。随后，2D 特征图被重新转换回 1D patch tokens。最后，重塑后的 1D patch tokens x_{1D}^{n+1} 与 Transformer 块的输出(公式(9))相加。得到的和记为 x_{ours}^{n+1}，然后作为输入传递到下一个块。整个过程如下所示：

$$x_{\mathrm{2D}}^{n}=\mathrm{Reshape}_{\mathrm{1D}\to\mathrm{2D}}\left(x^{n}\right) \tag{6}$$

$$x_{\mathrm{1D}}^{n+1}=\mathrm{Reshape}_{\mathrm{2D}\to\mathrm{1D}}\left(x_{\mathrm{2D}}^{\prime n}\right) \tag{7}$$

$$x_{ours}^{n+1}=x^{n+1}+x_{\mathrm{1D}}^{n+1} \tag{8}$$

$$x_{\mathrm{2D}}^{\prime n}=\mathrm{SWConv}\left(\mathrm{BN}\left(\mathrm{GELU}\left(x_{\mathrm{2D}}^{n}\right)\right)\right) \tag{9}$$

在所提出的模型设计中，SWConv 模块作为辅助机制，监督本文提出的 Transformer 架构，二者形成互补关系。具体而言，每个 Transformer 架构都由 SWConv 模块进行监督，以进一步捕捉可能被忽视的局部细节。在该框架中，Transformer 架构作为核心组件，而轻量化的 SWConv 模块则高效提取局部信息，从而显著提升整体性能。与某些复杂的混合架构相比，本文所提出的方法在保持简洁性的同时，展现了高效性和灵活性。

