

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2024)04-0908-14

论文引用格式: Sun Y, Ding J W, Zhang Q and Deng Q Y. 2024. Images super-resolution reconstruction of transposed self-attention with local feature enhancement. Journal of Image and Graphics, 29(04):0908-0921(孙阳, 丁建伟, 张琪, 邓琪瑶. 2024. 局部特征增强的转置自注意力图像超分辨率重建. 中国图象图形学报, 29(04):0908-0921)[DOI:10.11834/jig.230320]

局部特征增强的转置自注意力图像超分辨率重建

孙阳, 丁建伟*, 张琪, 邓琪瑶

中国人民公安大学信息安全学院, 北京 100038

摘要: 目的 超分辨率(super resolution, SR)重建任务通过划分窗口引入自注意力机制进行特征提取, 获得了令人瞩目的成绩。针对划分窗口应用自注意力机制时会限制图像信息聚合范围、制约模型对特征信息进行建模的问题, 本文基于转置自注意力机制构建全局和局部信息建模网络捕捉图像像素依赖关系。**方法** 首先采用轻量的基线模型对特征进行简单关系建模, 然后将空间维度上的自注意力机制转换到通道维度, 通过计算交叉协方差矩阵构建各像素点之间的长距离依赖关系, 接着通过引入通道注意力块补充图像重建所需的局部信息, 最后构建双门控机制控制信息在模型中的流动, 提高模型对特征的建模能力及其鲁棒性。**结果** 实验在5个基准数据集Set5、Set14、BSD100、Urban100、Manga109上与主流方法进行了比较, 在不同比例因子的SR任务中均获得了最佳或者次佳的结果。与SwinIR(image restoration using swin Transformer)在 $\times 2$ 倍SR任务中相比, 在以上5个数据集上的峰值信噪比分别提升了0.03 dB、0.21 dB、0.05 dB、0.29 dB和0.10 dB, 结构相似度也获得了极大提升, 同时视觉感知优化十分明显。**结论** 所提出的网络模型能够更充分地对特征信息全局关系进行建模, 同时也不会丢失图像特有的局部相关性。重建图像质量明显提高, 细节更加丰富, 充分说明了本文方法的有效性与先进性。

关键词: 图像超分辨率重建; 自注意力机制; 深度学习; 图像复原; 门控网络

Images super-resolution reconstruction of transposed self-attention with local feature enhancement

Sun Yang, Ding Jianwei*, Zhang Qi, Deng Qiyao

School of Information Network Security, People's Public Security University of China, Beijing 100038, China

Abstract: Objective Research on super-resolution image reconstruction based on deep learning techniques has gained exceptional progress in recent years. In particular, when the development of traditional convolutional neural networks reached a bottleneck, Transformer, which performs extremely well in natural language processing, was introduced to approximate super-resolution image reconstruction. However, the computational complexity of Transformer is related to the square of the width and height of the input image, leading to the inability to migrate Transformer to low-level computer vision tasks fully. Recent methods, such as image restoration using Swin Transformer (SwinIR), have achieved excellent performance by dividing windows, performing self-attention within the windows and interacting the information between the windows. However, this method of dividing windows increases the computational burden as the window size increases. Moreover, the window division method cannot model the global information of images completely, resulting in partial loss of

收稿日期: 2023-05-30; 修回日期: 2023-08-30; 预印本日期: 2023-09-06

* 通信作者: 丁建伟 jwding@ppsuc.edu.cn

基金项目: 中央高校基本科研业务费专项资金资助(2022JKF02011)

Supported by: Fundamental Research Funds for the Central Universities (2022JKF02011)

information. To solve the above problems, we model the long-range dependencies of images by constructing a Transformer block while maintaining a moderate level of the number of parameters. Excellent super-resolution reconstruction performance is achieved by constructing global dependencies of features. **Method** The proposed super-resolution network based on self-attention (SRTSA) consists of four main stages: a shallow feature extraction module, a deep feature special extraction module, an image upsampling module, and an image reconstruction module. The shallow feature extraction part consists of a 3×3 convolution. The deep feature extraction part mainly consists of a global and local information extraction block (GLIEB). Our proposed GLIEB performs simple relational modeling through a sufficiently lightweight nonlinear activation free block (NAFBlock). Although dropout can improve the robustness of the model, we discard the dropout layer to avoid losing other information before modeling the feature information globally. In the global modeling of feature information using the transposed self-attention mechanism, we keep the features with positive effects on image reconstruction and discard the features with negative effects by replacing the softmax activation function in the self-attention mechanism with the ReLU activation function, which makes the reconstructed global dependencies more robust. Given that an image includes global and local information, a residual channel attention module is used to supplement the local information and enhance the expressive ability of the model. Furthermore, a new dual-channel gating mechanism is introduced to control the flow of information in the model to improve the modeling capability of the model for features and its robustness. The image upsampling module uses subpixel convolution to expand the features to the target dimension, and the reconstruction module employs a 3×3 convolution to obtain the final reconstruction results. For the loss function, although many loss functions have been proposed to optimize model training, to demonstrate the advancement and effectiveness of our model, we use the same L1 loss function as that of SwinIR to supervise the model training. The L1 loss function can provide a stable gradient that allows the model to converge quickly. In the image training phase, 800 images from the DIV2K dataset are used for training. The 800 training images are randomly rotated or horizontally flipped to expand the dataset, and 16 LR image blocks of size 48×48 pixels are used as input in each iteration. The Adam optimizer is used for training. **Result** We test on five datasets commonly used in super-resolution tasks, namely, Set5, Set14, Berkeley segmentation dataset 100 (BSD100), Urban100, and Manga109, to demonstrate the effectiveness and robustness of the proposed method. We also compare the proposed method with SRCNN, VDSR, EDSR, RCAN, SAN, HAN, NLSA, and SwinIR networks in terms of objective metrics. These networks are supervised using the L1 loss function during the training process. The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are calculated on the Y channel of the YCbCr space of the output image to measure the image reconstruction effect. Experimental results show that the PSNR and SSIM values obtained our method are both optimal. In the $\times 2$ super-resolution tasks, compared with those of SwinIR, the PSNR of the proposed method is improved by 0.03 dB, 0.21 dB, 0.05 dB, 0.29 dB, and 0.10 dB, and the SSIM is enhanced by 0.000 4, 0.001 6, 0.000 9, and 0.002 7 on four datasets, except Manga109. The reconstruction effect demonstrates that SRTSA can recover more detailed information and more texture structure compared with most methods. From the attribution analysis of the model using local attribution maps (LAM), SRTSA uses a larger range of pixels in the reconstruction process compared with other methods, such as SwinIR, which fully illustrates the global modeling capability of SRTSA. **Conclusion** The proposed super-resolution image reconstruction algorithm based on a transposed self-attention mechanism can fully model the global relationship of feature information without losing the local relationship of features by converting the global relationship modeling in the spatial dimension into a channel dimension for global relationship modeling. It also contains global and local information, which effectively improves the image super-resolution reconstruction performance. The excellent PSNR and SSIM on five datasets and the significantly high quality of the reconstructed images with rich details and sharp edges fully demonstrate the effectiveness and advancedness of the proposed method.

