

Studies Advanced in Image Super-resolution Reconstruction based on Deep Learning

Hongjie Cao *

Department of Resources and Environment, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China

* Corresponding author email: 2020070903011@std.uestc.edu.cn

Abstract. Image super-resolution reconstruction (SR) has always been a research hotspot in the computer vision community, whose purpose is to convert low-resolution (LR) images into high-resolution (HR) images by algorithms without changing the hardware. Early SR works were mainly based on manual features and mathematical statistical models, relying on interpolation to achieve high-resolution image reconstruction. Thanks to the rapid development of convolutional neural networks, SR based on deep learning has achieved breakthroughs in both accuracy and speed. This paper first introduces the basic principles of image super-resolution reconstruction, including the basic framework and loss function of SR. Then, we introduce the representative super-resolution reconstruction methods such as SRCNN, SRGAN, and TTSR. We also introduce the common public datasets and evaluation indexes, and compare the performance of several SR methods to analyze their advantages and disadvantages. Finally, we summarize the existing challenges in the image super-resolution reconstruction, and give a look out for the future development direction of SR.

Keywords: Image Super-resolution Reconstruction; GAN; Deep Learning.

1. Introduction

Image super-resolution reconstruction (SR) is a technology that converts low-resolution (LR) images into high-resolution (HR) images. Its purpose is to solve the problem of blurred image and low quality. The SR is widely used in remote sensing images [1], security monitoring [2], medical imaging [3] and other fields.

Before the emergence of deep learning, traditional SR methods are mainly divided into the following three types: methods based on interpolation, methods based on degradation model, and methods based on learning. The traditional methods mainly depend on the construction of constraint terms and the accuracy of image registration to achieve the reconstruction effect, but they are not suitable for the SR with large magnification. With the wide application of deep learning in various fields, it has also been introduced into SR, and achieved remarkable results.

This paper consists of follows: Section 2 introduces the basic fundamentals of SR. Section 3 briefly describes some deep learning models of SR. Section 4 introduces some datasets and evaluation methods of SR. Section 5 compares some models. Section 6 is a summary of this paper.

2. Fundamentals

This section introduces the basic theory of SR from three aspects: definition, frame and loss function.

2.1 SR Definition

In real life, due to the influence of equipment, environment and other degradation factors, the images we get are usually LR images. Equation (1) is the definition of LR image, where $D(\cdot)$ represents the degradation function, and θ represents the degradation factor.

$$I_{LR} = D(I_{HR}, \theta) \quad (1)$$

Then, the reconstruction function is shown in (2), where $\hat{I}_{HR}$ stands for reconstructed image, $F(\cdot)$ stands for super resolution reconstruction model, ϵ stands for parameters in the model.

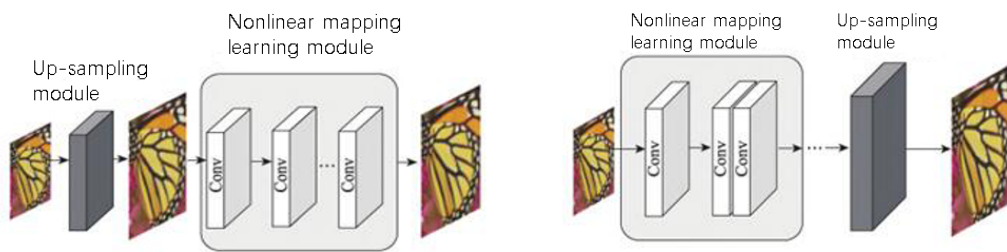
$$\hat{I}_{HR} = F(I_{LR}, \epsilon) \quad (2)$$

2.2 SR Framework

Depending on the location of the up-sampling module, the SR method can be summarized into four frameworks, which will be introduced in detail in the following subsections.

(1) Pre-up-sampling framework. As shown in Fig 1(a), The pre-up-sampling uses a traditional up-sampling algorithm to obtain a higher resolution image, and then uses a deep neural network to refine it.

(2) Post-up-sampling framework. As shown in Fig 1(b), the framework will place the up-sampling module in the back part of the network, which improves the utilization efficiency of computing resources. At the the same time, most convolution calculations under this framework are carried out in low-dimensional space, and the end-to-end learnable upper sampling layer is used for amplification, which further releases the computational power of convolution and reduces the complexity of the model.



