华东交通大学学报
Journal of East China Jiaotong University

基于渐进式交叉学习的单张图像超分辨率

曾康利[1]，吕亚楠[2]，余 鹰[1]，董文会[1]，岳远昊[1]

（1.华东交通大学信息与软件工程学院，江西 南昌 330013；2.华东交通大学理学院，江西 南昌 330013）

摘要：单张图像超分辨率是一个具有挑战性的不适定问题。当前基于卷积神经网络的方法存在性能瓶颈，而 Transformer 模型虽能通过全局建模提升性能,却受限于计算复杂度难以实现效率平衡。为此,提出了一种渐进式交叉学习网络(CLNet)，通过集成超稠密空洞残差模块（UD2B）与增强型 Transformer 模块（ETB），构建协同工作的渐进式架构。UD2B 通过多尺度空洞卷积聚合高低频特征以增强局部表征，ETB 则通过跨通道自注意力建立长程依赖关系以捕获全局上下文。还提出了跨特征与跨层级注意力融合模块（C2AFB），通过自适应学习实现多层次特征的有效融合。在多个基准数据集上的实验表明，CLNet 在客观指标与视觉感知质量上均优于现有先进方法，实现了性能与效率的较好平衡。

关键词：Transformer；特征融合；交叉学习；注意力机制

中图分类号：TP391　　**文献标志码：**A

Progressive Cross-Learning for Single-Image Super-Resolution

Zeng Kangli[1], Lv Yanan[2], Yu Ying[1], Dong Wenhui[1], Yue Yuanhao[1]

(1. School of Computer Science and Engineering, East China Jiaotong University, Nanchang 330013, China；2. School of Science, East China Jiaotong University, Nanchang 330013, China)

Abstract: Single-image super-resolution is a challenging ill-posed problem. Current methods based on convolutional neural networks face performance bottlenecks, while Transformer models, though capable of improving performance through global modeling, struggle to achieve computational efficiency due to their high computational complexity. Therefore, we propose a progressive Cross-Learning Network (CLNet) that integrates ultra-dense dilated residual blocks (UD2B) with enhanced Transformer blocks (ETB) to construct a synergistic progressive architecture. UD2B aggregates high- and low-frequency features through multi-scale dilated convolutions to enhance local representations, while ETB establishes long-range dependencies via cross-channel self-attention to capture global context. We also introduce a Cross-Feature and Cross-Level Attention Fusion Block (C2AFB) that achieves effective fusion of multi-level features through adaptive learning. Experiments on multiple benchmark datasets demonstrate that CLNet outperforms existing methods in both objective metrics and visual perceptual quality, achieving a favorable balance between performance and efficiency.

Key words: Transformer; feature fusion; cross-learning; attention mechanism

单图像超分辨率（SISR，Single Image Super-Resolution）作为图像增强领域的关键任务，旨在从低分辨率图像中重建出具有清晰纹理的高分辨率图像。近年来，由于高分辨率图像在视觉合理性方面的显著优势，SISR 在计算机视觉中受到广泛关注，并成功应用于医学影像、卫星遥感及安防监控等多个领域。

随着深度学习的发展，基于卷积神经网络的 SISR 方法取得了显著进展。例如，Dong 等人[1]首次提出使用三层卷积构建超分辨率网络；随着 ResNet[2]的提出，残差结构被广泛引入到超分辨率任务，并

收稿日期：2025−11−26

基金项目：国家自然科学基金项目(No. 62501237，62462033，62163016)；江西省自然科学基金项目(No.20252BAC200016，20242BAB25092)；江西省职业早期青年科技人才培养专项项目（No.20244BCE52163）

衍生出多种变体。然而，现有 CNN 方法主要聚焦于局部特征提取，难以建模全局信息，导致重建图像易出现伪影等问题。同时，受限于卷积操作的局部感受野，这些方法在性能与复杂度之间难以取得平衡。近年来，Transformer 模型因其强大的全局建模能力受到关注，最初在自然语言处理中表现优异[3]，随后被引入计算机视觉领域。例如，Wang 等人[5]提出高效超分 Transformer，通过轻量级结构和特征分离策略降低内存消耗；另有研究[4,5,6]借助预训练策略进一步提升模型性能。视觉 Transformer 的核心优势在于其自注意力机制能够建模长程依赖与高阶空间交互，但其高计算复杂度仍是超分辨率任务中的主要挑战。

因此，本文通过协同融合 CNN 与 Transformer 特征，构建了更深更宽的网络以提升 SISR 性能。提出一种联合 CNN 与 Transformer 的交叉学习网络（CLNet），该模型分为四个阶段，每一阶段均融合局部与全局特征进行交叉注意力学习。具体地，设计了超稠密空洞残差卷积块（UD2B），用于优化多层卷积特征，并引入通道注意力机制进行特征重加权。UD2B 中包含多个稠密扩张残差块（D2B），通过密集连接与残差学习可以扩展感受野并加速收敛。此外，Transformer 模块由多个增强型 Transformer 块（ETB）级联组成，各阶段通过调整注意力头数与层数形成从粗到细的渐进式特征优化机制，充分挖掘网络深度与宽度以恢复细节信息。通过交叉特征与跨层注意力融合，所提出方法能够同时建模全局与局部特征关系，在扩展感受野的同时增强上下文表征，从而提升重建质量。

1 基于 CNN 和 Transformer 的交叉学习网络

1.1 总体框架

CLNet的总体框架如图1所示，主要由三个主要模块组成：浅层特征提取模块、非线性映射学习模块和重构模块。非线性映射学习模块分为四个阶段，每个阶段由局部特征学习模块和全局特征学习模块组成。为简单起见，在下面的符号中，X_{LR}、Y_{SR}和Y_{HR}分别表示原始LR输入、CLNet的SR输出和真实HR图像。

