Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230567

Image super-resolution techniques using deep neural networks

Meilin Guo

Australian National University

U7094624@anu.edu.au

Abstract. Super-resolution (SR) based on deep convolutional neural networks is a rapidly developing field with many real-world applications. In this paper, we examine cutting-edge super-resolution neural networks in-depth using freshly released difficult datasets to test single-image SR. We present a taxonomy that divides existing techniques into six categories, including upsampling, residual, recursive, dense connection, attention-based, and loss function designs. This taxonomy is applicable to deep learning-based SR networks. The comprehensive analysis shows that in the past few years, the accuracy has increased steadily and rapidly, while the complexity of the model and the accessibility of large-scale information have also increased accordingly. It has been noted that the present techniques have greatly outperformed the past techniques that were indicated as benchmarks. On this basis, this paper will put forward some suggestions for future research.

Keywords: Image Super-resolution (SR), Deep Learning, Convolutional Neural Networks (CNNs), Computer Vision, Survey

1. Introduction

Image super-resolution (SR) is a method of producing a raster image with a higher resolution than its source. One or more images or frames could be the source. The source for this study is a single raster image, and the focus is single-frame super-resolution. It is a significant branch of image processing techniques in computer vision. HR images with superior landscape details and constituent objects are valuable and necessary for many hardware devices, i.e., large-scale monitors, high-definition televisions and cell phones, cameras, etc. In addition, single image super-resolution (SISR) is also useful in other domain-specific artificial intelligence (AI) tasks like object detection [1], forensics [2], medical imaging [3], image interpretation in remote sensing [4], and face recognition [5].

However, after many years of research, the classic topic of SR is still regarded as a difficult research subject in computer vision along with some limitations. First, image SR is an unpredictable inverse problem. In particular, there are several options available to get an SR image for the LR image [6]. Therefore, reliable prior knowledge is typically needed to define the solution space. Second, the up-sampling scale factors are mostly limited, constrained by the integer scales (x2, x3, x4, or more), and the same image SR of different factors are treated as the independent tasks. However, in practical situations, it is common for the image SR to use the user-specified scale factor. Consequently, it provides a new direction for solving the generality of the image SR, like the reconstruction in an unconstrained space. Meta-SR [7] and LIIF-SR [8] show great performance in targeting this problem. Third, with the upscaling factor rising, the problem complexity and the reconstruction time both rise significantly. A higher scaling factor makes it considerably more difficult to recover the details and further results in the replication of inaccurate information (or artifacts). Additionally, evaluating the quality of the output SR image is not easy since quantitative metrics, such as peak signal to noise ratio (PSNR) and structural similarity (SSIM), only objectively reflect the SR image quality without considering the human subjective perception.

The earliest image super-resolution (SR) was basically done using traditional direct image processing methods, but with the rapid development of neural networks, it is much more efficient and accurate to training complex models. Traditional methods including interpolation-based methods like the nearest neighbor, bilinear and bicubic interpolation and reconstruction-based methods, use existing pixels to generate new ones so that they cannot offer any additional information for the SR image and the lost information cannot be restored. Reconstruction-based methods typically impose specific knowledge priors or limitations on an inverse reconstruction problem [9]. Although traditional algorithms have been around for decades, deep learning-based models outperform most of them. Therefore, most current techniques depend on data-driven deep learning models to rebuild the necessary information for precise image SR. Therefore, a subset of deep learning neural networks is designed to automatically learn the relationships between input and output directly from the data in such a scenario. This study attempts to provide a thorough introduction to the area of SISR with a primary focus on deep learning-based techniques.

Many deep learning-based models are different in many ways. In this paper, our contribution is: (1) We summarize the state-of-the-art approaches based on supervised deep learning in recent years from three major aspects: upsampling methods, network design architectures, and loss functions. (2) We give a full analysis of advantages and disadvantages of each. (3) We offer a comprehensive experimental evaluation of models on several public datasets for image SR. (4) We provide some insights and suggestions for potential future directions.

2. SR model design methodologies

The state-of-the-art approaches based on supervised deep learning in recent years include two major aspects – upsampling frames and network designs. Details are demonstrated in this section.

2.1. Upsampling

Many image SR models are proposed based on the rise of deep learning networks. These models can be divided into three frames by the location of upsampling modules in the models, including pre-upsampling, post-upsampling, and progressive upsampling. These are demonstrated separately as below.

