Image and Vision Computing 158 (2025) 105519

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Towards trustworthy image super-resolution via symmetrical and recursive artificial neural network

Mingliang Gao [a], Jianhao Sun [a], Qilei Li [b,*], Muhammad Attique Khan [c], Jianrun Shang [a], Xianxun Zhu [d], Gwanggil Jeon [e,*]

[a] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, Shandong, China
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom
[c] Department of Artificial Intelligence, College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia
[d] School of Communication and Information Engineering, Shanghai University, Shanghai, China
[e] Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea

ARTICLE INFO

Keywords:
Image super-resolution
Symmetrical CNN
Recursive Transformer
Long-range dependencies

ABSTRACT

AI-assisted living environments by widely apply the image super-resolution technique to improve the clarity of visual inputs for devices like smart cameras and medical monitors. This increased resolution enables more accurate object recognition, facial identification, and health monitoring, contributing to a safer and more efficient assisted living experience. Although rapid progress has been achieved, most current methods suffer from huge computational costs due to the complex network structures. To address this problem, we propose a symmetrical and recursive transformer network (SRTNet) for efficient image super-resolution via integrating the symmetrical CNN (S-CNN) unit and improved recursive Transformer (IRT) unit. Specifically, the S-CNN unit is equipped with a designed local feature enhancement (LFE) module and a feature distillation attention in attention (FDAA) block to realize efficient feature extraction and utilization. The IRT unit is introduced to capture long-range dependencies and contextual information to guarantee that the reconstruction image preserves high-frequency texture details. Extensive experiments demonstrate that the proposed SRTNet achieves competitive performance regarding reconstruction quality and model complexity compared with the state-of-the-art methods. In the ×2, ×3, and ×4 super-resolution tasks, SRTNet achieves the best performance on the BSD100, Set14, Set5, Manga109, and Urban100 datasets while maintaining low computational complexity.

1. Introduction

AI-assisted living environments widely leverage image super-resolution techniques to enhance the clarity of visual inputs for devices such as smart cameras and medical monitors. This improved resolution enables more precise object recognition, facial identification, and health monitoring, thereby contributing to a safer and more efficient assisted living experience. The affordability and hardware-agnostic nature of image super-resolution (SR) technology have contributed to its widespread use in applications such as video surveillance, military reconnaissance, medical imaging, and intelligent transportation. As a low-level vision task, Image SR is fundamentally ill-posed due to the inherent difficulty in recovering high-resolution (HR) image information from their degraded low-resolution (LR) counterparts [1,2].

Recent advancements in image SR have been driven by CNN-based methods, which leverage their robust feature extraction abilities to achieve excellent performance. Dong et al. [3] made a pioneering contribution by introducing the super-resolution CNN (SRCNN) model, which exhibited outstanding performance. In the follow-up work, with the aid of some powerful learning strategies and mechanisms, *e.g.,* residual learning, attention mechanism, and feedback mechanism, numerous excellent CNN-based methods have been achieved, such as residual dense network (RDN) [4], residual channel attention network (RCAN) [5] and gated multiple feedback network (GMFN) [6].

However, the intricate network structures and large number of parameters in these methods make their implementation difficult in practical scenarios with restricted storage and computing resources. Therefore, an efficient and lightweight model is urgently needed. To this end, some efficient deep learning-based image SR methods that utilized group convolutional network, recursive network, and information distillation mechanism were proposed, *e.g.,* DRCN [7], IDN [8], IMDN [9].

* Corresponding authors.
E-mail addresses: mlgao@sdut.edu.cn (M. Gao), 23504030570@stumail.sdut.edu.cn (J. Sun), qilei@ieee.org (Q. Li), attique.khan@ieee.org (M.A. Khan), 21404020515@stumail.sdut.edu.cn (J. Shang), zhuxianxun@shu.edu.cn (X. Zhu), gjeon@inu.ac.kr (G. Jeon).

https://doi.org/10.1016/j.imavis.2025.105519
Received 30 September 2024; Received in revised form 27 February 2025; Accepted 18 March 2025
Available online 29 March 2025

Despite their ability to reduce computational expense, these methods inevitably compromise performance, especially in reconstructing high-frequency details.

Constrained by the local receptive field, it is difficult for CNN to extract global information [10]. Owing to the significant success of Transformer technology, widespread attention has been paid to the Transformer in the computer vision community [11]. Benefiting from the powerful long-range dependence learning ability, the Transformer can retain more long-distance texture information, which is essential for obtaining high-quality reconstruction images. Transformers have high computational costs, while CNNs possess irreplaceable feature extraction capabilities. We aim to address this by integrating a computationally efficient Transformer unit (IRT) to capture global context. Meanwhile, the powerful local feature extraction ability of CNN-based units (S - CNN) is maintained to conserve details. Although CNN-based methods excel at extracting local features, they struggle to capture long-range dependencies crucial for restoring fine high-frequency details. This limitation results in blurry or artifact-ridden reconstructions, especially in textures and edges. In contrast, Transformer-based methods, while proficient at capturing long-range interactions, tend to have high computational complexity. It makes them unsuitable for resource-constrained environments. Existing hybrid methods attempt to combine the strengths of CNNs and Transformers. However, many approaches still face challenges in achieving the optimal balance between performance and efficiency. Specifically, we need methods that can effectively integrate both local and global information without introducing excessive computational overhead.

To this end, we propose a Symmetrical and Recursive Transformer Network (SRTNet) for efficient image SR. The proposed SRTNet mainly consists of three parts, namely symmetrical CNN (S-CNN) unit, improved recursive Transformer (IRT) unit, and reconstruction module. As mentioned above, CNN plays an irreplaceable role in feature extraction, and the proposed S-CNN unit further strengthens the ability of local feature extraction. Specifically, we designed a local feature enhancement (LFE) module composed of multiple cascaded feature distillation attention in attention (FDAA) blocks. The LFE module contains rich residual connections, which can efficiently transfer feature information. Meanwhile, the symmetrical LFE modules adopt a parameter-sharing mechanism to speed up the feature fusion process. With the help of Attention in Attention (A^2) block, the FDAA block can obtain a wider range of features. Besides, the multi-branch structure is also conducive to enlarging the receptive field. To capture the global similarity features, we introduce the efficient Transformer to learn the image's long-range dependence. Moreover, introducing a recursive mechanism can fully train the Transformer without causing a sharp increase in computational cost. We compared SRTNet with other state-of-the-art (SOTA) methods on the BSD100 [12], Set14 [13], Set5 [14], Manga109 [15], and Urban100 [16] datasets. Experimental results show that SRTNet achieved the best performance at ×2, ×3, and ×4 resolutions while maintaining relatively low computational complexity.

To summarize, the contributions of this paper are as follows:

1. We design an S-CNN unit for local feature extraction and fusion. In the S-CNN unit, an LFE module and FDAA block are employed for feature extraction and utilization.
2. We propose an IRT unit to utilize the global information and refine the high-frequency texture details without increasing the computational cost.
3. We introduce an SRTNet, which achieves efficient image SR by integrating the S-CNN and IRT units. Comprehensive experiments substantiate the efficacy and performance of the SRTNet in terms of reconstruction quality and model complexity.

