

From Beginner to Master: A Survey for Deep Learning-based Single-Image Super-Resolution

Juncheng Li[†], Zehua Pei[†], and Tieyong Zeng^{*}

Abstract—Single-image super-resolution (SISR) is an important task in image processing, which aims to enhance the resolution of imaging systems. Recently, SISR has made a huge leap and has achieved promising results with the help of deep learning (DL). In this survey, we give an overview of DL-based SISR methods and group them according to their targets, such as reconstruction efficiency, reconstruction accuracy, and perceptual accuracy. Specifically, we first introduce the problem definition, research background, and the significance of SISR. Secondly, we introduce some related works, including benchmark datasets, upsampling methods, optimization objectives, and image quality assessment methods. Thirdly, we provide a detailed investigation of SISR and give some domain-specific applications of it. Fourthly, we present the reconstruction results of some classic SISR methods to intuitively know their performance. Finally, we discuss some issues that still exist in SISR and summarize some new trends and future directions. This is an exhaustive survey of SISR, which can help researchers better understand SISR and inspire more exciting research in this field. An investigation project for SISR is provided in <https://github.com/CV-JunchengLi/SISR-Survey>.

Index Terms—Image super-resolution, single-image super-resolution, SISR, survey, overview.

1 INTRODUCTION

IMAGE super-resolution (SR), especially single-image super-resolution (SISR), is one kind of image transformation task and has received increasing attention in academic and industry. As shown in Fig. 1, SISR aims to reconstruct a super-resolution (SR) image from its degraded low-resolution (LR) one. It is widely used in various computer vision applications, including security and surveillance image, medical image reconstruction, video enhancement, and image segmentation.

Many SISR methods have been studied long before, such as bicubic interpolation and Lanczos resampling [1] which are based on interpolation. However, SISR is an inherently ill-posed problem, and there always exist multiple HR images corresponding to one original LR image. To solve this issue, some numerical methods utilize prior information to restrict the solution space of the reconstruction, such as edge-based methods [2] and image statistics-based methods [3]. Meanwhile, there are some widely used learning methods, such as neighbor embedding methods [4] and sparse coding methods [5], which assume that there exists a transformation between LR and HR patches.

Recently, deep learning (DL) [6] has demonstrated better performance than traditional machine learning models in many artificial intelligence fields, including computer vision [7] and natural language processing [8]. With the rapid development of DL techniques, numerous DL-based methods have been proposed for SISR, continuously prompting the State-Of-The-Art (SOTA) forward. Like other image transformation tasks, the SISR task can generally be divided

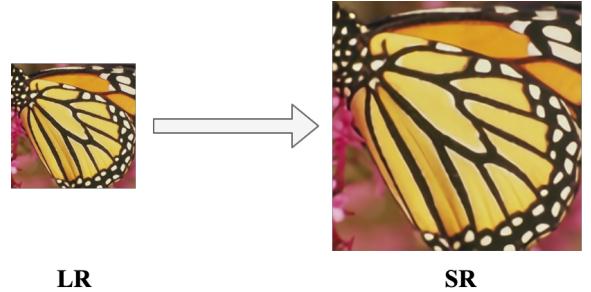


Fig. 1. SISR aims to reconstruct a super-resolution (SR) image from its degraded low-resolution (LR) one.

into three steps: feature extraction and representation, non-linear mapping, and image reconstruction [9]. In traditional numerical models, it is time-consuming and inefficient to design an algorithm satisfying all these processes. On the contrary, DL can transfer the SISR task to an almost end-to-end framework incorporating all these three processes, which can greatly decrease manual and computing expense [10]. Additionally, given the ill-posed nature of SISR which can lead to unstable and hard convergence on the results, DL can alleviate this issue through efficient network architecture and loss functions design. Moreover, modern GPU enables deeper and more complex DL models to train fast, which show greater representation power than traditional numerical models.

It is well known that DL-based methods can be divided into supervised and unsupervised methods. This is the simplest classification criterion, but the range of this classification criterion is too large and not clear. As a result, many technically unrelated methods may be classified into the same type while methods with similar strategies may be classified into completely different types. Different from

• *: Corresponding author.

• †: Contribute equally to this work and are co-first authors.

• J. Li, Z. Pei, and T. Zeng are with the Center for Mathematical Artificial Intelligence (CMAI), Department of Mathematics, The Chinese University of Hong Kong. (E-mails: cujunchengli@gmail.com, pzehua2000@gmail.com, zeng@math.cuhk.edu.hk.)

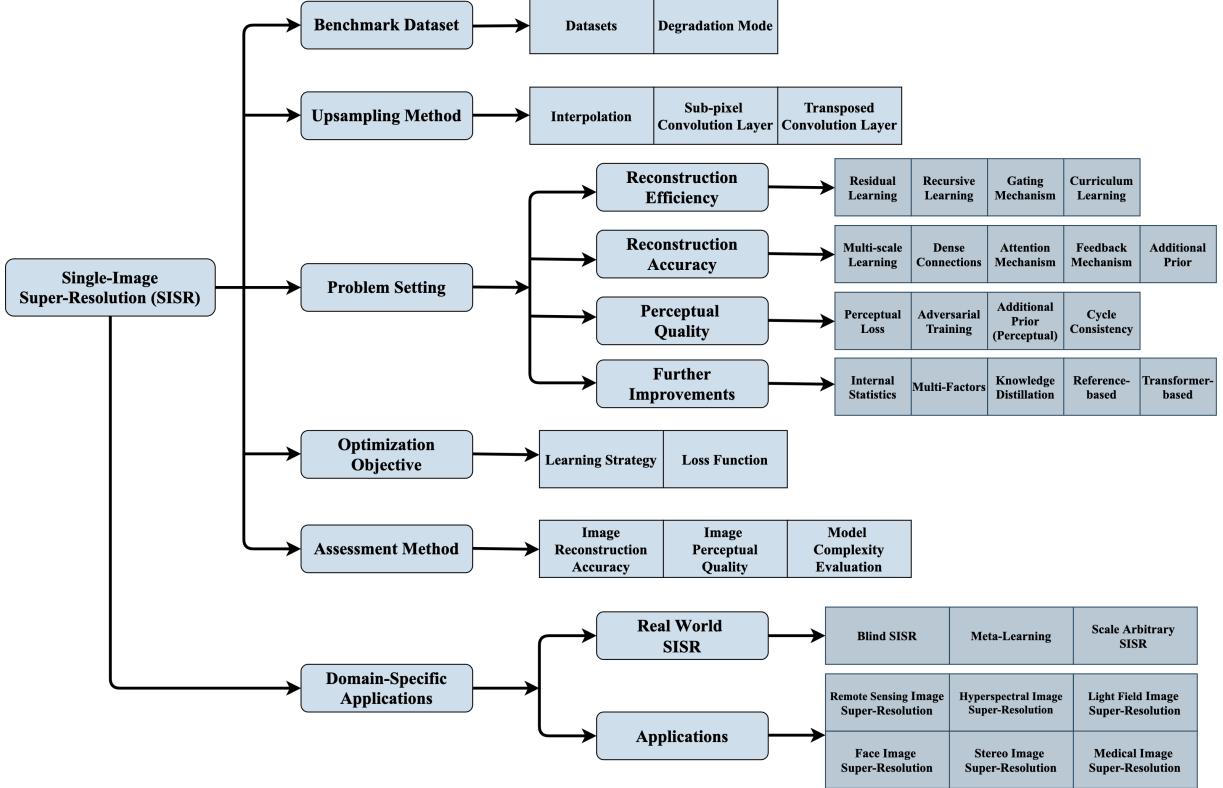


Fig. 2. The content and taxonomy of this survey. In this survey, we divide the DL-based SISR methods into four categories, which are classified according to their specific targets. Among them, the dark gray blocks are the focus methods in this survey.

previous SISR surveys [11], [12] that use supervision as the classification criterion or introduce the methods in a pure literature way, in this survey, we attempt to give a comprehensive overview of DL-based SISR methods and categorize them according to their specific targets. In Fig. 2, we show the content and taxonomy of this survey. Obviously, we divide the DL-based SISR methods into four categories: reconstruction efficiency methods, reconstruction accuracy methods, perceptual quality methods, and further improvement methods. This target-based survey has a clear context hence it is convenient for readers to consult. Specifically, in this survey, we first introduce the problem definition, research background, and significance of SISR. Then, we introduce some related works, including benchmark datasets, upsample methods, optimization objectives, and assessment methods. After that, we provide a detailed investigation of SISR methods and provide the reconstruction results of them. Finally, we discuss some issues that still exist in SISR and provide some new trends and future directions. Overall, the main contributions of this survey are as follows:

(1). We give a thorough overview of DL-based SISR methods according to their targets. This is a new perspective that makes the survey has a clear context hence it is convenient for readers to consult.

(2). This survey covers more than 100 SR methods and introduces a series of new tasks and domain-specific applications extended by SISR in recent years.

(3). We provide a detailed comparison of reconstruction results, including classic, latest, and SOTA SISR methods, to help readers intuitively know their performance.

(4). We discuss some issues that still exist in SISR and summarize some new trends and future directions.

2 PROBLEM SETTING AND RELATED WORKS

2.1 Problem Definition

Image super-resolution is a classic technique to improve the resolution of an imaging system, which can be classified into single-image super-resolution (SISR) and multi-image super-resolution (MISR) according to the number of the input LR images. Among them, MISR has gradually developed into video super-resolution (VSR). Compared with MISR/VSR, SISR is much more challenging since MISR/VSR have extra information for reference while SISR only has information of a single input image for the missing image features reconstruction.

Define the low-resolution image as $I_x \in \mathbb{R}^{h \times w}$ and the ground-truth high-resolution image as $I_y \in \mathbb{R}^{H \times W}$, where $H > h$ and $W > w$. Typically, in a SISR framework, the LR image I_x is modeled as $I_x = D(I_y; \theta_D)$, where D is a degradation map $\mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{h \times w}$ and θ_D denotes the degradation factor. In most cases, the degradation process is unknown. Therefore, researchers are trying to model it. The most popular degradation mode is:

$$D(I_y; \theta_D) = (I_y \otimes \kappa) \downarrow_s + n, \quad (1)$$

where $I_y \otimes \kappa$ represents the convolution between the blur kernel κ and the HR image I_y , \downarrow_s is a subsequent down-sampling operation with scale factor s , and n is usually the additive white Gaussian noise (AWGN) with standard

deviation σ . In the SISR task, we need to recover a SR image I_{SR} from the LR image I_x . Therefore, the task can be formulated as $I_{SR} = \mathcal{F}(I_x; \theta_{\mathcal{F}})$, where \mathcal{F} is the SR algorithm and $\theta_{\mathcal{F}}$ is the parameter set of the SR process.

Recently, researchers have converted the SISR into an end-to-end learning task, relying on massive training data and effective loss functions. Meanwhile, more and more DL-based models have been proposed due to the powerful representation power of CNN and its convenience in both forward and backward computing. Therefore, the SISR task can be transformed into the following optimization goal:

$$\hat{\theta}_{\mathcal{F}} = \arg \min_{\theta_{\mathcal{F}}} \mathcal{L}(I_{SR}, I_y) + \lambda \Phi(\theta), \quad (2)$$

where \mathcal{L} denotes the loss function between the generated SR image I_{SR} and the HR image I_y , $\Phi(\theta)$ denotes the regularization term, and λ is the trade-off parameter that is used to control the percentage of the regularization term.

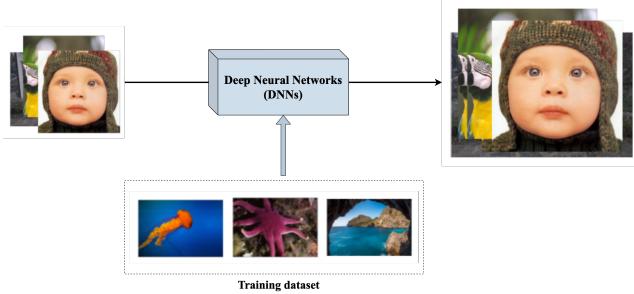


Fig. 3. The training process of data-driven based deep neural networks.

2.2 Benchmarks Datasets

Data is always essential for data-driven models, especially the DL-based SISR models, to achieve promising reconstruction performance (Fig. 3). Nowadays, industry and academia have launched several available datasets for SISR.

2.2.1 Training and Test Datasets

Recently, many datasets for the SISR task have been proposed, including BSDS300 [17], DIV2K [16], and Flickr2K [15]. Meanwhile, there are also many test datasets that can be used to effectively test the performance of the models, such as Set5 [25], Set14 [26], Urban100 [27], and Manga109 [28]. In Table 1, we list a series of commonly used datasets and indicate their detailed attribute.

