基于ViT-CNN特征增强的图像超分辨率

高志荣[a],孙清清[bc*],熊承义[bc],李帆[bc],郑瑞华[bc]
(中南民族大学 a.计算机科学学院; b.电子信息工程学院; c.智能无线通信湖北省重点实验室,武汉 430074)

摘 要 卷积神经网络(CNN)可以提取图像的局部相关特征,视觉Transformer(ViT)则侧重于捕获图像的远距离依赖关系,二者有效结合能够改进图像的重构质量.研究了一种基于ViT-CNN特征增强的图像超分辨率(SR)网络.具体来说,网络包含了基于ViT的SR分支与基于CNN的梯度分支,SR分支主要用于提取图像特征域中的全局相关性,而梯度分支则专注于图像梯度域中的局部依赖关系.通过对两种信息的融合与渐进增强,获得高倍放大的重构图像.此外,在网络的学习阶段引入了梯度损失及渐进训练策略,有效降低了网络的训练难度并增强了训练的稳定性.在多个公开数据集上的大量实验结果验证了所提方法在改善重构系统性能方面的有效性.

关键词 图像超分辨率;卷积神经网络;视觉Transformer;特征融合
中图分类号 TP391.4 **文献标志码** A **文章编号** 1672-4321(2025)02-0253-07
doi:10.20056/j.cnki.ZNMDZK.20250215

Image super-resolution based on feature enhancement with ViT-CNN

GAO Zhirong[a],SUN Qingqing[bc*],XIONG Chengyi[bc],LI Fan[bc],ZHENG Ruihua[bc]
(South-Central Minzu University, a. College of Computer Science; b. College of Electronic and Information Engineering; c. Hubei Key Lab of Intelligent Wireless Communication, Wuhan 430074, China)

Abstract The effective combination of Convolution Neural Network (CNN) which extract the local correlation features of images and Vision Transformer (ViT) which focuses on capturing the remote dependence of images can improve the quality of image reconstruction. A network of image super-resolution based on feature enhancement with ViT-CNN is studied. Specifically, the network includes ViT-based SR branch and CNN-based gradient branch, which extract the global correlation in the image feature domain and the local dependency in the image gradient domain respectively. Through the fusion and gradual enhancement of the two kinds of information, the reconstructed image with large factor is obtained. In addition, by introducing gradient loss and progressive training strategy, the difficulty of training is effectively reduced and the stability of training is enhanced. A large number of experimental results on multiple public datasets demonstrate the effectiveness of the proposed method in improving the performance of the reconstruction system.

Keywords image super resolution; Convolution Neural Network; Vision Transformer; feature fusion

图像超分辨率(Super-Resolution,SR)是一种图像处理技术,旨在从退化的低分辨率图像(Low Resolution, LR)中恢复出高分辨率图像(High Resolution,HR).在许多领域中,通过高效且节省成本的SR技术来获得HR图像是很有必要的,例如视频监控[1]、医疗成像[2]和卫星成像[3]等.

近年来,由于深度学习(Deep Learning,DL)和卷积神经网络(Convolutional Neural Network,CNN)表现出的强大学习能力,各类基于CNN的图像SR方法被大量提出[4-10].虽然这类方法相比较传统方法在重构性能的提升上显现出了巨大优势,但由于图像和卷积核之间的操作通常与图像内容无关,使得基于固定卷积核运算的特征提取不能自适应于不同的图像.此外,受卷积核大小的约束,基于CNN的方法难以捕获图像中的远距离依赖关系,从而严重限制了其特征表达能力以及在图像重构方面的性能.最近,一种基于自注意力机制的Transformer结构[11]被提出并在自然语言处理领域中表现突出.受

收稿日期 2022-11-25 **通信作者** 孙清清,研究方向:图像超分辨率,E-mail:290955543@qq.com
作者简介 高志荣(1972-),女,副教授,博士,研究方向:图像处理与机器视觉,E-mail:gaozhirong@mail.scuec.edu.cn
基金项目 多谱信息处理技术国家重点实验室基金资助项目(6142113210303);中央高校基本科研业务专项资金资助项目(CZY21013)

此启发,DOSOVITSKIY等[12]提出了用于图像处理领域的视觉Transformer(Vision Transformer, ViT),并成为了Transformer在计算机视觉领域应用的里程碑式工作.当前,ViT在SR领域中也得到了极大的关注[13-16],其利用Transformer的自注意力机制,对图像中存在的远距离依赖关系进行建模,有效捕获了图像中的全局相关性,为进一步提升SR性能注入了新的活力.实际上,纯粹的CNN或ViT结构均有其对应的优势与缺陷,前者侧重于捕获图像的局部依赖关系,而后者则更多地关注图像的全局相关性.因此,如何有效地将两者结合并发挥各自优势成为了当前SR领域研究的热点问题.

