Multimedia Systems (2025) 31:333
https://doi.org/10.1007/s00530-025-01910-6

REGULAR PAPER

Dual attention transformer with adaptive frequency enhancement for real-world Chinese–English scene text image super-resolution

Yanbin Liu[1] · Qin Shi[1] · Ziming Zhu[1] · Xiaofeng Ling[1] · Yu Zhu[1,2]

Received: 1 March 2025 / Accepted: 16 June 2025 / Published online: 21 August 2025

Abstract

Scene text image super-resolution (STISR) has achieved remarkable performance on the pure English dataset, TextZoom. Nevertheless, existing STISR models are primarily designed for fixed-size English text images, limiting their ability to reconstruct structurally complex characters like Chinese characters. Due to the squared computational complexity of standard self-attention, existing methods usually restrict the self-attention calculation within local windows, leading to a limited receptive field. In this paper, we propose a dual attention transformer with adaptive frequency enhancement (DA^2FE) model which alternates between two complementary window attention mechanisms. Specifically, dense window attention (DWA) facilitates interactions between neighboring tokens, which is beneficial for learning local features. Sparse window attention (SWA) establishes associations between spaced tokens, enabling effective global information extraction. Additionally, we incorporate a parallel depth-wise convolution (DWConv) branch to establish cross-window relations. Subsequently, a spatial-channel interaction module is employed to facilitate the bi-directional interaction between the window attention branch and the DWConv branch. Furthermore, we design a feed-forward network with adaptive frequency enhancement (FFN-AFE), which introduces a learnable quantitative matrix in the frequency domain to adaptively select and enhance significant frequency information. Finally, the output features from multiple layers are aggregated and refined to provide more comprehensive information for SR reconstruction. Comparative experiments with advanced methods on the Real-CE dataset demonstrate our superior performance in terms of objective indicators and subjective visual results for both $2\times$ and $4\times$ STISR tasks. Furthermore, DA^2FE exhibits excellent results on natural image super-resolution datasets, further demonstrating its broad applicability.

Keywords Real-world text image super-resolution · Dual-branch network · Dense and sparse window attention · Spatial-channel interaction · Adaptive frequency enhancement

1 Introduction

Image super-resolution (SR) aims to reconstruct a degraded low-resolution (LR) image into a high-resolution (HR) image with intricate detail information and distinct texture features. Scene text image super-resolution (STISR) is a subfield of SR, focusing on the restoration of text images. Influenced by the shooting environment and imaging equipment, text images in real scenes may have blurry character contours and insufficient detail information, leading to poor text recognition accuracy. Therefore, it is of great research significance to adopt reasonable pre-processing methods to enhance the clarity of text images before text recognition.

Early STISR methods [1–3] are typically trained on synthetic datasets, where LR images are obtained from corresponding HR images through specific degradation

Yanbin Liu and Qin Shi have contributed equally to this work.

Communicated by Chenggang Yan.

✉ Yu Zhu
zhuyu@ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

[2] Shanghai Engineering Research Center of Internet of Things for Respiratory Medicine, Shanghai 200032, China

Fig. 1 Comparison between HR, synthetic LR and real LR text images in RealCE dataset. 'HR' and 'Real LR' are captured with different focal lengths in real scenes. 'Synthetic LR' denotes the bicubic down-sampled image of 'HR'. Evidently, the real LR images are much more challenging than the synthetic LR images

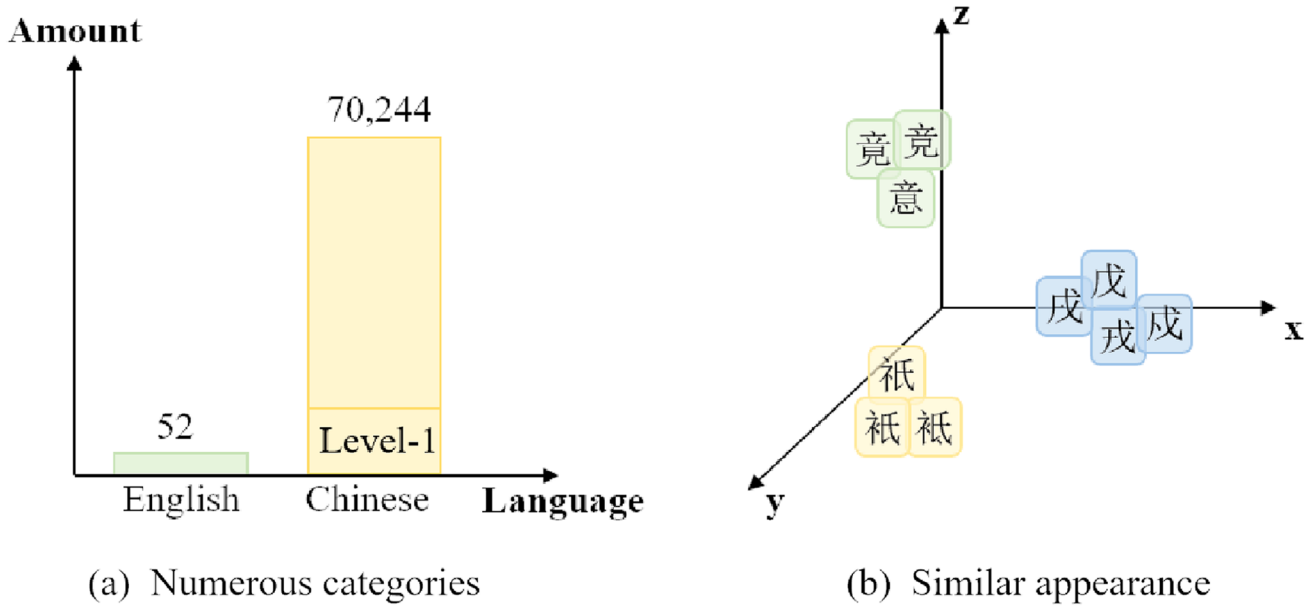

Fig. 2 Characteristics of Chinese characters. Compared to English characters, Chinese characters possess a greater variety of categories and intricate stroke structures

operations (e.g., blurring, downsampling). As shown in Fig. 1, the degradation forms in real scenes are more complex and diverse, thus the artificially predefined degradation methods cannot accurately simulate the real degradation situations. Consequently, the STISR models trained on the synthetic datasets exhibit poor generalization ability and struggle to address the intricate degradation issues encountered in real-world scenarios.

Wang et al. [4] proposed the first real-world STISR dataset, termed TextZoom. However, TextZoom is limited to English text images, where the character set is limited to alphanumeric symbols ('0' to '9', 'a' to 'z') and exhibits relatively simplistic stroke structures. Moreover, TextZoom focuses on the reconstruction of fixed-size (16×64) LR text images, with the super-resolution (SR) outputs being 32×128. As a result, SR models trained on TextZoom struggle to reconstruct structurally complex characters, such as Chinese characters, and fail to generalize to text images of varying sizes. To address these limitations, Ma et al. [5] proposed a new real-world STISR dataset, namely Real-CE, which contains a large number of Chinese text images with more diverse sizes. Known from Fig. 2a, the number of Chinese characters is 70,244 (including 3,755 commonly used Level-1 characters) which is much larger than the scale of English characters ('a'-'z', 'A'-'Z'). Furthermore, Chinese characters are composed of multiple strokes and complex structures. As shown in Fig. 2b, many Chinese characters are very similar in appearance and a slight deviation in strokes carries entirely different meanings, such as '戊', '戊' and '戊'.

Existing SR networks primarily adopt either CNN-based or Transformer-based architectures. While CNN-base methods [6–9] excel at local feature extraction, they suffer from restricted receptive fields and cannot effectively model global dependencies. In contrast, Transformer-based methods have revolutionized the image super-resolution field by leveraging global modeling capabilities. To mitigate the quadratic computational complexity of standard self-attention [10], recent works [11–14] generally resort to local window attention [15], where neighboring tokens are grouped into windows for self-attention computation. Nevertheless, the tokens in each window are sourced from a dense area of

feature map, leading to a restricted receptive field. To overcome this limitation, we propose sparse window attention, which performs sparse sampling at a certain interval size to to efficiently capture long-range dependencies.

In this paper, we propose a novel architecture termed DA^2FE for the real-world Chinese-English STISR task. The main part of DA^2FE employs two complementary window attention mechanisms: dense window attention (DWA) and sparse window attention (SWA). DWA enables interactions among neighbouring tokens, facilitating the learning of local features. Conversely, SWA establishes correlations between spaced tokens, effectively capturing global dependencies. Alternating between DWA and SWA can effectively enhance both local and global feature modeling abilities with lower computational complexity. Notably, both DWA and SWA primarily focus on the interactions within local windows. Consequently, we incorporate a parallel depth-wise convolution (DWConv) branch to establish connections across windows. Furthermore, we propose a spatial channel interaction (SCI) module to facilitate the bi-directional interaction between window attention branch and DWConv branch, adaptively adjusting feature maps in both the spatial and channel dimensions. To enhance crucial frequency information, a feed-forward network with adaptive frequency enhancement (FFN-AFE) is designed, introducing a learnable quantitative matrix in the frequency domain to adaptively determine the importance of each frequency. Finally, by integrating features from various layers, we can provide more comprehensive and abundant information for SR reconstruction.

The main contributions are summarized as follows:

(1) A Dual Attention Transformer with Adaptive Frequency Enhancement (DA^2FE) model is proposed for Chinese-English STISR task. In the deep feature extraction stage, the proposed dense window attention (DWA) and sparse window attention (SWA) are alternately used, respectively for extracting local features and establishing global associations.
(2) The SCI module interacts between the window attention branch and the DWConv branch to facilitate bi-directional information exchange across both the channel and spatial dimensions.
(3) The proposed FFN-AFE incorporates a learnable quantization matrix within the frequency domain, enabling it to adaptively select and enhance significant frequency information, thus facilitating the extraction of high-frequency details.
(4) Extensive experiments on RealCE dataset demonstrate that DA^2FE outperforms previous methods across all evaluation metrics, and achieves excellent visual results in reconstructing Chinese characters with complex structures. Moreover, DA^2FE exhibits broad applicability on general SR datasets.

