

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2024)04-0875-15

论文引用格式: Bi X P, Chen S and Zhang L F. 2024. Blueprint separable convolution Transformer network for lightweight image super-resolution. Journal of Image and Graphics, 29(04):0875-0889(毕修平, 陈实, 张乐飞. 2024. 轻量级图像超分辨率的蓝图可分离卷积Transformer网络. 中国图象图形学报, 29(04):0875-0889)[DOI:10.11834/jig.230225]

轻量级图像超分辨率的蓝图可分离卷积Transformer网络

毕修平¹, 陈实¹, 张乐飞^{1,2*}

1. 武汉大学计算机学院国家多媒体软件工程技术研究中心, 武汉 430072; 2. 湖北珞珈实验室, 武汉 430079

摘要: 目的 图像超分辨率重建的目的是将低分辨率图像复原出具有更丰富细节信息的高分辨率图像。近年来, 基于Transformer的神经网络在图像超分辨率重建领域取得了令人瞩目的性能, 然而, 这些网络往往参数量巨大、计算成本较高。针对该问题, 设计了一种轻量级图像超分辨率重建网络。**方法** 提出了一种轻量级图像超分辨率的蓝图可分离卷积Transformer网络(blueprint separable convolution Transformer network, BSTN)。基于蓝图可分离卷积(blueprint separable convolution, BSConv)设计了蓝图前馈神经网络和蓝图多头自注意力模块。然后设计了移动通道注意力模块(shift channel attention block, SCAB)对通道重点信息进行加强, 包括移动卷积、对比度感知通道注意力和蓝图前馈神经网络。最后设计了蓝图多头自注意力模块(blueprint multi-head self-attention block, BMSAB), 通过蓝图多头自注意力与蓝图前馈神经网络以较低的计算量实现了自注意力过程。**结果** 本文方法在4个数据集上与10种先进的轻量级超分辨率方法进行比较。客观上, 本文方法在不同数据集上取得了不同程度的领先, 并且参数量和浮点运算量都处于较低水平。当放大倍数分别为2、3和4时, 在Set5数据集上相比SOTA(state-of-the-art)方法, 峰值信噪比(peak signal to noise ratio, PSNR)分别提升了0.11 dB、0.16 dB和0.17 dB。主观上, 本文方法重建图像清晰, 模糊区域小, 具有丰富的细节。**结论** 本文所提出的蓝图可分离卷积Transformer网络BSTN以较少的参数量和浮点运算量达到了先进水平, 能获得高质量的超分辨率重建结果。

关键词: 图像超分辨率; 轻量级模型; Transformer; 深度学习; 注意力机制

Blueprint separable convolution Transformer network for lightweight image super-resolution

Bi Xiuping¹, Chen Shi¹, Zhang Lefei^{1,2*}

1. National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China; 2. Hubei LuoJia Laboratory, Wuhan 430079, China

Abstract: **Objective** Image super-resolution aims to enhance the resolution and quality of low-resolution images, making them more visually appealing and suitable for human or machine recognition. By utilizing a series of degraded low-resolution images with coarse details, the objective is to reconstruct high-resolution images with finer details. The applications of super-resolution algorithms are vast and encompass areas, such as object detection, medical pathological analysis, remote sensing satellite images, and security monitoring. The promising prospects of these applications have led to an

收稿日期: 2023-04-28; 修回日期: 2023-07-03; 预印本日期: 2023-07-10

* 通信作者: 张乐飞 zhanglefei@whu.edu.cn

基金项目: 国家自然科学基金项目(62122060); 湖北珞珈实验室专项基金项目(220100014)

Supported by: National Natural Science Foundation of China (62122060); Special Fund of Hubei LuoJia Laboratory (220100014)

increased recognition of the importance of image super-resolution algorithms among researchers. With the advancement of deep learning in computer vision, this method has been successfully applied to image super-resolution, leading to significant achievements. However, the substantial number of parameters and the computational requirements of super-resolution models result in slow running speeds, limiting their practicality in real-world development and generation, particularly in mobile and edge devices. To address this issue, several lightweight super-resolution models have been proposed. Among these models, the Transformer-based approach stands out because it provides rich detail information in reconstructed images. However, this type of model still suffers from computational redundancy and large model size. To overcome these challenges, this study presents a novel lightweight super-resolution network based on the Transformer architecture.

Method A blueprint separable convolution Transformer network (BSTN) is proposed for lightweight image super-resolution. BSTN is divided into three parts: shallow feature extraction, deep feature extraction, and image reconstruction. In the shallow feature extraction stage, a 3×3 standard convolution operation is employed to extract low-level features from the input image. This initial feature extraction step helps capture basic image information, which is directly transmitted to the tail of the network to provide residual information via the long skip connection. The deep feature extraction component is composed of four successive residual attention Transformer groups (RATGs). The key elements within this stage are the shift channel attention module (SCAB) and the blueprint multi-head self-attention block (BMSAB). SCAB and BMSAB are combined to form the hybrid attention Transformer module (HATB). Two HATBs are connected together with a residual connection, and a standard convolution operation is applied to follow the two HATBs to construct the RATG. The blueprint feed-forward neural network is first designed for effectively suppressing low-information features and retaining only relevant and useful information. Then, the blueprint feed-forward neural network is introduced into the two aforementioned attention modules to efficiently extract the significant deep features for super-resolution. SCAB consists of three major components: shift convolution, contrast-aware channel attention, and blueprint feed-forward neural networks. Shift convolution reduces the number of network parameters and performs spatial information aggregation, enabling effective information fusion across different regions of the image. The contrast-aware channel attention mechanism focuses on important channel information, enhancing the representation of crucial features. BMSAB consists of a blueprint multi-head self-attention and a blueprint feed-forward neural network. This module allows for the extraction of self-attention with reduced computational complexity while suppressing low-information features through the blueprint feed-forward neural network. Finally, the shallow features extracted in the earlier stage and the deep features obtained from the RATGs are added together. The combined features are then processed using pixel shuffle, a technique that rearranges features to increase their spatial resolution. This final step generates the reconstructed high-resolution image with improved quality and detail. By utilizing the designed architecture and specific components, the proposed lightweight super-resolution network achieves effective feature extraction, self-attention calculation, and image reconstruction, addressing the challenges of parameter redundancy and large model size commonly encountered in Transformer-based super-resolution models. Our method is implemented using PyTorch on NVIDIA RTX 3090 GPU. The training datasets used in this study are DIV2K and Flickr2K, which consist of 800 and 1 000 images, respectively. Batch size is set to 32, and the patch size of the training data is set to 48×48 pixels. The initial learning rate is set to 5×10^{-4} and updated with an Adam optimizer by using a cosine descent strategy, while the total iteration is 10^6 .

Result The proposed method is compared with 11 state-of-the-art approaches on 4 datasets. In accordance with the quantitative results, the proposed method has achieved varying degrees of improvement in different magnifications and datasets, while parameter size and floating-point operations are at low levels. When the magnification factor is 2, the peak signal to noise ratio (PSNR) of this model is ranked first place on Set5, Set14, BSD100, and Urban100. It performs well on Set5 and Set14, surpassing the second-best model by 0.11 dB and 0.08 dB, respectively. When the magnification factor is 3, the PSNR also ranks first place, surpassing Set5 and Urban100 by 0.16 dB and 0.06 dB, respectively. When magnification is 4, it still ranks first place and outperforms the second-place models by 0.17, 0.05, and 0.04 dB on Set5, BSD100, and Urban100, respectively. In accordance with the qualitative results, the reconstructed image of the proposed method is clear, the blurred area is small, and details are rich.

Conclusion A large number of comparative experiments and ablation studies demonstrate that the proposed EBST not only achieves state-of-the-art super-resolution results with excellent quantitative and visual performance, but it also has fewer parameters and floating-

point operations. In particular, the proposed blueprint separable multi-head self-attention can effectively perform self-attention in Transformer blocks through a concise structure. The proposed blueprint feed-forward neural network can focus on helpful information and filter out useless information for super-resolution, resulting in high efficiency and low cost. It can be seamlessly integrated into other modules. Although our method performs well, its advantages in terms of lightweight models are in evident and should be further enhanced.

