Modeling and Simulation 建模与仿真, 2023, 12(4), 3657-3672
Published Online July 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.124336

Hans汉斯

CNN-Transformer混合模型在计算机视觉领域的研究综述

戴洋毅[1,2]，何　康[1,2]，瑚　琦[1,2*]，黄　凯[1]

[1]上海理工大学光电信息与计算机工程学院，上海
[2]上海理工大学上海市现代光学系统重点实验室，上海

收稿日期：2023年5月5日；录用日期：2023年7月10日；发布日期：2023年7月17日

摘　要

近年来，CNN-Transformer混合模型在计算机视觉领域的研究已经成为热点话题之一。这种模型可以结合卷积神经网络(Convolutional Neural Network, CNN)和Transformer各自的优势，提高模型在多种计算机视觉任务中的性能。首先对CNN与Transformer分别进行简述并分析其优缺点，然后通过介绍与分析近几年国内外表现出色的CNN-Transformer混合模型，对多种常见的混合方式进行分类阐述，这些方法旨在发挥卷积神经网络在局部特征提取方面的优势以及Transformer在全局信息建模方面的优势。最后，对CNN-Transformer混合模型在计算机视觉领域以及其他领域未来所面对的挑战和发展趋势进行展望。

关键词

计算机视觉，卷积神经网络，Transformer，混合模型，深度学习

Review of CNN-Transformer Hybrid Model in Computer Vision

Yangyi Dai[1,2], Kang He[1,2], Qi Hu[1,2*], Kai Huang[1]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[2]Shanghai Key Laboratory of Modern Optical Systems, University of Shanghai for Science and Technology, Shanghai

Received: May 5th, 2023; accepted: Jul. 10th, 2023; published: Jul. 17th, 2023

Abstract

In recent years, research on CNN-Transformer hybrid models in computer vision has become one

*通讯作者。

文章引用: 戴洋毅, 何康, 瑚琦, 黄凯. CNN-Transformer 混合模型在计算机视觉领域的研究综述[J]. 建模与仿真, 2023, 12(4): 3657-3672. DOI: 10.12677/mos.2023.124336

of the hottest topics. This type of model combines the advantages of Convolutional Neural Networks (CNN) and Transformers to improve the performance of various computer vision tasks. First, the pros and cons of CNN and Transformer are briefly introduced and analyzed. Subsequently, various common hybrid methods are elaborated through the introduction and analysis of outstanding CNN transformer hybrid models from national and international research in recent years. These methods aim to leverage the local feature extraction capabilities of Convolutional Neural Networks and the global information modeling capabilities of Transformers. Finally, the paper looks at the challenges and development trends facing CNN-Transformer hybrid models in computer vision and other fields in the future.

Keywords

Computer Vision, CNN, Transformer, Hybrid Model, Deep Learning

1. 引言

近年来，深度学习技术快速发展，并在不同领域中取得了优异的成果。CNN 和 Transformer 是两种被广泛使用的深度学习模型，其中 CNN 具有强大的图像特征提取和分层表示学习能力，在图像分类、目标检测、语义分割等计算机视觉任务中表现出色。Transformer 最初应用于自然语言处理领域，随后被引入计算机视觉领域，其通过自注意力机制可以捕获长距离依赖关系，具有出色的全局建模能力。尽管 CNN 和 Transformer 在各自的领域中都有出色的表现，但也存在一些局限性。为了克服这些局限性，越来越多的研究者开始探索如何将 CNN 和 Transformer 进行结合，设计出可以将两者优势互相补充的 CNN-Transformer 混合模型。这些混合模型的出现为计算机视觉领域带来了新的思路和方法，也为实现更加高效和准确的图像处理任务提供了新的途径。

本文将分别对 CNN 和 Transformer 进行介绍，包括它们的原理和优缺点。然后，我们将介绍 CNN-Transformer 混合模型的基本原理和设计思路，并分析和总结当前涌现的一些常见的混合方法。最后，我们将探讨 CNN-Transformer 混合模型在计算机视觉领域中的应用前景，并为未来的研究提供参考与见解。总之，CNN-Transformer 混合模型作为一种新兴的模型，具有广泛的应用前景，其不断发展和创新，将推动计算机视觉领域的快速发展和进步。

2. CNN 简述

CNN (Convolutional Neural Network)，即卷积神经网络，是一种前馈神经网络。它通过使用卷积层来提取图像中的特征，主要用于处理和分析具有网格状结构的数据。在计算机视觉(Computer Vision, CV)领域，CNN 模型得到了广泛的应用。

CNN 的发展可以追溯到 1980 年代，当时 LeCun 等人提出了 LeNet [1]，并在 MNIST 数据集上取得了较好的表现。随后，CNN 在 1990 年代取得了进一步的发展，并在计算机视觉领域得到广泛应用。其中，AlexNet [2]、ResNet [3]、VGG [4]和 Inception-ResNet [5]等网络的出现大大提高了图像分类的准确率。之后，研究者在 ResNet 的基础上提出了 ResNeXt [6]、DenseNet [7]、SENet [8]、EfficientNet [9]及 ConvNeXt [10]等模型。此外，为了适应计算资源有限的硬件平台，研究者们还开发了一些轻量级的模型，如

SqueezeNet [11]、MobileNet 系列[12] [13] [14]、GhostNet 系列[15] [16]以及 ShuffleNet 系列[17] [18]等。

CNN 具有局部感知性强、鲁棒性强、可拓展性强等优点，但它也存在一些局限性，如：1) 缺乏对全局信息的感知力，导致其对长序列的处理不佳；2) CNN 使用卷积操作提取特征，会导致输入数据的位置信息丢失；3) 当卷积核和图像尺寸较大时，计算量较大，需要较高的计算资源；4) 由于参数共享机制，对于一些需要考虑细节的任务，如图像超分辨率任务上表现不佳，等。

3. Transformer 简述

3.1. 引言

Transformer [19]是一种基于自注意力机制的序列建模方法。最初主要用于自然语言处理(Natural Language Processing, NLP)领域，后来也在计算机视觉领域得到了广泛应用。

ViT [20] (Vision Transformer)是 Transformer 在计算机视觉领域的首次应用，它将图像分类问题转化为序列建模问题。随后，Swin Transformer [21]通过引入基于滑动窗口的自注意力机制，结合了局部感受野，提高了计算效率和准确率。该模型在多种计算机视觉任务上取得了显著的性能提升。Swin Transformer V2 [22]进一步优化了原始 Swin Transformer 的结构，提高了模型性能和训练稳定性。Han 等人[23]在他们的文章中详细介绍了 Transformer 在计算机视觉领域的最新研究进展。