基于上述背景,本文研究了一种基于ViT-CNN特征增强的图像超分辨率网络.具体来说,本文构建了多级的双分支网络,包括基于ViT的SR分支与基于CNN的梯度分支.SR分支主要用于提取图像特征域中的全局相关性,而梯度分支则专注于图像梯度域中的局部依赖关系.网络还包含了融合特征信息和梯度信息的融合模块,利用梯度分支生成的HR梯度图来指导图像重构过程,更好地保留图像的结构信息,消除伪影缓解结构变形,从而生成视觉效果良好的高倍放大图像.此外,本文还采用了渐进式训练策略以降低高倍放大任务的训练难度,提升训练过程的稳定性.实验结果表明了本文方法在改善重构系统性能方面的有效性.

1 相关工作

1.1 图像超分辨率

近年来,基于深度学习的图像超分辨率方法在计算机视觉领域中取得了显著进展.DONG等[4]首次将卷积神经网络用于SR任务,提出了一个利用3层卷积实现LR和HR图像对之间非线性映射的超分辨率卷积神经网络(SRCNN).KIM等[5]提出了极深的超分辨率网络(VDSR),堆叠了20层卷积并使用了跳转连接,进一步改善了SR的重构性能.之后,各类网络通过加深网络深度和设计精巧的网络结构来提高SR的性能.为了进一步改善重构图像的视觉效果,MA等[17]提出了一种基于梯度指导的结构保留超分辨率方法(SPSR),引入了图像的梯度信息,并用来指导SR的重构过程,改善了重构图像结构失真的问题.对于高倍放大的超分辨率任务,直接地上采样操作一般难以获得满意的效果,而且存在网络训练难以稳定等问题,因此渐进超分辨率策略被广为采用.LAI等[6]提出了一种渐进式SR方法,上采样遵循拉普拉斯金字塔原理,通过对输入图像每次执行2倍上采样,逐步重建出高倍放大的重构图像.此外,LAI等[8]进一步改进了他们的方法,采用深度更广的递归结构和多尺度训练.然而在高倍放大任务中,输入图像的特征信息几乎完全丢失,且CNN难以捕获图像中的远距离依赖关系和较弱的纹理细节,使得网络很难有效地恢复出视觉效果良好的重构图像.

1.2 视觉Transformer

最近,Transformer在自然语言处理(Natural Language Processing, NLP)领域中受到了广泛的关注,其核心为自注意力机制,能够有效地捕获句中单词之间的全局相关性.Transformer在NLP领域的突破引发了计算机视觉领域学者的极大兴趣.DOSOVITSKIY等[11]提出了用于图像分类任务的Vision Transformer(VIT),将切块后的图像转换为序列的形式以适应Transformer的输入,取得了可观的效果.CHEN等[13]提出了一种适用于低级视觉任务的通用Transformer预训练模型(IPT),对于不同任务连接不同的尾部模型,在图像超分辨率、去噪、去雾等任务中均取得了不错的成绩.不同于文本信息,图像的像素点多,全局自注意力的计算复杂度为像素点数量的平方,庞大的计算成本限制了其在视觉任务中的发展.为了解决该问题,LIU等[18]提出Swin Transformer(SwinT),使用不重叠的窗口来划分原始尺寸的特征图,只在每个窗口内执行区域自注意力计算,使得计算复杂度降低到了像素点数量的线性比例.然而,Transformer与CNN相比缺少归纳偏置(inductive bias),需要使用大规模的数据集来进行训练.此外,虽然Transformer能够高效地捕获全局信息,但在获取局部信息方面仍存在局限性,可能会阻碍图像中纹理细节的恢复.

2 提出的方法

本文提出了一种基于ViT-CNN特征增强的图像超分辨率,利用基于ViT的SR分支与基于CNN的梯度分支构建多级网络,逐步实现图像的高倍重构.其中,SR分支主要提取图像特征域中的全局相关性,梯度分支则专注于图像梯度域中的局部依赖关系.随后,对两种信息进行融合,利用HR梯度图来指导图像重构的过程,提升重构图像的质量.