2.1.1. Pre-upsampling. It is difficult to learn HR images directly from LR images. It is a direct solution to optimize and reconstruct high-quality details through CNN by sampling from traditional methods, such as the bicubic interpolation [10]. Dong et al. proposed the first deep learning-based image SR model, super-resolution convolutional neural network (SRCNN) [11], as Fig. 1a shows, which outperformed traditional algorithms. The CNN simply must improve the HR image after upsampling using the conventional approach, considerably reducing the learning challenge. An interpolated image of any size can be used as input, and the effect is comparable to that of a single-scale model. The side effects of the up-sampling method are noise amplification, blurring, time, and space cost of calculation in high dimensional space.

2.1.2. Post-upsampling. To increase computational effectiveness and fully utilise deep learning technologies, researchers suggest carrying out as many as procedures in low-dimensional space and then perform upsampling operations as Fig. 1b shows. The benefit of this strategy is that the expensive feature extraction procedure only operates in low-dimensional space, which greatly reduces the computational amount and spatial complexity. Consequently, this framework has emerged as one of the most widely used frameworks and has been widely used in recent years. Faster Region-CNN (FSRCNN) [12] and

Efficient Sub-pixel Convolutional Network (ESPCN) [13] are deep learning models that first used a post-upsampling frame. Certainly, it resulted in a reduction in the number of operations compared to SRCNN. However, if there are not enough layers after the upsampling process, the overall performance will degrade. Furthermore, they cannot be trained on multiple scales because the input image size is different for each upsampling scale [14].

2.1.3. Progressive upsampling. The SR framework for post-upsampling still has certain flaws. First, the learning complexity is greatly raised since upsampling is done in only one step, especially for high scaling factors (4 or 8). Second, if a requirement for multi-scale SR arises, a totally new SR model is required for each scaling factor, which significantly slows down the pace of trials. As shown in Fig. 1c, the progressive upsampling architecture was proposed to solve previous difficulties and was initially utilized by the Laplacian pyramid SR network (LapSRN) [15]. This system uses a cascade of CNN to rebuild HR pictures in a step-by-step manner. At each step, the pictures are upsampled to a higher resolution and improved using CNN. This paradigm has been used in other research, such as MS-LapSRN [16] and progressive SR (ProSR) [17], with comparable findings.

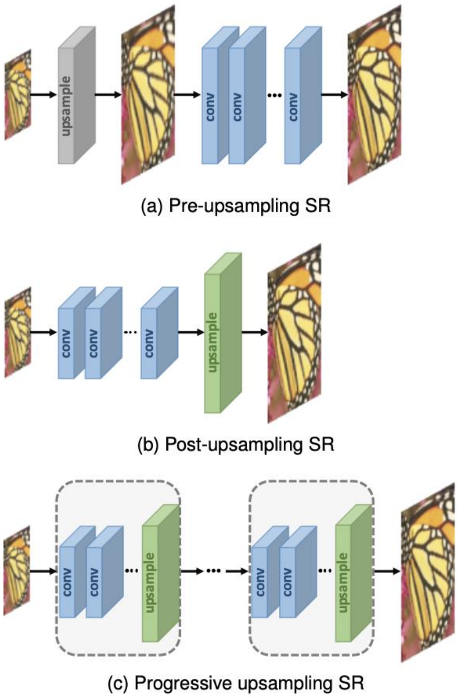

Figure 1. SR model upsampling frames based on deep learning. The output size is represented by the cube size. Predefined upsampling is represented by the grey ones, whereas learnable upsampling and convolutional layers are represented by the green and blue ones, respectively. Stackable modules are represented by the blocks bounded by dashed boxes [18].

2.2. Deep neural network (DNN) design

Network architecture is one of the most important components of deep learning. Researchers in the SR field use a range of network design approaches based on four SR frameworks to construct models. In this section, we break down these learning models into basic network design concepts or techniques, present them, and examine their benefits and drawbacks one by one.

2.2.1. Residual learning. Prior to He et al .'s proposal of ResNet [19], residual learning was extensively employed by SR models [20, 21], as seen in Fig. 2a. ResNet was offered as a method of learning residuals rather than a thorough mapping. Residual learning employs skip connections to prevent gradients from disappearing, creating extremely deep networks. Its importance was initially proven in the context of the picture categorization [19]. Several networks [22] have recently used residual learning to improve SR performance. Algorithms in this technique learn residual or high-frequency differences from the input to the ground truth [6]. Interestingly, despite the usage of only three layers, the result surpasses the non-deep learning techniques outlined earlier.

