研究论文

基于并行Transformer和CNN的图像压缩感知重构网络

张新岩[1],祝勇俊[1,2,3*],吴宏杰[1],周凡利[3]

摘要 图像压缩感知是一种能够在低采样率下实现高效信号采样与重构的技术,但在实现高质量图像重构时,面临局部与全局特征难以有效融合的问题。为此,提出一种结合Transformer与卷积神经网络(convolutional neural networks,CNN)优点的图像压缩感知重构框架(transformer-CNN mixture transformer,TCMformer)。该框架充分利用CNN的局部建模能力和Transformer的全局特征捕捉能力;设计了一种特征融合模块(TCM Block),有效桥接局部与全局特征,从而提升特征表示效率;同时,为降低模型复杂度并控制计算成本,框架采用基于窗口的Transformer结构,通过分块实现高效的全局建模。此外,引入渐进式重建策略,利用多尺度特征图逐步优化重建质量。实验结果表明,TCMformer在峰值信噪比、结构相似性和视觉效果上相较于主流的压缩感知重构算法表现更优,为实现高质量的图像重建提供了一种有效的解决方案。

关键词 压缩感知;Transformer;卷积神经网络;图像重建

压缩感知(compressive sensing,CS)理论[1]表明,当信号在某些变换域中具有稀疏性时,可以从远低于奈奎斯特采样定理要求的采样测量中恢复信号。这一特性显著降低了对采样率的要求,有效减轻数据存储和传输带宽的压力[2]。基于此优势,压缩感知在单像素相机[3]、磁共振成像[4]、快照压缩成像[5]等领域获得了广泛应用并取得了成功。

在压缩感知方法中,原始信号$x \in \mathbb{R}^N$通过快速采样获得线性随机测量$y = \boldsymbol{\Phi}x \in \mathbb{R}^M$。其中,$\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$是测量矩阵,且$M \ll N$,M/N表示压缩感知的采样率。由于未知数$x$的维度$N$远大于测量值$y$的维度$M$,该逆问题通常是不适定的。为解决此问题,传统的压缩感知重建算法[6-7]通常利用信号在特定变换域中的稀疏性。将不适定的L_0范数优化问题转化为L_1范数凸优化问题,或采用迭代方法逐步逼近原始信号。这些方法虽然在理论上具有可行性和可解释性,但依赖迭代计算,导致计算成本较高。此外,传统方法对稀疏性假设的依赖限制了其在非稀疏信号中的应用,难以适应实际应用中多样化信号的需求。

近年来,随着深度学习(deep learning,DL)的兴

1. 苏州科技大学电子与信息工程学院,苏州 215009
2. 南京航空航天大学电子信息工程学院,南京 211106
3. 苏州同元软控信息技术有限公司,苏州 215123

收稿日期:2023-12-03;修回日期:2024-05-16
基金项目:国家自然科学基金项目(62073231)
作者简介:张新岩,硕士研究生,研究方向为图像处理,电子信箱:2213041047@post.usts.edu.cn;祝勇俊(通信作者),高级实验师,研究方向为图像信号处理、智能楼宇与智慧交通,电子信箱:zyj@mail.usts.edu.cn
引用格式:张新岩, 祝勇俊, 吴宏杰, 等. 基于并行Transformer和CNN的图像压缩感知重构网络[J]. 科技导报, 2025, 43(2): 108-116;
doi: 10.3981/j.issn.1000-7857.2023.12.01823

起[8],推动了多种基于数据驱动的CS深度神经网络模型的发展,这些模型在重建质量和恢复速度方面表现出色[9]。

现有基于深度学习的CS方法主要分为2类,第一类是深度展开方法[10],通过深度神经网络模拟传统迭代优化算法。如Zhang等[11]提出ISTA-Net,将经典的迭代收缩阈值算法展开为多层卷积神经网络(CNN),并引入传统优化算法中的数学先验,显著提高了重构效率和质量;第二类是基于人工设计参数约束的前馈网络方法[12-13],通过CNN进行图像重建。这类方法相较于传统方法在重建性能上有所提升,但局限于CNN的局部感知特性和权重共享机制,难以捕捉全局信息。为此,Sun等[14]引入非局部先验引导网络,通过融合非局部特征,提高图像重建质量。Sun等[15]提出双路径注意网络DPA-Net,将图像重建分为结构路径和纹理路径。尽管上述方法一定程度上改善了重建效果,但处理复杂特征时效果有限。

与基于卷积的深度神经网络不同,Transformer[16]因自注意力机制擅长建模全局上下文信息,已在自然语言处理(NLP)和计算机视觉任务中展现强大性能,如图像分类[17]、图像处理[18]和图像生成[19],并成为CNN的潜在替代方案。TransCS[20]首次尝试将Transformer应用于CS任务,以迭代收缩阈值算法为基础,采用定制的Transformer为核心结构,并结合CNN处理网络的输入和输出数据,验证了两者结合的潜力。然而,现有方法有些仅单独使用卷积或注意力机制,有些则简单地替换特定模块,这些设计均未能充分发挥两者的互补优势;此外,局部卷积与全局表征的自注意力机制的有效融合仍缺乏系统验证,其潜在的性能优势尚未完全挖掘。

