Journal of Engineering, Project, and Production Management
2026, 16(7), 2026-0061

Face Replacement Method for Film and Television Drama Videos Based on Generative Adversarial Network and Task Decomposition Strategy

Ludongdong Shan[1], Li Zhou[2], and Yuelin Du[3]
[1] Associate Professor, Hunan University of Media and Communications, China
[2] Department Head, Indoor Choir in the School of Art, Shijiazhuang Institute of Technology, China, E-mail: zhouulii@outlook.com (corresponding author).
[3] Director, School of Art, Shijiazhuang Institute of Technology, China

Production Management
Received May 29, 2026; revised July 14, 2026; accepted July 23, 2026
Available online August 12, 2026

Abstract: The current face replacement technology faces problems such as insufficient detection accuracy and poor image quality when processing high-resolution videos. A face replacement method for film and television videos, grounded in improved multi-task Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GANs), is proposed to enhance face detection accuracy and the naturalness and realism of replacement images. An improved Multi-Task Convolutional Neural Network (MTCNN) algorithm that incorporates depthwise separable convolution and a feature pyramid structure is studied. A generator-discriminator structure based on an improved generative adversarial network is designed, with residual blocks introduced to enhance performance. The research findings indicate that the raised generative adversarial network achieves a Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of 0.10, a peak signal-to-noise ratio close to 25.0dB, a structural similarity index close to 1.0, and an emotion retention rate of 85%. Meanwhile, its computing resource consumption is 19%, and its robustness is 90%, indicating excellent performance. The proposed method for facial replacement in film and television dramas can raise the efficiency and quality of facial replacement, providing technical support for post-production and video content creation, as well as new ideas and methods for related technology research.

Keywords: GAN; task decomposition strategy; face replacement; film and television drama videos; facial detection.

DOI 10.32738/JEPPM-2026-0061

1. Introduction

As digital technology advances rapidly, video processing for film and television dramas has gradually become a research hotspot in the field of multimedia. Efficient and natural face replacement in videos has significant applications in video content creation and digital entertainment. As deep learning technology rises, especially with the widespread application of Generative Adversarial Networks (GANs), facial replacement technology has advanced significantly. However, current technology still faces many challenges when processing high-resolution videos (Deuze and Beckett 2022; Rowe et al., 2022). In face detection, the Multi-task Convolutional Neural Network (MTCNN) has achieved strong results, but its detection speed and accuracy still need improvement for high-resolution videos. In addition, the core of face replacement technology lies in generating realistic replacement images. Despite their strong image-generation capabilities, traditional GAN architectures frequently struggle to balance image quality and naturalness of replacement when attempting complex face-replacement tasks (Xia et al., 2022; Huang and Jafari, 2023).

GANs have received widespread attention in image generation for their powerful generative capabilities. Sun et al. (2022) proposed a new GAN for generating high-resolution 3D medical images. This method reduced memory requirements and ensured anatomical consistency by using a hierarchical structure during training and by generating random sub-volumes of low- and high-resolution images simultaneously. The findings showed that this method outperformed existing techniques for generating medical images and had good potential for data augmentation and clinical feature extraction. Alkishri et al. (2024) designed a method grounded on high-frequency Fourier mode analysis to address the challenge of detecting deepfake images. This method analyzed differences in high-frequency features between real images and deep-network-generated

images to reveal systematic deficiencies in the latter. The findings indicated that this method could significantly reduce the detection rate of forged images, but the adversarial party could eliminate their feature fingerprints by modifying the GAN component, thereby increasing detection difficulty. Shi et al. (2023) propose a new GAN method to restore spatially varying bidirectional reflectance distribution functions from a single image when the lighting and geometry are unknown. This method adopts a single adversarial framework and introduces attention modules to guide the network to focus on details, while combining graph loss to avoid excessive attention to highlights. Experiments denoted that this method outperformed existing techniques on both public and real datasets, providing a better solution for restoring the appearance of objects from a single image. Shao et al. (2023) proposed a GAN-based method guided by dual thresholds for smart fault diagnosis in rotating machinery under speed fluctuations and limited sample sizes. This method was employed to produce high-quality infrared thermal imaging to facilitate fault diagnosis. This method uses a Loss Function (LF) that combines the Wasserstein distance with a gradient penalty, and an attention mechanism to extract global thermal correlation features. The findings denoted that this method was superior to other comparative methods in diagnosing rotating machinery faults with speed fluctuations and limited samples.

Significant progress has recently been made in deep learning-based image replacement methods. Liaoet al. (2023) proposed a system based on facial muscle movement characteristics to address the challenge of identifying compressed deepfake videos on social networks. This method extracted facial landmarks from consecutive frames, constructed facial muscle motion features, and used evidence theory to fuse multiple forensic knowledge to output detection results. The findings indicated that this method was superior to existing techniques for compressed video detection, and its performance in real social network environments was comparable to that of uncompressed videos. To address inefficient feature extraction and classification in facial recognition systems, Chaabane et al. (2022) proposed a facial recognition method based on statistical features and the Support Vector Machine (SVM) algorithm. The results demonstrated that this method achieved 99.37% recognition accuracy on 400 test images and was faster than the traditional SVM algorithm and other existing methods. This verifies the superiority of statistical features in the SVM algorithm for face recognition. In response to the problems of insufficient accuracy and high resource consumption in current video face replacement technologies, Wu and Wang (2025) proposed an automated face replacement method that integrates an improved MTCNN with a GAN. The research results show that the average value of the similarity index of replacement under different frame numbers is as high as 0.994, the mean peak signal-to-noise ratio is 35.65, and it performs optimally in the tests of structural similarity and state error, proving its excellent application effect in improving the naturalness of replacement and maintaining facial features

In summary, despite significant progress in face detection and replacement technology, there are still many limitations in high-resolution video processing, robustness in complex scenes, and optimization of computing resources. There is still room for improvement in generating image quality and replacing naturalness using existing methods. Therefore, this study proposes a method for replacing human faces in TV dramas and films based on the improved MTCNN and GAN, aiming to solve the following three problems: (1) How to improve the detection accuracy and speed of multi-scale faces in high-resolution TV videos. (2) How to maintain the structural similarity and original facial expression features of the generated images during the replacement process. (3) How to reduce the computational resource consumption of the generation model while ensuring the quality of the replacement.

The theoretical contributions of the research are as follows: The task decomposition strategy separates the detection and generation processes into independent sub-problems that can be optimized separately, thereby reducing the complexity of joint optimization; Deep separable convolution, feature pyramid, and residual blocks form a detection-generation collaborative mechanism through U-Net skip connections, filling the information gap in end-to-end substitution; This strategy is mapped into parallel processes, establishing the constraint relationship between detection accuracy and generation quality for later scheduling, and connecting with engineering management theory.