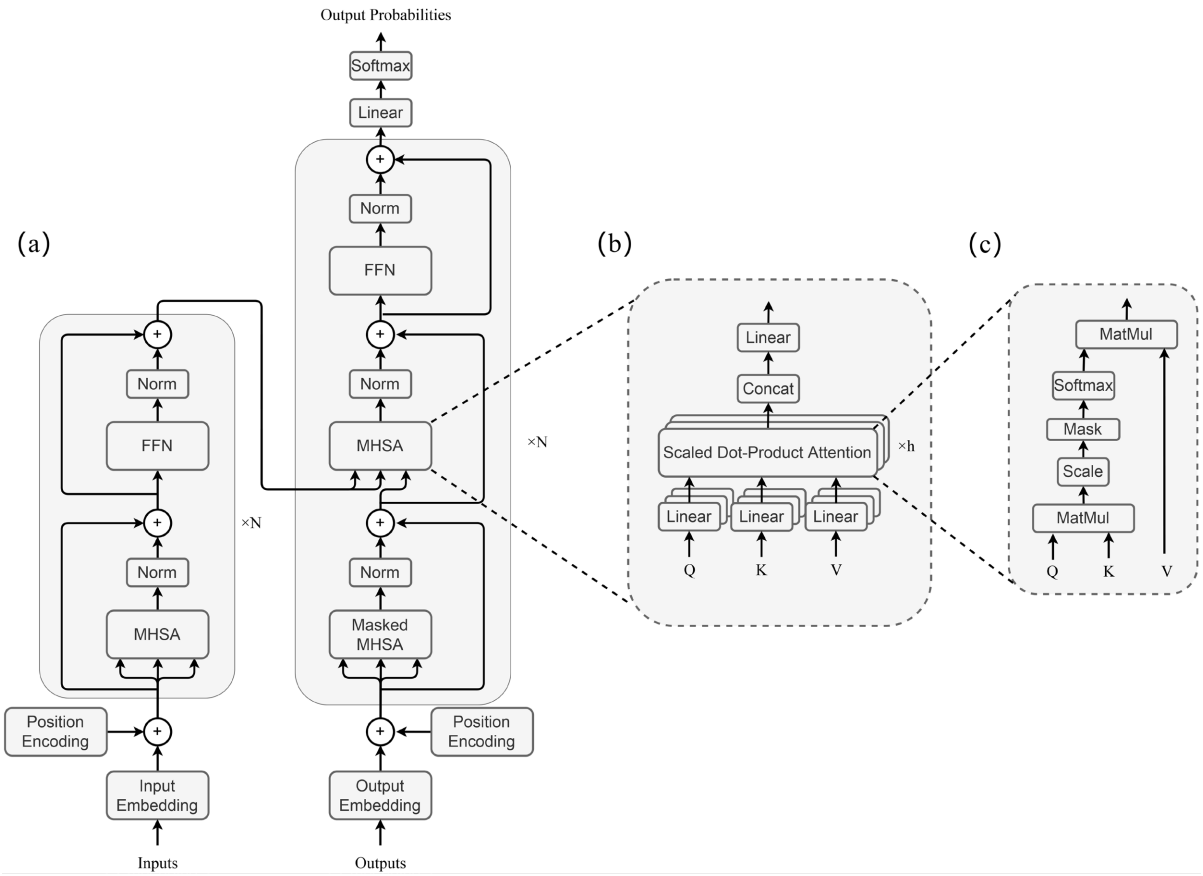

Figure 1. (a) Overall structure of transformer (b) multi-head self-attention layer (c) single-head self-attention layer
图 1. (a) Transformer 整体结构 (b) 多头自注意力层结构 (c) 单头自注意力层结构

本章将介绍原始 Transformer 及 ViT 的核心原理和组成部分，为后续章节提供关键的理论基础，以求

更好地理解混合模型的构建和优化过程，以及它们在实际应用中所展现出的潜力。

3.2. NLP 领域的 Transformer

Vaswani 等人在 2017 年首次提出 Transformer 模型并将其应用于 NLP 领域中的机器翻译任务，其结构如图 1(a)所示。它由多组编码器与解码器模块堆叠而成，编码器负责生成输入序列并然后传送到解码器中，随后利用其中的上下文信息生成输出序列。编码器模块和解码器模块均由多头自注意力层、前馈神经网络层、层归一化及残差连接层(多条恒等映射)构成，其中多头自注意力层结构由多个平行的单头自注意力层组成，结构分别如图 1(b)及图 1(c)所示。Transformer 的关键组成部分包括：

1) 位置编码：为了在输入时保留语句中单词序列的顺序关系，Transformer 利用位置编码(Position Encoder, PE)模块为输入序列中的每个单词添加相对或绝对位置信息。

2) 自注意力机制：Transformer 引入了缩放点积注意力机制(Scaled Dot-Product Attention)，即自注意力机制(Self-Attention)，该机制可以将输入向量映射到多个不同的子空间中，从而获取更丰富的特征信息。多头自注意力机制是将输入序列的维度划分为多个头，并为每个头定义不同的权重矩阵，以此将输入向量映射到多个不同的子空间中，使模型可以从更多角度进行学习并获取更丰富的特征信息。

3) 前馈神经网络：前馈神经网络(Feedforward Neural Network, FNN)层由两个线性变换层和一个 ReLU 激活层组成，该模块连接在每个编码器和解码器的多头自注意力模块之后。

3.3. CV 领域的 VisionTransformer

ViT 最大程度保留了 Transformer 的原始结构，并将其用于计算机视觉领域的图像分类任务。为了使用 Transformer 处理二维的图像，ViT 提出补丁嵌入层，将输入的二维图像划分为一系列扁平的二维图像块，并将其作为输入序列。随后经过位置嵌入层，为每个输入序列添加位置信息。同时为了方便后续将模型应用于图像分类任务，ViT 借鉴了 BERT [24]中的[class]标记，用一个可学习的特征表示整个图像的特征信息，最后将经处理后的输入序列传送至编码器中。

ViT 由于模型结构简单易懂，可扩展性强且效果好而成为了 Transformer 在 CV 领域中应用的里程碑，但 ViT 仍然存在部分缺点，如：1) 对数据集要求高：由于缺乏类似 CNN 的归纳偏置特性，导致 Transformer 在数据量不足的情况下缺乏足够的泛化能力，难以达到理想的效果；2) 对硬件要求高：由于自注意力机制的计算复杂度高，且其计算量的增速是图片大小增速的二次方，因此对硬件要求非常高；3) 参数多；4) 缺少空间归纳偏置，对空间信息不敏感；5) 绝对位置编码导致迁移性差；6) 模型训练困难，等。

4. 结合 CNN 与 Transformer 的常见方法

4.1. 引言

CNN 在图像处理中有着良好的表现，能够处理复杂的图像特征，其在处理局部特征方面表现出色，但是在处理全局信息时表现较弱。而 Transformer 在 NLP 领域中具有很好的表现，能够处理序列数据的建模和生成，其在处理全局信息方面具有优势，但是对于局部信息处理能力相对较弱。通过将 CNN 与 Transformer 进行结合，可以有效地捕捉与处理图像中的局部和全局信息，从而提高模型的性能和效果。本节将对几种常见的混合方法进行介绍。

4.2. 基于架构设计参考

深度学习模型的架构设计是一个非常关键的问题，因为它在一定程度上可以决定模型的性能和训练效率。如 3.3 节中所介绍，原始的 ViT 模型是一种完全基于 Transformer 结构的视觉模型，相比传统的