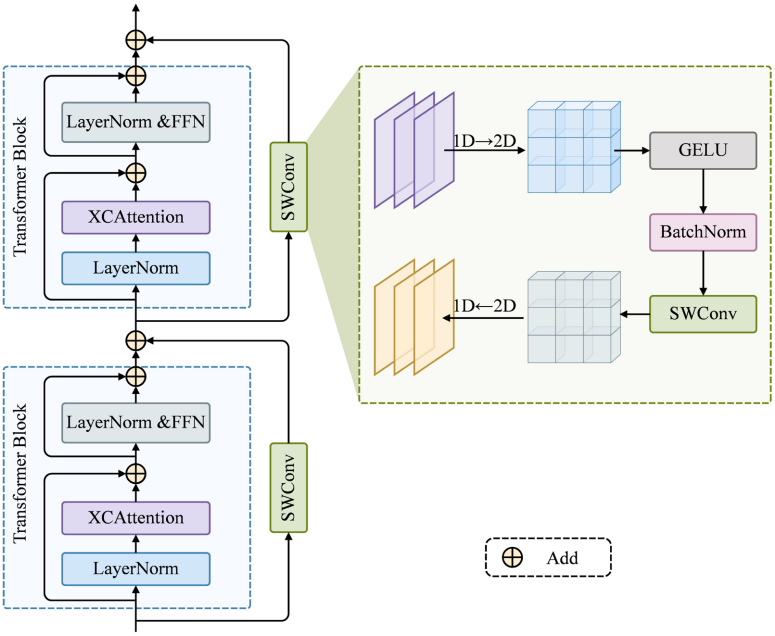

Figure 4. Transformer and SWConv architecture
图 4. Transformer 和 SWConv 架构

2.4. 变体

除了基本架构外，基于核心结构本文还设计了几种变体，如图 5 所示。

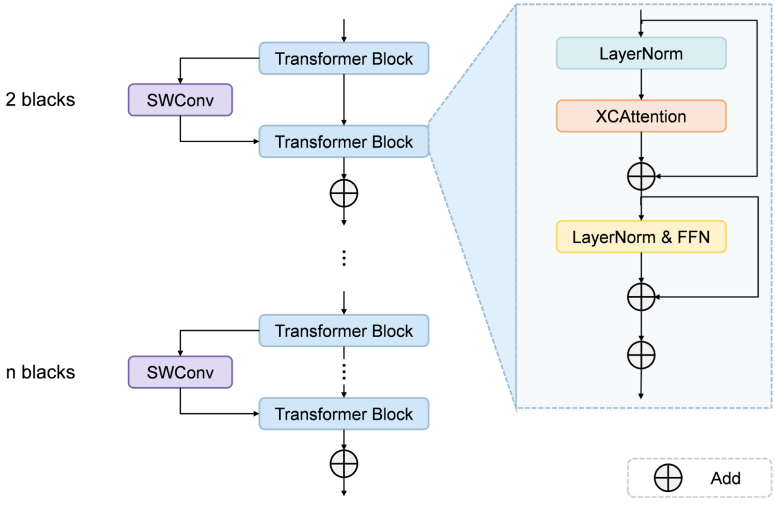

Figure 5. Transformer architecture variants
图 5. Transformer 结构变体

在基本架构中，SWConv 模块绕过每个 Transformer 块；而在其他变体中，该模块则绕过多个 Transformer 块。这些变体在处理更深层的视觉 Transformer 时表现出显著优势，有效减少了参数数量和计算成本。此外，在多阶段 Transformer 架构中，特征图的尺寸会随着阶段的增加而减小，而特征维度则会增大。为了确保空间卷积模块的输入和输出尺寸一致，应该将其绕过范围限制在每个阶段内部，避免模块跨越多个阶段绕过 Transformer 块。

3. 复杂度计算及分析

本文所提出的轻量级模块应用于每个 Transformer 块，并且将 Transformer 中的多头注意力机制替换为 XCA Attention。该方法旨在提高模型效率的同时，保持强大的特征提取能力。与一些将卷积层插入 Transformer 块的方法不同，本文所提出的模块独立于 Transformer 块，可以作为即插即用模块，适用于大多数现有的视觉 Transformer 模型。参数的增加取决于 Transformer 模型的深度和维度。因为该模块独立于每个 Transformer 块且不共享参数，导致较深的 Vision Transformer 模型可能会引入更多参数。然而，与 Transformer 主干相比，参数的增加是微不足道的。例如，在实验中使用的 ViT-Tiny 模型，由 12 个维度为 192 的 Transformer 块组成，3 × 3 深度卷积核的额外参数约为 12 × 192 × (3 × 3 + 1) = 23,040 (0.023 M)，与大约 550 万参数的主干相比，可以忽略不计。此外，由于图像的 patch 大小为 16，且图像被调整为 224，特征图的尺寸为 14 × 14。

对于 ViT-Tiny 模型，增加的计算成本大致为 12 × 192 × (14 × 14) × (3 × 3) = 4,064,256 (0.004 G)，与总计 1.14 G FLOPs 相比微不足道。计算中忽略了 BatchNorm 的参数和计算，因为它们对模型的影响较小。

在实验中，所提出的方法有时甚至能够减少参数和 FLOPs 的数量，因为在小数据集训练时，一些模块和位置嵌入可以被移除。参数和 FLOPs 的增加是最小的，并且高度依赖于模型的层数和维度。此外，它们还取决于所使用的 ViT 模型的架构变体。

一些混合架构将卷积网络融合到 Transformer 结构中，随着卷积网络成为 Transformer 架构的重要组成部分，导致参数和计算量大幅增加。此外，这些方法通常是为特定的 Transformer 架构设计的，因此在其他模型中不具备通用性。与此不同，本文所提出的方法旨在便于集成到各种视觉 Transformer 模型中。复杂度分析表明，该方法引入的额外开销可以忽略不计，大部分参数和计算量仍来自 Transformer 结构。然而，尤其在小数据集上的性能提升是显著的。

4. 实验评估

4.1. 数据集介绍

为了验证所提方法的有效性，本文选择了 ViT-Tiny 和 ViT-Small 模型，并在经典数据集上进行了实验，参数设置参考了文献[18]。具体来说，ViT-Tiny 和 ViT-Small 的特征维度分别为 192 和 384，二者均采用 MLP 比率为 4，MLP 层维度分别为 768 和 1536。所有实验均使用 AdamW 优化器[19]，共进行 100 次训练迭代，其中包括 20 个热身训练轮次，权重衰减系数为 0.05。在三种小数据集 CIFAR10、CIFAR100、Tiny-Image 上的实验中，批量大小设为 128，并使用单个 NVIDIA 4090 Ti GPU 进行训练。学习率遵循余弦衰减策略，初始学习率为 5e−4。输入图像调整为 224，所有 ViT 模型的 patch 大小设为 16。正则化和数据增强策略，包括 colorjitter [20]、AutoAugment [21]、Random Erasing [22]、MixUp [23]和 CutMix [24]。所有实验均从头开始在各自的数据集上训练，未使用任何额外的数据集。