2.1 网络结构

如图1所示,整体网络结构主要由三部分组成:浅层特征提取模块(Shallow Features Extraction Block,

SFEB）、SwinT-CNN 混合模块（SwinT-CNN Hybrid Block，STCHB）和重构模块（Reconstruction Block，RB）.

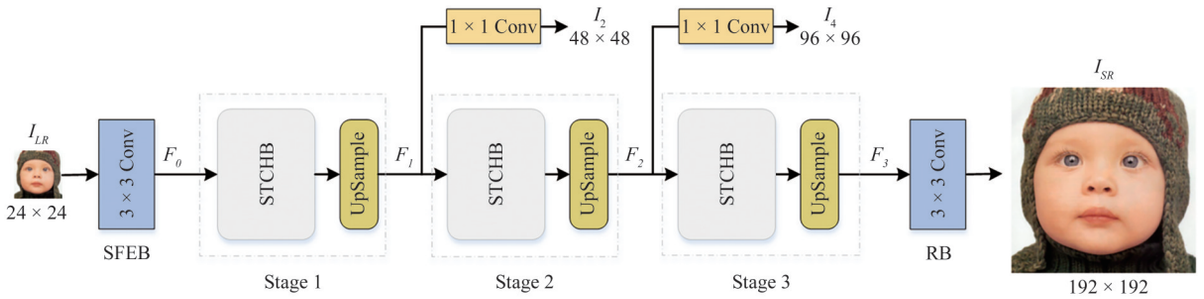


图 1　网络的整体结构框图

Fig. 1　Block diagram of the whole network structure

首先，浅层特征提取模块（SFEB）包含 1 个卷积核大小为 3 × 3，步长为 1，填充为 1 的卷积层，用于从给定的 LR 图像 $I_{\mathrm{LR}} \in \mathbb{R}^{H \times W \times 3}$（$H$ 为图像的高度，W 为图像的宽度，3 为图像的通道数）中提取浅层特征 $F_0 \in \mathbb{R}^{H \times W \times 64}$：

$$F_0 = H_{\mathrm{SFEB}}(I_{\mathrm{LR}}), \tag{1}$$

其中，$H_{\mathrm{SFEB}}(\cdot)$ 表示 SFEB.

随后，将网络分为三个阶段来实现 8 倍放大. 每个阶段中使用 SwinT-CNN 混合模块（STCHB）进行深度特征提取，再送入上采样层（UpSample）执行 2 倍上采样操作：

$$F_i = H^i_{\mathrm{UP}}\left(H^i_{\mathrm{STCHB}}\left(F_{i-1}\right)\right), \tag{2}$$

其中：$H^i_{\mathrm{STCHB}}(\cdot)$ 表示第 i 个 STCHB，$H^i_{\mathrm{UP}}(\cdot)$ 表示第 i 个上采样层，$F_i \in \mathbb{R}^{2^i H \times 2^i W \times 64}$ 表示第 i 个阶段的输出，i=1，2，3. 更多关于 STCHB 的细节将在第 2.2 节给出. 此外，在前两个阶段中，使用卷积核为 1 × 1，步长为 1，填充为 0 的卷积层，将上采样后得到的中间特征图转换为对应的 RGB 图像 $I_2 \in \mathbb{R}^{2H \times 2W \times 3}$ 和 $I_4 \in \mathbb{R}^{4H \times 4W \times 3}$：

$$I_2 = H_{\mathrm{conv1}}\left(F_1\right), I_4 = H_{\mathrm{conv2}}\left(F_2\right), \tag{3}$$

其中：$H_{\mathrm{convi}}(\cdot)$ 表示第 i 个 1 × 1 卷积层，i=1，2.

最后，重构模块（RB）包含了 1 个卷积核大小为 3 × 3，步长为 1，填充为 1 的卷积层，用来获得最终的重构图像 $I_{\mathrm{SR}} \in \mathbb{R}^{8H \times 8W \times 3}$：

$$I_{\mathrm{SR}} = H_{\mathrm{SR}}\left(I_{\mathrm{LR}}\right) = H_{\mathrm{RB}}\left(F_3\right), \tag{4}$$

其中：$H_{\mathrm{RB}}(\cdot)$ 表示 RB，$H_{\mathrm{SR}}(\cdot)$ 表示整个网络.

2.2　SwinT-CNN 混合模块

联合具有捕获全局特征能力的 ViT 和具有局部归纳特性的 CNN 可以提高网络的综合实力. 提出了 SwinT-CNN 混合模块（STCHB），如图 2（a）所示，通过对图像特征域中与图像梯度域中的信息的增强与融合，提升网络的重构质量.

