International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-024-02277-2

ORIGINAL ARTICLE

Joint features-guided linear transformer and CNN for efficient image super-resolution

Bufan Wang[1] · Yongjun Zhang[1] · Wei Long[1] · Zhongwei Cui[2]

Received: 27 December 2023 / Accepted: 3 July 2024

Abstract

Integrating convolutional neural networks (CNNs) and transformers has notably improved lightweight single image super-resolution (SISR) tasks. However, existing methods lack the capability to exploit multi-level contextual information, and transformer computations inherently add quadratic complexity. To address these issues, we propose a **J**oint features-**G**uided **L**inear **T**ransformer and CNN **N**etwork (JGLTN) for efficient SISR, which is constructed by cascading modules composed of CNN layers and linear transformer layers. Specifically, in the CNN layer, our approach employs an inter-scale feature integration module (IFIM) to extract critical latent information across scales. Then, in the linear transformer layer, we design a joint feature-guided linear attention (JGLA). It jointly considers adjacent and extended regional features, dynamically assigning weights to convolutional kernels for contextual feature selection. This process garners multi-level contextual information, which is used to guide linear attention for effective information interaction. Moreover, we redesign the method of computing feature similarity within the self-attention, reducing its computational complexity to linear. Extensive experiments shows that our proposal outperforms state-of-the-art models while balancing performance and computational costs.

Keywords Image super-resolution · Multi-level contextual information · Linear self-attention · Lightweight network

1 Introduction

Single image super-resolution (SISR) aims to restore high-resolution (HR) images from their corresponding low-resolution (LR) counterparts. This technique holds paramount importance in diverse applications such as remote sensing [1], medical imaging [2], hyperspectral imaging [3], and surveillance [4]. Despite its significance, the inherent ill-posedness of SISR renders accurate image restoration challenging. The advent of convolutional neural networks (CNNs) introduced a transformative approach, facilitating a direct mapping from LR to HR images. Dong et al. [5] pioneered this arena with their SRCNN model, surpassing conventional methods. This led to the proliferation of CNN models, further refining SISR methodologies. Nonetheless, their extensive computational demands hinder efficient deployment on edge devices, as shown in Fig. 1.

To address the computational and storage constraints of edge devices, recent research on SISR is inclining towards the development of lightweight neural network architectures. CNN-based strategies have prevailed, frequently integrating residual or densely connected blocks, often paired with attention mechanisms, to enhance performance. For instance, PAN [6] combined attention mechanisms with residual learning, improving performance. However, CNNs, with their inherent limitations in extracting local features, struggled with long-distance dependencies. Consequently, researchers turned to the transformer for SISR tasks. SwinIR [7] stood out by incorporating transformer components, establishing a solid baseline for image restoration, and emphasizing transformers' potential. ESRT [8] blended CNNs and transformers to develop a lightweight

✉ Yongjun Zhang
zyj6667@126.com

Bufan Wang
gs.wangff21@gzu.edu.cn

Wei Long
lwsch5940@163.com

Zhongwei Cui
zhongweicui@gznc.edu.cn

1 State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, Guizhou, China

2 School of Mathematics and Big Data, Guizhou Education University, Guiyang 550018, China

Published online: 09 July 2024

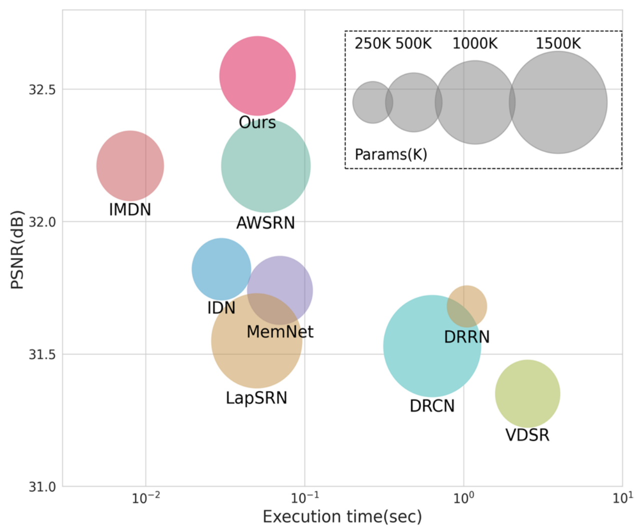

Fig. 1 Model inference time comparison on Set5 dataset(x4)

architecture, leading to an efficient SISR solution. Yet, both CNNs and Transformers often overlook neighborhood and contextual features. Earlier studies posited that neurons should adapt based on behavior. [9] devised a dynamic convolution mechanism adjusting weights per contextual cues. [10] unveiled a context-gated convolution, adaptively altering convolutional kernel weights via context. Recognizing that pixels in images aren't isolated, it's understood that they interact with their surroundings. Notably, the intrinsic quadratic complexity of transformers posed computational challenges. [11] addressed this by representing self-attention as a linear dot product of kernel feature maps, thus reducing complexity to linear levels. [12] utilized a binarization paradigm, approximating linear complexity attention mechanisms through binary code dot products.

Inspired by prior work, we propose the joint feature-guided linear transformer and CNN for an efficient image super-resolution network (JGLTN). Our approach integrates multi-level contextual feature to guide the self-attention mechanism and refines feature similarity calculations, aiming to reduce the transformer's complexity from quadratic to linear. JGLTN comprises several CNN layers and linear transformer layers cascades, each consisting of a CNN layer and a linear transformer layer. Within the CNN layer, we introduce an inter-scale feature integration module (IFIM). This module utilizes the latent information mining component (LIMC) to extract features, emphasizing valuable information while discarding redundancies. For the linear transformer layer, we put forth the joint feature-guided linear attention (JGLA), anchored by the multi-level contextual feature aggregation (MCFA) block, to effectively integrate adjacent, extended regional, and contextual features. To further optimize self-attention, we revisit feature similarity computations, ensuring maintained linear complexity. Our primary contributions are summarized as follows:

(1) We introduce a latent information mining component (LIMC) that filters redundant information and flexibly learns local data. Concurrently, we specially design an inter-scale feature integration module (IFIM) to meticulously combine LIMC, ensuring cross-scale feature learning.
(2) We develop a joint feature-guided linear attention (JGLA), utilizing the designed multi-level contextual feature aggregation (MCFA) to synthesize local and extended regional features and adaptively adjust the weights of modulation convolution kernels, enabling the selection of required contextual information. This approach facilitates self-attention in guiding the information exchange. Additionally, we revisit the feature similarity computation method in attention mechanisms, reducing the computational complexity of self-attention to linear complexity.
(3) We construct a joint features-guided linear transformer and CNN for efficient image super-resolution network (JGLTN). Experiments on five benchmark datasets show that our approach achieves an ideal balance between the computational cost and performance of the model.

2 Related work

2.1 Efficient SISR model

Deep learning models such as EDSR [13], ENLCN [14], and DAT [15] have showcased superior performance SISR. However, their high parameter complexity and computational demands limited their practical deployment. As a result, recent research emphasized lightweight SR models. For example, CARN [16] enhanced efficiency through cascaded residual networks, IMDN [17] leveraged distillation and fusion modules for feature aggregation, and RFDN [18] applied channel splitting and fusion residual connection strategies. Similarly, LatticeNet [19] simplified model complexity with its lattice block and backward feature fusion; RepSR [20] refined SR by reintroducing BN and employing structural re-parameterization; FDIWN [21] integrated wide-residual distillation connection and self-calibration fusion to capture multi-scale details. LatticeNet-CL [22] enhanced performance using its novel lattice block (LB) and a contrastive loss, while GASSL [23] innovated in structured pruning with a sparse structure alignment technique. AsConvSR [24] offered a divide-and-conquer tactic in SR by modulating convolution kernels based on input features, culminating in a swift and compact super-resolution network.

The advent of attention mechanisms significantly advanced lightweight SISR tasks. MAFFSRN [25] incorporated multi-attention blocks (MAB) into a feature extraction group (FFG) to bolster feature fusion. Drawing from attention mechanisms, PAN [6] pioneered a pixel attention (PA) strategy, enabling SR with reduced parameters. This method seamlessly merged PA attention into the primary and reconstruction branches, yielding two innovative building blocks. Similarly, A2N [26] employed an attention-in-attention strategy, which enabled more proactive pixel attention adjustments and enhanced the utilization and comprehension of attention tasks within SISR. PFFN [27] unveiled a progressive attention module to optimize the potential of feature mapping by broadening the receptive field of individual layers. RLFN [28] presented a distinctive residual local feature network, refining feature aggregation and re-examining the contrastive loss. FMEN [29] devised an enhanced residual block paired with a sequential attention branch, accelerating network inference. However, many current CNN-based lightweight SISR models might treat all features uniformly, neglecting the critical nuances of finer details, which could compromise the network's reconstruction proficiency.

