Scholar Publishing Group
International Journal of World Medicine
https://doi.org/10.38007/IJWM. 2026.070106.
ISSN 2789-7133 Vol. 7, Issue 1: 52-65

Unsupervised Super-Resolution Analysis of CT Images Based on Hierarchical Attention Transformer

Wenhui Xia[1,a], Huibo Wang[2,b,*], Tao Bian[1,c], Wenjiao Deng[1,d], Yunhe Li[1,e]

[1]*School of Electronics and Electrical Engineering, Zhaoqing University, Zhaoqing 526000, Guangdong, China*

[2]*Shenzhen Skyworth Display Technologies Co., Ltd., Shenzhen 518057, Guangdong, China*

[a]*202408540530@stu.zqu.edu.cn,* [b]*443957950@qq.com,* [c]*202408540517@stu.zqu.edu.cn,* [d]*3331372640@qq.com,* [e]*liyunhe@zqu.edu.cn*

[*]*Corresponding author*

Keywords: Super-Resolution Analysis, CT Image, Swin Transformer, Hierarchical Attention, Deep Learning

Abstract: Increase the resolution of the CT image to improve the detection accuracy of small tumours and other tumours, and thus enhance diagnostic precision. How to build a training dataset when paired high-resolution and low-resolution CT images are not available is what I will explore in this study. First of all, only low-resolution images are available; therefore, a degradation network and noise injection are employed to create domain-consistent low-resolution data. This way, the training data will be close to what you would find in real image pairs. Based on the above, I introduce a hierarchical attention transformer-based super-resolution network, DeHAT-SR. Shifted Window Self-Attention from Swin Transformer is employed to address the issue of long-range dependency. A network that adds channel attention and a shifted window cross-aggregation module can increase the receptive field and gather more pixel information for sharper super-resolution reconstruction. DeHAT-SR is a relatively stable adversarial training system that includes a generator, a discriminator and a feature extractor for super-resolution tasks. Thus, four CT images can be reconstructed at a high quality. The experiments have shown this. NIQE, BRISQUE and PIQE are no-reference image quality indicators; on 4x super-resolved CT images, DeHAT-SR outperforms the top methods such as EDSR, RCAN, ESRGAN and SwinIR. In addition, the visual comparison shows that DeHAT-SR produces images with finer details and is generally more perceptually pleasing.

1. Introduction

Among the main kinds of medical images, CT is relatively common. You can see inside the body of a person without injury to help diagnose their illness[1]. High-resolution (HR) medical images are required at this time to show fine structural details that help doctors precisely locate any diseases[2].

Therefore, the quality of the image must be high; otherwise, a low-resolution CT scan will fail to detect a disease in time.

There are now some actual deficiencies. CT systems are hardware and software that include the size of the X-ray focal spot, detector pixel size, reconstruction algorithms, etc. None of them are ideal. Even the standard CT machines today cannot achieve the ideal resolution we desire for early-stage disease detection. The New Hardware is not simple either. Commercial CT devices are technologically advanced but also have high construction and operation costs. Therefore, it is not the case that hospitals can simply add new modules to improve the resolution every few years. There is also a problem of radiation. X-rays are relatively high-risk, so clinics use a low-dose model. They reduce the tube voltage, current and scan time to decrease exposure. Reducing the dose will always result in a lower quality of the scan [3]. You will have a safer operation, but the pictures will not be very clear.That is to say, the medical applications of CT scans require high safety and accuracy. Therefore, a new kind of high-resolution image reconstruction method has been developed for the lower-resolution data obtained under the new safe protocol. This is the section that will directly affect patient care, and it will not go away anytime soon.

In the past ten years, many new directions have emerged in the field of super-resolution reconstruction for medical images. At the same time, the extent and accuracy of the research will also be extended. Yang and his team[4] started strong and released a supervised super-resolution method based on dictionary learning. Gu[5]and Osendorfer[6]took a different way; their systems used sparse sensing theory. Babcock [7] has proposed compressed sensing in the foundation for super-resolution.Deep Learning Has Been Great. Dong [8]proposed SRCNN and has made convolutional neural networks a general method for image super-resolution. Therefore, there has been some development. Both Lim's EDSR [9] and Zhang's RCAN [10] significantly improved the quality of reconstructed images by adding residual learning and channel attention mechanisms. Then GANs were released [11]. Ledig and his team [12]applied GANs to the problem of super-resolution to introduce a new image reconstruction method, SRGAN. Wang [13]was not finished yet. By modifying the structure of SRGAN, he introduced ESRGAN and used residual-in-residual dense blocks and skips batch normalisation. This way to refine the perceptual loss function by using pre-activation feature values, keep the luminance stable, and restore texture details.Based on the RRDB structure of ESRGAN, Jiang [14] and Zhang [15] have proposed dedicated super-resolution networks for CT images. Their efforts have achieved the goal of solid construction for clinical imaging.

Convolutional Neural Networks have difficulty addressing the problem of long-range dependencies in images. The local receptive fields limit what they can see at one time. Recently, Transformers have been updated by adding self-attention mechanisms. It has become a popular application of natural language processing and has been extended to computer vision. Dosovitskiy and his colleagues[16] proposed the Vision Transformer (ViT) and demonstrated that a standard Transformer model could perform well in the task of image classification. Based on the above, Liu and his team designed the Swin Transformer [17]. To reduce the high computational cost of ViT, a hierarchical structure and a shifted-window self-attention mechanism were introduced. The above innovations have made Swin Transformer a very powerful backbone network for many applications in computer vision.

