

From Early Models to Modern Techniques: A Deep Learning Survey on Single Image Super-Resolution

Haorui Li*

Faculty of Arts, McGill University, Montreal, H8N 0H5 Quebec, Canada

Abstract. The primary goal of Single Image Super-Resolution (SISR), a fundamental yet challenging computer vision task with several practical applications in domains such as surveillance, medical imaging, and remote sensing, is to reconstruct a high-resolution (HR) image from a single low-resolution (LR) input. The performance of SISR has been greatly improved by the advent of deep learning, specifically Convolutional Neural Networks (CNNs) and Transformer architectures. An extensive review of deep learning-based SISR techniques is presented in this study. Begin by formulating the SISR problem and discussing prevalent evaluation metrics that balance distortion (e.g., PSNR) and perceptual quality (e.g., SSIM, LPIPS). Subsequently, classifying and analyzing key methodologies across five categories: interpolation-based and traditional models, CNN-based architectures (e.g., SRCNN, VDSR, EDSR), GAN-based frameworks (e.g., SRGAN, ESRGAN, Real-ESRGAN), attention-enhanced networks (e.g., RCAN), and Transformer-based approaches (e.g., SwinIR, HAT). In each category, the theoretical framework, design innovations, and corresponding advantages and limitations are explored. By showing architectural design strategies and training paradigms, this review highlights a structured understanding of the significant evolution from early CNNs to sophisticated GANs and Transformers in SISR, serving as a reference for future model development and practical deployment.

1 Introduction

Single image super-resolution (SISR) is a highly appealing challenge in computer vision. It demonstrates the technique of generating high-resolution (HR) pictures derived from low-resolution (LR) photographs. Since HR images contain richer detail information, which significantly enhances the performance of image analysis, understanding, recognition, and subsequent extension or processing tasks. In today's world, digital images are ubiquitous, and obtaining HR images is critical and crucial for numerous visual applications. Super-resolution (SR), as a classic inverse problem in computer vision, has attracted significant interest owing to its extensive applications in fields such as medical imaging and aerial photography, video surveillance, facial recognition, and mobile photography [1].

* Corresponding author: lihaorui917@gmail.com

However, in daily practical situations, HR images are not always easily obtained, they are often limited by constraints such as the diffraction limit of optical systems, the physical characteristics of sensors, bandwidth limitations, storage costs, and the performance of early imaging devices, etc. These factors frequently result in the acquisition of only LR images, where LR images typically suffer from issues such as detail loss, edge blurring, aliasing effects, and noise—severely limiting their usability and effectiveness in critical applications.

By enhancing visual quality and preserving fine details, SR technology is essential for scenarios that require high-fidelity image analysis or low-bandwidth image acquisition. As a foundational yet highly challenging task within computer vision and image processing, this technology possesses profound theoretical significance alongside extensive practical value.

Single image SR technology is essentially an inverse problem. When the input value is an LR image, theoretically, there can be an infinite number of HR images as output, because the original image can be obtained after the HR images are degraded. This uncertainty is one of the major challenges faced by SR reconstruction. Specifically, the main difficulties include information loss, complex degradation process and trade-off between perceptual quality and fidelity.

Traditional SR approaches are primarily based on interpolation such as bicubic, Lanczos or model-based optimization methods using priors such as edge smoothness or sparsity [2] [3]. Consequently, based on the development of shallow learning such as Neighbor Embedding, Sparse Coding and Anchored Neighborhood Regression significantly improve reconstruction quality by learning the mapping relationship between LR and HR image blocks.

The subject of SR has seen a revolution due to the emergence of deep learning, particularly convolutional neural networks (CNNs). Dong et al. proposed SRCNN [4], the first CNN-based model for SR, demonstrating significant performance improvements over classical methods. Subsequent advancements included more intricate architectures like VDSR, EDSR, and RCAN, Later advancements introduced generative adversarial networks (GANs) to prioritize perceptual realism, as seen in SRGAN which was proposed by Ledig et al. [5], ESRGAN proposed by Wang et al. [6], and Real-ESRGAN proposed by Wang et al. [7]. Attention mechanisms proposed by Zhang et al. [8] and, more recently, SwinIR, a transformer-based model proposed by J. Liang et al. [9], and HAT which was proposed by Chen et al. [10], have further pushed the boundaries of SR performance in both fidelity and visual quality.

This review aims to provide a structured methodological overview of deep learning-based SR techniques. It focuses on representative methods categorized into five groups: interpolation-based and traditional algorithms, CNN-based methods, GAN-based models, attention-enhanced networks, and transformer-based architectures. Each category is analysed in terms of its architectural design, optimization strategy, performance, and limitations.

