Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2025.KL24798

Multi-Scale Convolution-Aided Transformer-Based Medical Image Super-Resolution

Ziyi Zhou

School of Computer Science and Technology, Anhui University, Hefei, China
e22201033@stu.ahu.edu.cn

Abstract. The focus of this study is on the multi-scale super-resolution technology for medical images, with the intention of enhancing the fine texture features of the Transformer model. With the evolution of deep learning, medical image super-resolution has turned into a tool for boosting diagnostic accuracy and aiding clinical decision-making, making it a crucial area of research. However, although the Transformer model performs well in dealing with long-distance dependencies, it is sensitive to image details and texture information. The capturing ability is relatively insufficient. This can lead to the neglect of some small but crucial texture features during the high-resolution reconstruction process, thus affecting the final super-resolution result. In order to make up for this shortcoming, this study will introduce an additional convolutional neural network structure. The framework operates in concert to fully leverage the advantages of CNN in local detail extraction.

Keywords: Medical image super-resolution, Transformer

1. Introduction

As medical technology continues to evolve and progress, medical imaging has become an indispensable auxiliary tool in modern medical diagnosis. The quality of medical images has a crucial impact on disease diagnosis. Nevertheless, the medical images acquired in practice frequently encounter the issue of low resolution (LR) due to hardware device limitations or image capturing constraints[1]. To counteract this, methods for improving the determination of medical images, known as super-resolution[2], have been developed.

Super-resolution (SR) methods try to retrieve a high-definition images from a unclear images image. Since it can obtain high-quality medical images which are beneficial to downstream diagnosis tasks, SR has been widely studied[3][4]. For example, in gan-circle, You et al.[4] replaced the ReLU in residual blocks employing LeakyReLU for nonlinear processing to enhance super-resolution. Based on this model, Jiang et al.[3] used skip connections to iteratively learn the output of the previous layer with 16 identical residual blocks, and applied parallel 1×1 convolutional operations to diminish the size of the output from each hidden layer, making network training smoother.

Even though these techniques have reached satisfactory performance, they often focus on local regions while ignoring global semantics and contextual information. Additionally, they suffer from inadequate feature utilization and due to their single attention source. For dealing with these matters, In this research paper, we put forward a multi-scale convolution-aided transformer-based medical image super-resolution (MSCT-SR) network. Inspired by visual transformers (ViT)[5][6] in

DOI: 10.54254/2755-2721/2025.KL24798

modeling long-range contextual information, we propose a customized transformer for medical image super-resolution. However, considering their lack of adequate fusion and transformation of different information in the feature maps during the decoding process, we propose a novel fusion module and plug it into the decoder of ViT. This module can enhance the model's expressive ability and feature extraction capabilities by introducing a gating mechanism. Besides, there still exist limitations that need to be addressed when applying visual transformers to medical image super-resolution. Transformers often apply relatively large and coarse-grained patches, and thus are not conducive to learning precise edge features. Extracting precise edges from small objects is difficult, and inaccuracy reconstruction of edges may directly affect doctors' observations and diagnoses. To tackle this problem, we apply a convolution based network to extract the detail and edge information. The Transformer based network and convolution based network are alongside simultaneously record images' global and local data.

Since we utilize two branches, i.e., Transformer based branch and convolution based branch, we need to fuse the representations of the two branches. Consequently, we introduce a unique Pyramid Multi-Scale Feature Fusion (PMS-FN) module designed to efficiently merge the two representations. Our inspiration stems from the principle of multi-scale fusion.[7], we first extract features of different scales of each branch, and integrate features of the two branches in each scale respectively. Subsequently, we merge the representations across various scales. This method enables us to incorporate semantic information from high-level features into the lower-level images, thereby effectively reconstructing both the low-stage details and the high-stage semantics. After applying the PMS-FN, we utilize an upsampling layer to achieve the super-resolution images.

The contributions of this thesis are summarized as follows:

- We present a new MSCT-SR framework proposal, which simultaneously apply Transformer and convolutional network to extract the global and local features form medical images. In the Transformer branch, we design a new fusion module via a gating mechanism to boost the feature extraction. To address the issue that the transformer is not good at learning precise edge features, we propose a convolutional network to extract detail and edge information.
- To effectively fuse the representations of Transformer and convolutional network, we design an innovative PMS-FN which is capable of effectively reconstructing high-level semantic and low-level detail information for super-resolution.
- Our experiments on various benchmark medical image datasets have yielded results indicating that our method outperforms other leading super-resolution methods.

2. Related Works

In this section, we supply a concise overview of the related research on image super-resolution.

The concept of super-resolution was initially proposed by Gerchberg[2]. With the advancement of technology, super-resolution technology has been redefined as a method aimed at restoring low-resolution images (with limited information) to obtain structurally rich high-definition images, thereby solving the problem of information loss in low-resolution images. Traditional image interpolation algorithms increase the resolution of the image simply through enhancement the pixel size[8]. While this approach is fast and simple to carry out, it lacks a deep understanding of the complex structures and textures in images, making it difficult to accurately restore the lost high-frequency information.

