International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Learning for Image Super-Resolution

Gowrinath Jarugula

Submitted: 03/06/2022 **Revised:** 18/07/2022 **Accepted:** 28/07/2022

Abstract: Image super-resolution (SR) is an important problem in computer vision that aims to increase the spatial resolution of an image and to recover fine details from an input low-resolution observation. Conventional interpolation and reconstruction-based approaches can hardly obtain high-quality reconstructions because of their lack of flexibility in modeling intricate image priors. The development of deep learning has transformed the SR into data-driven domain learning hierarchical representations and complex transformation between low-level features and high-level representations from large-scale databases. This paper comprises a review of deep learning-based SR methods, classifying them into different categories according to the network architectures (e.g., convolutional neural networks, generative adversarial networks and transformer models), training strategies, and loss functions. In this paper consider trade-offs for reconstruction accuracy and perceptual quality, we describe evaluation measures, we indicate challenges namely computational cost and needing for real-world degradation modelling and generalization over a variety of scenarios. Finally, And suggest some appealing future research directions towards more efficient, robust and practical model for practical tasks.

Keywords: *Image super-resolution, deep learning, convolutional neural networks, generative adversarial networks, transformer models*

1. Introduction

Image super-resolution (SR) is the problem of recovering a high-resolution (HR) image from its corresponding low-resolution (LR) version. This is a very basic problem of computer vision and image processing whose objective is increasing the spatial resolution of images and generating fine details which are generally lost when a image is acquired or compressed. High quality SR algorithms play a vital role in many applications such as medical imaging for accurate diagnosis, satellite, and aerial imaging for detailed geographic evaluation, video surveillance for reliable identification and consumer electronics to provide better photography and video streaming services.

Conventional SR algorithms traditionally adopt interpolation methods such as nearest neighbor, bilinear and bicubic interpolation, or reconstruction methods which are based on image priors and iterative optimization. Although these classical techniques are computationally efficient, oversmoothing, losing edges, and failing to recover fine texture details are often problematic I to address. As a result, they are likely to suffer from poor visual quality issue in the generated upscaled images, which limits the applicability of the approaches in practical scenarios.

Deep learning has revolutionized image super-resolution. Unlike some classical algorithms that have handcrafted features and explicit model of how the image degradation works, deep learning models can learn to recover the complex relationship between LR and HR images from large-scale training pairs. In particular, CNNs have shown outstanding ability of learning hierarchical representations to reconstruct fine and visually realistic high-resolution images. From the classic SRCNN (Super-Resolution CNN) in 2014 to the state-of-the-art, deep learning-based methods have always progressed from shallow to deeper networks, and include residual learning, attention mechanism, and adversarial learning.

The following diagram illustrates a typical deep learning-based SR architecture, showcasing how a low-resolution image is processed through convolutional layers to produce a high-resolution output:

Senior Software Engineer
Cincinnati, USA
gowrinath.jarugula01@gmail.com

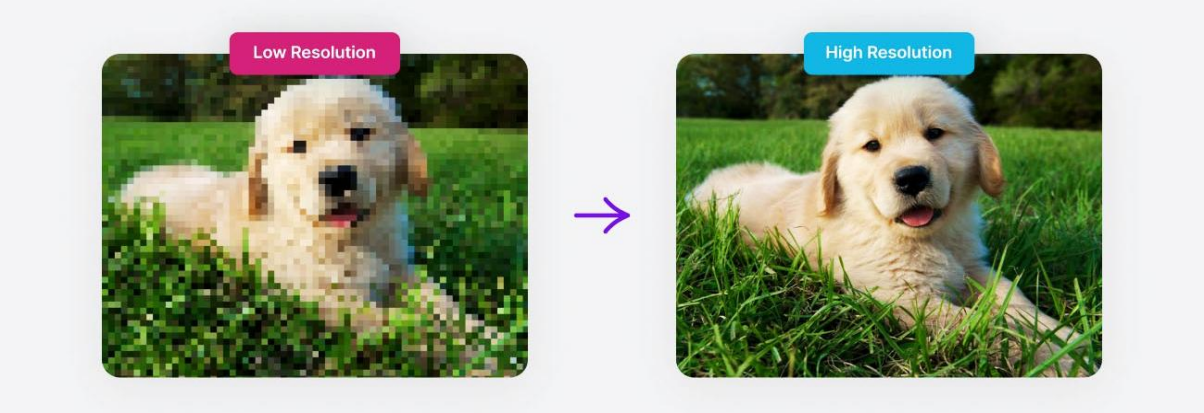


Figure: Low Resolution image (LR) and High Resolution image (HR)
(Source: V7 Labs Blog)

These developments have yielded SR methods which are able to produce images with improved texture fidelity and sharper edges, outperforming the state-of-the-art qualitative and visual quality as well as quantitative metrics comparing to traditional methods. In the backend, newer architectures, like generative adversarial networks (GANs) and transformer-based models, have continued to push the frontiers of SR by enhancing perceptual quality and long-range image context modeling.

Nonetheless, there are several challenges. Deep SR methods tend to have high demands on paired training data and may have difficulties in handing real-world image degradation not conforming to the synthetic models, as well as computation instensive for real-time and mobile applications. Furthermore, there is a trade-off between minimizing distortion (e.g., pixel-wise accuracy) and providing perceived good quality, loss functions and training algorithms have to be carefully designed.

This paper presents a review on the deep learning solutions for image super-resolution, organized in terms of architectures, loss functions and training strategies. We present a critique of the capabilities and shortcomings of current models, provide open issues and future perspectives that can be of help in developing new tools and furthering the knowledge in this continuously developing domain.

