Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29437

Single Image Super-Resolution with Deep Learning

Yuan Li

Chongqing University, Chongqing, China
liyuan0490@gmail.com

Abstract. Single Image Super-Resolution (SISR) aims to reconstruct high-resolution images from low-resolution images. With the breakthrough of deep learning technology, especially the introduction of the convolutional neural Network (CNN), significant progress has been made in the field of single-image super-resolution. Although single-image super-resolution technology has developed rapidly, there are still several limitations. First, high-performance models are often computationally complex and difficult to apply in real-time. Second, most methods rely on the idealized degradation assumption, which is difficult to adapt to the complex noise and blur of real scenes. Third, the commonly used evaluation metrics are not completely consistent with human eye perception. Fourth, the generalization ability of the model is limited, and the cross-domain application performance is insufficient. This paper systematically reviews the evolution of this technology since the application of deep learning, focusing on five types of representative models. By comparing and analyzing the architecture innovation and performance of various models, this paper summarizes their respective advantages and limitations, and looks forward to the future development direction of higher accuracy, stronger generalization, and better perceptual quality.

Keywords: Super Score, Deep Learning, Evolution, Transformer

1. Introduction

Single-image Super-resolution (SISR) refers to the technology of reconstructing high-resolution images from low-resolution images, which is of great value in remote sensing imaging, medical imaging, and other fields. However, traditional interpolation and sparse coding methods are difficult to effectively recover image details and textures. SISR has ushered in a breakthrough, with the SRCNN proposed by Dong et al introducing a convolutional neural network into SISR for the first time, which improves the reconstruction quality and far exceeds the traditional methods in various indicators [1]. Due to the limited depth and small receptive field of the network, it is unable to reconstruct complex textures. Different magnifications need to be trained separately, which is insufficient in adaptability. Subsequently, the reconstruction performance of SISR is further improved by introducing deeper convolutional networks and residual structures [2,3]. In recent years, the introduction of new architectures such as attention mechanisms and the Transformer has enhanced the representation ability and generalization performance of the model. To this end, this paper systematically reviews the research progress in this field, focusing on the design characteristics and performance of representative models, analyzing their advantages and

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29437

limitations, and looking forward to future trends, so as to provide a reference for engineering practice and follow-up research.

2. Convolutional neural network and its representative model (SRCNN)

Convolutional neural networks are the beginning of deep learning for image super-resolution. The representative model, Super-Resolution Convolutional Neural Network (SRCNN), uses the most basic three-layer convolutional structure. The first layer extracts feature blocks from the interpolated magnified low-resolution image, and the second layer maps the features to the high-dimensional space. The third layer is responsible for reconstructing and generating high-resolution images.The innovation of SRCNN is to directly learn the end-to-end nonlinear mapping without relying on manually designed features. Experiments show that compared with traditional bicubic interpolation, neighborhood embedding, and sparse coding methods, SRCNN achieves higher PSNR and SSIM indicators on standard datasets. For example, for magnification ×3, the PSNR of SRCNN on the Set5 dataset reaches 32.75 dB, which is significantly better than that of bicubic interpolation (29.69 dB). The model parameters of SRCNN are only about 57K, which is a very small computational overhead at that time [1].

Although SRCNN is a pioneer, its limitations are also obvious. The shallow network leads to a small receptive field, and it is difficult to reconstruct complex textures. The model needs to be trained separately for each scale, which is inflexible. For the enhancement of the model's expressive capability, improved models such as Fast Super-Resolution Convolutional Neural Network (FSRCNN) have been proposed. By extracting features in the low-resolution space and using deconvolution layers to amplify them. The inference speed is accelerated [4]. In addition, deep recursive models such as Deeply-Recursive Convolutional Network (DRCN) increase the effective network depth to more than 16 layers by means of parameter sharing, which alleviates the parameter explosion problem caused by directly deepening the network [5]. In general, the models at the CNN stage lay a foundation for subsequent, more complex structures. They prove that deep networks can learn high-frequency details required for image super-resolution, but also expose problems such as difficulties in deep training and insufficient capture of long-term dependencies. This prompts research to enter the residual network stage to seek performance breakthroughs at a deeper level.

