A Systematic Survey of Deep Learning-based Single-Image Super-Resolution

JUNCHENG LI, Shanghai University, China
ZEHUA PEI, The Chinese University of Hong Kong, China
WENJIE LI, Beijing University of Posts and Telecommunications, China
GUANGWEI GAO, Nanjing University of Posts and Telecommunications, China
LONGGUANG WANG, Aviation University of Air Force, China
YINGQIAN WANG, National University of Defense Technology, China
TIEYONG ZENG, The Chinese University of Hong Kong, China

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 design targets. 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 at https://github.com/CV-JunchengLi/SISR-Survey.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Reconstruction**; **Neural networks**.

Additional Key Words and Phrases: Image super-resolution, single-image super-resolution, SISR, survey.

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 academia and industry. As shown in Fig. 1, SISR aims to reconstruct a high-resolution (HR) image from its degraded low-resolution (LR) one. It is widely used in various computer vision applications, including security and surveillance images, medical image reconstruction, video enhancement, and image segmentation.

Many SISR methods have been studied long before, such as bicubic interpolation and Lanczos resampling [42], which are based on interpolation. However, SISR is an inherently ill-posed problem, and multiple HR images corresponding to the same LR image always exist. To solve this issue, some

Authors' addresses: Juncheng Li, Shanghai University, Shanghai, China, junchengli@shu.edu.cn; Zehua Pei, The Chinese University of Hong Kong, Hong Kong, China, pzehua2000@gmail.com; Wenjie Li, Beijing University of Posts and Telecommunications, Beijing, China, lewj2408@gmail.com; Guangwei Gao, Nanjing University of Posts and Telecommunications, Nanjing, China, csggao@gmail.com; Longguang Wang, Aviation University of Air Force, Changchun, China, wanglongguang15@nudt.edu.cn; Yingqian Wang, National University of Defense Technology, Changsha, China, wangyingqian16@nudt.edu.cn; Tieyong Zeng, The Chinese University of Hong Kong, Hong Kong, China, zeng@math.cuhk.edu.hk.

0360-0300/2024/2-ART $15.00
https://doi.org/

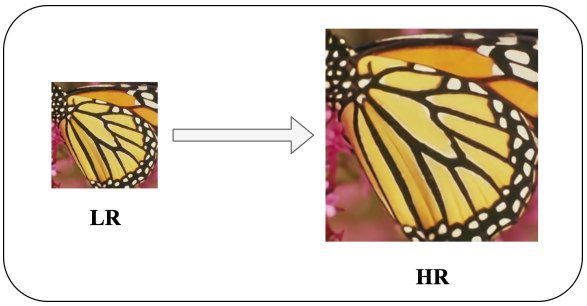


Fig. 1. SISR aims to reconstruct a high-resolution (HR) image from its degraded low-resolution (LR) one.

numerical methods (e.g., edge-based methods [77] and image statistics-based methods [85]) utilize prior information to restrict the solution space. Meanwhile, there are some widely used learning-based methods, such as neighbor embedding methods [18] and sparse coding methods [202], which learn a transformation between LR and HR patches.

Recently, deep learning (DL) [90] has demonstrated better performance than traditional machine learning models in many artificial intelligence fields, including computer vision [86] and natural language processing [31]. 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 into three steps: feature extraction and representation, non-linear mapping, and image reconstruction [38]. 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 expenses [40]. 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 shows 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 previous SISR surveys [6, 190] 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 image super-resolution methods and categorize them according to their specific targets. In Fig. 2, we show the content and taxonomy of this survey. We divide these methods into three categories: Simulation SISR, Real-World SISR, and Domain-Specific Applications. Additionally, we divide Simulation SISR methods into three categories: Efficient Network / Mechanism Design Methods, Perceptual Quality Methods, and Additional Information Utilization Methods, according to their specific targets. 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:

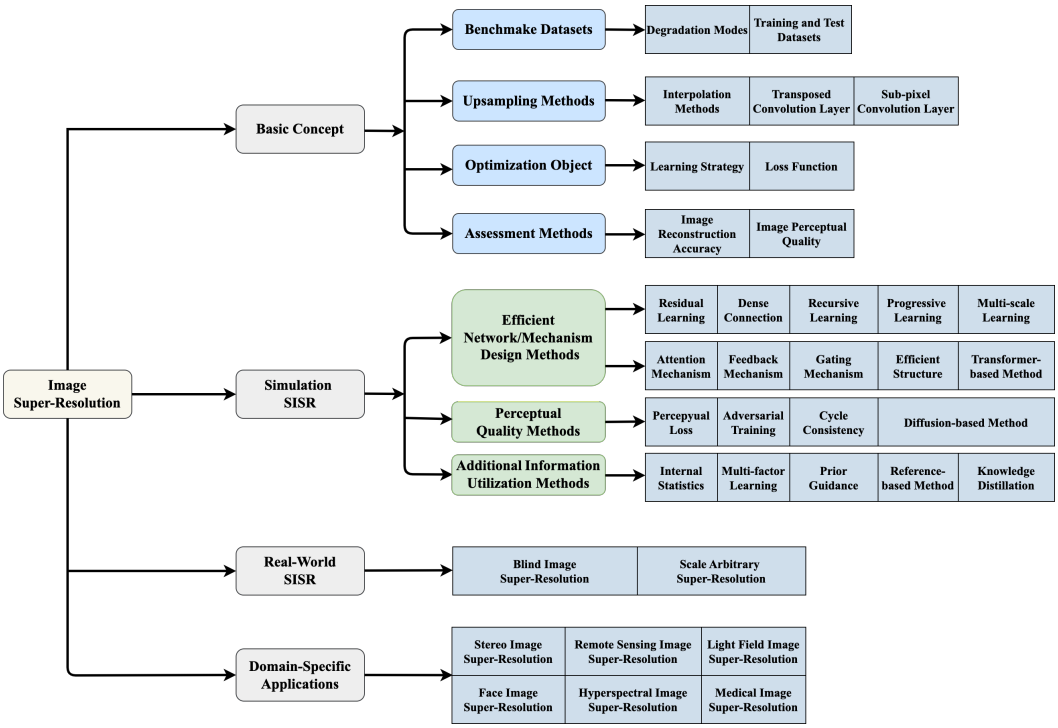


Fig. 2. The content and taxonomy of this survey. In this survey, we divide image super-resolution methods into three categories: Simulation SISR, Real-World SISR, and Domain-Specific Applications.

(1). We give a thorough overview of DL-based SISR methods according to their targets. This is a new perspective that makes the survey clear in context and convenient.

(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 look forward to the future trend and direction of SR.

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 input LR images. Compared with MISR, SISR is much more challenging since MISR has 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 an SISR framework, the LR image I_x is modeled as $I_x = \mathcal{D}(I_y; \theta_{\mathcal{D}})$, where $\mathcal{D}$ is a degradation map $\mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{h \times w}$ and $\theta_{\mathcal{D}}$ 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:

$$\mathcal{D}(I_y;\theta_{\mathcal{D}}) = (I_y \otimes \kappa)\downarrow_s + n, \tag{1}$$

where $I_y \otimes \kappa$ represents the convolution between the blur kernel κ and the HR image I_y, $\downarrow_s$ is a subsequent downsampling 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 an 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, SISR task can be transformed into the following optimization goal:

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

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 weight of the regularization term.

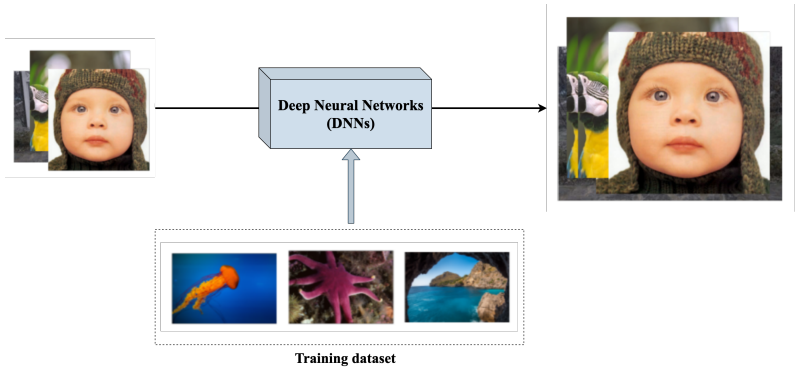


Fig. 3. The training process of data-driven based deep neural networks.

2.2 Benchmark Datasets

Data is always essential for data-driven models, especially in 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 ***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 ones. 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, the 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,

Table 1. Benchmarks datasets for SISR.

Name	Usage	Amount	Format	Description
General-100 [39]	Train	100	BMP	Common images with clear edges but fewer smooth regions
T91 [202]	Train	91	PNG	Common Images
WED [127]	Train	4744	MAT	Common images
Flickr2K [160]	Train	2650	PNG	2K images from Flickr
FFHQ [82]	Train	70000	PNG	A high-quality image dataset of human faces
CelebA-HQ [92]	Train/Val	30000	PNG	A GAN Synthetic data of human faces
CelebA [118]	Train/Val	202600	JPG	a large-scale face attributes dataset
DRealSR [194]	Train/Val	31970	PNG	a benchmark with diverse real-world degradation processes
DIV2K [2]	Train/Val	1000	PNG	High-quality dataset for CVPR NTIRE competition
BSDS300 [130]	Train/Val	300	JPG	Common images
BSDS500 [7]	Train/Val	500	JPG	Common images
RealSR [12]	Train/Val	100	PNG	100 real-world low and high resolution image pairs
OutdoorScene [176]	Train/Val	10624	PNG	Images of outdoor scenes
City100 [20]	Train/Test	100	RAW	Common images
Flickr1024 [184]	Train/Test	100	RAW	Stereo images used for Stereo SR
SR-RAW [230]	Train/Test	7*500	JPG/ARW	Raw images produced by real-world computational zoom
PIPAL [79]	Test	200	PNG	Perceptual image quality assessment dataset
Set5 [8]	Test	5	PNG	Common images, only 5 images
Set14 [214]	Test	14	PNG	Common images, only 14 images
BSD100 [130]	Test	100	JPG	A subset of BSDS500 for testing
Urban100 [73]	Test	100	PNG	Images of real-world structures
Manga109 [46]	Test	109	PNG	Japanese manga
L20 [161]	Test	20	PNG	Common images, very high-resolution
PIRM [9]	Test	200	PNG	Common images, datasets for ECCV PIRM competition

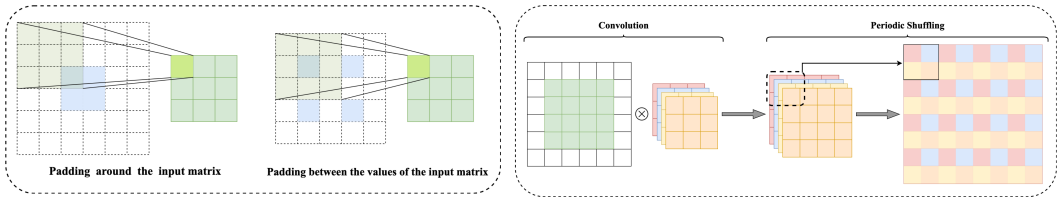
which is essentially a bicubic downsampling operation. For **BD**, the HR images are blurred by a Gaussian kernel of size 7×7 with standard deviation *1.6* and then downsampled with a scaling factor of 3. To obtain LR images under DN mode, the Bicubic downsampling is performed on the HR image with a scaling factor of 3, and then the Gaussian noise with a noise level of 30 is added to the image.