Among these datasets, DIV2K [16] is the most widely used dataset for model training, which is a high-quality dataset that contains 800 training images, 100 validation images, and 100 test images. Flickr2k is a large extended dataset, which contains 2650 2K images from Flickr. RealSR [19] is the first truly collected SISR dataset with paired LR and HR images. In addition to the listed datasets, some datasets widely used in other computer vision tasks are also used as supplementary training datasets for SISR, such as ImageNet [31] and CelebA [32]. In addition, combining multiple datasets (e.g., DF2K) for training to further improve the model performance has also been widely used.

2.2.2 Degradation Mode

Due to the particularity of the SISR task, it is difficult to construct a large-scale paired real SR dataset. Therefore, researchers often apply degradation patterns on the aforementioned datasets to obtain corresponding degraded images to construct paired datasets. However, images in the real world are easily disturbed by various factors (e.g., sensor noise, motion blur, and compression artifacts), resulting in the captured images being more complex than the simulated images. In order to alleviate these problems and train a more effective and general SISR model, some works model the degradation mode as a combination of several operations (Eq. 1). Based on this degradation formula, three most widely used degradation modes have been proposed: BI, BD, and DN. Among them, **BI** is the most widely used degraded mode to simulate LR images, which is essentially a bicubic downsampling operation. For **BD**, the HR images are blurred by a Gaussian kernel of size 7×7 with standard deviation 1.6 and then downsampled with the scaling factor of $\times 3$. To obtain **DN** mode LR images, the bicubic downsampling is performed on the HR image with scaling factor $\times 3$, and then the Gaussian noise with noise level = 30 is added into the image.

2.3 Upsampling Methods

The purpose of SISR is to enlarge a smaller size image into a larger one and to keep it as accurate as possible. Therefore, enlargement operation, also called upsampling, is an important step in SISR. The current upsampling mechanisms can be divided into four types: pre-upsampling SR, post-upsampling SR, progressive upsampling SR, and iterative up-and-down sampling SR. In this section, we will talk about several kinds of upsampling methods that support these upsampling mechanisms.

2.3.1 Interpolation Methods

Interpolation is the most widely used upsampling method. The current mainstream of interpolation methods includes Nearest-neighbor Interpolation, Bilinear Interpolation, and Bicubic Interpolation. Being highly interpretable and easy to implement, these methods are still widely used today. Among them, **Nearest-neighbor Interpolation** is a simple and intuitive algorithm that selects the nearest pixel value for each position to be interpolated, which has fast execution time but has difficulty in producing high-quality results. **Bilinear Interpolation** sequentially performs linear interpolation operations on the two axes of the image. This method can obtain better results than nearest-neighbor interpolation while maintaining a relatively fast speed. **Bicubic Interpolation** performs cubic interpolation on each of the two axes. Compared with Bilinear, the results of Bicubic are smoother with fewer artifacts, but slower than other interpolation methods. Interpolation is also the mainstream method for constructing SISR paired datasets, and is widely used in the data pre-processing of CNN-based SISR models.

2.3.2 Transposed Convolutional Layers

As shown in Fig. 4, researchers usually consider two kinds of transposed convolution operations: one adds padding around the input matrix and then applies the convolution

TABLE 1
Benchmarks datasets for single-image super-resolution (SISR).

Name	Usage	Amount	Format	Description
General-100 [13]	Train	100	BMP	Common images with clear edges but fewer smooth regions
T91 [5]	Train	91	PNG	Common Images
WED [14]	Train	4744	MAT	Common images
Flickr2K [15]	Train	2650	PNG	2K images from Flickr
DIV2K [16]	Train/Val	1000	PNG	High-quality dataset for CVPR NTIRE competition
BSDS300 [17]	Train/Val	300	JPG	Common images
BSDS500 [18]	Train/Val	500	JPG	Common images
RealSR [19]	Train/Val	100	Train/Val	100 real world low and high resolution image pairs
OutdoorScene [20]	Train/Val	10624	PNG	Images of outdoor scenes
City100 [21]	Train/Test	100	RAW	Common images
Flickr1024 [22]	Train/Test	100	RAW	Stereo images used for Stereo SR
SR-Raw [23]	Train/Test	7*500	JPG/ARW	Raw images produced by real world computational zoom
PIPAL [24]	Test	200	PNG	Perceptual image quality assessment dataset
Set5 [25]	Test	5	PNG	Common images, only 5 images
Set14 [26]	Test	14	PNG	Common images, only 14 images
BSD100 [17]	Test	100	JPG	A subset of BSDS500 for testing
Urban100 [27]	Test	100	PNG	Images of real world structures
Manga109 [28]	Test	109	PNG	Japanese manga
L20 [29]	Test	20	PNG	Common images, very high-resolution
PIRM [30]	Test	200	PNG	Common images, datasets for ECCV PIRM competition

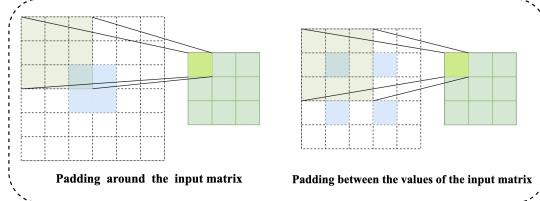


Fig. 4. Two kinds of transposed convolutional layers.

operation, the other adds padding between the values of the input matrix followed by the direct convolution operation. The latter is also called fractionally strided convolution, since it works like doing convolution with a stride less than one. In the transposed convolutional layer, the upsampling level is controlled by the size of padding and it is essentially the opposite of the operation of the normal convolutional layer. Transposed convolutional layer is first proposed in FSRCNN [13] and widely used in DL-based SISR models.

2.3.3 Sub-pixel Convolutional Layer

In ESPCN [33], Shi *et al.* proposed an efficient sub-pixel convolutional layer. Instead of increasing the resolution by directly increasing the number of LR feature maps, sub-pixel first increases the dimension of LR feature maps, i.e., the number of the LR feature maps, and then a periodic shuffling operator is used to rearrange these points in the expanded feature maps to obtain the HR output (Fig. 5). In detail, the formulation of the sub-pixel convolutional layer can be defined as follow:

$$I_{SR} = f^L(I_x) = \mathcal{PS}(W_L * f^{L-1}(I_x) + b_L), \quad (3)$$

where \mathcal{PS} denotes the periodic shuffling operator, which transfers a $h \times w \times C \cdot r^2$ tensor to a tensor of shape $rh \times rw \times C$, and $rh \times rw$ is explicitly the size of HR image, C is the dimension of operating channels. In addition, the

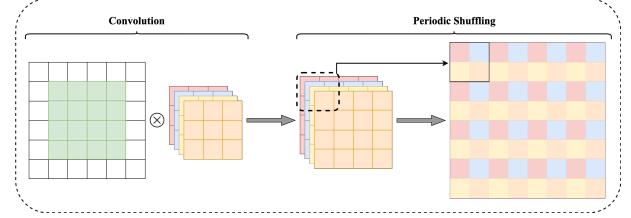


Fig. 5. Principle of the sub-pixel convolutional layer.

convolutional filter W_L has the shape $n_{L-1} \times r^2 C \times K_L \times K_L$, where n_L is the number of feature maps in the $L - 1$ layer. Compared with the transposed convolutional layer, the sub-pixel convolutional layer shows better efficiency, so it is also widely used in DL-based SISR models.

2.4 Optimization Objective

Evaluation and parameter up-gradation are the important steps in all DL-based models. In this section, we will introduce the necessary procedures during the model training.

2.4.1 Learning Strategy

According to different strategies, the DL-based SISR models can be mainly divided into supervised learning methods and unsupervised learning methods.

Supervised Learning: In supervised learning SISR, researchers compute the reconstruction error between the ground-truth image I_y and the reconstructed image I_{SR} :

$$\hat{\theta}_{\mathcal{F}} = \arg \min_{\theta_{\mathcal{F}}} \mathcal{L}(I_{SR}, I_y). \quad (4)$$

Alternatively, researchers may sometimes search for a mapping Φ , such as a pre-trained neural network, to transform the images or image feature maps to some other space and then compute the error:

$$\hat{\theta}_{\mathcal{F}} = \arg \min_{\theta_{\mathcal{F}}} \mathcal{L}(\Phi(I_{SR}), \Phi(I_y)). \quad (5)$$

Among them, \mathcal{L} is the loss function which is used to minimize the gap between the reconstructed image and ground-truth image. According to different loss functions, the model can achieve different performances. Therefore, an effective loss function is also crucial for SISR.

Unsupervised Learning: In unsupervised learning SISR, the way of evaluation and parameter up-gradation is changing by different unsupervised learning algorithms. For example, ZSSR [34] uses the test image and its downscaling images with the data augmentation methods to build the “training dataset” and then applies the loss function to optimize the model. In CinCGAN [35], a model consists of two CycleGAN [36] is proposed, where parameters are upgraded through optimizing the generator-adversarial loss, the cycle consistency loss, the identity loss, and the total variation loss together in each cycle.

2.4.2 Loss Function

In the SISR task, the loss function is used to guide the iterative optimization process of the model by computing some kind of error. Meanwhile, compared with a single loss function, researchers find that a combination of multiple loss functions can better reflect the situation of image restoration. In this section, we will briefly introduce several commonly used loss functions.

Pixel Loss: Pixel loss is the simplest and most popular type among loss functions in SISR, which aims to measure the difference between two images on pixel basis so that these two images can converge as close as possible. It mainly includes the L1 loss, Mean Square Error (MSE Loss), and Charbonnier loss (a differentiable variant of L1 loss):

$$\mathcal{L}_{L1}(I_{SR}, I_y) = \frac{1}{hwc} \sum_{i,j,k} |I_{SR}^{i,j,k} - I_y^{i,j,k}|, \quad (6)$$

$$\mathcal{L}_{MSE}(I_{SR}, I_y) = \frac{1}{hwc} \sum_{i,j,k} (I_{SR}^{i,j,k} - I_y^{i,j,k})^2, \quad (7)$$

$$\mathcal{L}_{Char}(I_{SR}, I_y) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{(I_{SR}^{i,j,k} - I_y^{i,j,k})^2 + \epsilon^2}, \quad (8)$$

where, h , w , and c are the height, width, and the number of channels of the image. ϵ is a numerical stability constant, usually setting to 10^{-3} . Since most mainstream image evaluation indicators are highly correlated with pixel-by-pixel differences, pixel loss is still widely sought after. However, the image reconstructed by this type of loss function usually lacks high-frequency details, so it is difficult to obtain excellent visual effects.

Content Loss: Content loss is also called perceptual loss, which uses a pre-trained classification network to measure the semantic difference between images, and can be further expressed as the Euclidean distance between the high-level representations of these two images:

$$\mathcal{L}_{Cont}(I_{SR}, I_y, \phi) = \frac{1}{h_l w_l c_l} \sum_{i,j,k} (\phi_{(l)}^{i,j,k}(I_{SR}) - \phi_{(l)}^{i,j,k}(I_y)), \quad (9)$$

where ϕ represents the pre-trained classification network and $\phi_{(l)}(I_{HQ})$ represents the high-level representation extracted from the l layer of the network. h_l , w_l , and c_l are the height, width, and the number of channels of the feature

map in the l th layer respectively. With this method, the visual effects of these two images can be as consistent as possible. Among them, VGG [37] and ResNet [38] are the most commonly used pre-training classification networks.

Adversarial Loss: In order to make the reconstructed SR image more realistic, Generative Adversarial Networks (GANs [39]) have been proposed and introduced into various computer vision tasks. Specifically, GAN is composed of a generator and a discriminator. The generator is responsible for generating fake samples, and the discriminator is used to determine the authenticity of the generated samples. For example, the discriminative loss function based on cross-entropy is proposed by SRGAN [38]:

$$\mathcal{L}_{Adversarial}(I_x, G, D) = \sum_{n=1}^N -\log D(G(I_x)), \quad (10)$$

where $G(I_{LQ})$ is the reconstructed SR image, G and D represent the Generator and the Discriminator, respectively.

