Lightweight Image Super-Resolution with Superpixel Token Interaction

摘要

基于Transformer的方法在单图像超分辨率（SISR）任务上取得了令人印象深刻的结果。然而，当应用于整个图像时，自注意机制在计算上是昂贵的。因此，目前的方法将低分辨率的输入图像分成小块，这些小块被单独处理，然后融合以生成高分辨率图像。然而，这种常规的规则块划分过于粗糙并且缺乏可解释性，导致在注意操作期间的伪影和非相似结构干扰。为了应对这些挑战，我们提出了一种新的超级令牌交互网络（SPIN）。我们的方法采用超像素聚类局部相似像素，形成可解释的局部区域，并利用超像素内的注意，使局部信息的交互。它是可解释的，因为只有相似的区域相互补充，不相似的区域被排除在外。此外，我们设计了一个超像素交叉注意模块，以促进信息传播通过替代的超像素。大量的实验表明，所提出的SPIN模型表现出优于最先进的SR方法的准确性和轻量级。

关键词：super-resolution，transformer，superpixel

1 引言

单图像超分辨率(SISR)是计算机视觉中的一项重要任务，其目的是提高低分辨率图像的分辨率和视觉质量。SISR的目标是从给定的低分辨率图像（LR）图像生成高分辨率(HR)图像，这在需要高质量图像的应用中特别有用，例如医学成像、监视和数字摄影。

自Dong等人的开创性工作以来 [6]，已经开发了许多神经网络来解决从低分辨率输入重建高质量图像的挑战。一些基于CNN的方法使用更深入和更复杂的架构来实现更好的性能。然而，这些方法带来了增加的计算资源和更高的成本的权衡，这可能限制其应用场景。

注意机制 [30]已经被证明对高水平视觉任务和低水平领域都有显著的影响，包括超分辨率(SR)。注意机制允许网络选择性地聚焦于输入的相关区域，这可以提高SR输出的质量。利用注意机制，转换器已被应用于SR任务，如SwinIR [19]和ESRT [22]。这些模型突出了全局特征提取能力在SISR中的重要性。此外，为了提高效率，ELAN [34]提出了一种分组自我注意模型，并在计算块之间的关联度时共享权重。然而，注意机制具有较高的计算复杂度和内存消耗，需要将大图像分成小块进行单独处理。虽然这种策略提高了基于变压器的模型的效率，但也带来了一些问题。基于固定形状划分块会导致连续结构的分裂，这意味着需要使用其他区域中的相似信息来增强图像细节。此外，应用在每个块内的局部注意机制在计算中涉及不相关的区域，导致不希望得到的推断。

为了解决这些问题，我们提出了一种新的方法，该方法将局部和全局注意机制与精细的超像素划分相结合。我们首先对输入图像的像素进行基于CNN的浅层特征提取，然后进行局

部聚类，将相邻像素分组为超像素。然后通过基于相似度的超像素聚类得到局部区域，并分别对其进行局部特征提取。不同于以前使用固定形状块划分的方法 [19, 34]，这种方法只用于提高并行计算效率，我们的区域划分策略更具解释性，允许对输入图像进行更灵活和自适应的划分，并防止连续结构的分裂。然后，我们引入了超像素交叉注意模块，通过超像素的替代来实现远程信息交互。此外，我们还设计了一种应用于超像素像素的超像素内注意(ISPA)机制，将原来的注意操作扩展到常规图像区域。这确保了局部注意机制信息交互发生在相似的区域，消除了干扰和无关的计算。这两种注意力机制相互交错，在局部和全局特征提取中相互配合。如图 1所示，所提出的自旋在峰值信噪比和模型尺寸之间有很好的权衡。我们的贡献概述如下：

(A)我们提出了一种新的超分辨率模型，该模型将超像素聚类与变换结构相结合，得到了一个更具解释力的框架。

(B)我们提出了超像素内注意力(ISPA)和超像素交叉注意力(SPCA)模块，它们在超像素内部和超像素之间运行，在保持捕获远程依赖的能力的同时，允许在不规则区域进行计算。