The innovation points of this method are reflected in three aspects: in the detection stage, deep separable convolution is used instead of standard convolution, reducing the parameter quantity by approximately 46% and increasing the detection speed by 43%; in the generation stage, U-Net skip connections and residual blocks are used instead of the traditional encoder-decoder structure, enabling the cross-level transfer of low-level textures and high-level semantics; the task decomposition strategy separates face replacement from the overall process into two independent optimization sub-problems of detection and generation, which is more flexible in the design of the loss function compared to end-to-end joint training.

2. Methods and Materials

The study proposes a face-replacement method for film and television drama videos based on an improved MTCNN and GAN. The complex video face replacement process is decomposed into multiple ordered subtasks using a task decomposition strategy, including video frame extraction, face detection and alignment, feature extraction, generation of replacement images by the generator, and discriminator evaluation of image authenticity, to achieve efficient and natural face replacement effects.

2.1. Face Detection in Film and Television Drama Videos Based on Task Decomposition Strategy and MTCNN

Film and television drama videos usually contain many frames, each of which may include multiple faces. To efficiently process video data, it is necessary to decompose the video into a sequence of single-frame images. Using a video decoder, a video can be decomposed into a series of frames, each with a two-dimensional image. The frame rate determines the number of frames extracted per second (Lei et al. 2024; Li et al. 2025). The total frame rate can be calculated from the video's total duration, as given in Eq. (1).

$$N = f \times T \tag{1}$$

In Eq. (1), N means the total frame rate of the video.f means the video's frame rate, the number of frames per second.T represents the total duration of the video. After extracting video frames, the face detection task can be further decomposed into two subtasks: face detection and face alignment. The goal of the former is to locate the facial region in each image frame, while face alignment ensures that the detected face has a uniform orientation and scale within the image, facilitating subsequent feature point localization and face replacement operations (AlAmir and AlGhamdi 2022). The task decomposition strategy is shown in Fig. 1.

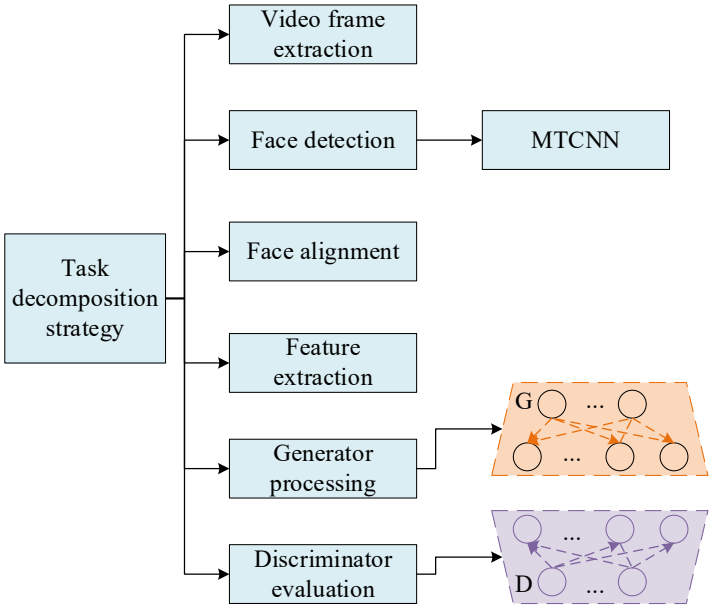


Fig. 1. Process diagram of task decomposition strategy

In Fig. 1, the task decomposition strategy breaks facial replacement into ordered subtasks: frame extraction, improved MTCNN-based detection and alignment, feature extraction, GAN-based generation and discrimination, and final integration back into video frames. To achieve accurate and efficient facial recognition, an optimized XYZ model is used as the main recognition technique in the study. The XYZ model is a hierarchical neural network structure that contains three sequentially executed phases: the primary recognition phase, the detail optimization phase, and the final output phase. Each stage assumes specific responsibilities to collaboratively achieve facial recognition and feature point determination (Zhou et al. 2022; Buragadda et al. 2022). The primary recognition stage is the first stage of MTCNN and is responsible for rapidly generating candidate face frames. Given an input image, the primary recognition stage generates a series of candidate frames using a sliding window, each containing positional information and a confidence level. The position information represents the coordinates of the upper left corner, width and height of the candidate box, while the confidence level denotes the probability that the candidate box contains a face. The output of the primary recognition stage is shown in Eq. (2).

$$\{B_i = (x_i, y_i, w_i, h_i, p_i) \mid i = 1,2, \dots, M\} \tag{2}$$

In Eq. (2), B_i indicates the ith frame in the face frame set obtained after the primary recognition stage processing. The x_i, y_i, w_i, and h_i respectively represent the top left and top right coordinates, width, and height of the face frame. Thep_i represents the confidence level of the face frame. The detail optimization stage is the second stage of MTCNN and is responsible for further screening and precise positioning of the candidate frames generated by the primary recognition stage. The detail optimization stage receives the candidate frames output from the primary recognition stage and performs classification and regression to remove misdetected candidate frames and adjust their positions. The output of the detail optimization phase is a set of more accurate candidate boxes, each containing updated position information and a confidence level. The output of the detail optimization phase is shown in Eq. (3).

$$\{B_r^{'} = (x_r^{'}, y_r^{'}, w_r^{'}, h_r^{'}, p_r^{'}) | r = 1,2, \dots, K\} \tag{3}$$

In Eq. (3), $B_r^{'}$ denotes the rth frame in the set of face frames obtained after processing; $x_r^{'}$, $y_r^{'}$, $w_r^{'}$, and $h_r^{'}$ denote the coordinates of the upper-left and upper-right corners of the face frames, as well as the width and height of the frames; and $p_r^{'}$ indicates the credibility score of the frame. The final output stage is the third stage of MTCNN and is responsible for precise positioning and feature point detection of the candidate frames produced by the detail optimization stage. The final output stage receives the candidate frames from the detail optimization stage and further adjusts their position and size while detecting the positions of key facial points, such as the eyes, nose and mouth. The final output stage produces a set of final face frames, each containing the final position information, confidence level and coordinates of key points. The final output is given in Eq. (4).

$$\{B_l^{''} = (x_l^{''}, y_l^{''}, w_l^{''}, h_l^{''}, p_l^{''}, \{P_{l,1}, P_{l,2}, \dots, P_{l,5}\}) \mid l = 1,2, \dots, L\} \tag{4}$$

In Eq. (4), $B_l^{''}$ represents the lth frame in the face frame set obtained after final output stage processing; $x_l'', y_l'', w_l'', h_l''$ represent the top left and top right coordinates, width, and height of the face frame, respectively; p_l'' represents the confidence level of the face frame; $\{P_{l,1}, P_{l,2}, \ldots, P_{l,5}\}$ represents the coordinates of key points in the face frame.

To improve MTCNN's performance in video face detection for film and television dramas, the study optimizes its network architecture. Depth separable convolution (DSC) replaces the regular convolution operation, effectively reducing the number of operations and the parameter scale. DSC splits the regular convolution into two parts: depth convolution and point-by-point convolution (PPC). Deep convolution processes each input channel separately, and then PPC uses a 1-by-1 convolution kernel to integrate the outputs across channels. This decomposition dramatically increases the detection speed while maintaining the detection accuracy (Zeng et al. 2022; Wang et al. 2024). The DSC's structure is denoted in Fig. 2.