为提高图像压缩感知重建质量,提出一种端到端的图像压缩感知混合框架TCMformer,通过自适应采样与混合重建,有效提升重建质量。采样阶段使用可学习矩阵逐块测量图像,重建阶段结合初始重建与Transformer-CNN混合重建,充分利用局部和全局信息。同时引入渐进重建策略处理多尺度特征图,显著减少内存开销和计算复杂度。与CNN方法相比,TCMformer模型具有自注意力机制、擅长处理远程特征、特征融合和渐进式重建等优势。

1　并行Transformer和CNN图像压缩感知重构网络

在传统CNN和Transformer结合应用于图像压缩感知的研究基础上,提出一种端到端的创新性混合框架TCMformer,如图1所示。该网络架构主要包含3部分:采样、初始重建和Transformer-CNN混合重建。在采样阶段通过可学习的采样矩阵自适应捕捉图像特征;在初始重建阶段,利用卷积层进行快速图像重建;在混合重建阶段,创新性地引入Transformer和CNN特征融合模块,充分发挥局部卷积与全局自注意力的互补优势,为图像压缩感知任务提供高效优质的解决方案。

1.1　采样与初始重建子网络

对于图像信号,直接对整图进行采样会带来较大的计算负担,因此,Gan等[21]提出了基于块的图像压缩感知算法(block-based compressive sensing, BCS),该算法通过对图像分块处理,有效降低采样端和重构端的计算压力。首先将图像$\boldsymbol{X}\in\mathbb{R}^{H\times W}$划分为块$\boldsymbol{X}_i\in\mathbb{R}^{H_p\times W_p}$,再将块$\boldsymbol{X}_i$分解为$B\times B$大小的非重叠子块,则子块的数量是$\frac{H_p}{B}\times\frac{W_p}{B}$,并对每个子块进行矢量化处理,随后通过测量矩阵$\boldsymbol{\Phi}$进行采样。假设子块$\boldsymbol{X}_{ij}$是输入块$\boldsymbol{X}_i$的块$j$,则相应的测量值$\boldsymbol{Y}_{ij}$由$\boldsymbol{Y}_{ij}=\boldsymbol{\Phi}\boldsymbol{X}_{ij}$

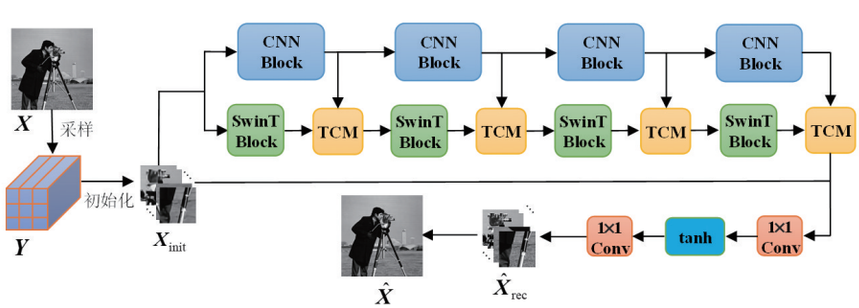


图1　在CS比率为25%时TCMformer系统框图

获得,其中 $\boldsymbol{\Phi} \in \mathbb{R}^{m \times B^2}$,$\frac{m}{B^2}$ 表示每个子块的采样率。最后通过堆叠每个子块获得输入块 $\boldsymbol{X}_i$ 的测量值 $\boldsymbol{Y}_i \in \mathbb{R}^{\frac{H_p}{B} \times \frac{W_p}{B} \times m}$。在传统的BCS算法中,采样过程是使用维度为 $M \times N$ 的随机高斯矩阵作为采样矩阵,由于高斯矩阵不可学习,且不具备自适应特性,严重影响了采样质量,同时基于分块的采样打破了图像中的像素间依赖性,导致明显的块效应。因此,本实验采用适当大小的卷积核与合适步长的卷积操作来替代传统的采样过程,从而提升采样质量并减少块效应。采样过程可表示为:

$$\boldsymbol{Y}_{ij} = \boldsymbol{W}_B \otimes \boldsymbol{X}_{ij} \tag{1}$$

式中,$\boldsymbol{W}_B$ 表示由 m 个大小为 $B \times B$ 的卷积核组成的无偏差卷积层,步长等于 B。对 $\boldsymbol{X}_i$ 进行卷积运算后,得到最终的总CS测量值 $\boldsymbol{Y}_i$。