The rest of this paper is arranged as follows. Section 2 reviews related work on super-resolution based on CNN and Transformer. Section 3 describes the proposed method in detail. Section 4 presents the experimental setup and results. Section 5 concludes the paper and discusses future work.

2. Related work

2.1. Traditional CNN-based image SR

CNN-based method for Image SR approach has quickly risen to prominence in this domain, leveraging the robust feature extraction capabilities of CNNs [17]. These methods generally consist of three steps, *i.e.,* feature representation and expression, SR reconstruction and non-linear mapping. Dong et al. [3] pioneered the use of CNNs in SR with the development of SRCNN. To address the challenge of a confined receptive field resulting from an inadequate network depth, Kim et al. [18] introduced the concept of residual learning and proposed the very deep super resolution (VDSR). Subsequently, Kim et al. [7] expanded the receptive field and built the deeply recursive convolutional network (DRCN) via recursive supervision and skip-connection. Some other networks refine shallow features by introducing feedback mechanisms, such as GMFN [6] and GAMA [19]. However, the methods' applicability to practical scenarios is limited due to their intricate nature and time-consuming processes. High computational complexity and numerous parameters mean that more storage space and stronger processing capabilities are required, which presents challenges for deployment on resource-constrained devices. Complex models typically require longer inference times, which is unacceptable for real-time applications. Furthermore, traditional CNN methods struggle to capture long-range dependencies in images, which are crucial for recovering fine high-frequency texture details. Although some methods attempt to use larger convolution kernels or dilated convolutions to enlarge the receptive field, they often increase computational costs and have limited effectiveness. Although these traditional CNN-based methods laid the foundation for deep learning in SR, they have limitations in capturing long-range dependencies. This has prompted researchers to explore more efficient image super-resolution methods, such as those based on lightweight CNN architectures.

2.2. Efficient CNN-based image SR

Significant attention has been focused on exploring lightweight and efficient image SR methods characterized by a scarcity of parameters and computational load. Hui et al. [8] built an information distillation network (IDN) that can achieve high execution speed with the support of a stacked information distillation block and group convolutional network. To achieve superior performance, Hui et al. [9] introduced IMDN, a lightweight information multi-distillation network incorporating an information multi-distillation block and adaptive cropping strategy. Subsequently, Liu et al. [20] built a residual feature distillation network (RFDN) with feature distillation connections. Ahn et al. [21] introduced the CARN, a cascaded residual network, which employs both global and local cascaded connections. Zhao et al. [22] designed a pixel attention network (PAN) by combining self-calibration convolution with pixel attention. Wang et al. [23] introduced Attentive Auxiliary Features Network (A^2F) which utilizes projection units to combine auxiliary features with current features. However, these methods fall short of ideal performance due to their limited receptive fields and lack of global information, which can result in issues like artifacts in the reconstructed images. Although these efficient CNN-based methods reduce the computational burden, they often compromise reconstruction quality, especially in restoring fine details. To further improve the performance of image super-resolution while maintaining low computational cost, researchers have turned their attention to methods that combine Transformers.

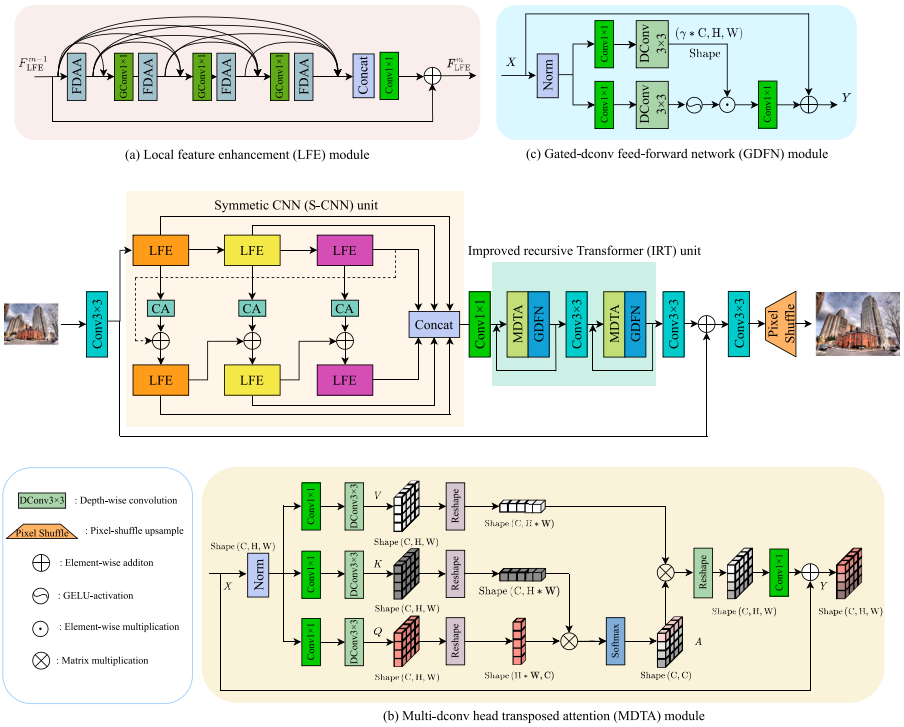

Fig. 1. Architecture of the proposed SRTNet for image SR. (a) Local Feature Enhancement (LFE) module: It is used in the S-CNN unit, where cascaded FDAA blocks are employed to extract and refine local features. (b) Multi-dconv head transposed attention (MDTA) module: It captures long-range dependencies efficiently using a transposed attention mechanism. (c) Gated-dconv feed-forward network (GDFN) module: It refines features from MDTA, focusing on high-frequency details.

2.3. CNN and Transformer-based image SR

Leveraging the strength of Transformers in capturing long-range dependencies, numerous hybrid approaches combining Transformers and CNNs have been proposed for image SR. SwinIR, proposed by Liang et al. [24], achieves high-quality SR reconstruction by leveraging the capability of Swin Transformer to learn long-range image dependencies. To effectively leverage both local and non-local image priors for superior image SR, Fang et al. [25] presented HNCT, a hybrid network incorporating both transformer and CNN. Lu et al. [2] built the coding layer in the Transformer to build an effective SR Transformer (ESRT) and adopted an efficient multi-head attention mechanism to reduce computational cost. Gao et al. [26] introduced a Lightweight Bimodal Network (LBNet) that enables the seamless integration of CNN and Transformer components for improving performance. Inspired by LBNet [26], we intend further to explore the efficient integration of CNN and Transformer to achieve high-quality image SR. However, relying solely on Transformer also has some limitations. The computational complexity of the Transformer is generally high when processing high-resolution images. In addition, Transformer is not as effective as CNN in capturing local detail information. Therefore, how to effectively combine the advantages of CNN and Transformer while overcoming their respective limitations to achieve higher-quality image super-resolution is the focus of this study.

3. The proposed method

3.1. Network framework

The diagram in Fig. 1 outlines the architecture of SRTNet. It encompasses three fundamental elements, *i.e.*, symmetrical CNN (S-CNN) unit, recursive Transformer (RT) unit, and reconstruction module. Specifically, the S-CNN unit is utilized to extract and enhance local features efficiently. The IRT unit is employed to recover high-frequency detail features across long distances. The reconstruction module is utilized for image SR reconstruction.