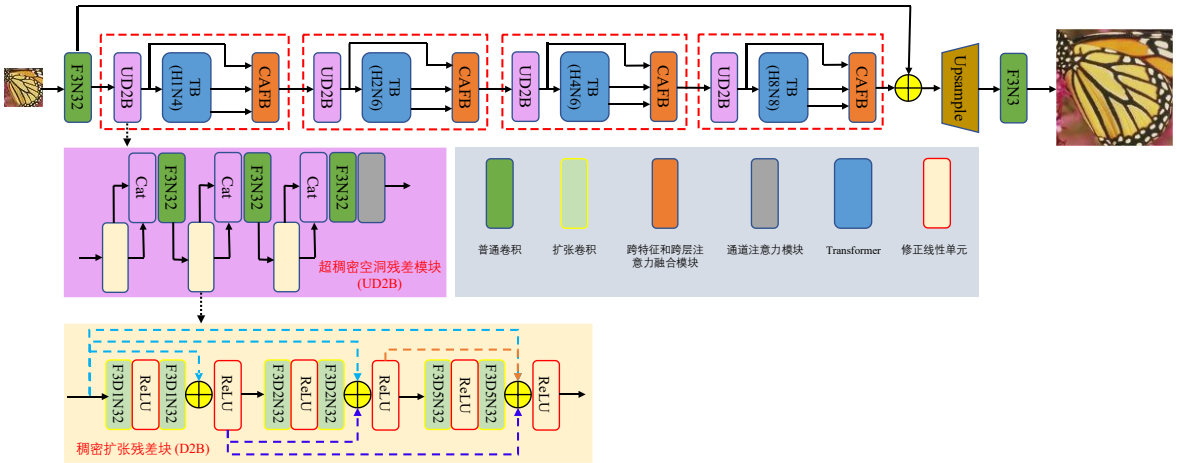

图 1 CLNet 总体框架示意图

Fig.1 Overall framework diagram of CLNet

在浅层特征提取模块中，先使用3×3卷积从输入的X_{LR}中提取初始化特征，$F_0 = \delta\left(f_{SFEM}\left(x_{LR}\right)\right)$，其中 f_{SFEM} 表示一个卷积核为3×3的卷积操作，δ 表示ReLU激活函数。为避免深度网络引发的梯度消失

问题并降低模型优化难度，将F0同时输入残差学习分支与骨干（非线性映射模块）分支。针对残差学习分支，采用等效映射实现跳跃连接；对于骨干分支，则采用四级级联结构逐步构建更深更宽的网络架构，

$$\begin{aligned}
F_1^{s1} &= f_{C2AFB}^{s1}\left(\left[f_{UD2B}^{s1}\left(F_0\right), f_{ETB}^{s1}\left(f_{UD2B}^{s1}\left(F_0\right)\right)\right]\right)\\
F_2^{s2} &= f_{C2AFB}^{s2}\left(\left[f_{UD2B}^{s2}\left(F_1^{s1}\right), f_{ETB}^{s2}\left(f_{UD2B}^{s2}\left(F_1^{s1}\right)\right)\right]\right)\\
F_3^{s3} &= f_{C2AFB}^{s3}\left(\left[f_{UD2B}^{s3}\left(F_2^{s2}\right), f_{ETB}^{s3}\left(f_{UD2B}^{s3}\left(F_2^{s2}\right)\right)\right]\right)\\
F_4^{s4} &= f_{C2AFB}^{s4}\left(\left[f_{UD2B}^{s4}\left(F_3^{s3}\right), f_{ETB}^{s4}\left(f_{UD2B}^{s4}\left(F_3^{s3}\right)\right)\right]\right)
\end{aligned} \tag{1}$$

式中：$F_i^{s_i}$ 表示第i层(i=1、2、3、4)的输出。$F_{UD2B}^{s_i}$ 和 $F_{ETB}^{s_i}$ 分别表示第i层的CNN和Transformer函数。$[\bullet]$ 表示拼接操作，$F_{C2AFB}^{s_i}$ 表示第i个C2AFB函数。然后，将分层特征 $F_4^{s_4}$ 和原始特征 F_0 上采样后的特征进行元素方式相加并应用最后一层卷积来获得最终的SR结果，$y_{HR}=f_{last}\left(\kappa\left(F_0+F_4^{s_4}\right)\right)$，其中 f_{last} 是最后一层3×3卷积，κ 表示子像素操作。

1.2 基于超稠密扩张残差块的局部特征学习

由于模型中的每一层特征在期望的SR结果中都有或多或少的重要作用，本文引入了UD2B来提炼互补性的注意力上下文特征，以实现高质量的SR重建。具体地说，UD2B包含两个基本块：三个密集扩张残差模块(D2B)和一个通道关注层。对于D2B，这里有多个基本的扩张残差模块。因此，第i层DR的输出 f_{DR}^i 可以表示为：

$$F_{DR}^i=\begin{cases}\delta\left(f_{D2}\left(\delta\left(F_{D1}\left(F_{DR}^{i-1}\right)\right)\right)+F_{DR}^{i-1}\right) & i>0\\ F_{input}, & i=0\end{cases} \tag{2}$$

式中：f_{D1} 和 f_{D2} 分别表示每个DR模块中第一个和第二个扩张卷积。F_{input} 表示D2B的输入特征，δ 表示RELU激活功能。通过密集连接，D2B的输出为 $F_{D2B}=\delta\left(f_D\left(F_{DR}^i\right)+F_{DR}^{i-1}+\cdots++F_{DR}^{i-n}\right)$，其中，$f_D$ 是最后一个DR模块的第二个扩张卷积层，n 表示DR模块的总个数。

为了进一步提取有效的表征，本文用D2B代替之前描述的DR，并在最终输出中跟随一个通道注意力层来增强有价值特征的注意力。因此，最终学习的局部特征是

$$F_{UD2B}=CA\left(\delta\left(f_C\left(F_{D2B}^i\right)+F_{D2B}^{i-1}+\cdots+F_{D2B}^{i-n}\right)\right) \tag{3}$$