2.2 Vision transformers

Recently, the Vision Transformer (ViT) has demonstrated robust potential in low-level visual tasks such as image denoising [30], deblurring [31], enhancement [32], and dehazing [33]. The exceptional ability of ViT to capture long-range information has also been evident in SISR tasks. Precisely, the HAT [34], by integrating channel attention with window self-attention strategies, had further enhanced pixel reconstruction accuracy. DAT [15] alternated between spatial and channel attention within transformer blocks and introduced an SGFN network to integrate intra-module features. While transformer-based approaches have achieved significant results, their deployment in lightweight SISR tasks remains challenging.

Consequently, recent research is dedicated to integrating transformers into lightweight SISRs. SwinIR [7], based on the Swin Transformer [35], designed a reconstruction network comprising multiple residual swin transformers and optimized it to a lightweight version, demonstrating impressive reconstruction outcomes. ESRT [8] developed an efficient multi-head transformer structure for SISR, which significantly reduced memory consumption, thereby enhancing feature representation capabilities. To further capture long-distance dependencies, ELAN [36] introduced a novel multi-scale self-attention mechanism that employed different window sizes for attention computation. NGswin [37], built on SwinIR, proposed a method that interacted via sliding window self-attention to extend degradation areas, integrating N-Gram into SISR and achieving an efficient SR network. These transformer-based lightweight networks have further advanced the performance of SISR tasks.

Introducing the Transformer increased computational demands due to the quadratic complexity of the self-attention mechanism. Addressing this issue, researchers have explored various linear ViT methodologies that capitalize on linear attention to reduce complexity to a linear magnitude. Notable studies [11, 38–42] have adopted kernel-based linear attention strategies to bolster ViT efficacy. These methods eschewed the softmax function, refined computational efficiency by reordering the self-attention computation, and leveraged either kernel functions or comprehensive self-attention matrices. Specifically, [11, 38] managed to attain a linear complexity, preserving a performance on par with the conventional ViT by transforming the softmax-based self-attention's exponential term using kernel functions and modifying the calculation sequence. Additionally, [41, 42] harnessed low-rank approximation and sparse attention to further optimize linear ViT's proficiency. In this paper, we reassess feature similarity computation for token similarities, creating a linear self-attention mechanism. This mechanism bypasses the softmax procedure and reaches linear computational complexity, curtailing computational expenses while upholding superior performance.

3 Approach

3.1 Network architecture

We propose a joint feature-guided linear transformer and CNN for efficient SISR (JGLTN). As depicted in Fig. 2, JGLTN comprises multiple CNN layers cascaded with linear transformer layers, and it further integrates two reconstruction modules. The CNN layers primarily focus on extracting beneficial cross-scale features while filtering out redundant ones. On the other hand, the linear transformer layers emphasize amalgamating adjacent, extended regional, and contextual features. This ensures the transformer not only boasts global modeling prowess but also excels in joint feature modeling, all while reducing the computational complexity of the transformer to a linear scale.

Given an input I_{LR}, the size is initially modified through a convolution f_{conv} to obtain the shallow feature I_0.

$$I_0 = f_{\text{conv}}(I_{\text{LR}}) \tag{1}$$

Subsequently, I_0 is utilized as the input. The process involving n cascaded CNN layers and linear transformer layer can be denoted as:

$$I_n = \zeta^n(\zeta^{n-1}(...(\zeta^1(I_0)))) \tag{2}$$

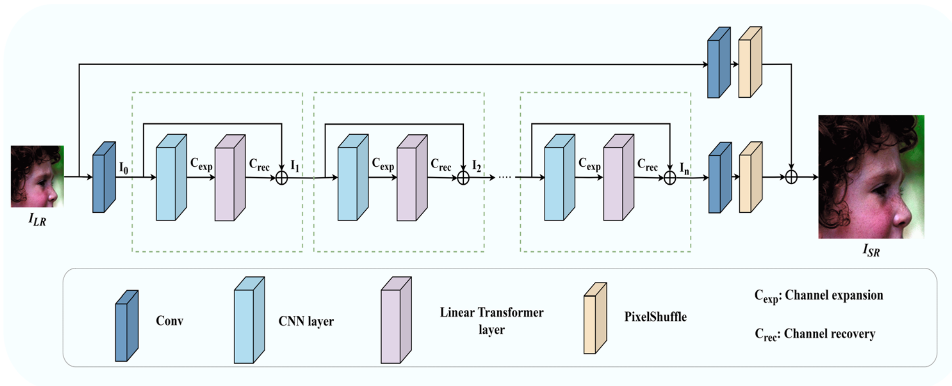

Fig. 2 Overall framework of proposed networks (JGLTN). It consists of a cascade of CNN layers and linear transformer layers

where each ζ^i can be represented as the i-th CNN layer f^i_{CNN} and channel expansion, as well as the i-th linear transformer layer f^i_{LT} and channel recovery operations, collectively depicting the entire process as:

$$\zeta^i = C_{rec}(f^i_{LT}(C_{\exp}(f^i_{CNN}(I^{i-1})))) \tag{3}$$

where I^{i-1} denotes the output from the (i-1)-th CNN layers and linear transformer layer, and $C_{\exp}$ and C_{rec} respectively represent channel expansion and channel recovery, specifically through convolutional layers that alter the channel dimensions of features. In JGLTN, $C_{\exp}$ expands the channel number to 144, while C_{rec} reduces the channel count back to 48. Ultimately, both I_n and I_{LR} are subjected to convolutional upsampling to finalize the image reconstruction.

$$I_{SR} = f_{rec}(I_n) + f_{rec}(I_{LR}) \tag{4}$$

where f_{rec} represents a reconstruction module that encompasses a PixelShuffle operation and a 3x3 convolution operation, I_{SR} stands for the reconstructed high-resolution image.

For the proposed JGLTN network, its loss function L can be expressed as:

$$\begin{aligned} L(\Theta) &= \arg\min_{\Theta} \frac{1}{m} \sum_{i=1}^{m} \left\| f_{JGLTN}(I^i_{LR}) - I^i_{HR} \right\|_1 \\ &= \arg\min_{\Theta} \frac{1}{m} \sum_{i=1}^{m} \left\| I^i_{SR} - I^i_{HR} \right\|_1 \end{aligned} \tag{5}$$

where $f_{JGLTN}(\cdot)$ represents the network model we proposed, I^i_{LR} is the input low-resolution image, I^i_{HR} is its corresponding high-resolution ground truth image, Θ denotes the learnable parameters within the network, and m signifies the number of image pairs in the dataset.

3.2 CNN layer

Within this layer, we present the inter-scale feature integration module (IFIM) designed to capture intricate feature details. As depicted in Fig. 3, the primary components of inter-scale feature integration module include two latent information mining component (LIMC) and a mechanism for cross-scale feature learning.

3.2.1 Inter-scale feature integration module

CNN-based models are notably adept at feature extraction. However, they often incorporate superfluous features. Furthermore, the receptive field of CNNs is inherently restricted, but utilizing features from multiple scales proves crucial for understanding intricate details. To tackle this, we introduce the inter-scale feature integration module (IFIM). This module amalgamates a latent information mining component (LIMC) with a cross-scale learning mechanism and stride convolution. Such a synthesis guarantees the model's proficiency in pinpointing and assimilating valuable features across varied scales.

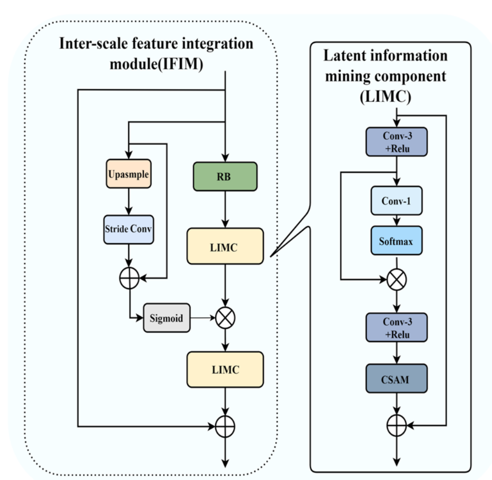

Fig. 3 Structure of the inter-scale feature integration module (IFIM), which is composed of the latent information mining component (LIMC), is used to filter unnecessary features and extract valuable information

Specifically, inter-scale feature integration module (IFIM) bifurcates the input into two branches. One branch retains the original feature dimensions, making selections from the native features. The other branch expands its receptive field to learn features within a larger spatial context. We articulate this entire process as

$$I_{IFIM} = f_{LIMC}(f_{LIMC}(f_{RB}(I_0)) \times (f_{sig}(f_{stride}(f_{CL}(I_0)) + I_0))) + I_0 \tag{6}$$

Where f_{RB} represents a residual block. For the shallow feature input, I_0, features are first extracted through a foundational residual block, followed by feature selection using our proposed latent information mining component (LIMC). f_{CL} corresponds to a learnable upsampling [43]. For f_{CL}, a mask is predicted through a convolution layer, representing the relationship between each original resolution pixel and its neighboring pixels; the original pixels are weighted sums of their neighborhood pixels, with weights provided by this mask. Subsequently, an unfold operation is used to expand the tensor. Essentially, this operation extracts a 3x3 neighborhood around each point, stretches these neighborhoods into one-dimensional vectors, and reshapes them to be weighted by the mask, allowing each pixel and its neighborhood information to be weighted by the corresponding weights in the mask. Finally, a summing operation aggregates the weighted neighborhood information of each pixel, resulting in a high-resolution output. Thus, it guides the transformation of features in the original feature map within a larger receptive field. This is followed by a stride convolution f_{stride} with a stride of 2, residual connection, and a Sigmoid function f_{sig}. f_{LIMC} is the latent information mining component (LIMC), further detailed in the ensuing segment. The cross-scale learning mechanism integrates information across different scales, enabling the network to better understand the structure within images.

