Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Exploring Deep Learning Strategies and Prospective Developments

Yi Sun *

School of Telecommunications Engineering, Xidian University, Xi'an, China

* Corresponding Author Email: 21009100110@stu.xidian.edu.cn

Abstract. Single Image Super-Resolution (SISR) is a process designed to transform Low-Resolution (LR) images into High-Resolution (HR) counterparts. This technology finds critical applications in various sectors, including gaming, photography, and medical imaging. With the advent and widespread success of deep learning, this approach has been increasingly applied in the realm of SISR. Deep learning-based SISR models are primarily categorized into three types based on their nonlinear module structures: Convolutional Neural Network (CNN)-based models, Generative Adversarial Network (GAN)-based models, and Transformer-based models. This paper presents a comprehensive overview of several emblematic models within each category. An in-depth analysis and comparison of their structural nuances and experimental outcomes are provided. This comparison elucidates how enhancements in network architectures and refined loss function optimizations contribute substantially to advancements in performance. Concluding with an analysis of current models, the paper outlines potential avenues for future exploration and development in the field of SISR, indicating a promising trajectory for further technological advancements.

Keywords: Single image super-resolution; deep learning; convolutional neural networks; generative adversarial networks; transformer.

1. Introduction

Single Image Super-Resolution technology generates High-Resolution images from Low-Resolution counterparts by analyzing LR image contents, deciphering mapping relationships between the two, and predicting and recreating missing information. SISR is an essential field in image processing and computer vision, aiming to enhance image quality without additional hardware. This technology has applications in various specialized fields, as will be discussed in section 4. The integration of deep learning into SISR commenced as the deep learning domain progressed. The inaugural deep learning-based SISR system, the Super-Resolution Convolutional Neural Network (SRCNN), was introduced by Dong et al. in 2014. SRCNN outperformed traditional models that did not use convolutional neural networks, excelling in Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM), and testing time. SRCNN's success demonstrated deep learning's potential in SISR [1]. Subsequently, researchers have proposed numerous models, achieving significant performance enhancements. Present SISR models primarily include Convolutional Neural Networks, Generative Adversarial Networks (GAN), and Transformer-based models, each differing in their approach to learning nonlinear mappings. CNNs offer robust feature extraction and simpler network structures, but the images they generate can lack perceptual realism, often appearing overly smooth. GANs, through adversarial training, can produce more realistic and refined images that align closely with human perception. However, GAN-based models are complex and challenging to train. Transformers, utilizing self-attention techniques, recognize long-range dependencies in images. Although Transformer-based models can achieve high performance, they also require substantial computational resources and GPU memory. Each structure has its own merits and limitations, contributing differently to the enhanced quality of generated images. This paper will further discuss select typical models.

2. Methods

2.1. CNN-based Models

CNN was proposed as early as 1989. However, without the advancement of computer technology and the growth of training samples, its use has not become widespread until recent days. CNN consists of convolutional layers, activation functions, and pooling layers. Thanks to advancements in computer technology, CNN is now widely used in the field of SISR. The first algorithm to use CNN in the SISR domain was SRCNN.

In SRCNN, there are three distinct layers. From the LR image, a collection of feature maps is extracted by the first convolutional layer, which then represents each of them as a high-dimensional vector. These vectors are then nonlinearly translated onto a second layer's high-dimensional vector. The ultimate high-resolution image is produced by combining predictions made by earlier layers. In experiments, three sets of convolutional kernels were chosen: $1\times9\times9\times64$, $64\times1\times1\times32$, and $32\times5\times5\times1$. SRCNN has a quite simple structure, but it achieved breakthroughs in reconstruction effects, leading deep learning research in the SISR domain.

Based on SRCNN, researchers further improve model performance through optimizations in network architecture, loss functions, and other aspects. Fast Super-Resolution Convolutional Neural Networks (FSRCNN) design a CNN structure which has a shape like an hourglass to make the generating progress faster [2]. FSRCNN can directly get a map from the original images to generated images without any pre-processing by using deconvolution filters. FSRCNN uses smaller convolutional kernels and a more profound network architecture, enabling enhanced efficiency. With these improvements, FSRCNN can be more than 40 times faster than SRCNN with better performance. Very Deep Convolutional Networks (VDSR) uses a very deep convolutional network to enhance model performance [3]. By increasing the depth of the network, VDSR achieves high performance, and it also finds that a network's performance improves with depth. But the number of parameters would also become larger as the depth improves, and the model becomes more difficult to train. SRResNet introduces residual networks to improve learning efficiency and network performance [4]. With the help of residual networks, SRResNet solves the problem of high computational consumption and memory use present in VSDR. EDSR improves the residual networks by eliminating the Batch Normalization (BN) layers, increasing the model size, and adopting residual scaling to enhance the stability of the training process [5].

