计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009723] [CSTR: 32024.14.csa.009723]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于特征层次递进融合的轻量级图像超分辨率网络①

张　豪, 马　冀, 袁　江

(华北电力大学 控制与计算机工程学院, 北京 102206)
通信作者: 张　豪, E-mail: zhangh@ncepu.edu.cn

摘　要: 近年来, 随着深度学习技术的发展, 卷积神经网络 (convolutional neural network, CNN) 和 Transformer 在图像超分辨率 (super-resolution, SR) 领域取得了显著的进展. 但是, 对于图像全局特征的提取, 过去的方法大多采用的是堆叠单个算子重复计算来逐步扩大感受野的方式. 为了更好地利用全局信息, 提出了对局部、区域和全局特征进行显式建模. 具体来说, 通过通道注意增强卷积、基于划分窗口的 Transformer 和 CNN 的双分支并行架构、标准的 Transformer 和划分窗口的 Transformer 双分支并行架构, 以一种层次递进的方式对图像的局部信息、区域与局部信息、全局与区域信息进行提取和融合. 此外, 设计了一种层次特征融合方式来对 CNN 分支提取到的局部信息和划分窗口的 Transformer 提取到的区域信息进行特征融合. 大量的实验表明, 所提网络在轻量级 SR 领域实现了更好的结果. 例如, 在 Manga109 数据集的 4 倍放大实验中, 该网络的峰值信噪比 (PSNR) 相较于 SwinIR 提升了 0. 51 dB.

关键词: 图像超分辨率; Transformer; 卷积神经网络; 层次特征融合; 全局特征提取

引用格式: 张豪,马冀,袁江.基于特征层次递进融合的轻量级图像超分辨率网络.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9723.html

Lightweight Image Super-resolution Network Based on Progressive Fusion of Hierarchical Feature

ZHANG Hao, MA Ji, YUAN Jiang

(School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: In recent years, with the development of deep learning techniques, convolutional neural network (CNN) and Transformers have made significant progress in image super-resolution. However, for the extraction of global features of an image, it is common to stack individual operators and repeat the computation to gradually expand the receptive field. To better utilize global information, this study proposes that local, regional, and global features should be explicitly modeled. Specifically, local information, regional-local information, and global-regional information of an image are extracted and fused hierarchically and progressively through channel attention-enhanced convolution, a dual-branch parallel architecture consisting of a window-based Transformer and CNN, and a dual-branch parallel architecture consisting of a standard Transformer and a window-based Transformer. In addition, a hierarchical feature fusion method is designed to fuse the local information extracted from the CNN branch and the regional information extracted from the window-based Transformer. Extensive experiments show that the proposed network achieves better results in lightweight SR. For example, in the 4× upscaling experiments on the Manga109 dataset, the peak signal-to-noise ratio (PSNR) of the proposed network is improved by 0.51 dB compared to SwinIR.

Key words: image super-resolution; Transformer; convolutional neural network (CNN); hierarchical feature fusion; global feature extraction

① 收稿时间: 2024-06-05; 修改时间: 2024-06-28; 采用时间: 2024-07-11; csa 在线出版时间: 2024-11-15

单图像超分辨率 (single image super-resolution, SISR) 是一个经典的低级视觉任务, 旨在通过生成细节信息将低分辨率 (low-resolution, LR) 图像恢复为对应的高分辨率 (high-resolution, HR) 图像. 它在遥感、视频监控和医学成像等领域应用广泛. 由于超分辨率 (super-resolution, SR) 任务本身是一个反向求解任务, 即给定的 LR 图像存在多个解, SISR 仍然存在挑战.

近 10 年来, 得益于深度学习技术的快速发展, 卷积神经网络 (convolutional neural network, CNN) 技术在 SR 领域的应用越来越多, 并取得了显著的效果. 基于 Transformer[1]的方法[2−4]通过有效地提取长距离依赖关系, 相比于基于 CNN 的方法[5−7]表现出了显著的性能提升. 自然图像包含全局、区域和局部范围层次特征结构[8]. 对于高质量的图像恢复, 充分利用退化图像中表现出的多尺度信息是至关重要的. CNN 适用于对边缘等仅覆盖数十个像素 (例如 3×3 卷积) 的局部特征进行建模. 而对覆盖几十个甚至上百个像素的区域特征进行建模, 具有窗口注意力机制的 Transformer (例如窗口为 8×8、16×16) 相比于 CNN 更加适合. 对于全局特征的建模, 很多方法[3,6,9,10]并没有使用单个计算模块显式地获取全局依赖关系, 而是通过重复计算 CNN 或 Transformer 单个算子逐步传播特征实现的. 然而, 采用单个算子简单堆叠的方式, 忽略了 SR 任务中迫切需要的图像的多尺度纹理特征. 尤其是对于轻量级模型, 因为它无法堆叠足够多的层. 此外, 均匀算子的堆叠被证明是低效的, 并且存在相互作用范围的过早饱和. 本文致力于设计一个在保证计算复杂度的同时使用单个计算模块显式地对图像局部、区域、全局层次特征进行建模的轻量级 SR 网络.