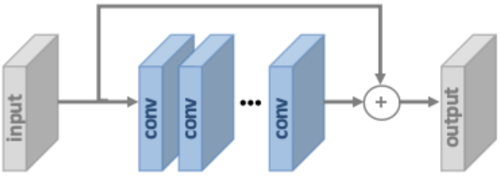

Figure 2. Residual learning block [18].

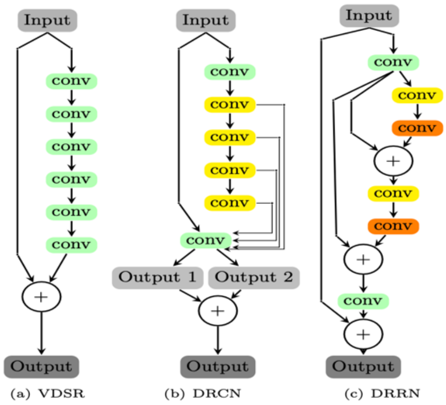

Figure 3. VDSR, DRCN, and DRRN models. The sharing parameters are shown by the same shade of yellow or orange [9].

Kim et al., motivated by incredibly deep networks' achievements, introduced two models: Very Deep Convolutional Network (VDCN) [23] and Deeply-Recursive Convolutional Network (DRCN) [24], both of which stack 20 convolutional layers, as illustrated in Fig. 3. (a, b). Tai et al. employed both global and local residual connections in their DRRN to achieve greater benefits from residual learning [25]. As shown in Fig.3 (c) [23], the identity branch makes use of global residual learning.

2.2.2. Recursive learning. Recursive learning, which involves applying the same modules numerous times in a recursive pattern, is incorporated into the SR field aiming to learn higher-level traits without introducing overpowering parameters, as shown in Fig. 4. The basic purpose of this design is to gradually break down the complex SR tasks into a collection of easier to solve task.

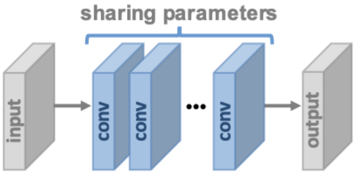

Figure 4. Recursive learning block [18].

DRCN, as discussed before, also applies the same smaller convolution nets, embedding net, inference net, and reconstruction net, multiple times, as Fig. 5 shows. Each recursion increases the size of the receptive field. The inference net produces HR feature maps, while the reconstruction net converts to grayscale or color images. Tai et al. also employed both global and local residual connections in their

Deep Recursive Residual Network (DRRN) to achieve greater benefits from residual learning [25], as Fig.5 shows.

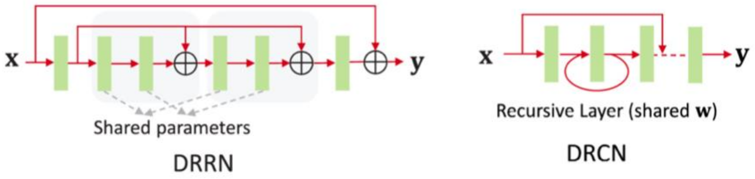

Figure 5. A glimpse of DRRN and DRCN architecture. The green blocks are convolution layers (generally followed by ReLU). ⊕ represents element-wise addition [6]

2.2.3. Dense connections. After the achievement of the DenseNet [26] proposed for image recognition, dense CNN layer connections have also been proposed to increase performance in image SR field, as Fig. 6 shows. In a l^{th} layer dense block ($l \geq 2$), each layer receives its inputs from the feature maps of all preceding layers, while the feature maps are fed into the following layers, producing in $l \cdot (l-1)/2$ connections [18]. The major goal of this design is to utilize hierarchical cues accessible as the model goes deeper to gain more flexibility and enrich feature representations.

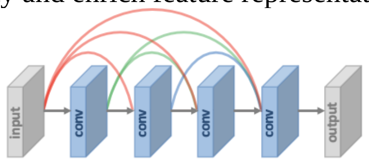

Figure 6. Dense connections [18].

The very first network that utilized densely connected networks to image SR is SRDenseNet [27]. It learnt the architecture from DenseNet [26], where a layer works on the output of each preceding layer. The vanishing gradient problem is avoided, intensive models may be learned, and the training process is sped up with this information transfer. Another net that uses this idea is the Residual Dense Network (RDN) [28], which combines the concepts of residual skip connections and dense connections. The suggested method is resistant to a variety of picture deterioration seen in LR photos and recovers significantly upgraded SR images.