(C)实验表明，与现有的轻量级随机共振方法相比，该方法具有更好的随机共振重建性能。

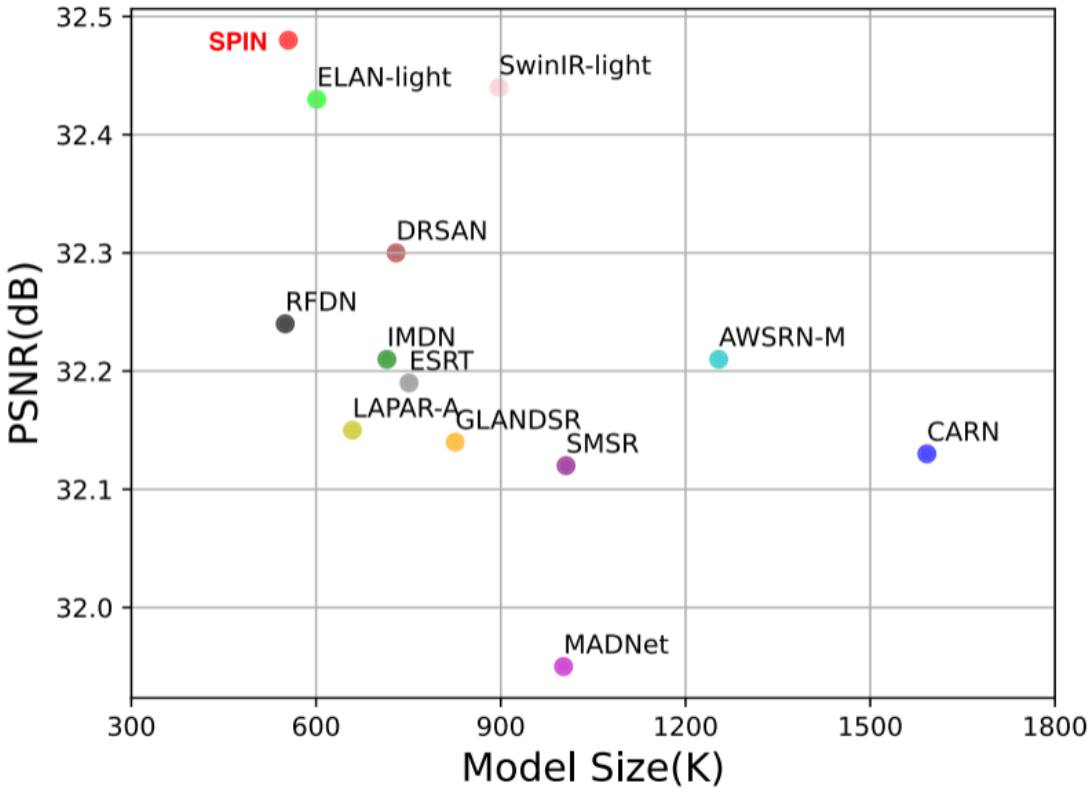

图 1. Set5上×4超分辨率的PSNR和模型参数

2 相关工作

2.1 超分辨率深度神经网络

随着深度学习的最新进展，基于神经网络的方法已成为单图像超分辨率（SR）的主流解决方案。SRCNN [6]使用三层CNN网络从其双三次下采样的低分辨率（LR）图像重建高分辨率（HR）图像。为了进一步提高准确性，最近基于CNN的方法采用了更复杂和有效的结构。例如，Kim等人 [16]应用基于CNN的深度架构和残差学习来提高SR准确性。注意机制也被引入到SR中，以提取最重要和信息量最大的特征。例如，Zhang等人 [35]使用通道注意力机制，而Hu等人 [11]将空间注意力与SR中的通道注意力结合起来。此外，受ViT [7]在高级视觉任务中的成功启发，Chen等人 [5]将Transformer引入SR，但它需要大量参数。为了减小模型大小，SwinIR [19]将Swin Transformer [21]框架应用于SR，将整个图像划分为固定大小为8×8的小窗口，并在应用多头注意力机制时移动窗口。虽然上述这些方法在提取信息特征方面是有效的，但它们可能需要大量参数。

2.2 轻量级超分辨率方法

轻量级是深度SR模型的关键考虑因素，并且已经提出了许多方法来提高其效率。例如，FSRCNN [6]和ESPCN [27]利用后上采样技术来减少计算负担，而卡恩 [1]使用组卷积和级联机制来提高效率，但性能受损。IMDN [13]应用三步蒸馏来提取特征和切片操作来划分提取的特征，但带来了可扩展性。LatticeNet [23]引入了计算复杂度低的格块。BSRN [18]设计了一个深度可分离卷积来降低模型复杂度，并利用注意力机制来提高SR重建性能。同时，提出了基于轻量级transformer的SR方法来降低模型复杂度，例如，通过使用基于窗口的注意力[19]并采用移位卷积和分组自注意力 [34]来减少计算的令牌。虽然这些方法是轻量级和高效的，SR重建的质量仍然保持改进的空间。

