Network Modeling Analysis in Health Informatics and Bioinformatics (2024) 13:34
https://doi.org/10.1007/s13721-024-00469-y

ORIGINAL ARTICLE

Medical image super-resolution via transformer-based hierarchical encoder–decoder network

Jianhao Sun[1] · Xiangqin Zeng[2] · Xiang Lei[3] · Mingliang Gao[1] · Qilei Li[4] · Housheng Zhang[1] · Fengli Ba[1]

Received: 11 March 2024 / Revised: 22 May 2024 / Accepted: 29 May 2024

Abstract

Medical image super-resolution (SR) has emerged as an effective means to enhance the resolution of medical images. Nevertheless, many existing methods still face the issue of insufficient representation of high-frequency features. To address this problem, we propose a Transformer-based hierarchical Encoder–Decoder Network (THEDNet). The THEDNet incorporates an advanced transformer to extract features at various hierarchical dimensions. Specifically, the Encoder and Decoder units are equipped with an Efficient Multi-scale Attention (EMA) module for capturing long-range interdependencies among features. By leveraging an enhanced transformer architecture, THEDNet can capture long-range feature interdependencies and create the final high-resolution images. Experiments are conducted on two medical CT image datasets, and comparative results verify the effectiveness of the proposed THEDNet.

Keywords Super-resolution(SR) · Integration of attention mechanism · Transformer

1 Introduction

Medical images hold immense significance across multiple domains including radiology, oncology, and cardiology. They can offer indispensable diagnostic information for medical professionals. Notably, CT images provide valuable insights into the precise localization of lesions. However, the inherent limitations of the hardware equipment often result in lower image resolutions, which makes it difficult to visualize finer details. This limitation hampers the ability of doctors to assess and diagnose medical conditions accurately. The super-resolution (SR) techniques can increase the resolution of medical images and thus provide more detailed information.

Currently, the mainstream SR methods include interpolation algorithms Khaledyan et al. (2020), sparse representation algorithms Ayas and Ekinci (2020), and deep learning-based algorithms Wang et al. (2020). The image SR method based on deep learning is the most popular method in the domain. This method usually reconstructs an image using a Convolutional Neural Network (CNN) or Generative Adversarial Network (GAN). It can learn the complex structure and features of an image. Dong et al. (2015) proposed the Super-Resolution Convolutional Neural Network (SRCNN) and adopted the deep learning model for image super-resolution (SR). On this basis, Dong et al. (2016) proposed an augmented iterative model called Fast Super-Resolution Convolutional Neural Networks (FSRCNN). The FSRCNN integrated non-convolutional layers and employed smaller convolutional kernels and additional mapping layers. Ledig et al. Ledig et al. (2017) proposed Super-Resolution Using a Generative Adversarial Network (SRGAN). It leveraged perceptual and adversarial loss mechanisms to generate highly realistic HR images. Lim et al. (2017) employed residual networks from the SRGAN model and proposed enhanced deep residual networks for single image super-resolution. Residual networks and attention mechanisms have been used for image SR recently. Zhang et al. (2018)

Jianhao Sun and Xiangqin Zeng contributed equally to this work.

✉ Mingliang Gao
mlgao@sdut.edu.cn

Fengli Ba
sdut_flba@163.com

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China

2 Zibo Central Hospital, Zibo 255020, China

3 Zhiyang Innovation Co., Ltd., Jinan 250101, China

4 School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

Published online: 24 June 2024

proposed the Residual Dense Network (RDN)by extracting a large amount of feature information from low-resolution (LR) images. Li et al. (2019) proposed a super-resolution network with a feedback mechanism called Feedback Network for Image Super-Resolution (SRFBN). The close connection between the upsampling and downsampling layers enhanced high-dimension information. Chen et al. (2020) proposed a feedback-adaptive weighted dense network to extract more useful high-dimension features in medical images. Li Li et al. (2019) proposed a Gated Multiple Feedback Network (GMFN) by utilizing multi-scale information and augmentation with the attention mechanism.

Dosovitskiy et al. (2021) employed the Transformer model in image SR. The Transformer model incorporated an attention mechanism to capture global information in images. To improve the resolution and quality of the images more effectively, Lei et al. (2021) proposed the Transformer-based Enhancement Network (TransENet). Nevertheless, the lack of high-frequency features and long-term feature interdependencies affect the quality of the generated images. To address this problem, we introduce the THEDNet. The THEDNet is designed in a transformer schema with a series of hierarchical Encode–Decoders. Meanwhile, an external attention (EA) Guo et al. (2021) module and efficient multi-scale attention (EMA) Ouyang et al. (2023) module are built into each encoder and decoder module. Enhancing high-dimensional feature representation entails integrating external attention modules, hierarchical feature extraction, and cross-attention fusion. The EA module enhances feature extraction by leveraging external memory units. It refines the attention map to capture essential high-dimensional features more accurately. Hierarchical feature extraction is facilitated by attention-fused encoders and decoders that process features at different levels. It can ensure the effective utilization of both low- and high-dimensional features. In the decoding process, the cross-attention mechanism integrates encoded low-dimensional features with high-dimensional representations. This improves the representation of high-dimensional features and thus enhances image detail recovery. The contributions are summarized as follows:

1. We proposed a Transformer-based hierarchical Encoder–Decoder Network for medical image SR.
2. We designed an attention-fused encoder–decoder framework to improve feature representation capability and exploit the long-term feature interdependencies.
3. We carried out experiments on two publicly accessible medical image datasets and the results verified the efficiency of the proposed method.