CNN 视觉模型，ViT 在模型结构和特征表示方面存在不足。

首先，ViT 将图像分为若干个固定大小的补丁，然后对其进行特征提取和分类，但这样会导致原始 ViT 模型对输入图像的尺寸比较敏感而且无法充分利用图像的全局信息，对模型的分类性能有较大影响。其次，原始 ViT 缺乏 CNN 中多层卷积和池化等操作，因此在提取图像特征时可能会受到一定的限制，导致对于一些纹理、形状等细节信息的提取能力相对较弱。

为了解决这些问题，研究者们开始探索将 CNN 和 Transformer 结构进行融合。最直接的想法就是将 CNN 常见的多尺度金字塔、残差连接等结构引入 ViT 中，从而提高模型对于输入图像尺寸的适应性和特征提取能力。

CNN 的多尺度金字塔结构可以更好地处理多尺度问题，提高模型的多尺度表示能力、局部特征提取能力、特征重用能力、鲁棒性和可解释性等方面的性能。Wang 等人便是在保留 ViT 原始特点的情况下，参考并引入了 CNN 架构中的金字塔结构来改进 ViT 在处理高分辨率图像时的效果，提出了 Pyramid Vision Transformer 模型[25]，即 PVT。不同于 VGG、ResNet 等模型使用不同的卷积步幅来获取多尺度特征图，PVT 采用了渐进收缩策略，其利用补丁嵌入层和加入了空间缩减机制的编码器实现对特征图尺度的灵活调整，这使得 PVT 既可以像传统 CNN 模型一样生成多尺度特征图，又降低了计算成本。

UNeXt [26]和 Uformer [27]分别采用了不同的方法参考 UNet [28]架构来优化 Transformer 模型在计算机视觉领域的应用。UNet 多用于图像分割任务，它的结构可以更好地提取和利用不同尺度下的特征，同时可以通过跳跃连接将不同尺度下的特征相互关联，提高模型的表现和性能。

UNeXt 中的编码器部分采用了类似 UNet 的结构，将输入图像逐步下采样到多个低分辨率特征图，然后通过多个解码器分支逐步上采样恢复分割结果。与 UNet 不同的是，UNeXt 中的解码器部分采用了 Transformer 模型，以求实现更好的特征表示和上下文建模能力。通过这种方式，UNeXt 可以在保持 UNet 优良分割效果的同时，提高模型对图像语义信息的表示和理解能力。

Uformer 在模型中引入了类似于 UNet 的跨层连接结构，以促进高层次和低层次特征的信息流动。该模型中的每个 Transformer 块都包含了一个多分辨率的注意力模块，该模块将不同层次的特征图进行特征融合，并通过一个反卷积模块进行上采样，以恢复分割结果。通过这种方式，Uformer 可以更好地利用低层次特征的详细信息和高层次特征的抽象语义信息，从而提高模型在图像分割任务中的性能。

ResNet 中的残差连接结构可以将不同层级的特征相互关联，提高模型的表现和性能。CSWin Transformer [29]模型采用了多层 Transformer 结构，并在每个模块之间引入了类似于残差连接结构，其通过参照并学习 ResNet 的架构实现了模型效果的优化，提高了模型的表达能力和学习效率。

HRFormer [30]则是在 HRNet [31]的多分辨率特征融合架构的基础上引入了 Transformer，这使得该模型既具有多分辨率特征融合架构的优势，又能更好地建模长距离依赖关系，捕捉到更全局的语义信息。

综上，这些模型通过参照 CNN 的架构设计，对原始 ViT 的不足进行改进。基于架构设计参考所提出的 CNN-Transformer 混合模型能更高效地提取特征，在不同的任务和数据集上具备了更好的可拓展性。注意力机制的引入，使得模型对光照、遮挡等复杂场景具备更好的鲁棒性。但同时也很容易造成模型计算成本高、参数量大、可解释性差等劣势。因此在基于架构参考设计混合模型时，仍需要在实际研究中对相应细节进行改进。

4.3. 基于知识蒸馏

知识蒸馏[32] [33] [34] (Knowledge Distillation, KD)是一种模型压缩方法，其核心思想是利用效果好且泛化能力强的复杂模型作为教师网络，去指导训练表达能力较弱的小模型，即学生网络，使学生网络也具备教师网络的泛化能力，实现知识迁移。因此基于知识蒸馏实现 CNN 与 Transformer 各自优势的结

合也是一个可行的方法。

Touvron 等人[35]提出的 DeiT 便是采用了知识蒸馏的方法对 ViT 进行优化，其在输入序列中加入了蒸馏令牌，与分类令牌、图像块令牌一起输入编码器中进行交互计算，同时利用泛化能力较强的 CNN 模型作为教师网络得到一组标签，然后计算编码器输出的蒸馏令牌与教师网络的输出标签的交叉熵损失，这种方法称为硬蒸馏策略。通过这种方法，DeiT 成功将 CNN 模型的归纳偏置特性引入了 Transformer 模型中，降低了 Transformer 模型对数据量的需求，提升了训练速度，同时取得了更好的性能和效果。

4.4. 基于串并联拼接

CNN 擅长有效地提取图像的局部特征，具有更好的泛化能力和更快的收敛速度。Transformer 则更擅长捕捉全局的语义信息，在大数据集上可以获得很好地效果。相比于单独使用 CNN 或 Transformer，将它们进行串联或并联的拼接，可以获得更好的表现。研究者提出了许多基于结构拼接的思想的混合模型，并对不同的拼接方式进行了分析研究。

Carion 等人[36]采用了先 CNN 再 Transformer 的串联拼接方式，提出了 DETR，首次将 Transformer 用于图像目标检测。他们首先利用 CNN 网络学习输入图像的二维特征，然后将 CNN 网络提取出的低分辨率的特征图重塑为一系列特征序列并对其进行位置编码，将结果输入到 Transformer 中进行学习，得到分类标签和预测框。这种方法有效减小了 Transformer 的输入尺寸，方便 Transformer 快速学习输入图像的全局特征，提高了模型的学习速度以及整体性能。

Dai 等人[37]采用了同样的拼接策略，提出了 CoAtNet，他们认为卷积块更擅长获取局部先验，因此应设计在 Transformer 模块之前。他们分别利用基于深度卷积改进的反向残差瓶颈 MBConv 模块和相对自注意力模块为模型引入 CNN 和 Transformer 的特性，并成功将 CNN 的平移同变性(Translation Equivariance)和 Transformer 的输入自适应加权(Input-Adaptive Weighting)、全局感受野(Global Receptive Field)的优点融合到单一架构中，获得了兼具泛化能力、模型容量和模型效率的混合模型。

Beal 等人[38]则是选择先 Transformer 再 CNN，提出了 ViT-FRCNN。该模型在 ViT 后串行拼接了 Faster R-CNN [39]作为模型的目标检测网络，然后将 ViT 的输出重组并传输至到目标检测网络的残差块，得到最终的分类标签和预测框，实现利用 Transformer 完成图像目标检测任务的目的。这种方法的成功也证明了 Transformer 主干可以保留足够的空间信息用于目标检测。