Key words: image super-resolution reconstruction; self-attention; deep learning; image restoration; gated networks

0 引言

单图像超分辨率(single image super-resolution, SISR)技术旨在将低分辨率(low resolution, LR)图像重建为高分辨率(high resolution, HR)图像。相较于LR图像来说,HR图像具有更丰富的细节信息和更友好的视觉感知,图像质量和可用性都更高。SISR技术是图像复原领域的一个热点研究问题,通常应用于刑侦、遥感和医学诊断等领域。SISR技术具有非常典型的不适定(ill-posed)特性,即一个LR输入对应许多可能的HR输出,映射空间太大,映射对象不唯一,具有非常大的挑战性。

常见的超分辨率技术主要分为以下3类:基于插值的方法、基于重建的方法和基于学习的方法。基于插值的方法主要依赖于图像先验信息和统计模型等进行插值计算,具有简单易用和计算复杂度低的优点。然而插值得到的图像容易出现锯齿状的伪影,并且性能严重依赖于原始图像的质量。基于重建的方法通过融合同一场景下的多个低分辨率图像中的高频信息生成高分辨率图像。但是在实际场景中,难以收集足够多的低分辨率图像。此外,基于重建的方法对先验的依赖较强,生成的结果容易出现伪影或失真问题。基于学习的方法对大量HR和LR图像对的学习来构建高低分辨率字典,并通过深度学习网络进行图像重建。在这个过程中,算法从大量数据对中学习图像的特征和结构信息,从而实现更加准确的重建效果。相较于基于插值和基于重建的方法,基于学习的方法能够获得更多的高频信息,从而得到更清晰的纹理细节和更丰富的信息特征,在准确性和计算效率等方面具有很大的优势,因此受到越来越多的关注。

Dong 等人(2014)提出的 SRCNN(convolutional neural network for image super-resolution)采用经典的端到端结构,通过简单的3层卷积结构在保证图像质量的同时又可以实现高分辨率图像的快速重建,开创了基于学习的超分辨率技术的先河。Shi 等人(2016)提出的 ESPCN(efficient sub-pixel convolutional neural network)方法提高了基于学习的方法的计算效率,获得了良好的重建效果。Kim 等人(2016)采用深度残差网络的思想,通过构建具有20层的深度网络 VDSR(accurate image super-

resolution using very deep convolutional network),获得了比SRCNN更好的重建图像质量,证明加深网络深度有助于学习图像特征。Zhang 等人(2017)提出了深度卷积神经网络 DnCNN(denoising convolutional neural network),通过学习图像残差进行去噪。Lim 等人(2017)通过舍弃 SRResNet(super-resolution residual network)(Ledig 等,2017)中残差块的批归一化(batch normalization, BN)层并优化网络结构,使图像重建效果得到大幅提升。Zhang 等人(2018a)首次将通道注意力(channel attention, CA)机制引入到SR任务中,提出的嵌套残差结构(residual in residual, RIR)成为目前SR任务中残差块堆叠的标准范式,其所提出的 RCAN(image super-resolution using very deep residual channel attention network)网络通过通道注意力机制区分不同通道特征对重建图像的贡献程度,自适应调整通道权重,图像重建质量获得极大程度的提升。

基于 Transformer 的方法在自注意力(self-attention, SA)的建模方面具有出色的能力,因此在处理输入数据时可以更好地捕捉数据之间的关系,在自然语言处理领域大放异彩。Dosovitskiy 等人(2021)提出的 ViT(vision Transformer)首次将广泛应用于自然语言处理(natural language processing, NLP)任务中的 Transformer(Vaswani 等,2017)迁移至视觉任务中。其首先将输入图像经过一个嵌入层转换为一组一维向量,并通过多个自注意力层学习图像中的信息。SwinIR(image restoration using Swin Transformer)(Liang 等,2021)将 Swin Transformer(Liu 等,2021)引入到SR任务中,将重建图像质量提高到了新的高度。通过将输入图像分割为很多小块,然后将每个小块作为 Swin Transformer 的输入来处理图像,通过跨窗口交流保持对像素之间长距离关系的建模,显著减少了模型的计算成本,同时这也是 Transformer 首次应用于SR任务中。刘花成等人(2022)提出了一种用于模糊图像超分辨重建的 Transformer 融合网络(Transformer fusion network, TFN),采用了双分支策略,在对模糊图像进行重建的同时能有效去除部分模糊特性。邱德粉等人(2023)根据红外图像的特性并针对深度网络中存在的特征冗余问题和计算效率问题,在超分辨率模块使用通道拆分策略,提出了一种使用高分辨率可见光图像引导红外图像进行超分辨率的神经网络模

型。王美华等人(2022)使用Transformer模型计算图像深层语义信息,提出了TDATE单幅图像去雨算法。UFormer(U-shaped Transformer for image restoration)(Wang等,2022)应用 8×8 的局部窗口并引入U-Net(Ronneberge等,2015)架构捕捉全局信息。ViT将自注意力引入到视觉任务中,并在高级任务中表现优异。但在应用到图像超分辨率这类低级计算机视觉中,考虑到需要逐像素进行操作,其复杂性与输入特征图的尺寸呈平方关系,尤其是对于像素更多、分辨率更高的图像而言,需要的计算成本极高,这限制了Transformer在低级计算机视觉任务中的使用。Mei等人(2021)提出的NLSA(non-local sparse attention)通过将图像划分为不重叠的块,独立计算注意力,但这样会引入伪影并降低恢复图像的质量。SwinIR虽然也通过划分窗口进行自注意力的计算,但是其通过移位机制与其他窗口建立联系。然而SwinIR的计算量会随着窗口大小的增加而急速增长,同时局部窗口会限制上下文的聚合范围,不利于各像素之间长距离依赖关系的建模。

针对目前Transformer在SR任务中存在的一系列问题,本文提出一种全新的基于转置自注意力的超分辨率神经网络(super-resolution network based on transposed self-attention, SRTSA),主要的贡献点包括:1)提出了一个全局和局部信息提取模块(global and local information extraction block, GLEIB),该模块通过计算交叉协方差矩阵构建各像素点之间的长距离依赖关系对全局信息进行建模,同时引入通道注意力结构对局部信息进行建模。通过不同层级提取

图像信息,提高模型的学习能力。2)在GLEIB中构建一个全新的双门控深度卷积前馈网络(double gated-dconv feed-forward network, DGDFN),控制信息在模型中的流动,提高模型对特征的建模能力及其鲁棒性。3)提出的SRTSA网络通过对图像的全局信息和局部信息进行建模,使用门控网络控制信息流动以进一步改善网络的学习能力。在5个基准数据集上的测试结果验证了本文方法的先进性和有效性,在SR任务中展现出极大的优势。

1 本文方法

1.1 网络结构

本文提出的SRTSA网络结构如图1所示,主要由浅层特征提取模块、深层特征提取、上采样模块、重建模块4部分组成。定义输入图像为 $I_{LR} \in \mathbf{R}^{H \times W \times C}$,输出图像 $I_{SR} \in \mathbf{R}^{H \times W \times C}$, H, W, C 分别为图像的高度、宽度、通道数。

首先,使用一个 3×3 卷积从输入图像 I_{LR} 中提取浅层特征 F_0 ,具体为

$$F_0 = H_{SF}(I_{LR}) \quad (1)$$

式中, $H_{SF}(\cdot)$ 表示卷积核大小为 3×3 的卷积层,其可以将图像空间映射到更高的维度上。

然后,将 F_0 送入深层特征提取模块中,可得

$$F_{DF} = H_{DF}(F_0) \quad (2)$$

式中, $H_{DF}(\cdot)$ 表示深层特征提取模块,由 M 个全局和局部信息提取组(global and local information extraction groups, GLEIG)和1个卷积层堆叠而成, GLEIG