图像的梯度图揭示了图像中需要突出关注的结构区域[17]. 为了增强重构图像中的纹理细节信息，利用梯度提取模块（Gradient Extraction Block，GEB）计算相邻像素的差值来获得输入图像 F_{i-1} 的梯度图 F_{g}：

$$F_{\mathrm{h}}(x,y) = F(x,y) - F(x-1,y), \tag{5}$$

$$F_{\mathrm{v}}(x,y) = F(x,y) - F(x,y-1), \tag{6}$$

$$F_{\mathrm{g}} = \nabla F = \sqrt{F_{\mathrm{h}}^2 + F_{\mathrm{v}}^2}, \tag{7}$$

其中：$F_{\mathrm{h}}(x,y)$ 和 $F_{\mathrm{v}}(x,y)$ 分别表示图像中坐标为 (x,y) 的像素点的水平方向和垂直方向的梯度值，∇ 表示计算图像的梯度.

由于梯度图中大部分区域的数值都接近于 0，只在轮廓边缘具有较大的数值，因此梯度分支（Gradient Branch）中使用更关注于局部依赖关系的 CNN 对梯度图进行增强. 使用 4 个 EDSR[7] 中提出的残差块（Residual Block，ResBlock）作为梯度分支的基础模块：

$$F_{\mathrm{c}} = H_{\mathrm{G}}\left(F_{\mathrm{g}}\right) = H^i_{\mathrm{Res}}\left(F_{\mathrm{g}}\right), \tag{8}$$

其中：$H_{\mathrm{G}}(\cdot)$ 表示梯度分支，$H^i_{\mathrm{Res}}(\cdot)$ 表示第 i 个残差块，F_{c} 表示 STCHB 模块内梯度分支的输出.

除了使用基于 CNN 的梯度分支在梯度域中捕获局部依赖关系之外，还使用了基于 ViT 的 SR 分支在图像特征域中考虑全局信息. 受 Swin Transformer 作用于视觉任务时能够极大降低模型计算复杂度的启发，使用 4 个 Swin Transformer 中的 Residual Swin Transformer Block（RSTB）作为 SR 分支的基本模块，用来捕获特征图中的全局依赖关系：

$$F_{\mathrm{t}} = H_{\mathrm{S}}\left(F_{i-1}\right) = H^i_{\mathrm{RSTB}}\left(F_{i-1}\right), \tag{9}$$

其中：$H_{\mathrm{S}}(\cdot)$ 表示 SR 分支，$H^i_{\mathrm{RSTB}}(\cdot)$ 表示第 i 个 RSTB，F_{t} 表示 STCHB 模块内 SR 分支的输出.

随后，使用融合模块（Fusion Block，FB）对两种特征进行融合，使得生成的 HR 梯度图能够为特征重构过程补充额外的结构信息，如图 2（b）所示. 首

先将两分支的输出F_t和F_c进行拼接,随后使用1×1卷积层构建通道融合模块(channel-wise fusion)以专注于通道维度上的融合,最后利用1×1卷积层对通道进行调整后得到输出F_{fusion}:

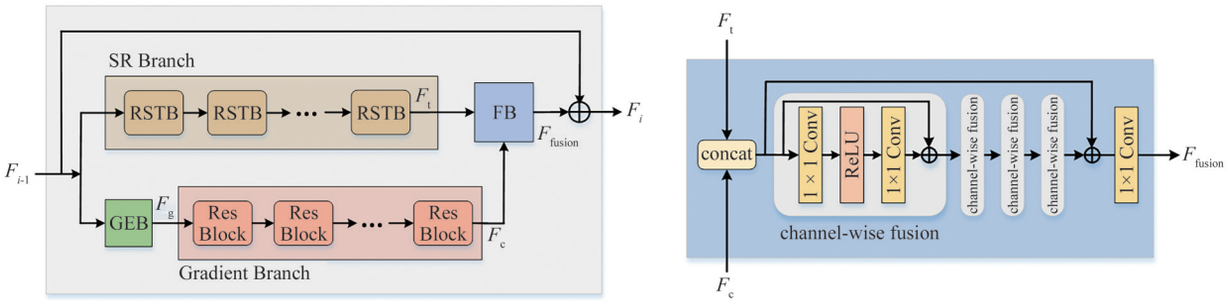
$$F_{\text{fusion}} = H^i_{\text{fusion}}\left(H_{\text{concat}}\left(F_t, F_c\right)\right), \tag{10}$$

其中:$H_{\text{concat}}(\cdot)$表示通道维度上的拼接,$H^i_{\text{fusion}}(\cdot)$表示第$i$个通道融合模块.

