Iwin Transformer: 使用交错窗口的分层视觉 Transformer

Simin Huo, Ning Li

***Abstract*—我们介绍了 Iwin Transformer，这是一种新颖的无位置嵌入的分层视觉 Transformer。通过创新的交错窗口注意力和深度可分离卷积的协作，该模型可以直接从低分辨率微调到高分辨率。该方法使用注意力连接远距离的 token，并应用卷积连接相邻的 token，从而在单个模块内实现全局信息交换，克服了 Swin Transformer 需要两个连续块来逼近全局注意力的限制。在视觉基准测试上的大量实验表明，Iwin Transformer 在图像分类（在 ImageNet-1K 上实现 87.4% 的 top-1 准确率）、语义分割和视频动作识别等任务中表现出强大的竞争力。我们还验证了 Iwin 中的核心组件作为一个独立模块的有效性，它可以无缝替换类条件图像生成中的自注意力模块。Iwin Transformer 引入的概念和方法有可能激发未来的研究，例如视频生成中的 Iwin 3D 注意力。代码和模型可在 https://github.com/cominder/Iwin-Transformer. 找到。**

***Index Terms*—Iwin Transformer, interleaved window attention, position-embedding-free.**

I. 介绍

V 识别 变压器 (ViTs) [1] 基本上通过借用自然语言模型中的变压器架构 [2] 改变了计算机视觉。与卷积神经网络 (CNN) [3] 依赖于局部感受野捕捉图像特征不同，ViTs 利用自注意机制获取全局依赖性，在视觉任务中展现了卓越的表现。然而，其相对于输入序列长度 N 的二次计算复杂度 $\mathcal{O}(N^2)$ 带来了显著的可扩展性挑战，特别是在计算机视觉中日益普遍的高分辨率图像处理应用中。

为了应对视觉 Transformers（ViTs）中存在的二次复杂性挑战，并在保持性能的同时提高其效率，已经提出了多种方法。层次结构设计，例如 PVT [4] 和 Twins [5]，利用多尺度特征金字塔逐步减少空间维度。混合卷积神经网络-Transformer 架构比如 ConViT [6] 和 CoAtNet [7] 结合卷积操作与自注意机制，以发挥两种模式的优势。有效的 Token 融合策略，例如 TokenLearner [8]，动态聚合 tokens 以减少序列长度，而像 Reformer [9] 示例的稀疏注意模式则利用局部敏感哈希仅关注相关 tokens。此外，像 Performer [10] 这样高效的实现通过核方法逼近注意力，实现线性复杂度。采用多种策略来缓解视觉 transformers 的计算需求。

应对这些挑战的最有前途的方法之一是 Swin Transformer [11]，它引入了一种具有偏移窗口自注意力的分层架构。通过将注意力计算限制在局部窗口内，并通过窗口偏移机制实现跨窗口连接，Swin Transformer 成功地将相对于图像大小的二次复杂度降低为线性复杂度。这个优雅的设计保持了模型捕获局部和全局依赖关系的能力，同时显著提高了计算效率。此外，Swin Transformer 采用了逐级合并更深层图像块的分层结构，生成类似于传统 CNN 骨干网络的多尺度特征图，这促进了其在各种视觉任务（包括目标检测和语义分割）中的应用。Swin Transformer 在各个基准上的出色性能确立了其作为高效视觉 Transformer 设计的一个里程碑，并展示了基于窗口的注意力机制在大规模视觉应用中的可行性。

尽管其设计具有开创性并且表现出色，Swin Transformer 仍存在一些显著的局限性。首先，由于注意力计算中所需的复杂掩码操作，移窗机制引入了额外的计算开销，这使得实现更为复杂并降低了硬件效率。其次，Swin 的架构需要两个连续的 Transformer 模块：一个带有常规窗口，另一个带有移窗，以实现全局信息交换，这导致了计算冗余，因为某些特征被多次处理。在生成内容（AIGC）的时代，这种双模块要求特别具有挑战性，其中需要将诸如文本提示之类的条件信息注入模型中；在这种严格的双模块结构中，文本和图像之间跨注意力的最佳位置并不明显，这解释了 Swin 在现代文本到图像扩散模型中有限的采纳。此外，如 Swin Transformer v2 [12] 所承认的，当模型在高分辨率输入上进行微调时，面临可扩展性问题。对于较大的窗口，相对位置编码的双三次插值会导致显著的性能下降，进而需要引入复杂的替代方案，如对数间隔连续位置偏移（Log-CPB）。这种对复杂位置编码方案的依赖最终阻碍了模型的扩展能力和更广泛的适用性。

为了在保持基于窗口注意力的计算效率的同时克服这些限制，我们引入了交错窗口 Transformer（Iwin Transformer）。Iwin 的关键创新在于结合了深度可分离卷积及交错窗口机制，在应用窗口注意力之前重排特征，使得每个窗口包含来自图像不同区域的像素。这种巧妙的方法使得在单个 Transformer 块中实现全局信息的交互，而无需像 Swin 那样复杂的掩码操作。此外，卷积引入了对视觉任务有益的归纳偏置，并提供了隐式的位置信息。这种混合方法不仅增强了特征表示，还减少了对显式位置编码的依赖，解决了 Swin 的一个关键限制。结合的设计使得 Iwin 能够以大约一半的计算成本实现两个连续 Swin 块的等效全局感受野，这对于高分辨率视觉应用特别有利，并且更易于与生成模型中的文本条件机制集成。

本文的主要贡献总结如下：

1) 交错窗口注意力：我们提出了一种新颖的重塑-转置-重塑（RTR）操作，该操作系统地将特征序列重新排序为交错模式，以应用窗口自注意力，然后恢复原始空间排列。该机制实现了线性复杂度。
2) 混合注意力卷积模块：我们巧妙地将深度可分离卷积与交错窗口注意力结合在一起，创建了一个计算效率高的模块，充分利用了这两种机制的互补优势。
3) 理论分析：我们提供数学证明，Iwin 通过混合注意力-卷积模块实现全局信息交换。
4) 无需位置嵌入：Iwin Transformer 不需要显式的位置编码，保证了其在不降低性能的情况下能在不同输入分辨率上具有强大的扩展性，从而克服了之前 transformer 架构中的一个关键限制。
5) 全面的实证验证：我们提供了大量实验证据表明，Iwin 在包括图像分类、语义分割和视频识别等各种

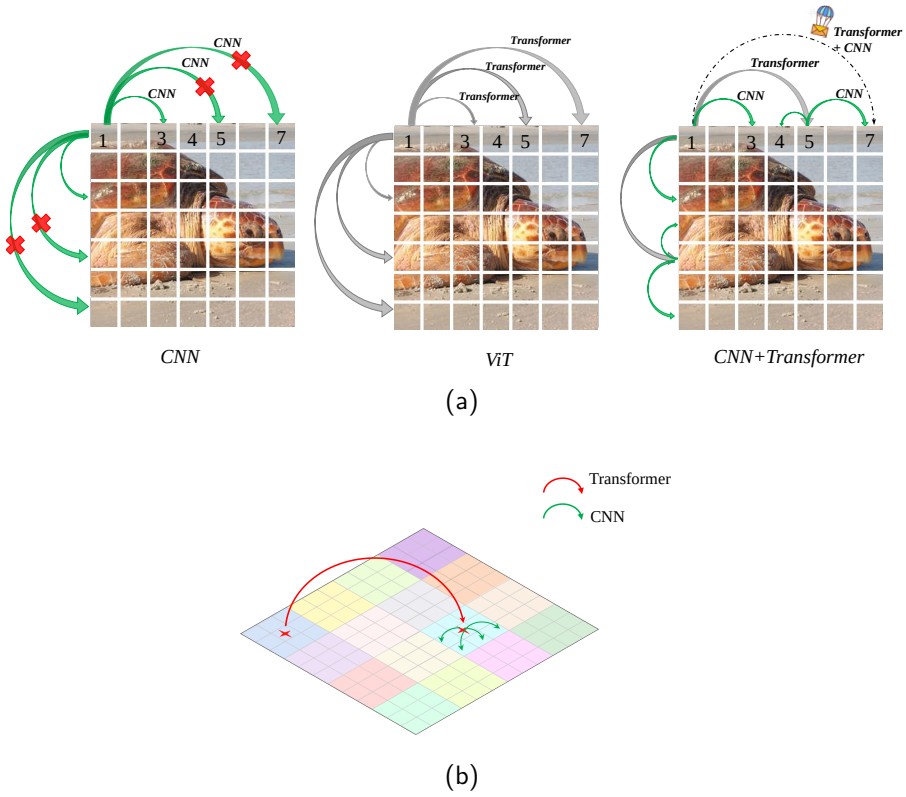


Fig. 1. 所提出模式的图示。(a) 中，CNN 内的标记 1 只能与附近的标记 3 交互，无法远距离接触标记 7。因此，CNN 只能捕捉局部特征。相比之下，ViT 中的标记 1 可以与任何标记关联，从而能够捕捉全局特征，但复杂度是二次的。在第三种提出的 CNN+Transformer 模式中，标记 1 首先通过注意力机制与短距离内的标记 5 连接，而标记 5 通过卷积与标记 7 相关联。就这样，尽管标记 1 和标记 7 相距甚远，仍能间接通信。(b) 展示了所提出的 CNN+Transformer 模式的直观俯视图。

视觉任务中保持或提升了 Swin 的性能，证明了其作为视觉骨干网的有效性。

6) 可扩展性到其他领域：在我们的讨论中，我们提供了将 Iwin 的交错窗口注意力扩展到一维用于大型语言模型和三维用于视频生成的线索，提供了一种相对于传统的三维全注意力和时空注意力机制的第三种替代方案，具有潜在的计算效率益处。

受到 Vision Transformers（ViTs）成功的启发，Transformer 架构在计算机视觉研究中引起了极大关注。然而，它们都面临一个共同问题，即二次复杂性的高计算开销。为了在保持良好性能的同时提高 transformer 结构的效率，已经提出了多种方法。一些工作通过结合 CNN 和 Transformer 来利用这两种结构的优点。其他方法则专注于修改 ViTs 的结构以更好地适应视觉任务。以下小节简要回顾了这些相关工作，并按其方法进行分类。

A. 线性和稀疏注意力

自注意力操作在序列长度方面引入了二次计算复杂度，对高分辨率的视觉输入提出了重大挑战。Linformer [21] 通过注意力矩阵的低秩分解实现了线性复杂度，将 $N \times N$ 注意力矩阵分解为两个较小的矩阵。Performer [10] 引入了通过正交随机特征（FAVOR+）进行快速注意力，使用随机特征映射来近似注意力核。Luna [22] 通过引入一组固定长度的投影嵌入来作为注意力计算的中间表示，提出了线性统一嵌套注意力。BigBird [23] 结合了随机、窗口和全局注意力模式以保持线性复杂度。Longformer [24] 采用扩张滑动窗口注意力与选择的全局注意力标记。Sparse Transformer [25] 引入了通过结构化稀疏性降低复杂度的分解注意力模式。[26], [27] 选择性地为最相关的标记对进行注意力计算。这些方法共同表明，对于有效的视觉表示学习，完全的全局注意力通常是没有必要的，这使得在不显著降低性能的情况下实现更高效的 transformer 设计成为可能。

B. 层次化视觉 *Transformer*

分层视觉变换器采用多尺度特征表示，在提高计算效率的同时保持建模能力。金字塔视觉变换器（PVT）[4] 引入了一个逐步缩小的金字塔结构，通过空间-缩减注意力在更深的层中减少序列长度。Swin Transformer [11] 提出了一个分层架构，其中包含移位窗口，将自注意力计算限制在局部窗口，并通过层之间的窗口移动建立跨窗口连接。这种设计将计算复杂度从与图像大小相关的二次方减少到线性。MViT [20] 采用池化注意力，逐步扩大通道容量，同时降低空间分辨率。Twins 结合了局部分组的自注意力和全局子采样注意力，以高效平衡局部和全局交互。CSWin [28] 利用十字形窗口自注意力分别捕捉水平和垂直依赖性。这些分层设计在目标检测和语义分割等密集预测任务中特别有效，在这些任务中，多尺度特征表示是必不可少的。