For example, owing to the shortage of consideration for structural information contained medical images, traditional interpolation algorithms may introduce artifacts[9]. Deep learning methods can effectively learn high-frequency information in images by learning image mapping relationships from a large amount of data. Therefore, the mainstream approach for image super-resolution reconstruction

DOI: 10.54254/2755-2721/2025.KL24798

is rooted in deep learning. For instance, Oktay et al.[10] proposed a versatile training strategy to integrate prior knowledge from a novel regularization model into neural networks for super-resolution. Hong et al.[11] applied single-image super-resolution techniques that leverage deep convolutional neural networks are applied to PET imaging.

Besides the convoultional based methods, Transformer also has been used for super-resolution. Transformer was originally suggested by Vaswani et al.[12] in 2017. In image super-resolution, Transformer has also been widely studied[13]. For example, Chen et al.[13] have put forth a pre-training model that leverages the Transformer framework for image processing. This model is amenable to fine-tuning on smaller datasets and can be directly utilized for specialized tasks like denoising and enhancing image resolution.

3. Method

In this section, we will offer a comprehensive review of our MSCT-SR network. We will consider ×4 super-resolution as a representative example, which can be easily adapted to other magnifications. Figure 1 depicts the architecture of the MSCT-SR network. It mainly comprises three distinct modules: The Global Feature Extraction (GFE) module, which incorporates a specialized transformer designed for extracting global information, the Local Feature Extraction (LFE) module, which leverages CNN to capture detailed and edge information, thereby capturing the local details of the image, and the Pyramid Multi-scale Feature Fusion (PMS-FN) module, which is adept at effectively integrating the feature information extracted by the first two modules. Given a LR image, we first upsample it to a $\times 2$ and $\times 4$ scales, and then feed the two rescaled images together with the original image into both the Transformer and CNN branches to extract features. Finally, we used a residual block and an upsampling layer to reconstruct the SR image from the feature map obtained by the PMS-FN module. Hereafter, a more detailed explanation of these modules will be provided.

3.1. Global Feature Extraction Module

Within this module, we leverage a Transformer to obtain global feature information from images. The encoding approach of this transformer's encoder is analogous to the Vision Transformer (ViT) as outlined in [14]. In more detail, Initially, we segment the input image into multiple fixed-size image patches. Let $I_r \in R^{C \times H \times W}$ represent the input image, where where C, and H, and W denote the quantity of channels, the vertical dimension, and the horizontal dimension, correspondingly. We evenly divide I_r into a series of flat image patches of size $S \times S$. Then, the sequence undergoes a trainable linear mapping, which transforms it into a latent embedding space. To increase the feature dimension and facilitate feature fusion, inspired by[15], we adopt two consecutive 3×3 convolutions with a stride of 2 and batch normalization to implement the process of linear projection to patch embeddings. In addition, to ensure the preservation of positional information, we adopt standard trainable one-dimensional positional embeddings and combine them with patch embeddings. Ultimately, this combined embedding is inputted into the encoder for processing.

The traditional Transformer encoder[12] is composed of multiple Transformer blocks, each of which incorporates a Multi-Layer Perceptron (MLP), two Layer Normalization (LN) components and a Multi-head Self-Attention mechanism (MSA). To achieve a more comprehensive extraction of image information, the encoder of our global feature extraction module mainly includes an attention mechanism (combining local self-attention and conventional attention), a multi-layer perceptron (MLP), normalization, dropout, and stochastic depth techniques to extract global semantic information from images.