Prior Loss Apart from the above loss functions, some prior knowledge can also be introduced into SISR models to participate in high-quality image reconstruction, such as sparse prior, gradient prior, and edge prior. Among them, gradient prior loss and edge prior loss are the most widely used prior loss functions, which are defined as follows:

$$\mathcal{L}_{TV}(I_{SR}) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{(I_{SR}^{i,j+1,k} - I_y^{i,j,k})^2 + (I_{SR}^{i+1,j,k} - I_y^{i,j,k})^2}, \quad (11)$$

$$\mathcal{L}_{Edge}(I_{SR}, I_y, E) = \frac{1}{hwc} \sum_{i,j,k} |E(I_{SR}^{i,j,k}) - E(I_y^{i,j,k})|. \quad (12)$$

where E is the image edge detector, and $E(I_{SR}^{i,j,k})$ and $E(I_y^{i,j,k})$ are the image edges extracted by the detector. The purpose of the prior loss is to optimize some specific information of the image toward the expected target so that the model can converge faster and the reconstructed image will contain more texture details.

2.5 Assessment Methods

The image quality assessment (IQA) can be generally divided into objective methods and subjective methods. Objective methods commonly use a specific formulation to compute the results, which are simple and fair, thus become the mainstream assessment method in SISR. However, they can only reflect the recovery of image pixels from a numerical point of view and are difficult to accurately measure the true visual effect of the image. In contrast, subjective methods are always based on human subjective judgments and more related to evaluate the perceptual quality of the image. Based on the pros and cons of the two types of methods mentioned above, several assessment methods are briefly introduced in the following with respect to the aspects of image reconstruction accuracy, image perceptual quality, and reconstruction efficiency.

2.5.1 Image Reconstruction Accuracy

The assessment methods applied to evaluate image reconstruction accuracy are also called *Distortion measures*, which are full-reference. Specifically, given a distorted image \hat{x} and a ground-truth reference image x , full-reference distortion

quantifies the quality of \hat{x} by measuring its discrepancy to x [40] using different algorithms.

Peak Signal-to-Noise Ratio (PSNR): PSNR is the most widely used IQA method in the SISR field, which can be easily defined via the mean squared error (MSE) between the ground truth image $I_y \in \mathbb{R}^{H \times W}$ and the reconstructed image $I_{SR} \in \mathbb{R}^{H \times W}$:

$$MSE = \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (I_y(i, j) - I_{SR}(i, j))^2, \quad (13)$$

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right), \quad (14)$$

where MAX is the maximum possible pixel of the image. Since PSNR is highly related to MSE, a model trained with the MSE loss will be expected to have high PSNR scores. Although higher PSNR generally indicates that the construction is of higher quality, it just considers the per-pixel MSE, which makes it fail to capture the perceptual differences [41].

Structural Similarity index measure (SSIM): SSIM [42] is another popular assessment method that measures the similarity between two images on perceptual basis, including structures, luminance, and contrast. Different from PSNR, which calculates absolute errors on the pixel-level, SSIM suggests that there exists strong inter-dependencies among the pixels that are spatially close. These dependencies carry important information related to the structures perceptually. Thus the SSIM can be expressed as a weighted combination of three comparative measures:

$$\begin{aligned} SSIM(I_{SR}, I_y) &= (l(I_{SR}, I_y)^\alpha \cdot c(I_{SR}, I_y)^\beta \cdot s(I_{SR}, I_y)^\gamma) \\ &= \frac{(2\mu_{I_{SR}}\mu_{I_y} + c_1)(2\sigma_{I_{SR}I_y} + c_2)}{(\mu_{I_{SR}}^2 + \mu_{I_y}^2 + c_1)(\sigma_{I_{SR}}^2 + \sigma_{I_y}^2 + c_2)}. \end{aligned} \quad (15)$$

where l , c , and s represent luminance, contrast, and structure between I_{SR} and I_y , respectively. $\mu_{I_{SR}}$, μ_{I_y} , $\sigma_{I_{SR}}^2$, $\sigma_{I_y}^2$, and $\sigma_{I_{SR}I_y}$ are the average(μ)/variance(σ^2)/covariance(σ) of the corresponding items.

A higher SSIM indicates higher similarity between two images, which has been widely used due to its convenience and stable performance on evaluating the perceptual quality. In addition, there are also some variants of SSIM, such as Multi-Scale SSIM, which is conducted over multiple scales by a process of multiple stages of subsampling.

2.5.2 Image Perceptual Quality

Since the visual system of humans is complex and concerns many aspects to judge the differences between two images, i.e., the textures and flow inside the images, methods which pursue absolutely similarity differences (PSNR/SSIM) will not always perform well. Although distortion measures have been widely used, the improvement in reconstruction accuracy is not always accompanied by an improvement in visual quality. In fact, researchers have shown that the distortion and perceptual quality are at odds with each other in some cases [40]. The image perceptual quality of an image \hat{x} is defined as the degree to which it looks like a natural image, which has nothing to do with its similarity to any reference image.

Mean Opinion Score (MOS): MOS is a subjective method that can straightforwardly evaluate perceptual quality. Specifically, a number of viewers rate their opinions on the quality of a set of images by Double-stimulus [43], i.e., every viewer has both the source and test images. After all the viewers finishing ratings, the results are mapped onto numerical values and the average scores will be the final MOS. MOS is a time-consuming and expensive method as it requires manual participation. Meanwhile, MOS is also doubted to be unstable, since the MOS differences may be not noticeable to the users. Moreover, this method is too subjective to guarantee fairness.

Natural Image Quality Evaluator (NIQE): NIQE [44] is a completely blind image quality assessment method. Without the requirement of knowledge about anticipated distortions in the form of training examples and corresponding human opinion scores, NIQE only makes use of measurable deviations from statistical regularities observed in natural images. It extracts a set of local (quality-aware) features from images based on a natural scene statistic (NSS) model, then fits the feature vectors to a multivariate Gaussian (MVG) model. The quality of a test image is then predicted by the distance between its MVG model and the MVG model learned from a natural image:

$$D(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{((\nu_1 - \nu_2)^T (\frac{\Sigma_1 + \Sigma_2}{2})^{-1} (\nu_1 - \nu_2))}, \quad (16)$$

where ν_1 , ν_2 and Σ_1 , Σ_2 are the mean vectors and covariance matrices of the HR and SR image's MVG model. Notice that, a higher NIQE index indicates lower image perceptual quality. Compared with MOS, NIQE is a more convenient perceptual-evaluation method.

Ma: Ma *et al.* [45] proposed a learning-based no-reference image quality assessment. It is designed to focus on SR images, while other learning-based methods are applied to images degraded by noise, compression, or fast fading rather than SR. It learns from perceptual scores based on human subject studies involving a large number of SR images. And then it quantifies the SR artifacts through three types of statistical properties, i.e., local/global frequency variations and spatial discontinuity. Then these features are modeled by three independent learnable regression forests respectively to fit the perceptual scores of SR images, \hat{y}_n ($n = 1, 2, 3$). The final predicted quality score is $\hat{y} = \sum_n \lambda_n \cdot \hat{y}_n$, and the weight λ is learned by minimizing $\lambda^* = \arg \min_{\lambda} (\sum_n \lambda_n \cdot \hat{y}_n - y)^2$.

Ma performs well on matching the perceptual scores of SR images, but it is still limited compared with other learning-based no-reference methods, since it can only assess the quality degradation arising from the distortion types on which they have been trained.

PI: In the 2018 PIRM Challenge on Perceptual Image Super-Resolution [30], perception index (PI) is first proposed to evaluate the perceptual quality. It is a combination of the no-reference image quality measures Ma and NIQE:

$$PI = \frac{1}{2}((10 - Ma) + NIQE). \quad (17)$$

A lower PI indicates better perceptual quality. This is a new image quality evaluation standard, which has been greatly promoted and used in recent years.

Apart from the aforementioned evaluation methods, some new methods have also been proposed over these years. For example, Zhang *et al.* [46] proposed *Ranker* to learn the ranking orders of NR-IQA methods (i.e., NIQE) on the results of some perceptual SR models. Zhang *et al.* [47] introduced a new dataset of human perceptual similarity judgments. Meanwhile, a perceptual evaluation metric, Learned Perceptual Image Patch Similarity (LPIPS), is constructed by learning the perceptual judgement in this dataset. In summary, how to measure the perceptual quality of SR images more accurately and efficiently is an important issue that needs to be explored.

2.5.3 Reconstruction Efficiency

Although designing deeper networks is the easiest way to obtain better reconstruction performance, it cannot be ignored that these models will also bring more parameters, execution time, and computational costs. In order to broaden the practical application of SISR, we need to consider the trade-off between the model performance and model complexity. Therefore, it is important to evaluate the reconstruction efficiency by the following basic assessments.

Model Size: The model size is related to the storage that the devices need to store the data. A model containing more parameters is harder for the device with limited hardware to run it. Therefore, building lightweight models is conducive to the promotion and application of the algorithm. Among all the indicators, the parameter quantity of the model is the most intuitive indicator to measure the model size.

Execution Time: Usually, a lightweight model tends to require a short execution time, but the emergence of complex strategies such as the attention mechanism has broken this balance. In other words, when some complex operations are introduced into the model, a lightweight network may also require a long execution time. Therefore, it is critically important to evaluate the execution time of the model.

Mult-Adds: The number of multiply-accumulate operations, or Mult-Adds, is always used to measure the model computation since operations in the CNN model are mainly multiplications and additions. The value of Mult-Adds is related to the speed or the time needed to run the model.

In summary, the trade-off between the model performance and model complexity is still need to be concerned.

3 SINGLE-IMAGE SUPER-RESOLUTION

3.1 Benchmark framework for DL-based SISR

In 2014, Dong *et al.* [9] proposed the Super-Resolution Convolutional Neural Network (SRCNN). SRCNN is the first CNN-based SISR model. It shows that a deep CNN model is equivalent to the sparse-coding-based method, which is an example-based method for SISR. Recently, more and more SISR models treat it as an end-to-end learning task. Therefore, building a deep neural network to directly learn the mapping between LR and HR images has become the mainstream method in SISR. Motivated by SRCNN, CNN-based SISR methods are blooming and constantly refreshing the best results.

According to different targets, we divide the DL-based SISR models into four categories: reconstruction efficiency methods, reconstruction accuracy methods, perceptual quality methods, and further improvement methods.

3.2 Reconstruction Efficiency Methods

The problem of low accuracy caused by hardware limitations raises the demand for research on efficient SISR models. Therefore, designing lightweight SISR models that can achieve the same or even better performance than their cumbersome counterparts is urgently needed. In this section, we will discuss some methods that contribute to efficient network structure design.

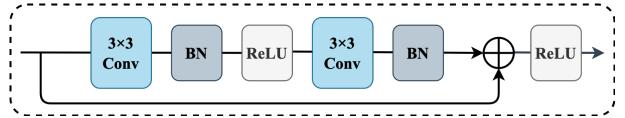


Fig. 6. Sketch of residual learning architecture / residual block.

3.2.1 Residual Learning

In SRCNN, researchers find that better reconstruction performance can be obtained by adding more convolutional layers to increase the receptive field. However, directly stacking the layers will cause vanishing/exploding gradients and degradation problem [48]. Meanwhile, adding more layers will lead to a higher training error and more expensive computational cost.

In ResNet [49], He *et al.* proposed a residual learning framework, where a residual mapping is desired instead of fitting the whole underlying mapping (Fig. 6). In SISR, as LR image and HR image share most of the same information, it is easy to explicitly model the residual image between LR and HR images. Residual learning enables deeper networks and remits the problem of gradient vanishing and degradation. With the help of residual learning, Kim [50] proposed a very deep super-resolution network, also known as VDSR. For the convenience of network design, the residual block [49] has gradually become the basic unit in the network structure. In the convolutional branch, it usually has two 3×3 convolutional layers, two batch normalization layers, and one ReLU activation function in between. It is worth noting that the batch normalization layer is often removed in the SISR task since EDSR [51] points out that the batch normalization layer consumes more memory but will not improve the model performance.

Global and Local Residual Learning: Global residual learning is a skip-connection from input to the final reconstruction layer, which helps improve the transmission of information from input to output and reduce the loss of information to a certain extent. However, as the network becomes deeper, a significant amount of image details are inevitably lost after going through so many layers. Therefore, the local residual learning is proposed, which is performed in every few stacked layers instead of from input to output. In this approach, a multi-path mode is formed and rich image details are carried and also helps gradient flow. Furthermore, many new feature extraction modules have introduced the local residual learning to reinforce strong learning capabilities [52], [53]. Of course, combining local residual learning and global residual learning is also highly popular now [38], [51], [53].

Residual Scaling: In EDSR [51], Lim *et al.* found that increasing the feature maps, i.e., channel dimension, above a

certain level would make the training procedure numerical unstable. To solve such issues, they adopted the residual scaling [54], where the residuals are scaled down by multiplying a constant between 0 and 1 before adding them to the main path. With the help of this residual scaling method, the model performance can be further improved.