C. 混合卷积神经网络-*Transformer* 架构

混合架构结合了卷积操作和自注意力，以利用这两种范式的优势。ConViT [6] 引入了一种门控位置自注意力机制，可以从卷积平滑过渡到 transformer 行为。CoAtNet [7] 在相对注意框架中统一了深度卷积和自注意力，在早期阶段

采用卷积，在后期阶段使用自注意力。MobileViT [19] 将卷积的局部处理与 transformer 的全局处理集成用于轻量模型。LocalViT [29] 通过深度卷积增强视觉 transformer，为自注意力层引入了局部性。这些混合方法通常可以在较小的数据集上比纯 transformer 架构实现更好的参数效率和性能，同时保持复杂视觉任务所必需的全局建模能力。

D. 动态计算策略

动态计算策略根据输入复杂度调整计算资源。DynamicViT [30] 引入了一种令牌稀疏化框架，该框架根据令牌的重要性评分逐步修剪冗余令牌，随着网络的加深减少序列长度。A-ViT [31] 采用自适应计算深度，允许不同的令牌根据其复杂度在不同的层退出网络。Token Merging (ToMe) [32] 在整个网络中动态合并相似的令牌，在减少序列长度的同时保留信息。Adaptive Token Sampling [33] 引入了可学习模块，这些模块根据输入采样重要令牌，在减少计算的同时保留关键信息。这些方法通过在最需要的地方分配计算资源，实现更高效的处理，使之特别适合于复杂度变化的实时应用。

E. 与先前工作的差异

与 Swin Transformer [11] 需要两个连续的块通过常规和移位窗口模式来建立跨窗口连接不同，Iwin Transformer 通过交错窗口注意力和深度可分离卷积的协同组合在单个块内实现全局信息交换。任何标记都可以通过中介连接与其他标记交互，类似于一个平面组织结构，个人可以直接联系他人而不需要通过多人。另一个好处是，交错窗口注意力和深度可分离卷积组成的模块可以无缝替代生成模型中的标准注意力模块，而不影响后续与文本条件的交叉注意力操作。由于 Swin 的双块依赖性，它无法做到这一点。

与早期的混合 CNN-Transformer 架构如 LocalViT [29] 和 MobileViT [19] 相比，这些架构通常只是将卷积用作捕获局部特征的动机。在 Iwin Transformer 中，深度可分离卷积和交错窗口注意力是相互依存的组件，形成了一个连贯的信息处理单元。深度可分离卷积建立了一些通过交错窗口注意力未能建立的令牌之间的连接。另一个优点是，由于卷积自然携带位置信息，Iwin 不再需要显式的位置信息编码。这使得 Iwin 成为一个无需位置嵌入的 transformer，允许在低分辨率上训练的模型在调整到高分辨率时轻松微调，同时保持性能。

总之，Iwin 是一种不依赖位置嵌入的 transformer，通过交织窗口注意力与深度可分离卷积的协同组合实现全局信息交换。

II. 方法论

A. 总体架构

图 7 展示了 Iwin Transformer 架构的概况。详细配置如表 III 所示。Iwin Transformer 遵循类似于 Swin Transformer [11] 的分层架构，在四个阶段中逐渐减少空间分辨率，同时扩大通道维数。对于给定的输入图像，Iwin 首先通过一个补丁分割模块将其划分为不重叠的补丁。每个补丁被视为一个标记或特征，然后架构在阶段 $\{S_1, S_2, S_3, S_4\}$ 中分别以分辨率 $\{H/4, H/8, H/16, H/32\}$ 和通道 $\{C, 2C, 4C, 8C\}$ 处理特征。尽管在本研究中使用了金字塔结构，Iwin 的核心模块也可用于平面结构中。

B. 交错窗口注意力

交错窗口注意力（IWA），如图 2 所示，是 Iwin Transformer 的核心创新。与标准窗口注意力不同，它将特征图均匀地划分为不重叠的窗口，IWA 在窗口划分之前重新排列特征图，使得来自不同区域的代币被分组到同一窗口中进行注意力计算。

对于一个特征图 $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$，其中 H 和 W 是空间维度，C 是通道维度，IWA 的操作如下：

1) 重新排列：输入特征图经过重新排列，使来自不同区域的标记被分组到同一个窗口内
2) 注意：标准的多头自注意力机制被应用于重新排列的标记
3) 还原：标记被恢复到它们原本的空间排列

假设窗口大小为 $M \times M$，我们首先将特征图划分为不重叠的 $M \times M$ 个窗口，每个窗口包含 $H_g \times W_g$ 个标记，其中 $H_g = H/M$ 和 $W_g = W/M$ 分别是沿高度和宽度方向的窗口数量。

1) 重排： 这种重新排列可以表示为：对于位置 (i, j) 的一个标记，其在重新排列的特征图中的新位置是：

$$\begin{aligned} i' &= (i \bmod H_g) \times M + \lfloor i/H_g \rfloor \\ j' &= (j \bmod W_g) \times M + \lfloor j/W_g \rfloor \end{aligned} \tag{1}$$

幸运的是，如图 2 和算法 1 中清晰展示的，我们可以通过 RTR（重塑-转置-重塑）操作优雅地实现这个过程。

随后，新特征图被平均分为 $H_g \times W_g$ 个不重叠的窗口，每个窗口包含 $M \times M$ 个标记。

对于最初位于位置 (i, j) 的标记，它将被分配到由 $(window\_row, window\_col)$ 表示的窗口：

$$\begin{aligned} window\_row &= \lfloor i'/M \rfloor = \lfloor ((i \bmod H_g) \times M + \lfloor i/H_g \rfloor)/M \rfloor \\ &= \lfloor (i \bmod H_g) + \lfloor i/H_g \rfloor /M \rfloor \\ &= i \bmod H_g \quad (\text{since } \lfloor i/H_g \rfloor /M < 1\ ) \end{aligned} \tag{2}$$

注意，$i \bmod H_g$ 的范围是从 0 到 $H_g - 1$，这正好对应于窗口网格中窗口的行索引。类似地，$window\_col = j \bmod W_g$ 的范围是从 0 到 $W_g - 1$，对应于窗口的列索引。

这确保具有相同 $(i \bmod H_g, j \bmod W_g)$ 的标记被分组到同一个窗口中进行注意力计算。这意味着，如果 (i_1, j_1) 和 (i_2, j_2) 处的两个标记在同一个窗口中，则它们必须满足：

$$i_1 \bmod H_g = i_2 \bmod H_g \text{ and } j_1 \bmod W_g = j_2 \bmod W_g \tag{3}$$

2) 自注意力： 在每个窗口中，我们应用标准的自注意力机制 [2]：

$$\begin{aligned} \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V \\ \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \end{aligned} \tag{4}$$

其中 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$ 是可学习的投影矩阵，而 d_k 是键向量的维度。

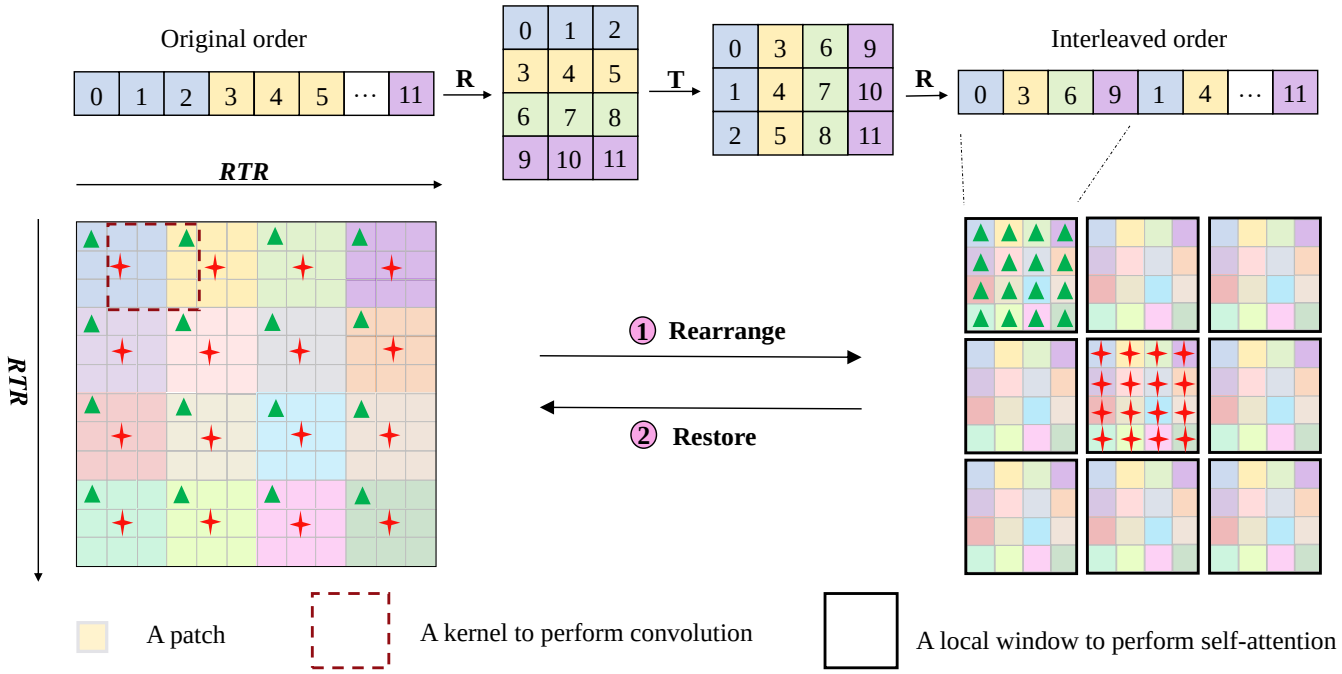


Fig. 2. 对 Iwin 注意力的示例。在左图中，绿色三角形和红色星星表示的 tokens 通过卷积在原始图像中连接。在右图中，所有表示 tokens 的绿色三角形通过 RTR（重构-转置-重构）操作和窗口分割被分配到相同的窗口中，执行窗口注意力以在它们之间建立连接。所有表示 tokens 的红色星星做同样的事情。结果是交错序列上的全局卷积和窗口注意力共同有效地近似标准的全局注意力，这意味着在原始图像中的任何 tokens 之间建立了连接。

*3) 恢复：*最后，使用逆 RTR 将标记恢复到其原始空间排列，这也是如图 2 和算法 1 所示的重塑-转置-重塑操作，以实现 $(i', j') \mapsto (i, j)$：

$$\begin{aligned} i &= (i' \bmod M) \times H_g + \lfloor i'/M \rfloor \\ j &= (j' \bmod M) \times W_g + \lfloor j'/M \rfloor \end{aligned} \tag{5}$$

Algorithm 1 重排和恢复操作的伪代码，以类似于 PyTorch 的风格。

```
# H_num_win: H // window_size, which means the number of windows
      along H
# W_num_win: w // window_size, which means the number of windows
      along W

def rearrange(x, H_num_win, W_num_win):
    B, H, W, C = x.shape

    x = x.reshape(B, -1, H_num_win, W, C).transpose(1, 2)
    x = x.reshape(B, -1, W, C)

    x = x.reshape(B, H, -1, W_num_win, C).transpose(2, 3)
    x = x.reshape(B, H, -1, C)
    return x


def restore(x, H_num_win, W_num_win):
    B, H, W, C = x.shape

    x = x.reshape(B, H, W_num_win, -1, C).transpose(2, 3)
    x = x.reshape(B, H, -1, C)

    x = x.reshape(B, H_num_win, -1, W, C).transpose(1, 2)
    x = x.reshape(B, -1, W, C)
    return x
```

