计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):85−94 [doi: 10.15888/j.cnki.csa.009600]

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

基于 Swin Transformer 的遥感图像超分辨率重建①

孔　锐, 冉友红

(暨南大学 智能科学与工程学院, 珠海 519070)
通信作者: 冉友红, E-mail: 2060920862@qq.com

摘　要: 由于遥感图像中的物体具有不确定性, 同时不同图像之间的特征信息差异较大, 导致现有超分辨率方法重建效果差, 因此本文提出一种结合 Swin Transformer 和 N-gram 模型的 NG-MAT 模型来实现遥感图像超分辨率. 首先, 在原始 Transformer 计算自注意力的分支上并联多注意力模块, 用于提取全局特征信息来激活更多像素. 其次, 将自然语言处理领域的 N-gram 模型应用到图像处理领域, 用三元 N-gram 模型来加强窗口之间的信息交互. 本文提出的方法在所选取的数据集上, 峰值信噪比在放大因子为 2、3、4 时达到了 34.68 dB、31.03 dB、28.99 dB, 结构相似度在放大因子为 2、3、4 时达到了 0.926 6、0.844 4、0.773 4, 实验结果表明, 本文提出的方法各个指标都优于其他同类方法.

关键词: Swin Transformer; 超分辨率; N-gram; 遥感图像

引用格式: 孔锐,冉友红.基于 Swin Transformer 的遥感图像超分辨率重建.计算机系统应用,2024,33(9):85–94. http://www.c-s-a.org.cn/1003-3254/9600.html

Super-resolution Reconstruction of Remote Sensing Image Based on Swin Transformer

KONG Rui, RAN You-Hong

(School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China)

Abstract: Due to the uncertainty of objects in remote sensing images and significant differences in feature information between different images, existing super-resolution methods yield poor reconstruction results. Therefore, this study proposes an NG-MAT model that combines the Swin Transformer and the N-gram model to achieve super-resolution of remote sensing images. Firstly, multiple attention modules are connected in parallel on the branch of the original Transformer to extract global feature information for activating more pixels. Secondly, the N-gram model from natural language processing is applied to the field of image processing, utilizing a trigram N-gram model to enhance information interaction between windows. The proposed method achieves peak signal-to-noise ratios of 34.68 dB, 31.03 dB, and 28.99 dB at amplification factors of 2, 3, and 4, respectively, and structural similarity indices of 0.926 6, 0.844 4, and 0.773 4 at the same amplification factors on the selected dataset. Experimental results demonstrate that the proposed method outperforms other similar methods in various metrics.

Key words: Swin Transformer; super-resolution; N-gram; remote sensing image

1　引言

图像超分辨率是计算机视觉领域中一个具有挑战性的任务, 其目的是从低分辨率图像中提取特征信息, 然后重建出高分辨率图像. 图像超分辨率在生活中具有广泛的应用, 如医学成像[1,2]、遥感图像处理[3,4]、安防监控[5]等. 图像超分辨率不仅能提高图像的质量, 更重要的是服务于其他计算机视觉任务, 如目标检测、图像分割等. 这些任务中, 图像质量越好, 实验结果往

① 收稿时间: 2024-03-05; 修改时间: 2024-04-03; 采用时间: 2024-04-10; csa 在线出版时间: 2024-07-26
CNKI 网络首发时间: 2024-07-29

往越理想.

遥感图像是通过卫星、航空器、无人机等设备获取的地球表面的图像. 通过对遥感图像的分析, 能够为环境监测、地形地貌分析、城市规划、水资源管理、自然灾害监测等提供巨大的帮助. 然而, 由于硬件设备的性能有限、大气扰动和云覆盖以及图像传输过程中的压缩等因素, 遥感图像的分辨率通常较低, 直接分析原始遥感图像效果可能不理想. 因此在分析之前需要对原始图像进行处理, 以充分挖掘图像中的信息, 提高分辨率, 使其更适用于各种实际应用场景. 故可以使用图像超分辨率来解决这个问题, 充分提取原始图像的高频信息重建出高分辨率图像, 然后再应用到其他视觉任务.

由于图像退化的原因不确定, 所以很难找到统一的方法来重建高分辨率图像, 因此图像超分辨率是一个不适定问题 (ill-posed problem). 基于插值的传统方法根据邻近像素点的值来获得高分辨率图像, 虽然计算复杂度低, 但是一些高频信息容易被忽略, 导致重建图像太平滑, 视觉效果很差. 随着深度学习的发展, 越来越多的研究者致力于使用深度学习方法来实现图像超分辨率的任务. Dong 等首次提出基于神经网络的图像超分辨率模型 SRCNN[6], 实验结果表明其重建效果远好于传统方法, 自此图像超分辨率进入了深度学习时代. 先后出现了 EDSR[7]、VDSR[8]、RCAN[9]等基于 CNN 的图像超分辨率模型, 这些模型结构更复杂、网络层数更多, 能够提取更多高频信息, 因此重建的图像视觉效果更佳. 生成对抗网络 (generative adversarial network) 出现之后, 研究人员发现这种结构非常适合图像超分辨率任务, SRGAN[10]、ESRGAN[11]、Rank-SRGAN[12]等基于 GAN 的图像超分辨率模型都取得了不错的效果. Transformer 在自然语言处理中的成功让人们开始思考其是否可用于图像处理之中, ViT[13]、Swin Transformer[14]的出现证明了 Transformer 应用到计算机视觉领域的可行性. SwinIR[15]、ESRT[16]、HNCT[17]、HAT[18]等基于 Transformer 的超分辨率重建模型开创了 SR 的新纪元, 在各个常用测试集上都取得了最好的效果.

遥感图像超分辨率重建主要有两种思路: 一是直接将自然图像超分辨率方法用于遥感图像超分辨率, 另一种是根据遥感图像的特点设计更适合遥感图像超分辨率的模型. 实验结果表明, 第 1 种方法的效果不太理想, 因此更多研究者倾向于研究新的方法. Haut 等[19]提出一种卷积生成模型, 该模型先学习低分辨率图像到高分辨率图像的映射关系, 然后将数据对称投影到目标分辨率, 保证低分辨率输入图像的重建约束. DRGAN[20]根据遥感图像的特点设计了一个密集残差网络作为 GAN 中的生成器网络, 这种结构可以充分提取低分辨率图像的层次特征, 同时根据具有梯度惩罚的 Wasserstein GAN[21]修改了损失函数并更改了判别器网络的结构, 以便训练过程更稳定. MHAN[3]中提出了用于恢复细节的高阶注意力 (HOA) 机制, 同时为了充分利用分层特征, 作者引入了频率感知来连接特征提取和特征细化网络. Jia 等[4]将注意力机制与生成对抗网络结合提出了多注意力 GAN (MA-GAN), 其中的 PCRD 模块可以自动学习和调整残差的比例, AUP 模块使用像素注意力 (PA) 来实现图像的重建.

