Advances in Applied Mathematics 应用数学进展, 2024, 13(4), 1842-1852
Published Online April 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.134173

Hans汉斯

基于多级轴向加性网络的轻量级单图超分辨率

邹观哲*，黄可言

南京邮电大学理学院，江苏 南京

收稿日期：2024年3月28日；录用日期：2024年4月23日；发布日期：2024年4月30日

摘　要

信息技术发展日新月异，视觉信息的质量广受重视，图像超分辨率技术正因此经过了长久的迭代。但作为一个不适定问题，这项技术仍将是一个长久的难题。随着自注意力机制的出现及引入，传统卷积神经网络方法逐渐在性能上落后。然而，包含自注意力的方法通常计算成本高昂，或是只能为节约计算成本在性能上妥协。因此，本文提出了一种多级轴向加性网络，很好地平衡了性能与成本。具体来说，我们首先设计了一种多级轴向注意力模块，在注意力机制内实现了轴向窗口的模式。然后，我们提出了一种高效的加性注意力，使注意力计算免于矩阵乘法运算。同时，我们还构建了一个轻量级的超分辨率网络MLAAN。最后，我们在五个基准数据集上评估了所提出的MLAAN的效果。在与SOTA方法的对比中，MLAAN在参数量较少的前提下体现了优越的超分辨率性能。

关键词

单图像超分辨率，轻量级网络，多级轴向加性网络(MLAAN)，多级轴向注意力模块(MLAAB)

Lightweight Single Image Super-Resolution with Multi-Level Axial Additive Network

Guanzhe Zou*, Keyan Huang

College of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Mar. 28th, 2024; accepted: Apr. 23rd, 2024; published: Apr. 30th, 2024

Abstract

The importance of visual data has been increasingly emphasized due to the swift advancement of information technology nowadays. As an ill-posed problem, Single Image Super-Resolution continues to present an enduring challenge even after years of progression. Massive self-attention based

*通讯作者。

文章引用: 邹观哲, 黄可言. 基于多级轴向加性网络的轻量级单图超分辨率[J]. 应用数学进展, 2024, 13(4): 1842-1852.
DOI: 10.12677/aam.2024.134173

methods proposed have shown performance exceeding traditional Convolutional Neural Networks based methods. However, methods including self-attention either suffer from large computational cost, or have to compromise on the weakened ability on capturing information thanks to modification on attention. We propose a Multi-Level Axial Additive Network with well-balanced trade-off in this work. Specifically, we first elaborate a Multi-Level Axial Attention Block enabling axial window patterns within attention. Then we present an effective additive attention that eliminates the need for expensive matrix multiplication operations in attention. We also construct a Feature Extraction Module base on shift-convolution to extract local features. We evaluate the efficacy of our proposed MLAAN on five benchmark datasets and show that it significantly enhances the super-resolution performance of the network. Our experimental results demonstrate state-of-the-art performance in lightweight SISR while using a low number of parameters.

Keywords

Single Image Super-Resolution, Lightweight Network, Multi-Level Axial Additive Network (MLAAN), Multi-Level Axial Attention Block (MLAAB)

1. 引言

单图像超分辨(SISR)的目标是从相应的低分辨率(LR)图像中推算出还原的高分辨率(HR)图像。在监控成像、自动驾驶和医疗成像等各种计算机视觉应用中，该技术都发挥着重要作用。作为一个不适定问题，即使经过多年的发展，SISR 仍然是一个持久的挑战。人们提出了许多基于卷积神经网络(CNN)的方法[1] [2]，以直接学习 LR 和 HR 图像对之间的映射。例如，Dong [3]等率先提出了基于 CNN 的初始模型 SRCNN。SRCNN 仅用三个卷积层，就成功超越了传统方法。之后，研究人员开始设计更多具有更复杂架构的网络，基于 CNN 的 SISR 方法取得了长足进步[4]。

尽管这些模型已经取得了显著的成就，但其庞大的计算成本却一直阻碍着它们的传播与实际应用投产。为了扩大 SISR 的应用范围，必须在性能和计算成本的平衡中求进。因此，学界与业界都把目光投向了轻量级的超分辨率方法，这些方法既有不俗的效果，又能将计算开支降到最低。DRRN [5]利用循环网络结构在不增加参数的情况下增加了网络深度。然而，由于牺牲了性能和计算量，这种方法的相对实际效率并未提升。随着人们对 SISR 研究的钻研，高效的网络结构设计分化出了多种路径，包括神经结构搜索(NAS)、多尺度结构和通道分组策略。CARN [6]通过级联局部和全局特征来恢复 HR 图像，速度和精度都很高。IMDN [7]通过引入多重信息蒸馏模块来聚合和提炼特征。RFDN [8]在 IMDN 的基础上进一步改进，加入了特征蒸馏连接(FDC)。BSRN [9]引入了蓝图可分离卷积，以更小的模型实现了更好的重建效果。但是，卷积核通常用于提取局部特征，这意味着对图像中长距离依赖关系彻的彻底忽视。

在这方面，Transformer 是一个出色的替代品。近年来，凭借它强大的全局建模能力，视觉 Transformer (ViT)在许多视觉任务中崭露头角。近期有研究将 Transformer 引入了 SISR 领域。Swin IR [10]利用移位窗口方案对长距离依赖关系进行建模，证明了 ViT 在 SISR 领域同样潜能巨大。ESRT [11]将 CNN 和 Transformer 结合，构建了高效的轻量级模型。ELAN [12]进一步简化了网络，避免了上游视觉任务庞大冗杂的网络。然而，现有的大多数基于 Transformer 的方法都使用密集注意力策略或移动窗口策略。本质上，感受野依