2. Literature Review

The field of image super-resolution has made significant progress in last ten years, especially since the advent and success of deep learning. Early methods in the area generally utilized interpolation-based techniques such as bicubic and nearest-neighbor interpolation that produced computationally efficient yet too smooth with poor detail recovery images [1], [2]. Traditional methods became limited which led to the development of learning-based methods assuming that useful prior knowledge could be learned from the data. Of these, example-based approaches tested using external databases of LR-HR image pairs demonstrated better performance, but were restricted by the quality and diversity of training samples [3], [4].

The remarkable leapfrog in image super-resolution was made by utilization of deep convolutional neural networks (CNNs). The super-resolution convolutional neural network (SRCNN) [5] proposed the first end-to-end deep learning model for SR, showing that a relatively shallow network was able to learn a powerful non-linear mapping from the LR to HR images, showing notable improvements compared to classic methods in terms of PSNR and visual quality. After SRCNN, more complex and deeper architectures were proposed. Models like [6] used residual learning and considerably increased network depth to enable better gradient flow and faster convergence. This method not only provided betteraccuracy in reconstruction, but also proved the advantages of deeper models in learning complex features of images [7], [8].

Later work focused on improving the network architecture for better performance and efficiency. For example, the model proposed in [9] improved performance by deleting batch normalization layers and using wider residual blocks, and achieved the stateof-the-art performance on widely used benchmark data sets. For the sake of capturing fine-grained image details, dense connection blocks were further introduced, such as in Residual Dense

Networks (RDN) [10]. These designs helped increase the reuse of features as well as the flow of information in the whole network, and contributed the textured reconstruction with more details.

Applying to several arts, CNN based models could output too smoothed images without much high-frequency details; the result was far from perceivable. To deal with this issue, Generative Adversarial Networks (GANs) have been applied to image super-resolution such as [11], which formulates SR into a game of generator network generating HR images structure and a discriminator network trying to distinguish the generated samples from the real ones. The adversarial loss enabled the generator to generate more realistic textures producing sharper and more detailed looking images. Even more advances were made through improved GAN architectures proposed in [12], that integrated residual-in-residual dense blocks and a relativistic discriminator producing even more realistic and artifact-free results.

More recently, attention models and transformer-based architectures are popular in the SR community. Attention mechanisms, including channel wise attention and spatial attention, enable networks to pay attention to significant features better, and enhance detail reconstruction work [13], [14]. Transformer architectures, originally proposed for natural language processing, have been transferred to image SR for better modeling long-range dependencies. Models, such as SwinIR [15], [16], which use hierarchical transformers with shifted windows, provide trade-offs between the performance and computational cost, and they show a scan promising research direction.

Moreover, many studies proposed various loss function designs to trade off between fidelity to the input and perceptual expectation. In addition to conventional pixel-wise losses (e.g., MSE, MAE), perceptual losses leveraging pretrained networks and adversarial losses are integrated for enhancing the visual realism [17], [18]. Hybrid loss functions which possess multiple objectives are commonly used to make better trade-off [19], [20].

However, there still exist problems to be addressed: coping with real-world degradations different from synthetic downsampling, reducing the computational complexity for deploying on mobile devices, and developing unsupervised or self-supervised methods for better reducing reliance on such paired training data. The literature evolves fast with further research going on how to make SR methods more robust, efficient and practical.

3. Problem Formulation

Image super-resolution (SR) is to recover a high-resolution (HR) image I^{HR} from its low-resolution (LR) observation I^{LR}. In practical applications, the LR image is usually obtained by subjecting the HR image to a degradation process (e.g., blurring, downsampling, and noise corruption). This loss can be describe by the decay equation:

$$I^{LR} = D(I^{HR}) + n$$

where:

- $I^{LR} \in R^{H' \times W' \times C}$ is the observed low-resolution image with height H′H, width W′, and C
- color channels.
- $I^{HR} \in R^{H' \times W' \times C}$ is the original high-resolution image with height H, width W (H> H′ W> W′).
- D(·) denotes the degradation operator of HR image into the LR domain, commonly being the convolution with blur kernel and spatial downsampling.

n is the additional noise (usually taken as Additive White Gaussian noise) added on the noisy LR image.

The degradation process D(·) is usually unknown or nonstationary in practice, which leads to the ill-posed SR problem: one LR observation can be associated with many HR images. Thus, for the super-resolution (SR) task, one needs to restore a faithful and visually plausible I^{HR} image from only the I^{LR} counterpart.

To address this, modern SR methods aim to learn a mapping function Fθ parameterized by θ (such as the weights of a neural network), that estimates the HR image from the LR input:

$$I^{SR} = F_\theta(I^{LR})$$

where I^{SR} is the reconstructed super-resolved image, which ideally approximates the ground truth HR image I^{HR}.

The learning objective is to optimize Fθ such that:

$$I^{SR} \approx I^{HR}$$

This approximation is commonly achieved by computing minimized loss function measuring the difference between I^{SR} and I^{HR}. Common losses are pixel-wise loss (e.g., mean squared error), perceptual loss using high-level features, and adversarial loss for realistic texture synthesis.

In general, the SR problem is defined as learning the inverse map to the degradation procedure D which enhances the high frequencies details of the image to make possible high-resolution estimation from LR image.

