TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 16/22 pp197–206
DOI: 10.26599/TST.2022.9010071
Volume 29, Number 1, February 2024

Transformer and GAN-Based Super-Resolution Reconstruction Network for Medical Images

Weizhi Du and Shihao Tian*

Abstract: Super-resolution reconstruction in medical imaging has become more demanding due to the necessity of obtaining high-quality images with minimal radiation dose, such as in low-field magnetic resonance imaging (MRI). However, image super-resolution reconstruction remains a difficult task because of the complexity and high textual requirements for diagnosis purpose. In this paper, we offer a deep learning based strategy for reconstructing medical images from low resolutions utilizing Transformer and generative adversarial networks (T-GANs). The integrated system can extract more precise texture information and focus more on important locations through global image matching after successfully inserting Transformer into the generative adversarial network for picture reconstruction. Furthermore, we weighted the combination of content loss, adversarial loss, and adversarial feature loss as the final multi-task loss function during the training of our proposed model T-GAN. In comparison to established measures like peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), our suggested T-GAN achieves optimal performance and recovers more texture features in super-resolution reconstruction of MRI scanned images of the knees and belly.

Key words: super-resolution; image reconstruction; Transformer; generative adversarial network (GAN)

1 Introduction

Computers have fueled our reliance on images. From the first X-rays of a tumor to the latest magnetic resonance imaging (MRI) scans, images have become integral to the practice of every field in medicine. The process of image acquisition is affected, and often limited, by many aspects, such as the equipment, environment, and cost. For instance, to reduce the radiation exposed on human body, computed tomography (CT) is required to decrease its beam's energy, resulting in scanned images with lower spatial resolution. In the medical diagnostic process, low-quality images can affect the pathological assessment by clinical experts and auxiliary computers, among others[1–4]. Calcification, for example, is a common symptom of most breast cancers, but calcification is small and difficult to detect. As a result, the low intensity variation between pathological tissue and healthy areas makes the diagnostic process cumbersome. In the diagnostic retinal images of fundus, many lesions cover extremely tiny areas and can be shown as microaneurysms or hemorrhages. Also, there are parts that may not be clearly visible such as soft exudates, certain neointima formation, etc.[5] Therefore, super-resolution reconstruction for medical images has become an essential role in clinical applications.

There were two main types of image super-resolution reconstruction techniques: single image super-resolution (SISR), where a high-resolution image was acquired from a single low-resolution image, and reference-based image super-resolution (RefSR), where a high-resolution image was synthesized from multiple low-

• Weizhi Du is with the Arts & Sciences College, Washington University in St. Louis, St. Louis, MO 63130, USA. E-mail: d.weizhi@wustl.edu.

• Shihao Tian is with the Department of Electric and Computing Engineering, Cornell University, Ithaca, NY 14850, USA. E-mail: st689@cornell.edu.

∗ To whom correspondence should be addressed.

Manuscript received: 2022-10-17; revised: 2022-12-21; accepted: 2022-12-29

resolution images. Among them, the goal of SISR often required optimizing the mean square error between high-resolution (HR) and SR pixels, however, the use of mean put error often led to edge blurring due to the uncomfortable nature of super-resolution (illposed). The main reason was that the high-resolution texture of a single image was often over corrupted and a large amount of information was lost, leading to unrecoverable textures[6]. While generating a large number of adversarial samples from images based on generative adversarial networks could alleviate such problems[7], the resulting hallucinations and artifacts posed a greater challenge to image super-resolution tasks.

RefSR had been shown to be promising in providing reference (Ref) images with similar content to the lowresolution (LR) input to recover high-resolution (HR) image details[8]. A large number of RefSR methods had produced visually more pleasing results compared to SISR methods. Currently, RefSR is mainly used to make full use of the Ref image information by methods such as image aligning and "patch matching". References [9–11] aligned LR and Ref images from different perspectives.

In Ref. [9], landmark aligned LR and Ref by a global registration while minimizing energy; In Ref. [10], LR and Ref images needed to be pre-aligned first, however, non-uniform warping operation was used to enhance Ref images by matching LR and Ref feature maps to obtain super resolution; In Ref. [11], the method used optical flow to align LR and Ref pictures at different scales and connected them to the decoder's relevant layers. However, the quality of the alignment between LR and Ref had a significant impact on the performance of these approaches. In addition, alignment methods such as optical flow required a large computational cost, making it difficult to be popularized in practical. On the other hand, Refs. [6, 8, 12, 13] used a "patch matching" approach to search for suitable reference information in the Ref image to complement the information and thus obtained super resolution. In Ref. [12], the gradient features in the downsampled Ref were searched to match the LR patch; in Ref. [8], the features in the CNN were used instead of gradient features to match the patch of Ref and LR, while the LR image was expanded using the SISR method; In Ref. [13], features in VGG were used to match the patch of Ref and LR, and super-resolution was obtained by swapping similar texture features. In Ref. [6], a texture transformer network was used to feature-match Ref and LR, where the low-resolution (LR) and reference (Ref) images were represented as queries and keywords in the Transformer, respectively. This setup allowed the LR and the Ref images to learn features together, i.e., deep feature correspondences that could improve accuracy in texture can be detected through an attention mechanism. However, when the reference image was less sharp, the quality of the RefSR might receive a devastating impact, resulting in impaired performance of the algorithm.

