Introduction

The resolution is an important indicator for reflecting the richness of image information. Images with a high resolution (HR) are highly desired and can offer meaningful details that are critical in various applications. Single image super-resolution (SR) aims to enlarge a low-resolution (LR) image into the corresponding HR image. They propose a novel discriminator that only penalises at the scale of the patch and, thus, has relatively few parameters to train. The generator of cGAN is an encoder–decoder with skip connections to shuttle the shared correlation between images. This type of methods needs to train on quantitative data contaminated by thick clouds and shadows, which leads to unsatisfactory results with a blurry or block effect.

Learning-based approaches can better exploit the potential correlation between images. This type of methods needs to train on the given image set in advance, and then recovers the missing high-frequency details using the prior knowledge learned. Li [2] expanded dictionary learning into the multitemporal recovery of quantitative data contaminated by thick clouds and shadows, which is effective on the visual effect. And Li et al. [3] also proposed to utilise both the spectral and temporal information to investigate the missing information reconstruction of remote sensing data, which producing a more stable reconstruction results; However, the traditional learning-based SR including sparse representation [4–7] and local linear regression [8–9] still cannot capture high-level features or low-level features of natural images. In the field of robust data representation, some innovative and efficient image feature extraction methods have been proposed, for example Zhang et al. [10, 11] proposed enhanced two nuclear- and L2, 1-norm regularised 2D neighbourhood preserving projection methods and robust block-diagonal adaptive locality-constrained latent representation to propagate and exchange useful information between salient features and coefficient. For the SR problem, both extractions of high-level and low-level features of images and simulation of mapping relations between features should be considered. Traditional algorithms cannot well fit the mapping relations between features. Therefore, currently, the deep learning method using a convolutional neural network (CNN) has become the mainstream in the SR field, which no longer focuses on reasoning speed, but on whether HR can be well recovered, even in the case of large amplification factor. The deep learning method learns parameters of convolution kernel in the neural network instead of an external dictionary, which is better than interpolation-based methods and learning-based methods in generalisation ability and feature extraction. Dong et al. [12] were the first to introduce a simple, super-resolution CNN (SRCNN) with three convolutional layers where the SR task was split into three steps: patch extraction, non-linear mapping, and reconstruction. All the steps in the pipeline were then integrated innovatively into a unified CNN scheme. Built based on SRCNN, Kim et al. [13] proposed a very deep convolutional network for SR (VDSR) by extending the architecture to 20 levels and suggested that as the depth increases, the performance improves significantly. Zhang et al. [14] developed a spatial–temporal–spectral framework based on a deep CNN (ST-S-CNN), which is qualified for reconstructing small missing areas or regions with regular texture. Wei et al. [15] proposed a deep residual Pan-sharpening neural network for the robust and high-quality fusion of panchromatic images and multi-spectral images, achieving the highest spatial–spectral unified accuracy. Recently, a residual network (ResNet) by He et al. [16] has been widely used in SR tasks, through which one can benefit from a considerably deep network without suffering from several drawbacks caused by training a deep model, such as gradient vanishing and non-convergence. In contrast to both SRCNN and VDSR where the LR input was initially enlarged into a coarse HR version by interpolation and then fed into the network, the computationally efficient sub-pixel convolutional network (ESPCN) proposed by Shi et al. [17] carried out the convolution operation on the original LR image directly and upscaled the size only in the last layer, bringing about a speed increase as compared to the convolution on HR images.

Hongxia Gao1, Zhanhong Chen1, Binyang Huang1, Jiahe Chen2, Zhifu Li3

1 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510000, People's Republic of China
2 School of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing 102200, People's Republic of China
3 School of Mechanical and Electric Engineering, Guangzhou University, Guangzhou 510000, People's Republic of China

E-mail: scuthuangby@163.com

Abstract: Generative adversarial network (GAN) is one of the most prevalent generative models that can synthesise realistic high-frequency details. However, a mismatch between the input and the output may arise when GAN is directly applied to image super-resolution. To alleviate this issue, the authors adopted a conditional GAN (cGAN) in this study. The cGAN discriminator attempted to guess whether the unknown high-resolution (HR) image was produced by the generator with the aid of the original low-resolution (LR) image. They propose a novel discriminator that only penalises at the scale of the patch and, thus, has relatively few parameters to train. The generator of cGAN is an encoder–decoder with skip connections to shuttle the shared low-level information directly across the network. To better maintain the low-frequency information and recover the high-frequency information, they designed a generator loss function combining adversarial loss term and L1 loss term. The former term is beneficial to the synthesis of fine-grained textures, while the latter is responsible for learning the overall structure of the LR input. The experiments revealed that the proposed method could generate HR images with richer details and less over-smoothness.

