第 5 章　基于 CNN 和 Transformer 的低剂量 CT 图像去噪方法

本章提出一种基于 Transformer 和 CNN 相结合的低剂量 CT 图像去噪方法。网络的编码器部分基于 CNN，主要用于提取图像细节信息。在解码器部分，我们提出了一种双路 Transformer 块(DPTB)，它分别通过两条路径提取跳跃式连接的输入特征和前一级的输入特征。DPTB 能较好地恢复去噪图像的细节和结构信息。为了更好地关注网络浅层提取的特征图像的关键区域，我们还在跳跃连接部分提出了多特征空间注意块(MSAB)。对于建立的方法进行了实验研究，并与现有网络进行了对比。结果表明，该方法在 PSNR、SSIM、RMSE 等评价指标上均能有效去除 CT 图像中的噪声，提高图像质量，并优于现有模型。该方法在梅奥诊所低剂量 CT 挑战赛数据集上的 PSNR 为 28.9720，SSIM 为 0.8595，RMSE 为 14.8657。对于 QIN_LUNG_CT 数据集上的不同噪声水平 σ(15、35、55)，我们的方法也取得了更好的性能。

5.1　Transformer 模型理论基础

近年来，Transformer 模型在众多语言处理任务中展现出卓越的性能，如文本分类和机器翻译等，引起了学术界的广泛关注。基于 Transformer 架构的模型，如 BERT[114]、RoBERTa[115] 和 T5[116]，因其在处理大规模数据集时展现出的卓越可扩展性而备受瞩目。特别地，Switch Transformer 模型的规模已扩展至 1.6 万亿参数，体现了 Transformer 在深度学习领域的深远影响。此外，学者们也开始探索 Transformer 在计算机视觉领域的应用潜力，其能够有效建模输入序列元素间的长期依赖关系，不同于传统卷积网络，Transformer 凭借最小的归纳偏差和对集合函数的自然适应性，在多模态学习任务中显示出优异表现。Transformer 架构的核心是自注意力机制，它能够捕捉序列内元素间的相互关系，与仅能捕获短期依赖的循环网络形成对比。Transformer 完全基于注意力机制构建，通过多头注意力模块的创新优化了并行处理能力，包含编码器和解码器，以及由多头注意层、前馈神经网络、残差连接和层归一化组成的 Transformer 模块，这些模块共同作用生成高质量的输出序列，如图 5-1 所示。