2.2.4. Attention mechanism. For image SR, the previously stated network architectures assign a consistent value to all spatial positions and channels. However, in some cases, it might be more beneficial to only focus on a few elements of a specific layer. Attention-based models provide flexibility because it recognizes that not all features are equally important for SR. Recent attention-based models have demonstrated considerable progress for image SR when combined with deeper networks.

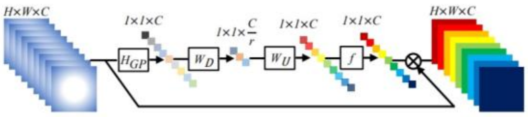

Figure 7. Channel attention block [29].

A deep CNN design that was just recently developed is the residual channel attention network (RCAN) [29]. Moreover, it has been the deepest model for image SR field. As seen in Fig. 7, it included a channel attention mechanism within the residual block. By averaging across a spatial dimension of H*W, the inputs of shape H*W*C are squeezed into the channel descriptor, yielding the output shape of 1*1*C. Channel descriptors are placed by gating the activation of a sigmoid function and multiplying it element by element with the input, allowing the user to adjust the amount of information is sent to the next layer for each channel.

3. Loss function

A loss function offers a guide to measure error for model optimization[18]. Researchers first used the pixel-wise loss to evaluate reconstruction quality, but they discover that it is not always the case that lower pixel-wise loss matches high human perceptional quality. Therefore, several loss functions (such as adversarial loss [30] and content loss [31]) are used to provide more accurate and high-quality results by better designs for the reconstruction error.

In this section, $\hat{I}$ denotes the target HR image and I denotes the generated HR image.

3.1 Pixel loss

Pixel-wise loss always refers to distance l_1 or distance l_2 (MSE) or a combination of both[32].

$$L_{pixel-l1}\left(\hat{I},I\right)=\frac{1}{hwc}\sum_{i,j,k}|\hat{I}_{i,j,k}-I_{i,j,k}|$$

$$L_{pixel-l2}\left(\hat{I},I\right)=\frac{1}{hwc}\sum_{i,j,k}(\hat{I}_{i,j,k}-I_{i,j,k})^2$$

where h, w and c are the height, width, and the number of channels of LR images respectively. [18] Practically speaking, the l_1 loss outperforms and converges more than the l_2 loss [14], [22], [33]. These measurements only capture information at the local pixel level, hence the generated pictures don't always produce perceptually results, where too smooth images and poor perceptual quality might arise.

3.2 Content loss

Content loss is firstly introduced to assess the perceptual quality of images [31], [34].

$$L_{content}\left(\hat{I},I;\varphi,l\right)=\frac{1}{h_lw_lc_l}\sqrt{\sum_{i,j,k}(\varphi^{(l)}{}_{i,j,k}(\hat{I})-\varphi^{(l)}{}_{i,j,k}(I))^2}$$

where h_l, w_l and c_l indicate the height, width and the number of channels on layer l, resepectively. [18] In fact, the leant understanding from more complex and concrete image features from the model φ is evaluated by content loss. In contrast to pixel-wise loss, which forces the output image $\hat{I}$ and the target image I to match the same pixels, the content loss encourages it to be perceptually like the target image I. As a result, it generates visually more perceptible images [35],[36], where the Visual Geometry Group (VGG) net [37] and ResNet [19] are widely used pre-trained models.

3.3 Texture loss

Texture loss, also known as style reconstruction loss, is incorporated into SR because the consistent requirements for the image color, texture and other aspects with the target image. The texture matching loss is measured as the l_1 loss between gram matrices generated from deep features [6].The Gram matrix is defined as $G^{(l)}\epsilon\mathbb{R}^{c_l\times c_l}$,where $G_{i,j}^{(l)}$ is the inner product of the feature maps i and j on layer l. [18]

$$G_{i,j}^{(l)}(I)=vec\left(\varphi_i^{(l)}(I)\right)vec\left(\varphi_j^{(l)}(I)\right)$$

where $vec()$ indicates the operation on vectors and $\varphi_i^{(l)}(I)$ indicates the i-th channel of the feature maps on layer l of image I. [18] Then the texture loss is indicated by:

$$L_{texture}\left(\hat{I},I;\varphi,l\right)=\frac{1}{c_l}\sqrt{\sum_{i,j}(G_{i,j}^{(l)}(\hat{I})-G_{i,j}^{(l)}(I))^2}\ [18]$$

The EnhanceNet [38] developed by Sajjadi et al. uses texture loss to produce significantly more realistic textures and visually better results. Even yet, the process of choosing a suitable patch size is still empirical. Due to the fact that texture statistics are averaging over locations with different textures, too little or too large of a patch might create artefacts in textured sections or the entire image.