自然图像中通常包含丰富的颜色、纹理和确定的结构, 遥感图像则恰恰相反. 遥感图像在不同的场景中差异较大, 如城镇、森林等场景的图像内容丰富、纹理细节清晰, 但是沙漠、海洋等场景的图像颜色单一、高低频信息不明确, 这种差异给超分辨率任务带来了巨大的挑战. 其次, 基于 Transformer 的方法中间特征往往会出现块伪影 (blocking artifacts), 移位窗口机制并不能很好地实现跨窗口信息交互[18].

为了解决上述问题, 本文提出了基于多注意力和 N-gram 模型的 Transformer 模型 NG-MAT, 如图 1 所示.

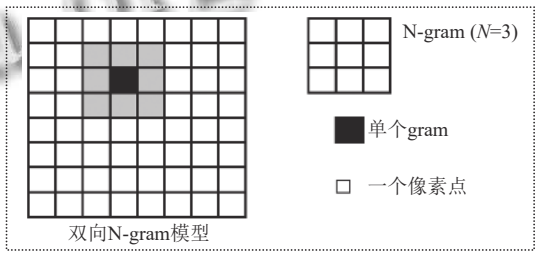

图 1 图像中的 N-gram 模型 (N=3)

本文提出的 NG-MAT 模型主要由 MAB 和 NGB 两个模块构成. MAB 模块结合了通道注意力、空间注意力和自注意力机制, 能够充分提取图像中的局部信息和全局信息, 全局特征信息反映了图像的轮廓特征, 是图像的整体属性, 局部特征反映了图像的细节特征, 是图像的区域属性, 因此这种结构能适应遥感图像复杂的场景; 此外, 本文还使用 N-gram 模块来提取相邻窗口内的特征, 实现窗口之间的特征交互. 实验

结果表明, 本文提出的模型在定量对比和视觉效果对比上, 都取得了最好的效果. 本文的主要贡献可归纳如下.

(1) 将 NLP 领域的三元 N-gram 语言模型应用到图像处理领域. 本文用 NGB 模块提取遥感图像的 N-gram context, 实现跨窗口信息交互.

(2) 提出 MAM 模块从两个不同维度提取图像的特征信息. 将通道注意力和空间注意力通过并联方式结合, 能够从两个不同维度提取图像的全局特征信息.

(3) 在原始 Transformer 计算自注意力的分支上并联 MAM 模块, 同时提取遥感图像的全局特征和局部特征.

本文第 2 节介绍本文提出的 NG-MAT 模型的结构. 第 3 节为本文的实验结果及分析. 第 4 节是结论.

2 本文方法

2.1 模型整体结构

如图 2(a) 所示, 本文提出的 NG-MAT 主要由 3 部分构成: 浅层特征提取模块 (shallow feature extraction module)、深层特征提取模块 (deep feature extraction module) 和图像恢复模块 (image reconstruction). 这种结构和已有的一些超分辨率结构一样 (SwinIR、HAT). 具体来说, 对于低分辨率输入图像$I_{\mathrm{LR}} \in R^{H\times W\times C}$, 首先, 用卷积神经网络提取浅层特征$F_S \in R^{H\times W\times D}$.

$$F_S = H_{\mathrm{SFE}}(I_{\mathrm{LR}}) = W_S * I_{\mathrm{LR}} \tag{1}$$

其中, C, D, H_{SFE}分别表示输入特征维度、浅层特征模块输出维度和具有权重为W_S的卷积神经网络.

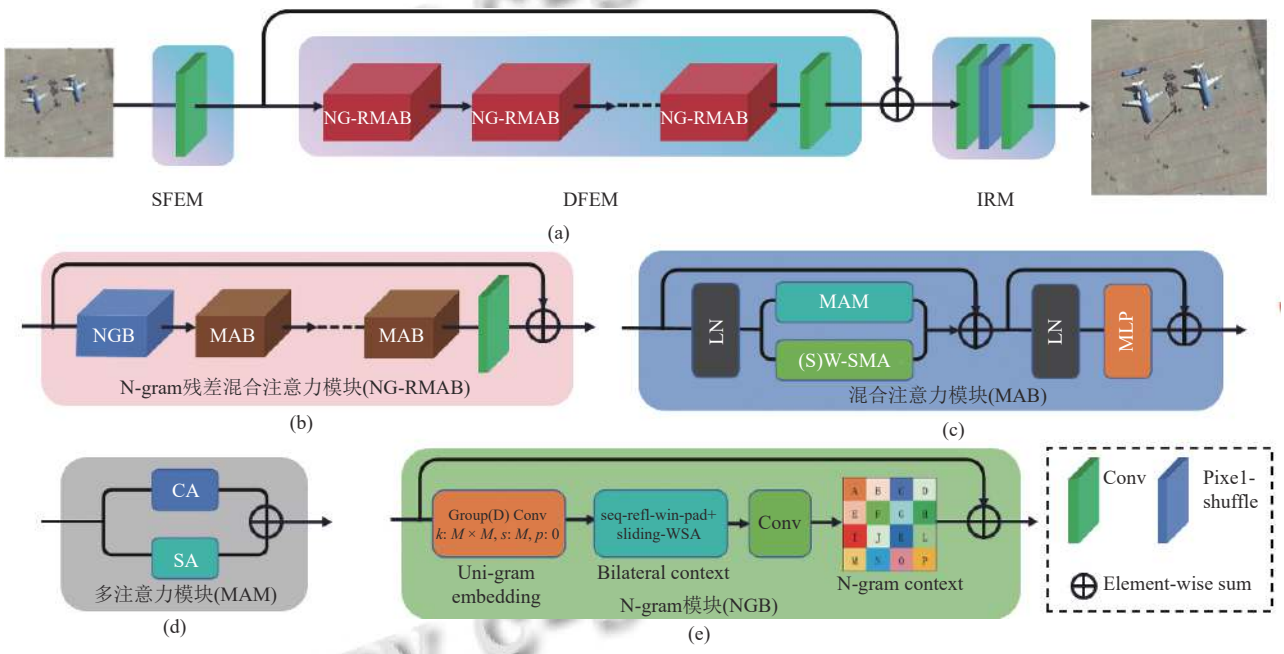

图 2 NG-MAT 模型整体架构

然后将提取到的浅层特征F_S输入到由几个 NG-RMAB 模块和卷积神经网络构成的模块中提取深层特征$F_D \in R^{H\times W\times D}$, 这个过程可以表示为式 (2):

$$F_D = H_{\mathrm{DFE}}(F_S) = H_{\mathrm{NG\text{-}RMAB}}(F_S)_{\times n} \tag{2}$$