3.2.2 Latent information mining component

Building on our prior research in the SISR task, we observe a tendency for reconstructed images to exhibit over-smoothing. This problem emerges mainly because the model indiscriminately processes all input image features, neglecting that certain features are pivotal for recovering image texture details. To address this, we introduce the latent information mining component (LIMC), which emphasizes texture details, as illustrated in Fig. 3. The underlying principle of latent information mining component (LIMC) is to filter out redundant features by generating masks, thereby better preserving vital image details. Initially, input features are processed through convolutions and activation functions, yielding mixed features. Following this, a 1x1 convolution is utilized to modify the channel count of these mixed features to three. Weights are assigned to each channel via a softmax operation, efficiently discarding superfluous features.

$$I'_{CR} = f_{CR}(I_{CR}) \tag{7}$$

$$I_{LIMC} = f_{CSAM}(f_{CR}(f_{SM}(f_{1\times1}(I'_{CR})) \times I'_{CR})) + I_{CR} \tag{8}$$

where f_{CR} denotes the combination of convolution and activation functions, the output feature I'_{CR} is produced. Subsequently, a 1x1 convolution, represented by $f_{1\times1}$, generates a mask. After a softmax operation f_{SM}, this mask filters out superfluous features. The resulting weight mask is then multiplied with the input I'_{CR} to retain sections of the original feature that contribute significantly to texture detail. Finally, an additional convolution layer and a channel spatial attention mechanism f_{CSAM} using 3D convolution are introduced to refine features further. CSAM utilizes the channel-spatial attention from the HAN [44], incorporating responses from all dimensions of the feature maps. A 3D convolutional layer is used to capture the joint channel and spatial features within the feature maps, generating an attention map. This is achieved by applying a 3D convolutional kernel to the data cube formed by multiple adjacent channels of the input features. The 3D convolutional kernels, sized 3x3x3 with a stride of 1, convolve with three consecutive groups of channels, each interacting with a set of 3D convolutional kernels, to produce three sets of channel-spatial attention maps. Through this process, CSAM is capable of extracting robust representations to describe inter-channel and intra-channel information across consecutive channels. A residual connection complements this to ensure the effective operation of the latent information mining component (LIMC). As a result, within inter-scale feature integration module (IFIM), latent information mining component (LIMC) retains the most valuable features via the weight mask.

3.3 Linear transformer layer

The powerful global feature learning of transformers presents a novel approach for the SISR task. However, transformers currently employed in SISR often overlook adjacent and extended regional features, and contextual features while also having a high computational complexity. To address these issues, we propose a joint-feature guided linear attention (JGLA). Within JGLA, we employ a multi-level contextual feature aggregation (MCFA) to obtain joint features. This block carefully considers adjacent, extended regional, and contextual features, optimizing the surrounding and contextual information of a given pixel during feature reconstruction. Furthermore, refining the self-attention mechanism, we streamlined the vector similarity computation, redesigning the self-attention calculations and effectively transitioning the complexity to a linear scale.

3.3.1 Multi-level contextual feature aggregation

Transformers are powerful at global feature modeling, both overlook the significance of adjacent, extended regional, and contextual features. Typically, in traditional transformers, the self-attention mechanism derives self-attention values from three linear layers following linear embedding. Our objective is for our transformer to handle adjacent, extended regional, and contextual features. As a result, we have modified the linear embedding layer within the transformer, replacing it with our proposed multi-level contextual feature aggregation (MCFA).

Multi-level contextual feature aggregation (MCFA) is adept at harnessing adjacent, extended regional, and contextual features. Specifically, the MCFA comprises a 3x3 convolution, a dilated convolution, and a context modulation convolution (CMC), as illustrated in Fig. 4. For the input feature I_{Input}, the entire process can be expressed as:

$$I_{MCFA} = f_{CMC}(f_{relu}(Concat(f_{adj}(I_{Input}), f_{ext}(I_{Input})))) \tag{9}$$

where f_{adj} represents a standard 3x3 convolution, which learns local features from adjacent feature vectors, f_{ext} is a dilated convolution, capturing a broader receptive field to grasp surrounding features better. After concatenating f_{adj} and f_{ext}, a ReLU operation is applied.

Subsequently, global contextual information is adaptively learned through f_{CMC}. In the CMC, for input features of size $c \times h \times w$, a max pooling operation first reduces the dimensions to $k \times k$. A shared-parameter linear layer projects the spatial location information into a vector of size $(k \times k)/2$, from which new channel weights are generated. To alleviate the time-consuming core modulation caused by a large number of channels, the concept of grouped convolution is applied to the linear layer, resulting in an output dimension o and facilitating channel interaction. Subsequently, another linear layer produces tensor outputs in two directions, $o \times 1 \times k \times k$ and $1 \times c \times k \times k$. These two tensors are added element-wise to form our modulated convolutional kernel, resized to $o \times c \times k \times k$, to simulate the convolution kernel in actual convolution operations. The simulated modulated convolutional kernel is then multiplied by an adaptive multiplier W and reshaped. W is an adaptive multiplier, matching the size of the tensor, and upon multiplication with the tensor, it transforms into a set of trainable type parameters that are bound to the module, allowing the weights of W to be automatically learned and modified during training to optimize the model. For the input X_{input}, the unfold function extracts sliding features of size $k \times k$, linking the context between different pixel features, resulting in a feature map size of $X'_{input} \in k^2 c \times hw$. Ultimately, a fully comprehended modulated convolutional kernel W' can use the input features to obtain context guidance, adaptively capturing the required contextual information for key pixels.

3.3.2 Joint-feature guided linear attention

To reduce the computational complexity of the self-attention mechanism, we introduce the linear transformer. The essence of linear attention lies in decomposing the similarity measure function into distinct kernel embeddings, denoted as $S(Q, K) \approx \Phi(Q)\Phi(K)^T$. Consequently, leveraging the properties of matrix calculations, we can rearrange the computation order to $\Phi(Q)(\Phi(K)^T V)$. The complexity of the attention mechanism is no longer the token length but depends on the feature dimension. Here, the feature dimension is much smaller than the token length. We redesign the

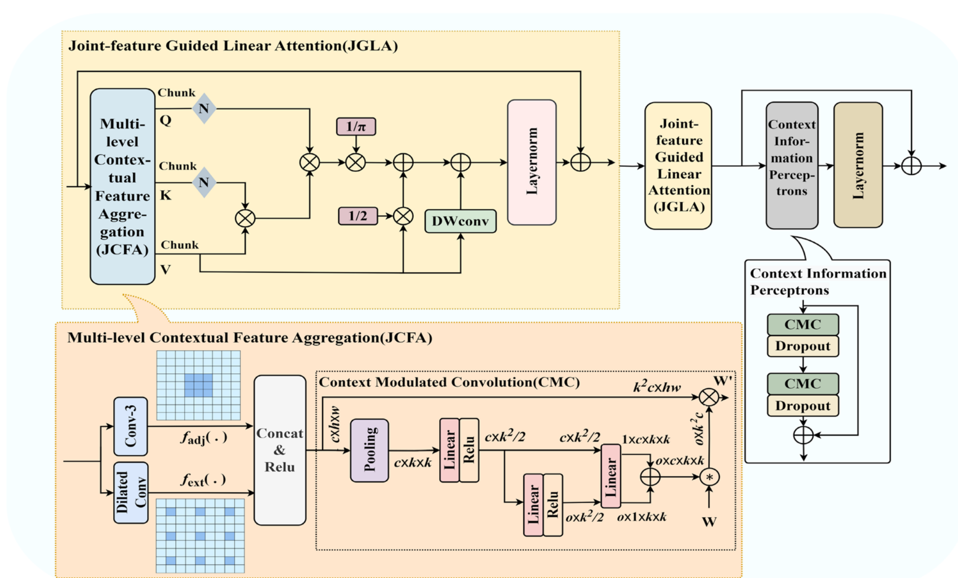

Fig. 4 Structure of the linear transformer layer, including multi-level contextual feature aggregation (MCFA), joint feature guided linear attention (JGLA) and contextual information perceptron