2.3 图像处理中的像素聚类

像素聚类是图像处理中一项得到充分研究的任务，深度学习方法的最新进展在这一领域取得了重大进展。像素聚类的一种常见方法是使用CNN来生成像素级嵌入，将相似的像素分组在一起。例如，Liu等人 [20]开发了一个深度亲和网络，它学习像素级的亲和度来聚类像素。类似地，Sun等人 [28]提出了一种学习像素级表示的网络，以聚类图像块。除了使用CNN来生成像素级嵌入之外，聚类算法还可以应用于CNN特征，将相似的像素分组到聚类中。J 'egou等人 [15]介绍了一种一次性聚类方法，该方法使用CNN特征生成初始聚类，然后使用聚类算法进一步细化。Li等人 [17]提出了一种弱监督聚类方法，该方法使用CNN特征和稀疏标记方案将像素聚类到对象区域。这些方法利用了CNN和聚类算法的强大功能，在图像处理任务中实现了更准确、更高效的像素聚类。最近，人们对使用图卷积网络（GCN）进行像素聚类越来越感兴趣。GCN能够通过构建图像的图形表示来对图像中像素之间的依赖性进行建模，其中每个像素是一个节点，并且边缘表示像素之间的关系。与传统的CNN相比，这使得GCN能够捕获像素之间更复杂和非局部的相互作用。例如，Zeng等人 [32]提出了一种基于GCN的高光谱图像分类框架，该框架使用两种聚类策略来利用多跳相关性。第一种聚类策略根据相似像素的光谱相似性对相似像素进行分组，而第二种聚类策略根据像素的空间相邻

性对像素进行分组。虽然像素聚类在各种图像处理任务中表现出了良好的效果，但它尚未有效地应用于超分辨率应用。

3 本文方法

所提出的模型的架构如图2所示，其主要由所提出的超像素交互（SPI）块组成。在SPI块之前，我们利用一个编码器，这是一个3 × 3卷积，嵌入低分辨率图像I_{LR}到一个高维特征空间。给定编码器，我们可以得到浅特征x_{emb}：

$$x_{emb} = f_{encoder}(I_{LR})$$

其中$f_{encoder}$表示所提出的模型的编码器。

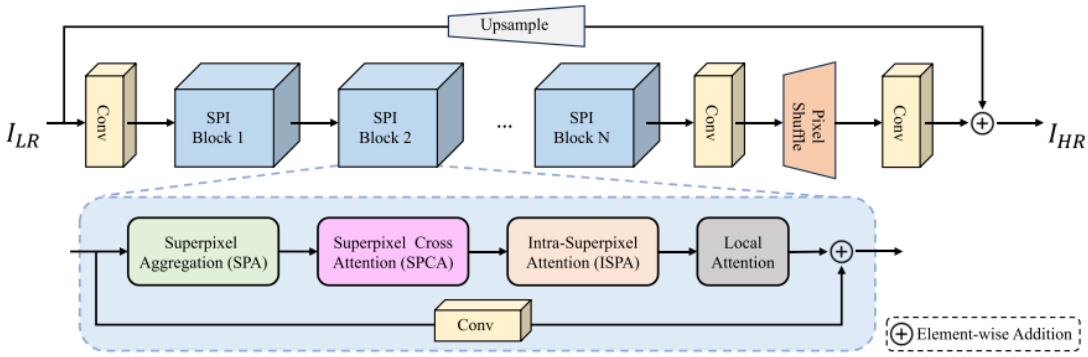

图 2. SPIN网络结构图

然后，我们在编码器上堆叠K个SPI块，以提取包含输入图像丰富的低级和高级信息的更深层次的特征。每个SPI块包括四个组件：超像素聚合（SPA）、超像素交叉注意（SPCA）、超像素内注意（ISPA）和局部注意。