然受到限制，需要堆叠大量模块才可能有效提取全局信息。为此，我们采用了轴向窗口注意力策略。由于相邻区域间的像素元素通常比远距离的像素元素相互作用更强，我们将注意力划分，分配到局部窗口和两个轴上。通过调节这三个部分的权重，实现了对局部与全局注意力的粒度调整。值得注意的是，在我们的架构中，局部窗口、水平轴向和垂直轴向的自注意力是以一种免于额外计算开支的并行模式计算的。

同时，我们还提出了一个用于 SISR 的多级轴向加性网络(MLAAN)，以基于 ViT 的架构组成。我们首先设计了了多级轴向注意力模块(MLAAB)，使提出的轴向窗口模式能够以轻量级的方式收集全局特征。然后，我们提出了一个高效的加性注意力模块(EAA)，在注意力计算中将繁杂的矩阵乘法运算实现取代。我们还构建了一个作用于局部的特征提取模块(FEM)，其中引入了移位卷积层和 GELU 激活函数。总之，本文的主要贡献可以概括为以下三个方面：

(1) 为SISR任务提出了一种结构简洁但功能强大的网络MLAAN。在轻量化的网络中巧妙地利用ViT的全局建模能力，在抑制计算成本的同时显著地提高了性能。并且通过标准数据集上的实验，定性和定量地论述了网络的优越性。

(2) 设计了一种多级轴向注意力模块 MLAAB，在注意力机制内实现了轴向窗口模式，使全局依赖提取的轻量化成为可能。

(3) 设计了一种高效的加性注意力模块 EAA，让注意力中繁杂的矩阵乘法运算得以被取代，进一步降低了计算复杂度。

2. 网络结构设计

以 MLAAB 为骨干单元，我们提出了 MLAAN，网络结构如图 1 所示。该网络有着清晰简洁的结构，先后包括了浅层特征提取模块和深层特征提取模块，两部分的输出之间进行残差连接，合并后的输出再输入到上采样模块进行重建。

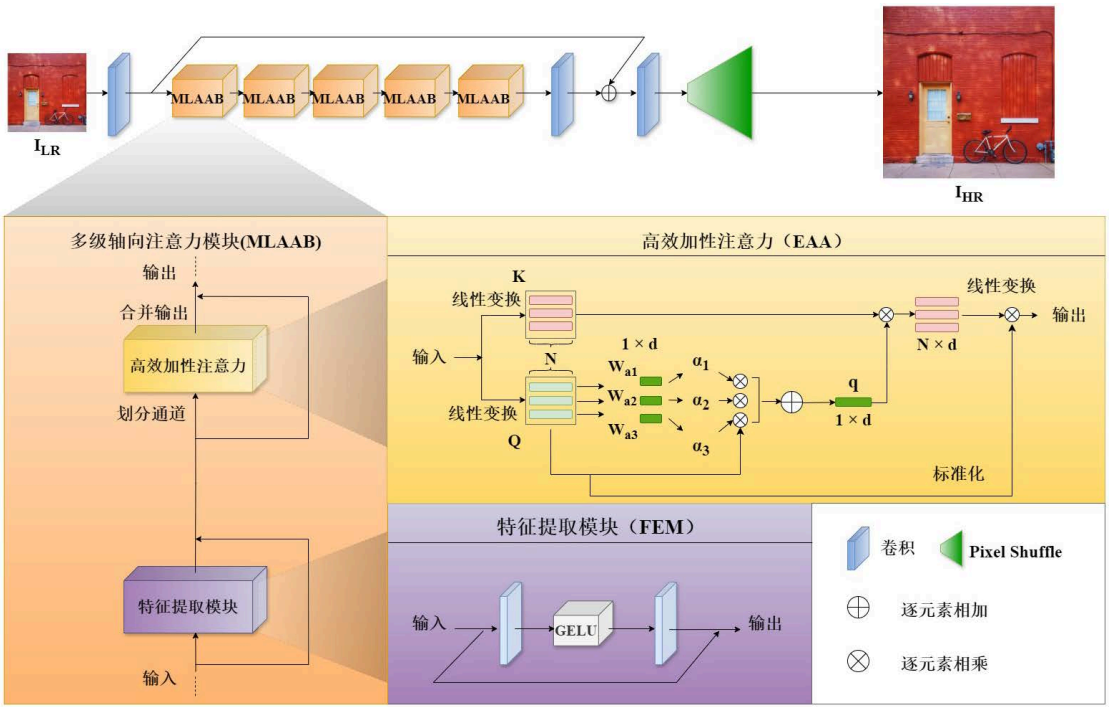

Figure 1. The overall architecture of the proposed MLAAN
图 1. 所提出的 MLAAN 的主体网络架构

2.1. 注意力机制

在 Transformer 中，核心单元是多头自注意力(MHSA)。所谓自注意，是指将输入 $\boldsymbol{X}\in\mathbb{R}^{N\times C}$ 分别线性投影为查询矩阵 $\boldsymbol{Q}\in\mathbb{R}^{N_q\times C}$，键矩阵 $\boldsymbol{K}\in\mathbb{R}^{N_{kv}\times C}$，以及值矩阵 $\boldsymbol{V}\in\mathbb{R}^{N_{kv}\times C}$，注意力函数会将每个 $\boldsymbol{Q}$ 矩阵转换为 $\boldsymbol{V}$ 矩阵的求和权重。该权重是通过 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的归一化点积确定的。上述计算可表示为如下矩阵运算：