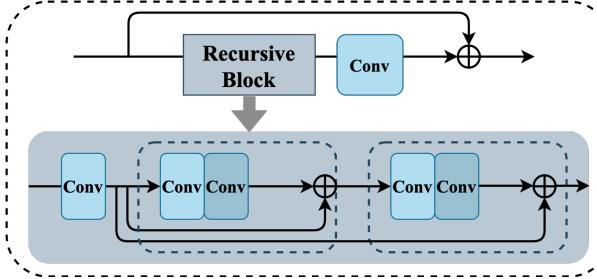


Fig. 7. The model structure of DRRN, where the shaded part denotes the recursive block and the parameters in the dashed box are sharing.

3.2.2 Recursive Learning

In order to obtain a large receptive field without increasing model parameters, recursive learning is proposed for SISR, where the same sub-modules are repeatedly applied in the network and they share the same parameters. In other words, a recursive block is a collection of recursive units, where the corresponding structures among these recursive units share the same parameters. For instance, the same convolutional layer is applied 16 times in DRCN [55], resulting in a 41×41 size receptive field. However, too many stacked layers in recursive learning based model will still cause the problem of vanishing/exploding gradient. Therefore, in DRRN [56], the recursive block is conducted based on residual learning (Fig. 7). Recently, more and more models introduce the residual learning strategy in their recursive units, such as MemNet [57], CARN [58], and SRRFN [59].

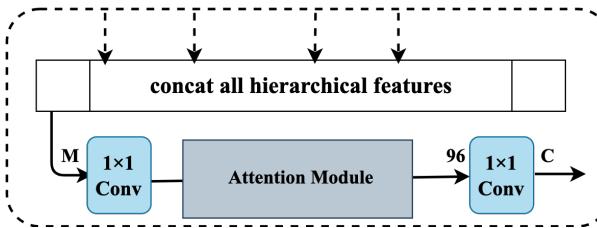


Fig. 8. The structure of the hierarchical feature distillation block (HFDB).

3.2.3 Gating Mechanism

Skip connection in the above residual learning tends to make the channel dimension of the output features extremely high. If such a high dimension channel remains the same in the following layers, the computational cost will be terribly large and therefore will affect the reconstruction efficiency and performance. Intuitively, the output features after the skip connection should be efficiently re-fused instead of simply concatenated.

To solve this issue, researchers recommend using the gating mechanism to adaptively extract and learn more efficient information. Most of the time, a 1×1 convolutional

layer is adopted to accomplish the gating mechanism, which can reduce the channel dimension and leave more effective information. In SRDenseNet [60] and MSRN [52], such 1×1 convolutional layer acts as a bottleneck layer before the reconstruction module. In MemNet [57], it is a gate unit at the end of each memory block to control the weights of the long-term memory and short-term memory. Note that the gate is not only able to serve as bottlenecks placed at the end of the network, but also be continuously conducted in the network. For example, in MemNet [57], IDN [61], and CARN [62], the gating mechanism is used in both global and local region. Sometimes, it can be combined with other operations, such as attention mechanism, to construct a more effective gate module to achieve feature distillation. For instance, Li *et al.* proposed a hierarchical feature distillation block (Fig. 8) by combining 1×1 convolutional layer and attention mechanism in MDCN [63].

3.2.4 Curriculum Learning

Curriculum learning refers to gradually increasing the difficulty of the learning task. For some sequence prediction tasks or sequential decision-making problems, curriculum learning is used to reduce the training time and improve the generalisation performance. Since SISR is an ill-posed problem which is always confronted with great learning difficulty due to some adverse conditions such as large scaling factors, unknown degradation kernels, and noise, it is suitable to utilize curriculum learning to simplify the learning process and improve the reconstruction efficiency.

In LapSRN [64], curriculum learning is applied to progressively reconstruct the sub-band residuals of high-resolution images. In ProSR [65], each level of the pyramid is gradually blended in to reduce the impact on the previously trained layers and the training pairs of each scale are incrementally added. In SRFBN [66], the curriculum learning strategy is applied to solve the complex degradation tasks, where targets of different difficulties are ordered to learn it progressively. With the help of curriculum learning, complex problems can be decomposed into multiple simple tasks, hence accelerating model convergence and obtaining better reconstruction results.

3.3 Reconstruction Accuracy Methods

The quality of the reconstructed SR image is always the main concern in SISR. In this section, we will introduce some classic methods and strategies that can help improve the reconstruction accuracy of SISR models.

3.3.1 Multi-scale Learning

As we all know, rich and accurate image features are essential for SR image reconstruction. Meanwhile, plenty of research works [64], [67], [68] have pointed out that images may exhibit different characteristics at different scales and thus making full use of these features can further improve model performance. Inspired by the inception module [68], Li *et al.* [52] proposed a multi-scale residual block (MSRB, Fig. 9) for feature extraction. MSRB integrates different convolution kernels in a block to adaptively extract image features at different scales. After that, Li *et al.* [63] further optimized the structure and proposed a more efficient

multi-scale dense cross block (MDCB) for feature extraction. MDCB is essentially a dual-path dense network that can effectively detect local and multi-scale features.

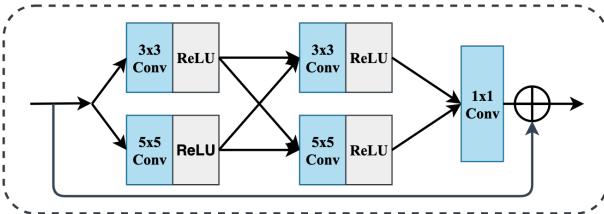


Fig. 9. The structure of multi-scale residual block (MSRB [52]).

Recently, more and more multi-scale SISR models have been proposed. For instance, Qin *et al.* [69] proposed a multi-scale feature fusion residual network (MSFFRN) to fully exploit image features for SISR. Chang *et al.* [70] proposed a multi-scale dense network (MSDN) by combining multi-scale learning with dense connection. Cao *et al.* [71] developed a new SR approach called multi-scale residual channel attention network (MSRCAN), which introduced the channel attention mechanism into the MSRB. All the above examples indicate that the extraction and utilization of multi-scale image features are of increasing importance to further improve the quality of the reconstructed images.

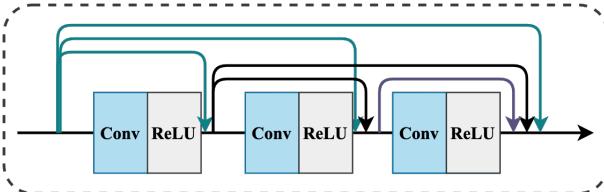


Fig. 10. The structure of a simple dense connection module.

3.3.2 Dense Connection

Dense connection mechanism was proposed in DenseNet [72], which is widely used in the computer vision tasks in recent years. Different from the structure that only sends the hierarchical features to the final reconstruction layer, each layer in the dense block receives the features of all preceding layers (Fig. 10). Short paths created between most of the layers can help alleviate the problem of vanishing/exploding gradients and strengthen the deep information flow through layers, thereby further improving the reconstruction accuracy.

Motivated by the dense connection mechanism, Tong *et al.* introduced it into SISR and proposed the SR-DenseNet [60]. SRDenseNet not only uses the layer-level dense connections, but also the block-level one, where the output of each dense block is connected by dense connections. In this way, the low-level features and high-level features are combined and fully used to conduct the reconstruction. In RDN [73], dense connections are combined with the residual learning to form the residual dense block (RDB), which allows low-frequency features to be bypassed through multiple skip connections, making the main branch focusing on learning high-frequency information. Apart

from aforementioned models, dense connection is also applied in MemNet [57], RPMNet [74], MFNet [75], etc. With the help of dense connection mechanism, the information flow among different depths of the network can be fully used, thus provides better reconstruction results.

3.3.3 Attention Mechanism

Attention mechanism can be viewed as a tool that can allocate available resources to the most informative part of the input. In order to improve the efficiency during the learning procedure, some works are proposed to guide the network to pay more attention to the regions of interest. For instance, Hu *et al.* [76] proposed a squeeze-and-excitation (SE) block to model channel-wise relationships in the image classification task. Wang *et al.* [77] proposed a non-local attention neural network for video classification by incorporating non-local operations. Motivated by these methods, attention mechanism has also been introduced into SISR.

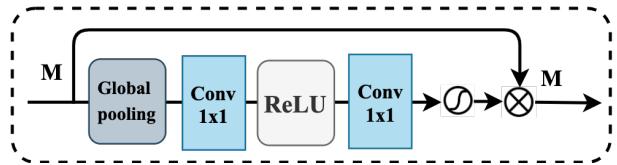


Fig. 11. The principle of channel attention mechanism (CAM).

Channel Attention: In SISR, we mainly want to recover as much valuable high-frequency information as possible. However, common CNN-based methods treat channel-wise features equally, which lacks flexibility in dealing with different types of information. To solve this problem, many methods [53], [78] introduce the SE mechanism in the SISR model. For example, Zhang *et al.* [53] proposed a new module based on the SE mechanism, named residual channel attention block (RCAB). As shown in Fig. 11, a global average pooling layer followed by a Sigmoid function is used to rescale each feature channel, allowing the network to concentrate on more useful channels and enhancing discriminative learning ability. In SAN [79], second-order statistics of features are explored to conduct the attention mechanism based on covariance normalization. A great number of experiments have shown that the second-order channel attention can help the network obtain more discriminative representations, leading to higher reconstruction accuracy.

Non-Local Attention: When CNN-based methods conduct convolution in a local receptive field, the contextual information outside this field is ignored, while the features in distant regions may have a high correlation and can provide effective information. Given this issue, non-local attention has been proposed as a filtering algorithm to compute a weighted mean of all pixels of an image. In this way, distant pixels can also contribute to the response of a position in concern. For example, the non-local operation is conducted in a limited neighborhood to improve the robustness in NLRN [80]. A non-local attention block is proposed in RNAN [81], where the attention mechanisms in both channel- and spatial-wise are used simultaneously in its mask branch to better guide feature extraction in the trunk branch. Meanwhile, a holistic attention network is proposed in HAN [82], which consists of a layer attention

module and a channel-spatial attention module, to model the holistic interdependence among layers, channels, and positions. In CSNLN [83], a cross-scale non-local attention module is proposed to mine long-range dependencies between LR features and large-scale HR patches within the same feature map. All these methods show the effectiveness of the non-local attention, which can further improve the model performance.

3.3.4 Feedback Mechanism

Feedback mechanism refers to carrying a notion of output to the previous states, allowing the model to have a self-correcting procedure. It is worth noting that the feedback mechanism is different from recursive learning since in the feedback mechanism the model parameters are keeping self-correcting and do not share. Recently, feedback mechanism has been widely used in many computer vision tasks [84], [85], which is also beneficial for the SR images reconstruction. Specifically, the feedback mechanism allows the network to carry high-level information back to previous layers and refine low-level information, thus fully guide the LR image to recover high-quality SR images.

In DBPN [86], iterative up- and down-sampling layers are provided to achieve an error feedback mechanism for projection errors at each stage. In DSRN [87], a dual-state recurrent network is proposed, where recurrent signals are exchanged between these states in both directions via delayed feedback. In SFRBN [66], a feedback block is proposed, in which the input of each iteration is the output of the previous one as the feedback information. Followed by several projection groups sequentially with dense skip connections, low-level representations are refined and become more powerful high-level representations.

3.3.5 Additional Prior

Most methods tend to build end-to-end CNN models to achieve SISR since it is simple and easy to implement. However, it is rather difficult for them to reconstruct realistic high-frequency details due to plenty of useful features have been lost or damaged. To solve this issue, priors guided SISR framework has been proposed. Extensive experiments have shown that with the help of image priors, the model can converge faster and achieve better reconstruction accuracy. Recently, many image priors have been proposed, such as total variation prior, sparse prior, and edge prior.

Motivated by this, Yang *et al.* integrated the edge prior with recursive networks and proposed a Deep Edge Guided Recurrent Residual Network (DEGREE [88]) for SISR. After that, Fang *et al.* proposed an efficient and accurate Soft-edge Assisted Network (SeaNet [89]). Different from DEGREE, which directly applies the off-the-shelf edge detectors to detect image edges, SeaNet automatically learns more accurate image edges from the constructed Edge-Net. Meanwhile, the authors pointed out that the more accurate priors introduced, the greater improvement in performance.