上述两步操作之后, 使用残差连接融合F_S和F_D. 最后将得到的特征输入图像恢复模块得到高分辨率图像.

$$I_{\mathrm{SR}} = H_{\mathrm{IR}}(F_S + F_D) \tag{3}$$

每个 NG-RMAB 模块由一个 N-gram 模块 (NGB) 和几个混合注意力模块 (MAB) 构成. 图像恢复模块中使用 Pixel-shuffle 方法来放大特征的尺寸, 使用卷积神经网络来改变输入输出特征的维度.

本文采用 L1 loss 来优化模型参数, 训练的目标是最小化损失函数L. I_{SR}、I_{GT}分别为表示重建的高分辨率图像和真实高分辨率图像, i 表示第 i 组图像对.

$$L = \|I_{\mathrm{SR}} - I_{\mathrm{GT}}\|_1 = \frac{1}{N}\left\|I^i_{\mathrm{SR}} - I^i_{\mathrm{GT}}\right\|_1 \tag{4}$$

2.2 N-gram 模块 (NGB)

NLP 中的 N-gram 模型: N-gram 在 NLP 领域内是

一种语言统计模型, 即一个序列中下一个位置出现某个词的概率只取决于固定窗口中已经确定的词. 常见的 N-gram 模型有 3 种: 一元模型 (unigram model)、二元模型 (bigram model) 和三元模型 (trigram model), 主要的区别在于可见窗口的不同, 其中二元模型有前向和反向两种, 本文以文本“这本书非常有用”来解释 3 种模型的区别, 如表 1 所示.

表 1 NLP 中的 N-gram 模型

模型	示例
一元模型	(这), (本), (书), (非), (常), (有), (用)
二元模型	前向: (这本), (本书), (书非), (非常), (常有), (有用) 反向: (这本), (本书), (书非), (非常), (常有), (有用)
三元模型	(这本书), (本书非), (书非常), (非常有), (常有用)

图像处理中的 N-gram 模型: 图像中的 N-gram 应该具有和 NLP 中相同的意义. 因此可以将一个 gram 定义为 Swin Transformer 里不重叠的小窗口, N-gram 则表示由 $N\times N$ 个 gram 组成的区域, 其中的一个 gram 相当于 NLP 中那个待确定的词, 其余的 gram 相当于 NLP 中已经确定的词. 当 N=3 时, 图像中的双向 N-gram 如图 1 所示.

本文的 N-gram (N=3) 模块计算过程如图 2(e) 所示. 首先使用卷积核大小为 $M\times M$、步长为 M 的群卷积将输入特征转换成单一 gram 嵌入 (uni-gram embedding). 然后采用 seq-refl-win-pad 的填充方式填充窗口的上下左右区域, 使用移动窗口自注意力提取双向 N-gram 内容, 这里的双向指垂直方向和水平方向, 再用 3×3 的卷积改变特征维度和特征图的数量得到最终的 N-gram 内容. 最后使用残差连接的方式将 N-gram 内容融合到输入特征中, 即 N-gram 内容中的某个值被同等添加到同一位置 (被相同的字符标记) 的 M^2 个像素. 这个过程使得 Swin Transformer 在窗口内计算自注意力时, 能考虑到相邻窗口内的一部分特征信息, 使最终的超分辨率重建图像视觉效果更好.

2.3 混合注意力模块 (MAB)

Swin Transformer 的出现解决了 ViT 计算复杂度高的问题, 这种方法是将输入图像划分为若干个局部窗口, 然后在这些局部窗口内计算自注意力, 这说明 Transformer 提取的是窗口的局部信息, 而各个窗口之间缺乏信息交互. 为了解决这个问题, 本文提出了图 2(c) 的结构, 在原始 Swin Transformer 计算自注意力的分支上并联一个 MAM (在第 2.4 节中介绍) 模块来增加各个窗口之间的交互, 将其称为 mixed attention block (MAB). 和原始 Swin Transformer 一样, 本文的模型交替计算 window-based multi-head self-attention (W-MSA) 和 shifted window-based multi-head self-attention (SW-MSA). 为了避免各个窗口之间的交互信息过度影响窗口内的自注意力, 使用参数 θ 来控制其大小. 这个过程可以用式 (5) 表示.

$$\begin{cases} X_{\mathrm{SA}} = H_{\mathrm{(S)W\text{-}MSA}}(LN(X_i)) \\ X_{\mathrm{MAM}} = H_{\mathrm{MAM}}(LN(X_i)) \\ X_1 = X_{\mathrm{SA}} + \theta X_{\mathrm{MAM}} + X_i \\ X_o = H_{\mathrm{MLP}}(LN(X_1)) + X_1 \end{cases} \tag{5}$$

其中, $X_1, X_{\mathrm{SA}}, X_{\mathrm{MAM}}$ 为中间特征信息; X_i 和 X_o 分别为 MAB 的输入、输出特征; $H_{\mathrm{MLP}}(\cdot)$ 表示多层感知机, 是 Swin Transformer 的基本组成部分.

Swin Transformer 中基于窗口的自注意力计算过程如下: 对于大小为 $H\times W\times D$ 的输入图像, 首先通过窗口划分将其形状变为 $\frac{HW}{M^2}\times M^2\times D$, 输入图像被划分为 $\frac{HW}{M^2}$ 个大小为 $M\times M$ 窗口; 然后在每一个窗口内计算自注意力, 对于局部窗口内的特征 $X_{\mathrm{LW}}\in R^{M^2\times D}$, Swin Transformer 会同时计算 h 次自注意力, h 为自注意力头的个数, 因此每次计算的特征为 $X_h\in R^{M^2\times d}$, 其中 $d=\frac{D}{h}$.