2.2.2 ***Training and Test Datasets.*** Recently, many datasets for the SISR task have been proposed, including BSDS300 [130], DIV2K [2], and Flickr2K [160]. Meanwhile, there are also many test datasets that can be used to effectively test the performance of the models, such as Set5 [8], Set14 [214], Urban100 [73], and Manga109 [46]. In Table 1, we list a series of commonly used datasets and indicate their detailed attribute.

Among these datasets, DIV2K [2] 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 [12] is the first truly collected real-world 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 [34] and CelebA [118]. In addition, combining multiple datasets (e.g., DF2K) for training to further improve the model performance has been also widely used.

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 introduce several upsampling methods that support these upsampling mechanisms.



(a) Two kinds of transposed convolutional layers (b) Principle of the sub-pixel convolutional layer

Fig. 4. Upsampling methods: (a) transposed convolutional layers (b) sub-pixel convolutional layer.

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 DL-based SISR models.

2.3.2 ***Transposed Convolutional Layers.*** As shown in Fig. 4 (a), researchers usually consider two kinds of transposed convolution operations: one adds padding around the input matrix and then applies the convolution operation, and 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 performing convolution with a sub-pixel level stride. In the transposed convolutional layer, the upsampling level is controlled by the size of the padding, which is essentially opposite to the operation of the normal convolutional layer. The transposed convolutional layer is first proposed in FSRCNN [39] and is widely used in DL-based SISR models.

2.3.3 ***Sub-pixel Convolutional Layer.*** In ESPCN [151], 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. 4 (b)). In detail, the formulation of the sub-pixel convolutional layer can be defined as follows:

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

where $\mathcal{PS}$ denotes the periodic shuffling operator, which transfers an $h \times w \times C \cdot r^2$ tensor to a tensor of shape $rh \times rw \times C$, and $rh \times rw$ is explicitly the size of the HR image, C is the number of channels. In addition, the convolutional filter W_L has the shape $n_{L-1} \times r^2C \times K_L \times K_L$, 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 exhibits better efficiency and thus is 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.*** In this work, we use a common division method in the SR field, that is, whether paired LR-HR images are used for model training. It is worth noting that the HR image here refers to the additional introduced high-resolution image, not the image itself. In addition, learning strategy has no clear definitions in SISR. According to this criteria, the DL-based SISR models can be mainly divided into supervised learning methods and unsupervised learning methods.

Supervised Learning: In SISR, we often call the method of using pairs of LR-HR images for training a supervised learning paradigm. In simulated SISR, LR images are often obtained by downsampling HR images. In real SISR, LR images and HR images are obtained by adjusting the zoom of the camera. In general, the LR and HR images of this type of method have a one-to-one correspondence, and researchers compute the reconstruction error between the ground-truth image I_y and the reconstructed image I_{SR}:

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

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 other space and then compute the error:

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

Among them, $\mathcal{L}$ is the loss function that is used to minimize the distance between the reconstructed image and the ground-truth image. By using different loss functions, the model can achieve different performance. Therefore, an effective loss function is also crucial for SISR.

Unsupervised Learning: The simulated paired images have poor versatility, while the real paired images are difficult to collect. To address this issue, some methods began to try to no longer use paired LR-HR images for training. We often call this type of method an unsupervised learning method. This type of unsupervised method no longer uses paired LR-HR images for training but uses unpaired LR-HR images (GAN-based method) or itself (self-supervised learning method) for training. For example, ZSSR [152] uses the test image and its downscaling versions with the data augmentation approaches to build the "training dataset" and then applies the loss function to optimize the model. In addition, weakly-supervised learning also belongs to the unsupervised learning strategy. Among them, some researchers first learn the HR-to-LR degradation and use it to construct datasets for training the model, while other researchers design cycle-in-cycle models to learn the LR-to-HR and HR-to-LR mappings simultaneously. For instance, CinCGAN [210] consists of two CycleGAN [238], where one cycle is adopted for translating between the real LR and synthetic LR images while the other is used between the real LR and HR images.

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 combining multiple loss functions can better reflect the situation of image restoration. In this section, we briefly introduce several commonly used loss functions.

Pixel Loss: Pixel loss is the simplest and most popular loss function in SISR, which aims to measure the difference between two images on a 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 the L1 loss):

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

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

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

where, h, w, and c are the height, width, and the number of channels of the image. ϵ is a numerical stability constant, usually being set to 10^{-3}. Since most mainstream image evaluation indicators are highly correlated with pixel-by-pixel differences, pixel loss is still widely used. However, the images reconstructed by this type of loss function usually lack high-frequency details and thus perform inferior in visual effects.

Content Loss: Content loss is also termed 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:

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

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. By using this loss, the visual effects of these two images can be as consistent as possible. Among them, VGG [153] and ResNet [91] are the most commonly used pre-training classification networks.

Adversarial Loss: To make the reconstructed SR image more realistic, Generative Adversarial Networks (GANs [56]) have been introduced into the SISR task. 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 [91]:

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

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:

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

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

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.

Fourier Space Loss: The design of perceptual losses predominantly focuses on the spatial domain. However, SR is tightly coupled to the frequency domain, as only high frequencies are removed during the downsampling process. To solve this problem, Fuoli *et al.* [47] propose a novel Fourier Space Loss by calculating the frequency components with the Fast Fourier Transform (FFT) for direct emphasis on the frequency content. Firstly, the image is transformed into Fourier space by applying the Fast Fourier transform (FFT). Then, the method calculates the amplitude difference F_f, $|.|$ and phase difference, $\angle$ of all frequency components between output image and ground truth image. The averaged differences are computed as the total frequency loss as follows:

$$L_f, |.| = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \Big| \mid \hat{Y} \mid_{u,v} - \mid Y \mid_{u,v} \Big|, \tag{13}$$

$$L_f, \angle = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \Big| \angle \hat{Y}_{u,v} - \angle Y_{u,v} \Big|, \tag{14}$$

$$L_f = \frac{1}{2} L_f, |.| + \frac{1}{2} L_f, \angle, \tag{15}$$

where $\hat{Y}_{u,v}$ represents the spectrum of the recovered image, and $Y_{u,v}$ represents the spectrum of the ground truth image.

Mixed Loss: In SISR, there are also some classic combinations of loss functions that are widely used to guide the network towards generating high-quality HR images. These combinations aim to balance the quality, details, and visual perception of the generated image. Here are some commonly used classic combinations of loss functions.

L1 + Perceptual Loss: combining L1 loss with perceptual loss, such as the feature loss based on VGG networks, can generate images that are clearer and have better details. This combination can effectively reduce noise and distortion in the image; L1 + TV Loss: combining L1 loss with total variation (TV) loss can generate images with good edge and texture details. TV loss helps to reduce blocky artifacts in the image; Content Loss + Adaptive Loss: combining Content loss with adaptive loss can generate images with better visual coherence. Adaptive loss can adjust the loss weights based on the content of the image.

The choice of loss function combinations depends on the specific requirements of the SISR task, such as the desired balance between perceptual quality and computational efficiency. In practical applications, researchers may adjust the weights of the loss functions based on experimental results to find the combination that best suits a specific task.

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 becoming 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 are more related to evaluating 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 used for image reconstruction accuracy evaluation 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 [10] 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 $I_{SR} \in \mathbb{R}^{H \times W}$:

$$MSE = \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (I_y(i,j) - I_{SR}(i,j))^2, \tag{16}$$

$$PSNR = 10 \cdot \log_{10}(\frac{MAX^2}{MSE}), \tag{17}$$

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 [188].

Structural Similarity Index Measure (SSIM): SSIM [189] is another popular assessment method that measures the similarity between two images on a perceptual basis, including structures, luminance, and contrast. Different from PSNR, which calculates absolute pixel-level errors, SSIM suggests that there exist strong inter-dependencies among the spatially adjacent pixels. These dependencies carry important information related to the structures perceptually. Therefore, the SSIM can be expressed as a weighted combination of three comparative measures:

$$\begin{aligned} SSIM(I_{SR}, I_y) &= (l(I_{SR}, i_y)^\alpha \cdot c(I_{SR}, I_y)^\beta \cdot s(I_{SR}, I_y)^\gamma) \\ &= \frac{(2\mu_{I_{SR}}\mu_{I_y} + c_1)(2\sigma_{I_{SR}I_y} + c_2)}{(\mu_{I_{SR}}^2 + \mu_{I_y}^2 + c_1)(\sigma_{I_{SR}}^2 + \sigma_{I_y}^2 + c_2)}, \end{aligned} \tag{18}$$

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 value, variance, and covariance of the corresponding items, respectively.

A higher SSIM indicates higher similarity between two images, which has been widely used due to its convenience and stable performance in evaluating 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 that pursue absolutely similar 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 [10]. 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, several volunteers rate their opinions on the quality of a set of images by Double-stimulus [135], i.e., every volunteer has both the source and test images. After all the volunteers finish 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 since 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.

Learned Perceptual Image Patch Similarity (LPIPS): LPIPS [220] is a popular metric used to measure the perceived differences between differnet images, which not only focuses on the structure and content of an image but also reflects the sensitivity of the human eye to image differences. Specifically, the feature layers of different images are first extracted using a pre-trained model(e.g., VGG [153]), and then the LPIPS value can be obtained by calculating the weighted summed distance between the different feature spaces:

$$LPIPS(I_{SR}, I_y) = \sum_{l=1}^{N} \left\| \omega_l . \left(\phi_l \left(I_{SR} \right), \phi_l \left(I_y \right) \right) \right\|_2, \tag{19}$$

where l is the lth feature layer of the pre-trained model, N is the total number of feature layers of the pre-trained model, ω_l is the weight used to weigh the lth feature layers, ϕ_l is the lth feature extraction layer in the pre-trained model, and $\|\|_2$ is the L2 paradigms. However, LPIPS is obtained by learning from DL models, so its performance is affected by the training data, leading to the fact that LPIPS may lack generalization ability in some cases.

Deep Image Structure and Texture Similarity (DISTS): DISTS [36] is the first complete reference image quality model that explicitly tolerates texture resampling, and it utilizes injective differentiable functions constructed from CNN to convert images to a multi-scale hyper-complete representation, a representation in which the spatial average of the feature maps captures the texture appearance and matches human ratings of image quality.

$$l(I_{SR}^{(i)}, I_y^{(i)}) = \frac{2\mu_{I_{SR}}^{(i)} \mu_{I_y}^{(i)} + c_1}{\left(\mu_{I_{SR}}^{(i)}\right)^2 + \left(\mu_{I_y}^{(i)}\right)^2 + c_1}, \tag{20}$$

$$s(I_{SR}^{(i)}, I_y^{(i)}) = \frac{2\sigma_{I_{SR}I_y}^{(i)} + c_2}{\left(\sigma_{I_{SR}}^{(i)}\right)^2 + \left(\sigma_{I_y}^{(i)}\right)^2 + c_2}, \tag{21}$$

$$DISTS(I_{SR}, I_y, \alpha, \beta) = 1 - \sum_{i=0}^{m} \sum_{j=1}^{n_i} \left(\alpha_{ij} l(I_{SR}^{(i)}, I_y^{(i)}) + \beta_{ij} s(I_{SR}^{(i)}, I_y^{(i)}) \right), \tag{22}$$

where $\mu_{I_{SR}}^{(i)}$, $\mu_{I_y}^{(i)}$, $\sigma_{I_{SR}}^{(i)}$, $\sigma_{I_y}^{(i)}$ and $\sigma_{I_{SR}I_y}^{(i)}$ denote average value and variances of $I_{SR}^{(i)}$ and $I_y^{(i)}$, and covariance between $I_{SR}^{(i)}$ and $I_y^{(i)}$, respectively. c_1 and c_2 are two small positive constants. And $\{\alpha_{ij}, \beta_{ij}\}$ are learnable weights, satisfying $\sum_{i=0}^{m} \sum_{j=1}^{n_j} (\alpha_{ij} + \beta_{ij}) = 1$. And despite its beneficial mathematical properties, the DISTS metric is still highly non-convex and therefore requires more iterations to recover from random noise using stochastic gradient descent methods than metrics such as SSIM.