(a) Framework of Pre-up-sampling

(b) Framework of Post-up-sampling

Fig 1. Comparison of the Pre-up and Post-up sampling framework

(3) Progressive up-sampling framework. Fig 2 is the structure of Progressive up-sampling framework. In this framework, image amplification is carried out step by step, and the generated images continue to be input until the target resolution is reached.

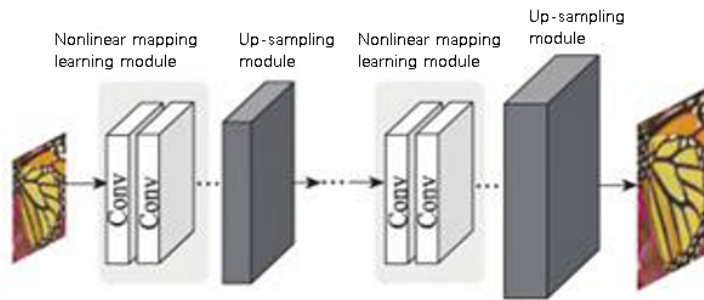


Fig 2. Progressive up-sampling framework

(4) Iterative up-and-down sampling framework. In this framework, through repeated LR-HR mapping study, the model can fully learn the mapping relationship (See Fig. 3). However, due to the complex structure and unclear design standards of this kind of framework, further exploration is needed.

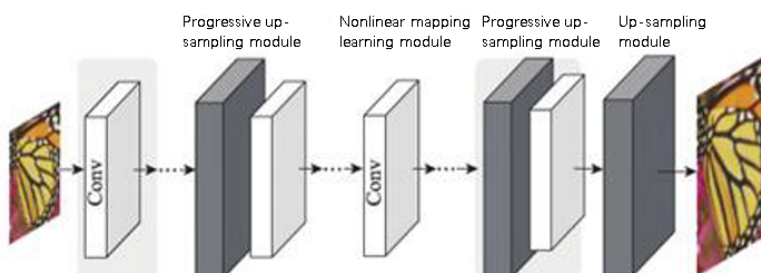


Fig 3. Iterative up-and-down sampling framework

2.3 Loss Function

In SR, the loss function measures the reconstruction quality of HR images by calculating errors. The effect of reconstruction can be improved by reasonable use of multiple loss functions. Several common loss functions are described next.

(1) Pixel loss function. Pixel loss functions include L1 and L2 loss functions, including L1 and L2 loss functions, whose formulas are shown in Equations (1) and (2) respectively. The pixel loss function can improve the convergence speed of the model, but it lacks the consideration of the image, so the reconstructed image is often excessively smooth.

$$L_1(I_{SR}, I_{HR}) = \frac{1}{hwc} \sum_m |I_{SR}^m - I_{HR}^m| \quad (3)$$

$$L_2(I_{SR}, I_{HR}) = \frac{1}{hwc} \sum_m (I_{SR}^m - I_{HR}^m)^2 \quad (4)$$

(2) Content loss function. The content loss function is to evaluate the image quality in the pursuit of human perception and experience. The pre-trained classification network is usually used to evaluate the difference between two images, so as to enhance the visual reality of the reconstructed image. Equation (5) is the calculation formula of content loss function.

$$L_{Cont}(I_{SR}, I_{HR}, \phi) = \frac{1}{hwc} \sum_m (\phi_{(i)}^m(I_{SR}) - \phi_{(i)}^m(I_{HR})) \quad (5)$$

(3) Texture loss function. Texture losses are mainly from the area of style transfer and are therefore also known as common style losses. It uses the correlation between different channels of the reconstructed image and the original image to measure the similarity of two images. Equation (6) is the calculation formula of texture loss function.

$$L_{tt}(I_{SR}, I_{HR}, G) = \frac{1}{C^2} \sqrt{\sum_m (G_{(i)}^m(I_{SR}) - G_{(i)}^m(I_{HR}))^2} \quad (6)$$

(4) Adversarial loss function. Adversarial losses come from generative adversarial networks, as shown in Equation (7). The counter loss function can enhance the counter between discriminator and generator and make the generated picture more realistic.

$$L_{Ad}(I_{LR}, G, D) = \sum_{n=1}^N -\log D(G(I_{LR})) \quad (7)$$

3. Deep Learning Model in SR

In this section, several popular SR methods based on deep learning will be introduced, because many subsequent SR algorithms are to improve these methods or a reasonable combination of use.