2 Related work

2.1 CNN-based image super-resolution

CNNs have held a pivotal role in SR tasks, benefitting by their multifaceted capabilities. As a pioneering work, SRCNN Dong et al. (2015) used a sparse coding formula to map patches from LR to HR. Dong et al. (2016) proposed the FSRCNN by optimizing the SR network structure and hyperparameter settings. Li et al. (2019) built the Super-Resolution Feedback Network (SRFBN). They reconstructed high-resolution images by enriching the correspondence between low and high-dimension representations. CNN has also played a crucial role in medical images SR. The medical image SR can be divided into two types, *i.e.,* 2D medical image SR and 3D medical image SR. Gu et al. (2020) focused on specific areas through different information channels in 2D medical CT images. Georgescu et al. (2020) proposed an end-to-end CNN model for 3D CT or MRI scan images SR. Ran et al. (2023) introduced a general CNN fusion framework termed GuidedNet. This framework leveraged the multi-scale information of high-resolution guidance images to fuse low-resolution images. Shang et al. (2024) combined CNNs with diffusion probabilistic models to predict low-frequency information and employed DPM to generate high-frequency details.

2.2 Transformer-based image super-resolution

Transformer architecture has been gradually applied in image SR. Yang et al. (2020) proposed the texture transformer network for Image Super-Resolution. It utilized the transformer to learn the features between the LR and reference images. Subsequently, Yoo et al. (2022) introduced a model that combines CNN and transformer through an aggregation approach. Liu et al. (2021) built the shifted window transformer model to address the challenges posed by hierarchical feature maps and shifted window attention. Lei et al. (2021) introduced Transformer-based multistage Enhancement (TransENet) for image SR. It can exploit the synergy between different feature dimensions. Yan et al. (2021) presented a transformer-based MRI image SR model based on amalgamating features from multiple layers. Han et al. (2023) combined the CNN and swin transformer Liu et al. (2021) for medical image SR.

2.3 Attention-based super-resolution

The attention mechanism intensified the focus on important areas to enhance the performance of the SR model. Zhang et al. (2018) introduced channel attention to applying distinct

weights to different channels for feature extraction. Shang et al. (2022) presented channel-space attention to enable information exchange across various channel dimensions. Du et al.utilized the convolutional block attention module et al. Wooet al. (2018) and self-augmented attention module to incorporate more multi-scale and layer features for generating HR images. Mei et al. (2020) constructed a cross-scale non-local attention module to unveil numerous cross-scale feature correlations in a single LR image. Lu et al. (2020) introduced a local residual dense attention module for medical image super-resolution models. This module enhanced the network's learning ability while simplifying the training. Mehri et al. (2021) established a two-fold attention module to enhance the extraction of the channel and spatial attention mechanism information in adaptive residual blocks.

3 Methodology

3.1 Overview of THEDNet

The architecture of the proposed THEDNet is illustrated in Fig. 1.

It consists of four modules, *i.e.*, , feature extraction module, attention fusion encoder module, attention fusion decoder module, and upsampling layer. AF-Encoder$_n$ denotes the n−*th* attention fusion encoder, and AF-Decoder$_n$ denotes the n−*th* attention fusion decoder. THEDNet incorporates a four-layer encoder and a single-layer decoder. Initially, a convolutional layer converts the LR image I_{LR} into the feature space. The result will be directed to the feature extraction module. This module aims at capturing high-frequency details from distinct regions in medical images. Specifically, we assign three feature extraction modules for the low-dimensional stage.

$$f_n = \mathrm{FE}_n(f_{n-1}) = \mathrm{FE}_n(\mathrm{FE}_{n-1}(\ldots \mathrm{FE}_1(f_0)\ldots)), \quad (1)$$

where FE$_n$ denotes the n−*th* feature extraction module.

The residual block constitutes the primary element of the feature extraction module. The structure of the feature extraction module is illustrated in Fig. 2.

The input to the feature extraction module includes a series of feature maps generated by the convolutional layer, which encodes low-resolution (LR) image information. The output consists of feature maps that have been processed to contain more abstract and higher-level information. After extracting low-dimensional features, we use a sub-pixel layer to transform these features into a higher-dimensional space. Then, we use them as input for the improved transformer model. Three AF-Encoders are adopted to handle low-dimensional features. Meanwhile, An AF-Encoder