式中：CA 是通道注意力层。F_{D2B}^i 表示第i个D2B的输出。f_c 是最后一个D2B的第二层卷积操作。通过残差学习和密集连接网络，可以集成后续的所有多级特征。

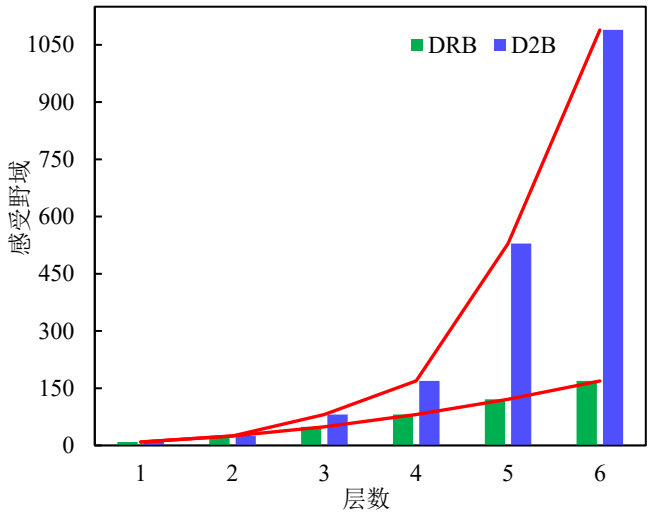

图 2 DRB 和 D2B 的感受野域比较

Fig.2 Receptive field comparison of DRB and D2B

本文使用了六层膨胀卷积，其中包括三个DR。为了避免网格问题，这里的六个膨胀卷积的扩张率依次设置为[1，1，2，2，5，5]。在图2中，比较了DRB和D2B感受野域的变化趋势。可以清楚地看到，(1)随着网络层数的增加，感受野逐渐增加；(2)在所提出的方法中，D2B有效地获得了更大的感受野。

1.3 基于 Transformer 的全局特征学习

受到文献[4,5,6]的启发，本文使用由多头自注意力模块(MSB)和对称门控模块(SGB)组成的Transformer。如图3所示，在每个模块之前使用层归一化，并且在每个模块之后的每个块中也使用残差连接。总体而言，T_o 的输出可以表示为 $T_m = f_{MSB}(T_i), T_o = f_{SGB}(T_m)$，其中 T_i 和 T_m 分别表示Transformer的输入和中间层输出。 $f_{MSB}(\bullet)$ 和 $f_{SGB}(\bullet)$ 分别表示MSB和SGB函数。

1.3.1 多头自注意力模块

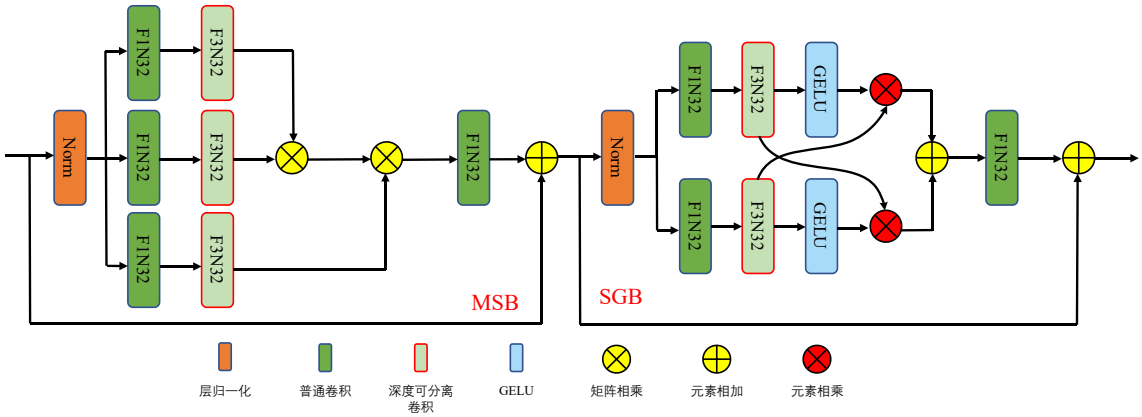

图 3 Transformer 结构示意图

Fig.3 Illustration of the structure of Escalated Transformer

传统的自注意力机制的关键要素是使用跨通道而不是空间维度，即计算跨通道协方差以生成隐含编码全局上下文的注意力。与文献[5]类似，本文分别聚合通道和空间上下文信息，并用深度可分离的卷积

代替普通卷积，因为它能够专注于恢复关键细节的局部上下文特征。总体而言，MSB流程定义为：

$$\begin{aligned}\hat{X} &= W_p Attention\left(\hat{Q},\hat{K},\hat{V}\right)+X \\ Attention\left(\hat{Q},\hat{K},\hat{V}\right) &= \hat{V}Soft\max\left(\hat{K}\hat{Q}/\alpha\right)\end{aligned} \tag{4}$$

式中：$\hat{Q},\hat{K},\hat{V}$ 是 Q,K,V 经过维度调整后的张量。α 是可学习的比例参数，用于在应用Softmax函数之前控制 $\hat{K}$ 和 $\hat{Q}$ 点积的幅度。

1.3.2 对称门控模块

如图3所示，本文使用对称双门控GELU来增强强相关性特征和过滤弱相关性特征，然后使用元素乘法来聚合有助于重建的重要信息。此外，与MSB设计一样，采用1×1卷积和3×3深度可分离卷积来丰富局部信息。因此，SGB可以表示为

$$\begin{aligned}&F_1 = DWConv_{3\times3}\left(Conv_{1\times1}\left(LN\left(F_{in}\right)\right)\right)\quad F_2 = DWConv_{3\times3}\left(Conv_{1\times1}\left(LN\left(F_{in}\right)\right)\right) \\ &F_{gate1} = GELU\left(F_1\right)\quad F_{gate2} = GELU\left(F_2\right) \\ &F_{out} = Conv_{1\times1}\left(F_1\odot F_{gate2} + F_2\odot F_{gate1}\right)+F_{in}\end{aligned} \tag{5}$$