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{C}}\right)\boldsymbol{V}.$$

这里引入了可变标量 $\sqrt{C}$ 以避免权重集中和梯度消失，通常根据输入的维度确定。对于视觉 Transformer 来说，$\boldsymbol{X}$ 是一个二维空间特征图 $N=H\times W$，其中 H 和 W 分别是特征图的高度和宽度。多头是指需要沿通道维度将输出划分为 h 段。不同注意力头的投射权重不同。上述计算可表述为：

$$\text{MHSA}\left(\boldsymbol{X}\right)=\text{Concat}\left(\mathbf{head}_0,\mathbf{head}_1,\ldots,\mathbf{head}_h\right)\boldsymbol{W}^{o},$$

$$\mathbf{head}_i=\text{Attention}\left(\boldsymbol{X}\boldsymbol{W}_i^{Q},\boldsymbol{X}\boldsymbol{W}_i^{K},\boldsymbol{X}\boldsymbol{W}_i^{V}\right),$$

其中 $\mathbf{head}_i\in\mathbb{R}^{N\times\frac{C}{h}}$ 是第 i^{th} 注意力头的输出。$\boldsymbol{W}_i^{Q},\boldsymbol{W}_i^{K},\boldsymbol{W}_i^{V}\in\mathbb{R}^{C\times\frac{C}{h}}$ 矩阵用于输入映射。通过对输入进行额外的线性变换得到的权重 $\mathbf{W}^{o}\in\mathbb{R}^{C\times C}$，用于各输出进行合并求和。

MHSA 中有 N 个 $\boldsymbol{Q}$，每个 $\boldsymbol{Q}$ 将处理 N 个键值对，因此复杂度为 $O\left(N^2\right)$。MHSA 的高复杂度给视觉任务的输入分辨率带来了不少限制。

2.2. 多级轴向窗口注意力(MLAW)

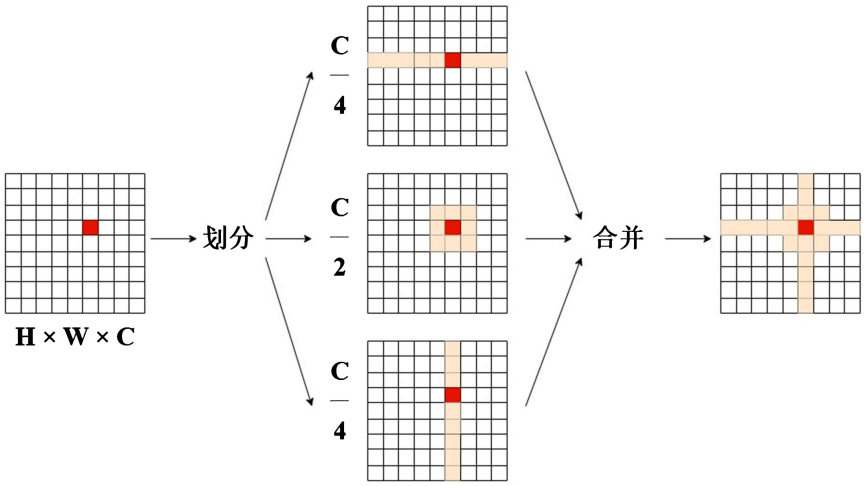

Figure 2. The basic scheme of MLAW. The input scale is set as (9, 9) for better understanding. ***H***, ***W***, ***C*** stand for input height, input width, and input channel respectively.
图 2. MLAW 的流程结构。为了清晰可视化将输入尺寸设为 9 × 9，***H***，***W***，***C*** 分别为输入的高、宽和通道

与 MHSA 中相同，输入特征 $\mathbf{X}\in\mathbb{R}^{(H\times W)\times C}$ 将首先线性投射到 K 个注意力头上，而此后每个注意力头在局部窗口或横轴或纵轴内进行自注意力运算，如图 2。

2.2.1. 轴向窗口注意力

在提出的水平轴向注意力中，$\mathbf{X}$ 被平均分割成不重叠的水平条状窗口 $\left[\mathbf{X}^1,\cdots,\mathbf{X}^H\right]$，每个窗口包含 $1\times W$ 个元素。形式上，假设第 k^{th} 注意力头的 $\mathbf{Q},\mathbf{K},\mathbf{V}$ 的维度都是 d_k，那么 k^{th} 注意力头的水平轴向注意

力输出定义为：

$$\boldsymbol{X}=\left[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^H\right],$$

$$\mathbf{Y}_k^i=\mathrm{MSA}\left(\mathbf{X}^i\mathbf{W}_k^Q,\mathbf{X}^i\mathbf{W}_k^K,\mathbf{X}^i\mathbf{W}_k^V\right),$$

$$\text{H-MSA}_k\left(\mathbf{X}\right)=\left[\mathbf{Y}_k^1,\mathbf{Y}_k^2,\cdots,\mathbf{Y}_k^H\right],$$