2 Related work

2.1 Natural image super resolution

The seminal work SRCNN [6] performs bicubic interpolation on the LR images and then adopts three convolutional layers to learning the mappings between LR images and HR images. Later on, an improved work named FSRCNN [7] is proposed for the drawback of slow training speed of SRCNN. VDSR [8] introduces the global residual structure to solve the problem of gradient vanishing and gradient explosion. RCAN [16] incorporates channel attention mechanism to enhance feature modeling ability. SRGAN [9] proposes adversarial loss and content perceptual loss to prevent the generation of overly smooth images. CRAFT [11] demonstrates that Transformer is adept at capturing low-frequency information but have limited capability in extracting high-frequency features. DAT [12] alternately utilizes spatial self-attention and channel self-attention to aggregate global features. HAT [13] combines window-based self-attention, channel attention, and overlapping cross-attention for SR task. CFAT [14] introduces additional triangular window attention to eliminate the boundary distortion problem. The above SR methods are primarily CNN-based and Transformer-based, encountering a trade-off between global modeling and efficient computation.

Besides, recent technological breakthroughs in other computer vision domains have also inspired novel methodologies for SR research. Specifically, HYMATOD [17] proposes a new hybrid multi-attention module to decrease noise and ambiguity in the conventional attention mechanism. CAERDCF [18] incorporates a selective spatial regularizer to safeguard essential object information while concurrently mitigating boundary effects. ASTABSCF [19] presents an adaptive spatially regularized technique for learning efficient spatial weight for a particular object.

2.2 Scene text image super resolution

Early works [1–3] are trained on synthetic datasets and cannot be well generalized to real-world text images. To solve this problem, Wang et al. [4] propose the first real-world text image super-resolution dataset, termed TextZoom, and simultaneously develop a new model termed TSRN. PCAN [20] designs a parallelly contextual attention block to model contextual dependencies between orthogonal features. TBSRN [21] designs a location-aware module and a content-aware module to further enhance the character

reconstruction effect. TPGSR [22] employs text recognition network CRNN [23] to extract text prior (TP) features which are then concatenate with image features to guide the SR reconstruction process. TATT [24] improves upon the feature fusion method in TPGSR by utilizing a self-attention mechanism instead of channel concatenation. MARCONet [25] exploits generative structural priors to restore precise strokes of Chinese characters. TextDiff [26] is the first work to adapt diffusion model for STISR task. DiffTSR [27] proposes a text diffusion model for text recognition, which can guide image diffusion model to generate text images with correct structures. Existing STISR methods are primarily designed for English text images, struggling to reconstruct structurally complex characters like Chinese characters, especially for real-world Chinese text images. In this paper, we propose a novel architecture for real-world Chinese-English STISR task, which alternately employs dense window attention and sparse window attention to capture global interactions with lower computational cost.

2.3 Scene text recognition

CRNN [23] utilizes a CNN for feature extraction, followed by a RNN for sequential context modeling. ASTER [28] introduces a spatial transformer network (STN) [29] to correct irregular text images and predict character sequences based on the attention mechanism. ABINet [30] constructs a text recognition network based on a visual model and a linguistic model. Yu et al. [31] propose a two-stage framework for Chinese text recognition (CTR), where a CLIP-like [32] pretrained model is responsible for learning the canonical representations of Chinese characters and the learned representations are used to supervise the CTR model. SegCTC [33] incoporates the advantages of recognition branches based on implicit segmentation and explicit segmentation for handwritten Chinese text recognition task. Current STR methods perform well in recognizing clear text images, but the recognition accuracy significantly decreases when applied to low-quality text images. Therefore, STISR can be adopted as a pre-processing method for STR, aiming to improve text recognition accuracy by reconstructing clear text content.

3 The proposed method

3.1 Overall architecture

As illustrated in Fig. 3, the proposed DA^2FE is mainly divided into four stages: shallow feature extraction, deep feature extraction, feature fusion and image reconstruction.

First, we use Canny edge detector [34] to extract LR text edge maps. Then the 3-channel RGB image $I_{LR} \in R^{H\times W\times 3}$ and the corresponding 1-channel text edge map $M_{LR} \in R^{H\times W\times 1}$ are concatenated to obtain the 4-channel input image, denoted as $I_{in} \in R^{H\times W\times 4}$. Canny edge map can be regarded as a prior label, enabling the SR network to distinguish between foreground and

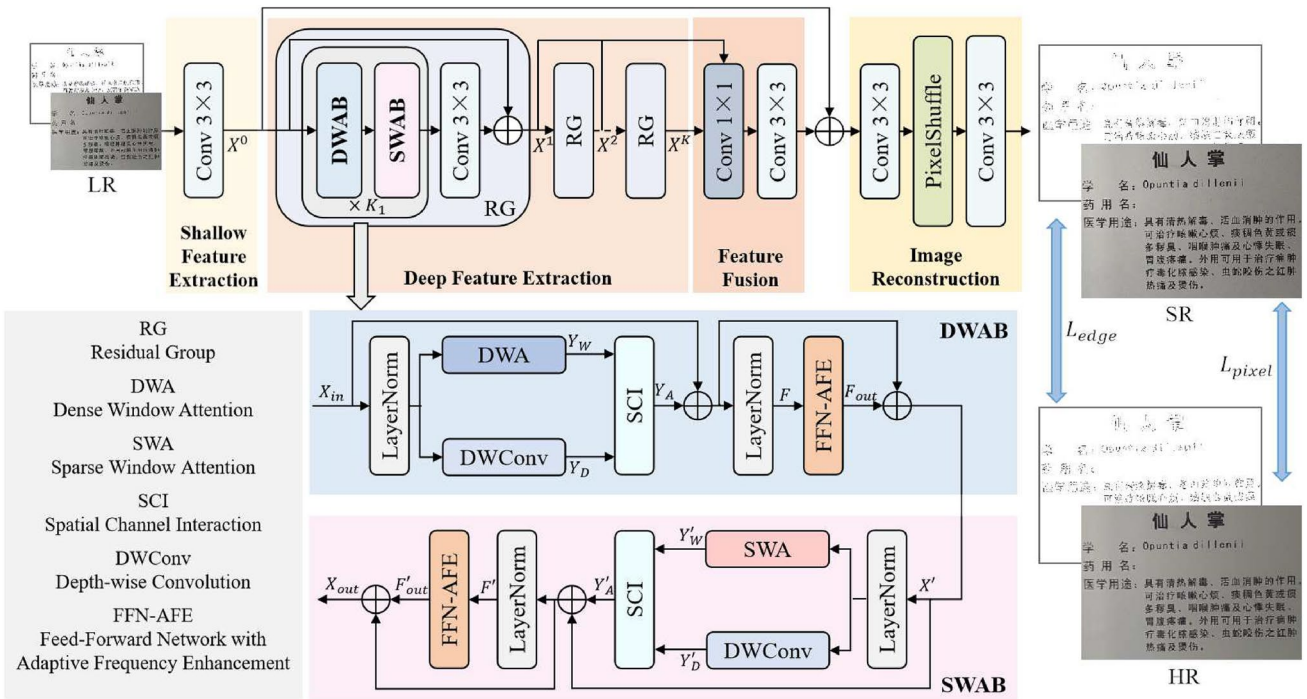

Fig. 3 The architecture of DA^2FE

background, thus focusing more on the reconstruction of foreground region. Even with noise or incompleteness, the edge map still provides coarse localization of text regions. As shown in Fig. 4, the character shape and structure in the original LR-HR image pair are blurry, but the corresponding Canny edge map can effectively enhance the text area. Then, a 3×3 convolutional layer is used to extract shallow features from I_{in}, and a series of residual groups (RGs) are employed to gradually extract deep features. Following, the output feature of each RG are concatenated along the channel dimension and refined by a 1×1 convolution and a 3×3 convolution. Finally, a pixel-shuffle layer is used to upsample the feature to the target size.

Specifically, each RG consists of K_1 pairs of attention blocks, namely dense window attention block (DWAB), sparse window attention block (SWAB). Both DWAB and SWAB adopt a parallel dual-branch architecture, which includes a window attention branch and a depth-wise convolution (DWConv) branch. Window attention concentrates on the interaction and aggregation within individual windows, while the incorporation of DWConv branch can establish connections across these windows. Additionally, DWConv operation can supplement local information for the window attention branch, which facilitates the capture of finer details. Subsequently, the spatial channel interaction (SCI) module is utilized to fuse and interact the two features extracted by window attention branch and DWConv branch. Finally, feed-forward network with adaptive frequency enhancement (FFN-AFE) is designed to adaptively enhance important frequency domain information.

3.2 Dual window attention

The standard self-attention (SA) mechanism calculates similarities among all pixels on the entire feature map, which can effectively model global dependencies but suffers from squared computational complexity. To mitigate the computational cost, existing works usually limit self-attention calculation within local windows. Nevertheless, the tokens of each window are usually sourced from a dense region of the image, which will result in a restricted receptive field. To address this issue, we design two complementary types of window attention, namely dense window attention (DWA) and sparse window attention (SWA). DWA calculates similarities between neighboring tokens, enabling the model to learn local features. In contrast, SWA learns associations between spaced tokens to supplement global information. The alternating application of DWA and SWA can effectively enhance global and local feature modeling capabilities with reduced computational complexity.

3.2.1 Process of DWA and SWA

As shown in Fig. 5, both DWA and SWA include three processes: window partition, window attention and window reverse.