采用学习到的卷积核代替传统的采样矩阵,可以有效提取图像的特征,使测量结果在之后的初始重建模块中更易于使用。

给定CS测量,传统的BCS通常通过 $\hat{\boldsymbol{X}}_{ij} = \boldsymbol{\Phi}^{\dagger} \boldsymbol{Y}_{ij}$ 来获得初始重建块,其中 $\hat{\boldsymbol{X}}_{ij}$ 是子块 $\boldsymbol{X}_{ij}$ 的重建,$\boldsymbol{\Phi}^{\dagger} \in \mathbb{R}^{p^2 \times m}$ 是 $\boldsymbol{\Phi}$ 的伪逆矩阵。在初始重建过程中,传统方法使用矩阵操作来恢复块,而本方法则使用卷积操作来替代 $\boldsymbol{\Phi}^{\dagger}$,并直接在 $\boldsymbol{Y}_i$ 上利用卷积层来恢复初始块。初始化首先采用核大小为 $1 \times 1 \times m$ 的 p^2 个卷积核将测量值 $\boldsymbol{Y}_i$ 的维度转换为 p^2。然后,采用子像素卷积层来获得初始块 $\hat{\boldsymbol{X}}_i$。通过卷积和子像素卷积的结合,直接获得每个初始重建块的张量输出,而不是传统方法中的向量形式,这使得处理更加高效。整个初始重建子网络可以表示为:

$$\boldsymbol{X}_{\text{init}} = \mathrm{F}_{\text{sub}}\left(\boldsymbol{W}_B \otimes \boldsymbol{X}\right) \tag{2}$$

式中,$\mathrm{F}_{\text{sub}}(\cdot)$ 表示子像素卷积层,$\boldsymbol{W}_B \otimes \boldsymbol{X}$ 表示对所有采样图像块的卷积操作。

1.2 Transformer-CNN混合重建子网络

Transformer-CNN混合重建子网络包括CNN单元、SwinT单元和特征融合模块。将初始重建 $\boldsymbol{X}_{\text{init}}$ 图像分别送入CNN单元和SwinT单元,每个单元又分别由4个CNN模块和SwinT模块组成。

1.2.1 CNN模块

图2为CNN模块结构图。在CNN模块中包含1个上采样层和2个卷积块。每个卷积块由1个卷积层后跟1个ReLU激活函数和1个批归一化层组成。卷积层的核大小为3×3,填充大小为1,输出通道大小与输入通道大小相同。因此,在卷积块之后,分辨率和通道大小保持一致。为了扩展到更高分辨率的特征,CNN模块在第1个模块外,添加1个上采样模块,上采样模块首先采用双三次上采样来提高先前特征的分辨率,然后使用1×1卷积层将维度降低到1/2。同时整个CNN模块采用残差连接,提高网络提取深层次特征的能力。

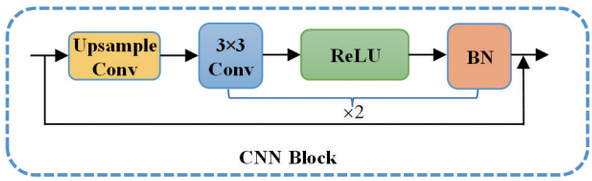


图2 CNN模块结构

1.2.2 SwinT模块

标准的Transformer将一系列序列作为输入,并在所有令牌之间计算全局自注意力。然而,如果将每个像素作为Transformer中的一个令牌用于CS重建,随着图像分辨率的增加,序列长度也会相应增加,导致计算复杂度呈指数级增长,尤其在处理高分辨率图像时,这种爆炸性的计算开销成为一个重要挑战。给定一个输入张量 $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$,其自注意力可以表示为:

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = \text{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\right)V \tag{3}$$

式中,$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q}$,$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{K}$,$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{V}$。为了简化分析,省略了归一化项,从式(3)中可以看出自注意力机制的复杂性主要来源于3个方面:(1) 生成 $\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$ 的张量,其复杂度为 $3HWC^2$;(2) 基于 $\boldsymbol{K}$-$\boldsymbol{Q}$ 点积生成注意力图,其复杂度为 $(HW)^2C$;(3) 加权求和过程,其复杂度为 $(HW)^2C$。可以看出在后2项中复杂度与空间尺寸呈二次关系。

为解决计算复杂度过高问题,TCMformer执行基于窗口的Swin Transformer。如图3所示是SwinT块利用自注意力机制,配合多层感知机(MLP)和层归一化(LayerNorm),为CS任务提供了一种结构化和高效的特征提取和重构框架。

给定Transformer的输入特征 $\mathrm{F}_t^j \in \mathbb{R}^{H_j \times W_j \times C_j}$,首先将 F_t^j 与可学习的位置编码 $\mathrm{E} \in \mathbb{R}^{H_j \times W_j \times C_j}$ 相加,然后将