每个块的输入特征首先通过SPA模块聚合成超像素。然后，ISPA模块捕获每个超像素内像素的依赖性和相互作用，而SPCA模块捕获长程像素之间的依赖性和相互作用。为了增强局部区域内像素之间的相互作用，我们在ISPA和SPCA模块之后利用了局部注意力模块，该模块使用基于窗口的注意力[24，20，21]。我们使用重叠的补丁，以加强功能的相互作用。形式上，对于第i个SPI块，整个过程可以公式化为：

$$s_i = f_{SPA}(x_{i-1})$$

$$x_i = x_{i-1} + f_{local}(f_{ISPA}(f_{SPCA}(x_{i-1}, s_i)))$$

其中s_i表示第i个SPI块中的超像素的特征，$f(\cdot)$表示每个单独分量的函数。在前人工作的基础上，利用剩余连接来简化整个训练过程。

在K个SPI块之后，我们采用3×3卷积层和像素混洗操作[34]来获得全局残差信息，将其添加到I_LR的上采样图像中以解决高分辨率图像I_SR。

3.1 SPA模块

与以往将输入图像或特征划分为规则块的方法不同，我们提出将输入特征划分为超像素。与可以容易地将连接区域裁剪成不同块的规则块相比，超像素分区可以在感知上将相似像素分组在一起，这可以描绘更精确的边界，降低生成模糊和不准确边界的风险。

与以往将输入图像或特征划分为规则块的方法不同，我们提出将输入特征划分为超像素。与可以容易地将连接区域裁剪成不同块的规则块相比，超像素分区可以在感知上将相似像素分组在一起，这可以描绘更精确的边界，降低生成模糊和不准确边界的风险。

具体来说，在超像素聚合的过程中，我们在SSN中利用基于软k均值的超像素算法 [14]。给定视觉标记$x \in R^{N\times C}$（其中N = H ×W是视觉标记的数量），假设每个标记$x(i) \in R^C$属于M个超像素$s \in R^{M\times C}$中的一个，使得有必要计算视觉标记和超像素标记之间的关联。

从形式上讲，超像素聚合的过程是一个类似期望最大化的过程，它包含总共T次迭代。首先，如图 3所示，我们通过对规则网格中的令牌进行平均来对初始超级令牌s_0进行采样，称为Patchify。假设网格大小为$H_S \times W_S$，则超级令牌的数量为$\frac{H\times W}{H_S\times W_S}$。对于第t次迭代，我们计算关联映射为：

$$A^t(ij) = e^{-\|x(i)-s^{t-1}(j)\|_2^2}$$

其中$A^t \in R^{N\times M}$是关联映射，$A^t(ij)$是第i行第j列的值。请注意，超像素聚合仅计算从每个标记到周围超像素的关联映射，这保证了超像素的局部性，使其在计算和内存方面也很有效[13]。

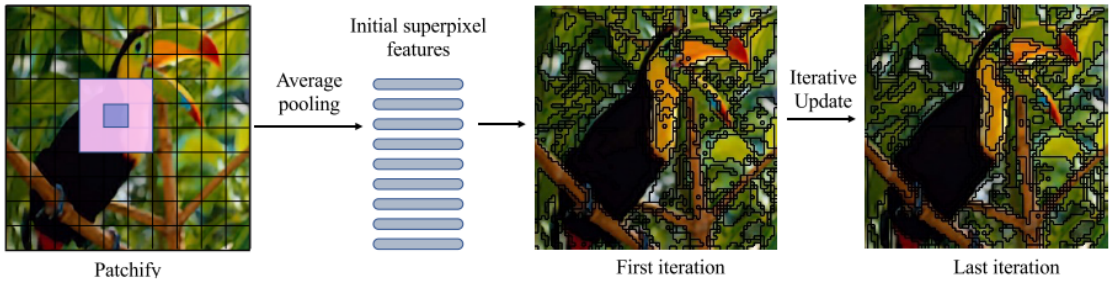

图 3. 超像素计算模块

之后，我们可以获得超像素s^t作为视觉标记的加权和，定义为：

$$s^t(j) = \frac{1}{z^t(j)} \sum_i A^t(ij)x(i)$$

其中$z^t(j) = \sum_i A^t(ij))$表示沿列的归一化项沿着。经过T次迭代，我们可以获得最终的关联映射A^T。为了简单起见，我们在下面的部分中省略了上标。

3.2 SPCA模块

由于超像素仅捕获局部区域中像素的局部性和互连，这可能缺乏捕获超分辨率的长程依赖性的能力。在这里，我们利用自我注意范式[37]通过超像素的替代来增强远程通信，这有助于利用特征之间的互补性来生成高质量的超分辨率图像。由于像素特征与所属的超像素特征高度相似，因此使超像素成为尽可能多地在像素之间传播信息的有希望的替代。