Yan 等人[40]设计的 ConT 块也采用了先 Transformer 再 CNN 的串行拼接策略，并在此基础上构建了 ConTNet。他们将标准 Transformer 编码器视为与卷积块相同的独立组件，然后将两个编码器与一个卷积核尺寸为 3 × 3 的卷积层串联，并将其命名为 ConT 块。图像进入网络后将首先经过一个卷积核尺寸为 7 × 7 的卷积层和一个最大池化层，然后通过由多个 ConT 块堆叠而成的框架，最后经过一个全局平均池化层和一个全连接层实现相关计算机视觉任务。这样既利用了 Transformer 强大的表征能力，又具备 CNN 的偏置归纳特性。该模型具备容易优化、鲁棒等优点，且不依赖于强大的数据增强和训练技巧，具有良好的迁移学习能力。

Mehta 等人[41]在 MobileViT 中采用了类似的设计策略。他们在该模型中参照 ViT 的结构设计了 MobileViT 块，该模块用更擅长处理全局信息的 Transformer 块替代了常规卷积中的矩阵乘法运算过程。然后将该模块与倒置残差块进行交替串联堆叠，实现对 CNN 和 Transformer 的优势的融合。MobileViT 块的主要思想是将 Transformer 视为卷积，这使得它可以同时获得卷积的归纳偏置特性和 Transformer 的全局性。与 ViT 相比，MobileViT 块既不会失去图像块的顺序信息，也不会丢失图像块中像素的空间位置信息。

随后 Mehta 等人[42]在 MobileViT 的基础上对其进行改进，提出了 MobileViTV2。虽然与一些轻量级

CNN 相比，MobileViT 的参数更少而且效果更好，但是由于使用了多头自注意力机制，导致该模型具有较高的延迟。基于此，他们提出了一种可分离自注意力机制替换了多头自注意力机制，提高了计算速度和内存效率。

不同于其他混合网络普遍采用的串联结构，Peng 等人[43]选择将 CNN 和 Transformer 进行并行拼接，提出了 Conformer，他们设计了一个特征耦合单元(Feature Coupling Unit, FCU)，通过该模块实现以并行交互的方式融合 CNN 网络所提取的局部特征和 Transformer 网络所提取的全局特征，将它们有效地整合成为最终的特征表示，并使得该模型在多个视觉识别任务中取得了优秀的性能。

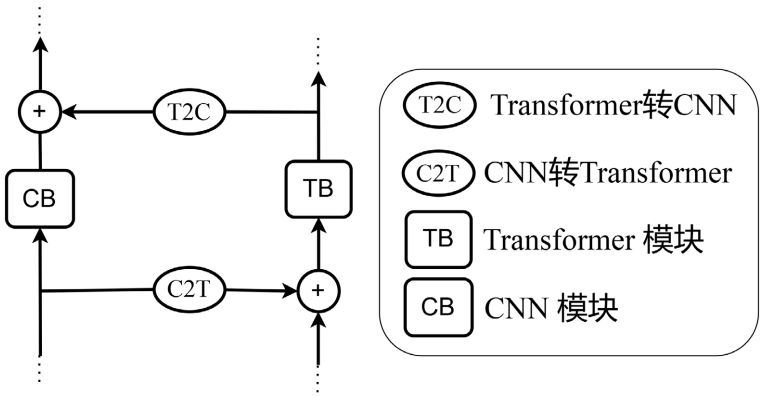

Figure 2. Bidirectional cross-bridging method
图 2. 双向交叉桥接方式

Chen 等人[44]在他们所提出的 Mobile-Former 中，采用了创新性的双向交叉桥接方式，其结构如图 2 所示。其中 CNN 网络模块由以 MobileNetV2 的倒置瓶颈块为参考提出的 Mobile 块堆叠而成，该模块以原始图像作为输入，利用深度卷积和点卷积提取局部特征。而以 Transformer 网络为参考提出的 Former 块则是由一个多头自注意力层和一个前馈神经网络组成，以可学习参数或令牌作为输入。这些标记是随机初始化的，每个标记代表了图像的全局先验信息，而 ViT 中的标记是图像块的线性映射，前者所使用的标记明显更少，因此计算量相比于 ViT 也更少，拥有较低的计算成本和更高的效率。CNN 部分的计算复杂度为 O (HWC^2)，Transformer 部分的计算复杂度为 O (M^{2}d + Md^2)，其中 M 和 d 分别为嵌入块的数量和维度。

该模型所提出的双向交叉注意力同时利用了 CNN 的局部性优势和 Transformer 的全局性优势，实现了局部特征和全局特征的融合。将局部特征映射表示为 X，全局标记表示为 Z，然后将它们分别拆分为 h 个头，即 $X=\left[\tilde{x}_1 \ldots \tilde{x}_h\right]$， $Z=\left[\tilde{z}_1 \ldots \tilde{z}_h\right]$，用于多头自注意力机制，则从 Mobile 块到 Former 块的桥接计算公式如公式(1)所示：

$$A_{X \to Z}=\left[Attention\left(\tilde{z}_i W_i^Q, \tilde{x}_i, \tilde{x}_i\right)\right]_{i=1:h} W^O \tag{1}$$

其中 W_i^Q 是第 i 个头的查询 Query 映射矩阵，W^O 是用于结合多头注意力的矩阵，Attention 函数代表标准注意力函数。

从 Former 块到 Mobile 块的桥接计算公式如公式(2)所示：

$$A_{X \to Z}=\left[Attention\left(\tilde{x}_i, \tilde{z}_i W_i^K, \tilde{z}_i W_i^V\right)\right]_{i=1:h'} \tag{2}$$

其中 W_i^K 是第 i 个头的 Key 映射矩阵，W_i^V 是第 i 个头的值 Value 映射矩阵。

值得注意的是，相比于标准自注意力机制，在 Mobile 端公式(1)的计算中去除了 Key 矩阵和 Value

矩阵，而在 Former 端公式(2)的计算中去除了 Query 矩阵，这样做的目的是节省计算量。Mobile 块和 Former 块之间通过双向交叉注意力组成一个 Mobile-Former 块。

Mobile-Former 的优点是计算效率高，表现力强，但也因为并行桥接了两种不同的模型导致参数共享效率不高，参数量大，在模型大小方面存在局限性。

Yoo 等人[45]所提出的 ACT 同样采用了双向桥接的方式实现并行结构，不同于 Mobile-Former 的交叉桥接设计，ACT 采用了双向同步桥接设计，其结构如图 3 所示。该模型融合 CNN 与 Transformer 的优点以解决超分辨率任务，其中 CNN 分支采用了 RCAN [46]模型中的卷积块残差通道注意力块(Residual Channel Attention Block, RCAB)。两种分支之间通过融合块(Fusion Block, FB)进行桥接实现特征融合，该模块将每步的输出作为融合块的输入，然后将结果与上一步的输出结合作为下一步的输入。这种方法在多个超分辨率数据集下取得了 SOTA (State Of The Arts)效果。