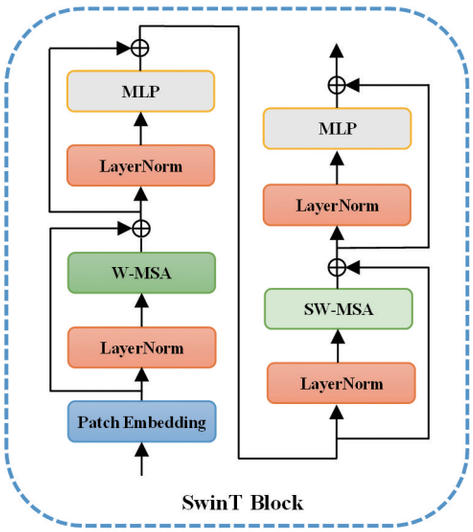


图3 SwinT模块结构

特征划分为$p\times p$个不重叠的窗口。在每$p\times p$个窗口中计算多头自注意力。在每个窗口中,特征$\mathrm{F}_{\mathrm{t}}^{\mathrm{win}}\in\mathbb{R}^{p^2\times\frac{C_j}{h}}$由自注意力计算,其中$h$是多头自注意中的头的数目。首先,查询、键和值矩阵计算为:

$$\begin{aligned}\boldsymbol{Q}&=\mathrm{F}_t^{\mathrm{win}}\times\boldsymbol{W}_Q\\\boldsymbol{K}&=\mathrm{F}_t^{\mathrm{win}}\times\boldsymbol{W}_K\\\boldsymbol{V}&=\mathrm{F}_t^{\mathrm{win}}\times\boldsymbol{W}_V\end{aligned}\tag{4}$$

式中,$\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$是大小为$C_j/h\times d$的投影矩阵。随后,自注意力可表示为:

$$\mathrm{O}\left(\mathrm{F}_t^{\mathrm{win}}\right)=\left[\sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt[2]{d}}+\mathrm{E}_r\right)\right]\boldsymbol{V}\tag{5}$$

式中,$\mathrm{O}\left(\mathrm{F}_t^{\mathrm{win}}\right)$表示自注意力操作,$\sigma(\cdot)$是Softmax函数,$\mathrm{E}_r$表示可学习的相对位置编码。多头自注意力通过并行执行h次自注意力操作,将每个头的输出连接起来,最终获得综合的输出。基于窗口的多头自注意力(MSA)通过局部化注意力范围显著降低了计算成本和图型处理器(GPU)内存消耗。MSA的输出通过由2个全连接层组成的MLP进行进一步处理,并通过高斯误差线性单元(GELU)激活函数进行非线性变换。在这一过程中,插入层归一化操作$\tau(\cdot)$,整个Transformer过程可以表述为:

$$\begin{aligned}\mathrm{F}_t^j&=\mathrm{F}_t^j+\mathrm{E}\\\mathrm{F}_a^j&=\mathrm{MSA}\left[\tau\left(\mathrm{F}_t^j\right)\right]+\mathrm{F}_t^j\\\mathrm{F}_a^j&=\mathrm{MLP}\left[\tau\left(\mathrm{F}_a^j\right)\right]+\mathrm{F}_a^j\end{aligned}\tag{6}$$

式中,F_a^j表示SwinT模块的特征。

1.2.3 特征融合模块

TCM模块以CNN模块和SwinT模块输出的特征作为输入,用于进一步提取和融合信息。由图4可以看出,TCM模块接收2种类型的特征:卷积特征F_c^j和Swin Transformer特征F_a^j。这些特征首先在通道维度上进行级联,形成合并后的特征F_{T}^j,并依次经过TCM模块的各个组件进行处理。首先,使用全局平均池化(global average pooling,GAP)对合并的特征进行空间维度的压缩,以提取全局上下文信息,将每个特征通道的空间信息缩减为单个数值。之后使用1×1卷积对特征进行通道压缩。通过ReLU激活函数引入非线性,增强模型对复杂特征的学习能力。随后,压缩后的特征通过1×1卷积扩展回原通道数,并通过Sigmoid激活函数将每个通道的值映射为权重σ。Sigmoid输出的权重与原来的特征$\mathrm{F}_{\mathrm{T}}^j$进行逐元素乘法,生成加权特征图。之后通过2个3×3卷积层和ReLU激活函数对加权特征图进行进一步处理,并通过残差连接保留了输入的原始特征。残差连接保证模块的输出不仅包含了经过各层处理后的特征,还包括了未经修改的原始特征。这样,即使模块内部的层没有学习到有效的特征转换,网络仍然能够利用输入的原始特征。最后使用卷积调整输出特征的通道数,生成最终的融合特征图。图4展示的TCM模块结构表明,该设计在结合各层处理后的特征基础上,有效提升了特征融合的稳健性和效率。整个特征融合模块可以表示为:

$$\mathrm{F}_{\mathrm{T}}^j=\mathrm{F}_c^j\oplus\mathrm{F}_a^j\tag{7}$$

$$\sigma=\mathrm{Sigmoid}\left\{\mathrm{Conv}_{1\times1}\left[\mathrm{ReLU}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{AvgPool}\left(\mathrm{F}_{\mathrm{T}}^j\right)\right)\right)\right]\right\}\tag{8}$$

$$\mathrm{F}^j=\mathrm{Conv}_{1\times1}\left\{\mathrm{F}_{\mathrm{T}}^j+\mathrm{Conv}_{3\times3}\left[\mathrm{ReLU}\left(\mathrm{Conv}_{3\times3}\left(\sigma\times\mathrm{F}_{\mathrm{T}}^j\right)\right)\right]\right\}\tag{9}$$