如图 4所示，给定超像素特征$s \in R^{M \times C}$，其中M表示超像素的数量，并且平坦化的像素特征$x \in R^{HW \times C}$。我们采用注意力机制[37]首先将像素信息传播到超像素。具体来说，我们使用线性投影来计算查询,键，值，如下所示：

$$Q^s = sW_q^s, K^x = xW_k^x, V^x = xW_v^x$$

其中$W_q^s \in R^{C \times D}, W_k^s \in R^{C \times D}, W_v^s \in R^{C \times C}$分别是根据查询、键和值的权重矩阵。可以通过首先计算查询和键之间的相似度并将其用作权重来聚合值来获得输出，其可以公式化为：

$$s_u = softmax(Q^s(K^x)^T/\sqrt{D})V^x$$

其中，$\sqrt{D}$是避免梯度消失的缩放因子，s_u是更新的超像素特征。请注意，与超像素聚合不同，此过程不考虑邻居限制，确保了远程信息的传播。

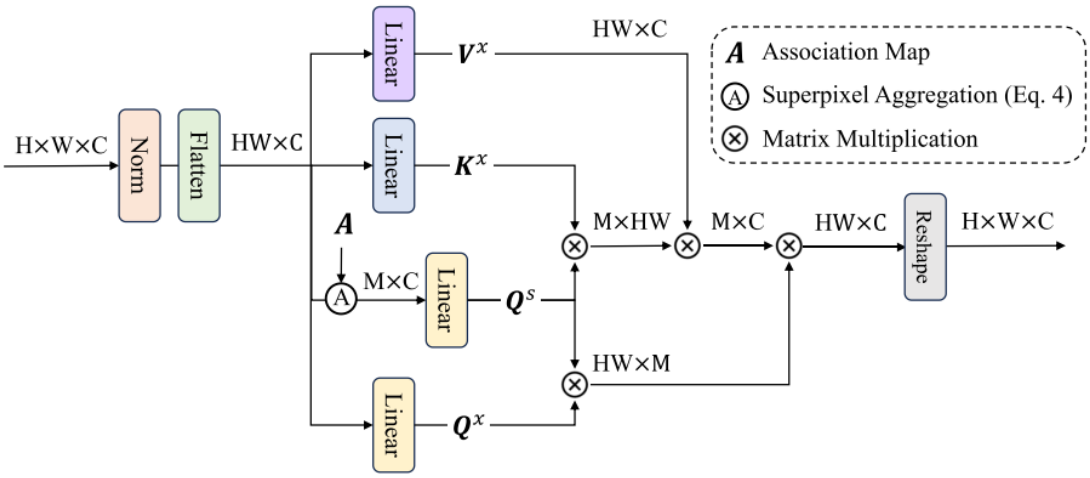

图 4. SPCA模块

一旦信息已经从像素传播到超像素，就有必要将聚合的信息分布回像素，以便实现像素之间的信息传播。在这里，我们进一步使用注意力机制。具体来说，我们利用另一个权重矩阵W_q^x从像素特征中获得查询。为了减少参数的数量，我们直接使用超像素特征Q^s作为键，更新的超像素特征作为值，并利用交叉注意将更新的超像素特征映射回像素级。与初始的Transformer块类似，我们也在上述过程后采用前馈网络（FFN）。我们的FFN包含一个层规范化 [2]层，之后我们利用特征门控 [26]来调制输入特征和通道注意力 [10]来提取全局信息。然后，使用两个全连接层和GELU [9]激活函数。

3.3 ISPA模块

给定关联图，提高超分辨率图像质量的直观方法是利用同一超像素内相似像素的互补性。为了实现这一点，我们需要获得每个超像素的对应像素。然而，不同的超像素可能包含不同数量的像素，这使得难以进行并行处理，并且还导致意外的存储器消耗，因为总是存在一些包括大量像素的超像素。

为了解决这个问题，如图 5所示，我们求助于关联图A^T并选择与每个超像素最相似的前N个像素。假设一个超像素的附属像素为$f = \{x(i)\}_N \in R^{N \times C}$，其中N表示所选像素的数

量。我们遵循标准的自我注意机制 [30]，即，当等式5和6进行超像素内关注，其包括用于查询、键和值投影的权重矩阵W_q^f, W_k^f, W_v^f。在超像素内交互之后，我们利用前N个选择过程中生成的索引，将细化的像素特征分散回图像中各自的位置。

