Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28945

Attention-Based Optimization of ESRGAN for Old Photo Restoration

Tongxin Ben[1], Nuo Xu[2*]

[1]*School of Computer Science, NJUPT, Nanjing University of Posts and Telecommunications, Nanjing, China*
[2]*School of Physical and Mathematical Sciences, Nanjing Tech University, Nanjing, China*
**Correponding Author. Email: promise@njtech.edu.cn*

Abstract. The excessive degradation of old photos poses significant challenges to the restoration of the image and human visual perception. To address this issue, we propose an optimized framework based on Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN). This method combines the attention mechanism of the Convolutional Block Attention Module (CBAM) with the face detection module to enhance the feature extraction of key channels and spatial locations to improve the repair effect of high-level semantic areas, such as faces, by weighting local losses. At the same time, we use L1 loss to replace traditional MSE loss, and combine perceived loss with fine-tuning strategies to improve the stability and convergence efficiency of model training. Finally, our research results on DIV2K and self-built old photo datasets show that the proposed model achieves an improvement of 0.17 in PSNR and 0.008 in SSIM over baseline ESRGAN, and the clarity and naturalness of face details have been greatly improved. In addition, the optimized model achieves higher scores in subjective evaluations, further demonstrating its superiority in old photo restoration.

Keywords: Old Photo, Restoration, ESRGAN, Attention Mechanism

1. Introduction

As an important medium for carrying historical memory, old photos often suffer from quality degradation problems such as blurring, scratches, low contrast, and loss of detail due to their age [1]. Traditional image restoration methods (such as filtering and interpolation) are difficult to effectively restore complex degraded information, especially in the detailed restoration of key areas such as faces.

In recent years, super-resolution (SR) techniques have developed rapidly, offering new approaches for image restoration [1]. In 2014, Goodfellow first proposed Generative Adversarial Networks (GANs), which quickly attracted attention due to their excellent effects and broad applicability, and became a focus of attention in artificial intelligence research [2]. In 2017, Ledig et al. developed the SuperResolution using a Generative Adversarial Network (SRGAN) method, which became the first neural network architecture to apply the GAN concept to image super-resolution generation tasks [3]. In 2018, XintaoWang et al. introduced an innovative Residual-in-

© 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.TJ28945

Residual Dense Block (RRDB) structure based on SRGAN, and proposed the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) model, which significantly improved the visual quality and stability of image reconstruction [4]. Today, ESRGAN has been widely adopted in various image enhancement and restoration tasks due to its superior generative capabilities.

However, ESRGAN still faces problems such as facial blurring, texture loss, and unclear edges when applied to severely degraded old photos. In addition, existing studies have shown that combining facial priors and using a multi-scale network structure can further improve restoration performance [5]. Therefore, motivated by these findings, this project explores how to combine these advanced mechanisms to make the optimized ESRGAN model more versatile and adaptable while maintaining high visual fidelity.

Therefore, we chose to optimize the structure of the original ESRGAN model and introduce an attention mechanism and a face detection module to improve the model's ability to repair key restoring regions, so as to achieve a more realistic and delicate restoration effect of old photos.

2. Background and motivation

2.1. ESRGAN model structure

Image generation based on ESRGAN is one of the widely used methods for image restoration and reproduction [6]. The Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) model consists of three core components: the generator, the discriminator, and the perceived loss, which are used to generate clearer, more natural high-resolution images from low-resolution inputs. They take on different roles, where the generator is based on the Residual-in-Residual Dense Block (RRDB) structure and is responsible for generating high-resolution images from low-resolution images. The Discriminator guides the generator to learn more realistic and natural image details; Perceptual Loss: Measures the difference between images in the high-level feature space to improve the subjective visual quality of the generated image.

2.2. CBAM attention mechanism

In the old photo restoration task, the traditional convolutional neural network is easily disturbed by background noise during feature extraction, resulting in the model's insufficient attention to key information areas. Therefore, this study introduces the Convolutional Block Attention Module (CBAM) on the basis of ESRGAN. Through the joint modeling of channel attention and spatial attention, the CBAM module enables the network to highlight important channels and key spatial areas in feature representation, thereby effectively improving the model's repair ability on key structures such as facial details and edge textures. Through this mechanism, the model can not only reduce the interference of some irrelevant information, but also enhance the attention to key semantic areas in old photos, so as to achieve a restoration effect that is more in line with human perception.