Natural Image Quality Evaluator (NIQE): NIQE [136] 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 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:

$$D(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{((\nu_1 - \nu_2)^T (\frac{\Sigma_1 + \Sigma_2}{2})^{-1} (\nu_1 - \nu_2))}, \tag{23}$$

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, respectively. Notice that, a higher NQIE index indicates lower image perceptual quality. Compared with MOS, NIQE is a more convenient perceptual evaluation method.

Ma: Ma *et al.* [126] 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 images. It learns from perceptual scores based on human subject studies involving a large number of SR images. Then, it quantifies the SR artifacts through three types of statistical properties, i.e., local/global frequency variations and spatial discontinuity. Afterward, 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 is still limited as 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.

Perception Index (PI): In the 2018 PIRM Challenge on Perceptual Image Super-Resolution [9], perception index (PI) is first proposed to evaluate perceptual quality. It is a combination of the no-reference image quality measures Ma and NIQE:

$$PI = \frac{1}{2}((10 - Ma) + NIQE). \tag{24}$$

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.* [221] 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.* [220] 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 judgment in this dataset. Ramsauer *et al.* [66] proposed Fréchet Inception Distance (FID), which quantifies the quality of SISR images by comparing the difference between the data distribution of SISR results and the true data distribution. In summary, how to measure the perceptual quality of SR images more accurately and efficiently is an important issue that needs to be explored.

3 IMAGE SUPER-RESOLUTION

In 2014, Dong *et al.* [38] 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 SR 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 SR. After that, CNN-based SR methods are blooming and constantly refreshing the best results.

In this part, we divide DL-based imgae super-resolution methods into three categories: Simulation SISR, Real-World SISR, and Domain-Specific Applications.

3.1 Simulation SISR

In recent years, the field of SISR has developed rapidly, and a large number of excellent models have emerged. However, it is worth noting that most of these models use simulated datasets for testing and training, we call this method simulated SISR. In other words, the low-resolution images used in this type of method are usually obtained by applying some fixed degradation

modes to the high-resolution images. This will affect the performance of the model in practical applications. However, it is undeniable that the emergence of these methods has enriched and promoted the development of SISR. According to different design targets, we divide these methods into three categories: efficient network/mechanism design methods, perceptual quality methods, and additional information utilization methods.

3.1.1 ***Efficient Network / Mechanism Design Methods.*** Most of the methods that have emerged in recent years focus on efficient and accurate network structure and mechanism design, which enable the model to achieve better performance with fewer parameters. In this section, we will discuss some methods that contribute to efficient and accurate network design.

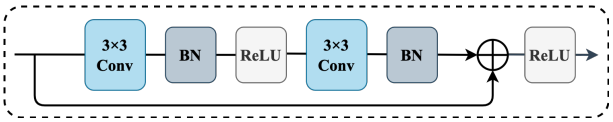


Fig. 5. Sketch of residual learning architecture / residual block.

Residual Learning: In SRCNN, researchers find that better results 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 problems [64]. Meanwhile, adding more layers will lead to a higher training error and more expensive computational costs.

In ResNet [65], He *et al.* proposed a residual learning framework, where a residual mapping is desired instead of fitting the whole underlying mapping (Fig. 5). In SISR, as the 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 *et al.* [83] proposed a very deep super-resolution network, also known as VDSR. For the convenience of network design, the residual block [65] has gradually become the basic unit in the network structure. The convolutional branch, 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 Lim *et al.* [108] point 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 reduces 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, 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 help gradient flow. Furthermore, many new feature extraction modules have introduced local residual learning to reinforce strong learning capabilities [98, 225]. Of course, combining local residual learning and global residual learning is also highly popular now [91, 108, 225].

Residual Scaling: In EDSR, Lim *et al.* [108] 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 technique [156], 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.

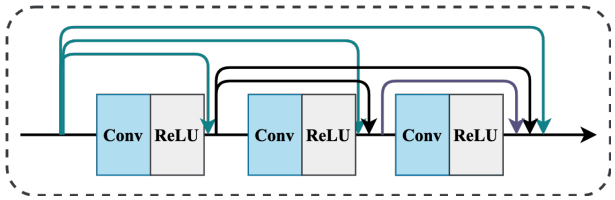


Fig. 6. The structure of the dense connection module.

Dense Connection: A dense connection mechanism was proposed in DenseNet [72], which is widely used in 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. 6). 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.* [162] proposed an SRDenseNet. SRDenseNet uses not only the layer-level dense connections but also the block-level ones, 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 [228], 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 the aforementioned models, the dense connection is also applied in MemNet [159], RPMNet [131], MFNet [150], etc. With the help of a dense connection mechanism, the information flow among different depths of the network can be fully used, thus yielding better reconstruction results.

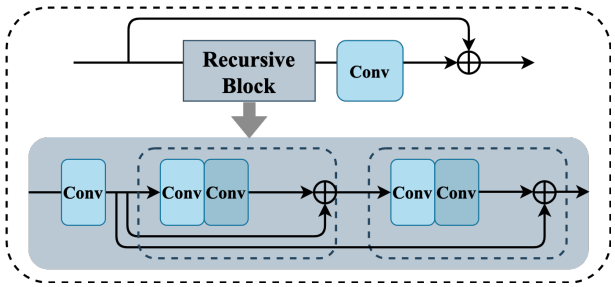


Fig. 7. The structure of DRRN, where the shaded part denotes the recursive block and the parameters in the dashed box are sharing.

Recursive Learning: To obtain a large receptive field without increasing model parameters, recursive learning is proposed for SISR, where the same sub-modules are repetitively applied in the network, and 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 [84], resulting in a 41×41 size receptive field. However, too many stacked layers in the recursive learning-based model will still cause the problem of vanishing/exploding gradient. Therefore, in DRRN [158], the recursive block is conducted based on residual learning (Fig. 7). Recently, more and more models have introduced the residual learning strategy in their recursive units, such as MemNet [159], CARN [3], and SRRFN [99].

Progressive Learning: Progressive learning refers to gradually increasing the difficulty of the learning task. For some sequence prediction tasks or sequential decision-making problems, progressive learning is used to reduce the training time and improve the generalization performance. Since SISR is an ill-posed problem that 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 progressive learning to simplify the learning process and improve the reconstruction efficiency.

In LapSRN [87], the method is applied to progressively reconstruct the sub-band residuals of high-resolution images. In ProSR [180], 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 [106], the strategy is applied to solve the complex degradation tasks, where targets of different difficulties are ordered for progressive learning. With the help of progressive learning, complex problems can be decomposed into multiple simple tasks, hence accelerating model convergence and obtaining better reconstruction results.

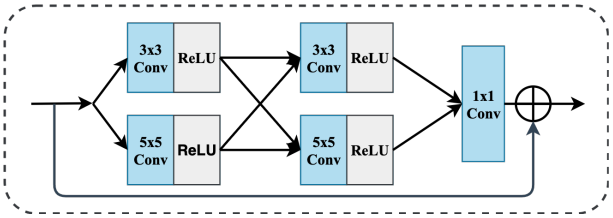


Fig. 8. The structure of multi-scale residual block (MSRB [98]).

Multi-scale Learning: Rich and accurate image features are essential for SR image reconstruction. Meanwhile, plenty of research works [29, 87, 157] 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 [29], Li *et al.* [98] proposed a multi-scale residual block (MSRB, Fig. 8) for feature extraction. MSRB integrates different convolution kernels in a block to adaptively extract image features at different scales. After that, Li *et al.* [97] further optimized the structure and proposed a more accurate 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.

Recently, more and more multi-scale SISR models have been proposed. For instance, Qin *et al.* [141] proposed a multi-scale feature fusion residual network (MSFFRN) to fully exploit image features for SISR. Chang *et al.* [17] proposed a multi-scale dense network (MSDN) by combining multi-scale learning with the dense connection. Cao *et al.* [15] 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.

Attention Mechanism: Attention mechanism can be considered as a tool that can allocate available resources to the most informative part of the input. 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.* [69] proposed a squeeze-and-excitation (SE) block to model channel-wise relationships in the image classification task. Wang *et al.* [174] proposed a non-local attention neural network for video classification by incorporating non-local operations. Motivated by these methods, an attention mechanism has also been introduced into SISR.

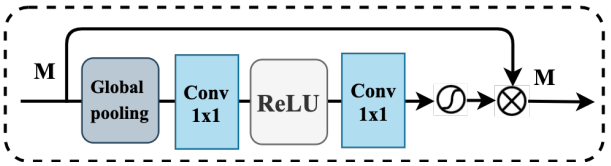


Fig. 9. The principle of channel attention mechanism (CAM).

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 [132, 225] introduce the SE mechanism in the SISR model. For example, Zhang *et al.* [225] proposed a new module based on the SE mechanism, named residual channel attention block (RCAB). As shown in Fig. 9, a global average pooling layer followed by a Sigmoid function is used to rescale each feature channel, allowing the network to concentrate on the more useful channels and enhancing discriminative learning ability. In SAN [33], second-order statistics of features are explored to conduct the attention mechanism based on covariance normalization. A great number of experiments have shown that 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 [113]. A non-local attention block is proposed in RNAN [227], 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 [138], 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 [134], 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. To mitigate the noise pollution caused by non-local attention, ENLCA [197] utilizes efficient non-local attenuation and sparse aggregation to focus on useful information with contrast learning to separate irrelevant features. All these methods show the effectiveness of non-local attention, which can further improve the model performance.

Feedback Mechanism: The 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 keep self-correcting and do not share. Recently, the feedback mechanism has been widely used in many computer vision tasks [14, 16], which is also beneficial for SR image reconstruction. Specifically, the feedback mechanism allows the network to carry high-level information back to previous layers and refine low-level information, thus fully guiding the LR image to recover high-quality SR images.

In DBPN [62], iterative up- and down-sampling layers are provided to achieve an error feedback mechanism for projection errors at each stage. In DSRN [61], a dual-state recurrent network is proposed, where recurrent signals are exchanged between these states in both directions via delayed feedback. In SFRBN [106], 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.

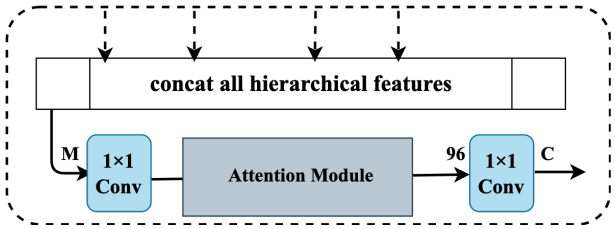


Fig. 10. The structure of the hierarchical feature distillation block (HFDB).

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 [162] and MSRN [98], such 1×1 convolutional layer acts as a bottleneck layer before the reconstruction module. In MemNet [159], 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 continuously conducted in the network. For example, in MemNet [159] and CARN [4], the gating mechanism is used in both global and local regions. Sometimes, it can be combined with other operations, such as the attention mechanism, to construct a more effective gate module to achieve feature distillation. For instance, Li *et al.* [97] proposed a hierarchical feature distillation block (Fig. 10) by combining 1×1 convolutional layer and attention mechanism.