Let $I_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$ and $I_{\mathrm{HR}} \in \mathbb{R}^{H\times W\times 3}$ represent the input image and output image. The model begins with a 3×3 convolutional layer to derive shallow feature $F_0 \in \mathbb{R}^{H\times W\times C}$, where $H\times W$ symbolizes the spatial dimension. C represents the number of channels. The process of shallow feature extraction is formulated as:

$$F_0 = f_{\mathrm{sf}}\left(I_{\mathrm{LR}}\right), \tag{1}$$

where $f_{\mathrm{sf}}(\cdot)$ denotes the operation of 3×3 convolution for shallow feature extraction. Then, inspired by LBNet [26], the S-CNN unit is employed to extract and enhance local features as:

$$F_{\mathrm{E}} = f_{\text{s-cnn}}\left(F_0\right), \tag{2}$$

where $f_{\text{s-cnn}}(\cdot)$ and F_{E} represent the output of S-CNN unit and the enhanced local features, respectively. The S-CNN unit comprises numerous pairs of LFE modules that share parameters, along with Channel Attention (CA) modules. More details are introduced in Section 3.2.

To reconstruct high-quality images incorporating global similarity features, we introduce an IRT designed to model the long-range dependencies. Since self-attention can significantly decrease computational requirements [27], we obtain an attention map enriched with crucial global context information by leveraging the MDTA module. The GDFN module is used to capture high-frequency texture features, and it can be formulated as:

$$F_{\mathrm{IRT}} = f_{\mathrm{rt}}\left(F_{\mathrm{E}}\right) = f_{\mathrm{gdfn}}\left(f_{\mathrm{mdta}}\left(F_{\mathrm{E}}\right)\right), \tag{3}$$

where $f_{\mathrm{rt}}(\cdot)$, $f_{\mathrm{gdfn}}(\cdot)$ and $f_{\mathrm{mdta}}(\cdot)$ symbolize the operations of IRT unit, GDFN module and MDTA module, respectively. F_{IRT} represents the feature enhanced by global information. After adding the shallow feature F_0 and the refined feature F_{IRT}, the proposed approach leverages a 3×3 convolutional layer in conjunction with a pixel-shuffle layer for image reconstruction. This process is formulated as:

$$I_{\mathrm{SR}} = f_{\mathrm{rm}}\left(F_{\mathrm{IRT}} + F_0\right) = f_{\mathrm{SRTNet}}\left(I_{\mathrm{LR}}\right), \tag{4}$$

where $f_{\mathrm{rm}}(\cdot)$ and $f_{\mathrm{SRTNet}}(\cdot)$ represent the reconstruction module and the proposed SRTNet, respectively. The addition of F_{IRT} and F_0 combines the locally extracted features (enhanced by the S-CNN unit) with

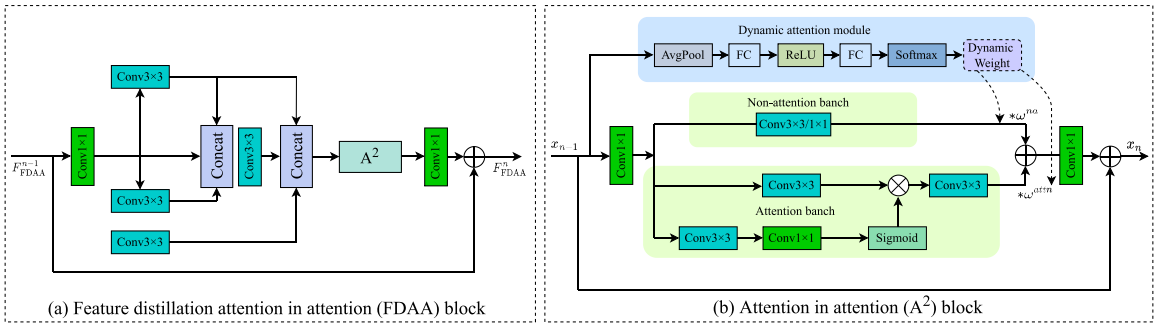

Fig. 2. Architecture of the proposed FDAA block and A^2 block. (a) Feature distillation attention in attention (FDAA) block: It features a multi-branch structure with different numbers of convolutional layers in each branch to capture features at multiple scales. (b) Attention in attention (A^2) block: It consists of an attention branch and a non-attention branch. The outputs of both branches are combined using dynamically learned weights.

the globally refined, high-frequency details captured by the IRT unit. The F_0 convolutional layer then learns to fuse these features and the pixel-shuffle layer upsamples the fused feature map to the desired high-resolution output. This effectively reconstructs the SR image.

The proposed SRTNet model is trained using the L_1 loss function. The L1 loss function calculates the sum of the absolute differences in pixel values between the predicted image and the target image. It is less sensitive to outliers and better preserves high-frequency texture information. The L_1 loss function is also more aligned with the human visual system's evaluation of image quality. Therefore, we chose the L_1 loss function to train the SRTNet model. For a training set $\left\{I^i_{\mathrm{LR}}, I^i_{\mathrm{HR}}\right\}_{i=1}^{N}$, can be expressed as:

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N}\left\|F_{\mathrm{SRTNet}}\left(I^n_{\mathrm{LR}}\right) - I^n_{\mathrm{HR}}\right\|_1, \tag{5}$$

where θ and $\|\cdot\|$ denote the parameter set of the proposed SRTNet and L_1 norm, respectively. N represents the number of images in the dataset.

3.2. Symmetrical CNN

As mentioned in Section 3.1, the S-CNN is adopted for local feature enhancement. It comprises a sequence of local feature enhancement (LFE) modules with shared parameters and Channel Attention (CA) modules. To effectively perform local feature fusion, two symmetrical LFE modules are built in a parameter-sharing mechanism. The design concept of this symmetric structure is that different regions in an image may share similar local feature patterns. By sharing parameters between the symmetric LFE modules, the model can learn more generalized feature representations, which improves feature extraction efficiency and generalization ability. Besides, a CA module is added between each pair of LFE modules to fully mine contributive feature information.

As depicted in Fig. 1, the S-CNN unit contains two branches. The initial feature F_0 is fed to LFE modules in the upper branch, and the outputs of these LFE modules will be used as part of the input of counterparts in the lower branch. The whole process can be formulated as

$$F^{\mathrm{u},1}_{\mathrm{LFE}} = f^{\mathrm{u},1}_{\mathrm{lfe}}\left(F_0\right), \quad i = 1, \tag{6}$$

$$F^{\mathrm{u},i}_{\mathrm{LFE}} = f^{\mathrm{u},i}_{\mathrm{lfe}}\left(F^{\mathrm{u},i-1}_{\mathrm{lfe}}\right), \quad i = 2, \ldots, n, \tag{7}$$

$$F^{\mathrm{l},1}_{\mathrm{LFE}} = f^{\mathrm{l},1}_{\mathrm{lfe}}\left(F^{\mathrm{l},n}_{\mathrm{lfe}} + f^{i}_{\mathrm{ca}}\left(F^{\mathrm{u},1}_{\mathrm{lfe}}\right)\right), \quad i = 1, \tag{8}$$