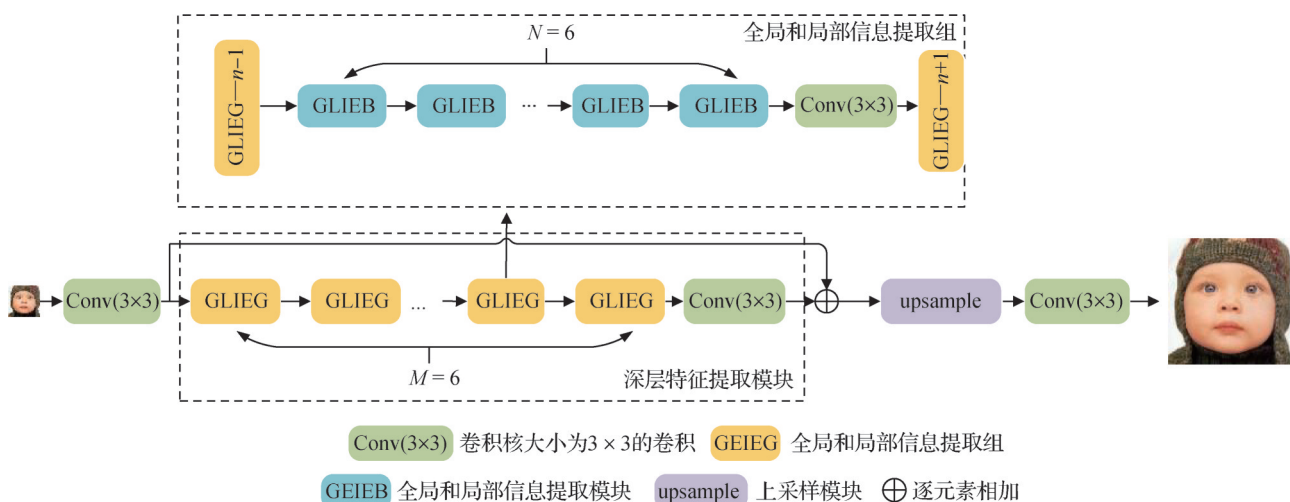


图1 SRTSA网络整体结构

Fig. 1 Overall architecture of the SRTSA network

由 N 个 GLEIB 和 1 个卷积层构成, M 和 N 的详细参数见 2.1 节。

深层特征 F_{DF} 主要负责图像的高频部分的恢复, 浅层特征 F_0 负责图像低频部分的重建, 因此需要将 F_0 与 F_{DF} 通过跳跃连接聚合再送到上采样模块中, 得到

$$F_{UP} = H_{UP}(F_0 + F_{DF}) \quad (3)$$

式中, $H_{UP}(\cdot)$ 表示由 Shi 等人 (2016) 提出的 ESPCN 一文中的亚像素卷积。它通过像素平移按给定上采样系数的比例采样。

经过上采样模块后将得到相应尺寸的特征图, 最后, 再将上采样特征 F_{UP} 送入重建模块中, 得到

$$I_{SR} = H_{REC}(F_{UP}) = H_{SRTSA}(I_{LR}) \quad (4)$$

式中, $H_{REC}(\cdot)$ 表示卷积核大小为 3×3 的卷积层, 用于将特征图转换为 RGB 空间的 SR 图像。 $H_{SRTSA}(\cdot)$ 表示整个 SRTSA 模型映射。

1.2 全局和局部信息提取模块

全局和局部信息提取模块 (GLEIB) 结构如图 2 所示, 该模块主要由 3 部分构成:

1) 全局信息建模模块。主要通过转置自注意力机制计算交叉协方差矩阵进行图像全局关系建模。考虑到如果只在通道维度建立长距离关系相关的话

会损失一些有用的空间纹理和结构信息, 这些信息对于重建高质量图像十分重要。因此在采用转置自注意力机制前使用一种轻量高效的特征提取块进行简单关系建模保存部分空间结构信息。

2) 局部信息建模模块。自然图像通常包含全局和局部信息结构, 在对全局信息进行建模的同时也需要注意到图像部分特性具有局部性, 因此使用通道注意力块补充局部信息, 增强模型的学习能力。

3) 双门控深度卷积前馈网络。DGDFN 通过在两个并行路径中采用门控机制抑制信息量较少的特征, 并将筛选后的信息进行融合, 进一步增强了模型对输入特征的建模能力, 提高模型的性能。

1.2.1 全局信息建模模块

Transformer 中的计算资源主要消耗于自身的自注意力层中。对于一幅 $H \times W$ 像素的图像来说, 其需要计算任意两个像素之间的相似度, 计算复杂度为 $O(W^2H^2)$, 将其应用于 SR 任务中显然是十分困难的。SwinIR 采用的 Swin Transformer 虽然通过划分窗口减轻了计算成本, 但是划分窗口限制了聚合上下文的范围, 不能有效连续提取特征。为了缓解这个问题, 采用 Zamir 等人 (2022) 提出的转置自注意力机制隐式构建全局注意力图。

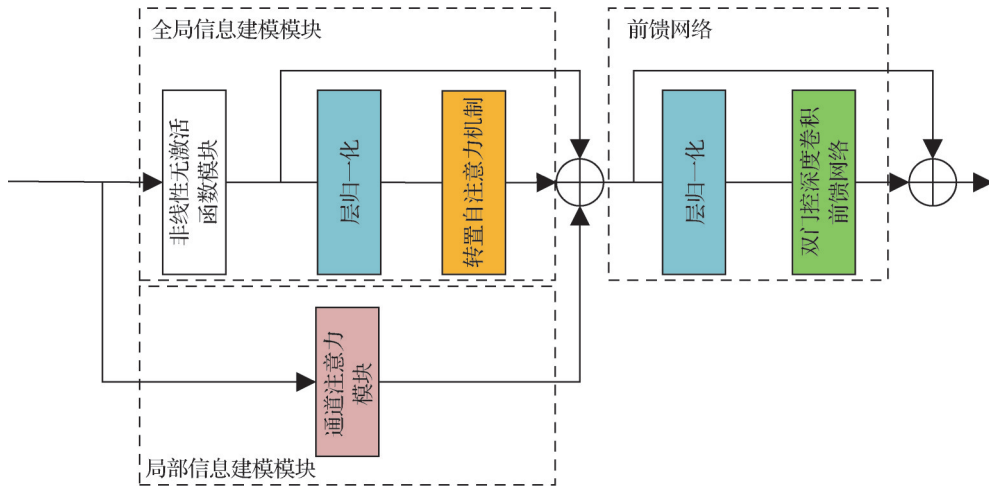


图2 全局和局部信息提取模块结构

Fig. 2 Architecture of the global and local information extraction block (GLEIB)

转置自注意力机制的结构如图 3 所示, 输入由 NAFBlock 输出的张量 $X \in \mathbf{R}^{H \times W \times C}$ 经过一个层归一化 (layer normalization, LN) 之后生成张量 $Y \in \mathbf{R}^{H \times W \times C}$ 。接着将 Y 输入到卷积核大小为 1×1

的卷积中聚合跨通道信息, 然后使用 3×3 大小的深度卷积生成 Q, K, V , 具体为

$$Q = W_d^Q W_p^Q Y \quad (5)$$

$$K = W_d^K W_p^K Y \quad (6)$$

$$V = W_d^V W_p^V Y \quad (7)$$

式中, $W_p(\cdot)$ 表示卷积核大小为 1×1 的卷积, $W_d(\cdot)$ 表示 3×3 的深度卷积。

接着通过 Reshape 操作重塑 Q 和 K , 使其生成大小为 $\mathbf{R}^{C \times C}$ 的转置注意力图, 而不是大小为 $\mathbf{R}^{HW \times HW}$ 的常规注意力图。整个过程可以表示为

$$X_{out} = W_p f_{Attention}(Q, K, V) + X_{in} \quad (8)$$

式中, $W_p(\cdot)$ 表示 1×1 大小的卷积, $f_{Attention}$ 为转置自注意力机制。

Li 等人 (2023) 提出经过 softmax 归一化生成的自注意力会影响特征聚合。根据 ReLU (rectified linear unit) 激活函数自身的特点, 可以在保留对图像重建呈积极效果的特征的同时去除负面特征, 这样极大地提升了网络模型重建图像的性能。因此本文搭

建的转置自注意力机制可以表示为

$$f_{Attention}(Q, K, V) = V \cdot \text{ReLU}\left(\frac{K \cdot Q^T}{\sqrt{d_k}}\right) \quad (9)$$

式中, $\sqrt{d_k}$ 是一个可学习的缩放参数, $\text{ReLU}(\cdot)$ 表示 ReLU 激活函数。

转置自注意力机制通过在通道维度对全局注意力图进行隐式建模很好地解决了计算复杂度与特征图空间维度呈二次方的关系, 但是考虑到如果只在通道维度建立长距离依赖关系的话会损失一些有用的空间纹理和结构信息, 这些信息对于重建高质量图像十分重要。因此在使用转置自注意力机制前需要对特征进行简单关系建模保存部分空间结构信息。