C. 深度可分离卷积

深度可分离卷积（DWConv） [34] 用于帮助建立某些不在相同注意窗口中的标记的缺失关系，并顺便提供隐式位置信息。

D. 下采样层

Iwin Transformer 采用标准卷积逐步减少空间分辨率并增加通道维度，遵循视觉骨干网络中常见的分层设计原则 [35] 。

$$\mathcal{D}(\boldsymbol{X}) = \mathrm{Conv}_{3\times3,\mathrm{stride}=2}(\boldsymbol{X}) \tag{6}$$

我们测试了四种降采样方法，如表 VII 所示：平均池化、补丁合并、标准卷积和深度可分离卷积。它们的表现都非常好，相互之间的性能差异仅为 0.2 %。我们选择了准确率最高的标准卷积。

E. *Iwin Transformer* 块

如图 3a 所示，Iwin Transformer 块由一个统一模块组成，并行集成了交错窗口多头自注意力（IW-MSA）和深度可分离卷积（DWConv）模块，随后是一个两层 MLP，其中包含层间的 GELU 激活 [36] 。每个统一模块和 MLP 之前都有一个 LayerNorm 层，并在每个之后都有残差连接。Iwin Transformer 块的前向传播可以表示为：

$$\begin{aligned} \boldsymbol{X}' &= \mathrm{LayerNorm}(\boldsymbol{X}) \\ \boldsymbol{X}'' &= \boldsymbol{X} + \text{IW-MSA}(\boldsymbol{X}') + \mathrm{DWConv}(\boldsymbol{X}') \\ \boldsymbol{X}''' &= \boldsymbol{X}'' + \mathrm{MLP}(\mathrm{LayerNorm}(\boldsymbol{X}'')) \end{aligned} \tag{7}$$

组合 IW-MSA 和 DWConv 的统一模块的计算成本如下：

$$\mathcal{O}_{Iwin} = \underbrace{\frac{HW}{M^2} \times 3M^2C^2}_{\text{QKV projection}} + \underbrace{2\frac{HW}{M^2} \times M^4C}_{\text{Attention computation}} + \underbrace{\frac{HW}{M^2} \times M^2C^2}_{\text{Output projection}} + \underbrace{HW \times C \times k^2}_{\text{Convolution}} = 4HWC^2 + (2M^2 + k^2)HWC \tag{8}$$

与 Swin 相比

$$\mathcal{O}_{Swin} = 4HWC^2 + 2M^2HWC \tag{9}$$

尽管 Iwin 相较于 Swin 引入了额外的 k^2HWC 计算，但这是非常值得的。这使得 Iwin 的层配置可以是 { 2, 2, 7, 2 }，而 Swin 依赖于两个连续的块来接近全局注意力，只能将其深度从 { 2, 2, 6, 2 } 增加到 { 2, 2, 8, 2 }。因此，Iwin 拥有很大的灵活性。此外，当 $M >> k$，M 主导公式 8 时，Iwin 和 Swin 的计算复杂度几乎相同。另外，作为独立模块的统一模块可以无缝替换某些生成模型中的自注意力模块。

F. 架构变体

参照 Swin Transformer [11]，我们构建了 Iwin-T、Iwin-S、Iwin-B 和 Iwin-L，其网络深度和宽度与 Swin 相同，以便进行公平比较。这也得到了我们关于 Iwin-T 的消融研究的支持，该研究表明，层配置 $\{2, 2, 6, 2\}$ 实现了最高精度（见表 VII）。因此，后续的 Iwin-S、Iwin-B 和 Iwin-L 都遵循与 Swin 相同的设置。对于 224、384、512 和 1024 的输入分辨率，窗口大小分别为 7、12、16 和 16。表 I 显示了不同变体在 ImageNet 上的模型大小、计算复杂度（FLOPs）和性能。

G. 全球信息交流

我们认为，当特征图中存在一条信息可以从一个位置 (i_1, j_1) 流动到另一个位置 (i_2, j_2) 的路径时，就实现了全局信息交换。

Iwin Transformer 的一个关键理论特性是其能够以线性计算复杂度实现全局信息交换。我们通过研究特征图中任意两个位置之间的信息流来分析这一特性。

Lemma 1 (Modular Property of Interleaved Window Attention): 在交错窗口注意中，当且仅当满足以下条件时，位于位置 (i_1, j_1) 和 (i_2, j_2) 的标记才位于相同的注意窗口中：

$$i_1 \bmod H_g = i_2 \bmod H_g \text{ and } j_1 \bmod W_g = j_2 \bmod W_g$$

Lemma 2 (Locality of Depthwise Separable Convolution): 对于核大小为 $K \times K$ 的深度可分离卷积，当且仅当满足以下条件时，位置 (i_1, j_1) 和 (i_2, j_2) 处的标记可以直接交换信息：

$$|i_1 - i_2| \le K \text{ and } |j_1 - j_2| \le K$$

基于这些引理，我们证明了以下定理：

Theorem 3 (Global Information Exchange Condition): 如果 $KM \ge \max(H, W)$，其中 K 是核大小，M 是窗口大小，那么 Iwin Transformer 块中的交错窗口注意力和深度可分离卷积的组合可以实现在特征图中任意两个位置 (i_1, j_1) 和 (i_2, j_2) 之间的信息交换。

考虑特征图中的任意两个位置 (i_1, j_1) 和 (i_2, j_2)，我们需要证明存在一条信息流从 (i_1, j_1) 到 (i_2, j_2) 的路径。

我们讨论三种情况：

情况 1：$(i_1 \bmod H_g = i_2 \bmod H_g)$ 和 $(j_1 \bmod W_g = j_2 \bmod W_g)$

在这种情况下，根据引理 **??**，位置 (i_1, j_1) 和 (i_2, j_2) 位于相同的注意力窗口中，因此它们可以通过注意力机制直接交换信息。

情况 2：$(|i_1 - i_2| \le K \quad and \quad |j_1 - j_2| \le K)$

在这种情况下，根据引理 **??**，位置 (i_1, j_1) 和 (i_2, j_2) 处于同一个卷积核中，因此它们可以通过卷积机制直接交换信息。

情况 3：否则（即，当 (i_1, j_1) 和 (i_2, j_2) 不在同一注意力窗口和卷积核中时）

在这种情况下，我们需要找到一个中间位置 (i_3, j_3) 来连接 (i_1, j_1) 和 (i_2, j_2)。

我们如下构建这样一个位置 (i_3, j_3)：

$$\begin{aligned} i_3 &= (i_1 \bmod H_g) + H_g \cdot \lfloor i_2/H_g \rfloor \\ j_3 &= (j_1 \bmod W_g) + W_g \cdot \lfloor j_2/W_g \rfloor \end{aligned} \tag{10}$$

现在我们有

$$\begin{aligned} i_3 \bmod H_g &= ((i_1 \bmod H_g) + H_g \cdot \lfloor i_2/H_g \rfloor) \bmod H_g \\ &= (i_1 \bmod H_g) \bmod H_g + (H_g \cdot \lfloor i_2/H_g \rfloor) \bmod H_g \\ &= (i_1 \bmod H_g) + 0 \\ &= i_1 \bmod H_g \end{aligned} \tag{11}$$

同样，$j_3 \bmod W_g = j_1 \bmod W_g$。这意味着根据引理 **??**，位置 (i_1, j_1) 和 (i_3, j_3) 位于相同的注意窗口中。

现在我们检查：

类似地，$|j_2 - j_3| < W/M$。

当 $KM \ge \max(H, W)$ 时，我们有 $|i_2 - i_3| \le K$ 和 $|j_2 - j_3| \le K$，因此位置 (i_3, j_3) 和 (i_2, j_2) 可以通过深度可分卷积直接交换信息。

因此，当 $KM \ge \max(H, W)$ 时，Iwin Transformer 模块可以在特征图中的任意两个位置之间进行信息交换。总是存在 (i_3, j_3) 使得

$$\begin{aligned} i_1 \bmod H_g &= i_3 \bmod H_g \\ j_1 \bmod W_g &= j_3 \bmod W_g \\ |i_2 - i_3| &\le K \\ |j_2 - j_3| &\le K \end{aligned} \tag{12}$$

这意味着位置 (i_1, j_1) 和 (i_3, j_3) 通过交错窗口注意力相连，而位置 (i_3, j_3) 和 (i_2, j_2) 通过深度可分离卷积相连，(i_1, j_1) 和 (i_2, j_2) 通过 (i_3, j_3) 作为中介桥梁建立了连接。

起初，我们遵循规则 $KM \ge \max(H, W)$，为每个阶段分配不同的卷积核大小。然而，在 Iwin-T 上的消融实验（见表 VII）显示，对于阶段 1、2 和 3 分别使用大小为 7、5 和 3 的卷积核会导致最差的性能相比于在各阶段使用一致的卷积核大小为 7、5 或 3，这表明一致的卷积核大小可带来更快的训练速度和更好的优化效果。这与 [37] 中的观察一致，即平衡的网络优于理论上最优但不平衡的配置。我们认为，随着网络的加深和下采样的增加，会增大有效感受野（ERF）[38]，从而导致 $K_{ERF} \cdot M \ge \max(H, W)$。

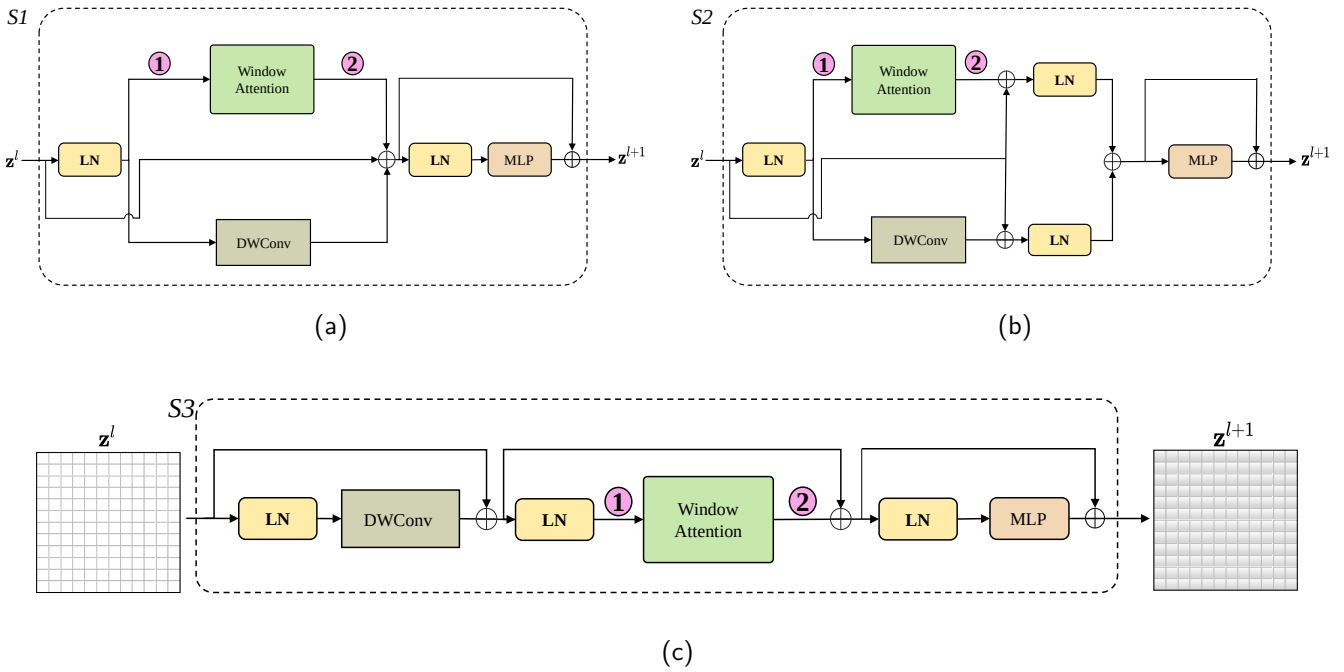


Fig. 3. Iwin 区块示意图。(a) S1 展示了一种并行结构，其中卷积和注意力结果直接结合，正如在本研究中实施的那样（最快）。(b) S2 是一种并行方案，具有对输入的独立卷积和注意力连接，表现最差。(c) S3 是串行配置，其中注意力输入接收卷积输出，表现略优于 S1，但需要额外进行一层归一化，增加计算量。