本文提出了全局区域局部特征融合组 (global regional local feature fusion group, GRLFFG), 以一种逐步扩大感受野的方式来进行局部信息、区域与局部信息、全局与区域信息的提取和聚合. 具体来说, GRLFFG 包括 3 个关键组件, 即局部特征提取块 (local feature extraction block, LFEB), 区域与局部特征融合块 (regional local feature fusion block, RLFFB), 全局与区域特征融合块 (global regional feature fusion block, GRFFB). LFEB 通过通道注意增强卷积提取局部特征. RLFFB 是基于 CNN 和划分窗口 Transformer 的双分支并行架构, 对于这种架构, 如何融合两条分支的特征是至关重要的. 有些方法[4,11]采用简单地直接相加融合的方式, Chen 等[9]提出了双分支交叉聚合的方式. 本文则提出了层次特征融合 (hierarchical feature fusion, HFF), 利用不同膨胀因子的卷积来提取多尺度信息的同时引入了层次相加策略, 以弥补膨胀卷积的网格效应带来的信息损失. 从而获得更精细的特征, 实现划分窗口的 Transformer 分支提取到的区域特征和 CNN 分支提取到的局部特征更好地融合. 此外, 为了增强卷积分支的表达能力, 本文还设计了特征增强块 (feature enhancement block, FEB). GRFFB 是基于划分窗口的 Transformer 和标准 Transformer 的双分支架构, 输入特征除了做基于划分窗口的自注意力提取区域信息外, 还要并行地做标准自注意力提取全局信息以对区域信息进行补充. 为了解决标准自注意力操作计算复杂度大的问题, 同时尽可能地利用到更多特征信息, 本文首先使用卷积进行特征通道压缩, 然后对 Key 和 Value 进行平均池化操作, 最后做自注意力计算. 卷积操作在降低计算复杂度的同时又扩大了池化感受野. 通过以上设计, 我们提出了一个新的基于 Transformer 和 CNN 的轻量级 SR 网络, 称为 GRLSR. 它呈现出全尺度特征 (即局部、区域和全局特征) 提取能力. 在较小的模型参数量下有着优异的性能.

本文主要贡献包括以下内容.

1) 设计了一个新颖的 Transformer 和 CNN 结合的结构, 显式地对局部、区域、全局信息进行建模, 实现了较好的 SR 性能.

2) 提出一种简单有效的层次特征融合方法. 通过使用具有不同膨胀因子的卷积和层次相加策略, 使得 Transformer 和 CNN 的输出特征有效结合, 一种简单的特征增强方法也被设计以提高 CNN 的表达能力.

3) 提出了全局与区域特征融合块, 实现了对全局特征显式建模, 起到对基于窗口划分的自注意力信息补充的作用.

4) 本文在不同尺度的实验上, 验证了模型的强大性能. 大量的实验表明, 本文提出的模型达到了较为先进的性能.

1 相关工作

1.1 单图像超分辨率

目前单图像超分辨率的主流方法分为两类: 基于 CNN[5−7]和基于 Transformer[2,3,12−14]. Dong 等[5]提出的 SRCNN 是第 1 个将 CNN 引入 SR 领域的工作. 自此,

出现了更多简洁有效的骨干网络. 例如, Lim 等[6]提出的 EDSR 去除了残差网络的批量归一化 (batch normalization, BN)[15]层, 实现了效果地显著提升. Zhang 等[16]提出的 RCAN 利用通道注意力来增强 SR 模型的表示能力. Zhang 等[16]提出的 RDN 在骨干网络中引入稠密连接也被证明是有效的. 为了在资源受限的设备下也能实现较好的重建质量, 一些方法[17−19]探索了轻量级的网络架构设计. Kim 等[20]提出的 DRCN 将递归操作应用于 SISR 任务之中, 大大减少了参数量. Anh 等[10]提出的 CARN 设计了一种在残差网络上实现级联机制的架构. Hui 等[21]提出的 IMDN 通过信息蒸馏, 逐步提取层次特征. Sun 等[22]提出的 HPUN 引入了高效且有效的下采样模块, 具体来说是利用 Pixel-Unshuffle 操作对输入特征进行下采样. 最近, 基于 Transformer 的 SISR 方法相较于基于 CNN 的方法取得了更加优异的效果. 其中, Liang 等[3]提出的 SwinIR 把 Liu 等[23]提出的 Swin Transformer 作为骨干网络实现了当时的 SOTA 性能.

1.2 视觉 Transformer

Transformer[1]首先在自然语言处理 (natural language processing, NLP) 领域取得了巨大突破. Dosovitskiy 等[12]提出了 ViT (vision Transformer), 它将 Transformer 应用到计算机视觉任务上. ViT 在各种高级视觉任务上都取得了令人震惊的结果, 比如目标检测, 分类等. 同时, 为了提高 ViT 的性能, 人们提出了许多有效的注意机制. Liu 等[23]提出的 Swin Transformer 使用局部窗口注意, 并通过移位操作实现窗口之间的交互. Dong 等[24]提出的 CSwin 介绍了十字形窗口注意机制. 受 Transformer 在高级任务中的成功启发, 一些工作试图将 Transformer 应用于低级视觉任务, 如 Chen 等[2]提出的 IPT. 然而, 自注意力的计算复杂度与图像大小成二次方关系, 无法直接应用标准的 Transformer. 为了降低计算复杂度, 很多方法被提了出来. 比如, Liang 等[3]提出的 SwinIR 基于 Swin Transformer 将图像划分成 8×8 的窗口, 在每个窗口内进行自注意力计算. Zhang 等[14]提出的 ELAN 设计了一个共享注意机制, 大大减小了计算量. Chen 等[4]提出的 CAT 设计了矩形窗口注意力, 具体来说是利用不同头部的水平和垂直矩形窗口注意力并行扩展注意力区域并聚合跨不同窗口的特征, 在不增加计算成本的情况下扩大了感受野. 一些方法也提出了将 CNN 和 Transformer 结合起来学习图像信息的局部和全局表示. 例如, Wang 等[25]提出的 OmniSR 是一种将 CNN 和 Transformer 串行的结构, 以逐步分层的方式编码上下文关系逐步扩大感受野. 而 Chen 等[9]提出的 DAT 是一种 CNN 和 Transformer 并行的结构, 在空间和通道两个维度上实现块间特征聚合.

2 方法

2.1 GRLSR 网络架构

如图 1(a) 所示, 所提出的 GRLSR 网络包括 3 个模块: 浅层特征提取、深层特征提取和图像重建.

浅层特征提取网络$H_{\mathrm{SF}}(\cdot)$是一个 3×3 的卷积层. 具体来说, 对于给定的低分辨率图像$\boldsymbol{I}_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times C_{\mathrm{in}}}$ (H, W, C_{in}是图像的高, 宽和输入通道数), 使用$H_{\mathrm{SF}}(\cdot)$来提取浅层特征$\boldsymbol{F}_0 \in \mathbb{R}^{H\times W\times C}$. 过程为:

$$\boldsymbol{F}_0 = H_{\mathrm{SF}}(\boldsymbol{I}_{\mathrm{LR}}) \tag{1}$$

其中, C是特征图的通道数. 卷积层提供了一种将输入图像空间映射到高维特征空间的简单方法, 它将输入图像的 RGB 三维颜色空间映射到高维特征空间.