Image super-resolution based on conditional generative adversarial network

Hongxia Gao1, Zhanhong Chen1, Binyang Huang1, Jiahe Chen2, Zhifu Li3

1 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510000, People's Republic of China
2 School of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing 102200, People's Republic of China
3 School of Mechanical and Electric Engineering, Guangzhou University, Guangzhou 510000, People's Republic of China

E-mail: scuthuangby@163.com

Abstract: Generative adversarial network (GAN) is one of the most prevalent generative models that can synthesise realistic high-frequency details. However, a mismatch between the input and the output may arise when GAN is directly applied to image super-resolution. To alleviate this issue, the authors adopted a conditional GAN (cGAN) in this study. The cGAN discriminator attempted to guess whether the unknown high-resolution (HR) image was produced by the generator with the aid of the original low-resolution (LR) image. They propose a novel discriminator that only penalises at the scale of the patch and, thus, has relatively few parameters to train. The generator of cGAN is an encoder–decoder with skip connections to shuttle the shared low-level information directly across the network. To better maintain the low-frequency information and recover the high-frequency information, they designed a generator loss function combining adversarial loss term and L1 loss term. The former term is beneficial to the synthesis of fine-grained textures, while the latter is responsible for learning the overall structure of the LR input. The experiments revealed that the proposed method could generate HR images with richer details and less over-smoothness.

Image super-resolution based on conditional generative adversarial network

Hongxia Gao1, Zhanhong Chen1, Binyang Huang1, Jiahe Chen2, Zhifu Li3

1 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510000, People's Republic of China
2 School of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing 102200, People's Republic of China
3 School of Mechanical and Electric Engineering, Guangzhou University, Guangzhou 510000, People's Republic of China

E-mail: scuthuangby@163.com

Abstract: Generative adversarial network (GAN) is one of the most prevalent generative models that can synthesise realistic high-frequency details. However, a mismatch between the input and the output may arise when GAN is directly applied to image super-resolution. To alleviate this issue, the authors adopted a conditional GAN (cGAN) in this study. The cGAN discriminator attempted to guess whether the unknown high-resolution (HR) image was produced by the generator with the aid of the original low-resolution (LR) image. They propose a novel discriminator that only penalises at the scale of the patch and, thus, has relatively few parameters to train. The generator of cGAN is an encoder–decoder with skip connections to shuttle the shared low-level information directly across the network. To better maintain the low-frequency information and recover the high-frequency information, they designed a generator loss function combining adversarial loss term and L1 loss term. The former term is beneficial to the synthesis of fine-grained textures, while the latter is responsible for learning the overall structure of the LR input. The experiments revealed that the proposed method could generate HR images with richer details and less over-smoothness.

Image super-resolution based on conditional generative adversarial network

Hongxia Gao1, Zhanhong Chen1, Binyang Huang1, Jiahe Chen2, Zhifu Li3

1 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510000, People's Republic of China
2 School of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, Beijing 102200, People's Republic of China
3 School of Mechanical and Electric Engineering, Guangzhou University, Guangzhou 510000, People's Republic of China