3.4 Adversarial loss

Due to their strong capacity for learning, GANs [39] have gained a lot of attention in recent years and have been given diverse visual tasks. The resulting generator might generate outputs matching the distribution of real data with sufficient iterative adversarial training, yet the discriminator is unable to tell the difference between produced and real data. Adopting adversarial learning in the context of image SR is simple. In that instance, all that is required is to consider the model as a generator and construct an additional discriminator to determine whether the HR image was generated. As a result, Ledig et al. [30] initially proposed that SRGAN could use the following adversarial loss based on cross entropy:

$$L_{gan\_ce\_g}\left(\hat{I}; D\right) = log\, D(\hat{I})$$

$$L_{gan\_ce\_d}\left(\hat{I}, I_s, ; D\right) = log\, D(\, I_s) log(1D\left(\hat{I}\right))$$

where $L_{gan\_ce\_g}$ and $L_{gan\_ce\_d}$ indicates, respectively, the adversarial loss of the generator (i.e., the SR model) and the discriminator D (i.e., a binary classifier) and I_s represents images randomly sampled from the ground truths [18].

From another perspective, images can be interpreted as sampled from a high-dimensional probability distribution, and that is a crucial connection between images and statistics. We use the probability distribution, which spans the pixels of photographs, to determine whether an image is unmodified. In such cases, the Kullback-Leibler Divergence measurement is used to quantify the difference between the ground truth distribution and generated distribution. It is supposed that human have learned the ground truth image as a natural distribution or a type of prior when the Kullback-Leibler Divergence between two distributions of the ground truth image and SR image reaches the minimum 0. Adversarial loss is referred to as a component of the perceptual loss in SRGAN [30], and the GAN-based model seeks to promote reconstructed images to have a distribution that is similar to that of the ground truth images. When dealing with the intricate manifold distributions in natural photographs, adversarial learning is helpful.

4. Evaluation and discussion

In this section, we compare the most recent accessible benchmark datasets, including Set5 [40], Set14 [41], BSD100 [42], Urban100 [43], DIV2K [44], and Manga109 [45].

4.1 Experiment setup

4.1.1. SR models **SRCNN.** SRCNN [11] is the initial CNN model for image SR with just traditional convolutional layers. A few subsequent initiatives in deep learning-based image SR have been inspired by this project, which undoubtedly represents the ground-breaking work in the field.

DRLN. Densely Residual Laplacian Network (DRLN) [46] is a recently proposed modular and hierarchal network. Through its modular architecture and cascading connections, the system is able to be exploited across a variety of connections due to its densely connected residual units. There is a replication of the structure in every block of the network.

SCN. This sparse coding network (SCN) mimics LISTA, which is a network that aims to be as compact as possible. By combining sparse coding[47] with deep CNNs, it provides a method that yields better results.

VDSR. Very Deep Super-Resolution (VDSR)[23], first proposed in [37], is inspired by a very popular deep CNN architecture, VGG net. To accelerate training, every network layer uses fixed-size convolutions. Their findings lend credence to the idea that deeper networks can learn generalizable representations for multi-scale SR and offer improved contextualization.

ESRGAN. Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [35] is extended based on SRGAN[35] by introducing dense blocks. A residual connection between each dense block is also created by reusing the outputs of dense blocks as inputs. To enforce residual learning, ESRGAN additionally contains a global residual connection. Additionally, the authors use a Relativistic GAN, an improved discriminator [48].

EDSR. Enhanced Deep Super-Resolution (EDSR) [22] is mainly modified upon the ResNet architecture [19] for image recognition. Upon removing Batch Normalization (for each residual block) and ReLU activation, they were able to achieve notable improvements (for the residual blocks not included in the analysis).

RCAN. Residual Channel Attention Network (RCAN) [29] provides different pathways for information to move between layers for information to pass from the beginning levels towards. It effectively illustrates the links between feature maps and enables the model to concentrate on more crucial and specific feature maps.

4.1.2. Datasets. **Set5** [42] is a traditional dataset that simply contains five test images of a baby, bird, butterfly, head, and woman.

Set14 [41] contains more categories than Set5 but only 14 images.