DOI: 10.54254/2755-2721/2025.KL24798

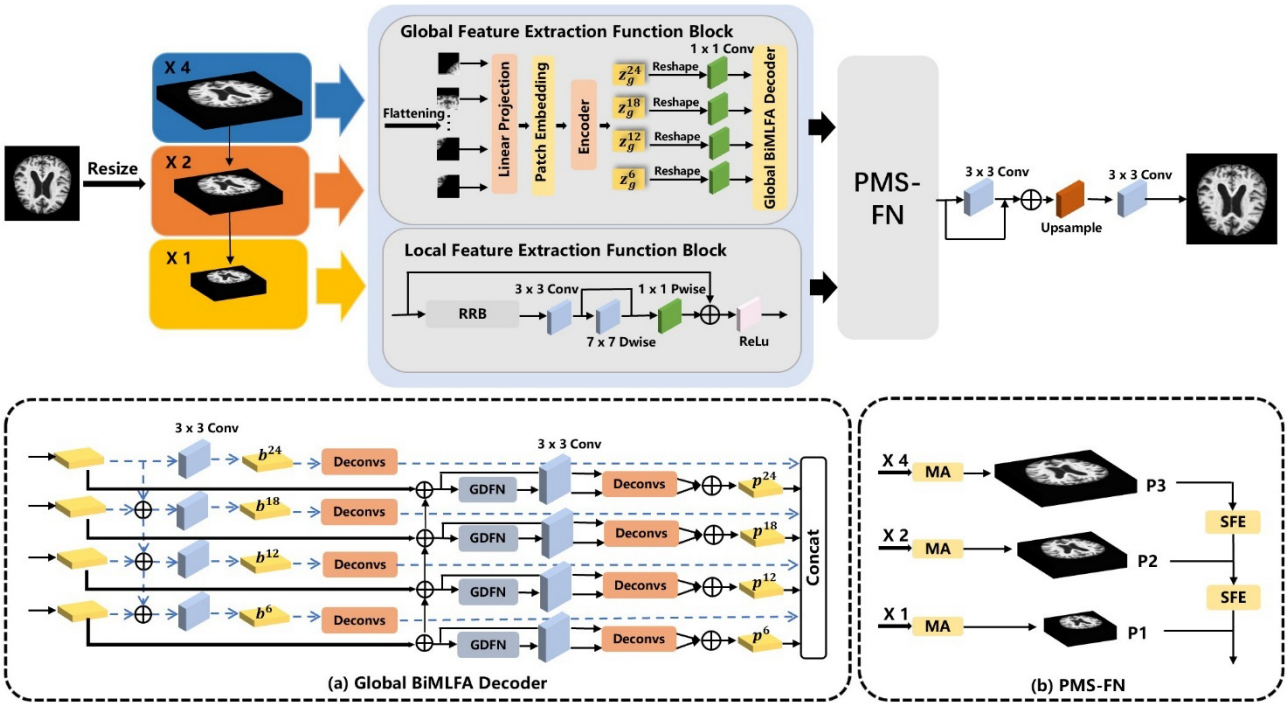

Figure 1: The architecture of MSCT-SR network.

Notice that traditional transformer-based methods for the vision task[5][6] do not perform sufficient fusion and transformation of different information in feature maps during the decoding process. To tackle this problem, we design a Global Bi-directional Multi-Level Feed-Forward Aggregation (BiMLFA) decoder, whose structure is shown in Figure 1(a). It applies the Gated-Dconv Feed-Forward Network (GDFN)[16] for high-resolution image restoration. It leverages spatially adjacent pixel positions to aid in learning image structures for effective image fusion. Therefore, we incorporate it into the decoder of ViT to achieve an efficient fusion of feature maps.

Firstly, we evenly distribute the N_g feature blocks into four distinct blocks and obtain the embedding feature blocks$\{ Z_6^g, Z_{12}^g, Z_{18}^g, Z_{24}^g \}$ from the final block of each group for input purposes. Subsequently, we transform these into three-dimensional features with the dimensions $H_{16} \times W_{16} \times C$. The formulaic representation is as follows:

$$\mu_r = D_e \left\{ Z_6^g, Z_{12}^g, Z_{18}^g, Z_{24}^g \right\} \tag{1}$$

μ_r is the three-dimensional features reshaped by the decoder, while is D_e set as the decoder.

With the top-down approach, we incorporate the identical design (a 1×1convolutional layer followed by a 3×3 convolutional layer) to each feature that has been restructured, generating four output features: $b_6, b_{12}, b_{18}, b_{24}$. They are subsequently inputted into deconvolutional layers, each succeeded by batch normalization (BN) and ReLU functions. Similarly, the foundational-to-peak method begins at the lowest tier(i.e., Z_6^g) and is splitted into two paths. The first path passes through a gated convolutional feedforward network followed by a 3×3 convolutional layer, gradually approaching the top layer (i.e., Z_{24}^g). The second path follows the same initial steps as path one but then also gradually approaches the top layer (i.e., Z_{24}^g), and subsequently undergoes a deconvolution and subsequent operations, yielding four output features $p_6, p_{12}, p_{18}, p_{24}$. Finally, the feature map of the global feature extraction module is formed by concatenating these eight sampled features into a unified tensor.

DOI: 10.54254/2755-2721/2025.KL24798

3.2. Local Feature Extraction Module

To address the limitations of the global feature extraction module in capturing image edges and details, we propose a local feature extraction module that employs convolution operations to extract local features from images. Since the RRB[17] structure, whose structure is shown in Figure 2(a), has the potential to uncover complex structures while maintaining computational efficiency, we apply it to extract features.

