Computer Science and Application 计算机科学与应用, 2023, 13(1), 93-103
Published Online January 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.131010

Hans汉斯

基于CNN和Transformer的轻量级超分辨率重建网络研究

李光明[1]，张　倩[2]，金　瑾[1]，何　嘉[1*]

[1]成都信息工程大学计算机学院，四川　成都
[2]活跃网络(成都)有限公司，四川　成都

收稿日期：2022年12月15日；录用日期：2023年1月10日；发布日期：2023年1月19日

摘　要

随着深度学习的发展，单图像超分辨率技术取得了长足的进步。然而，现有的大多数研究都专注于卷积神经网络来构建具有大量层数的更深层次的网络模型。这些方法难以应用于现实场景，因为它们不可避免的伴随着复杂操作所带来的计算和内存成本问题。为此，我们提出了一种用于单图像超分辨率重建的轻量级混合模型——轻量级融合CNN-Swin Transformer网络。具体来说，我们使用带有移动窗口的Swin Transformer块充分学习图像的长期依赖性，并构建了一个基于CNN的局部特征提取块来有效地提取图像的局部特征细节。同时，设计了一个多路径动态卷积块来学习图像的边缘特征。实验结果表明，与基于Transformer的单图像超分辨率模型相比，本文提出的模型取得了更好的结果。

关键词

单图像超分辨率重建，卷积神经网络，Swin Transformer，注意力机制，动态卷积

Research on Lightweight Super-Resolution Network Based on CNN and Transformer

Guangming Li[1], Qian Zhang[2], Jin Jin[1], Jia He[1*]

[1]School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan
[2]Active Network (Chengdu) Co., Ltd., Chengdu Sichuan

Received: Dec. 15th, 2022; accepted: Jan. 10th, 2023; published: Jan. 19th, 2023

*通讯作者。

文章引用: 李光明, 张倩, 金瑾, 何嘉. 基于CNN和Transformer的轻量级超分辨率重建网络研究[J]. 计算机科学与应用, 2023, 13(1): 93-103. DOI: 10.12677/csa.2023.131010

Abstract

With the development of deep learning, single image super-resolution technology has made great progress. However, most of the existing research focuses on convolutional neural networks to construct deeper network models with a large number of layers. These methods are difficult to apply to real-world scenarios because they inevitably come with computational and memory costs associated with complex operations. Therefore, we propose a lightweight hybrid model for super-resolution reconstruction of single image—lightweight fusion CNN-Swin Transformer network. Specifically, we use Swin Transformer block with shifted windows to fully learn the long-term dependence of the image, and build a CNN-based local feature extraction block to effectively extract the local feature details of the image. Meanwhile, a multipath dynamic convolution block is designed to learn the edge features of the image. Experimental results show that compared with the single image super-resolution model based on Transformer, the proposed model achieves better results.

Keywords

Single Image Super-Resolution, Convolutional Neural Network, Swin Transformer, Attention, Dynamic Convolution

1. 引言

单图像超分辨率重建(Single Image Super-Resolution, SISR)旨在从退化的低分辨率(Low Resolution, LR)图像中恢复相应的细节更丰富、视觉质量更好的高分辨率(High Resolution, HR)图像[1]。近年来，基于深度卷积神经网络(Convolutional Neural Network, CNN)的 SR 模型因其在恢复或生成图像高频细节方面的显著性能而流行起来。虽然这些方法已经取得了不错的性能，但由于计算成本高、存储空间大等原因，并不能很好地应用于实际生活中。因此，在保持网络轻量化的同时又能获得更好性能的模型就成为了新的探索方向。常用的策略之一是引入递归机制，如 DRCN [2]和 DRRN [3]。另一个是探索轻量化的网络结构，如 CARN [4]、IDN [5]、IMDN [6]、RFDN [7]等。这些模型都专注于构建更高效的网络结构，在一定程度上减少了模型参数的数量，但也导致了性能的下降，难以重建出边缘细节丰富的图像。

近年来，随着自然语言处理(Natural Language Processing, NLP)中 Transformer [8]技术的不断发展，如何将其应用于计算机视觉任务已成为一个热门话题。Transformer 可以对图像中的长期依赖性进行建模，这种强大的特征表示能力有助于恢复图像的纹理细节。虽然，基于 Transformer 的方法能够更好地提取图像中的长期依赖关系，但是 CNN 提取局部特征的能力仍是不可替代的，这些特征能够在不同的视角下保持自身的稳定性，有助于图像的理解和重建。因此，我们建议将 CNN 和 Transformer 融合，充分利用二者的优点，实现高效的 SR 图像重建。

为此，本文提出一个用于 SISR 的轻量级融合 CNN-Swin Transformer 网络(Lightweight Fusion CNN-Swin Transformer Network, LFCSTN)。在 LFCSTN 中，我们同时使用 CNN 和 Swin Transformer [9]来构建网络结构中的一个分支，称为轻量级融合 CNN-Swin Transformer 模块(Lightweight Fusion CNN-Swin Transformer