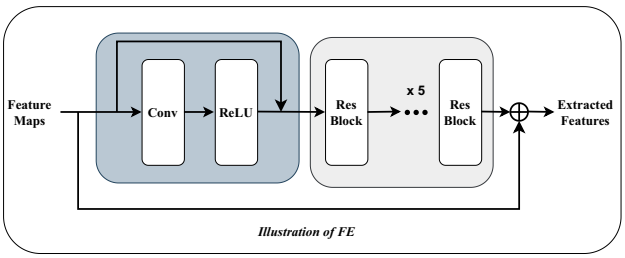

Fig. 2 Illustration of feature extraction module

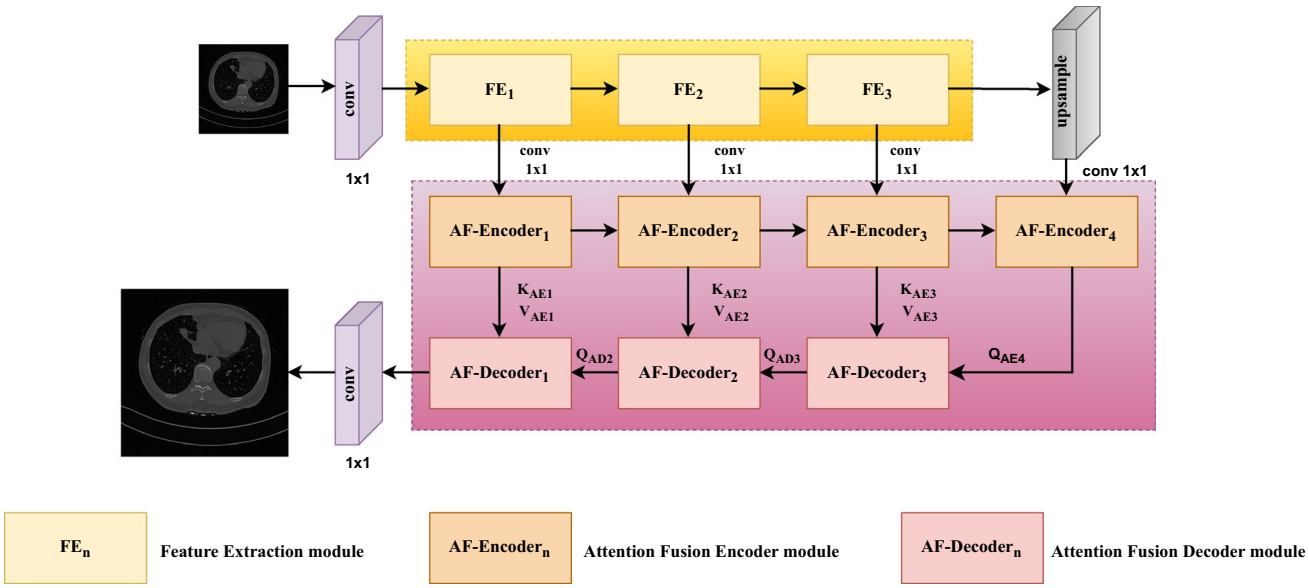

Fig. 1 Flowchart of the proposed THEDNet

is utilized to extract HR features by a 1×1 convolution. Subsequently, the features are fed into three AF-Decoders for enhancement. The AF-Decoder integrates multi-scale high-dimensional and low-dimensional features. It utilizes the encoded low-dimensional features as the keys (K) and values (V) inputs and performs cross-attention computation with high-dimensional features (used as the queries, Q). The fusion of low-dimensional features with high-dimensional representations occurs at various stages. This fusion enables the high-dimensional features to exploit the information from low-dimensional features effectively. Consequently, this integration leads to enhanced image detail recovery. Finally, after the feature enhancement stage, a single convolutional layer is adopted to produce the SR image I_{SR}.

3.2 Attention fusion encoder–decoder architecture

We emphasize the integration of attention modules within the network. This integration is aimed at substantially enhancing the feature extraction capability and expressiveness. As shown in Fig. 1, the low-latitude feature encoder module integrates three AF-Encoders, while a singular AF-Encoder constitutes the high-latitude counterpart. We employ a combination of four AF-Encoders and three AF-Decoders. Furthermore, we effectively capture long-range dependencies using the self-attention mechanism. As depicted in Fig. 3, AF-Encoder$_3$ and AF-Decoder$_3$ are employed as examples to illustrate their specific interactions. It provides a detailed account of the systematic transformation from input features to output data.

AF-Encoder: We transform the features of 3D images into 1D feature sequences. The initial step involves creating a 3D feature map. It is expressed as $f \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the map's height, width, and channel dimensions, respectively. Subsequently, the feature map f is partitioned into patches of size $P_H \times P_W$, where P_H and P_W denote the height and width of the patches, respectively. Each patch is transformed into a vector, $f_{pi} \in \mathbb{R}^{P_H \times P_W \times C}$, and arranged into a sequential series:

$$f_{\text{seq}} = \{f_{pi}\}_{i=1}^{N}, \tag{2}$$

$$N = \left(\frac{H \times W}{P_H \times P_W} \right), \tag{3}$$

where N represents the total number of patches. The resulting sequence f_{seq} is introduced as the input to the model. Each module does not incorporate positional embeddings. The network input can be represented as:

$$m_0 = [f_{p_1} W, f_{p_2} W, \ldots, f_{p_2} W], \tag{4}$$

where $W \in \mathbb{R}^{(P_H \times P_W \times C) \times D}$ represents the linear projection matrix, and D represents the dimension.