其中，$\mathbf{X}^i\in\mathbb{R}^{(1\times W)\times C},i\in\{1,2,\cdots,H\}$，MSA 表示多头注意力。$\mathbf{W}_k^Q\in\mathbb{R}^{C\times d_k},\mathbf{W}_k^K\in\mathbb{R}^{C\times d_k},\mathbf{W}_k^V\in\mathbb{R}^{C\times d_k}$ 分别代表 k^{th} 注意力头的 $\mathbf{Q},\mathbf{K},\mathbf{V}$ 的投影矩阵，$d_k=C/K$。类似可得垂直轴向注意力，其对第 k^{th} 注意力头的输出表示为 $\text{V-MSA}_k\left(\mathbf{X}\right)$。

对于局部窗口注意力，$\mathbf{X}$ 被平均分割成高度和宽度等于 M 的非重叠局部窗口 $\left[\mathbf{X}_m^1,\cdots,\mathbf{X}_m^N\right]$，每个窗口包含 $M\times M$ 元素。类似地，第 k^{th} 注意力头的局部窗口注意力输出定义为：

$$\mathbf{X}_m=\left[\mathbf{X}_m^1,\mathbf{X}_m^2,\cdots,\mathbf{X}_m^N\right],$$

$$\mathbf{Y}_k^i=\mathrm{MSA}\left(\mathbf{X}_m^i\mathbf{W}_k^Q,\mathbf{X}_m^i\mathbf{W}_k^K,\mathbf{X}_m^i\mathbf{W}_k^V\right),$$

$$\text{W-MSA}_k\left(\mathbf{X}\right)=\left[\mathbf{Y}_k^1,\mathbf{Y}_k^2,\cdots,\mathbf{Y}_k^N\right],$$

其中 $N=\left(H\times W\right)/\left(M\times M\right)$。

2.2.2. 粒度差异的构建

我们将 K 个注意力头分为三部分，给两个轴向窗口各分配 $K/4$ 个注意力头，给局部窗口分配 $K/2$ 个注意力头。通过这种刻画粒度的方式，我们为局部和稀疏全局特征设置了权重。第一组注意力头执行水平轴向注意力，第二组注意力头执行垂直轴向注意力，第三组注意力头执行局部窗口注意力。并行计算后，输出将被重新合并：

$$\text{head}_k=\begin{cases}\text{H-MSA}_k\left(X\right), & k=1,\cdots,K/4,\\ \text{V-MSA}_k\left(X\right), & k=K/4+1,\cdots,K/2,\\ \text{W-MSA}_k\left(X\right), & k=K/2+1,\cdots,K,\end{cases}$$

$$\mathrm{MLAW}\left(\mathbf{X}\right)=\mathrm{Concat}\left(\text{head}_1,\cdots,\text{head}_K\right)\mathbf{W}^O,$$

其中，$\mathbf{W}^O\in\mathbb{R}^{C\times C}$ 为投影矩阵，用于融合各注意力头的输出。与分别逐步实现轴向和窗口注意力相比，这种并行机制的计算复杂度更低。并且，可以通过改变各组注意力头的数量来实现不同的粒度。

2.3. 高效的加性注意力

此前，加性注意力机制在 NLP 中已被应用，通过元素乘法取代了点积运算，利用成对标记间的交互来获得全局信息。它与惯例方法相同，仍用 $\mathbf{Q},\mathbf{K},\mathbf{V}$ 的交互作用来编码输入序列上下文信息的相关性分数。在此之上，本文提出 EAA 只需通过加入线性投影层以聚焦于 $\mathbf{Q}$ 、$\mathbf{K}$ 间有效交互，就足以学习表征之间的关系(见图 1)。进一步简化了运算，在提高推理速度的条件下仍能稳健地提取特征。具体来说，输入 $\mathbf{X}$ 通过两个映射矩阵 $\mathbf{W}^Q,\mathbf{W}^K$ 转换成 $\mathbf{Q}$ 和 $\mathbf{K}$，其中 $\mathbf{Q},\mathbf{K}\in\mathbb{R}^{n\times d},\mathbf{W}^Q,\mathbf{W}^K\in\mathbb{R}^{d\times d}$，$n$ 是输入长度，d 是输入向量的维数。接下来，$\mathbf{Q}$ 矩阵可学习参数向量 $\mathbf{w}_a\in\mathbb{R}^d$ 相乘，学习 $\mathbf{Q}$ 的注意力权重，然后通过 Softmax 运算产生全局注意力查询向量 $\alpha\in\mathbb{R}^n$ 如下：

$$\alpha=\frac{\exp\left(\mathbf{Q}\cdot\mathbf{w}_a/\sqrt{d}\right)}{\sum_{j=1}^{n}\exp\left(\mathbf{Q}\cdot\mathbf{w}_a/\sqrt{d}\right)}.$$