2 Problem Formulation and Evaluation Metrics

2.1 Problem formulation

The goal of SISR is to use a given LR input to reconstruct an HR image. In the most common setting, SISR can be defined as a learning mapping function $F(\cdot)$ such that:

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

Where θ is the learnable parameters of the model, and I_{LR} stands for the input image with low resolution [4]. The transition from HR to LR is called the degradation, it is usually explained as:

$$I_{LR} = D(I_{HR}) + n \quad (2)$$

Where D is the degradation function including blur, downsampling and compression, and n is the additive noise. [11]. In most situations, the degradation function is assumed to be known and fixed. However, in real-life scenarios, it is often unpredictable and complex, which gives rise to blind or real-world super-resolution tasks.

2.2 Evaluation metrics

The evaluation of SR models relies on distortion-oriented and perception-oriented criteria. Researchers typically use peak signal-to-noise ratio (PSNR) to measure distortion, which is calculated as the logarithm of the ratio between the maximum possible pixel value and the mean square error (MSE) between the reconstructed image and the original image. Blau and Michaeli stated that although PSNR is popular and easy to use, research indicates that it poorly correlates with human visual perception, especially in textured or perceptually realistic reconstructed images [12].

In response to these shortcomings, the Structural Similarity Index Measure (SSIM) was introduced to better reflect human perception. SSIM considers the brightness, contrast, and similarity between the target and the reference image, thereby providing better consistency in visual quality assessment. However, SSIM still struggles to perform effectively when images consist of minor misalignments or require high perceptual realism but lack pixel-level fidelity.

Additionally, in recent years, perceptual metrics based on deep features have gained increasing attention. Zhang et al. introduced Learning Perceptual Image Patch Similarity (LPIPS), which evaluates the activation values of deep network layers for comparison between the output and the input, and it demonstrates excellent correlation with human judgments in image comparison tasks, particularly in GAN-based SR models [13].

Although PSNR and SSIM remain the standard in academic benchmarking, perceptual metrics such as LPIPS are increasingly being used to assess realism and visual quality, especially in tasks where the perceptual credibility of results is more important than pixel-level fidelity [12].

3 3. Methodologies

3.1 Interpolation-based and traditional methods

Before the widespread application of deep learning, SR was mainly addressed using traditional methods based on signal processing and statistical modeling. Among these, the most direct and computationally efficient method is interpolation-based. These methods estimate the values of missing high-resolution pixels by fitting curves or surfaces to the intensities of neighboring pixels. In particular, bicubic interpolation has become a widely accepted benchmark interpolation method in SR research due to its balance between speed and image smoothness [2].

Although interpolation performs well for upscaling smooth areas, it fails to replicate images with high-frequency features and frequently produces outputs that are too smooth or fuzzy, particularly at high magnification settings. Hence, this limitation motivated the

development of reconstruction-based methods. These methods formulate SR as an inverse problem by attempting to reconstruct an HR image by minimizing the loss function, which includes regularization terms such as total variation [3].

3.2 CNN-based methods (SRCNN, VDSR, EDSR)

The rise of deep learning, particularly through convolutional neural networks (CNNs), has indicated a significant change in single-image super-resolution (SISR) research. CNNs no longer need reliance on manually constructed prior knowledge or models but can directly discover the complete mappings between LR and HR data. A pioneering effort in this area is the super-resolution convolutional neural network (SRCNN) introduced by Dong et al., [4]. In CNN-based super-resolution models, the basic concept involves learning a mapping function \mathcal{F}_θ that outputs a HR version by given an LR original input through multiple layers of convolutional filters and non-linear activations

$$\hat{I}_{HR} = \mathcal{F}_\theta(I_{LR}) \quad (3)$$

Where the function is parameterized by θ . It learns the mapping using a straightforward three-layer CNN architecture from LR images upsampled by bicubic interpolation to HR images. Kim et al. expanded on this by proposing the Very Deep Super Resolution (VDSR) network [14]:

$$\hat{I}_{HR} = I_{LR} \uparrow + \mathcal{R}_\theta(I_{LR} \uparrow) \quad (4)$$

In this formula, \mathcal{R}_θ is the residual function and $I_{LR} \uparrow$ denotes bicubic interpolation. This formulation accelerates convergence and allows training of deeper networks. It employs a 20-layer convolutional architecture combined with residual learning, significantly increasing the network depth.