式中：F_{in} 和 F_{out} 分别是输入和输出特征，$\odot$ 表示元素相乘，$LN(\bullet)$ 表示层归一化。

总而言之，SGB允许信息通过多层模块，从而使每一层都能够专注于其他层提供的精细细节，并逐层添加额外的信息。换句话说，SGB通过捕获层次信息使特征更加丰富。与传统的前馈网络相比，所提出的SGB采用对称的双门控机制，从而获得更准确的信息。

1.4 跨特征和跨层注意力融合

为了自适应地聚合不同层之间的有用信息进行重建，本文基于自注意力模块设计了一个跨特征和跨层注意力融合模块（C2AFB）,其结构如图4所示。这样的设计不仅利用了不同层之间的依赖关系来增强表征，而且允许通过自注意力机制来探索不同层特征的相关性，这是因为不同层的激活对特定语义信息有不同的响应。

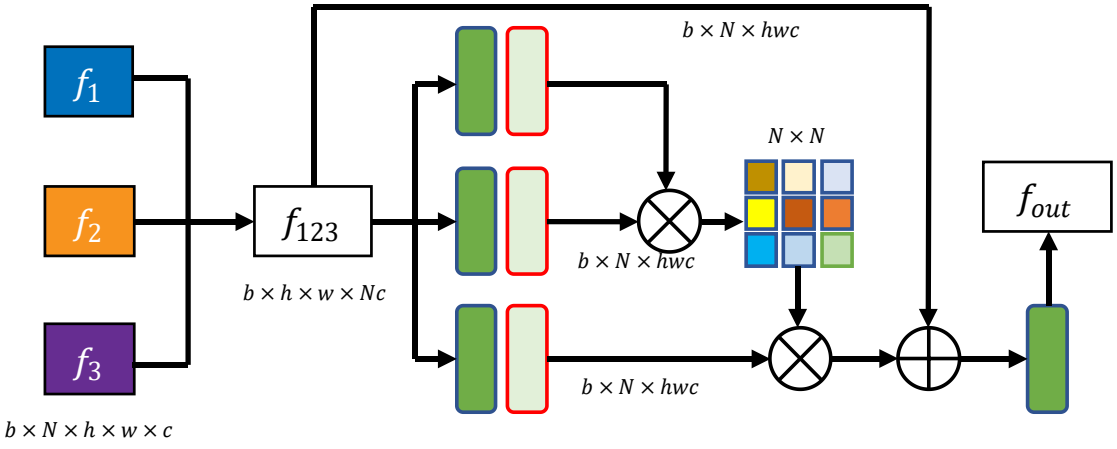

图 4 C2AFB 结构示意图

Fig.4 Illustration of the structure of C2AFB

与MSB一样，C2AFB使用1×1卷积和3×3深度可分离卷积。此外，级联多层次特征可以最大限度地逐层补充分层信息。具体的，对于输入的N个连续层的特征（文中N=3），先对其进行拼接操作并重塑为 f_{123}，随后通过自注意力机制得到融合之后的输出特征，

$$f_{out} = W_{1\times1} f_{C2AFB}(Q,K,V) + f_{123} \tag{6}$$

式中：Q、K、V分别是对输入特征通过 1×1 卷积和 3×3 深度卷积生成的查询、键和值。$W_{1\times1}$ 是 1×1 卷积。f_{out} 是聚焦于网络中信息丰富层的输出特征。

2 实验结果

2.1 数据集和评估指标

本文的训练和验证数据来自高质量(2K分辨率)图像数据集DIV2K[7]，其中前800张图像用于训练，其余图片用于验证。对于测试，本文使用了五个基准测试数据集，包括Set5、Set14、B100、Urban100和Manga109[8]。此外，还利用双三次插值模型得到了LR图像。本文实验结果以PSNR和SSIM[9]作为定量评价指标。

2.2 实验结果比较与分析

2.2.1 定量分析

表1报告了所提出方法与11种先进的SISR方法在五个基准数据集上的定量比较。结果表明，CLNet获得了最好的性能评估，尤其是对于大尺度因子。例如，CLNet在比例因子×4的PSNR分别提高了0.21 dB、0.09 dB、0.05 dB、0.10 dB和0.34 dB。值得注意的是，CLNet的性能表现优于两个比较典型的模型：HGSRCNN和Cross-SRN。其中，HGSRCNN利用不同类型的结构信息，设计了一种用于图像超分辨率的异构组卷积神经网络。Cross-SRN侧重于以多尺度特征融合的方式探索和保存结构信息。对于manga109而言，其他SISR方法受其自身网络结构的影响，对精细边缘和纹理的恢复更具挑战性。然而，CLNet利用具有不同膨胀因子的UD2B结构来有效缓解感受野域的限制，而ETB则弥补了全局信息的缺乏。总之，本文提出的交叉学习框架可以充分结合全局和局部特征，从输入数据集中准确地学习关键层次信息。

如表2所示，本文还比较了基于Transformer的方法(LBNet和ESRT)。可以明显看出，CLNet取得了具有竞争力的性能指标。值得注意的是，尽管LBNet能够细化细节和生成自然纹理，但本文方法在×3尺度因子的性能仍然优于它，其中PSNR提升了0.15 dB，SSIM提升了0.0041。此外，尽管ESRT使用级联的CNN和Transformer架构来探索深层特征，但在大多数指标上仍低于所提出的方法。一方面，CLNet具有独特的双门控前向传播模块，允许获取更重要的信息。另一方面，CLNet结合CNN和Transformer的联合特征，在一个更深更广的网络中提取纹理和细节，从而提高重建图像的质量。