3. Residual network and its representative model (EDSR)

The introduction of a residual network solves the difficult problem of deep network training. In the field of Super-Resolution, Kim et al. first applied the residual idea to the Very Deep Super-Resolution (VDSR) model [2]. After VDSR, researchers began to explicitly introduce residual blocks into the network structure to build deeper models. Lim et al. 's Enhanced Deep Residual Network (EDSR) is a typical representative of a deep residual network. EDSR adapts the SRResNet architecture, a residual network-based super-resolution model. The key enhancements involve deepening the network and removing the batch normalization layers from the residual blocks. The batch normalization layer will destroy the contrast distribution of the image in the image super-resolution, which is not conducive to the recovery of high-frequency details. Therefore, EDSR removes batch normalization to preserve color and contrast information. EDSR is constructed by stacking a large number of residual blocks, each of which contains two ReLU convolutional layers and the identity hop connection, followed by a high-resolution reconstruction module at the end. In order to stabilize the training, EDSR multiplied each residual branch output by a scaling factor to reduce its amplitude and used a larger receptive field convolution kernel. With these designs, EDSR

DOI: 10.54254/2755-2721/2026.TJ29437

refreshes the reconstruction accuracy of all the baselines at that time after training on the DIV2K dataset. For example, for ×4 superresolution, EDSR achieves a PSNR of 32.46 dB on Set5 and about 26.64 dB on Urban100, both of which are better than previous models. The number of parameters of EDSR is about 43 million, which belongs to the larger model at that time [3]. Nevertheless, EDSR demonstrated that reconstruction performance can be substantially improved by deep residual networks, providing a strong baseline for subsequent studies.

It is worth mentioning that some studies have optimized the efficiency of residual networks. For example, Blueprint Separable Residual Network (BSRN) introduces efficient convolution operators (such as BSConv) and channel attention module. It achieves the most advanced performance with far fewer parameters than large models such as Residual Channel Attention Network[6]. In addition, MemNet, Deep Recursive Residual Network (DRRN), and other models realize parameter sharing through the recursive residual module, which not only maintains the depth but also reduces the number of parameters. In general, the residual network alleviates the difficulties of deep training through skip connections, so that the network can dig more deeply into image detail features. On this basis, further improvement direction is to explore denser connection methods and attention mechanisms, so as to improve performance while controlling model complexity.

4. Complex scene robust network and its representative model (CASGCN)

Super-resolution based on deep learning can achieve remarkable performance, but in complex scenes, the robustness of image super-resolution methods is often poor. In practical applications, super-resolution algorithms often face complex and variable image scenes, such as complex textures, repetitive patterns, and noise interference. This puts forward higher requirements for the global modeling ability and robustness of the model. In order to improve the reconstruction quality in complex scenes, researchers have introduced new technologies such as graph neural networks and grouped convolution, and developed a number of network structures that are robust to complex scenes. The Channel Attention and Spatial Graph Convolutional Network (CASGCN) proposed by Yang and Qi et al. [7] is a representative model in this direction.

The core idea of CASGCN is to integrate graph convolutional networks into regular convolutional networks to explicitly model the global relationship between image pixels. The basic unit of CASGCN is called the CASG module, which contains a parallel channel attention branch and a spatial convolution integration branch. In the Channel branch, a channel Attention mechanism similar to the Residual Channel Attention Network is introduced to weight the feature channels, selectively highlight useful features, and suppress irrelevant information. In the spatial branch, the Graph Convolutional Network (GCN) is introduced. Firstly, the Gram matrix is used to dynamically generate the adjacency matrix of the graph, so that the graph convolution can break through the limitations of the convolution kernel and establish the connection between pixels in the whole graph. Then, the graph convolution operation is performed on the image features to capture the self-similar global correlation between any two pixels in the image. Due to the dynamic adjacency matrix, CASGCN can capture the long-distance dependence of images without additional learning parameters.