In this paper, our objective is to unleash the potential of RefSR by generating 71 reference images with more texture details through generative adversarial networks, and to discover deep feature correspondences by using the Transformer framework to perform joint feature learning between LR and reference (Ref) images. The comparing result could be found in Fig. 1, which

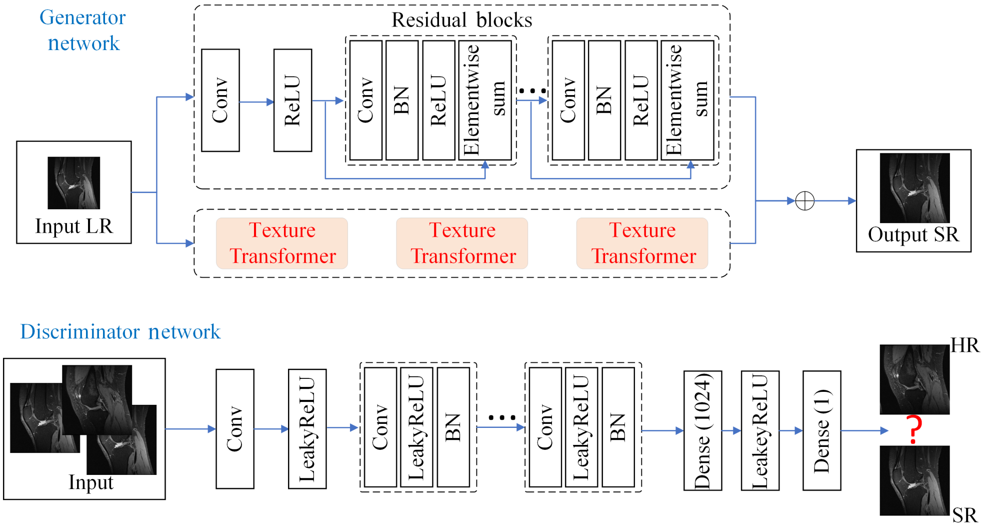


Fig. 1 Schematic diagram of the model framework.

demonstrate the effectiveness of our method on medical images. The main contributions of this paper are as follows:

- Propose a generative adversarial network (GAN) framework to recover the detailed information of the original photo from the severely downsampled (low-resolution) image to obtain a high-resolution image using the powerful generative power of GAN networks.
- Introduce Transformer architecture to extract learnable texture features.

2 Model Description

In this section, we would review previous classical algorithms for building single image super-resolution (SISR) for GAN-based framework in a single image and classical methods for RefSR in the Transformer framework for reference-based image super-resolution, which are the most relevant to our work.

2.1 GANs in image reconstruction

In image super-resolution reconstruction tasks, generative adversarial networks (GANs)[7] had emerged as an effective method for enhancing the perceptual quality of upsampled images[7, 14–16]. GANs were an effective min-max two-player game[7]. The generator G captured the data distribution, while the discriminator D continuously distinguishes whether the samples were from the training dataset or not. GANs could produce more aesthetically attractive images without supervised input using this powerful method. However, due to their intrinsic instability, the initial GANs were difficult to train. Wasserstein GANs (WGANs) used weight clipping to ensure that D is in the space of 1-Lipschit[17]; the improved training of Wasserstein GANs (WGANGP)[18] could use gradient penalties to encourage D to learn smoother decision boundaries; and Gulrajani et al.[19] proposed a weighted combination of WGANs and WGAN-GP loss terms to form a complex loss function that facilitates the model to generate. However, the above GANs were only applied on SISR while limited to image datasets with single-scale upsampling at relatively low target resolution[20]. A multi-scale GAN-enhanced SISR approach was proposed in Ref. [8], which was progressive in both architecture and training, similar to what was done in course learning, simulating the learning process from easy to hard; Lim et al.[21] proposed a framework for recovering its fine texture details when super-resolution at magnification factors. This framework proposed a perceptual loss function, which consisted of an adversarial loss and a content loss. The calculation of adversarial loss allowed the reconstructed image to be pushed towards the natural image stream shape, while constructing a discriminator network for distinguishing the super-resolution image from the original photo-realistic image.