式中,F_c^j表示卷积模块的特征,$\oplus$表示通道维度级联。

输出投影模块由2个卷积层和1个tanh激活函数组成,该函数将TCM模块输出的特征F映射到单通道重建块,再将重建图像块与初始重建图像块求和以获得最终图像块$\hat{\boldsymbol{X}}_{\mathrm{rec}}$,合并所有图像块以获得最终重建图像$\hat{\boldsymbol{X}}$。

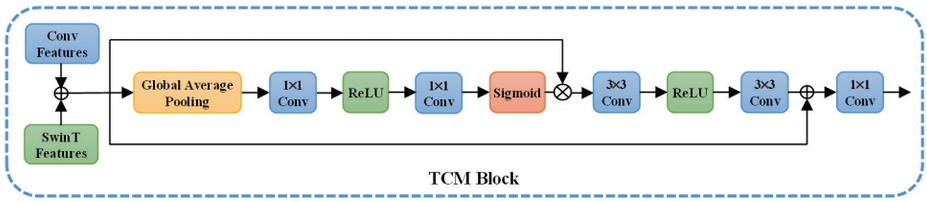


图4 TCM模块结构

1.3 损失函数

通过最小化输出重建图像$\hat{\boldsymbol{X}}$和真实图像$\boldsymbol{X}$之间的均方误差(MSE)来优化TCMformer的参数,损失函数ι定义如下

$$\iota = \left\| \hat{\boldsymbol{X}} - \boldsymbol{X} \right\|_2^2 \tag{10}$$

TCMformer采用基于块的重建方法,而损失函数是在整个图像范围内进行计算的。因此,块效应可以在不引入额外去块操作的情况下得到有效抑制。

2 实验结果与分析

2.1 实验设置

2.1.1 数据集和指标

为保证实验结果的公平性,选用了图像CS领域公认的BSD500数据集[22],该数据集包括200张训练图像和200张测试图像,共计400张图像。由于训练视觉Transformer需要大量的数据样本,因此采用数据增强技术扩充训练集,具体操作为:首先将训练图像随机裁剪成固定大小96×96的子图像,然后通过旋转、翻转等操作进一步增强训练集,最终将增强后的图像进行灰度化处理,得到89600个子图像作为网络的训练数据。在测试阶段,使用2个广泛使用的基准数据集:Set11[23]和Urban100[24]。Set11数据集包含11张灰度图像,Urban100数据集包含100张高分辨率的、具有丰富纹理的城市图像。对于彩色图像,重建结果均在亮度通道上进行评估,采用峰值信噪比(peak signal to noise ratio,PSNR)和结构相似性指数(structural similarity index measure,SSIM)为评价指标。

2.1.2 训练详情

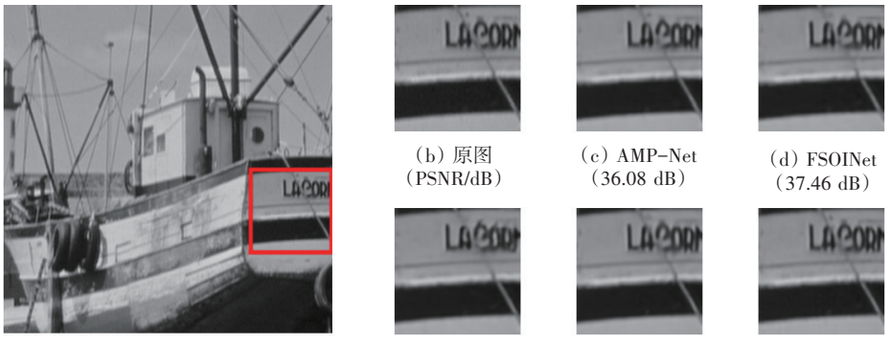
训练图像被裁剪成96×96图像作为输入,即H_p=W_p=96。将采样过程中的采样卷积核大小设置为B=16,即16×16的卷积层,步长为16。为所有SwinT块的窗口多头自注意力的窗口大小设置为p×p=8×8。每个SwinT块由L=4个Swin Transformer网络堆叠而成。实验在一块NVIDIA 3060Ti显卡上,使用Pytorch框架进行模型训练,并采用Adam优化器进行优化。学习速率初始设定为2×10^{-4},并通过余弦衰减策略进行调整,经过100个epoch将学习速率降至5×10^{-5},训练的前3个epoch作为warm-up阶段,学习率从0线性增长为2×10^{-4}。余弦衰减策略的核心思想是在每个训练周期中,根据余弦函数的周期性变化规律调整学习率。通过控制余弦函数的参数,能够有效控制学习率的变化速度和周期,从而实现更好的训练效果。