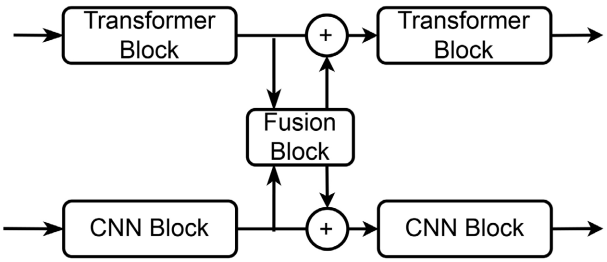

Figure 3. Bidirectional synchronous bridging method
图 3. 双向同步桥接方式

通过对多个具有代表性的混合模型进行介绍和分析，我们发现基于串并联拼接的混合方式可以结合 CNN 在处理图像时的优势以及 Transformer 在建模序列数据时的优势，进一步提高模型的表现。此外，通过对不同的拼接方法进行探究，我们可以发现在不同的任务场景下，采用不同的混合方式可以获得更好的性能。值得注意的是，由于混合模型具有更复杂的结构，因此训练和推理时间相对较长，且需要更高的计算资源。因此，对于具体任务的选择和模型结构的设计，需要根据具体场景和需求来确定。

4.5. 基于局部替换

在 CNN 和 Transformer 混合模型的研究中，一种被广泛探讨的混合方式是替换网络中的特定模块，以结合 CNN 和 Transformer 的优势。通过替换不同的模块，可以使得混合模型更加灵活，同时也可以更好地结合 CNN 和 Transformer 各自的优点，如：

1) 对嵌入块的替换

一种常见的策略是替换嵌入块。例如，ViT_C [47]使用多个 3 × 3 卷积和一个 1 × 1 卷积的堆叠取代了 ViT 中的 16 × 16 卷积，显著提高了模型在 ImageNet-1k 数据集上的表现。相比之下，Hassani 等人[48]提出的 CVT 和 CCT 模型则通过改变输入数据的处理方式，提高了模型对局部空间信息的敏感性，减少了对位置编码的依赖。其中 CVT 采用了一种基于自注意力机制的序列池化方式，该方式通过融合每个块标志的嵌入用于分类，直接将编码层产生的含顺序信息的序列进行池化。CCT 在 CVT 的基础上进一步对输入数据进行处理，将图像经 CNN 网络处理后再划分块标志，然后输入 CVT 中，卷积层所提供的归纳偏置特性使得模型可以更好的保留局部的空间信息。这些方法都有效，但它们在性能、稳定性、泛化能力和计算成本等方面存在差异。例如，ViT_C 对学习率和优化器选择非常鲁棒，但 CCT 模型在小数据集上的表现更好，且计算成本更低。

2) 对前馈层的替换

另一种策略是替换 Transformer 的前馈层。例如，LocalViT [49]通过比较并发现了前馈网络和反向残差块结构的相似性，将反向残差块中所使用的深度卷积引入前馈网络中，从而为其引入局部性。这种方法的优点是它可以改进模型的性能，但可能会增加模型的复杂性。

3) 对自注意力层的替换

d'Ascoli 等人[50]提出了 ConViT，他们认为在 Transformer 中引入 CNN 的归纳偏置特性有利也有弊，在大数据集上进行训练时，固有偏置归纳特性会限制模型的上限，因此他们引入了一个门控位置自注意力层(Gated Positional Self-Attention, GPSA)，该模块可以根据上下文信息自主决定是否表现为卷积层，从而实现可控的归纳偏置。Srinivas 等人[51]提出的 BoTNet 将 ResNet 的瓶颈层中尺寸为 3 × 3 的卷积替换为多头自注意力模块，并将其命名为 BoT 块，然后用 BoT 块替换 ResNet 框架中最后三个瓶颈层。此方法原理简单，但在实例分割和目标检测任务中显著提高了基线，且具备很小的训练与推理开销，是一个值得研究的主干架构。

总的来说，基于局部替换的混合模型为研究者提供了一种有效的策略，可以结合 CNN 和 Transformer 的优点，同时保留模型的灵活性。然而，选择最适合的方法需要考虑多个因素，包括任务类型、数据集大小、模型复杂性以及计算成本等。

5. 多层次混合 CNN 与 Transformer

如 4.2 节至 4.5 节所述，在 CNN-Transformer 混合模型的设计过程中，研究者们探索了不同的混合方式以实现更好的性能表现。然而，许多最新的混合网络设计并不仅仅使用一种混合方法，而是采用多种方法的组合，本章将介绍这些多种方式混合的网络设计，并分析它们如何结合 CNN 与 Transformer 的优势以提高模型性能表现。

Graham 等人[52]参考 LeNet，为 ViT 引入卷积组件，提出了 LeViT。该模型采用了多分辨率金字塔结构，用注意力机制实现下采样，同时采用了一种可学习且具有不变性的注意力偏差，取代传统 ViT 中的位置嵌入。此外，他们删除了分类标记，而是参考卷积网络，利用一个激活映射上的平均池化来代替它，有效减少了第一层网络中的特征数量。LeViT 实现了传统 ViT 在宽度和空间分辨率方面的缩小。

Wu 等人[53]提出的 CvT 也通过在 ViT 中引入卷积来提高性能和效率，主要提出了两个改进：1) 包含一种新的卷积令牌嵌入(Convolutional Token Embedding, CTE)的 Transformer 层次结构；2) 一种使用卷积映射的 Transformer 块(Convolutional Projection for Attention, CPA)。通过以上两种改进，该模型成功将 CNN 网络的一些理想特性，如平移、缩放和失真不变性引入了 ViT 网络中，使得模型可以捕捉局部信息以及局部信息之间的空间关系，并在减少计算量和参数量的同时，保持了 Transformer 网络自身的优点，如动态注意力、全局感受野、泛化能力强等。结果表明，以合适的方式移除 ViT 中的位置编码模块可以在简化结构设计的同时保持网络的性能，并使其更适合处理不同分辨率的输入。

Yuan 等人[54]在 CeiT 中首先设计了一个基于卷积的 Image-to-Tokens (I2T)模块，该模块从生成的低级特征中提取补丁。随后将编码器中的前馈网络替换为局部增强前馈层(Locally Enhanced Feed-Forward, LeFF)，它将令牌重新拼接位特征图，然后采用深度可分离卷积提取局部信息，再将其结果重新映射为令牌，这促进了空间维度上相邻令牌之间的相关性。最后，类似于特征金字塔，它在 Transformer 顶部附加了一个逐层分类标签自注意力模块(Layer-wise Class token Attention, LCA)，将全部的分类令牌作为输入共同参与结果预测，以获得多尺度的表征。通过以上三种改进，CeiT 成功将 CNN 网络的不变性和局部性引入 Transformer 中，使模型既具备 CNN 网络提取低级特征和局部信息的能力，也具有 Transformer 网络获取远程依赖关系的优势。与之类似的是，Jeevan P 等人[55]在 CXV 也在网络最前面添加了卷积层以引入卷积网络的归纳先验特性，替换掉了 ViT 中的分类令牌和位置编码嵌入，然后使用线性自注意力机制