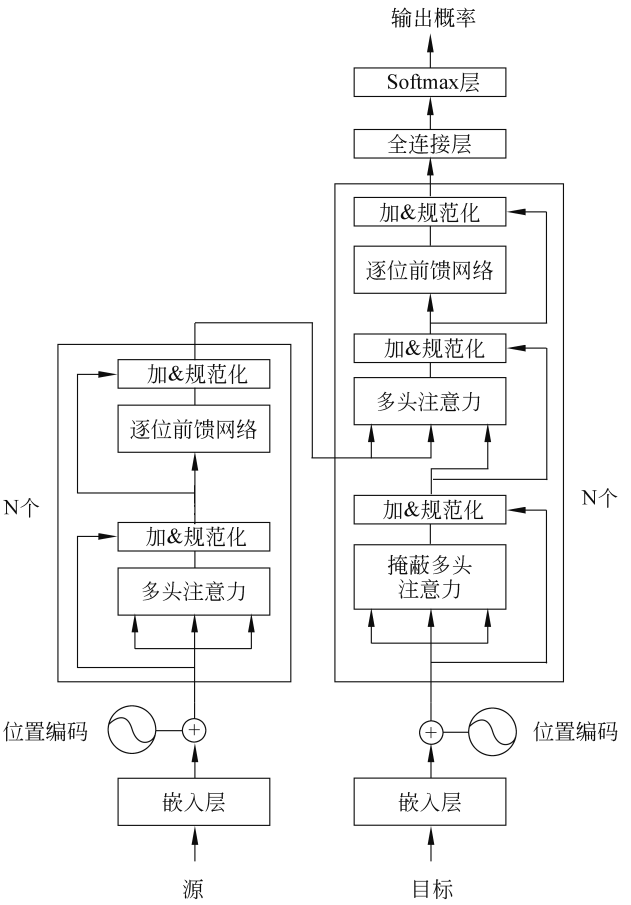

图 5-1　Transformer 架构

5.1.1　Transformer 中的自注意机制

自注意机制致力于评估序列中每个元素与其他元素的相关性，比如预测哪些词汇可能在句子中共同出现。作为 Transformer 架构的关键组成部分，自注意力使得模型能够通过捕捉序列内所有元素之间的相互作用，有效应用于各种结构化预测任务中。它主要通过汇总整个输入序列的全局信息，进而实现对序列中每个元素的更新。这种机制通过考查序列内元素间的动态关联性，使得模型不仅能捕获局部特征，也能理解和利用长距离依赖，进而增强模型对复杂序列数据的理解能力。假设一个含 n 个元素的序列（X_1，X_2，…，X_n），$X \in \mathbb{R}^{n \times d}$，其中 d 表示每个元素的嵌入维度。自注意力机制通过对每个元素的全局上下文信息进行编码，得到所有 n 个元素之间的相互作用。这是通过定义 3 个

可学习的权重矩阵，即查询矩阵 Queries（$W^Q \in \mathbb{R}^{d\times d_q}$）、键矩阵 Keys（$W^K \in \mathbb{R}^{d\times d_k}$）和值矩阵 Values（$W^V \in \mathbb{R}^{d\times d_v}$）来进行变换实现的。首先将输入序列 X 投影到这些权重矩阵上，得到 $\boldsymbol{Q}=XW^Q$、$\boldsymbol{K}=XW^K$ 还有 $\boldsymbol{V}=XW^V$。自注意层的输出 $Z\in\mathbb{R}^{n\times d}$ 可用公式（5-1）表示为

$$Z=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_q}}\right)\boldsymbol{V} \tag{5-1}$$

对于序列中一个给定的元素，自注意会计算所有 K 和 Q 的点积，然后使用 Softmax 操作对其进行归一化，以得到注意力分数。然后，每个元素被更新成为序列中所有元素的加权和，其中权重值根据注意力分数给出。

5.1.2　掩码自注意力

一个标准的自注意层会关注序列中所有的元素。为了训练出用来预测序列的下一个元素的 Transformer 模型，解码器中的自注意层使用掩码来屏蔽后续的元素。这是通过和掩码 $\boldsymbol{M}\in\mathbb{R}^{(n\times n)}$ 的元素级乘法运算来完成的，其中 $\boldsymbol{M}$ 是一个上三角矩阵。掩码自注意被定义为以下公式：

$$\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_q}}\circ\boldsymbol{M}\right) \tag{5-2}$$

其中，$\circ$ 代表哈达玛（Hadamard）乘积。预测序列中的一个元素时，后续元素的注意分数在掩码自注意中一般会被设置为零。

5.1.3　多头注意力

多头注意力机制通过并行地运用多个自注意层增强了基础自注意层的性能，允许模型在不同的表示子空间中捕获信息。与单头注意力相比，它解决了对特定位置关注的限制，使模型能够同时关注序列中多个位置的信息。这是通过为每个头分配一组唯一的查询、键和值权重矩阵实现的，这些矩阵在初始训练阶段通过随机初始化，并在训练过程中学习，以将输入向量映射到不同的表示子空间。通过这种方式，多头注意力能够丰富模型的表示能力，提高对序列数据的理解深度和复杂性处理能力。

为了更清晰地说明这一步骤，假设给定一个输入向量和头部数量 h，首先将输入向量转换为 3 组不同的向量：Query、Key 和 Value。每一组中都存在 h 个向量，它们的维数为 $d_{q'}=d_{k'}=d_{v'}=d_{\mathrm{model}}/h=64$。然后，来自不同输入的向量被打包成 3 组不同的矩阵：$\{Q_i\}_{i=1}^h$、$\{K_i\}_{i=1}^h$ 和 $\{V_i\}_{i=1}^h$。其过程如下所示：

$$\mathrm{MultiHead}(\boldsymbol{Q}',\boldsymbol{K}',\boldsymbol{V}')=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_n)\boldsymbol{W}^{\mathrm{o}} \tag{5-3}$$

其中，$\boldsymbol{Q}'$ 由 $\{Q_i\}_{i=1}^h$ 拼接得到，$\boldsymbol{K}'$ 和 $\boldsymbol{V}'$ 也类似，而 $\boldsymbol{W}^{\mathrm{o}}\in\mathbb{R}^{d_{\mathrm{model}}\times d_{\mathrm{model}}}$ 是投影的权重矩阵。