4.2. 改进 ViT 模型的分类性能实验

为验证本文提出模型相较于原始 ViT 模型的有效性，在 CIFAR10 数据集上对二者的准确率和损失值

进行比较，其曲线分别如图 6、图 7 所示。如图 6 所示，当训练轮次为 60 轮左右时，本文提出的模型准确率曲线逐步趋于稳定，而原始 ViT 模型的准确率曲线还呈现上升趋势，直到 80 轮左右时才达到收敛。最终改进 ViT 模型的准确率达 95.7%，原始 ViT 模型准确率为 93.90%。如图 7 所示，随着训练轮次的增加，模型的损失值均不断减小，并在 70 轮左右时逐步趋于稳定，但改进 ViT 的损失值曲线下降速度更快，最终达到 0.101，而原始 ViT 的最终损失值为 0.51，说明本文提出的模型更有效地最小化了训练误差，具有更好的学习能力。因此可以证明，本文的改进策略是合理可行的，提高了模型的稳定性，达到较好的训练效果。

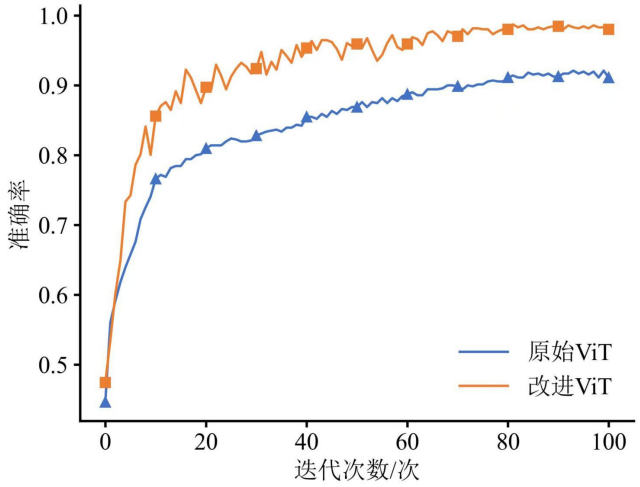

Figure 6. Accuracy comparison curve
图 6. 准确率对比曲线

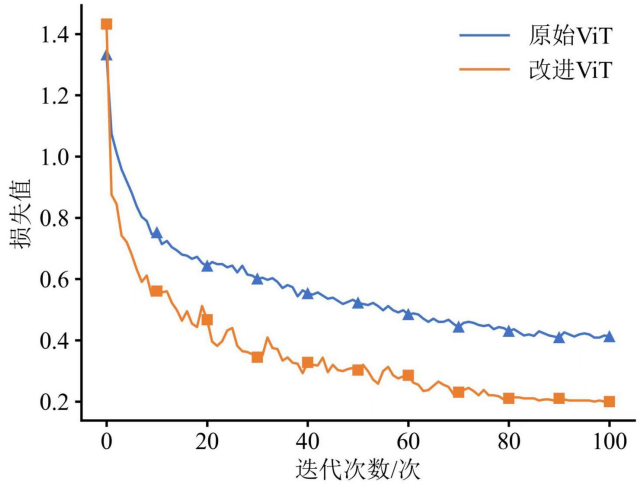

Figure 7. Loss value comparison curve
图 7. 损失值对比曲线

4.3. 消融实验

为了验证本文对 Vision Transformer 的各个改进策略的有效性和性能效果，开展了消融实验和对比实验，在相同的训练参数和配置环境下，对 SWConv 和新的 Transformer 结构进行测试，判断其在 ViT 模型上的有效性，消融实验和对比实验的结果如表 1 所示。在相同的训练参数和配置环境下，对 SWConv