While CNN-based models have advanced significantly in terms of SSIM or PSNR, there are still some problems such as image over-smoothing and limited receptive fields. Moreover, due to CNN models typically requiring deep structures and a great number of parameters, the training cost is high.

2.2. GAN-based Models

Images generated by CNN-based models tend to be overly smooth and often lose too many high-frequency details. To solve this problem, Ledig et al. first introduced GAN into SISR, proposing the Super-Resolution Generative Adversarial Network (SRGAN). SRGAN uses a generator and a discriminator to generate HR images [6]. The generator is used to generator HR images from LR images. SRGAN used SRResNet as the generator. By employing the adversarial loss, the produced images and the actual HR ones can be distinguished by the discriminator. The discriminator and generator undergo alternating training to raise the quality of the images generated. Moreover, SRGAN changes the loss function. Models before SRGAN tend to use mean-square error (MSE) to acquire higher PSNR or SSIM. Although this can be useful to get higher scores, it will lead to overly smooth images. SRGAN adopted a loss based on feature maps of the VGG network to solve this problem. Through these ways, SRGAN can generate more visually impressive images. Experiments found that although this loss function resulted in lower PSNR and SSIM compared to MSE, it significantly outperformed MSE in terms of mean-opinion-score (MOS). The high MOS indicates that images generated by SRGAN would be more realistic to human eyes.

SRGAN suffers from problems such as difficult training process and model instability. Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) significantly improves the quality of generated images through improvements in network architecture and loss functions. ESRGAN removes the BN layers and introduces Residual-in-Residual Dense Blocks (RDDB) in order to boost efficiency and lower computational complexity. ESRGAN improves the discriminator. The discriminator would predict relative realness rather than the absolute true or false. ESRGAN also improves the loss function. ESRGAN uses the VGG features before activation as perceptual loss rather than after activation, which contributes to the model's effectiveness in generating better images with more precise details and more realistic textures. All these improvements make ESRGAN a better-performing and easier-to-train model.

GAN-based models can generate images that are more realistic to human eyes, but they also suffer from some problems, such as model complexity, training instability, large computational consumption, and difficulty in assessing the quality of generated images.

2.3. Transformer-based Models

Transformer is a network design that was presented in 2017 and is based on an attention mechanism. Transformer entirely abandons the use of recurrence and convolutions and achieves great success in translation tasks. Transformer was initially proposed for language processing tasks. Transformer can catch global dependencies among input sequences by using self-attention mechanisms. This design enables the Transformer to effectively address long-distance dependency problems and achieves great performance. Due to the success, Transformer has also been applied to the field of SISR.

Liang et al. introduced Transformer to the field of SISR for the first time with their strong model, SwinIR, which is based on Swin Transformer [7]. Swin Transformer has the benefit of both CNN and Transformer. SwinIR is composed of three parts: shallow feature extraction, deep feature extraction, where Multiple Residual Swin Transformer Blocks (RSTB) are used in, and high-quality image reconstruction. Before the residual connection, RSTB is made up of a convolutional layer and some Swin Transformer layers. This construction not only enhances the translational equivariance of SwinIR but also allows for the combination of different levels of features. SwinIR achieves great performance in experiments, significantly out performances traditional CNN-based models. However, due to the use of Transformer, SwinIR consumes much computational resources and GPU memory. Efficient Super-Resolution Transformer (ESRT) was suggested to reduce resource consumption. ESRT combines a Lightweight CNN Backbone (LCB) and a Lightweight Transformer Backbone (LTB) [8]. To minimize resource usage, the feature map's size is modified using LCB. LTB consists of several Efficient Transformers (ET). ET use a specifically designed Efficient Multi-Head Attention (EMHA) mechanism to use less GPU memory. Compared to SwinIR, ESRT can achieve similar PSNR an SSIM while consume much less GPU memory. In the experiment, SwinIR consumes 6996M GPU memory while ESRT only consumes 4191M GPU memory. Efficient Mixed Transformer (EMT) proposes the Mixed Transformer Block (MTB) with several Transformer layers [9]. In MTB, the self-attention mechanisms in some layers are replaced by Pixel Mixers (PM). PM can use pixel shifting operations to enhance the local knowledge aggregation. EMT also uses striped window for SA (SWSA) to improve the global dependency modelling progress. By implementing these methods, EMT requires fewer parameters while achieving better performance compared to previous models.