表 1　基于 CNN 的单张图像超分辨率方法在比例因子为 2 的定量比较

Tab.1 Quantitative comparison of CNN-based single image super-resolution methods for scale factors ×2

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM

SeaNet-baseline[10]	37.99	0.9607	33.60	0.9174	32.18	0.8995	32.08	0.9276	38.48	0.9768
MRFN[11]	37.98	0.9611	33.41	0.9159	32.14	0.8997	31.45	0.9221	38.29	0.9759
FilterNet[12]	37.86	0.9610	33.34	0.9150	32.09	0.8990	31.24	0.9200	/	/
HDN[13]	37.75	0.9590	33.49	0.9150	32.03	0.8980	31.87	0.9250	/	/
CFSRCNN[14]	37.79	0.9591	33.51	0.9165	32.11	0.8988	32.07	0.9273	/	/
ESRGCNN[15]	37.79	0.9589	33.48	0.9166	32.08	0.8978	32.02	0.9222	/	/
HGSRCNN[16]	37.80	0.9591	33.56	0.9175	32.12	0.8984	32.21	0.9292	/	/
Cross-SRN[17]	38.03	0.9606	33.62	0.9180	32.19	0.8997	32.28	0.9290	38.75	0.9773
Ours	38.16	0.9612	33.72	0.9200	32.32	0.9016	32.68	0.9336	39.21	0.9782

表 2　基于 CNN 的单张图像超分辨率方法在比例因子为 3 的定量比较

Tab.2 Quantitative comparison of CNN-based single image super-resolution methods for scale factors ×3

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SeaNet-baseline[10]	34.36	0.9280	30.34	0.8428	29.09	0.8053	28.17	0.8527	33.40	0.9444
MRFN[11]	34.21	0.9267	30.03	0.8363	28.99	0.8029	27.53	0.8389	32.82	0.9396
FilterNet[12]	34.08	0.9250	30.03	0.8370	28.95	0.8030	27.55	0.8380	/	/
HDN[13]	34.24	0.9240	30.23	0.8400	28.96	0.8040	27.93	0.8490	/	/
CFSRCNN[14]	34.24	0.9256	30.27	0.8410	29.03	0.8035	28.04	0.8496	/	/
GLADSR[18]	34.41	0.9272	30.37	0.8418	29.08	0.8050	28.24	0.8537	/	/
ESRGCNN[15]	34.24	0.9252	30.29	0.8413	29.05	0.8036	28.14	0.8512	/	/
HGSRCNN[16]	34.35	0.9260	30.32	0.8413	29.09	0.8042	28.29	0.8546	/	/
Cross-SRN[17]	34.43	0.9275	30.33	0.8417	29.09	0.8050	28.23	0.8535	33.65	0.9448
Ours	34.60	0.9285	30.49	0.8444	29.21	0.8082	28.54	0.8600	33.97	0.9472

表 3　基于 CNN 的单张图像超分辨率方法在比例因子为 4 的定量比较

Tab.3 Quantitative comparison of CNN-based single image super-resolution methods for scale factors ×4

方法	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SeaNet-baseline[10]	32.18	0.8948	28.61	0.7822	27.57	0.7359	26.05	0.7896	30.44	0.9088
MRFN[11]	31.90	0.8916	28.31	0.7746	27.43	0.7309	25.46	0.7654	29.57	0.8962
FilterNet[12]	31.74	0.8900	28.27	0.7730	27.39	0.7290	25.53	0.7680	/	/
HDN[13]	32.23	0.8960	28.58	0.7810	27.53	0.7370	29.09	0.7870	/	/
CFSRCNN[14]	32.06	0.8920	28.57	0.7800	27.53	0.7333	26.03	0.7824	/	/
GLADSR[18]	32.14	0.8940	28.62	0.7813	27.59	0.7361	26.12	0.7851	/	/
ESRGCNN[15]	32.02	0.8920	28.57	0.7801	27.57	0.7348	26.10	0.7850	/	/
HGSRCNN[16]	32.13	0.8940	28.62	0.7820	27.60	0.7363	26.27	0.7908	/	/

Cross-SRN[17]	32.24	0.8954	28.59	0.7817	27.58	0.7364	26.16	0.7881	30.53	0.9081
Ours	32.49	0.8981	28.72	0.7844	27.66	0.7395	26.37	0.7942	30.91	0.9129

表 4　基于 Transformer 方法的定量比较

Tab.4 Quantitative comparison of Transformer-based methods

方法	Scale	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LBNet[6]		34.47	0.9277	30.38	0.8417	29.13	0.8061	28.42	0.8559	33.82	0.9460
ESRT[5]	x3	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
Ours		34.60	0.9285	30.49	0.8444	29.21	0.8082	28.54	0.8600	33.97	0.9475
LBNet[6]		32.29	0.8960	28.68	0.7832	27.62	0.7382	26.27	0.7906	30.76	0.9111
ESRT[5]	x4	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
Ours		32.49	0.8981	28.72	0.7844	27.66	0.7395	26.37	0.7942	30.91	0.9129