$$F^{\mathrm{l},i}_{\mathrm{LFE}} = f^{\mathrm{l},i}_{\mathrm{lfe}}\left(F^{\mathrm{l},i-1}_{\mathrm{lfe}} + f^{i}_{\mathrm{ca}}\left(F^{\mathrm{u},i}_{\mathrm{lfe}}\right)\right), \quad i = 2, \ldots, n, \tag{9}$$

where $f^{\mathrm{u},i}_{\mathrm{lfe}}(\cdot)$ symbolizes the ith LFE module in the upper branch and $f^{\mathrm{l},i}_{\mathrm{lfe}}(\cdot)$ represents the corresponding LFF module in the lower branch. $F^{\mathrm{u},i}_{\mathrm{LFE}}$ denotes the outputs of the upper branch, while $F^{\mathrm{l},i}_{\mathrm{LFE}}$ represent the outputs of upper and lower branches. $f^{i}_{\mathrm{ca}}(\cdot)$ denotes the CA module between ith pair of LFE modules. Specifically, the two symmetrical LFE modules adopt the parameter-sharing mechanism, *i.e.*, $f^{\mathrm{u},i}_{\mathrm{lfe}}(\cdot) = f^{\mathrm{l},i}_{\mathrm{lfe}}(\cdot)$. Subsequently, the outputs from all LFE modules are concatenated, and each FDAA block is preceded by a 1×1 group convolutional layer. Finally, the most contributive features are transmitted to the IRT unit to excavate the global contextual features further.

LFE module. The S-CNN unit relies significantly on the essential contributions of the LFE module. As shown in Fig. 1(a), the LFE module has a similar network architecture with residual dense block (RDB) [28]. The difference is that we introduce an FDAA block at the head of the RDB to enhance the feature extraction capability. Although the ReLU activation function is simple and efficient, its hard cutoff at zero may lead to information loss. The FDAA module better preserves and refines feature information through its multi-branch structure and attention mechanism. This multi-branch design promotes feature diversity, with each branch learning a slightly different representation of the input. Thus it leads to a richer and more comprehensive feature set. At the same time, it can also avoid the loss of negative value information caused by the characteristics of the ReLU activation function. Then, we replace the ReLU activation function layer in RDB with an FDAA block to further transfer the feature information. Furthermore, for dimensionality reduction, a 1×1 group convolutional layer is incorporated before FDAA block. The procedure of the LFE module is formulated as

$$F^{1}_{\mathrm{FDAA}} = f^{1}_{\mathrm{fdaa}}\left(F^{m-1}_{\mathrm{LFE}}\right), \tag{10}$$

$$F^{2}_{\mathrm{FDAA}} = f^{2}_{\mathrm{fdaa}}\left(f^{1}_{\mathrm{gc}}\left(\left[F^{m-1}_{\mathrm{LFE}}, F^{1}_{\mathrm{FDAA}}\right]\right)\right), \tag{11}$$

$$F^{3}_{\mathrm{FDAA}} = f^{3}_{\mathrm{fdaa}}\left(f^{2}_{\mathrm{gc}}\left(\left[F^{m-1}_{\mathrm{LFE}}, F^{1}_{\mathrm{FDAA}}, F^{2}_{\mathrm{FDAA}}\right]\right)\right), \tag{12}$$

$$F^{m}_{\mathrm{LFE}} = F^{m-1}_{\mathrm{LFE}} + f_{1\times1}\left(\left[F^{m-1}_{\mathrm{LFE}}, F^{1}_{\mathrm{FDAA}}, F^{2}_{\mathrm{FDAA}}, F^{3}_{\mathrm{FDAA}}\right]\right), \tag{13}$$

where $f^{n}_{\mathrm{fdaa}}(\cdot)$ and F^{n}_{FDAA} denote the nth ($n = 1, 2, 3$) FDAA block in LFE module and its output, respectively. $f^{i}_{\mathrm{gc}}(\cdot)$ is the ith ($i = 1, 2$) group convolutional layer in front of FDAA block. $f_{1\times1}(\cdot)$ represents the 1×1 convolutional layer. F^{m}_{LFE} is the output of the LFE module.

FDAA block. As illustrated in Fig. 2(a), the FDAA block adopts a multi-branch network structure specially designed for feature distillation and refinement. Besides, the Attention in Attention (A^2) block [29] is introduced to utilize a wider range of feature information. To obtain different scale features, we apply different numbers of convolution layers in the two branches of FDAA block to change the receptive field size. Then, the A^2 block accentuates high-contributing information while mitigating redundant information. After passing through the global residual path, the input features are combined with the output of the A^2 block.

A^2 block. The A^2 block proposed by Chen et al. [29] was proved to enhance the capacity of the attention network without increasing

the parameter count significantly. As illustrated in Fig. 2(b), the A^2 block features the attention branch and the non-attention branch. While the former focuses on refining significant features, the latter acts to retain detailed information that the attention branch may disregard. The A^2 block specifically incorporates a dynamic attention module that discards non-essential attention features and optimizes the branches. We x_n denote the input to the A^2 block and x_{n+1} denote the output of the A^2 block. Formally, this operation can be defined as

$$x_{n+1} = f_{1\times1}\left(\omega_n^{\mathrm{a}} \times x_n^{\mathrm{a}} + \omega_n^{\mathrm{na}} \times x_n^{\mathrm{na}}\right), \tag{14}$$

where ω_n^{a} denotes the weight of the attention branch, and ω_n^{na} denotes the weight of the non-attention branch. It is worth noting that we have weighted summation $\omega_n^{\mathrm{a}} + \omega_n^{\mathrm{na}} = 1$. $f_{1\times1}(\cdot)$ represents a 1×1 convolution operation. $x^{\mathrm{a}}n$ represents the output of the attention branch, while $x^{\mathrm{na}}n$ denotes the output of the non-attention branch. Similarly, dynamic weights can be defined as

$$\omega_n = f_{\mathrm{da}}\left(x_n\right), \tag{15}$$

where $f_{\mathrm{da}}(\cdot)$ denotes the function of the dynamic attention module. Specifically, it adjusts the dynamic weighted contributions of the attention and non-attention branches via weighted summation. With the help of the dynamic attention module, the vital information in the final refined feature is highlighted while the redundancy is reduced.

3.3. Improved recursive transformer

The reconstruction of a high-quality image relies heavily on global information. However, lightweight networks are constrained by their depth and typically lack sufficient receptive fields to capture global information. To overcome this limitation, the RT unit was proposed in LBNet [26]. In this work, we proposed an improved recursive Transformer (IRT) to get long-range texture features. Different from the LBNet [26], the MDTA and GDFN modules are utilized to form the Transformer. The MDTA module facilitates high-frequency detail recovery by incorporating long-range texture information. MDTA uses a transposed attention mechanism to flatten the spatial dimensions before performing attention computation. This reduces computational complexity while retaining the ability to model long-range dependencies. The GDFN module manages the exchange of information across hierarchical levels, and it facilitates the specialization of each level towards details that are complementary to other levels. By combining the MDTA and GDFN modules, IRT can capture global context information and high-frequency texture features. The recursive mechanism enhances the expressive power of the IRT, which enables information to flow between multiple IRT units. Since each recursive step processes the same feature map size, the recursive mechanism does not significantly increase computational cost. Ablation studies in Section 4.4 are conducted to extensively evaluate the impact and effectiveness of SRTNet.