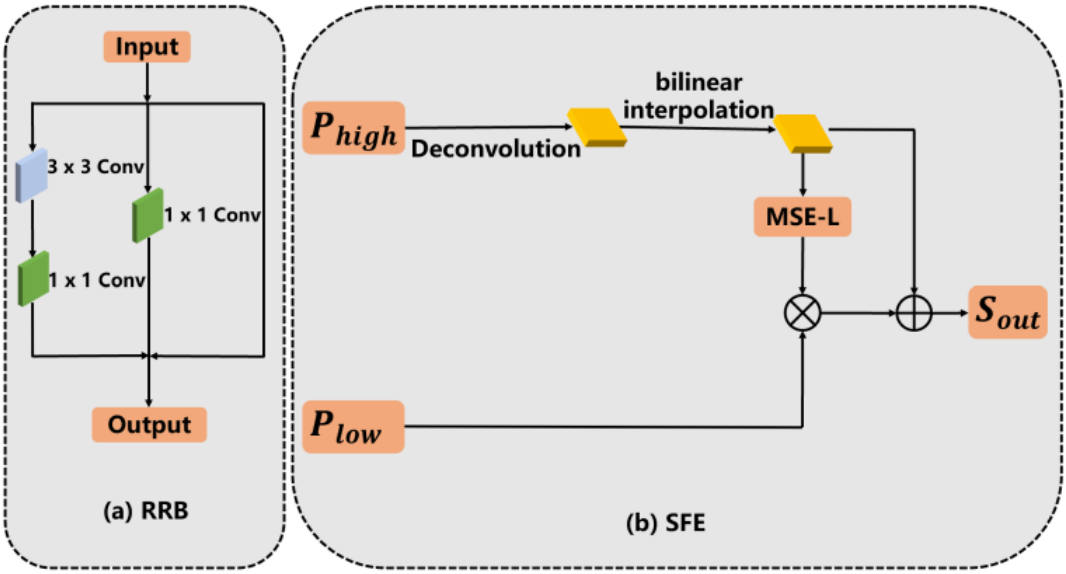

Figure 2: Figure 2(a) is the RRB structure, and figure 2(b) is the SFE structure.

To this end, we first pass the image through the RRB structure to capture the complex details of the image, and then we apply a single-layer 3×3 convolution to extract features. Research conducted previously has revealed that as the count of 3×3 convolutions layered on top of each other grows, the dimensions of the efficient receptive area tends to decrease to a certain extent[18]. To expand the efficient perception area to a certain degree, we process the image with a 7×7 depthwise convolution to achieve a broader receptive area, followed by a 1×1pointwise convolution and subsequent activation functions.

3.3. Pyramid Multi-scale Feature Fusion (PMS-FN) module

To fuse the features extracted from the two branches more effectively, we propose a Multi-content Aggregation (MA) attention mechanism, which is shown in Figure 3(a). Since medical images differ from natural images, it is important for medical image networks to focus on reconstructing the region of interest features, for instance textures and edges, when extracting medical image characteristics.

The existing methods often rely solely on single-scale content features[19]. Large-scale features contain a lot of detailed information, like structural layouts, while small-scale features may lack this detail and only provide simple local representations. This leads to difficulties for generation methods that depend only on single-scale features, making it hard to comprehensively understand the complex composition of the tissue in images and reproduce intricate details. Inspired by[7][19], we propose a Pyramid Multi-Scale Feature Fusion (PMS-FN) to promote multi-level feature fusion. Figure 1(b) shows PMS-FN. In more detail, the features obtained from the transformer branch, denoted as r_i, and the features obtained from the convolutional branch, denoted as f_i, are concatenated to form a channel information feature I_c. To enhance the adaptive selective channel fusion capability, we use a Mean Squared Error (MSE) 3(b) attention mechanism on I_c, which first combines average pooling and max

DOI: 10.54254/2755-2721/2025.KL24798

pooling processes, followed by a 3 × 3 convolution to generate the feature S_c. This goes through the same pooling process, an activation function, and a convolution layer with an activation function. The attention mechanism produces a global channel-aware vector W_C, which is used to weight the channel information feature Ic through inter-channel multiplication. Finally, after a residual connection, we apply a 1 × 1 convolution to less total channels in I_d to obtain the output I.