Multiple CASG modules are stacked in series to form the main body of CASGCN. The output features of each module are enriched in both channel and spatial dimensions, thereby empowering each pixel to integrate globally relevant information. Experimental results show that CASGCN achieves excellent performance under various degradation models and is superior to the traditional Residual Channel Attention Network in terms of quantitative indicators and visual quality [8]. Especially on images containing complex textures and repetitive patterns, CASGCN can reconstruct

DOI: 10.54254/2755-2721/2026.TJ29437

more coherent and natural detailed textures by means of global graph convolution. Yang et al. report that on typical super-resolution benchmark datasets, the peak signal-to-noise ratio and structural similarity of CASGCN are significantly improved compared with the Residual Channel Attention Network and other models, and it is more robust to different types of degradation (e.g., blur and noise). This series of work shows that the combination of a graph neural network and an attention mechanism can expand the receptive field of a convolutional neural network and effectively model the long-range correlation inside the image, which is an effective way to improve the super-resolution performance of complex scenes.

It is worth mentioning that complex scene robust networks are not limited to one implementation of CASGCN. Other models, such as SRGAT (Introducing Graph Attention) and DIPNet (Two-branch Attention fusing complex background), also strive to enhance the adaptation ability of the network to complex content. However, in general, CASGCN has achieved highly representative results in complex and variable image super-resolution tasks through the innovative design of channel attention + spatial graph convolution, which is regarded as one of the milestones of robust super-resolution networks for complex scenes.

5. Attention mechanism network and its representative model (RCAN)

Deep convolutional neural network (CNN) has a significant effect in extracting high-frequency details. However, with the increase in the number of network layers, how to make the model focus on more meaningful features has become a new topic. The attention mechanism emphasizes key information and suppresses irrelevant information by allocating weights, which provides a new idea for further improving the performance of super-resolution. In single-image super-resolution, the classical model that introduces Attention is the Residual Channel Attention Network (RCAN).

RCAN was proposed by Zhang et al., and its basic unit is the Residual Channel Attention Block (RCAB). RCAB firstly extracts local residual features through several convolutions, and then models the correlation between channels through a series of global average pooling and two fully connected layers to generate an attention weight vector with the same length as the number of channels. These weights are used to recalibrate each channel of the residual feature, so that the discriminative channel features are amplified and the redundant features are suppressed. The weighted features are then added to the module input (residual connection) to obtain the output of RCAB. In this way, RCAN incorporates channel attention on the basis of the deep residual network, which significantly improves the reconstruction accuracy.

RCAN consists of multiple RCABs in series to form a deep network, and uses long and short hop connections after each module to maintain stability. The number of parameters is about 16M, which is comparable to the Enhanced Deep Residual Network, but RCAN achieves higher accuracy due to the addition of attention. For example, on ×4 superresolution, RCAN achieves peak signal-to-noise ratios of 32.63 dB and 28.87 dB on Set5 and Set14 datasets, respectively. Compared with the Enhanced Deep Residual Network (EDN) of 32.46 dB and 28.80 dB, it is slightly improved. On Urban100 with complex texture, RCAN is also slightly better than the Residual Dense Network [8]. More importantly, RCAN restores image details and contrast better in subjective vision. In tasks such as remote sensing image super-classification, RCAN can reconstruct details such as building edges more accurately with less color distortion.

In general, the attention mechanism network effectively improves the reconstruction effect of the deep super-resolution model by explicitly modeling the importance of features. As a milestone of this kind of method, RCAN enables the super-resolution model to have the ability of "focusing" for the first time, which improves the objective indicators and visual quality. With the development of

DOI: 10.54254/2755-2721/2026.TJ29437

the attention module, it is also integrated with residual, dense connections, and other technologies, and a stronger and more efficient, and compact network structure is continuously born.

6. Transformer hybrid network and its representative model (ACT)

In recent years, the advent of Vision Transformers (Vits) has provided a new paradigm for low-level vision tasks. Through the self-attention mechanism, Transformers can capture the correlation between locations in an image without distance restrictions, which is particularly beneficial for modeling global dependencies. Pure Transformer architectures (e.g., IPT, SwinIR) have shown strong performance in super-resolution, even surpassing traditional convolutional neural network attention models. However, Transformers also have problems, such as a lack of prior knowledge of local details and high computational overhead. To this end, some studies have begun to explore the hybrid architecture of a convolutional neural network and Transformer fusion, hoping to combine the advantages of both. Yoo et al. 's Aggregation Connection Transformer (ACT) is one of them [9].