As depicted in Fig. 1, the IRT unit comprises two 3×3 convolution layers and two efficient Transformers. The Transformer consists of MDTA and GDFN modules. The architecture of MDTA module and GDFN modules are illustrated in Fig. 1(b) and (c). Then, the function of the IRT unit is defined as,

$$F_{\mathrm{IRT}} = f_{3\times3}\left(F_{\mathrm{et2}}^{r}\left(f_{3\times3}\left(F_{\mathrm{et1}}^{r}\left(F_{\mathrm{E}}\right)\right)\right)\right), \tag{16}$$

where $f_{3\times3}(\cdot)$ denotes the 3×3 convolution layer and $F_{\mathrm{et}}^{r}(\cdot)$ symbolizes the operator of efficient Transformer with the recursive connection.

MDTA module. The MDTA module is a three-branch network with pixel convolution and deep-wise convolution at the front of each branch. Given the input feature $X \in \mathbb{R}^{H\times W\times C}$, the query ($Q$), key ($K$) and value ($V$) are formulated as:

$$Q = F_{\mathrm{dc}}^{1}\left(F_{\mathrm{pc}}^{1}\left(F_{\ln}(X)\right)\right), \tag{17}$$

$$K = F_{\mathrm{dc}}^{2}\left(F_{\mathrm{pc}}^{2}\left(F_{\ln}(X)\right)\right), \tag{18}$$

$$V = F_{\mathrm{dc}}^{3}\left(F_{\mathrm{pc}}^{3}\left(F_{\ln}(X)\right)\right), \tag{19}$$

where $F_{\ln}(\cdot)$ represents the layer normalization, $F_{\mathrm{pc}}(\cdot)$ denotes the pixel convolution and $F_{\mathrm{dc}}(\cdot)$ is the deep-wise convolution. $\hat{Q} \in \mathbb{R}^{HW\times C}$, $\hat{K} \in \mathbb{R}^{C\times HW}$, and $\hat{V} \in \mathbb{R}^{HW\times C}$ can be acquired after performing the reshape operation on Q, K, and V respectively. Subsequently, a transposed attention map $A \in \mathbb{R}^{C\times C}$ and output $Y \in \mathbb{R}^{C\times H\times W}$ can be obtained via a series of matrix multiplications. The complete operation is formulated as:

$$A = \mathrm{Softmax}\left(\hat{K} \cdot \hat{Q}/\alpha\right), \tag{20}$$

$$Y = F_{\mathrm{pc}}\left(F_{\mathrm{rs}}\left(A \cdot \hat{V}\right)\right) + X, \tag{21}$$

where Softmax and $F_{\mathrm{rs}}(\cdot)$ symbolize the Softmax function and reshape operation, respectively. α means learnable scaling parameters. MDTA module can perform local and non-local correlation pixel interactions and help the network retain more comprehensive contextual information.

GDFN module. To further capture the rich high-frequency detail information, we adopt the GDFN module behind the MDTA module. The MDTA module can perform controlled feature changes that allow each layer to focus on different detailed information. The GDFN module further refines the features, which enhances the high-frequency details necessary for clear image reconstruction. The MDTA module captures long-range dependencies and provides global contextual information. With the input feature $X \in \mathbb{R}^{H\times W\times C}$, the operation of GDFN module can be defined as

$$X_{\mathrm{GDFN}}^{1} = \delta\left(F_{\mathrm{dc}}\left(F_{\mathrm{pc}}\left(F_{\ln}(X)\right)\right)\right), \tag{22}$$

$$X_{\mathrm{GDFN}}^{2} = F_{\mathrm{dc}}\left(F_{\mathrm{pc}}\left(F_{\ln}(X)\right)\right), \tag{23}$$

$$Y_{\mathrm{GDFN}} = X_{\mathrm{GDFN}}^{1} \odot X_{\mathrm{GDFN}}^{2}, \tag{24}$$

$$Y = F_{\mathrm{pc}}\left(Y_{\mathrm{GDFN}}\right), \tag{25}$$

where $\delta(\cdot)$ means the operation of GELU non-linearity and $\odot$ represents element-wise multiplication. With the help of the MDTA and GDFN modules, the proposed IRT unit enables both global information learning and efficient capture of high-frequency texture details.

4. Experiments and discussion

4.1. Datasets and evaluation metrics

The training phase of SRTNet utilizes the DIV2K [30] dataset. The DIV2K dataset is a high-quality image dataset specifically designed for image restoration tasks. The dataset consists of 800 training images, 100 validation images, and 100 test images, with resolutions that range from 1080p to 2K. This helps the model learn feature representations at different scales and improves its generalization ability. Compared to other commonly used super-resolution datasets, the DIV2K dataset has higher image quality It is closer to real-world images, which allows for better training of more practical super-resolution models. Subsequently, in the evaluation phase, five widely recognized datasets, *i.e.,* BSD100 [12], Set14 [13], Set5 [14], Manga109 [15] and Urban100 [16] are employed. The BSD100 [12] dataset comprises a diverse array of images. It encompasses a spectrum from natural scenery to specific subjects like botanical elements, individuals, and culinary items. The Set14 [13] dataset contains 14 images and the Set5 [14] dataset contains five images. The Manga109 [15] dataset contains 109 cartoons drawn by professional Japanese cartoonists and is generally only used to test performance. The Urban100 [16] dataset comprises 100 challenging images of urban landscapes with different frequency band details. BSD100 and Set5/Set14 primarily focus on

natural scenes and general images. They are suitable for testing the algorithm's ability to process natural images. Manga109 features a more artistic and complex image style, while Urban100 is suitable for testing the algorithm's enhancement of architectural and street details. These differences reflect the emphasis of each dataset in their respective application scenarios.

The objective results are evaluated using two metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [31]. The PSNR evaluates the quality of image reconstruction by analyzing the ratio of peak signal power to noise power, which is correlated with lower distortion levels at higher PSNR values. The SSIM assesses the similarity between two images by comparing their luminance, contrast, and structure, with values nearing 1 indicating a greater similarity between the images. The calculation methods for PSNR and SSIM are as follows:

$$\mathrm{PSNR}\,(x,y) = 10 \cdot \log 10 \times \left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2} \right), \tag{26}$$

where L represents the maximum pixel, and N denotes the number of all pixels in I_{LR} and I_{HR}.

$$\mathrm{SSIM}(x,y) = \frac{2\mu_x\mu_y + k_1}{\mu_x^2 + \mu_y^2 + k_1} \cdot \frac{\sigma_{xy} + k_2}{\sigma_x^2 + \sigma_y^2 + k_2}, \tag{27}$$

where x and y represent two images. $\sigma_{x,y}$ symbolizes the covariance between x and y. μ and σ represent the average value and variance. k_1 and k_2 denote constant relaxation terms.

We also evaluate computational complexity [23,32] by employing Multi-Adds and model parameters.