5.1.4 前馈网络

在原生的 Transformer 架构中，前馈网络被应用在每个编码器和解码器的自注意层之后。它由两个线性变换层和一个非线性激活函数组成。可以用公式(5-4)表示：

$$\mathrm{FFN}(X)=\boldsymbol{W}_2\sigma(\boldsymbol{W}_1X) \tag{5-4}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为两个线性变换层的两个参数矩阵，σ 为非线性激活函数。

在 Transformer 的架构中，在编码器和解码器中的每一层都进行了残差连接，这可以加强信息传递，实现更高的性能。接着，在残差连接之后进行层归一化。经过这些操作后得到的输出可以表示为公式(5-5)：

$$\mathrm{LayerNorm}(\boldsymbol{X}+\mathrm{Attention}(\boldsymbol{X})) \tag{5-5}$$

其中，$\boldsymbol{X}$ 为自注意层的输入，Query、Key 和 Value 矩阵都来自相同的输入矩阵 $\boldsymbol{X}$。

解码器中的最后一层是用来将向量栈转换回一个单词。这是通过一个线性层和一个 Softmax 层来实现的。线性层将该向量投影到一个具有 d_{word} 维数的对数向量中，其中 d_{word} 是词汇表中的单词数。然后使用 Softmax 层将对数向量转换为概率。

5.1.5 计算机视觉中的 Transformer

受到 Transformer 模型在 NLP 领域成功的启发，研究人员开始探索其在图像处理任务中的应用潜力。图像处理任务的复杂性，尤其是其高维度、噪声水平和冗余信息的特性，使得图像生成建模成为一项挑战。Dosovitskiy[154] 等提出了(ViT)，这是一种创新的尝试，将图像分割成一系列 16 像素×16 像素的补丁，类比于 NLP 中的词汇，进而将这些补丁视作序列数据输入到 Transformer 结构中，如图 5-2 所示。ViT 的提出是 Transformer 模型在计算机视觉领域应用的重要里程碑，并激发了大量后续研究。该模型尽量保持了原始 Transformer 设计的核心原则，其架构展示了如何有效地将 Transformer 应用于视觉领域的任务中。

为了处理二维图像，输入图像 $X\in\mathbb{R}^{h\times w\times c}$ 被重塑为一个扁平的二维补丁 $X_p\in\mathbb{R}^{n\times(p^2\cdot c)}$ 序列，其中 c 为通道数。(h,w) 为原始图像的分辨率，(p,p) 为每个图像补丁的分辨率。因此，ViT 的有效序列长度为 $n=hw/p^2$。因为 Transformer 模型在其所有层中使用恒定的宽度，因此线性投影将图像补丁映射为固定长度的向量输入 Transformer，此操作称为补丁嵌入。输入 Transformer 时，会在嵌入补丁的序列中应用一个可训练的向量，用于对图像分类。此外，可训练的位置向量也会被添加到补丁嵌入中，以保留位置信息。值得注意的是，ViT 只使用了标准 Transformer 的编码器部分。