Efficient Structure: There is no doubt that increasing the depth of the model is the easiest way to improve the model performance. However, due to the huge computational overhead of deep and large models, it is difficult to be applied to mobile devices with limited computing capabilities. To address this issue, more and more lightweight and efficient SISR methods have been proposed in recent years. For instance, Ahn *et al.* [3] designed an architecture (CARN) that implements a cascading mechanism upon the residual network, which achieved fast, accurate, and lightweight SR. Hui *et al.* [75] proposed a novel Information Distillation Network (IDN) with lightweight parameters and computational complexity by using the information distillation strategy. After that, the author further proposed a Lightweight Information Multi-Distillation Network (IMDN) by constructing the cascaded information multi-distillation blocks. Liu *et al.* [114] proposed a RFDN, enhances the efficiency of single image super-resolution (SISR) by incorporating a lighter feature distillation connection operation. Zhou *et al.* [234] have developed VapSR, which refines attention mechanisms to create a more efficient super-resolution network. Li *et al.* [155] introduced ShuffleMixer, a technique that investigates the use of large convolutions and channel splitting shuffle operations to make the network more mobile-compatible. Li *et al.* [105] proposed a Blueprint Separable Residual Network (BSRN) containing two efficient designs, blueprint separable convolution and more effective attention modules. Li *et al.* [101] proposed a novel Cross-receptive Field Guided

Transformer (CFGT) to enable the selection of contextual information required for reconstruction by using a modulated convolutional kernel. In addition, some hardware-friendly SISR methods have emerged. For example, Luo *et al.* [121] proposed an Individual Kernel Sparsity (IKS) method for memory-efficient and sparsity-adjustable image SR, which enables deep networks can be deploymented in memory-limited devices. Ye *et al.* [204] proposed a Hardware-friendly Scalable SR (HSSR) with progressively structured sparsity. This model can cover multiple SR models with different sizes by a single scalable model, without extra retraining or post-processing. Lin *et al.* [109] proposed a Memory-friendly Scalable dynamic SR (MSSR) lightweight model via rewinding, which can be easily generalized to different SR models. Choi *et al.* [28] introduced a NGswin, which boosts performance in SISR by broadening the receptive field of window-based self-attentive methods. Wang *et al.* [166] proposed a Omni-SR, enhancing the capabilities of lightweight models by replicating pixel interactions across both spatial and channel dimensions. Li *et al.* [102] introduced a DLGSANet, which streamlines SISR efficiency by employing sparse global self-attention modules to pinpoint the most pertinent similarity values.

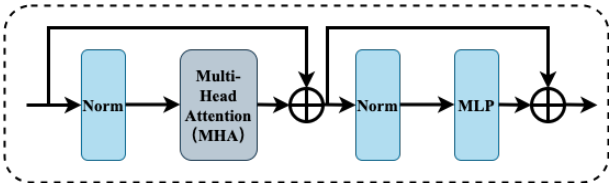


Fig. 11. The structure of classic Transformer. The key component is the multi-head attention (MHA) module.

Transformer-based Method: The key idea of the Transformer is the "self-attention" mechanism, which can capture long-term information between sequence elements. Recently, Transformer [164] (Fig. 11) has achieved brilliant results in NLP tasks. For example, the pre-trained deep learning models (e.g., BERT [35], GPT [144]) have shown effectiveness over conventional methods. Inspired by this, more and more researchers have begun to explore the application of Transformers in computer vision tasks and have achieved breakthrough results in many tasks. In image restoration, Transformer is often used to capture the global information of the image to further improve the quality of the reconstructed image.

In recent years, more and more Transformer-based models have been proposed. For example, Chen *et al.* proposed the Image Processing Transformer (IPT [22]) 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 [22] 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 [107] for image restoration based on the Swin Transformer [117]. Specifically, the Swin Transformer blocks (RSTB) are proposed for feature extraction and DIV2K+Flickr2K is used for training. To improve the lack of direct interaction between different windows in SwinIR. Zamir [212] *et al.* proposed Restormer to reconstruct high-quality images by embedding CNNs within Transformer and performing local-global learning at multiple scales. Chen *et al.* proposed CAT [27] to extend the attention region and aggregate features across different windows. Then, to activate more of the pixels that Transformer focuses on, Chen *et al.* proposed HAT [24], which uses overlapping cross-attention modules in conjunction with a pre-training strategy to enhance Transformer model potential. Li [103] *et al.* proposed GRL to explicitly model the image hierarchy at global, regional, and local scales by integrating various attentions within the Transformer. As

for the application on the lightweight SISR model, Lu *et al.* [119] proposed an Efficient Super-Resolution Transformer (ESRT) for fast and accurate SISR which achieves competitive results with fewer parameters and low computing costs. Zhang *et al.* [222] proposed ELAN with a shared self-attention mechanism to reduce model complexity and accelerate the Transformer-based model. Wang *et al.* [191] proposed the Uformer, a general and superior U-shaped Transformer, which can reduce the computational complexity on high-resolution feature map while capturing local context and multi-scale features. Zamir *et al.* [212] proposed an efficient Restormer that can capture long-range pixel interactions while remaining applicable to large images. Li *et al.* [101] proposed a Cross-receptive Focused Inference Network (CFIN) that can incorporate contextual modeling to achieve good performance with limited computational resources. Zhu *et al.* [239] designed an Attention Retractable Frequency Fusion Transformer (ARFFT) to strengthen the representation ability and extend the receptive field to the whole image. Li *et al.* [100] proposed a concise and powerful Pyramid Clustering Transformer Network (PCTN) for lightweight SISR. Chen *et al.* [26] proposed a novel Dual Aggregation Transformer (DAT) for SISR, which aggregates features across spatial and channel dimensions, in the interblock and intra-block dual manner. Zhou *et al.* [236] proposed a SRFormer, elevates the performance of window-based Transformer approaches by effectively integrating self-attentive channel and spatial information. Li *et al.* [103] achieves optimal performance across multiple scenarios by developing GRL, a hierarchical Transformer-based model for image upscaling that operates on global, regional, and local scales. ATDSR, brought forth by Zhang *et al.* [218], enriches the SR Transformer with an auxiliary set of adaptive token dictionaries, thereby enhancing the precision of SISR. Adaptive token sparsifcation transformer (AdaFormer) proposed by Luo *et al.* [120] speeds up model inference for images by incorporating sparsity strategies. Although the performance of the Transformer-based method has greatly improved, the attention mechanism used in Transform will occupy a large amount of GPU memory. Therefore, how to further reduce the GPU memory of Transformer-based methods is worth further exploration.

3.2 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.* [10] 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 the good perceptual quality of the image.

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, the content loss is widely used to keep the image consistent with the target [91, 176], 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 losses need to be considered.

Texture Loss: Texture loss, also called style reconstruction loss, is proposed by Gatys *et al.* [53, 54], 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:

$$\mathcal{L}_{texture}^{\phi,j}(I_{SR}, I_y) = ||G_j^{\phi}(I_{SR}) - G_j^{\phi}(I_y)||_F^2. \tag{25}$$

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 [80].

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.* [142] proposed a targeted perceptual loss that penalized images at different semantic levels based on the labels of object, background, and boundary. Therefore, more realistic textures and sharper edges can be obtained to reconstruct realistic SR images.

Adversarial Training: In 2014, the Generative Adversarial Networks (GANs) were proposed by Goodfellow *et al.* [56], which has been widely used in compute vision tasks, such as style transfer and image inpainting. The GANs consist 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.* [91] proposed the Super-Resolution Generative Adversarial Network (SRGAN). In SRGAN, the generator G is essentially an SR model that is 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 real and natural SR images. The generative loss $\mathcal{L}_{Gen}(I_x)$ can be defined as:

$$\mathcal{L}_{Gen} = -\log D_{\theta_D}(G_{\theta_G}(I_x)), \tag{26}$$

and the loss in terms of the discriminator is:

$$\mathcal{L}_{Dis} = -\log(D_{\theta_D}(I_y)) - \log(1 - D_{\theta_D}(G_{\theta_G}(I_x))). \tag{27}$$

Therefore, we need to solve the following problem:

$$\begin{aligned} \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I_{y\sim p_{data}(I_y)}}(\log D_{\theta_D}(I_y)) + \\ \mathbb{E}_{I_{x\sim p_G(I_x)}}(\log(1 - D_{\theta_D}(G_{\theta_G}(I_x)))). \end{aligned} \tag{28}$$

In SRGAN [91], the generator is the SRResNet and the discriminator uses the architecture proposed by Radford *et al.* [143]. In ESRGAN [177], 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 [139], 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 [177], Wang *et al.* adopted the Relativistic GAN [81], 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. Wang *et al.* [178] proposed a novel GAN inversion framework that utilizes the powerful generative ability of StyleGAN-XL, which shows preferable quantitative and qualitative results in SISR.

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.* [238] 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 is 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 perform image degradation first without paired data can help generate more realistic images [11]. In CinCGAN [210], a cycle-in-cycle network is proposed, where the noisy and blurry input is mapped to a noise-free LR domain first and then upsampled with a pre-trained model. In DRN [60], 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 model is trained with both paired and unpaired data.

Diffusion-based Method: Derived from the recent inspiration in the denoising diffusion probability model (DDPM) [68], a new conditional image generation method is incorporated into the SISR task. Compared with the GAN-based SISR method, the diffusion model-based SISR methods [52, 96, 147, 149] have better fidelity and reduce the generation of artifacts.

SRDiff [96] is the first diffusion-based SISR model, which provides diverse and realistic SISR predictions by gradually converting Gaussian noise into SISR images with LR as the input condition through Markov chains. SR3 [147] iteratively refines the pure Gaussian noise input using a model trained for denoising at various noise levels. Compared to GAN-based methods, it can output more realistic photos. IDM [52] integrates implicit neural representation and denoising diffusion model end-to-end and employs implicit neural representation to learn continuous image resolution representation during decoding. DR2 [193] utilizes DDPM to coarsely reduce more complex low-quality face images and then uses the enhancement module to fully restore them to high-resolution (HR) face images. There is also a class of methods that aim to utilize the prior diffusion-based models to aid SISR. For example, StableSR [167] and DiffBIR [110] achieve real-world SISR by fine-tuning with prior knowledge from a pre-trained text-to-image diffusion model, such as Stable diffusion [146]. DiffIR [198] utilizes a pre-trained model trained on ground-truth images to incorporate the prior into the SISR model, which can result in accurate estimates using fewer iterations than traditional DDPM. However, diffusion-based SISR models still need a large number of new samples and the slow convergence rate of the model limits their use scenarios. Therefore, how to overcome these drawbacks is still worthy of study.

3.3 Information Utilization 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. Although the current SISR model has made a significant breakthrough, how to use the information inside and outside of the image to further improve the performance of the model is still worth exploring.

Internal Statistics: In [241], 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 patch 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 [152], 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 the 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 [148], 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.

Multi-factor 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. Surprisingly, researchers found that this method can fully exploit the inter-scale correlation between different upsampling factors, which can further improve model performance.

In LapSRN [88], 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 [108], 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 [97]. 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.

