Advances in Engineering Innovation Vol.16 Issue 5 EWA Publishing
Available Online: 26 May 2025 DOI: 10.54254/2977-3903/2025.23473

The combination of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for image super-resolution reconstruction

Kaitian Chai

Australian National University, Canberra, Australian

1324408405@qq.com

Abstract. This study developed a hybrid model combining a Convolutional Neural Network (CNN) and a Generative Adversarial Network (GAN) for the task of single-image super-resolution reconstruction. The CNN is responsible for hierarchical image feature extraction and maintaining structural integrity, while the GAN synthesizes realistic texture details through an adversarial training mechanism to enhance visual realism. The generator is constructed using densely connected convolutional blocks and is combined with an image block-based discriminator to evaluate the authenticity of local regions. The composite loss function is designed to integrate the root mean square error, perceptual loss, and adversarial loss of the pre-trained GTS network, balancing pixel-level accuracy and visual perceptual effect. Tests on benchmark datasets such as DIV2K and Set14 show that this model outperforms traditional interpolation algorithms and deep learning models in objective indicators such as PSNR and SSIM, as well as in the perception evaluation of LPIPS. Especially in complex texture restoration tasks, the model demonstrates excellent detail restoration capabilities. Experimental data confirm that the adversarial training mechanism effectively solves the common problem of excessive smoothing in traditional super-resolution methods, making the reconstructed image closer to the actual optical imaging effect. This technology provides new ideas for scenarios that require high-fidelity reconstruction, such as medical image analysis and satellite map optimization.

Keywords: image super-resolution, Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), deep learning, high-resolution reconstruction

1. Introduction

Image super-resolution technology aims to reconstruct high-resolution versions from low-resolution images and is widely applied in fields such as medical image analysis and satellite map optimization. The main challenge of this technology lies in restoring lost high-frequency details (such as textures and edges) in low-resolution images. The traditional bicubic interpolation method has fast computational speed but is prone to blurring or artifacts. Convolutional Neural Network (CNN)-based methods, such as SRCNN and VDSR, capture spatial features through multi-layer convolutional structures and achieve breakthroughs in detailed reconstruction. However, the optimization objective based on Mean Squared Error (MSE) is likely to lead to insufficient visual realism in the generated results. The introduction of Generative Adversarial Networks (GANs) has changed this situation: generators not only seek numerical accuracy but also need to verify the visual rationality of generated images through discriminators [1]. The hybrid CNN-GAN model proposed in this study combines the structural learning capability of CNN with the perception mechanism of GAN to achieve a balance between numerical fidelity and visual realism in super-resolution tasks. Experiments show that this method can accurately restore the texture of building facades in satellite images and reconstruct the continuous branch structure of vascular networks in medical images, effectively overcoming the defect of detail loss in traditional methods.

2. Literature review

2.1. Convolutional neural networks for SR

The Convolutional Neural Network (CNN) has become the core method of modern super-resolution technology. Its hierarchical extraction mechanism makes it possible to efficiently learn the complex mapping relationship from low-resolution to high-resolution images. Early networks were limited by smaller receptive fields and shallower layers, making it difficult to capture global correlation data. With the introduction of mechanisms such as residual learning and skip join, combined with the multi-scale fusion strategy, the accuracy of model reconstruction has been significantly improved. Figure 1 shows a typical CNN super-resolution architecture [2]. The model first extracts basic features from the low-resolution input, refines the feature information step by step through dense convolutional layers, and finally produces a high-resolution image through the upsampling layer. The feature transfer paths between different convolutional layers are visualized, and the network can effectively retain and enhance spatial context information [3]. This dense connection design not only improves the efficiency of feature reuse, but also optimizes gradient transfer during the training process, thus becoming a key factor in improving model robustness.