深层特征提取网络由K个级联的 GRLFFG 和卷积构成. 特征提取过程可以描述为:

$$\begin{cases} \boldsymbol{F}_i = H^i_{\mathrm{GRLFFG}}(\boldsymbol{F}_{i-1}),\ i = 1,2,\cdots,K \\ \boldsymbol{F}_{DF} = CONV(\boldsymbol{F}_K) \end{cases} \tag{2}$$

其中, $H^i_{\mathrm{GRLFFG}}(\cdot)$表示第$i$个 GRLFFG, $CONV(\cdot)$表示最后一个卷积层.

将通过浅层特征提取网络得到的浅层特征$\boldsymbol{F}_0$和通过深层特征提取网络得到的深层特征$\boldsymbol{F}_{\mathrm{DF}}$通过一个全局残差连接[26]进行融合, 然后送入图像重建网络H_{REC}. 过程为:

$$\boldsymbol{I}_{\mathrm{SR}} = H_{\mathrm{REC}}(\boldsymbol{F}_{\mathrm{DF}} + \boldsymbol{F}_0) \tag{3}$$

其中, H_{REC}是图像重建网络, 来对图像进行上采样, 由卷积层和亚像素卷积层[27]构成. 最后, 输出 SR 图像.

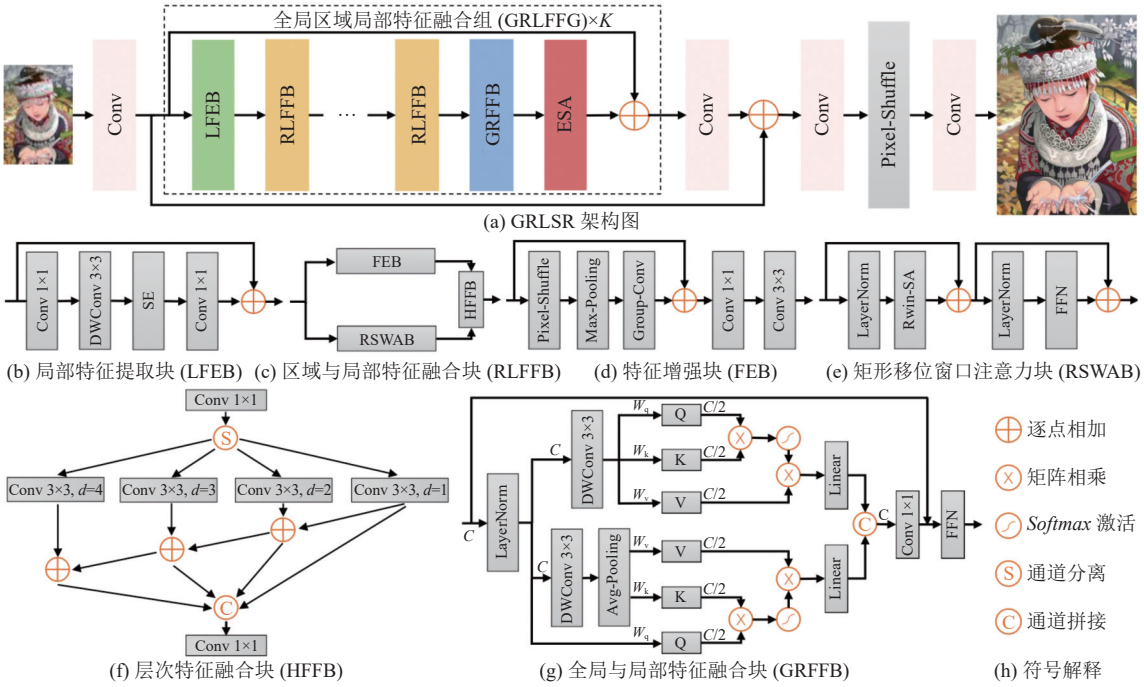
2.2 全局区域局部特征融合组 (GRLFFG)

如图 1(a) 所提出的全局区域局部特征融合组 (GRLFFG) 由局部特征提取块 (LFEB), N个级联的区域与局部特征融合块 (RLFFB), 全局与区域特征融合块 (GRFFB), 增强的空间注意力模块 (ESA) 和一个残差连接构成.

2.2.1 局部特征提取块 (LFEB)

如图 1(b) 所示, 该块参考 EfficientNet[28]中的

MBConv 块, 在用于升维的 1×1 卷积和 3×3 的深度可分离卷积之间添加一个 SE 模块[29]以自适应地重新加权通道级特征, 最后的 1×1 卷积用于调整输出维度. 该块旨在聚合局部上下文信息并提高网络的可训练性[25].



(a) GRLSR 架构图

(b) 局部特征提取块 (LFEB)　(c) 区域与局部特征融合块 (RLFFB)　(d) 特征增强块 (FEB)　(e) 矩形移位窗口注意力块 (RSWAB)

(f) 层次特征融合块 (HFFB)　(g) 全局与局部特征融合块 (GRFFB)　(h) 符号解释

图 1　GRLSR、LFEB、RLFFB、FEB、RSWAB、HFFB、GRFFB 结构图

2.2.2　区域与局部特征融合块 (RLFFB)

如图 1(c) 所示, 该块采用的是 Transformer 和 CNN 并行的架构. Transformer 可以有效地捕获全局信息并对像素之间的长距离依赖关系进行建模. 然而, CNN 的归纳偏置 (平移不变性和局部性) 在图像恢复任务中仍是必不可少的. 它可以聚合局部特征, 比如角和边缘. 为了用局部性来补充 Transformer 并实现区域和局部信息聚合, 在计算自注意力时添加一个独立的卷积操作块: 特征增强块 (FEB). FEB 块的具体实现如图 1(d) 所示. 在图 2 中对 FEB 块进行了更详尽的描述, 使用 4 种不同的颜色来标识原始图中的特征, 输入特征首先经过 Pixel-Shuffle 操作进行 2 倍上采样, 得到的高分辨率图像包含原始特征的完整信息. 为了在卷积操作之前能够更好地提取局部特征, 在 Pixel-Shuffle 之后使用了非线性操作 Max-Pooling, 它作用在原始特征图的不同通道上, 以保留原始特征的主要信息. 其窗口大小为 2×2, 以保证输入与输出特征在空间维度上匹配. 然后使用分组卷积 (Group-Conv) 来保证输出特征在通道维度上和输入特征保持一致. 在进入深度可分离卷积之前, 使用一个残差连接将提取到的图像信息添加到原始特征中, 以实现增强特征的作用.