取代了原始的自注意力机制以减少 GPU 的负荷并提高模型在小数据集上的性能。

EdgeViTs [56]和 CMT [57]都采用了类似于 ResNet 的多层级结构。其中，EdgeViTs 引入了一个基于自注意力机制和卷积的最优集成的具备高成本效益的 Local-Global-Local (LGL)信息交换瓶颈块。LGL 主要由三阶段组成，分别为基于深度卷积和点卷积的局部聚合模块(Local Aggregation)、全局稀疏自注意力模块(Global Sparse-Self-Attention)和基于转置卷积的局部传播模块(Local Propagation)，其中稀疏自注意力模块的使用降低了自注意力机制的复杂度，在精度与延迟之间实现了更好的平衡。EdgeViTs 的提出使得基于自注意力的视觉模型在准确性和设备效率的均衡情况下能够与最好的轻量 CNN 模型竞争。而在 CMT 中，输入图像首先通过卷积预处理层进行细粒特征提取，然后将结果输送进一系列堆叠的 CMT 模块中进行表示学习。CMT 模块由一个基于深度卷积的局部感知单元(Local Perception Unit, LPU)、一个轻量级的多头自注意力模块(Lightweight Multi-Head Self-Attention, LMHSA)和一个参考反向残差前馈网络(Inverted Residual Feed-Forward Network, IRFFN)组成，用于同时捕获全局和局部结构信息提高网络的表示能力。其中局部感知单元通过使用深度卷积提取局部信息，为模型提供了平移不变性。轻量级多头自注意力模块利用两个深度可分离卷积分别对 Key 和 Value 的生成进行下采样处理，减小了多头自注意力机制的计算量。反向残差前馈网络参考 MobileNetV2 中的反向残差块，在原始前馈网络中加入了深度可分离卷积增强局部信息的提取能力，而残差结构促进了梯度的传播能力。此外，CMT 参考 EfficientNet，提出了一种适用于 Transformer 架构的复合缩放策略，使模型可以在计算成本与性能之间取得更好的平衡。

Zhang 等人[58]在其提出的 ParC-Net 中引入了一种轻量型的位置感知循环卷积(Position aware circular Convolution, ParC)，它既具备 ViT 结构的全局感受野，又拥有卷积的局部性。ParC 结构主要包含三个改动：1) 采用循环填充代替传统的零填充，增大感受野以提取全局特征；2) 引入位置编码嵌入，消除使用循环卷积时对空间结构的影响，保持输出特征对空间位置信息的敏感度；3) 引入插值函数，对于不同的输入分辨率动态生成相应大小的卷积核和位置编码嵌入。他们将 ParC 与 SENet 相结合，构建了一个类似于 Meta-Former [59]结构的纯卷积结构 ParC 块，该结构显著降低了计算成本，同时保留了能够提取全局特征的特点。经过参考 CoAtNet 和 MobileViT 的成功经验，ParC-Net 的外框架搭建也采用了分叉结构。他们保留了 MobileViT 中浅层阶段的 MobileNetV2 块，用 ParC 块替换了深层阶段的 ViT 块，使其转变为一个纯卷积网络。ParC-Net 在图像分类、目标检测及语义分割任务中均取得了最好的整体性能表现。

Li 等人[60]提出了 Next-ViT，它以创新的方式融合了卷积与 Transformer。他们设计了基于多头卷积注意力(Multi-Head Convolutional Attention, MHCA)的 NCB 模块(Next Convolution Block)用以提取局部特征，而 NTB 模块(Next Transformer Block)负责提取全局特征，同时作为一个轻量级的高低频信号混合器以增强建模能力。此外，传统混合策略一般选择在网络前期采用卷积块，然后在网络后期堆叠 Transformer 块，然而采用这种混合策略的模型很容易在分割和检测等下游任务上达到性能饱和。为了克服传统混合策略的不足，他们提出了一种新的混合策略 NHS (Next Hybrid Strategy)，该策略创造性地以(N + 1)*L 混合范式堆叠 NCB 模块和 NTB 模块，大大的减少了 Transformer 块的比例，显著提升了模型在下游任务上的性能，实现了高效部署。

Maaz 等人[61]提出的 EdgeNeXt 也采用金字塔型结构，然后引入了自适应卷积编码器和分割深度转置注意力编码器(Split Depth-wise Transpose Attention, SDTA)，后者将输入张量分割为多个信道组，利用深度卷积和跨信道维度的自注意力机制来隐式增加感受野并编码多尺度特征。他们通过使用互协方差注意力机制，将原始的空间维度的计算转换为特征通道维度计算，既有效地提取了全局特征，又显著减少了计算复杂度。EdgeNeXt 在图像分类、目标检测及语义分割任务中实现了很好的效果。此外，它在减小了模型大小和参数量的同时，也保持了较少的运算量，使得该轻量化模型能够方便地部署在边缘设备上。

Pan 等人通过将卷积和自注意力机制分别进行分解，研究它们之间的共性与异性，并对二者的内在联系加以利用，提出了 ACmix [62]。对于卷积操作，他们将传统卷积核分解为多个 1 × 1 卷积，然后进行移位和求和操作；对于自注意力机制，他们将 Query、Key 和 Value 的投影视为多个 1 × 1 卷积，然后对自注意力权值和聚合值进行计算。通过共享卷积与自注意力机制的第一阶段的计算资源，显著减少了计算复杂度，成功地将两种不同范式进行了结合。因此，ACmix 兼具自注意力机制和卷积的优点的同时，又拥有更少的计算复杂度。

这些网络的设计思想各具特色，通过组合不同的混合方式在不同程度上融合了 CNN 和 Transformer 的优势，对于特定的任务和应用场景都有一定的优势和适用性。

6. 性能评价

6.1. 数据集及常用评价指标

6.1.1. ImageNet

ImageNet 数据集是一个大型视觉数据库，旨在推动计算机视觉和深度学习领域的研究。该数据集包含超过 1400 万张标注过的高分辨率图片，涵盖了 2 万多个类别。ImageNet-1k 是 ImageNet 数据集的一个子集，它包含了 1000 个类别。该数据集常用于训练和评估计算机视觉中的图像分类和物体检测算法。由于 ImageNet-1k 的类别数量较少，与整个 ImageNet 数据集相比，它的规模更小、计算复杂度更低，因此更适合用于研究和实验。常用的评价指标为 Top-1 Accuracy，表示模型预测的类别与真实类别相同的比例。在实际应用中，Top-1 准确率越高，表示模型的分类性能越好。

6.1.2. COCO

COCO (Common Objects in Context)由微软于 2014 年发布，是一个用于图像识别、物体检测、分割和关键点检测等计算机视觉任务的大型数据集。该数据集包含了 33 万张高质量的图像，涵盖了 91 种常见物体类别，共计超过 200 万个标注框。这些类别包括不同的动物、食物、交通工具等。COCO 数据集的图片特点是场景复杂、多样，物体之间存在大量的遮挡关系。常用的评价指标为平均精度(Average Precision, AP)。