3.1 SRCNN

SRCNN [4] is the first work to apply convolutional neural networks to the field of SR, which was proposed by Dong in 2014. Fig.4 is the structure of SRCNN. The structure can be divided into three modules. The following is a brief analysis of these three modules.

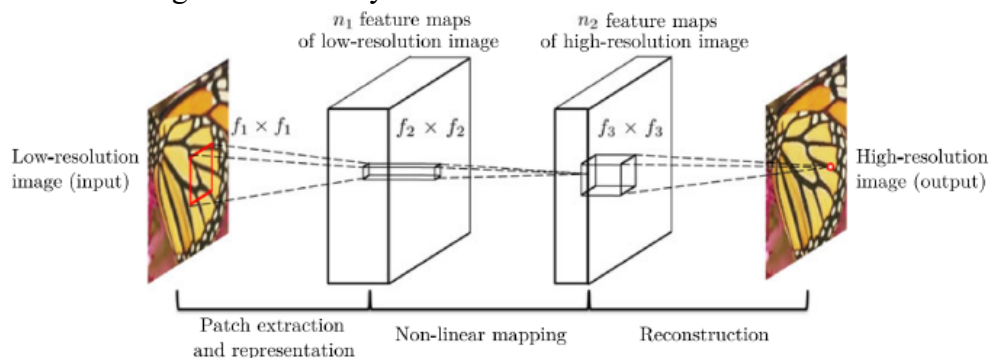


Fig 4. The framework of SRCNN

(1) Patch extraction and representation. The purpose of this process is to obtain a series of feature maps by input image y . This convolutional layer is shown in Equation (8):

$$f_1(y) = \max(0, \omega_1 * y + b_1) \quad (8)$$

where ω_1 and b_1 represent the weight and bias of the filter respectively. This is actually a convolution and activation operation.

(2) Non-linear mapping. This operation fits the nonlinear mapping between LR -HR image, and is shown in the Equation (9):

$$f_2(y) = \max(0, \omega_2 * f_1(y) + b_2) \quad (9)$$

(3) Reconstruction. This operation consolidates the results of the previous operations to produce the most realistic HR image, which is shown in the following equation:

$$f(y) = \omega_3 * f_2(y) + b_3 \quad (10)$$

3.2 SRGAN

With the development of SR, researchers have found that although the evaluation index of the reconstructed HR image is getting higher and higher, the generated image is too smooth and loses more information. The reconstructed image lacks the sense of reality of the human eye, and the actual application effect is not good.

In order to solve these problems, researchers proposed SRGAN [5]. The framework of generative network part of SRGAN is shown in Fig.5. The generative network part of SRGAN network is a deep network composed of B Resnet blocks. The main purpose of Resnet is to use skip connection to enhance the flow of information across layers and to prevent gradient disappearance as the network deepens. The discriminant network part of SRGAN is a CNN network.