Liang et al. [18] introduced the Swin Transformer into image restoration to disrupt image super-resolution and built the SwinIR network. SwinIR is better than the others because it employs residual Swin Transformer blocks that combine window-based self-attention and residual connections, and it surpasses the performance of traditional CNN-based methods in the main super-resolution benchmarks. SwinIR is not perfect, and for a long time, no one paid much attention to its deficiencies. Chen and others [19] have done so. They have investigated this in greater detail

and, through an all-encompassing study of interpretability, have come to the conclusion that SwinIR does not utilise the general pixel features fully during reconstruction. Therefore, they introduced the Hybrid Attention Transformer (HAT) to integrate channel attention and a shifted window cross-aggregation module. This combination has expanded the receptive field significantly and increased the performance of super-resolution. Others were still doing so. Based on the foundation of SwinIR, Conde et al. [20] released Swin2SR and incorporated the improvements from Swin Transformer V2 in the architecture. The end? Sharper reconstruction accuracy and improved visual quality.Finally, Chen and others[21]proposed the Dual Aggregation Transformer (DAT). DAT is a dual-aggregation method that enhances features in both the spatial and channel dimensions simultaneously; therefore, it can also compete effectively with other image super-resolution models.

Transformer-based architectures are altering the direction of natural image restoration and are now also being applied to medical imaging. Wang and others proposed CT former [22], the first pure Transformer employed for low-dose CT denoising. Chu's group is next [23]. They have developed HorusEye, a self-supervised foundation model based on the Swin Transformer encoder. This model is for X-ray tomography, and many other uses for CT image reconstruction have been developed. Japanese scholars have further advanced this with Swin2SR [24]. Apply the above super-resolution model to reconstruct lung CT images. More than a thousand people have been conducting experiments to find that the super-resolution images are better than traditional CT scans for detecting pulmonary nodules and have increased the accuracy of diagnosis. This area is likely to see some development.

The original algorithms ESRGAN, RCAN and EDSR generally start with high-resolution images and then produce low-resolution ones by performing bicubic downsampling [25] at specific scaling factors. Thus, they create their training datasets. However, this way will lose a lot of the frequency-specific structure in the original image, and when upscaling, everything will appear too smooth or even completely blurred. Transformer-based methods are good at learning global features, but most researchers have been limited to supervised learning. There is not much work yet on unsupervised CT super-resolution, and even less so without access to matched image pairs.

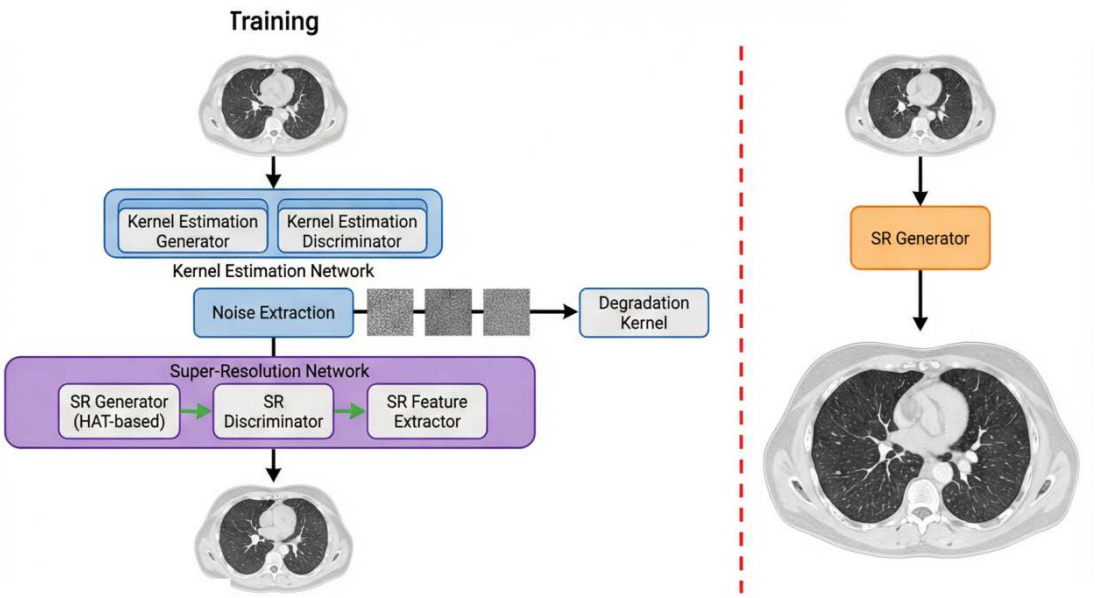

Figure 1: Architecture Diagram of the Degradation-prior Hierarchical Attention Super-Resolution Network (DeHAT-SR)

To solve the above problems directly, first of all, we built datasets of high- and low-resolution image pairs that closely matched the appearance and characteristics of real-world images. Our kernel estimation network will be used for CT image preprocessing. It explores the internal patch distribution of CT scans, extracts characteristic features, and adds noise to them. Take the original high-resolution image and degrade it to create a low-resolution version; then, by assembling pairs that match the distribution of natural images, we can do so.I_{LR} I_{HR} I_{LR}

Drawing inspiration from SwinIR [18], HAT [19], and VGG-19 [26],we craft DeHAT-SR, a hierarchical attention Transformer-based super-resolution network. It builds on Swin Transformer's window-based self-attention, integrating channel attention and shifted window cross-aggregation for dual-domain feature fusion.

The generator branch of a trained network is used in the Degradation-prior Hierarchical Attention Super-Resolution Network (DeHAT-SR) to reconstruct CT images at 4x super-resolution, and this deep learning framework is designed to improve image quality by focusing on degradation-prior information.The architecture and workflow of DeHAT-SR are depicted in Figure 1,showcasing its effectiveness in enhancing CT image resolution.