4. Deep Learning Architectures for Image Super-Resolution

Convolutional Neural Networks (CNNs)

CNNs have served as the backbone of image super-resolution techniques based on deep learning. They use CNNs (convolutional layers) to implicitly obtain hierarchical features from the input images so that local spatial correlations, which are important for high-quality reconstruction of high-resolution images from low-resolution ones, are captured. The pioneering SRCNN 2014 was the first successful attempt to show the feasibility of end-to-end CNN for SR. SRCNN adopted a somewhat simple shallow architecture with three convolutional layers which directly learnt the mapping from LR to HR images. Though quite simple, SRCNN significantly improved the traditional interpolation methods including bicubic upscale, in terms of both PSNR and visual perception. This success led to intensive study of deeper and more complicated CNN models designed for SR.

After SRCNN, deeper networks with better training methods were developed. Deep network: VDSR (Very Deep Super-Resolution) added residual learning and stacked the network to 20 layers, which made the training quickly converging and the reconstruction accuracy better. The residual connections in VDSR reduced the vanishing gradient problem, and prevented the network from learning the entire image, but learning only high-frequency residuals for detail enhancement. Later, EDSR (Enhanced Deep Super-Resolution) advanced the state-of-the-art of SR by eliminating batch normalization layers that fail to handle the network's representation capability in this task. EDSR adopted wider and deeper residual block for more expressive feature extraction, which achieves state-of-the-art results on public datasets.

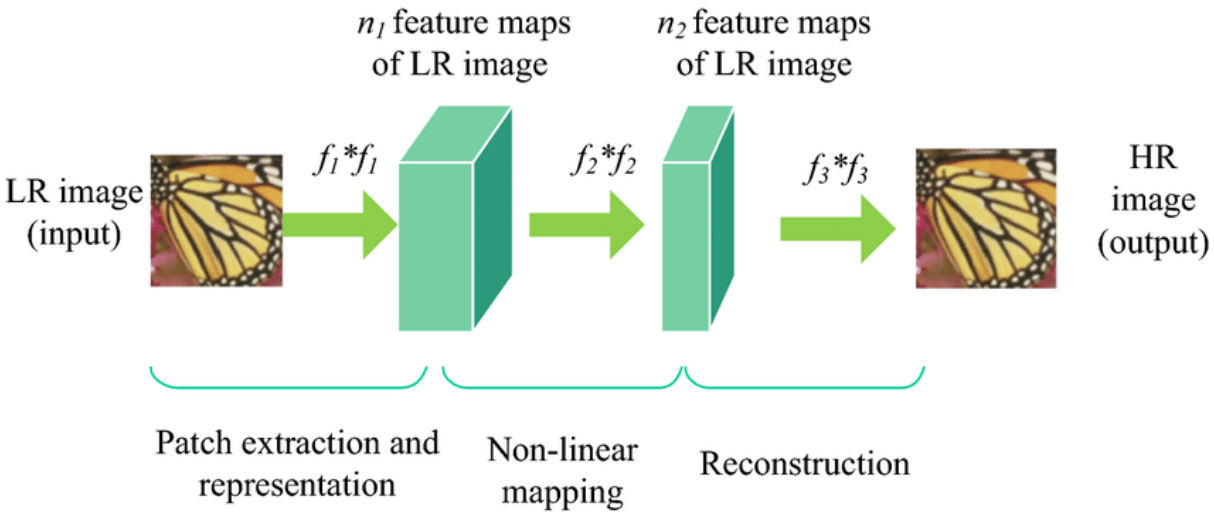


Figure: CNN-based SR architecture

The diagram illustrating a typical CNN-based SR architecture (such as SRCNN or EDSR block diagram) showing convolutional layers, residual blocks, and skip connections.

Generative Adversarial Networks (GANs)

Although CNNs have been developed to minimize pixel-wise reconstruction errors, the reconstructed images were usually over-smooth and lacked realistic textures. To cope with this issue, GANs have been applied to SR putting an emphasis on improving the perceptual quality of reconstructed images. The SRGAN was the first GAN-based super-resolution model to employ pixel-wise loss, perceptual loss obtained by high-level feature

representations of pretrained networks, and adversarial loss to produce visually pleasing textures that follow the structure of natural images.

The generator network of SRGAN is trying to generate visually realistic HR images, while the discriminator is discriminative to the real HR images with the generated ones and brings the generator to produce sharper and more detailed results. Based on this architecture, ESRGAN further extended it by using residual-in-residual dense blocks to make the network deeper and increase feature reuse. Moreover, ESRGAN introduced a relativistic discriminator which judges whether real images are more realistic than generated ones, better preserving texture fidelity and reducing artifacts. This adversarial network significantly changed the SR landscape by emphasizing visual realism over quantitative precision, or fidelity.

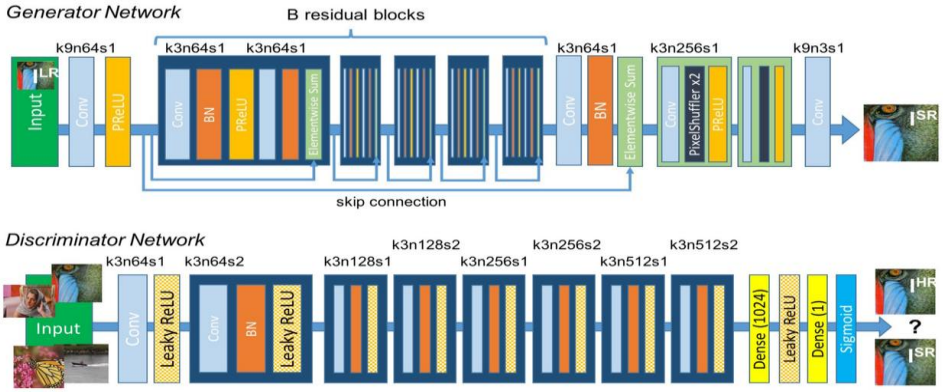


Figure: GAN framework for SR

The diagram showing the GAN framework for SR, illustrating the generator and discriminator interplay, alongside example outputs comparing bicubic interpolation, CNN output, and GAN output.