The traditional reference learning based image superresolution reconstruction model (RefSR) took the highresolution image as the reference (Ref), so that the relevant texture was transferred to the low-resolution (LR) image. Currently, RefSR mainly made full use of the image information of Ref by methods such as image aligning and “patch matching”. However, most of the traditional methods fed all swapped features equally into the main network, neglecting to transfer high-resolution textures from the reference image using attention mechanisms, thus limiting the application of these methods in challenging situations[6, 8]. Following that, Ref. [6] proposed a new texture transformation network for image super-resolution (TTSR), in which the LR and Ref pictures were represented as queries and keywords in the transformation, respectively. A learnable texture extractor for deep neural networks (DNNs), a relevance embedding module, a hard-attention module for texture transfer, and a soft-attention module for texture synthesis were all part of TTSR, which was tuned for the image production task. This approach supported cooperative feature learning across LR and Ref pictures, allowing attention to uncover deep feature correspondences and produce accurate texture features. The suggested texture converter could also be stacked in a cross-scale manner, allowing texture recovery at multiple magnification levels (e.g., from 1× to 4×). The method, however, relied on a high-resolution image as a reference (Ref), while in practice, a huge number of low-resolution images were frequently obtained. Therefore, we used GANs to enhance the quality of low-resolution images from the perspective of low-resolution images as a reference, used the enhanced images for information complementation, and used the attention mechanism to transfer the effective features of different images to LR images to achieve RefSR reconstruction in complex environments.

2.2 Deep-learning models based on Transformer and GAN

In fact, SISR aims to learn the non-linear mapping relations between LR and HR images. In general, this non-linear mapping can be expressed as

$$y = \phi(k \times x + n) \tag{1}$$

where ϕ is the non-linear compression operator, k represents the convolution operation, n represents the random noise, y represents the degenerated LR image, and x is real HR image. In general, Eq. (1) can be simplified as

$$y = Dx \tag{2}$$

where D is the degeneracy matrix representing the down-sampling operation. Since the conditions for the discomfort inverse issue expressed in single image superresolution are not sufficient, x can not be recovered in the simple way that

$$x = D^{-1}y \tag{3}$$

Fortunately, deep learning based models have made a huge success in image processing fields. Many researchers have applied these deep learning models to reconstruct HR images from LR images, which actually learns an implicit mapping between LR and HR.

$$\hat{x} = F(y) \tag{4}$$

in which $\hat{x}$ represents the reconstructed high resolution image corresponding to the ground-truth image x. Technicality, deep neural network models minimize the optimization objective mainly by training a network F as

$$\frac{1}{N}\sum_{i=1}^{N}(F(y_i) - x_i)^2 \tag{5}$$

where N is the number of training samples. In general, this type of deep learning model can be represented as

$$\hat{x} = F_d(\cdots F_3(F_2(F_1(y)))) \tag{6}$$

where d denotes the number of layers of the deep network (number of convolutional layers).

2.3 Proposed framework

Rather than just increasing the network depth, the key goal is to increase the performance of SISR neural networks by selecting optimal internal mechanisms. The generative network and the adversarial network are the two key components of our proposed model, as depicted in Fig. 1. The residual learning channel and the texture Transformer channel are two elements of the generative network.

Eventually, we intend to train a generative function G that estimates its corresponding HR image from a given LR input image. To achieve this, we propose a generative network consisting of two channels, residual learning and texture Transformer. Here we use θ_G to denote all parameters of the generative network as well as the bias term that is learned by optimizing a particular SR reconstruction loss l^{SR}. Specifically, for a given training HR image I_n^{HR} and its corresponding LR image I_n^{LR}, the objective of the generative network is

$$\hat{\theta}_G = \arg\min \frac{1}{N} l^{\mathrm{SR}}(G_\theta(I_n^{\mathrm{LR}}), I_n^{\mathrm{HR}}) \tag{7}$$

Multiple residual learning blocks and deconvolution blocks make up the residual learning channel (as shown in Fig. 1). Because of their success in image classification, convolutional operations are now frequently utilized in deep learning, and several studies have transferred CNNs to SISR. These CNN-based SR techniques, on the other hand, rarely consider whether convolutional processes are appropriate for the SISR mechanism. The majority of them just apply CNN models to SISR from image classification tasks. The main objective of SISR is to figure out how LR and HR images are related. For the mapping relationship between LR and HR images, it can be represented by a simple linear degenerate model as follows:

$$y = x \times k \tag{8}$$

The convolution theorem states that spatial convolution can be converted to frequency-domain multiplication.

$$\mathcal{F}(y) = \mathcal{F}(x) \cdot \mathcal{F}(k) \tag{9}$$

where $\mathcal{F}(\cdot)$ is the Fourier transform and $\cdot$ is the corresponding element multiplication. Thus, in the Fourier domain, x can be expressed as

$$x = \mathcal{F}^{-1}(\mathcal{F}(y)/\mathcal{F}(k)) = \mathcal{F}^{-1}(1/\mathcal{F}(k)) * y \tag{10}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform and $*$ denotes the convolution operation. Thus, the true HR image can be recovered from the low-resolution image y by a pseudo-inverse calculation, i.e.,

$$x = k\dagger * y,$$

where $\dagger *$ denotes the deconvolution operation.