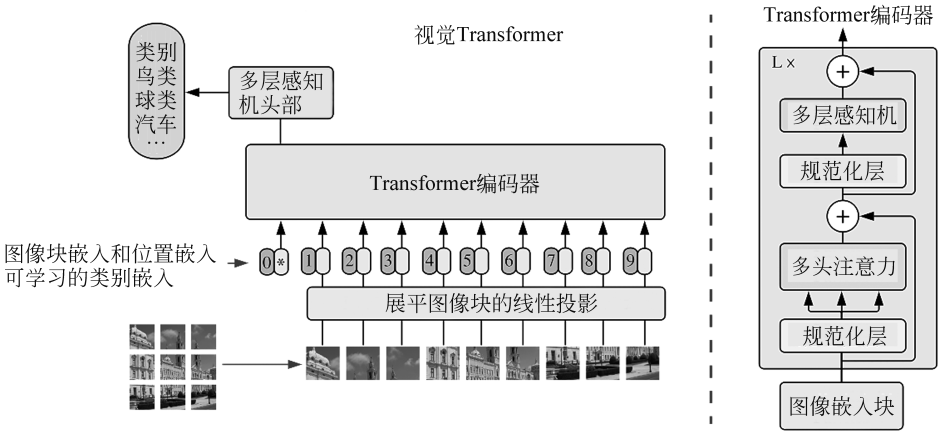

图 5-2　视觉 Transformer

根据 ViT 的范式，一系列不同的 ViT 变体被提出来，以提高视觉任务的性能。主要方法如下。

1. 增强局部性处理能力

原始 ViT 模型虽擅长捕捉图像区块间的远程依赖关系，但在局部特征提取方面表现不足，仅将二维区块直接映射为一维向量，忽略了局部特征的重要性。近期，学界开始着力增强模型对局部信息的处理能力。例如，Transformer-in-Transformer（TNT）[117]模型进一步细分每个图像补丁为更小的子块，并通过内部 Transformer 与外部 Transformer 相结合的方式增强了对局部与全局信息的处理。此外，T2T[118]等方法通过局部特征聚合进一步提升局部信息的建模效果，这些进展体现了在 ViT 结构中融合局部与全局信息交换的重要性。

2. 自注意力机制的改进

自注意力层是 Transformer 架构核心，实现图像区块间的全局互动。近期研究致力于优化此机制，以提高效率和效果，如 DeepViT[119]通过建立不同注意力头之间的交流，增强层间互动和注意力多样性。XCiT[120]则是通过特征通道而非单独区块进行自注意力计算，实现高效处理高分辨率图像。自注意力机制的计算复杂性和效果精度是当前和未来研究的关键点。

3. 网络架构设计创新

在 CNN 领域，网络架构设计是关键因素之一。虽然 ViT 最初仅采用了重复的

Transformer 块堆叠，但现代 ViT 架构设计已经成为研究热点。例如，采用金字塔结构的 PVT[121] 和 PiT[122]，以及通过神经架构搜索发掘优化 Transformer 结构的 Scaling-ViT[123] 和 GLiT[124]，均显示了借鉴 CNN 设计经验对视觉 Transformer 创新的价值。

5.2 基于 CNN 和 Transformer 的低剂量 CT 图像去噪网络

5.2.1 整体网络架构

近年来，尽管基于 CNN 的图像去噪技术已取得显著进展，但研究主要集中于利用卷积层提取特征信息以及优化网络结构。然而，这类方法存在局限，特别是卷积层仅能处理局部区域信息，限制了其在提取高级特征时的能力，导致了对多层卷积网络的依赖。

近期，基于 Transformer 的研究在 NLP 领域迅速增加，并已成为该领域的主流方法。鉴于其卓越的性能，这种方法也引起了计算机视觉领域研究人员的广泛关注。尽管如此，无论是基于 CNN 还是基于 Transformer 的现有方法均存在缺陷。正如先前所述，基于 CNN 的技术由于受到感知范围的限制，在提取包含长期空间依赖性的上下文信息方面能力不足。而基于 Transformer 的方法在捕捉细节信息方面表现不佳，可能会影响医生对病变的诊断判断。据此，本章提出一种结合 CNN 和 Transformer 优势的 LDCT 图像去噪策略，旨在有效利用图像的细节与全局信息。

众所周知，U-Net 模型已被广泛证实为一种有效的图像信息提取框架。本章设计的方法借鉴了 U-Net 的结构，采用了优雅的 U 型设计，如图 5-3 所示，由编码器、瓶颈层、解码器以及跳跃连接构成。编码器基于卷积块构建，而解码器核心采用本章创新性提出的双路径 Transformer。首先，输入带噪声的图像至编码器，通过卷积块提取特征，特征信息既向下一层传递，又通过跳跃连接传入深层。跳跃连接将浅层特征直接传递至对应深层，以增强局部信息提取效率，其中引入多尺度空间注意力模块，以进一步优化性能。为了将特征传入 Transformer 模块，采用补丁嵌入技术将图像分割成 4×4 的非重叠补丁，通过设置与补丁大小相同的卷积步长和核心，定义输出通道以确定嵌入向量大小，随后展开 H、W 维度并置于首维。鉴于 Transformer 深度的增加，可能导致计算成本剧增且难以收敛，本章仅在瓶颈层和解码器中使用两个连续的 Transformer 块进行特征提取。瓶颈层使用标准 Transformer 模块处理编码器输出的特征，而解码器的每一层则由两个连续的双路径 Transformer 组成，分别处理跳跃连接的特征和上一层的输出。上采样通过补丁扩展技术重构特征图，实现原始分辨率的 2 倍上采样。最终，通过一个 4 倍上采样的补丁扩展层恢复输入分辨率，并通过线性投影层对上采样特征进行处理，以产出图像去噪预测结果。