BSD100 [42] is another traditional dataset with 100 test photos that Martin et al. proposed. The dataset contains a variety of images including natural images and object-specific images such as food and people.

Urban100 [43] , a more recent dataset, has the same number of images as BSD100, and is proposed by Huang et al[42] but concentrated on man-made architectures.

DIV2K [44] is a dataset for NTIRE challenge. Approximately 800 images are included in the training module and 100 images are included in the testing module.The image quality is of 2K resolution. To maintain privacy, the results of all algorithms are only published on validation images, as the test set may not be accessed by the general public.

Manga109 [45] is the most recent improvement for assessing SR methods. A manga volume is represented by 109 images in the dataset. Until 2010, these manga were unavailable for commercial use due to their professional drawing by Japanese artists.

4.1.3. Metrics. The two major methods of evaluating the quality of SR photos are subjective evaluation based on human perception and objective evaluation based on quality measures. The two major methods of evaluating the quality of SR photos are subjective evaluation based on human perception and objective evaluation based on quality measures. Overall, the former measures just variations at the pixel level and is more in line with the realistic requirement. However, subjective evaluation has the following limitations. (i) Personal preferences might easily influence the outcome of the examination. (ii) The evaluation procedure is not automatable and frequently expensive. In comparison, objective evaluation is easier to utilise, although the outcomes of various assessment measures may not always be as consistent as those of subjective evaluation. Commonly used metrics for evaluating the objective quality of super-resolved images are PSNR, SSIM [49].

a) Peak Signal-to-Noise Ratio (PSNR)

A PSNR measurement is one of the most widely used measurements of reconstruction quality. (e.g., image compression and image inpainting). A measure of pixel value is defined based on the maximum value of a pixel (referred to as L) and the mean squared error between two images in an image SR.The PSNR between the reconstruction $\hat{I}$ and the target image I with N pixels is defined as follows:

$$PSNR = 10 \cdot log_{10}(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I(i) - \hat{I}(i))^2})$$

where L equals to 255 in general cases using 8-bit representations. [18] PSNR does not consider human perceptions when representing reconstruction quality in real scenes, since it only considers pixel-wise differences. This can often lead to poor representation of reconstruction quality in real scenes. As a result of the requirement to compare with literature publications and the absence of a completely correct measure of perception, PSNR is still the most used evaluation criteria for SR models.

Since PSNR only considers pixel-wise differences, it frequently performs poorly when attempting to depict the reconstruction quality in real scenes. As people concern more about visual effects, PSNR is not sufficient for a full measurement. Despite this, PSNR is still the most popular evaluation criterion for SR models because it must be compared to published literature and there aren't any completely accurate perceptual metrics.

b) SSIM index

SSIM index [49] is suggested for measuring the image quality as a combination of luminance, contrast, and structures, considering that the image structure extraction is highly suited for the human visual system (HVS) [50]. The luminance μ_I and contrast σ_I are calculated as the mean and standard deviation of the image intensity, respectively, for an image I of N pixels. To be more specific, the comparisons are carried out in conjunction in the areas of luminance, contrast, and structures as

$$SSIM = [l(X,\hat{X})]^{\alpha}[c(X,\hat{X})]^{\beta}[s(X,\hat{X})]^{\gamma}$$

where $l(X,\hat{X}) = \frac{2\mu_X\mu_{\hat{X}}+C_1}{\mu_X{}^2+\mu_{\hat{X}}{}^2+C_1}$, $c(X,\hat{X}) = \frac{2\sigma_X\sigma_{\hat{X}}+C_2}{\sigma_X{}^2+\sigma_{\hat{X}}{}^2+C_2}$ and $s\alpha$, β, and γ are weighting parameters. μ_X and σ_X denote the mean and standard deviation of X, respectively. The same as $\mu_{\hat{X}}$ and $\sigma_{\hat{X}}$ of $\hat{X}$. $\sigma_{X\hat{X}}$ is the covariance between $\hat{X}$ and X. C_1, C_2, and C_3 are constants. Additionally, the equation above can be simplified whenever $\alpha = \beta = \gamma = 1$ and $C_3 = \frac{C_2}{2}$ as [49]:

$$SSIM = \frac{(2\mu_X\mu_{\hat{X}}+C_1)(2\sigma_{X\hat{X}}+C_2)}{(\mu_X{}^2+\mu_{\hat{X}}{}^2+C_1)(\sigma_X{}^2+\sigma_{\hat{X}}{}^2+C_2)}$$

In comparison, SSIM [49] reflects more accurate visual quality than PSNR. When a ground truth image is available, PSNR and SSIM [49] are typically joint to evaluate the restored image's quality.

c) Mean opinion score (MOS)

Mean opinion score (MOS) is a popular subjective image quality assessment technique, in which human raters rate the perceptual quality of the testing images. Specifically, the ratings range from 1 (poor) to 5. (good). Additionally, the arithmetic mean of all ratings is used to calculate the final MOS.