2.3. Local weighted enhancement mechanism

In human perception, the face is the most critical area in old photos, but traditional super-resolution methods often have blurred facial features or unnatural structures when restoring faces. To solve this shortcoming, we introduce a lightweight face detection algorithm and implement an enhancement mechanism for local weights in the loss function based on the detection results, so that the model pays more attention to the face area during training, thereby improving the clarity and restoration

DOI: 10.54254/2755-2721/2026.TJ28945

ability of details such as eyes and lips. This method can not only better maintain the consistency of the overall structure, but also enhance the authenticity and naturalness of the facial area, which is more in line with the human perception of the priority of the face, and provides a more practical improvement solution for the restoration of old photos.

3. Proposed methods

To improve the performance and robustness of ESRGAN in old photo restoration tasks, we propose a series of optimization strategies based on the original framework.

First, it introduces a face detection module to automatically identify face areas in the image during the preprocessing stage and weight them to make the model training more focused on key content.

Secondly, the L1 loss function is used to replace the traditional MSE loss, which better maintains the image structure details at the pixel level and improves the stability of the training process. To enhance the performance and robustness of ESRGAN in old photo restoration tasks, this study proposes a series of optimization strategies on top of the original framework.

In addition, the CBAM attention mechanism is embedded in the RRDB module of ESRGAN, and the model's ability to express important texture areas is enhanced through joint modeling of channel and spatial attention. In addition, we fine-tune and reinforce the pre-trained model to reduce training costs and accelerate convergence while retaining the original performance advantages.

Finally, a dynamic loss-weighting strategy is introduced in the face area or other high-semantic regions to guide the model to pay more attention to the recovery of key areas.

Through the above comprehensive optimizations, it has successfully improved the overall structure preservation, detail restoration, and performance of key areas of the model, providing an effective solution for high-quality restoration of old photos.

4. Experiments

4.1. Data

To validate the effectiveness of the proposed method, we constructed the experimental dataset by combining public datasets with a self-collected dataset. The publicly available DIV2K dataset of high-resolution (HR) images was used as the training source, where low-resolution (LR) samples were generated by applying Bicubic downsampling (×4) to the HR images, ensuring reproducibility and standardization [7, 8].

Meanwhile, to improve the model’s adaptability to real-world degradations, we collected and curated a large number of authentic historical old photographs, building a self-constructed dataset for supplementary training and validation. In addition, to emphasize the importance of facial regions in old photo restoration, we established a face-annotated subset consisting of approximately 500 historical photos with human faces. These were annotated using automatic detection followed by manual correction to provide high-quality labels for targeted evaluation of facial restoration performance.

4.2. Training details

It first loaded the original ESRGAN as the baseline model, using the official implementation of RRDBNet as the generator in combination with adversarial training and perceptual loss. This

DOI: 10.54254/2755-2721/2026.TJ28945

provided a stable and widely validated starting point, enabling reproducible and quantifiable comparisons. The baseline was tested on the old photo dataset to evaluate restoration performance.

Building on this baseline, we embedded the Convolutional Block Attention Module (CBAM) into the RRDB blocks, connecting channel and spatial attention modules to the RRDB outputs to form the RRDB-CBAM structure. CBAM was applied to the output features of each RRDB before merging with the residual path, resulting in an improved generator design capable of selectively emphasizing critical textures and structural regions while reducing attention dispersion.

The improved model was then fine-tuned on the DIV2K dataset, training on paired HR images and their Bicubic ×4 downsampled LR counterparts. This process aligned the model distribution on clean, standardized pairs, providing a solid foundation for subsequent transfer to real old photos. Consistent with the baseline, identical data augmentation methods and evaluation protocols were maintained, while carefully balancing adversarial loss weights to prevent over-sharpening and artifacts.