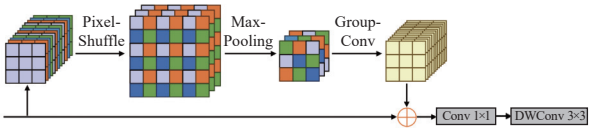


图 2　特征增强块 (FEB)

而对于 Transformer 分支, 如图 1(e) 所示, 本文参考 CAT[4]的设计, 使用划分窗口的移位矩形注意力机制, 并行地计算不同头部的水平和垂直矩形窗口注意力来扩展注意力区域并聚合跨不同窗口的特征. 为了更好地融合不同分支的特征, 如图 1(f) 所示, 本文设计了层次特征融合块 (HFFB). 利用多尺度特征表示对实现高质量的图像超分辨率是至关重要的, 一些方法[30,31]通过使用不同大小的卷积核来提取多尺度特征. 而本文采用的是具有不同膨胀系数的 3×3 卷积, 它具有更小的参数量. 为了弥补膨胀卷积存在的感受野中很多像素没有利用上的缺陷, 本文提出了一种层次相加的方法. 具体来说, 融合特征被划分为多份, 每份通过具有不同膨胀因子的卷积[32], 以提取多尺度的信息. 为了得到更精细的特征, 每一份输出特征经过层次相加之后再次融合. 如图 3 所示, 4 种颜色代表了 4 个膨胀系数, 通过依次相加前面的特征, 逐步扩大了感受野. 最后的 1×1 卷积用来融合不同尺度的特征并调整通道数.

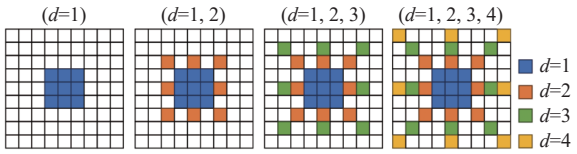


图 3 层次特征融合 (HFF)

2.2.3 全局与区域特征融合块 (GRFFB)

为了进一步扩大感受野, 提取到更大范围的特征, 在 RLFFB 之后设计了一个全局与区域特征融合块. 如图 1(g) 所示, 该块采用的是双分支并行自注意力架构, 一条分支做的是基于划分窗口的区域自注意力, 为了弥补划分窗口的自注意力带来的信息损失同时扩大感受野, 我们希望使用标准的 Transformer 来提取全局特征. 但是标准的 Transformer 计算量过大, 不适用于轻量级的超分任务, 同时受图像跨尺度相似性[8]的启发, 本文在 Key 和 Value 上先做平均池化 (Avg-Pooling) 操作来降低后续自注意力的计算量, 然后再进行投影操作, 而对于 Q 则是进行标准的投影操作使其具有完整信息. 它的实现如下:

$$\begin{cases} Q^a=\text{DWConv}(X)\cdot W_q,\ Q^b=\text{DWConv}(X)\cdot W'_q \\ K^a=\text{DWConv}(X)\cdot W_k,\ K^b=AVG(\text{DWConv}(X))\cdot W'_k \\ V^a=\text{DWConv}(X)\cdot W_v,\ V^b=AVG(\text{DWConv}(X))\cdot W'_v \\ Y^a=Softmax(Q^a\cdot(K^a)^{\text{T}})\cdot V^a \\ Y^b=Softmax(Q^b\cdot(K^b)^{\text{T}})\cdot V^b \\ Y=Concat(Y^a,Y^b) \end{cases} \tag{4}$$

其中, DWConv(·)表示逐深度卷积操作, $AVG(\cdot)$表示的是平均池化操作, $Concat(\cdot)$表示通道拼接操作. 对图像进行通道分离的操作会损失信息, 不利用图像的恢复任务, 因此对于以上两个分支, 使用卷积进行通道压缩, 保证原始图像的完整信息. 对以上两个分支得到的区域特征和感受野更大的全局特征进行通道拼接操作之后, 再用 1×1 卷积进行特征融合.

本文提出的架构可以对局部、区域和全局范围内的图像层次结构进行建模, 并且是以一种局部特征、区域与局部特征融合、全局与区域特征融合依次递进的方式来提取特征的. 在每个 GRLFFG 的末端, 使用了 ESA 块和残差连接来更好地增强模型的表达能力.

3 实验

3.1 实验设置

● 数据集和指标: 本文使用 DIV2K[33]的 800 张图片作为训练数据集. 在 5 个标准数据集 Set5[34]、Set14[35]、BSD100[36]、Urban100[37]、Manga109[38]上测试本文模型的性能. 本文在不同的缩放因子: ×2, ×3, ×4 下进行了实验. 在 YCbCr 空间的 Y 通道上, 使用峰值信噪比 (peak signal to noise ratio, PSNR)[39]和结构相似性 (structural similarity index, SSIM)[39]来评估模型的性能.

● 实施细节: 按照一般设置, 本文使用双 3 次下采样从原始 HR 图像中获取相应的 LR 图像. 在训练过程中, 将训练图像随机裁剪为 64×64 的图像片, 每批次随机输入 32 个图像片, 总训练迭代为 500k. 采用 Adam 优化器[40]来最小化$\boldsymbol{L}_1$损失, 其中$\beta_1=0.9$和$\beta_2=0.999$. 初始学习率设置为2×10^{-4}, 在 250k、400k、450k、475k 次迭代后学习率降为之前的一半. 此外, 在训练过程中, 利用随机旋转和水平翻转进行数据增强. 本文的模型基于 PyTorch[41]框架实现, 在 4 张 NVIDIA 2080Ti GPU 上进行训练. 在 GRLSR 中, GRLFFG 的个数为 3, 每个 GRLFFG 中的 RLFFB 的数量依次为 4、4、2. GRLFFG 中的特征通道、注意力头数和多层感知机 (multi-layer perceptron, MLP) 的扩展因子分别设置为 60、6 和 2.