Key words: image super-resolution; lightweight model; Transformer; deep learning; attention mechanism

0 引言

单图像超分辨率(single image super-resolution, SISR)旨在从一幅细节粗糙的低分辨率(low-resolution, LR)图像中恢复出细节精致、美观的高分辨率(high-resolution, HR)图像,是计算机低级视觉和图像处理中非常重要的任务(Wang 等,2021)。其核心要求在于:在充分保留低分辨率图像像素信息的基准下,提高低分辨率图像的像素密度,使其在人眼视觉或后续其他高级计算机视觉任务中有更好的效果。因此作为其他计算机视觉任务的基础,图像超分辨率重建广泛地应用在目标检测(Girshick 等,2016)、红外图像(邱德粉 等,2023)、医学图像(高媛 等,2018)、移动设备以及边缘设备的目标识别(王素玉和沈兰荪,2007)等领域(Wang 等,2021)。图像超分辨率重建方法丰富,但总体来说可以分为两大类:早期的传统方法和基于深度学习的方法。过去几十年里,大量传统方法相继提出以解决图像超分辨率重建问题,主要分为:基于插值的方法、基于重构的方法和基于浅层学习的方法。Jian 和 Lam (2015)提出了一种基于稀疏表示的人脸重建方法,可以保留人脸图像中的高频细节,增强了人脸重建的视觉效果。但是相比于基于深度学习的方法,传统方法重建的图像质量差,难以满足现在的生产生活需求。目前,基于深度学习的图像超分辨率重建主要通过搭建深度学习网络,通过训练网络让其学习低分辨率图像和高分辨率图像之间的映射关系,然后通过该关系将低分辨率图像重建为高分辨率图像。现实生活中低分辨率和高分辨率图像都是常见且广泛的,但是成对出现的低分辨率图像和高分辨率图像却难以获得。于是研究人员先采集高分辨率图像作为真实数据(ground truth),然后对该图像进行双三次下采样,将其退化为低分辨率图像,由此组成网络的训练对。Transformer 网络由 Google 团队

(Vaswani 等,2017)提出,目的是提高机器翻译的效率和质量,其优秀能力让研究人员开始思考如何将其引入图像超分辨率领域。Yang 等人(2020)提出了一种新颖的纹理 Transformer 超分辨率网络 TTSR(texture Transformer network for image super-resolution),它可以利用参考图像中的高质量纹理来增强低分辨率图像的细节。其核心思想是将低分辨率图像和参考图像分别表示为 Transformer 中的查询和关键字,然后通过注意力机制来计算两者之间的相似度,并从参考图像中提取最相关的纹理信息。但是 TTSR 依赖于参考图像的质量和相似度,如果参考图像与低分辨率图像之间没有足够的纹理匹配,那么 TTSR 可能无法有效地提取和传输纹理信息。此外 TTSR 的计算复杂度较高,因为它需要在多个尺度上进行注意力计算,并且使用了多个 Transformer 模块导致模型体积过大。为了降低 Transformer 注意力的计算消耗,提出了 Swin Transformer(Liu 等,2021)以解决该问题。这是一种新型 Transformer 架构,采用移位窗口计算自注意力。之后,Liang 等人(2021)首次将 Swin Transformer 引入图像复原领域,提出了 SwinIR 网络并在图像超分辨率重建上取得了巨大成功。但是这些 Transformer 超分辨率模型的参数和计算量大,不适合于边缘设备和移动设备上。尽管提出了许多基于 Transformer 的轻量化网络来解决该问题,如 SCET(self-calibrated efficient Transformer)(Zou 等,2022)采用 self-calibrated attention 的新型自注意力减少计算量;ESRT(efficient super-resolution Transformer)(Lu 等,2022)将 Q (query)、 K (key)、 V (value)划分为小组分别计算注意力然后再拼接以减少计算量;LBNet(lightweight bimodal network)(Gao 等,2022)使用通道减少策略来减少模型的参数,并引入了通道注意和空间注意机制来重新权衡从不同分支提取的特征信息;NGSwin(n -gram swin Transformer)(Choi 等,2023)引入 N -Gram 上下文,在不牺牲性能的情况下减少计算量。但这些

方法都使用普通卷积,该卷积计算量大,同时需要较大的参数量,尤其当通道数量增多时,计算量和参数量也急剧膨胀。而且,基于Transformer的模型大多使用多层感知机(multilayer perceptron, MLP)作为前馈神经网络,参数量较大,同时随着多头自注意力头部数量的增加导致计算量,参数量过大。

针对该问题,本文提出了一种轻量级图像超分辨率重建的蓝图可分离卷积Transformer网络,网络结构分为浅层特征提取部分、深层特征提取部分以及上采样重建3部分,该网络使用移动卷积在进行特征提取时降低所需参数量,同时还能实现空间信息的聚合,并使用对比度感知通道注意力加强对重要特征信息的提取。本文还基于蓝图可分离卷积设计了一种蓝图前馈网络,蓝图可分离卷积(Haase和Amthor, 2020)先进行逐点卷积再进行逐通道卷积,可以有效地减少参数量和计算量,同时保持良好的性能。蓝图前馈网络可以抑制信息较少的特征,只允许有用的信息通过该模块。Transformer中所用到的多头自注意力机制需要巨大的参数量,导致大规模的计算开销。因此本文提出了一种微型的多头蓝图自注意力,它可以设置头部数量,同时对其进行限制,防止计算开销过大,使用蓝图可分离卷积进行 Q 、 K 、 V 提取,降低了所需参数量,该模块能够聚合局部和非局部像素交互,高效地关注图像的重点部分。实验结果表明,该模型以较少的参数量和浮点运算量达到了优秀的超分辨率图像重建结果。

本文主要工作和贡献如下:1)提出了一种轻量级图像超分辨率的蓝图可分离卷积Transformer网络,该网络参数量小,浮点运算量小,并实现了良好的超分辨率重建性能。在基准数据集上的大量实验表明,该方法以较少的参数实现了先进的超分辨率重建结果。2)设计了一种高效的蓝图前馈神经网络来关注有利于超分辨率重建的特征信息,相比于传统的前馈网络MLP,蓝图前馈网络占用计算资源少,增强网络特征提取的能力更强,可以简单地应用于各个网络模块。3)设计了一种高效的多头蓝图自注意力来进行局部和非局部的信息交互,相比于传统的自注意力机制,参数量更少,能有效地减少计算开销。

1 相关工作

1.1 单图像超分辨率

自2014年以来,学者提出了许多SISR方法来表示LR图像和HR图像之间的映射。Dong等人(2014)提出了第1个基于深度学习的超分辨率方法SRCNN(super-resolution convolutional neural network),吸引了更多研究人员将深度学习与超分辨率结合起来。为了解决深度网络中训练困难的问题, Kim等人(2016a)首先引入了残差学习和递归学习来有效地开发非常深的网络,以获得更好的超分辨率性能。他们提出的VDSR(super-resolution using very deep convolutional networks)通过将更多的卷积层与残差学习堆叠在一起,构建了更深的网络。然而,深度残差网络会产生巨大的参数。DRCN(deeply-recursive convolutional network)(Kim等, 2016b)利用递归卷积在没有扩张更多参数的情况下增加了感受野,DRRN(deep recursive residual network)(Tai等, 2017)在较少参数的情况下进一步增强了接受野。Wang等人(2022)提出了一种利用多源参考图像和跨尺度双重残差融合机制来提高低分辨率人脸图像质量的方法,通过将人脸图像分为面部特征区域和背景区域来补偿高频细节。随着Transformer的出现, Liang等人(2021)在Swin Transformer的基础上提出了一种用于图像恢复的Transformer架构,实现了显著的改进,刷新了当前的性能。Li等人(2023)Swin Transformer提出了一种超分辨率网络HST(hierarchical swin Transformer),该网络能处理复杂的混合失真,重建高质量的图像。Yang等人(2023)提出了一种新的图像超分辨率方法,该方法使用双梯度回归引导来提高图像质量。虽然上述方法在性能上取得了很大的进步,但大多数方法都带来了较高的计算成本,这使得研究人员需要寻找更有效的超分辨率方法。

1.2 轻量化超分辨率模型

越来越多的轻量化超分辨率模型相继提出。Lim等人(2017)提出的EDSR(enhanced deep super-resolution network)相比于SRCNN极大地提升了图像重建质量。Ahn等人(2018)提出的CARN(cascading residual network)使用级联机制和多个 1×1 卷积减少计算消耗。Hui等人(2019)基于通道分割策略