然后，根据学习到的注意力权重对 **Q** 矩阵进行池化，最终得到一个单一的全局查询向量 $\mathbf{q}\in\mathbb{R}^d$，如下所示：

$$\mathbf{q}=\sum_{i=1}^{n}\alpha_i*\mathbf{Q_i}.$$

接下来，通过元素乘积对全局查询向量 **q** 和键矩阵 $\mathbf{K}\in\mathbb{R}^{n\times d}$ 的交互进行编码，从而形成全局上下文($\mathbb{R}^{n\times d}$)。该矩阵与 MHSA 中的注意力矩阵相似，能捕捉全局的元素信息，在学习输入间的相关性上足够敏锐。不过，与 MHSA 相比，它的计算成本相对较低，复杂度与输入长度呈线性关系。受 Transformer 架构的启发，我们在 **Q** 、**K** 交互中加入线性变换，从而学习输入的隐藏层表征。高效加性注意力的输出 $\hat{\mathbf{X}}$ 可以描述为：

$$\hat{\mathbf{X}}=\hat{\mathbf{Q}}+\mathbf{T}(\mathbf{K}*\mathbf{q}),$$

其中，$\hat{\mathbf{Q}}$ 表示归一化查询矩阵，**T** 表示线性变换。

2.4. 复杂度分析

对于 MLAW，输入特征的大小为 $H\times W\times C$，窗口大小为 (M,M)，标准的全局自注意力的计算复杂度为：

$$\Omega(\text{Global})=4HWC^2+2(HW)^2C,$$

而并行计算的 MLAW 的计算复杂度为：

$$\Omega(\text{MLAW})=4HWC^2+HWC*\left(\frac{1}{2}H+\frac{1}{2}W+M^2\right),$$

与全局计算相比，这显然可以减轻计算和内存负担，因为 $2HW>>\left(\frac{1}{2}H+\frac{1}{2}W+M^2\right)$ 总是成立的。

至于 EAA，其中学习 **Q** 、**K** 矩阵的映射与交互的参数部分的计算复杂度为 $O(n\cdot d)$，而此后的逐元素乘积复杂度也同为 $O(n\cdot d)$。与传统自注意力的 $O(n^2\cdot d)$ 复杂度相比，EAA 在节约计算上显出优越性，且在图像的大输入尺寸的背景下更为显著。

2.5. 多级轴向注意力模块

在上述内容的基础上，构建了 FEM 作为 MLAAB 的局部特征提取基础单元，我们在 FEM 中使用了两个移位卷积，并采用了 GELU 激活函数。这个基本卷积单元的结构如图 1 所示。至此，MLAAB 可定义为：

$$\widehat{\mathbf{X}^l}=\text{FEM}\left(\text{LN}\left(\mathbf{X}^i\right)\right)+\mathbf{X}^i,\mathbf{X}^l=\text{MLAW}\left(\text{LN}\left(\widehat{\mathbf{X}^l}\right)\right)+\widehat{\mathbf{X}^l},$$

其中 $\widehat{\mathbf{X}^l},\mathbf{X}^l$ 分别表示第 l 个 MLAAB 模块中的 FEM 模块和 MLAW 模块的输出特征。

2.6. 损失函数

在训练阶段，给定一个 LR-HR 训练集 $\left\{I_i^{LR},I_i^{HR}\right\}_{i=1}^{N}$，MLAAN 的损失函数可以表示为：

$$Loss(\theta)=\arg\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}F_{MLAAN}\left(I_{LR}^i\right)-I_{HR1}^i,$$

其中，$F_{MLAAN}(\cdot)$ 表示我们提出的 MLAAN，θ 表示 MLAAN 的参数集，N 表示训练集中 LR-HR 图像对

的数量。

3. 实验结果与讨论

3.1. 数据集和评估标准

在训练阶段，我们采用了 DF2K 数据集，包含 DIV2K 和 Flickr2K 数据集，共 3450 张图像。至于模型的评估，我们使用 5 个 SISR 标准数据集进行，分别为 Set5、Set14、BSD100、Urban100 和 Manga109。为了呈现明晰的量化结果，我们采用了峰值信噪比(PSNR)和结构相似度(SSIM)作为数值指标。具体来说，SR 图像的 PSNR 和 SSIM 是在 YCbCr 色彩空间的亮度 Y 通道上计算的。

3.2. 训练细节

为了得到符合 SISR 条件的训练数据，我们首先应用双三次插值法进行 HR 图像下采样。每个训练批次由 16 个随机抽取的 48×48 像素块组成，每个输入随机进行旋转或水平翻转的数据增强。我们通过 Adam 优化器对模型进行了 1000 次训练，动量参数为 0.9，损失函数为 $L1$。初始学习率为 2×10^{-4}，每经 200 个 epoch 减半。该网络的训练使用了英伟达 RTX3080Ti GPU 和 PyTorch 框架。

3.3. 实验结果对比

本节，我们通过与最先进的轻量级 SR 模型的对比，展示了所们提出的模型的有效性。首先，我们直接呈现了量化结果(PSNR 和 SSIM)和计算成本，这是轻量级网络通常的关注重点。其次，我们从 SR 任务的实际目的出发，展示了视觉效果并进行了定性评估。

3.3.1. 定量比较

在表 1 和表 2 中，我们将我们的网络与其他先进的 SISR 模型进行了不同缩放尺度上的比较，其中包括 VDSR、EDSR、SRMDNF、CARN、IMDN、ESRT、SwinIR-light。最好的结果均用粗体标出。可见，我们的 MLAAN 在各标准测试集上取得了可观的结果，在 PSNR 和 SSIM 上可与这些最先进的模型相媲美。此外，值得注意的是，我们提出的方法在实验的各尺度上基本优于众多类似的基于或融合了 ViT 的模型：ESRT 和 SwinIR-light。这主要得益于我们的模型在整合局部和长距离依赖方面取得了很好的平衡。因此，重建的图像包含丰富的结构细节，局部细节连贯，看起来比其他模型的图像更自然。