4.2. Implementation details

To effectively train the proposed SRTNet, high-resolution images are downsampled via bicubic interpolation. We apply random rotations of $90° \times n$ (n = 1, 2, 3) and horizontal mirroring to augment the training sample set. In the training phase, the data is randomly cropped into image patches measuring 48×48. However, no resizing operations are applied during the training phase before the image is randomly cropped to 48×. The Adam [33] algorithm is utilized to enhance the model, with the initial learning rate being set to 2×10^{-4}. The choice of learning rate is based on the setting for the learning rate in LBNet [26], as well as related empirical experience. We found that it effectively learns image features while ensuring stable convergence of the model. We conducted our experiments using the PyTorch framework, with training and testing performed on two NVIDIA 3090 Ti GPUs in parallel. The experimental setup utilized CUDA Toolkit 11.4, cuDNN 8.2.2, and Python 3.8. We use six LFE modules to build the final model, which is termed SRTNet in this work.

4.3. Comparisons with SOTA methods

This section presents a comparative analysis of SRTNet against 13 state-of-the-art methods. We evaluate their performance on five benchmark datasets., *i.e.,* BSD100 [12], Set14 [13], Set5 [14], Manga109 [15] and Urban100 [16]. Among these methods, SRCNN [3], VDSR [18], DRCN [7], FSRCNN [34], and DRRN [35] are traditional CNN-based SR methods. CARN [21], IDN [8], IMDN [9], PAN [22], RFDN [20] and A^2F-M [23] are efficient CNN-based SR methods. LBNet [26] and ESRT [2] are the transformer CNN-based SR methods. Objective image reconstruction quality results for scaling factors of ×2, ×3, and ×4 are presented in Tables 1, 2, and 3, respectively.

With a scaling factor of ×2, SRTNet achieves the highest performance scores, as evident in Table 1. Specifically, compared with SRCNN [3] and VDSR [18], the SRTNet improves the PSNR by 0.61 dB and 1.48 dB, respectively. PSNR and SSIM evaluations demonstrate that the SRTNet outperforms DRCN [7]. SRTNet outperforms all other methods in terms of PSNR and SSIM scores for a scale factor of ×3, as demonstrated in Table 2. Specifically, the results indicate that SRTNet surpasses SOTA lightweight CNN-based SR methods, *i.e.,* PAN [22], RFDN [20] and A^2F-M [23]. As shown in Table 3, the proposed SRTNet ranks first in all datasets except Urban100, where it is second in PSNR and SSIM. At a scaling factor of ×4, SRTNet demonstrates superior overall performance compared to Transformer-based lightweight methods, specifically LBNet [26] and ESRT [2].

Furthermore, the visual analysis of trade-offs between performance, Multi-Adds and Parameters are depicted in Fig. 3. SRTNet is positioned in a favorable region of the chart, which demonstrates an excellent trade-off. It achieves high PSNR, comparable to or surpassing many computationally intensive methods. Meanwhile, it maintains relatively low Multi-Adds and parameters. This demonstrates that SRTNet strikes an effective balance between performance and model complexity.

The merits of the SRTNet are further substantiated by a subjective analysis of reconstructed SR image quality. Fig. 4 visually compares the results obtained using the different methods discussed. It shows that the proposed SRTNet successfully reconstructs images with fine-grained texture details. As shown in the architectural image "img_005", the Bicubic method produces a relatively blurred image. While the SRCNN and FSRCNN methods restore some details, they still exhibit noticeable noise and artifacts. The reconstruction quality of the CARN, IDN, IMDN, and PAN methods is improved but still inferior to SRTNet. The ESRT method achieves comparable detail restoration in certain areas to SRTNet, but it is still slightly inferior in overall sharpness and texture fidelity. As illustrated by the "baby" image from Set5, the images reconstructed by the methods of SRCNN [3], IDN [8], and CARN [21] have certain degrees of distortion and aliasing. Although the reconstruction results of the IMDN [9], PAN [22] and ESRT [2] possess good outlines, they lack fine high-frequency details, particularly sharp edges. In contrast, the image reconstructed by the SRTNet has a stronger ability to maintain fine texture features, benefiting from the global information learning ability of the IRT unit.

4.4. Ablation studies

Ablation study on LFE module. To improve feature representation, the LFE module incorporates multiple FDAA blocks in a cascaded configuration. To demonstrate the effectiveness of FDAA block, we use some feature extraction modules widely used in lightweight SR algorithms, including RCAB [5], IMDB [9], SCPA [22] and FRDAB [26], to replace the proposed FDAA block for ablation study. Table 4 reports the comparison results for Parameters, Multi-Adds, PSNR, and SSIM. Among the blocks tested, the FDAA block achieved the highest PSNR and SSIM scores with a minimal increase in parameters. These results substantiate the role of the FDAA block in significantly improving image reconstruction accuracy. Compared to the results of the RCAB block, IMDB block, SCPA block, and FRDAB block, the inclusion of only the FDAA increases the PSNR by 3.04%, 2.12%, 1.45%, and 0.29%, respectively. The multi-branch structure and attention mechanism of the FDAA module can effectively extract and fuse local features.

Ablation study on FDAA block. To evaluate the advantages of the FDAA block further, we conducted an ablation study on the feature fusion scheme located at the end. At the end of the FDAA block for ablation investigation, we add spatial attention (SA), channel attention (CA), pixel attention (PA), and attention in attention (A^2) block and a combination of them. The ablation study results are summarized in Table 5. The results indicate that integrating the A^2 block enables the model to achieve optimal performance, with only a marginal increase in parameters. Compared to the results of the CA block, SA block, PA block, and the combination of CA and SA blocks, the inclusion of the A^2 block increases the PSNR by 1.51%, 2.18%, 0.38%, and 1.39%, respectively. The SSIM increases by 1.62%, 2.37%, 0.41%, and 1.53%, respectively. This suggests that the A^2 block can more effectively utilize

Table 1
A comparative analysis of image reconstruction quality on datasets that utilize the scaling factor of ×2. The best result is marked in red, while the next best result is indicated in blue.

Method	BSD100		Set14		Set5		Manga109		Urban100	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Bicubic	29.56	0.8431	30.24	0.8688	33.66	0.9299	30.80	0.9339	26.88	0.8403
SRCNN [3]	31.36	0.8879	32.45	0.9067	36.66	0.9542	35.60	0.9663	29.50	0.8946
VDSR [18]	31.90	0.8960	33.03	0.9124	37.53	0.9587	37.22	0.9750	30.76	0.9140
FSRCNN [34]	31.53	0.8920	32.63	0.9088	37.00	0.9558	36.67	0.9710	29.88	0.9020
DRRN [35]	32.05	0.8973	33.23	0.9136	37.74	0.9591	37.88	0.9749	31.23	0.9188
DRCN [7]	31.85	0.8942	33.04	0.9118	37.63	0.9588	37.55	0.9732	30.75	0.9133
IDN [8]	32.08	0.8985	33.30	0.9148	37.83	0.9600	38.01	0.9749	31.27	0.9196
CARN [21]	32.09	0.8978	33.52	0.9166	37.76	0.9590	38.36	0.9765	31.92	0.9256
IMDN [9]	32.19	0.9177	33.63	0.8996	38.00	0.9605	38.88	0.9774	32.17	0.9283
PAN [22]	32.18	0.8997	33.59	0.9181	38.00	0.9605	38.70	0.9773	32.01	0.9273
RFDN [20]	32.16	0.9184	33.68	0.8994	38.05	0.9606	38.88	0.9773	32.12	0.9278
A^2F-M [23]	32.18	0.8996	33.67	0.9184	38.04	0.9607	38.87	0.9774	32.27	0.9294
LBNet [26]	32.16	0.8994	33.65	0.9177	38.05	0.9607	38.88	0.9775	32.29	0.9290
SRTNet (Ours)	32.27	0.9009	33.81	0.9213	38.14	0.9619	39.06	0.9778	32.57	0.9321

Table 2
A comparative analysis of image reconstruction quality on datasets that utilize the scaling factor of ×3. The best result is marked in red, while the next best result is indicated in blue.