和新的 Transformer 结构进行测试，判断其在 ViT 模型上的有效性，通过对表格中结果的分析，当引入 SWConv 和新的 Transformer 结构时，模型的性能指标在各个数据集上最低提高了约 1.5%，这说明 SWConv 模块和新的 Transformer 结构在提升模型性能上起到了积极作用；当模型单独使用 SWConv 模块后，模型的表现进一步提升，尤其是准确率达到了 95.6%。这一结果表明，SWConv 通过引入局部性和尺度不变性两种归纳偏置，减少了特征冗余和提高了模型对全局与局部特征信息的提取。当 SWConv 模块和新的 Transformer 结构同时使用时，模型的性能达到最高，准确率在 CIFAR10、CIFAR100、Tiny-image 数据集上分别提升 1.8%、3.32%、5%，充分验证了模型改进的有效性，有效证明了两者的结合能够最大化地提升 ViT 的识别效果。

Table 1. Ablation study
表 1. 消融实验

Model	CIFAR10			CIFAR100			Tiny-image		
	Acc	Par	Flops	Acc	Par	Flops	Acc	Par	Flops
ViT-Tiny	93.90%	5.5 M	1.26 G	73.68%	5.5 M	1.26 G	59%	5.6 M	1.26 G
ViT-Tiny w/oPE	87.80%	5.5 M	1.26 G	64.41%	5.5 M	1.26 G	53.15%	5.6 M	1.26 G
ViT-Tiny-X	95.7% (+1.8%)	5.5 M	1.14 G (–0.12 G)	77% (+3.32%)	5.5 M	1.14 G (–0.12 G)	64% (+5%)	5.6 M	1.14 G (–0.12 G)
ViT-Tiny-X w/oPE	95.6% (+1.7%)	5.5 M	1.14 G (–0.12 G)	76.4% (+2.72%)	5.5 M	1.14 G (–0.12 G)	63.2% (+4.2%)	5.6 M	1.14 G (–0.12 G)
ViT-Small	95.09%	21.7 M	4.61 G	73.90%	21.7 M	4.61 G	60.90%	21.8 M	4.61 G
ViT-small w/oPE	89.20%	21.7 M	4.61 G	66%	21.7 M	4.61 G	53.98%	21.8 M	4.61 G
ViT-small-X	96.4% (+1.31%)	21.7 M	4.37 G (–0.24 G)	79.6% (+5.7%)	21.7 M	4.37 G (–0.24 G)	66.7% (+5.8%)	21.8 M	4.37 G (–0.24 G)
ViT-small-X w/oPE	96.31% (+1.22%)	21.7 M	4.37 G (–0.24 G)	79.7% (+5.8%)	21.7 M	4.37 G (–0.24 G)	66.4% (+5.5%)	21.8 M	4.37 G (–0.24 G)

为了改进 ViT 模型结构的作用，本文还在 ViT-small 模型上再次做了消融实验，实验结果显示，当引入 SWConv 和新的 Transformer 结构时，模型的性能指标在各个数据集上最低提高了约 1.31%，当模型单独使用 SWConv 模块后准确率达到了 96.31%，当 SWConv 模块和新的 Transformer 结构同时使用时，模型的性能达到最高，准确率在 CIFAR10、CIFAR100、Tiny-image 数据集上分别提升 1.31%、5.7%、5.8%，再次验证了模型改进的有效性。

本文提出的 SWConv 模块与改进的 Transformer 结构在多个数据集上的表现均优于原始 ViT 模型，无论是单独引入 SWConv 模块，还是与新的 Transformer 结构联合使用，均显著提升了模型的准确率和特征提取能力。尤其在联合使用的情况下，模型在 CIFAR10、CIFAR100 和 Tiny-ImageNet 等数据集上均实现了性能提升，进一步验证了两者在局部建模能力与全局信息交互方面的协同优势。

5. 总结

本文提出了一种新的方法，通过引入 SWC 模块绕过部分 Transformer 块，使 Vision Transformer 模型能够有效捕获全局和局部信息，同时保持较低的计算开销。此外，ViT 模型的结构得到了改进，降低了计算复杂度，并显著提升了特征表示能力，使模型能够更高效地捕获多尺度特征。大量实验结果表明，采用该方法的小型 ViT 模型，在小数据集上的图像分类任务中，超越了具有更多参数和计算复杂度的大型