While deeper networks improve SR performance, they often introduce training difficulties and redundancy. To solve this problem, Lim et al. proposed the Enhanced Deep Super-Resolution Network (EDSR) [15].

$$\hat{I}_{HR} = I_{LR} \uparrow + \sum_{i=1}^N ResBlock_i(\cdot) \quad (5)$$

where each $ResBlock_i$ is a residual block without batch normalization. It removes unnecessary modules such as batch normalization and adopts a deep residual block architecture with a larger channel width.

Overall, CNN-based methods represent the first-generation deep learning SISR model. They can be seen as very strong baselines and have motivated the subsequent improvements in the future.

3.3 GAN-based methods

While CNN-based approaches show outstanding results in pixel-level accuracy, they often result in outputs that are overly smooth, failing to capture realistic texture nuances since they rely on the mean squared error (MSE) loss function. This limitation has driven the integration of generative adversarial networks (GANs) into super-resolution models, aiming to enhance perceptual quality by generating images that more closely resemble human visual perception. GAN-based SR models extend CNN formulations by introducing an adversarial learning

framework composed of a discriminator D and a generator G . The discriminator separates created and genuine HR images, while the generator reconstructs the HR image from the LR input.

The Super Resolution GAN (SRGAN), a revolutionary SISR contribution, was proposed by Ledig et al. The generator learns the mapping:

$$\hat{I}_{HR} = G_{\theta}(I_{LR}) \quad (6)$$

Training is driven by an adversarial loss \mathcal{L}_{adv} and a content loss $\mathcal{L}_{content}$. The total generator loss used is

$$\mathcal{L}_G = \sum_{i \in \{content, adv\}} w_i \mathcal{L}_i \quad (7)$$

where $w_{content} = 1$, $w_{adv} = \lambda$, and $\mathcal{L}_{content}$ is often defined using perceptual features from a pre-trained network, λ balances the two terms. Although SRGAN often led to slightly lower PSNR scores compared to traditional CNN-based methods, it produced images with sharper edges and more realistic textures.

To solve artifacts and instability difficulties during training, Wang et al. [6] introduced ESRGAN (Enhanced SRGAN), which builds upon SRGAN. ESRGAN replaces the original residual blocks with residual-overlaid residual dense blocks (RRDB) to enhance feature learning and network stability.

$$\mathcal{L}_{adv}^{rel} = -\mathbb{E}_{I_{HR}} \left[\log \left(D(I_{HR} - \hat{I}_{HR}) \right) \right] - \mathbb{E}_{\hat{I}_{HR}} \left[\log \left(1 - D(\hat{I}_{HR} - I_{HR}) \right) \right] \quad (8)$$

This encourages the generator to produce results that are relatively more realistic than the real ones, improving visual texture.

Despite these advancements, both SRGAN and ESRGAN rely on synthetic LR-HR image pairs generated using bicubic downsampling, which limits their generalization ability on real images. Wang et al. suggested Real-ESRGAN as a more workable substitute to close this gap [7]. Instead of relying solely on bicubic degradation, Real-ESRGAN introduces a generalized degradation model that synthesizes LR images through a combination of multiple degradation steps, defined as:

$$I_{LR} = D_2(K * D_1(I_{HR}) + n) + n' \quad (9)$$

Here, D_1 and D_2 are downsampling operations, blur kernel is denoted by K , while other noise types, such as Gaussian or Poisson, are represented by n, n' . This formula models complex, layered degradations such as camera blur, JPEG compression, and sensor noise. The generator is then trained in a GAN framework to invert this degradation process in a blind setting.

By simulating diverse real-world degradations during training and maintaining the perceptual fidelity of ESRGAN, Real-ESRGAN achieves strong generalization to natural images without requiring access to real HR-LR image pairs. It has become a widely adopted baseline for blind SR tasks in practical applications.

3.4 Attention-based methods

As SR networks grew deeper and more complex, improving the efficiency and selectivity became more and more important. Attention mechanisms, originally derived from natural

language processing, are now gaining attention in computer vision as a dynamic approach to information feature prioritization. Zhang et al.'s residual channel attention network (RCAN) is a significant advancement in this area [8]. The deep residual-in-residual (RIR) architecture is modified to include a new channel attention (CA) module. This CA mechanism models the interdependencies between channels using a “squeeze-and-excitation” structure. For an input feature map $X \in \mathbb{R}^{C \times H \times W}$, the attention weight vector $\alpha \in \mathbb{R}^C$ is computed via:

$$\alpha = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot v)), v = \text{GAP}(X) \quad (10)$$

The sigmoid activation function σ is applied to normalize the response into the range $[0,1]$, yielding the channel-wise attention vector α . W_1 and W_2 are learnable linear layers, and GAP stands for global average pooling. The recalibrated output is $\hat{X}_c = \alpha_c \cdot X_c$ for each channel $c \in [1, C]$.