2. High-Low Resolution Image Pair Data Preprocessing

2.1 Degradation Model

The corresponding natural high-resolution (HR) CT image and low-resolution (LR) CT image can be approximated as:

$$I_{LR} = (I_{HR} * D_{Es}) \downarrow_s + N_{in} \quad (1)$$

The degradation kernel, injected noise and scaling factor all have a significant impact on the resulting high-quality super-resolution image; thus, accurate parameters for these elements are required to produce realistic high-resolution images from low-resolution ones, and the quality of the final images will be directly affected by the authenticity and consistency of these parameters.D_{Es} N_{in} s D_{Es} N_{in}

2.2 Injected Noise

Downsampling images removes high-frequency details in the images, and when noise is added subsequently, it does not look natural. It is not the kind of noise that will be generated by a real CT scan. Therefore, we take noise patches directly from the high-resolution CT images. Thus, the noise will be in line with what one expects in a real clinical environment. Now we cannot simply take any patch and call it "noise"; we set two key parameters for our noise patches: the minimum mean value and the maximum variance. Not all of these are the same. Fine-tune them for each set of images, and different situations have different noise characteristics. We are not doing so by chance. A large number of comparative experiments are conducted to find suitable values for the above parameters. It works as follows: Randomly sample images from the training set and extract patches that meet the criteria. Then we gather these patches to form a custom noise patch dataset. At the time of simulating image degradation, we take noise patches from this dataset and directly insert them according to Eq. (1) to generate noise that is suitable for each image. Thus, the noise will be in agreement with what we observe in the actual CT images. I_{CT} I_{HR} m_{min} d_{max} I_{CT} SET_{train} SET_{noise} SET_{noise}

2.3 Kernel Estimation Degradation Network.

A generative adversarial network called the Kernel Estimation Network is employed to train the degradation kernel in Equation (1), and low-resolution images that resemble real scans are produced. There is no paired high-resolution and low-resolution CT dataset for supervised learning; only the original set of CT images is available. Therefore, we train the Kernel Estimation Network in an unsupervised manner to learn the degradation kernel accurately. Each training step corresponds to one image. After training converges, we fix the generator module and use it to learn degradation kernels for new images.

Randomly sample images from different rounds during training. Thus, a range of estimated kernels has been obtained and stored in a dataset. To simulate degradation, we choose a kernel from the above set and replace it with Eq. (1) to generate images that look like actual degradation.$D_{Es}\ SET_{train}\ I_{CT}\ D_{Es}\ D_{Es}\ SET_{train}\ SET_{ker}\ D_{Es}\ SET_{ker}$

The Kernel Estimation Network takes a randomly selected source image as input. The generator branch downsamples the input image using a scaling factor to generate a degraded image. At the same time, a cropped patch is randomly extracted from the original image. The role of the discriminator is to judge the multi-scale consistency of the internal pixel patch distributions between the degraded image and the randomly cropped image; thus, it helps to evaluate the consistency of images at different scales within the network and improve the quality of the generated images.$I_{CT}\ I_{CT}\ s\ I_{de}\ I_{cr}\ I_{de}\ I_{CT}\ I_{de}\ I_{cr}$

The objective function of the Kernel Estimation Network is defined as:

$$G^{*}(I_{CT}) = \arg\min_{G}\max_{D}\{E_{patches(I_{CT})}\left[|D(I_{CT}) - 1| + \left|D\left(G(I_{CT})\right)\right|\right] + l_{K}\} \quad (2)$$

In Kernel Estimation Generative Adversarial Networks (KEGANs), the Generator and Discriminator models are used, denoted as G and D, and a loss function denoted as L is employed. The subscript indicates the degradation in kernel estimation, and thus a more comprehensive loss function formulation is achieve.$G\ D\ l_{K}\ K$

$$l_{K} = \alpha_{S} l_{S} + \alpha_{M} l_{M} + \alpha_{SQ} l_{SQ} + \alpha_{CE} l_{CE} \quad (3)$$

Here, l_{S}, l_{M}, l_{SQ}, and l_{CE} are defined as:

$$l_{S} = \left|1 - \sum_{i,j} k_{i,j}\right| \quad (4)$$

$$l_{M} = \sum_{i,j} \left|k_{i,j} \cdot m_{i,j}\right| \quad (5)$$

$$l_{SQ} = \sum_{i,j} \left|k_{i,j}\right|^{\frac{1}{2}} \quad (6)$$

$$l_{CE} = (x_{0}, y_{0}) - \frac{\sum_{i,j} k_{i,j} \cdot (i,j)}{\sum_{i,j} k_{i,j}} \quad (7)$$

Marks the parameter value for each node in the degradation kernel here. The variable v serves as an all-pass constant for the node weights and increases exponentially with the distance from the centre of. These are the indices of the centre coordinates for the kernels. The four are treated as fixed constants and are not expected to change. The subscripts indicate which pixel you are looking at horizontally and vertically. Now, if you are using images with a size of 256×256, both are integers ranging from 1 to 256.

The Kernel Estimation Generator functions as an image downsampling model with a fully linear activation design and a multi-layer linear convolutional structure to enhance the convergence speed of the network and thus avoid physically implausible optimisation solutions. After training, the degradation function layer of the generator is extracted and used as the learned degradation kernel

to improve the general quality of the downsampling process.

Kernel Estimation Discriminator determines the degree of similarity between the generator's degraded image and patches from the original image. Several stacked convolutional layers are used inside; there is no pooling, but spectral normalisation, batch normalisation and ReLU are present. Instead of one number, there is a probability heatmap output. Each pixel shows the probability that that spot originated from the real image. To train the discriminator, we use pixel-wise mean squared error between this predicted heatmap and the actual label map. Thus, the model is able to distinguish between genuine and counterfeit at a relatively local level.