E-mail: scuthuangby@163.com

Abstract: Generative adversarial network (GAN) is one of the most prevalent generative models that can synthesise realistic high-frequency details. However, a mismatch between the input and the output may arise when GAN is directly applied to image super-resolution. To alleviate this issue, the authors adopted a conditional GAN (cGAN) in this study. The cGAN discriminator attempted to guess whether the unknown high-resolution (HR) image was produced by the generator with the aid of the original low-resolution (LR) image. They propose a novel discriminator that only penalises at the scale of the patch and, thus, has relatively few parameters to train. The generator of cGAN is an encoder–decoder with skip connections to shuttle the shared low-level information directly across the network. To better maintain the low-frequency information and recover the high-frequency information, they designed a generator loss function combining adversarial loss term and L1 loss term. The former term is beneficial to the synthesis of fine-grained textures, while the latter is responsible for learning the overall structure of the LR input. The experiments revealed that the proposed method could generate HR images with richer details and less over-smoothness.
SR is a severely ill-posed problem, particularly when it comes to handling large magnification factors as many reasonable HR outputs may meet the criteria given the LR input. Because of the lack of the strong prior knowledge required for the problem, most of the above solutions yield unsatisfactory results with an overly smooth texture and unrealistic details. Considering that essential information is probably missing in the LR image, the highly challenging SR issue shifts from recovering details to synthesising visually convincing details, which is called the generative model. For instance, the SR using a generative adversarial network (GAN) approach developed by Ledig et al. [18] is a representative application of the generative model, consisting of a generator network and a discriminator network. The generator component is designed to fool the discriminator, which is trained to distinguish between real images and synthesised ones. As the discriminator's objective function decreases, synthesised images become closer to real ones in detail texture, so there is no phenomenon of excessive smoothness. Although a generative adversarial network (GAN) provides a promising direction, it has some deficiencies, among which the mode collapse [19] arising from the failure to capture the diversity of the training data may result in a mismatch between the LR inputs and the HR outputs. For example, it is possible that the outputs of an SR model remain almost unchanged even though the inputs change. At this point, although the HR outputs may have realistic details, they are unacceptable because of the inconsistency with the inputs.

In this work, we proposed a new SR method to match the output with the input while maintaining the desirable realistic details. Firstly, we used a conditional GAN (cGAN) in the SR task, where the discriminator of cGAN took both unknown HR image and LR image as the input. Here, LR input served as an auxiliary condition for the decision to exclude real HR images unrelated to LR one. Secondly, we developed an encoder–decoder structure with skip connections, which was conducive to the persistence of information flow and worked well in reducing global mismatch and synthesising local finer details. Thirdly, we designed a generator loss function composed of adversarial loss term and L1 loss term, in order to keep the overall structure of HR output consistent with LR input. Finally, PatchGAN discriminator was developed to determine the generated HR images' authenticity. There is a confrontational relationship between the generator and the discriminator because the purpose of training a generator network is to make its synthesised image fool the discriminator as far as possible, while the purpose of training a discriminator is to recognise the fake image produced by the generator. As training progresses, the synthesised images can reach the level of real images and make it difficult for the discriminator to distinguish between real and fake. The objective of GAN can be defined as follows:

\[
G^* = \arg \min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}_x[I(x)] - \mathbb{E}_z[I(G(z))].
\]

However, the traditional GAN discriminator only pays attention to whether the HR input image obeys the real distribution and does not care whether it matches the original LR image. Therefore, if any irrelevant real HR image in the training set is passed to the discriminator, it will still be judged as true. To approach this problem, cGAN guides the judgments using the original LR image as a conditional input of the discriminator. Fig. 1 shows a schematic representation of the cGAN discriminator.

The objective of cGAN can then be expressed as follows:

\[
\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{z}[\log(1 - D(x, G(z))].
\]

As is shown, the discriminator accepts two inputs where the conditional input \(x\) offers additional information to constrain the model and the other input \(y\) is the image to be discriminated. As such, an unsupervised discriminator of GAN is then changed into supervised in cGAN.

2.2 Structure of the generator

2.2.1 Encoder–decoder model: The proposed generator adopts an encoder–decoder structure employed in many deep learning tasks for feature encoding and decoding [20–22]. The structure consists of an encoder and a decoder. The convolution in the encoder is a downsampling operation that compresses the input data to extract high-level features (such as its global structure). After the input data are encoded by multiple encoder units, they are compressed into a compact vector representing high-level features and reach the bottleneck where it cannot be compressed anymore. As the downsampling continues, the low-level information, such as detailed textures in the input image, is missing. These low-level data are normally pixel-wise and correspond to the high-frequency part of the information to be recovered. The role of the decoder unit is to generate pixels to fill in the specific details based on the bottleneck feature vector, thus restoring the missing high-frequency information in the original LR image. To do so, the convolution in the decoder is an upsampling operation called deconvolution, which is exactly the inverse operation of the encoder. The size of the output data is in a stepped increase through multiple decoders and finally reaches the target size of the HR image. This stage corresponds to the process of a continuous enrichment of the detailed information. The downsampling operation of the encoder and the upsampling operation of the decoder are symmetrical to each other. Therefore, the encoder–decoder generator exhibits a hourglass shape.