6.1.3. ISIC

ISIC (International Skin Imaging Collaboration)数据集是一个皮肤病变图像的大型数据集，主要用于皮肤癌检测和分类任务。ISIC 数据集包含多种皮肤病变类型，如良性痣、基底细胞癌、鳞状细胞癌和黑色素瘤等。该数据集的图片来源于多个不同的数据库和临床研究，涵盖了各种年龄、性别、肤色和皮肤病变部位。这使得 ISIC 数据集具有较高的多样性，对算法具有较强的泛化挑战。常用的评价指标为 F1 分数(F1-score)：精确度和召回率的调和平均值，用于衡量模型在精确度和召回率之间的平衡。

6.2. 效果对比

表 1 展示了不同混合模型在 ImageNet-1K 数据集上的性能对比。这些模型采用了多种不同的结构，包括 CNN、Transformer 和多种采用混合架构的模型。部分混合模型的性能超过了纯 CNN 和纯 Transformer 模型。在参数量和计算量方面，混合模型展示了不同的权衡。其中大部分模型在保持较高准确率的同时，成功减少了参数数量和计算量。表 2 展示了几种模型在 COCO 数据集上的目标检测任务的混合模型的性能对比。表 3 展示了在 ISIC 数据集上进行医学图像分割任务的混合模型的性能对比。相比于 CNN 经典模型 UNet，混合模型 UNeXt 在参数量和计算量方面表现出明显优势，同时在 F1 指标上达到了更高分数。

Table 1. Performance comparison of hybrid models on the ImageNet-1K dataset
表 1. 混合模型在 ImageNet-1K 数据集上的性能对比

任务	模型	结构	#Params(M)	Flops(G)	Top-1 Acc (%)
图像分类	ResNet-101 [3]	CNN	45	7.9	79.8
	ViT-B [20]	Transformer	86.6	17.6	77.9
	PVT-S [25]	架构参考	24.5	3.8	79.8
	CSWin-S [29]	架构参考	35	6.9	83.6
	HRT-B [30]	架构参考	50.3	13.7	82.8
	DeiT-B [35]	知识蒸馏	86	17.5	81.8
	CoAtNet-0 [37]	串联拼接	25	4.2	81.6
	ConTNet-B [40]	串联拼接	39.6	6.4	81.8
	MobileViT-S [41]	串联拼接	5.6	-	78.4
	MobileViTV2-2.0 [42]	串联拼接	10.6	4	80.4
	ConFormer-S [43]	并联拼接	37.7	10.6	83.4
	Mobile-Former-508M [44]	并联拼接	14	0.508	79.3
	ViTc-1GF [47]	嵌入块替换	17.8	4	79.1
	CCT [48]	嵌入块替换	22.36	11.06	80.67
	LocalViT-S [49]	前馈层替换	22.4	4.6	80.8
	ConViT-S [50]	自注意力层替换	27	5.4	81.3
	BoTNet-S1-50 [51]	自注意力层替换	20.8	8.54	84.7
	LeViT-384 [52]	架构参考 + 局部替换	39.1	2.353	82.6
	CvT-21 [53]	架构参考 + 局部替换	32	7.1	82.5
	CeiT-S [54]	嵌入块替换 + 前馈层替换	24.2	4.5	82
	Edgevits-S [56]	架构参考 + 局部替换	11.1	1.9	81
	CMT-S [57]	架构参考 + 局部替换	25.1	4	83.5
	ParC-Net-S [58]	架构参考 + 局部替换	5	3.5	78.6
	Next-ViT-S [60]	局部替换 + 创新混合策略	31.7	5.8	82.5
	EdgeNeXt-S [61]	架构参考 + 局部替换	5.6	1.93	78.8
	Swin-ACmix-S [62]	创新混合模块	51	9	83.5

Table 2. Performance comparison of hybrid models on the COCO dataset
表 2. 混合模型在 COCO 数据集上的性能对比

任务	模型	结构	#Params(M)	Flops(G)	AP(%)
目标检测	Faster R-CNN [39]	CNN	60	150	35
	RetinaNet [63]	CNN	36	100	36
	DETR [36]	串联拼接	41	86	42
	ViT-B/16-FRCNN [38]	串联拼接	-	-	36.6

Table 3. Performance comparison of hybrid models on the ISIC dataset
表 3. 混合模型在 ISIC 数据集上的性能对比

任务	模型	结构	#Params(M)	Flops(G)	数据集	指标	数值(%)
医学图像分割	UNet [28]	CNN	31.13	55.84	ISIC	F1	84.03 ± 0.87
	UNeXt [26]	架构参考	1.47	0.57	ISIC	F1	89.70 ± 0.96

基金项目

国家自然科学基金(项目号：61975125)。

参考文献

[1] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[3] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[4] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[5] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231

[6] Xie, S., Girshick, R., Dollár, P., Tu, Z.W. and He, K.M. (2017) Aggregated Residual Transformations for Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5987-5995. https://doi.org/10.1109/CVPR.2017.634

[7] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/CVPR.2017.243

[8] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[9] Tan, M. and Le, Q. (2019) Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. http://proceedings.mlr.press/v97/tan19a.html?ref=jina-ai-gmbh.ghost.io

[10] Liu, Z., Mao, H., Wu, C.Y., *et al*. (2022) A ConvNet for the 2020s. *Proceedings of the* 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 11966-11976. https://doi.org/10.1109/CVPR52688.2022.01167

[11] Iandola, F.N., Han, S., Moskewicz, M.W., *et al*. (2016) SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and < 0.5 MB Model Size. arXiv: 1602.07360.

[12] Howard, A.G., Zhu, M., Chen, B., *et al*. (2017) Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861.

[13] Sandler, M., Howard, A., Zhu, M.L., Zhmoginov, A. and Chen, L.C. (2018) Mobilenetv2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the* 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[14] Howard, A., Sandler, M., Chu, G., *et al*. (2019) Searching for MobileNetV3. *Proceedings of the* 2019 *IEEE*/*CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 1314-1324. https://doi.org/10.1109/ICCV.2019.00140

[15] Han, K., Wang, Y., Tian, Q., *et al*. (2020) Ghostnet: More Features from Cheap Operations. *Proceedings of the* 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 1577-1586. https://doi.org/10.1109/CVPR42600.2020.00165

[16] Tang, Y., Han, K., Guo, J., *et al*. (2022) GhostNetV2: Enhance Cheap Operation with Long-Range Attention. arXiv:

2211.12905.

[17] Zhang, X.Y., Zhou, X.Y., Lin, M.X. and Sun, J. (2018) ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6848-6856. https://doi.org/10.1109/CVPR.2018.00716

[18] Ma, N., Zhang, X., Zheng, H.T. and Sun, J. (2018) ShuffleNet v2: Practical Guidelines for Efficient CNN Architecture Design. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, *European Conference on Computer Vision*, Springer, Cham, 122-138. https://doi.org/10.1007/978-3-030-01264-9_8

[19] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017.