局部特征 $X_h\in R^{M^2\times d}$ 的 Q、K、V 矩阵可由式 (6) 计算得出.

$$Q = W_q * X_h, K = W_k * X_h, V = W_v * X_h \tag{6}$$

为了减少内存的消耗, 不同的局部窗口采用共享的映射矩阵 W_q, W_k, W_v, 基于窗口的自注意力 $Atten(Q,K,V)$ 可由式 (7) 计算.

$$Atten(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \tag{7}$$

其中, $B\in R^{M^2\times M^2}$ 为可学习的相对位置偏差, 可以依据文献[22]计算得出.

2.4 多注意力模块 (MAM)

MAM 是本文提出的一种多注意力结构, 由通道注意力和空间注意力构成, 它从两个不同的维度提取输入图像的全局特征信息, MAM 模块结构如图 2(d) 所示. 通道注意力和空间注意力的计算如图 3 所示.

通道注意力: 对于输入特征 $F_{\mathrm{in}}\in R^{H\times W\times D}$, 首先在空间维度上用最大池化和平均池化提取两类不同的特

征, 然后将提取的特征经共享 MLP 层处理之后得到特征 $F^c_{\max} \in R^{1\times1\times D}$ 和 $F^c_{\text{avg}} \in R^{1\times1\times D}$, 最后经过特征融合和 element-wise product 操作将这个特征映射到输入特征的维度, 得到通道注意力 $F_{\text{CA}} \in R^{H\times W\times D}$.

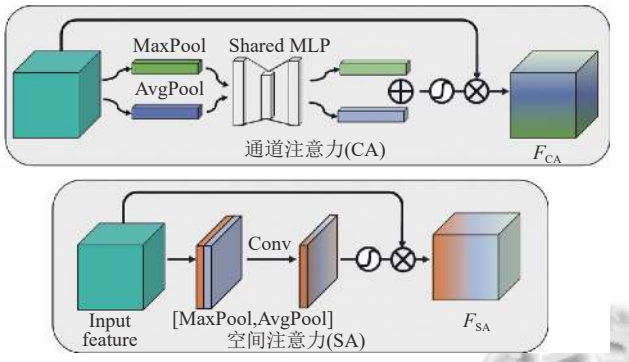

图 3 通道注意力和空间注意力结构

空间注意力: 对于输入特征 $F_{\text{in}} \in R^{H\times W\times D}$, 首先使用最大池化和平均池化操作来聚合特征图的通道信息, 可以得到两类特征 $F^s_{\max} \in R^{H\times W\times1}$ 和 $F^s_{\text{avg}} \in R^{H\times W\times1}$, 然后用 3×3 卷积层融合两类特征, 最后经过 element-wise product 操作将这个特征映射到输入特征的维度, 得到空间注意力 $F_{\text{SA}} \in R^{H\times W\times D}$. 故 MAM 模块可以用式 (8) 来描述.

$$\begin{cases} F_{\text{CA}} = (F^c_{\max} + F^c_{\text{avg}}) \otimes F_{\text{in}} \\ F_{\text{SA}} = H_{\text{Conv}}([F^s_{\max}, F^s_{\text{avg}}]) \otimes F_{\text{in}} \\ F_{\text{MAM}} = F_{\text{CA}} + F_{\text{SA}} \end{cases} \tag{8}$$

3 实验分析

3.1 数据集和实验设置

数据集: 本文实验选择 AID[23]和 NWPU-RESISC45[24] 两个数据集作为训练集和测试集, 它们在目标检测和图像超分辨率等任务中被广泛使用. AID 数据集包含 30 个场景的遥感图像, 每个场景约 200–420 张, 一共有 10 000 张图像, 每张大小为 600×600. 为了充分利用图像的信息, 将 AID 数据集中的图像有重叠的裁剪为大小为 264×264 的子图像用作训练集. NWPU-RESISC45 数据集包含 45 个场景的遥感图像, 每个场景 700 张, 一共有 31 500 张图像, 每张大小为 256×256. 将每个类中图像名称序号后缀为 28、69、55、60、74、97 抽取出来, 一共 1 890 张图像, 一半作为验证集, 一半作为测试集; 剩下的 29 160 张图像作为训练集.

实验设置: 本文提出的模型 NG-MAT 在 NVIDIA GeForce RTX 4080, 显存为 16 GB 的设备上训练. 与 SwinIR 一样, 我们将模型的 NG-RMAB 模块和 MAB 模块的个数设置为 6, 在窗口内计算多头注意力时的 h 设置为 6, 窗口大小设置为 8. MAB 中的权重参数 θ 设置为 0.01, N-gram 模块中的 N 设置为 3. 模型放大因子 upscale 有 3 个: 2/3/4 分别表示将图像放大 4 倍、9 倍、16 倍. 学习率设置为 2E–4, 采用 Adam ($\beta_1 = 0.9, \beta_2 = 0.99$) 优化器更新梯度. batch_size 设置为 4, 一共训练 250 000 个 iteration, 分别在 iteration 为 100 000、150 000、200 000、2 250 000 时将学习率减小一半.

3.2 消融实验

本节所有的实验都是在 upscale 为 2 的条件下进行, 同时 PSNR/SSIM 值为验证集的推理结果.

不同 θ 对实验结果的影响: 我们在第 2.3 节中提到, 用 θ 来控制 MAM 模块的权重, 进行特征融合. 为了探索不同的 θ 权重对模型的影响, 本文取 θ 值为 {0, 0.01, 0.1, 1} 进行实验, 除 θ 值不同外, 其他实验设置与第 3.1 节中一样, 实验结果如表 2 所示. θ 越大表示 MAM 模块提取的特征占比越大, $\theta = 0$ 表示不采用 MAM 模块. 由结果可知, 加入 MAM 模块能够提高模型重建效率, 并且 θ 为 0.01 时, 效果最好. 这说明适当的全局特征能够提高模型效率, 加强窗口之间的交互, 但是当权重太大时, 模型效率会逐渐降低.