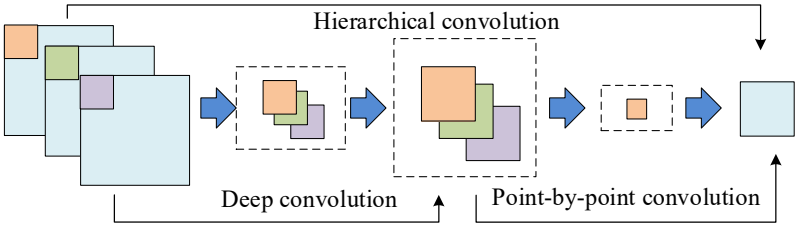


Fig. 2. Diagram of DSC structure

In Fig. 2, the hierarchical convolution technique splits the conventional convolution process into two simplified steps: deep convolution and PPC. In the deep convolution process, each input channel operates independently with its corresponding convolution kernel, avoiding redundant computation across channels and significantly reducing the parameter scale and computational burden. This architecture effectively minimizes the total number of parameters and the model's computational complexity, making the network structure more compact and efficient. PPC significantly improves the computational efficiency while maintaining the network performance. In addition, the study introduces a feature pyramid structure to better handle multi-scale face detection. The feature pyramid can effectively represent the multi-scale features of an image by extracting features at different scales, thereby reducing computational resource consumption (Fung et al. 2023; Zhang et al. 2022). The feature pyramid structure is denoted in Fig. 3.

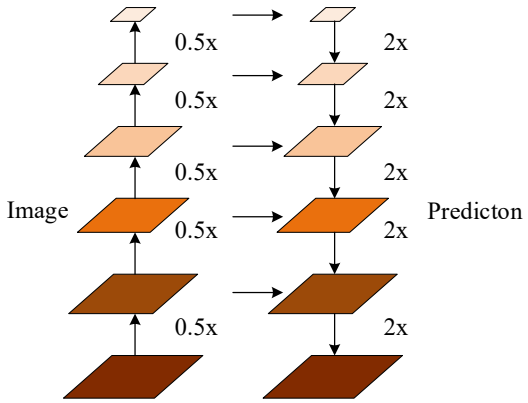


Fig. 3. Feature pyramid structure diagram

Depthwise separable convolution is integrated into each convolutional layer across the three stages of MTCNN, replacing the standard 3×3 convolution with a series of depthwise and pointwise convolutions. In this combination, the depthwise convolution performs spatial filtering on each channel, while the pointwise convolution performs channel fusion using 1×1 kernels. This replacement only affects the internal convolution operation and does not change the input and output dimensions and the number of feature maps of each stage. The feature pyramid structure is appended after the third stage of MTCNN. It takes the three intermediate feature maps of different scales (with scales of 1/8, 1/16, and 1/32 of the input image) from the third stage, uniformly reduces the number of channels through 1×1 convolution, and then performs 2x upsampling and element-wise addition fusion from top to bottom. The fused feature maps at each scale are independently fed into the final detection branch to predict large, medium, and small-scale faces, respectively.

The generator U-Net consists of three encoders and three decoders. The encoders consist of convolution, batch normalization, and ReLU, and the downsampling halves the size layer by layer; the decoders consist of transposed convolution, batch normalization, and ReLU, and receive features of the same scale from the corresponding encoder layers through skip connections. The discriminator is a stack of four convolutional layers, with the first two layers downsampling with a stride of 2 and the last two with a stride of 1. The output is a 1×1 convolution followed by a Sigmoid, producing an N×N×1 confidence map that emphasizes local discrimination. Compared with SimSwap, FaceShifter and DeepFaceLab, the proposed GAN differs in the following ways: the generator of SimSwap lacks multi-scale feature fusion; FaceShifter uses a global discriminator; the generator of DeepFaceLab lacks skip connections. The generator in this study is based on U-Net

and includes skip connections and residual blocks; the discriminator is a PatchGAN (with a 70×70 receptive field), which differs from the global discriminator.

2.2. Facial Replacement in Films and Television Videos Based on Improved GAN

After face recognition in movie and drama videos is complete, the next step is to perform efficient, realistic face replacement. The goal of face replacement is to replace the features of the original face image with those of another face while preserving the image's naturalness and realism. To achieve this purpose, an optimized adversarial network-based technique is used, and the structures of the generative and evaluation models are adjusted to meet the requirements of complex face replacement in film and drama videos. An adversarial network is an efficient image generation model that comprises two parts: a generator and an evaluator. The generative model aims to create real images, while the evaluation model is used to distinguish the generated images from real images (Wang et al. 2023; Preethi and Mamatha 2023). To improve image generation quality and the effectiveness of face replacement, the standard adversarial network architecture is enhanced, particularly in the structures of the generation and evaluation models. The improved GAN framework is shown in Fig. 4.

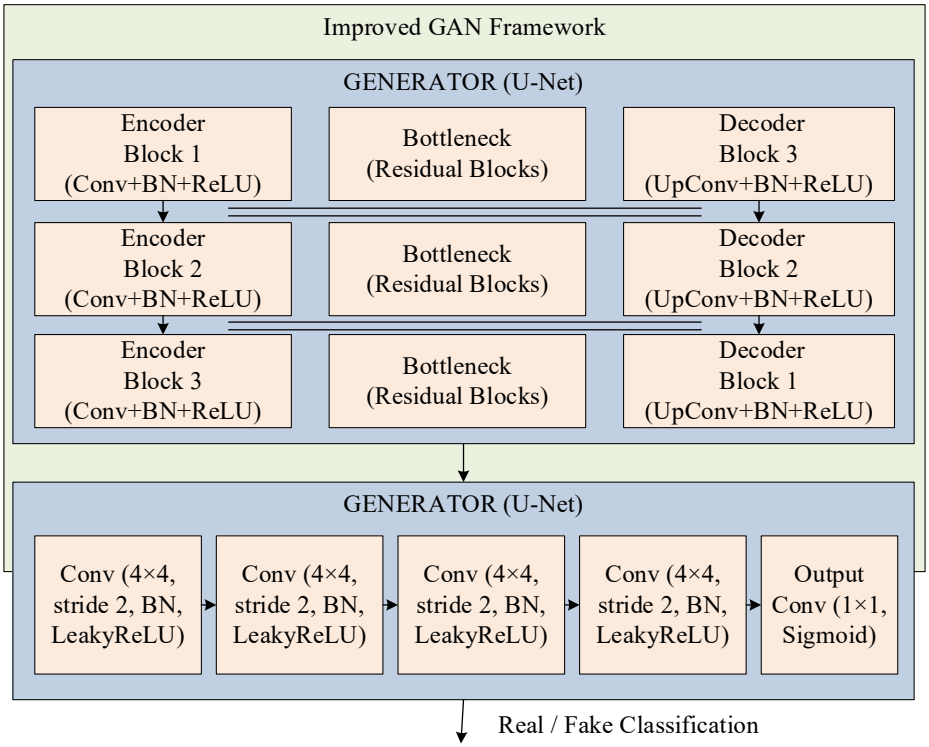


Fig. 4. Diagram of the improved GAN framework

The entire network adopts an end-to-end joint training method, in which the generator and the discriminator are optimized simultaneously, without a separate pre-training stage. The generator loss consists of three components: adversarial loss (LSGAN), L1 pixel reconstruction loss, and SSIM structural similarity loss, with weights set to 1, 100, and 5, respectively; the discriminator loss is a binary cross-entropy loss used to distinguish real from generated images. The total loss is the weighted sum of the two components' losses and is updated synchronously by backpropagation to update the parameters of both components. The LF of the generator and discriminator is shown in Eq. (5).