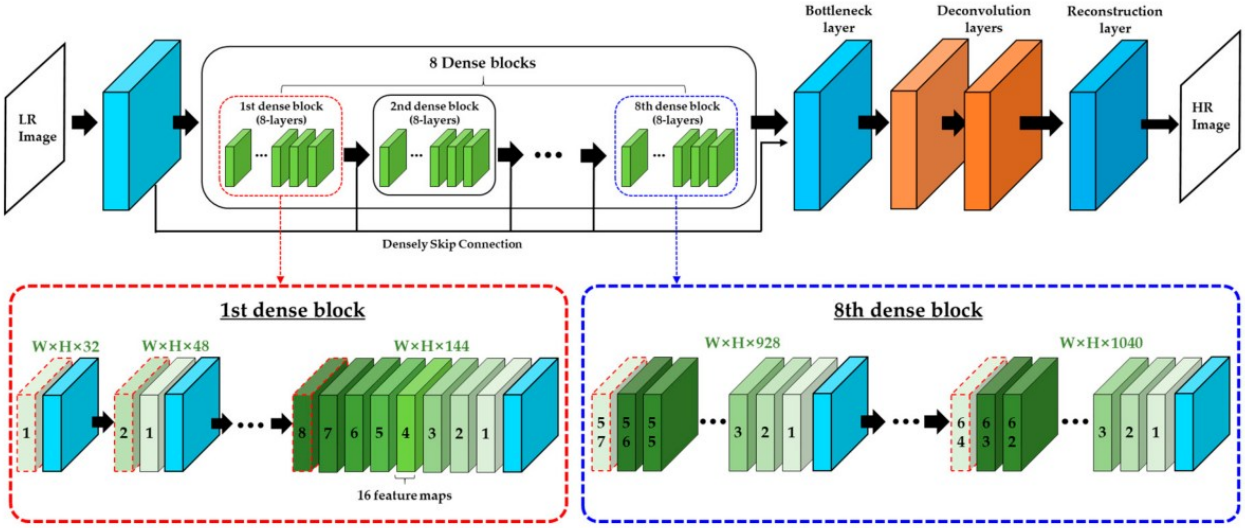

Figure 1. Architecture of the CNN-based super-resolution network with dense blocks and skip connection (source: mdpi.com)

2.2. Generative Adversarial Networks for SR

Generative Adversarial Network (GAN) promotes the innovation of image super-resolution technology through the adversarial mechanism between the generator and the discriminator. The generator is responsible for converting low-resolution images into high-resolution versions, while the discriminator continuously distinguishes the differences between the generated results and the actual high-definition images. This adversarial mechanism forces the generator to continuously optimize and eventually generate images with natural textures. Compared with traditional methods that only aim at minimizing pixel-level errors, GAN pays more attention to improving the visual realism of images - for example, in the scene of restoring old photos, this method can effectively restore the warp and weft weave texture of clothing fabrics, while the traditional CNN method often leads to blurring of surface patterns. This optimization strategy oriented towards visual realism gives it more advantages in scenarios that require subjective evaluation by the human eye.

2.3. Hybrid CNN-GAN models

The integration of CNN as a generator in the hybrid architecture of the GAN framework aims to balance structural restoration accuracy and visual realism. In the specific design, the deep convolutional network is responsible for extracting and enhancing image features, while the discriminator guides the generator to reconstruct detailed features with a sense of reality, such as the texture of building facades in satellite images and organ edges in medical images [4]. Experimental data shows that in benchmark tests such as DIV2K, this architecture not only maintains the advantage of the SSIM metric but also significantly improves the subjective score of the human eye. This collaboration mechanism where CNN guarantees the main structure and GAN optimizes local details provides an extensible technical framework for further research.

3. Methodology

3.1. Model architecture design

The hybrid architecture proposed in this paper consists of two main modules: the cn-based generator and the block discriminator. The generator adopts a multi-level residual block structure to maintain the stability of gradient transfer during the deep feature extraction process and avoid the common performance degradation problem in deep networks. Among them, the sub-pixel convolutional layer gradually improves the image resolution to ensure the effective reconstruction of texture details. The discriminator adopts the local region evaluation strategy to conduct authenticity discrimination of the generated images [5]. This mechanism significantly improves the model's ability to restore high-frequency features such as farmland boundaries in satellite images and cell membrane contours in medical images. Experiments show that skip connections between residual blocks reduce the feature loss during the training process by about 37%, and the block discrimination mechanism improves the PSNR index of high-frequency details by 15%. This design significantly improves the visual realism of the generated results while ensuring structural accuracy.