表 2 不同权重因子 θ 对模型的影响

指标	0	0.01	0.1	1
PSNR (dB)	34.65	34.75	34.70	34.68
SSIM	0.928 0	0.929 4	0.928 7	0.928 5

NG-RMAB 模块数量对模型的影响: 在本文提出的 NG-MAT 模型中, 最重要的是深度特征提取部分中的 NG-RMAB 模块, 这一部分直接决定了重建图像是否具有丰富的细节和纹理特征. 根据深度学习领域的经验, NG-RMAB 模块的个数越多, 效果可能会越好, 但是这样会导致网络参数量增加, 增加了模型训练和应用的难度, 因此需要找到一个折中的选择. 选取 NG-RMAB 模块的个数为 {3, 4, 5, 6, 7} 进行实验, 除了 NG-RMAB 模块的个数外, 其他实验设置与第 3.1 节中相同, 分别对比 PSNR、SSIM、训练时间、模型参数量 4 个指标, 实验结果如图 4 所示. 一共训练 50 000 个 iteration, PSNR、SSIM 为验证集的推理结果.

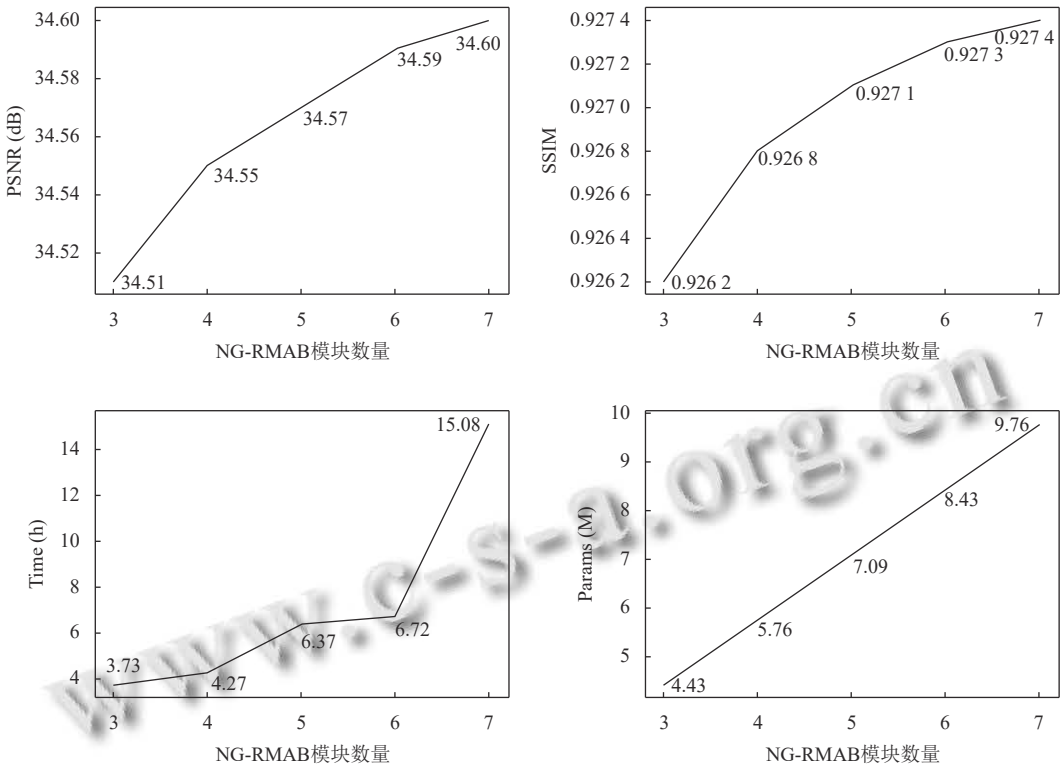

图 4 不同 NG-RMAB 数量对模型效率的影响

分析实验结果可知, PSNR、SSIM、参数量 Params 基本与 NG-RMAB 的数量呈线性关系. 训练时间与 NG-RMAB 的数量则有明显的差异, NG-RMAB 个数为 6 时训练时间为 6.72 h, 但 NG-RMAB 个数为 7 时, 训练时间急剧增加到 15.08 h.

综上可知, NG-RMAB 个数为 6 时, 能够实现模型效率的最大化, 同时尽可能减少模型的训练时间和参数量.

NGB 和 MAM 模块对模型的影响: 本文设计了实验来验证所提出模块 NGB 和 MAM 的有效性. 除了是否加入 NGB 模块或 MAM 模块外, 其他实验设置与第 3.1 节中保持一致. 实验结果如表 3 所示. 由实验结果可知, 分别加入 NGB 模块、MAM 模块都提高了 PSNR/SSIM 值, 当两个模块同时都加入时, PSNR 和 SSIM 分别提高了 0.19 dB 和 0.002 5. 因此, 实验结果证明了本文提出模块的有效性.

3.3 与其他方法的比较

我们将 NG-MAT 与其他遥感图像超分辨率模型进行了比较, 包括 bicubic interpolation (双三次插值)、SRCNN[6]、EDSR[7]、VDSR[8]、SRGAN[10]、RCAN[9]、SwinIR[15], 其中 bicubic interpolation 是传统超分辨率方法, SRCNN、EDSR、VDSR、RCAN 是基于 CNN 的方法, SRGAN 是基于生成对抗网络的方法, SwinIR 是基于 Transformer 的方法, 它们代表了超分辨率算法发展的历程. 为了保证公平性, 每个模型的训练环境及训练数据集均保持一致.

表 3 NGB 和 MAM 模块的有效性

MAM	NGB	PSNR (dB)	SSIM
×	×	34.56	0.926 9
×	√	34.65	0.928 0
√	×	34.67	0.928 3
√	√	34.75	0.929 4

定量比较: 表 4–表 6 为不同放大因子时, 各个模型在不同场景的遥感图像测试集上的定量对比结果. 由于篇幅的限制, 每个表格中只列出了 9 个场景的测试结果, 除此之外, 最后的两行平均值 (9) 和平均值 (45) 分别表示表格中前 9 个场景的平均测试结果和整个测试集 45 个场景的平均测试结果. 为了尽量测试所有的场景, 不同的放大因子时选取不同的场景进行测试, 测试最好的结果用加粗表示.