Table 1. Average PSNR/SSIM comparison with SISR models on **×3** scale
表 1. 与其他 SISR 方法在**×3** 缩放尺度上的 PSNR/SSIM 均值比较

方法	参数量	Set5	Set14	BSD100
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
VDSR	666K	33.66/0.9213	29.77/0.8314	28.82/0.7976
EDSR	1,555K	34.37/0.9270	30.28/0.8417	29.09/0.8052
SRMDNF	1,528K	34.12/0.9254	30.04/0.8382	28.97/0.8025
CARN	1,592K	34.29/0.9255	30.29/0.8407	29.06/0.8034
IMDN	703K	34.36/0.9270	30.32/0.8417	29.09/0.8046
ESRT	770K	34.42/0.9268	30.43/0.8433	29.15/0.8063
SwinIR-light	886K	34.62/**0.9289**	30.54/0.8463	29.20/0.8082
MLAAN (Ours)	706K	**34.64**/0.9286	**30.58**/**0.8465**	**29.22**/**0.8086**

Table 2. Average PSNR/SSIM comparison with SISR models on **×4** scale
表 2. 与其他 SISR 方法在**×4** 缩放尺度上的 PSNR/SSIM 均值比较

方法	参数量	BSD100	Urban100	Manga109
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
VDSR	666 K	27.29/0.7251	25.18/0.7524	28.83/0.8870
EDSR	1,518 K	27.57/0.7357	26.04/0.7849	30.35/0.9067
SRMDNF	1,552 K	27.49/0.7337	25.68/0.7731	30.09/0.9024
CARN	1,592 K	27.58/0.7349	26.07/0.7837	30.47/0.9084
IMDN	715 K	27.56/0.7353	26.04/0.7838	30.45/0.9075
ESRT	751 K	27.69/0.7379	26.39/0.7962	30.75/0.9100
SwinIR-light	897 K	27.69/0.7406	26.47/0.7980	**30.92**/0.9151
MLAAN (Ours)	714 K	**27.73**/**0.7411**	**26.55**/**0.8003**	**30.92**/**0.9155**

Table 3. Overall capacity comparison with lightweight SISR models
表 3. 与其他轻量级 SISR 方法的综合性能比较

方法	网络类型	参数量	FLOPs (x4)	PSNR
VDSR	CNN	0.67 M	612.6 G	31.35
LapSRN	CNN	**0.25 M**	149.4 G	31.54
IMDN	CNN	0.70 M	**40.9 G**	32.21
ESRT	hybrid	0.68 M	67.7 G	32.19
SwinIR	Transformer	0.90 M	218.8 G	32.44
MLAAN (Ours)	Transformer	0.71 M	107.3 G	**32.48**

我们还进一步探究了所提出模型的计算成本，并与其他方法进行了比较。出于提高性能，并尽可能削减参数量、复杂度和推理速度等项的考量，我们的 MLAAN 实现了适当的平衡，如表 3 所示。从表中可以看出，首先 MLAAN 在性能上显著超出上述方法。同时作为一种基于 ViT 的模型，MLAAN 在参数量方面与多数 CNN 和混合模型相当，性能则是有着巨大提升。每秒浮点运算次数(FLOPs)上，明显优于传统 CNN 方法，和较为新进的 CNN 和混合网络相比作为 ViT 模型仍有优化空间。最重要的是，MLAAN 在参数量、FLOPs 和性能方面全面优于同样基于 ViT 的 SwinIR。综合而言，MLAAN 比 SOTA 方法实现了更好的性能_开支平衡。

3.3.2. 定性比较

此外，我们还在图 3 中提供了我们的 MLAAN 与其他 SISR 方法的直观视觉对比。就图 3 中的上图而言，多数比较方法重建的 SR 图像都含有严重的伪影，图像上的线条模糊不清。相比之下，MLAAN 重建的 SR 图像更加贴近原始图像，线条和色块更加清晰分明。不仅如此，其他方法存在过度平滑的问题，丢失了许多图像中的高频细节。基于混合的方法和其他基于 ViT 的方法可以缓解但不能完全克服这些现象，而 MLAAN 呈现出了最好的效果。在此之上 MLAAN 如实地重建了图像的结构，其他效果较清晰的图像则都在图像的右部出现了失真。对于下图来说，与上图的情况一样，其他方法要么存在过度平滑的问题，要么无法稳健地整合全局信息，导致视觉上的失调，图像出现了不同程度的扭曲。然而，我们的 MLAAN 一方面克服了过度平滑的问题，呈现出相对更清晰的边缘，另一方面还能重建出符合原始图像结构的图像。很明显，我们的 MLAAN 可以重建具有更精确纹理细节和边缘的高质量图像。这进一步证明了所提出的 MLAAN 的有效性。