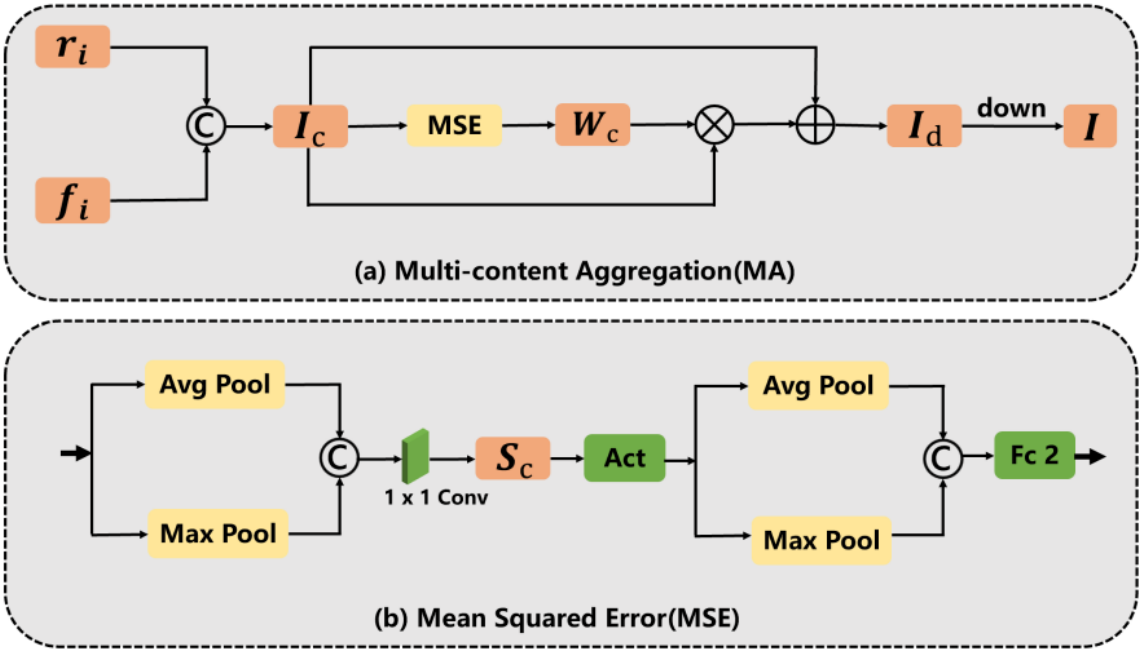

Figure 3: Figure 3(a) is the Multi-Content Aggregation (MA) structure, and figure 3(b) is the Mean Squared Error (MSE) structure.

In the final part, we proposed an SFE strategy to integrate semantic information from high-level features into low-level images, whose structure is shown in Figure 2(b). which can effectively reconstruct low-stage and high-stage specifics. Specifically, we use the multi-scale features P_1, P_2, and P_3 outputted by the Multi-Content Aggregation. First, we performed a deconvolution on P_3 , followed by bilinear interpolation to scale it to the size of P_2, resulting in P_{31}. To better fuse features from different scales, we passed it through MSE-L (which differs from MSE as it lacks a subsequent pooling and convolutional process) to generate features P_{32}. Then, we multiplied features P_2 and P_{32} element-wise, and finally concatenated the resulting feature with P_{31} to form feature S_{32}. The same steps were repeated for feature S_{32} and P_1.

4. Loss Function

In the MSCT-SR, we assess the pixel-by-pixel loss for the SR image I_{sr} and the ground truth HR images I_{gt} with

$$\mathcal{L}_{\mathrm{SR}} = \frac{1}{D}\sum_{i=1}^{D}\left\| I_{sr}^{(i)} - I_{gt}^{(i)} \right\|_1 \tag{2}$$

$I_{sr}^{(i)}$ is the actual high-resolution image of the ground truth of the i-th image in the training set and $I_{gt}^{(i)}$ is the MSCT-SR image of the i-th image in the training set . D is the count of images within the training dataset.

DOI: 10.54254/2755-2721/2025.KL24798

5. Experiments

5.1. Data Sets

We utilize three standard medical image datasets to assess the efficacy of our approach., including ADNI-MRI1 , ADNI-PET2 , and Demented-MRI3. The original ADNI-PET and ADNI-MRI data are 3D images which slice the images in ADMI-MRI into 5100 2D images at intervals of 0.01 mm along the axial plane and slice the images in ADMI-PET into 12000 2D images in the same way. The Demented-MRI dataset consists of 2D magnetic resonance imaging (MRI) images. This data set is categorized into three classes: non-dementia, mild dementia, and very mild dementia, with 890 MRI images in each category, totaling 2670 images. Each of the three datasets is haphazardly partitioned into training, validation, and test subsets, maintaining a proportion of 8:1:1.

5.2. Experimental Setup and Implementation Details

For the ADNI-MRI and ADNI-PET datasets, we resize the images to 128×128 and 148×148, respectively. In our approach, we established the batch sizes for the Demented-MRI, ADNI-PET, and ADNI-MRI datasets at 8, 4, and 8, respectively. The training epochs are established at 800, employing Adam as the optimization algorithm with a learning rate of 0.0001.

Our evaluation of reconstruction quality employs both the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Higher values for these two metrics indicate better reconstruction performance. We downsize the original data with a reduction factor of 4. Our approach employs Bicubic interpolation as the standard method. In addition, we evaluate our approach against the following advanced SR techniques: RCAN[20], IMDN[21], A²F[22], SWD[23], SwinIR[24], ELAN[25], SRDA[26], and HiT-SR[27]. For all other comparison methods, we used the authors' released code and adopted the hyperparameter settings recommended in the respective literature.

5.3. Experimental Results

Table 1 displays the quantitative comparison results of PSNR and SSIM for all methods across three datasets. For most datasets presented in the figure, our method outperforms others inStandard indicators. This demonstrates the effectiveness of our method.

Figures 6, 5 and 4 display the qualitative performance of different methods on the Demented-MRI datasets, ADNI-PET, and ADNI-MRI datasets, respectively. In the zoomed-in section of the upper left corner of Figure 4, it can be seen that the super-resolution images produced by our method contain more details and are closer to the original real images. Similarly, from the tissue lines in the upper left corner of Figure 6 and the overall tissue line direction in Figure 5, we can see that our method generates super-resolution images with clearer reconstructed tissue lines that match the original real images line distribution. Therefore, these images indicate that our network can reconstruct high-resolution image details, proving its effectiveness and superiority.