Window Partition. As depicted in Fig. 5, DWA samples tokens from adjacent positions on the feature map and divides the feature map into windows of size $S \times S$, with the number of windows being $\frac{HW}{S^2}$. SWA samples tokens at a certain interval (denoted by I) on the feature map, and divides the feature map into windows of size $\frac{H}{I} \times \frac{W}{I}$, with the number of windows being I^2. Consequently, the alternating application of DWA and SWA provides interactions

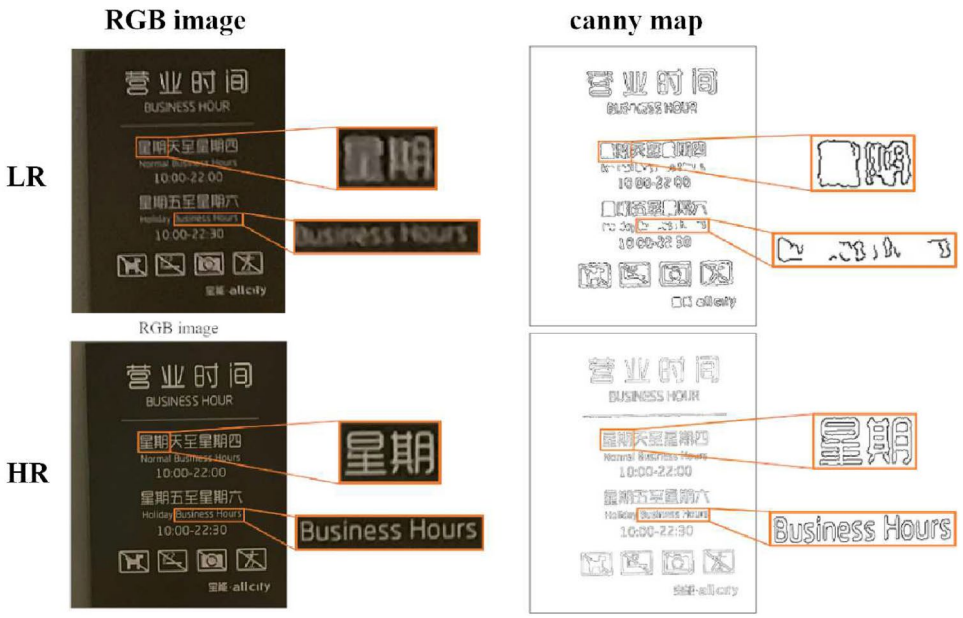

Fig. 4 RGB images (left) and their Canny edge maps (right). In the Canny edge maps, the pixel values of the character contour regions are set to 0, while those of the background regions are set to 1. Using the text edge map as prior knowledge enables the network to effectively perceive text structure and strokes

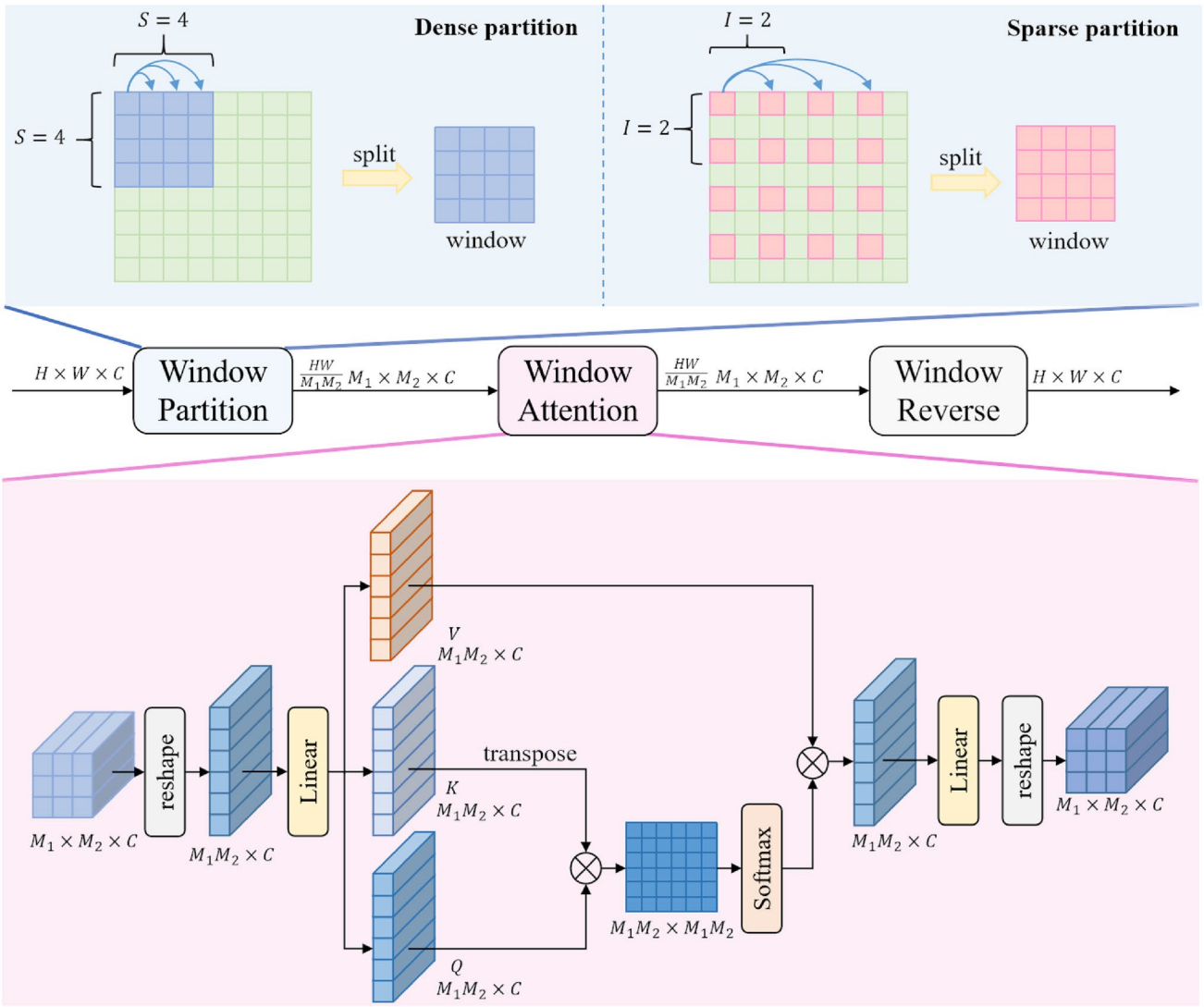

Fig. 5 The schematic diagram of DWA and SWA. Both of them encompass three processes: window partition, window attention and window reverse. The only difference between DWA and SWA lies in their distinct window partitioning strategies. Specifically, DWA conducts dense sampling of neighbouring tokens to delineate a set of windows, whereas SWA performs sparse sampling at a certain interval size

for tokens from both dense regions and sparse regions, resulting in a wider receptive field. The steps of window partition can be expressed as follows:

$$\{X_i\} = WinPartition(X_{in}) \quad i = 1, 2, \ldots, N \tag{1}$$

where M_1 and M_2 respectively denote the height and width of window X_i, and N represents the number of windows.

For DWA, $M_1 = S$, $M_2 = S$, $N = \frac{HW}{M_1 M_2} = \frac{HW}{S^2}$; For SWA, $M_1 = \frac{H}{I}$, $M_2 = \frac{W}{I}$, $N = \frac{HW}{M_1 M_2} = I^2$. Figure 6 is a schematic diagram of multiple windows divided by DWA and SWA on the test image. Taking a 9×9 feature map as an illustration, DWA aggregates 3×3 adjacent tokens into windows, resulting in a total of 9 windows. For SWA, the interval size of sparse sampling is configured as 3, yielding 9 windows.

Window Attention. After window partition operation, self-attention calculation is conducted separately within each window. For the *i*-th window, denoted as $X_i \in R^{M_1 \times M_2 \times C}$, the first step is to flatten it into a vector of length $C \times 1$, with the number of vectors being $M_1 \times M_2$. Then, the self-attention is calculated in parallel within different feature subspaces. Assuming that the number of heads is *h* which means that the features are divided into *h* groups along the channel dimension, with each group having a channel size of $d = \frac{C}{h}$. Then we describe the steps of self-attention calculation for the *m*-th head of the *i*-th window.

First, the query matrix (Query), key matrix (Key) and value matrix (Value) are extracted through three parallel linear mapping (Linear) layers:

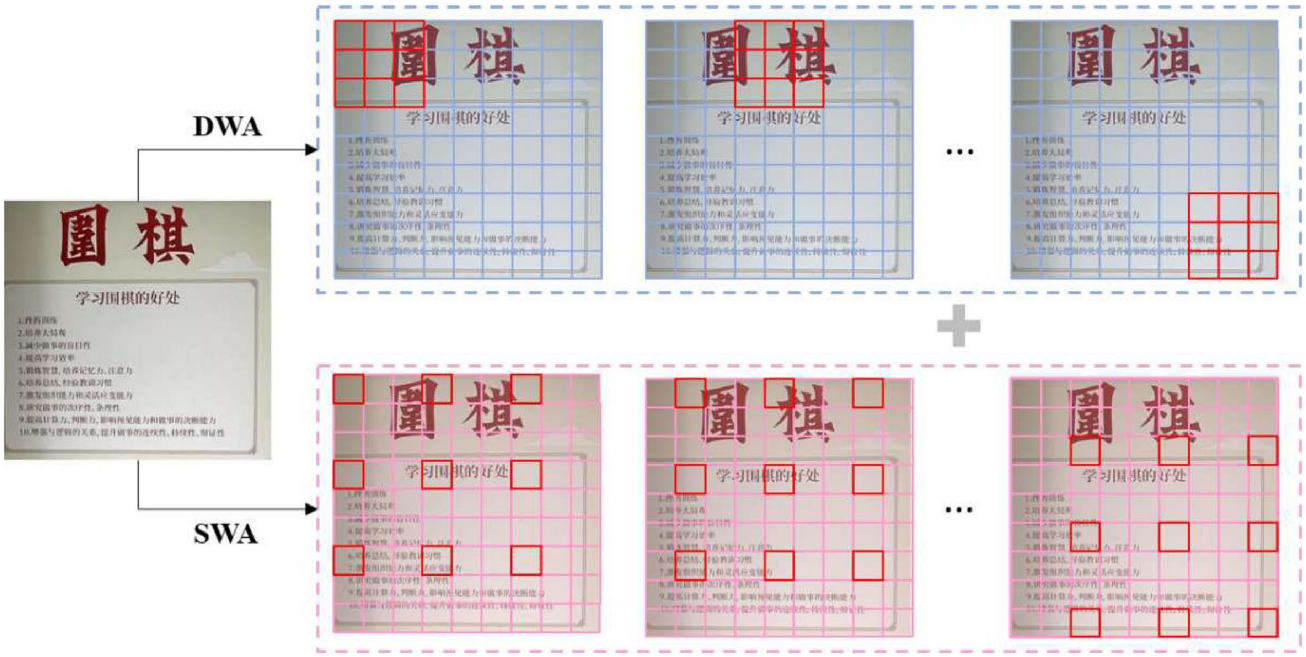

Fig. 6 Illustration of the window partitioning strategies for a 9×9 feature map. DWA combines 3×3 neighboring tokens into windows, resulting in 9 windows. SWA performs sparse sampling with an interval of 3, yielding 9 windows