Module, LFCSTM)。该模块主要由局部特征提取块(Local Feature Extraction Block, LFEB)、Swin-Transformer块(Swin-Transformer Block, STB)和增强空间注意力(Enhanced Spatial Attention, ESA) [10]组成。对于 CNN 部分，我们专注于局部特征提取，LFEB 主要由级联的卷积层、GeLU 激活函数和高效通道注意力块(Efficient Channel Attention, ECA) [11]组成。对于 Transformer 部分，我们利用 Swin-Transformer 的移动窗口机制来学习图像的长期依赖关系。然后，我们使用增强空间注意力(ESA)来进行上下文信息融合，从而利用学习到的局部信息和全局信息进一步细化纹理细节。同时，为了保留特征图的细节和边缘，我们基于 Chen 等人提出的动态卷积[12]思想设计了一个多路径动态卷积块(Multi-path Dynamic Convolution Block, MDCB)作为我们的边缘增强策略，这有助于图像的最终恢复质量。本文的主要工作：

1) 我们提出了一种新的轻量级混合模型用于 SISR 任务，将 CNN 和 Swin Transformer 相结合，有效的融合了图像丰富的局部和非局部特征。

2) 我们研究了动态卷积技术，并提出了一种面向边缘的多路径动态卷积块作为整体网络结构的一个分支，在可接受的计算成本内，有效提高了模型的性能，并取得了更自然的视觉效果。

2. 相关工作

2.1. 基于 CNN 的 SISR

得益于 CNN 强大的特征表示能力，近年来基于 CNN 的 SISR 方法取得了很大的进展。例如，2014 年，Dong 等人提出的 SRCNN [13]首次将 CNN 应用于 SISR，并在当时取得了极具竞争力的效果。2017 年，Lim 等人在 EDSR [14]中提出了一种增强残差块来训练深度模型，并且去掉了批量归一化层。为了建立更有效的 SISR 模型，RCAN [15]提出了一种具有残差结构和通道注意力机制的深度残差网络。除了这些深度网络，近年来也提出了许多轻量级的 SISR 模型。例如，Ahn 等人利用级联机制提出了 CARN [4]。Zheng 等人利用群卷积提出了 IDN [5]，结合短期和长期特征，对模型大小进行压缩。然后，他们改进了 IDN 的模型结构并提出 IMDN [6]，引入信息多蒸馏块，有效的提取了分级特征。Liu 等人进一步将 IMDN 的信息多蒸馏块改进为残差特征蒸馏块提出了 RFDN [7]。虽然这些模型都取得了较好的结果和视觉表现，但它们都是纯基于 CNN 的模型。这意味着它们只能提取局部特征，不能学习图像的全局信息，并不利于图像纹理和边缘细节的还原。

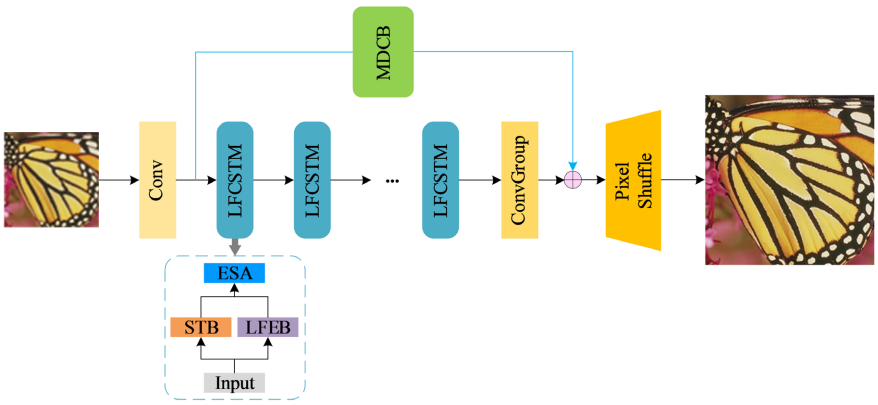

Figure 1. The complete architecture of lightweight fusion CNN-Swin Transformer network (LFCSTN)
图 1. 轻量级融合 CNN-Swin Transformer 网络(LFCSTN)整体结构

2.2. 基于 Transformer 的 SISR

Transformer 在自然语言处理(NLP)中的突破启发了研究者们在计算机视觉任务中使用自注意力(Self At-

tention, SA) [8]机制。Transformer 中的 SA 机制可以有效地捕捉序列元素之间的长期信息，并在一些高级视觉任务中取得了令人映像深刻的结果，如图像分类，图像检测和分割。其中，ViT [16]是第一个用 Transformer 代替标准 CNN 的工作。为了生成序列元素，ViT 将 2D 图像块扁平化成一个向量并将它们输入到 Transformer 中。Chen 等人提出的 IPT [17]是一个非常大的预训练模型，用于基于视觉 Transformer 的各种低级视觉任务中。Liang 等人提出的 SwinIR [18]将 Swin Transformer 直接迁移到图像恢复任务中，并取得了很好的效果。Lu 等人提出的 ESRT [19]通过轻量级 Transformer 和特征分离策略减少了 GPU 内存消耗。但这些模型都没有充分考虑将 CNN 和 Transformer 融合，很难在模型大小和性能之间达到最佳的平衡。