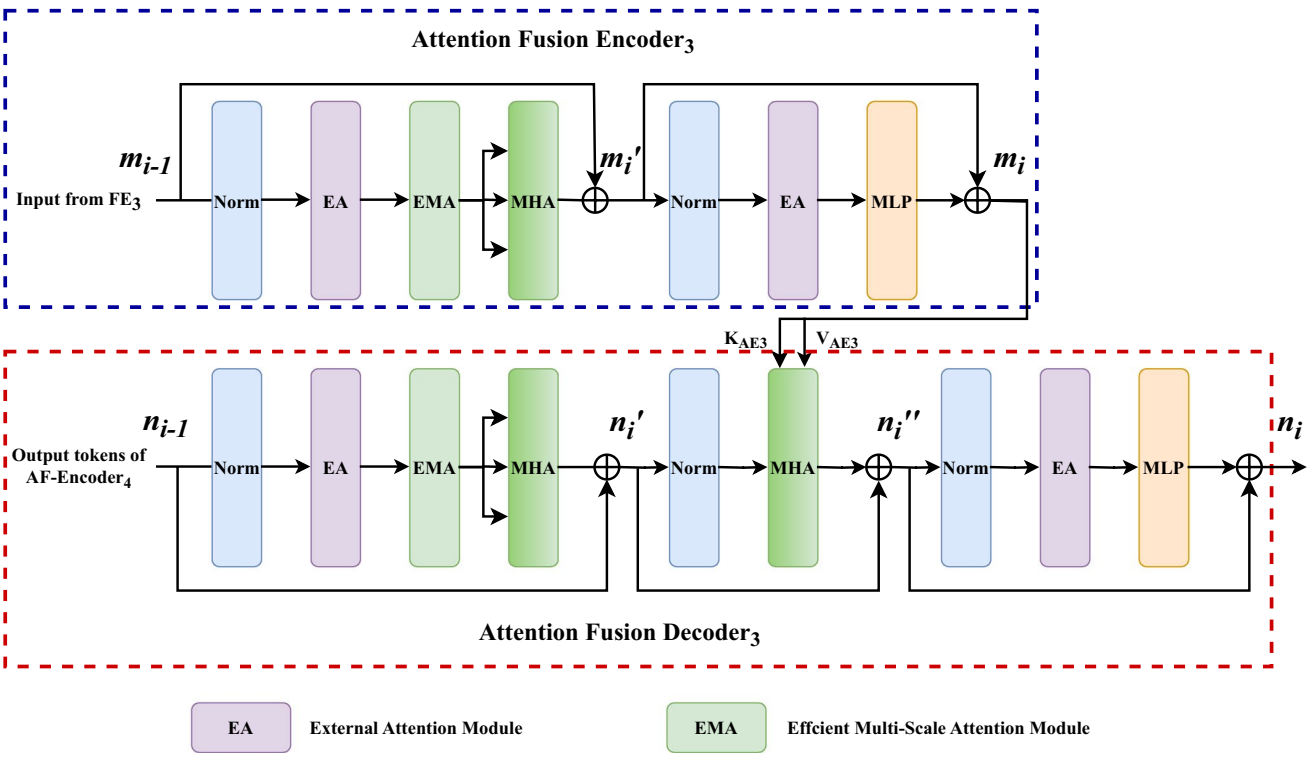

Fig. 3 Illustration of attention fusion encoder–decoder module

The architecture of the encoder expands upon the original Transformer structure Vaswani et al. (2017). It incorporates External Attention (EA) via Two Linear Layers and the EMA module with cross-spatial learning. This augmentation integrates a multi-head self-attention (MHA) mechanism with EMA and EA. This structure is reinforced by the multilayer perceptron (MLP). EA computes attention between input pixels and external memory units using the following formula:

$$\begin{aligned} A_{\text{map}} &= \text{Norm}\left(fM_k^T\right), \\ f_{\text{out}} &= A_{\text{map}} M_v. \end{aligned} \tag{5}$$

A_{map} is the attention matrix. We utilize M_k and M_v as K and V, respectively, which improves network performance.

EMA enhances efficiency and semantic information distribution by partitioning the channel dimension into sub-features for parallel processing. It employs three parallel pathways to extract attention-weight descriptors. Two pathways capture long-range dependencies along one dimension while preserving positional information in the other. A 1x1 convolution then generates channel attention maps. Two-dimensional global average pooling is applied to both branches to encode global spatial information. The global pooling operation is formulated as:

$$AvgPool_c = \frac{1}{H \times W} \sum_{q}^{H} \sum_{p}^{W} x_c(p, q), \tag{6}$$

where x_c denotes the input features at $c-th$ channel and (p,q) is the pixel index in the feature map. The third pathway utilizes a 3×3 convolution to capture multi-scale features and expand the feature space. To optimize the learning efficiency of crucial input data features, layer normalization (LN) Ba et al. (2016) precedes each module. And it incorporates a local residual structure. Figure 3 illustrates the architectures of the AF-Encoder$_3$ and AF-Decoder$_3$. The overall computational process within the AF-Encoder is formulated as:

$$\begin{aligned} m_i' &= \text{MSA}(\text{EMA}(\text{EA}(\text{Ln}(m_{(i-1)})))) + m_{(i-1)} \\ m_i &= \text{MLP}(\text{EA}(\text{Ln}(m_i'))) + m_i', \end{aligned} \tag{7}$$

In the MLP, one of the two layers utilizes the GeLU Hendrycks and Gimpel (2023) activation function. The AF-Encoder module converts local features extracted from different image stages into a digital format. It can efficiently extract essential information. Then, the AF-Decoder amalgamates the embedded representations and combines them to generate image features with augmented dimensions.

AF-Decoder: In contrast to the previously mentioned encoder, the AF-Decoder incorporates a Multi-Scale Attention (MSA) module. This component facilitates the output features from the AF-Encoder and the input features of the AF-Decoder. It merges the K and V from the Encoder's output with the Q from the previous Decoder layer's output. Subsequently, it projects K, V, and Q into a new feature space and employs multiple sets of Scaled Dot-Product Attention to capture correlations from various aspects. Finally, it combines the attention outcomes and produces the final result. MSA serves as the cornerstone for AF-Decoder functionality. The extensive computational framework of the AF-Decoder is formulated as:

$$\begin{aligned} n_i' &= \text{MSA}(\text{EMA}(\text{EA}(\text{Ln}(\text{n}_{(\text{i}-1)})))) + n_{(i-1)} \\ n''_i &= \text{MSA}(\text{Ln}(n'_{(i-1)}), \text{Ln}(n_0)) + n'_{(i-1)} \\ n_i &= \text{MLP}(\text{EA}(\text{Ln}(n_i''))) + n_i'' \\ n_0 &= \left[f_{E_1}, f_{E_2}, \ldots, f_{E_N}\right] \\ \left[f_{D_1}, f_{D_2}, \ldots, f_{D_N}\right] &= m_{L_d}, \end{aligned} \tag{8}$$

where f_{Di} is the output of the decoder and f_{Ei} is the output of the encoder. L_d denotes the number of layers in the decoder.