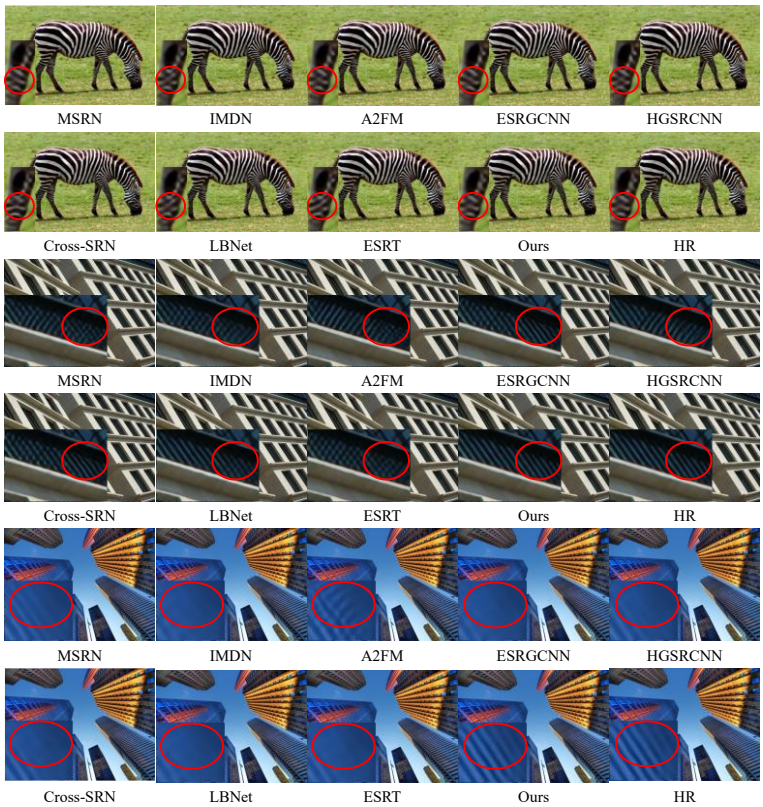

图 5　不同方法在 Set14 和 Urban100 数据集上进行 4 倍超分辨率的可视化结果比较

Fig.5 Visual comparison of different methods on the Set14 and Urban100 datasets with x4

2.2.2 定性分析

如图5所示，可视化了在Set14和Urban100数据集上几个典型样本的×4SR结果的比较。大多数对比方法重建的图像由于缺乏重要细节，导致超分辨率结果不如CLNet重建结果。具体地说，由于网络结构的限制，先前提出的MSRN和IMDN生成了不完整的图像轮廓。虽然近几年的A2FM、ESRGCNN、HGSRCNN和Cross-SRN可以集中在全局轮廓上，但它们产生的图像缺少细节或纹理失真。然而，CLNet

可以通过交叉学习策略来有效恢复纹理。此外，与LBNet和ESRT相比，CLNet的重建图像结构也更加完整和丰富。以图6中第三行和第四行的可视化结果为例，LBNet和ESRT的结果显示出严重的模糊和明显的伪影，而CLNet恢复了清晰的边缘和正确的砖块轮廓。分析表明，CLNet具有强大的纹理恢复和细节重建能力。

2.3 消融实验

2.3.1 UD2B的有效性分析

为了验证UD2B对所提出方法的有效性，以相同的方式进行了五个可比较的移除或添加模块的消融实验。这里的基线是以RB为基本单元，包含6个3×3卷积，其他4个模型也保持6层结构。通过添加密集连接，RB升级为DRB。此外，用膨胀卷积代替了DRB中的普通卷积，就形成了D2B。最后，UD2B通过密集连接集成D2B，并级联一个通道注意力层。表3列出了五个数据集的定量结果。

表 5　不同配置在五个数据集上的 PSNR 和 SSIM 比较
Tab.5 Comparison of PSNR and SSIM for different configurations on five datasets

Model (x2)		Set5	Set14	B100	Urban100	Manga109
RB	PSNR	37.95	33.48	32.20	32.17	38.75
	SSIM	0.9604	0.9177	0.8998	0.9285	0.9772
DRB	PSNR	38.05	33.49	32.24	32.39	38.92
	SSIM	0.9607	0.9189	0.9005	0.9304	0.9776
D2B	PSNR	38.06	33.56	32.26	32.43	38.98
	SSIM	0.9607	0.9195	0.9009	0.9310	0.9777
UD2B(CLNet)	PSNR	38.16	33.72	32.32	32.68	39.21
	SSIM	0.9612	0.9200	0.9016	0.9336	0.9782

如表5所示，通过在残差网络中引入密集连接，其PSNR提升了0.01-0.22 dB。与以前的模型相比，所设计的D2B在PSNR指标上提高了0.06-0.26 dB。特别需要指出的是，基于UD2B构建的方案在所有对比架构中表现最为优异。这主要是因为本文方法采用的渐进式扩张卷积策略、密集连接机制与注意力模块的协同作用，这使其能够有效获取足够大的感受野，从而保障了模型的卓越表现。

2.3.2 H和N的有效性分析

众所周知，Transformer中有两个关键参数：注意力头数(Head)与层数(Number)，它们共同影响着网络的深度与宽度。因此设置了五种参数配置方案，其Head与Number取值分别为：{[1,1,1,1], [1,1,1,1]}、{[1,1,1,1], [4,4,4,4]}、{[1,1,1,1], [4,6,6,8]}、{[4,4,4,4], [4,6,6,8]}以及{[1,2,4,8], [4,6,6,8]}。如图6所示，通过Set5、Urban100和manga109三个数据集上×3超分辨率实验，对比了不同配置下模型的PSNR与SSIM。实验表明，随着注意力头数与层数的增加，模型性能呈现上升趋势。然而，参数规模的增大会不可避免地影响模型效率，这显然不可取。综合平衡性能与效率，本文最终将参数设定为Head=[1,2,4,8]且Number=[4,6,6,8]。