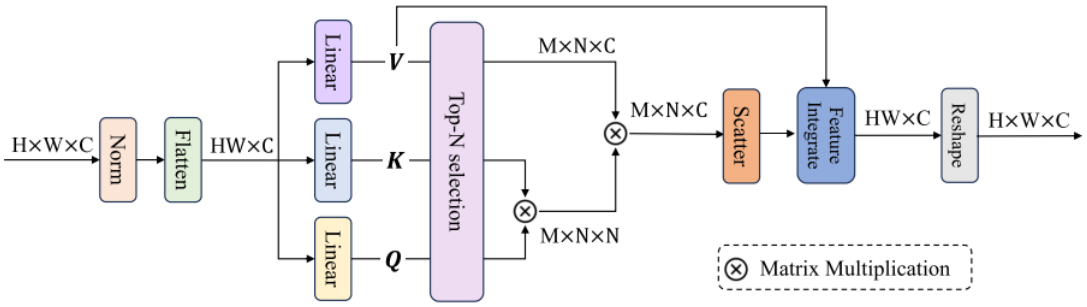

图 5. ISPA模块

前N个选择可能导致一些“被忽略”的像素，即，这些像素不包括在任何超像素中。对于那些“被忽略的”像素，我们利用值投影W_v^f来投影它们以获得更新的特征，然后将其与通过像素内交互更新的那些像素集成。与SPCA模块类似，我们在ISPA模块之后采用相同的FFN。

4 复现细节

4.1 与已有开源代码对比

小波变换是一种有效的时频分析方法。考虑到小波变换的可逆性和保留全部信息的能力，小波变换在CNN结构中被用来提高各种视觉任务的性能。例如，Bae等人 [3]验证了在小波子带上学习CNN表示有助于图像恢复的任务。DWSR [8]将低分辨率的小波子带作为输入来恢复图像超分辨率任务中丢失的细节。利用多层小波变换在不丢失信息量的情况下扩大图像的接收范围，用于图像恢复。原论文中最后一个注意力模块使用的是常规的自注意力，所以对其进行修改以达到增强的效果，具体来说是将其更换为Wave-Vit [31]中的Wave-Attention，细节将在创新点小结详细讨论。

4.2 实验环境搭建

我们使用DIV2K [29]作为训练集，这是一个包含各种自然场景图像的高清数据集。该数据集包括900张高分辨率图像，前800张图像用于训练，最后100张图像用于验证。在RCAN [35]之后，使用双重三重下采样方法生成LR样本。此外，我们在五个常用的基准测试中评估了我们的方法，包括Set5 [4]，Set14 [33]，BSDS100 [24]，Urban100 [12]和Manga109 [25]。

在训练过程中，初始学习率被设置为5e-4，并且训练过程在1000个epoch之后停止。使用的优化器是Adam优化器，$\beta 1$为0.9，$\beta 2$为0.999。为了训练模型，我们使用随机旋转90度，180度，270度和水平翻转来进行数据增强。在最终模型中，所有块的输出通道都设置为40。我们将SPI块的数量设置为8，并采用不同的初始补丁来跨各种SPI块进行超像素聚合，跨度从12到24。

对于评价，我们主要使用常用的评价指标，包括峰值信噪比（PSNR）和结构相似性（SSIM）。我们遵循RCAN [44]，在将RGB转换为YCbCr格式后测量Y通道上的指标。