The encoder-decoder structure simulates the human brain's cognitive process. Human beings abstract the received external information for high-level perception, which is similar to refining information to form a low-rank vector in the encoding stage. In addition, some prior knowledge can be added to assist in the information abstraction. Therefore, each encoder in the pipeline is equivalent to a link to the human memory chain. In contrast, the decoding stage corresponds to the process of recalling, which uses prior knowledge to decode the stored low-rank information. In summary, the training of the encoder–decoder model corresponds to the acquisition of the human capabilities to process information comprehensively.

First, we enlarged the 64 × 64 LR image into a 256 × 256 image by interpolation and used that as the input of the generator. Considering that SR is a low-level visual task, we did not adopt a pooling layer because it was prone to missing the detailed information. The proposed generator consisted of eight encoders
and eight decoders. The size of all the convolution kernels was $4 \times 4$, and the convolution stride was 2. Both the downsampling factor in the encoders and the upsampling factor in the decoders were 2. The encoder and decoder units were in the form of a ‘convolution-BatchNorm-activation: function’ [23], where the first encoder unit and the last decoder unit did not have BatchNorm. The encoder’s activation function used leaky-ReLU with a slope of $0.2$, while the decoder took ReLU and the last decoder unit adopted a Tanh activation function. Meanwhile, we applied the dropout module with a $50\%$ dropout rate in the first three decoder units. The proposed generator based on the encoder–decoder structure is shown in Fig. 2.

The eight encoder units are listed as follows: (see equation below). The eight decoder units are listed as follows: (see equation below). Let $\text{conv}k$ or $\text{deconv}k$ represent the convolution-BatchNorm-activation function layer with $k$ filters.

### 2.2.2 Skip connections

We introduced skip connections into the encoder–decoder structure in view of the following considerations. First, the LR image and its HR counterpart differed in the richness of details, but both shared the same high-level information. Therefore, we proposed to skip connections to shuttle this high-level information to relieve the potential mismatch in the SR results. Moreover, the skip connections were conducive to the cross-layer transmission of low-level pixel-wise information over the bottleneck of the encoder–decoder. For example, no significant difference was observed in smoothing areas. When synthesising the HR image, the content information in these areas was directly transferred from the LR input, which is exactly what the skip connections do. Lastly, the skip connections could enhance memory persistence to solve the problem of the loss of structural information in the transmission process as the network deepened.

Specifically, the output of the encoder unit was connected to its corresponding decoder unit, i.e. the output of the $i$th layer and the output of the $(n - i + 1)$th layer, where $n$ is the number of layers in the generator network. The proposed encoder–decoder generator network with skip connections is shown in Fig. 3.

### 2.3 Structure of the PatchGAN discriminator

Generative models in deep learning are generally divided into two categories – those generating an entire image [24, 25] and those generating an image patch. Commonly, the former method is only effective for small images. The latter method, usually called the Markov model, assumes that the pixels in a different patch of the same image are independent of each other. Therefore, the entire image can be regarded as a Markov random field (MRF). The Markov model can better recover the local high-frequency information and synthesise finer details by learning the statistical characteristics of the pixels in the local image patches. On the basis of the Markov model, we propose a PatchGAN discriminator, which is different from the traditional discriminators in that it makes decisions at the patch scale but not on the entire image. In our implementation, the output of the PatchGAN discriminator was a $30 \times 30$ matrix where each element value ($0$ to $1$) indicated how realistic the corresponding $70 \times 70$ patch of the unknown image was. Sixteen patches were obtained by convolution on the image, and the stride is set to 62. Each path matches a region of the corresponding LR images. PatchGAN could reduce the parameters that needed training (the parameters required for training were reduced from 838 to 686 MB for full-image processing). Therefore, it was more lightweight and easier to train. Furthermore, the fixed-size PatchGAN discriminator could be used for images of any size.