$$\begin{cases} L_A = -\mathbb{E}_{u\sim q(u)}[log\,J\,(K(u))] + \mu\mathbb{E}_{y\sim q(y)}[\|y - K(M(y))\|_1] \\ L_B = -\mathbb{E}_{y\sim q(y)}[log\,J\,(y)] - \mathbb{E}_{u\sim q(u)}[log(\,1 - J(K(u)))]n \end{cases} \tag{5}$$

In Eq. (5), L_A denotes the LF of the generative model; L_A denotes the LF of the discriminative model.$K(u)$ denotes the image generated by the generative model.$J(y)$ denotes the discriminative model's judgment on the real image.$J(K(u))$ denotes the judgment of the discriminative model on the generative image.μ denotes the weight parameter.$M(y)$ denotes the encoding of the real image by the encoder, and $\|y - K(M(y))\|_1$ denotes the distance between the generative image and the real image. The RB addresses the issues of gradient vanishing and exploding in deep networks by introducing "skip connections" to directly add input to output. Fig. 5 shows the enhanced RB structure.

In Fig. 5, the RB structure adds the input features and the features processed by the convolutional layer to form skip connections. The output of the RB is specifically shown in Eq. (6).

$$z = g(u) + u \tag{6}$$

In Eq. (6), z means the RB's output; $g(u)$ refers to the convolution operation in the RB; u denotes the input of the RB. The overall process of face replacement in film and television dramas includes video frame extraction, face detection, face alignment, feature extraction, generator-generated replacement image generation, and discriminator evaluation of image authenticity. The specific process is denoted in Fig. 6.

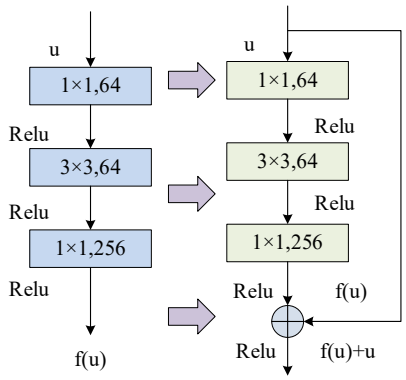


Fig. 5. Improved residual block structure

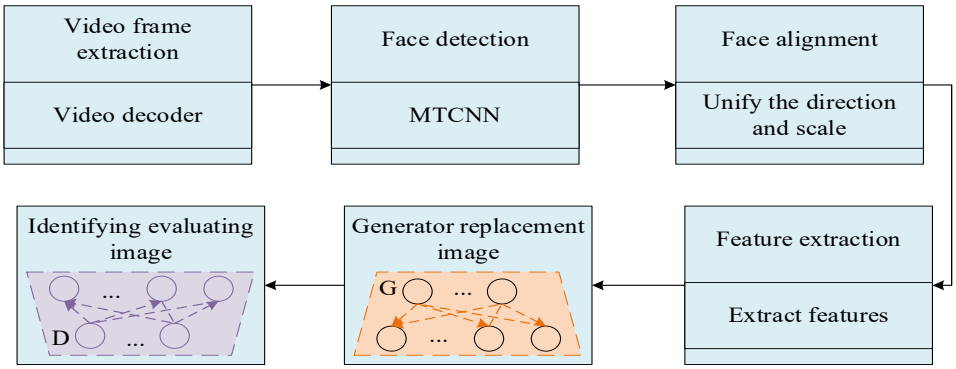


Fig. 6. General flowchart of video face replacement

In Fig. 6, individual frames are first extracted from film and television drama videos to form an image sequence. Secondly, the improved MTCNN algorithm is used to detect faces in these image frames and identify facial regions. Then, it aligns the detected facial regions to unify the face's direction and scale. Then, it extracts features from the aligned facial images. Afterward, the features of the source and target face images are fed into an improved generator to produce a replaced face image. Finally, a discriminator is used to evaluate the authenticity of the generated image and to provide feedback that optimizes the generator.

3. Results

A comprehensive evaluation of the raised method's effect was conducted, and the performance of the improved GAN on face-replacement tasks was verified through experiments. The face replacement effect of the improved GAN was analyzed to demonstrate the method's practical application in replacing faces in film and television drama videos.

3.1. The Algorithm Performance Evaluation of Improved GAN

The experimental parameter configuration follows the common practice for training generative adversarial networks. The Adam optimizer is adopted for its adaptive learning rate and momentum mechanisms, which are well-suited to non-stationary targets and high-dimensional parameter spaces. The learning rate is set at 0.0003, a relatively conservative value that helps stabilize the adversarial training dynamics between the generator and the discriminator. The batch size is set to 32 to strike a balance between gradient estimation accuracy and memory constraints, ensuring stable convergence under 128 GB of memory. The total number of training iterations is set to 2000, which is sufficient to make the loss curves flatten on both datasets. Data preprocessing includes histogram equalization to enhance contrast and median filtering to reduce noise, both of which are helpful for extracting more robust features in the subsequent detection and generation stages. The selection of datasets included the publicly available facial detection dataset Wide Face and the facial recognition dataset VGGFace. During training, the Wide Face dataset contains approximately 32,000 images, and the VGGFace dataset contains approximately 260,000 images. Both datasets were divided into training and validation sets in a 70% to 30% ratio, resulting in a training set of approximately 22,400 images and a validation set of approximately 9,600 images for Wide Face, and a training set of approximately 182,000 images and a validation set of approximately 78,000 images for VGGFace. To enhance the model's generalization ability, data augmentation was applied to the training data, including random rotation (±15°), random scaling (0.8 to 1.2×), and color jitter (brightness, contrast, and saturation, each ±10%). The settings for the experiment parameters are denoted in Table 1.

Table 1. Experimental parameter settings table

Parameter name	Parameter value
Processor model	Intel(R) Core(TM) i8-8900X CPU@3.50GHz
Graphics card model	NVIDIA GeForce GTX 2080
Memory size	128G
Training times	2000
Batch sample quantity	32
Optimizer	Adam
Learning rate	0.0003
Data preprocessing	Histogram equalization, median filtering smoothing

In Table 1, data preprocessing used histogram equalization to enhance image contrast and median filtering to denoise the image. The processor was Intel (R) Core (TM) i8-8900X CPU@3.50GHz, the graphics card was NVIDIA GeForce GTX 2080, and the memory was 128GB. To evaluate the effect of the enhanced MTCNN on face detection tasks, it was compared with traditional Haar feature cascade classifiers, Single Shot MultiBox Detector (SSD), and YOLOv5. The comparison findings are denoted in Table 2.