Multistage reinforcement: THEDNet employs multiple encoders and decoders to enhance multistage features. Specifically, the input features for the three low-dimensional feature encoders are derived from $FE_i(i = 1,2,3)$. The AF-Encoder$_4$ module supplies the upsampled high-dimensional feature encoding. This feature serves as Q in the AF-Decoder. The features are encoded as K and V by the AF-Encoder$_i(i = 1,2,3)$. The fusion process takes place in the hybrid input multi-head self-attention (MHSA) module of the decoder. It is denoted as:

$$\begin{aligned} \text{SA}\,(Q, K, V) &= \text{softmax}\left(QK^T/\sqrt{d_k}\right)V, \\ \text{MH}(Q, K, V) &= \text{concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^O, \\ \text{head}_j &= \text{SA}\left(QW_j^Q, KW_j^K, VW_j^V\right), \end{aligned} \tag{9}$$

where SA(·) denotes the dot-product attention function. d_k represents the feature dimension in the decoders and j denotes the number of heads in the MSA module. Additionally, W_j^Q, W_j^K, W_j^V, and W^O represent projection matrices. The subscripts of the $Q/K/V$ variables are determined by the numbering of the AF-Encoder.

3.3 Loss function

The L_1 loss function is introduced to evaluate the difference between LR and SR images. It induces the model to generate images with a relatively smooth appearance and demonstrates resilience to outliers. Based on the given I_{LR} image and the corresponding SR image I_{SR}, the loss function is denoted as:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|I_{\mathrm{SR}}^{i} - I_{\mathrm{LR}}^{i}\right| \tag{10}$$

where i denotes the $i-th$ image and n is the total number of training datasets.

4 Experiments results and analysis

4.1 Dataset

We performed the experiments on two medical Low-Dose CT (LDCT) image datasets McCollough et al. (2020) and QIN-LUNG dataset Kalpathy-Cramer et al. (2015). The LDCT dataset encompasses CT scans acquired from 299 clinical patients. We derived two datasets from LDCT images using variations in X-ray dosage and windowing techniques. The LDCT-A and LDCT-B. LDCT-A consists of full-dose images, while LDCT-B consists of quarter-dose images. LDCT-A is based on a bone window (window width of 1500, window level of 450), and LDCT-B is based on a mediastinum window (window width of 300, window level of 40). LDCT-A comprises 1822 images for training and 450 for testing. LDCT-B includes 892 training images and 240 testing images. Additionally, the QIN-LUNG dataset consists of CT scans from 47 lung cancer patients. It utilizes 328 images for training and 150 for testing.

4.2 Experimental settings

We investigate the SR of medical images at three scale factors: ×2, ×3, and ×4. The original images in each dataset are utilized as high-resolution (HR) images. We use the bicubic downsampling method to create corresponding LR images. We extracted 48 × 48 blocks from the LR images in the training phase. Additionally, corresponding reference blocks were randomly extracted from the HR dataset. We incorporate random rotations of 90° ×n (n=1, 2, 3) and horizontal mirroring to expand the training sample set. Additionally, we integrated the inverse projection technique fromGlasner et al. (2009); Shocher et al. (2018) into the network. This mitigates the influence of block effects on generating the final HR image results. To enhance the training of the substantial dataset, we employ the Adam Kingma and Ba (2014) optimizer. Parameters for the Adam optimizer are specified as $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-8}, and we utilize a batch size of 16 throughout the training phase. We employed PyTorch and conducted the experiments on two parallel NVIDIA 3080Ti GPUs.

4.3 Evaluation metrics

The objective results are evaluated using two metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM)Wang et al. (2004). The PSNR gauges image reconstruction quality by analyzing the ratio of peak signal power to noise power. Higher PSNR values correlate with lower levels of distortion Sara et al. (2019). The SSIM assesses the similarity between two images by comparing brightness, contrast, and structure. SSIM values closer to 1 indicate a higher degree of resemblance between the images Thung and Raveendran (2009); Li et al. (2024). They are calculated as:

$$\mathrm{PSNR}(x,y) = 10\cdot\log 10\times\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2}\right), \tag{11}$$

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

$$\mathrm{SSIM}(x,y) = \frac{2\mu_x\mu_y + k_1}{\mu_x^2 + \mu_y^2 + k_1}\cdot\frac{\sigma_{xy} + k_2}{\sigma_x^2 + \sigma_y^2 + k_2}, \tag{12}$$

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

4.4 Comparative analysis

To evaluate the effectiveness of the proposed THEDNet, we conducted comparative experiments with SRCNN Dong et al. (2015), FSRCNN Dong et al. (2016), SRGAN Ledig et al. (2017), RDN Zhang et al. (2018), SRFBNN Li et al. (2019), GMFN Li et al. (2019), and TransENet Lei et al. (2021) methodologies. Objective evaluations were conducted on the LDCT-A, LDCT-B, and QIN-LUNG test sets for three scale factors: ×2, ×3, and ×4.

Table 1 presents a comparative analysis of PSNR and SSIM scores among the LDCT-A, LDCT-B, and QIN-LUNG datasets. The experimental results indicate that THEDNet performs best in PSNR on the LDCT-B and QIN-LUNG datasets across different scale factors. On the LDCT-A dataset, THEDNet ranks first in PSNR for ×3 and ×4 scale factors, and second place for the ×2 scaling factor. Specifically, we compare the proposed THEDNet with SRCNN at the ×2 scale factor. The average improvement of PSNR and SSIM values for THEDNet is 4.36 dB and 0.0025, respectively. Similarly, compared with the FSRCNN, the PSNR and SSIM values have increased by 4.01 dB and 0.0023 dB, respectively. Compared with the baseline, TransEnet, the PSNR and SSIM values are improved by 3.09 dB and 0.0010,

Table 1 Comparative results on the LDCT-A, LDCT-B, and QIN-LUNG datasets. (The best result is highlighted in **bold**.)