表 4 在 NWPU-RESISC45 数据集上 NG-MAT 与 7 种方法的定量比较 (upscale=2)

场景	指标	bicubic interpolation	SRCNN	EDSR	VDSR	SRGAN	RCAN	SwinIR	NG-MAT
教堂	PSNR (dB)	23.81	27.16	27.28	27.35	28.34	29.15	30.31	**30.55**
	SSIM	0.6952	0.8642	0.8764	0.8834	0.8831	**0.8973**	0.8865	0.8911
商业区	PSNR (dB)	25.37	27.75	27.53	28.05	29.37	29.61	32.74	**32.97**
	SSIM	0.7653	0.8870	0.8886	0.9041	0.9034	0.8912	0.9261	**0.9294**
高速公路	PSNR (dB)	27.42	29.46	29.97	29.81	30.64	31.95	34.12	**34.40**
	SSIM	0.7409	0.8648	0.8832	0.8803	0.8789	0.8961	0.9092	**0.9133**
田径场	PSNR (dB)	26.80	29.11	28.75	29.46	30.81	30.44	33.35	**33.54**
	SSIM	0.7884	0.9037	0.8989	0.9173	0.9164	0.9018	0.9273	**0.9296**
湖泊	PSNR (dB)	31.89	32.43	32.28	35.30	32.62	33.81	38.14	**38.21**
	SSIM	0.8398	0.9148	0.9146	0.9204	0.9093	0.9157	0.9448	**0.9455**
中型住宅	PSNR (dB)	25.58	27.98	28.09	28.27	29.42	29.58	31.58	**31.75**
	SSIM	0.7237	0.8737	0.8759	0.8904	0.8825	0.8779	0.8947	**0.8971**
闲置住宅	PSNR (dB)	28.09	29.52	29.31	29.73	30.57	31.59	33.20	**33.35**
	SSIM	0.7077	0.8629	0.8657	0.8740	0.8586	0.8717	0.8746	**0.8790**
阳台	PSNR (dB)	30.13	31.07	31.01	31.41	32.95	33.23	37.05	**37.20**
	SSIM	0.8088	0.8995	0.9014	0.9107	0.9156	0.9107	0.9399	**0.9418**
湿地	PSNR (dB)	30.03	31.74	31.68	31.74	32.63	34.02	36.14	**36.21**
	SSIM	0.7859	0.8987	0.8985	0.8997	0.9001	0.9086	0.9241	**0.9249**
平均值 (9)	PSNR (dB)	27.68	29.61	29.57	29.76	30.78	31.48	34.07	**34.24**
	SSIM	0.7618	0.8850	0.8897	0.8982	0.8931	0.8964	0.9141	**0.9169**
平均值 (45)	PSNR (dB)	27.54	29.71	29.52	29.97	30.52	31.28	34.49	**34.68**
	SSIM	0.7801	0.8943	0.8943	0.9055	0.9046	0.9040	0.9239	**0.9266**

表 5 在 NWPU-RESISC45 数据集上 NG-MAT 与 7 种方法的定量比较 (upscale=3)

场景	指标	bicubic interpolation	SRCNN	EDSR	VDSR	SRGAN	RCAN	SwinIR	NG-MAT
机场	PSNR (dB)	24.83	24.90	28.66	29.24	28.68	28.82	30.56	**30.76**
	SSIM	0.6602	0.7298	0.8212	0.8278	0.8204	0.8262	0.8476	**0.8528**
桥梁	PSNR (dB)	27.57	28.26	31.24	31.41	31.25	31.12	33.60	**33.82**
	SSIM	0.7894	0.8317	0.8834	0.8883	0.8840	0.8879	0.9080	**0.9122**
丛林	PSNR (dB)	22.95	21.79	25.74	26.06	25.72	25.78	27.36	**27.46**
	SSIM	0.5325	0.6832	0.7158	0.7194	0.7134	0.7192	0.7392	**0.7460**
云	PSNR (dB)	29.94	31.10	35.99	35.53	35.99	36.05	37.77	**37.84**
	SSIM	0.8285	0.8714	0.9150	0.9069	0.9148	0.9159	0.9292	**0.9302**
沙漠	PSNR (dB)	31.47	30.32	35.36	35.54	35.37	35.42	37.04	**37.11**
	SSIM	0.7838	0.8402	0.8832	0.8879	0.8831	0.8839	0.9015	**0.9026**
森林	PSNR (dB)	26.46	26.78	28.75	28.83	28.76	28.87	30.35	**30.40**
	SSIM	0.5757	0.7042	0.7310	0.7314	0.7310	0.7371	0.7657	**0.7684**
港口	PSNR (dB)	20.96	21.33	23.56	23.92	23.62	23.76	25.76	**26.13**
	SSIM	0.6754	0.7595	0.8134	0.8164	0.8184	0.8214	0.8430	**0.8555**
工业区	PSNR (dB)	23.99	24.74	28.67	28.43	28.63	28.91	30.67	**31.00**
	SSIM	0.6628	0.7350	0.8371	0.8365	0.8360	0.8445	0.8633	**0.8715**
道路	PSNR (dB)	27.84	28.19	31.87	31.97	31.69	32.14	34.32	**35.11**
	SSIM	0.7557	0.7708	0.8593	0.8607	0.8571	0.8653	0.8859	**0.8953**
平均值 (9)	PSNR (dB)	26.22	27.16	29.98	30.17	29.97	30.09	31.94	**32.18**
	SSIM	0.6960	0.7707	0.8288	0.8311	0.8283	0.8335	0.8537	**0.8594**
平均值 (45)	PSNR (dB)	25.32	25.24	28.89	28.98	28.88	29.03	30.80	**31.03**
	SSIM	0.6645	0.7452	0.8092	0.8132	0.8086	0.8146	0.8379	**0.8444**

从测试结果可以看出, 传统方法 bicubic interpolation 的性能总是比深度学习方法差, 本文提出的 NG-MAT 模型几乎在各个场景都取得了最好的结果. 具体来说, 当放大因子 upscale=2 时, NG-MAT 在除了教堂场景之外的其他场景都取得了最好的结果; 在 9 个场景的平均测试中, PSNR、SSIM 值分别达到了 34.24 dB, 0.916 9, 相较于 SwinIR 模型提高了 0.17 dB 和 0.002 8; 在 45 个场景的测试中, PSNR 和 SSIM 也都取得了最好的结果. 当放大因子 upscale=3 时, NG-MAT 在各个场景都是表现最好的; 在整个测试集上的平均 PSNR

为 31.03 dB, 平均 SSIM 为 0.844 4, 分别比第 2 最佳值高了 0.23 dB 和 0.006 5. 当放大因子 upscale=4 时, 可以发现各个方法的性能均显著性下降, 但是 NG-MAT 在整个测试集上的性能仍然是最好的, 平均 PSNR 和 SSIM 分别为 28.99 dB 和 0.773 4. 总的来说, NG-MAT 在各种放大因子时都取得了比其他模型更好的效果, 测试结果验证了模型的有效性.