和多特征蒸馏模块在 IDN (information distillation network) 上改进提出了 IMDN (information multi-distillation network), 该模型以较少的参数量得到了不错的重建结果。之后, Li 等人 (2020) 提出的 LAPAR-A (linearly-assembled pixel-adaptive regression) 使用邻域范围的滤波来对输入图像中的像素进行处理, 得到高分辨率图的对应像素, 该模型重建图像质量高且速度较快。Zhao 等人 (2020) 提出了一种基于像素注意力机制的轻量级超分辨率卷积神经网络 PAN (pixel attention network), 该网络结构简单而有效, 能够在保证图像质量的前提下, 大大减少计算资源的消耗。Luo 等人 (2020) 提出了一种轻量级的图像超分辨率模型 LatticeNet (lattice network), 在保持高精度的同时, 减少了参数量和计算量。雷鹏程等人 (2020) 融合空间注意力机制和分层特征融合机制, 提出一种轻量级的超分辨率网络, 在使用较少参数情况下获得了理想的重建结果。Park 等人 (2023) 提出了一种轻量的超分辨率网络 DRSAN (dynamic residual self-attention network), 该网络使用名为 DRSA (dynamic residual self-attention) 的新型自注意力机制以减少计算量。随着 Transformer 在超分辨率领域的应用, 不少基于 Transformer 的轻量化超分辨率模型也逐渐出现, 并表现出更优秀的图像重建性能。Zou 等人 (2022) 提出了一种高效的 Transformer 图像超分辨率网络 SCET, 该网络注重精简结构, 减少计算成本和内存存储。Lu 等人 (2022) 提出了一种快速而精确的图像超分辨率模型 ESRT, 其能以较低的计算成本提取深度特征。此外, Choi 等人 (2023) 使用 N -Gram 积并基于 SwinIR (image restoration using swin Transformer) 的原理提出了轻量超分辨率模型 Ngswin, 该模型相比 SwinIR 参数量和计算量都有极大的压缩, 同时也取得了不错的图像重建结果。但这些方法使用普通卷积, 该卷积计算量大。此外, 大多方法使用了参数量较多的 MLP 前馈神经

网络。

2 方法

2.1 网络整体结构

图 1 是本文提出的蓝图可分离卷积 Transformer 网络 (blueprint separable convolution Transformer network, BSTN) 的结构图。它由浅层特征提取、深层特征提取以及图像重建 3 部分组成。首先通过一个普通 3×3 卷积进行浅层特征提取, 然后将提取的浅层特征送入 4 个残差注意力 Transformer 组 (residual attention Transformer groups, RATG) 进行深层特征提取, 每个 RATG 多头注意力的头部数量不一样, 由浅入深分别设置为 1、2、4、8, 保证多头注意力计算不占用太多计算资源。然后使用一个残差连接将浅层特征与深层特征相加, 接着通过一个 3×3 普通卷积后进行像素重组得到重建的高分辨率图像。用 I_0 和 I_1 分别表示网络的输入图像和输出图像。将 I_0 输入网络后, 先进行浅层特征提取, 该过程可以表示为

$$f_0 = H_0(I_0) \quad (1)$$

式中, $H_0(\cdot)$ 表示 3×3 普通卷积操作。然后将浅层特征送入第 1 个残差注意力 Transformer 组, 提取到的特征继续送入下一个残差注意力 Transformer 组, 以此进行深层特征提取, 具体过程可表示为

$$f_1 = R_1(f_0) \quad (2)$$

$$\vdots$$

$$f_k = R_k(f_{k-1}) \quad (3)$$

式中, R_i 表示第 i 个 RATG, f_i 表示由第 i 个 RATG 提取的特征信息。将提取的深层特征信息与浅层信息相加再通过一个 3×3 普通卷积进行整合, 然后通过像素重组得到重建的高分辨率图像, 该过程可以表示为

$$f_m = ADD(f_0 + f_k) \quad (4)$$

$$f'_m = H_1(f_m) \quad (5)$$

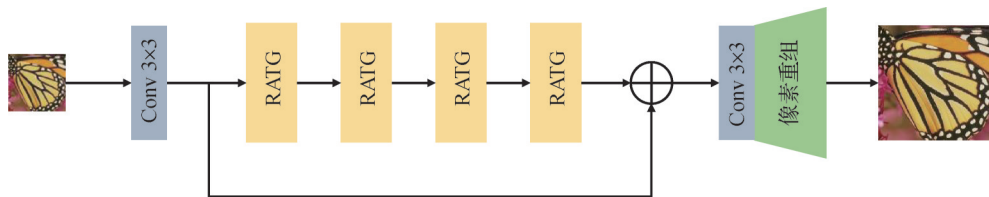


图1 蓝图可分离卷积Transformer网络结构图

Fig. 1 The structure of blueprint separable convolution Transformer network

$$I_1 = P(f'_m) \quad (6)$$

式中, f_m 表示合并的特征信息, $ADD(\cdot)$ 为张量相加操作, $H_1(\cdot)$ 表示 3×3 普通卷积操作。 f'_m 表示通过 3×3 普通卷积整合后的特征, $P(\cdot)$ 表示像素重组操作, 该方法可以避免在重建过程中可能出现的棋盘效应以及人造的伪影等问题, 最后得到重建后的高分辨率图像 I_1 。

2.2 基于移位算子和分组卷积思想的移动卷积

移位算子是精细模型设计中一种无参数量的方案。它进行了一种无参数量、无计算量的移位操作, 作为空间卷积的替代, 移位操作可以将权重矩阵或特征图中的元素按照一定的方向和步长进行平移, 从而实现空间信息的聚合。分组卷积是一种卷积操作的变种, 它将输入特征图和输出特征图按照深度方向分成若干组, 然后每组之间进行独立的卷积。分组卷积可以减少参数量和计算量, 同时增加网络的多样性和非线性。使用移位算子进行分组卷积就是移动卷积的核心思想, 图2表示移动卷积的卷积过程。

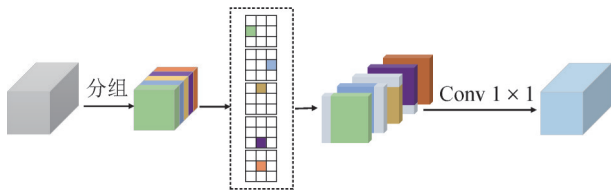


图2 移动卷积细节图

Fig. 2 The details of shift convolution

首先, 初始化一个 3×3 值为 0 的卷积权重, 将通道平均分为 5 组, 每组大小记为 g 。以二维视角来看, 权重每个点都可以用横坐标与纵坐标组成的坐标对表示, 以 $(0, 0)$ 为原点。其中 $[0, g]$ 通道权重 $(0, 1)$ 位置设为 1, 表示关注上方位信息; $[g, 2g]$ 通

道权重 $(1, 0)$ 位置设为 1, 关注左方位信息; $[2g, 3g]$ 通道权重 $(1, 2)$ 位置设为 1, 关注右方位信息; $[3g, 4g]$ 通道权重 $(2, 1)$ 位置设为 1, 关注下方位信息; $[4g, 5g]$ 通道权重 $(1, 1)$ 位置设为 1, 关注全局信息。按输入通道数量分组, 进行分组卷积。最后通过一个 1×1 卷积对通道信息进行混合。总体来看, 移动卷积可以看做是一种结构化稀疏的卷积模块, 它可以有效地减少冗余信息, 并提高网络性能。而且使用移动卷积对特征图进行维度升维后再降维, 这个过程使用的参数量远小于普通的 3×3 卷积。因此将移动卷积引入超分辨率网络, 既有利于降低网络参数量、加快推理速度, 同时又有利于减少冗余信息, 提高网络性能。

2.3 蓝图前馈神经网络

蓝图可分离卷积 (blueprint separable convolution, BSConv) (Haase 和 Amthor, 2020) 将每个卷积核表示为 1 个二维的蓝图内核和 1 个一维的权重向量, 其中蓝图内核用于提取空间信息, 权重向量用于混合通道信息。蓝图可分离卷积可以有效地减少参数量和计算量, 同时保持良好的性能。如图3所示, 从卷积过程来讲, 蓝图可分离卷积先进行逐点卷积, 再进行逐通道卷积。其中, 逐点卷积可表示为 1 个 1×1 的卷积, 逐通道卷积可表示为分组卷积。蓝图可分离卷积计算量 C 和参数量 P 计算为

$$C = M \times N \times D_F \times D_F + D_K \times D_K \times N \times D_F \times D_F \quad (7)$$

$$P = M \times N + D_K \times D_K \times N \quad (8)$$

式中, $M, N, D_F \times D_F, D_K \times D_K$ 分别表示输入数据的通道数、输出特征图的通道数、输出特征图大小和卷积核大小。而普通卷积 C 计算成本和参数量 P 计算为