3. Super-Resolution Network

This paper proposes a hierarchical attention Transformer-based generative adversarial network, termed DeHAT-SR, to accomplish 4× super-resolution reconstruction from low-resolution CT images to high-resolution counterparts.

By pre-processing the HR-LR image pairs we constructed, we have obtained training samples that are closer in appearance and feel to real images. The above are the samples for the super-resolution network. It is not the case that ESRGAN and other methods only use a convolutional generator. Our super-resolution generator is a hierarchical attention design based on Swin Transformers, which can capture global feature relationships through self-attention in a way that convolutions cannot. The form of the model is a generator-discriminator architecture. Instead of a generator and a discriminator in the usual setup, we will do something more extraordinary. A feature extractor is used to introduce perceptual loss and increase the sharpness of low-frequency components in the image. Therefore, we constructed the feature extractor based on VGG-19 [26].

The three parts of the loss function for the super-resolution network are pixel-wise [13], perceptual and adversarial, and together they aim to improve the quality of the output image by enhancing both resolution and sharpness for better visual perception.

$$l_{SR} = \alpha_x l_x + \alpha_c l_c + \alpha_a l_a \tag{8}$$

A pixel-wise loss function with constant coefficients is used to calculate the L1 distance between the generated image and the ground-truth image; thus, content loss can be determined at the pixel level without a high AIGC rate.α_x α_c α_a l_x $\hat{I}_{HR}$ I_{HR}

$$l_x = E_{I_{LR}} \| G(I_{LR}) - I_{HR} \|_1 \tag{9}$$

A Perceptual Loss Function is used to quantify the differences in content and style between two images. It is composed of two components: a content-focused feature reconstruction loss and a style reconstruction loss. These two parts are weighted by coefficients, denoted as alpha and beta. The feature map extracted from a particular convolutional layer of a perceptual feature extractor is represented by phi, and its dimensions are channels, height and width. The squared Frobenius norm is denoted as ||. Thus, both the perceived deviation of content and style in images can be expressed numerically.l_c l_f l_t λ_f λ_t $\phi_j(I)$ j I $C_j \times H_j \times W_j$ $\|\cdot\|_F^2$

$$l_c = \lambda_f l_f + \lambda_t l_t \tag{10}$$

$$l_f = \frac{1}{C_j H_j W_j} \| \phi_j\big(G(I_{LR})\big) - \phi_j(I_{HR}) \|_2^2 \tag{11}$$

$$l_t = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \| \frac{\phi_j(G(I_{LR}))_{h,w,c}}{\phi_j(G(I_{LR}))_{h,w,c'}} - \frac{\phi_j(I_{HR})_{h,w,c}}{\phi_j(I_{HR})_{h,w,c'}} \|_F^2 \tag{12}$$

The adversarial loss function l_a is used to enhance the texture details of the generated image, making it appear more realistic:

$$l_a = \sum_{n=1}^{N} - D\left(G(I_{LR})\right) \tag{13}$$

3.1 Swin Transformer-based Super-Resolution Generator

The core part of the proposed framework is the super-resolution generator; it is not the residual-in-residual dense block (RRDB) structure used in ESRGAN, but rather based on the hierarchical attention Transformer (HAT) architecture from a previous study[19]. The detailed design of the generator is shown in Figure 2, and its new way of improving image resolution is presented here.

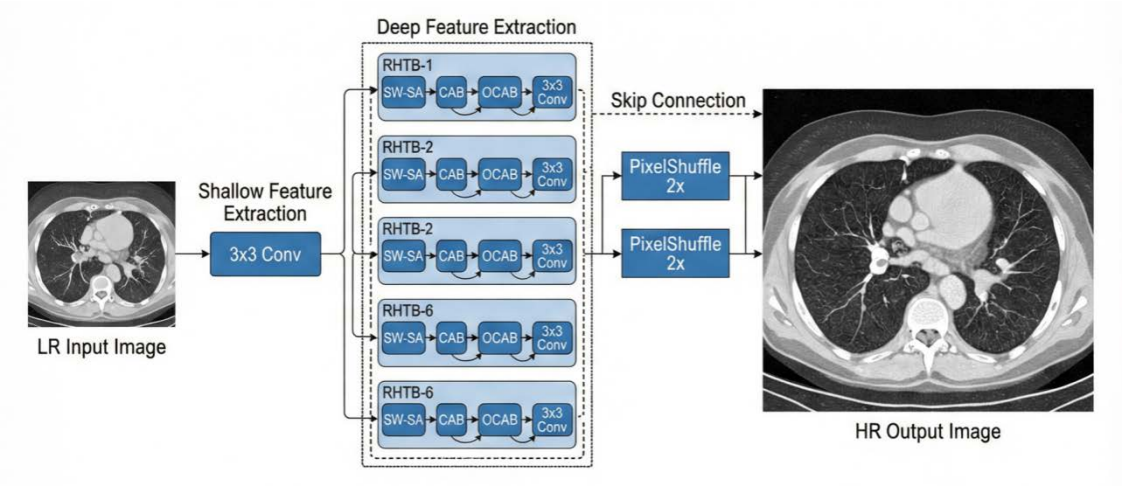

Figure 2 Structure of the super-resolution generator (HAT-SR-G)

The three components of the generator are: a shallow feature extraction module, a deep feature extraction module, and a high-quality image reconstruction module.