Method	BSD100		Set14		Set5		Manga109		Urban100	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Bicubic	27.21	0.7385	27.55	0.7742	30.39	0.8682	26.95	0.8556	24.46	0.7349
SRCNN [3]	28.41	0.7863	29.30	0.8215	32.75	0.9090	30.48	0.9117	26.24	0.7989
VDSR [18]	28.82	0.7976	29.77	0.8314	33.66	0.9213	32.01	0.9340	27.14	0.8279
FSRCNN [34]	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080	31.10	0.9210
DRRN [35]	28.80	0.7963	29.76	0.8311	33.82	0.9226	32.24	0.9343	27.15	0.8276
DRCN [7]	28.95	0.8004	29.96	0.8349	34.03	0.9244	32.71	0.9379	27.53	0.8378
IDN [8]	28.95	0.8013	29.99	0.8354	34.11	0.9253	32.71	0.9381	27.42	0.8359
CARN [21]	29.06	0.8034	30.29	0.8407	34.29	0.9255	33.50	0.9440	28.06	0.8493
IMDN [9]	29.06	0.8034	30.29	0.8407	34.29	0.9255	33.50	0.9440	28.06	0.8493
PAN [22]	29.11	0.8050	30.36	0.8423	34.40	0.9271	33.61	0.9448	28.11	0.8511
RFDN [20]	29.09	0.8050	30.34	0.8420	34.41	0.9273	33.67	0.9449	28.21	0.8525
A^2F-M [23]	29.11	0.8054	30.39	0.8427	34.50	0.9278	33.66	0.9453	28.28	0.8546
ESRT [2]	29.15	0.8063	30.43	0.8433	34.42	0.9268	33.95	0.9455	28.46	0.8574
LBNet [26]	29.13	0.8061	30.38	0.8417	34.47	0.9277	33.82	0.9460	28.42	0.8559
SRTNet (Ours)	29.17	0.8070	30.45	0.8442	34.60	0.9287	33.96	0.9468	28.46	0.8582

Table 3
A comparative analysis of image reconstruction quality on datasets that utilize the scaling factor of ×4. The best result is marked in red, while the next best result is indicated in blue.

Method	BSD100		Set14		Set5		Manga109		Urban100	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Bicubic	25.96	0.6675	26.00	0.7027	28.42	0.8104	24.89	0.7866	23.14	0.6577
SRCNN [3]	26.90	0.7101	27.50	0.7513	30.48	0.8626	27.58	0.8555	24.52	0.7221
VDSR [18]	27.29	0.7251	28.01	0.7674	31.35	0.8838	28.83	0.8870	25.18	0.7524
FSRCNN [34]	26.98	0.7150	27.61	0.7550	30.72	0.8660	27.90	0.8610	24.62	0.7280
DRRN [35]	27.23	0.7233	28.02	0.7670	31.53	0.8854	28.93	0.8854	25.14	0.7510
DRCN [7]	27.38	0.7284	28.21	0.7720	31.68	0.8888	29.45	0.8946	25.44	0.7638
IDN [8]	27.41	0.7297	28.25	0.7730	31.82	0.8903	29.41	0.8942	25.41	0.7632
CARN [21]	27.58	0.7349	28.60	0.7806	32.13	0.8937	30.47	0.9084	26.07	0.7837
IMDN [9]	27.56	0.7353	28.58	0.7811	32.21	0.8948	30.45	0.9075	26.04	0.7838
PAN [22]	27.59	0.7363	28.61	0.7822	32.13	0.8948	30.51	0.9095	26.11	0.7854
RFDN [20]	27.57	0.7360	28.61	0.7819	32.24	0.8952	30.58	0.9089	26.11	0.7858
A^2F-M [23]	27.58	0.7364	28.62	0.7828	32.28	0.8955	30.57	0.9100	26.17	0.7892
ESRT [2]	27.69	0.7379	28.69	0.7833	32.19	0.8947	30.75	0.9100	26.39	0.7962
LBNet [26]	27.62	0.7382	28.68	0.7832	32.29	0.8960	30.76	0.9111	26.27	0.7906
SRTNet (Ours)	27.72	0.7384	28.71	0.7848	32.38	0.8969	30.85	0.9124	26.29	0.7922

Table 4
Evaluating the impact of various feature extraction modules within the LFE Module through ablation analysis. (The best result is indicated in **bold**.)

Scale	RCAB [5]	IMDB [9]	SCPA [22]	FRDAB [26]	FDAA	Params	Multi-Adds	PSNR↑	SSIM↑
×4	✓	✗	✗	✗	✗	228K	23.7G	29.94	0.9002
×4	✗	✓	✗	✗	✗	295K	31.3G	30.21	0.9043
×4	✗	✗	✓	✗	✗	520K	9.1G	30.41	0.9068
×4	✗	✗	✗	✓	✗	742K	38.9G	30.76	0.9111
×4	✗	✗	✗	✗	✓	918K	35.1G	**30.85**	**0.9124**

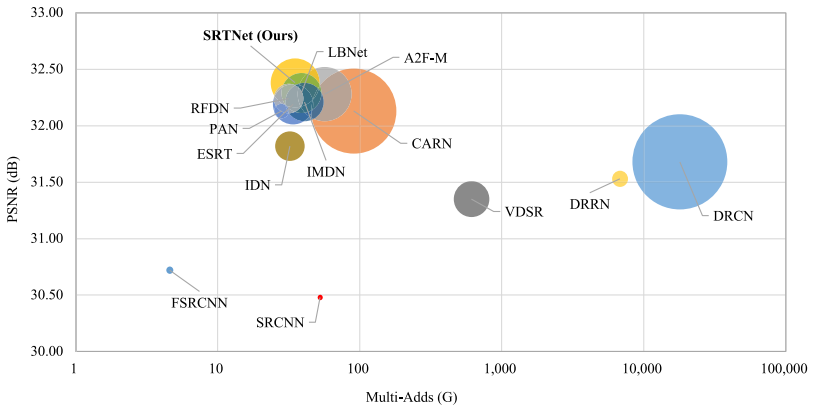

Fig. 3. Trade-off among PSNR, Multi-Adds, and Parameters on Set5 dataset with a scaling factor of ×4.

Fig. 4. Qualitative comparison on Set5, BSD100 and Urban100 datasets with scale factor of ×4.

different types of attention information, which enhances its feature representation ability.