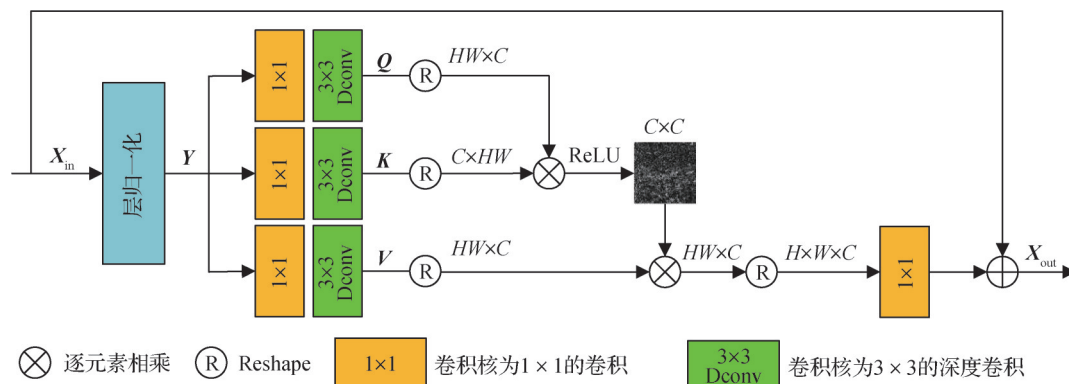


图3 转置自注意力机制结构图

Fig. 3 Architecture of transposed self-attention

Chen 等人 (2022) 通过揭示 GELU (Gaussian error linear unit)、GLU (gated linear unit) 和通道注意力之间的联系, 通过移除或替换非线性激活函数 (例如 sigmoid, ReLU, GELU), 提出了一个适用于图像复原领域的简单基线模型非线性无激活函数模块 (nonlinear activation free block, NAFBlock)。该模型通过将 Transformer 内较为成熟的模块和机制应用于卷积神经网络中, 提升了卷积神经网络的学习能力。NAFBlock 的网络结构如图 4 所示, 其主要由两个部分组成: 移动卷积 (mobile convolution, MBConv) 模块和具有两个全连接层的前馈网络 (feedforward network, FFN)。在 MBConv 和 FFN 之前都加入了层归一化 (layer norm, LN) 用于标准化输入数据, 提高模型的泛化性能。同时对两个模块使用了残差连接, 加速网络训练和收敛。整个过程可以表示为

$$X_{mid} = \text{MBConv}(\text{LN}(X_{in})) + X_{in} \quad (10)$$

$$X_{out} = \text{FFN}(\text{LN}(X_{mid})) + X_{mid} \quad (11)$$

NAFBlock 通过引入 SimpleGate 单元来替代非线性激活函数, 给定输入 $X \in \mathbf{R}^{H \times W \times C}$, 将特征沿通道维度分割成两个特征 X_1 和 X_2 , $X_1, X_2 \in \mathbf{R}^{H \times W \times C/2}$, 然后使用线性门控来计算输出, 具体计算为

$$f_{\text{SimpleGate}}(X) = X_1 \odot X_2 \quad (12)$$

式中, $\odot$ 表示逐元素相乘。由于简化的 SimpleGate 引入了非线性, 所以常用的激活函数就不需要再加入到网络中。

需要注意的是, NAFBlock 分别在 MBConv 和 FFN 中采用正则化 (dropout) 技术随机丢弃部分神经元以达到减少模型过拟合、增强模型泛化性能的作用。但是 GLEIB 将经过 NAFBlock 进行简单关系建模后的特征送入转置自注意力机制中进行全局信息

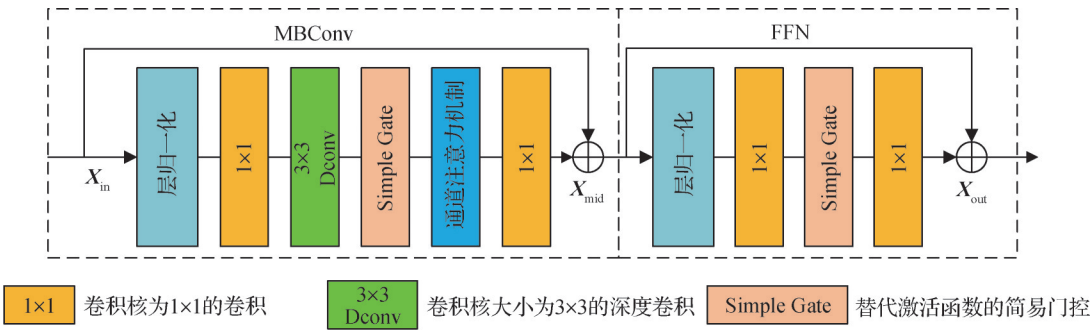


图4 非线性无激活函数模块结构图
Fig. 4 Architecture of nonlinear activation free block

建模,为防止全局信息建模时特征信息不全,本文不使用NAFBlock中的dropout层以使得转置自注意力机制对所有特征进行全局关系建模,转而将舍弃部分内容增强模型泛化性能的任务交给转置自注意力机制中的ReLU激活函数。

1.2.2 局部信息建模模块

自然图像包含全局信息和局部信息,通常使用自注意力机制对图像的全局依赖关系进行建模,但是只使用自注意力机制对图像信息进行建模会丢失图像特有的局部相关性。图像局部信息只涵盖了几个像素,典型局部特征表现为边缘、局部色彩等,因此使用通道注意力块(channel attention block, CAB)

对图像局部信息进行补充。

CAB的结构如图5所示,由两个标准的卷积层组成,两者之间使用GELU激活函数,并且还包含一个通道注意力。CAB通过压缩和扩展通道特征,以及通过通道注意力模块的自适应调整,实现了对输入特征的处理和提取。这个过程有助于提取重要的特征信息,并为后续模型提供更有代表性的特征表示。整个过程可以表示为

$$X_{out} = CA(X_{mid}) \cdot X_{mid} \tag{13}$$

$$X_{mid} = W^1(\sigma(W^2(X_{in}))) \tag{14}$$

式中, $\sigma(\cdot)$ 表示GELU激活函数, $W(\cdot)$ 表示 3×3 大小的卷积, $CA(\cdot)$ 表示通道注意力机制。

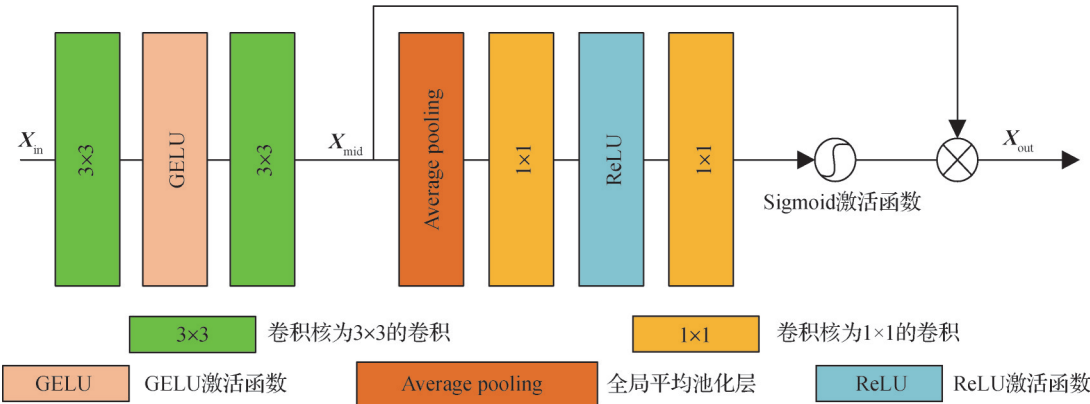


图5 通道注意力块结构图
Fig. 5 Architecture of channel attention block

1.2.3 双门控深度卷积前馈网络

在传统Transformer中,FFN通常包含两个线性变换层和一个非线性激活函数。第1个线性变换层用于扩展特征通道,第2个用于将通道减回原始输入维度。为了更好地捕捉输入图像的特征信息,本文在Restormer(Zamir等,2022)的基础上提出一个