The discriminator took two inputs – one was the HR image to be discriminated, and the other was an LR conditional input (after interpolation magnification). The discriminator was composed of multiple encoder units. Each unit adopted the convolution-BatchNorm-activation function structure. The activation function of the middle units used leaky-relu with a slope of $0.2$, and the last unit used the sigmoid activation function. The entire convolution kernel size was $4 \times 4$. The convolution stride of the middle layers was 2, and the last layer's stride was 1. The proposed discriminator network is shown in Fig. 4.

The concat unit combines the $256 \times 256 \times 3$ interpolated LR input and the $256 \times 256 \times 3$ unknown HR input into a $256 \times 256 \times 6$ vector. The five encoder units are listed as follows:

$$\text{conv}64 - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  
$$\text{conv64} - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  
$$\text{conv64} - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  
$$\text{conv64} - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  
$$\text{conv64} - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  
$$\text{conv64} - \text{conv}128 - \text{conv}256 - \text{conv}512 - \text{conv}1.$$  

### 2.4 Training

According to the standard training strategy in the original GAN article [24], we performed gradient descent in turn between the generator network and the discriminator network, i.e. alternating between one step of gradient descent on the generator and then one
step on the discriminator. The loss function of the proposed discriminator network was slightly modified on the basis of (2), as shown below:

\[
\mathcal{L}_{\text{discrim}} = -E_x[\log(D(x, y) + \epsilon)] - E_{(x, G(x))}[\log(1 - D(x, G(x)) + \epsilon)].
\]  

where \( \epsilon \) is used to prevent the logarithm term from being 0; its value was \( 1 \times 10^{-12} \) in this study.

The loss function of the proposed generator network consisted of two items – the adversarial loss term and the L1 loss term. The adversarial loss took the form in [24], as shown below:

\[
\mathcal{L}_{\text{gen-loss, gan}} = -E_x[\log(D(x, G(x)) + \epsilon)].
\]  

Both the PatchGAN architecture and the adversarial loss term were designed for high-frequency information characterisation, which was far from sufficient to complete the SR task. Considering the special nature of the SR problem, in addition to generating high-frequency information such as local fine textures, we had to learn low-frequency information such as global structures. The learning of such low-frequency information was completed by the L1 loss term, which is shown below:

\[
\mathcal{L}_{\text{gen-loss, L1}} = E_x[||y - G(x)||].
\]  

Therefore, the overall loss function of the generator composed of two parts is shown in the following equation:

\[
\mathcal{L}_{\text{gen-loss}} = \lambda_{\text{gan}} \cdot \mathcal{L}_{\text{gen-loss, gan}} + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{gen-loss, L1}},
\]  

where \( \lambda_{\text{gan}} \) and \( \lambda_{\text{L1}} \) are the weights used to balance the effect of each item. In the experiments, \( \lambda_{\text{gan}} \) and \( \lambda_{\text{L1}} \) were set to 1 and 1.0, respectively. The formula above implies that the task of the generator was to not only generate a realistic image to fool the discriminator, but also make the synthesised image as close as possible to the original HR image.

The training set was derived from Microsoft’s COCO dataset [26]. To optimise the proposed model, we adopted minibatch gradient descent and Adam’s solver [27] with the initial learning rate of 0.0002 and the momentum parameters of \( \beta_1 = 0.5 \) and \( \beta_2 = 0.999 \). The epoch number was set to 50.

3 Experiments

3.1 Comparison of SR results with other baselines

To assess the effectiveness of the proposed method, we compared the experimental results with the results of SRGAN, SRMD [28] and DBPN [29] in addition to those of the traditional interpolation BICUBIC. All the deep models were trained on the same dataset. We first chose several images (called coco_1, coco_2, and coco_3) from the COCO dataset, changed the size to 256 × 256 as the ground-truth HR image, and then resized them to 64 × 64 as the LR input images. The magnification factor was 4.