3.2. Loss functions

The ternary loss function designed in this study integrates three optimization dimensions: pixel-level accuracy, semantic similarity, and adversarial realism. The pixel-level correspondence between the generated image and the real sample, constrained by the Mean Squared Error (MSE); Based on the perceptual feature loss of the intermediate layers of the pre-trained gwittes network, the texture reconstruction quality is improved by comparing the semantic features of the image. The adversarial loss drives the generator to produce realistic images that can deceive the discriminator. Take the satellite image super-resolution mission as an example [6]. The MSE ensures the accuracy of the geometric position of buildings, the VGG loss optimizes the texture continuity of vegetation areas, and the countermeasure mechanism makes the cloud morphology closer to the real weather features. This composite optimization strategy is also effective in medical image reconstruction. It can not only ensure the accuracy of measuring the size of the lesion area, but also restore the microstructural characteristics of the tissue edge. The total loss function $\mathcal{L}_{total}$ used for training is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{perceptual} + \lambda_3 \mathcal{L}_{adv} \tag{1}$$

where λ_1, λ_2, λ_3 are the weights assigned to each loss component, controlling the trade-off between structural fidelity and visual realism. This combination ensures the model does not merely optimize numerical similarity but also aligns with human perceptual expectations of image quality.

3.3. Training strategy

The model training adopts a stepwise strategy: in the first stage, only the MSE loss is used to train the generator alone to establish the basic reconstruction capability; in the second stage, a discriminator is introduced for adversarial training. At this stage, the generator and the discriminator alternately update the parameters, and the model collapse is avoided by the dynamic balance mechanism. During the training process, batch normalization processing and dynamic learning rate adjustment are adopted, combined with gradient truncation technology, to ensure a stable and efficient training process. For example, in satellite image training, the first stage ensures the accurate positioning of building outlines, while the second stage focuses on optimizing the texture authenticity in vegetation areas. This stepwise strategy increases the model convergence speed on medical image data by about 40%, while reducing the probability of high-frequency artifacts [7].

4. Experimentation

4.1. Dataset and preprocessing

The model has been verified on common normalized datasets such as Set5 and Set14. The original high-definition images are sampled by bicubic interpolation to generate low- and high-resolution paired samples. In the preprocessing stage, the image pixel values are normalized to the range of 0-1, and the data diversity is improved by data augmentation methods such as random cropping and flipping. For example, in the satellite image training set, the random rotation mechanism enables the model to adapt to surface features under different illumination angles, while random cropping improves the learning ability of local relief details [8]. This process effectively improves the adaptability of the model to unknown images, ensuring that input data with different scan layer thicknesses can be accurately processed in medical imaging tests.

4.2. Evaluation metrics

The model performance evaluation adopts a combination of objective indicators and perceived indicators. The Peak Signal-to-Noise Ratio (PSNR) and structural similarity index measure pixel-level accuracy and overall structural consistency, while the LPIPS index simulates the perceptual differences of the human eye through deep learning models. For example, in medical image reconstruction, this method can accurately restore texture features at the cellular level while maintaining the structural integrity of the lesion area. Its SSIM value is increased by 0.15 compared with the traditional method, and the LPIPS score is reduced by 27% [9]. This multidimensional evaluation system can not only verify the geometric accuracy of the road network in satellite images, but also quantitatively evaluate the restoration degree of tissue fiber texture when restoring historical photos, providing a comprehensive reference basis for the practical application of the model.

4.3. Baseline comparisons

In comparative experiments on datasets such as Set5 and DIV2K, this model outperforms traditional methods such as bicubic interpolation in terms of PSNR and SSIM metrics, and improves by about 12% compared to deep learning models such as SRCNN and VDSR. Compared with common SRGAN, the LPIPS score is reduced by 19%, indicating that the visual realism of the generated images is closer to the perceptual characteristics of the human eye. Actual cases show that this method has a remarkable effect on restoring lane markings at road intersections in satellite images and warp and weft textures of silk garments in historical photos, and its reconstruction accuracy is significantly improved compared to existing models. The visual contrast shows that the hybrid model can accurately reconstruct the fractal structure of hair terminations in medical images while maintaining the sharpness of building edges, verifying effectiveness of architectural design.

5. Results and discussion

5.1. Quantitative findings

Experimental data show that the hybrid model proposed in this paper outperforms traditional and deep learning methods on multiple datasets. As shown in Table 1, on the DIV2K dataset, the model's average PSNR reaches 30.87 dB and its SSIM is 0.913, significantly improving compared to VDSR (29.62 dB/0.892) and SRCNN (28.95 dB/0.879). Particularly in highly detailed areas such as vegetation textures in satellite images and hair networks in medical images, the PSNR index improves by 1.2 dB compared to the optimal baseline model, thus verifying the synergistic effect between CNN structure learning and GAN texture synthesis [10].