$$C = M \times N \times D_K \times D_K \times D_F \times D_F \quad (9)$$

$$P = M \times N \times D_K \times D_K \quad (10)$$

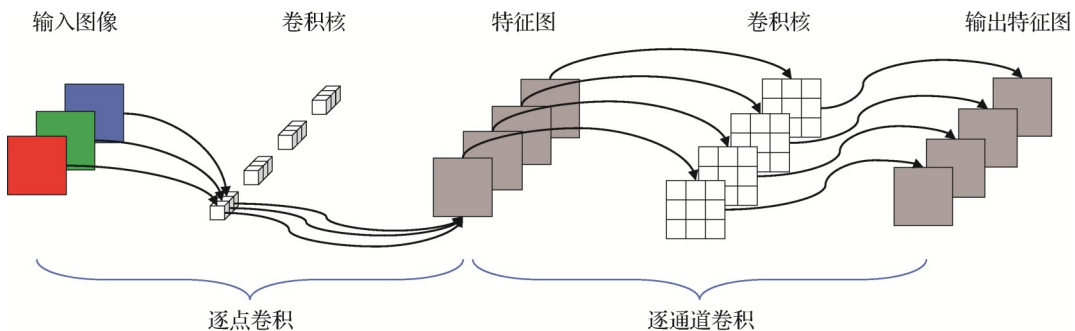


图3 蓝图可分离卷积细节图

Fig. 3 The details of blueprint separable convolution

与普通卷积相比,蓝图可分离卷积参数量和计算量都有大幅减少。

前馈神经网络是一种人工神经网络,它的节点之间的连接不形成循环。它由多个层组成,每个层包含若干个节点。每个节点接收前一层的输出,并进行线性变换和非线性激活,然后输出给下一层。在Transformer中,前馈神经网络是一个重要的组成部分,它接受自注意力层的输出,并进行两次线性变换,增加了模型的表达能力。前馈神经网络可以对复杂的过程进行拟合,弥补了自注意力机制可能存在的不足。前馈神经网络也可以提高模型的并行性,因为它不依赖于序列信息,可以同时处理多个位置的输入。Transformer中常用的前馈神经网络是多层感知机(MLP),它由多个全连接层和非线性激活函数组成。在Transformer中,每个编码器和解码器的子层都包含一个MLP,它对自注意力层的输出进行两次线性变换,并使用ReLU(rectified linear unit)或GELU(Gaussian error linear unit)作为激活函数。MLP有助于增加模型的表达能力和非线性,以及对不同维度的特征进行混合。但是MLP网络结构和参数很难确定,需要大量的试验和调整;训练过程耗时长,计算量大,容易陷入局部最优解。因此本文设计了一种更轻量,更容易使用的蓝图前馈神经网络。其结构如图4所示。输入的数据先进行层标准化(layer normalization, LN),相比于批标准化(batch normalization, BN),LN是沿特征维度进行归一化,适用性强。然后分为两个分支,两个分支都使用蓝图可分离卷积进行维度扩张,其中一条分支使用GELU激活函数,GELU是一种非单调激活函数,有助于保持小的负值,从而稳定网络梯度流。然后将

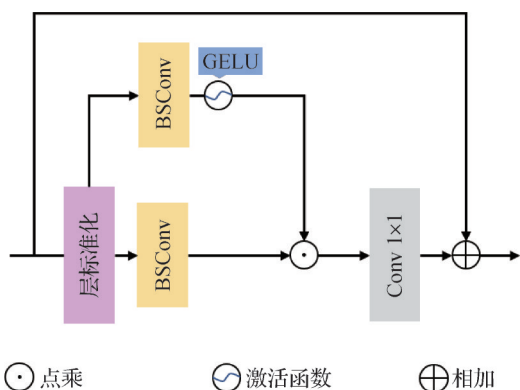


图4 蓝图前馈网络结构图

Fig. 4 The structure of blueprint feed-forward network

两个分支相乘后再由1个 1×1 卷积整合得到输出特征。这样可以实现可控的特征变换,即抑制低信息特征,仅保留有用信息。

2.4 蓝图多头自注意力

SwinIR使用了Swin Transformer的滑动窗口自注意力,它在一个限定大小的窗口中计算自注意力,同时结合移动窗口使得相邻的窗口能进行信息交互。SwinIR使用了多尺度的滑动窗口自注意力,即在不同层级上使用不同大小和步长的窗口,这样可以增加模型的灵活性和表征能力。相比于最初Transformer注意力机制,SwinIR所采用的计算方式降低了计算复杂度,节约了33%的参数量,然而这种方法训练时间长且占用显存大,不利于实际应用。因此本文设计了一种更加轻量的蓝图多头自注意力,结构如图5所示,其表示头部数为1时的自注意力计算过程。假设输入数据形状为 $X \in \mathbb{R}^{C \times H \times W}$,表示输入数据通道为 C ,高为 H ,宽为 W 。首先进行通过层标准化,对特征维度进行归一化。然后用蓝图可分离卷积得到查询(Q)、键(K)、值(V)矩阵,接着进行矩阵变维得到 $Q \in \mathbb{R}^{HW \times c}$ 、 $K \in \mathbb{R}^{C \times HW}$ 和 $V \in \mathbb{R}^{HW \times c}$ 。对 Q, K 进行 L_p 范数正则化,将输入张量在指定维度上除以该维度对应的范数,让输入张量在指定维度上具有相同的尺度和方向。之后 K 与 Q 相乘再通过激活函数得到注意力矩阵 $A \in \mathbb{R}^{C \times C}$ 。假设头部数量设置为 h ,将会生成 h 个 Q 与 K ,分别为 $Q_i \in \mathbb{R}^{HW \times c}$ 、 $K_i \in \mathbb{R}^{C \times HW}$,其中 $c = C/h$ 。然后成对相乘再经由激活函数得到 h 个注意力矩阵 $A' \in \mathbb{R}^{C \times C}$ 。将每个注意力矩阵与 V 相乘再进行拼接,此时形状为 $\mathbb{R}^{h \times HW \times c}$,经过矩阵变维得到带有注意力权重的特征矩阵 $\hat{X} \in \mathbb{R}^{C \times H \times W}$ 。最后由一个卷积整合得到输出 $Y \in \mathbb{R}^{C \times H \times W}$ 。蓝图多头自注意力,不需要进行窗口移动,同时使用蓝图可分离卷积提

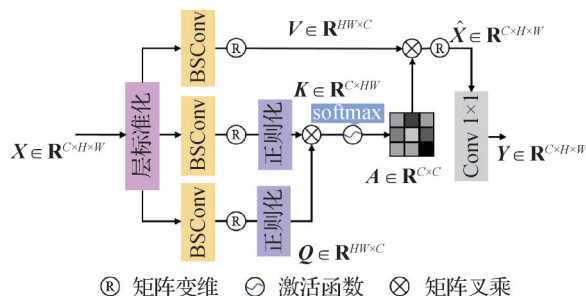


图5 蓝图多头自注意力结构图

Fig. 5 The structure of blueprint multi-head self-attention

取 Q 、 K 、 V 速度快,使用参数量少。

2.5 残差注意力Transformer组

综合以上技术与模块,本文方法设计了一种混合注意力Transformer模块HATB(hybrid attention Transformer module),结构如图6所示。它分为两个部分,首先是移动通道注意力模块(shift channel attention block, SCAB),将输入数据先进行层标准化,然后通过第1个移动卷积进行通道扩展,因为移动卷积使用移位算子进行分组卷积,在进行通道扩展时所使用的参数远少于普通 3×3 卷积。使用ReLU激活函数,保持非负单元通过,再通过第2个移动卷积对通道进行回缩。之后,利用对比度感知通道注意力(contrast-aware channel attention, CCA)(Hui等,2019)加强对不同通道信息的关注,残差连接将之前的特征信息结合起来。随后通过蓝图前馈神经网络得到移动通道注意力模块输出。移动通道注意

力模块输出的特征信息作为输入特征送入第2个部分,即蓝图多头自注意力模块(blueprint multi-head self-attention block, BMSAB)。先进行多头自注意提取,然后通过残差连接将输入特征与注意力特征矩阵相加送入蓝图前馈神经网络,抑制信息量较少的特征,只允许有用的信息进一步通过网络,再通过一个残差连接将之前特征信息相加作为蓝图多头自注意力模块的输出。由此,完成一个混合注意力Transformer模块的特征提取过程。而一个残差注意力Transformer组(RATG)由两个混合注意力Transformer模块和一个 3×3 普通卷积通过残差结构组合而成。残差注意力Transformer组是网络中的最核心部分,也是参数量占有的绝大部分,负责深层特征提取,直接关系到重建后的图像质量。本文方法通过精细化的模型设计,引入微型注意力机制,让其既有优秀的图像重建能力,同时也维持了较低的参数量和计算量。