These figures reveal that the HR images produced by the proposed method were more visually realistic and more detailed in the texture areas. The results of the other algorithms either had artefacts or were too smooth, which was more noticeable in the texture areas, making the generated HR image less realistic and unnatural.

With the traditional interpolation method, the deep learning methods considerably enhanced the clarity of the images and worked far better, particularly in terms of the edge-preserving ability. The SRGAN, based on a shallow convolutional network, reduced the blurry effect in BICUBIC. The DBPN made the local texture clearer. However, the results of SRMD and DBPN appeared to be over-smooth in the detail areas. Although SRGAN reduced smoothness, it introduced more artefacts.

The proposed method effectively restored the texture details of the lawn around the dog shown in Fig. 5, while the other methods failed to maintain the details, resulting in a serious information loss and affecting the visual perception. Similarly, the finely textured structure of leaves shown in Fig. 6 and the toy’s fuzz shown in Fig. 7 were better recovered. This was attributed to the fact that the proposed adversarial loss enabled the model to learn high-frequency information well and avoid the over-smoothing in the textured areas. Note that although SRGAN could recover some details, it introduced a very large number of artefacts, indicating that the method still lacked in maintaining the overall structure. Additionally, the proposed model worked well in smooth areas, reflecting its ability to maintain low-frequency information, which was exactly what the L1 loss item brought.

In addition to being more detailed in the texture areas, the synthesised image was closer to the real in colour distribution. Other methods, particularly SRMD and DBPN have an obvious ‘sense of oil painting.’ In Fig. 6, the umbrella surface of the proposed method had a strong colour richness, while the colour of the other umbrellas was too single. This was attributed to the fact that the loss function of SRMD and DBPN was based on the mean square error (MSE), which tended to average the pixel values and impaired the colour richness. Furthermore, the transition of different colours was not natural in the other approaches. This jumping phenomenon in colour is called the ‘step effect’ here. For example, there are step effects in the comparison results at the umbrella ridge shown in Fig. 6 and at the intersection of the black and white hair of the dog shown in Fig. 5, which increased the blockiness in the generated image. The proposed approach effectively reduced the step effect and made the colour transition more natural and realistic, while still keeping the edges sharp.

Summarily, the proposed method had a strong edge-preserving ability and could effectively generate texture details and reduce the over-smoothness and artefacts. The colour transition in the synthesised image was more natural, and the step effect was reduced. The experiments demonstrated that the proposed algorithm could successfully recover the high-frequency information while maintaining the low-frequency information well.

Table 1 shows the average peak signal-to-noise ratio (PSNR) and structural self-similarity (SSIM) metrics [30] of the various methods for the above images.

According to the objective assessments, DBPN based on the MSE loss achieved a higher PSNR value. The deeper the network was, the better was the PSNR value. This was attributed to the fact that the goal of minimising the MSE loss was consistent with the goal of maximising the PSNR value of the generated image. The PSNR value of the proposed algorithm was the lowest, even lower than that of the BICUBIC interpolation. This was attributed to the fact that the proposed adversarial loss term was designed to make the synthesised image fool the discriminator, which did not necessarily lead to the highest PSNR value. Although the MSE loss term favoured the higher PSNR, it tended to average the details and cause smoothness. The synthesised image with high PSNR did not necessarily achieve a satisfactory subjective visual effect as reflected by the PSNR. As described in [18, 22], for GAN and the other generative models, as the details of the images were all synthesised, the traditional objective assessments such as PSNR
and SSIM could not accurately evaluate the quality of the generated images.

In order to express the image quality more objectively, we designed the Laplace sum (LS) to measure image contrast and detail retention ability. It is defined as follows:

\[
LS = \sum_{m=1}^{M-1} \sum_{n=1}^{N-1} |G_{mn}|
\]

where \( G_{mn} = 8I_{mn} - (I_{m-1,n-1} + I_{m-1,n} + I_{m-1,n+1} + I_{m,n-1} + I_{m,n+1} + I_{m+1,n-1} + I_{m+1,n+1}) \). \( I \) is the input image, \( M, N \) are the width and height of the images. Table 2 shows the LS of various methods for the above images.