3. 本文方法

3.1. LFCSTN 整体结构

如图 1 所示，LFCSTN 主要有四个部分组成：浅层特征提取、轻量级融合 CNN-Swin Transformer 模块(LFCSTM)、多路径动态卷积块(MDCB)和图像重建。我们将 I_{LR} 和 I_{SR} 分别定义为输入 LR 图像和重建 SR 图像。首先，我们假设给定一个低分辨率图像 $I_{LR} \in R^{H\times W\times C}$，其中 H 和 W 分别为 LR 图像的高度和宽度，C 表示特征通道数。我们使用 3 × 3 卷积层从 I_{LR} 中提取浅层特征 $F_{sf} \in R^{H\times W\times C}$ 为

$$F_{sf} = f_{conv3}\left(I_{LR}\right) \tag{1}$$

其中，$f_{conv3}(\cdot)$ 表示卷积层，F_{sf} 表示提取的浅层特征。然后将提取出来的浅层特征分别作为 LFCSTM 和 MDCB 输入

$$F_n = f_L^n\left(f_L^{n-1}\left(\cdots\left(f_L^1\left(F_{sf}\right)\right)\right)\right) \tag{2}$$

$$F_{ef} = f_M\left(F_{sf}\right) \tag{3}$$

其中，f_L^n 表示第 n 个 LFCSTM 模块的映射，F_n 表示第 n 个 LFCSTM 的输出，f_M 表示 MDCB 的映射，F_{ef} 表示 MDCB 提取到的边缘特征。然后，我们将 F_n 输入到一个卷积组中，进一步提取图像的深层特征

$$F_{df} = f_{CG}\left(F_n\right) \tag{4}$$

其中，F_{df} 表示提取的深层特征，$f_{CG}(\cdot)$ 表示用于提取图像深层特征的卷积组，主要由 3 × 3 卷积层和 LeakyReLU 激活函数组成。最后将 F_{df} 和 F_{ef} 同时输入图像重建模块得到 SR 图像 $I_{SR} \in R^{H\times W\times C}$。

$$I_{SR} = f_p\left(F_{df} + F_{ef}\right) \tag{5}$$

其中 $f_p(\cdot)$ 代表亚像素卷积层。关于 LFCSTM 和 MDCB 更详细的部分，我们将在之后的小节进行介绍。

3.2. Swin Transformer 块(STB)

STB 作为 LFCSTM 中的 Transformer 分支部分，结构如图 2(b)所示。Swin Transformer 的构建方法是将 Transformer 块中的标准多头自注意力(Multi-Head Self-Attention, MSA) [8]模块替换为基于移动窗口的模块，其它层保持不变。首先，我们假设给定大小为 $H\times W\times C$ 的浅层特征 F_{sf} 作为 STB 的输入，Swin Transformer 将输入重塑为 $\frac{HW}{M^2}\times M^2\times C$ 的特征，方法是将输入划分为不重叠的 $M\times M$ 的局部窗口，其中 $\frac{HW}{M^2}$ 是窗口的总数。然后，再为每个窗口分别计算局部注意力，得到局部窗口特征 $F_{lwf} \in R^{M^2\times C}$。对于局部窗口特征 F_{lwf}，查询、键和值矩阵：Q、K 和 V 的计算公式为