The centre of the above framework is a super-resolution generator. The generator in this work is based on HAT's hierarchical attention Transformer architecture [19] (as shown in Figure 2), not the RRDB structure of ESRGAN. It first has a shallow feature extraction module: only a 3×3 convolutional layer that raises the low-resolution input image into a high-dimensional feature space. Then, the deep feature extraction module will be used to add a number of Residual Hierarchical Transformer Blocks (RHTBs) sequentially. Each RHTB is a combination of several hierarchical attention aggregation sub-modules and is followed by a 3×3 convolution. To help the gradient flow more smoothly during training, residual connections are added at various points in the network.

The first innovation of this new generator is an improved hierarchical attention aggregation sub-module. Unlike SwinIR, which employs only the standard Shifted Window Self-Attention (SW-SA), our method adds both a Channel Attention Block (CAB) and an Overlapped Cross-Aggregation Block (OCAB) to each sub-module. Thus, this design can extract richer features and perform comprehensive aggregation at both the spatial and channel levels simultaneously.

Shifted Window Self-Attention (SW-SA) is a new type of self-attention that divides the input feature map into several local windows for self-attention calculations. To enhance the flow of information among these windows, regular and shifted window divisions are used in the following layers, and then self-attention is performed independently within each window to strengthen feature interaction and optimize information processing; thus, its performance as a neural network has been improved. X

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{14}$$

Where, are the query, key, and value matrices respectively, is the feature dimension, and is the relative position bias.

A Channel Attention Block (CAB) is required to enable the model to learn the importance of all channels in a feature map. Global average pooling and a Multi-Layer Perceptron (MLP) are used for recalibrating feature responses efficiently, and thus, adaptive channel attention weights are generated to improve the overall performance of the model.

$$\mathrm{CAB}(X) = X \otimes \sigma\left(W_2\delta\left(W_1\mathrm{GAP}(X)\right)\right) \tag{15}$$

Where σ and δ denote the Sigmoid and ReLU activation functions respectively, W_1 and W_2 are the weight matrices of the MLP, and GAP denotes the global average pooling operation.

An Overlapping Cross-Aggregation Block (OCAB) has been introduced to address the shortcomings of the traditional window-based self-attention mechanism; therefore, this paper aggregates feature maps from cross-shaped windows to improve inter-window communication and effectively extend the receptive field; it increases the number of pixels included in the reconstruction process and raises the overall accuracy.

The enhanced-image-reconstruction-module now employs a PixelShuffle layer to improve the quality of the image; detailed features are moved from the low-resolution domain to the high-resolution domain, and finally, a 3x3 convolutional layer produces a high-quality final image.

3.2 Super-Resolution Discriminator

PatchGAN[27] is a type of architecture that employs a uniform receptive field arrangement across all of its discriminators' convolutional layers. Consequently, the output at each layer is determined by a local patch of a fixed size in space, and thus only considers local feature details to enhance overall performance.

3.3 Super-Resolution Feature Extractor

Implement the perceptual loss function to enhance the low-frequency components of the generated images and improve their visual realism. Introduce a dedicated super-resolution feature extractor inspired by VGG-19 [26] into the generative adversarial network to boost the quality of image reconstruction significantly, thus advancing the field of image generation.

4. Experiments and Results

We have developed this way to handle the reconstruction problems of CT medical images. The QIN LUNG CT dataset [28] will be used for training and validation. It is a set of preoperative diagnostic CT scans of cancer patients, with a total of 47 cases and 3,954 images at a resolution of 512×512 pixels. The above high-resolution scans are used to construct the training and validation subsets. The H. Lee Moffitt Cancer Center[29-30] created this dataset and deleted all patient identifiers to protect medical privacy.

It is difficult to obtain real-world pairs of high- and low-resolution CT images, and public datasets lack a sufficient number of good matched samples; therefore, we have built our own test set. We collected 60 pairs, and for each, there was a 512×512 high-resolution CT scan and a corresponding 128×128 low-resolution version. The test set will be used to determine how effective the super-resolution methods are according to the PSNR and SSIM criteria. In our paired evaluation, we take the 128×128 images, upscale them by a factor of 4, and then directly compare the results with the ground-truth 512×512 images. DS_{TE}

Therefore, this paper points out the problem that high-quality reference images are not always available and carries out no-reference image quality assessment by enlarging the original 512×512 test images by a factor of 4 to evaluate the efficiency of the proposed method in a practical setting. Three well-known no-reference IQA indices - NIQE [31], BRISQUE [32], and PIQE [33] - are employed to comprehensively assess the perceptual quality of the enhanced images, and thus we hope to demonstrate that this method can improve image quality practically without needing high-resolution reference images.

The MATLAB functions NIQE, BRISQUE and PIQE produce evaluation scores between 0 and 100; a lower score indicates higher perceived quality, and thus these scores are used to assess image quality.

In the initial stage of data preparation before the actual training process, various methods such as RCAN, EDSR and ESRGAN are used to generate low-resolution images for both the training and validation sets via the bicubic technique. At the same time, SwinIR and DeHAT-SR employ kernel estimation networks and noise injection to create their own low-resolution images. Unlike the above two methods, the bicubic approach does not require training and simply upsamples the images through interpolation; thus, no extra data processing steps are needed, and the entire image enhancement process is simplified.DS_{TR} DS_{VA}

A typical configuration of the test computer is presented below: Intel i7-12700K CPU, 64GB of memory (RAM), and an NVIDIA GeForce RTX 3090 graphics card. Matlab's internal interpolation function is used for upsampling in the bicubic method. The other methods are DeHAT-SR, SwinIR, RCAN, EDSR, and ESRGAN, and they run on PyTorch. All of them use modules from the "xinntao/BasicSR" project on GitHub [34]. For DeHAT-SR, we have also added the Transformer module from the "XPixelGroup/HAT" repository [19].