2.2 性能比较

为便于比较,本研究在2个广泛使用的测试集上评估了TCMformer的性能,并与7种最新的基于深度学习的方法进行比较,包括CSNet$^+$、ISTA-Net$^+$、DPA-Net、AMP-Net、FSOINet[25]、CASNet[26]和TransCS。在Set11数据集上,关于5个CS比率的PSNR/SSIM重建性能总结于表1。从表1可以看出,本方法在所有比率下均取得最佳的PSNR和SSIM性能。其平均PSNR性能分别优于CSNet$^+$、ISTA-Net$^+$、DPA-Net、AMP-Net、FSOINet、CASNet和TransCS 2.63、3.50、3.76、1.10、0.14、0.26和1.01。同时,TCMformer的平均SSIM分别提高0.0166、0.0394、0.0276、0.0109、0.0006、0.0011和0.0069。图5为当CS比率为25%时,Set11数据集中Boats图像的重建结果与PSNR值,其中图5(a)为原图像,图5(b)为局部放大图像,图5(c)~(f)为目前主流方法重建后的局部放大细节,图5(g)为本研究方法。可以看出,TCMformer相较于其他方法呈现了更清晰的边缘和更精细的细节。而其他方法则恢复出较为模糊的纹理。一种可能的解释是,图像中的文字区域纹理相对模糊,其他方法更关注局部特征,而TCMformer则通过Trans-

表1 Set11数据集中不同采样率下不同算法重构图像的PSNR(dB)/SSIM对比

方法	10%采样率	25%采样率	30%采样率	40%采样率	50%采样率	平均
CSNet+	28.28/0.8690	33.17/0.9420	34.36/0.9529	36.67/0.9676	38.58/0.9763	34.21/0.9416
ISTA-Net+	26.49/0.8036	32.44/0.9237	33.70/0.9382	36.02/0.9579	38.07/0.9706	33.34/0.9188
DPA-Net	27.66/0.8530	32.38/0.9311	33.35/0.9425	35.21/0.9580	36.80/0.9685	33.08/0.9306
AMP-Net	29.40/0.8779	34.63/0.9481	36.03/0.9586	38.28/0.9715	40.34/0.9804	35.74/0.9473
FSOINet	30.46/0.9023	35.80/0.9595	37.00/0.9665	39.14/0.9764	41.08/0.9832	36.70/0.9576
CASNet	30.36/0.9014	35.67/0.9591	36.92/0.9662	39.04/0.9760	40.93/0.9826	36.58/0.9571
TransCS	29.54/0.8877	35.06/0.9548	35.62/0.9588	38.46/0.9737	40.49/0.9815	35.83/0.9513
TCMformer	30.71/0.9033	35.95/0.9602	37.15/0.9671	39.21/0.9768	41.16/0.9838	36.84/0.9582



(a) Boats (b) 原图 (PSNR/dB) (c) AMP-Net (36.08 dB) (d) FSOINet (37.46 dB) (e) CAS-Net (36.57 dB) (f) TransCS (35.06 dB) (g) TCMformer (37.95 dB)

图5 25%采样率下图像Boats(Set11)的重构图像视觉效果与PSNR值对比

former-CNN混合模块有效地利用长程依赖关系,实现了更加精确的图像恢复。

此外,表2展示了在Urban100数据集上TCMformer与其他方法的比较,该数据集包含更多的高分辨率图像,具有更丰富的图像分布。结果表明,TCMformer在所有采样率下均实现了更优的重建质量。

表2 Urban100数据集中不同采样率下不同算法重构图像的PSNR(dB)/SSIM对比

方法	10%采样率	25%采样率	30%采样率	40%采样率	50%采样率	平均
CSNet+	24.99/0.7979	29.23/0.9070	30.35/0.9256	32.28/0.9408	34.22/0.9588	30.21/0.9060
ISTA-Net+	23.51/0.7201	28.91/0.8834	30.15/0.9070	32.19/0.9362	34.37/0.9571	29.83/0.8808
DPA-Net	24.55/0.7841	28.80/0.8944	29.47/0.9034	31.09/0.9311	32.08/0.9447	29.20/0.8915
AMP-Net	26.04/0.8151	30.89/0.9202	32.19/0.9365	34.37/0.9578	36.33/0.9712	31.96/0.9202
FSOINet	27.53/0.8627	32.62/0.9430	33.84/0.9540	35.93/0.9688	37.80/0.9777	33.54/0.9412
CASNet	27.46/0.8616	32.20/0.9396	33.37/0.9511	35.48/0.9669	37.45/0.9777	33.19/0.9394
TransCS	26.72/0.8413	31.72/0.9330	31.95/0.9483	35.22/0.9648	37.20/0.9761	32.56/0.9327
TCMformer	27.70/0.8631	32.68/0.9445	34.10/0.9550	36.21/0.9672	37.95/0.9788	33.73/0.9417