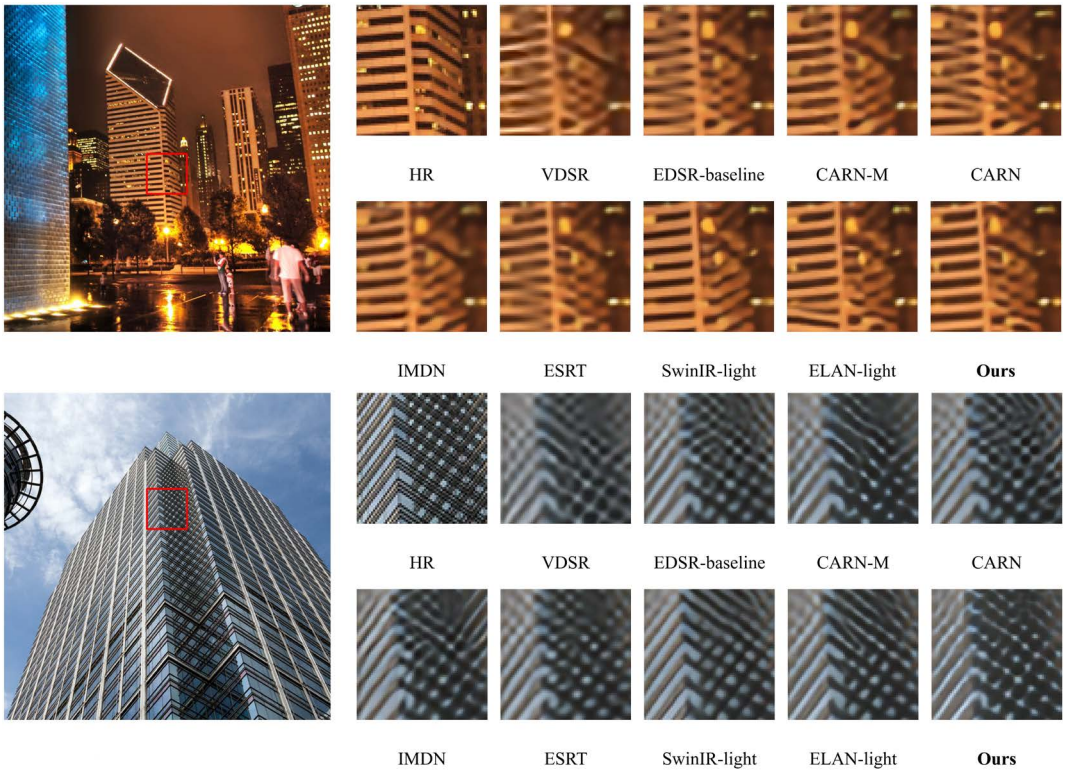

Figure 3. The Qualitative comparison of **×4** image SR on the Urban100 dataset
图 3. 在 Urban100 数据集**×4** 缩放尺度上的定性比较

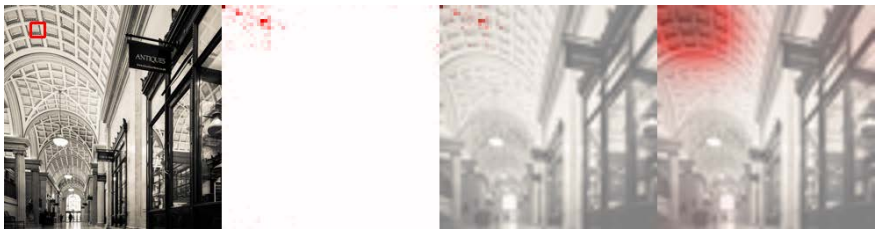

Figure 4. An example of LAM results of MLAAN
图 4. MLAAN 的 LAM 结果样本

不仅如此，我们还在图 4 中展示了 MLAAN 的局部归因图(LAM)结果。图中红色高亮部分代表了最左侧图像中，红框部分输入重建所建立的依赖范围。可见，网络对全局依赖建模的范围相当广泛。由此，LAM 进一步证明了 MLAAN 能够在大范围内聚集像素信息，从而重构图像细节。

3.4. 消融实验

在本节中，我们进行了一系列研究来探究各模块的作用，以进一步展示模型的有效性。

3.4.1. 表征之间的相似性

首先，我们引入了中心核对齐(CKA)，以直观展示 MLAW 学习到的表征模式。具体来说，CKA 结果显示了神经元之间的相似性。CKA 得分越高，表示相似神经元越多，神经元间传递的信息就越少。如图 5 所示，MLAW 的轴向窗口内部神经元相似性明显更低，这表明我们的 MLAW 的模式能有效提取这些范围内的信息。

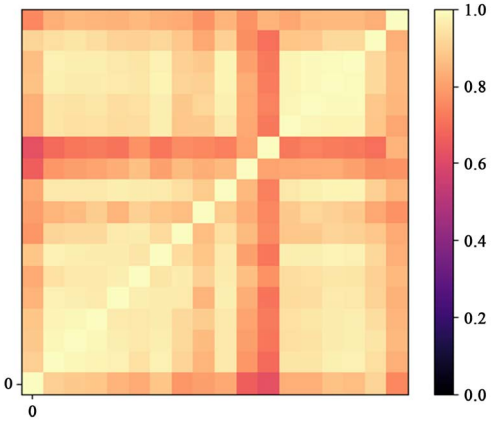

Figure 5. An example of LAM results of MLAAN
图 5. MLAAN 的 LAM 结果样本

3.4.2. MLAW 和 EAA 的有效性

为了验证所提出的 MLAW 与 EAA 的有效性，我们在实验中进行了两模块的消融对比，如表 4。从表中我们可以清楚地看到，MLAW 的引入使重建质量显著提升，参数的膨胀也远不及上文中对比的基于 ViT 的网络严重。这表明，MLAW 可以提高模型的学习能力，从而在有效抑制计算成本升高的情况下提高模型性能。而 EAA 的引入，相比传统自注意力，显然使参数量和推理延时下降了，性能上则在 Set5 的 SSIM 上有微小的下降，但在 PSNR 上略有提高。该结果完美符合了轻量级网络的要求，在把握性能和开支平衡的基础上向前跃进。