The DeHAT-SR method establishes its kernel estimation and super-resolution networks as before. Kernel estimation network is used to assign loss function coefficients to k_i, b_i, and v_i. Both the generator and the discriminator use the ADAM optimizer, and their hyperparameters are the same as above. Training for 3000 epochs, starting with a learning rate of 0.002, and reducing it by a factor of ten every 750 iterations.The loss coefficients for the super-resolution network are 1, 2, and 3. Generator and Discriminator use the same ADAM optimizer and hyperparameters: = and. Training is carried out for 120,000 iterations here, and a 0.0001 learning rate is used at the start, then reduced by cosine annealing.The window size of the hierarchical attention Transformer in the super-resolution generator is 8×8, it has 6 attention heads, 180 feature dimensions, and 6 cascaded RHTB blocks. Each RHTB has 6 hierarchical attention aggregation sub-modules. This Layout Drives the Network's Performance and Structure.

In the noise extraction process of image pair preprocessing, the parameters for the minimum mean and maximum variance of noise blocks are set to and, respectively.

RCAN is the same as the one in [10]. The generator uses the ADAM optimizer with the same beta values and runs for 300,000 epochs at a constant learning rate of 0.0001. EDSR follows [9], also uses the ADAM optimizer for the generator, and has the same betas, but trains for 400,000 epochs at a learning rate of 0.0001. ESRGAN is based on [13]. Both the generator and discriminator use the ADAM optimizer, matching betas again, training for 300,000 epochs at 0.0001. SwinIR takes its configuration from [18]. Its generator is run with the ADAM optimizer and the same beta values as the others, but here, the number of training epochs is 250,000 and the starting learning rate is 0.0002. SwinIR is also used for learning rate scheduling by cosine annealing. $\beta_1 = 0.9\ \beta_2 = 0.99$ $\beta_1 = 0.9\ \beta_2 = 0.99$ $\beta_1 = 0.9\ \beta_2 = 0.99$ $\beta_1 = 0.9\ \beta_2 = 0.99$

Most super-resolution tasks in actual clinical practice aim to increase the resolution of high-resolution CT scans, but there is a problem: there is no ground-truth image for comparison. Therefore, we cannot directly judge the output by comparing it with reference images. Instead, only

no-reference image quality assessment (IQA) indicators can be used, such as NIQE [31], BRISQUE [32], and PIQE [33]. Except for the basic, untrainable biCubic baseline, all the comparison models—RCAN, EDSR, ESRGAN, SwinIR and our DeHAT-SR—were trained independently with the same parameter settings mentioned earlier. Training is finished; now we will use all these models to upscale the 512×512 images in the test set to 2048×2048. Table 1 shows the lowest and mean scores for NIQE, BRISQUE and PIQE, and how each method performed.

Figures 3, 4, and 5 present the distributions of the no-reference image quality assessment metrics for all generated images.

All the indices of NIQE, BRISQUE and PIQE were calculated in MATLAB. All the scores are positive numbers in the range [0, 100], and a lower value indicates better image quality. As shown in Table 1 and Figure 3, according to NIQE results, the DeHAT-SR method achieved the highest score in assessing how natural the spatial structure was in most of the test images. SwinIR has developed some independently, but RCAN and EDSR lag slightly behind SwinIR. As shown in Table 1 and Figure 4, BRISQUE performs well; DeHAT-SR has once again led the field and surpassed all other methods in most of the test images. SwinIR comes close for a few samples, but ESRGAN remains steady and does not lead the other deep-learning methods. Now, if you look at the PIQE metric (Table 1 and Figure 5) for image quality based on local features, all the above methods such as biCubic, EDSR, and RCAN are in the same range. ESRGAN is better than the others, SwinIR is slightly higher than ESRGAN, and DeHAT-SR has the highest PIQE score overall. Therefore, generally speaking, all DeHAT-SRs have shown good perceptual qualities.

Table 1. Minimum and average values of metrics including NIQE, BRISQUE, and PIQE.

Method	NIQE_MIN	NIQE_MEAN	BRISQUE_MIN	BRISQUE_MEAN	PIQE_MIN	PIQE_MEAN
BiCubic	6.28	7.61	47.92	49.73	82.94	87.05
EDSR	4.47	5.42	42.63	46.85	76.48	82.17
RCAN	3.35	4.71	41.72	45.83	71.83	78.96
ESRGAN	5.43	7.86	35.61	48.78	32.58	50.41
SwinIR	2.53	3.68	32.26	37.15	24.18	31.28
DeHAT-SR	0.87	1.91	25.08	30.36	14.85	19.21

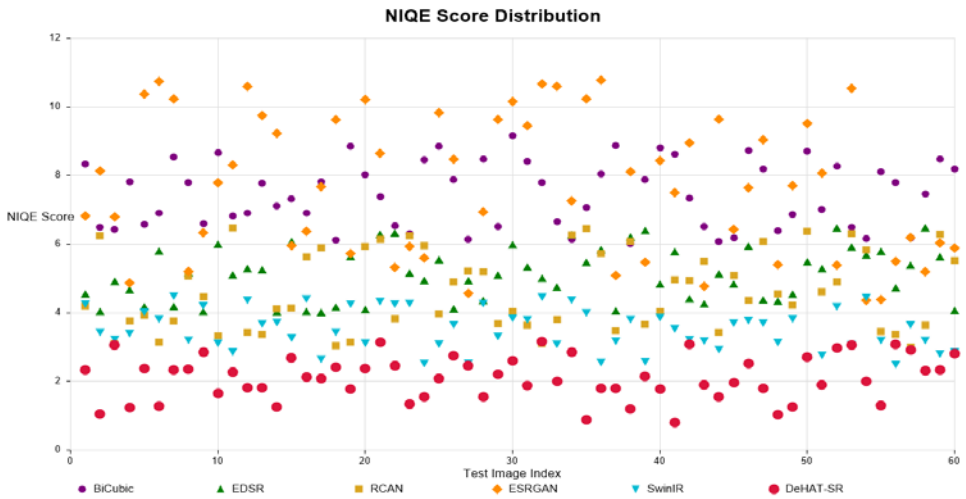

Figure 3. Distribution of NIQE evaluation metrics for generated images.