3.4 Perceptual Quality Methods

Most methods simply seek to reconstruct SR images with high PSNR and SSIM. However, the improvement in reconstruction accuracy is not always accompanied by an

improvement in visual quality. Blau *et al.* [90] pointed out that there was a perception-distortion trade-off. It is only possible to improve either perceptual quality or distortion, while improving one must be at the expense of the other. Hence, in this section, we provide methods to ease this trade-off problem, hoping to provide less distortion while maintaining good perceptual quality of the image.

3.4.1 Perceptual Loss

Although pixel-wise losses, i.e., L1 and MSE loss, have been widely used to achieve high image quality, they do not capture the perceptual differences between the SR and HR images. In order to address this problem and allow the loss functions to better measure the perceptual and semantic differences between images, content loss, texture loss, and targeted perceptual loss are proposed. Among them, content loss has been widely used to obtain more perceptual and natural images [20], [38], [91], which has been introduced in Sec. 2.4.1. Apart from obtaining more similar content, the same style, such as colors, textures, common patterns, and semantic information are also needed. Therefore, other perceptual loss need to be considered.

Texture Loss: Texture loss, also called style reconstruction loss, is proposed by Gatys *et al.* [92], [93], which can make the model reconstruct high-quality textures. The texture loss is defined as the squared Frobenius norm of the difference between the Gram matrices $G_j^\phi(x)$ of the output and the ground truth images:

$$\mathcal{L}_{\text{texture}}^{\phi,j}(I_{SR}, I_y) = \|G_j^\phi(I_{SR}) - G_j^\phi(I_y)\|_F^2. \quad (18)$$

With the help of the texture loss, the model tends to produce images that have the same local textures as the HR images during training [94].

Targeted Perceptual Loss: The conventional perceptual loss estimates the reconstruction error for an entire image without considering semantic information, resulting in limited capability. Rad *et al.* [95] proposed a targeted perceptual loss that penalized images at different semantic levels on the basis of the labels of object, background, and boundary. Therefore, more realistic textures and sharper edges can be obtained to reconstruct realistic SR images.

3.4.2 Adversarial Training

In 2014, the Generative Adversarial Networks (GANs) was proposed by Goodfellow *et al.* [39], which has been widely used in compute vision tasks, such as style transfer and image inpainting. The GANs consists of a generator and a discriminator. When the discriminator is trained to judge whether an image is true or false, the generator aims at fooling the discriminator rather than minimizing the distance to a specific image, hence it tends to generate outputs that have the same statistics as the training set.

Inspired by GAN, Ledig *et al.* proposed the Super-Resolution Generative Adversarial Network (SRGAN [38]). In SRGAN, the generator G is essentially a SR model that trained to fool the discriminator D , and D is trained to distinguish SR images from HR images. Therefore, the generator can learn to produce outputs that are highly similar to HR images, and then reconstruct more perceptual and

natural SR images. Following this approach, the generative loss $\mathcal{L}_{Gen}(I_x)$ can be defined as:

$$\mathcal{L}_{Gen} = -\log D_{\theta_D}(G_{\theta_G}(I_x)), \quad (19)$$

and the loss in terms of discriminator is:

$$\mathcal{L}_{Dis} = -\log(D_{\theta_D}(I_y)) - \log(1 - D_{\theta_D}(G_{\theta_G}(I_x))). \quad (20)$$

Therefore, we need to solve the following problem:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I_y \sim p_{data}(I_y)} (\log D_{\theta_D}(I_y)) + \mathbb{E}_{I_x \sim p_G(I_x)} (\log(1 - D_{\theta_D}(G_{\theta_G}(I_x)))) \quad (21)$$

In SRGAN [38], the generator is the SRResNet and the discriminator uses the architecture proposed by Radford *et al.* [96]. In ESRGAN [97], Wang *et al.* made two modifications to the SRResNet: (1) replace the original residual block with the residual-in-residual dense block; (2) remove the BN layers to improve the generalization ability of the model. In SRFeat [98], Park *et al.* indicated that the GAN-based SISR methods tend to produce less meaningful high-frequency noise in reconstructed images. Therefore, they adopted two discriminators: an image discriminator and a feature discriminator, where the latter is trained to distinguish SR images from HR images based on the intermediate feature map extracted from a VGG network. In ESRGAN [97], Wang *et al.* adopted the Relativistic GAN [99], where the standard discriminator was replaced with the relativistic average discriminator to learn the relatively realistic between two images. This modification helps the generator to learn sharper edges and more detailed textures.

3.4.3 Additional Prior (Perceptual)

In Sec. 3.3.5, we have introduced the applications of prior knowledge in the CNN-based SISR models. In this section, we will show the benefits of using additional priors in GAN-based models. The target of all the introduced additional priors is to improve the perceptual quality of the reconstructed SR images.

For example, the semantic categorical prior is used to generate richer and more realistic textures with the help of spatial feature transform (SFT) in SFTGAN [20]. With this information from high-level tasks, similar LR patches can be easily distinguished and more natural textual details can be generated. In SPSR [100], the authors utilized the gradient maps to guide image recovery to solve the problem of structural distortions in the GAN-based methods. Among them, the gradient maps are obtained from a gradient branch and integrated into the SR branch to provide structure prior. With the help of gradient maps, we know which region should be paid more attention to, so as to guide the image generation and reduce geometric distortions.

3.4.4 Cycle Consistency

Cycle consistency assumes that there exist some underlying relationships between the source and target domains, and tries to make supervision at the domain level. To be precise, we want to capture some special characteristics of one image collection and figure out how to translate these characteristics into the other image collection. To achieve this, Zhu *et al.* [36] proposed the cycle consistency mechanism,

where not only the mapping from the source domain to the target domain is learned, but also the backward mapping is combined. Specifically, given a source domain X and a target domain Y , we have a translator $G : X \rightarrow Y$ and another translator $F : Y \rightarrow X$ that trained simultaneously to guarantee both an *adversarial loss* that encourages $G(X) \approx Y$ and $F(Y) \approx X$ and a *cycle consistency loss* that encourages $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$.

In SISR, the idea of cycle consistency has also been widely discussed. Given the LR images domain X and the HR images domain Y , we not only learn the mapping from LR to HR but also the backward process. Researchers have shown that learning how to do image degradation first without paired data can help generate more realistic images [101]. In CinCGAN [35], a cycle in cycle network is proposed, where the noisy and blurry input is mapped to a noise-free LR domain firstly and then upsampled with a pre-trained model and finally mapped to the HR domain. In DRN [102], the mapping from HR to LR images is learned to estimate the down-sampling kernel and reconstruct LR images, which forms a closed-loop to provide additional supervision. DRN also gives us a novel approach in unsupervised learning SR, where the deep network is trained with both paired and unpaired data.

3.5 Further Improvement Methods

In the aforementioned part, we have introduced the way to design an efficient SISR model, as well as obtaining high reconstruction accuracy and high perceptual quality for SR images. Though current SISR models have made a significant breakthrough in achieving a balance between reconstruction accuracy and perceptual quality, it still remains a hot topic to explore more effective models.

3.5.1 Internal Statistics

In [103], Zontak *et al.* found that some patches exist only in a specific image and can not be found in any external database of examples. Therefore, SR methods trained on external images can not work well on such images due to the lack of patches information, while methods based on internal statistics may have a good performance. Meanwhile, Zontak *et al.* pointed out that the internal entropy of patches inside a single image was much smaller than the external entropy of patches in a general collection of natural images. Therefore, using the internal image statistics to further improve model performance is a good choice.

In ZSSR [34], the property of internal image statistics is used to train an image-specific CNN, where the training examples are extracted from the test image itself. In training phase, several LR-HR pairs are generated by using data augmentation, and a CNN is trained with these pairs. In test time, the LR image I_{LR} is fed to the trained CNN as input to get the reconstructed image. In this process, the model makes full use of internal statistics of the image itself for self-learning. In SinGAN [104], an unconditional generative model with a pyramid of fully convolutional GANs is proposed to learn the internal patch distribution at different scales of the image. To make use of the recurrence of internal information, they upsampled the LR image several times (depending on the final scale) to obtain the final SR output.

3.5.2 Multi-factors Learning

Typically, in SISR, we often need to train specific models for different upsampling factors and it is difficult to arise at the expectation that a model can be applied to multiple upsampling factors. To solve this issue, some models have been proposed for multiple upsampling factors, such as LapSRN [105], MDSR [51], and MDCN [63].

In LapSRN [105], LR images are progressively reconstructed in the pyramid networks to obtain the large-scale results, where the intermediate results can be taken directly as the corresponding multiple factors results. In [51], Lim *et al.* found the inter-related phenomenon among multiple scales tasks, i.e., initializing the high-scale model parameters with the pre-trained low-scale network can accelerate the training process and improve the performance. Therefore, they proposed the scale-specific processing modules at the head and tail of the model to handle different upsampling factors. To further exploit the inter-scale correlation between different upsampling factors, Li *et al.* further optimized the strategy in MDCN [63]. Different from MDSR which introduces the scale-specific processing strategy both at the head and tail of the model, MDCN can maximize the reuse of model parameters and learn the inter-scale correlation.

3.5.3 Knowledge Distillation

Knowledge distillation refers to a technique that transfers the representation ability of a large (Teacher) model to a small one (Student) for enhancing the performance of the student model. Hence, it has been widely used for network compression or to further improve the performance of the student model, which has shown the effectiveness in many computer vision tasks. Meanwhile, there are mainly two kinds of knowledge distillation, soft label distillation and feature distillation. In soft label distillation, the softmax outputs of a teacher model are regarded as soft labels to provide informative dark knowledge to the student model [106]. In feature distillation, the intermediate features maps are transferred to the student model [107], [108].

Inspired by this, some works introduce the knowledge distillation technique to SISR to further improve the performance of lightweight models. For instance, in SRKD [109], a small but efficient student network is guided by a deep and powerful teacher network to achieve similar feature distributions to those of the teacher. In [110], the teacher network leverage the HR images as privileged information and the intermediate features of the decoder of the teacher network are transferred to the student network via feature distillation, so that the student can learn high frequencies details from the Teacher which trained with the HR images.

3.5.4 Reference-based SISR

In contrast to SISR where only a single LR image is used as input, reference-based SISR (RefSR) takes a reference image to assist the SR process. The reference images can be obtained from various sources like photo albums, video frames, and web image searches. Meanwhile, there are several approaches proposed to enhance image textures, such as image aligning and patch matching. Recently, some RefSR methods [111], [112] choose to align the LR and reference images with the assumption that the reference image possesses similar content as the LR image. For instance, Yue *et*

al. [111] conducted global registration and local matching between the reference and LR images to solve an energy minimization problem. In CrossNet [112], optical flow is proposed to align the reference and LR images at different scales, which are later concatenated into the corresponding layers of the decoder. However, these methods assume that the reference image has a good alignment with the LR image. Otherwise, their performance will be significantly influenced. Different from these methods, Zhang *et al.* [23] applied patch matching between VGG features of the LR and reference images to adaptively transfer textures from the reference images to the LR images. In TTSR [113], Yang *et al.* proposed a texture transformer network to search and transfer relevant textures from the reference images to the LR images based on the attention mechanisms.

3.5.5 Transformer-based SISR

The key idea of Transformer is the self-attention mechanism, which can capture long-term information between sequence elements. Recently, Transformer [114] has achieved brilliant results in NLP tasks. For example, the pre-trained deep learning models (e.g., BERT [115], GPT [116]) have shown effectiveness over conventional methods. Inspired by this, more and more researchers have begun to explore the application of Transformer in computer vision tasks and have achieved breakthrough results many tasks.

Nowadays, some researchers try to introduce Transformer to image restoration tasks. For example, Chen *et al.* proposed the Image Processing Transformer (IPT [117]) which was pre-trained on large-scale datasets. In addition, contrastive learning is introduced for different image processing tasks. Therefore, the pre-trained model can efficiently be employed on the desired task after finetuning. However, IPT [117] relies on large-scale datasets and has a large number of parameters (over 115.5M parameters), which greatly limits its application scenarios. To solve this issue, Liang *et al.* proposed the SwinIR [118] for image restoration based on the Swin Transformer [119]. Specifically, the Swin Transformer blocks (RSTB) is proposed for feature extraction and DIV2K+Flickr2K are used for training. Moreover, Lu *et al.* [120] proposed an Efficient Super-Resolution Transformer (ESRT) for fast and accurate SISR. It is worth noting that ESRT is a lightweight model, which achieves competitive results with fewer parameters and low computing costs. Transformer is a powerful technology, but how to use fewer parameters and datasets to effectively train the model is still worth exploring.