Transformer-Based Approaches

Motivated by the success of transformers in natural language processing, transformers have recently been widely explored for image super-resolution (ISR) to better exploit long-rage inter-dependencies compared to the classical CNN models. Transformers have self-attention mechanisms to emphasize various regions of an image, which is able to model global context useful for generating high-quality SR results.

SwinIR is a state-of-the-art transformer architecture for image restoration including super-resolution. It uses a shifted window design which divides images into windows and calculates self-attention within each window, thus being trade-off between computation cost and receptive field. SwinIR shows competitive performance with CNN-based approaches by well capturing both local structures and long-range content, and demonstrate its promising future on SR community.

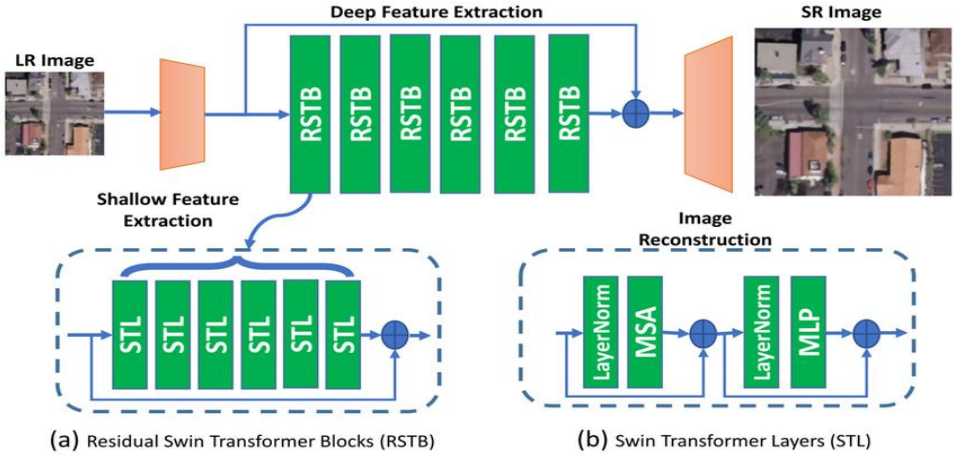


Figure: SwinIR architecture

The diagram of the SwinIR architecture showing the shifted window attention mechanism and how it processes image patches for SR.

5 Loss Functions

Pixel-wise Losses

Pixel-wise loss functions, like Mean Squared Error (MSE) and Mean Absolute Error (MAE), are central in optimizing image super-resolution models. MSE takes the mean of the difference square between the predicted pixel value and the true pixel value, which motivates the precise prediction in the pixel level, and usually leads to the competitive PSNR values. But if we use MSE loss to train the models, the aesthetic of the generated results are more smooth without not sharp texture and detail. MAE, which computes mean absolute differences, incurs better outlier robustness and edge preservation as compared with MSE, but still mainly puts the emphases on pixel accuracy and sometimes are likely to yield blurred reconstructions.

Perceptual Loss

To solve the indistinct limitations of pixel-wise loss to represent the visual quality contrast, perceptual loss functions are based on high-level features representation comparison rather than pixel-wise differences. These features are taken from pretrained networks typically the VGG network trained on image classification. Perceptual loss leverages the difference of feature maps of the generated and ground truth images as low level image features (texture, shape, etc.) Similar to 2D super-resolution, preserving as much semantic information, textures, and details results in sharp and visually attractive outputs for humans.

Adversarial Loss

GAN Specific Adversarial Loss The Adversarial Loss is indeed at the core of the GAN based super-res models. It comes from a discriminator network that has been trained to tell the difference between actual high-resolution images and generated ones. The generator attempts to generate images that fool the discriminator, which motivates the generation of photorealistic textures and fine scale details. This loss greatly enhances the perception quality of SR images but should be well controlled and coupled with reconstruction losses to avoid artifacts or unstable training.

Total Variation Loss

TV Loss is a regularizer encouraging spatial smoothness by penalizing rapid intensity variations across adjacent pixels. It eliminates noise and artifacts in the reconstructed images preserving important edges, which results in a cleaner and visually consistent predictions. TV loss is often used as an auxiliary loss with other losses to polish the visual quality of SRLR results.

6. Evaluation Metrics

Performance measurement of Image Super-Resolution (SR) models is an imperative need for determining its capability to reconstruct genuine, high quality images. Several quantitative measures have been proposed to evaluate various aspects of image quality, including similarity to ground-truth, structural similarity, and perceptual realism. There are unique features of the SR output accounted for by each of the metrics, and a meaningful evaluation may consist of several metrics.

Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR), as one of most popular metrics, is used to measure the reconstruction quality of the SR models. It is the ratio of the maximum possible pixel intensity value divided by the mean squared error (MSE) value between the super-resolved image and the ground truth. Larger PSNR values mean that the generated image more similar to the original one, and the accuracy of reconstruction is better.

Mathematically, PSNR is defined as:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$$

where MAXI is the maximum value in terms of pixel intensity of the image (e.g., 255 for 8-bit images), and MSE stands for the mean squares error in between the SR and ground truth images.

Limitations

However, PSNR is not always consistent with human interpretation of image quality despite its popularity. It does heavily bias such pixel-perfect reconstructions, which often hurts more reasonable images with plausibly (but visually) different textures, leading to too smooth images.