$$Q_i^m = X_i W_m^Q, \quad K_i^m = X_i W_m^K, \quad V_i^m = X_i W_m^V \tag{2}$$

where $Q_i^m \in R^{C\times d}$, $K_i^m \in R^{C\times d}$ and $V_i^m \in R^{C\times d}$ respectively represent the Query, Key and Value for the *m*-th head of the *i*-th window.

Next, we calculate the attention as follows:

$$Y_i^m = Attention\left(Q_i^m, K_i^m, V_i^m\right) = Softmax\left(\frac{Q_i^m \left(K_i^m\right)^T}{N} + B\right) V_i^m \tag{3}$$

where *T* represents the matrix transpose operation. *B* denotes the learnable position encoding. Y_i^m represents the attention result for the *m*-th head of the *i*-th window.

After calculating *h* groups of attention in parallel, the attention results $\{Y_i^m\}$ are concatenated along the channel dimension:

$$Y_i = Concat\left(Y_i^1, Y_i^2, \ldots, Y_i^h\right) \tag{4}$$

where $Y_i \in R^{M_1\times M_2\times C}$ represents the final attention result of the *i*-th window.

Window Reverse. The calculation results of *N* windows are denoted as $\{Y_1, Y_2, \ldots, Y_N\}$. Finally, the *N* windows are recombined in the original partition order to obtain the final output $Y_W \in R^{H\times W\times C}$.

$$Y_W = WinReverse(\{Y_i\}), \quad i = 1, 2, \ldots, N \tag{5}$$

In conclusion, both DWA and SWA can be regard as a type of local window attention strategy since the interactions of tokens are restricted in local regions. Notably, SWA can interact with more distant tokens, compensating for the lack of global information. With the alternating application of DWA and SWA, the model can capture both local and global dependencies simultaneously.

3.2.2 Computational complexity

In this session, we analyze the computational complexity of standard SA, DWA and SWA on the feature map $X \in R^{H\times W\times C}$. The standard SA mechanism calculates similarities among all pixels on the entire feature map which can effectively model long-range dependencies, but suffers from quadratic computational complexity. The computational complexity of standard SA is shown as follows:

$$\Omega(SA) = 4HWC^2 + 2(HW)^2C \tag{6}$$

As shown in Fig. 5, DWA conducts dense sampling of neighbouring tokens to form a group of windows. The entire feature map is split into several windows, with each window containing $S \times S$ tokens. Subsequently, self-attention calculation is conducted within each window for $S \times S$ times. The computational complexity of DWA is presented as follows:

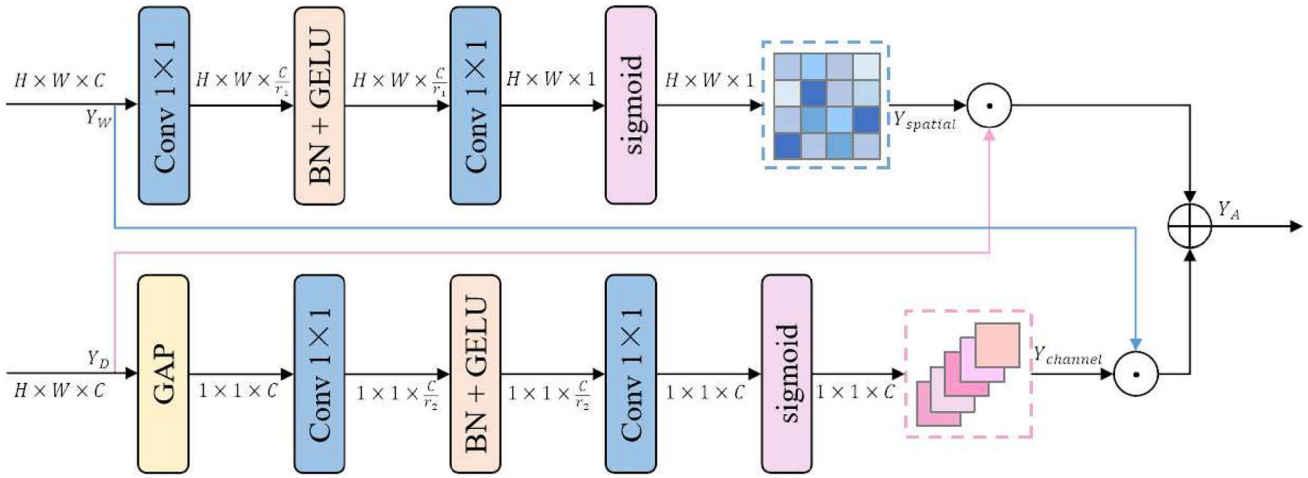

Fig. 7 The structure of SCI module. It contains two parallel branches, respectively for generating a channel attention map $Y_{channel}$ and spatial attention map $Y_{spatial}$. Then, the outputs of the two branches are element-wise multiplied with the input of another branch to perform a bi-directional interation

$$\begin{aligned}\Omega(DWA) &= \left(4S^2C^2 + 2S^4C\right) \times \frac{H}{S} \times \frac{W}{S} \\ &= 4HWC^2 + 2S^2HWC\end{aligned} \tag{7}$$

Similarly, SWA performs sparse sampling at a certain interval (denoted by I). The entire feature map is split into several windows, with each window containing $\frac{H}{I} \times \frac{W}{I}$ tokens. Subsequently, self-attention calculation is carried out within each window for I^2 times. The computational complexity of SWA is shown as follows:

$$\begin{aligned}\Omega(SWA) &= \left(4\frac{H}{I} \times \frac{W}{I}C^2 + 2\left(\frac{H}{I} \times \frac{W}{I}\right)^2 C\right) \times I \times I \\ &= 4HWC^2 + 2\frac{H}{I}\frac{W}{I}HWC\end{aligned} \tag{8}$$

Since $S^2 \ll HW$ and $\frac{H}{I}\frac{W}{I} < HW$ in practical applications, DWA and SWA can effectively reduce the computational complexity compared with the standard SA.

3.3 Spatial channel interaction

DWAB and SWAB are alternately used in RGs respectively for learning local information and extracting global dependencies. Both of them adopt a parallel dual-branch architecture, consisting of a window attention branch and a depth-wise convolution (DWConv) branch. The window attention mechanism computes similarities within local windows and shares weights across the channel dimension, inherently limiting its ability to effectively exploit cross-channel information. In contrast, DWConv splits the input feature into multiple single-channel feature maps and performs convolution operation channel by channel. Compared with standard convolution, DWConv prioritizes channel-specific information while significantly reducing computational overhead. However, this approach fails to account for inter-channel correlations at identical spatial locations. Based on this, we introduce a spatial channel interaction (SCI) module, which performs bi-directional interaction between the features extracted by window attention and DWConv. In other words, SCI module is designed to enhancinge the modeling ability of DWA and SWA in

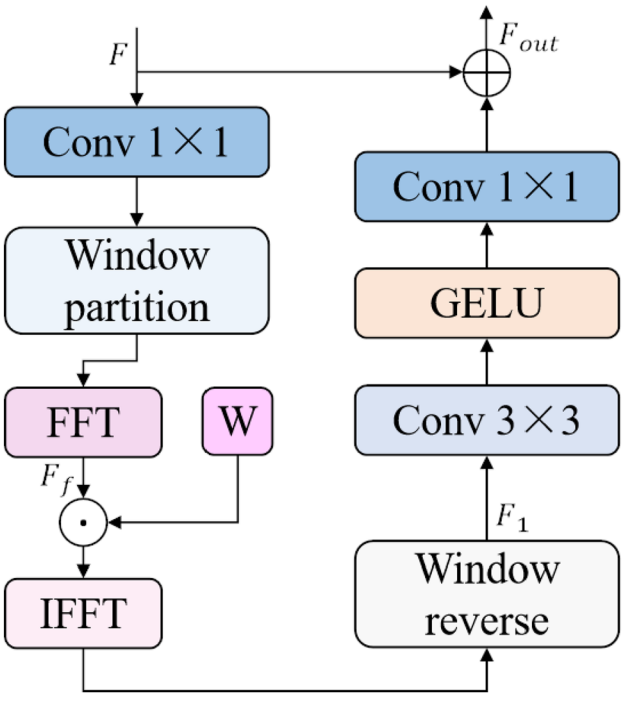

Fig. 8 The structure of FFN-AFE. FFT is introduced to generate spectral feature F_f, which is element-wise multiplied with a learnable quantization matrix W to adaptively determine the significance of each frequency

the channel dimension, as well as the modeling ability of DWConv in the spatial dimension.