similarity computation between vector pairs, formulating a linear attention apt for the SISR task. This offers a substitute for conventional softmax-based attention, enhancing performance outcomes. Specifically, the dot product of two vectors, v_1 and v_2, is defined as:

$$v_1 \cdot v_2 = \|v_1\| \cdot \|v_2\| \cos\theta \tag{10}$$

where θ represents the angle between vector pairs, consequently, the angle between vector pairs can be expressed as:

$$\theta(v_1, v_2) = \arccos\left(\frac{\langle v_1, v_2\rangle}{\|v_1\| \cdot \|v_2\|}\right) \tag{11}$$

where $< \cdot >$ denotes the inner product and $\|\cdot\|$ signifies the Euclidean norm. Thus, the range of θ is $[0, \pi]$. We incorporate this angular calculation into the similarity computation between the Q and K values in the attention mechanism, articulated as:

$$S(Q,K) = 1 - \frac{1}{\pi} \cdot \theta(Q,K) \tag{12}$$

We confine the output range of $S(Q, K)$ to [0, 1]. If the distance between Q and K decreases, the angular distance θ also decreases, approaching θ, making $S(Q, K)$ near 1. On the other hand, as the distance between Q and K grows, θ increases, leading $Sim(Q, K)$ to approach 0, which signifies a diminished similarity between Q and K.

Substituting Eq. 10 into Eq. 11, we obtain:

$$S(Q,K) = 1 - \frac{1}{\pi} \arccos\left(\frac{\langle Q, K\rangle}{\|Q\| \cdot \|K\|}\right) \tag{13}$$

To simplify Eq. 10, we employ trigonometric calculations and infinite series expansion.

$$\begin{aligned} \mathrm{S}(Q,K) &= 1 - \frac{1}{\pi}\left(\frac{\pi}{2} - \arcsin(Q \cdot K^T)\right) \\ &= \frac{1}{2} + \frac{1}{\pi} \cdot (Q \cdot K^T) \\ &+ \frac{1}{\pi} \cdot \sum_{t=1}^{\infty} \frac{(2t)!}{2^{2t}(t!)^2(2t+1)} (Q \cdot K^T)^{2t+1} \end{aligned} \tag{14}$$

In this configuration, $(Q \cdot K^T)$ denotes a normalization operation $\left(\frac{\langle Q,K\rangle}{\|Q\| \cdot \|K\|}\right)$. As observed from the above equation, the first term can serve as a similarity measure in linear attention with a complexity of $O(n)$. The latter term, being of a higher order, introduces greater complexity. To tailor our linear attention mechanism for the SISR task, we propose employing a linear expansion approach. Given that our Q and K are near zero, we suggest retaining only the first linear term and discarding the higher-order terms. This ensures linear complexity conservation while sidestepping additional complexity induced by elevated order elements. Moreover, extant studies suggest that softmax-based attention can generate a full-rank attention map, mirroring model feature diversity. However, linear attention cannot yield a full-rank attention map [45]. As a countermeasure, we introduce a DWconv, reviving the rank of the attention matrix and ensuring feature diversity.

Therefore, our attention mechanism can be described as:

$$\begin{aligned} L &= S(Q,K) \cdot V \\ &\approx \frac{1}{2} \cdot V + \frac{1}{\pi} \cdot Q \cdot (K^T \cdot V) + f_{DW} \cdot V \end{aligned} \tag{15}$$

Given that $\frac{1}{2} \cdot V + \frac{1}{\pi} \cdot Q \cdot (K^T \cdot V)$ is the linear term from Eq. 13 and f_{DW} represents a depthwise separable convolution to achieve a full-rank attention matrix, thereby enriching feature diversity. Consequently, the final transformer layer consists of two joint feature-guided linear attention (JGLA) and a context information perceptron, as shown in Fig. 4. For the context information perceptron, we retain the context modulation convolution (CMC), substituting the multi-layer perceptron from the traditional transformer.

4 Experiments

4.1 Datasets and metric

Consistent with previous works, we utilize the DIV2K [53] dataset for our training, which consists of 800 training images. We further validate the effectiveness of our model using five public benchmark datasets for testing, including Set5 [54], Set14 [55], B100 [56], Urban100 [57], and Manga109 [58]. Additionally, the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are utilized as metrics to evaluate the image restoration quality on the Y channel.

4.2 Implementation details

For training, we apply x2, x3, and x4 upscaling factors. The batch size is set to 16. Additionally, for data augmentation, we apply random rotations of 90°, 180°, 270°, and randomly crop a patch of size 48x48 as input. We set the initial learning rate at $5e^{-4}$ and employ the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$. The model features an input channel count 48, with channel adjustments after each CNN layers and linear transformer layers invocation. Furthermore, the training architecture consists of eight CNN layers and eight linear transformer layers. All experiments are conducted on an Nvidia RTX A6000 GPU.

4.3 Comparison with advanced lightweight SISR models

In this section, we compare our approach with state-of-the-art lightweight SISR methods, including [16–19, 23, 25, 26, 29, 46–51]. We will analyze from two perspectives: qualitative analysis and quantitative analysis.

4.3.1 Quantitative evaluations

We validate the quantitative performance of our model by comparing it with state-of-the-art models on five benchmark datasets at x2, x3, and x4 scales, as presented in Table 1. We adhere to standardized methods for the uniform calculate the number of parameters. For the calculation of Multi-Adds, we are consistent with other methods and set the HR image size to 1280x720. The results highlight the superior performance of our method over advanced models. Notably, JGLTN consistently ranks among the top across all datasets. At the x4 scale, our approach exceeds the LatticeNet-CL model by margins of 0.25dB on Set5, 0.14dB on Urban100, and 0.19dB on Manga109. Such enhanced performance is attributed to the seamless integration of CNN and transformer, efficiently capturing adjacent, extended regional, and contextual information. Additionally, adapting linear attention for SISR tasks simplifies the model's design while ensuring excellent performance.

Furthermore, we conducted a comparative analysis of our model and advancing transformer-based models, as illustrated in Table 2. To demonstrate the comparison more vividly between our method and advanced transformer-based approaches, we have introduced the use of average PSNR/SSIM for evaluation. This allows for a rapid and informative assessment of the algorithm's effectiveness by examining the average performance. Compared to lightweight models such as SwinIR, ESRT, and NGswin, our approach demonstrates equivalent efficacy, maintaining a similar scale of parameters and computational resources. It is worth noting that compared to SwinIR, JGLTN still maintains a comparable number of parameters with lower Multi-Adds. It is noteworthy that SwinIR uses a pretrained model for initialization and sets the patch size to 64x64 during training. Extensive experiments have shown that larger patch sizes yield better results. However, JGLTN utilizes a patch size of 48x48. Moreover, SwinIR employs an additional dataset (Flickr2K [59]) for training, which is crucial for further enhancing model performance. To ensure a fair comparison with methods like ESRT, we did not use this external dataset in our work. The results of JGLTN on some datasets even surpass those of SwinIR, with higher average PSNR, while still maintaining comparable parameters and fewer Multi-Adds.

4.3.2 Qualitative evaluations

To more thoroughly examine the efficacy of our model, we performed a qualitative comparative analysis alongside the current advancing models. As depicted in Figs. 5 and 6, prior methods manifest issues such as boundary-blurring, excessive smoothing, line distortions, and, in some instances, alterations to the original image structure, significantly diminishing its visual appeal. In stark contrast, our proposed model preserves the original image structure, maintains sharp boundaries, and retains intricate texture details. Specifically, the $'ppt3'$ in Fig. 5, our method provides more pronounced edge information than representative CNN and transformer methods. Furthermore, within $'img\_78'$ of the Urban100 dataset, while numerous advanced techniques compromise the original image structure, our approach successfully maintains the essential original information, yielding a visually superior result. Therefore, JGLTN not only upholds a high PSNR metric but also ensures enhanced visual outcomes.

4.4 Ablation study

4.4.1 The effectiveness of CNN layer

Effectiveness of latent information mining component (LIMC): We validate the effectiveness of the LIMC within the JGLTN model. Experiments are conducted by excluding the LIMC module from inter-scale feature integration module (IFIM). Notably, to expedite the experiments, the model omits the linear transformer layers. Table 3 displays the comparative results between configurations with and without LIMC. The results reveal that although the parameter count decreases with the removal of LIMC, model performance experiences a decrease of 1.03dB on the Urban100 dataset and 2.19dB on the Manga109 dataset. This underscores the role of the LIMC component in retaining valuable information and filtering out unnecessary features.