Prior Guidance: 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, a 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.* [203] integrated the edge prior with recursive networks and proposed a Deep Edge Guided Recurrent Residual Network (DEGREE) for SISR. After that, Fang *et al.* [43] proposed an efficient and accurate Soft-edge Assisted Network (SeaNet). 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 EdgeNet. Meanwhile, they find that more accurate priors can lead to more significant performance. Additionally, image priors are also beneficial for GAN-based models. 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[176]. With this information from high-level tasks, similar LR patches can be easily distinguished and more natural textual details can be generated. In SPSR [125], 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 image generation and reduce geometric distortions. In FeMaSR [19], the authors use discrete features obtained by VQ-GAN [208] pre-training in HR images as prior information to performing image recovery by matching distorted LR image features with distortion-free HR features from the pre-trained HR prior.

Reference-based Method: 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 [211, 232] chose 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.* [211] conducted global registration and local matching between the reference and LR images to solve an energy minimization problem. In CrossNet [232], 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.* [230] 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 [201], Yang *et al.* proposed a texture transformer network to search and transfer relevant textures from the reference images to the LR image.

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 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 [67]. In feature distillation, the intermediate features maps are transferred to the student model [1, 5].

Inspired by this, some works introduce the knowledge distillation technique to SISR to further improve the performance of lightweight models. For instance, in SRKD [51], 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 [93], the teacher network leverages 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-frequency details from the Teacher which is trained with the HR images. Subsequently, JDSR [122] explored a joint distillation learning that effectively improves the distillation performance of lightweight models by using distillation of HR's privileged information in conjunction with internal self-distillation. CSD [175] combines the contrast learning and distillation tasks to further reduce the solution space of SISR. In addition, to solve the model compression problem for unsupervised issues, [224] used a generator to synthesize training samples close to the original data after using a progressive distillation scheme to improve student model performance.

3.4 Real-World Image Super-Resolution

The degradation modes are complex and unknown in real-world scenarios [21], 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, simulation SISR models exhibit poor performance when handling real-world images. Meanwhile, most of the aforementioned models can only be applied to some specific integer upsampling factors. This greatly limits the practical application and promotion of these models. To solve these problems, some interesting methods have been proposed. Based on the problems they intend to solve, we divide them into two major categories: Blind Image Super-Resolution and Scale Arbitrary Super-Resolution.

3.4.1 ***Blind Image Super-Resolution.*** Blind SISR has attracted increasing attention due to its significance in real-world applications, which aim to super-resolve LR images with unknown degradation. It is worth noting that blind SISR has no clear definitions. More details about blind SISR can be found in [111]. In this work, we simply divided them into two categories: explicit

degradation modeling methods and implicit degradation modeling methods, according to the ways of degradation modeling.

Explicit Degradation Modeling: Blind SISR methods with explicit modeling of the degradation process are mainly based on the classical degradation model, where the blur kernel and additive noise are two main degradation factors. According to whether the degradation process is estimated, this type of method can be further divided into two categories: image-specific adaptation without degradation estimation and image-specific adaptation with degradation estimation. Among them, the first type of method often uses an external method to perform degradation estimation before the SR process, thus adapting the framework to the blind setting. The second type of method often uses an internal module for degradation estimation and outputs the degradation representation. For example, Zhang *et al.* [217] proposed a simple and scalable deep CNN framework (SRMD) for multiple degradations learning. In SRMD, the concatenated LR image and degradation maps are taken as input of the network to achieve image super-resolution under different degradations. Based on SRMD, Xu *et al.* [200] proposed the UDVD, which uses dynamic convolution to process different degradations in different areas in the image. This type of method often relies on reliable degradation estimation methods to quickly obtain satisfactory SR output. Hence, a method that incorporates degradation estimation into the SR framework will obtain more stable and reliable results. Towards filling this gap, growing attention has been paid to image-specific adaptation method with degradation estimation. This type of method combines degradation estimation and SR processes into a unified model, in which kernel estimation is the main research work. For example, in IKC [57], the iterative kernel correction procedure is proposed to help the blind SISR task find more accurate blur kernels. Inspired by it, Luo *et al.* [124] adopted an alternating optimization algorithm and proposed a Deep Alternating Network (DAN) to estimate blur kernel and restore SR image in a single network, which makes the restorer and estimator well compatible with each other, and thus achieves good results in kernel estimation. Although such methods are more robust than using off-the-shelf estimation algorithms, such iterative schemes often consume more inference time and may lead to SR failure due to large estimation errors. To address this issue, some works introduced more accurate degradation estimation methods. In [171], the author suggested learning abstract representations to distinguish various degradations in the representation space and introduced a Degradation-Aware SR (DASR) network with flexible adaption to various degradations based on the learned representations. There are also some methods proposed to estimate more realistic kernels from real images. For instance, in [12], 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.

Implicit Degradation Modeling: Blind SISR methods with implicit modeling of degradation process aim to model the degradation through learning with external dataset. This type of method usually learns data distribution by a GAN framework, and one or more discriminators are used to distinguish generated images from real ones. For example, Yuan *et al.* [210] proposed an unsupervised image SR using Cycle-in-Cycle Generative Adversarial Networks (CinCGAN). CinCGAN first mapped the noisy and blurry input to a noise-free low-resolution space, and then the intermediate image was up-sampled with a pre-trained model. Finally, these two modules are fine-tuned in an end-to-end manner to get SR output. Bulat *et al.* [11] believed that low-resolution images in the real world constitute a specific distribution in high-dimensional space, and use a generative adversarial network to generate low-resolution images consistent with this distribution from high-resolution images. After that, Yuan *et al.* [210] and Maeda *et al.* [129] further proposed a unified framework, which can simultaneously learn the generation of pseudo-low-resolution images and the reconstruction of high-resolution images, achieving better results in actual scenes. Wei *et al.* [195] further considered the domain difference between pseudo-low-resolution images and real low-resolution images, and proposed a domain adaptation mechanism to improve model performance. Wolf *et*

al. [196] proposed the DeFlow framework, which uses the stochastic modeling ability of the flow model to enhance the diversity of pseudo-low-resolution images, and further improves the image super-resolution performance in real scenes.

3.4.2 ***Scale Arbitrary.*** 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 [70] and Meta-USR [71]. 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.* [173] 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.

3.5 Domain-Specific Applications

The technology of image super-resolution has been widely used in many application scenarios. As shown in Fig. 12, we introduce various applications of SR in this section.

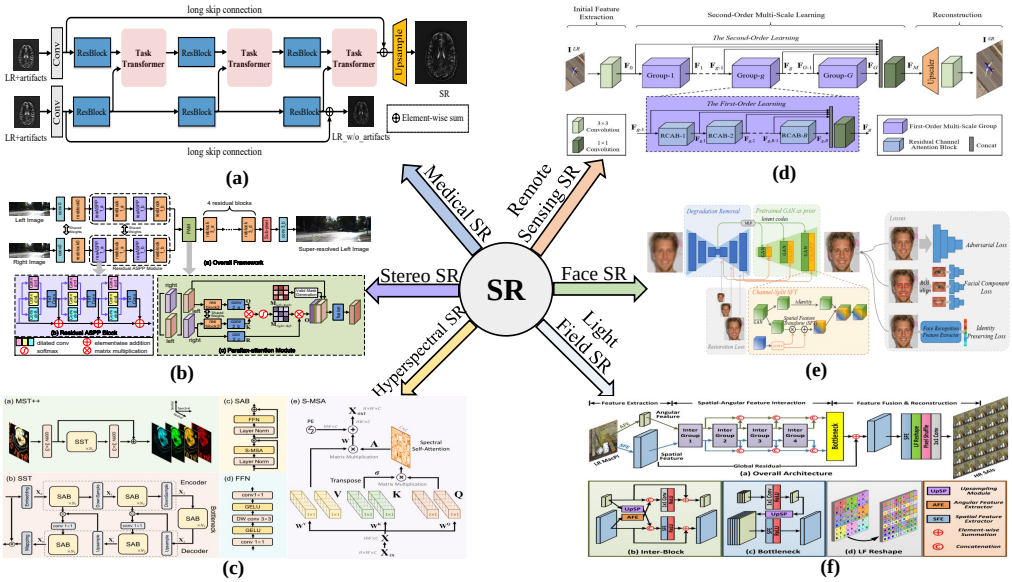


Fig. 12. Examples of various popular SR tasks. (a) T2Net [44] for Medical SR, (b) PASSRNet [169] for Stereo SR, (c) MST++ [13] for Hyperspectral SR, (d) SMSR [41] for Remote Sensing SR, (e) GFPGAN [175] for Face SR, (f) LF-InterNet [185] for Light Field SR.

3.5.1 ***Stereo Image Super-Resolution.*** The dual camera has been widely used to estimate depth information. Meanwhile, stereo imaging can also be applied in image restoration. In this task, we

have two images with a disparity much larger than one pixel. Therefore, full use of these two images can enhance spatial resolution.

In StereoSR [76], 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, 172] 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 [184], a dataset named Flickr1024 is proposed for stereo image super-resolution, which consists of 1024 high-quality stereo image pairs. In [205], 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 [187], a symmetric bi-directional parallax attention module and an inline occlusion handling scheme are proposed to effectively interact with cross-view information. In [32], 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. In [30], in addition to extracting single image features from the left and right views separately using NAFNet [23], a stereo cross-attention module is introduced to fuse the image features from the left and right views.

3.5.2 ***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 [63] 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 [58], a Deep Residual Squeeze and Excitation Network (DRSEN) are proposed to overcome the problem of the high complexity of remote sensing image distribution. In [215], 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 [41], the authors developed a Dense-Sampling Super-Resolution Network (DSSR) to explore the large-scale SR reconstruction of the remote sensing imageries. In [94], the authors proposed a new Hybrid-scale Self-similarity Exploitation Network (HSENet), which can simultaneously exploit single and cross-scale similarities for high-quality image reconstruction; In [181], Wang *et al.* proposed a Multi-scale Enhancement Network (MEN), which uses multi-scale features of remote sensing images to enhance the network's reconstruction capability; In [116], Liu *et al.* proposed a Dual Learning-based Graph Neural Network (DLGNN), in which the graph neural network (GNN) is utilized to consider the self-similarity patches in remote sensing imagery by aggregating cross-scale neighboring feature patches. All these methods achieve excellent results in remote sensing image super-resolution.

3.5.3 ***Light Field Image Super-Resolution.*** A 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 has become 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 [185]. To solve this issue, SR technology is introduced to achieve a good balance between spatial and angular resolution.

In [206], 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. To reduce the dependence of accurate depth or disparity information as priors for the light-field image super-resolution, Sun *et al.* [179] proposed a bidirectional recurrent convolutional neural network and an implicitly multi-scale fusion scheme for SR images reconstruction. In [185], 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 [186]) to handle the disparity problem of LF image SR. In [183], Wang *et al.* further proposed a generic light field disentangling mechanism to achieve state-of-the-art performance in spatial SR, angular SR and disparity estimation, respectively. In [163], Duong *et al.* proposed a light field SR model via joint spatial-angular and epipolar information, which can simultaneously exploit information from three different types of 4D LF representation.

3.5.4 ***Face Image Super-Resolution***. Face image super-resolution is the most famous field in which SR technology to domain-specific images. Due to the potential applications in facial recognition systems such as security and surveillance, face image SR has become an active area of research.

Recently, DL-based methods have achieved remarkable progress in face image SR. In [233], 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 [240], Zhu *et al.* proposed to jointly learn face hallucination and facial spatial correspondence field estimation. In [209], spatial transformer networks are used in the generator architecture to overcome problems related to the misalignment of input images. In [37, 216], 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 [49], 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. These methods [59, 235, 237] for reconstructing high-quality face images with photo-realistic textures from very low-resolution inputs are mainly based on the generative prior of GAN.