In this session, we introduce the SCI module following DWA as an example. As depicted in Fig. 7, the input of SCI module consists of two parts: the output of DWA denoted by $Y_W \in R^{H\times W\times C}$ and the output of DWConv denoted by $Y_D \in R^{H\times W\times C}$. SCI module contains two parallel branches: channel attention branch and spatial attention branch. The channel attention branch aims to enhance the feature representation ability of the window attention in the channel dimension, while the spatial attention branch offers effective spatial guidance for DWConv.

The specific process is as follows: Y_W is fed into the spatial attention branch to generate a spatial attention map $Y_{spatial} \in R^{H\times W\times 1}$ with pixel values ranging from 0 to 1. Simultaneously, Y_D is also fed into the channel attention branch to generate a channel attention map $Y_{channel} \in R^{1\times 1\times C}$. Subsequently, the outputs of the two branches are element-wise multiplied with the inputs of another branch. Specifically, $Y_{channel}$ is element-wise multiplied with the feature map Y_W, which adds channel attention weights to Y_W. Similarly, $Y_{spatial}$ is element-wise multiplied with the feature map Y_D, which adds spatial attention weights to Y_D. The results of the two parallel branches are then added to obtain the final output Y_A. The specific process is shown as follows:

$$\begin{aligned} Y_{Spatial} = SA\left(Y_W\right), \;\; Y_{Channel} = CA\left(Y_D\right) \\ Y_A = Y_{Channel} \odot Y_W + Y_{Spatial} \odot Y_D \end{aligned} \tag{9}$$

where *SA* and *CA* respectively represent the spatial attention branch and the channel attention branch. $\odot$ represents element-wise multiply.

3.4 Feed-forward network with adaptive frequency enhancement

The Transformer architecture [10] employs a feed-forward network (FFN) to to improve the features extracted by self-attention. Consequently, designing an effective FFN is of great significance to facilitate high-quality image reconstrcution. To enhance the focus on frequency features, we introduce discrete fourier transform (DFT) which converts the input features with spatial dimension of $H \times W$ to continuous frequencies. The calculation process is as follows:

$$F(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} f(x,y)\cdot e^{-i2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)} \tag{10}$$

where $f(x, y)$ denotes the pixel value at the coordinate (x, y) in the spatial domain and $F(u, v)$ represents the complex value at the coordinates (u, v) in the frequency spectrum.

Each point in the frequency spectrum is calculated from all the pixels in the spatial domain. As a result, the frequency spectrum contains rich global information, which is beneficial for extracting long-range contextual dependencies. Given that the DFT result of real vector is conjugate symmetric, it is sufficient to compute only within half of the feature map. As a widely adopted technique, fast fourier transform (FFT) decomposes DFT into sub-problems and performs recursive computations, which can significantly reduce computational complexity.

In this paper, we design a feed-forward network with adaptive frequency enhancement (FFN-AFE), which introduces a learnable quantization matrix in the frequency domain to adaptively select and preserve crucial frequency information, thus enhancing the ability to extract high-frequency components corresponding to character edge details. Although the FFN-AFE module and Canny edge guidance employ distinct implementation mechanisms, they yield similar outcomes, as both are designed to strengthen character regions. When the Canny detection is suboptimal (e.g., in low-contrast or noisy regions), the learnable quantization matrix automatically compensates by amplifying the intrinsic frequency signatures of text structures.

The structure of the FFN-AFE is illustrated in Fig. 8. First, a 1×1 convolution layer is employed to expand the channel dimension to $\gamma \cdot C$, where γ is the channel expansion coefficient. Subsequently, we partition the feature map with the window size of 8×8 to effectively reduce the computational complexity. Following, FFT is performed in each window separately to obtain the corresponding spectral feature F_f. Besides, a learnable frequency domain matrix W is initialized with all elements set to 1. During the training process, W is adaptively updated and element-wise multiplied with the spectral features F_f, which can adaptively enhance important frequency information. Next, inverse fast fourier transform (IFFT) is performed in each window to transform the feature map from the frequency domain to the spatial domain. After that, the windows are pasted back to their original locations to restore the original size of the feature map. Then, a 3×3 convolution layer and GELU activation function are used to enhance the feature extraction ability and nonlinear representation ability of the network. Subsequently, a 1×1 convolutional layer is used to restore the original channel count. Finally, the output is added to the input feature F to obtain the final output. The specific process is shown below:

Table 1 Quantitative comparison with various methods for both $\times 2$ and $\times 4$ SR tasks

Methods	Type	$\times 2$				$\times 4$			
		ACC(%)	NED	PSNR	SSIM	ACC(%)	NED	PSNR	SSIM
TSRN [4]	STISR	28.54	0.4809	18.99	0.5233	23.16	0.4159	18.11	0.4850
TPGSR [22]	STISR	30.07	0.4913	18.83	0.5562	23.26	0.4123	18.07	0.4758
TBSRN [21]	STISR	31.81	0.5294	19.01	0.5366	25.27	0.4444	18.33	0.4826
TATT [24]	STISR	31.27	0.5240	19.06	0.5772	23.50	0.4342	17.96	0.4904
TextDiff [26]	STISR	31.82	0.5368	19.22	0.5779	24.47	0.4408	18.26	0.4929
DiffTSR [27]	STISR	32.16	0.5446	19.46	0.5887	25.56	0.4528	18.64	0.4968
SRResNet [46]	General SR	34.99	0.6996	20.72	0.7360	28.79	0.6361	20.22	0.7224
RRDB [43]	General SR	34.94	0.7003	21.10	**0.7535**	29.20	0.6421	20.23	0.7231
RCAN [16]	General SR	34.84	0.7006	20.98	0.7435	28.79	0.6321	20.33	0.7232
ELAN [44]	General SR	35.08	0.6922	21.16	0.7480	29.53	0.6404	20.39	**0.7299**
CFAT [14]	General SR	35.19	0.7012	21.09	0.7479	29.70	0.6436	20.41	0.7248
MambaIR [45]	General SR	35.34	0.7029	21.14	0.7472	30.01	0.6478	20.44	0.7254
Ours	SR	**35.67**	**0.7062**	**21.19**	0.7486	**30.70**	**0.6599**	**20.45**	0.7251

Table 2 Comparion of model complexity for $\times 4$ SR task. The output size is set to 640×640 for FLOPs calculation

Methods	RCAN	ELAN	CFAT	MambaIR	Ours
Params(M)	15.59	8.312	22.07	16.7	1.76
FLOPs(G)	407	284	587	439	62.8
ACC(%)	28.79	29.53	29.70	30.01	30.7
NED	0.6321	0.6404	0.6436	0.6478	0.6599

$$\begin{aligned} F_f &= FFT\left(P(Conv_{1\times1}(F))\right) \\ F_1 &= P^{-1}\left(IFFT(W \odot F_f)\right) \\ F_{out} &= Conv_{1\times1}\left(GELU(Conv_{3\times3}(F_1))\right) + F \end{aligned} \tag{11}$$

where P represents window partition, and P^{-1} represents the reverse operation of P, namely window paste. $\odot$ denotes element-wise multiply.

3.5 Loss function

In this paper, we adopt two types of loss including pixel loss L_{pixel} and edge-aware loss L_{edge}. L_{pixel} calculates the L_1 distance between SR image I_{SR} and HR image I_{HR}:

$$L_{pixel} = \|I_{HR} - I_{SR}\|_1 \tag{12}$$

While pixel loss can effectively force SR images to have high PSNR values, it fails to guarantee good visual perception quality. Incorporating high-level supervision during the learning process can effectively improve the perceptual quality of SR images [5]. Consequently, we additionally introduce the edge-aware loss L_{edge} which measures the difference between HR edge map M_{HR} and SR edge map M_{SR} from the perspective of the feature domain. Specifically, we employ the pre-trained VGG19 network [35] to extract edge features from M_{HR} and M_{SR}. In addition, VGG19 is also employed to extract image features from I_{HR} and I_{SR}. Subsequently, image features are element-wise multiplied with edge features to enhance the text regions. L_{edge} is formulated as follows:

$$L_{edge} = \|F(I_{HR}) \cdot F(M_{HR}) - F(I_{SR}) \cdot F(M_{SR})\|_1 \tag{13}$$

where $F(\cdot)$ is the pre-trained VGG19 network. $F(I_{HR})$ and $F(I_{SR})$ respectively represent the image features extracted from HR images and SR images. $F(M_{HR})$ and $F(M_{SR})$ respectively denote the edge features extracted from the HR Canny edge map and SR Canny edge map.

The overall loss can be calculated as follows:

$$L = L_{pixel} + \alpha \cdot L_{edge} \tag{14}$$

where α represents the weight of loss term. According to the ablation experimental results, we set $\alpha = 1$.