Table 4. Evaluate the effectiveness of MLAW and EAA
表 4. MLAW 和 EAA 的有效性验证

缩放比	MLAW	EAA	参数量	延时	Set5 PSNR/SSIM
×4		√	518K	584.37 ms	32.35/0.8962
	√		703K	849.52 ms	32.47/**0.8982**
	√	√	714K	639.16 ms	**32.48**/0.8980

3.4.3. MLAAB 数量的影响

为了整体调整模型大小与性能间的平衡，我们考察了 MLAAB 数量对模型的影响，在五个数据集上比较了了 PSNR 和 SSIM。从表 5 中可见，当 MLAAB 数量为 12 时，模型拟合程度明显不足；而与 18 的数量相比，模型在 24 个 MLAAN 块下依然有显著的性能提升，而参数量恰好符合轻量级的条件。当数量进一步增加时，模型性能的提高成本将难以承受。因此，我们将模型中的 MLAAB 数量设定为 24。

Table 5. Evaluate the model capacity under different numbers of MLAAB
表 5. MLAAB 数量的影响评估

模块数	12	18	24
参数量	601 K	653 K	714 K
Set5	32.06/0.8964	32.39/0.8977	**32.48/0.8980**
Set14	28.68/0.7837	28.80/0.7864	**28.83/0.7870**
B100	27.63/0.7384	27.71/0.7408	**27.73/0.7411**
Urban100	26.29/0.7930	26.48/0.7986	**26.55/0.8003**
Manga109	30.69/0.9109	30.86/0.9142	**30.92/0.9155**

4. 结论

近年，图像超分辨率技术经过了快速的迭代。但作为一个不适定问题，这项技术仍将是一个长久的难题。随着自注意力机制的出现及引入，关键点逐渐落在了性能与计算成本的平衡上。本文中，我们提出了一种轻量级的多级轴向加性网络(MLAAN)。具体来说，我们首先设计了多级轴向注意力模块(MLAAB)，在注意力机制内实现了轴向窗口的模式以整合全局特征。然后，我们提出了一种高效的加性注意力(EAA)，注意力计算免于繁杂的矩阵乘法运算。同时，我们还构建了一个轻量级的超分辨率网络MLAAN。最后，我们在五个基准数据集上评估了所提出的 MLAAN 的效果。在与 SOTA 方法的对比中，MLAAN 在参数量较少的前提下体现了优越的超分辨率性能。

参考文献

[1] 周登文, 李文斌, 李金新, 黄志勇. 一种轻量级的多尺度通道注意图像超分辨率重建网络[J]. 电子学报, 2022(10): 2336-2346.

[2] 江明. 基于深度学习的图像超分辨率轻量级算法研究[D]: [硕士学位论文]. 赣州: 江西理工大学, 2022. https://link.cnki.net/doi/10.27176/d.cnki.gnfyc.2022.000740

[3] Dong, C., Loy, C.C., He, K. and Tang, X. (2015) Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[4] Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y. (2018) Residual Dense Network for Image Super-Resolution. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2472-2481. https://doi.org/10.1109/CVPR.2018.00262

[5] Tai, Y., Yang, J. and Liu, X. (2017) Image Super-Resolution via Deep Recursive Residual Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, 21-26 July 2017, 3147-3155. https://doi.org/10.1109/CVPR.2017.298

[6] Ahn, N., Kang, B. and Sohn, K.A. (2018) Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. *Computer Vision - ECCV* 2018, Munich, 256-272. https://doi.org/10.1007/978-3-030-01249-6_16

[7] Hui, Z., Gao, X., Yang, Y. and Wang, X. (2019) Lightweight Image Super-Resolution with Information Multi-Distillation Network. *Proceedings of the* 27*th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 2024-2032. https://doi.org/10.1145/3343031.3351084

[8] Liu, J., Tang, J. and Wu, G. (2020) Residual Feature Distillation Network for Lightweight Image Super-Resolution. *Computer Vision-ECCV* 2020 *Workshops*, Glasgow, UK, 23-28 August 2020, 41-55. https://doi.org/10.1007/978-3-030-67070-2_2

[9] Li, Z., Liu, Y., Chen, X., Cai, H., Gu, J., Qiao, Y. and Dong, C. (2022) Blueprint Separable Residual Network for Efficient Image Super-Resolution. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, LA, 19-20 June 2022, 833-843. https://doi.org/10.1109/CVPRW56347.2022.00099

[10] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L. and Timofte, R. (2021) Swinir: Image Restoration Using Swin Transformer. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, BC, 11-17 October 2021, 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[11] Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L. and Zeng, T. (2022) Transformer for Single Image Super-Resolution. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, LA, 19-20 June 2022, 457-466. https://doi.org/10.1109/CVPRW56347.2022.00061

[12] Zhang, X., Zeng, H., Guo, S. and Zhang, L. (2022) Efficient Long-Range Attention Network for Image Super-Resolution. *Computer Vision - ECCV* 2022, Tel Aviv, 649-667. https://doi.org/10.1007/978-3-031-19790-1_39