因此，我们相信如果网络足够深，在经过足够多的连续 Iwin Transformer 块后，最初较小的核大小可以扩展到足够大，以至于在某个深度及更深处，模型能够看到整个世界。

III. 实验

我们在 ImageNet-1K 图像分类 [39]、COCO 目标检测 [40]、ADE20K 语义分割 [41]、Kinetics-400 [42] 视频识别以及类条件图像生成上进行实验。接下来，我们首先将提出的 Iwin Transformer 与之前的最新技术进行比较。然后，我们对 Iwin Transformer 的重要设计元素进行消融研究。

A. *ImageNet-1K* 上的图像分类

对于图像分类任务，我们在 ImageNet-1K [39] 上对提出的 Iwin Transformer 进行了基准测试，该数据集中包含来自 1000 个类别的 128 万张训练图像和 5 万张验证图像。实验设置严格遵循 Swin Transformer [11] 的设置。我们在两种场景下报告了单次裁剪的 top-1 准确率：从头开始在 ImageNet-1K 上训练，以及在包含 1420 万张图像和 22000 个类别的 ImageNet-22K 上进行预训练，然后在 ImageNet-1K 上进行微调。

从 ImageNet-1K 数据集开始训练。我们使用 AdamW 优化器 [43] 进行 300 个周期训练，其中包含 20 个周期的线性预热以及余弦学习率衰减。我们将批量大小设为 512，权重衰减为 0.05，并使用 DeiT [44] 中的大部分增强策略。初始学习率设定为 0.0005，且 Iwin-T、Iwin-S 和 Iwin-B 的 drop path 率分别为 0.2、0.3 和 0.5。

在 ImageNet-22K 上进行预训练，训练持续 90 个周期，并有 5 个周期的预热。我们使用 4096 的批大小，初始学习率为 $1.25e-4$，权重衰减为 0.05。然后对预训练模型在 ImageNet-1K 上进行微调 10 到 30 个周期，使用分辨率为 224×224 的图像，批大小为 1024，恒定学习率为 $2e-05$，权重衰减为 $1e-8$。

跨分辨率微调。Iwin Transformer 的一个关键优势是它可以简单地转移到更高的分辨率。与需要分阶段微调或插值绝对/相对位置偏差的传统方法不同，我们在 224^2 分辨率上预训练的模型可以直接在更高的分辨率（如 384^2、512^2 和 1024^2）上进行微调。唯一需要做的就是改变超参数窗口大小。

在这种跨分辨率微调过程中，我们调整窗口大小以匹配输入分辨率——224 对应 7，384 对应 12，512 和 1024 对应 16——同时保持架构不变。微调过程运行 30 个周期，使用恒定的学习率 $2e-05$，权重衰减为 $1e-8$，以及 5 到 10 个周期的预热。

a) 结果： 表格 I 展示了 Iwin Transformer 的竞争性能和分辨率适应性综合结果。

ImageNet-1K。对于在分辨率为 224×224 的 ImageNet-1K 上从头开始训练的模型，Iwin Transformer 变体始终显示出具有竞争力的 Top-1 准确率。Iwin-T 达到 82.0 %，匹敌 ConvNeXt-T（82.1 %），并以 0.7 % 的优势超过 Swin-T（81.3 %）。更深的变体 Iwin-S 取得最高 83.4 %，超越 Swin-S（83.0 %）0.4 % 和 ConvNeXt-S（83.1 %）0.3 %，同时在参数（51.6M 对 50.0M）和 FLOPs（9.0G 对 8.7G）方面与 Swin-T 和 ConvNeXt-T 相似。通过增加嵌入维度，Iwin-B 达到 83.5 %，接近 PVT-v2-B4 的 83.6 %，并位于 Swin-B 的 83.3 % 和 ConvNeXt-B 的 83.8 % 之间。Iwin 模型（例如，Iwin-T 为 729 img/s）的吞吐量略低于 Swin 和 ConvNeXt（例如，Swin-T 为 758 img/s 和 ConvNeXt-T 为 775 img/s）。

ImageNet-22K。通过利用大规模的 ImageNet-22K 预训

练，Iwin 显著提升了其性能。预训练的 Iwin-B 在分辨率为 224 × 224 时，实现了令人印象深刻的 85.5 % 的 Top-1 准确率。当微调到分辨率为 384 × 384 时，它达到了坚实的 86.6 %，超越了 Swin-B (86.4 %) 0.2 %，并与精心配置的 ConvNeXt-B (86.8 %) 紧密比肩。最大的变体 Iwin-L 在 224 × 224 时展示了尖端的 86.4 % 的性能，并在 384 × 384 时达到了令人印象深刻的 87.4 %，与顶级模型如 Swin-L (87.3 %) 和 ConvNeXt-L (87.5 %) 竞争得不相上下。

跨分辨率微调。Iwin 模型具有直接从 224^2 预训练状态微调到更高分辨率的能力，这一特性在从头开始在 ImageNet-1K 上训练的模型中特别显著。例如，Iwin-S 在其初始训练于 224^2 后，可以微调以在 384^2 达到令人印象深刻的 84.3 % 的 Top-1 准确率（相比其 224^2 基线 83.4 % 提高了 0.9 %），并在 512^2 达到 84.4 %（提升 1.0 %）。同样地，Iwin-B 在 384^2 达到 84.9 %（相比其 224^2 基线 83.5 % 提高了 1.4 %），超越了 Swin-B (84.5%) 在 384^2 的 0.4 %，并且紧密接近 ConvNeXt-B (85.1%)。此外，Iwin-B 在更高分辨率下表现出强劲性能，在 512^2 达到 85.1 %（提升 1.6 %），甚至在 1024^2 达到显著的 85.0 %（提升 1.5 %）。

对于在 ImageNet-22K 上预训练的模型，这种优势同样明显，展示了 Iwin 的无缝适应性。Iwin-B 从 224^2 微调到 384^2 (86.6 % ，相较于其 224^2 基线的 85.5 % 提升了 1.1 %)，在 384^2 上直接超越了 Swin-B (86.4%) 0.2 % 。它在 512^2 (86.1 %) 和 1024^2 (85.6 %) 上继续表现强劲。同时，Iwin-L 从 224^2 无缝过渡到 384^2 (87.4 % ，提升了 1.0 %)，与 Swin-L (87.3%) 和 ConvNeXt-L (87.5%) 表现齐平。

这种高分辨率的微调能力归功于 Iwin 的交错窗口注意力与深度可分离卷积的协作，该协作使模型能够在不依赖位置编码的情况下获得全局视野。这种在高分辨率下的强大性能以及协作在像高分辨率图像和视频生成等模型中取代标准注意力的潜力，比起与 Swin 相比在吞吐量上的小差距要更加重要。

B. 在 COCO 上的目标检测

我们遵循 Swin [11] 的设置，并使用 MMDetection [54] 工具箱在 COCO [40] 上评估 Iwin 主干网络以进行目标检测和实例分割任务，使用 Mask R-CNN [52] 和级联 Mask R-CNN [53] 。我们采用多尺度训练、AdamW 优化器和在 ImageNet-1K 上预训练的模型。

如表中所示，Iwin Transformer 在各种框架和训练计划中始终表现不如 Swin Transformer。在 Mask-RCNN 3 × 计划中，Swin-T 的边界框 AP 达到 46.0，超过了 Iwin-T 的 44.7。当使用具有 3 × 计划的 Cascade Mask-RCNN 时，Swin-T 达到 50.4 AP ^{box} ，比 Iwin-S 达到的 49.4AP^{box} 高 1.0 AP。值得注意的是，在这种设置下，Iwin-S 的性能相较于 Iwin-T 并未显示出任何提升。

为了找出差距的原因，我们通过在每个验证周期绘制 (AP ^{box}) 比较了 Iwin-T 和 Swin-T 在 Cascade Mask-RCNN 3 × 计划下的表现。这两个模型都遵循相同的学习率计划，带有分步衰减。

最初，Iwin-T 展现出具有竞争力的性能，与 Swin-T 表现接近，并在前 27 个 epoch 中偶尔超过它。然而，在第 28 个 epoch 时，两者出现了显著的差异，这正好与第一个主要学习率衰减同时发生。Swin-T 的 AP 从 45.9 跃升至 49.2，而 Iwin-T 仅显示出适度的提升（从 46.2 提高到 48.7），此后一直落后。学习率的下降似乎对 Swin 更有利。

为了弥补这一性能差距，我们为 Iwin 模型探索了几种替代的训练配置，包括不同的学习率策略（例如，余弦退火）和具有相对位置编码的架构增强（参见消融实验表 X）。尽管进行了这些努力，我们仍无法在 COCO 目标检测任务中匹配或超过 Swin 的性能，这是 Iwin Transformer 唯一没有超越 Swin 的基准，表明这是一个特定任务的优化挑战，而不是一般架构上的缺陷。这个任务留待未来研究解决。