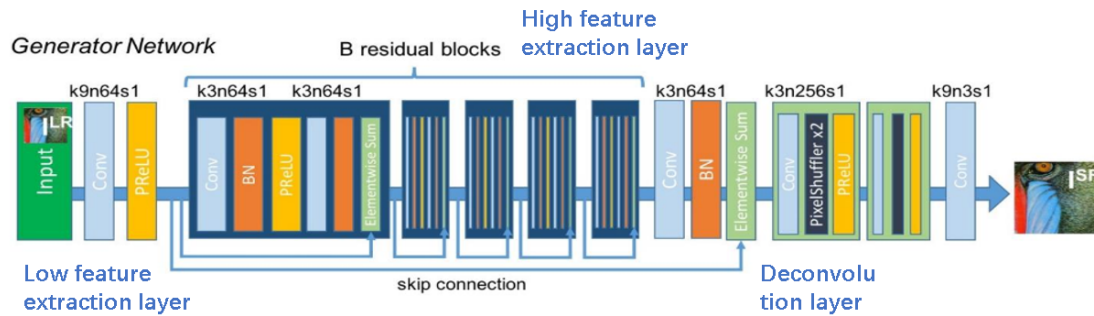


Fig 5. The framework of generative network part of SRGAN

SRGAN is pioneering of shifting attention from quantitative index to image perceptual quality and proposed the perceptual loss function. The loss function is expressed by Equation (11), where the content loss is represented by I_X^{SR} , and the adversarial loss is represented by $10^{-3}I_{Gen}^{SR}$. The content loss function is the feature vector of the preformed VGG network. The main purpose is to use the VGG network to extract the features of HR images and the images generated by the generator, and compare the differences of the corresponding feature maps. The content loss function is shown in Equation (12). Equation (13) is the expression of the adversarial loss function. The purpose is to make the results of the generated network produce higher discriminant value to deceive the discriminant network.

$$I^{SR} = I_X^{SR} + 10^{-3}I_{Gen}^{SR} \quad (11)$$

$$I_{VGG/i,j}^{SR} = \frac{1}{W_{ij}H_{ij}} \sum_{x=1}^{W_{ij}} \sum_{y=1}^{H_{ij}} (\varphi_{ij}(I^{HR})_{x,y} - \varphi_{ij}(G_{\theta G}(I^{LR}))_{x,y})^2 \quad (12)$$

$$I_{Gen}^{SR} = \sum_{n=1}^N -\log D_{\theta D}(G_{\theta G}(I^{LR})) \quad (13)$$

3.3 TTSR

The structure of the texture transformer [6] is shown in Fig.6. Texture transformer consists of four structures: the learnable texture extractor implemented, the relevance embedding module, the hard-attention module for texture conversion, and the soft-attention module for texture synthesis. In addition, the entire module can be further stacked on multiple ladders, which enables texture recovery from different scales. TTSR makes the best use of the information of the reference image. It can

effectively search and transfer high-resolution texture features to low-resolution images, and solve the problem of texture blur and texture distortion.

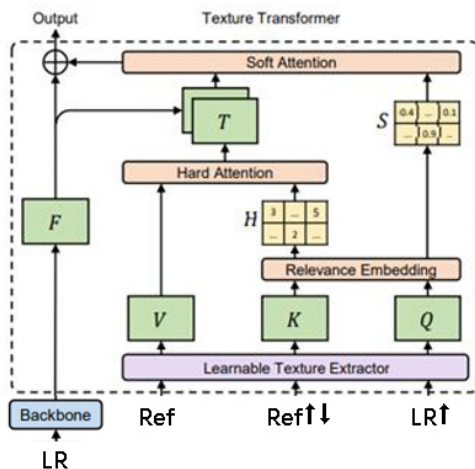


Fig 6. TTSR network structure

In addition, the texture converter can be stacked across scales with the CSFI module to achieve more accurate feature representation. At present, the qualitative and quantitative evaluation results of TTSR are superior to other methods. TTSR have a broad space for development in image generation tasks in the future.

4. Experiments and Performance Comparison

4.1 Common Dataset

Dataset is one of the essential basic elements of deep learning algorithm modeling, and the amount of data and the quality of data will directly affect the effectiveness of the target model. At present, there are a variety of data sets for the training and testing of SR models, which are different in the number of images, the quality of images, the resolution of images, the diversity of images and the imaging environment of images.

Table.1 shows the data sets commonly used in the SR field, and lists the number of images, image format, image type, image average resolution, and category information contained in the images. Urban100[7], BSDS100[8] and Manga109[9] are often used as benchmark datasets for model reconstruction effect evaluation. At the same time, researchers often use multiple data sets to participate in the training and testing of the model to pursue better reconstruction results.

Table 1. Details of common SR datasets

Dataset	Image number	Image format	Image type	usage	Image category information
BSDS100[10]	100	PNG	Unpaired	Val	Animals, people, buildings, etc.
BSDS500[8]	500	JPG	Unpaired	Train/ Val	Animals, people, buildings, etc.
General-100[11]	100	PNG	Unpaired	Train	Animals, food, plants, etc.
CelebA-HQ[12]	30000	PNG	Unpaired	Train	HD face
DIV2K[13]	1000	PNG	Paired	Train/ Val	People, animals, scenery, etc.
Manga109[9]	109	PNG	Unpaired	Val	Cartoon image
Urban100[7]	100	PNG	Unpaired	Val	urban buildings
Flickr2K[14]	2650	PNG	Unpaired	Train	People, animals, scenery, etc.