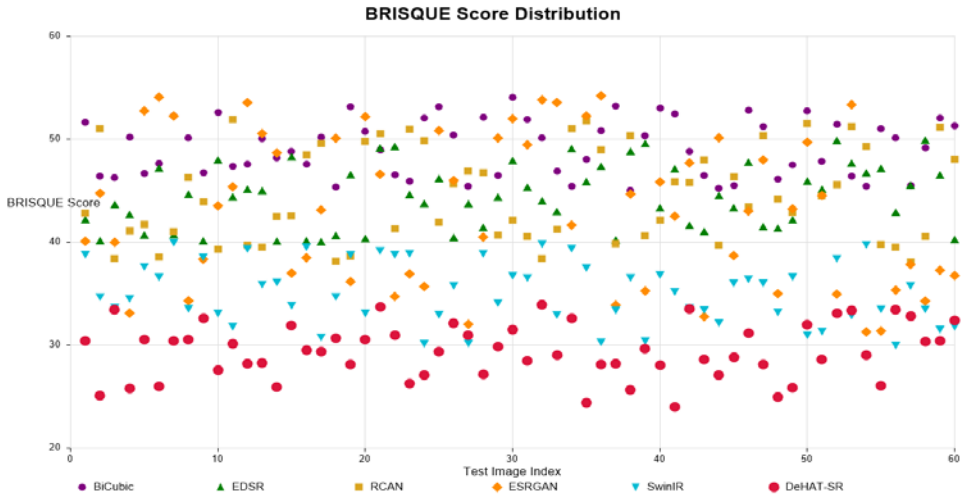

Figure 4. Distribution of BRISQUE evaluation metrics for generated images.

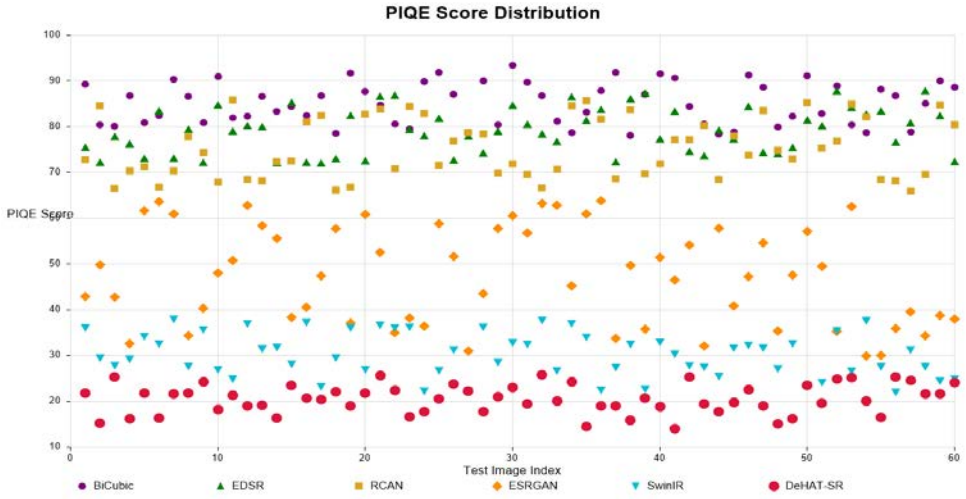

Figure 5. Distribution of PIQE evaluation metrics for generated images.

DeHAT-SR performs very well in the general no-reference IQA test and has good perceptual qualities. It is a hierarchical attention Transformer structure that effectively addresses spatial and channel feature interactions, enables the network to use long-range contextual information for super-resolution reconstruction, and thus produces high-resolution CT images with better texture details and precise edge structures compared with other methods overall.

Place the CT images side by side, and then DeHAT-SR is shown. As shown above, the division of tissue and the richness of texture in this image are both more distinct and more lifelike. ESRGAN can enhance some of the detailed textures, but it is prone to generating odd artifacts and tends to be overly sharp. SwinIR has relatively smooth images and does not contain the high-frequency details that DeHAT-SR can restore. RCAN and EDSR have rather blurry results, and most of the fine textures have been lost.

5. Conclusion

This paper studies computed tomography (CT) and investigates how to reconstruct sharp, high-resolution images from the relatively low-resolution output of CT scanners. Rather than using the usual ways, such as convolutional neural networks (e.g., RCAN, EDSR, ESRGAN) or even the Transformer-based approach of SwinIR, the authors have taken a different path. They build high- and low-resolution image pairs using a custom pipeline with a degradation network and added noise. The Goal? Generate low-resolution images that actually look and behave like real CT scans, and thus the training data will have a natural distribution of images rather than just artificial downgrades. Based on the heavyweights of deep learning such as SwinIR, HAT and VGG-19, the authors present the Degradation-aware Hierarchical Attention Transformer Super-Resolution network, or DeHAT-SR. This model takes the window self-attention backbone from Swin Transformer and adds some extra modules, namely a channel attention mechanism and a shifted window cross-aggregation block. Combine spatial and channel-level feature aggregation to obtain more detailed information at both levels. Result: A relatively large effective receptive field and images that can capture the detailed texture and edge information required for CT analysis.