最后,第i个STCHB的最终输出F_i为:

$$F_i = H^i_{\text{STCHB}}\left(F_{i-1}\right) = F_{\text{fusion}} + F_{i-1}, \tag{11}$$

其中,$H^i_{\text{STCHB}}(\cdot)$表示第$i$个SwinT-CNN混合模块.



(a) STCHB (b) Fusion Block (FB)

图2 TCHB的内部结构框图

Fig. 2 Block diagram of the whole STCHB structure

3 实验

3.1 训练方法

为了生成高质量的8倍放大图像,网络采用了渐进式训练方法.网络总共训练600个epoch,第一阶段训练100个epoch,利用SFEB和STCHB生成2倍放大图像,并将其与目标图像进行比较.第二阶段训练200个epoch,对第一阶段的输出进行处理,获得4倍放大输出后再次与对应的目标图像进行比较.第三阶段训练300个epoch,重复上述过程来获得最终的8倍放大输出图像.该方法允许网络对每个分辨率下的图像都加以损失约束,有效、稳定地恢复出高倍放大重构图像.

3.2 损失函数

为了使网络在梯度域中学习到更多的信息,在像素级损失的基础上增加了梯度损失,因此总的损失函数定义为:

$$L_{\text{all}} = L_{\text{sr}} + \alpha L_{\text{grad}}, \tag{12}$$

其中:L_{sr}与L_{grad}分别表示像素级损失和梯度损失,α为损失权重.

选择$L1$损失函数来最优化所提网络.给定N对图像作为训练集,可表示为$\left\{I^i_{\text{LR}}, I^i_{\text{HR}}\right\}^N_{i=1}$.每对图像对包含1幅LR图像和对应的HR图像,优化目标L_{sr}和L_{grad}表示为:

$$L_{\text{sr}} = \frac{1}{N}\sum_{i=1}^{n}\left\| H_{\text{SR}}\left(I^i_{\text{LR}}\right) - I^i_{\text{HR}}\right\|_1, \tag{13}$$

$$L_{\text{grad}} = \frac{1}{N}\sum_{i=1}^{n}\left\| \nabla\left(H_{\text{SR}}\left(I^i_{\text{LR}}\right)\right) - \nabla\left(I^i_{\text{HR}}\right)\right\|_1, \tag{14}$$

其中:$H_{\text{SR}}(\cdot)$表示超分辨率网络,∇表示提取图像梯度.

3.3 实验设置和训练数据

本文所提网络使用的训练集来自DIV2K[19]数据集中的HR图像,其中的800张用于训练,100张用于验证.在训练阶段,将每幅HR图像随机分割成大小为192×192的图像块,再分别进行不同尺度因子(2×、4×和8×)的下采样操作,获得对应的LR图像.随后将所有图像随机旋转90°、180°、270°和水平翻转来增加数据的多样性.每次迭代时将8个大小为24×24的LR图像块作为网络的输入.网络的学习率为10^{-4},损失权重$\alpha = 0.5$,通过Adam来优化,参数为:β_1=0.9,β_2=0.999,ε=10^{-8}.

在测试阶段,为了评估模型的性能,选用五个标准数据集:Set5[20],Set14[21],BSD100[22],Urban100[22]和Manga109[23]进行测试.在图像YcbCr颜色空间中的Y通道上计算PSNR和SSIM指标,指标越高表示图像质量越优.所有实验均使用了Pytorch深度学习框架以及两块NVIDIA GTX-1080TI显卡.

3.4 实验结果

在五个标准数据集上将本文方法与其他方法在同等条件下进行比较,包括SRCNN[4]、VDSR[5]和EDSR[7],采用渐进式网络结构的LapSRN[6],引入图像梯度信息的SPSR[17]和基于transformer结构的SwinIR[16].表1列出了在8倍放大下所有方法的对比结果,最好结果用加粗黑体标出.

从表1可以看出,本文所提方法具有较强的可比性.以Set5[20]数据集为例,本文方法与

EDSR[7]和 SPSR[17]相比，在 PSNR 指标上分别获得了 0.48 dB 和 0.15 dB 的性能提升，且参数量分别减少了 36.88 M 和 17.86 M. 而相较于 SRCNN[4]、VDSR[5]和 LapSRN[6]来说，虽然本文方法的参数量有所增加，但是在 PSNR 上分别获得了 1.7、1.31、和 0.88 dB 的性能提升. 在其余 4 个数据集上，本文方法也比上述方法在重构性能上有不同程度的改进.