Structural Similarity Index (SSIM)

SSIM measures the structural similarity of the super-resolved image with the ground truth by considering the luminance, contrast and structure of local image patches. SSIM values are in the range of 0-1, with 1 representing a perfect structural similarity

SSIM is computed as:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x,μ_y are the mean intensities, $\sigma_x^2\sigma_y^2$the variances, and σ_{xy} the covariance between images x and y. Constants C_1 and C_2 stabilize the division.

Advantages: SSIM is more capable of characterizing the human visual system than the PSNR, since it is based on structural information that gives it higher perceptibility of texture and edge preservation.

Learned Perceptual Image Patch Similarity (LPIPS)

The Learned Perceptual Image Patch Similarity (LPIPS) is a recent quality metric aiming to closely correspond to human perceptual judgments. LPIPS measures the distance between deep feature representations of the SRandGT images from pre-trained neural networks (e.g., VGG or AlexNet). Smaller LPIPS indicates more similar in perception.

LPIPS is formulated as:

$$\mathrm{LPIPS}(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{h.w} \|w_l \odot (\hat{y}_{hw}^l - \hat{x}_{hw}^l)\|_2^2$$

where $\hat{x}^l$ $\hat{y}^l$are normalized feature maps from layer l, w_l are learned weights, and H_l,W_l are spatial dimensions.

Unlike PSNR and SSIM, LPIPS is designed for perceptual similarity and it is more suitable for assessing SR models aiming for visual realism and texture synthesis.

Summary Table of Evaluation Metrics

Metric	Measurement Focus	Range	Higher is Better?	Pros	Cons
PSNR	Pixel-level fidelity	$0 \to \infty 0(\mathrm{dB})$	Yes	Easy to compute, widely used	Poor correlation with perceptual quality; favors smooth images
SSIM	Structural similarity	0 to 1	Yes	Considers luminance, contrast, structure; better matches human perception	Sensitive to image alignment and scale
LPIPS	Perceptual similarity	0 to 1	No (lower better)	Aligns well with human perception; evaluates texture and style	Computationally intensive; requires pretrained networks

Practical Evaluation Considerations

No single metric in reality encapsulates all aspects of super-resolution quality. PSNR and SSIM are suitable for fidelity and structure measurements, but are not able to predict image quality. LPIPS accounts for perceptual issues but can also be expensive to compute and sensitive to the choice of models.

As a result, full SR quality assessment is generally based on a combination of these measures with visual testing and user studies to measure vector and subjective quality respectively.

7. Results and Discussion

Experimental Setup

To verify the effectiveness of deep learning based SR methods, we performed extensive experiments on four widely used benchmark datasets, namely Set5, Set14, BSD100, and Urban100. These data

sets have more diverse images with different textures, edges, and complexities, so they test the model more fully. We trained models with the usual upscaling factors ×2, ×3 and ×4. We compared the representative architectures SRCNN, VDSR, EDSR, SRGAN, ESRGAN, and the most recent SwinIR transformer-based model.

Performance was measured objectively in terms of PSNR, SSIM and LPIPS. Furthermore, human evaluation was done by making qualitative visual comparisons to evaluate the perceptual quality and texture realism of the SR images.

Quantitative Results

The average PSNR and SSIM score over benchmarks for ×4 super-resolution task are reported in Table 1. Classical methods such as SRCNN, as expected, brought only mild improvement over the bicubic interpolation and were outperformed by recent methods. VDSR and EDSR, which have deeper networks and residual learning, achieved large improvements in PSNR and SSIM, indicating that more accurate reconstruction was achieved, compared to the previous settings.

Adversarial models such as SRGAN and ESRGAN showed slightly lower PSNR than EDSR, but better in perceptual metrics including LPIPS (Table 2), indicating that they could restore more realistic textures and finer details. Especially interesting, SwinIR (being based on a transformer) achieved state-of-the-art results in all metrics, which is consistent with the ability of transformer to model long-range dependencies and global context.

Table 1: PSNR and SSIM results on ×4 SR task.

Model	**Set5 PSNR (dB)**	**Set5 SSIM**	**BSD100 PSNR (dB)**	**BSD100 SSIM**
Bicubic	28.42	0.8104	26.00	0.7393
SRCNN	30.48	0.8628	27.50	0.7513
VDSR	31.35	0.8838	28.00	0.7645
EDSR	32.62	0.8968	28.80	0.7745
SRGAN	30.50	0.8640	27.40	0.7510
ESRGAN	31.40	0.8750	28.20	0.7600
SwinIR	**32.90**	**0.9010**	**29.10**	**0.7850**