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[21] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[22] Liu, Z., Hu, H., Lin, Y., *et al*. (2022) Swin Transformer v2: Scaling up Capacity and Resolution. *Proceedings of the* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 11999-12009. https://doi.org/10.1109/CVPR52688.2022.01170

[23] Han, K., Wang, Y., Chen, H., *et al*. (2022) A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 87-110. https://doi.org/10.1109/TPAMI.2022.3152247

[24] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[25] Wang, W., Xie, E., Li, X., *et al*. (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 548-558. https://doi.org/10.1109/ICCV48922.2021.00061

[26] Valanarasu, J.M.J. and Patel, V.M. (2022) UNeXt: Mlp-Based Rapid Medical Image Segmentation Network. 25*th International Conference of Medical Image Computing and Computer Assisted Intervention—MICCAI* 2022, Singapore, 18-22 September2022, 23-33. https://doi.org/10.1007/978-3-031-16443-9_3

[27] Wang, Z., Cun, X., Bao, J., *et al*. (2022) Uformer: A General U-Shaped Transformer for Image Restoration. *Proceedings of the* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 17662-17672. https://doi.org/10.1109/CVPR52688.2022.01716

[28] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. 18*th International Conference of Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[29] Dong, X., Bao, J., Chen, D., *et al*. (2022) Cswin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *Proceedings of the* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 12114-12124. https://doi.org/10.1109/CVPR52688.2022.01181

[30] Yuan, Y., Fu, R., Huang, L., *et al*. (2021) HRFormer: High-Resolution Transformer for Dense Prediction. arXiv: 2110.09408.

[31] Sun, K., Xiao, B., Liu, D. and Wang, J.D. (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 5686-5696. https://doi.org/10.1109/CVPR.2019.00584

[32] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv: 1503.02531.

[33] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.

[34] 黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624-653.

[35] Touvron, H., Cord, M., Douze, M., et al. (2021) Training Data-Efficient Image Transformers & Distillation through Attention. *Proceedings of the* 38*th International Conference on Machine Learning*, Virtual Event, 18-24 July 2021, 10347-10357.

[36] Carion, N., Massa, F., Synnaeve, G., *et al*. (2020) End-to-End Object Detection with Transformers. 16*th European Conference of Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[37] Dai, Z., Liu, H., Le, Q.V. and Tan, M.X. (2021) CoAtNet: Marrying Convolution and Attention for All Data Sizes. 35*th Conference on Neural Information Processing Systems*, Virtual, 6-14 December 2021, 3965-3977.

[38] Beal, J., Kim, E., Tzen, E., *et al*. (2020) Toward Transformer-Based Object Detection. arXiv: 2012.09958.

[39] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv: 1506.01497.

[40] Yan, H., Li, Z., Li, W., *et al*. (2021) Contnet: Why Not Use Convolution and Transformer at the Same Time? arXiv: 2104.13497.

[41] Mehta, S. and Rastegari, M. (2021) MobileVit: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. arXiv: 2110.02178.

[42] Mehta, S. and Rastegari, M. (2022) Separable Self-Attention for Mobile Vision Transformers. arXiv: 2206.02680.

[43] Peng, Z., Huang, W., Gu, S., *et al*. (2021) Conformer: Local Features Coupling Global Representations for Visual Recognition. *Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 357-366. https://doi.org/10.1109/ICCV48922.2021.00042

[44] Chen, Y., Dai, X., Chen, D., *et al*. (2022) Mobile-Former: Bridging Mobilenet and Transformer. *Proceedings of the* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 5260-5269. https://doi.org/10.1109/CVPR52688.2022.00520

[45] Yoo, J., Kim, T., Lee, S., *et al*. (2023) Rich CNN-Transformer Feature Aggregation Networks for Super-Resolution. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 2-7 January 2023, 4945-4954. https://doi.org/10.1109/WACV56688.2023.00493

[46] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B. and Fu, Y. (2018) Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, Cham, 294-310. https://doi.org/10.1007/978-3-030-01234-2_18

[47] Xiao, T., Singh, M., Mintun, E., *et al*. (2021) Early Convolutions Help Transformers See Better. arXiv: 2106.14881.

[48] Hassani, A., Walton, S., Shah, N., *et al*. (2021) Escaping the Big Data Paradigm with Compact Transformers. arXiv: 2104.05704.

[49] Li, Y.W., Zhang, K., Cao, J.Z., Timofte, R. and Van Gool, L. (2021) LocalViT: Bringing Locality to Vision Transformers. arXiv: 2104.05707.

[50] D'Ascoli, S., Touvron, H., Leavitt, M.L., *et al*. (2021) ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. arXiv: 2103.10697.

[51] Srinivas, A., Lin, T.Y., Parmar, N., *et al*. (2021) Bottleneck Transformers for Visual Recognition. *Proceedings of the* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 16514-16524. https://doi.org/10.1109/CVPR46437.2021.01625

[52] Graham, B., El-Nouby, A., Touvron, H., *et al*. (2021) LeViT: A Vision Transformer in Convnet's Clothing for Faster Inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 12239-12249. https://doi.org/10.1109/ICCV48922.2021.01204

[53] Wu, H., Xiao, B., Codella, N., *et al*. (2021) CvT: Introducing Convolutions to Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 22-31. https://doi.org/10.1109/ICCV48922.2021.00009

[54] Yuan, K., Guo, S., Liu, Z., *et al*. (2021) Incorporating Convolution Designs into Visual Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 559-568. https://doi.org/10.1109/ICCV48922.2021.00062

[55] Jeevan, P. (2022) Convolutional Xformers for Vision. arXiv: 2201.10271.

[56] Pan, J., Bulat, A., Tan, F., *et al*. (2022) EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision Transformers. 17*th European Conference of Computer Vision—ECCV* 2022, Tel Aviv, 23-27 October 2022, 294-311. https://doi.org/10.1007/978-3-031-20083-0_18

[57] Guo, J., Han, K., Wu, H., *et al*. (2022) CMT: Convolutional Neural Networks Meet Vision Transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 12165-12175. https://doi.org/10.1109/CVPR52688.2022.01186

[58] Zhang, H., Hu, W. and Wang. X. (2022) ParC-Net: Position Aware Circular Convolution with Merits from Convnets and Transformer. 17*th European Conference of Computer Vision—ECCV* 2022, Tel Aviv, 23-27 October 2022, 613-630. https://doi.org/10.1007/978-3-031-19809-0_35

[59] Yu, W., Luo, M., Zhou, P., *et al*. (2022) Metaformer Is Actually What You Need for Vision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 10809-10819. https://doi.org/10.1109/CVPR52688.2022.01055

[60] Li, J., Xia, X., Li, W., *et al*. (2022) Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios. ArXiv: 2207.05501.

[61] Maaz, M., Shaker, A., Cholakkal, H., *et al*. (2023) Edgenext: Efficiently Amalgamated CNN-Transformer Architecture

for Mobile Vision Applications. *Computer Vision—ECCV* 2022 *Workshops*, Tel Aviv, 23-27 October 2022, 3-20. https://doi.org/10.1007/978-3-031-25082-8_1

[62] Pan, X., Ge, C., Lu, R., *et al*. (2022) On the Integration of Self-Attention and Convolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 805-815. https://doi.org/10.1109/CVPR52688.2022.00089

[63] Lin, T.Y., Goyal, P., Girshick, R., He, K.M. and Dollár, P. (2017) Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324