表 1　不同重构算法在 8×放大倍数下的 PSNR（dB）和 SSIM 的比较

Tab. 1　Comparison of PSNR（dB）and SSIM of different reconstruction algorithms at 8× scale factor

Method	Params	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN[4]	57 K	25.33	0.690	23.76	0.591	24.13	0.566	21.29	0.544	22.37	0.682
VDSR[5]	667 K	25.72	0.711	24.21	0.609	24.37	0.576	21.54	0.560	22.83	0.707
LapSRN[6]	1.31 M	26.15	0.738	24.35	0.620	24.54	0.586	21.81	0.581	23.39	0.735
EDSR[7]	43 M	26.55	0.754	24.66	0.626	24.63	0.588	22.08	0.620	24.27	0.768
SPSR[17]	23.98 M	26.88	0.773	24.82	0.635	24.75	0.572	22.31	0.608	24.33	0.772
SwinIR[16]	11.75 M	27.01	0.772	24.94	0.638	24.79	0.587	22.49	0.615	**24.58**	0.707
Ours	6.12 M	**27.03**	**0.776**	**24.96**	**0.642**	**24.83**	**0.595**	**22.51**	**0.623**	24.52	**0.779**

而相较于 SwinIR[16]，在 PSNR 指标上，所提方法仅在 Manga109[23]数据集上略低于 SwinIR[16]. 为了更好地说明两种方法的区别，在 Manga109[23]数据集上，对两种方法在不同的方面进行了比较. 由表 2 可知，虽然本文方法在性能指标上略低于 SwinIR[16]，但本文方法需要的重构时间更少，且参数量更小. 相比较于表 2 中的其他方法，本文所提方法的 PSNR 值最高，除 LapSRN[6]外，拥有最短的重构时间和最低的参数量，但 PSNR 值比 LapSRN[6]高出了 1.13 dB.

表 2　Manga109 数据集上，不同算法在速度、参数和性能之间的比较

Tab. 2　Comparison of speed, parameter, performance of different algorithms on Manga109

	LapSRN[6]	EDSR[7]	SPSR[17]	SwinIR[16]	Ours
重构时间/s	**0.15**	2.79	2.34	1.84	0.71
模型参数/M	**1.31**	43	23.98	11.75	6.12
PSNR/dB	23.39	24.27	24.33	**24.58**	24.52

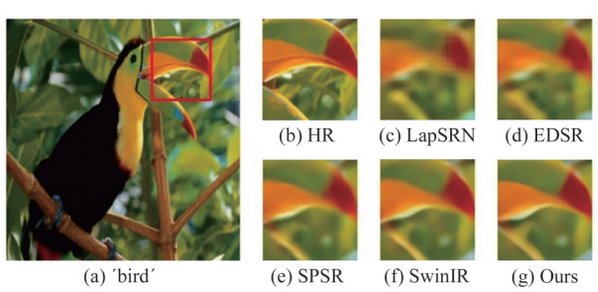
为了验证两分支中 ResBlock 和 RSTB 数量选择的合理性，分别将数量设置为同样的 3 个、4 个和 5 个并分析对网络性能的影响. 如表 3 所示，随着模型数量的增加，PSNR 值更高，这是因为深度网络具有良好的非线性表征能力，但网络的参数也在逐渐增大. 相较于使用 4 个 ResBlock 和 RSTB，使用 3 个时，网络的性能较差，使用 5 个时，网络的参数较大且性能仅有轻微的提升. 因此决定将模型的数量设定为 4.

为了证明本文方法的重构图像视觉效果更佳，给出了不同数据集上本文方法和其他方法得到的 8 倍放大重构图像，如图 3、图 4 和图 5 所示. 在图 3 中，本文方法重构出的图像在鸟喙处更加尖锐，且轮廓更具可辨性. 在图 4 中，本文方法重构出的字符更加清晰，其他方法均有较严重的模糊和重影. 在图 5 中，本文方法得到了分界线更加分明的墙壁图像.