LS is actually an indicator based on gradient information, which can be used to measure the retention of an image in contrast and details. The intuitive analysis shows that the higher the image contrast and the more details, the higher the corresponding LS value. The experimental results show that our method achieves the

Fig. 5 SR results for coco_1 obtained using
(a) BICUBIC, (b) SRGAN, (c) SRMD, (d) DBPN, (e) Proposed method, (f) Ground-truth HR image

Fig. 6 SR results for coco_2 obtained using
(a) BICUBIC, (b) SRGAN, (c) SRMD, (d) DBPN, (e) Proposed method, (f) Ground-truth HR image
highest LS score, which is in line with the subjective visual master of the human eye.

3.2 Ablation experiment: effect of network structure and objective function form

3.2.1 Effect of skip connections: Fig. 8a shows the effect of the skip connections on the synthesised images. There are many artefacts in the image without skip connections. Their presence could be attributed to the fact that the low-frequency information reflecting the global structure had not been effectively learned, resulting in the discontinuity of the image structure.

3.2.2 Effect of objective function form: The proposed objective function in (6) is composed of the adversarial loss $\mathcal{L}_{\text{gen_loss_gan}}$ and the L1 loss $\mathcal{L}_{\text{gen_loss_L1}}$. We discuss their effect on the result by excluding these two terms from the objective function. Fig. 8b shows the experimental results. It can be readily seen that the adversarial loss item was beneficial to the detail enhancement. After removing this item, the image appeared to be very smooth. In contrast, the removal of the L1 loss item caused more artefacts and a serious loss of structural information.

4 Conclusions

In this paper, we investigated image SR on the basis of cGAN. A new generator network is proposed with the encoder–decoder symmetrical structure. To better implement the cross-layer transmission of the low-level information between the input and the output, skip connections were added to the proposed network. A PatchGAN discriminator is also designed to reduce the parameters to be trained, which makes the model more lightweight and easier to train, and is beneficial to the high-frequency texture restoration. In order to effectively capture both low-level and high-level features, adversarial loss item and L1 loss item were employed together to form the proposed generator network’s objective function. The former item was conducive to the characterisation of high-frequency information, such as the textural details, while the latter was used to capture low-frequency information. The proposed model worked well for synthesising HR images with realistic textures and less over-smoothness. Moreover,
the colour distributions of synthesised images are more natural with less step effects. The experimental results show that the proposed method could simultaneously maintain the low-frequency information and restore the high-frequency information.

5 Acknowledgment

This work was supported by the Fundamental Science and Technology Program of Guangzhou, China #201707010054, with less step effects. The experimental results show that the proposed method could simultaneously maintain the low-frequency information and restore the high-frequency information.

5 References

This work was supported by the Fundamental Science and Technology Program of Guangzhou, China #201707010054, with less step effects. The experimental results show that the proposed method could simultaneously maintain the low-frequency information and restore the high-frequency information.

6 References

[1] Nasrollahi, K., Moeslund, B.: ‘Super-resolution: a comprehensive survey’, Mach. Vis. Appl., 2014, 25, (6), pp. 1423–1468

[2] Li, X., Shen, H., Zhang, L., et al.: ‘Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning’, IEEE Trans. Geosci. Remote Sens., 2014, 52, (11), pp. 7086–7098

[3] Li, X., Shen, H., Zhang, L., et al.: ‘Sparse-based reconstruction of missing information in remote sensing images from spectral/temporal complementary information’, ISPRS J. Photogramm. Remote Sens., 2015, 106, pp. 1–15

[4] Yang, J., Wright, J., Huang, T., et al.: ‘Image super-resolution as sparse representation of raw image patches’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2008, pp. 1–8

[5] Yang, J., Wright, J., Huang, T., et al.: ‘Image super-resolution via sparse representation’, IEEE Trans. Image Process., 2010, 19, (11), pp. 2861–2873

[6] Li, J., Yuan, Q., Shen, H., et al.: ‘Hyperspectral image super-resolution by spectral mixture analysis and spatial–spectral group sparsity’, IEEE Trans. Geosci. Remote Sens., 2013, 51, (7), pp. 3998–4008