ACT adopts a two-branch hybrid architecture, including a parallel convolutional branch and a Transformer branch, which interact with each other several times during the reconstruction process to learn from each other's strengths. Specifically, the convolutional branch consists of several residual convolutional blocks, which are good at extracting local structure and edge information. The Transformer branch stacks multiple layers of self-attention Transformer blocks, which are good at capturing long-range dependencies and repetitive textures. Between these two trunks, the authors design a feature Fusion Block, which concatenates and fuses the features extracted by the convolutional branch and the Transformer branch at each stage, and transmits information to each other through operations such as 1×1convolution, so that the convolutional features and self-attention features complement each other. The structure of this series-parallel hybrid connection allows the network to pay attention to both local sharp details and global consistency.

In addition, ACT introduces a Cross-Scale Token Attention (CSTA) module inside the Transformer branch. CSTA divides the feature Token into two in the channel dimension, one part keeps the original scale, and the other part undergoes re-tokenization to form a larger scale token. Then, the cross-self-attention calculation is performed on the tokens of different scales, so that the Transformer can efficiently model the association between patches of different scales. Through the CSTA module, ACT overcomes the limitation that the standard ViT can only pay attention to a fixed scale, and incorporates multi-scale information without significantly increasing the amount of computation.

In summary, the Transformer hybrid network represents the latest trend in the evolution of image super-resolution models. By fusing the local inductive bias of convolutional neural networks and the global modeling ability of Transformers, models such as ACT achieve a balance between performance and efficiency, which provides a new paradigm for the design of future super-resolution networks.

7. Performance comparison and visual analysis

The performance of the above representative models on standard benchmark datasets is shown in Table 1. In this paper, the commonly used Set5, Set14, and Urban100 datasets are selected, and the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) indicators of different models are compared by taking super-resolution reconstruction with a magnification factor of 4× as an example. At the same time, the number of model parameters is given to evaluate their complexity.

DOI: 10.54254/2755-2721/2026.TJ29437

Table 1. Comparative performance of different models on the ×4 super-resolution task

Model	Number of parameters	Set5 PSNR/SSIM	Set14 PSNR/SSIM	Urban100 PSNR/SSIM
Bicubic interpolation	–	28.43/0.810	26.00/0.702	23.14/0.657
SRCNN[1]	57K	30.48/0.862	27.50/0.751	24.52/0.722
VDSR[2]	665K	31.35/0.884	28.01/0.767	25.18/0.752
EDSR[3]	43M	32.46/0.897	28.80/0.788	26.64/0.803
RCAN[8]	16M	32.63/0.898	28.87/0.788	26.82/0.808
CASGCN[7]	15M	32.60/0.897	28.88/0.789	26.79/0.809
ACT[9]	11.8 M	33.10/0.907	29.20/0.798	27.72/0.830

As can be seen from the table, the performance of the deep learning model outperforms the traditional bicubic interpolation method by a large amount. For example, SRCNN outperforms Bicubic by about 2.0 dB in PSNR on Set5. With the evolution of the network, the performance of the model continues to improve. VDSR achieves higher PSNR than SRCNN through deep convolution and residual learning. EDSR further pushes the PSNR of Set5 to more than 32.5 dB by using a deep residual network. The recent Transformer fusion model ACT leads the previous generation CNN model in all aspects, with ACT achieving the highest 33.1dB (Set5) and 27.7dB (Urban100). It shows the advantage of Transformer self-attention for recovering complex textures. It is worth noting that models such as ACT control the number of parameters at about 12 million through a hybrid architecture while improving the effect, which is much lower than the early EDSR. This reflects that in recent years, model design has paid more attention to efficiency and practical application requirements while pursuing performance.