1https://adni.loni.usc.edu/

2https://adni.loni.usc.edu/

3https://www.heywhale.com/mw/dataset/\6245d687e1d37c0017029c05/content

Table 1: Different methods comparison on medical datasets

Methods	Demented - MRI		ADNI - PET		ADNI - MRI	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	22.95	0.795	30.92	0.894	25.15	0.727

DOI: 10.54254/2755-2721/2025.KL24798

Table 1: (continued).

IMDN[21]	27.86	0.931	34.34	0.943	31.11	0.910
RCAN[20]	30.53	0.953	35.19	0.951	34.32	0.950
SWD[23]	30.15	0.935	32.32	0.930	31.56	0.917
A²F[22]	29.69	0.946	**38.20**	0.953	33.36	0.940
SwinIR[24]	30.15	0.952	36.14	0.953	33.57	0.944
ELAN[25]	27.61	0.926	34.14	0.943	30.86	0.904
SRDA[26]	30.69	0.956	35.24	0.951	31.72	0.922
HiT-SR[27]	26.42	0.910	33.97	0.940	29.61	0.887
MSCT-SR	**30.86**	**0.957**	37.63	**0.959**	**34.37**	**0.955**

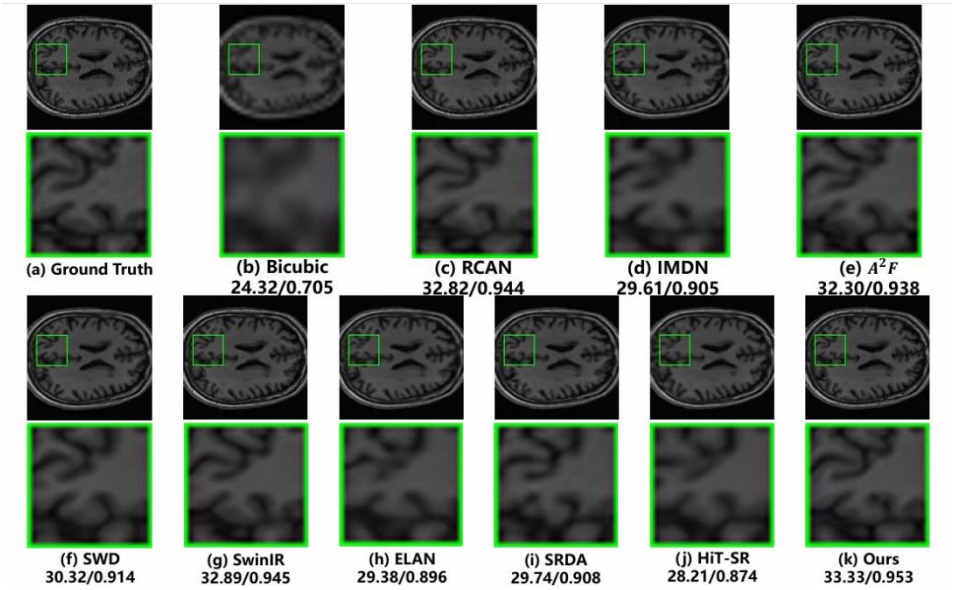

Figure 4: Conduct a qualitative comparative analysis on the ADNI-MRI data set.

5.4. Ablation Study

We performed some ablation studies to exhibit the efficacy of our engineered modules. We refer to our base network as GFE, which consists of only a global feature extraction module within a single-scale SR network. LFE is the local feature extraction module. And MA is a multi-content aggregation attention mechanism that fuses both. Finally, PMS-FN is the pyramid multi-scale feature fusion module integrates a multi-content aggregation attention mechanism.

Table 2 presents the quantitative results. The local feature extraction module, when compared to the baseline SR model, enhanced performance to a certain extent. Subsequently, by incorporating the multi-content aggregation attention mechanism, we achieved further improved performance. Furthermore, with the addition of the pyramid multi-scale feature fusion module, the performance improved further, demonstrating its effectiveness.

Figures 7 shows the visual results. From the zoomed-in view of the lines in the top left corner, we can see that contrasted to the baseline super-resolution (SR) model, the local feature extraction module refines the tissue to some extent. Then, by adding the multi-content aggregation attention mechanism, we achieve a more cohesive line structure. Additionally, after adding the pyramid