3.2 实验结果

为了证明所提出的 GRLSR 模型的有效性, 本文在×2、×3、×4 尺度下与几个先进的轻量级 SISR 方法进行了比较. 包括: EDSR[6]、CARN[10]、IMDN[21]、MAFFSRN[42]、LatticeNet[18]、LAPAR[43]、SMSR[44]、HPUN[22]、ESRT[13]、ELAN[14]、LBNet[45]、SwinIR[3]、STSN[46]、Omni-SR[25]、CRAFT[47]和 MSRA-SR[48].

● 定量比较: 表 1 展示了不同的轻量级 SISR 方法在 5 个基准数据集上的性能. 可以看出, 对于不同倍数的 SR 任务, 在 5 个基准数据集上 GRLSR 都实现了最好的结果. 特别是对于目前轻量级 SISR 最好方法之一的 MSRA-SR[48], 在相同的参数量下, GRLSR 在×4 尺度下的 Set5、Urban100、Manga109 数据集上分别比 MSRA-SR 高了 0.11 dB、0.16 dB 和 0.35 dB. 在×2 和×3 上相比 MSRA-SR 也有较大的性能提升. 相比经典的 SwinIR 方法, 本文提出的 GRLSR 有更小的参数量, 同时性能有显著的提升, 在 3 个不同尺度下的 Urban100 和 Manga109 数据集上, 平均提升了 0.43 dB 和 0.47 dB.

表 1 不同 SISR 方法在×2、×3、×4 尺度下的参数量、PSNR 和 SSIM

放大倍数	模型	参数量 (k)	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM
×2	EDSR-baseline	1 370	37.99	0.960 4	33.57	0.917 5	32.16	0.899 4	31.98	0.927 2	38.54	0.976 9
	CARN	1 592	37.76	0.959 0	33.52	0.916 6	32.09	0.897 8	31.51	0.931 2	—	—
	IMDN	694	38.00	0.960 5	33.63	0.917 7	32.19	0.899 6	32.17	0.928 3	38.88	0.977 4
	MAFFSRN-L	790	38.07	0.960 7	33.59	0.917 7	32.23	0.900 5	32.38	0.930 8	—	—
	LatticeNet	756	38.15	0.961 0	33.78	0.919 3	32.25	0.900 5	32.43	0.930 2	—	—
	LAPAR-A	548	38.01	0.960 5	33.62	0.918 3	32.19	0.899 9	32.10	0.928 3	38.67	0.977 2
	SMSR	985	38.00	0.960 1	33.64	0.971 9	32.17	0.899 0	32.19	0.928 4	38.76	0.977 1
	ELAN-light	582	38.17	0.961 1	33.94	0.920 7	32.30	0.901 2	32.76	0.934 0	39.11	0.978 2
	SwinIR-light	878	38.14	0.961 1	33.86	0.920 6	32.31	0.901 2	32.76	0.934 0	39.12	0.978 3
	STSN	881	38.19	0.961 1	33.78	0.919 9	32.30	0.901 3	32.68	0.933 6	39.13	0.977 8
	Omni-SR	772	38.22	0.961 3	33.98	0.921 0	32.36	0.902 0	33.05	0.936 3	39.28	0.978 4
	CRAFT	737	38.23	0.961 5	33.92	0.921 1	32.33	0.901 6	32.86	0.934 3	39.39	0.978 6
	MSRA-SR	769	38.23	0.961 4	34.01	0.921 1	32.33	0.901 7	32.98	0.935 8	39.24	0.978 3
	ours	769	**38.26**	**0.961 6**	**34.13**	**0.922 6**	**32.38**	**0.902 3**	**33.24**	**0.937 9**	**39.45**	**0.978 7**
×3	EDSR-baseline	1 555	34.37	0.927 0	30.28	0.841 7	29.09	0.805 2	28.15	0.852 7	33.45	0.943 9
	CARN	1 592	34.29	0.925 5	30.29	0.840 7	29.06	0.803 4	27.38	0.840 4	—	—
	IMDN	703	34.36	0.927 0	30.32	0.841 7	29.09	0.804 6	28.17	0.851 9	33.61	0.944 5
	MAFFSRN-L	807	34.45	0.927 7	30.40	0.843 2	29.13	0.806 1	28.26	0.855 2	—	—
	LatticeNet	765	34.53	0.928 1	30.39	0.842 4	29.15	0.805 9	28.33	0.853 8	—	—
	LAPAR-A	594	34.36	0.926 7	30.34	0.842 1	29.11	0.805 4	28.15	0.852 3	33.51	0.944 1
	SMSR	993	34.40	0.927 0	30.33	0.841 2	29.10	0.805 0	28.25	0.853 6	33.68	0.944 5
	ESRT	770	34.42	0.926 8	30.43	0.843 3	29.15	0.806 3	28.46	0.857 4	33.95	0.945 5
	ELAN-light	590	34.61	0.928 8	30.55	0.846 3	29.21	0.808 1	28.69	0.862 4	34.00	0.947 8
	LBNet	736	34.47	0.927 7	30.38	0.841 7	29.13	0.806 1	28.42	0.855 9	33.82	0.946 0
	SwinIR-light	886	34.62	0.928 9	30.54	0.846 3	29.20	0.808 2	28.66	0.862 4	33.98	0.947 8
	STSN	888	34.62	0.929 2	30.54	0.846 6	29.22	0.809 0	28.59	0.862 1	34.11	0.948
	Omni-SR	780	34.70	0.929 4	30.57	0.846 9	29.28	0.809 4	28.84	0.865 6	34.22	0.948 7
	CRAFT	744	34.71	0.929 5	30.61	0.846 9	29.24	0.809 3	28.77	0.863 5	34.29	0.949 1
	MSRA-SR	777	34.65	0.929 1	30.60	0.847 0	29.24	0.809 3	28.86	0.866 4	34.29	0.948 9
	ours	777	**34.76**	**0.930 2**	**30.70**	**0.848 5**	**29.32**	**0.811 0**	**29.13**	**0.871 1**	**34.56**	**0.950 4**
×4	EDSR-baseline	1 518	32.09	0.893 8	28.58	0.781 3	27.57	0.735 7	26.04	0.784 9	30.35	0.906 7
	IMDN	715	32.21	0.894 8	28.58	0.781 1	27.56	0.735 3	26.04	0.783 8	30.45	0.907 5
	MAFFSRN-L	830	32.20	0.895 3	28.62	0.782 2	27.59	0.737 0	26.16	0.788 7	—	—
	LatticeNet	777	32.30	0.896 2	28.68	0.783 0	27.62	0.736 7	26.25	0.787 3	—	—
	LAPAR-A	659	32.15	0.894 4	28.61	0.781 8	27.61	0.736 6	26.14	0.787 1	30.42	0.907 4
	SMSR	1 006	32.12	0.893 2	28.55	0.780 8	27.55	0.735 1	26.11	0.786 8	30.54	0.908 5
	HPUN-L	734	32.31	0.896 2	28.73	0.784 2	27.66	0.738 6	26.27	0.791 8	30.77	0.910 9
	ESRT	751	32.19	0.894 7	28.69	0.783 3	27.69	0.737 9	26.39	0.796 2	30.75	0.910 0
	ELAN-light	601	32.43	0.897 5	28.78	0.785 8	27.69	0.740 6	26.54	0.798 2	30.92	0.915 0
	LBNet	742	32.29	0.896 0	28.68	0.783 2	27.62	0.738 2	26.27	0.790 6	30.76	0.911 1
	SwinIR-light	897	32.44	0.897 6	28.77	0.785 8	27.69	0.740 6	26.47	0.798 0	30.92	0.915 1
	STSN	898	32.46	0.898 2	28.76	0.786 0	27.68	0.740 5	26.39	0.797 1	30.93	0.914 2
	Omni-SR	792	32.49	0.898 8	28.78	0.785 9	27.71	0.741 5	26.64	0.801 8	31.02	0.915 1
	CRAFT	753	32.52	0.898 9	28.85	0.787 2	27.72	0.741 8	26.56	0.799 5	31.18	0.916 8
	MSRA-SR	789	32.46	0.898 4	28.86	0.787 6	27.72	0.741 9	26.65	0.803 7	31.08	0.915 7
	ours	789	**32.57**	**0.899 7**	**28.88**	**0.788 4**	**27.75**	**0.743 4**	**26.81**	**0.808 8**	**31.43**	**0.918 7**