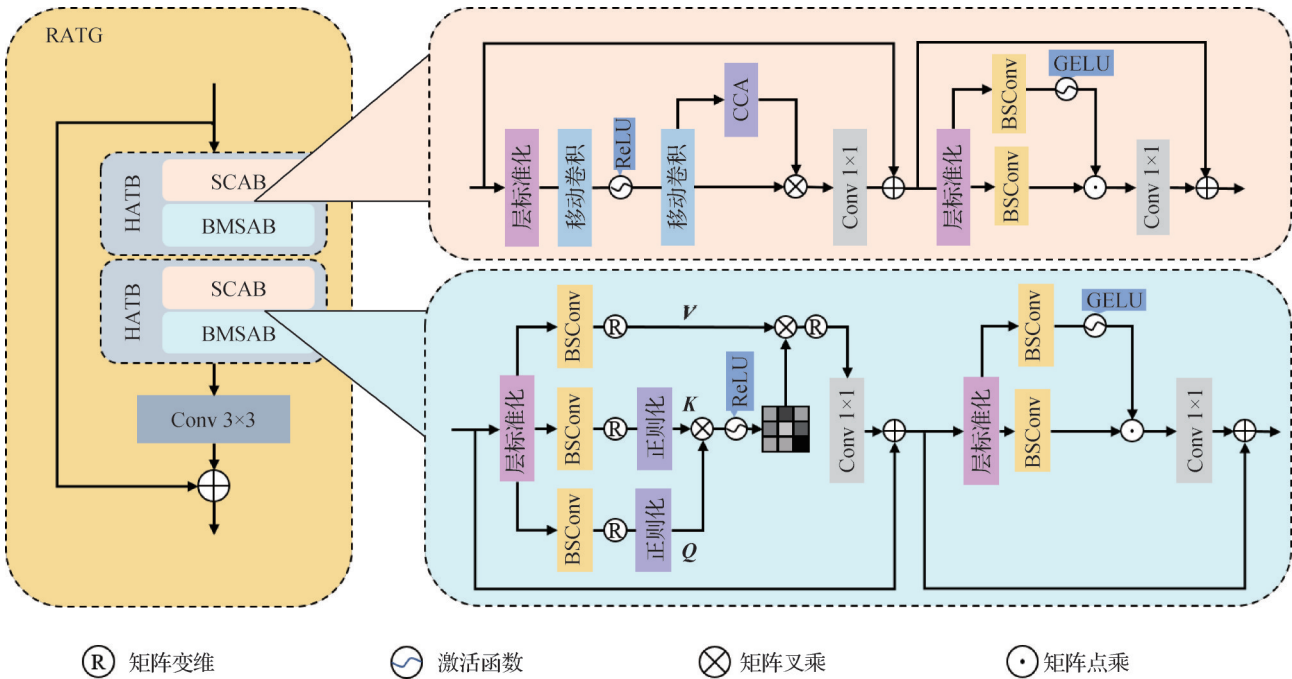


图6 残差注意力Transformer组
Fig. 6 Residual attention Transformer group

3 实验

3.1 实验设计

3.1.1 数据集

在模型训练时,选用DIV2K和Flickr2K中的801号到1 800号图像作为训练集。DIV2K数据集是一个

用于单图像超分辨率的数据集,共包含1 000幅高清高分辨率图像,分为训练集800幅,验证集100幅和测试集100幅,图像种类多样,包括动物、人像和建筑等,其中测试集并没有公开。Flickr2K数据集是一个用于图像超分辨率重建的数据集,包含2 650幅图像。此外选取了4个常用测试集进行测试,分别为Set5、Set14、BSD100和Urban100。

3.1.2 实现细节

BSTN 网络深层特征提取部分由 4 个残差注意力 Transformer 组成。其中特征通道大小为 48, 每批训练数据大小为 32, 每批输入的低分辨率图像大小为 48。详细软硬件信息见表 1。

此外, 为了增加网络的鲁棒性, 本文对输入的图片进行数据增强, 具体方法为: 将输入的图片进行 90°、180° 和 270° 的随机旋转和水平翻转。本文模型采用 Adam (Kingma 和 Ba, 2017) 优化器进行参数更新, 优化器参数为 $\beta_1 = 0.9, \beta_2 = 0.99$, 初始学习率设置为 5×10^{-4} , 使用余弦学习率下降策略对学习率进行更新。它通常与“重启”结合使用, 即当学习率达到最小值时, 再次提升到最大值 (一般与原始最大值不同), 然后再次衰减, 总共迭代训练 10^6 次。所有超参数均在一个 yml 文件中配置。最后本文网络使用 PyTorch 框架进行编写, 并在一张 NVIDIA RTX 3090 显卡上进行训练, 训练时间约为 46 h。

表 1 实验平台软硬件环境

Table 1 Software and hardware environment of the experimental platform

类别	型号/版本号
CPU 型号	Intel(R) Xeon(R) Gold 6240C
显卡型号	NVIDIA GeForce RTX 3090
显存	24 576 M
操作系统	Ubuntu 20.04.1
Python	3.8.13
PyTorch	1.13.0
torchsummaryX	1.3.0
CUDA	11.7

3.1.3 评价指标

本文采用峰值信噪比 (peak signal-to-noise ratio, PSNR) 和结构相似性 (structural similarity, SSIM) 作为评价指标, 这也是目前超分辨率领域中常用的客观评价指标。同时, 定量结果在 YCbCr 空间上的 Y 通道上计算。

3.2 对比实验

为了证明本文方法的优越性, 本节将与领域内先进方法 SRCNN (Dong 等, 2014)、VDSR (Kim 等, 2016a)、CARN (Ahn 等, 2018)、EDSR (Lim 等, 2017)、IMDN (Hui 等, 2019)、PAN (Zhao 等, 2020)、LAPAR-

A (Li 等, 2020)、Lattice (Luo 等, 2020)、ESRT (Lu 等, 2022)、NGSwin (Choi 等, 2023) 和 LBNNet (Gao 等, 2022) 进行对比。本文方法与其他先进方法在 Urban100 上放大 4 倍时, PSNR、参数数量和浮点运算量的可视化对比见图 7, 可以发现本文方法将参数数量和浮点运算量都维持较低水平, 同时 PSNR 取得最佳。

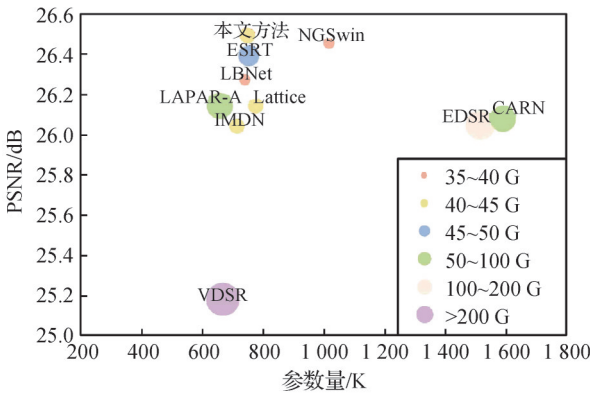


图 7 PSNR、参数量、浮点运算量对比图

Fig. 7 Comparison between PSNR, parameters and flops

3.2.1 客观实验结果

各方法在数据集 Set5、Set14、BSD100 和 Urban100 上的参数量、浮点运算量以及测试结果见表 2—表 4。其中, 浮点运算量采用大小为 $3 \times 1280 \times 720$ 像素图像计算。SRCNN 是第 1 个基于深度学习的图像超分辨率重建网络, 可以发现自 SRCNN 提出以来, 各方法在重建结果上都有了比较大的提升。因为 SRCNN 只用了两个卷积层, 因此参数量较少, 但是它先将图像放大再进行特征提取, 浮点运算量大。在其他方法中, PAN 的参数量最小, 但由于它并没有很好地挖掘深层特征, 所以在 4 个数据集上表现较差。总体上看, 本文方法参数量和计算量都处于一个适中位置, 与 Transformer 类方法相比处于一个较低位置。

以 2 倍和 4 倍为例, 从 2 倍超分辨率结果来看, 本文方法在 PSNR 上取得最佳, 在 Set5 和 Set14 上领先明显, 分别优于性能第 2 的方法 0.11 dB 和 0.08 dB。在 Urban100 上 SSIM 略低于 NGSwin, 取得第 2 名, 但在其他数据集取得最佳, 这是因为 NGSwin 使用了移动窗口, 因此能获得较多的结构信息。而相比 NGSwin, 本文方法在重建图像时保留了较多的细节信息, 但是损失了一些结构信息。在 4 倍超分辨率上, 本文方法延续了优异的表现, PSNR 在各数据集