3.5.5 ***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[145]. The hyperspectral system is often compromised due to the limitations of the amount of 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 [133], a 3D fully convolutional neural network is proposed to extract the feature of hyperspectral images. In [104], 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 [45], an unsupervised CNN-based method is proposed to effectively exploit the underlying characteristics of the hyperspectral images. In [78], Jiang *et al.* proposed a group convolution and progressive upsampling framework to reduce the size of the model and make it feasible to obtain stable training results under small data conditions. In [112], 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. In [13], the quality of reconstructed images is improved from coarse to fine by using the spectral-wise multi-headed self-attention, which is based on the HSI spatially sparse while spectrally selfsimilar nature to compose the basic unit. In [219], Zhang *et al.* proposed an efficient Transformer for hyperspectral image super-resolution via a novel and efficient SCC-kernel-based self-attention method.

3.5.6 ***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 about the human body. In recent years, many DL-based methods have also been proposed for medical image SR.

For instance, Chen *et al.* proposed a Multi-level Densely Connected Super-Resolution Network (mDCSRN [25]) with GAN-guided training to generate high-resolution MR images, which can train and infer quickly. In [182], a 3D Super-Resolution Convolutional Neural Network (3DSRCNN) is proposed to improve the resolution of 3D-CT volumetric images. In [231], 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 [140], Peng *et al.* introduced a Spatially-Aware Interpolation Network (SAINT) for medical slice synthesis to alleviate the memory constraint that volumetric data posed. In [44], Feng *et al.* proposed a Task Transformer Network (T2Net) to allow the network to share representation and feature transfer between the two tasks of reconstruction and super-resolution. In [55], Georgescu *et al.* performed medical image super-resolution using a multimodal low-resolution input and propose a novel multimodal multi-head convolutional attention mechanism for multi-contrast medical image SR.

All of these methods are the cornerstone of building the smart medical system and have great research significance and value.

4 RECONSTRUCTION RESULTS

To help readers intuitively know the performance of the aforementioned SISR models, we provide a detailed comparison of the reconstruction results of these models. Specifically, we collect 53 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. 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 or large models; (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.

5 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 point out some challenges and summarize some promising future directions.

5.1 Lightweight SISR for Edge Devices

With the huge development of the 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, and smart cameras). This is because the model size and computational costs of these models still exceed the limits of edge devices. Therefore, exploring lightweight SISR models that can be practical in use for edge devices has great research significance and commercial value. To achieve this, more efficient network structures and mechanisms are worthy of further exploration. Moreover, it is also necessary to use technologies like network binarization [128] and network

Table 2. PSNR/SSIM comparison on Set5 (×4), Set14 (×4), and Urban100 (×4). Meanwhile, the training datasets and the number of model parameters are provided. It is worth noting that the upper part of the table is lightweight models with parameters less than 1M (M=million) and they are sorted in ascending order by PSNR results on Set5. Meanwhile, the best results are **highlighted**.

Models	Set5 PSNR/SSIM	Set14 PSNR/SSIM	Urban100 PSNR/SSIM	Training Datasets	Parameters
SRCNN [207]	30.48/0.8628	27.50/0.7513	24.52/0.7221	T91+ImageNet	57K
ESPCN [151]	30.66/0.8646	27.71/0.7562	24.60/0.7360	T91+ImageNet	20K
FSRCNN [39]	30.71/0.8660	27.59/0.7550	24.62/0.7280	T91+General-100	13K
VDSR [83]	31.35/0.8838	28.02/0.7680	25.18/0.7540	BSD+T91	665K
LapSRN [87]	31.54/0.8855	28.19/0.7720	25.21/0.7560	BSD+T91	812K
DRRN [158]	31.68/0.8888	28.21/0.7721	25.44/0.7638	BSD+T91	297K
MemNet [159]	31.74/0.8893	28.26/0.7723	25.50/0.7630	BSD+T91	677K
AWSRN-S [165]	31.77/0.8893	28.35/0.7761	25.56/0.7678	DIV2K	588K
IDN [75]	31.82/0.8903	28.25/0.7730	25.41/0.7632	BSD+T91	678K
NLRN [113]	31.92/0.8916	28.36/0.7745	25.79/0.7729	BSD+T91	330K
ECBSR [223]	31.92/0.8946	28.34/0.7817	25.81/0.7773	DIV2K	682K
CARN-M [3]	31.92/0.8903	28.42/0.7762	25.62/0.7694	DIV2K	412K
SMSR [168]	32.12/0.8932	28.55/0.7808	26.11/0.7868	DIV2K	1006K
RFDN [114]	32.18/0.8948	28.58/0.7812	26.04/0.7848	DIV2K	441K
ESRT [119]	32.19/0.8947	28.69/0.7833	26.39/0.7962	DIV2K	751K
IMDN [74]	32.21/0.8949	28.58/0.7811	26.04/0.7838	DIV2K	715K
FDIWN [48]	32.23/0.8955	28.66/0.7829	26.28/0.7919	DIV2K	664K
MAFFSRN [137]	32.24/0.8952	28.61/0.7819	26.11/0.7858	DIV2K	550K
MSFIN [192]	32.28/0.8957	28.57/0.7813	26.13/0.7865	DIV2K	682K
LBNet [50]	32.29/0.8960	28.68/0.7832	26.27/0.7906	DIV2K	742K
LatticeNet-CL [123]	32.30/0.8958	28.65/0.7822	26.19/0.78555	DIV2K	777K
HPUN-L [154]	32.31/0.8962	28.73/0.7842	26.27/0.7918	DIV2K	734K
ELAN [222]	32.43/0.8975	**28.78**/**0.7858**	**26.47**/**0.7980**	DIV2K	601K
SwinIR-light [107]	32.44/0.8976	28.77/**0.7858**	**26.47/0.7980**	DIV2K	886K
CFIN [101]	**32.49**/**0.8985**	28.74/0.7849	26.39/0.7946	DIV2K	699K
DSRN [61]	31.40/0.8830	28.07/0.7700	25.08/0.7470	T91	1.2M
DRCN [84]	31.53/0.8838	28.02/0.7670	25.14/0.7510	T91	1.8M
MADNet [89]	31.95/0.8917	28.44/0.7780	25.76/0.7746	DIV2K	1M
SRMD [217]	31.96/0.8925	28.35/0.7787	25.68/0.7731	BSD+DIV2K+WED	1.6M
SRDenseNet [162]	32.02/0.8934	28.50/0.7782	26.05/0.7819	ImageNet	2.0M
SRResNet [91]	32.05/0.8910	28.49/0.7800	—-/—-	ImageNet	1.5M
MSRN [98]	32.07/0.8903	28.60/0.7751	26.04/0.7896	DIV2K	6.3M
CARN [3]	32.13/0.8937	28.60/0.7806	26.07/0.7837	BSD+T91+DIV2K	1.6M
SeaNet [43]	32.33/0.8970	28.81/0.7855	26.32/0.7942	DIV2K	7.4M
CRN [3]	32.34/0.8971	28.74/0.7855	26.44/0.7967	DIV2K	9.5M
EDSR [108]	32.46/0.8968	28.80/0.7876	26.64/0.8033	DIV2K	43M
RDN [228]	32.47/0.8990	28.81/0.7871	26.61/0.8028	DIV2K	22.6M
DBPN [62]	32.47/0.8980	28.82/0.7860	26.38/0.7946	DIV2K+Flickr2K	10M
SRFBN [106]	32.47/0.8983	28.81/0.7868	26.60/0.8015	DIV2K+Flickr2K	3.63M
MDCN [97]	32.48/0.8985	28.83/0.7879	26.69/0.8049	DIV2K	4.5M
RNAN [227]	32.49/0.8982	28.83/0.7878	26.61/0.8023	DIV2K	7.5M
SRRFN [99]	32.56/0.8993	28.86/0.7882	26.78/0.8071	DIV2K	4.2M
RCAN [226]	32.63/0.9002	28.87/0.7889	26.82/0.8087	DIV2K	16M
SAN [33]	32.64/0.9003	28.92/0.7888	26.79/0.8068	DIV2K	15.7M
HAN [138]	32.64/0.9002	28.90/0.7890	26.85/0.8094	DIV2K	16.1M
RFANet [115]	32.66/0.9004	28.88/0.7894	26.92/0.8112	DIV2K	11M
ENLCN [197]	32.67/0.9004	28.94/0.7892	27.12/0.8141	DIV2K	43.6M
DRN-S [60]	32.68/0.9010	28.93/0.7900	26.84/0.8070	DIV2K+Flickr2K	4.8M
CRAN [229]	32.72/0.9012	29.01/0.7918	27.13/0.8167	DIV2K	14.9M
SwinIR [107]	32.92/0.9044	29.09/0.7950	27.45/0.8254	DIV2K+Flickr2K	11.8M
CAT-A [27]	33.08/0.9052	29.18/0.7963	27.89/0.8339	DIV2K+Flickr2K	16.6M
GRL-B [103]	33.10/**0.9094**	29.37/**0.8058**	**28.53**/**0.8504**	DIV2K+Flickr2K	20.2M
HAT-L [24]	**33.30**/0.9083	**29.38**/0.8001	28.37/0.8447	DIV2K+Flickr2K	40.2M

quantization [95] to further reduce the model size. Therefore, combining lightweight SISR models with model compression schemes has great application value.

5.2 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 results.

5.3 New Loss Functions and Assessment Methods

In the past, most 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 an 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.

5.4 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, integrate image super-resolution with high-level tasks has become a hot topic in recent years. To achieve this, Zangeneh *et al.* [213] proposed a novel nonlinear coupled mapping architecture using two deep convolutional neural networks to project the low and high resolution face images into a common space to achieve low-resolution face recognition; Wang *et al.* [170] proposed a dual super-resolution learning method for semantic segmentation, which integrate image super-resolution with semantic segmentation into a end-to-end model; Xiang *et al.* [199] boosted high-level vision with joint compression artifacts reduction and super-resolution. Although these methods combine SR with high-level tasks and achieve good results, they focus more on the results of high-level vision tasks and ignore the use of feedback from other tasks to further improve the quality of SR images. 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 SISR and other tasks can promote and learn from each other.

5.5 Efficient and Accurate Real SISR

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

5.6 Efficient and Accurate Scale Arbitrary SISR

SISR has seen its applications in diverse real-life scenarios and users. Currently, most DL-based SISR models can only be applied to one or a limited number of multiple upsampling factors. 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. 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, it is of great significance to explore a simple and flexible CNN-based accurate scale-arbitrary SISR model like Bicubic.

5.7 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 most 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 model construction. In the future, it will be a trend to further optimize the existing SISR models based on prior knowledge and the characteristics of the domain-specific images.

6 CONCLUSION

In this survey, we provided a comprehensive overview of DL-based SISR 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.

7 ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China under Grant 2021YFA1003004, in part by the National Natural Science Foundation of China under Grants 62301306, 62301601, in part by the Science and Technology Commission of Shanghai Municipality under Grant 23ZR1422200, 23YF1412800.

REFERENCES

[1] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. 2015. Fitnets: Hints for thin deep nets. (2015).

[2] Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*.

[3] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. 2018. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*.