This adaptive mechanism enhances channel-wise discriminative learning by amplifying informative feature channels and suppressing irrelevant ones.

3.5 Transformer-based methods

With the development of CNN models, they have become dominant in the SISR field due to their powerful local inductive bias and computational efficiency. However, these models are unable to model long-range dependencies due to their restricted receptive field. Therefore, recent research has explored the application of Transformers in SISR.

The SwinIR model (Swin Transformer for Image Restoration), put forth by Liang et al. [9], is a leading effort in this shift. SwinIR applies typical multi-head self-attention (MHSA) within each of the non-overlapping windows created by partitioning the input image. The self-attention is calculated as follows:

$$SA(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d}}QK^\top\right)V \quad (11)$$

The query (Q) is compared with the key (K) through an inner product, scaled by the factor $\frac{1}{\sqrt{d}}$ to stabilize gradients when the dimension d is large. SwinIR adapts the hierarchical Swin Transformer architecture to SR tasks by introducing shifted window-based self-attention. This mechanism allows the model to efficiently capture both local and non-local dependencies without incurring the quadratic complexity of global attention.

Chen et al. developed the Hierarchical Aggregation Transformer (HAT) to build on these advancements [10], which further enhances representation learning for high-fidelity image restoration. HAT introduces a cross-layer aggregation mechanism that adaptively merges outputs from earlier transformer blocks. The hierarchical aggregation at layer l can be formalized as:

$$\hat{X}^{(l)} = \mathcal{A}^{(l)}(\hat{X}^{(1)}, \hat{X}^{(2)}, \hat{X}^{(3)}, \dots, \hat{X}^{(l-1)}) \quad (12)$$

where $\mathcal{A}^{(l)}$ denotes the attention-based aggregation module at layer l . This design enhances feature reuse, enables multi-scale learning, and strengthens both local and global structure recovery in high-fidelity image restoration.

Overall, integrating visual Transformers into SR models (as exemplified by SwinIR and HAT) has significantly improved the oriented benchmarks of distortion and perception. Their capacity to simulate long-range dependencies marks a major advancement over traditional

CNN-based architectures and lays the foundation for the next generation of high-quality image restoration models.

4 Conclusion

This paper provides an overview of deep learning-based methods of SISR, specifically focusing on the architectural evolution and design strategies of recent models. The SR problem is first described, followed by a discussion of popular evaluation measures. Subsequently, SR methods are categorized into five major classes: interpolation-based techniques, CNN-based methods, GAN-based frameworks, attention-enhanced networks, and Transformer-based architectures.

This paper analyses the core modelling principles, structural innovations, and performance trade-offs of each category of methods. CNN-based models like SRCNN, VDSR, and EDSR primarily focused on optimizing pixel-level accuracy, while GAN-based models such as SRGAN and ESRGAN aimed to enhance perceptual realism through adversarial learning. Attention-based methods introduced mechanisms for dynamic feature reweighting, and recent transformer-based models like SwinIR and HAT extended the modelling capacity by capturing long-range dependencies through self-attention.

By describing these methods, highlighting the strengths and limitations of each approach, and detailing how specific design choices impact fidelity and perceived quality, this paper provides a comprehensive reference for understanding the methodological landscape of SR and offers valuable insights for future model development.

References

1. Z. Wang, J. Chen, S.C. Hoi, Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3365–3387 (2020)
2. R. Keys, Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **29**(6), 1153–1160 (2003)
3. J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **19**(11), 2861–2873 (2010)
4. C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 184–199 (2014)
5. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, ... W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4681–4690 (2017)
6. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, ... C. Change Loy, ESRGAN: Enhanced super-resolution generative adversarial networks, in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0 (2018)
7. X. Wang, L. Xie, C. Dong, Y. Shan, Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1905–1914 (2021)
8. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481 (2018)

9. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: Image restoration using Swin Transformer, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1833–1844 (2021)
10. X. Chen, X. Wang, J. Zhou, Y. Qiao, C. Dong, Activating more pixels in image super-resolution transformer, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 22367–22377 (2023)
11. K. Zhang, W. Zuo, L. Zhang, Deep plug-and-play super-resolution for arbitrary blur kernels, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1671–1681 (2019)
12. Y. Blau, T. Michaeli, The perception-distortion tradeoff, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6228–6237 (2018)
13. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 586–595 (2018)
14. J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1646–1654 (2016)
15. B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 136–144 (2017)