表2 不同方法的2倍超分辨率客观评价结果

Table 2 Results of objective results of 2 × super-resolution for different methods

方法	参数量/K	浮点 运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM
SRCNN(Dong 等,2014)	8	52.7	36.66	0.954 2	32.45	0.906 7	31.36	0.887 9	29.50	0.894 6
VDSR(Kim 等,2016a)	666	612.6	37.53	0.958 7	33.03	0.912 4	31.90	0.896 0	30.76	0.914 0
CARN(Ahn 等,2018)	1 592	222.8	37.76	0.959 0	33.52	0.916 6	32.09	0.897 8	31.92	0.925 6
EDSR(Lim 等,2017)	1 370	316.2	37.99	0.960 4	33.57	0.917 5	32.16	0.899 4	31.98	0.927 2
IMDN(Hui 等,2019)	694	158.8	38.00	0.960 5	33.63	0.917 7	32.19	0.899 6	32.17	0.928 3
PAN(Zhao 等,2020)	261	70.5	38.00	0.960 5	33.59	0.918 1	32.18	0.899 7	32.01	0.927 3
LAPAR-A(Li 等,2020)	548	171.0	38.01	0.960 5	33.62	0.918 3	32.19	0.899 9	32.10	0.928 3
Lattice(Luo 等,2020)	756	169.5	38.06	0.960 7	33.70	0.918 7	32.20	0.899 9	32.25	0.928 8
ESRT(Lu 等,2022)	677	118.1	38.03	0.960 0	33.75	0.918 4	32.25	0.900 1	32.58	0.931 8
NGSwin(Choi 等,2023)	998	140.4	38.05	0.961 0	33.79	0.919 9	32.27	0.900 8	32.53	0.932 4
LBNet(Gao 等,2022)	731	153.2	38.05	0.960 7	33.65	0.917 7	32.16	0.899 4	32.30	0.929 1
本文	736	163.6	38.17	0.961 6	33.87	0.920 1	32.28	0.901 0	32.60	0.931 8

注:加粗字体表示各列最优结果。

表3 不同方法的3倍超分辨率客观评价结果

Table 3 Results of objective results of 3 × super-resolution for different methods

方法	参数量/K	浮点 运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM
SRCNN(Dong 等,2014)	8	52.7	32.75	0.909 0	29.30	0.821 5	28.41	0.786 3	26.24	0.798 9
VDSR(Kim 等,2016a)	666	612.6	33.66	0.921 3	29.77	0.831 4	28.82	0.797 6	27.14	0.827 9
CARN(Ahn 等,2018)	1 592	118.8	34.29	0.925 5	30.29	0.840 7	29.06	0.803 4	28.06	0.849 3
EDSR(Lim 等,2017)	1 555	162.3	34.37	0.927 0	30.28	0.841 7	29.09	0.805 2	28.15	0.852 7
IMDN(Hui 等,2019)	703	71.5	34.36	0.927 0	30.32	0.841 7	29.09	0.804 6	28.17	0.851 9
PAN(Zhao 等,2020)	261	39.0	34.40	0.927 1	30.36	0.842 3	29.11	0.805 0	28.11	0.851 1
LAPAR-A(Li 等,2020)	544	114	34.36	0.926 7	30.34	0.842 1	29.11	0.805 4	28.15	0.852 3
Lattice(Luo 等,2020)	765	76.3	34.40	0.927 2	30.32	0.841 6	29.10	0.804 9	28.19	0.851 3
ESRT(Lu 等,2022)	770	61.8	34.42	0.926 8	30.43	0.843 3	29.15	0.806 3	28.46	0.857 4
NGSwin(Choi 等,2023)	1 007	66.6	34.52	0.928 2	30.53	0.845 6	29.19	0.807 8	28.52	0.860 3
LBNet(Gao 等,2022)	736	68.4	34.47	0.927 7	30.38	0.841 7	29.13	0.806 1	28.42	0.855 9
本文	742	72.2	34.68	0.929 6	30.54	0.845 1	29.23	0.808 1	28.62	0.860 4

注:加粗字体表示各列最优结果。

上都取得不同程度领先,其中在Set5上领先明显,领先第2的NGSwin 0.17dB。SSIM只在Set14上落后于NGSwin,在其他数据集上都取得领先。另外,本

文方法只用了NGSwin约74%的参数量就实现了整体优于NGSwin的重建性能。ESRT提出一种高频过滤模块提取高频信息,本文方法使用了与它相同的

表 4 不同方法的 4 倍超分辨率客观评价结果

Table 4 Results of objective results of 4 × super-resolution for different methods

方法	参数量/K	浮点运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM
SRCNN(Dong 等,2014)	8	52.7	30.48	0.862 6	27.50	0.751 3	26.90	0.710 1	24.52	0.722 1
VDSR(Kim 等,2016a)	666	612.6	31.35	0.883 8	28.01	0.767 4	27.29	0.725 1	25.18	0.752 4
CARN(Ahn 等,2018)	1 592	90.9	32.13	0.893 7	28.60	0.780 6	27.58	0.734 9	26.07	0.783 7
EDSR(Lim 等,2017)	1 518	114.2	32.09	0.893 8	28.58	0.781 3	27.57	0.735 7	26.04	0.784 9
IMDN(Hui 等,2019)	715	40.9	32.21	0.894 8	28.58	0.781 1	27.56	0.735 3	26.04	0.783 8
PAN(Zhao 等,2020)	272	28.2	32.13	0.894 8	28.61	0.782 2	27.59	0.736 3	26.11	0.785 4
LAPAR-A(Li 等,2020)	659	94.0	32.15	0.894 4	28.61	0.781 8	27.61	0.736 6	26.14	0.787 1
Lattice(Luo 等,2020)	777	43.6	32.18	0.894 3	28.61	0.781 2	27.57	0.735 5	26.14	0.784 4
ESRT(Lu 等,2022)	751	47.7	32.19	0.894 7	28.69	0.783 3	27.69	0.737 9	26.39	0.796 2
NGSwin(Choi 等,2023)	1 019	36.4	32.33	0.895 5	28.78	0.785 9	27.66	0.739 6	26.45	0.796 3
LBNet(Gao 等,2022)	742	38.9	32.29	0.896 0	28.68	0.783 2	27.62	0.738 2	26.27	0.790 6
本文	751	41.8	32.50	0.898 5	28.78	0.785 2	27.71	0.740 3	26.49	0.796 5

注:加粗字体表示各列最优结果。

参数量,但在重建结果上全面超过了ESRT。由此可以看出,本文方法具有更好的图像重建能力。综合来看,本文方法作为Transformer架构的模型,以较少的参数和计算量取得了显著的超分辨率客观评价结果。

3.2.2 主观实验结果

图8—图10分别展示了不同方法在2倍、3倍以及4倍超分辨率重建下的结果,其中GT为原始高分辨率图像,本文方法表现出了更好的超分辨率重建性能。本文以2倍和4倍的结果为例来进行对比分析。

图8上半部分为Set14中图像ppt3上的2倍超分辨率的重建结果,该图像包含了大量文字。观测其他方法重建出的图像,可以发现它们在文字上存在着模糊和平滑的问题,尤其在“your point”单词上的差异非常明显。尽管LBNet的重建结果相对于其他方法更好,但是在“point”一词的字母“i”上仍然不如本文方法清晰;图8下半部分为Urban100中图像img012上的2倍超分辨率的重建结果,该图像包含了建筑物中大量重复的纹理和细节。可以发现,IMDN、Lattice 和 ESRT 等方法产生了不同程度的模糊和大量与GT图像方向不一致的线条,而本文方法没有产生与GT图像不一致的线条,且重建图像更加

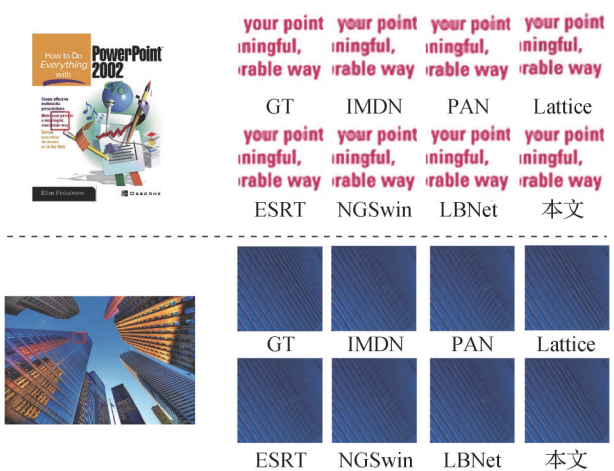


图 8 2 倍超分辨率可视化对比
Fig. 8 Comparison of different methods for 2× super-resolution visualization