视觉效果对比: 图 5–图 7 给出了放大因子为 2、3、4 时, NG-MAT 与其他基础模型超分辨率重建的视觉对比. 图 5–图 7 中对应数据集中的图像分别为中型住宅 655、道路 497、环形路口 369. 我们可以观察到, 与其他基础模型相比, NG-MAT 对于图像纹理和细节的重建效果更接近真实图像, 甚至视觉效果更好, 例如当放大因子 upscale=2 时, 本文提出的模型重建的图像与真实图像相比噪声更少, 纹理细节丰富. 但是当放大因子变大时, 各种超分辨率模型的重建效果均会变差, 主要是因为原始低分辨率图像中含有的高频信息太少, 难以达到真实图像的视觉效果, 但是 NG-MAT 重建图像的视觉效果仍然优于其他超分辨率方法. 视觉效果的比较进一步验证了本文提出的 NG-MAT 的有效性.

表 6 在 NWPU-RESISC45 数据集上 NG-MAT 与 7 种方法的定量比较 (upscale=4)

场景	指标	bicubic interpolation	SRCNN	EDSR	VDSR	SRGAN	RCAN	SwinIR	NG-MAT
棒球场	PSNR (dB)	25.45	28.37	26.51	27.12	27.51	28.44	30.56	**30.77**
	SSIM	0.668 1	0.760 8	0.766 5	0.774 6	0.794 3	0.794 2	0.821 9	**0.828 0**
篮球场	PSNR (dB)	23.04	26.14	25.12	26.23	25.41	26.48	27.45	**27.83**
	SSIM	0.530 7	0.677 4	0.681 2	0.693 2	0.703 6	0.704 2	0.720 7	**0.738 7**
海滩	PSNR (dB)	24.53	27.62	27.42	27.34	27.54	27.92	28.62	**28.67**
	SSIM	0.610 2	0.759 6	0.783 8	0.768 6	0.784 2	**0.785 2**	0.742 1	0.744 3
环形农田	PSNR (dB)	26.95	29.68	29.43	30.07	30.31	29.76	32.42	**32.70**
	SSIM	0.716 2	0.816 2	0.809 7	0.835 4	0.842 1	0.776 2	0.856 8	**0.863 2**
路口	PSNR (dB)	20.16	22.86	23.27	23.29	23.08	23.58	24.73	**25.37**
	SSIM	0.464 6	0.675 7	0.668 6	0.697 6	0.682 6	0.689 1	0.709 7	**0.732 3**
草地	PSNR (dB)	29.82	30.47	30.76	30.43	31.14	31.07	33.03	**33.09**
	SSIM	0.631 1	0.706 2	0.689 7	0.707 4	0.742 3	0.731 8	0.759 2	**0.760 9**
河流	PSNR (dB)	25.52	28.36	28.63	28.44	28.93	28.45	30.45	**30.52**
	SSIM	0.604 0	0.759 4	0.773 4	0.754 1	0.768 4	**0.794 7**	0.785 7	0.789 0
环形路口	PSNR (dB)	21.83	24.21	24.36	23.89	24.48	24.45	26.42	**26.68**
	SSIM	0.517 8	0.696 5	0.703 2	0.701 6	0.720 5	**0.756 1**	0.725 3	0.737 5
海冰	PSNR (dB)	24.26	27.24	27.14	27.02	27.16	27.52	30.60	**30.71**
	SSIM	0.666 1	0.795 4	0.801 8	0.798 7	0.798 5	0.806 8	0.829 2	**0.831 9**
平均值 (9)	PSNR (dB)	25.62	27.07	26.87	27.16	27.31	27.34	29.36	**29.59**
	SSIM	0.601 1	0.744 3	0.741 5	0.749 6	0.761 9	0.759 4	0.772 3	**0.780 6**
平均值 (45)	PSNR (dB)	23.74	26.12	26.08	26.16	26.43	26.05	28.76	**28.99**
	SSIM	0.578 7	0.738 7	0.732 4	0.743 5	0.754 4	0.750 7	0.764 0	**0.773 4**

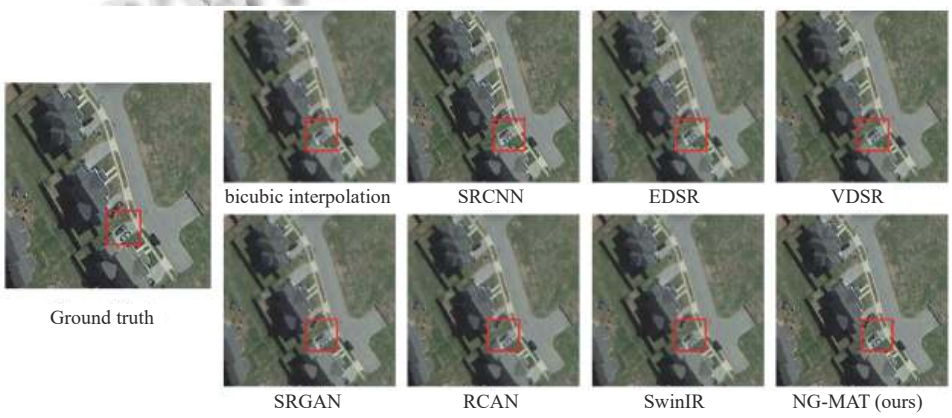

图 5 在 NWPU-RESISC45 数据集上 NG-MAT 与其他 SR 方法的视觉对比 (upscale=2)