Usually, the deconvolution kernel $k\dagger$ is hard to obtain. Therefore, we construct multiple residual learning blocks and a deconvolution block to implement the deconvolution operation. Specifically, we use a convolution kernel with a small size 3×3 and 64 feature mappings as the convolution layer followed by a batch normalization layer, while employing the ReLU function as the activation function. A residual learning mechanism is introduced (constant mapping) in order to avoid information loss and also to eliminate the gradient disappearance and gradient explosion phenomena. Finally we use the deconvolution layer (step size $= 0.5$) proposed by Zheng et al.[13] to improve the resolution of the input image.

For the texture Transformer channel, similar to the setup in Ref. [6] (shown in Fig. 2), LR, LR↑, and Ref

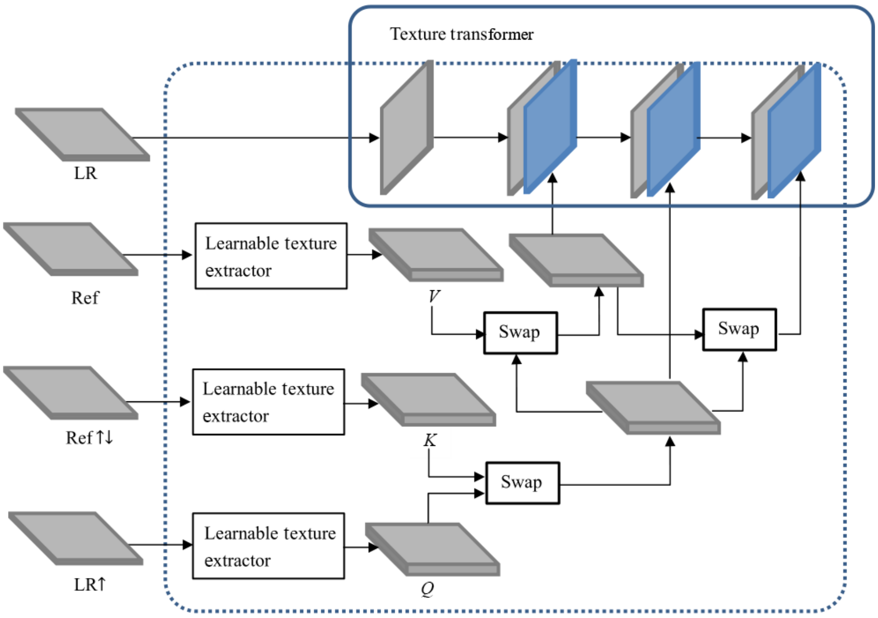

Texture transformer
LR
Ref
Learnable texture extractor
V
Swap
Swap
Ref↑↓
Learnable texture extractor
K
Swap
LR↑
Learnable texture extractor
Q

Fig. 2 Schematic diagram of the texture Transformer strategy.

denote the input image, the 4× double-triple upsampling input image, and the reference image, respectively. We apply the double triple downsampling and upsampling in turn, using the same factor 4× on Ref to obtain Ref ↑↓, with the domain consistent with LR↑. The texture converter accepts the LR features generated by Ref, Ref↑↓, and LR↑ trunk and outputs a synthetic feature map which is further used to generate HR predictions. The texture converter consists of four components: a learnable texture extractor (LTE), a correlation embedding module (CEM), a hard attention module (HA) for feature transfer, and a soft attention module (SA) for feature synthesis.

LTE mainly uses an end-to-end model to train the learning parameters such that the images of LR and Ref are able to perform joint feature learning and therefore capture more accurate texture features. LTE mainly extracts the texture features of the following three images and notates them as Q (query), K (key), and V (value): $Q = \text{LTE}\,(\text{LR}\uparrow)$, $K = \text{LTE}\,(\text{Ref}\downarrow\uparrow)$, and $V = \text{LTE}\,(\text{Ref})$, where LTE () denotes the output of the learnable texture extractor. After extracting the texture features, the RE establishes the matching relationship between LR and Ref images by estimating the similarity between Q and K. First, Q and K are expanded into a number of patches (patches), which are used to compute normalized inner products to obtain the correlation between each patch; similarly, HA transfers features for the most relevant positions in each Q and V. As a technique to fully merge LR and Ref related information, SA employs a soft attention mechanism in which relevant texture transfers are amplified and less relevant texture transfers are avoided. In conclusion, the texture converter can effectively convert key HR texture characteristics in the reference image to LR texture features, allowing for more accurate texture production.

2.4 Loss function

The perceptual loss function l^{SR} definition guides the optimization direction of the generative network and is critical to the performance of the model. We use mean squared error (MSE) to l^{SR} modeling and express the perceptual loss as a weighted sum of content loss and adversarial loss components using features extracted from the texture Transformer channel, as follows:

$$l^{\text{SR}} = l(X^{\text{SR}}) + 10^{-3} l(\text{Gen}^{\text{SR}}) \tag{11}$$

where $l(X^{\text{SR}})$ is content loss, and $l(\text{Gen}^{\text{SR}})$ is the adversarial loss.

2.4.1 Content loss

Traditional content loss is often based on pixel-wise MSE loss, e.g.,

$$l_{\text{MSE}}^{\text{SR}} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{\text{HR}} - G_{\theta_G}(I^{\text{LR}})_{x,y})^2 \tag{12}$$

where W and H refer to the width and height of the input image, respectively.