TABLE I
IMAGENET-1K 上分类准确率。吞吐量基于 PYTORCH 框架和 V100 GPU 进行测试

Method	Image Size (px)	Param (M)	FLOPs (G)	Throughput (img/s)	Top-1 Acc (%)
(a) ImageNet-1K trained models					
DeiT-Small/16 [44]	224^2	22.0	4.6	406	79.9
T2T-ViT-14 [45]	224^2	22.0	5.2	-	81.5
PVT-Small [46]	224^2	24.5	3.8	794	79.8
Twins-S [5]	224^2	24.0	2.9	979	81.7
PVT-v2-B2 [47]	224^2	25.4	4.0	664	82.0
PoolFormer-S36 [48]	224^2	31.0	5.1	764	81.4
Swin-T [11]	224^2	29.0	4.5	758	81.3
ConvNeXt-T [49]	224^2	29	4.5	775	82.1
Iwin-T(ours)	224^2	30.2	4.7	729	**82.0**
T2T-ViT-19 [45]	224^2	39.2	8.9	-	81.9
PVT-Medium [46]	224^2	44.2	6.7	511	81.2
Twins-B [5]	224^2	56.0	8.6	433	83.2
PVT-v2-B3 [47]	224^2	45.2	6.9	443	83.2
PoolFormer-M36 [48]	224^2	56.0	9.0	494	82.1
Swin-S [11]	224^2	50.0	8.7	437	83.0
ConvNeXt-S [49]	224^2	50	8.7	447	83.1
Iwin-S(ours)	224^2	51.6	9.0	410	**83.4**
Iwin-S(ours)	384^2	51.6	27.7	142	**84.3**
Iwin-S(ours)	512^2	51.6	52.0	78	**84.4**
Iwin-S(ours)	1024^2	51.6	207.9	20	**83.8**
DeiT-Base/16 [44]	224^2	86.6	17.6	273	81.8
T2T-ViT-24 [45]	224^2	64.1	14.1	-	82.3
PVT-Large [46]	224^2	61.4	9.8	357	81.7
Twins-L [5]	224^2	99.2	15.1	271	83.7
PVT-v2-B4 [47]	224^2	62.6	10.1	298	83.6
PoolFormer-M48 [48]	224^2	73.0	11.8	337	82.5
Swin-B [11]	224^2	88.0	15.4	287	83.3
Swin-B [11]	384^2	88.0	47.0	85	84.5
ConvNeXt-B [49]	224^2	89	15.4	292	83.8
ConvNeXt-B [49]	384^2	89	45.0	96	85.1
Iwin-B(ours)	224^2	91.2	15.9	271	**83.5**
Iwin-B(ours)	384^2	91.2	48.3	78	**84.9**
Iwin-B(ours)	512^2	91.3	89.5	51	**85.1**
Iwin-B(ours)	1024^2	91.3	358.2	12	**85.0**
(b) ImageNet-22K pre-trained models					
R-101x3 [50]	384^2	388	204.6	-	84.4
R-152x4 [50]	480^2	937	840.5	-	85.4
ViT-B/16 [51]	384^2	86	55.4	86	84.0
ViT-L/16 [51]	384^2	307	190.7	27	85.2
Swin-B [11]	224^2	88	15.4	278	85.2
Swin-B [11]	384^2	88	47.0	85	86.4
ConvNeXt-B [49]	224^2	89	15.4	292	85.8
ConvNeXt-B [49]	384^2	89	45.1	96	86.8
Iwin-B(ours)	224^2	91.2	15.9	271	**85.5**
Iwin-B(ours)	384^2	91.2	48.3	79	**86.6**
Iwin-B(ours)	512^2	91.2	89.5	51	**86.1**
Iwin-B(ours)	1024^2	91.2	358.2	12	**85.6**
Swin-L [11]	224^2	197	34.5	145	86.3
Swin-L [11]	384^2	197	103.9	46	87.3
ConvNeXt-L [49]	224^2	198	34.4	147	86.6
ConvNeXt-L [49]	384^2	198	101.0	50	87.5
Iwin-L(ours)	224^2	204.3	35.4	138	**86.4**
Iwin-L(ours)	384^2	204.3	106.6	43	**87.4**

C. ADE20K 上的语义分割

我们使用 MMSegmentation [56] 工具箱中的 UperNet [55] ，在 ADE20K [41] 语义分割基准上评估 Iwin 骨干网

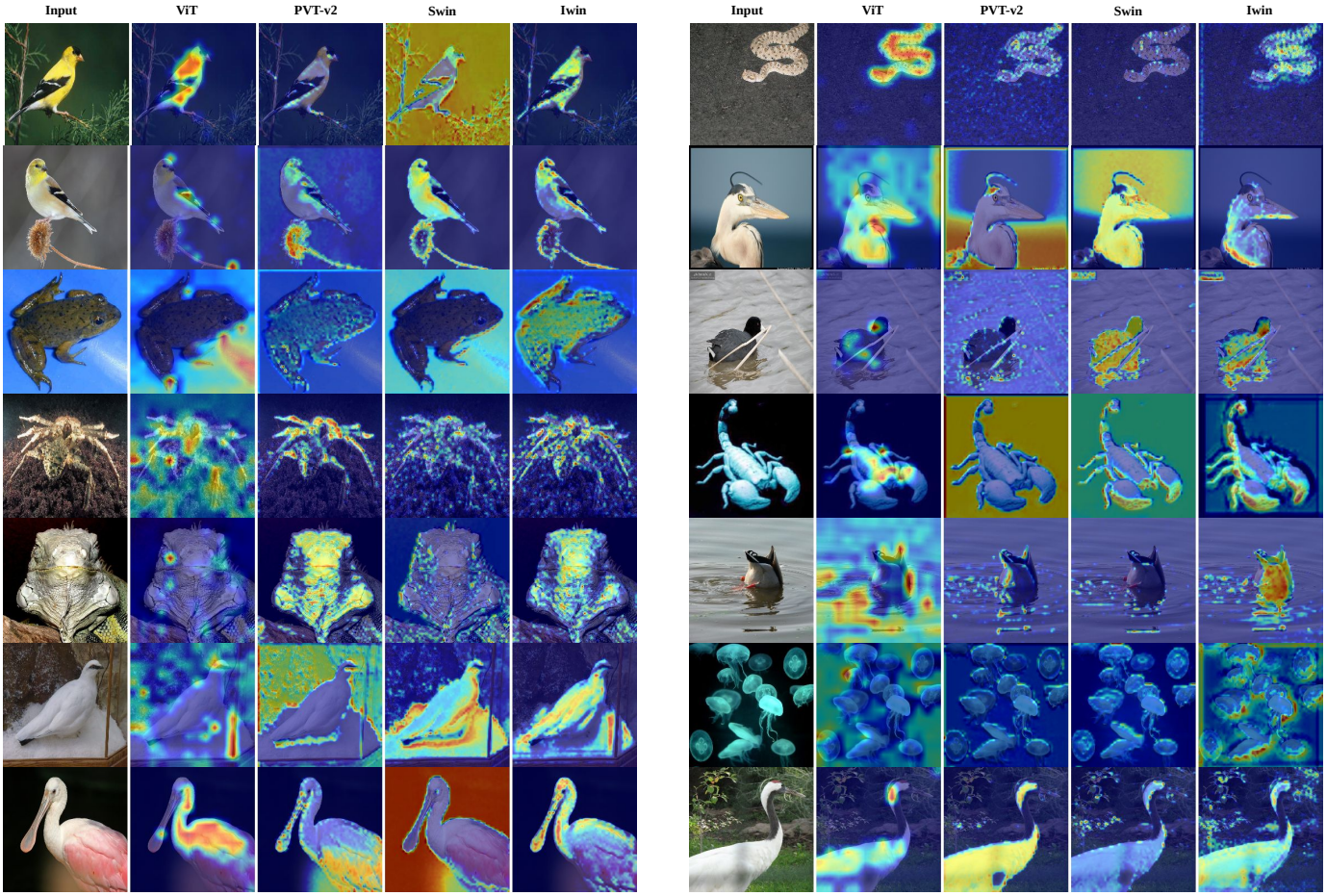


Fig. 4. 热图的可视化。左栏显示输入图像，而后续列显示来自原生 VIT、PVTv2、Swin 和 Iwin（我们的方法）的结果。结果表明 Iwin 能够有效地将激活集中在目标对象上。

络。实验设置遵循 Swin [11] 。训练在 160K 次迭代上进行，总批量为 16（在 8 个 GPU 上，每个 GPU 上 2 张图像）。

如表 IV 所示，Iwin-B 实现了 48.9 % 的 mIoU，以 0.8 % 的优势超越了 Swin-B 的 48.1 % mIoU，同时保持几乎相同的 FLOPs（1189G 对 1188G）和参数量（124.8M 对 121.0M）。该性能接近于领先的 ConvNeXt-B，其在类似计算成本下录得 49.1 % mIoU。对于较小的模型，Iwin-T 实现了 44.7 % 的 mIoU，略微超过了 Swin-T 的 44.5 % mIoU，而 FLOPs 可比。这些结果表明 Iwin 在语义分割任务中的有效性和竞争力。

D. 在 Kinetics-400 上的视频识别

根据 Video Swin Transformer [58] 的设置，我们提出了 Video Iwin Transformer 用于在 Kinetics-400 [42] 数据集上使用 MMaction2 [59] 工具箱进行动作识别。Video Iwin Transformer 使用其在 ImageNet 上预训练的 Iwin 模型进行初始化。与 Video Swin Transformer 从 2D 到 3D 复杂的适配（包括窗口平移、遮罩和相对定位）不同，Iwin 模型只需在时间维度上添加一个可学习的绝对位置编码，其他组件保持不变。如图 9 所示，在 Iwin 3D Attention 中，跨所有帧的交错窗口中的令牌被收集用于注意力计算，而深度可分离卷积则保持其在单个帧上的 2D 操作。

根据表格 V，Iwin 模型在性能和效率上优于 Swin。在可比的模型规模下，Iwin-T 的 Top-1 准确率达到 79.1 %，Top-5 准确率为 93.8 %，略微超过了 Swin-T 的 78.8 % 和 93.6 %。更重要的是，Iwin-T 的计算成本显著降低，仅需 74 GFLOPs，而 Swin-T 需要 88 GFLOPs，减少了 15.9 %。这表明 Iwin-T 在超越 Swin-T 性能的同时，还提供了更高的计算效率。对于更大规模的 Iwin-S，虽然其 Top-1 和 Top-5 准确率（分别为 80.0 % 和 94.1 %）低于 Swin-S（80.6 % 和 94.5 %），但其计算成本（140 GFLOPs）仍然显著低于 Swin-S（166 GFLOPs），反映出大约 15.7 % 的下降。这表明 Iwin 在保持竞争性能的同时显著减少了计算成本。

E. 图像生成

我们遵循 LightningDiT [65] 的设置，并构建了一个 FlashDiT 模型，以验证 Iwin 中关键组件在 ImageNet 分类条件图像生成任务中的有效性。我们用提出的交错窗口注意力和深度可分离卷积的组合替换了标准的自注意力。我们去掉了位置编码，并将卷积核大小设为 3×3，窗口大小设为 4×4，这基于 16×16 的潜在特征图尺寸（由 256×256 输入图像经过 16 倍 $\times$ 下采样得到），其中 $3 \times 4 = 12$ 接近于 16。

如表 VI 所示，我们提出的 FlashDiT 在图像生成方面表现出效率。在仅需 56 次训练迭代中，它就实现了竞争力的

TABLE II
使用 MASK-RCNN 和 CASCADE MASK-RCNN 的 COCO 目标检测和分割结果。FLOPs 是基于图像尺寸 (1280, 800) 计算的。

Backbone	FLOPs	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}
Mask-RCNN 1 × schedule						
PVTv2-B1	-	41.8	64.3	45.9	38.8	61.2
Swin-T	267G	43.7	66.6	47.7	39.8	63.3
Iwin-T	268G	42.2	65.3	45.8	38.9	62.1
Iwin-S	358G	43.7	67.0	47.4	40.0	63.9
Mask-RCNN 3 × schedule						
PVTv2-B2	-	47.8	-	-	43.1	-
Swin-T	267G	46.0	68.1	50.3	41.6	65.1
ConvNeXt-T	262G	46.2	67.9	50.8	41.7	65.0
Iwin-T	268G	44.7	67.2	48.8	40.9	64.1
Iwin-S	358G	45.5	67.5	49.6	41.0	64.3
Cascade Mask-RCNN 1 × schedule						
Swin-T	745G	48.1	67.1	52.2	41.7	64.4
Iwin-T	747G	47.2	66.1	51.3	40.9	63.5
Cascade Mask-RCNN 3 × schedule						
ResNet-50	739G	46.3	64.3	50.5	40.1	61.7
X101-32	819G	48.1	66.5	52.4	41.6	63.9
X101-64	972G	48.3	66.4	52.3	41.7	64.0
PVTv2-B2	788G	51.1	69.8	55.3	44.4	-
Swin-T	745G	50.4	69.2	54.7	43.7	66.6
ConvNeXt-T	741G	50.4	69.1	54.8	43.7	66.5
Iwin-T	747G	49.4	68.4	53.5	42.9	65.8
Swin-S	838G	51.9	70.7	56.3	45.0	68.2
ConvNeXt-S	827G	51.9	70.8	56.5	45.0	68.4
Iwin-S	837G	49.4	68.1	53.3	43.0	65.6

Fig. 5. 在 COCO2017 上的目标检测可视化。最左边的列显示了输入图像。由左到右，依次显示了基于 PVTv2、基于 Swin 和基于 Iwin 的 Mask R-CNN 生成的结果。