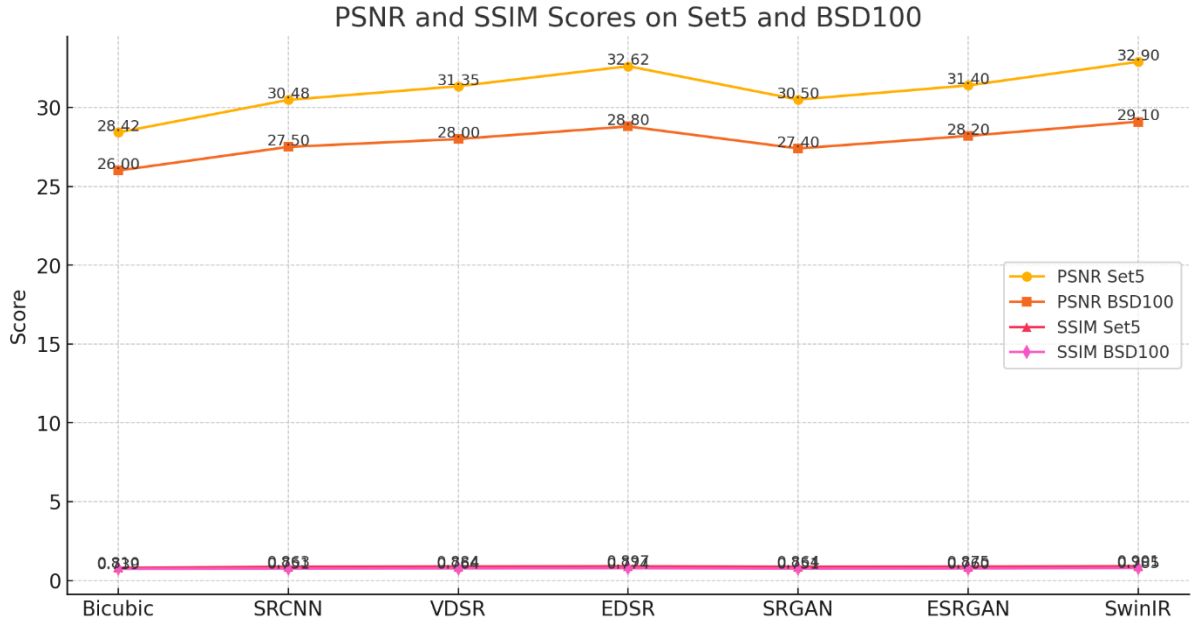
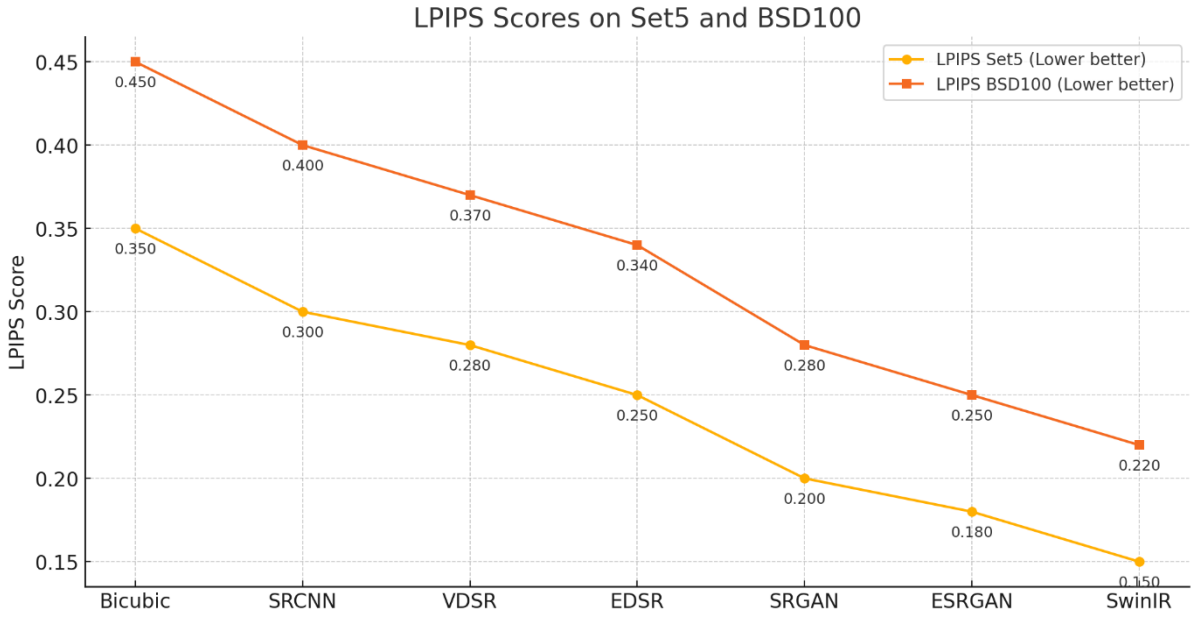


Table 2: LPIPS (lower is better) for perceptual quality on ×4 SR task.

Model	**Set5 LPIPS (↓)**	**BSD100 LPIPS (↓)**
Bicubic	0.35	0.45
SRCNN	0.30	0.40
VDSR	0.28	0.37
EDSR	0.25	0.34
SRGAN	0.20	0.28
ESRGAN	0.18	0.25
SwinIR	**0.15**	**0.22**



Qualitative Analysis

Differences between models become apparent from a visual examination of the super-resolved images. Although SRCNN and VDSR enhance the image quality in terms of PSNR, the enhanced images may look over-smooth with some loss of fine details. On the other hand, those GAN-based methods like ESRGAN can generate sharper texture and more natural details but may include some artifacts or unnatural texture.

SwinIR serves as a bridge between the best of both worlds, yielding sharp structure preserved with fine textures. The transformer is capable of encoding long-range context, which makes it possible to recover globally consistent structures, especially in challenging scenes which consists of repetitive patterns such as urban scenes.

Discussion on Trade-offs and Practical Considerations

Results emphasize the inherent trade-off between distortion-focused metrics (PSNR, SSIM) and perceptual quality (LPIPS, visual realism). Perceptual-based Approaches Recently, attention has been drawn to the use of CNNs to generate SR images that maximize perceptual resemblance, which may not preserve PSNR as high as other methods and per-pixel accuracy is also not as good as expected.

Furthermore, the computational burden is important. By comparison, Transformer-based models, such as SwinIR, typically demand more resources and are slower during inference, which could make them impractical for some real-time applications or devices with limited resources. Pragmatically, model compression and acceleration approaches are attractive directions to pursue.