全新的双门控深度卷积前馈网络(DGDFN),其结构如图6所示。

DGDFN引入深度卷积对空间相邻像素的信息进行编码,丰富局部信息。在每条路径上都通过GELU激活函数以实现双门控机制。在两个并行路径中将通过门控机制处理后的特征进行逐元素相乘

操作,这样可以过滤掉两个路径中相应位置上特征向量中的较小值,从而减少信息量较少的特征的影响。之后逐元素求和操作将对应位置的特征值相加,以合并两个路径中有用的信息,同时由于双门控机制具有更强的非线性建模能力,因此可以学习更复杂的模式,提高模型的泛化性和建模能力。整个过程可以表示为

$$X_{out} = W_p^0 Gating(X_{in}) + X_{in} \quad (15)$$

$$Gating(X_{in}) = \sigma(W_d^1 W_p^1(LN(X_{in}))) \odot W_d^2 W_p^2(LN(X_{in})) + \sigma(W_d^2 W_p^2(LN(X_{in}))) \odot W_d^1 W_p^1(LN(X_{in})) \quad (16)$$

式中, $\sigma(\cdot)$ 表示 GELU 激活函数,即门控机制, $W_p(\cdot)$ 表示卷积核大小为 1×1 的卷积, $W_d(\cdot)$ 表示 3×3 大小的深度卷积, $\odot$ 表示逐元素相乘。

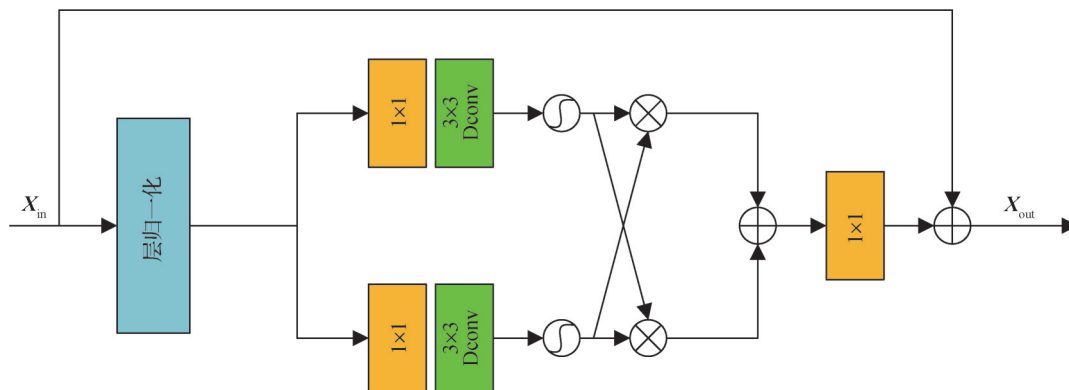


图6 双门控深度卷积前馈网络结构图

Fig. 6 Architecture of double gated-dconv feed-forward network

1.3 损失函数

虽然 L_1 损失函数、 L_2 损失函数、感知损失以及对抗损失等多种损失函数的组合可以提高重建图像的视觉效果,但为了验证 SRTSA 网络的有效性,本文采用了与上述相关工作相同的损失函数—— L_1 损失函数进行优化。相较于其他损失函数, L_1 损失函数提供了稳定的梯度,并有助于模型快速收敛,使用 L_1 损失函数进行优化是验证 SR 网络性能的可靠方式。给定包含 N 个 LR-HR 图像对的训练集 $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$, 通过最小化 L_1 损失优化 SRTSA 的模型参数

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \|H_{SRTSA}(I_{LR}^i) - I_{HR}^i\|_1 \quad (17)$$

式中, θ 表示 SRTSA 网络模型的参数。

2 实验结果及分析

2.1 数据集及训练细节

本文使用 DIV2K 数据集 (Agustsson 和 Timofte, 2017) 的 800 幅训练图像对 SRTSA 网络进行训练,采用双三次插值法 (Bicubic) 得到 $\times 2$ 、 $\times 3$ 、 $\times 4$ 共 3 个比例因子的低分辨率图像。为了增加训练图像的数据多样性,将这 800 幅训练图像水平翻转或者随机旋转 90° 、 180° 、 270° 。在每次迭代中将 16 个 48×48 像

素的低分辨率图像送入网络 (batch size = 16), 使用 Adam 优化器进行训练,其中, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, 初始学习率 $l_r = 5 \times 10^{-4}$, 总共迭代 500 000 次,迭代到 200 000 次时学习率减半。

使用具有不同特性的 5 个标准测试集验证网络的有效性: Set5 (Bevilacqua 等, 2012)、Set14 (Zeyde 等, 2012)、BSD100 (Berkeley segmentation dataset 100) (Martin 等, 2002)、Urban100 (Huang 等, 2015)、Manga109 (Matsui 等, 2017), 并在输出图像的 YCbCr 空间 Y 通道上计算峰值信噪比 (peak signal-to-noise ratio, PSNR) 和结构相似度 (structural similarity, SSIM) 指标评价 SR 结果。使用 PyTorch 框架在两张 NVIDIA GeForce RTX 3090 显卡上训练模型。在整个网络中,为保持与 SwinIR 模型进行公平对比,设置 GLEIG 数 $M = 6$, GLEIB 数为 $N = 6$, CAB 输出缩放权重 $\alpha = 0.05$, 转置自注意力特征通道设置为 120, 多头数设置为 6。

2.2 实验结果分析

为了展示 SRTSA 模型的先进性,本文与 Bicubic 算法、SRCNN、VDSR、EDSR (enhanced deep residual network for single image super-resolution)、RCAN、SAN (second-order attention network for single image

super-resolution) (Dai 等, 2019)、HAN (single image super-resolution via a holistic attention network) (Niu 等, 2020)、NLSA、SwinIR 等模型在×2、×3、×4 共 3 个比例因子上对 PSNR 指标和 SSIM 指标进行对比。较高的 PSNR 值表示重建效果更好, SSIM 越接近 1 表示 SR 图像与 HR 图像在结构上更为相似。

Chu 等人(2022)指出在图像复原工作中由于计算资源限制,一般都将要修复的图像裁剪成小块(Patch)送入到模型中进行训练。而在推理过程,一般则直接将需要修复的图像送入到网络中进行复原。这种模式存在着训练与推理过程的不一致性,对利用全局信息的模型会产生负面影响。因此在测试过程中采用 Chu 等人(2022)提出的测试时间局部转换器(test-time local converter, TLC)维持训练与推理过程的一致性。

按照 2.1 节的实验设置对模型进行完整训练,得到×2、×3、×4 共 3 个不同比例因子的 PSNR 和 SSIM,分别展示在表 1—表 3 中。可以看到,在×4 倍

SR 任务中, SRTSA 较 SwinIR 在 Set5、Set14、BSD100 和 Urban100 上, PSNR 分别提升 0.07 dB、0.01 dB、0.03 dB 和 0.08 dB, SSIM 也有较大提升。在×3 倍 SR 任务中, SRTSA 较 SwinIR 在 Set5、BSD100、Urban100 和 Manga109 上, PSNR 分别提升 0.02 dB、0.04 dB、0.15 dB 和 0.02 dB, SSIM 在 5 个测试集上都获得最佳指标。在×2 倍 SR 任务中, SRTSA 较 SwinIR 在 Set5、Set14、BSD100、Urban100 和 Manga109 上, PSNR 分别提升 0.03 dB、0.21 dB、0.05 dB、0.29 dB 和 0.10 dB, SSIM 在除 Manga109 以外的 4 个测试集上, 分别提升 0.000 4、0.001 6、0.000 9 和 0.002 7,足以说明 SRTSA 的优势十分明显。

SRTSA 网络在 BSD100 和 Urban100 这两个包含丰富的结构和纹理信息且具有更高复杂度的测试集上的优势更为显著。说明 SRTSA 网络可以在复杂场景中保持高水平的细节保留和纹理重建能力,具有较强的结构感知能力和泛化性能,可以适应不同类型的图像数据并提供高质量的超分辨率重建图像。