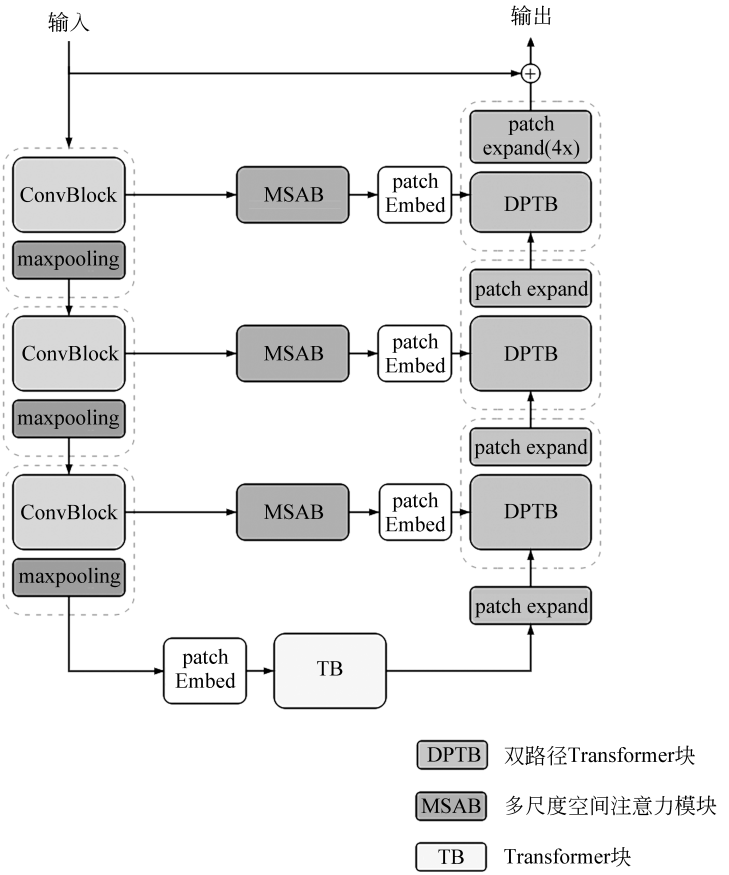

图 5-3　基于 CNN 和 Transformer 的低剂量 CT 图像去噪整体网络架构(见彩图)

5.2.2　基于 CNN 的编码器

在本章的网络架构中,CNN 被选为编码器路径的特征提取器,以便于生成输入的特征映射,目的是在网络底层中最大限度地保留图像的局部细节信息。这种设计选择允许我们在解码路径中有效地利用中间层的高分辨率 CNN 特征图,通过跳跃连接将这些特征图传递给解码路径,从而实现细节信息的精准还原。

具体而言,编码器路径沿用了 U-Net 网络的标准架构,由 3 个连续的卷积块和下采样层组成。每个卷积块内包含两个卷积层及其后接的 ReLU 激活层,卷积核大小设为 3×3,步长为 1,填充为 1,以保持特征图的空间维度。卷积层的通道数分别为 96、192 和 384,以逐层加深网络的特征提取能力。下采样则采用 2×2 的最大池化操作,步幅设置为 2,旨在减少特征图的空间尺寸,同时增加特征的抽象层级,为捕捉更广泛的上下文信息做

好准备。

5.2.3 多尺度空间注意模块

在本章的网络设计中，通过在编码器和解码器之间引入跳跃连接，实现了相同级别层的直接连接。跳跃连接主要有两个显著的优点：首先，它有助于缓解梯度消失问题，从而保证了即便是深层网络也能够稳定训练。其次，尽管随着编码过程中的逐步下采样可能导致图像信息的丢失，进而影响解码器在细节信息恢复上的能力，跳跃连接通过将编码器阶段的特征图信息直接融合至对应的解码器层，能够在一定程度上补偿这种信息损失，尤其增强结构细节的补充。

对于 LDCT 去噪问题，保持图像中的细节信息至关重要。医生分析 CT 图像时，细致的病灶区域细节对于准确诊断疾病状态非常关键。因此，我们强调在 CT 图像去噪过程中，必须给予图像的详细特征和局部信息以特别关注，以确保诊断的准确性和可靠性。