For facial detail optimization, we employed local weighted loss combined with perceptual loss and L1 loss. Since faces are the most perceptually sensitive regions, this weighting strategy guided the model to allocate more capacity toward restoring details such as eyes, lips, and edges. Facial bounding boxes and keypoints were obtained using RetinaFace, and a binary mask M was generated for localized weighting.

After convergence, the fine-tuned model ensures consistency with the inference script and compatibility with baseline weights for reproducible experiments. Finally, a unified testing script was developed to evaluate both baseline and improved models. On the DIV2K validation set, PSNR and SSIM metrics were computed, while side-by-side visual comparisons with cropped regions were saved to illustrate differences in detail restoration. This provided both quantitative improvements and qualitative validation of the benefits of attention mechanisms and local weighting strategies in old photo restoration.

4.3. Experimental setup

To assess the effectiveness of the optimized ESRGAN model in old photo restoration, we conducted evaluations on both the public DIV2K validation set and the self-constructed historical photo dataset. All experiments were implemented using the PyTorch framework, with training performed on an NVIDIA RTX 3060 GPU, batch size of 16, and an initial learning rate of 1e-4. During testing, all images were kept at their original resolution to avoid the influence of additional rescaling on the results.

5. Evaluation

5.1. Comparison with existing methods

This study compares the performance of the original ESRGAN model and the optimized CBAM-ESRGAN model on the same set of historical photos, as shown in Figure 1.

On the self-constructed historical photo dataset, 20 volunteers were invited to conduct a subjective evaluation (scored on a scale of 1–5, with higher scores indicating better visual quality). The results show that the optimized model achieved an average score of 4.36, an improvement of 0.42 over the baseline. Most participants agreed that the optimized model demonstrated superior performance in noise removal, detail completion, and maintaining overall naturalness, especially in face restoration, where it better aligns with human visual perception.

DOI: 10.54254/2755-2721/2026.TJ28945



(a)Original model (b) Optimized model

Figure 1. Comparison of restoration results between the original and optimized models (photo/picture credit: original)

In image super-resolution research, selecting appropriate objective evaluation metrics is crucial for assessing model performance. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are the two most widely used metrics in this field.

First, PSNR, as a traditional full-reference metric, directly reflects the pixel-level error between the reconstructed image and the original high-resolution image [9]. Although PSNR does not fully capture human subjective perception, it is simple to compute and easy to compare, making it widely applied in image restoration and super-resolution tasks.

Second, SSIM is more consistent with the characteristics of the human visual system (HVS) [10]. By incorporating luminance, contrast, and structural information in addition to pixel differences, SSIM provides a better evaluation of perceptual image quality. In recent super-resolution studies, SSIM has increasingly become an important metric alongside PSNR, often used to analyze structural fidelity and visual consistency.

For these reasons, this study adopts PSNR and SSIM as benchmark metrics in the experimental comparisons, as shown in Table 1.

Table 1. Quantitative results and evaluation of different ESRGAN variants

Model Version	PSNR	SSIM	Visual Performance
Original ESRGAN	27.85	0.764	Stable overall restoration with acceptable textures.
CBAM-ESRGAN	28.02	0.772	Slight improvement in local details

From the quantitative perspective, the improved CBAM-ESRGAN outperforms the original model in both PSNR and SSIM, with improvements from 27.85 to 28.02 and from 0.764 to 0.772, respectively. This indicates slight enhancements in pixel-level reconstruction quality and structural similarity. In terms of visual performance, CBAM-ESRGAN achieves better local detail restoration.

5.2. Limitations

Despite the improvements achieved by the optimized model, several limitations remain. First, the generalization ability of the model may not be fully realized due to the limited scale and quality of the datasets used, which constrains its performance on unseen data. Second, although the introduction of the CBAM attention mechanism enhances feature selection, its effectiveness may

DOI: 10.54254/2755-2721/2026.TJ28945

require integration with more sophisticated network architectures or multi-scale feature fusion mechanisms to achieve a more comprehensive improvement. Moreover, the locally weighted loss strategy exhibits high sensitivity to hyperparameters during training, meaning that its stability and effectiveness still rely on careful hyperparameter tuning and training strategy design. These issues highlight clear directions for future research.