Algorithm	Scale	LDCT-A		LDCT-B		QIN-LUNG	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SRCNN Dong et al. (2015)	×2	43.79	0.9822	33.82	0.9488	34.20	0.9352
FSRCNN Dong et al. (2016)	×2	44.14	0.9837	34.39	0.9502	35.65	0.9362
SRGAN Ledig et al. (2017)	×2	39.80	0.9652	32.92	0.9463	27.55	0.8426
RDN Zhang et al. (2018)	×2	44.53	0.9841	35.05	0.9508	37.15	0.9401
SRFBN Li et al. (2019)	×2	47.32	0.9878	35.41	0.9523	38.49	0.9800
GMFN Li et al. (2019)	×2	**48.66**	**0.9886**	35.42	0.9529	38.58	**0.9802**
TransENet Lei et al. (2021)	×2	45.06	0.9824	33.02	0.9412	33.42	0.9188
Ours	×2	48.15	0.9847	**36.20**	**0.9635**	**41.00**	0.9625
SRCNN Dong et al. (2015)	×3	39.12	0.9633	29.86	0.8072	31.85	0.8578
FSRCNN Dong et al. (2016)	×3	38.87	0.9623	30.19	0.8116	32.28	0.8612
SRGAN Ledig et al. (2017)	×3	–	–	–	–	–	–
RDN Zhang et al. (2018)	×3	44.70	0.9668	31.79	0.8853	33.24	0.8911
SRFBN Li et al. (2019)	×3	44.16	**0.9804**	31.75	0.8843	34.55	0.9512
GMFN Li et al. (2019)	×3	44.80	0.9630	31.84	0.8856	34.55	**0.9516**
TransENet Lei et al. (2021)	×3	42.07	0.9705	29.64	0.8622	30.31	0.8729
Ours	×3	**46.03**	0.9428	**32.14**	**0.9035**	**37.86**	0.9406
SRCNN Dong et al. (2015)	×4	36.63	0.9465	28.46	0.8337	27.48	0.8381
FSRCNN Dong et al. (2016)	×4	37.06	0.9363	28.49	0.8215	27.55	0.8668
SRGAN Ledig et al. (2017)	×4	35.99	0.9308	27.92	0.8306	24.44	0.8097
RDN Zhang et al. (2018)	×4	40.78	0.9546	29.83	0.8346	30.43	0.8462
SRFBN Li et al. (2019)	×4	41.05	0.9714	30.06	0.8398	31.78	0.9226
GMFN Li et al. (2019)	×4	42.55	**0.9748**	30.02	0.8386	31.70	**0.9237**
TransENet Lei et al. (2021)	×4	40.01	0.9602	28.09	0.8072	28.24	0.8359
Ours	×4	**44.62**	0.9216	**30.12**	**0.8597**	**34.31**	0.8995

respectively. In LDCT-B test sets across different scale factors, THEDNet performs best regarding PSNR and SSIM. Moreover, compared with SRFBN which utilizes feedback mechanisms, the PSNR scores demonstrated incremental improvements of 0.79 dB, 0.39 dB, and 0.06 dB. The SSIM scores improved by 0.0112, 0.0192, and 0.0199, respectively. On the QIN-LUNG test set, THEDNet surpassed the competitors in PSNR metrics. Compared with GMFN, which employs a many-to-many feedback connection mechanism, THEDNet improves the PSNR by 2.42 dB, 3.31 dB, and 2.61 dB for scale factors ×2, ×3, and ×4, respectively. Compared to TransENet, THEDNet improves PSNR by 22.68%, 24.91%, and 21.49% for ×2, ×3, and ×4 scale factors.

To demonstrate the effect of the proposed model visually, we provide a subjective comparison of ablation experiments. We chose the ×4 scale factor for testing on three datasets, LDCT-A, LDCT-B, and QIN-LUNG. Comparative results are shown in Fig. 4.

Figure 4a–c display the SR results obtained for distinct anatomical regions: left kidney, spleen, and lung. Specifically, Fig 4a indicates that the left kidney scans generated by SRFBN, GMFN, and THEDNet exhibit clearer details than those displayed by FSRCNN and SRGAN. Each bone exhibits distinct contours, and its clarity allows for more accurate observation of bone morphology and structure. This aids doctors in identifying potential abnormalities or injuries. Figure 4b illustrates the contrast of details in the spleen region in the upper abdominal soft tissue window. THEDNet can distinctly exhibit details of the spleen and the associated blood vessels. Simultaneously, it can also present a clearer structure of the ribs. Figure 4c showcases the qualitative results of the QIN-LUNG dataset. This CT lung window scan image enhances lung structure and detail. The texture and anatomical structure of lung tissue are more prominently visible in the image. This provides robust support for the detection of pulmonary lesions. It shows that THEDNet can capture details in the esophagus and the upper part of the thoracic vertebrae. Compared to other methods, the THEDNet outperforms other competitors in enhancing the depth and hierarchy perception of the image.

4.5 Ablation study

To demonstrate the effectiveness of EA and EMA, we performed ablation experiments on test set QIN-LUNG at a scale factor of ×4. Table 2 displays the quantitative evaluation results. The baseline model indicates the absence of integration of both EMA and EA attention modules.