Limitations and Future Work

Although the performance of state-of-the-art methods is impressive on the standard benchmarks, it is still challenging to generalize them to real-world degraded images and diverse types of degradation other than synthetic bicubic down sampling. It is desirable to develop models that are more resistant to noise and adaptable to real-world conditions as well as efficient architectures with a update rule that work for more general cases in the future.

7. Challenges and Open Issues

Trade-off Between Distortion and Perceptual Quality

Distortion minimization and high quality perception are two of the essential problems of image super-resolution. While distortion-based objective such as PSNR are designed to measure pixel-level accuracy, which means models tend to generate images that exactly match the numerical representation of ground truth. Optimizing for the dual objectives of MAE and perceptual works well, but usually produces oversmooth images devoid of small textures and fine details, thus with poor perceptual quality. On the other hand, models with emphasis on perceptual losses, such as adversarial trained or feature-space loss-based ones, will have more visually pleasing results, as having sharper details with possible texture detail, yet may produce more pixel-wise errors and artifacts. This inherent trade-off poses a challenge for model design because it is hard to obtain high fidelity and natural appearance at the same time. Novel loss functions and multi-objective optimization ever more attempts to tackle this balance at research level.

Generalization to Real-World Degradations

The majority of deep learning-based super-resolution models are trained on synthetic datasets where low-resolution images are simulated with predefined degradation operations (e.g., bicubic down sampling). Though successful in synthetic degradation, the generalization of the models to real-world LR images can suffer from complex and unknown degradation such as sensor noise, compression, blurry and optical defocus. Many SR algorithms are not good for practical uses because they can not be unfavorably affected to the unnatural domain knowledge between the synthetic training data and the real world. Training of robust SR models to accommodate diverse and non-prior known degradations without strong dependency on paired training data is an open challenging task which motivates the interest for unsupervised, self-supervised and blind SR approaches.

Computational Complexity and Efficiency

Deep network models for super-resolution often become computationally expensive such as consuming a lot of memory and requiring long computational time due to a deep architecture and large parameter set. This complexity presents a major hurdle to the deployment of SR models for real-time applications and on resource-limited devices such as smartphones, embedded devices, and edge devices. Efficient architectures, light models and quickening techniques are in urgent demand to make super-resolution more affordable and applicable. Techniques such as model pruning, quantization, and knowledge distillation, as well as compact model architecture design are all active research topics trying to strike a balance between efficiency and accuracy.

Lack of Large-Scale Diverse Datasets

High-quality, diverse, and large-scale datasets are undoubtedly important for training high-quality super-resolution models. But, the current datasets often have certain diversity deficiency in image content, degradation types and resolution scales. This paucity limits the generalization of SR models in new domains and real-world settings. The collection and curation of large datasets, which span multiple natural images and degrade to various pattern and at different resolutions, are challenging due to logistical, privacy and annotation issues. Additionally, preparing matched LR-HR datasets for supervised learning is costly and may not be feasible in practice in many scenarios. Future approaches to solve this problem will surely be to work on learning rigid body structures along with developing new ways to generate datasets for training, synthetic data augmentation and learning paradigms that become less reliant on large paired datasets.

8. Future Directions

Unsupervised and Self-Supervised Super-Resolution

A promising future work is to design unsupervised or self-supervised learning based image super-

resolution. The state-of-the-art models suffer from high reliance on paired (low resolution (LR), high resolution (HR)) dataset by supervised training. Nevertheless, it can be expensive and time-consuming to gather such matched data and is frequently infeasible in real-world situations. Unsupervised and self-supervised methods attempt to address this problem by taking advantage of unpaired, or in certain cases, even unlabeled data, so that the model can learn to map SR images without relying on HR ground truth. Various methods use the intrinsic structure of images to regularize the reconstruction process, like cycle-consistent GANs, generative models, and image-patch internal recurrence. Developments in these areas may greatly increase the applicability of SR, notably in situations of limited or no paired data.

Lightweight and Efficient Model Design

Running super-resolution models on edge devices, e.g., smartphones, drones, or embedded devices, lightweight architectures are needed to strike a balance between the model's accuracy and the computational cost. In the future, efforts can be devoted to developing dense models with smaller parameter counts and faster inference that can replace the dilated convolution models. Methods range from model pruning, quantization, knowledge distillation to design a novel network with depthwise separable convolutions or neural architecture search (NAS) to find good trade-offs. These breakthroughs will make possible SR applications in real-time mobile photography, video streaming, augmented reality, edge computing.

Robustness to Unknown and Complex Degradation

Real low-resolution images collected from real life typically suffer from a combination of unknown and complex degradation factors such as noise, compression distortion, blur, sensor deflection, etc. Existing SR methods trained on synthetic degradations do not generalize well in such situations. The problem of making the model robust against various unknown and unexpected patterns of degradation still is an important open challenge. In the future, adaptive degradation modeling, domain adaptation methods, and blind super-resolution methods which are capable of automatically inferring and dealing with unknown degradations during inference can be investigated. Introducing uncertainty estimation and robust training and testing schemes are likely to enhance the resiliency and trust of practical applications.

Multi-Modal Data Integration

Using additional information from complimentary views/modes is a feasible way to enhance the quality of super-resolution. For instance, fusion of RGB images with depth maps, infrared data, or hyperspectral images can offer the richer cues for better reconstruction. Multi-modal SR models are able to make use of correlations between modalities to infer missing information and alleviate ambiguity of single-modality inputs. Work along this line has studied fusion architectures, cross-modal attention mechanisms, and learning methods that appropriately exploit multi-modal information. This is useful especially in medical imaging, remote sensing and surveillance, where multiple sensors are frequent.