4.3 创新点

和原论文相比，做了如下改进，将原论文中局部注意力模块更改为Wave-Attention，如图6所示。

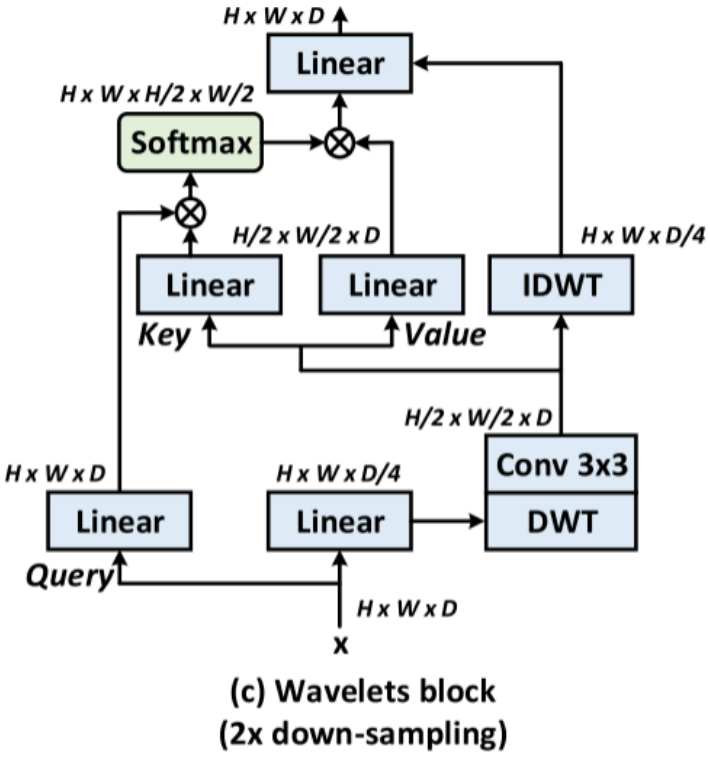

图 6. 小波注意力模块

该注意力的思想为将原始图像进行小波变换后得到尺寸为原图二分之一的高频图和低频图作为自注意力的键和值，进一步降低计算复杂度同时将小波变换后的特征图逆变换后与自注意力结果图拼接作为补充信息。同时还将Wave-Attention中的批归一化更改为适合超分辨率的层归一化，同时将注意力块扩大为原来的两倍以在计算复杂度相近的情况下增大感受野。具体代码如图 7， 8所示。

```python
class WaveAttention(nn.Module):
    def __init__(self, dim, num_heads, sr_ratio):
        super().__init__()
        self.num_heads = num_heads
        head_dim = dim // num_heads
        self.scale = head_dim**-0.5
        self.sr_ratio = sr_ratio

        self.dwt = DWT_2D(wave='haar')
        self.idwt = IDWT_2D(wave='haar')
        self.reduce = nn.Sequential(
            nn.Conv2d(dim, dim//4, kernel_size=1, padding=0, stride=1),
            # nn.BatchNorm2d(dim//4),
            LayerNorm2d(dim//4),
            nn.ReLU(inplace=True),
        )
        self.filter = nn.Sequential(
            nn.Conv2d(dim, dim, kernel_size=3, padding=1, stride=1, groups=1),
            # nn.BatchNorm2d(dim),
            LayerNorm2d(dim),
            nn.ReLU(inplace=True),
        )
        self.kv_embed = nn.Conv2d(dim, dim, kernel_size=sr_ratio, stride=sr_ratio) if sr_ratio > 1 else nn.Identity()
        self.q = nn.Linear(dim, dim)
        self.kv = nn.Sequential(
            nn.LayerNorm(dim),
            nn.Linear(dim, dim * 2)
        )
        self.proj = nn.Linear(dim+dim//4, dim)
        self.apply(self._init_weights)
```

图 7. 类初始化代码

```python
def forward(self, x, H, W):

    B, N, C = x.shape
    q = self.q(x).reshape(B, N, self.num_heads, C // self.num_heads).permute(0, 2, 1, 3)

    x = x.view(B, H, W, C).permute(0, 3, 1, 2)
    x_dwt = self.dwt(self.reduce(x))
    x_dwt = self.filter(x_dwt)

    x_idwt = self.idwt(x_dwt)
    x_idwt = x_idwt.view(B, -1, x_idwt.size(-2)*x_idwt.size(-1)).transpose(1, 2)

    kv = self.kv_embed(x_dwt).reshape(B, C, -1).permute(0, 2, 1)
    kv = self.kv(kv).reshape(B, -1, 2, self.num_heads, C // self.num_heads).permute(2, 0, 3, 1, 4)
    k, v = kv[0], kv[1]

    attn = (q @ k.transpose(-2, -1)) * self.scale
    attn = attn.softmax(dim=-1)
    x = (attn @ v).transpose(1, 2).reshape(B, N, C)
    x = self.proj(torch.cat( tensors: [x, x_idwt], dim=-1))
    return x
```

图 8. 前向传播代码

5 实验结果分析

图 9是具体的复现结果和改进后的结果,通过表中数据可知复现的比较成功，甚至在一些指标中超越了原文。但是注意到在将局部注意力更改为小波注意力的实验结果上性能出现了退化，经过仔细思考，发现小波注意力实际上是降低了参数量，在同样尺寸的Transformer块，