Table 2. Performance comparison of MTCNN with other detection algorithms

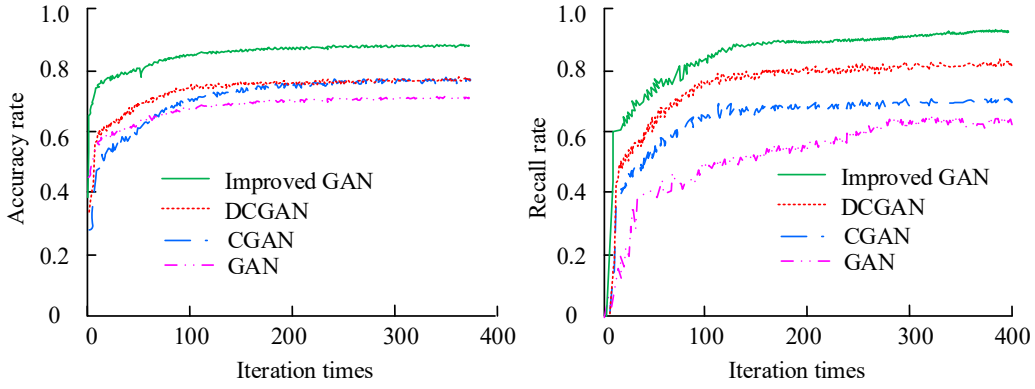
Detection algorithm	P	R	Mean average precision (mAP)	Detection time (ms/ frame)
Haar cascade	0.75	0.60	0.65	150
SSD	0.85	0.75	0.80	50
YOLOv5	0.88	0.80	0.84	30
MTCNN	0.90	0.85	0.88	40
Improved MTCNN	0.93	0.90	0.92	35

In Table 2, the improved MTCNN achieved a detection accuracy of 0.93, a recall rate of 0.90, an mAP of 0.92, and a detection time of 35ms/frame. The improved MTCNN achieved higher accuracy and efficiency for face detection in high-resolution videos and better met the needs of face replacement in film and television drama videos. The improved GAN was compared and analyzed with the original GAN, Conditional GAN (CGAN), and Deep Convolutional GAN (DCGAN). The precision (P), recall (R), and F1 values of several algorithms are denoted in Fig. 7.

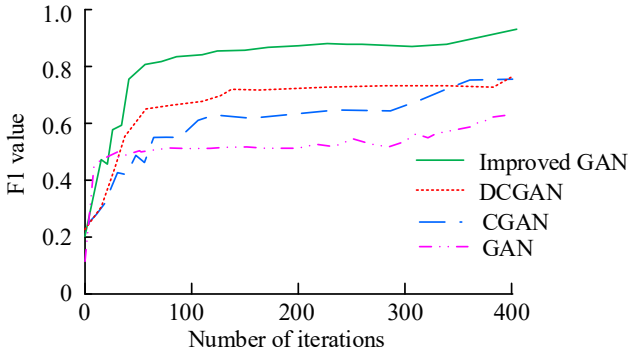
In Fig. 7(a), the accuracy of the enhanced GAN approached 0.9 after about 100 iterations; the accuracy of DCGAN reached about 0.85 after about 150 iterations; the accuracy of CGAN reached about 0.8 after about 200 iterations; and the accuracy of GAN was about 0.75 after about 300 iterations. In Fig. 7(b), the recall rate of the improved GAN rapidly increased to nearly 1.95 after about 50 iterations. The recall rate of DCGAN was about 0.9 after about 100 iterations, the recall rate of CGAN was about 0.85 after about 150 iterations, and the recall rate of GAN was about 0.8 after about 300 iterations. In Fig. 7(c), the improved GAN achieved an F1 value of approximately 0.85 after about 50 iterations and maintained stable growth to 0.9 after 400 iterations. This indicated that the improved GAN achieved faster convergence and higher overall performance on face replacement tasks, demonstrating its superiority and practicality. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of several algorithms are shown in Fig. 8.

In Fig. 8(a), the improved GAN maintained the lowest MAE value throughout the entire process, decreasing from approximately 0.11 to around 0.09, demonstrating its superiority in generating image quality. In Fig. 8(b), the improved GAN also demonstrated optimal performance, with RMSE values decreasing from 0.11 to around 0.10, DCGAN decreasing from 0.12 to around 0.11, CGAN decreasing from 0.14 to around 0.11, and GAN decreasing from 0.13 to around 0.11. These results further confirmed the effectiveness of the improved GAN in reducing errors and improving the quality of the image generated. The LFs of several algorithms are shown in Fig. 9.

Fig. 9(a) shows the variation in LF values across iteration numbers for four GAN algorithms on the Wide Face dataset. The LF value of the improved GAN decreased the fastest, from approximately 10^2 to around 10^-4, demonstrating its high efficiency and superior performance during training. Fig. 9(b) showcases the variation of LF values with iteration times for four algorithms on the VGGGFace dataset. The improved GAN also demonstrated the fastest convergence speed, with the LF value decreasing from approximately 10 to around 10-6. The improved GAN could quickly reduce the LF value on different data sets, demonstrating its stability and effectiveness during the training process.

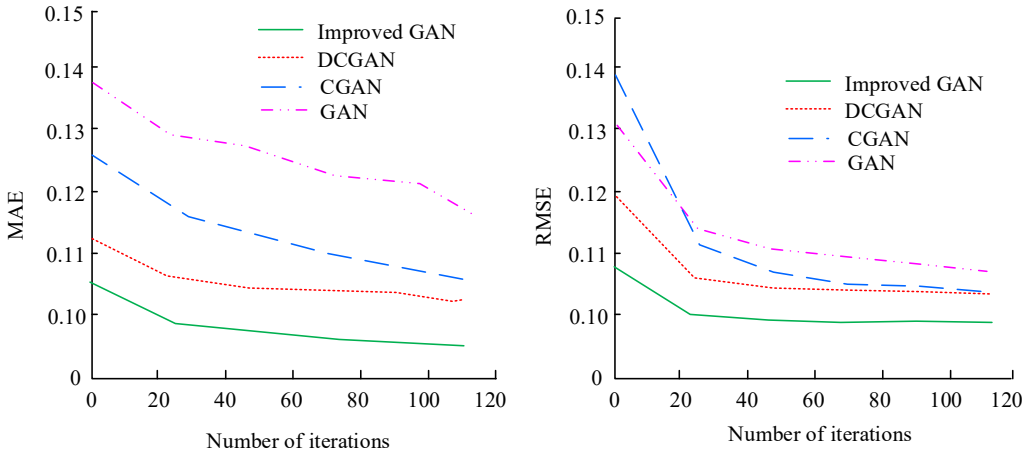


(a) The accuracy rates of the four algorithms (b) Recall rates of the four algorithms