$$Q = F_{lwf}P_Q \tag{7}$$

$$K = F_{lwf} P_K \tag{8}$$

$$V = F_{lwf} P_V \tag{9}$$

其中 P_Q、P_K、和 P_V 是在不同窗口间共享的投影矩阵，$Q,K,V \in R^{M^2 \times d}$。

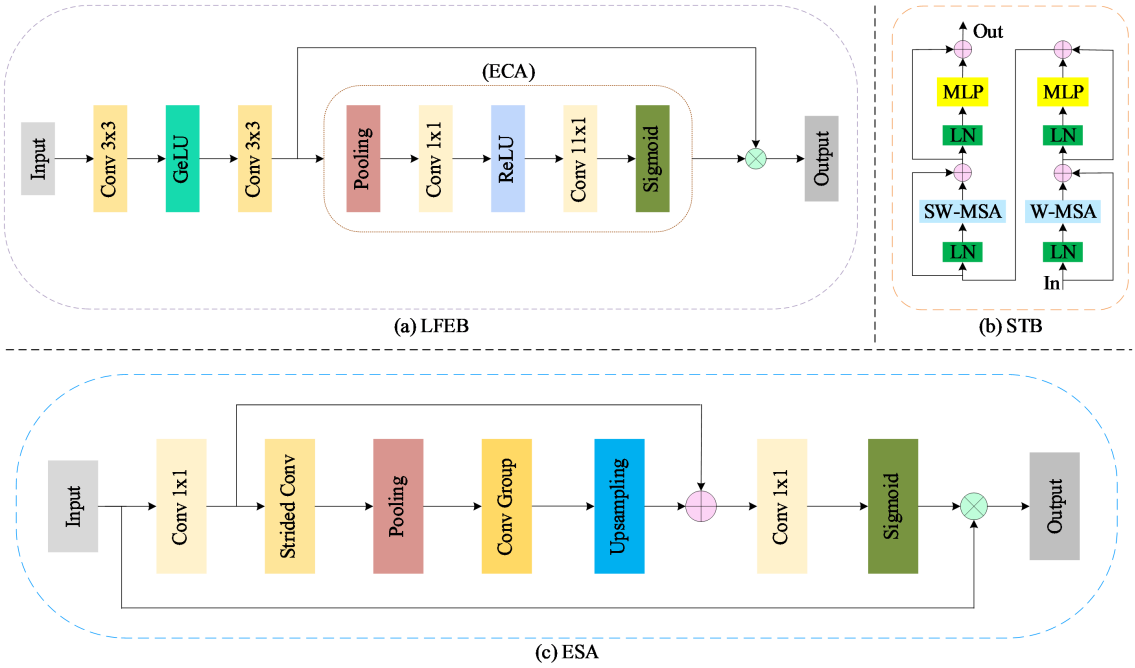

Figure 2. (a) Local feature extraction block (LFEB); (b) Swin Transformer block (STB); (c) Enhanced spatial attention (ESA)
图 2. (a) 局部特征提取块(LFEB)；(b) Swin Transformer 块(STB)；(c) 增强空间注意力(ESA)

然后，通过自注意力机制在局部窗口内计算注意力矩阵

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \tag{10}$$

其中 B 是可学习的相对位置编码，d 为 $\frac{Q}{K}$ 维度。

接下来，使用一个多层感知机(Multi-Layer Perceptron, MLP)层进行进一步的特征转换，MLP 由两个全连接(Fully Connected, FC)层和 GeLU 激活函数组成。在每个 MSA 模块和 MLP 层之前应用了层归一化(Layer Norm, LN)层，并对两个模块使用了残差连接，整个过程可以表示为

$$F_{lwf} = MSA\left(LN\left(F_{lwf}\right)\right) + F_{lwf} \tag{11}$$

$$F_{lwf} = MLP\left(LN\left(F_{lwf}\right)\right) + F_{lwf} \tag{12}$$

基于窗口的自注意力模块会缺乏跨窗口的连接，这限制了网络的整体性能。因为了引入跨窗口连接，同时保持非重叠窗口高效的计算能力，Swin Transformer 的作者提出了一种移动窗口分区的方法。其中，W-MSA 和 SW-MSA 分别表示使用规则和移动窗口的多头自注意力，同时移动窗口分区则意味着在分区之前将特征移动 $\left(\left\lfloor \frac{M}{2} \right\rfloor, \left\lfloor \frac{M}{2} \right\rfloor\right)$ 像素。

3.3. 局部特征提取块(LFEB)

LFEB 作为 LFCSTM 的 CNN 分支部分，其主要作用是对输入图像的浅层特征 F_{sf} 做进一步提取，逐步细化提取的特征，得到局部特征 F_{lf}。在卷积层部分，我们选择使用 GeLU 激活函数，相比于 ReLU 激活函数，GeLU 激活函数引入了随机正则的思想，直观上更符合自然的认识。为了进一步提升特征的表达能力，考虑到模型的性能和复杂度，我们在卷积层之后引入了计算复杂度低且能保持高性能的高效通道注意力(ECA)机制，具体结构如图 2(a)所示。同时，为了缓解梯度消失的问题，我们在卷积层之后加入了残差连接，LFEB 的整体流程可以表示为

$$F_{lf1} = f_{conv3}\left(GELU\left(f_{conv3}\left(F_{sf}\right)\right)\right) \tag{13}$$

$$F_{lf2} = f_{eca}\left(F_{lf1}\right) \tag{14}$$

$$F_{lf} = F_{lf1} \times F_{lf2} \tag{15}$$

其中，$f_{eca}(\cdot)$ 为 ECA 机制。当得到局部特征 F_{lf} 后，我们利用增强空间注意力(ESA)块将 F_{lf} 和 F_{lwf} 进行特征融合，从而得到更精确的特征，这一个步骤可以表示为