4 Experiments and analysis

4.1 Datasets and evaluation metrics

STISR Dataset. For the real-world Chinese-English STISR task, we train and test models on the Real-CE dataset [5], where text images are captured using iPhone 11 pro and iPhone 12 pro with varying focal lengths (13 mm, 26 mm and 52 mm). In the same shooting scene, image pairs captured by 13 mm and 26 mm, as well as those captured by 26 mm and 52 mm are utilized for $\times 2$ SR task. In contrast, the pairs captured by 13mm and 52mm are employed for $\times 4$ SR task. The training set contains 1,935 LR-HR image pairs, and the test set contains 783 image pairs, 522 of which are used for $\times 2$ SR task and the rest 261 pairs are used for $\times 4$ SR task. Additionally, Real-CE also provides annotations including text detection boxes and text labels, enabling the

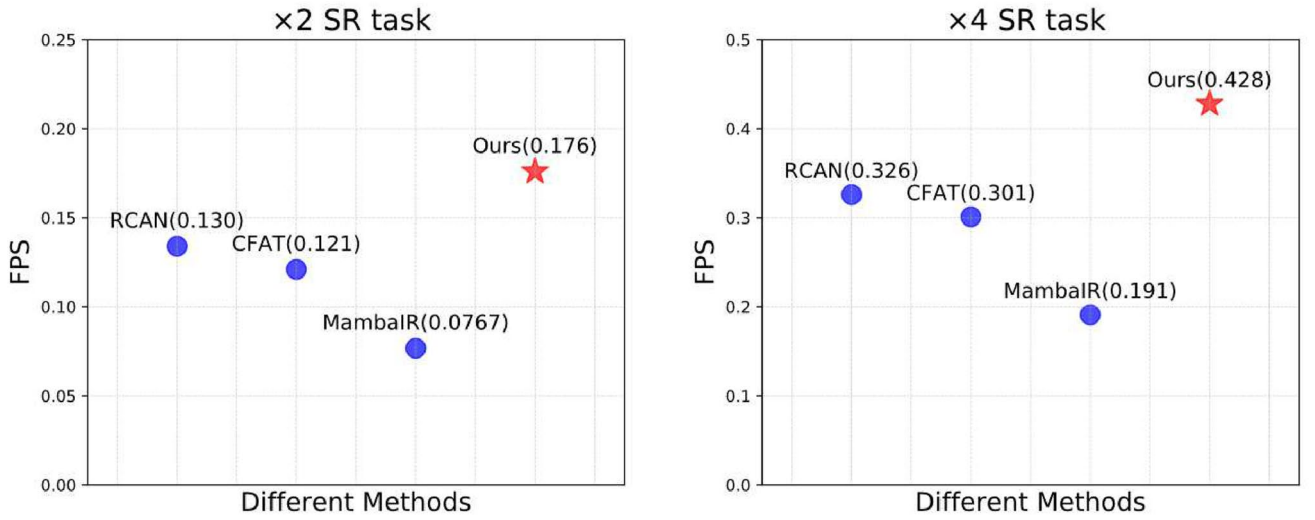

Fig. 9 Comparison of inference speed on Real-CE test set for both $\times 2$ and $\times 4$ SR tasks. The inference speed is measured in frames per second (FPS)

evaluation of STISR performance from the text recognition perspective.

General SR Datasets. To validate the effectiveness of our method for general SR task, we also train our model on DIV2K [36] and Flickr2K [37] datasets, and then evaluate on five common SR benchmarks, including Set5 [38], Set14 [39], BSD100 [40], Urban100 [41] and Manga109 [42].

Evaluation Metrics. For the STISR task, we adopt text recognition accuracy (ACC), normalized edit distance (NED), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as evaluation metrics. Notably, ACC and NED are calculated by CRNN [23]. For the general SR task, we follow previous works and adopt PSNR and SSIM as evaluation metrics.

4.2 Implementation details

We adopt Pytorch 2.0.1 as the experimental framework and conduct experiments on an Nvidia RTX A6000 with 48GB of memory. The model is trained and evaluated on the benchmark dataset Real-CE [5]. During the training phase, LR images are cropped into 64×64 patches to serve as the input. In contrast, the entire LR images are taken as input during testing. The training batch size is configured as 32 and the learning rate is set to 0.0002. We employ the Adam optimizer with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. The model is trained for 500,000 iterations in total. The number of channels is configured as $C = 64$. In deep feature extraction stage, we stack 6 residual groups (RGs), with each RG containing 2 pairs of DWAB and SWAB. The window size of DWA is set to 8×8 and the number of heads in self-attention (SA) mechanism is configured as 4. In the SCI module, the channel compression coefficients are defined as: $\gamma_1 = 16$, $\gamma_2 = 8$. For FFN-AFE, the channel expansion coefficient γ is set to 2.

4.3 Comparisons with state-of-the-arts

4.3.1 Evaluations on STISR dataset

In this section, we compare our method with other advanced SR methods on Real-CE dataset in terms of quantitative indicators, visual results and model complexity for both $\times 2$ and $\times 4$ SR task. We choose classic general SR methods including SRResNet [9], RRDB [43], RCAN [16], ELAN [44], CFAT [14], MambaIR [45] and state-of-the-art STISR methods such as TSRN [4], TPGSR [22], TBSRN [21], TATT [24], TextDiff [26], DiffTSR [27]. For a fair comparison, we modify their implementation to handle $\times 2$ and $\times 4$ SR tasks. All the above methods are trained and evaluated on Real-CE dataset under the same experimental settings, including learning rates and the number of training epochs. For model architectures and parameter settings, we follow the officially released implementations specified in the original papers.

Quantitative Metrics. Table 1 compares the quantitative indicators of various SR methods on Real-CE dataset. For $\times 2$ SR task, our method achieves the highest value in PSNR, text recognition accuracy(ACC) and normalized edit distance(NED) while reaches the sub-optimal value in SSIM, which is 0.0049 lower than the first place. As the scale factor increases, more details need to be reconstructed, rendering $\times 4$ SR task more challenging. For $\times 4$ SR task, our method outperforms other existing methods in terms of ACC, NED and PSNR. Notably, the ACC and NED of our method exceed the sub-optimal values by 0.69% and 0.0121

Fig. 10 Visual comparison with various methods for $\times 2$ SR task on Real-CE dataset. To highlight the contrast effect, a yellow rectangular frame is employed to mark the local area on each image and the marked area is magnified for better visibility. The text below each image represents the recognition result by CRNN, with the incorrectly recognized characters highlighted in red

Fig. 11 Visual comparison with various methods for $\times 4$ SR task on Real-CE dataset. To highlight the contrast effect, a yellow rectangular frame is employed to mark the local area on each image and the marked area is magnified for better visibility. The text below each image represents the recognition result by CRNN, with the incorrectly recognized characters highlighted in red

respectively. The results show that general SR methods perform significantly better than STISR methods on Real-CE dataset. The reason is that general SR methods take the whole images as input, while the STISR methods are designed for cropped text images only containing text lines, making it difficult to adapt to the Real-CE dataset. In conclusion, the proposed method shows great improvements across all indicators for both $\times 2$ and $\times 4$ SR tasks, especially in terms of ACC and NED, which focus on measuring the reconstruction quality of text regions. This confirms that our method can effectively reconstruct the semantic information of the text region, thus enhancing the performance of downstream text recognition tasks.

Model Complexity. Table 2 presents a comparison of model parameters and FLOPs among various classic SR methods for the $\times 4$ task, with the output size set to 640×640 for FLOPs calculation. Compared to previous methods, our DA^2FE achieves a stepwise decrease in model complexity, but demonstrates the optimal performance on the Real-CE dataset. Figure 9 presents a comparative analysis of inference speed (measured in frames per second, FPS) among our DA^2FE, RCAN [16], CFAT [14] and MambaIR [45]. As depicted in Fig. 9, our method achieves significantly higher FPS compared to competing methods for both $\times 2$ and $\times 4$ SR tasks. Notably, the $\times 4$ SR task exhibits faster inference speeds than the $\times 2$ task. This performance gap occurs because all input images for $\times 4$ task are captured with a focal length of 13mm, whereas the input for $\times 2$ task includes images captured at both 13mm and 26mm. Since longer focal lengths yield higher resolution images, thereby increasing computational cost and reducing inference speed.

Visual Results. Figure 10 and Fig. 11 compare the visual results of our method with a Transformer-based method CFAT [14] and a Mamba-based method MambaIR [45] on the Real-CE test set for $\times 2$ and $\times 4$ SR tasks. To highlight the contrast effect, a yellow rectangular frame is employed to mark the local area on each image and the marked area is magnified for better visibility. Addtionally, CRNN [23] is

Table 3 PSNR/SSIM metrics on general SR datasets for $\times 4$ task

Methods	Types	Params	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN [16]	Classic SR	15.59M	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
ELAN [44]	Classic SR	8.312M	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
SwinIR [15]	Classic SR	11.90M	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
SRformer [47]	Classic SR	10.40M	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
MambaIR [45]	Classic SR	16.7M	33.03	0.9046	29.20	0.7961	27.98	0.7503	27.68	0.8287	32.32	0.9272
SwinIR-light [15]	Light SR	0.93M	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN-light [44]	Light SR	0.64M	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SRformer-light [47]	Light SR	0.873M	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
SMFANet [48]	Light SR	0.496M	32.51	0.8985	28.87	0.7872	27.74	0.7412	26.56	0.7976	31.29	0.9163
Ours	Light SR	1.76M	32.82	0.9036	29.04	0.7928	27.89	0.7484	27.34	0.8236	31.97	0.9248

applied to recognize the text within the yellow box, with the incorrectly recognized characters marked in red. The results indicate that the text areas reconstructed by CFAT and MambaIR suffer from varying degrees of distortion, making it difficult to accurately reconstruct characters with complex stroke structures. For instance, in the first row of Fig. 10, both the outputs of CFAT and MambaIR are severely blurry when attempting to restore the character '惰', making it to be mistakenly recognized as '情' by CRNN. Similarly, in the second row of Fig. 11, CFAT and MambaIR reconstruct the character '含' into the closely related characters '合' and '台' respectively. In contrast, the proposed method can avoid the above errors and effectively improve the clarity of multi-line Chinese text, thereby improving the text recognition accuracy.

4.3.2 Evaluations on general SR datasets

Our DA^2FE is mainly designed for real-world Chinese-English text images. To verify its effectiveness on natural images, we compared it with five classic SR methods, including RCAN [16], ELAN [44], SwinIR [15], SRformer [47] and MambaIR [45], as well as four light SR methods, namely SwinIR-light [15], ELAN-light [44], SRformer-light [47] and SMFANet [48] on general SR datasets. Our DA^2FE maintains a parameter count comparable to that of lightweight SR methods while achieving significantly superior SR performance. Conversely, although its performance metrics are slightly lower than those of classic SR methods, it requires substantially fewer parameters. In conclusion, for the general SR task, our method strikes a favorable balance between model complexity and SR performance.

4.4 Ablation study

In ablation studies, we train and evaluate models on Real-CE dataset for $\times 2$ STISR task.