4.2 Evaluation Methods

Using image quality assessment indicators can better evaluate the results of image super-resolution reconstruction. At present, there are two main types of evaluation: objective evaluation and subjective

evaluation. The results of subjective evaluation and objective evaluation are often different. Objective evaluation is from the mathematical point of view. The image quality can be directly obtained from the data but does not take into account human sensory feelings. Subjective evaluation is based on human subjects to compare and evaluate the quality of the reconstructed images. This evaluation method can reflect the visual experience of the human eye. Therefore, in the actual use process, it should be based on the situation.

(1) PSNR. The PSNR [15] is one of the most commonly used objective indicators of evaluation in the field of scientific research. In SR, it is defined by the mean square error between the maximum pixel value and the two images. The larger the PSNR is, the better the image quality is. Equation.14 is the calculation formula of PSNR, where M_1 is the maximum pixel value, M_2 is the mean square of the corresponding pixel difference between two images.

$$\text{PSNR} = 10 \log \left(\frac{M_1^2}{M_2} \right) \quad (14)$$

Because PSNR is related to MSE and does not take into account the visual perception of human beings, although some models can achieve relatively high PSNR indicators, the actual effect in the real scene is still not satisfactory.

(2) SSMI. SSMI [16] is a similarity index for comparing brightness, contrast and structural details between images. The value range of SSMI is [0,1]. The larger the structural similarity of the image is, the better the image quality is. SSMI is expressed by the following equation:

$$L(X, Y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (15)$$

$$C(X, Y) = \frac{2\mu_x\mu_y + C_2}{\mu_x^2 + \mu_y^2 + C_2} \quad (16)$$

$$S(X, Y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (17)$$

$$\text{SSIM}(X, Y) = L(X, Y)^\delta \times C(X, Y)^\epsilon \times S(X, Y)^\rho \quad (18)$$

Among them, μ_x and μ_y denote the pixel mean of X and Y, σ_x and σ_y denote the pixel standard deviation of X and Y, σ_{xy} denotes the covariance of X and Y. C_1 , C_2 , C_3 are constants.

(3) MOS. MOS [17] is a commonly used method of subjective image quality assessment. It performs visual perception and scoring on the clarity, contrast, color, and appearance details of the image by participating in the evaluation. However, this method depends on people's subjective feelings, so there will be human bias and some limitations.

4.3 Performance Comparison

In this section, we will compare the PSNR and SSMI of some typical SR methods in Manga109, Urban100 and Sun80 datasets on a 4-fold magnification factor, and analyze them quantitatively. ESRGAN [18] is an improvement of SRGAN. It improves the network architecture and loss function, so that the recovery effect at the detail texture is better.

Table 2. Performance comparison of different methods on various datasets

Methods	Manga109		Urban100		Sun80	
	PSNR	SSMI	PSNR	SSMI	PSNR	SSMI
<i>SRCNN</i> [4]	27.12	0.850	24.41	0.738	28.26	0.810
<i>SRGAN</i> [5]	25.12	0.802	24.07	0.729	26.76	0.725
<i>ESRGAN</i> [18]	23.53	0.797	20.91	0.620	24.18	0.651
<i>TTSR</i> [6]	30.09	0.907	25.87	0.784	30.02	0.814

We can see from the data in Table 2 that TTSR has the best performance on both PSNR and SSMI data, indicating that the TTSR method has the best image reconstruction effect from a quantitative perspective. The performance of SRGAN and ESRGAN on both data is not ideal, because SRGAN and ESRGAN more focus on the human eye perception of image. This also shows that the

reconstruction effect of the SR method cannot be completely quantitatively judged, and the visual experience of the human eye should also be paid attention to.