$$F_{n} = f_{esa}\left(F_{lf} + F_{lwf}\right) \tag{16}$$

其中 $f_{esa}(\cdot)$ 表示 ESA 机制，结构如图 2(c)所示。至此，我们得到了 LFCSTM 的输出 F_{n} 以作为后续模块的输入。

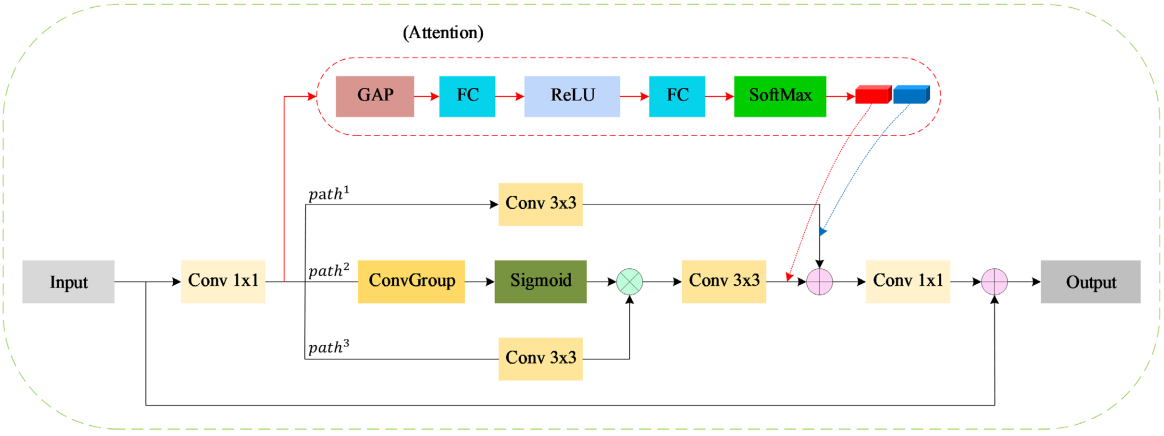

Figure 3. Multi-path Dynamic Convolutional Block
图 3. 多路径动态卷积块(MDCB)

3.4. 多路径动态卷积块(MDCB)

针对参数量较少的轻量级网络，为了更好的学习图像的边缘信息，这需要对模型进行更精细的设计。Lu 等人在 ESRT 中提出了使用高频滤波模块(High-Frequency Filtering Module, HFM) [19]来保留特征图的细节和边缘，这不可避免的增加了模型的计算复杂度和特征冗余，从而导致视觉上不自然的 SR 图像。因此，我们基于动态卷积的思想，设计出了一个多路径动态卷积块(MDCB)作为整体模型的边缘增强分支，目的是有效的提取重要的边缘信息，同时过滤掉无用的特征。我们提出的 MDCB 以可接受的计算成本和参数量获得了比 ESRT 更好的性能和视觉效果。如图 3 所示，给定大小为 $H \times W \times C$ 的浅层特征 F_{sf} 作为输入

$$F_{tf} = f_{conv1}\left(F_{sf}\right) \tag{17}$$

其中 $f_{conv1}(\cdot)$ 表示 1 × 1 卷积层，我们利用 1 × 1 卷积层作为过渡层，得到过渡特征 F_{tf}，这便于后续更有效的提取特征。然后，我们将过渡特征 F_{tf} 作为后续三条路径的输入，其中 path[1] 和 path[3] 由单个的 3 × 3 卷积层组成，path[2] 为卷积组，由级联的 3 × 3 卷积层、LeakyReLU 函数、1 × 1 卷积层和 Sigmoid 函数组成。值得注意的是，path[2] 和 path[3] 将通过逐元素相乘的方式进行特征聚合，再通过一个 3 × 3 卷积层输入到动态卷积部分。至此，我们通过多路径进行特征提取的流程可以表示为

$$F_{tf1} = f_{path1}\left(F_{tf}\right) \tag{18}$$

$$F_{tf23} = f_{conv3}\left(f_{path2}\left(F_{tf}\right)\times\left(f_{path3}\left(F_{tf}\right)\right)\right) \tag{19}$$

其中，F_{tf1} 和 F_{tf23} 分别表示经 path[1]、path[2] 和 path[3] 得到的特征信息，$f_{path1}(\cdot)$、$f_{path2}(\cdot)$ 和 $f_{path3}(\cdot)$ 分别表示 path[1]、path[2] 和 path[3] 路径。在获得 F_{tf1} 和 F_{tf23} 后，我们采用动态卷积来进行边缘信息的学习。如图 3 中 Attention 部分所示，利用注意力机制得到接下来每个卷积核的权重。Attention 部分与 SENet [20]类似，不同点在于最后采用 SoftMax 函数来生成两个自适应权重，动态的调节卷积核参数来提取图像的边缘特征。给定输入过渡特征 F_{tf}，动态卷积的具体操作如下

$$w_1, w_2 = f_{soft}\left(FC\left(ReLU\left(FC\left(f_{gap}\left(F_{tf}\right)\right)\right)\right)\right) \tag{20}$$

$$F_{ef} = f_{conv1}\left(F_{tf23}\times w_1 + F_{tf1}\times w_2\right) + F_{sf} \tag{21}$$

其中，$f_{gap}(\cdot)$ 表示全局平均池化(Global Average Pooling, GAP)操作，$FC(\cdot)$ 表示全连接层，$ReLU(\cdot)$ 表示 ReLU 函数，$f_{soft}(\cdot)$ 表示 SoftMax 函数。w_1 和 w_2 分别表示两个自适应权重值。最后，我们通过一个残差连接得到边缘特征 F_{ef} 用于 SR 图像重建。

4. 实验

4.1. 数据集和评价指标

在本文研究中，我们使用包含 800 张训练图像的 DIV2K [21]数据集来训练我们的 LFCSTN 模型，并使用了 5 个基准数据集，包括 Set5 [22]、Set14 [23]、BSD100 [24]、Urban100 [25]和 Manga109 [26]进行性能比较。本文采用 PSNR 和 SSIM 作为评价指标，评价 SR 图像在 YCbCr 颜色空间 Y 通道上的性能。为了进行性能的比较，我们也提供了模型的参数数量作为参考。