However this loss tends to make the model ignore the high frequency content information during training, making the solution to the problem of overly smooth textures (OST) not ideal. Instead of relying on pixel loss, our proposed model is based on the difference between the features extracted by the texture Transformer channel, and then we define the texture-based Transformer loss as the difference between the reconstructed image $G_{\theta_G}(I^{\text{LR}})$and the reference image I^{HR} as the Euclidean distance between the feature representations of the reconstructed image and the reference image.

$$l^{\text{SR}}_{\text{VGG}/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{\text{HR}})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{\text{LR}}))_{x,y})^2 \tag{13}$$

2.4.2 Adversarial loss

The content loss describes the difference between the reconstructed image and the reference image, while we also need to consider the loss incurred when reconstructing the image using GAN. When reconstructing an image using GAN, we need to cheat the discriminator network to obtain a more insurgent image, while generating a loss based on the probability that the discriminator produces a natural sample over all training samples defined as

$$l^{\text{SR}}_{\text{Gen}} = \sum_{n=1}^{N} -\log R_{\theta_{R,n}}(G_{\theta_G}(I^{\text{LR}})) \tag{14}$$

where $R_{\theta_{R,n}}(G_{\theta_G}(I^{\text{LR}}))$ denotes the reconstructed image $G_{\theta_G}(I^{LR})$ is the estimated probability of the natural HR image.

3 Experiment

In this section, the proposed model is analyzed in comparison with bicubic interpolation and some typical deep CNN based image super-resolution reconstruction model frameworks, including enhanced deep residual networks for single image super-resolution (EDSR)[21], and wide activation for efficient and accurate image super-resolution (WDSR)[22]. EDSR and WDSR won the international competitions NTIRE 2017 and NTIRE 2018 image high-resolution competitions, respectively.

3.1 Validation indicator

We conducted experiments on a number of benchmark medical image datasets. For a fair quantitative comparison, we use peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM)[23] for SR framework assessment, and the evaluation indicators of PSNR and SSIM are calculated as follows:

$$\text{MSE} = \frac{1}{M^2} \sum_{i-1}^{M} \sum_{j-1}^{M} (a(i,j) - b(i,j))^2 \tag{15}$$

$$\text{PSNR} = 10 \cdot \log_{10}(\frac{\text{MAX}^2}{\text{MSE}}) \tag{16}$$

$$\beta_{\text{SIM}} = \frac{(2\mu_a\mu_b + c_1)(2\sigma_{ab} + c_2)}{(\mu_a^2 + \mu_b^2 + c_1)(\sigma_a + \sigma_b + c_2)} \tag{17}$$

where a is generated image, b is ground truth image, M is image size, and MAX is gray scale's maximum value; μ and σ denote the mean and variance, respectively, and σ_{ab} denotes the covariance of the two images; two constants $c_1 = (0.01 - \text{MAX})^2$ and $c_2 = (0.03 - \text{MAX})^2$ were calculated according to the SSIM convention.

3.2 Dataset and implementation details

To validate the effectiveness of our model in real medical images, we selected separate datasets of MRI scans of the knee and abdomen datasets‡ for comparison tests. MRI imaging methods are completely different from CT images and natural images in general, and each pixel's value in MRI images has no particular physical meaning. Before training and testing, zero-mean normalization had to be applied to each MRI image (i.e., the normalization calculation is to use each value to subtract the mean and then divide by standard deviation). The low-resolution MRI image slices were obtained by averaging the 4 × 4 pooling over the original high-resolution MRI image slices. We set $\lambda_1 = 5 \times 10^{-2}, \lambda_2 = 5\times^{10^{-3}}$ and $\lambda_{L_1} = 10^{-2}$ for training the loss function in the proposed super-resolution model and iterative optimization using the Adam optimizer[24] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the initial learning rate is set to 10^{-4}.

It should be noted that we used knee MRI images, abdominal MRI images, and chest CT images[25] as the training set. To enlarge the training set, we crop the single original image into multiple small images of the same size while the downsampling factor is set to 4 to obtain low-resolution input images. Following that, the suggested depth model is trained using the obtained lowresolution dataset as well as the original high-resolution dataset.

3.3 Super-resolution reconstruction results of MRI images

First we choose the MRI images of knee and abdomen for testing. After the reconstruction process, the PNSR/SSIM test results for the knee MRI test images for all comparing methods are shown in Table 1. The PNSR/SSIM test results for the abdominal MRI images are shown in Table 2. It is worth mentioning that all

‡ http://mridata.org/about

Table 1 Reconstruction results of each comparing algorithm on MRI images of the knee.