3 消融实验与分析

3.1 模块消融

为验证所提架构的有效性,在25%采样率下对网络中的不同模块进行消融实验。(1) SwinT/T:将SwinT Block替换为普通Transformer Block,(2) w/o TCM Block:将TCM模块替换成简单的通道拼接,(3) SwinT:将特征融合阶段的CNN Block替换成

SwinT Block。表3展示了在25%采样率下,重建Set11数据集的PSNR、SSIM和参数量对比。通过比较实验结果,采用基于窗口的Transformer模块对参数量的改变巨大。引入TCM模块后,网络的性能有明显的提升,分析原因是TCM模块能够更好地提取图像的局部细节信息。然而,将特征融合阶段的CNN模块改成SwinT Block时,网络的性能有所下降,这是因为在通道缩减过程中,Transformer网络相比于CNN网络更容易丢失信息,从而影响重建质量。

表3 在CS比率为25%时重建Set11数据集图像变体网络的PSNR、SSIM、参数量比较

方法	PSNR/dB	SSIM	参数量/M
SwinT/T	35.82	0.9632	7.21
w/o TCM Block	34.63	0.9521	6.67
SwinT	35.33	0.9528	6.98
TCMformer	35.95	0.9602	6.85

3.2 复杂性分析

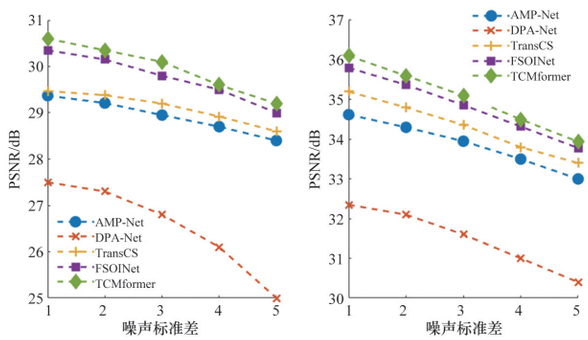
在许多实际应用中,计算成本和模型大小至关重要。因此,对不同方法在压缩感知比率为50%时,用于重建256×256图像的参数量、模型大小和FLOPs进行比较(表4)。考虑到TCMformer使用了Transformer和CNN相结合的模型,其总参数量仍然比使用双通道CNN结构的DPA-Net低30%。与其他方法相比,TCMformer的FLOPs是最小的。尽管运行时间有所增加,但实验结果清楚地表明,该方法在所有采样率和不同数据集上的定性和定量评估中,始终优于现有方法。另一方面,TCMformer仍然比大多数经典的迭代重建方法快得多,因为传统的图像CS方法通常需要几秒钟到几分钟的时间来重建256×256图像。未来的工作将进一步优化其运行时间。

表4 在CS比率为50%下重建 图像的参数、时间和FLOPs比较

方法	参数量/M	时间/s	FLOPs/G
DPA-Net	9.78	0.0339	106.36
AMP-Net	1.53	0.0322	23.97
TransCS	2.28	0.4258	38.38
TCMformer	7.31	2.4512	20.13

3.3 噪声敏感性测试

在实际的应用中,重建模型可能受到噪声的影响,然而,目前尚无公开的真实数据集能够完全适用于这类CS重建方法。为此,为了评估所提模型的鲁棒性,首先向Set11数据集中的原始图像添加具有不同噪声水平的高斯噪声。然后,TCMformer及其他方法以含噪图像作为输入,并在压缩比为10%和25%时对图像进行采样和恢复。图6展示了在不同标准差噪声下,各方法的PSNR值。实验结果表明,TCMformer对噪声干扰具有很强的鲁棒性。



(a) 采样率=10% (b) 采样率=25%

图6 不同噪声水平下各算法的PSNR值比较

4 结论

本文提出了一种新的图像压缩感知混合网络TCMformer,通过结合CNN与Transformer的优势,显著提升图像重建性能。该方法不仅在图像恢复质量上取得了突破,同时还提高了特征表达能力并有效降低了计算复杂度。此外,设计的特征融合模块为CNN和Transformer之间的特征信息融合提供了有效的机制。本研究不仅为图像压缩感知领域提供了一种新的解决方案,也为如何在其他信号处理任务中有效结合局部和全局特征提供了潜在的思路。TCMformer在图像重建方面取得了显著成效,但在处理大规模数据集和视频重建任务时,计算开销较大,限制了其实际应用。未来的研究将进一步优化模型的计算效率,拓宽TCMformer的应用范围,并为其他复杂任务中的信号重建提供新的思路与方法。

参考文献(References)

[1] Donoho D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[2] 张华, 曹良才, 金国藩, 等. 压缩成像技术的应用与挑战[J]. 科技导报, 2018, 36(10): 20-29.

[3] Duarte M F, Davenport M A, Takhar D, et al. Single-pixel imaging via compressive sampling[J]. IEEE Signal Processing Magazine, 2008, 25(2): 83-91.