清晰。上述分析表明,本文方法在2倍超分辨率重建时优于其他对比方法。在进行4倍超分辨率重建时,退化图像有着更小的空间分辨率,因此在重建高分辨率图像的过程中需要恢复更多的信息,难度更大。

图10上半部分展示了放大倍数为4时不同方法在Urban中图像img058上的重建结果。可以看到,IMDN、PAN 和 NGSWin 等方法尤其表现出严重的模

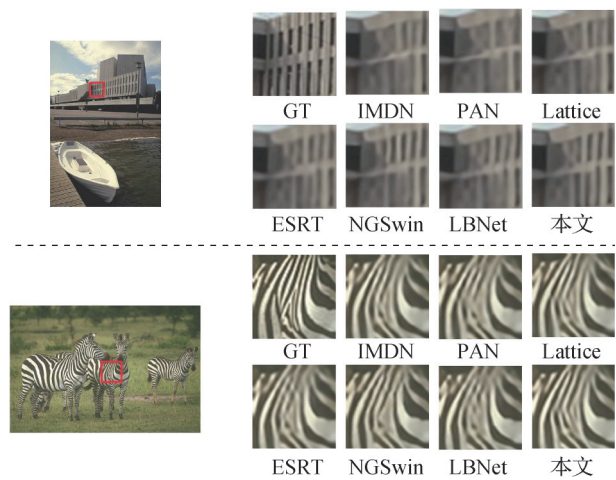


图9 3倍超分辨率可视化对比
Fig. 9 Comparison of different methods for 3× super-resolution visualization

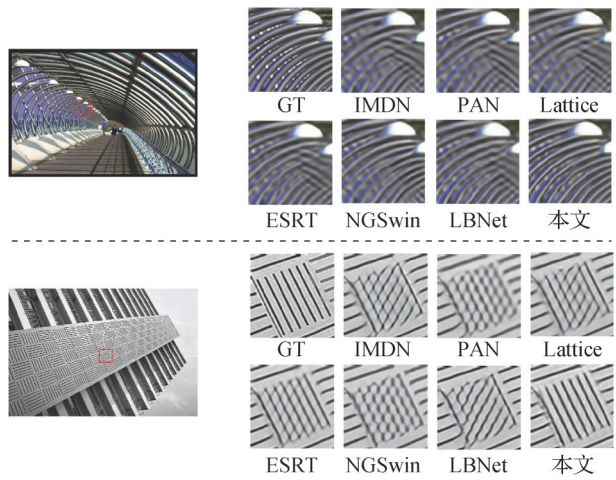


图10 4倍超分辨率可视化对比
Fig. 10 Comparison of different methods for 4× super-resolution visualization

糊,并产生了大量错乱的线条。而本文方法可以恢复出与GT图像几乎一致的线条和纹理,相比于其他

对比方法有着更好的重建性能。图10下半部分展示了放大倍数为4时不同方法在Urban100中图像img092上的重建结果。可以很明显看出,其他对比方法重建出的图像都出现了不同程度的线条错乱和模糊,比如PAN产生了波动和模糊的线条结构,ESRT在GT图像的基础上产生了多余的倾斜线条,LBNet产生了与GT图像方向不一致的线条等。而本文方法产生的图像模糊区域小、更加清晰且线条基本拟合GT图像。

3.3 消融实验

3.3.1 RATG数量变化的性能分析

残差注意力Transformer组是网络最主要部分,其包含了网络的大部分参数和计算量,通过对其增删可改变模型大小,但是残差注意力Transformer组数量的改变自然也会对模型的性能产生影响。为此,本文进行了残差注意力Transformer组数量变化的消融实验,其结果见表5。表5中数量为3,表示网络中使用了3个残差注意力Transformer组,其他数量以此类推。因为超分辨率模型训练时间较长,故本文将所有消融实验迭代训练次数设置为 5×10^5 。每个实验采用的训练数据集相同,超参数相同。可以发现RATG数量为3的模型参数量和浮点运算量最低,但是这样带来的影响就是PSNR和SSIM客观指标的下降。如果将RATG数量增加到5后,模型参数增加至923 K,浮点运算量增加到50.8 G,但是除了在Urban100上比数量为4的模型有微小提升,在其他测试集上都有不同程度的退化。因此证明将残差注意力Transformer组的数量设置为4能取得更好的超分结果,同时在参数量和浮点运算量维持一个较低的水平。

表5 不同数量RATG的4倍超分辨率结果
Table 5 Results of 4 × super-resolution for different number of RATGs

数量	参数量/K	浮点 运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
3	579	31.7	32.35	0.897 5	28.70	0.7835	27.65	0.7386	26.31	0.7919
4	715	41.8	32.43	0.898 3	28.75	0.7848	27.68	0.7393	26.39	0.7939
5	923	50.8	32.42	0.898 6	28.75	0.7847	27.66	0.7391	26.39	0.7943

注:加粗字体表示各列最优结果。

3.3.2 蓝图前馈神经网络有效性分析

Transformer模型中常用的前馈神经网络为MLP

网络,但是由于MLP通常有以下几个缺点:容易过拟合,训练速度慢,参数量大。为了证明蓝图前馈神经

网络(blueprint feed-forward neural network, BFFN)的有效性,本实验将在BSTN的基础框架上将蓝图前馈神经网络替换为MLP网络,并命名为BSTN_1,然后以相同的数据集和超参数进行训练。实验结果出现较大的退化,这是因为MLP训练困难,通常需要多次实验才能达到一个较好的结果。为了验证是否因学

习率过大,出现振荡导致结果不佳,将学习率更改为 10^{-4} 进行训练,该方法命名为BSTN_2。结果见表6。可以发现,使用MLP替换BFFN后,浮点运算量有较为明显的降低,但无论是BSTN_1还是BSTN_2在各个测试集上的超分结果都有巨大退化,同时参数量也有部分增加。由此证明蓝图前馈神经网络的有效性。

表6 替换不同前馈神经网络的4倍超分辨率结果

Table 6 Results of 4 × super-resolution for different feed-forward neural network in BSTN

方法	参数量/K	浮点运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
BSTN_1	795	22.8	32.07	0.893 8	28.49	0.778 6	27.50	0.733 9	25.89	0.779 0
BSTN_2	795	22.8	32.07	0.894 3	28.52	0.779 4	27.52	0.734 6	25.89	0.779 9
BSTN	715	41.8	32.43	0.898 3	28.75	0.784 8	27.68	0.739 3	26.39	0.793 9

注:加粗字体表示各列最优结果。

3.3.3 移动卷积有效性分析

移动卷积可以在减少卷积计算量和参数量的同时更广泛地获取特征空间信息,为了证明移动卷积的这一优点,在本消融实验中将移动卷积替换为普通 3×3 卷积,然后在相同的数据集,以相同的超参数迭代训练 5×10^5 次。在Set5、Set14、BSD100和Urban100数据集上进行测试,得到

PSNR和SSIM结果,如表7所示。可以明显看出,使用普通 3×3 卷积替换移动卷积后,参数量和浮点运算量急剧膨胀,但是在各测试集上的表现反而退化不少,由此证明了移动卷积的有效性,同时也说明参数量和浮点运算量的增加不一定促使结果变好,往往更加优秀的结构设计更能保证最后的结果。

表7 移动卷积4倍超分辨率的消融实验结果

Table 7 Results of 4 × super-resolution ablation experiments for shift convolution

数量	参数量/K	浮点运算量/G	Set5		Set14		BSD100		Urban100	
			PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
普通 3×3 卷积	1 914	109.7	32.39	0.897 7	28.68	0.783 4	27.64	0.738 6	26.22	0.789 2
移动卷积	715	41.8	32.43	0.898 3	28.75	0.784 8	27.68	0.739 3	26.39	0.793 9

注:加粗字体表示各列最优结果。

4 结 论

Transformer引入到超分辨率领域以来,取得了巨大成功,相比于卷积神经网络,在重建质量上有较大的提升。但是随之而来的是庞大的参数量和计算消耗,这导致Transformer结构的超分辨率模型难以应用到计算资源有限、显存小,且要求实时的任务中。为解决该痛点,本文提出了一种轻量级图像超分辨率的蓝图可分离卷积Transformer网络(BSTN)。针对传统前馈神经网络MLP训练过程耗时长,计算

量大,容易陷入局部最优解的问题,本文提出了蓝图前馈神经网络进行替代,同时使用计算量、参数量较小的蓝图可分离卷积和移动卷积对模型进行压缩。另外,基于蓝图可分离卷积提出了一种轻量的蓝图多头自注意力模块,以较低的计算量实现了自注意力计算过程。