4.2. 训练细节

在训练过程中，我们采用水平翻转和垂直翻转进行数据增强。与此同时，模型使用 Adam 优化器进行训练，$\beta_1 = 0.9$，$\beta_2 = 0.999$，学习率设置为 5×10^{-4}。损失函数采用 L_1 损失函数，与 L_2 损失函数相比，它可以产生更清晰的图像。然后，我们在 PyTorch 框架下使用一个 NVIDIA TITAN RTX 显卡进行模型的训练。

4.3. 与各 SISR 模型对比

在表 1 中，我们将 LFCSTN 与其他 SISR 模型进行了比较：VDSR [27]、LapSRN [28]、EDSR [14]、CARN [4]、IMDN [6]、RFDN-L [7]、ESRT [19]。其中，CARN、IMDN、RFDN-L、ESRT 为轻量级 SISR 模型。显然，我们的 LFCSTN 在所有放大倍数下都获得了最好的客观评价指标结果。可以看出，虽然 CARN 在性能上接近我们的 LFCSTN，但它的参数数量几乎是 LFCSTN 的三倍。同时，RFDN-L 和 ESRT 的参数数量接近 LFCSTN，但 LFCSTN 取得了比它们更好的结果。此外，我们可以观察到 LFCSTN 在放大倍数×3 上取得了最好的结果，这证明了我们的模型的有效性。

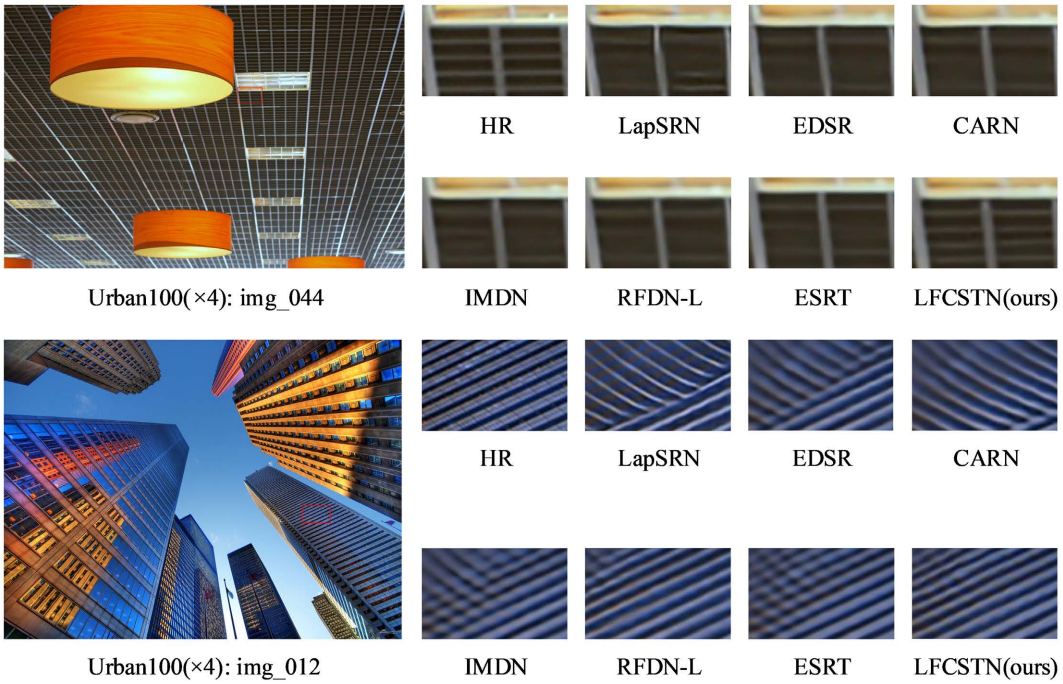

Figure 4. Visual comparison with SISR models
图 4. 与各 SISR 模型的可视化比较

Table 1. Qualitative and quantitative comparisons with SISR models show the best results in bold
表 1. 与各 SISR 模型定性定量比较，最好的结果已加粗表示

Method	Scale	Params	Set5	Set14	BSD100	Urban100	Manga109
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
VDSR	×2	665K	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
LapSRN		812K	37.52/0.9591	33.08/0.9130	31.08/0.8950	30.41/0.9101	37.27/0.9740
EDSR		1370K	37.99/0.9604	33.57/0.9175	32.16/0.8994	31.98/0.9272	38.54/0.9769
CARN		1592K	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
IMDN		694K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
RFDN-L		626K	38.08/0.9606	33.67/0.9190	32.18/0.8996	32.24/0.9290	38.95/0.9773
ESRT		677K	38.03/0.9600	33.75/0.9184	32.25/0.9001	**32.58/0.9318**	**39.12/0.9774**
LFCSTN (ours)		**554K**	**38.08/0.9613**	**33.92/0.9219**	**32.28/0.9018**	32.47/0.9018	39.02/0.9780
VDSR	×3	665K	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279	32.01/0.9310
LapSRN		812K	33.82/0.9227	29.87/0.8320	28.82/0.7980	27.07/0.8280	32.21/0.9350
EDSR		1555K	34.37/0.9270	30.28/0.8417	29.09/0.8052	28.15/0.8527	33.45/0.9439
CARN		1592K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
IMDN		703K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
RFDN-L		633K	34.47/0.9280	30.35/0.8421	29.11/0.8053	28.32/0.8547	33.78/0.9458
ESRT		770K	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
LFCSTN (ours)		**561K**	**34.52/0.9286**	**30.53/0.8460**	**29.18/0.8088**	**28.48/0.8592**	**33.96/0.9469**