Ablation study on efficient Transformer. An ablation study was conducted to evaluate the impact of the Transformer's components on the effectiveness of our IRT. We adopt the GDFN module and MDTA module to form the IRT. Multi-Layer Perception (MLP) and Multi-Head Attention (MHA) are also widely used in lightweight Transformer-based methods [2,26]. The comparative results of two different Transformers are provided in Table 6. It proves that IRT exceeds the schema of MHA+MLP in PSNR, SSIM, and Multi-Adds. Compared to the MHA+MLP model, the multi-addition mode of IRT reduces by 9.77%, while the PSNR and SSIM increase by 0.28% and 0.1%, respectively. This suggests that the combination of the MDTA and GDFN modules can more effectively capture long-range dependencies and high-frequency

Table 5
Evaluating the impact of feature fusion schemes on FDAA block performance through ablation analysis. (The best result is indicated in **bold**).

Scale	CA	SA	PA	A^2	Params	Multi-Adds	PSNR↑	SSIM↑
×4	✓	✗	✗	✗	393K	28.2G	25.83	0.7776
×4	✗	✓	✗	✗	389K	28.3G	25.66	0.7719
×4	✓	✓	✗	✗	394K	28.3G	25.86	0.7783
×4	✗	✗	✓	✗	545K	37.2G	26.12	0.7870
×4	✗	✗	✗	✓	623K	**27.3G**	**26.22**	**0.7902**

Table 6
Ablation analysis on various Transformers (The best result is indicated in **bold**.)

Transformer	Params	Multi-Adds	PSNR↑	SSIM↑
MHA+MLP	742K	38.9G	32.29	0.8960
MDTA+GDFN (IRT)	918K	**35.1G**	**32.38**	**0.8969**

texture information. At the same time, the fewer additions in IRT indicate that it also has an advantage in computational efficiency.

5. Conclusion

This work introduces a symmetrical and recursive transformer network (SRTNet) for efficient image SR. The proposed SRTNet incorporates two key components, namely the symmetrical CNN (S-CNN) unit and the improved recursive Transformer (IRT) unit. The SRTNet operates in two stages: initially, local features are extracted using the S-CNN unit, followed by the capture of long-range dependencies and contextual information via the IRT unit. In the S-CNN unit, we build the local feature enhancement (LFE) module and feature distillation attention in attention (FDAA) block to expand the scope of the S-CNN unit's receptive field and enhance the feature extraction ability. In the IRT unit, we introduce two efficient Transformers to obtain global information and refine texture details to improve model reconstruction capability. The efficacy of SRTNet is validated through extensive experimentation on five benchmark datasets. The results indicate that SRTNet achieves highly competitive performance, both in accuracy and model complexity, relative to a broad range of state-of-the-art (SOTA) methods. The lightweight nature of SRTNet makes it promising for deployment on resource-constrained devices, which will help drive the application of image super-resolution technology in a wider range of practical scenarios. Despite the promising results, SRTNet is still limited in its generalizability. Given that it is trained on datasets with bicubic degradation, the performance on other types of degradation might vary and require further study to achieve blind SR. Future research will be focused on exploring more advanced attention mechanisms to further enhance the feature representation capabilities of the model. Additionally, more efficient lightweight techniques will be explored to reduce model complexity while maintaining performance.

CRediT authorship contribution statement

Mingliang Gao: Project administration, Methodology. **Jianhao Sun:** Methodology. **Qilei Li:** Formal analysis. **Muhammad Attique Khan:** Validation. **Jianrun Shang:** Software. **Xianxun Zhu:** Resources, Writing – review & editing. **Gwanggil Jeon:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] H. Huang, L. Shen, C. He, W. Dong, H. Huang, G. Shi, Lightweight image super-resolution with hierarchical and differentiable neural architecture search, 2021, arXiv abs/2105.03939.

[2] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, T. Zeng, Transformer for single image super-resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2022, pp. 456–465.

[3] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13, Springer, 2014, pp. 184–199.

[4] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[5] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 286–301.

[6] Q. Li, Z. Li, L. Lu, G. Jeon, K. Liu, X. Yang, Gated multiple feedback network for image super-resolution, 2019, arXiv abs/1907.04253.

[7] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1637–1645.

[8] Z. Hui, X. Wang, X. Gao, Fast and accurate single image super-resolution via information distillation network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 723–731.

[9] Z. Hui, X. Gao, Y. Yang, X. Wang, Lightweight image super-resolution with information multi-distillation network, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019.

[10] W. Zou, T. Ye, W. Zheng, Y. Zhang, L. Chen, Y. Wu, Self-calibrated efficient transformer for lightweight super-resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2022, pp. 929–938.

[11] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. PP (2020) 1–1.

[12] D.R. Martin, C.C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vol. 2, 2001, pp. 416–423, vol.2.

[13] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7, Springer, 2012, pp. 711–730.

[14] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi-Morel, Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding, BMVA Press, 2012.

[15] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, K. Aizawa, Sketch-based manga retrieval using manga109 dataset, Multimedia Tools Appl. 76 (2016) 21811–21838.

[16] J.-B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 5197–5206.

[17] X. Wang, J. Jiang, M. Gao, Z. Liu, C. Zhao, Activation ensemble generative adversarial network transfer learning for image classification, J. Electron. Imaging 30 (2021) 013016–013016.

[18] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Springer, 2016, pp. 1646–1654.

[19] J. Shang, X. Zhang, G. Zhang, W. Song, J. Chen, Q. Li, M. Gao, Gated multi-attention feedback network for medical image super-resolution, Electronics 11 (21) (2022).

[20] J. Liu, J. Tang, G. Wu, Residual feature distillation network for lightweight image super-resolution, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 41–55.

[21] N. Ahn, B. Kang, K.-A. Sohn, Fast, accurate, and lightweight super-resolution with cascading residual network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 252–268.

[22] H. Zhao, X. Kong, J. He, Y. Qiao, C. Dong, Efficient image super-resolution using pixel attention, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 56–72.
[23] X. Wang, Q. Wang, Y. Zhao, J. Yan, L. Fan, L. Chen, Lightweight single-image super-resolution network with attentive auxiliary feature learning, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[24] J. Liang, J. Cao, G. Sun, K. Zhang, L.V. Gool, R. Timofte, SwinIR: Image restoration using swin transformer, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW, 2021, pp. 1833–1844.
[25] J. Fang, H. Lin, X. Chen, K. Zeng, A hybrid network of CNN and transformer for lightweight image super-resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2022, pp. 1102–1111.
[26] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, T. Zeng, Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer, 2022, arXiv abs/2204.13286.
[27] S.W. Zamir, A. Arora, S.H. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 5718–5729.
[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y.R. Fu, Residual dense network for image super-resolution, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.
[29] H. Chen, J. Gu, Z. Zhang, Attention in attention network for image super-resolution, 2021, arXiv abs/2104.09497.
[30] E. Agustsson, R. Timofte, NTIRE 2017 challenge on single image super-resolution: Dataset and study, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2017, pp. 1122–1131.
[31] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[32] A. Muqeet, J. Hwang, S. Yang, J. Kang, Y. Kim, S.-H. Bae, Multi-attention based ultra lightweight image super-resolution, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 103–118.
[33] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations, ICLR, 2015.
[34] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, Springer, 2016, pp. 391–407.
[35] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2790–2798.