表 1 各模型在不同数据集上的表现 (×2)
Table 1 Performance of each model on different datasets (scale×2)

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	33.66	0.929 9	30.24	0.868 8	29.56	0.843 1	26.88	0.840 3	30.80	0.933 9
SRCNN	36.66	0.954 2	32.45	0.906 7	31.36	0.887 9	29.50	0.894 6	35.60	0.966 3
VDSR	37.05	0.956 0	33.05	0.913 0	31.90	0.896 0	30.77	0.914 0	37.22	0.975 0
EDSR	38.11	0.960 2	33.92	0.919 5	32.32	0.901 3	32.93	0.935 1	39.10	0.977 3
RCAN	38.27	0.961 4	34.11	0.921 6	32.41	0.902 6	33.34	0.938 4	39.43	0.978 6
SAN	38.31	0.962 0	34.07	0.921 3	32.42	0.902 8	33.10	0.937 0	39.32	0.979 2
HAN	38.27	0.961 4	34.16	0.921 7	32.41	0.902 7	33.35	0.938 5	39.46	0.978 5
NLSA	38.34	0.961 8	34.08	<u>0.923 1</u>	32.43	0.902 7	33.42	<u>0.939 4</u>	39.59	<u>0.978 9</u>
SwinIR	<u>38.35</u>	<u>0.962 0</u>	<u>34.14</u>	0.922 7	<u>32.44</u>	<u>0.903 0</u>	<u>33.40</u>	0.939 3	<u>39.60</u>	0.979 2
SRTSA(本文)	38.38	0.962 4	34.35	0.924 3	32.49	0.903 9	33.69	0.942 0	39.70	0.978 8

注:加粗字体和下划线字体分别表示每列最优结果和次优结果。

图 7 展示了 Urban100 数据集 044_img. png 图像通过不同算法模型重建得到的图像。可以看到, Bicubic 算法所重建图像十分模糊, VDSR、RDN (Zhang 等, 2018b)等网络只能重建出图像较为模糊的轮廓且细节信息基本丢失, RCAN、SAN、SwinIR 可以重建出较为清晰的轮廓,但对于纹理结构的重建

并不完善, SRTSA 网络所重建的图像不但轮廓清晰, 结构合理, 并且对于细节纹理的恢复也最接近 HR 图像。

图 8 展示了 Urban100 数据集 092_img. png 图像经过不同模型重建后的效果对比。Bicubic 插值法得到的结果呈现出非常模糊的特征, 几乎无法重建

表 2 各模型在不同数据集上的表现 (×3)
Table 2 Performance of each model on different datasets (scale×3)

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	30.39	0.868 2	27.55	0.774 2	27.21	0.738 5	24.46	0.734 9	26.95	0.855 6
SRCNN	32.75	0.909 0	29.30	0.821 5	28.41	0.786 3	26.24	0.798 9	30.48	0.911 7
VDSR	33.67	0.921 0	29.78	0.832 0	28.83	0.799 0	27.14	0.829 0	32.01	0.934 0
EDSR	34.65	0.928 0	30.52	0.846 2	29.25	0.809 3	28.80	0.865 3	34.17	0.947 6
RCAN	34.74	0.929 9	30.64	0.848 1	29.32	0.811 1	29.08	0.870 2	34.43	0.949 8
SAN	34.75	0.930 0	30.59	0.847 6	29.33	0.811 2	28.93	0.867 1	34.30	0.949 4
HAN	34.75	0.929 9	30.67	0.848 3	29.32	0.811 0	29.10	0.870 5	34.48	0.950 0
NLSA	34.85	<u>0.930 6</u>	30.70	0.848 5	29.34	0.811 7	29.25	0.872 6	34.57	0.950 8
SwinIR	<u>34.89</u>	0.931 2	30.77	<u>0.850 3</u>	<u>29.37</u>	<u>0.812 4</u>	<u>29.29</u>	<u>0.874 4</u>	<u>34.74</u>	<u>0.951 8</u>
SRTSA (本文)	34.91	0.931 2	<u>30.75</u>	0.850 4	29.41	0.813 6	29.44	0.876 3	34.76	0.951 9

注:加粗字体和下划线字体分别表示每列最优结果和次优结果。

表 3 各模型在不同数据集上的表现 (×4)
Table 3 Performance of each model on different datasets (scale×4)

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic	28.42	0.810 4	26.00	0.702 7	25.96	0.667 5	23.14	0.657 7	24.89	0.786 6
SRCNN	30.48	0.862 8	27.50	0.751 3	26.90	0.710 1	24.52	0.722 1	27.58	0.855 5
VDSR	31.35	0.883 0	28.02	0.768 0	27.29	0.725 1	25.18	0.754 0	28.83	0.887 0
EDSR	32.46	0.896 8	28.80	0.787 6	27.71	0.742 0	26.64	0.803 3	31.02	0.914 8
RCAN	32.63	0.900 2	28.87	0.788 9	27.77	0.743 6	26.82	0.808 7	31.22	0.917 3
SAN	32.64	<u>0.900 3</u>	28.92	0.788 8	27.78	0.743 6	26.79	0.806 8	31.18	0.916 9
HAN	32.64	0.900 2	28.90	0.789 0	27.80	0.744 2	26.85	0.809 4	31.42	0.917 7
NLSA	32.59	0.900 0	28.87	0.789 1	27.78	0.744 4	26.96	0.810 9	31.27	0.918 4
SwinIR	<u>32.72</u>	0.902 1	<u>28.94</u>	<u>0.791 4</u>	<u>27.83</u>	<u>0.745 9</u>	<u>27.07</u>	<u>0.816 4</u>	31.67	0.922 6
SRTSA (本文)	32.79	0.902 1	28.95	0.791 6	27.86	0.747 2	27.15	0.817 1	<u>31.63</u>	<u>0.922 2</u>

注:加粗字体和下划线字体分别表示每列最优结果和次优结果。

出图像的细节。基于深度学习的其他方法在重建过程中虽然保留了主要的轮廓信息,但在纹理结构和细节方面却存在明显的扭曲和伪影。而使用SRTSA网络进行重建的图像则具有纹理清晰、边缘锐利以及细节丰富等优点,通过恢复更多的信息获得更好的结果。

图 8 中,不同方法重建结果差异性较大是因为 SR 任务是一个典型的不适定问题,LR 图像无法唯

一确定对应的 HR 图像。VDSR、EDSR 等 SR 模型受感受野较小、模型学习能力较弱等限制,无法充分考虑图像的全局结构恢复细节信息,在生成 SR 图像时都出现了将竖状条纹重建为斜状条纹、交叉条纹或者出现伪影等错误范式。相比之下,SRTSA通过对全局信息进行建模,获得更多纹理结构信息,极大地缓解了不适定问题,重建得到的 SR 图像也与 HR 图像更为相似。

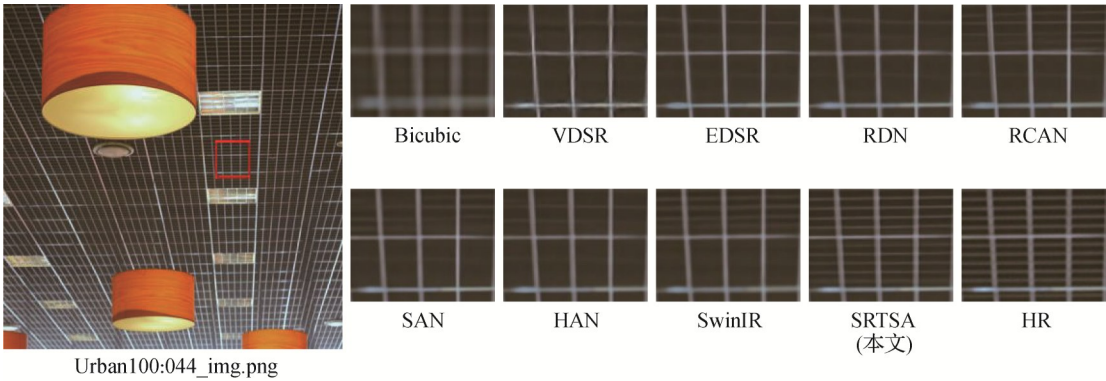


图7 Urban100数据集中044_img. png的测试结果(×4)
Fig. 7 Test results of 044_img. png in Urban100 dataset (scale×4)