4 DOMAIN-SPECIFIC APPLICATIONS

4.1 Real-World SISR

The degradation modes are complex and unknown in real-world scenarios, where downsampling is usually performed after anisotropic blurring and sometimes signal-dependent noise is added. It is also affected by the in-camera signal processing (ISP) pipeline. Therefore, SISR models trained on bicubic degradation exhibit poor performance when handling real-world images. Moreover, all the aforementioned models can only be applied to some specific integral upsampling factors, but it is essential to develop scale arbitrary SISR models for different practical applications.

Recently, some datasets and new technologies have been proposed for real SISR. In [19], the RealSR dataset is proposed, where paired LR-HR images on the same scene are captured by adjusting the focal length of a digital camera. Meanwhile, a Laplacian Pyramid based Kernel Prediction Network (LP-KPN) is trained with this dataset to learn per-pixel kernels to recover SR images. After that, a series of real image pairs-based methods [121], [122], [123] are proposed. However, this dataset are post-processed and difficult to collect in large quantities, which still limits the model performance. Otherwise, some new technologies have been proposed, such as unsupervised learning [124], [125], self-supervised learning [34], [126], zero-shot learning [34], [127], meta-learning [128], [129], blind SISR, and scale arbitrary SISR [130], [131]. In this part, we introduce the latter three methods due to their impressive foresight and versatility.

4.1.1 **Blind SISR**

Blind SISR has attracted increasing attention due to its significance in real-world applications, which aims to super-resolved LR images with unknown degradation. According to the ways of degradation modelling, they can be simply divided into two categories: explicit degradation modeling methods and implicit degradation modeling methods. Among them, explicit degradation modeling methods can be further divided into two categories according to whether they use the kernel estimation technology. For instance, Zhang *et al.* proposed a simple and scalable deep CNN framework for multiple degradation (SRMD [132]) learning. In SRMD, the concatenated LR image and degradation maps are taken as input after the dimensionality stretching strategy. In DPSR [133], deep super-resolver can be used as a prior with a new degradation model, in order to handle LR images with arbitrary blur kernels. After that, UDVD [134], AMNet [135], USRNet [136], and a series of blind SISR methods are proposed by using the degradation map as an additional input for SR images reconstruction. In contrast, some blind SISR methods pay attention to the kernel estimation along with the SR process [137], [138], [139], [140]. For example, in IKC [137], the iterative kernel correction procedure is proposed to help the blind SISR task to find more accurate blur kernels. In DAN [138], Luo *et al.* adopted an alternating optimization algorithm to estimate blur kernel and restore SR image in a single network, which makes the restorer and estimator be well compatible with each other, and thus achieves good results in kernel estimation. However, the reconstruction accuracy of the above methods greatly depends on the accuracy of the degradation mode estimation. To address this issue, more implicit degradation modeling methods are proposed [35], [141], [142], which aim to implicitly learn the potential degradation modes by the external datasets.

4.1.2 **Meta-Learning**

It is hard for artificial agents to quickly adapt to new things/data like human intelligence, since it is challenging to integrate the prior experience with a few more new information. Meta-learning, or learning to learn, is the mechanism proposed for the learning-based problems, which is usually used in few-shot/zero-shot learning and transfer learning. In meta-learning, the trained model quickly learns

a new task in large task space, where the test samples are used to optimize the meta-learner, therefore the model can quickly adapt with the help of the meta-learner when it encounters new tasks. In SISR, considering the lack of real paired samples, we hope that the model can be trained on simulated paired datasets and then transfer the learned experience to the real SISR task. To address this issue, Soh *et al.* proposed the MZSR [128]. In MZSR, a novel training scheme based on meta-transfer learning is proposed to learn an effective initial weight for fast adaptation to new tasks with the zero-shot unsupervised setting, thus the model can be applied to the real-world scenarios and achieve good results. In [129], Park *et al.* proposed an effective meta-learning method to further improve the model performance without changing the architecture of conventional SISR networks. This method can be applied to any existing SISR models and effectively handle unknown SR kernels. In [143], Hu *et al.* proposed the first unified super-resolution network for arbitrary degradation parameters with meta-learning, termed Meta-USR [143].

4.1.3 **Scale Arbitrary SISR**

In real application scenarios, in addition to processing real images, it is also important to handle arbitrary scale factors with a single model. To achieve this, Hu *et al.* proposed two simple but powerful methods termed Meta-SR [130] and Meta-USR [143]. Among them, Meta-SR is the first SISR method that can be used for arbitrary scale factors and Meta-USR is an improved version that can be applied to arbitrary degradation mode (including arbitrary scale factors). Although Meta-SR and Meta-USR achieve promising performance on non-integer scale factors, they cannot handle SR with asymmetric scale factors. To alleviate this problem, Wang *et al.* [131] suggested learning the scale-arbitrary SISR model from scale-specific networks and developed a plug-in module for existing models to achieve scale-arbitrary SR. Specifically, the proposed plug-in module uses conditional convolution to dynamically generate filters based on the input scale information, thus the networks equipped with the proposed module achieve promising results for arbitrary scales with only a single model.

4.2 Remote Sensing Image Super-Resolution

With the development of satellite image processing, remote sensing has become more and more important. However, due to the limitations of current imaging sensors and complex atmospheric conditions, such as limited spatial resolution, spectral resolution, and radiation resolution, we are facing huge challenges in remote sensing applications.

Recently, many methods have been proposed for remote sensing image super-resolution. For example, a new unsupervised hourglass neural network is proposed in [144] to super-resolved remote sensing images. The model uses a generative random noise to introduce a higher variety of spatial patterns, which can be promoted to a higher scale according to a global reconstruction constraint. In [145], a Deep Residual Squeeze and Excitation Network (DRSEN) is proposed to overcome the problem of the high complexity of remote sensing image distribution. In [146], a mixed high-order attention network (MHAN) is proposed,

which consists of a feature extraction network for feature extraction and a feature refinement network with the high-order attention mechanism for detail restoration. In [147], the authors developed a Dense-Sampling Super-Resolution Network (DSSR) to explore the large-scale SR reconstruction of the remote sensing imageries.

4.3 Hyperspectral Image Super-Resolution

In contrast to human eyes that can only be exposed to visible light, hyperspectral imaging is a technique for collecting and processing information across the entire range of electromagnetic spectrum [148]. The hyperspectral system is often compromised due to the limitations of the amount of the incident energy, hence there is a trade-off between the spatial and spectral resolution. Therefore, hyperspectral image super-resolution is studied to solve this problem.

In [149], a 3D fully convolutional neural network is proposed to extract the feature of hyperspectral images. In [150], Li *et al.* proposed a grouped deep recursive residual network by designing a group recursive module and embedding it into a global residual structure. In [151], an unsupervised CNN-based method is proposed to effectively exploit the underlying characteristics of the hyperspectral images. In [152], Jiang *et al.* proposed a group convolution and progressive upsampling framework to reduce the size of the model and made it feasible to obtain stable training results under small data conditions. In [153], a Spectral Grouping and Attention-Driven Residual Dense Network is proposed to facilitate the modeling of all spectral bands and focus on the exploration of spatial-spectral features.

4.4 Light Field Image Super-Resolution

Light field (LF) camera is a camera that can capture information about the light field emanating from a scene and can provide multiple views of a scene. Recently, the LF image is becoming more and more important since it can be used for post-capture refocusing, depth sensing, and de-occlusion. However, LF cameras are faced with a trade-off between spatial and angular resolution [154]. In order to solve this issue, SR technology is introduced to achieve a good balance between spatial and angular resolution.

In [155], a cascade convolution neural network is introduced to simultaneously up-sample both the spatial and angular resolutions of a light field image. Meanwhile, a new light field image dataset is proposed for training and validation. In order to reduce the dependence of accurate depth or disparity information as priors for the light-field image super-resolution, Sun *et al.* [156] proposed a bidirectional recurrent convolutional neural network and an implicitly multi-scale fusion scheme for SR images reconstruction. In [154], Wang *et al.* proposed a spatial-angular interactive network (LF-InterNet) for LF image SR. Meanwhile, they designed an angular deformable alignment module for feature-level alignment and proposed a deformable convolution network (LF-DFnet [157]) to handle the disparity problem of LF image SR.

4.5 Face Image Super-Resolution

Face image super-resolution is the most famous field in which apply SR technology to domain-specific images.

Due to the potential applications in facial recognition systems such as security and surveillance, face image super-resolution has become an active area of research.

Recently, DL-based methods have achieved remarkable progress in face image super-resolution. In [158], a dubbed CPGAN is proposed to address face hallucination and illumination compensation together, which is optimized by the conventional face hallucination loss and a new illumination compensation loss. In [159], Zhu *et al.* proposed to jointly learn face hallucination and facial spatial correspondence field estimation. In [160], spatial transformer networks are used in the generator architecture to overcome problems related to misalignment of input images. In [161], [162], the identity loss is utilized to preserve the identity-related features by minimizing the distance between the embedding vectors of SR and HR face images. In [163], the mask occlusion is treated as image noise and a joint and collaborative learning network (JDSR-GAN) is constructed for the masked face super-resolution task.

4.6 Medical Image Super-Resolution

Medical imaging methods such as computational tomography (CT) and magnetic resonance imaging (MRI) are essential to clinical diagnoses and surgery planning. Hence, high-resolution medical images are desirable to provide necessary visual information of the human body. Recently, many methods have been proposed for medical image super-resolution

For instance, Chen *et al.* proposed a Multi-level Densely Connected Super-Resolution Network (mDCSRN [164]) with GAN-guided training to generate high-resolution MR images, which can train and inference quickly. In [165], a 3D Super-Resolution Convolutional Neural Network (3DSR-CNN) is proposed to improve the resolution of 3D-CT volumetric images. In [166], Zhao *et al.* proposed a deep Channel Splitting Network (CSN) to ease the representational burden of deep models and further improve the SR performance of MR images. In [167], Peng *et al.* introduced a Spatially-Aware Interpolation Network (SAINT) for medical slice synthesis to alleviate the memory constraint that volumetric data posed. All of these methods are the cornerstone of building the smart medical system and have great research significance and value.

4.7 Stereo Image Super-Resolution

The dual camera has been widely used to estimate the depth information. Meanwhile, stereo imaging can also be applied in image restoration. In the stereo image pair, we have two images with disparity much larger than one pixel. Therefore, full use of these two images can enhance the spatial resolution.

In StereoSR [168], Jeon *et al.* proposed a method that learned a subpixel parallax prior to enhancing the spatial resolution of the stereo images. However, the number of shifted right images is fixed in StereoSR, which makes it fail to handle different stereo images with large disparity variations. To handle this problem, Wang *et al.* [169], [170] proposed a parallax-attention mechanism with a global receptive field along the epipolar line, which can generate reliable correspondence between the stereo image pair and

improve the quality of the reconstructed SR images. In [22], a dataset named Flickr1024 is proposed for stereo image super-resolution, which consists of 1024 high-quality stereo image pairs. In [171], a stereo attention module is proposed to extend pre-trained SISR networks for stereo image SR, which interacts with stereo information bi-directionally in a symmetric and compact manner. In [172], a symmetric bi-directional parallax attention module and an in-line occlusion handling scheme are proposed to effectively interact crossview information. In [173], a Stereo Super-Resolution and Disparity Estimation Feedback Network (SSRDE-FNet) is proposed to simultaneously handle the stereo image super-resolution and disparity estimation in a unified framework.

5 RECONSTRUCTION RESULTS

In order to help readers intuitively know the performance of the aforementioned SISR models, we provide a detailed comparison of reconstruction results of these models. According to the number of model parameters, we divide SISR models into two types: lightweight models and large models. Note that we call model with parameters less than 1000K as lightweight model and model with parameters more than 1M (M=million) as large model. Specifically, we collect 44 representative SISR models, including the most classic, latest, and SOTA SISR models.

In TABLE 2 we provide the reconstruction results, training datasets, and model parameters of these models (lightweight models and large models are separated by the bold black line). According to the results, we can find that: (1) using a large dataset (e.g., DIV2K+Flickr2K) can make the model achieve better results; (2) it is not entirely correct that the more model parameters, the better the model performance. This means that unreasonably increasing the model size is not the best solution; (3) Transformer-based models show strong advantages, whether in lightweight models (e.g., ESRT [120]) or large models (e.g., SwinIR [118]); (4) research on the tiny model (parameters less than 1000K) is still lacking. In the future, it is still important to explore more discriminative evaluation indicators and develop more effective SISR models.

6 REMAINING ISSUES AND FUTURE DIRECTIONS

It is true that the above models have achieved promising results and have greatly promoted the development of SISR. However, we cannot ignore that there are still many challenging issues in SISR. In this section, we will point out some of the challenges and summarize some promising trends and future directions.