DOI: 10.54254/2755-2721/2025.KL24798

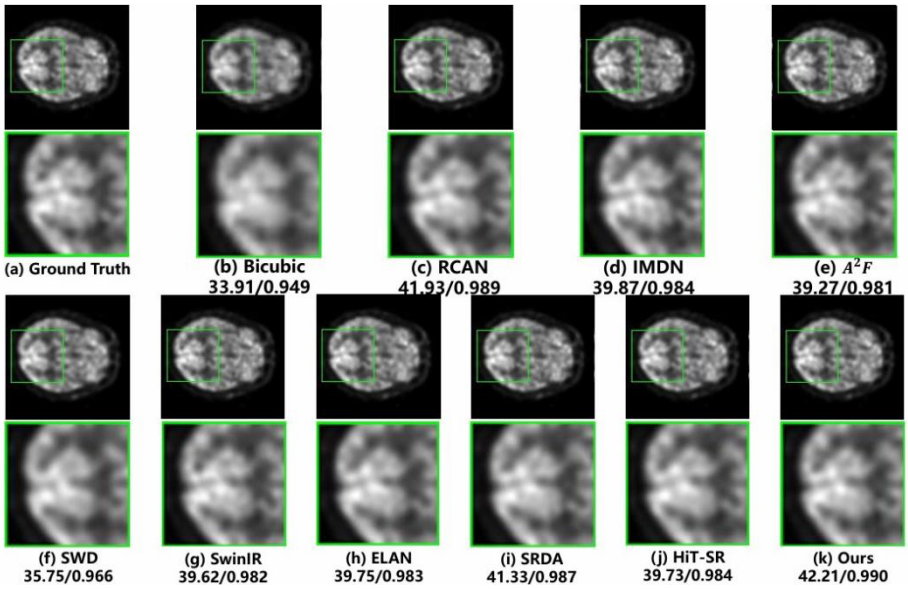

Figure 5: Conduct a qualitative comparative analysis on the ADNI-PET data set.

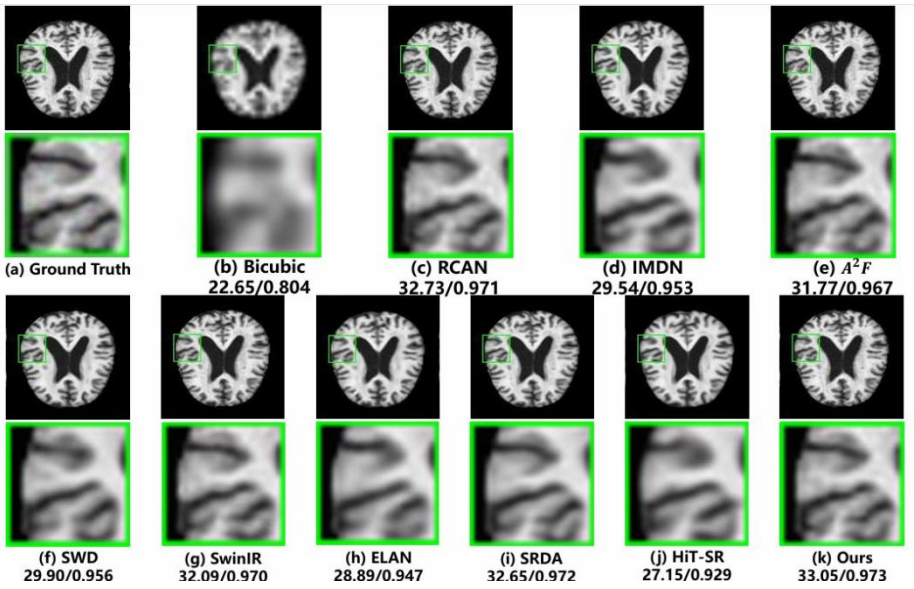

Figure 6: Conduct a qualitative comparative analysis on the Demented-MRI data set.

multi-scale feature fusion module, we obtain tissue lines that are closer to the original image, proving its effectiveness.

Table 2: Ablation Study

Methods	Demented - MRI	
	PSNR	SSIM
GFE	25.22	0.697
GFE + LFE	27.72	0.935
GFE + LFE + MA	30.67	0.948
GFE + LFE + PMS-FN	**30.86**	**0.957**