4.4.1 Effect of DWA and SWA

Interval size in SWA. Sparse window attention (SWA) performs sparse sampling on the feature map at a specific interval size (denoted by I) to form a set of windows, and then conducts self-attention calculation separately within each window. When $I = 1$, SWA will be transferred to standard self-attention. Generally, a smaller value of I means a wider receptive field but incurs a higher computational cost. As shown in Table 4, we assign a group of values {8, 6, 4, 2} for I to observe the changes in SR performance and model complexity. As the sampling interval (I) decreases, FLOPs gradually increase, while the number of model parameters remains unchanged. When I decreases from 8 to 4, various

Table 4 Impact of different interval sizes in SWA. We compare the model complexity for ×2 SR task, and the output size is set to 320 × 320 for FLOPs calculation

Interval size (I)	Params(M)	FLOPs(G)	ACC(%)	NED	PSNR	SSIM
8	1.25	33.79	35.05	0.7042	20.83	0.7375
6	1.25	35.20	35.19	0.7025	20.84	0.7425
4	1.25	38.22	**35.43**	0.7055	**20.99**	**0.7435**
2	1.25	55.94	35.42	**0.7057**	20.98	0.7431

Table 5 Impacts of DWA and SWA

DWA	SWA	ACC(%)	NED	PSNR	SSIM
✓		34.39	0.6978	20.36	0.7401
	✓	34.42	0.6989	20.46	0.7418
✓	✓	**35.43**	**0.7055**	**20.99**	**0.7435**

SR performance metrics improve significantly. The reason is that, with a smaller interval size for sparse sampling, each token can interact with a greater number of tokens, which can more effectively extract global dependencies. When I further reduces from 4 to 2, although NED increase slightly, ACC, PSNR and SSIM decline. Considering the trade-off between SR performance and computational complexity, I is set to 4.

DWA and SWA. Table 5 demonstrates the necessity for simultaneous usage of dense window attention(DWA) and sparse window attention (SWA). We employ three different experimental settings, respectively incorporating 4×DWA, 4×SWA and 2 pairs of alternating DWA and SWA in each RG. As we can see, the model using SWA alone is better than using DWA alone across all indicators. The reason is that although DWA can effectively reduce the computational complexity of standard self-attention, it falls short in learning global features. In contrast, SWA exhibits a strong ability to extract global information, which compensates for the limitations of DWA to a certain degree. The alternating usage of DWA and SWA can provide interactions for tokens from not only dense regions but also sparse regions of an image to obtain a wider receptive field, thus achieving the optimal performance across various evaluation metrics. The results validates that both local context and global sparse interaction contribute to improving SR performance.

4.4.2 Effect of other propose modules

In this session, we will incorporate other modules one by one, with the experimental results shown in Table 6. Specifically, model-A indicates the model that only alternately employs DWA and SWA, model-B is obtained by integrating the SCI module to model-A, model-C further substitutes FFN with FFN-AFE, and model-D is equipped with the FF module to form the complete model.

When SCI, FFN-AFE and FF are progressively integrated, both the model parameters and FLOPs increase slightly, while SR performance improves accordingly. Compared to model-A, model-B improves ACC by 0.03% and PSNR by 0.02dB. As shown in Fig. 12, there are some artifacts on the two characters '请听' reconstructed by model-A. In contrast, model-B can effectively reduce artifacts and improve clarity of SR images. Model-C builds upon model-B by replacing FFN with FFN-AFE, resulting in further improvements of 0.13dB in PSNR and 0.15% in ACC. As depicted in Fig. 13, when reconstructing the two characters '障碍'

Table 6 Impacts of SCI, FFN-AFE and FF. We compare the model complexity for ×2 SR task, and the output size is set to 320 × 320 for FLOPs calculation

Model	DWA+SWA	SCI	FFN-AFE	FF	Params(M)	FLOPs(G)	ACC(%)	NED	PSNR	SSIM
model-A	✓				1.25	38.22	35.43	0.7055	20.99	0.7435
model-B	✓	✓			1.31	38.94	35.46	0.7058	21.01	0.7436
model-C	✓	✓	✓		1.55	45.39	35.61	0.7060	21.14	0.7452
model-D	✓	✓	✓	✓	1.61	46.99	**35.67**	**0.7062**	**21.19**	**0.7486**

Fig. 12 Effects of SCI

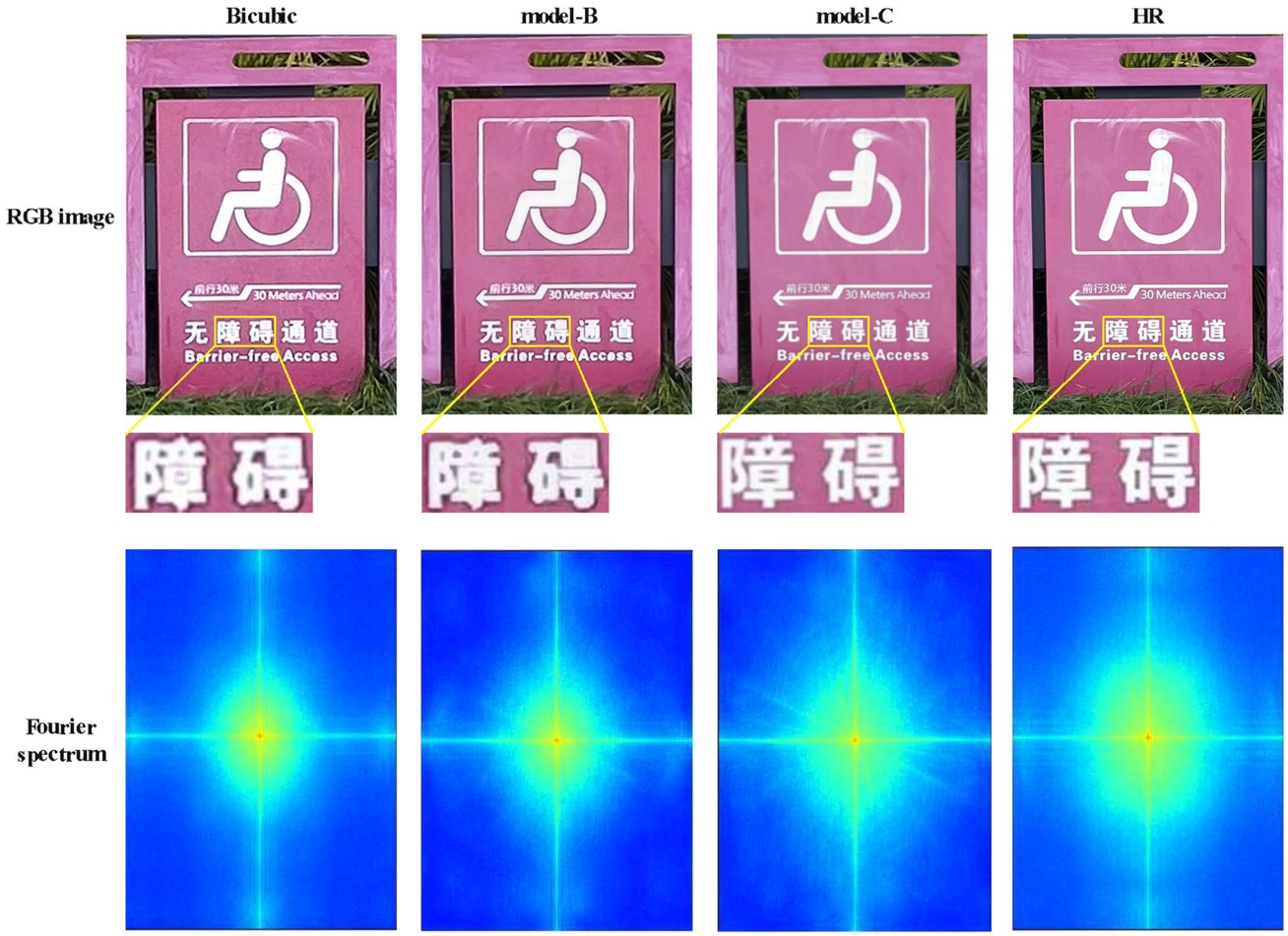

Fig. 13 Effects of FFN-AFE. The Fourier spectrum demonstrates that model-C with FFN-AFE can activate more high-frequency components which correspond to character edge details

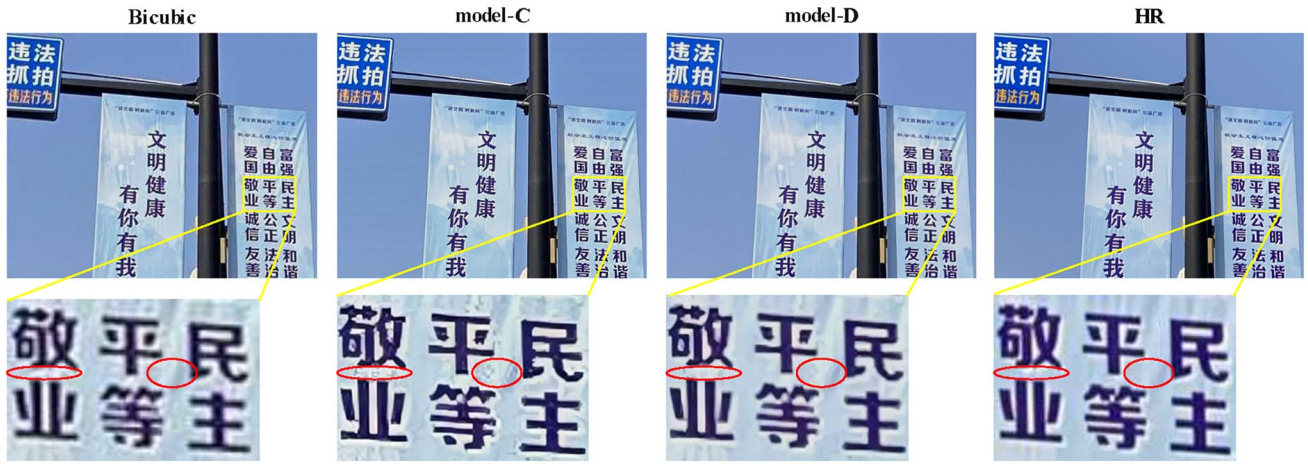

Fig. 14 Effects of FF

with numerous strokes and complex structure, many strokes in the SR output of model-B are stuck together. Conversely, model-C can restore distinct stroke outlines. Additionally, Fig. 13 displays the Fourier spectrum, demonstrating that FFN-AFE can activate more high-frequency components, thus effectively restoring character edge details. Model-D builds upon model-C by integrating the FF module, demonstrating the optimal performance across all metrics. As

Table 7 Effects of different loss balance coefficients

α	ACC(%)	NED	PSNR	SSIM
0	34.98	0.6913	20.89	0.7398
0.1	35.12	0.7021	21.10	0.7430
0.5	35.28	0.7034	21.12	0.7463
1	**35.67**	**0.7062**	21.19	**0.7486**
5	35.57	0.7057	**21.21**	0.7483

indicated by the red circle in Fig. 14, the SR image generated by model-C is interfered by background noise. In contrast, model-D can effectively suppress the background noise. This verifies that fusing features from different layers can generate more comprehensive and diverse features for the final reconstruction step, thus further enhancing the visual quality of SR images. In conclusion, it is reasonable and acceptable to trade off an moderate increase in model complexity for an improvement in SR performance.