No reference-free image quality assessment metrics were used in the quantitative experiments, and DeHAT-SR outperformed both the traditional bicubic method and some of the latest ones, such as RCAN, EDSR, ESRGAN and SwinIR. DeHAT-SR has higher scores for statistical regularity in the spatial-domain natural scene model, scene statistics based on locally normalised luminance coefficients, and local feature-based image quality. As shown in the figure above, the reconstruction results of DeHAT-SR are relatively smooth and aesthetically pleasing.

Train the network by generating its own training data through self-degradation of the existing high-resolution CT images. As a result, DeHAT-SR has four times the super-resolution reconstruction accuracy for the above original high-resolution CT scans. So what? Images that are relatively clearer and can aid in diagnosis. It will thus be of some use for DeHAT-SR.

There will be many other subjects in the future. Next, we will extend this way to super-resolution reconstruction of 3D volumetric CT images. Three-Dimensional attention mechanisms can be employed to address the problem of complexity in volumetric medical data. The other is self-supervised pre-training. Tap into large pools of unlabelled CT data to improve the generalisation capability of the model. Finally, I will use DeHAT-SR to conduct super-resolution experiments on several kinds of medical images. It will be shown that this way is effective.

Funding

This work was supported by the Fund of Natural Science Foundation of Guangdong Province of China (Grant No.2025A1515010733), the General University Key Field Special Project of Guangdong Province, China (Grant No. 2024ZDZX1004).

References

[1] *T. M. Buzug, "Computed tomography," in Springer handbook of medical technology: Springer, 2011, pp. 311-342.*

[2] *A. Hassan, S. A. Nazir, and H. Alkadhi, "Technical challenges of coronary CT angiography: today and tomorrow," European journal of radiology, vol. 79, no. 2, pp. 161-171, 2011.*

[3] *H. M. Marshall, R. V. Bowman, I. A. Yang, K. M. Fong, and C. D. Berg, "Screening for lung cancer with low-dose computed tomography: a review of current status," Journal of thoracic disease, vol. 5, no. Suppl 5, p. S524, 2013.*

[4] *J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in 2008 IEEE conference on computer vision and pattern recognition, 2008: IEEE, pp. 1-8.*

[5] *S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1823-1831.*

[6] *C. Osendorfer, H. Soyer, and P. Van Der Smagt, "Image super-resolution with fast approximate convolutional sparse coding," in International Conference on Neural Information Processing, 2014: Springer, pp. 250-257.*

[7] *H. P. Babcock, J. R. Moffitt, Y. Cao, and X. Zhuang, "Fast compressed sensing analysis for super-resolution imaging using L1-homotopy," Optics express, vol. 21, no. 23, pp. 28583-28596, 2013.*

[8] *C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 2, pp. 295-307, 2015.*

[9] *B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136-144.*

[10] *Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286-301.*

[11] *I. J. Goodfellow et al., "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.*

[12] *C. Ledig 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, 2017, pp. 4681-4690.*

[13] *X. Wang et al., "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0-0.*

[14] *X. Jiang, Y. Xu, P. Wei, and Z. Zhou, "CT Image Super Resolution Based On Improved SRGAN," in 2020 5th International Conference on Computer and Communication Systems (ICCCS), 15-18 May 2020, pp. 363-367.*

[15] *X. Zhang, C. Feng, A. Wang, L. Yang, and Y. Hao, "CT super-resolution using multiple dense residual block based GAN," Signal, Image and Video Processing, vol. 15, no. 4, pp. 725-733, 2021.*

[16] *A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.*

[17] *Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012-10022.*

[18] *J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2021, pp. 1833-1844.*

[19] *X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22367-22377.*

[20] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration," in Proceedings of the European Conference on Computer Vision Workshops, 2022.

[21] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12202-12211.

[22] F. Wang, H. Li, R. Zhang, et al., "CTformer: Convolution-free Token2Token dilated vision transformer for low-dose CT denoising," Physics in Medicine & Biology, vol. 68, 2023.

[23] Y. Chu, L. Zhou, G. Luo, et al., "HorusEye: A self-supervised foundation model for generalizable X-ray tomography restoration," Nature Computational Science, 2026.

[24] Comparison of conventional 512-matrix CT images with Swin2SR-based 2048-matrix images in the visualization and diagnosis of lung nodules, Japanese Journal of Radiology, 2026.

[25] R. Keys, "Cubic convolution interpolation for digital image processing," IEEE transactions on acoustics, speech, and signal processing, vol. 29, no. 6, pp. 1153-1160, 1981.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[27] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," arXiv preprint arXiv:1803.07422, 2018.

[28] K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," Journal of digital imaging, vol. 26, no. 6, pp. 1045-1057, 2013.

[29] D. Goldgof et al., Data From QIN_LUNG_CT, doi: https://doi.org/10.7937/K9/TCIA.2015.NPGZYZBZ.

[30] J. Kalpathy-Cramer, S. Napel, D. Goldgof, and B. Zhao, QIN multi-site collection of Lung CT data with Nodule Segmentations,

[31] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," IEEE Signal processing letters, vol. 20, no. 3, pp. 209-212, 2012.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Transactions on image processing, vol. 21, no. 12, pp. 4695-4708, 2012.

[33] N. Venkatanath, D. Praneesh, M. Lamb, B. J. Philip, and S. S. Iyengar, "Blind image quality evaluation using perception based local features," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 2207-2211.

[34] X. Wang et al., "BasicSR: Open source image and video restoration toolbox," https://github.com/XPixelGroup/BasicSR, 2022.