在去噪任务中，维持图像结构的完整性以及防止细节损失和边缘平滑是一项主要挑战，特别是在医学图像处理领域，器官和组织的结构纹理信息对于诊断至关重要。以肺部 CT 图像中的空洞征象为例，这类细节信息是诊断肺癌的重要依据。常规的 CNN 处理此类问题时存在局限性，主要原因在于卷积操作的局部性及下采样操作的限制。卷积操作因尺寸固定，仅能处理等尺寸的特征，缺乏全局性的信息关联。同时，池化操作导致的特征图分辨率变化限制了对不同尺寸对象信息的准确提取。

多尺度信息，即不同层次上生成的特征图，对于细粒度图像细节的恢复具有重要意义。低层次的特征图保留了较高分辨率的图像几何信息，如边缘和纹理，尽管其感受野较小导致缺乏足够的上下文信息处理能力。而深层特征虽然在噪声鲁棒性和上下文信息提取方面表现出色，但较低的分辨率可能破坏图像结构。因此，网络结构设计需要综合考虑这些特点，利用多尺度特征的优势。U-Net 及其编码器—解码器结构便是一个融合多尺度特征，以提升性能的典型例子。尽管如此，同一层次上编码器和解码器获取的特征信息并非完全相同，而简单地通过跳跃连接融合浅层与深层信息并不能保证优化后效果的最佳化。这暗示了在融合多尺度特征时需采取更精细的策略，以确保最终图像的去噪效果，同时保留重要的结构和细节信息。

在本章中，我们提出了一种基于多尺度信息的空间注意力模块，旨在通过空间注意力机制增强输入特征中关键区域的权重，从而更有效地恢复这些区域的细节信息。该模块的设计包含以下几个关键步骤：首先，对跳跃连接后的输入特征在通道维度上进行压缩，具体操作为执行最大值池化和平均值池化，分别提取通道上的最大值和平均值，以生成两个通道数为 1 的特征图。这两个特征图随后被合并，形成通道数为 2 的特征图，该特征图再通过卷积滤波器处理，以降低通道数至 1。此后，应用 Sigmoid 激活函数产生权重图，该权重图与模块的初始输入相乘，以实现空间注意力加权。

为了捕获不同尺度的特征信息，该模块采用了 3 种不同大小的卷积滤波器，分别是 3×3、5×5 和 7×7，以此构成一系列连续的空间注意力模块。通过这样的设计，模块能够提取和强化来自不同空间范围内的特征信息，形成一个多尺度空间注意力模块。该模块的结构如图 5-4 所示，它通过对特征图进行精细的空间加权处理，旨在优化医学图像去噪过程中的细节信息恢复，从而有助于保持图像结构的完整性并提高去噪效果。