4.4.3 Loss balance

In this session, we conduct an ablation study for the selection of loss balance coefficient α in Eq. (14). As presented in Table 7, we set a group of α from 0 to 5 to observe the performance change. The results reveal that all evaluation indicators are at the lowest values when edge-aware loss is not incorporated ($\alpha = 0$). When α is set to 0.01, all indicators are significantly improved, which demonstrates that edge-aware loss plays an significant role in our task. As α gradually increases from 0.01 to 1, various indicators also show an upward trend. However, when α is increased to 5. SSIM, ACC and NED all decline. Considering all indicators comprehensively, α is set to 1.

5 Discussion

The proposed DA^2FE method achieves state-of-the-art performance across multiple evaluation criteria, demonstrating significant improvements in both text reconstruction quality (measured by ACC and NED) and image fidelity (evaluated via PSNR and SSIM). Besides, DA^2FE demonstrates robust performance on natural image datasets, highlighting its generalizability beyond text-specific applications. Nevertheless, DA^2FE demonstrates limited effectiveness when processing severely degraded Chinese characters exhibiting high inter-class similarity. To address this limitation, our future work will focus on integrating a text recognition network to to extract semantic context from input images, which will be incorporated into the SR reconstruction pipeline to improve the restoration quality of challenging text instances. Furthermore, in subsequent research, we plan to explore lightweight diffusion model implementations (e.g., latent diffusion architectures) to achieve high-quality image super-resolution at an acceptable computational cost.

6 Conclusion

In this paper, we propose a Dual Attention Transformer with Adaptive Frequency Enhancement (DA^2FE) network, which is applied to the real-world Chinese-English STISR task. The main part of our network alternates between two complementary window attention mechanisms which utilize different window partition strategies. Specifically, dense window attention (DWA) samples a fixed number of neighbouring tokens on the feature map to form a set of windows, and then calculates self-attention within each window, facilitating the learning of local features. In contrast, sparse window attention (SWA) samples tokens at a certain interval, thereby establishing global dependencies. Additionally, spatial-channel interaction (SCI) module performs bi-directional interaction between the window attention branch and depth-wise convolution (DWConv) branch across the channel and spatial dimensions. To enhance crucial frequency information, we design a feed-forward network with adaptive frequency enhancement (FFN-AFE), which adaptively determine the importance of each frequency by introducing a learnable quantization matrix in the frequency domain. Finally, the outputs of each residual group (RG) are fused together to provide comprehensive and rich information for SR reconstruction. Extensive experiments on Real-CE dataset demonstrate that DA^2FE outperforms previous methods in both objective evaluation indicators and subjective visual results. Moreover, DA^2FE achieves excellent results on general SR datasets, validating its extensive applicability.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant (62170110, 62476088), and the Shanghai Automotive Industry Science and Technology Development Foundation under Grant 2304.

Author contributions Yanbin Liu and Qin Shi designed the algorithm, conducted experiments and wrote the original paper. Ziming Zhu, Xiaofeng Ling and Yu Zhu contributed to refining the experiments and reviewing the paper.

Data availability The data supporting the findings of this study are openly available to the public.

Declarations

Conflicts of interest The authors declare no competing interests.

Ethical and informed consent This work does not involve experimental procedures with human subjects or animals.

References

1. Dong, C., Zhu, X., Deng, Y., et al.: Boosting optical character recognition: a super-resolution approach. arXiv preprint arXiv:1506.02211 (2015)
2. Wang, W., Xie, E., Sun, P., et al.: Textsr: content-aware text super-resolution guided by recognition. arXiv preprint arXiv:1909.07113 (2019)
3. Mou, Y., Tan, L., Yang, H., et al.: Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, pp 158–174, (2020)
4. Wang, W., Xie, E., Liu, X., et al.: Scene text image super-resolution in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, Springer, pp 650–666, (2020)
5. Ma, J., Liang, Z., Xiang, W., et al.: A benchmark for chinese-english scene text image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 19452–19461 (2023)
6. Dong, C., Loy, C.C., He, K., et al.: 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, pp 184–199, (2014)
7. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, pp 391–407, (2016)
8. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
9. Zhou, C., Li, Q., Li, C., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419 (2023)
10. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, (2017)
11. Li, A., Zhang, L., Liu, Y., et al.: Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12514–12524, (2023)
12. Chen, Z., Zhang, Y., Gu, J., et al.: Dual aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12312–12321, (2023)
13. Liu, Y., Zhang, Y., Wang, Y., et al.: A survey of visual transformers. IEEE Trans. Neural Netw. Learn. Syst. (2023)
14. Ray, A., Kumar, G., Kolekar, M.H.: Cfat: Unleashing triangular windows for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 26120–26129, (2024)
15. Liang, J., Cao, J., Sun, G., et al.: Swinir: image restoration using Swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1833–1844, (2021)
16. Zhang, Y., Li, K., Li, K., et al.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301, (2018)
17. Moorthy, S., Sachin Sakthi, K.S., Arthanari, S., et al.: Hybrid multi-attention transformer for robust video object detection. Eng. Appl. Artif. Intell. **139**, 109606 (2025)
18. Kuppusami Sakthivel, S.S., Moorthy, S., Arthanari, S., et al.: Learning a context-aware environmental residual correlation filter via deep convolution features for visual object tracking. Mathematics **12**(14), 2279 (2024)
19. Arthanari, S., Moorthy, S., Jeong, J.H., et al.: Adaptive spatially regularized target attribute-aware background suppressed deep correlation filter for object tracking. Signal Process. Image Commun. **136**, 117305 (2025)
20. Zhao, C., Feng, S., Zhao, B.N., et al.: Scene text image super-resolution via parallelly contextual attention network. In: Proceedings of the 29th ACM International Conference on Multimedia, pp 2908–2917, (2021)
21. Chen, J., Li, B., Xue, X.: Scene text telescope: text-focused scene image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12026–12035, (2021)
22. Ma, J., Guo, S., Zhang, L.: Text prior guided scene text image super-resolution. IEEE Trans. Image Process. **32**, 1341–1353 (2023)
23. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
24. Ma, J., Liang, Z., Zhang, L.: A text attention network for spatial deformation robust scene text image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5911–5920, (2022)
25. Li, X., Zuo, W., Loy, C.C.: Learning generative structure prior for blind text image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10103–10113, (2023)
26. Liu, B., Yang, Z., Wang, P., et al.: Textdiff: mask-guided residual diffusion models for scene text image super-resolution. arXiv preprint arXiv:2308.06743 (2023)
27. Zhang, Y., Zhang, J., Li, H., et al.: Diffusion-based blind text image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 25827–25836, (2024)
28. Shi, B., Yang, M., Wang, X., et al.: Aster: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2035–2048 (2018)
29. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural Inf. Process. Syst. **28** (2015)
30. Fang, S., Xie, H., Wang, Y., et al.: Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7098–7107, (2021)
31. Yu, H., Wang, X., Li, B., et al.: Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11943–11952, (2023)
32. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR, pp. 8748–8763 (2021)
33. Huang, J., Peng, D., Li, H., et al.: Segctc: offline handwritten Chinese text recognition via better fusion between explicit and implicit segmentation. In: International Conference on Document Analysis and Recognition, Springer, pp. 332–349, (2023)
34. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 679–698 (1986)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Timofte, R., Agustsson, E., Van Gool, L., et al.: 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, (2017)
37. Lim, B., Son, S., Kim, H., et al.: 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 (2017)
38. Bevilacqua, M., Roumy, A., Guillemot, C., et al.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
39. 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, Springer, pp 711–730 (2012)
40. Martin, D., Fowlkes, C., Tal, D., et al.: 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, IEEE, pp. 416–423, (2001)
41. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206, (2015)
42. Matsui, Y., Ito, K., Aramaki, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. **76**, 21811–21838 (2017)
43. Wang, X., Yu, K., Wu, S., et al.: Esrgan: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp 0–0, (2018)
44. Zhang, X., Zeng, H., Guo, S., et al.: Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision, Springer, pp 649–667, (2022)
45. Guo, H., Li, J., Dai, T., et al.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision, Springer, pp 222–241, (2025)
46. Ledig, C., Theis, L., Huszár, F., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4681–4690, (2017)
47. Zhou, Y., Li, Z., Guo, C.L., et al.: Srformer: Permuted self-attention for single image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12780–12791 (2023)
48. Zheng, M., Sun, L., Dong, J., et al.: Smfanet: a lightweight self-modulation feature aggregation network for efficient image super-resolution. In: European Conference on Computer Vision, Springer, pp 359–375, (2024)