Based on the quantitative and qualitative results, it can be seen that the evolution of network structure is very important to improve the performance of super-resolution. Simple convolutional networks have been replaced by deeper residual/dense networks, which in turn have been surpassed by new models combining attention and transformers. This trend shows that the ability of global information extraction and critical detail attention are the keys to high-performance super-resolution models. At present, the most advanced ACT model combines CNN and Transformer, which has the ability of both local detail recovery and global dependency modeling, and thus performs outstandingly in various aspects. However, it should also be noted that different models have their own advantages and disadvantages. For example, RCAN performs well in color fidelity, and the Transformer model has obvious advantages in dealing with repetitive textures but has relatively large computational complexity. In practical applications, the accuracy, efficiency, and complexity of the model should be weighed according to the requirements of the scene.

8. Conclusion

SRCNN introduces the end-to-end convolutional neural network into the field of single-image super-resolution for the first time. Its model structure is lightweight and runs fast. The deep network based on residual learning (such as VDSR/EDSR) effectively alleviates the training difficulty and greatly improves the reconstruction accuracy. Among them, EDSR refreshes various indicators by removing the batch normalization layer and expanding the network scale, but the corresponding amount of parameters and calculation is also significantly increased. Densely connected networks such as RDN use multi-layer convolutional features to achieve hierarchical information fusion, which further improves the reconstruction effect. The attention mechanism significantly enhances

DOI: 10.54254/2755-2721/2026.TJ29437

the feature representation ability: RCAN adopts the residual-channel attention structure to adaptively highlight key information. CASGCN incorporates spatial graph convolution to explicitly model long-range dependencies between pixels, which makes it more robust in complex texture recovery. The latest CNN-Transformer hybrid models (such as ACT) balance local details and global consistency and achieve leading performance on multiple benchmark datasets, but their self-attention branch usually brings higher computation and memory overhead. A promising direction for future SISR research is the development of diffusion-based models, such as SinSR, which introduces a one-step diffusion process, significantly improving inference speed while maintaining or surpassing state-of-the-art performance. SinSR has demonstrated remarkable results on both synthetic and real-world datasets, and its potential for real-time applications makes it an exciting area for future exploration in the field of super-resolution .

Looking into the future, single-image super-resolution models will continue to improve in terms of network structure design, generalization ability, adaptive degradation modeling, and perceptual quality optimization. On the one hand, more efficient hybrid architectures and adaptive normalization schemes will be explored to reduce the computational cost while maintaining or improving the performance. On the other hand, it is necessary to enhance the adaptability of the model to unknown degradation in real scenes, such as domain adaptation, unsupervised learning, or test-time adaptation, to build robust degradation estimation and inverse mapping. Finally, in the aspect of perceptual quality optimization, more advanced perceptual loss and no-reference image quality assessment metrics should be introduced to balance the traditional objective indicators and subjective visual effects. In general, the design of future super-resolution models will pay more attention to balance accuracy, efficiency, generalization, and perceptual quality, and provide more comprehensive and reliable support for high-definition imaging.

References

[1] Dong, C., Loy, C.C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In ECCV (pp. 184–199).
[2] Kim, J., Lee, J.K., & Lee, K.M. (2016). Accurate image super-resolution using very deep convolutional networks. In CVPR (pp. 1646–1654).
[3] Lim, B., Son, S., Kim, H., Nah, S., & Lee, K.M. (2017). Enhanced deep residual networks for single image super-resolution. In CVPR Workshops (pp. 136–144).
[4] Dong, C., Loy, C.C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In ECCV (pp. 391–407).
[5] Kim, J., Lee, J.K., & Lee, K.M. (2016). Deeply-recursive convolutional network for image super-resolution. In CVPR (pp. 1637–1645).
[6] Li, Y., Tang, J., & Wu, F. (2022). Blueprint separable residual network for efficient image super-resolution. In CVPR Workshops (pp. 3061–3065).
[7] Yang, Y., & Qi, X. (2021). Image super-resolution via channel attention and spatial graph convolutional network. Pattern Recognition, 112, 107798.
[8] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In ECCV (pp. 286–301).
[9] Yoo, J., Kim, T., Lee, S., Kim, S.H., Lee, H., & Kim, T.H. (2023). Aggregated Contextual Transformers for Image Super-Resolution. In CVPR (pp. 28139–28149).