[4] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. 2018. Image super-resolution via progressive cascading residual network. In *CVPRW*.
[5] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *CVPR*.
[6] Saeed Anwar, Salman Khan, and Nick Barnes. 2020. A deep journey into super-resolution: A survey. *Comput. Surveys* 53, 3 (2020), 1–34.
[7] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2010), 898–916.
[8] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*.
[9] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*.
[10] Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *CVPR*.
[11] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. 2018. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *ECCV*.
[12] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*.
[13] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. 2022. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPR*.
[14] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. 2015. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*.
[15] Feilong Cao and Huan Liu. 2019. Single image super-resolution via multi-scale residual channel attention network. *Neurocomputing* 358 (2019), 424–436.
[16] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2016. Human pose estimation with iterative error feedback. In *CVPR*.
[17] Chia-Yang Chang and Shao-Yi Chien. 2019. Multi-scale dense network for single-image super-resolution. In *ICASSP*.
[18] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. 2004. Super-resolution through neighbor embedding. In *CVPR*.
[19] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. 2022. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACMMM*.
[20] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. 2019. Camera lens super-resolution. In *CVPR*.
[21] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E Sheriff, and Ce Zhu. 2022. Real-world single image super-resolution: A brief review. *Information Fusion* 79 (2022), 124–145.
[22] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *CVPR*.
[23] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple baselines for image restoration. In *CECCV*.
[24] X Chen, X Wang, J Zhou, and C Dong. 2023. Activating More Pixels in Image Super-Resolution Transformer. (2023).
[25] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. 2018. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network. In *MICCAI*.
[26] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. 2023. Dual aggregation transformer for image super-resolution. In *ICCV*.
[27] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. 2022. Cross Aggregation Transformer for Image Restoration. (2022).
[28] Haram Choi, Jeongmin Lee, and Jihoon Yang. 2023. N-gram in swin transformers for efficient lightweight image super-resolution. In *CVPR*.
[29] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*.
[30] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. 2022. NAFSSR: stereo image super-resolution using NAFNet. In *CVPR*.
[31] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*.
[32] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. 2021. Feedback Network for Mutually Boosted Stereo Image Super-Resolution and Disparity Estimation. *ACMMM* (2021).
[33] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *CVPR*.
[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[36] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2020), 2567–2581.

[37] Berk Dogan, Shuhang Gu, and Radu Timofte. 2019. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*.

[38] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *ECCV*.

[39] Chao Dong, Chen Change Loy, and Xiaoou Tang. 2016. Accelerating the super-resolution convolutional neural network. In *ECCV*.

[40] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. 2011. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing* 20, 7 (2011), 1838–1857.

[41] Xiaoyu Dong, Longguang Wang, Xu Sun, Xiuping Jia, Lianru Gao, and Bing Zhang. 2020. Remote Sensing Image Super-Resolution Using Second-Order Multi-Scale Networks. *IEEE Transactions on Geoscience and Remote Sensing* 59, 4 (2020), 3473–3485.

[42] Claude E. Duchon. 1979. Lanczos Filtering in One and Two Dimensions. *Journal of Applied Meteorology and Climatology* 18, 8 (1979), 1016–1022.

[43] Faming Fang, Juncheng Li, and Tieyong Zeng. 2020. Soft-edge assisted network for single image super-resolution. *IEEE Transactions on Image Processing* 29 (2020), 4656–4668.

[44] Chun-Mei Feng, Yunlu Yan, Huazhu Fu, Li Chen, and Yong Xu. 2021. Task transformer network for joint MRI reconstruction and super-resolution. In *MICCAI*.

[45] Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. 2019. Hyperspectral image super-resolution with optimized RGB guidance. In *CVPR*.

[46] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Manga109 dataset and creation of metadata. In *MANPU*.

[47] Dario Fuoli, Luc Van Gool, and Radu Timofte. 2021. Fourier space losses for efficient perceptual image super-resolution. In *ICCV*.

[48] Guangwei Gao, Wenjie Li, Juncheng Li, Fei Wu, Huimin Lu, and Yi Yu. 2022. Feature distillation interaction weighting network for lightweight image super-resolution. In *AAAI*.

[49] Guangwei Gao, Lei Tang, Fei Wu, Huimin Lu, and Jian Yang. 2023. JDSR-GAN: Constructing An Efficient Joint Learning Network for Masked Face Super-Resolution. *IEEE Transactions on Multimedia* 25 (2023), 1505–1512.

[50] Guangwei Gao, Zhengxue Wang, Juncheng Li, Wenjie Li, Yi Yu, and Tieyong Zeng. 2022. Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. *IJCAI* (2022).

[51] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. 2018. Image super-resolution using knowledge distillation. In *ACCV*.

[52] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. 2023. Implicit Diffusion Models for Continuous Super-Resolution. (2023).

[53] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. *NeurIPS*.

[54] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[55] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Andreea-Iuliana Miron, Olivian Savencu, Nicolae-Cătălin Ristea, Nicolae Verga, and Fahad Shahbaz Khan. 2023. Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution. In *CVPR*.

[56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *NeurIPS* (2014).

[57] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. 2019. Blind super-resolution with iterative kernel correction. In *CVPR*.

[58] Jun Gu, Xian Sun, Yue Zhang, Kun Fu, and Lei Wang. 2019. Deep residual squeeze and excitation network for remote sensing image super-resolution. *Remote Sensing* 11, 15 (2019), 1817.

[59] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. 2022. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*.

[60] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. 2020. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*.

[61] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. 2018. Image super-resolution via dual-state recurrent networks. In *CVPR*.

[62] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. 2020. Deep back-projectinetworks for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (2020), 4323–4337.

[63] Juan Mario Haut, Ruben Fernandez-Beltran, Mercedes E Paoletti, Javier Plaza, Antonio Plaza, and Filiberto Pla. 2018. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote sensing* 56, 11 (2018), 6792–6810.
[64] Kaiming He and Jian Sun. 2015. Convolutional neural networks at constrained time cost. In *CVPR*.
[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[66] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS* (2017).
[67] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[68] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).
[69] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*.
[70] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. 2019. Meta-SR: A magnification-arbitrary network for super-resolution. In *CVPR*.
[71] Xuecai Hu, Zhang Zhang, Caifeng Shan, Zilei Wang, Liang Wang, and Tieniu Tan. 2020. Meta-USR: A unified super-resolution network for multiple degradation parameters. *IEEE Transactions on Neural Networks and Learning Systems* 32, 9 (2020), 4151–4165.
[72] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*.
[73] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *CVPR*.
[74] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. 2019. Lightweight image super-resolution with information multi-distillation network. In *ACMMM*.
[75] Zheng Hui, Xiumei Wang, and Xinbo Gao. 2018. Fast and accurate single image super-resolution via information distillation network. In *CVPR*.
[76] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. 2018. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*.
[77] Jian Sun, Zongben Xu, and Heung-Yeung Shum. 2008. Image super-resolution using gradient profile prior. In *CVPR*.
[78] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. 2020. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging* 6 (2020), 1082–1096.
[79] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. 2020. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*.
[80] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.
[81] Alexia Jolicoeur-Martineau. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734* (2018).
[82] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.
[83] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*.
[84] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Deeply-recursive convolutional network for image super-resolution. In *CVPR*.
[85] K. I. Kim and Y. Kwon. 2010. Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 6 (2010), 1127–1133.
[86] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS*.
[87] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.
[88] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.
[89] Rushi Lan, Long Sun, Zhenbing Liu, Huimin Lu, Cheng Pang, and Xiaonan Luo. 2020. Madnet: A fast and lightweight network for single-image super resolution. *IEEE Transactions on Cybernetics* 51, 3 (2020), 1443–1453.
[90] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[91] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

[92] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *CVPR*.
[93] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. 2020. Learning with privileged information for efficient image super-resolution. In *ECCV*.
[94] Sen Lei and Zhenwei Shi. 2021. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–10.
[95] Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. 2020. Pams: Quantized super-resolution via parameterized max scale. In *ECCV*.
[96] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479 (2022), 47–59.
[97] Juncheng Li, Faming Fang, Jiaqian Li, Kangfu Mei, and Guixu Zhang. 2020. MDCN: Multi-scale dense cross network for image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 7 (2020), 2547–2561.
[98] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. 2018. Multi-scale Residual Network for Image Super-Resolution. In *ECCV*.
[99] Juncheng Li, Yiting Yuan, Kangfu Mei, and Faming Fang. 2019. Lightweight and Accurate Recursive Fractal Network for Image Super-Resolution. In *ICCVW*.
[100] Meng Li, Bo Ma, and Yulin Zhang. 2023. Lightweight Image Super-Resolution with Pyramid Clustering Transformer. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
[101] Wenjie Li, Juncheng Li, Guangwei Gao, Weihong Deng, Jiantao Zhou, Jian Yang, and Guo-Jun Qi. 2023. Cross-receptive focused inference network for lightweight image super-resolution. *IEEE Transactions on Multimedia* 26 (2023), 864–877.
[102] Xiang Li, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. 2023. DLGSANet: lightweight dynamic local and global self-attention networks for image super-resolution. In *ICCV*.
[103] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*.
[104] Yong Li, Lei Zhang, Chen Dingl, Wei Wei, and Yanning Zhang. 2018. Single hyperspectral image super-resolution with grouped deep recursive residual network. In *BigMM*.
[105] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. 2022. Blueprint separable residual network for efficient image super-resolution. In *CVPR*.
[106] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. 2019. Feedback network for image super-resolution. In *CVPR*.
[107] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image restoration using swin transformer. In *ICCVW*.
[108] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*.
[109] Jin Lin, Xiaotong Luo, Ming Hong, Yanyun Qu, Yuan Xie, and Zongze Wu. 2023. Memory-Friendly Scalable Super-Resolution via Rewinding Lottery Ticket Hypothesis. In *CVPR*.
[110] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070* (2023).
[111] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. 2022. Blind image super-resolution: A survey and beyond. *IEEE transactions on pattern analysis and machine intelligence* 45, 5 (2022), 5461–5480.
[112] Denghong Liu, Jie Li, and Qiangqiang Yuan. 2021. A Spectral Grouping and Attention-Driven Residual Dense Network for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing* 59, 9 (2021), 7711–7725.
[113] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. 2018. Non-local recurrent network for image restoration. *NeurIPS* (2018).
[114] Jie Liu, Jie Tang, and Gangshan Wu. 2020. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*.
[115] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. 2020. Residual feature aggregation network for image super-resolution. In *CVPR*.
[116] Ziyu Liu, Ruyi Feng, Lizhe Wang, Wei Han, and Tieyong Zeng. 2022. Dual learning-based graph neural network for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.
[117] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *CVPR* (2021).
[118] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*.
[119] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. 2021. Transformer for Single Image Super-Resolution. *CVPRW* (2021).