Effectiveness of inter-scale feature integration module (IFIM): We evaluate the impact of cross-scale leairning(CL) to ascertain the efficacy of IFIM. Integration of CL results in an average increase of 0.04dB in PSNR value across five datasets, as Table 4 illustrates. This fact underscores the crucial role of CL in cross-scale feature learning within IFIM. For a more comprehensive validation of IFIM, we conduct a comparative analysis with state-of-the-art CNN modules, replacing IFIM with advanced modules such as RCAB, HPB, and LB. To expedite the experiment, we exclude linear transformer layers. Despite a marginal parameter increase, IFIM demonstrates a nearly 0.1dB improvement in PSNR over these advanced CNN modules, as indicated in Table 4. This performance attests to the capability of IFIM to learn valuable features across various scales.

Table 1 Performance of our method compared with state-of-the-art SR methods with BI degradation for ×2,×3 and ×4 image super-resolution on benchmark datasets

Scale	Methods	Params	Multi-adds	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
2	VDSR [46]	655K	612.6G	37.53/0.9587	33.03/0.9124	31.90/0.8960	30.76/0.9140	37.22/0.9750
	IDN [47]	553K	124.6G	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196	38.01/0.9749
	CARN [16]	1592K	222.8G	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
	IMDN [17]	694K	158.8G	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
	MADNet [48]	878K	187.1G	37.85/0.9600	33.38/0.9161	32.04/0.8979	31.62/0.9233	- / -
	MAFFSRN-L [25]	790K	154.4G	38.07/0.9607	33.59/0.9177	32.23/0.9005	32.38/0.9308	- / -
	RFDN [18]	534K	123.0G	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
	LatticeNet+ [19]	756K	165.5G	38.15/0.9610	33.78/0.9193	32.25/0.9005	32.43/0.9302	- / -
	SMSR [49]	985K	351.5G	38.00/0.9601	33.64/0.9179	32.17/0.8990	32.19/0.9284	38.76/0.9771
	A2N [26]	1036K	247.5G	38.06/0.9608	33.75/0.9194	32.22/0.9002	32.43/0.9311	38.87/0.9769
	DRSAN [50]	690K	159.3G	38.11/0.9609	33.64/0.9185	32.21/0.9005	32.35/0.9304	- / -
	FMEN [29]	748K	172.0G	38.10/0.9609	33.75/0.9192	32.26/0.9006	32.48/0.9311	38.95/**0.9778**
	LatticeNet-CL [51]	756K	169.5G	38.09/0.9608	33.70/0.9188	32.21/0.9000	32.29/0.9291	- / -
	GASSL-B [23]	689K	158.2G	38.08/0.9607	33.75/0.9194	32.24/0.9005	32.29/0.9298	38.92/0.9777
	Ours	900K	115.5G	**38.19/0.9612**	**33.85/0.9197**	**32.27/0.9010**	**32.59/0.9319**	**39.08**/0.9776
3	VDSR [46]	665K	612.6G	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279	32.01/0.9340
	IDN [47]	553K	56.3G	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359	32.71/0.9381
	CARN [16]	1592K	118.8G	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.43/0.9427
	IMDN [17]	703K	71.5G	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
	MADNet [48]	930K	88.4G	34.16/0.9253	30.21/0.8398	28.98/0.8023	27.77/0.8439	- / -
	MAFFSRN-L [25]	807K	68.5G	34.45/0.9277	30.40/0.8432	29.13/0.8061	28.26/0.8552	- / -
	RFDN [18]	541K	55.4G	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525	33.67/0.9449
	LatticeNet+ [19]	765K	76.3G	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538	- / -
	SMSR [49]	993K	156.8G	34.40/0.9270	30.33/0.8412	29.10/0.8050	28.25/0.8536	33.68/0.9445
	A2N [26]	1036K	117.5G	34.47/0.9279	30.44/0.8437	29.14/0.8059	28.41/0.8570	33.78/0.9458
	DRSAN [50]	740K	76.0G	34.50/0.9278	30.39/0.8437	29.13/0.8065	28.35/0.8566	- / -
	FMEN [29]	757K	77.2G	34.45/0.9275	30.40/0.8435	29.17/0.8063	28.33/0.8562	33.86/0.8462
	LatticeNet-CL [51]	765K	76.3G	34.46/0.9275	30.37/0.8422	29.12/0.8054	28.23/0.8525	- / -
	GASSL-B [23]	691K	70.4G	34.47/0.9278	30.39/0.8430	29.15/0.8063	28.27/0.8546	33.77/0.9455
	Ours	913K	52.6G	**34.64/0.9285**	**30.52/0.8456**	**29.22/0.8087**	**28.59/0.8601**	**33.98/0.9473**
4	VDSR [46]	665K	612.6G	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8870
	IDN [47]	553K	32.36G	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8942
	CARN [16]	1592K	90.9G	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.42/0.9070
	IMDN [17]	715K	40.9G	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
	MADNet [48]	1002K	54.1G	31.95/0.8917	28.44/0.7780	27.47/0.7327	25.76/0.7746	- / -
	MAFFSRN-L [25]	830K	38.6G	32.20/0.8953	28.62/0.7822	27.59/0.7370	26.16/0.7887	- / -
	RFDN [18]	550K	31.6G	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
	LatticeNet+ [19]	777K	43.6G	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873	- / -
	SMSR [49]	1006K	89.1G	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
	A2N [26]	1047K	72.4G	32.30/0.8966	28.71/0.7842	27.61/0.7374	26.27/0.7920	30.67/0.9110
	DRSAN [50]	730K	49.0G	32.30/0.8954	28.66/0.7838	27.61/0.7381	26.26/0.7920	- / -
	FMEN [29]	769K	44.2G	32.24/0.8955	28.70/0.7839	27.63/0.7379	26.28/0.7908	30.70/0.9107
	LatticeNet-CL [51]	777K	43.6G	32.30/0.8958	28.65/0.7822	27.59/0.7365	26.19/0.7855	- / -
	GASSL-B [23]	694K	39.9G	32.17/0.8950	28.66/0.7835	27.62/0.7377	26.16/0.7888	30.70/0.9100
	HPUN-L [52]	734K	39.7G	32.38/0.8969	28.72/0.7847	27.66/0.7393	26.36/0.7947	30.83/0.9124
	Ours	931K	30.7G	**32.55/0.8980**	**28.78/0.7858**	**27.69/0.7412**	**26.42/0.7962**	**30.89/0.9139**

'-' denotes the results are not reported. The best and second best results are highlighted in bold and underlined

Table 2 Performance of our method compared with state-of-the-art transformer-based methods on benchmark datasets

Methods	Params	Multi-adds	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM	Average PSNR/SSIM
SwinIR [7]	897K	49.6G	32.44/0.8976	28.77/0.7858	27.69/0.7406	**26.47/0.7980**	**30.92/0.9151**	29.26/**0.8274**
ESRT [8]	751K	67.7G	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100	29.14/0.8244
NGswin [37]	1019K	36.4G	32.33/0.8963	28.78/**0.7859**	27.66/0.7396	26.45/0.7963	30.80/0.9128	29.20/0.8262
Ours	931K	30.7G	**32.55/0.8980**	**28.78**/0.7858	**27.69/0.7412**	26.42/0.7962	30.89/0.9139	**29.27**/0.8270

The best and second best results are highlighted in bold and underlined

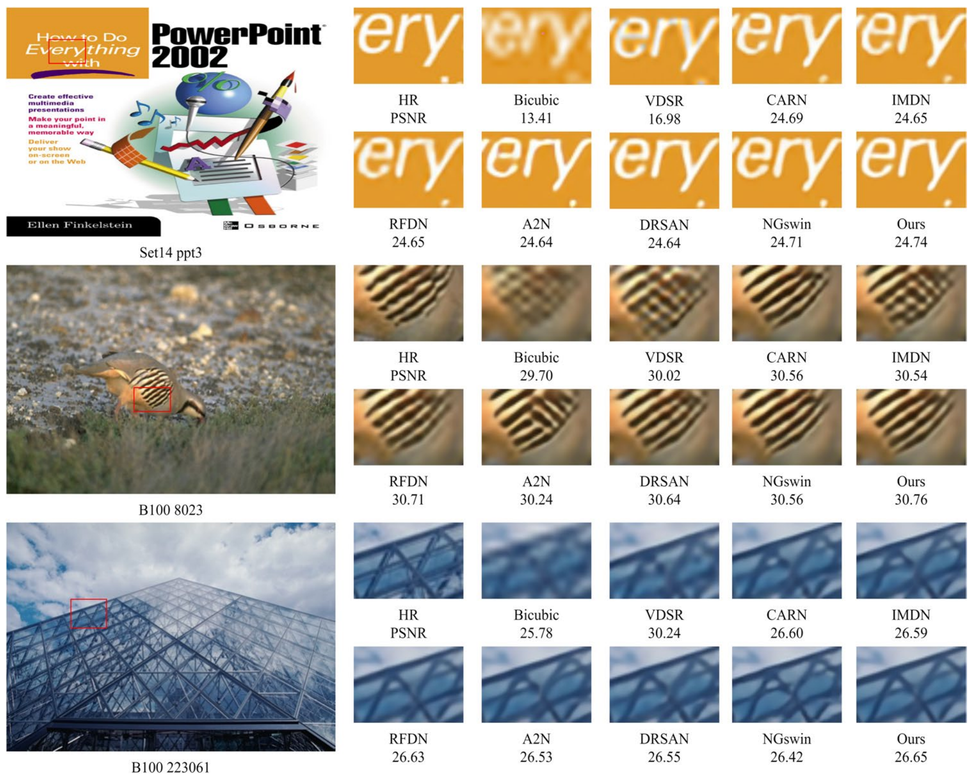

Fig. 5 Qualitative comparison of our JGLTN with recent state-of-the-art lightweight image SR methods for the ×4 SR on Set14 and B100 datasets. The performance of each image patch is shown in the following