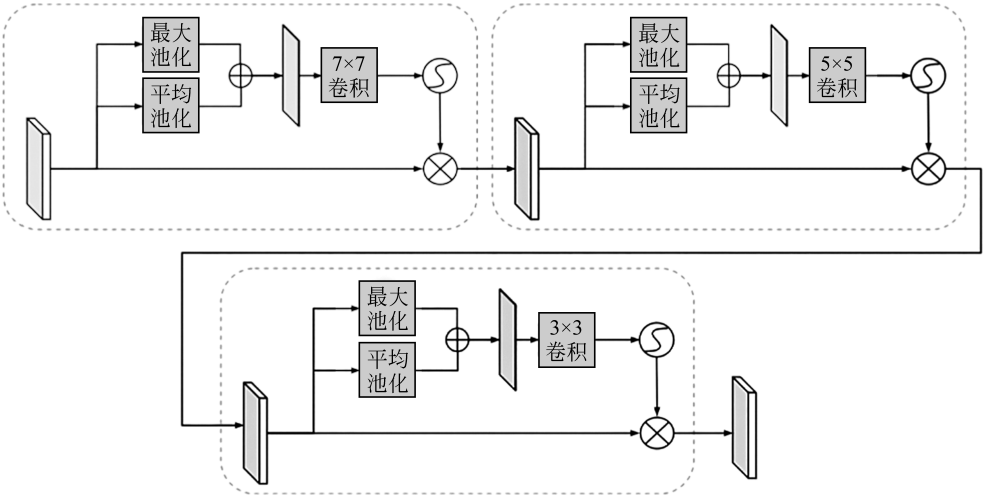

图 5-4　多尺度空间注意模块

5.2.4　双路径 Transformer 模块

将 Transformer 模型从语言处理领域成功迁移到视觉任务中的一个主要挑战源于语言与视觉信息处理之间的本质差异。在语言处理中，Transformer 模型以单词令牌作为基本处理单元，这些令牌具有固定的规模和相对较小的变化范围。相比之下，视觉信息以像素为基础，像素的范围和复杂性远大于文字令牌，特别是在处理图像时，像素分辨率显著高于文本数据。这导致基于 Transformer 的模型面临两个主要问题：首先，固定规模的令牌不适应视觉领域中的变化范围；其次，图像的高分辨率使得像素级任务（如语义分割、图像去噪）的处理变得尤为困难。特别是，Transformer 模型中的自注意力机制需要计算每个像素与图像中所有其他像素之间的关系，这在高分辨率图像上导致计算复杂度与图像大小呈平方级增长。

针对这一挑战，学者提出了一种名为 swin Transformer 的通用 Transformer 主干结构。swin Transformer 通过构建层次化的特征映射，以及计算复杂度与图像大小呈线性关系的策略，有效应对了上述问题。不同于 ViT 模型在自注意力计算中包含每个像素点与其他所有像素点的相关性，swin Transformer 采用了一种渐进式的策略，从小尺寸图像

块开始，逐步在更深层次合并相邻的图像块，以此构建层次化的表示。通过在分割的图像的非重叠窗口内局部计算自注意力，实现了计算复杂度与图像大小的线性关系，其中每个窗口内包含的补丁数量固定。图 5-5(a)展示了 swin Transformer 的结构，而(b)图对比展示了 ViT 的结构，直观地说明了 swin Transformer 如何通过局部注意力和层次化策略优化计算效率与性能。

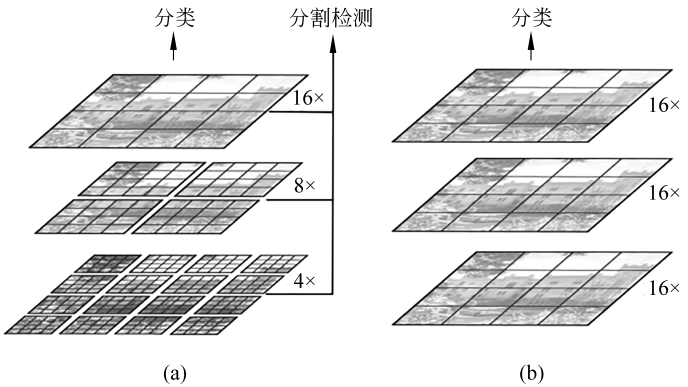

图 5-5　swin Transformer 和 ViT 对比

swin Transformer 的独特之处在于其对连续自注意层中窗口分区进行的移位操作，如图 5-6 所示。通过移位窗口，swin Transformer 能够连接相邻层中的窗口，从而实现层间的有效信息流通，显著增强模型的建模能力。这种移位策略确保了窗口内的所有查询补丁与相同的键集共享信息，有利于优化硬件中的内存访问效率，其中查询和键代表自注意层中的投影向量。与此相对，早期采用滑动窗口的自注意方法因查询像素对应的键集各不相同，导致在实际应用中难以高效地执行内存访问。通过这种创新的设计，swin Transformer 在保持计算效率的同时提高了模型对图像细节的捕获能力，为视觉领域的各种任务提供了一种强大的模型基础。

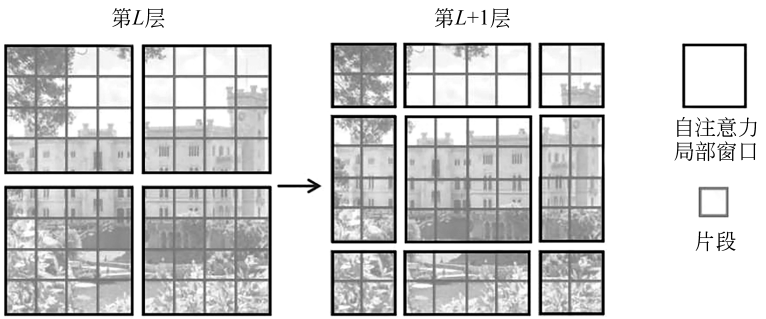

图 5-6　swin Transformer 的移位窗口操作