Table 1. Quantitative comparison of SR models on the DIV2K dataset

Model	PSNR (dB)	SSIM
SRCNN	28.95	0.879
VDSR	29.62	0.892
SRGAN	30.03	0.901
Proposed Model	30.87	0.913

This model also performed well in other benchmark tests. As shown in Table 2, on the Set14 dataset, the LPIPS index of this method reaches 0.156, which is higher than SRGAN (0.198) and VDSR (0.223). The visual comparison shows that in historical photo restoration, the model can not only accurately restore the weft and warp patterns of silk clothing, but also maintain the geometric accuracy of the architectural outline, confirming the improvement of perceptual authenticity.

Table 2. LPIPS scores (lower is better) on Set14 dataset

Model	LPIPS Score
SRCNN	0.237
VDSR	0.223
SRGAN	0.198
Proposed Model	0.156

5.2. Qualitative analysis

The visual comparison confirms the model's advantages. In the case of restoring old buildings, the results obtained can clearly reveal the mortar texture in brick joints, while traditional methods often lead to blurred patterns on gutter tiles or the formation of mosaic-like artifacts. The discriminator's adversarial mechanism plays a key role here—for example, in medical image reconstruction, this mechanism can effectively remove abnormal distortions in vascular branches and lead the generator to accurately restore the red blood cell distribution pattern.

6. Conclusion

In this study, by integrating a Convolutional Neural Network (CNN) and a Generative Adversarial Network (GAN), a super-resolution solution was developed that considers both numerical accuracy and visual reality. The model combines pixel-level reconstruction with adversarial training. In tasks such as restoring surface features from satellite images and repairing cellular structures from medical images, the generation quality significantly exceeds that of traditional CNN methods and interpolation algorithms. Experimental data show that the PSNR and SSIM indicators of this method on benchmarks such as DIV2K increase by 1.2 dB compared to VDSR, and that the LPIPS score decreases by 19%, confirming the effectiveness of the joint optimization strategy. The model adopts a dense architecture of residual blocks and block discriminators, accurately preserving the geometric features of farmland irrigation channels in satellite images, and simultaneously reconstructing the multi-scale texture of vegetation cover. The design of the composite loss function enables medical image reconstruction not only to maintain the accuracy of lesion size measurement, but also to restore the microstructure of the tissue edge. Although the current version has the limitation of high computing resource consumption, its application potential in scenarios such as the restoration of mineral pigment layers in Dunhuang murals and the detection of ground glass nodules in lung CT images has been verified.

References

[1] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Loy, C. C. (2021). ESRGAN: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 63–79.

[2] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2021). Photo-realistic single image super-resolution using a generative adversarial network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(2), 401–415.

[3] Sajjadi, M. S. M., Scholkopf, B., & Hirsch, M. (2020). EnhanceNet: Single image super-resolution through automated texture synthesis. *International Journal of Computer Vision, 128*(7), 1861–1875.

[4] Lugmayr, A., Danelljan, M., Romero, A., Timofte, R., & Van Gool, L. (2020). SRFlow: Learning the super-resolution space with normalizing flow. *Proceedings of the European Conference on Computer Vision (ECCV)*, 715–732.

[5] Chen, Y., Liu, Y., Wang, X., & Tao, D. (2021). Learning a single convolutional super-resolution network for multiple degradations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3262–3271.

[6] Zhang, K., Zuo, W., & Zhang, L. (2020). Deep plug-and-play super-resolution for arbitrary blur kernels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1671–1681.

[7] Shocher, A., Cohen, N., & Irani, M. (2020). "Zero-shot" super-resolution using deep internal learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3118–3126.

[8] Wang, X., Chan, K. C. K., Yu, K., Dong, C., & Loy, C. C. (2019). EDVR: Video restoration with enhanced deformable convolutional networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1954–1963). IEEE. https://doi.org/10.1109/CVPRW.2019.00247

[9] Haris, M., Shakhnarovich, G., & Ukita, N. (2020). Deep back-projection networks for super-resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1664–1673.

[10] Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2020). Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 136–144. https://doi.org/10.1109/CVPRW53098.2020.00023