DOI: 10.54254/2755-2721/2025.KL24798

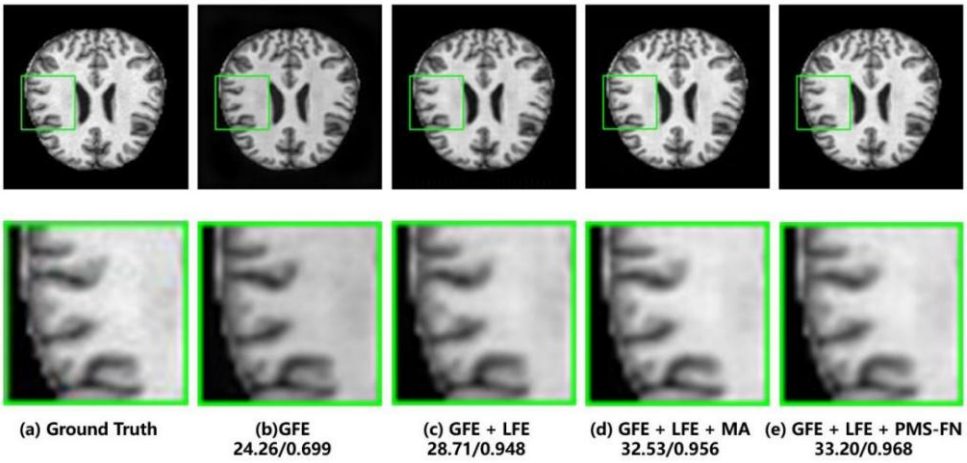

Figure 7: Ablation Quantitative Comparison on the Demented - MRI Data set

5.5. Conclusion

Here, we present a network that combines a Transformer architecture with multi-scale convolution techniques to achieve super-resolution in medical imaging. We applied a Transformer to extract global information and then introduced a local extraction convolution module to capture the local information in the representation learning. Then, we carefully designed a novel pyramid Multi-scale feature integration unit to effectively reconstruct the high-level semantical and low-level detailed information for super-resolution. At long last, we carried out performing in-depth experiments across various benchmarks medical image datasets. The results of these experiments clearly showcased the efficacy and the excellence of our suggested approach.

References

[1] John A. Kennedy, Ora Israel, Alex Frenkel, Rachel Bar-Shalom, and Haim Azhari. Super-resolution in PET imaging. IEEE Trans. Medical Imaging, 25(2):137–147, 2006.

[2] R. W. Gerchberg. Super-resolution through error energy reduction. Journal of Modern Optics, 21:709–720, 1974.

[3] Xin Jiang, Mingzhe Liu, Feixiang Zhao, Xianghe Liu, and Helen Min Zhou. A novel super-resolution ct image reconstruction via semi-supervised generative adversarial network. Neural Computing and Applications, 32:14563 – 14578, 2020.

[4] Chenyu You, Yi Zhang, Xiaoliu Zhang, Guang Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, Punam K. Saha, and Ge Wang. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). IEEE Transactions on Medical Imaging, 39:188–203, 2018.

[5] Christos Matsoukas, Johan Fredin Haslum, Magnus P Soderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?ArXiv, abs/2108.09038, 2021.

[6] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. Edter: Edge detection with transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1392–1402, 2022.

[7] Yifei Chen, Chenyan Zhang, Ben Chen, Yiyu Huang, Yifei Sun, Changmiao Wang, Xianjun Fu, Yuxing Dai, Feiwei Qin, Yong Peng, and Yu Gao. Accurate leukocyte detection based on deformable-detr and multi-level feature fusion for aiding diagnosis of blood diseases. Computers in biology and medicine, 170:107917, 2024.

[8] Mei-Juan Chen, Chin-Hui Huang, and Wen-Li Lee. A fast edge-oriented algorithm for image interpolation. Image Vis. Comput., 23:791–798, 2005.

[9] Hongyu Hou, Qunchao Jin, Guixu Zhang, and Zhi Li. Ct image quality enhancement via a dual-channel neural network with jointing denoising and super-resolution. Neurocomputing, 492:343–352, 2022.

[10] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias P. Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy J. W. Dawes, Declan P. O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert.Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging, 37:384–395, 2017.

DOI: 10.54254/2755-2721/2025.KL24798

[11] Xiang Hong, Yunlong Zan, Fenghua Weng, Weijie Tao, Qiyu Peng, and Qiu Huang. Enhancing the image quality via transferred deep residual learning of coarse pet sinograms. IEEE Transactions on Medical Imaging, 37(10):2322–2332, 2018.
[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
[13] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12294–12305, 2020.
[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929, 2020.
[15] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoyue Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. In International Joint Conference on Artificial Intelligence, 2022.
[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming–Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5718–5729, 2022.
[17] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1692–1701, 2023.
[18] Qing Xu, Wenting Duan, and Nana He. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. Computers in biology and medicine, 154:106626, 2022.
[19] Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In AAAI Conference on Artificial Intelligence, 2023.
[20] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Raymond Fu. Image super-resolution using very deep residual channel attention networks. ArXiv, abs/1807.02758, 2018.
[21] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. Proceedings of the 27th ACM International Conference on Multimedia, 2019.
[22] Xuehui Wang, Qing Wang, Yuzhi Zhao, Junchi Yan, Lei Fan, and Long Chen. Lightweight single-image super-resolution network with attentive auxiliary feature learning. In Asian Conference on Computer Vision, 2020.
[23] Zhen Chen, Xiaoqing Guo, Chen Yang, Bulat Ibragimov, and Yixuan Yuan. Joint spatial-wavelet dual-stream network for super-resolution. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 184–193, Cham, 2020. Springer International Publishing.
[24] Jingyun Liang, Jie Cao, Guolei Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 1833–1844, 2021.
[25] Xindong Zhang, Huiyu Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In European Conference on Computer Vision, 2022.
[26] Jingwei Wang, Peng Zhou, Xianjun Han, and Yanming Chen. Medical image super-resolution via diagnosis-guided attention. In 2023 IEEE International Conference on Multimedia and Expo (ICME), pages 462–467, 2023.
[27] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In European Conference on Computer Vision, 2024.