To further enhance the performance of the optimized CBAM-ESRGAN, several strategies can be explored. One promising direction is the introduction of multi-scale attention mechanisms, such as asymmetric self-attention. According to a study by J. Yang [11], combining positional attention with asymmetric convolution has demonstrated improved perception of multi-level details and enhanced generalization performance in related tasks. Additionally, incorporating facial priors (e.g., landmarks or semantic segmentation maps) may strengthen face restoration quality. Transfer learning could also be employed to enable rapid model adaptation and fine-tuning with a small number of real historical photo samples

6. Future work

Beyond its effectiveness in old photo restoration, the proposed optimized ESRGAN model demonstrates strong cross-domain adaptability. This model demonstrates considerable potential for application across various domains related to image enhancement and image generation. For instance, in cultural heritage preservation scenarios such as historical document restoration, film reel repair, and art archive reconstruction, the model can improve image detail levels and enhance visual quality, thereby supporting the digital reproduction and study of cultural relics.

In fields such as public security surveillance, facial recognition, and news image archiving, the model's capability to enhance facial details also proves valuable. By improving key regions in blurred images, the model can assist in recognition and subsequent information mining, thereby increasing the usability and analytical depth of visual data.

Furthermore, the attention mechanism and local region weighting strategy proposed in this study exhibit strong generalizability, making them applicable to tasks such as video super-resolution, medical image enhancement, and multimodal image generation. With the integration of temporal modeling mechanisms, future research could explore frame-to-frame consistency in video restoration or leverage semantic guidance to achieve more precise content generation.

In addition, with the advancement of model compression techniques, this type of model has the potential to evolve toward lightweight deployment. Such development would enable image restoration under resource-constrained conditions, making it feasible for edge devices and mobile platforms.

7. Conclusion

This study, driven by practical application needs, systematically optimized and experimentally validated the ESRGAN model for the task of old photo restoration. By successfully reproducing the original ESRGAN model and innovatively integrating the CBAM attention mechanism, the study achieved effective module embedding and feature fusion, significantly enhancing the model's ability to perceive critical features. On this basis, a complete training and evaluation workflow was established, employing a combination of quantitative and qualitative verification methods.

Experimental results demonstrate that the optimized model achieves substantial improvement in detail quality on test images, particularly showing notable enhancement in the restoration of facial regions. The textures are restored more clearly and naturally, effectively validating the improvement

DOI: 10.54254/2755-2721/2026.TJ28945

and practical value of the proposed method. This research provides an effective solution for old photo restoration and holds positive practical significance for the preservation and digital restoration of historical images.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Wang, M. (2024). Research on image super-resolution reconstruction and color restoration of ancient calligraphy and painting based on generative adversarial networks (Master's thesis, Beijing University of Posts and Telecommunications, Beijing, China).
[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144.
[3] Ledig, C., Theis, L., Huszár, F. (2017). Photo-realistic single-image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 105–114). IEEE.
[4] Wang, X. T., Yu, K., Dong, C. (2018). Recovering realistic texture in image super-resolution by deep spatial feature transform. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 606–615). IEEE.
[5] Hou, D., Bian, D., Luo, Z. (2025). A lightweight super-resolution reconstruction method for phase contrast interferometric microscopy images. Optics Communications, 591, 132050.
[6] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Loy, C. C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. arXiv.
[7] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. arXiv. https: //arxiv.org/abs/1807.06521
[8] Agustsson, E., & Timofte, R. (2017, July 20). DIV2K: DIVerse 2K resolution high quality images. Retrieved August 25, 2025, from https: //data.vision.ee.ethz.ch/cvl/DIV2K/
[9] Singh, A., & Rodriguez, C. R. (2023). Peak signal-to-noise ratio in image quality evaluation. In Digital Image Enhancement and Reconstruction. ScienceDirect. https: //www.sciencedirect.com/topics/engineering/peak-signal-to-noise-ratio
[10] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600–612.
[11] Chu, J. (2025). Remote sensing image object detection method based on position attention and asymmetric convolution. Geomatics and Spatial Information Technology, 48(1), 108–111.