[4] Zhu L, Wu X, Sun Z Y, et al. Compressed-sensing accelerated 3-dimensional magnetic resonance cholangiopancreatography: Application in suspected pancreatic diseases[J]. Investigative Radiology, 2018, 53(3): 150-157.

[5] Yuan X, Brady D J, Katsaggelos A K. Snapshot compressive imaging: Theory, algorithms, and applications[J]. IEEE Signal Processing Magazine, 2021, 38(2): 65-88.

[6] Zhang J, Zhao D B, Gao W. Group-based sparse representation for image restoration[J]. IEEE Transactions on Image Processing, 2014, 23(8): 3336-3351.

[7] Dong W S, Shi G M, Li X, et al. Compressive sensing via nonlocal low-rank regularization[J]. IEEE Transactions on Image Processing, 2014, 23(8): 3618-3632.

[8] 山世光, 阚美娜, 刘昕, 等. 深度学习: 多层神经网络的复兴与变革[J]. 科技导报, 2016, 34(14): 60-70.

[9] Shi W Z, Jiang F, Liu S H, et al. Image compressed sensing using convolutional neural network[J]. IEEE Transactions on Image Processing, 2019, 29: 375-388.

[10] Zhang Z, Liu Y, Liu J, et al. AMP-Net: Denoising-based deep unfolding for compressive image sensing[J]. IEEE Transactions on Image Processing, 2020, 30: 1487-1500.

[11] Zhang J, Ghanem B. ISTA-net: Interpretable optimization-inspired deep network for image compressive sensing[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1828-1837.

[12] Kabkab M, Samangouei P, Chellappa R. Task-aware compressed sensing with generative adversarial networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1-8.

[13] Chen J W, Sun Y B, Liu Q S, et al. Learning memory augmented cascading network for compressed sensing of images[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 513-529.

[14] Sun Y B, Yang Y, Liu Q S, et al. Learning non-locally regularized compressed sensing network with half-quadratic splitting[J]. IEEE Transactions on Multimedia, 2020, 22 (12): 3236-3248.

[15] Sun Y B, Chen J W, Liu Q S, et al. Dual-path attention network for compressed sensing image reconstruction[J]. IEEE Transactions on Image Processing, 2020, 29: 9482-9495.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[18] Wang Z D, Cun X D, Bao J M, et al. Uformer: A general U-shaped transformer for image restoration[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 17683-17693.

[19] Jiang Y, Chang S, Wang Z. Transgan: Two pure transformers can make one strong gan, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.

[20] Shen M H, Gan H P, Ning C, et al. TransCS: A transformer-based hybrid architecture for image compressed sensing [J]. IEEE Transactions on Image Processing, 2022, 31: 6991-7005.

[21] Gan L. Block compressed sensing of natural images[C]//Proceedings of 15th International Conference on Digital Signal Processing. Piscataway, NJ: IEEE, 2007: 403-406.

[22] Arbeláez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898-916.

[23] Kulkarni K, Lohit S, Turaga P, et al. ReconNet: Non-iterative reconstruction of images from compressively sensed measurements[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 449-458.

[24] Dong W S, Wang P Y, Yin W T, et al. Denoising prior driven deep neural network for image restoration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2305-2318.

[25] Chen W J, Yang C L, Yang X. FSOINET: Feature-space optimization-inspired network for image compressive sensing[C]//Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 2460-2464.

[26] Chen B, Zhang J. Content-aware scalable deep compressed sensing[J]. IEEE Transactions on Image Processing, 2022, 31: 5412-5426.

A parallel Transformer-CNN network for image compression sensing reconstruction

ZHANG Xinyan[1], ZHU Yongjun[1,2,3*], WU Hongjie[1], ZHOU Fanli[3]

1. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China
2. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
3. Suzhou Tongyuan Software & Control Technology Company, Suzhou 215123, China

Abstract Compressive sensing (CS) for images is an effective technique for signal sampling and reconstruction at low sampling rates. However, achieving high-quality image reconstruction remains to be challenging due to the difficulty in effectively integrating local and global features. To address this issue, a novel image compressive sensing reconstruction framework, a combination of Transformer and the strengths of convolutional neural networks (CNN), namely Transformer-CNN Mixture Transformer (TCMformer) is proposed. The framework leverages CNN for efficient local modeling and Transformers for capturing global context. A dedicated feature fusion module (TCM Block) is designed to bridge local and global features, enhancing the efficiency of feature representation. Additionally, to reduce model complexity and computational cost, the framework adopts a window-based Transformer structure, enabling efficient global modeling through partitioned operations. Furthermore, a progressive reconstruction strategy is introduced to optimize reconstruction quality by utilizing multi-scale feature maps. Experimental results demonstrate that TCMformer significantly outperforms state-of-the-art CS reconstruction methods in terms of peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and visual quality, particularly under low sampling rates. This work provides an effective and practical solution for achieving high-quality image reconstruction in compressive sensing applications.

Keywords compressive sensing; Transformer; convolutional neural networks; image reconstruction

（责任编辑 傅雪）