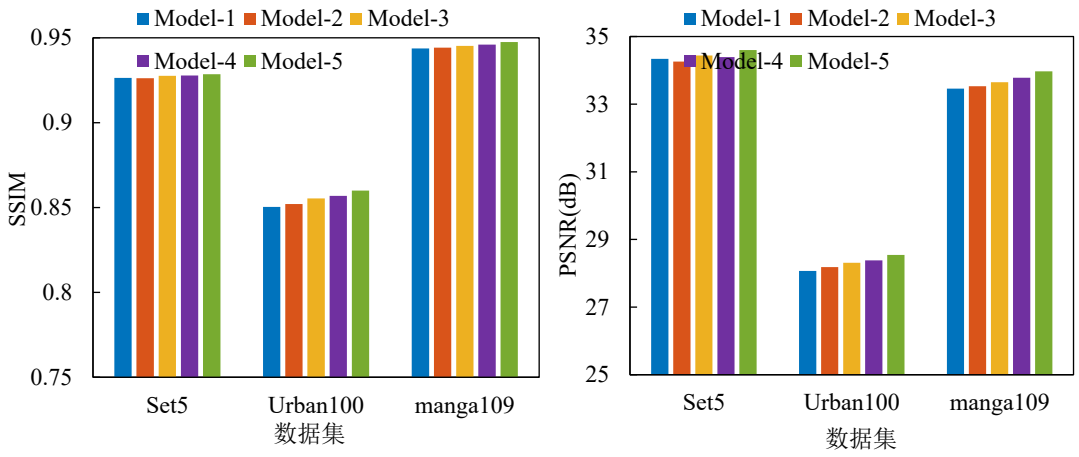

图 6 在 Set5、Urban100 和 manga109 数据集上不同 H 和 N 配置的×3 重建性能比较

Fig.6 Comparison of the reconstruction performance of ×3 for different configurations of H and N on the three datasets (Set5, Urban100 and manga109)

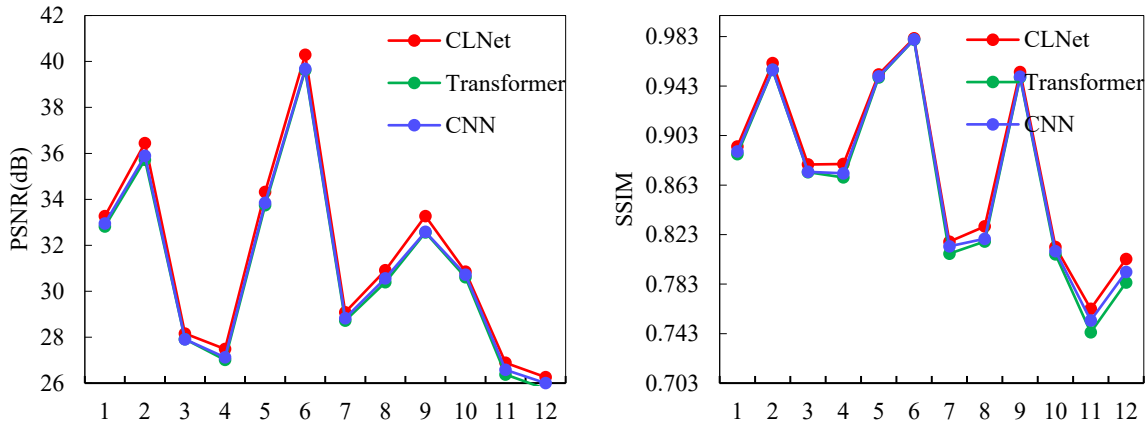

图 7 B100 数据集中 12 张样本图像重建结果的 PSNR 和 SSIM 比较

Fig. 7 Comparison of PSNR and SSIM for Reconstructed Results of 12 Sample Images in the B100 dataset

2.3.3 CNN *VS.* Transformer

为了更好验证局部学习和全局学习的有效性，本文将CLNet中的CNN与Transformer分离，建立基于单特征学习框架进行消融实验。如图7所示，相较于单特征学习方法，所提出的CLNet带来了显著性能提升。CNN架构能有效学习局部特征表示，而后续处理需要捕获全局信息。Transformer架构能够提供具有全局形状和结构信息的特征表示。此外，通过跨特征与跨层注意力机制实现的聚合学习，进一步增强了整体建模能力。

表 6 不同模型的参数，PSNR 和 SSIM 的比较

Tab.6 Parameter number, PSNR and SSIM comparisons on different model

Model	Param.(M)	PSNR	SSIM
ViT	5.60	30.66	0.9103
MSB + GCFN	1.44	30.85	0.9118
MSB + SGB (CLNet)	1.44	30.91	0.9129

表6实验结果表明，本文提出的ETB模块（由MSB与SGB构成）相较于标准Transformer在PSNR指标

上实现了0.25 dB的显著提升。与GDFN的单门控机制相比，本文引入的双门控机制也展现出性能优势。通过引入门控深度卷积，进一步增强了SGB模块的性能。此外，还在Urban100数据集上对比了不同配置下处理100张图像所需的可训练参数量。如表6所示，所提出的CLNet相比标准Transformer具有更高效率，可训练参数量降低74%。

2.3.4 C2AFB的有效性分析

为了验证C2AFB模块的有效性，通过替换或者保留该模块的方式进行了消融实验，具体设置如下：（1）不含C2AFB；（2）特征相加（Add）；（3）特征拼接+卷积（Cat+Conv）；（4）完整C2AFB。表6所示的PSNR与SSIM结果表明，C2AFB对模型性能提升具有显著贡献。当采用具有跨特征交互能力的C2AFB时，模型性能得到显著改善；而简单相加操作带来的性能增益较为有限。

表 7　C2AFB 的消融实验
Tab.7　Ablation study on C2AFB

Dataset	Set5	B100	Urban100
Without	32.05/0.8931	27.52/0.7337	25.92/0.7803
Add	32.13/0.8942	27.57/0.7360	26.09/0.7860
Cat + Conv	32.37/0.8972	27.61/0.7375	26.19/0.7892
With (C2AFB)	32.49/0.8981	27.66/0.7395	26.37/0.7942

2.3.5 讨论

为了深入研究CLNet的有效性，本文在Set5数据集上对比了该方法与前沿方法在推理时间、可训练参数量及PSNR指标方面的表现。如图8所示，尽管提出的CLNet取得了最优的PSNR指标，但其在推理效率与模型复杂度方面的性能仍有待提升，这将是后续研究的重点方向。