● 定性比较: 如图 4、图 5 所示, 本文提供了不同轻量级 SISR 方法在×2、×4 尺度下的 Urban100 数据集的视觉比较. 可以观察到, 本文提出的 GRLSR 包含了更多的细粒度细节, 而其他方法在复杂区域生成了更多的模糊边缘或者伪影. 例如, 在×4 尺度下的图像“img004.png”中, 可以清晰地看到本文的方法恢复了更准确的结构和更清晰的纹理细节. 得益于在模型中提出的全局与区域特征融合块, 对于划分窗口的自注意力起到了补充作用, 扩大了感受野, 因此在像“img100.png”这种具有长条的纹理特征图片上, 本文的模型有更好的

恢复效果. 本文提出的模型集成了 CNN 和 Transformer 的优势, 并以一种逐步扩大感受野的方式来提取特征, 可以有效地提取图像的局部信息、区域信息和全局信息. 实验结果证明了本文所提出方法的合理性和有效性.

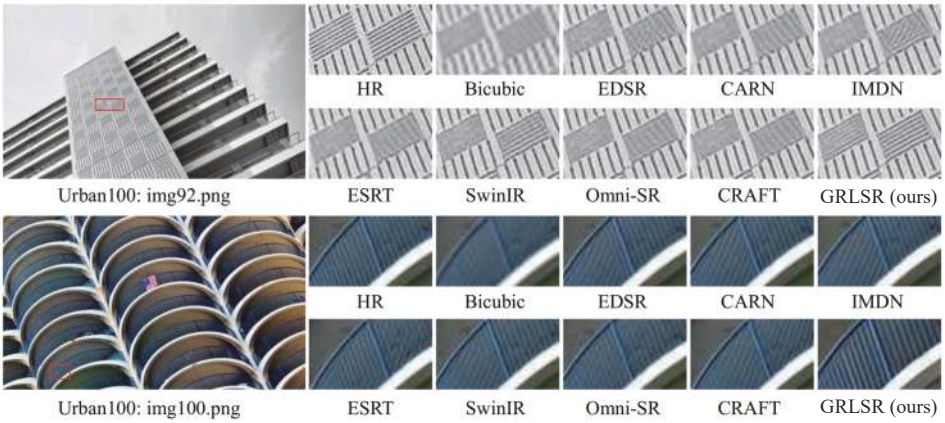


图 4　2 倍 SR 的视觉效果比较

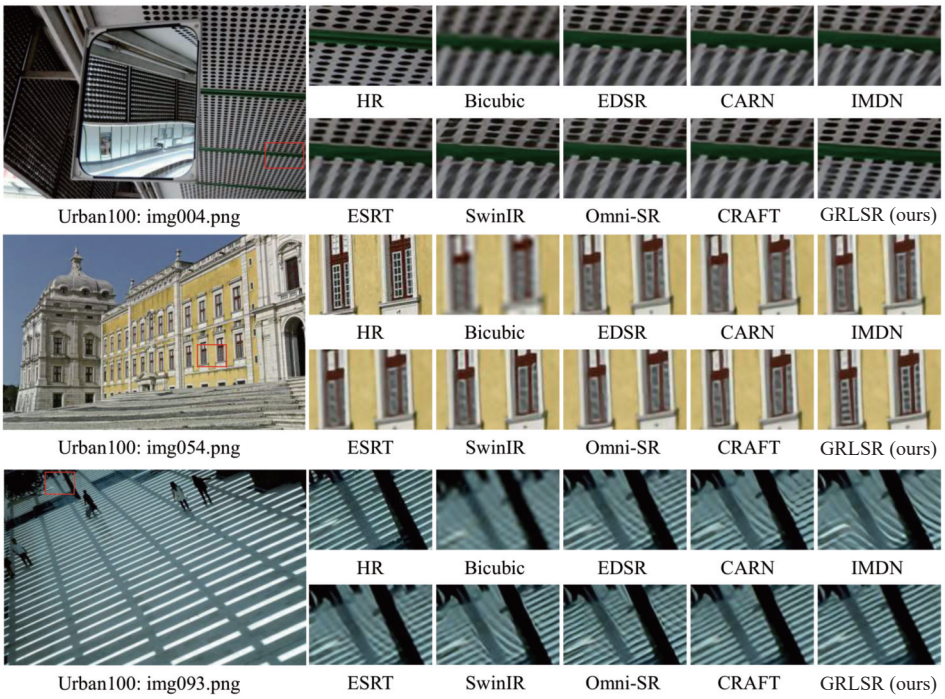


图 5　4 倍 SR 的视觉效果比较

3.3　消融实验

对于消融实验, 将批次大小设置为 16, 在 DIV2K 数据集上训练 500k 次迭代, 在×4 尺度的 Urban100 和 Manga109 数据集上进行测试. 本文通过从整个模型架构中删除单个模块来分析模块的有效性, 或者删除某些模块的一些具体操作来验证设计的合理性. 但为了保证消融实验的公平性, 即进行消融实验的模型参数量与原模型的参数量大小尽可能保持一致. 因此对于参数量较大的模块, 在删除之后我们加深了原有网络的层数.

● 局部特征提取块 (LFEB): 对于注重细节的 SR 任务来说, 局部信息的提取也是至关重要的. LFEB 中应用了通道注意力, 可以选择性地强调与任务更相关的特征, 同时减少不太相关的特征. 它的应用可以使模型更好地提取到局部信息. 实验结果如表 2 所示, 可以看到删除 LFEB 块之后, 在 Urban100 和 Manga109 数据集上 PSNR 值都降低了 0.08 dB.

● 区域与局部特征融合块 (RLFFB): 对于 RLFFB,

本文需要验证提出的特征增强块 (FEB) 和层次特征融合块 (HFFB) 的有效性. 对于 FEB, 1×1 和 3×3 卷积之前的操作是为了提取图像特征的主要信息, 通过 Pixel-Shuffle 和 Max-Pooling 这种简单的操作组合, 实现了更好地增强特征表示的效果. 对于消融实验, 只删除 1×1 和 3×3 卷积之前的操作, 卷积分支还保留 1×1 和 3×3 卷积. 