(c) The F1 values of the four algorithms

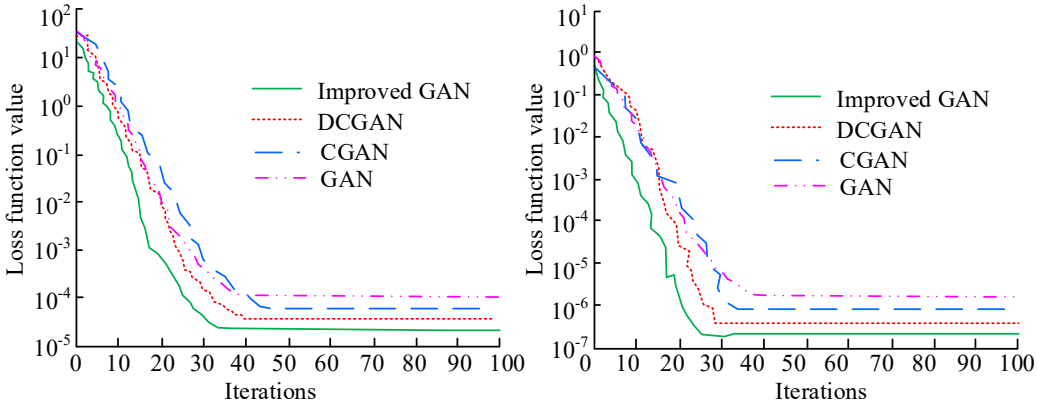
Fig. 7. Precision, recall and F1 value of several algorithms



(a) MAE (b) RMSE

Fig. 8. MAE and RMSE of several algorithms

To verify the effectiveness of each improvement module, ablation experiments were set up. The depth separable convolution, feature pyramid structure, and residual block were sequentially removed, and the detection accuracy and PSNR of the generated images were compared on the Wide Face and VGGFace datasets. After removing the depth-separable convolution, the detection time increased from 35ms/frame to 62ms/frame, and the mAP decreased from 0.92 to 0.87, indicating that this module primarily improves detection efficiency while maintaining accuracy. After removing the feature pyramid structure, the recall rate for small-scale faces (less than 32×32 pixels) decreased from 0.88 to 0.76, and the mAP decreased to 0.84, indicating that this module has a significant contribution to multi-scale face detection. After removing the residual block, the PSNR of the generated images decreased from 25.0dB to 23.6dB, and the SSIM decreased from 0.98 to 0.94, indicating that the residual block has a prominent role in improving the quality of the generated images. When all three modules were removed simultaneously, the mAP decreased to 0.79, and the PSNR decreased to 22.8dB, further demonstrating that the collaborative effect of each module is crucial for overall performance.

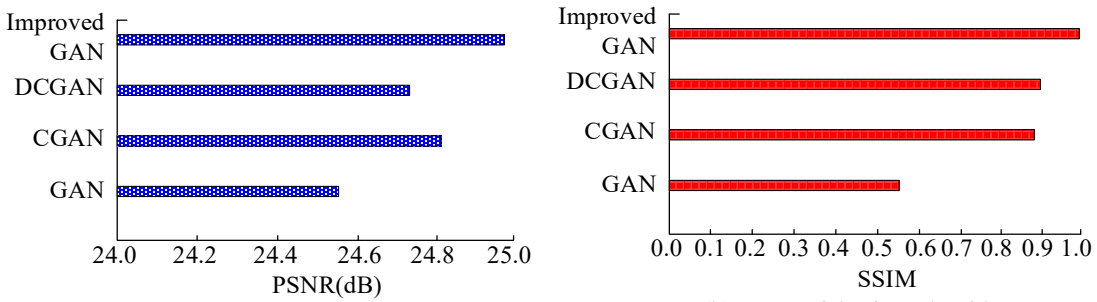


(a) Dataset of Wider Face (b) Dataset of VGGGFace

Fig. 9. Loss functions of several algorithms

3.2. Analysis of Face Replacement Effect Based on Improved GAN

To evaluate the effect of the raised face replacement method, multiple publicly available facial datasets from film and television drama videos were selected for experimentation. These datasets contained diverse facial samples across races, ages, genders, and expressions to ensure the broad applicability of the experimental results. The Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) of several algorithms are denoted in Fig. 10.



(a) PSNR of the four algorithms (b) SSIM of the four algorithms

Fig. 10. PSNR and SSIM of several algorithms

In Fig. 10(a), the improved GAN performed best in terms of PSNR, approaching 25.0dB, followed by CGAN with a PSNR of about 24.8dB. The PSNR value of DCGAN was lower than that of CGAN, about 24.7dB, while the PSNR value of GAN was 24.3dB. This indicated that the improved GAN had a significant advantage in generating PSNR images and could produce higher PSNR images that more closely resemble real images. In Fig. 10(b), the improved GAN also achieved the highest SSIM values, approaching 1.0, indicating that its generated image structure was most similar to the real image. The improved GAN also had significant advantages in maintaining SSIM in generated images and better preserving the structural information of the original image. The actual application results of several algorithms for face replacement in film and television dramas are shown in Table 3.

In Table 3, the improved GAN performed the best in terms of user satisfaction, visual quality, and emotional retention, with values of 4.65 dB, 38 dB, and 85%, respectively, indicating that its generated facial images were more appreciated by users and could better preserve the original emotions during the replacement process. Although the improved GAN had a slightly longer generation time of 120 milliseconds, it performed well in terms of resource consumption at only 15% and showed the best robustness at 90%, indicating that it could operate stably under different conditions. In terms of diversity, the improved GAN also scored the highest, at 8, indicating it generated a wider variety of facial images. Both GAN and DCGAN with improved resolution could reach 2048x2048, meeting high-resolution requirements. The quantitative comparison of the facial replacement effects using different methods is shown in Fig. 11.

In Fig. 11, the three indicators of the Improved GAN are all the best. The Precision is approximately 0.91, the Recall is approximately 0.95, and the F1-score is approximately 0.92. These values are significantly higher than those of other methods. This indicates that the improved GAN achieves significantly higher classification accuracy, recall, and overall performance in the face replacement task than the traditional GAN model, thereby verifying its effectiveness.