Although Transformer-based models have incredible performance, they require great number of computational costs and GPU memory. Additionally, Transformer-based models tend to be complex and exhibit poorer training stability and convergence speed.

3. Results and Analysis

Table 1. Experiment results of typical models.

Models	Scale	Set5		Set14		BSD100		Urban100	
		PSNR/SSIM							
SRCNN	×2	33.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
	×3	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989
	×4	30.09	0.8627	27.18	0.7861	26.26	0.7291	24.52	0.7221
FSRCNN	×2	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020
	×3	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080
	×4	30.71	0.8657	27.59	0.7535	26.98	0.7398	24.62	0.7280
VDSR	×2	37.53	0.9588	33.03	0.9124	31.90	0.8960	30.76	0.9140
	×3	33.68	0.9201	29.86	0.9312	28.83	0.7966	27.15	0.8315
	×4	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524
SRResNet	×4	32.05	0.9019	28.49	0.8184	27.58	0.7620	26.07	0.7839
SRGAN	×4	29.40	0.8472	26.02	0.7397	25.16	0.6688	——	——
EDSR	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351
	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653
	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033
SRDenseNet	×4	32.02	0.8934	28.50	0.7782	27.53	0.7337	26.05	0.7819
RCAN	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384
	×3	34.74	0.9299	30.51	0.8461	29.32	0.8111	29.09	0.8702
	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087
ESRGAN	×4	32.73	0.9011	28.99	0.7917	27.85	0.7455	27.03	0.8153
HAN [9]	×2	38.33	0.9617	34.24	0.9224	32.45	0.9030	33.53	0.9398
	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705
	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094
SwinIR[10]	×2	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340
	×3	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624
	×4	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980
ESRT[11]	×3	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574
	×4	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962
EMT [12]	×3	34.80	0.9303	30.71	0.8489	29.33	0.8113	29.16	0.8716
	×4	32.64	0.9003	28.97	0.7901	27.81	0.7441	26.98	0.8118

Table 1 lists the PSNR and SSIM performance of some typical models on common datasets: Set5, Set14, BSD100, and Urban100. Dataset Set5 is a small dataset containing 5 high-quality images. Dataset Set14 includes 14 images, larger than dataset Set5. Dataset Set5 and Set 14 are often used for benchmark in SISR models. Dataset BSD100 is a part of Berkeley Segmentation Dataset and contains 500 images. Among the 500 images, 100 images are used for tests of tasks like SISR. Dataset Urban100 contains 100 images of urban sights such as buildings and city streets. Analyzing Table 1 leads to the following conclusion.

As the network structure deepens and the model architecture becomes more and more complex, the performance of the models significantly improves. VDSR uses 20 weight layers to improve performance and proves that the deeper the network is, the more performance it would have.

Combining various network modules can enhance the network performance. SRResNet introduces residual networks into SISR. SRDenseNet uses dense connected convolutional networks to enhance performance. The channel attention mechanism proposed by RCAN has the ability to adjust the channel-wise feature scale based on the interdependencies between the channels. The mechanism can solve the problem of decreased representational capability of CNNs due to the equal treatment of low-frequency information.

Improving the loss function can effectively boost the image's data metrics and texture details. In order to address the issue of very smooth images, SRGAN employs a loss function based on feature maps of the VGG network as opposed to MSE. It has been demonstrated that ESRGAN employs the VGG features prior to perceptual loss, as opposed to post activation, to offer more robust supervision for texture recovery and brightness consistency.