Continued

VDSR	×4	665K	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8809
LapSRN		812K	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
EDSR		1518K	32.09/0.8938	28.58/0.7813	27.57/0.7357	26.04/0.7849	30.35/0.9067
CARN		1592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
IMDN		715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
RFDN-L		643K	32.28/0.8957	28.61/0.7818	27.58/0.7363	26.20/0.7883	30.61/0.9096
ESRT		751K	32.19/0.8947	28.69/0.7833	**27.69/0.7379**	26.39/0.7962	30.75/0.9100
LFCSTN (ours)		**571K**	**32.34/0.8972**	**28.72/0.7851**	27.67/0.7412	**26.49/0.7981**	**30.97/0.9142**

在图 4 中，我们还提供了 LFCSTN 与其他 SISR 模型在放大倍数×4 上的视觉比较。视觉上容易观察到，RFDN-L 的效果是和我们最接近的，但其重建出的 SR 图像仍然存在着部分伪影。而我们的 LFCSTN 重建的 SR 图像则包含更少的伪影和更丰富的纹理细节，尤其是边缘和线条。这得益于我们提出的面向边缘的 MDCB 带来的提升，它可以学习到图像更多的边缘信息。这进一步验证了我们提出的 LFCSTN 的有效性。

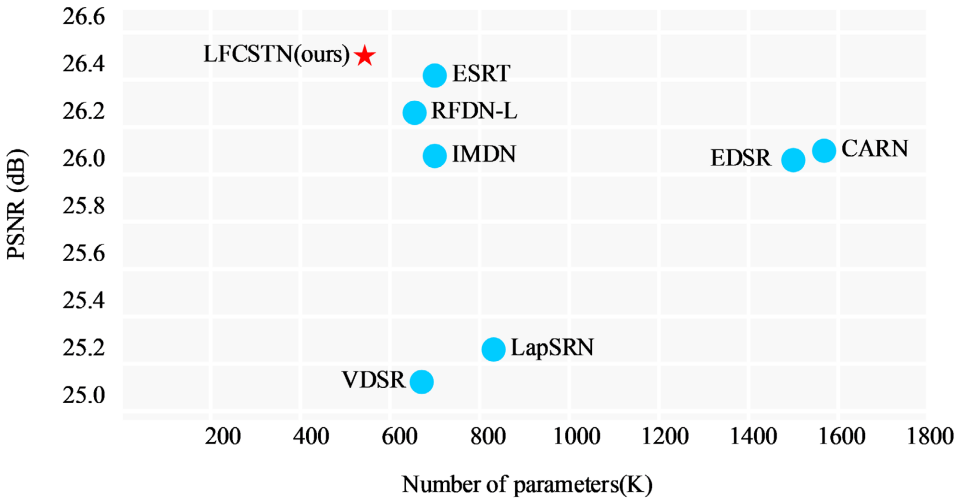

Figure 5. The trade-off between the number of model parameters and performance on Urban100 (×4)
图 5. 在 Urban100 (×4)上模型参数数量和性能之间的比较

此外，如图 5 所示，我们还提供了模型参数数量和性能之间的权衡分析可视化。显然，我们可以观察到 LFCSTN 在模型的大小和性能之间实现了更好的平衡。

4.4. 消融实验

为了验证我们提出的 MDCB 以及 LFCSTM 组件的有效性，我们设计了 3 组对比实验。在表 2 中，我们分析了使用和不使用各组件的模型的性能。可以看出，在边缘信息最丰富的 Urban100 数据集上，我们的 LFCSTN 如果去掉 MDCB，PSNR 值明显从 26.49 下降到 26.25，这说明我们提出的 MDCB 可以很好的处理和学习复杂图像的边缘信息，从而提升模型的重建效果。

5. 结论

在这项工作中，我们为SISR 任务提出了一种轻量级融合CNN-Swin Transformer 网络(LFCSTN)。LFCSTN

首先通过 STB 来学习图像特征的长期依赖性，再利用 LFEB 来提取图像的局部信息，并通过增强空间注意力机制(ESA)进行特征融合。另外，我们提出了一个多路径动态卷积块(MDCB)以较低的计算成本来学习图像的边缘特征，并取得了较好的视觉效果。大量实验表明，我们的 LFCSTN 不仅在模型性和计算成本之间取得了最好的结果，而且重建出的 SR 图像的具有清晰的纹理和边缘细节，整体画面更佳真实且自然。

Table 2. Effects of components on model performance on Urban100 (×4)
表 2. 在 Urban100 (×4)上各组件对模型性能影响

Scale	MDCB	ESA	LFEB	PSNR/SSIM
	×			**26.25/0.7913**
×4	×	×		26.16/0.7890
	×	×	×	26.15/0.7884

基金项目

四川省科技厅项目(编号：2021Z005)。

参考文献

[1] Ledig, C., Theis, L., Huszar, F., *et al*. (2016) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 105-114. https://doi.org/10.1109/CVPR.2017.19

[2] Kim, J., Lee, J.K. and Lee, K.M. (2016) Deeply-Recursive Convolutional Network for Image Super-Resolution. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1637-1645. https://doi.org/10.1109/CVPR.2016.181

[3] Ying, T., Jian, Y. and Liu, X. (2017) Image Super-Resolution via Deep Recursive Residual Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2790-2798.