5. Discussion

Although SR has made great progress., there are still a large number of problems in this field. Based on the research on SR, the future improvement and development directions of SR include:

(1) Improve loss function. At the beginning, SRCNN only used pixel loss function and L1 loss function. However, with the continuous development of the SR field, more and more loss functions are added to the model, the loss function is becoming more and more complex, and the quality of the reconstructed image is getting better and better. If the complexity of the loss function can be improved, the training effect of the model will be better and better, and the quality of the reconstructed image will be greatly improved.

(2) Effective combination of multiple network modules. When deep learning was first applied to SR, only three convolutional layers were used. Subsequently, more and more network models are used in SR, the depth of the model is increasing, and the network performance is also greatly improved. Therefore, the reasonable combination of multiple networks can make the reconstruction effect better and better.

(3) Rational use of evaluation methods. Generally, PSNR / SSMI can reflect the quality of image reconstruction. However, it can be seen from SRGAN and ESRGAN that quantitative indicators do not fully explain the quality of reconstructed images. Therefore, finding a reasonable way to use the evaluation index can be of great help to image reconstruction.

(4) Improved up-sampling method. At present, the up-sampling method used by SR will have problems such as uneven amplification and color error, and the magnification is also limited. Therefore, designing an efficient and appropriate up-sampling method can improve the effect of model reconstruction.

6. Conclusion

This paper reviews SR based on deep learning. Firstly, the basic framework and loss function of SR are introduced. Then, SRCNN, SRGAN and TTSR are briefly described respectively. Then, some datasets of SR and three evaluation methods of PSNR, SSMI and MOS are introduced. Then the PSNR and SSMI of some models are compared, and the TTSR is the best. Finally, some improvements are proposed for the SR method of deep learning. At present, there are still some problems in SR. In the future, the application field of SR will be more and more extensive, and the challenges will be more and more. Look forward to the excellent performance of SR in various fields in the future.

References

- [1] Yuan QQ, Li Y, Li JC, et al. Remote sensing image super-resolution via regional spatially adaptive total variation model. Proceedings of 2014 IEEE Geoscience and Remote Sensing Symposium. Quebec City: IEEE, 2014. 3073–3076.
- [2] ZHANG L P, ZHANG H Y, SHEN H F, et al. A super-resolution reconstruction algorithm for surveillance images [J]. Signal Processing, 2010, 90(3): 848-859.
- [3] GREENSPAN H. Super- resolution in medical imaging[J]. The Computer Journal, 2009, 52(1): 43-63.
- [4] CHAO D, CHEN C L, HE K M, et al. Learning a deep convolutional network for image super-resolution[C]//LNCS 8692: Proceedings of the 13th European Conference on Computer Vision, Zurich, Sep 6-12, 2014. Cham: Springer, 2014: 184-199.
- [5] LEDIG C, THEIS L, HUSZAR F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 105-114.

- [6] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5791-5800.
- [7] Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5197–5206.
- [8] Martin D, Fowlkes C, Tal D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver: IEEE, 2001. 416–423.
- [9] Ogawa T, Otsubo A, Narita R, et al. Object detection for comics using Manga109 annotations. arXiv: 1803.08670, 2018.
- [10] Arbeláez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898–916. [doi: 10.1109/TPAMI.2010.161]
- [11] Dong C, Loy CC, He KM, et al. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295–307. [doi: 10.1109/TPAMI.2015.2439281]
- [12] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 2020, 64: 131–148. [doi: 10.1016/j.inffus.2020.06.014]
- [13] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1122-1131.
- [14] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1132-1140.
- [15] Almohammad A, Ghinea G. Stego image quality and the reliability of PSNR. Proceedings of the 2010 2nd International Conference on Image Processing Theory, Tools and Applications. Paris: IEEE, 2010. 215-220.
- [16] Marabini R, Sorzano COS, Matej S, et al. 3-D reconstruction of 2-D crystals in real space. IEEE Transactions on Image Processing, 2004, 13(4): 549–561. [doi: 10.1109/TIP.2003.822620]
- [17] Viswanathan M, Viswanathan M. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. Computer Speech & Language, 2005, 19 (1) 55–83.
- [18] WANG X T, YU K, WU S X, et al. ESRGAN: enhanced super-resolution generative adversarial networks [C] // LNCS 11133: Proceedings of the 15th European Conference on Computer Vision Workshops, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 63-79.