6.1 Lightweight SISR for Edge Devices

With the huge development of smart terminal market, research on lightweight SISR models has gained increasing attention. Although existing lightweight SISR models have achieved a good balance between model size and performance, we find that they still cannot be used in edge devices (e.g., smartphones, smart cameras). This is because the model size and computational costs of these models are still exceed the limits of edge devices. Therefore, exploring

lightweight SISR models that can be practical in use for the edge devices has great research significance and commercial value. To achieve this, more efficient network structure and mechanisms are worthy of further exploration. Moreover, it is also necessary to use technologies like network binarization [184] and network quantization [185] to further reduce the model size. In the future, it is worth combining the lightweight SISR models with model compression schemes to achieve the usage of SISR on edge devices.

6.2 Flexible and Adjustable SISR

Although DL-based SISR models have achieved gratifying results, we notice a phenomenon that the structure of all these models must be consistent during training and testing. This greatly limits the flexibility of the model, making the same model difficult to be applied to different applications scenarios. In other words, training specially designed models to meet the requirements of different platforms in necessary for previous methods. However, it will require a great amount of manpower and material resources. Therefore, it is crucial for us to design a flexible and adjustable SISR model that can be deployed on different platforms without retraining while keeping good reconstruction results.

6.3 New Loss Functions and Assessment Methods

In the past, most of SISR models relied on L1 loss or MSE loss. Although some other new loss functions like content loss, texture loss, and adversarial loss have been proposed, they still cannot achieve a good balance between reconstruction accuracy and perceptual quality. Therefore, it remains a important research topic to explore new loss functions that can ease the perception-distortion trade-off. Meanwhile, some new assessment methods are subjective and unfair. Therefore, new assessment methods that can efficiently reflect image perception and distortion at the same time are also essential.

6.4 Mutual Promotion with High-Level Tasks

As we all know, high-level computer vision tasks (e.g., image classification, image segmentation, and image analysis) are highly dependent on the quality of the input image, so SISR technology is usually used for pre-processing. Meanwhile, the quality of the SR images will greatly affect the accuracy of these tasks. Therefore, we recommend using the accuracy of high-level CV tasks as an evaluation indicator to measure the quality of the SR image. Meanwhile, we can design some loss functions related to high-level tasks, thus we can combine the feedback from other tasks to further improve the quality of SR images. On the other hand, we find that the two-step method of pre-processing the image using the SISR model is inefficient, which cannot fully use the potential features of the image itself, resulting in poor model performance. Therefore, we recommend exploring SISR models that can interact with high-level CV tasks, thus SISR and other tasks can promote and learn from each other.

6.5 Efficient and Accurate Real SISR

Real SISR is destined to become the future mainstream in this field. Therefore, it will inevitably become the focus

TABLE 2

PSNR/SSIM comparison on Set5 ($\times 4$), Set14 ($\times 4$), and Urban100 ($\times 4$). Meanwhile, the training datasets and the number of model parameters are provided. Sort by PSNR of Set5 in ascending order. Best results are **highlighted**. Please zoom in to see details.

Models	Set5 PSNR/SSIM	Set14 PSNR/SSIM	Urban100 PSNR/SSIM	Training Datasets	Parameters
SRCNN [174]	30.48/0.8628	27.50/0.7513	24.52/0.7221	T91+ImageNet	57K
ESPCN [33]	30.66/0.8646	27.71/0.7562	24.60/0.7360	T91+ImageNet	20K
FSRCNN [13]	30.71/0.8660	27.59/0.7550	24.62/0.7280	T91+General-100	13K
VDSR [50]	31.35/0.8838	28.02/0.7680	25.18/0.7540	BSD+T91	665K
LapSRN [64]	31.54/0.8855	28.19/0.7720	25.21/0.7560	BSD+T91	812K
DRRN [56]	31.68/0.8888	28.21/0.7721	25.44/0.7638	BSD+T91	297K
MemNet [57]	31.74/0.8893	28.26/0.7723	25.50/0.7630	BSD+T91	677K
AWSRN-S [175]	31.77/0.8893	28.35/0.7761	25.56/0.7678	DIV2K	588K
IDN [61]	31.82/0.8903	28.25/0.7730	25.41/0.7632	BSD+T91	678K
NLRN [80]	31.92/0.8916	28.36/0.7745	25.79/0.7729	BSD+T91	330K
CARN-M [58]	31.92/0.8903	28.42/0.7762	25.62/0.7694	DIV2K	412K
MAFFSRN [176]	32.24/0.8952	28.61/0.7819	26.11/0.7858	DIV2K	550K
RFDN [177]	32.18/0.8948	28.58/0.7812	26.04/0.7848	DIV2K	441K
ESRT [120]	32.19/0.8947	28.69/0.7833	26.39/0.7962	DIV2K	751K
IMDN [178]	32.21/0.8949	28.58/0.7811	26.04/0.7838	DIV2K	715K
MSFIN [179]	32.28/0.8957	28.57/0.7813	26.13/0.7865	DIV2K	682K
DSRN [87]	31.40/0.8830	28.07/0.7700	25.08/0.7470	T91	1.2M
DRCN [55]	31.53/0.8838	28.02/0.7670	25.14/0.7510	T91	1.8M
MADNet [180]	31.95/0.8917	28.44/0.7780	25.76/0.7746	DIV2K	1M
SRMD [132]	31.96/0.8925	28.35/0.7787	25.68/0.7731	BSD+DIV2K+WED	1.6M
SRDenseNet [60]	32.02/0.8934	28.50/0.7782	26.05/0.7819	ImageNet	2.0M
SRResNet [38]	32.05/0.8910	28.49/0.7800	—/—	ImageNet	1.5M
MSRN [52]	32.07/0.8903	28.60/0.7751	26.04/0.7896	DIV2K	6.3M
CARN [58]	32.13/0.8937	28.60/0.7806	26.07/0.7837	BSD+T91+DIV2K	1.6M
SeaNet [89]	32.33/0.8970	28.81/0.7855	26.32/0.7942	DIV2K	7.4M
CRN [58]	32.34/0.8971	28.74/0.7855	26.44/0.7967	DIV2K	9.5M
EDSR [51]	32.46/0.8968	28.80/0.7876	26.64/0.8033	DIV2K	43M
RDN [73]	32.47/0.8990	28.81/0.7871	26.61/0.8028	DIV2K	22.6M
DBPN [86]	32.47/0.8980	28.82/0.7860	26.38/0.7946	DIV2K+Flickr2K	10M
SRFBN [66]	32.47/0.8983	28.81/0.7868	26.60/0.8015	DIV2K+Flickr2K	3.63M
MDCN [63]	32.48/0.8985	28.83/0.7879	26.69/0.8049	DIV2K	4.5M
RNAN [81]	32.49/0.8982	28.83/0.7878	26.61/0.8023	DIV2K	7.5M
SRRFN [59]	32.56/0.8993	28.86/0.7882	26.78/0.8071	DIV2K	4.2M
IGNN [181]	32.57/0.8998	28.85/0.7891	26.84/0.8090	DIV2K	48M
NLSA [182]	32.59/0.9000	28.87/0.7891	26.96/0.8109	DIV2K	41M
RCAN [183]	32.63/0.9002	28.87/0.7889	26.82/0.8087	DIV2K	16M
SAN [79]	32.64/0.9003	28.92/0.7888	26.79/0.8068	DIV2K	15.7M
HAN [82]	32.64/0.9002	28.90/0.7890	26.85/0.8094	DIV2K	16.1M
IPT [117]	32.64/—	29.01/—	27.26/—	ImageNet	115.5M
RFANet [177]	32.66/0.9004	28.88/0.7894	26.92/0.8112	DIV2K	11M
DRN-S [102]	32.68/0.9010	28.93/0.7900	26.84/0.8070	DIV2K+Flickr2K	4.8M
RRDB [97]	32.73/0.9011	28.99/0.7917	27.03/0.8153	DIV2K+Flickr2K	16.7M
DRN-L [102]	32.74/0.9020	28.98/0.7920	27.03/0.8130	DIV2K+Flickr2K	9.8M
SwinIR [118]	32.92/0.9044	29.09/0.7950	27.45/0.8254	DIV2K+Flickr2K	11.8M

of researchers in the next few years. On the one hand, a sufficiently large and accurate real image dataset is critical to Real SISR. To achieve this, in addition to the manual collection, we recommend using generative technology to simulate the images, as well as using the generative adversarial network to simulate enough degradation modes to build the large real dataset. On the other hand, considering the difficulty of constructing real image dataset, it is important to develop unsupervised learning-based SISR, meta learning-based SISR, and blind SISR. Among them, unsupervised learning can make the models get rid of the dependence on dataset, meta learning can help models migrate from simulated datasets to real data with simple fine-tuning, and blind SISR can display or implicitly learn

the degradation mode of the image, and then reconstruct high-quality SR images based on the learned degradation mode. Although plenty of blind SISR methods have been proposed, they always have unstable performance or have strict prerequisites. Therefore, combining them may bring new solutions for real SISR.

6.6 Efficient and Accurate Scale Arbitrary SISR

SISR has seen its applications in diverse real-life scenarios and users. Therefore, it is necessary to develop a flexible and universal scale arbitrary SISR model that can be adapted to any scale, including asymmetric and non-integer scale factors. Currently, most DL-based SISR models can only be applied to one or a limited number of multiple upsampling

factors. Although a few scale arbitrary SISR methods have also been proposed, they tend to lack the flexibility to use and the simplicity to be implemented, which greatly limits their application scenarios. Therefore, exploring a CNN-based accurate scale arbitrary SISR model as simple and flexible as Bicubic is crucial to the spread of SISR technology.

6.7 Consider the Characteristics of Different Images

Although a series of models have been proposed for domain-specific applications, most of them directly transfer the SISR methods to these specific fields. This is the simplest and feasible method, but it will also inhibit the model performance since they ignore the data structure characteristics of the domain-specific images. Therefore, fully mining and using the potential prior and data characteristics of the domain-specific images is beneficial for efficient and accurate domain-specific SISR models construction. In the future, it will be a trend to further optimize the existing SISR models based on the prior knowledge and the characteristics of the domain-specific images.

7 CONCLUSION

In this survey, we have given a comprehensive overview of DL-based single image super-resolution methods according to their targets, including reconstruction efficiency, reconstruction accuracy, perceptual quality, and other technologies that can further improve model performance. Meanwhile, we provided a detailed introduction to the related works of SISR and introduced a series of new tasks and domain-specific applications extended by SISR. In order to view the performance of each model more intuitively, we also provided a detailed comparison of reconstruction results. Moreover, we provided some underlying problems in SISR and introduced several new trends and future directions worthy of further exploration. We believe that the survey can help researchers better understand this field and further promote the development of this field.