Image No.	Bicubic		EDSR		WDSR		T-GAN	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	32.92	0.9033	34.56	0.9405	34.68	0.9468	35.26	0.9526
2	31.58	0.8802	33.26	0.9369	34.12	0.9405	35.03	0.9502
3	31.88	0.8902	33.38	0.9354	33.87	0.9378	35.16	0.9514
4	28.56	0.8682	32.24	0.9215	33.18	0.9289	34.17	0.9453
5	31.21	0.8722	33.12	0.9324	34.14	0.9465	34.98	0.9487
Average	31.23	0.8828	33.312	0.9333	33.998	0.9401	34.92	0.9496

Table 2 Reconstruction results of each comparing algorithm on abdominal MRI images.

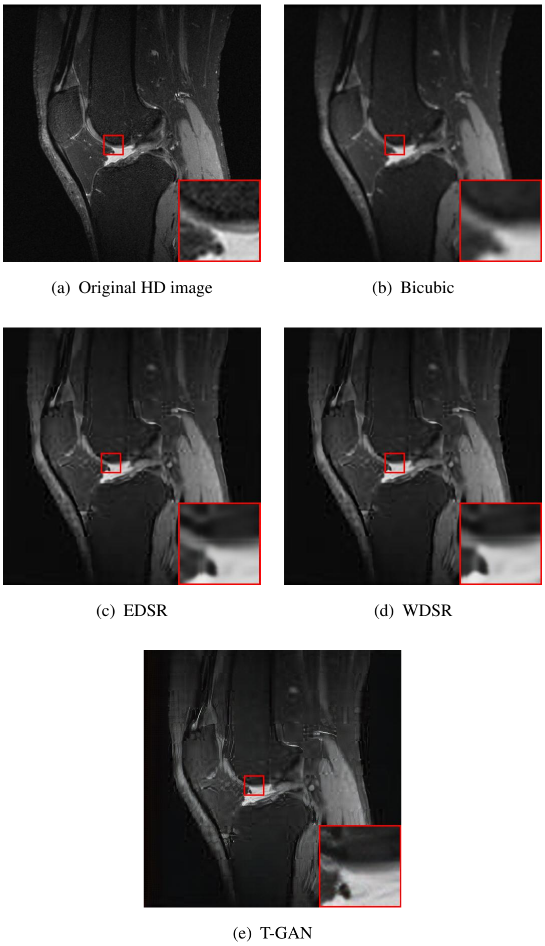
Image No.	Bicubic		EDSR		WDSR		T-GAN	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	31.28	0.8842	33.86	0.9026	34.23	0.9242	35.13	0.9327
2	31.34	0.8901	34.15	0.9158	34.98	0.9384	34.62	0.9318
3	30.78	0.8807	32.98	0.9005	33.86	0.9276	34.76	0.9294
4	32.16	0.8986	34.87	0.9327	34.98	0.9487	35.13	0.9401
5	30.84	0.8872	32.57	0.9124	33.74	0.9305	34.28	0.9389
6	29.47	0.8804	32.12	0.9118	33.68	0.9316	34.19	0.9391
Average	30.98	0.8869	33.43	0.9126	34.25	0.9335	34.69	0.9353

metrics were calculated on cropped photos in order to eliminate the impact of non-subject areas. The quantitative results show that for knee MRI images, our proposed T-GAN model achieves the best performance on the PSNR/SSIM metrics. For abdominal MRI images, our model essentially achieves optimal performance, with individual image WDSR slightly outperforming our model. The experimental results proves that our model is more suitable for medical image super-resolution reconstruction than the existing deep learning based image super-segmentation models.

We likewise give the visualization comparison results for each comparison algorithm, as shown in Figs. 3 and 4. It can be seen that the reconstructed images based on bicubic interpolation and deep learning based EDSR and WDSR both show oversmoothing phenomenon. In contrast, our T-GAN performs better for the reconstruction of detail information due to the texture Transformer structure. Also, Figs. 3 and 4 clearly show that our proposed T-GAN provides the best reconstruction of details, with very low amount of artifacts and noise. The reconstructed images based on bicubic interpolation and deep learning based EDSR both exhibit some loss of detail information due to the loss of some salient image features during the filtering process.