表 3　模块数量对网络性能的影响

Tab. 3　Influence of the number of modules on network performance

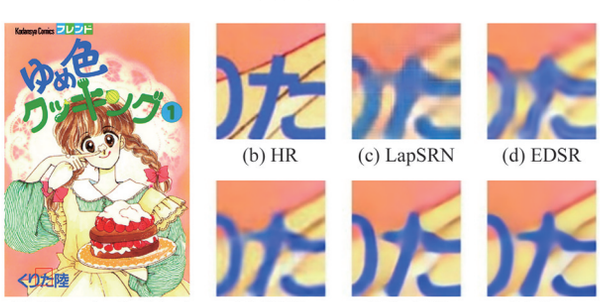
	模块的数量		
数量/个	3	4	5
网络参数/M	**4.23**	6.12	8.07
PSNR/dB	26.92	27.03	**27.04**



(a) 'bird'　(b) HR　(c) LapSRN　(d) EDSR　(e) SPSR　(f) SwinIR　(g) Ours

图 3　8 倍放大下不同算法对'bird'的重构结果

Fig. 3　Reconstructed results of image 'bird' by different algorithms at 8 scale factor



(a) 'YumeiroCooking'　(b) HR　(c) LapSRN　(d) EDSR　(e) SPSR　(f) SwinIR　(g) Ours

图 4　8 倍放大下不同算法对'YumeiroCooking'的重构结果

Fig. 4　Reconstructed results of image 'YumeiroCooking' by different algorithms at 8 scale factor

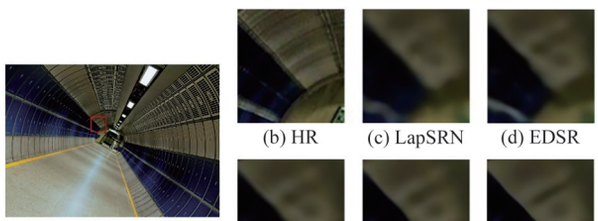


图5 8倍放大下不同算法对'BSD100_078'的重构结果

Fig. 5 Reconstructed results of image 'BSD100_078' by different algorithms at 8 scale factor

3.5 消融实验及分析

为了验证STCHB组成架构的有效性,本文对STCHB的架构在8倍放大下Set5[20]数据集上进行了消融实验,如表4所示.

表4 STCHB架构的消融实验

Tab. 4 Ablation experimental of STCHB structure

不同的STCHB组成架构			
组成架构	all CNN	all Transformer	CNN+Transformer
PSNR/dB	26.46	26.36	**27.03**

从表4中可以看出,当STCHB中两分支的基本组成模块全部使用CNN或Transformer时,都造成了网络重构性能不同程度的下降.这是因为此时的网络只侧重于捕获局部信息或全局信息,而忽略了另一部分信息的重要性.而STCHB结合了Transformer和CNN二者的优势,提高了网络的综合实力,重构性能有了明显的提升.

为了验证渐近结构和梯度分支的有效性,本文验证了在8倍放大下Set5[20]数据集上渐进结构和梯度分支对重构性能的影响,如表5所示.从表中可以看出,将渐近放大改为直接放大或移除梯度分支,都造成了网络重构性能不同程度的下降.这是因为渐近结构缓解了网络训练的难度,而梯度分支能够为图像重构过程提供额外的结构信息.基于此,网络将二者都保留,以得到更佳的重构效果.

表5 渐近结构和梯度信息的消融实验

Tab. 5 Ablation experimental of gradient information and progressive structure

渐近结构和梯度信息的不同组合方式				
渐近结构	×	√	×	√
梯度分支	×	×	√	√
PSNR/dB	26.73	26.82	26.86	**27.03**

为了验证梯度损失的有效性,对损失权重α进行不同的取值,在8倍放大下Set5[20]数据集上,验证梯度损失对网络性能的影响,如表6所示.从表中可以看出,相较于仅使用单一像素级损失($\alpha = 0$),增加梯度损失后重构质量明显提高,并且当$\alpha = 0.5$时,可以获得最优的重构效果.这是因为在梯度域中增加了约束之后,网络能够从中学到更多的局部特征,为重构过程提供额外的结构信息,有助于恢复出视觉效果良好的SR图像.

表6 梯度损失的消融实验

Tab. 6 Ablation experimental of gradient loss

α的不同取值					
α	0	0.3	0.5	0.7	1
PSNR/dB	26.81	26.95	**27.03**	26.98	26.93

此外,为了验证梯度信息能够改善重构图像的视觉效果,让梯度分支同样作用于特征域,其他设置保持不变.图6给出了使用梯度信息与不使用梯度信息对Set14[21]数据集中'baboon'进行8倍放大后的结果.从图中可以看出,使用梯度信息重构出的图像中皮毛纹理细节更加清晰,伪影较少,且边缘之间更加分明.这表明梯度信息的引入可以帮助改善重构图像的视觉效果.