[4] Ahn, N., Kang, B. and Sohn, K.A. (2018) Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network.

[5] Zheng, H., Wang, X. and Gao, X. (2018) Fast and Accurate Single Image Super-Resolution via Information Distillation Network.

[6] Hui, Z., Gao, X., Yang, Y., *et al*. (2019) Lightweight Image Super-Resolution with Information Multi-Distillation Network. *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 2024-2032. https://doi.org/10.1145/3343031.3351084

[7] Liu, J., Tang, J. and Wu, G. (2020) Residual Feature Distillation Network for Lightweight Image Super-Resolution. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2356-2365. https://doi.org/10.1109/CVPR42600.2020.00243

[8] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need.

[9] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[10] Liu, J., Zhang, W., Tang, Y., *et al*. (2020) Residual Feature Aggregation Network for Image Super-Resolution. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2356-2365. https://doi.org/10.1109/CVPR42600.2020.00243

[11] Wang, Q., Wu, B., Zhu, P., *et al*. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155

[12] Chen, Y., Dai, X., Liu, M., *et al*. (2020) Dynamic Convolution: Attention Over Convolution Kernels. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11027-11036. https://doi.org/10.1109/CVPR42600.2020.01104

[13] Dong, C., Loy, C.C., He, K., *et al*. (2016) Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[14] Lim, B., Son, S., Kim, H., *et al*. (2017) Enhanced Deep Residual Networks for Single Image Super-Resolution. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Honolulu, 21-26 July 2017, 1132-1140. https://doi.org/10.1109/CVPRW.2017.151

[15] Zhang, Y., Li, K., Li, K., *et al*. (2018) Image Super-Resolution Using Very Deep Residual Channel Attention Networks. 15*th European Conference*, Munich, 8-14 September 2018, 294-310. https://doi.org/10.1007/978-3-030-01234-2_18

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[17] Chen, H., Wang, Y., Guo, T., *et al*. (2020) Pre-Trained Image Processing Transformer. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 12294-12305. https://doi.org/10.1109/CVPR46437.2021.01212

[18] Liang, J., Cao, J., Sun, G., *et al*. (2021) SwinIR: Image Restoration Using Swin Transformer. 2021 *IEEE*/*CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[19] Lu, Z., Li, J., Liu, H., *et al*. (2022) Transformer for Single Image Super-Resolution. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Orleans, 19-20 June 2022, 456-465. https://doi.org/10.1109/CVPRW56347.2022.00061

[20] Jie, H., Li, S. and Gang, S. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-23 June 2018, 7132-7141.

[21] Cai, J., Gu, S., Timofte, R., *et al*. (2019) NTIRE 2019 Challenge on Real Image Super-Resolution: Methods and Results. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Long Beach, 16-17 June 2019, 2211-2223.

[22] Bevilacqua, M., Roumy, A., Guillemot, C., *et al*. (2012) Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding. *Proceedings British Machine Vision Conference* 2012, Surrey, 3-7 September 2012, 135.1-135.10. https://doi.org/10.5244/C.26.135

[23] Zeyde, R., Elad, M. and Protter, M. (2010) On Single Image Scale-Up Using Sparse-Representations. *Curves and Surfaces—7th International Conference*, Avignon, 24-30 June 2010, 711-730.

[24] Martin, D., Fowlkes, C., Tal, D., *et al*. (2002) A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *IEEE International Conference on Computer Vision*, Vancouver, 7-14 July 2001, 416-423.

[25] Huang, J.B., Singh, A. and Ahuja, N. (2015) Single Image Super-Resolution from Transformed Self-Exemplars. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 5197-5206. https://doi.org/10.1109/CVPR.2015.7299156

[26] Aizawa, K., Fujimoto, A., Otsubo, A., *et al*. (2020) Building a Manga Dataset "Manga109" with Annotations for Multimedia Applications. *IEEE MultiMedia*, **27**, 8-18. https://doi.org/10.1109/MMUL.2020.2987895

[27] Kim, J., Lee, J.K. and Lee, K.M. (2016) Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *IEEE Conference on Computer Vision & Pattern Recognition*, Las Vegas, 27-30 June 2016, 1646-1654. https://doi.org/10.1109/CVPR.2016.182

[28] Lai, W.S., Huang, J.B., Ahuja, N., *et al*. (2017) Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5835-5843. https://doi.org/10.1109/CVPR.2017.618