Table 3. Practical application effects of several algorithms in face replacement in film and television drama videos

Indicator	Improved GAN	GAN	CGAN	DCGAN
User satisfaction	4.65	4.2	4.4	4.3
Generation time (milliseconds)	120	150	100	80
Replacement accuracy	94.5%	92.0%	93.5%	91.0%
Visual quality (dB)	38	36	37	35
Emotional retention	85%	80%	82%	78%
Computing resource consumption (%)	15	20	18	25
Robustness	90%	70%	75%	60%
Diversity	8	7	8	6
Resolution (pixels)	2048x2048	1024x1024	2048x2048	1080p

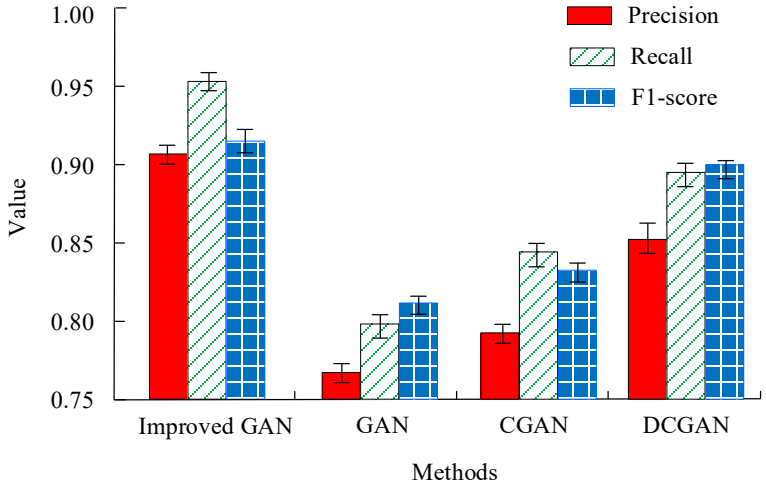


Fig. 11. Quantitative comparison of facial replacement effects using different methods

To verify the generalization ability of the proposed method in cross-dataset scenarios, the Flickr-Faces-HQ (FFHQ) dataset (approximately 70,000 images) and real TV drama video clips (containing continuous frame sequences under different lighting, posture, and occlusion conditions) were selected for testing. On the FFHQ dataset, the mAP of the improved face detection method was 0.90, and the PSNR of the generated images was 24.6 dB. Compared with the results on the Wide Face dataset, it decreased by 0.02 dB and 0.4 dB, respectively, indicating good cross-domain generalization performance. In the real video clip test, three sets of conditions were set: normal lighting (illuminance $\geq$ 300 lux), low lighting (illuminance $\leq$ 50 lux), and local occlusion (mask occlusion area $\geq$ 30%). The detection accuracy under normal lighting was 0.92, decreased to 0.85 under low lighting, and was 0.81 under occlusion conditions; the SSIM of the generated images under the three conditions was 0.97, 0.93, and 0.90, respectively, and the emotion retention rates were 84%, 78%, and 72%, respectively. The detection times under the three conditions were 38ms/frame, 42ms/frame, and 45ms/frame, respectively, and the resource consumption was 16%, 19%, and 22%, respectively. The overall performance degradation was controllable, indicating that this method is robust in complex real-world scenarios.

On the Wide Face (approximately 32,000 images) and VGGFace (approximately 260,000 images) benchmark datasets, the mAP values are 0.92 and 0.90, respectively, and the PSNR values are 25.0 dB and 24.7 dB, respectively. The difference between the two is less than 3%. For proprietary film and television data (approximately 120 minutes, approximately 180,000 frames), the mAP is 0.89, and the PSNR is 24.2 dB, with the decline kept within 5%. On FFHQ (70,000 images), the mAP is 0.90, and the PSNR is 24.6 dB. The scale range for the four sets of data spans 32,000 to 260,000 images, and the resolutions vary by specification. The fluctuation range of the detection accuracy and generation quality indicators does not exceed 5%, indicating that this method has stable performance across different data set scales and sources and good generalizability.

3.3.Engineering Management

At the production management level, this method has reduced the post-production time for single-camera face replacement from an average of 4.5 hours per person to approximately 18 minutes. The number of manual intervention points has been reduced from 12 to 3, and the processing throughput has reached approximately 3,200 frames per hour, which is about 15 times higher than the traditional frame-by-frame manual replacement process. This significantly reduces the manpower and time costs in film and television post-production.

At the deployment and integration level, this system is recommended for deployment on servers equipped with NVIDIA Tesla V100 GPUs (32GB of video memory) and 128GB of memory. The software environment is Ubuntu 18.04, CUDA

11.0 and PyTorch 1.8. It can interface with the standardized frame-sequence import interfaces of mainstream editing software (Adobe Premiere Pro, DaVinci Resolve). Operators need to have basic Python programming skills and an understanding of deep learning inference processes. There is no need for model re-training or parameter tuning.

In the post-production scheduling, a 15% GPU resource occupancy rate enables a single server to manage 6 replacement tasks. The 3200 frames-per-hour throughput capacity enables completion of approximately 40 standard shots (calculated at 90 seconds per shot) of replacement work within a single day. This transformation shifts the compositing process from a sequential waiting state after editing to a pre-intervention parallel process. This efficiency enables the production team to simultaneously initiate facial replacement during the pre-editing stage, reducing the overall post-production cycle by approximately 35%. It provides a quantitative basis for scheduling optimization and resource allocation.

Theoretically, the positive correlation between detection accuracy (mAP = 0.92) and generation quality (PSNR = 25.0 dB) supports the effectiveness of the task decoupling strategy: independently optimizing the detection and generation sub-problems can reduce the difficulty of joint optimization. Improved detection accuracy directly enhances the quality of the input features during the generation stage. In practice, the time for single-lens replacement has been reduced to 18 minutes, and the rework rate has dropped below 4%, transforming facial replacement from a sequential step after editing into a parallel process that can be intervened in advance, thereby directly aligning with resource scheduling decisions in production management.

4. Conclusion

The research aimed to address the problems of insufficient detection accuracy, poor image quality, and high computational resource consumption in face replacement in film and television dramas. A method based on improved MTCNN and GAN was proposed to achieve efficient, natural, and realistic video face replacement effects. The research results showed that the improved MTCNN algorithm performed well in the face detection stage, with a detection accuracy of 0.93 and a detection time of 35ms/frame, improving the efficiency and quality of face detection. In the face-replacement stage, the improved GAN architecture outperformed existing technologies across multiple key performance indicators. Its PSNR reached 25.0dB, and SSIM approached 1.0, demonstrating extremely high image quality and SSIM. In addition, the user satisfaction score was 4.65, and the emotional retention rate was 85%, indicating that this method has significant merit in maintaining original emotions and visual quality. Although the generation time was slightly longer at 120 milliseconds, computational resource consumption was only 15%, demonstrating efficient resource utilization and robustness. The proposed method, based on an improved MTCNN and GAN, can enhance face detection accuracy and improve image quality while reducing computational resource consumption. However, this method still has three limitations. First, robustness to extreme lighting and complex backgrounds needs further verification; second, training relies on a large amount of labeled face data, and the model's transferability in low-resource scenarios is limited. Third, the current architecture is designed for single-camera, single-person replacement. When extending to simultaneous replacement of multiple people or real-time streaming scenarios, the computational load and memory usage increase linearly, and the scalability needs to be optimized. Future research will focus on optimizing algorithm efficiency, further reducing generation time, and exploring potential applications in more complex scenarios.