性能，而之前的最先进模型如 DiT (1400) 和 MAR (800) 则需要更多的迭代次数。与这些复杂度为 32^2 或 16^2 的模型不同，FlashDiT 的计算复杂度仅为 $3^2 + 4^2 = 25$ ，同时保持了高生成质量，得到 3.08 的 gFID 和 223.2 的 IS，无需分类器指导。因此，FlashDiT 验证了 Iwin 中关键组件作为一个独立模块的有效性，能够无缝地替换生成模型中的自注意力模块。

F. 消融研究

我们广泛的消融研究，汇总于表格 VII，默认在 Iwin-T 和 ImageNet 上进行。

我们研究了整合深度可分离卷积（DWConv）和不同注意力机制的效果。如所示，DWConv 和 IW-MSA（交错窗口多头自注意力）合作取得了最佳性能，实现了% 82.0 的 Top-1 准确率。这验证了我们所提方法的优越性。

我们评估了网络阶段之间的各种下采样方法。结果表明，在我们最终的 Iwin 设计中采用的标准卷积 (Std Conv) 实现了 82.0% 的最高准确率。我们在标准卷积和提供最高吞吐量的平均池化之间权衡，但为了更高的准确率，我们选择了标准卷积。

我们研究了在不同阶段中不同卷积核大小对 DWConv 的影响。我们最终的配置使用固定的卷积核大小 { 3, 3, 3, None } ，达到了 82.0 % 的准确率，并具备 736 img/s 的吞吐量。虽然使用 { 5, 5, 5, None } 卷积核的准确率稍高，为 82.2 % ，但较小的 { 3, 3, 3, None } 卷积核在性能和计算效率（更高的吞吐量）之间提供了更有利的平衡。此外，为满足 $KM \geq \max(H, W)$ 而在不同阶段使用不同的卷积核大小并未产生最佳结果。这与 [37] 中的观察一致，平衡的网络配置优于理论上最优但不平衡的配置。

我们研究了网络四个阶段中块数量的不同分布。起初，我们探索了配置 { 4, 3, 2, 2 }，它旨在通过堆叠较小的卷积块来近似较大的内核尺寸（例如，用四个 3×3 卷积的 7×7 内核）。然而，这种设置导致最低的准确率 (80.5 %) 和吞吐量 (473 img/s)。相比之下，配置 { 2, 2, 6, 2 } 实现了最高的准确率 82.0 %，且吞吐量为 736 img/s。这个配置将更多的块分配给较深的阶段，证明了在保持最佳速度的同时，最大化性能是有益的。

我们探讨了位置编码对模型的影响。在 Iwin-T 模型中，使用相对位置嵌入实现了 82.4 % 的最高准确率。然而，在像 Iwin-S 这样更深的模型中，无位置嵌入的方法达到了 83.4 % 的最高 Top-1 准确率，超越了相对位置嵌入（83.3 % ）并且每秒处理更多图像 (410 img/s)。这表明在非常深

TABLE III
对于分辨率 224^2 ，WIN. SZ. 的详细架构规格是 12, 16, 16，而对于分辨率 384^2 ，512^2 ，1024^2 。

	downsp. rate (output size)	Iwin-T	Iwin-S	Iwin-B	Iwin-L
stage 1	4 × (56 × 56)	ker. 3 pad. 2 4 × 4, 96-d, LN	ker. 3 pad. 2 4 × 4, 96-d, LN	ker. 3 pad. 2 4 × 4, 128-d, LN	ker. 3 pad. 2 4 × 4, 192-d,
		win. sz. 7×7, ker. sz. 3×3, dim 96, head 3 × 2	win. sz. 7×7, ker. sz. 3×3, dim 96, head 3 × 2	win. sz. 7×7, ker. sz. 3×3, dim 128, head 4 × 2	win. sz. 7×7, ker. sz. 3×3, dim 192, head 6 × 2
stage 2	8 × (28 × 28)	ker. 3 pad. 2 , 192-d , LN	ker. 3 pad. 2 , 192-d , LN	ker. 3 pad. 2 , 256-d , LN	ker. 3 pad. 2 , 384-d , LN
		win. sz. 7×7, ker. sz. 3×3, dim 192, head 6 × 2	win. sz. 7×7, ker. sz. 3×3, dim 192, head 6 × 2	win. sz. 7×7, ker. sz. 3×3, dim 256, head 8 × 2	win. sz. 7×7, ker. sz. 3×3, dim 384, head 12 × 2
stage 3	16 × (14 × 14)	ker. 3 pad. 2 , 384-d , LN	ker. 3 pad. 2 , 384-d , LN	ker. 3 pad. 2 , 512-d , LN	ker. 3 pad. 2 , 768-d , LN
		win. sz. 7×7, ker. sz. 3×3, dim 384, head 12 × 6	win. sz. 7×7, ker. sz. 3×3, dim 384, head 12 × 18	win. sz. 7×7, ker. sz. 3×3, dim 512, head 16 × 18	win. sz. 7×7, ker. sz. 3×3, dim 768, head 24 × 18
stage 4	32 × (7 × 7)	ker. 3 pad. 2 , 768-d , LN	ker. 3 pad. 2 , 768-d , LN	ker. 3 pad. 2 , 1024-d , LN	ker. 3 pad. 2 , 1536-d , LN
		win. sz. 7×7, dim 768, head 24 × 2	win. sz. 7×7, dim 768, head 24 × 2	win. sz. 7×7, dim 1024, head 32 × 2	win. sz. 7×7, dim 1536, head 48 × 2

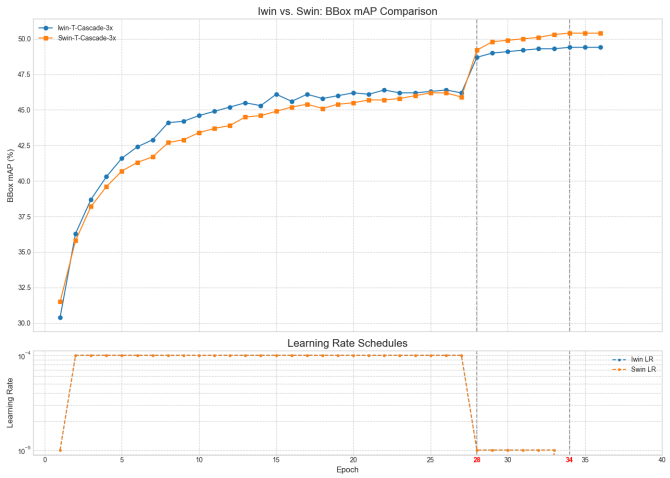


Fig. 6. Iwin-T 与 Swin-T 在 COCO 数据集（Cascade Mask-RCNN 3 × 调度）上的 BBox mAP 和学习率进展。

TABLE IV
ADE20K 语义分割任务的结果。FLOPs 是在输入尺寸为 512 × 2048 时测量的。

Backbone	Semantic FPN 80k			UperNet 160k		
	Param(M)	FLOPs(G)	mIoU(%)	Param(M)	FLOPs(G)	mIoU(%)
ResNet50 [35]	28.5	183	36.7	-	-	-
PVTv2 B2 [46]	29.1	165	45.2	-	-	-
ConvNeXt-T [49]	-	-	-	60.0	939	46.0
Swin-T [11]	-	-	-	59.9	945	44.5
Iwin-T(ours)	-	-	-	**61.9**	**946**	**44.7**
ResNet101 [35]	47.5	260	38.8	-	-	-
PVTv2 B3 [46]	49.0	224	47.3	-	-	-
ConvNeXt-S [49]	-	-	-	82.0	1027	48.7
Swin-S [11]	-	-	-	81.3	1038	47.6
Iwin-S(ours)	-	-	-	**83.2**	**1038**	**47.5**
ResNeXt101-64 × 4d [57]	86.4	-	40.2	-	-	-
PVTv2 B4 [46]	66.3	285	47.9	-	-	-
ConvNeXt-B [49]	-	-	-	122.0	1170	49.1
Swin-B [11]	-	-	-	121.0	1188	48.1
Iwin-B(ours)	-	-	-	**124.8**	**1189**	**48.9**

的网络中，位置嵌入可能是多余的甚至有害的。此外，在训练过程中，使用绝对或相对位置嵌入的模型相比于没有位置嵌入的模型需要更多的时间来学习。

如表 VIII 所示，我们的基准 Iwin-T 模型在相同设置下实现了 42.2 AP^{box} ，比 Swin-T 的 43.7 AP^{box} 低 1.5。改

TABLE V
在 KINETICS-400 上的比较。“VIEWS” 表示 # 时间片段 × # 空间裁剪。FLOPs 和参数的量级分别为千亿（10^9 ）和百万（10^6 ）。

Method	Pretrain	Top-1	Top-5	Views	FLOPs	Param
R(2+1)D [60]	-	72.0	90.0	10 × 1	75	61.8
I3D [61]	ImageNet-1K	72.1	90.3	-	108	25.0
NL I3D-101 [62]	ImageNet-1K	77.7	93.3	10 × 3	359	61.8
SlowFast R101+NL [63]	-	79.8	93.9	10 × 3	234	59.9
X3D-XXL [64]	-	80.4	94.6	10 × 3	144	20.3
MViT-B, 32×3 [20]	-	80.2	94.4	1 × 5	170	36.6
MViT-B, 64×3 [20]	-	81.2	95.1	3 × 3	455	36.6
ViViT-L/16x2 [16]	ImageNet-21K	80.6	94.7	4 × 3	1446	310.8
ViViT-L/16x2 320 [16]	ImageNet-21K	81.3	94.7	4 × 3	3992	310.8
Swin-T [11]	ImageNet-1K	78.8	93.6	4 × 3	88	28.2
Swin-S [11]	ImageNet-1K	80.6	94.5	4 × 3	166	49.8
Iwin-T(ours)	ImageNet-1K	79.1	93.8	4 × 3	74	29.8
Iwin-S(ours)	ImageNet-1K	80.0	94.1	4 × 3	140	51.1

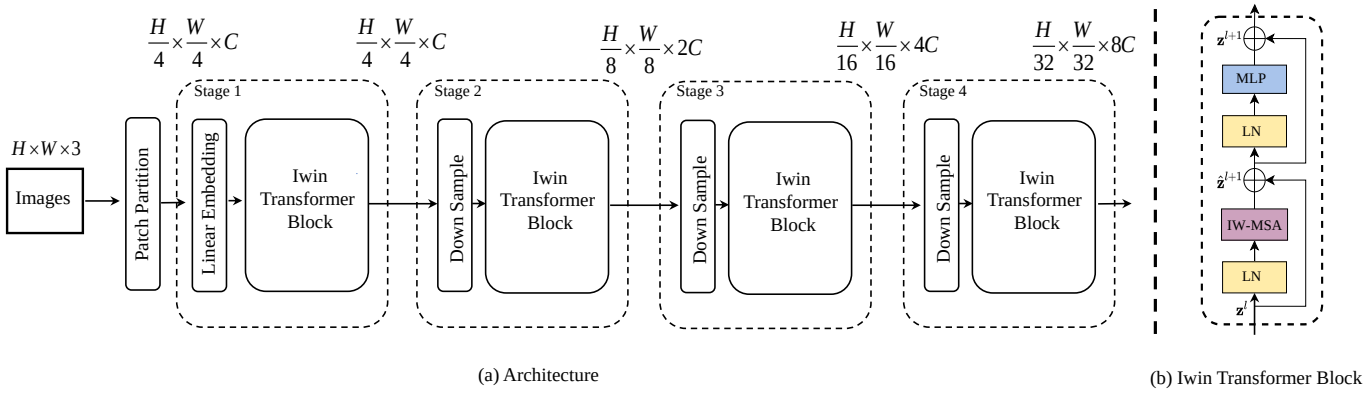
变学习率策略为更平滑的余弦退火或增加初始学习率均未能带来改进。然而，加入相对位置编码使 Iwin-T 模型有了显著的 0.7 AP $^{\text{box}}$ 的提升。受到这一改进的鼓舞，我们进一步研究了其对扩展版 Iwin-S 模型的影响。Iwin-S 成功弥合了性能差距，达到了 43.7 AP $^{\text{box}}$ ，与 Swin-T 相当。然而，将相对位置编码添加到 Iwin-S 导致性能略微下降至 43.5 AP $^{\text{box}}$ 。这些结果表明 Iwin 架构在 COCO 基准上遇到了复杂的特定任务优化挑战。