[7] Lorenzi, L., Melgani, F., Mercier, G.: ‘Missing-area reconstruction in multispectral images under a compressive sensing perspective’, IEEE Trans. Geosci. Remote Sens., 2013, 51, (7), pp. 3998–4008

[8] Timofte, R., De, S.V., Van, G.L.: ‘A+: adjusted anchored neighbourhood regression for fast super-resolution’, Asian Conf. on Computer Vision (ACCV), Singapore, 2014, pp. 111–126

[9] Yang, C.Y., Yang, M.H.: ‘Fast direct super-resolution by simple functions’. Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 561–568

[10] Zhang, Z., Li, F., Zhao, M., et al.: ‘Robust neighborhood preserving projection by nuclear/L2,1-norm regularization for image feature extraction’, IEEE Trans. Image Process., 2017, 26, (4), pp. 1607–1622

[11] Zhang, Z., Ren, J., Li, S., et al.: ‘Robust subspace discovery by block-diagonal adaptive locality-constrained representation’. arXiv: 1908.01266, 2019

[12] Dong, C., Chen, C.L., He, K., et al.: ‘Image super-resolution using deep convolutional networks’, IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38, (2), p. 295

[13] Kim, J., Lee, J.K., Lee, K.M.: ‘Accurate image super-resolution using very deep convolutional networks’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1646–1654

[14] Zhang, Q., Yuan, Q., Zeng, C., et al.: ‘Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network’. IEEE Trans. Geosci. Remote Sens., 2016, 56, (8), pp. 4274–4286

[15] Wei, Y., Yuan, Q., Shen, H., et al.: ‘Boosting the accuracy of multispectral image pan-sharpening by learning a deep residual network’, IEEE Geosci. Remote Sens., 2017, 14, (10), pp. 1795–1799

[16] He, K., Zhang, X., Ren, S., et al.: ‘Deep residual learning for image recognition’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778

[17] Shi, W., Caballero, J., Huszar, F., et al.: ‘Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1874–1883

[18] Ledig, C., Theis, L., Huszar, F., et al.: ‘Photo-realistic single image super-resolution using a generative adversarial network’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Puerto Rico, 2017, pp. 105–114

[19] Metz, L., Poole, B., Pfau, D., et al.: ‘Unrolled generative adversarial networks’. arXiv:1611.02163, 2016

[20] Pathak, D., Kr windy, P., Donahue, J., et al.: ‘Context encoders: feature learning by inpainting’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2536–2544

[21] Wang, X., Gupta, A.: ‘Generative image modelling using style and structure adversarial networks’. Euro. Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016, pp. 318–335

[22] Johnson, J., Alahi, A., Li, F.F.: ‘Perceptual losses for real-time style transfer and super-resolution’. Euro. Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016, pp. 694–711

[23] Ioffe, S., Szegedy, C.: ‘Batch normalization: accelerating deep network training by reducing internal covariate shift’. Int. Conf. on Machine Learning (ICML), Lille, France, 2015, pp. 448–456

[24] Goodfellow, I.J., Pouget, A.J., Mirza, M., et al.: ‘Generative adversarial nets’. Advances in Neural Information Processing Systems (NIPS), Montreal, Quebec, Canada, 2014, pp. 2672–2680

[25] Denton, E., Chintala, S., Zael, A., et al.: ‘Deep generative image models using a Laplacian pyramid of adversarial networks’. Advances in Neural Information Processing Systems (NIPS), Montreal, Quebec, Canada, 2015, pp. 1486–1494
Lin, T.Y., Maire, M., Belongie, S., et al.: ‘Microsoft COCO: common objects in context’. Euro. Conf. on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 740–755

Kingma, D.P., Ba, J.: ‘Adam: a method for stochastic optimization’. arXiv:1412.6980, 2014

Zhang, K., Zuo, W., Zhang, L.: ‘Learning a single convolutional super-resolution network for multiple degradations’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 3262–3271

Haris, M., Shakhnarovich, G., Ukita, N.: ‘Deep back-projection networks for super-resolution’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 1664–1673

Wang, Z., Bovik, A.C., Simoncelli, E.P.: ‘Structural approaches to image quality assessment’, in ‘Handbook of image and video processing’ (Academic Press, Cambridge, MA, USA, 2005), p. 18