Author Contributions

Ludongdong Shan contributed to conceptualization, methodology, software, validation, analysis, data collection and draft preparation. Li Zhou contributed to methodology, data collection, manuscript editing, supervision, project administration, and funding acquisition. Yuelin Du contributed to data collection, manuscript editing, visualization and supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

The research is supported by Hunan Province Social Science Achievements Evaluation Committee general projects (project number: XSP2023JYC172).

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors confirm that no AI tools were used in the preparation of this manuscript.

References

AlAmir, M. and AlGhamdi, M. (2022). The Role of Generative Adversarial Networks in Medical Image Analysis: An In-Depth Survey. *ACM Computing Surveys,* 55(5), 1-36.

Alkishr,i W., Widyarto, S., and Yousif, J. H. (2024). Evaluating the Effectiveness of a Gan Fingerprint Removal Approach in Fooling Deepfake Face Detection. *Journal of Internet Services and Information Security (JISIS),* 14(1), 85-103.

Buragadda, S., Rani, K. S., Vasantha, S. V., and Chakravarthi, M. K. (2022). Hcugan: Hybrid cyclic unetgan for generating augmented synthetic images of chest x-ray images for multi classification of lung diseases. *International Journal of Engineering Trends and Technology,* 70(2), 229-238.

Chaabane, S. B., Hijji, M., Harrab,i R., and Seddik, H. (2022). Face recognition based on statistical features and SVM classifier. *Multimedia Tools and Applications,* 2022, 81(6), 8767-8784.

Deuze, M. and Beckett, C. (2022). Imagination, algorithms and news: Developing AI literacy for journalism. *Digital Journalism,* 10(10), 1913-1918.

Fung, B. S., Chan, W. H., Lo, I. M., and Tsang, D. H. (2023). Freshwater microscopic algae detection based on deep neural network with GAN-based augmentation for imbalanced algal data. *ACS ES&T Water,* 4(3), 982-990.

Huang, G. and Jafari, A. H. (2023). Enhanced balancing GAN: Minority-class image generation. *Neural computing and applications*, 35(7), 5145-5154.

Lei, M. F., Zhang, Y. B., Deng, E., Ni, Y. Q., Xiao, Y. Z., Zhang, Y., and Zhang, J. J. (2024). Intelligent recognition of joints and fissures in tunnel faces using an improved mask region-based convolutional neural network algorithm. *Computer-Aided Civil and Infrastructure Engineering,* 39(8), 1123-1142.

Li W, Liu D, Li Y, Hou M, Liu J, Zhao Z, and Deng W. (2025). Fault diagnosis using variational autoencoder GAN and focal loss CNN under unbalanced data. Structural Health Monitoring, 24(3), 1859-1872.

Liao X, Wang Y, Wang T, Hu J, and Wu X. (2023). FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(12), 7236-7251.

Preethi, P. and Mamatha, H. R. (2023). Region-based convolutional neural network for segmenting text in epigraphical images. *Artificial Intelligence and Applications,* 1(2), 103-111.

Rowe. M., Nicholls, D. A., and Shaw, J. (2022). How to replace a physiotherapist: artificial intelligence and the redistribution of expertise. *Physiotherapy Theory and Practice,* 38(13), 2275-2283.

Shao, H., Li, W., Cai, B., Wan, J., Xiao, Y., and Yan, S. (2023). Dual-threshold attention-guided GAN and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation. *IEEE Transactions on Industrial Informatics,* 19(9), 9933-9942.

Shi, Z., Lin, X., and Song, Y. (2023). An attention-embedded GAN for SVBRDF recovery from a single image. *Computational Visual Media,* 9(3), 551-561.

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., and Batmanghelich, K. (2022). Hierarchical amortized GAN for 3D high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26(8), 3966-3975.

Wang, C., Sun, B., Du, K. J., Li,J. Y., Zhan, Z. H., Jeon, S. W., and Zhang, J. (2023). A novel evolutionary algorithm with column and sub-block local search for sudoku puzzles. *IEEE Transactions on Games*, 16(1), 162-172.

Wang, L., Yu, Q., Li, X., Zeng, H., Zhang, H., and Gao, H. (2024). A CBAM-GAN-based method for super-resolution reconstruction of remote sensing image. *IET Image Processing,* 18(2), 548-560.

Wu, Y. and Wang, X. (2025). MTCNN-UGAN: A Self-Attention Enhanced Face Replacement Pipeline for Film and Television Video Frames. *Informatica,* 49(5).

Xia, W., Zhang, Y., Yang, Y., Xue, J. H., Zhou, B., and Yang, M. H. (2022). Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence,* 45(3), 3121-3138.

Zeng, Y., Fu, J., Chao, H., and Guo, B. (2022). Aggregated contextual transformations for high-resolution image inpainting. *IEEE transactions on visualization and computer graphics,* 29(7), 3266-3280.

Zhang, K., Hu, H., Philbrick, K., Conte, G. M., Sobek, J. D., Rouzrokh, P., and Erickson, B. J. (2022). SOUP-GAN: Super-resolution MRI using generative adversarial networks. *Tomography,* 8(2), 905-919.

Zhou, Z., Li, Y., Li, J., Yu, K., Kou, G., Wang, M., and Gupta, B. B. (2022). Gan-siamese network for cross-domain vehicle re-identification in intelligent transport systems. *IEEE transactions on network science and engineering,* 10(5), 2779-2790.

Ludongdong Shan is an associate professor at Hunan University of Media and Communications, an AIGC application engineer (senior), a young outstanding teacher in Hunan Province, and a young domestic visiting scholar. Her research interests include the application of artificial intelligence in communication, new media communication theory, ideological and political teaching reform in courses, blended teaching models, course evaluation system construction, cultural tourism integration, and traditional culture construction.

Li Zhou earned her master's degree in Musicology (2008) from Hebei Normal University. She is currently an associate professor and head of the Indoor Choir at the School of Art, Shijiazhuang Institute of Technology. She is a member of the Hebei Musicians Association, a senior "Double-Qualified" teacher of vocational education in Hebei Province, a senior Orff music instructor, and a mental health instructor. She has been teaching for many years, has extensive experience and has been rated an "Excellent Teacher" by the college. She has repeatedly guided students to win awards in national and provincial competitions and has been honored as an "Excellent Instructor". She has published two monographs and compiled one textbook as the chief editor. In addition, she has published more than twenty papers in national and provincial journals. Her areas of interest include vocal singing, piano impromptu accompaniment, chorus and conducting, and phonation.

Yuelin Du obtained her Master of Arts degree from the Academy of Director in Kraków, Poland, in 2018. Currently, she serves as the teaching and research office director in the School of Art, Shijiazhuang Institute of Technology. She is an intermediate-level, double-qualified teacher of vocational education in Hebei Province, a member of the Hebei Musicians Association and the Shijiazhuang Musicians Association, and the Deputy Secretary-General of the Music Education Committee of the Shijiazhuang Musicians Association. She holds the College Teacher Qualification Certificate and the Enterprise Human Resource Manager Professional Qualification Certificate.