实验结果如表 3 所示, 在仅增加了 6k 参数量的基础上, 性能有了很好地提升, 比如在 Urban100 和 Manga109 数据集上 PSNR 分别提高了 0.07 dB 和 0.04 dB. 对于 HFFB, 本文设置了两组消融实验. 一组是对于不同膨胀因子的卷积, 本文利用不同膨胀因子的卷积来获取图像的多尺度特征. 为了验证这种方式的有效性, 消融实验用一个普通的组数为 4 的分组卷积替代, 来保证模型的参数量与原模型一致. 实验结果如表 4 所示, 本文的方法在 Urban100 和 Manga109 数据集上 PSNR 高了 0.07 dB 和 0.05 dB. 另外一组用来验证提出的层次相加方法, 如图 1(f) 所示的, 即每进行一次不同的膨胀因子卷积操作之后都加上之前卷积操作的结果, 用来逐步扩大感受野. 这种方式在不增加模型参数量下, 使性能得到了不错地提升. 实验结果如表 5 所示, 在 Urban100 和 Manga109 数据集上 PSNR 都提高了 0.06 dB.

表 2 GRLSR 包含和不包含 LFEB 在 4 倍 SR 的比较

模型	参数量 (k)	Urban100	Manga109
w/o LFEB	760	26.66	31.17
GRLSR	789	26.74	31.25

表 3 RLFFB 包含和不包含 FEB 在×4 SR 的比较

模型	参数量 (k)	Urban100	Manga109
w/o FEB	783	26.67	31.21
GRLSR	789	26.74	31.25

表 4 HFFB 中的膨胀卷积是否用分组卷积代替在×4 SR 的比较

模型	参数量 (k)	Urban100	Manga109
w/o HFFB	789	26.67	31.20
GRLSR	789	26.74	31.25

● 全局与区域特征融合块 (GRFFB): 对于 GRFFB, 需要验证提出的获取全局信息的 Transformer 块对于划分窗口 Transformer 的补充作用. 本文提出的 GRFFB 获取全局信息的分支通过对 Key、Value 进行 Avg-Pooling 操作之后进行标准的 Transformer 计算, 这种方式相比划分窗口的方式极大地扩大了感受野, 这对于 SR 任务也是至关重要的. 对于消融实验, 删除整个 GRFFB, 为了保证实验的公平性, 增加了 RLFFB 的数量, 从而使得模型的参数量尽可能和原模型保持一致, 同时也能证明本文设计的用于增加感受野的 GRFFB 的有效性. 实验结果如表 6 所示, 相比原模型, 在 Urban-100 和 Manga109 数据集上 PSNR 降低了 0.09 dB 和 0.14 dB. 消融实验的结果充分证明了 GRFFB 的有效性和必要性.