Furthermore, to validate the efficacy of inter-scale feature integration module (IFIM) in feature extraction and pinpoint its focus areas on the image, we executed a meticulous visual analysis. Figure 7a contrasts the output feature maps with and without the incorporation of IFIM. Without IFIM, the network seemingly emphasizes the flat regions interspersed among textures. Notably, texture details are paramount for image reconstruction. With the integration of IFIM, the network shifts its attention predominantly towards the intricate textures of the butterfly. This observation strongly attests to the enhanced feature extraction capabilities of IFIM.

4.4.2 The effectiveness of linear transformer layer

Effectiveness of multi-level contextual feature aggregation (MCFA): MCFA plays a crucial role in guiding joint features for the transformer. We validate its effectiveness by replacing MCFA with the linear embedding layer commonly used in traditional self-attention mechanisms. Table 5 demonstrates that MCFA outperforms the linear embedding layer regarding PSNR metrics while utilizing fewer parameters. Specifically, it achieves a 0.22dB improvement on the Set5 dataset and a 0.4dB enhancement on the Manga109 dataset. These results verify that MCFA effectively integrates adjacent, extended regional, and contextual information, providing optimal feature guidance for the linear transformer.

Effectiveness of joint feature-guided linear attention (JGLA): We compare JGLA with the self-attention mechanism prevalent in traditional transformers. According to the data in Table 5, JGLA achieves superior metrics while maintaining a consistent parameter count. Specifically, it outperforms them by 0.04dB on the Urban100 dataset and by 0.21dB on the Manga109 dataset. Given that JGLA is a form of linear attention, it operates with significantly lower

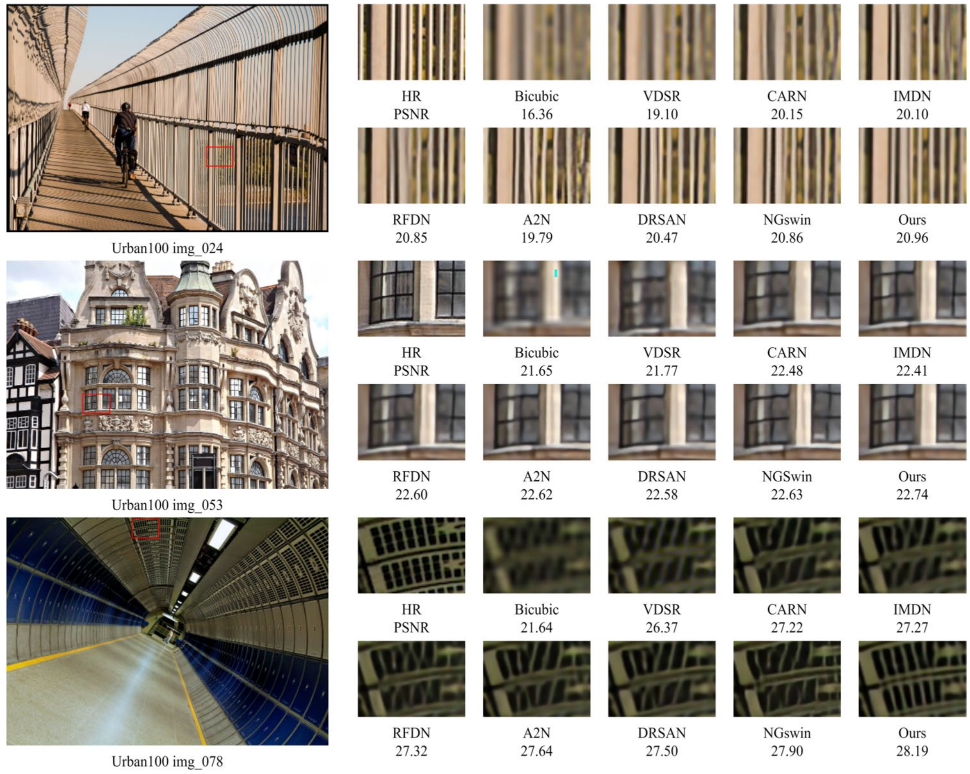

Fig. 6 Qualitative comparison of our JGLTN with recent state-of-the-art lightweight image SR methods for the ×4 SR on Urban100 Datasets. The performance of each image patch is shown in the following

Table 3 The effect of the latent information mining component (LIMC) in terms of PSNR score on the benchmark datasets for x4 SR

Modules	Params	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
w/o LIMC	132K	5.2G	30.40	27.42	26.82	24.30	27.12
w LIMC	215K	5.3G	31.59	28.21	27.31	25.33	29.31

Table 4 The effect of the inter-scale feature integration module (IFIM) in terms of PSNR score on the benchmark datasets for x4 SR

Modules	Params	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
w RCAB [46]	107K	6.1G	30.37	27.43	16.82	24.27	27.12
w HPB [8]	132K	7.6G	31.42	28.08	27.21	25.15	28.94
w LB [19]	197K	11.2G	31.50	28.16	27.29	25.30	29.24
w/o CL	205K	6.9G	31.54	28.18	27.30	25.29	29.24
w IFIM	215K	5.3G	31.59	28.21	27.31	25.33	29.31

Where w/o CL represents the comparison result of IFIM without cross-scale learning

Table 5 The effect of the multi-level contextual feature aggregation (MCFA) and joint feature-guided linear attention (JGLA) in terms of PSNR score on the benchmark datasets for x4 SR

Modules	Params	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
MCFA *	1775K	19.9G	32.33	28.63	27.61	26.29	30.49
JGLA*	931K	30.4G	32.48	28.72	27.67	26.38	30.68
JGLTN	931K	30.7G	32.55	28.78	27.69	26.42	30.89

Among them, MCFA* means replacing MCFA with the linear embedding layer in traditional self-attention. JGLA* means replacing JGLA with traditional self-attention based on softmax

Table 6 The effect of integrating CNN layer(CNN) with linear transformer layer(LT) in terms of PSNR score on the benchmark datasets for x4 SR

Modules	Params	GPU	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
w/o LT	722K	3888 M	19.7G	32.09	28.51	27.54	25.91	30.27
w/o CNN	330K	11470 M	17.9G	32.46	28.65	27.64	26.32	30.63
JGLTN	931K	13052 M	30.7G	32.55	28.78	27.69	26.42	30.89

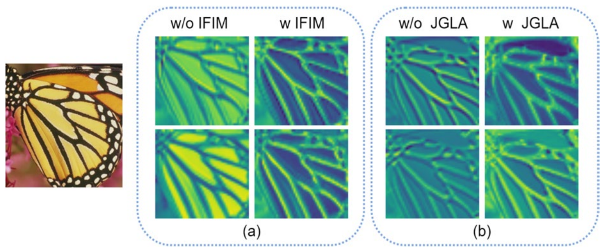

Fig. 7 Visual comparison of features with and without inter-scale feature integration module (IFIM) and joint feature-guided linear attention (JGLA)

complexity compared than the traditional self-attention mechanism. These findings demonstrate that JGLA not only works with linear complexity but also delivers performance superior to that of the traditional self-attention mechanism, establishing its suitability for the SISR task.

To further ascertain the specific regions emphasized by joint feature-guided linear attention (JGLA) within the network, a detailed visual analysis was conducted. Figure 7b contrasts the focal areas of features when integrating JGLA and when omitting it. Without JGLA, the model predominantly targets the contour information of the butterfly. With the introduction of JGLA, the network accentuates not only texture features but also the surrounding details and contextual information adjacent to the contour. These observations underscore the capacity of JGLA to consolidate surrounding details and contextual information, facilitating the reconstruction process and enhancing the effectiveness of joint feature-guided linear attention (JGLA).