IV. 讨论

我们相信某些工作领域可以从 Iwin Transformer 中受益并获得启发。

A. 迁移到大型语言模型

Iwin Transformer 的无位置嵌入设计原则为大语言模型（LLMs）的应用提供了有前景的机会。目前，大语言模型在很大程度上依赖位置嵌入来保存序列顺序信息。通过结合交错窗口注意力和深度可分离卷积，可能实现更自然的长度泛化。这种方法依赖于结构信息而非参数化位置信息，从而更易于进行长度外推。如图 10 所示，计算被分为两个部分：1D 因果深度可分离卷积和 1D 交错窗口因果注意力。二者均确保 token 仅与前面的 token 相关联，使 Iwin 1D 注意力具有因果性。此外，我们可以通过将两个窗口大小 M_1 和 M_2 设为相等，替换深度可分离卷积为普通窗口因果注意，从而得到窗口大小 $\sqrt{N}$ 。这将序列长度 N 的复杂度从 N^2 降低到 N 。



(a) Architecture (b) Iwin Transformer Block

Fig. 7. (a) Iwin Transformer（Iwin-T）的整体架构。(b) 单个 Iwin Transformer 块。IW-MSA 涉及并行应用交错窗口多头自注意力和深度可分离卷积。

TABLE VI
在 IMAGENET 上与最新模型进行图像生成的比较。

Method	Epoches	Params	256 × 256 w/o CFG					256 × 256 w/ CFG				
			gFID ↓	sFID ↓	IS ↑	Pre. ↑	Rec. ↑	gFID ↓	sFID ↓	IS ↑	Pre. ↑	Rec. ↑
AutoRegressive (AR)												
MaskGIT [?]	555	227M	6.18	-	182.1	0.80	0.51	-	-	-	-	-
LlamaGen [66]	300	3.1B	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58
VAR [67]	350	2.0B	-	-	-	-	-	1.80	-	365.4	0.83	0.57
MagViT-v2 [68]	1080	307M	3.65	-	200.5	-	-	1.78	-	319.4	-	-
MAR [69]	800	945M	2.35	-	227.8	0.79	0.62	1.55	-	303.7	0.81	0.62
Latent Diffusion Models												
MaskDiT [70]	1600	675M	5.69	10.34	177.9	0.74	0.60	2.28	5.67	276.6	0.80	0.61
DiT [18]	1400	675M	9.62	6.85	121.5	0.67	0.67	2.27	4.60	278.2	0.83	0.57
SiT [71]	1400	675M	8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59
MDTv2 [72]	1080	675M	-	-	-	-	-	1.58	4.52	314.7	0.79	0.65
REPA [73]	800	675M	5.90	-	-	-	-	1.42	4.70	305.7	0.80	0.65
LightningDiT [65]	64	675M	5.14	4.22	130.2	0.76	0.62	2.11	4.16	252.3	0.81	0.58
FlashDiT(Ours)	56	675M	7.88	5.91	116.2	0.73	0.60	3.08	6.00	223.2	0.78	0.57

B. 应用于生成模型

生成模型，例如去噪扩散模型，是 Iwin 设计能够提供显著优势的另一个领域。Iwin 没有位置嵌入，使得其能够毫无缝隙地适应各种分辨率，而无需调整参数或插值，这对于渐进生成策略至关重要。这个特性可以帮助图像生成模型创建更高分辨率或任意大小的图像。此外，由于 Iwin 中的深度可分离卷积引入的归纳偏差，Iwin 中观察到的更快的收敛特性可以减少扩散模型的训练时间。

我们提出的 Iwin 3D Attention 如图 9 所示，其在空间和时间域中形成窗口，与空间域中的 2D 深度可分离卷积结合，已在动作识别中证明了有效性。我们相信它可以作为视频生成的第三个选项，与 3D 完全注意力和时空注意力机制并列。虽然 3D 完全注意力是一个昂贵的理想解决方案，但时空注意力机制的串行结构可能导致时间注意力破坏空间注意力形成的分布，可能导致帧图像的不和谐。相比之下，Iwin 3D Attention 使用一次注意力操作和一次深度可分离卷积来建立视频中所有标记之间的关系，因此我们可以预期生成的视频质量会更高。

C. 局限性和未来工作

Iwin 在目标检测方面的表现不如 Swin，原因尚不清楚。由于计算资源有限，我们无法进行广泛的实验来找出有效的学习策略或通过精心优化来提升其目标检测能力。这项任务留待未来的研究。此外，我们没有验证 Iwin 是否遵循放缩定律。未来的工作将包括将提出的 Iwin 注意力及其变体扩展到大语言模型、图像生成和 3D 视频生成的应用中。

V. 结论

在本文中，我们介绍了 Iwin Transformer，这是一种新颖的位置嵌入无关的视觉 Transformer，利用了创新的交错窗口注意力和深度可分离卷积的结合。在多个视觉基准上的广泛实验评估表明，Iwin 在图像分类、语义分割和视频动作识别等任务中表现出竞争力。

最重要的是，使用注意力机制来捕捉长程依赖关系并使用卷积来抓取局部关系以构建全局连接的想法及其实现方法，可以为未来的工作提供灵感。Iwin Transformer 的核心组件可以直接应用于二维生成模型，并已在类别条件图像生成任务中证明其有效性，同时在通过 Iwin 三维注意力扩展到三维数据（如视频生成）方面展现了潜力。Iwin 一维注意力可能对于大型语言模型中的一维数据也同样有效，有待于未来的工作去验证。

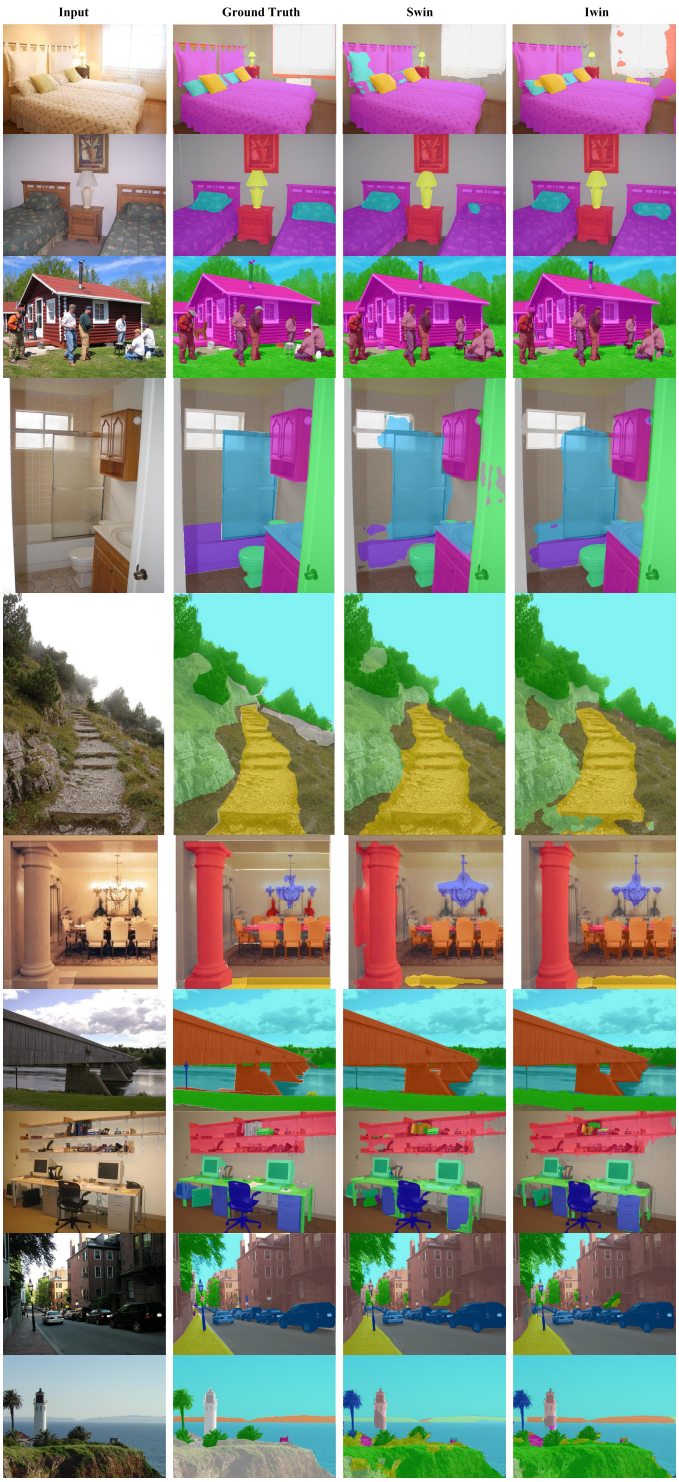


Fig. 8. ADE20K 数据集上的语义分割结果。第一列显示的是输入图像。从左到右依次为：真实标签、基于 Swin 的 UperNet 和基于 Iwin 的 UperNet。

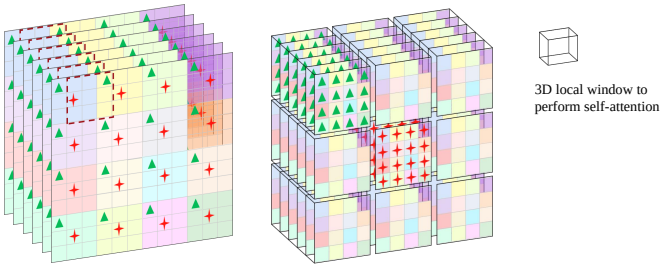


Fig. 9. Iwin 3D Attention 的示意图。在左侧图像中，如同在 2D 中一样，对每一帧执行深度可分离卷积。在右侧，执行 3D 中的交错窗口注意力。这有效地近似了 3D 全卷积，使视频中的任何标记之间的连接成为可能。唯一的缺点是在时间维度上需要位置编码。为了完全摆脱它，我们提出了两个选项：在时间维度执行 RTR 操作并配合一维卷积；第二个是增大时间维度的卷积核尺寸以跨越帧。由于论文篇幅限制，这部分将留待今后的工作中解决。

TABLE VII
关于各种架构组件的消融研究。FLOPs 为 Giga (10^9)，参数为 Mega (10^6)。

Setting	Param(M)	FLOPs(G)	Throughput(img/s)	Top-1 Acc(%)
Ablation on Attention and Convolution Combination				
DwConv	21.60	3.20	861	79.4
W-MSA	30.20	4.71	758	80.2
IW-MSA	30.20	4.71	756	80.4
DwConv + W-MSA	30.23	4.72	737	81.8
DwConv + IW-MSA	30.23	4.72	736	82.0
Ablation on Downsampling Methods				
DWConv	27.14	4.57	746	81.9
Avg Pooling	27.13	4.51	762	81.8
Patch Merging	28.29	4.51	741	81.8
Std Conv	30.23	4.72	736	82.0
Ablation on Kernel Size for Depthwise Convolution				
{ 7, 5, 3, None }	30.24	4.75	714	81.0
{ 7, 7, 7, None }	30.34	4.78	729	82.1
{ 5, 5, 5, None }	30.27	4.75	731	82.2
{ 3, 3, 3, None }	30.23	4.72	736	82.0
Ablation on Block Number Configuration				
{ 4, 3, 2, 2 }	23.79	4.43	473	80.5
{ 3, 3, 3, 3 }	32.56	4.80	588	81.8
{ 2, 2, 6, 2 }	30.23	4.72	736	82.0
Ablation on Position Embedding				
abs. pos.	30.53	4.72	735	82.1
rel. pos.	30.25	4.72	724	82.4
no pos.	30.23	4.72	736	82.0
rel. pos. (Iwin-S)	51.60	8.98	403	83.3
no pos. (Iwin-S)	51.60	8.98	410	83.4

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al. , "An image is worth 16x16 words: Transformers for image recognition at scale," International Conference on Learning Representations (ICLR) , 2021.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems (NeurIPS) , vol. 30, 2017.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems , vol. 25, 2012.