表 5 HFFB 包含和不包含层次相加在×4 SR 的比较

模型	参数量 (k)	Urban100	Manga109
w/o +	789	26.68	31.19
GRLSR	789	26.74	31.25

表 6 GRLSR 包含和不包含 GRFFB 在×4 SR 的比较

模型	参数量 (k)	Urban100	Manga109
w/o GRFFB	751	26.65	31.11
GRLSR	789	26.74	31.25

4 结论与展望

本文提出了一个轻量级的 SISR 网络架构: GRLSR, 实现了使用单个计算模块显式地对图像局部、区域、全局层次特征进行建模. 提出了全局区域局部特征融合组 (GRLFFG), 它包括 3 个关键组件, 即局部特征提取块 (LFEB), 区域与局部特征融合块 (RLFFB), 全局与区域特征融合块 (GRFFB), 它以提取局部信息、区域与局部信息、全局与区域信息一种逐步扩大感受野的方式进行 SR 任务. 此外本文提出了一个特征增强块 (FEB), 它以一种纯卷积的方式用于增强局部特征. 本文还提出了一个层次特征融合块 (HFFB), 它使用不同膨胀因子的卷积来获取多尺度特征. 对于 RLFFB, 本文使用了卷积与矩形窗口的 Transformer 双分支并行架构. 对于 GRFFB, 一条分支做划分窗口的 Transformer, 另外一条分支做标准的 Transformer, 但通过通道压缩和 Avg-Pooling 操作来降低计算量. 实验结果表明, 在轻量级的 SISR 框架中, 本文提出的 GRLSR 实现了更好的结果. 下一步研究的是在较低的计算量下能更好地提取全局特征.

参考文献

1 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

2 Chen HT, Wang YH, Guo TY, *et al*. Pre-trained image

processing Transformer. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12294–12305.

3 Liang JY, Cao JZ, Sun GL, *et al*. SwinIR: Image restoration using swin Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 1833–1844.

4 Chen Z, Zhang YL, Gu JJ, *et al*. Cross aggregation Transformer for image restoration. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1847.

5 Dong C, Loy CC, He KM, *et al*. Learning a deep convolutional network for image super-resolution. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 184–199.

6 Lim B, Son S, Kim H, *et al*. Enhanced deep residual networks for single image super-resolution. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1132–1140.

7 Zhang YL, Li KP, Li K, *et al*. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision and Pattern Recognition. Munich: Springer, 2018. 294–310.

8 Li YW, Fan YC, Xiang XY, *et al*. Efficient and explicit modelling of image hierarchies for image restoration. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 18278–18289.

9 Chen Z, Zhang YL, Gu JJ, *et al*. Dual aggregation Transformer for image super-resolution. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 12278–12287.

10 Ahn N, Kang B, Sohn KA. Fast, accurate, and lightweight super-resolution with cascading residual network. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 256–272.

11 Chen XY, Wang XT, Zhou JT, *et al*. Activating more pixels in image super-resolution Transformer. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22367–22377.

12 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

13 Lu ZS, Li JC, Liu H, *et al*. Transformer for single image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 456–465.

14 Zhang XD, Zeng H, Guo S, *et al*. Efficient long-range attention network for image super-resolution. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 649–667.

15 Ioffe S. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1942–1950.

16 Zhang YL, Tian YP, Kong Y, *et al*. Residual dense network for image super-resolution. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2472–2481.

17 Tai Y, Yang J, Liu XM. Image super-resolution via deep recursive residual network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2790–2798.

18 Luo XT, Xie Y, Zhang YL, *et al*. LatticeNet: Towards lightweight image super-resolution with lattice block. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 272–289.

19 Hui Z, Wang XM, Gao XB. Fast and accurate single image super-resolution via information distillation network. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 723–731.

20 Kim J, Lee JK, Lee KM. Deeply-recursive convolutional network for image super-resolution. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1637–1645.

21 Hui Z, Gao XB, Yang YC, *et al*. Lightweight image super-resolution with information multi-distillation network. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 2024–2032.

22 Sun B, Zhang YL, Jiang SY, *et al*. Hybrid pixel-unshuffled network for lightweight image super-resolution. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 2375–2383.

23 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

24 Dong XY, Bao JM, Chen DD, *et al*. CSWin Transformer: A general vision Transformer backbone with cross-shaped windows. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 12114–12124.

25 Wang H, Chen XH, Ni BB, *et al*. Omni aggregation networks for lightweight image super-resolution. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern

Recognition. Vancouver: IEEE, 2023. 22378–22387.
26 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
27 Shi WZ, Caballero J, Huszár F, *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1874–1883.
28 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.
29 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
30 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826.
31 周登文, 李文斌, 李金新, 等. 一种轻量级的多尺度通道注意图像超分辨率重建网络. 电子学报, 2022, 50(10): 2336–2346. [doi: 10.12263/DZXB.20201089]
32 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122, 2016.
33 Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1122–1131.
34 Bevilacqua M, Roumy A, Guillemot C, *et al*. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. Proceedings of the 2012 British Machine Vision Conference. Surrey: BMVA Press, 2012. 1–10. [doi: 10.5244/C.26.135]
35 Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. Proceedings of the 7th International Conference on Curves and Surfaces. Avignon: Springer, 2012. 711–730.
36 Martin D, Fowlkes C, Tal D, *et al*. 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: IEEE, 2001. 416–423.
37 Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5197–5206.
38 Matsui Y, Ito K, Aramaki Y, *et al*. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 2017, 76(20): 21811–21838. [doi: 10.1007/s11042-016-4020-z]
39 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]
40 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.
41 Paszke A, Gross S, Chintala S, *et al*. Automatic differentiation in PyTorch. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1–4.
42 Muqeet A, Hwang J, Yang SB, *et al*. Multi-attention based ultra lightweight image super-resolution. Proceedings of the 2020 European Conference on Computer Vision. Glasgow: Springer, 2020. 103–118.
43 Li WB, Zhou K, Qi L, *et al*. 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: Curran Associates Inc., 2020. 1708.
44 Wang LG, Dong XY, Wang YQ, *et al*. Exploring sparsity in image super-resolution for efficient inference. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4915–4924.
45 Gao GW, Wang ZX, Li JC, *et al*. 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: ijcai.org, 2022. 913–919.
46 Gendy G, Sabor N, Hou JC, *et al*. A simple Transformer-style network for lightweight image super-resolution. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 1484–1494.
47 Li A, Zhang L, Liu Y, *et al*. Feature modulation Transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 12480–12490.
48 Zhou XQ, Huang HB, He R, *et al*. MSRA-SR: Image super-resolution Transformer with multi-scale shared representation acquisition. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 12665–12676.

(校对责编: 张重毅)