4.4.3 The effectiveness of CNN layer and linear transformer layer

In this subsection, we undertake ablation experiments focusing on the CNN layers and the linear transformer layers. Models comprising exclusively CNN layers and solely linear transformer layers form the basis of this analysis. The data presented in Table 6 illustrate that dependence solely on CNN layers increases parameter usage, while exclusive reliance on linear transformers leads to higher GPU consumption. Employing either component in isolation falls short of matching the performance level of JGLTN, rendering them less viable for practical applications. For instance, on the Manga109 dataset, models utilizing only CNN or linear transformer layers underperform JGLTN by 0.62dB and 0.26dB, respectively. Therefore, a balanced integration of CNN and linear transformer layers optimizes model size, GPU consumption, and performance, enhancing suitability for real-world deployment.

Table 7 Comparison with SISR methods on RealSR

Scale	Methods	IMDN [17]	LP-KPN [60]	ESRT [8]	ours
x4	PSNR	28.68	28.65	28.78	**28.97**
	SSIM	0.815	0.820	0.815	**0.832**

The best results are highlighted in bold

4.5 Real-world image super-resolution

To further validate whether our model is applicable to the super-resolution of real images, we compared it with other lightweight models on the RealSR [60] dataset, including IMDN [17], LP-KPN [60], and ESRT [8]. Notably, LP-KPN is specifically designed for SR of real images. As shown in Table 7, our model outperforms the other methods on the RealSR dataset in terms of metrics, demonstrating that JGLTN is also suitable for SR of real images.

4.6 Model size analysis

In this section, we benchmark our model against others, considering PSNR, parameters, and Multi-Adds. We carry out these experiments on the Set5 dataset at x4 upscaling factor. As depicted in Fig. 8, while JGLTN does not possess the lowest parameter, it excels in the PSNR metric. Specifically, our model has fewer parameters than NGswin and A2N and marginally more than FMEN and LatticeNet-CL. However, it significantly surpasses these advanced methods in terms of PSNR. According to Fig. 9, JGLTN establishes an exemplary balance between PSNR and Multi-Adds, upholding a superior PSNR performance even when accounting for the minimal Multi-Adds. Our proposed method outperforms other models within similar parameter ranges, achieving a judicious balance between model complexity and performance. Therefore, we affirm that our approach is lightweight and efficient, ensuring a favorable balance between model size and performance.

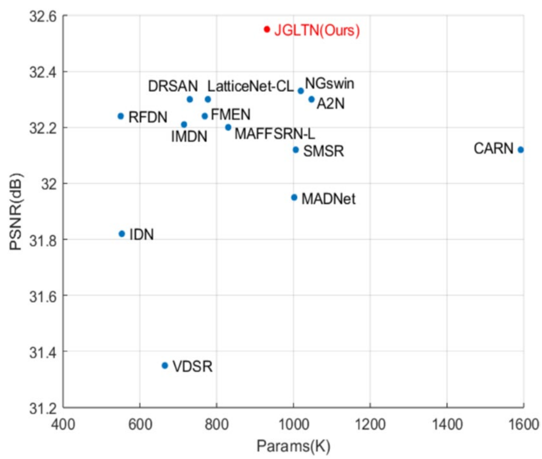

Fig. 8 Model performance and size comparison on Set5 (×4)

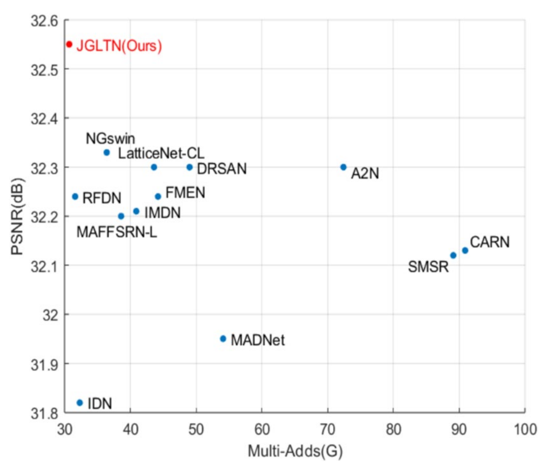

Fig. 9 Model performance and Multi-Adds comparison on Set5 (×4)

In the experiments, we set the number of CNN layers and linear transformer layers to 8 and explore the performance under conditions with fewer CNN layers and linear transformer layers. As illustrated in Fig. 10, when the number of CNN layers and linear transformer layers is increased, performance is consistently enhanced compared to the model variant with G=2 (where G represents the number of CNN layers and linear transformer layers). Consequently, JGLTN exhibits superior performance across diverse configurations, showcasing its effectiveness and scalability.

5 Conclusions

In this paper, we propose a lightweight super-resolution network termed the joint feature-guided linear transformer and CNN network (JGLTN). The structure of this network consists of a cascade of CNN layer and linear transformer layer, collectively termed CNN layers and linear transformer

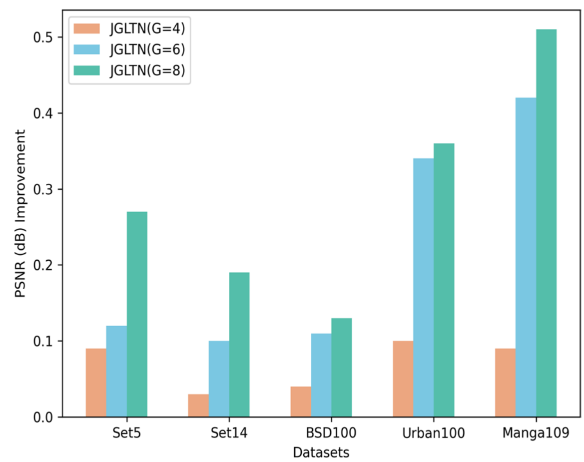

Fig. 10 PSNR improvement of JGLTN variants with different CNN layers and linear transformer layers number (G) over the smallest JGLTN (G=2) for ×4 SR

layers. The CNN layer incorporates an inter-scale feature integration module (IFIM), and the linear transformer layer encompasses Joint feature guided linear attention (JGLA). Specifically, IFIM aims to extract valuable feature information, while JGLA, integrating with the multi-level contextual feature aggregation (MCFA), brings together adjacent, extended regional, and contextual features to guide the linear attention. Regarding linear attention, we revisit inter-feature similarity calculations and reduce the quadratic computational complexity of self-attention to linear complexity. A wide range of experiments shows that the JGLTN network strikes an impressive balance between performance and computational costs. Future work will involve an in-depth exploration of the intrinsic mechanisms of the JGLTN network to identify more precise feature extraction methodologies and more efficient computational strategies to enhance its capacity for further SISR tasks.

Acknowledgements This work was supported by Natural science research project of Guizhou Provincial Department of Education, China (QianJiaoJi[2022]029,QianJiaoHeKY[2021]022).

Data availability Data will be made available on request.

Declarations

Conflict of 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.

References

1. Chen G, Jiao P, Hu Q, Xiao L, Ye Z (2022) Swinstfm: Remote sensing spatiotemporal fusion using swin transformer. IEEE Trans Geosci Remote Sens 60:1–18. https://doi.org/10.1109/TGRS.2022.3182809