[120] Xiaotong Luo, Zekun Ai, Qiuyuan Liang, Ding Liu, Yuan Xie, Yanyun Qu, and Yun Fu. 2024. AdaFormer: Efficient Transformer with Adaptive Token Sparsification for Image Super-resolution. In *AAAI*.
[121] Xiaotong Luo, Mingliang Dai, Yulun Zhang, Yuan Xie, Ding Liu, Yanyun Qu, Yun Fu, and Junping Zhang. 2022. Adjustable Memory-efficient Image Super-resolution via Individual Kernel Sparsity. In *ACMMM*.
[122] Xiaotong Luo, Qiuyuan Liang, Ding Liu, and Yanyun Qu. 2021. Boosting lightweight single image super-resolution via joint-distillation. In *ACMMM*.
[123] Xiaotong Luo, Yanyun Qu, Yuan Xie, Yulun Zhang, Cuihua Li, and Yun Fu. 2022. Lattice network for lightweight image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4826–4842.
[124] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. 2020. Unfolding the alternating optimization for blind super resolution. *NeurIPS* (2020).
[125] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. 2020. Structure-preserving super resolution with gradient guidance. In *CVPR*.
[126] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. 2017. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding* 158 (2017), 1–16.
[127] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. 2016. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing* 26, 2 (2016), 1004–1016.
[128] Yinglan Ma, Hongyu Xiong, Zhe Hu, and Lizhuang Ma. 2019. Efficient super resolution using binarized neural network. In *CVPRW*.
[129] Shunta Maeda. 2020. Unpaired image super-resolution using pseudo-supervision. In *CVPR*.
[130] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*.
[131] Kangfu Mei, Aiwen Jiang, Juncheng Li, Bo Liu, Jihua Ye, and Mingwen Wang. 2019. Deep residual refining based pseudo-multi-frame network for effective single image super-resolution. *IET Image Processing* 13, 4 (2019), 591–599.
[132] Kangfu Mei, Aiwen Jiang, Juncheng Li, Jihua Ye, and Mingwen Wang. 2018. An Effective Single-Image Super-Resolution Model Using Squeeze-and-Excitation Networks. In *NeurIPS*.
[133] Shaohui Mei, Xin Yuan, Jingyu Ji, Yifan Zhang, Shuai Wan, and Qian Du. 2017. Hyperspectral image spatial super-resolution via 3D full convolutional neural network. *Remote Sensing* 9, 11 (2017), 1139.
[134] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. 2020. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*.
[135] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.
[136] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a ?completely blind? image quality analyzer. *IEEE Signal Processing Letters* 20, 3 (2012), 209–212.
[137] Abdul Muqeet, Jiwon Hwang, Subin Yang, JungHeum Kang, Yongwoo Kim, and Sung-Ho Bae. 2020. Multi-attention based ultra lightweight image super-resolution. In *ECCV*.
[138] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. 2020. Single image super-resolution via a holistic attention network. In *ECCV*.
[139] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee. 2018. Srfeat: Single image super-resolution with feature discrimination. In *ECCV*.
[140] Cheng Peng, Wei-An Lin, Haofu Liao, Rama Chellappa, and S Kevin Zhou. 2020. SAINT: spatially aware interpolation network for medical slice synthesis. In *CVPR*.
[141] Jinghui Qin, Yongjie Huang, and Wushao Wen. 2020. Multi-scale feature fusion residual network for Single Image Super-Resolution. *Neurocomputing* 379 (2020), 334–342.
[142] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Srobb: Targeted perceptual loss for single image super-resolution. In *ICCV*.
[143] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
[144] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* (2019).
[145] Lee J Rickard, Robert W Basedow, Edward F Zalewski, Peter R Silverglate, and Mark Landers. 1993. HYDICE: An airborne system for hyperspectral imaging. In *Imaging Spectrometry of the Terrestrial Environment*, Vol. 1937. 173–179.
[146] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
[147] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

[148] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. Singan: Learning a generative model from a single natural image. In *ICCV*.
[149] Shuyao Shang, Zhengyang Shan, Guangxing Liu, and Jinglin Zhang. 2023. ResDiff: Combining CNN and Diffusion Model for Image Super-Resolution. (2023).
[150] Mingyu Shen, Pengfei Yu, Ronggui Wang, Juan Yang, Lixia Xue, and Min Hu. 2019. Multipath feedforward network for single image super-resolution. *Multimedia Tools and Applications* 78 (2019), 19621–19640.
[151] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*.
[152] Assaf Shocher, Nadav Cohen, and Michal Irani. 2018. ?zero-shot? super-resolution using deep internal learning. In *CVPR*.
[153] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[154] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. 2023. Hybrid pixel-unshuffled network for lightweight image super-resolution. (2023).
[155] Long Sun, Jinshan Pan, and Jinhui Tang. 2022. Shufflemixer: An efficient convnet for image super-resolution. (2022).
[156] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.
[157] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.
[158] Ying Tai, Jian Yang, and Xiaoming Liu. 2017. Image super-resolution via deep recursive residual network. In *CVPR*.
[159] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. 2017. Memnet: A persistent memory network for image restoration. In *CVPR*.
[160] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*.
[161] Radu Timofte, Rasmus Rothe, and Luc Van Gool. 2016. Seven ways to improve example-based single image super resolution. In *CVPR*.
[162] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. 2017. Image super-resolution using dense skip connections. In *ICCV*.
[163] Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, and Byeungwoo Jeon. 2023. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Transactions on Computational Imaging* 9 (2023), 350–366.
[164] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
[165] C Wang, Z Li, and J Shi. 2019. Lightweight image super-resolution with adaptive weighted learning network. *arXiv preprint arXiv:1904.02358* (2019).
[166] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. 2023. Omni aggregation networks for lightweight image super-resolution. In *CVPR*.
[167] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015* (2023).
[168] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. 2021. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*.
[169] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. 2020. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
[170] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. 2020. Dual super-resolution learning for semantic segmentation. In *CVPR*.
[171] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. 2021. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In *CVPR*.
[172] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2019. Learning parallax attention for stereo image super-resolution. In *CVPR*.
[173] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2021. Learning a single network for scale-arbitrary super-resolution. (2021).
[174] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *CVPR*.
[175] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *CVPR*.

[176] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*.
[177] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*.
[178] Yanbo Wang, Chuming Lin, Donghao Luo, Ying Tai, Zhizhong Zhang, and Yuan Xie. 2023. High-Resolution GAN Inversion for Degraded Images in Large Diverse Datasets. *AAAI* (2023).
[179] Yunlong Wang, Fei Liu, Kunbo Zhang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. 2018. LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing* 27, 9 (2018), 4274–4286.
[180] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. 2018. A fully progressive approach to single-image super-resolution. In *CVPRW*.
[181] Yu Wang, Zhenfeng Shao, Tao Lu, Changzhi Wu, and Jiaming Wang. 2023. Remote sensing image super-resolution via multiscale enhancement network. *IEEE Geoscience and Remote Sensing Letters* 20 (2023), 1–5.
[182] Yukai Wang, Qizhi Teng, Xiaohai He, Junxi Feng, and Tingrong Zhang. 2019. CT-image of rock samples super resolution using 3D convolutional neural network. *Computers & Geosciences* 133 (2019), 104314.
[183] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. 2022. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 425–443.
[184] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. 2019. Flickr1024: A large-scale dataset for stereo image super-resolution. In *ICCVW*.
[185] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. 2020. Spatial-angular interaction for light field image super-resolution. In *ECCV*.
[186] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. 2020. Light field image super-resolution using deformable convolution. *IEEE Transactions on Image Processing* 30 (2020), 1057–1071.
[187] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. 2021. Symmetric parallax attention for stereo image super-resolution. In *CVPR*.
[188] Zhou Wang and Alan C Bovik. 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine* 26, 1 (2009), 98–117.
[189] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.
[190] Zhihao Wang, Jian Chen, and Steven CH Hoi. 2020. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3365–3387.
[191] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022. Uformer: A general u-shaped transformer for image restoration. In *CVPR*.
[192] Zhengxue Wang, Guangwei Gao, Juncheng Li, Yi Yu, and Huimin Lu. 2021. Lightweight Image Super-Resolution with Multi-scale Feature Interaction Network. In *ICME*.
[193] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. 2023. DR2: Diffusion-based Robust Degradation Remover for Blind Face Restoration. In *CVPR*.
[194] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. 2020. Component divide-and-conquer for real-world image super-resolution. In *ECCV*.
[195] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. 2021. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*.
[196] Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2021. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *CVPR*.
[197] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. 2022. Efficient non-local contrastive attention for image super-resolution. In *AAAI*.
[198] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. Diffir: Efficient diffusion model for image restoration. *ICCV* (2023).
[199] Xiaoyu Xiang, Qian Lin, and Jan P Allebach. 2021. Boosting high-level vision with joint compression artifacts reduction and super-resolution. In *ICPR*.
[200] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. 2020. Unified dynamic convolutional network for super-resolution with variational degradations. In *CVPR*.
[201] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning texture transformer network for image super-resolution. In *CVPR*.
[202] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. 2010. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* 19, 11 (2010), 2861–2873.

[203] Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. 2017. Deep edge guided recurrent residual learning for image super-resolution. *IEEE Transactions on Image Processing* 26, 12 (2017), 5895–5907.
[204] Fangchen Ye, Jin Lin, Hongzhan Huang, Jianping Fan, Zhongchao Shi, Yuan Xie, and Yanyun Qu. 2023. Hardware-friendly Scalable Image Super Resolution with Progressive Structured Sparsity. In *ACMMM*.
[205] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. 2020. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters* 27 (2020), 496–500.
[206] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. 2017. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters* 24, 6 (2017), 848–852.
[207] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. 2015. Learning a deep convolutional network for light-field image super-resolution. In *ICCVW*.
[208] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved VQGAN. *arXiv preprint arXiv:2110.04627* (2021).
[209] Xin Yu and Fatih Porikli. 2017. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*.
[210] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. 2018. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*.
[211] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. 2013. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing* 22, 12 (2013), 4865–4878.
[212] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*.
[213] Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh. 2020. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications* 139 (2020), 112854.
[214] Roman Zeyde, Michael Elad, and Matan Protter. 2010. On single image scale-up using sparse-representations. In *ICCS*.
[215] Dongyang Zhang, Jie Shao, Xinyao Li, and Heng Tao Shen. 2020. Remote sensing image super-resolution via mixed high-order attention network. *IEEE Transactions on Geoscience and Remote Sensing* 59, 6 (2020), 5183–5196.
[216] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. 2018. Super-identity convolutional neural network for face hallucination. In *ECCV*.
[217] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*.
[218] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. 2024. Transcending the Limit of Local Window: Advanced Super-Resolution Transformer with Adaptive Token Dictionary. (2024).
[219] Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie Guo, Xinbo Gao, and Jing Zhang. 2023. ESSAformer: Efficient Transformer for Hyperspectral Image Super-resolution. In *CVPR*.
[220] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.
[221] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. 2019. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*.
[222] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. 2022. Efficient long-range attention network for image super-resolution. In *ECCV*.
[223] Xindong Zhang, Hui Zeng, and Lei Zhang. 2021. Edge-oriented convolution block for real-time super resolution on mobile devices. In *ACMMM*.
[224] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. 2021. Data-free knowledge distillation for image super-resolution. In *CVPR*.
[225] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*.
[226] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*.
[227] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. 2019. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082* (2019).
[228] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *CVPR*.
[229] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. 2021. Context reasoning attention network for image super-resolution. In *ICCV*.

[230] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. 2019. Image super-resolution by neural texture transfer. In *CVPR*.
[231] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. 2019. Channel splitting network for single MR image super-resolution. *IEEE Transactions on Image Processing* 28, 11 (2019), 5649–5662.
[232] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. 2018. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *ECCV*.
[233] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. 2015. Learning face hallucination in the wild. In *AAAI*.
[234] Lin Zhou, Haoming Cai, Jinjin Gu, Zheyuan Li, Yingqi Liu, Xiangyu Chen, Yu Qiao, and Chao Dong. 2022. Efficient image super-resolution using vast-receptive-field attention. In *ECCV*.
[235] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. 2022. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS* (2022).
[236] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. 2023. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*.
[237] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. 2022. Blind face restoration via integrating face shape and generative priors. In *CVPR*.
[238] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
[239] Qiang Zhu, Pengfei Li, and Qianhui Li. 2023. Attention Retractable Frequency Fusion Transformer for Image Super Resolution. In *CVPR*.
[240] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. 2016. Deep cascaded bi-network for face hallucination. In *ECCV*.
[241] Maria Zontak and Michal Irani. 2011. Internal statistics of a single natural image. In *CVPR*.