Although MOS score appears to be a reliable image quality assessment technique, it has certain inherent flaws, including rating of criteria as non-linearly perceived scales, biases, and volatility. In practice, many image SR models outperform others in terms of perceptual quality while doing badly in standard image quality assessment metrics (e.g., PSNR). Considering these circumstances, the MOS score provides a reliable method of accurately assessing the perceptual quality of an image. [16, 30, 36, 38].

d) Other Methods

Compliance with perceived quality is a fundamental problem for common picture quality metrics like PSNR. Images consequently become overly flattened and lack texture detail.

Numerous perceptual loss measures have been suggested to address this problem. More and more modern perceptual metrics, such as LPIPS [51] and PieAPP [52], are designed to mimic human perception of pictures in contrast to the more traditional ones, which were fixed, such as SSIM [49] and multi-scale SSIM [53]. Each of these measures suffered a specific failure. Therefore, there isn't a single perceptual metric that consistently performs well in all circumstances and accurately measures image quality. Therefore, a special challenge and workshop for perceptually sound image SR techniques (PIRM 2018) have been launched to promote development in this field [54].

4.2 Performance

To summarize the current performance of the image SR models based on deep learning, we offer a range of comparisons in Table 1. We use two image objective quality measurements: SSIM and PSNR to evaluate performance. The higher the measurement score is, the better the quality of the reconstructed image.

In general, ESRGAN performs better for 4× and both ESRGAN and DRLN perform better for 8× in terms of PSNR and SSIM. RCAN is following closely behind as well. However, many factors make it impossible to identify one approach to be clearly superior to the others, such as the network complexity, the depth of the model, the size of the training patch, the number of feature mappings, and others. Only by maintaining consistency across all the factors is a fair comparison feasible.

Table 1. Performance of State-of-the-Art models for PSNR and SSIM on widely used public datasets for 4x and 8x.

Scale	Method	SET5 [42]		SET14 [43]		BSD100 [42]		Urban100 [43]		Manga109 [45]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
4x	SRCNN	30.48	0.8628	27.45	0.7503	26.90	0.7108	24.51	0.7229	25.79	0.7310
	DRLN	32.63	0.9004	28.92	0.7900	27.83	0.7450	26.94	0.8122	**31.51**	**0.9194**
	SCN	31.04	0.8775	27.76	0.7623	27.11	0.7196	25.63	0.7469	27.94	0.8268
	VDSR	31.35	0.8838	28.05	0.7679	27.23	0.7256	25.21	0.7519	27.24	0.7953
	ESRGAN	**32.58**	**0.9747**	**30.24**	**0.8894**	**29.26**	**0.8460**	**27.89**	**0.8438**	29.74	0.8631
	EDSR	31.03	0.8648	27.87	0.7424	27.02	0.7079	24.82	0.7375	22.47	0.7034
	RCAN	32.61	0.9002	28.88	0.7889	27.75	0.7436	26.81	0.8087	31.22	0.9173
8x	SRCNN	25.03	0.6781	23.75	0.6214	24.23	0.5778	20.91	0.5559	22.49	0.7010
	DRLN	27.35	**0.7904**	25.67	**0.6541**	25.03	**0.6150**	23.04	**0.6472**	25.31	0.8004
	SCN	25.58	0.7068	23.97	0.6023	24.28	0.5696	21.73	0.5563	22.70	0.6898
	VDSR	25.92	0.7238	24.25	0.6179	24.53	0.5826	21.71	0.5709	23.16	0.7249
	ESRGAN	**27.58**	0.7902	**25.46**	**0.6541**	**25.14**	0.6076	**23.31**	0.6072	**25.49**	**0.8079**
	EDSR	26.93	0.7759	24.92	0.6421	24.82	0.5984	22.52	0.6222	24.70	0.7842
	RCAN	27.29	0.7876	25.33	0.6541	24.98	0.6056	23.05	0.6469	25.28	0.8040