大量实验证明了本文方法在Set5、Set14、BSD100和Urban100上都取得领先结果。在主观实验结果中,本方法重建图像细节信息更丰富,质量更高。

值得指出的是,虽然本文方法在图像重建方面取得了出色的结果,但在参数量和浮点运算量方面

的优势并不明显。因此,未来的工作可以在以下几个方面进行扩展和丰富,以增强轻量化图像超分辨率重建方法的效率和适用性。首先,需要优化模型架构,减少参数数量和计算复杂度。可以探索先进的模型压缩技术,研究针对超分辨率任务量身定制的新颖轻量化操作。其次,需要在不同的数据集和场景上进行广泛实验,全面评估所提方法的性能和泛化能力。重要的是要评估模型在各种图像类型下的稳健性,包括自然场景、医学图像和遥感数据等。最后,将本文模型部署至移动设备和边缘设备上,研究针对移动设备和边缘设备的模型压缩和优化技术,以进一步提高超分辨率重建过程的效率和实时性能。总之,未来的工作应着重优化本文模型,解决参数量和计算成本的挑战,并探索移动设备和边缘设备上的部署策略。通过进一步提高所提方法的效率和实用性,可以充分发挥其在实际超分辨率图像重建应用中的潜力。

参考文献(References)

- Ahn N, Kang B and Sohn K A. 2018. Fast, accurate, and lightweight super-resolution with cascading residual network//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 256-272 [DOI: 10.1007/978-3-030-01249-6_16]
- Choi H, Lee J and Yang J. 2023. N-Gram in swin Transformers for efficient lightweight image super-resolution//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2071-2081 [DOI: 10.1109/CVPR52729.2023.00206]
- Dong C, Loy C C, He K M and Tang X O. 2014. Learning a deep convolutional network for image super-resolution//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: IEEE: 184-199 [DOI: 10.1007/978-3-319-10593-2_13]
- Gao G W, Wang Z X, Li J C, Li W J, Yu Y and Zeng T Y. 2022. Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive Transformer//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI.org: 913-919 [DOI: 10.24963/ijcai.2022/128]
- Gao Y, Liu Z, Qin P L and Wang L F. 2018. Medical image super-resolution algorithm based on deep residual generative adversarial network. *Journal of Computer Applications*, 38(9): 2689-2695 (高媛, 刘志, 秦品乐, 王丽芳. 2018. 基于深度残差生成对抗网络的医学影像超分辨率算法. *计算机应用*, 38(9): 2689-2695) [DOI: 10.11772/j.issn.1001-9081.2018030574]
- Girshick R, Donahue J, Darrell T and Malik J. 2016. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1): 142-158 [DOI: 10.1109/TPAMI.2015.2437384]
- Haase D and Amthor M. 2020. Rethinking depthwise separable convolutions: how intra-kernel correlations lead to improved MobileNets//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14588-14597 [DOI: 10.1109/CVPR42600.2020.01461]
- Hui Z, Gao X B, Yang Y C and Wang X M. 2019. Lightweight image super-resolution with information multi-distillation network//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 2024-2032 [DOI: 10.1145/3343031.3351084]
- Jian M W and Lam K M. 2015. Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(11): 1761-1772 [DOI: 10.1109/TCSVT.2015.2400772]
- Kim J, Lee J K and Lee K M. 2016a. Accurate image super-resolution using very deep convolutional networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1646-1654 [DOI: 10.1109/CVPR.2016.182]
- Kim J, Lee J K and Lee K M. 2016b. Deeply-recursive convolutional network for image super-resolution//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1637-1645 [DOI: 10.1109/CVPR.2016.181]
- Kingma D P and Ba J L. 2017. Adam: a method for stochastic optimization [EB/OL]. [2023-04-13]. <https://arxiv.org/pdf/1412.6980.pdf>
- Lei P C, Liu C, Tang J G and Peng D L. 2020. Hierarchical feature fusion attention network for image super-resolution reconstruction. *Journal of Image and Graphics*, 25(9): 1773-1786 (雷鹏程, 刘丛, 唐坚刚, 彭敦陆. 2020. 分层特征融合注意力网络图像超分辨率重建. *中国图象图形学报*, 25(9): 1773-1786) [DOI: 10.11834/jig.190607]
- Li B C, Li X, Lu Y T, Liu S, Feng R Y and Chen Z B. 2023. HST: hierarchical swin Transformer for compressed image super-resolution//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer: 651-668 [DOI: 10.1007/978-3-031-25063-7_41]
- Li W B, Zhou K, Qi L, Jiang N J, Lu J B and Jia J Y. 2020. LAPAR: linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 20343-20355
- Liang J Y, Cao J Z, Sun G L, Zhang K, van Gool L and Timofte R. 2021. SwinIR: image restoration using swin Transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 1833-1844 [DOI: 10.1109/ICCVW54120.2021.00210]
- Lim B, Son S, Kim H, Nah S and Lee K M. 2017. Enhanced deep

- residual networks for single image super-resolution//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 1132-1140 [DOI: 10.1109/CVPRW.2017.151]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Liu Z S, Li J C, Liu H, Huang C Y, Zhang L L and Zeng T Y. 2022. Transformer for single image super-resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 456-465 [DOI: 10.1109/CVPRW56347.2022.00061]
- Luo X T, Xie Y, Zhang Y L, Qu Y Y, Li C H and Fu Y. 2020. LatticeNet: towards lightweight image super-resolution with lattice block//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 272-289 [DOI: 10.1007/978-3-030-58542-6_17]
- Park K, Soh J W and Cho N I. 2023. A dynamic residual self-attention network for lightweight single image super-resolution. IEEE Transactions on Multimedia, 25: 907-918 [DOI: 10.1109/TMM.2021.3134172]
- Qiu D F, Jiang J J, Hu X Y, Liu X M and Ma J Y. 2023. Guided Transformer for high-resolution visible image guided infrared image super-resolution. Journal of Image and Graphics, 28(1): 196-206 (邱德粉, 江俊君, 胡星宇, 刘贤明, 马佳义. 2023. 高分辨率可见光图像引导红外图像超分辨率的Transformer网络. 中国图象图形学报, 28(1): 196-206) [DOI: 10.11834/jig.220604]
- Tai Y, Yang J and Liu X M. 2017. Image super-resolution via deep recursive residual network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2790-2798 [DOI: 10.1109/CVPR.2017.298]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wang R, Jian M W, Yu H, Wang L and Yang B. 2022. Face hallucination using multisource references and cross-scale dual residual fusion mechanism. International Journal of Intelligent Systems, 37(11): 9982-10000 [DOI: 10.1002/int.23024]
- Wang S Y and Shen L S. 2007. Intelligent visual surveillance technology: a survey. Journal of Image and Graphics, 12(9): 1505-1514 (王素玉, 沈兰荪. 2007. 智能视觉监控技术研究进展. 中国图象图形学报, 12(9): 1505-1514) [DOI: 10.3969/j.issn.1006-8961.2007.09.001]
- Wang Z H, Chen J and Hoi S C H. 2021. Deep learning for image super-resolution: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3365-3387 [DOI: 10.1109/TPAMI.2020.2982166]
- Yang F Z, Yang H, Fu J L, Lu H T and Guo B N. 2020. Learning texture Transformer network for image super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5790-5799 [DOI: 10.1109/CVPR42600.2020.00583]
- Yang H L, Zhang Y J, Cui Z W, Xu Y J and Yang Y T. 2023. DGRN: image super-resolution with dual gradient regression guidance. Computers and Graphics, 110: 141-150 [DOI: 10.1016/j.cag.2022.12.005]
- Zhao H Y, Kong X T, He J W, Qiao Y and Dong C. 2020. Efficient image super-resolution using pixel attention//Proceedings of the Computer Vision——ECCV 2020 Workshops. Glasgow, UK: Springer: 56-72 [DOI: 10.1007/978-3-030-67070-2_3]
- Zou W B, Ye T, Zheng W X, Zhang Y C, Chen L and Wu Y. 2022. Self-calibrated efficient Transformer for lightweight super-resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, USA: IEEE: 929-938 [DOI: 10.1109/CVPRW56347.2022.00107]

作者简介

毕修平,男,硕士研究生,主要研究方向为图像超分辨率。

E-mail: bixiuping@whu.edu.cn

张乐飞,通信作者,男,教授,主要研究方向为图像处理和模式识别。E-mail: zhanglefei@whu.edu.cn

陈实,男,博士研究生,主要研究方向为图像复原。

E-mail: chenshi@whu.edu.cn