4. Application

SISR can be used in many different fields, including:

Medical Imaging: Improving the quality of medical photos is essential because LR images can substantially impede medical diagnosis. With the help of SISR, doctors can get much clear medical images. For example, a fast medical image super-resolution (FMISR) is proposed to help get HR medical images precisely and effectively [13].

Videos and Games: Due to insufficient transmission bandwidth or device performance, there tend not to be able to achieve high resolution while watching live streams or playing games. By using SISR, higher quality images can be obtained without consuming a lot of performance. For example, to increase game frame rates, NVIDIA created the Deep Learning Super Sampling (DLSS) technology, which produces scenes at a lower resolution and then uses a GPU to upscale them to the target resolution.

Photography: Through SISR, it's possible to generate higher resolution, clearer versions of photos or videos. For example, photos taken by smartphones in low light conditions or at long distances often lack clarity due to size constraints. By employing SISR, the computational resources of the phone can be fully used to generate photos that are significantly clearer and brighter. SISR can s significantly improve the overall photography experience, making high-quality imaging accessible to more people without professional equipment. Nowadays, many smartphone manufactures have employed SISR models in their products.

5. Possible Future Improvements

5.1. SISR under Unsupervised or Weakly Supervised Conditions

Nowadays most of the models are trained under fully supervised conditions, which require image pairs, LR and HR, as input. While the majority of LR images are obtained through fixed, solitary deterioration from HR photos, the LR images may not match the complex and various LR images in reality. This mismatch may affect the performance and application value of SISR models in real-world situations. Meanwhile, fully supervised model training consumes more human resources. SISR under Unsupervised or Weakly Supervised Conditions may help to solve these problems.

5.2. SISR Models in Embedded Devices

Embedded technology is becoming more and more prevalent in everyday life. SISR models can be quite useful in embedded devices. For example, images taken at night or at long distances can be processed into high-quality images. However, most of the models tend to be more and more complex and have many parameters in order to get high performance. They can get high PSNR or SSIM, but they are too complex and resource-consuming to be deployed on embedded devices. So, designing lightweight models is crucial to deploy them on embedded systems.

5.3. Improvements in Network Structure and Loss Functions

Research has indicated that deeper networks can function better, and that integrating different network modules can also improve network performance. SRResNet adds a residual network structure, and VDSR adopts an extremely deep network structure. Both achieve performance improvements through improvements in network architecture. SRGAN improved the loss function,

which greatly improved the rebuilt images' perceptual quality. Therefore, improving the loss function can also enhance model performance. With the advancement of computing capabilities, it is possible to create loss functions and network architectures that are deeper and more complicated. These models can capture better features in images, thus improving the accuracy and quality of super-resolution.

6. Conclusion

This review examines Single Image Super-Resolution technology within the context of deep learning, focusing on various models derived from Convolutional Neural Networks, Generative Adversarial Networks, and Transformers. It compares these models, highlighting differences in their nonlinear module structures. The analysis of model structures and empirical data reveals that enhancements in network architectures and loss functions can significantly boost both performance and efficiency of these models. Furthermore, the integration of diverse network modules is shown to be crucial in advancing model capabilities. Additionally, the paper delves into practical applications of SISR technology and presents several potential avenues for future exploration. It is hoped that the insights provided herein will serve as a valuable resource for ongoing research in the SISR domain.

Reference

[1] Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13 (pp. 184-199). Springer International Publishing.

[2] Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14 (pp. 391-407). Springer International Publishing.

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

[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. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

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

[6] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0).

[7] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844).

[8] Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., & Zeng, T. (2022). Transformer for single image super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 457-466).

[9] Zheng, L., Zhu, J., Shi, J., & Weng, S. (2024). Efficient mixed transformer for single image super-resolution. Engineering Applications of Artificial Intelligence, 133, 108035.

[10] Tong, T., Li, G., Liu, X., & GAO, Q. (2017). Image super-resolution using dense skip connections. In Proceedings of the IEEE international conference on computer vision (pp. 4799-4807).

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

[12] Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S. ... & Shen, H. (2020). Single image super-resolution via a holistic attention network. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16 (pp. 191-207). Springer International Publishing.

[13] Zhang, S., Liang, G., Pan, S., & Zheng, L. (2018). A fast medical image super resolution method based on deep learning network. IEEE Access, 7, 12319-12327.