5. Conclusion

SISR is a difficult scientific problem with significant real-world applications. Deep CNN-based solutions for image SR have rapidly expanded because of the remarkable success of deep learning methodologies. Exciting progress of network designs and learning technology has led to a wide range of methodological suggestions. This paper provides a comprehensive inspection of current SR techniques based on deep learning. We identify the following trends in the current art through thorough quantitative and qualitative comparisons: (a) Spatial details in an image are more accurately reserved by reconstruction error-based approaches, whereas GAN-based approaches typically produce visually pleasant outputs. (b) For large magnification rates (i.e., 8 or above), the present models typically produce less-than-ideal results. (c) the method with the best performance is usually more complicated and in-depth in calculation than its competitors. (d) The residual learning simplifies the learning process through signal decomposition, thus significantly improving the performance. On the whole, we can see that SR performance has improved significantly recently, and the complexity of the network has also increased. Surprisingly, there are still some problems in the most advanced methods, which restrict their applications in practice, such as inadequate evaluation metrics and high level of model complexity. We expect that this paper will promote more work on these pressing issues.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142-158, 2015.

[2] A. Swaminathan, M. Wu, and K. R. Liu, "Digital image forensics via intrinsic fingerprints," IEEE transactions on information forensics and security, vol. 3, no. 1, pp. 101-117, 2008.

[3] H. Greenspan, "Super-resolution in medical imaging," The computer journal, vol. 52, no. 1, pp. 43-63, 2009.

[4] T. Lillesand, R. W. Kiefer, and J. Chipman, Remote sensing and image interpretation. John Wiley & Sons, 2015.

[5] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 5, pp. 1034-1040, 2015.

[6] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," ACM Computing Surveys (CSUR), vol. 53, no. 3, pp. 1-34, 2020.

[7] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1575-1584.

[8] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8628-8638.

[9] V. K. Ha, J. Ren, X. Xu, S. Zhao, G. Xie, and V. M. Vargas, "Deep learning based single image super-resolution: A survey," in International Conference on Brain Inspired Cognitive Systems, 2018: Springer, pp. 106-119.

[10] R. Keys, "Cubic convolution interpolation for digital image processing," IEEE transactions on acoustics, speech, and signal processing, vol. 29, no. 6, pp. 1153-1160, 1981.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in European conference on computer vision, 2014: Springer, pp. 184-199.

[12] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in European conference on computer vision, 2016: Springer, pp. 391-407.

[13] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874-1883.

[14] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 252-268.

[15] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 624-632.

[16] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 11, pp. 2599-2613, 2018.

[17] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 864-873.

[18] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 10, pp. 3365-3387, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[20] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 1920-1927.

[21] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in Asian conference on computer vision, 2014: Springer, pp. 111-126.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136-144.

[23] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646-1654.

[24] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637-1645.

[25] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3147-3155.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.

[27] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4799-4807.

[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2472-2481.

[29] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286-301.

[30] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681-4690.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in European conference on computer vision, 2016: Springer, pp. 694-711.

[32] R. Timofte, S. Gu, J. Wu, and L. Van Gool, "Ntire 2018 challenge on single image super-resolution: Methods and results," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 852-863.

[33] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," IEEE Transactions on computational imaging, vol. 3, no. 1, pp. 47-57, 2016.

[34] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," Advances in neural information processing systems, vol. 29, 2016.

[35] X. Wang et al., "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0-0.

[36] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 606-615.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556," arXiv preprint arXiv:1409.1556, 2015.

[38] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4491-4500.

[39] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[40] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[41] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in International conference on curves and surfaces, 2010: Springer, pp. 711-730.

[42] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 2001, vol. 2: IEEE, pp. 416-423.

[43] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5197-5206.

[44] E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study (supplementary material)."

[45] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in Proceedings of the 1st international workshop on comics analysis, processing and understanding, 2016, pp. 1-5.

[46] S. Anwar and N. Barnes, "Densely residual laplacian super-resolution," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[47] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," IEEE transactions on image processing, vol. 19, no. 11, pp. 2861-2873, 2010.

[48] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," arXiv preprint arXiv:1807.00734, 2018.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600-612, 2004.

[50] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?," in 2002 IEEE International conference on acoustics, speech, and signal processing, 2002, vol. 4: IEEE, pp. IV-3313-IV-3316.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.

[52] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1808-1817.

[53] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, 2003, vol. 2: Ieee, pp. 1398-1402.

[54] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 185-200.