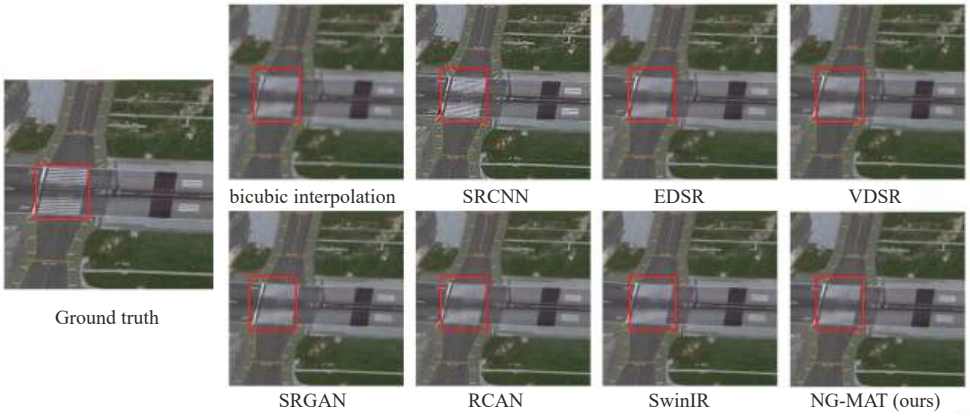

图 6 在 NWPU-RESISC45 数据集上 NG-MAT 与其他 SR 方法的视觉对比 (upscale=3)

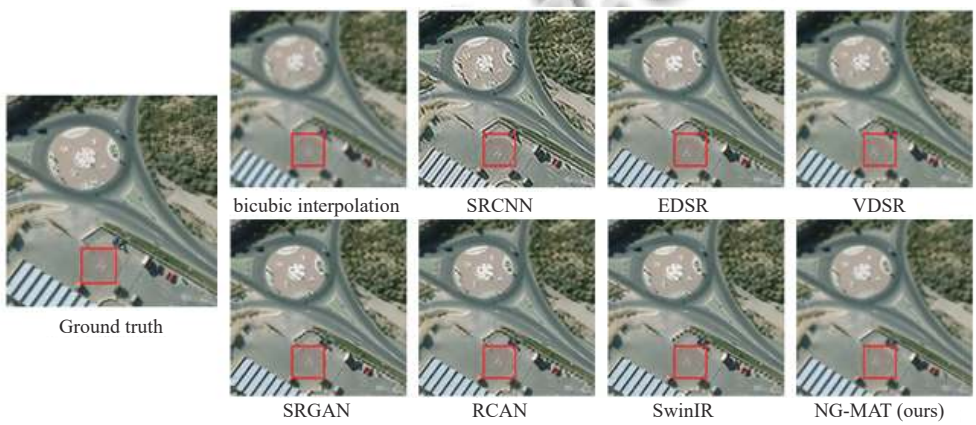

图 7 在 NWPU-RESISC45 数据集上 NG-MAT 与其他 SR 方法的视觉对比 (upscale=4)

4 结论

本文提出了一种基于 Swin Transformer 和 N-gram 模型的遥感图像超分辨率重建网络. 由于不同场景的遥感图像之间差异较大, 而卷积神经网络又不能很好地提取图像的全局特征信息, 因此本文采用 Transformer 作为超分辨率重建的基准模型, 同时引入 N-gram 模型来加强 Transformer 窗口之间的信息交互. 针对现有模型感受野有限的问题, 本文结合通道注意力、空间注意力和自注意力, 从 3 个不同维度提取图像的特征信息, 激活更多像素. 本文模型在 AID 和 NWPU-RESISC45 两个遥遥感数据集上进行训练、验证和测试, 分别通过定量对比和视觉对比验证了模型的有效性. 重建图像纹理细节也更丰富, 能够兼顾多个场景的差异, 有效解决了现有模型存在的一些的问题.

参考文献

1 Ahmad W, Ali H, Shah Z, *et al*. A new generative adversarial network for medical images super resolution. Scientific Reports, 2022, 12(1): 9533. [doi: 10.1038/s41598-022-13658-4]

2 Qiu DF, Zheng LX, Zhu JQ, *et al*. Multiple improved residual networks for medical image super-resolution. Future Generation Computer Systems, 2021, 116: 200–208. [doi: 10.1016/j.future.2020.11.001]

3 Zhang DY, Shao J, Li XY, *et al*. Remote sensing image super-resolution via mixed high-order attention network. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(6): 5183–5196. [doi: 10.1109/TGRS.2020.3009918]

4 Jia S, Wang ZH, Li QQ, *et al*. Multiattention generative adversarial network for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5624715.

5 Rasti P, Uiboupin T, Escalera S, *et al*. Convolutional neural network super resolution for face recognition in surveillance monitoring. Proceedings of the 9th International Conference on Articulated Motion and Deformable Objects. Palma de

Mallorca: Springer International Publishing, 2016. 175–184.
6 Dong C, Loy CC, He KM, *et al*. Learning a deep convolutional network for image super-resolution. Proceedings of the 13th European Conference Computer Vision. Zurich: Springer, 2014. 184–199.
7 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. 136–144.
8 Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convolutional networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1646–1654.
9 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 (ECCV). Munich: Springer, 2018. 286–301.
10 Ledig C, Theis L, Huszár F, *et al*. Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 105–114.
11 Wang XT, Yu K, Wu SX, *et al*. ESRGAN: Enhanced super-resolution generative adversarial networks. Proceedings of the 2018 Conference on Computer Vision —ECCV 2018 Workshops. Munich: Springer, 2018. 63–79.
12 Zhang WL, Liu YH, Dong C, *et al*. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3096–3105.
13 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.
14 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.
15 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. Montreal: IEEE, 2021. 1833–1844.
16 Lu ZS, Li JC, Liu H, *et al*. Transformer for single image super-resolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 456–465.
17 Fang JS, Lin HJ, Chen XY, *et al*. A hybrid network of CNN and Transformer for lightweight image super-resolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Orleans: IEEE, 2022. 1102–1111.
18 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.
19 Haut JM, Fernandez-Beltran R, Paoletti ME, *et al*. A new deep generative network for unsupervised remote sensing single-image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(11): 6792–6810. [doi: 10.1109/TGRS.2018.2843525]
20 Ma W, Pan ZX, Yuan F, *et al*. Super-resolution of remote sensing images via a dense residual generative adversarial network. Remote Sensing, 2019, 11(21): 2578. [doi: 10.3390/rs11212578]
21 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv: 1701.07875, 2017.
22 Wu K, Peng HW, Chen MH, *et al*. Rethinking and improving relative position encoding for vision Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10013–10021.
23 Cheng G, Han JW, Lu XQ. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017, 105(10): 1865–1883. [doi: 10.1109/JPROC.2017.2675998]
24 Xia GS, Hu JW, Hu F, *et al*. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965–3981. [doi: 10.1109/TGRS.2017.2685945]

(校对责编: 孙君艳)