2. Wang C, Lv X, Shao M, Qian Y, Zhang Y (2023) A novel fuzzy hierarchical fusion attention convolution neural network for medical image super-resolution reconstruction. Inform Sci 622:424–436. https://doi.org/10.1016/j.ins.2022.11.140
3. Ran R, Deng L-J, Jiang T-X, Hu J-F, Chanussot J, Vivone G (2023) Guidednet: a general cnn fusion framework via high-resolution guidance for hyperspectral image super-resolution. IEEE Trans Cybernet. https://doi.org/10.1109/TCYB.2023.3238200
4. Pang Y, Cao J, Wang J, Han J (2019) Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Trans Inform Forensics Secur 14(12):3322–3331. https://doi.org/10.1109/TIFS.2019.2916592
5. Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307. https://doi.org/10.48550/arXiv.1501.00092
6. Zhao H, Kong X, He J, Qiao Y, Dong C (2020) Efficient image super-resolution using pixel attention. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 56–72. Springer. https://doi.org/10.48550/arXiv.2010.01073
7. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844. https://doi.org/10.48550/arXiv.2108.10257
8. Lu Z, Li J, Liu H, Huang C, Zhang L, Zeng T (2022) Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 457–466. https://doi.org/10.48550/arXiv.2108.11084
9. Wu F, Fan A, Baevski A, Dauphin YN, Auli M (2019) Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430. https://doi.org/10.1145/3340531.3412118
10. Lin X, Ma L, Liu W, Chang S-F (2020) Context-gated convolution. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, pp. 701–718. Springer. https://doi.org/10.48550/arXiv.1910.05577
11. Katharopoulos A, Vyas A, Pappas N, Fleuret F (2020) Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning, pp. 5156–5165. PMLR. https://doi.org/10.48550/arXiv.2006.16236
12. Liu J, Pan Z, He H, Cai J, Zhuang B (2022) Ecoformer: Energy-saving attention with linear complexity. Adv Neural Inform Process Syst 35: 10295–10308. https://doi.org/10.48550/arXiv.2209.09004
13. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136–144. https://doi.org/10.48550/arXiv.1707.02921
14. Xia B, Hang Y, Tian Y, Yang W, Liao Q, Zhou J (2022) Efficient non-local contrastive attention for image super-resolution. In: Proceedings of the AAAI conference on artificial intelligence, 36: 2759–2767 (2022) https://doi.org/10.48550/arXiv.2201.03794
15. Chen Z, Zhang Y, Gu J, Kong L, Yang X, Yu F (2023) Dual aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12312–12321. https://doi.org/10.48550/arXiv.2308.03364
16. Ahn N, Kang B, Sohn K-A (2018) Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 252–268. https://doi.org/10.48550/arXiv.1803.08664
17. Hui Z, Gao X, Yang Y, Wang X (2019) Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th Acm International Conference on Multimedia, pp. 2024–2032. https://doi.org/10.48550/arXiv.1909.11856
18. Liu J, Tang J, Wu G (2020) 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, pp. 41–55. Springer. https://doi.org/10.48550/arXiv.2009.11551
19. Luo X, Xie Y, Zhang Y, Qu Y, Li C, Fu Y (2020) Latticenet: Towards lightweight image super-resolution with lattice block. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pp. 272–289. Springer. https://doi.org/10.1007/978-3-030-58542-6_17
20. Wang X, Dong C, Shan Y (2022) Repsr: Training efficient vgg-style super-resolution networks with structural re-parameterization and batch normalization. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2556–2564. https://doi.org/10.1145/3503161.3547915
21. Gao G, Li W, Li J, Wu F, Lu H, Yu Y (2022) Feature distillation interaction weighting network for lightweight image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 661–669. https://doi.org/10.48550/arXiv.2112.08655
22. Luo X, Qu Y, Xie Y, Zhang Y, Li C, Fu Y (2023) Lattice network for lightweight image restoration. IEEE Trans Pattern Anal Mach Intell 45(4):4826–4842. https://doi.org/10.1109/TPAMI.2022.3194090
23. Wang H, Zhang Y, Qin C, Van Gool L, Fu Y (2023) Global aligned structured sparsity learning for efficient image super-resolution. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2023.3268675
24. Guo J, Zou X, Chen Y, Liu Y, Liu J, Yan Y, Hao J (2023) Asconvsr: Fast and lightweight super-resolution network with assembled convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1582–1592. https://doi.org/10.48550/arXiv.2305.03387
25. Muqeet A, Hwang J, Yang S, Kang J, Kim Y, Bae S-H (2020) Multi-attention based ultra lightweight image super-resolution. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 103–118. Springer. https://doi.org/10.48550/arXiv.2008.12912
26. Chen H, Gu J, Zhang Z (2021) Attention in attention network for image super-resolution. arXiv preprint arXiv:2104.09497. https://doi.org/10.48550/arXiv.2104.09497
27. Zhang D, Li C, Xie N, Wang G, Shao J (2021) Pffn: Progressive feature fusion network for lightweight image super-resolution. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3682–3690. https://doi.org/10.1145/3474085.3475650
28. Kong F, Li M, Liu S, Liu D, He J, Bai Y, Chen F, Fu L (2022) Residual local feature network for efficient super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 766–776. https://doi.org/10.48550/arXiv.2205.07514
29. Du Z, Liu D, Liu J, Tang J, Wu G, Fu L (2022) Fast and memory-efficient network towards efficient image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 853–862. https://doi.org/10.48550/arXiv.2204.08397
30. Fan C-M, Liu T-J, Liu K-H(2022) Sunet: Swin transformer unet for image denoising., 2333–2337. https://doi.org/10.1109/ISCAS48785.2022.9937486
31. Tsai F-J, Peng Y-T, Lin Y-Y, Tsai C-C, Lin C-W (2022) Stripformer: Strip transformer for fast image deblurring. In: European Conference on Computer Vision, pp. 146–162. Springer. https://doi.org/10.48550/arXiv.2204.04627

32. Ma F, Sun J (2022) Crossuformer: A cross attention u-shape transformer for low light image enhancement. In: Proceedings of the Asian Conference on Computer Vision, pp. 928–943
33. Song Y, He Z, Qian H, Du X (2023) Vision transformers for single image dehazing. IEEE Transactions on Image Processing 32, 1927–1941. https://doi.org/10.48550/arXiv.2204.03883
34. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22367–22377 (2023). https://doi.org/10.48550/arXiv.2205.04437
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021). https://doi.org/10.48550/arXiv.2103.14030
36. Zhang X, Zeng H, Guo S, Zhang L (2022) Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision, pp. 649–667. Springer. https://doi.org/10.48550/arXiv.2203.06697
37. Choi H, Lee J, Yang J(2023) N-gram in swin transformers for efficient lightweight image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2071–2081 . https://doi.org/10.48550/arXiv.2211.11436
38. Cai H, Gan C, Han S (2022) Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 . https://doi.org/10.48550/arXiv.2205.14756
39. Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, Hawkins P, Davis J, Mohiuddin A, Kaiser L, et al. (2020) Rethinking attention with performers. arXiv preprint arXiv:2009.14794. https://doi.org/10.48550/arXiv.2009.14794
40. You, H., Xiong, Y., Dai, X., Wu, B., Zhang, P., Fan, H., Vajda, P., Lin, Y.C.: Castling-vit: Compressing self-attention via switching towards linear-angular attention at vision transformer inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14431–14442 (2023). https://doi.org/10.48550/arXiv.2211.10526
41. Chen, B., Dao, T., Winsor, E., Song, Z., Rudra, A., Ré, C.: Scatterbrain: Unifying sparse and low-rank attention. Advances in Neural Information Processing Systems 34, 17413–17426 (2021). https://doi.org/10.48550/arXiv.2110.15343
42. Dass, J., Wu, S., Shi, H., Li, C., Ye, Z., Wang, Z., Lin, Y.: Vitality: Unifying low-rank and sparse approximation for vision transformer acceleration with a linear taylor attention. In: 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 415–428 (2023). IEEE. https://doi.org/10.48550/arXiv.2211.05109
43. Lipson L, Teed Z, Deng J (2021) Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV), pp. 218–227. IEEE. https://doi.org/10.48550/arXiv.2109.07547
44. Niu B, Wen W, Ren W, Zhang X, Yang L, Wang S, Zhang K, Cao X, Shen H (2020) Single image super-resolution via a holistic attention network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pp. 191–207. Springer. https://doi.org/10.48550/arXiv.2008.08767
45. Han D, Pan X, Han Y, Song S, Huang G (2023) Flatten transformer: Vision transformer using focused linear attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5961–5971. https://doi.org/10.48550/arXiv.2308.00442
46. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301. https://doi.org/10.48550/arXiv.1807.02758
47. Hui Z, Wang X, Gao X (2018) Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 723–731 . https://doi.org/10.48550/arXiv.1803.09454
48. Lan R, Sun L, Liu Z, Lu H, Pang C, Luo X (2020) Madnet: A fast and lightweight network for single-image super resolution. IEEE Trans Cybernet 51(3):1443–1453. https://doi.org/10.1109/TCYB.2020.2970104
49. Wang L, Dong X, Wang Y, Ying X, Lin Z, An W, Guo Y (2021) Exploring sparsity in image super-resolution for efficient inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4917–4926. https://doi.org/10.48550/arXiv.2006.09603
50. Park K, Soh JW, Cho NI (2021) Dynamic residual self-attention network for lightweight single image super-resolution. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2021.3134172
51. Luo X, Qu Y, Xie Y, Zhang Y, Li C, Fu Y (2022) Lattice network for lightweight image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4), 4826–4842. https://doi.org/10.48550/arXiv.2112.08655
52. Sun B, Zhang Y, Jiang S, Fu Y (2023) Hybrid pixel-unshuffled network for lightweight image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2375–2383. https://doi.org/10.48550/arXiv.2203.08921
53. Timofte, R, Agustsson E, Van Gool L, Yang M-H, Zhang L (2017) Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 114–125. https://doi.org/10.1109/CVPRW.2017.149
54. Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. https://doi.org/10.5244/C.26.135
55. Zeyde, R., Elad, M., Protter, M.: 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, pp. 711–730 (2012). Springer. https://doi.org/10.1007/978-3-642-27413-8_47
56. Martin, D., Fowlkes, C., Tal, D., Malik, J.: 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, pp. 416–423 (2001). IEEE. https://doi.org/10.1109/ICCV.2001.937655
57. Huang J-B, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206. https://doi.org/10.1109/CVPR.2015.7299156
58. Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications 76, 21811–21838. https://doi.org/10.48550/arXiv.1510.04389
59. Agustsson E, Timofte R (2017) Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135. https://doi.org/10.1109/CVPRW.2017.150
60. Cai J, Zeng H, Yong H, Cao Z, Zhang L (2019) Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3086–3095. https://doi.org/10.48550/arXiv.1904.00523