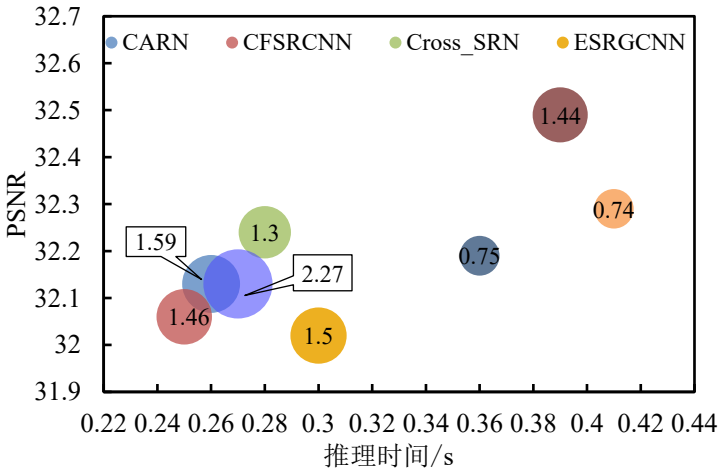

图 8 Set5 数据集上不同方法在 PSNR、推理时间和参数量之间关系的可视化分析
Fig. 8 Visualization of the trade-off between PSNR, inference time and trainable parameters for different methods on the Set5 dataset

3　结论

本文提出一种面向单张图像超分辨率的CNN与Transformer交叉学习网络（CLNet），通过渐进式跨特征与跨层级重组层次化特征，以增强模型的表达能力。具体而言，所设计的UD2B模块通过扩大特征

感知域提取有效局部特征；ETB模块采用跨通道而非空间维度的自注意力机制建模全局上下文；SGB模块引入双门控机制实现可控特征变换。为了高效聚合CNN与Transformer特征，跨特征与跨层级特征注意力融合模块通过自适应加权方式学习鲁棒的特征表示。在五个基准数据集上的大量实验表明，CLNet取得了具有竞争力的性能表现。在未来的研究中，针对实际应用场景中面临低光照、噪声干扰、极端分辨率缩放等情况，将重点关注多模态适配能力，融合红外、偏振等多源信息，增强复杂场景下的特征表达与方法泛化性。

参考文献：

[1] 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, 2015, 38(2): 295-307.
[2] AYOUB A, NAEEM E A, El-SHAFAI W, et al. Video quality enhancement using recursive deep residual learning network[J]. Signal, Image and Video Processing, 2023, 17(1): 257-265.
[3] ZENG K, WANG Z, LU T, et al. Self-attention learning network for face super-resolution[J]. Neural Networks, 2023, 160: 164-174.
[4] 严丽平,张文剥,宋凯,等.基于 Transformer 的交通标志检测模型研究[J].华东交通大学学报,2024,41(1):61-69.
YAN L P, ZHANG W B, SONG K, et al. Research on Traffic Sign Detection Model Based on Transformer[J].Journal of East China Jiaotong University,2024,41(1):61-69
[5] WANG Y, SHAO Z, LU T, et al. A lightweight distillation CNN-transformer architecture for remote sensing image super-resolution[J]. International Journal of Digital Earth, 2023, 16(1): 3560-3579.
[6] CHEN X, WU Y, CHEN J, et al. Efficient face image super‐resolution with convenient alternating projection network[J]. IET Signal Processing, 2023, 17(4): e12205.
[7] SONG H, HAN J, MA H, et al. Edge Priors Guided Deep Unrolling Network for Single Image Super-resolution[J]. Expert Systems with Applications, 2025: 131019.
[8] MATSUI Y, ITO K, ARAMAKI Y, et al. Sketch-based manga retrieval using manga109 dataset[J]. Multimedia tools and applications, 2017, 76(20): 21811-21838.
[9] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.
[10] FANG F, LI J, ZENG T. Soft-edge assisted network for single image super-resolution[J]. IEEE Transactions on Image Processing, 2020, 29: 4656-4668.
[11] HE Z, CAO Y, DU L, et al. MRFN: Multi-receptive-field network for fast and accurate single image super-resolution[J]. IEEE Transactions on Multimedia, 2019, 22(4): 1042-1054.
[12] LI F, BAI H, ZHAO Y. FilterNet: Adaptive information filtering network for accurate and fast image super-resolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1511-1523.
[13] JIANG K, WANG Z, YI P, et al. Hierarchical dense recursive network for image super-resolution[J]. Pattern Recognition, 2020, 107: 107475.
[14] TIAN C, XU Y, ZUO W, et al. Coarse-to-fine CNN for image super-resolution[J]. IEEE Transactions on Multimedia, 2020, 23: 1489-1502.
[15] TIAN C, YUAN Y, ZHANG S, et al. Image super-resolution with an enhanced group convolutional neural network[J]. Neural Networks, 2022, 153: 373-385.
[16] TIAN C, ZHANG Y, ZUO W, et al. A heterogeneous group CNN for image super-resolution[J]. IEEE transactions on neural networks and learning systems, 2022, 35(5): 6507-6519.
[17] LIU Y, JIA Q, FAN X, et al. Cross-SRN: Structure-preserving super-resolution network with cross convolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(8): 4927-4939.
[18] ZHANG X, GAO P, LIU S, et al. Accurate and efficient image super-resolution via global-local adjusting dense network[J]. IEEE Transactions on multimedia, 2020, 23: 1924-1937.

第一作者：曾康利（1992—），男，副教授，博士，硕士生导师，研究方向为人工智能、图像处理。E-mail：zkl_92825@163.com。

通信作者：吕亚楠（1993—），女，讲师，硕士，研究方向为图像处理。E-mail：lvyanan0312@163.com。