[4] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," IEEE International Conference on Computer Vision (ICCV) , pp. 568–578, 2021.

[5] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in Advances in Neural Information Processing Systems (NeurIPS) , vol. 34, 2021, pp. 9355–9366.

[6] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," International Conference on Machine Learning (ICML) , pp. 2286–2296, 2021.

[7] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," Advances in Neural Information Processing Systems (NeurIPS) , vol. 34, pp. 3965–3977, 2021.

[8] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: What can 8 learned tokens do for images and videos?" arXiv preprint arXiv:2106.11297 , 2021.

[9] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," arXiv preprint arXiv:2001.04451 , 2020.

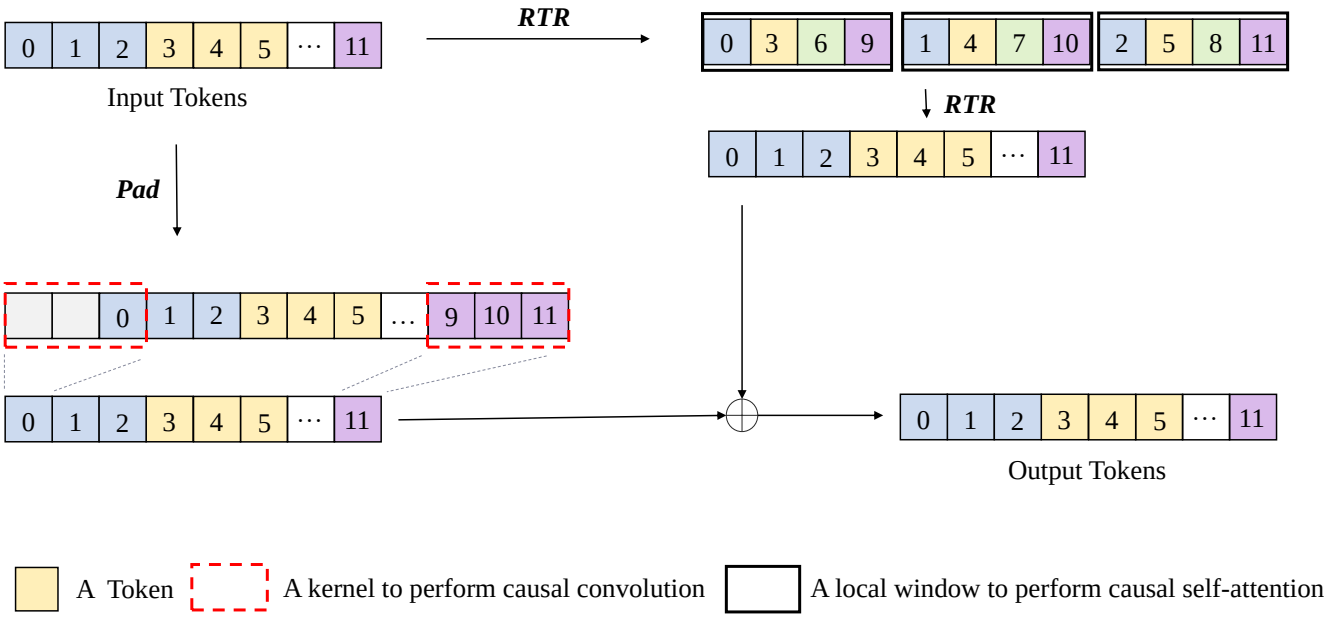


Fig. 10. Iwin 一维注意力的示意图展示了如何将 Iwin Transformer 概念应用于大型语言模型（LLM）。红色框代表因果深度可分离卷积，而黑色框表示窗内因果注意力。这两个过程都不会泄露未来信息。最终输出源自这两个因果操作的组合，遵循因果原则。这种方法可能也有助于解决 LLM 中长序列的高复杂性问题。

TABLE VIII
在 COCO 上针对 MASK R-CNN 1 × 计划对 IWIN-T 进行消融研究。“LR”表示初始学习率。“REL. POS.”表示相对位置编码。

Method	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Swin-T (lr=1e-4, step)	43.7	66.6	47.7	39.8	63.3	42.7
Iwin-T (lr=1e-4, step)	42.2	65.3	45.8	38.9	62.1	41.6
CosineAnnealing	42.0	65.2	45.7	38.7	61.9	41.2
Increase lr to 2e-4	42.2	64.9	46.2	39.1	62.0	41.9
rel. pos.	42.9	66.0	46.7	39.4	62.7	42.2
Iwin-S	43.7	67.0	47.4	40.0	63.9	42.5
Iwin-S + rel. pos.	43.5	66.7	47.4	40.1	63.4	42.7

[10] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser et al. , “Rethinking attention with performers,” International Conference on Learning Representations (ICLR) , 2021.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” IEEE International Conference on Computer Vision (ICCV) , pp. 10 012–10 022, 2021.

[12] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin transformer v2: Scaling up capacity and resolution,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp. 12 009–12 019, 2022.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” International Conference on Machine Learning (ICML) , 2021.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” European Conference on Computer Vision (ECCV) , pp. 213–229, 2020.

[15] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr et al. , “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp. 6881–6890, 2021.

[16] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “Vivit: A video vision transformer,” IEEE International Conference on Computer Vision (ICCV) , pp. 6836–6846, 2021.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159 , 2020.

[18] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in Proceedings of the IEEE/CVF international conference on computer vision , 2023, pp. 4195–4205.

[19] S. Mehta and M. Rastegari, “Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer,” International Conference on Learning Representations (ICLR) , 2022.

[20] H. Fan, B. Xiong, K. Mangalam, Y. Li, K. He, and J. Malik, “Multiscale vision transformers,” in Proceedings of the IEEE/CVF International Conference on Computer Vision , 2021, pp. 6824–6835.

[21] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: Self-attention with linear complexity,” arXiv preprint arXiv:2006.04768 , 2020.

[22] X. Ma, X. Kong, S. Wang, C. Zhou, J. May, H. Ma, and L. Zettlemoyer, “Luna: Linear unified nested attention,” Advances in Neural Information Processing Systems , vol. 34, pp. 2441–2453, 2021.

[23] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang et al. , “Big bird: Transformers for longer sequences,” Advances in Neural Information Processing Systems (NeurIPS) , vol. 33, pp. 17 283–17 297, 2020.

[24] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150 , 2020.

[25] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers,” arXiv preprint arXiv:1904.10509 , 2019.

[26] Y. Tay, D. Bahri, L. Yang, D. Metzler, and D.-C. Juan, “Sparse sinkhorn attention,” International Conference on Machine Learning (ICML) , pp. 9438–9447, 2020.

[27] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, “Efficient content-based sparse attention with routing transformers,” Transactions of the Association for Computational Linguistics , vol. 9, pp. 53–68, 2021.

[28] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, “Cswin transformer: A general vision transformer

backbone with cross-shaped windows," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp. 12 124–12 134, 2022.
[29] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," arXiv preprint arXiv:2104.05707 , 2021.
[30] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," Advances in Neural Information Processing Systems (NeurIPS) , vol. 34, 2021.
[31] H. Yin, A. Vahdat, J. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-vit: Adaptive tokens for efficient vision transformer," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp. 10 809–10 818, 2022.
[32] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," arXiv preprint arXiv:2210.09461 , 2022.
[33] M. Fayyaz, S. A. Koohpayegani, F. Rezaei, S. Somayaji, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," European Conference on Computer Vision (ECCV) , pp. 209–226, 2022.
[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition , 2017, pp. 1251–1258.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition , 2016, pp. 770–778.
[36] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," arXiv preprint arXiv:1606.08415 , 2016.
[37] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning . PMLR, 2019, pp. 6105–6114.
[38] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," Advances in neural information processing systems , vol. 29, 2016.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition . Ieee, 2009, pp. 248–255.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13 . Springer, 2014, pp. 740–755.
[41] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," International Journal of Computer Vision , vol. 127, pp. 302–321, 2019.
[42] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980
[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning (ICML) , 2021, pp. 10 347–10 357.
[45] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2021, pp. 558–567.
[46] W. Wang, E. Xie, X. Li, D. Fan, M.-M. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," arXiv preprint arXiv:2102.12122 , 2021.
[47] ——, "Pvtv2: Improved baselines with pyramid vision transformer," in Computational Visual Media (CVM) , 2022, pp. 100–110.
[48] W. Yu, C. Si, Z. Zhou, X. Tan, and J. Wang, "Metaformer is actually what you need for vision," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2021, pp. 6658–6668.
[49] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , 2022, pp. 11 976–11 986.
[50] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," 2020. [Online]. Available: https://arxiv.org/abs/1912.11370
[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929
[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision , 2017, pp. 2961–2969.
[53] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition , 2018, pp. 6154–6162.
[54] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," arXiv preprint arXiv:1906.07155 , 2019.
[55] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in Proceedings of the European conference on computer vision (ECCV) , 2018, pp. 418–434.
[56] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.
[57] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1611.05431
[58] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," arXiv preprint arXiv:2106.13230 , 2021.
[59] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.
[60] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition , 2018, pp. 6450–6459.
[61] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2017, pp. 6299–6308.
[62] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition , 2018, pp. 7794–7803.
[63] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in Proceedings of the IEEE/CVF international conference on computer vision , 2019, pp. 6202–6211.
[64] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , 2020, pp. 203–213.
[65] J. Yao, B. Yang, and X. Wang, "Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2025.
[66] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan, "Autoregressive model beats diffusion: Llama for scalable image generation," arXiv preprint arXiv:2406.06525 , 2024.
[67] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," arXiv preprint arXiv:2404.02905 , 2024.
[68] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, V. Birodkar, A. Gupta, X. Gu et al. , "Language model beats diffusion–tokenizer is key to visual generation," arXiv preprint arXiv:2310.05737 , 2023.

[69] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," arXiv preprint arXiv:2406.11838 , 2024.
[70] H. Zheng, W. Nie, A. Vahdat, and A. Anandkumar, "Fast training of diffusion models with masked transformers," arXiv preprint arXiv:2306.09305 , 2023.
[71] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie, "Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers," arXiv preprint arXiv:2401.08740 , 2024.
[72] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan, "Mdtv2: Masked diffusion transformer is a strong image synthesizer," arXiv preprint arXiv:2303.14389 , 2023.
[73] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, "Representation alignment for generation: Training diffusion transformers is easier than you think," arXiv preprint arXiv:2410.06940 , 2024.

Simin Huo received his B.S. degree from Nanjing University of Science and Technology, China, Nanjing, in 2018, and his M.S. degree from Bauman Moscow State Technical University, Russia, Moscow, in 2021. He is currently a Ph.D. student with the Department of Automatics, Shanghai Jiao Tong University. His main research interests include computer vision and deep learning , with a focus on efficient vision transformer and generation models.

Ning Li (Member, IEEE) received the B.S. and M.S. degrees from Qingdao University of Science and Technology, Qingdao, China, in 1996 and 1999, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2002. She is currently a Professor with the Department of Automation, Shanghai Jiao Tong University. Her research interests include modeling and control of complex systems, artificial intelligence, and big data analysis.