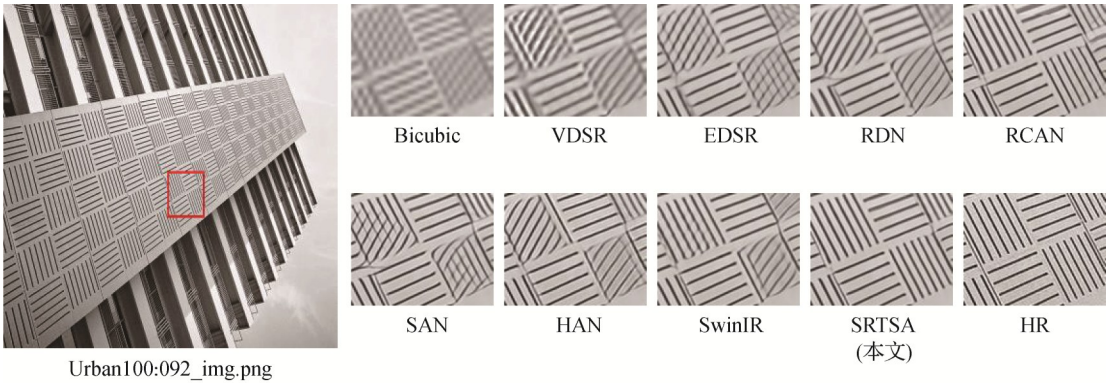


图8 Urban100数据集中092_img. png的测试结果(×4)
Fig. 8 Test results of 092_img. png in Urban100 dataset (scale×4)

2.3 局部归因分析

局部归因图(local attribution maps, LAM)是由Gu和Dong(2021)针对SR领域具有不可解释性而提出的归因分析方法,通过局部归因图的形式将对输出结果产生强烈影响的像素重点标注出来。通过这种方式,可以很直观地看到SR网络的有效性。为了直观地展示SRTSA的优势,本文将SRTSA和SwinIR同时使用LAM工具进行比较,比较结果如图9所示。在LAM中,作者提出扩散指数(diffusion index, DI)衡量LAM的结果,表示形式为

$$DI = (1 - GI) \times 100 \tag{18}$$

式中,GI表示基尼系数(Gini index, GI)。GI是国际上通用衡量一个国家或者地区收入差距的指标。在LAM中,GI是衡量LAM差距的指标,它可以反映LAM所涉及的像素范围。简单理解就是对于重建图像的某一部分来说,如果重建该部分只涉及少数像素点,则GI系数相对较高。因此DI和GI呈反比例关系,DI越大说明重建某一部分网络考虑到更多的像素点。

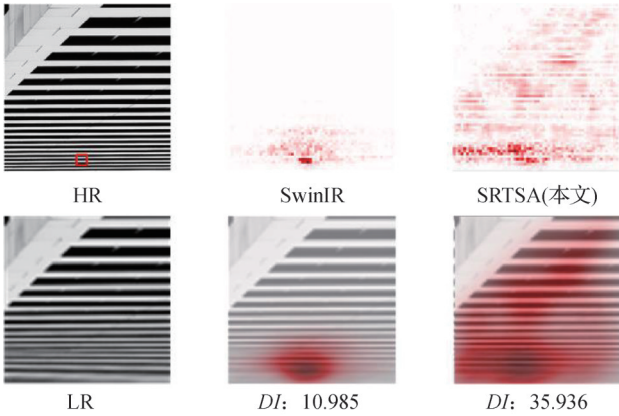


图9 不同模型的LAM结果
Fig. 9 LAM results for different models

从图9可以直观地看到,SRTSA生成SR图像时考虑到的像素范围远大于SwinIR所考虑的像素范围。实验结果和客观指标评价、主观视觉感知高度一致,从可解释性的角度证明了SRTSA的优越性。

2.4 消融实验

为了验证SRTSA网络各组成部分的有效性,本文对不同结构进行消融实验,并对实验结果进行对

比分析。1) 实验中把只有转置自注意力机制和门控前馈网络(GDFN)的模型作为 Baseline; 2) 在 Baseline 上引入 NAFBlock 以验证其对特征进行的简单关系建模是否可以补充转置自注意力机制损失的部分空间纹理和结构信息; 3) 在 Baseline 上引入 CAB 验证补充局部信息是否可以增强图像重建效果; 4) 同时引入 NAFBlock 和 CAB 验证对特征进行简单关系建模是否与对特征进行局部关系建模产生冲突; 5) 由于 DGDFN 相对独立, 在实验 4) 的基础上引入 DGDFN 即可验证其是否可以提高模型的建模能力。在消融实验中设置 batch size = 8, 其余均按照 2.1 节的设置重新训练网络。

在 Set5 测试集上的 PSNR 如表 4 所示。由表 4 中实验 2) 可知, 在 Baseline 模型的基础上引入 NAFBlock 模块之后 PSNR/SSIM 获得了 0.01 dB/0.000 3 的提升, 说明对特征进行简单关系建模可以保存部分空间结构信息。实验 3) 结果表明, 引入 CAB 模块后, PSNR 和 SSIM 分别提升了 0.02 dB 和 0.000 3, 说明 CAB 通过捕捉图像特有的局部相关性从而增强了网络的建模能力。在实验 4) 中, 同时引入以上两

个模块后, PSNR 和 SSIM 分别提升了 0.05 dB 和 0.000 7。由实验 5) 可知, 本文 DGDFN 的 PSNR 和 SSIM 分别提升了 0.03 dB 和 0.000 9, 说明 DGDFN 可以提高模型的泛化性能和建模能力, 同样也说明在 Transformer 模型中, 自注意力机制和前馈网络均发挥着不可或缺的作用。它们共同构成了模型的核心组件, 并相互协作以实现高效的特征提取和建模。

同时, 为了更直观地展示各个模块在 SRTSA 模型中所做出的贡献, 采用 LAM 工具分析模型的感受野, 过程性结果如图 10 所示。

表 4 在 Set5 (×4) 测试上的消融实验

Table 4 Ablation study on Set5 (scale×4)

实验	NAFBlock	CAB	DGDFN	PSNR/dB	SSIM
1)	×	×	×	32.67	0.901 1
2)	√	×	×	32.68	0.901 4
3)	×	√	×	32.69	0.901 4
4)	√	√	×	32.72	0.901 8
5)	√	√	√	32.75	0.902 0