REFERENCES

- [1] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology and Climatology*, vol. 18, 1979.
- [2] Jian Sun, Zongben Xu, and Heung-Yeung Shum, "Image super-resolution using gradient profile prior," in *CVPR*, 2008.
- [3] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, 2010.
- [4] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *CVPR*, 2004.
- [5] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, 2010.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, 2015.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012.
- [8] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML*, 2008.
- [9] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.
- [10] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, 2011.
- [11] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, 2019.
- [12] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, 2020.
- [13] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*, 2016.
- [14] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, 2016.
- [15] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *CVPRW*, 2017.
- [16] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPRW*, 2017.
- [17] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.
- [18] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, 2011.
- [19] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *ICCV*, 2019.
- [20] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *CVPR*, 2018.
- [21] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, "Camera lens super-resolution," in *CVPR*, 2019.
- [22] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *ICCVW*, 2019.
- [23] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *CVPR*, 2019.
- [24] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020.
- [25] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on non-negative neighbor embedding," in *BMVC*, 2012.
- [26] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *ICCS*, 2010.
- [27] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.
- [28] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *MANPU*, 2016.
- [29] R. Timofte, R. Rothe, and L. Van Gool, "Seven ways to improve example-based single image super resolution," in *CVPR*, 2016.
- [30] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *ECCVW*, 2018.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
- [32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.
- [33] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016.
- [34] A. Shocher, N. Cohen, and M. Irani, "?zero-shot? super-resolution using deep internal learning," in *CVPR*, 2018.
- [35] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *CVPRW*, 2018.
- [36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
- [37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [38] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-

- realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.
- [39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.
- [40] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *CVPR*, 2018.
- [41] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, 2009.
- [42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, 2004.
- [43] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, 2012.
- [44] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a ?completely blind? image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, 2012.
- [45] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, 2017.
- [46] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *ICCV*, 2019.
- [47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.
- [48] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *CVPR*, 2015.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [50] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.
- [51] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.
- [52] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *ECCV*, 2018.
- [53] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
- [54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.
- [55] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *CVPR*, 2016.
- [56] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017.
- [57] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *CVPR*, 2017.
- [58] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *ECCV*, 2018.
- [59] J. Li, Y. Yuan, K. Mei, and F. Fang, "Lightweight and accurate recursive fractal network for image super-resolution," in *ICCVW*, 2019.
- [60] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *ICCV*, 2017.
- [61] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *CVPR*, 2018.
- [62] N. Ahn, B. Kang, and K.-A. Sohn, "Image super-resolution via progressive cascading residual network," in *CVPRW*, 2018.
- [63] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "Mdcn: Multi-scale dense cross network for image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, 2020.
- [64] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*, 2017.
- [65] 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 *CVPRW*, 2018.
- [66] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *CVPR*, 2019.
- [67] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.
- [68] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017.
- [69] J. Qin, Y. Huang, and W. Wen, "Multi-scale feature fusion residual network for single image super-resolution," *Neurocomputing*, vol. 379, 2020.
- [70] C.-Y. Chang and S.-Y. Chien, "Multi-scale dense network for single-image super-resolution," in *ICASSP*, 2019.
- [71] F. Cao and H. Liu, "Single image super-resolution via multi-scale residual channel attention network," *Neurocomputing*, vol. 358, 2019.
- [72] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.
- [73] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018.
- [74] K. Mei, A. Jiang, J. Li, B. Liu, J. Ye, and M. Wang, "Deep residual refining based pseudo-multi-frame network for effective single image super-resolution," *IET Image Processing*, vol. 13, 2019.
- [75] M. Shen, P. Yu, R. Wang, J. Yang, L. Xue, and M. Hu, "Multipath feedforward network for single image super-resolution," *Multimedia Tools and Applications*, vol. 78, 2019.
- [76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.
- [77] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.
- [78] K. Mei, A. Jiang, J. Li, J. Ye, and M. Wang, "An effective single-image super-resolution model using squeeze-and-excitation networks," in *NeurIPS*, 2018.
- [79] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.
- [80] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," *arXiv preprint arXiv:1806.02919*, 2018.
- [81] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.
- [82] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *ECCV*, 2020.
- [83] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *CVPR*, 2020.
- [84] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *CVPR*, 2016.
- [85] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *ICCV*, 2015.
- [86] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for single image super-resolution," *arXiv preprint arXiv:1904.05677*, 2019.
- [87] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *CVPR*, 2018.
- [88] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Transactions on Image Processing*, vol. 26, 2017.
- [89] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, 2020.
- [90] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *CVPR*, 2018.
- [91] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.
- [92] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," 2015.
- [93] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
- [94] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.
- [95] M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, "Srobb: Targeted perceptual loss for single image super-resolution," in *ICCV*, 2019.

- [96] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [97] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *ECCV*, 2018.
- [98] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *ECCV*, 2018.
- [99] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.
- [100] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *CVPR*, 2020.
- [101] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *ECCV*, 2018.
- [102] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, "Closed-loop matters: Dual regression networks for single image super-resolution," in *CVPR*, 2020.
- [103] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *CVPR*, 2011.
- [104] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *ICCV*, 2019.
- [105] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*, 2017.
- [106] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [107] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *CVPR*, 2019.
- [108] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
- [109] Q. Gao, Y. Zhao, G. Li, and T. Tong, "Image super-resolution using knowledge distillation," in *ACCV*, 2018.
- [110] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *ECCV*, 2020.
- [111] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving web images," *IEEE Transactions on Image Processing*, vol. 22, 2013.
- [112] 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 *ECCV*, 2018.
- [113] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *CVPR*, 2020.
- [114] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [115] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [116] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, 2019.
- [117] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *CVPR*, 2021.
- [118] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCVW*, 2021.
- [119] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [120] Z. Lu, H. Liu, J. Li, and L. Zhang, "Efficient transformer for single image super-resolution," *arXiv preprint arXiv:2108.11084*, 2021.
- [121] Y. Shi, H. Zhong, Z. Yang, X. Yang, and L. Lin, "Ddet: Dual-path dynamic enhancement network for real-world image super-resolution," *IEEE Signal Processing Letters*, vol. 27, 2020.
- [122] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *ECCV*, 2020.
- [123] W. Sun, D. Gong, Q. Shi, A. van den Hengel, and Y. Zhang, "Learning to zoom-in via learning to zoom-out: Real-world super-resolution by generating and adapting degradation," *IEEE Transactions on Image Processing*, vol. 30, 2021.
- [124] K. Prajapati, V. Chudasama, H. Patel, K. Upla, R. Ramachandra, K. Raja, and C. Busch, "Unsupervised single image super-resolution network (usisresnet) for real-world data using generative adversarial network," in *CVPRW*, 2020.
- [125] G. Kim, J. Park, K. Lee, J. Lee, J. Min, B. Lee, D. K. Han, and H. Ko, "Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator," in *CVPRW*, 2020.
- [126] J. Kim, C. Jung, and C. Kim, "Dual back-projection-based internal learning for blind super-resolution," *IEEE Signal Processing Letters*, vol. 27, 2020.
- [127] M. Emad, M. Peemen, and H. Corporaal, "Dualsr: Zero-shot dual learning for real-world super-resolution," in *WACV*, 2021.
- [128] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *CVPR*, 2020.
- [129] S. Park, J. Yoo, D. Cho, J. Kim, and T. H. Kim, "Fast adaptation to super-resolution networks via meta-learning," in *ECCV*, 2020.
- [130] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *CVPR*, 2019.
- [131] L. Wang, Y. Wang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning a single network for scale-arbitrary super-resolution," 2021.
- [132] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *CVPR*, 2018.
- [133] —, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *CVPR*, 2019.
- [134] Y.-S. Xu, S.-Y. R. Tseng, Y. Tseng, H.-K. Kuo, and Y.-M. Tsai, "Unified dynamic convolutional network for super-resolution with variational degradations," in *CVPR*, 2020.
- [135] Z. Hui, J. Li, X. Wang, and X. Gao, "Learning the non-differentiable optimization for blind super-resolution," in *CVPR*, 2021.
- [136] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *CVPR*, 2020.
- [137] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *CVPR*, 2019.
- [138] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," *arXiv preprint arXiv:2010.02631*, 2020.
- [139] S. Y. Kim, H. Sim, and M. Kim, "Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment," in *CVPR*, 2021.
- [140] M. Yamac, B. Ataman, and A. Nawaz, "Kernelnet: A blind super-resolution kernel estimation network," in *CVPR*, 2021.
- [141] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *CVPR*, 2021.
- [142] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," *arXiv preprint arXiv:2107.10833*, 2021.
- [143] X. Hu, Z. Zhang, C. Shan, Z. Wang, L. Wang, and T. Tan, "Met/usr: A unified super-resolution network for multiple degradation parameters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, 2020.
- [144] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, 2018.
- [145] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep residual squeeze and excitation network for remote sensing image super-resolution," *Remote Sensing*, vol. 11, 2019.
- [146] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, 2020.
- [147] X. Dong, L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang, "Remote sensing image super-resolution using second-order multi-scale networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, 2020.
- [148] L. J. Rickard, R. W. Basedow, E. F. Zalewski, P. R. Silvergate, and M. Landers, "Hydice: An airborne system for hyperspectral imaging," in *Imaging Spectrometry of the Terrestrial Environment*, vol. 1937, 1993.
- [149] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3d full convolutional neural network," *Remote Sensing*, vol. 9, 2017.

- [150] Y. Li, L. Zhang, C. Ding, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *BigMM*, 2018.
- [151] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized rgb guidance," in *CVPR*, 2019.
- [152] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, 2020.
- [153] D. Liu, J. Li, and Q. Yuan, "A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, 2021.
- [154] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *ECCV*, 2020.
- [155] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, 2017.
- [156] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, 2018.
- [157] Y. Wang, J. Yang, L. Wang, X. Ying, T. Wu, W. An, and Y. Guo, "Light field image super-resolution using deformable convolution," *IEEE Transactions on Image Processing*, vol. 30, 2020.
- [158] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *AAAI*, 2015.
- [159] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *ECCV*, 2016.
- [160] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *CVPR*, 2017.
- [161] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *ECCV*, 2018.
- [162] B. Dogan, S. Gu, and R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," in *CVPRW*, 2019.
- [163] G. Gao, D. Zhu, H. Lu, Y. Yu, H. Chang, and D. Yue, "Robust facial image super-resolution by kernel locality-constrained coupled-layer regression," *ACM Transactions on Internet Technology*, vol. 21, 2021.
- [164] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li, "Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network," in *MICCAI*, 2018.
- [165] Y. Wang, Q. Teng, X. He, J. Feng, and T. Zhang, "Ct-image of rock samples super resolution using 3d convolutional neural network," *Computers & Geosciences*, vol. 133, 2019.
- [166] X. Zhao, Y. Zhang, T. Zhang, and X. Zou, "Channel splitting network for single mr image super-resolution," *IEEE Transactions on Image Processing*, vol. 28, 2019.
- [167] C. Peng, W.-A. Lin, H. Liao, R. Chellappa, and S. K. Zhou, "Saint: spatially aware interpolation network for medical slice synthesis," in *CVPR*, 2020.
- [168] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *CVPR*, 2018.
- [169] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *CVPR*, 2019.
- [170] L. Wang, Y. Guo, Y. Wang, Z. Liang, Z. Lin, J. Yang, and W. An, "Parallax attention for unsupervised stereo correspondence learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [171] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Processing Letters*, vol. 27, 2020.
- [172] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," in *CVPR*, 2021.
- [173] Q. Dai, J. Li, Q. Yi, F. Fang, and G. Zhang, "Feedback network for mutually boosted stereo image super-resolution and disparity estimation," *arXiv preprint arXiv:2106.00985*, 2021.
- [174] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.
- [175] C. Wang, Z. Li, and J. Shi, "Lightweight image super-resolution with adaptive weighted learning network," *arXiv preprint arXiv:1904.02358*, 2019.
- [176] A. Muqeet, J. Hwang, S. Yang, J. Kang, Y. Kim, and S.-H. Bae, "Multi-attention based ultra lightweight image super-resolution," in *ECCV*, 2020.
- [177] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *CVPR*, 2020.
- [178] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *ACMMM*, 2019.
- [179] Z. Wang, G. Gao, J. Li, Y. Yu, and H. Lu, "Lightweight image super-resolution with multi-scale feature interaction network," in *ICME*, 2021.
- [180] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "Madnet: A fast and lightweight network for single-image super resolution," *IEEE Transactions on Cybernetics*, vol. 51, 2020.
- [181] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," *arXiv preprint arXiv:2006.16673*, 2020.
- [182] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *CVPR*, 2021.
- [183] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
- [184] Y. Ma, H. Xiong, Z. Hu, and L. Ma, "Efficient super resolution using binarized neural network," in *CVPRW*, 2019.
- [185] H. Li, C. Yan, S. Lin, X. Zheng, B. Zhang, F. Yang, and R. Ji, "Pams: Quantized super-resolution via parameterized max scale," in *ECCV*, 2020.



Juncheng Li received the Ph.D. degree from the School of Computer Science and Technology, East China Normal University (ECNU), China, in 2021. He is currently a Postdoctoral Fellow at the Center for Mathematical Artificial Intelligence (CMAI), The Chinese University of Hong Kong (CUHK). His research interests include artificial intelligence and its applications to computer vision and image processing (e.g., image super-resolution, image denoising, image deblurring, image dehazing, and image enhancement).



Zehua Pei is now an undergraduate student at the Department of Mathematics, The Chinese University of Hong Kong (CUHK). He will receive his B.S. degree in 2022. His research interests include data mining, computer vision, optimization, and image processing (e.g., image super-resolution, image denoising, image dehazing, and image enhancement).



Tieyong Zeng received the B.S. degree from Peking University, Beijing, China, in 2000, the M.S. degree from cole Polytechnique, Palaiseau, France, in 2004, and the Ph.D. degree from the Universit of Paris XIII, Paris, France, in 2007. He is currently a Professor at the Department of Mathematics, The Chinese University of Hong Kong (CUHK). He is also the Director of the Center for Mathematical Artificial Intelligence, CUHK. He has published more than 100 papers. His research interests include image processing, optimization, artificial intelligence, scientific computing, computer vision, machine learning, and inverse problems.