ViT 模型。此外，计算机制的优化提高了训练效率和稳定性。但 Vision Transformer 的模型参数量较大，对硬件要求较高，难以部署在移动设备上。因此，未来工作希望在保证准确率的同时，降低模型的参数量，并尝试将该模型应用于其他小型数据集中，提高模型的泛化能力。

参考文献

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Un-terthiner, T., Dehghani, M., Minderer, M., Heigold, G. and Gelly, S. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[2] Dong, P., Niu, X., Tian, Z., Li, L., Wang, X., Wei, Z., *et al.* (2023) Progressive Meta-Pooling Learning for Lightweight Image Classification Model. *ICASSP* 2023—2023 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icassp49357.2023.10096973

[3] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and J'egou, H. (2021) Training Data-Efficient Image Transformers & Distillation through Attention. arXiv: 2012.12877.

[4] Wei, Z., Pan, H., Li, L.L., Lu, M., Niu, X., Dong, P. and Li, D. (2022) Convformer: Closing the Gap between CNN and Vision Transformers. arXiv: 2209.07738.

[5] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., *et al.* (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 548-558. https://doi.org/10.1109/iccv48922.2021.00061

[6] Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. (2020) Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv: 2010.04159.

[7] Qin, J., Wu, J., Xiao, X., Li, L. and Wang, X. (2022) Activation Modulation and Recalibration Scheme for Weakly Supervised Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 2117-2125. https://doi.org/10.1609/aaai.v36i2.20108

[8] Tay, Y., Dehghani, M., Bahri, D. and Metzler, D. (2020) Efficient Transformers: A Survey. arXiv: 2009.06732.

[9] Li, G., Wang, Y., Zhao, Q., Yuan, P. and Chang, B. (2023) PMVT: A Lightweight Vision Transformer for Plant Disease Identification on Mobile Devices. *Frontiers in Plant Science*, **14**, Article 1256773. https://doi.org/10.3389/fpls.2023.1256773

[10] He, F., Liu, Y. and Liu, J. (2024) ECA-ViT: Leveraging ECA and Vision Transformer for Crop Leaves Diseases Identification in Cultivation Environments. 2024 4*th International Conference on Machine Learning and Intelligent Systems Engineering* (*MLISE*), Zhuhai, 28-30 June 2024, 101-104. https://doi.org/10.1109/mlise62164.2024.10674238

[11] Wu, S., Sun, Y. and Huang, H. (2021) Multi-Granularity Feature Extraction Based on Vision Transformer for Tomato Leaf Disease Recognition. 2021 3*rd International Academic Exchange Conference on Science and Technology Innovation* (*IAECST*), Guangzhou, 10-12 December 2021, 387-390. https://doi.org/10.1109/iaecst54258.2021.9695688

[12] Sharma, S.K. and Vishwakarma, D.K. (2024) Classification of Banana Plant Leaves Based on Nutrient Deficiency Using Vision Transformer. 2024 5*th International Conference for Emerging Technology* (*INCET*), Belgaum, 24-26 May 2024, 1-6. https://doi.org/10.1109/incet61516.2024.10593120

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. arXiv: 1706.03762.

[14] Sajid, U., Chen, X., Sajid, H., Kim, T. and Wang, G. (2021) Audio-Visual Transformer Based Crowd Counting. 2021 *IEEE*/*CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 2249-2259. https://doi.org/10.1109/iccvw54120.2021.00254

[15] Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) Layer Normalization. arXiv: 1607.06450.

[16] Hendrycks, D. and Gimpel, K. (2016) Gaussian Error Linear Units (GELUS). arXiv: 1606.08415.

[17] Ioffe, S. and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the International Conference on Machine Learning* (*ICML*), Lille, 6-11 July 2015, 448-456.

[18] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. and Beyer, L. (2021) How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers. arXiv: 2106.10270.

[19] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv: 1412.6980.

[20] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[21] Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. and Le, Q.V. (2019) AutoAugment: Learning Augmentation Strategies from Data. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 113-123. https://doi.org/10.1109/cvpr.2019.00020

[22] Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y. (2020) Random Erasing Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 13001-13008. https://doi.org/10.1609/aaai.v34i07.7000

[23] Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D. (2017) MixUp: Beyond Empirical Risk Minimization. arXiv: 1710.09412.

[24] Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y. and Choe, J. (2019) CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6022-6031. https://doi.org/10.1109/iccv.2019.00612