注: “√”表示采用, “×”表示未采用。

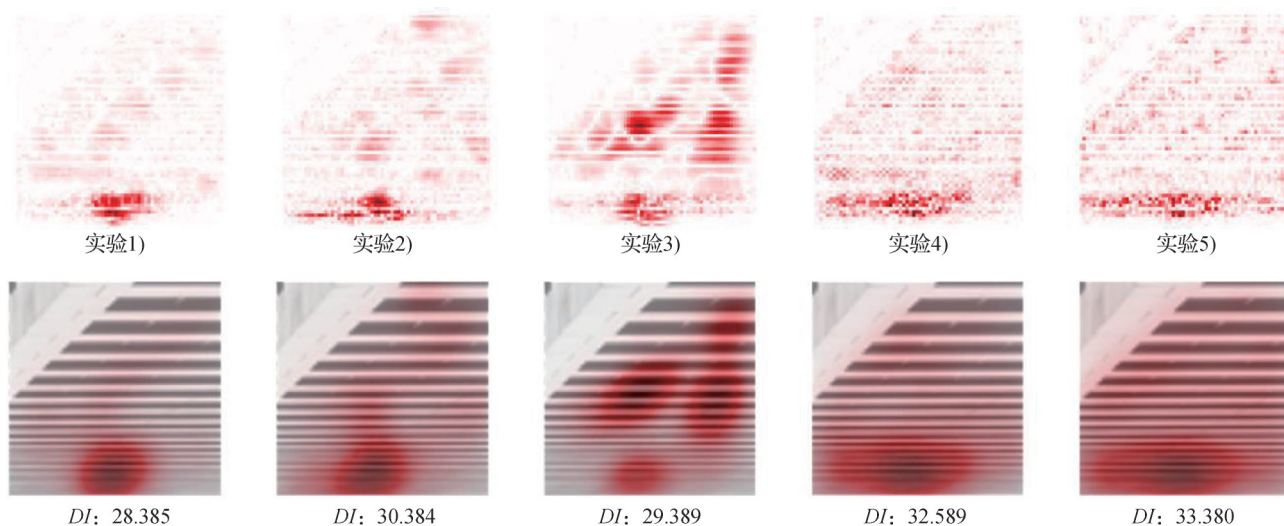


图 10 过程性结果展示

Fig. 10 Process results display

3 结 论

为了解决主流方法通过在窗口内使用自注意力机制, 会限制特征聚合范围而导致图像重建效果不佳的问题, 提出了基于转置自注意力机制的超分辨

率网络(SRTSA)。通过转置自注意力机制和通道注意力模块分别对全局特征和局部特征进行建模, 同时提出一个全新的双门控深度卷积前馈网络(DGDFN)提高模型的泛化性能。大量实验结果表明, 本文方法在客观指标、主观视觉和模型可解释性方面均优于目前主流的单图像超分辨率算法。然

而,本文算法也存在较多不足:1)用于局部信息增强的通道注意力块在大通道数时会使用较大参数量,使得模型推理速度变慢,后期将构建一个全新的轻量级局部信息增强模块对局部特征信息进行补充。2)尽管本文方法通过引入转置自注意力机制隐式构建全局上下文关系节省了较多的参数,但是全局信息建模网络结构较为冗余,使得网络复杂度更高。在未来的工作中,将设计一个更为简洁的自注意力机制用于图像全局信息的提取。

参考文献 (References)

- Agustsson E and Timofte R. 2017. NTIRE 2017 challenge on single image super-resolution: dataset and study//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA; IEEE: 1122-1131 [DOI: 10.1109/CVPRW.2017.150]
- Bevilacqua M, Roumy A, Guillemot C and Morel M L A. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding//Proceedings of the 23rd British Machine Vision Conference. Surrey, UK; BMVA Press: 1-12 [DOI: 10.5244/c.26.135]
- Chen L Y, Chu X J, Zhang X Y and Sun J. 2022. Simple baselines for image restoration//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel; Springer: 17-33 [DOI: 10.1007/978-3-031-20071-7_2]
- Chu X J, Chen L Y, Chen C P and Lu X. 2022. Improving image restoration by revisiting global information aggregation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel; Springer: 53-71 [DOI: 10.1007/978-3-031-20071-7_4]
- Dai T, Cai J R, Zhang Y B, Xia S T and Zhang L. 2019. Second-order attention network for single image super-resolution//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 11057-11066 [DOI: 10.1109/CVPR.2019.01132]
- 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; Springer: 184-199 [DOI: 10.1007/978-3-319-10593-2_13]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2021. An image is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. [2023-05-10]. <https://arxiv.org/pdf/2010.11929.pdf>
- Gu J J and Dong C. 2021. Interpreting super-resolution networks with local attribution maps//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 9195-9204 [DOI: 10.1109/CVPR46437.2021.00908]
- Huang J B, Singh A and Ahuja N. 2015. Single image super-resolution from transformed self-exemplars//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA; IEEE: 5197-5206 [DOI: 10.1109/CVPR.2015.7299156]
- Kim J, Lee J K and Lee K M. 2016. Accurate image super-resolution using very deep convolutional networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 1646-1654 [DOI: 10.1109/CVPR.2016.182]
- Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z H and Shi W Z. 2017. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 105-114 [DOI: 10.1109/CVPR.2017.19]
- Li X, Pan J S, Tang J H and Dong J X. 2023. DLGSANet: lightweight dynamic local and global self-attention networks for image super-resolution [EB/OL]. [2023-05-10]. <https://arxiv.org/pdf/2301.02031.pdf>
- 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 H C, Ren W Q, Wang R and Cao X C. 2022. A super-resolution Transformer fusion network for single blurred image. *Journal of Image and Graphics*, 27(5): 1616-1631 (刘花成, 任文琦, 王蕊, 操晓春. 2022. 用于单幅模糊图像超分辨率的Transformer融合网络. *中国图象图形学报*, 27(5): 1616-1631) [DOI: 10.11834/jig.210847]
- 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 (ICCV). Montreal, Canada: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Martin D, Fowlkes C, Tal D and Malik J. 2002. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics//Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada; IEEE: 416-423 [DOI: 10.1109/iccv.2001.937655]
- Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T and Aizawa K. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20): 21811-

- 21838 [DOI: 10.1007/s11042-016-4020-z]
- Mei Y Q, Fan Y C and Zhou Y Q. 2021. Image super-resolution with non-local sparse attention//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3516-3525 [DOI: 10.1109/CVPR46437.2021.00352]
- Niu B, Wen W L, Ren W Q, Zhang X D, Yang L P, Wang S Z, Zhang K H, Cao X C and Shen H F. 2020. Single image super-resolution via a holistic attention network//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 191-207 [DOI: 10.1007/978-3-030-58610-2_12]
- 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]
- Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Shi W Z, Caballero J, Huszár F, Totz J, Aitken A P, Bishop R, Rueckert D and Wang Z H. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1874-1883 [DOI: 10.1109/CVPR.2016.207]
- 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 M H, Ke F H, Liang Y, Fan Z and Liao L. 2022. 3D attention and Transformer based single image deraining network. *Journal of Image and Graphics*, 27(5): 1509-1521 (王美华, 柯凡晖, 梁云, 范衡, 廖磊. 2022. 融合3D注意力和Transformer的图像去雨网络. *中国图象图形学报*, 27(5): 1509-1521) [DOI: 10.11834/jig.210794]
- Wang Z D, Cun X D, Bao J M, Zhou W G, Liu G Z and Li H Q. 2022. Uformer: a general U-shaped Transformer for image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 17662-17672 [DOI: 10.1109/CVPR52688.2022.01716.]
- Zamir S W, Arora A, Khan S, Hayat M, Khan F S and Yang M H. 2022. Restormer: efficient Transformer for high-resolution image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5718-5729 [DOI: 10.1109/CVPR52688.2022.00564]
- Zeyde R, Elad M and Protter M. 2012. On single image scale-up using sparse-representations//Proceedings of the 7th International Conference on Curves and Surfaces. Avignon, France: Springer: 711-730 [DOI: 10.1007/978-3-642-27413-8_47]
- Zhang K, Zuo W M, Chen Y J, Meng D Y and Zhang L. 2017. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7): 3142-3155 [DOI: 10.1109/TIP.2017.2662206]
- Zhang Y L, Li K P, Li K, Wang L C, Zhong B N and Fu Y. 2018a. Image super-resolution using very deep residual channel attention networks//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 294-310 [DOI: 10.1007/978-3-030-01234-2_18]
- Zhang Y L, Tian Y P, Kong Y, Zhong B E and Fu Y. 2018b. Residual dense network for image super-resolution//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2472-2481 [DOI: 10.1109/CVPR.2018.00262]

作者简介

孙阳,男,硕士研究生,主要研究方向为图像超分辨率重建。
E-mail:834976314@qq.com

丁建伟,通信作者,男,副教授,主要研究方向为计算机视觉和人工智能安全。E-mail:jwding@ppsuc.edu.cn

张琪,女,讲师,主要研究方向为生物特征识别和计算机视觉。E-mail:qi.zhang@ppsuc.edu.cn

邓琪瑶,女,讲师,主要研究方向为图像生成与编辑。
E-mail:20236991@ppsuc.edu.cn