小波注意力键和值都只有原来的四分之一大小。虽然性能指标上并没有得到提升，但对于模型参数量降低，轻量化上取得一定效果。

	scale	set5 PSNR/SSIM	set14 PSNR/SSIM	b100 PSNR/SSIM	u100 PSNR/SSIM	manga109 PSNR/SSIM
SPIN(原文)	2	38.20/0.9615	33.90/0.9215	32.31/0.9015	32.79/0.9340	39.18/0.9784
SPIN(复现)	2	38.24/0.9615	33.92/0.9217	32.31/0.9013	32.78/0.9339	39.19/0.9785
SPIN+wave	2	38.15/0.9613	33.82/0.9201	32.26/0.9008	32.59/0.9318	39.04/0.9780
SPIN(原文)	3	34.65/0.9293	30.57/0.8464	29.23/0.8089	28.71/0.8627	34.24/0.9489
SPIN(复现)	3	34.65/0.9292	30.57/0.8460	29.23/0.8084	28.69/0.8625	34.21/0.9486
SPIN(原文)	4	32.48/0.8983	28.80/0.7862	27.70/0.7415	26.55/0.7998	30.98/0.9156
SPIN(复现)	4	32.44/0.8976	28.82/0.7855	27.70/0.7401	26.55/0.7988	30.94/0.9141

图 9. 复现及改进结果

上采样倍率为2的训练损失如图 10所示，在五个基准数据集的测试结果如图 11所示。由此可知模型训练过程中损失很快就到达了平稳的状态，很长一段时间都处于细微波动的状态，收敛花费较长时间。

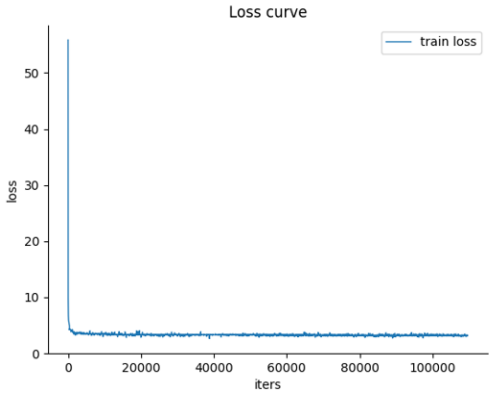

图 10. 训练损失

```
100%|                                                  | 5/5 [00:15<00:00,  3.10
100%|                                                  | 14/14 [00:32<00:00,  2.30
100%|                                                  | 100/100 [01:24<00:00,  1.19
100%|                                                  | 100/100 [05:20<00:00,  3.21
100%|                                                  | 109/109 [08:20<00:00,  4.59
[set5-X2], PSNR/SSIM: 38.24/0.9615 (Best: 38.24/0.9616, Epoch: 348/326)
[set14-X2], PSNR/SSIM: 33.92/0.9217 (Best: 33.94/0.9219, Epoch: 336/365)
[b100-X2], PSNR/SSIM: 32.31/0.9013 (Best: 32.31/0.9016, Epoch: 393/333)
[u100-X2], PSNR/SSIM: 32.78/0.9339 (Best: 32.78/0.9339, Epoch: 424/424)
[manga109-X2], PSNR/SSIM: 39.19/0.9785 (Best: 39.22/0.9785, Epoch: 395/323)
```

图 11. 测试集上结果

6 总结与展望

在本文中，我们复现了一种称为超级令牌交互网络（SPIN）的新方法，该方法利用超像素将局部相似像素分组为可解释的局部区域，采用内部像素的注意力，以促进不规则的局部超像素区域内的局部信息交互，而超像素的交叉注意力模块，通过超像素的替代促进远程信息交互。此外，所提出的方法提供了一个有前途的解决方案的挑战，处理整个图像与可解释的区域划分，未来可以寻找更好的聚类算法以寻找更好自注意力图像块以充分发挥Transformer的性能。

参考文献

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Woong Bae, Jaejun Yoo, and Jong Chul Ye. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 145–153, 2017.

[4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017.

[9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019.

[12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[13] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019.

[14] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018.

[15] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017.

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[17] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016.

[18] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–843, 2022.

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[20] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–703, 2018.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022.

[23] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 272–289. Springer, 2020.

[24] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

[25] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017.

[26] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11908–11915, 2020.

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[28] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 769–777, 2015.

[29] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision*, pages 328–345. Springer, 2022.

[32] Hao Zeng, Qingjie Liu, Mingming Zhang, Xiaoqing Han, and Yunhong Wang. Semi-supervised hyperspectral image classification with graph clustering convolutional networks. *arXiv preprint arXiv:2012.10932*, 2020.

[33] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012.

[34] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, pages 649–667. Springer, 2022.

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.