3.4 Super-resolution reconstruction of low-dose CT images

Instead of typical MRI images, the proposed image


(a) Original HD image

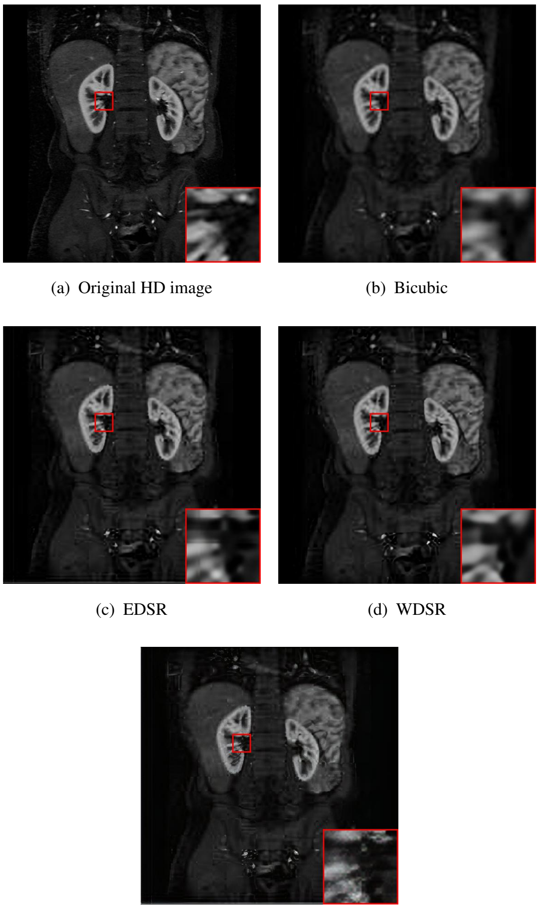
(b) Bicubic

(c) EDSR

(d) WDSR

(e) T-GAN

Fig. 3 Reconstruction results of each algorithm for MRI images of the knee.



(a) Original HD image (b) Bicubic

(c) EDSR (d) WDSR

(e) T-GAN

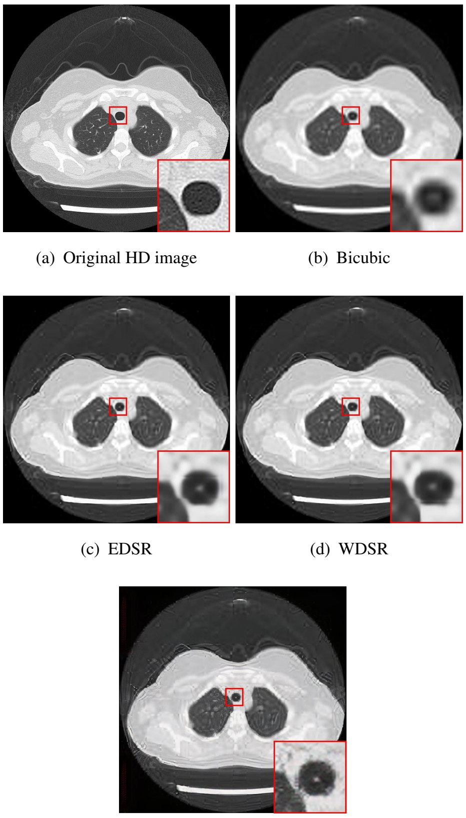
Fig. 4 Reconstruction results of each algorithm for abdominal MRI images.

reconstruction algorithm can also be applied to other medical pictures, such as X-ray scans and computed tomography (CT) scans. X-ray and CT scans are widely used in clinical applications such as noninvasive illness detection, anatomical imaging, and treatment planning. These imaging approaches, however, have some serious limitations and disadvantages. Because it requires high energy electromagnetic wave to pass through human body during imaging process, the radiation damage is unavoidable and high image precision often requires greater energy from the scanner. Low-dose CT (LDCT) is currently the clinically recommended strategy for preventing irreversible radiation harm to the body, however, it comes at the cost of getting CT pictures with low resolution or noise contamination. The spatial resolution is generally coarser than typical CT imaging which has a high signal-to-noise ratio. As a result, obtaining high-resolution scanned images with a lowdose CT scanner would be significantly beneficial to both the doctors and patients for diagnosis purpose.

In this section, we selected chest CT images of COVID-19 patients in an actual hospital[25] for our experiments. The visualization results of the experiments are shown in Figs. 5 and 6. The experimental results show that our proposed T-GAN is also applicable to the super-resolution reconstruction of low-dose CT images, and the high-resolution images obtained by our model have more detailed information compared with the baseline algorithm.

4 Conclusion

In this paper, we present a super-resolution model (TGAN) for medical pictures based on Transformer and generative adversarial network (GAN), with Tansformer approach and residual learning as two generator channels. The results suggest that our proposed T-GAN model can be employed directly for super-resolution MRI image reconstruction, and that our reconstruction methods preserve more texture information than generic image reconstruction algorithms. The findings of the experiments suggest that using the super-resolution reconstruction model to recover more picture details

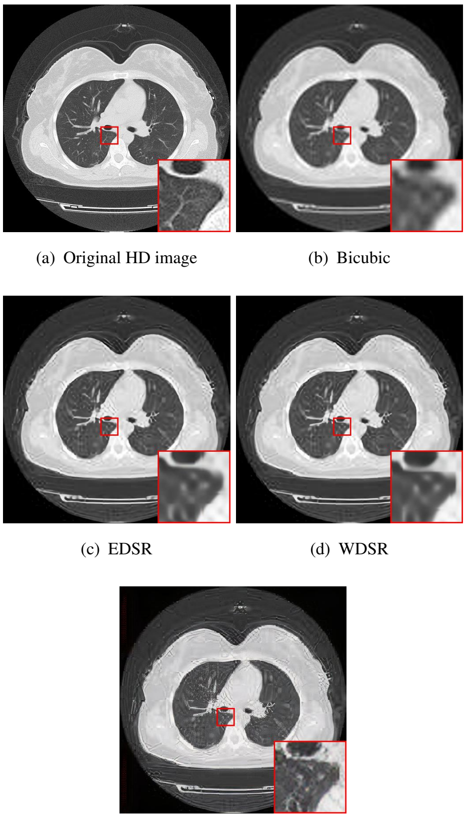


(a) Original HD image (b) Bicubic

(c) EDSR (d) WDSR

(e) T-GAN

Fig. 5 Reconstruction results of each algorithm for low-doze chest CT images: Case 1.