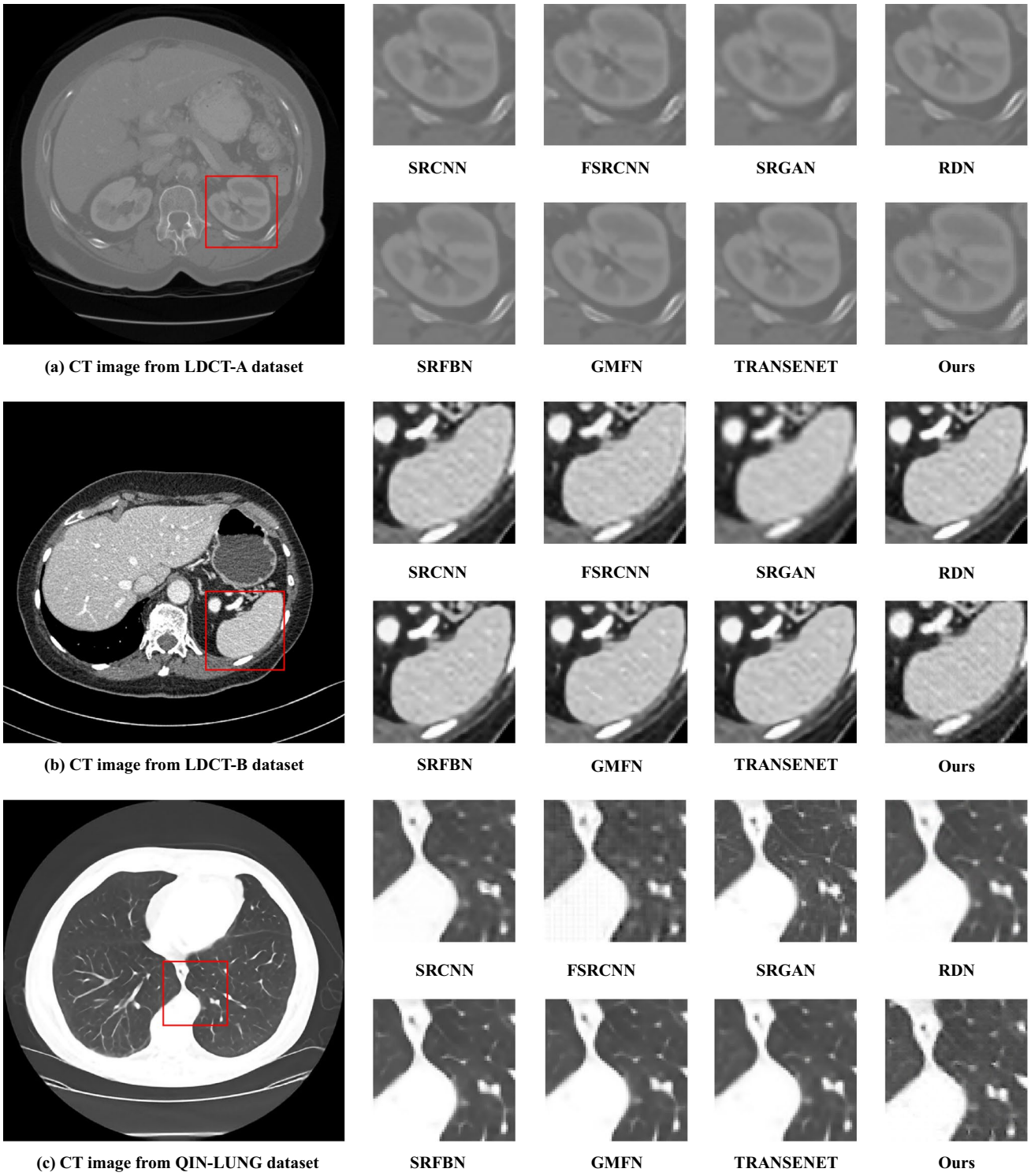

Fig. 4 Comparative analysis of CT image SR at ×4 scale factor against alternative methods

The separate introduction of the EMA module for fusion marginally enhances PSNR and SSIM. The EA module introduced separately for fusion reduces PSNR and SSIM. Notably, the simultaneous integration of EMA and EA modules leads to a significant improvement in PSNR and SSIM. This affirms the effectiveness of the proposed method in integrating MHA, EMA, and EA.

Table 2 Ablation analysis of the proposed model for various attentional fusion strategies

Scale	EMA	EA	PSNR ↑	SSIM ↑
×4	×	×	28.24	0.8359
×4	✓	×	28.68	0.8347
×4	×	✓	27.85	0.8321
×4	✓	✓	**31.84**	**0.8389**

Table 3 The result of different layers of encoder and decoder. (The best result is highlighted in **bold**.)

Encoder layers	Decoder layers	PSNR ↑	SSIM ↑
1	1	32.14	0.8076
1	2	31.82	0.8019
1	4	**34.31**	**0.8995**
1	8	32.63	0.8347
1	12	32.99	0.8776
2	1	32.42	0.8734
4	1	19.54	0.3842
8	1	31.85	0.8389
12	1	25.49	0.5277

We conducted a series of ablation studies to investigate the influence of the number of encoder and decoder layers on model performance. In Table 3, we present a comparison of the experimental results obtained on the QIN-LUNG data set with different numbers of layers. The scale factor is ×4. It can be observed that the model performs optimally when the number of encoder layers is 1 and the number of decoder layers is 4. It indicates that extracting features with high dimensions should be the focus of feature processing.

5 Conclusion

This work presents a Transformer-based hierarchical Encoder–Decoder with Attention Fusion Enhancement Network (THEDNet) for CT image super-resolution. THEDNet utilizes the Transformer architecture to enrich high-dimensional feature representations of the upsampling layer. This approach emphasizes attention fusion in both encoder and decoder by integrating external attention (EA) and efficient multi-scale attention (EMA). This aids in capturing long-distance interdependencies among features. We conducted comparative experiments on two publicly available medical image datasets. Ablation studies prove that the EA and EMA modules can improve the model's performance in PSNR and SSIM. Meanwhile, comparative results on QIN-LUNG and LDCT datasets verify that the THEDNet is comparable to the SOTA competitors.