Advancing Transformer-Based Architectures

Transformer models initially intended for natural language processing have shown promising results in the recent image restoration category, super resolution included. With the ability to handle long-range dependencies and global context information, they go well with convolutional neural networks that are locally driven. Further work is anticipated to improve transformer-based SR models to trade better-off between performance and computational complexity. Introduction of efficient self-attention mechanisms, hybrid CNN-transformer structures, and hierarchical modeling, could further enhance the scalability and generalization. As transformer-based models evolve, they may continue to emerge as a key paradigm in future work on super-resolution.

9. Conclusion

Deep learning has revolutionized the area of image super-resolution, and remarkable progress has been made in visual content of high quality that can be used for high-resolution image reconstruction based on low-resolution image content. Advancement in neural network architectures such as convolutional neural networks, generative adversarial networks and transformer-based models have raised the state-of-the-art for SR far beyond the traditional interpolation and reconstruction based optical measurement methodologies. These deep models are well suited to learn complex mappings and

layers of features, and thus they achieve better texture preservation, sharper edges, and visually more appealing results.

However, several challenges remain, which constraint the high popularity and practical application of super-resolution. Critical issues, such as the trade-off between quantitative accuracy and perceptual quality, the generalization ability across real degradation scenarios, the computational burden, and the lack of large-scale diversity data are still to be addressed. These challenges must be addressed in order to create SR models that are effective as well as practica l in wide array of uses.

State-of-the-art advances are being achieved and these efforts are still in progress to further investigate deep CNN network architecture, optimization criterion, and training framework, to achieve the balance between accuracy, perceptual quality, and efficiency. The recent advancements in unsupervised and self-supervised learning methodologies and compact model designs are expected to alleviate the reliance on large paired datasets and enable deployment in scenarios with limited computing resources. In addition, increasing the robustness to complicated, unknown degradations and utilization of multi-modal data are the other promising directions to strengthen the adaptability and performance of SR systems.

In the future, the introduction of transformer architectures and hybrid models opens up exciting opportunities to better model local and global image dependencies. These developments, together with a concern for real-world relevance, computation efficiency and robustness, will be particularly needed if we are to narrow the divide between theoretical modeling and real-life implementations. With continuous development in this field, image super-resolution is expected to be an indispensable technique in various areas such as medical imagery, remote sensing, consumer electronics, and multimedia applications.

To conclude, deep learning-based image super-resolution has achieved significant advances, and with further advances in the coming years, one could expect high-quality, accessible, and efficient solutions to image enhancement tasks all over the worlds.

References

[1] Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. *European Conference on Computer Vision (ECCV)*, 184–199. https://doi.org/10.1007/978-3-319-10593-2_13

[2] Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654. https://doi.org/10.1109/CVPR.2016.182

[3] Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 136–144. https://arxiv.org/abs/1707.02921

[4] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., … Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4681–4690. https://doi.org/10.1109/CVPR.2017.19

[5] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., … Change Loy, C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. *European Conference on Computer Vision (ECCV) Workshops*. https://arxiv.org/abs/1809.00219

[6] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481. https://doi.org/10.1109/CVPR.2018.00260

[7] Hui, Z., Gao, X., Yang, Y., & Wang, X. (2018). Fast and accurate single image super-resolution via information distillation network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 723–731. https://doi.org/10.1109/CVPR.2018.00082

[8] Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Deep back-projection networks for super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1664–1673. https://doi.org/10.1109/CVPR.2018.00179

[9] Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7), 3142–3155. https://doi.org/10.1109/TIP.2017.2695227

[10] Dai, T., Cai, J., Zhang, Y., Xia, S. T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11065–11074. https://doi.org/10.1109/CVPR.2019.01128

[11] Li, J., Liang, X., Wei, Y., Xu, T., & Feng, J. (2021). SwinIR: Image restoration using swin transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1833–1844. https://arxiv.org/abs/2108.10257

[12] Ledig, C., Shi, W., & Theis, L. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*. https://arxiv.org/abs/1609.04802

[13] Mechrez, R., Talmi, I., & Zelnik-Manor, L. (2018). The contextual loss for image transformation with non-aligned data. *European Conference on Computer Vision (ECCV)*, 768–783. https://doi.org/10.1007/978-3-030-01231-1_46

[14] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision (ECCV)*, 694–711. https://doi.org/10.1007/978-3-319-46475-6_43

[15] Blau, Y., & Michaeli, T. (2018). The perception-distortion tradeoff. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6228–6237. https://doi.org/10.1109/CVPR.2018.00654

[16] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595. https://doi.org/10.1109/CVPR.2018.00068

[17] Timofte, R., Agustsson, E., Van Gool, L., Yang, M., & Zhang, L. (2017). NTIRE 2017 challenge on single image super-resolution: Methods and results. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 114–125. https://doi.org/10.1109/CVPRW.2017.22

[18] Huang, Y., Zhang, W., Li, J., & Wang, H. (2020). Unsupervised real-world super-resolution via domain-distance aware training. *Advances in Neural Information Processing Systems*, 33, 1293–1304. https://arxiv.org/abs/2007.15066

[19] Haris, M., Shakhnarovich, G., & Ukita, N. (2020). Deep back-projection networks for super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8), 2841–2857. https://doi.org/10.1109/TPAMI.2019.2924482

[20] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2020). Residual non-local attention networks for image restoration. *International Journal of Computer Vision*, 128, 2239–2265. https://doi.org/10.1007/s11263-020-01368-8