(a) Original HD image (b) Bicubic

(c) EDSR (d) WDSR

(e) T-GAN

Fig. 6 Reconstruction results of each algorithm for low-doze chest CT images: Case 2.

from clinically collected low-resolution images is possible (e.g., LDCT, low-field MRI, and MRI spectral imaging).

References

[1] X. Lu, Z. Huang, and Y. Yuan, MR image super-resolution via manifold regularized sparse learning, *Neurocomputing*, vol. 162, pp. 96–104, 2015.

[2] A. Rueda, N. Malpica, and E. Romero, Single-image super-resolution of brain MR images using overcomplete dictionaries, *Med. Image Anal.*, vol. 17, no. 1, pp. 113–132, 2013.

[3] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, and J. Yang, Exponential wavelet iterative shrinkage thresholding algorithm for compressed sensing magnetic resonance imaging, *Inf. Sci.*, vol. 322, pp. 115–132, 2015.

[4] G. Zheng, G. Han, and N. Q. Soomro, An inception module CNN classifiers fusion method on pulmonary nodule diagnosis by signs, *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 368–383, 2020.

[5] X. Yang, S. Zhan, C. Hu, Z. Liang, and D. Xie, Super-resolution of medical image using representation learning, in *Proc. 2016 8th Int. Conf. Wireless Communications & Signal Processing* (*WCSP*), Yangzhou, China, 2016, pp. 1–6.

[6] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, Learning texture transformer network for image super-resolution, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 5790–5799.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proc. 27th Int. Conf. on Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.

[8] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, Image super-resolution by neural texture transfer, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 7974–7983.

[9] Y. Wang, Y. Liu, W. Heidrich, and Q. Dai, The light field attachment: Turning a DSLR into a light field camera using a low budget camera ring, *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 10, pp. 2357–2364, 2016.

[10] H. Yue, X. Sun, J. Yang, and F. Wu, Landmark image super-resolution by retrieving web images, *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4865–4878, 2013.

[11] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, CrossNet: An end-to-end reference-based super resolution network using cross-scale warping, in *Proc. European Conference on Computer Vision*, Munich, Germany, 2018, pp. 87–104.

[12] V. Boominathan, K. Mitra, and A. Veeraraghavan, Improving resolution and depth-of-field of light field cameras using a hybrid imaging, in *Proc. 2014 IEEE Int. Conf. on Computational Photography* (*ICCP*), Santa Clara, CA, USA, 2014, pp. 1–10.

[13] H. Zheng, M. Ji, L. Han, Z. Xu, H. Wang, Y. Liu, and L. Fang, Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution, in *Proc. British Machine Vision Conference*, London, UK, 2017, pp. 1–13.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 105–114.

[15] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, EnhanceNet: Single image super-resolution through automated texture synthesis, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 4501–4510.

[16] X. Wang, K. Yu, C. Dong, and C. C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 606–615.

[17] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, A fully progressive approach to single-image super-resolution, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, UT, USA, 2018, pp. 977–97709.

[18] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein generative adversarial networks, in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 214–223.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, Improved training of Wasserstein GANs, in *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 5769–5779.

[20] X. Zhu, L. Zhang, L. Zhang, X. Liu, Y. Shen, and S. Zhao, GAN-based image super-resolution with a novel quality loss, *Math. Probl. Eng.*, vol. 2020, p. 5217429, 2020.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, Enhanced deep residual networks for single image super-resolution, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Honolulu, HI, USA, 2017, pp. 1132–1140.

[22] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, Wide activation for efficient and accurate image super-resolution, arXiv preprint arXiv: 1808.08718, 2018.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[24] D. P. Kingma and L. J. Ba, Adam: A method for stochastic optimization, presented at Int. Conf. on Learning Representations, San Diego, CA, USA, 2015.

[25] H. Gunraj, L. Wang, and A. Wong, COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images, *Front. Med.* (*Lausanne*), vol. 7, p. 608525, 2020.

Shihao Tian received the PhD degree from Cornell University, USA in 2019, the MS degree from Cornell University, USA in 2015, and the BS degree from University of Virginia, USA in 2012. He is currently interested in applying AI to physics and scientific research. Also, he is exploring the potential of employing NLP technology to improve STEM education and facilitate the research progress of younger students. He is a member of IEEE and AAPT, and he serves as the judge of ISEF, CONRAD, and PUPC competitions.

Weizhi Du is a rising freshman at Washington University in St. Louis, St. Louis, MO, USA. He is interested in learning computer science and taking digital artwork.