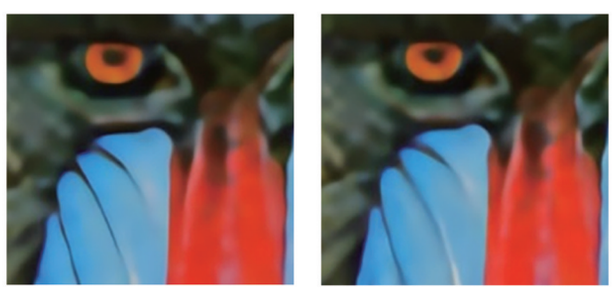


图6 8倍放大下有否梯度信息对'baboon'的重构视觉

Fig. 6 Reconstructed results of image 'baboon' without and with the gradient information at 8 scale factor

4 结语

研究了一种基于ViT-CNN特征增强的图像超分辨率方法,得到了高倍放大任务中重构图像质量的较好提升.通过基于ViT的SR分支与基于CNN的梯度分支的特征增强,有效提高了网络的学习能力.SR分支主要用于提取图像特征域中的全局相关性,而梯度分支则专注于图像梯度域中的局部依赖关系.通过引入梯度损失和渐进训练策略进一步提高了重构质量.实验结果验证了本文方法在提升重构性能方面的有效性.当然,本文方法也存在ViT与CNN融合方式较为简单的不足,因此如何更好地改进二者的融合方式以得到更好的效果,成为今后需要进一步深入探究的问题.

参 考 文 献

[1] ZHANG L, ZHANG H, SHEN H, et al. A super-resolution reconstruction algorithm for surveillance images[J]. Signal Processing, 2010, 90(3): 848-859.

[2] GREENSPAN H. Super-resolution in medical imaging[J]. The Computer Journal, 2009, 52(1): 43-63.

[3] SHERMEYER J, VAN ETTEN A. The effects of super-resolution on object detection performance in satellite imagery[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2019: 1432-1441.

[4] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]// European Conference on Computer Vision. Zurich: Springer, 2014: 184-199.

[5] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.

[6] LAI W S, HUANG J B, AHUJA N, et al. Deep Laplacian pyramid networks for fast and accurate super-resolution[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 5835-5843.

[7] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE, 2017: 1132-1140.

[8] LAI W S, HUANG J B, AHUJA N, et al. Fast and accurate image super-resolution with deep Laplacian pyramid networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2599-2613.

[9] PARK S, YOO J, CHO D, et al. Fast adaptation to super-resolution networks via meta-learning[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 754-769.

[10] 熊承义，李雪静，高志荣，等. 基于并行反向投影的图像超分辨率[J]. 中南民族大学学报(自然科学版)，2024，43(1)：53-60.

[11] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[C]//Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017: 1,2,4,5.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x6 words: Transformers for image recognition at scale[J]. arXiv Preprint arXiv: 2010.11929, 2020.

[13] CHEN H, WANG Y, GUO T, et al. Pre-trained image processing transformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 12294-12305.

[14] 熊承义，郑瑞华，高志荣，等. 结合多尺度多注意力的遥感图像超分辨率重构[J]. 中南民族大学学报(自然科学版)，2024，43(5)：692-700.

[15] Lu Z, Li J, Liu H, et al. Transformer for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 456-465.

[16] LIANG J, CAO J, SUN G, et al. SwinIR: Image restoration using swin transformer[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal: IEEE, 2021: 1833-1844.

[17] MA C, RAO Y, CHENG Y, et al. Structure-preserving super resolution with gradient guidance[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 7766-7775.

[18] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 9992-10002.

[19] AGUSTSSON E, TIMOFTE R. NTIRE 2017 challenge on single image super-resolution: Dataset and study[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE, 2017: 1122-1131.

[20] BEVILACQUA M, ROUMY A, GUILLEMOT C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[C]//Electronic Proceedings of the British Machine Vision Conference. Surrey: BMVC, 2012: 1-10.

[21] ZEYDE R, ELAD M, PROTTER M. On single image scale-up using sparse-representations[C]//Curves and Surfaces. International Conference on Curves and Surfaces. Avignon: Springer, 2010: 711-730.

[22] 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[C]//IEEE International Conference on Computer Vision. Vancouver: IEEE, 2001: 416-423.

[23] MATSUI Y, ITO K, ARAMAKI Y, et al. Sketch-based manga retrieval using manga109 dataset[J]. Multimedia Tools and Application, 2017, 76: 21811-21838.

（责编&校对　刘钊）