Acknowledgements This work is partly supported by the Project of Shandong Province Graduate High-Quality Education and Teaching Resources (No. SDJAL2023017).

Data availability Data sharing does not apply to this article, as no datasets were generated or analyzed during the current study.

Declaration

Conflict of interest The authors declare that they have no Conflict of interest.

References

Ayas S, Ekinci M (2020) Single image super resolution using dictionary learning and sparse coding with multi-scale and multi-directional gabor feature representation. Inf Sci 512:1264–1278

Ba JL, Kiros JR, Hinton GE (2016) Layer normalization

Chen L, Yang X, Jeon G, Anisetti M, Liu K (2020) A trusted medical image super-resolution method based on feedback adaptive weighted dense network. Artif Intell Med 106:101857

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

Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale

Georgescu MI, Ionescu RT, Verga N (2020) Convolutional neural networks with intermediate loss for 3d super-resolution of ct and mri scans. IEEE Access 8:49112–49124

Glasner D, Bagon S, Irani M (2009) Super-resolution from a single image. In: 2009 IEEE 12th international conference on computer vision, pp. 349–356. IEEE

Gu Y, Zeng Z, Chen H, Wei J, Zhang Y, Chen B, Li Y, Qin Y, Xie Q, Jiang Z et al (2020) Medsrgan: medical images super-resolution using generative adversarial networks. Multimed Tools Appl 79:21815–21840

Guo MH, Liu ZN, Mu TJ, Hu SM (2021) Beyond self-attention: External attention using two linear layers for visual tasks

Han X, Xie Z, Chen Q, Li X, Yang H (2023) Learning the degradation distribution for medical image superresolution via sparse swin transformer. Comput Graph

Hendrycks D, Gimpel K (2023)Gaussian error linear units (gelus)

Kalpathy-Cramer J, Napel S, Goldgof D, Zhao B (2015) Qin multi-site collection of lung ct data with nodule segmentations. Cancer Imaging Arch **10**, K9. https://wiki.cancerimagingarchive.net/display/Public/QIN+LUNG+CT

Khaledyan D, Amirany A, Jafari K, Moaiyeri MH, Khuzani AZ, Mashhadi N (2020) Low-cost implementation of bilinear and bicubic image interpolation for real-time image super-resolution. In: 2020 IEEE Global Humanitarian Technology Conference (GHTC), pp. 1–5. https://doi.org/10.1109/GHTC46280.2020.9342625

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network

Lei S, Shi Z, Mo W (2021) Transformer-based multistage enhancement for remote sensing image super-resolution. IEEE Trans Geosci Remote Sens 60:1–11

Li Q, Li Z, Lu L, Jeon G, Liu K, Yang X (2019) Gated multiple feedback network for image super-resolution. arXiv preprint arXiv:1907.04253

Li X, Pan J, Shang J, Souri A, Gao M (2024) An improved blind/referenceless image spatial quality evaluator algorithm for image quality assessment. Int J Comput Sci Eng 27(1):48–56

Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W (2019) Feedback network for image super-resolution

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

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022

Lu W, Song Z, Chu J (2020) A novel 3d medical image super-resolution method based on densely connected network. Biomedical Signal Processing and Control 62:102120. https://doi.org/10.1016/j.bspc.2020.102120.https://www.sciencedirect.com/science/article/pii/S1746809420302706

McCollough C, Chen B, Holmes D, Duan X, Yu Z, Xu L, Leng S, Fletcher J (2020) Low dose ct image and projection data. The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52758026

Mehri A, Ardakani PB, Sappa AD (2021) prnet: Multi-path residual network for lightweight image super resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2704–2713

Mei Y, Fan Y, Zhou Y, Huang L, Huang TS, Shi H (2020) mage super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5690–5699

Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, Huang Z (2023) Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096516

Ran R, Deng LJ, Jiang TX, Hu JF, Chanussot J, Vivone G (2023) Guidednet: A general cnn fusion framework via high-resolution guidance for hyperspectral image super-resolution. IEEE Trans Cybern

Sara U, Akter M, Uddin MS (2019) Image quality assessment through fsim, ssim, mse and psnr-a comparative study. J Comput Commun 7(3):8–18

Shang J, Zhang X, Zhang G, Song W, Chen J, Li Q, Gao M (2022) Gated multi-attention feedback network for medical image super-resolution. Electronics **11**(21). https://doi.org/10.3390/electronics11213554. https://www.mdpi.com/2079-9292/11/21/3554

Shang S, Shan Z, Liu G, Wang L, Wang X, Zhang Z, Zhang J (2024) Resdiff: Combining cnn and diffusion model for image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 8975–8983

Shocher A, Cohen N, Irani M (2018) 'zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3118–3126

Thung KH, Raveendran P (2009) Survey of image quality measures. In: 2009 international conference for technical postgraduates (TECHPOS), pp. 1–4. IEEE

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:2

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

Wang Z, Chen J, Hoi SC (2020) Deep learning for image super-resolution: a survey. IEEE Trans Pattern Anal Mach Intell 43(10):3365–3387

Woo S, Park J, Lee JY, Kweon IS (2018) Cbam:Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19

Yan C, Shi G, Wu Z (2021) Smir: A transformer-based model for mri super-resolution reconstruction. In: 2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE), pp. 1–6. https://doi.org/10.1109/ICMIPE53131.2021.9698880

Yang F, Yang H, Fu J, Lu H, Guo B (2020) Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Yoo J, Kim T, Lee S, Kim S, Lee H, Kim T (2022) Rich cnn-transformer feature aggregation networks for super-resolution. arXiv preprint arXiv:2203.07682

Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks

Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2472–2481