Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/135/2025.21086

Facial Deepfake Detection Based on Spactial Image Features

Xiuqi Cao[1,a,*]

[1]*School of Software Engineering, Chongqing University of Posts and Telecommunications, Nanshan Street, Chongqing, China*
a. 2022214128@stu.cqupt.edu.cn
**corresponding author*

Abstract: With the continuous development of deepfake technology, led by "ai in face" many depth of forged related technology began to fill in the line of sight of the public, it's not only in the entertainment industry provides extremely convenient function, but also brought including but not limited to information fraud many malignant events, so the depth of forged detection technology is particularly important. This article begins by reviewing the history of deep forgery, And the most common reproduction, editing, replacement and synthesis, Following up with examples and comparisons of data sets emerging in recent years, Then, the face depth forgery detection technology based on spatial image features based on attention network, based on autoencoder, based on vision Transformer and based on data enhancement of four directions are roughly introduced, Finally, the paper explores some of the problems in the field, and with the existing challenges, Some future development directions are proposed, The full text is also summarized.

Keywords: Facial Deepfake, Deepfake Detection, Information Security, Detection Technology

1. Introduction

Deepfake, a term derived from the combination of "deep," meaning deep learning, and "fake," meaning deception, represents the fusion of deep learning and forgery. It refers to a technology that uses deep learning techniques to manipulate videos, images, and audio. The technology emerged in the United States in 2017, when a user with the ID "deepfakes" posted a pornographic video of a female celebrity, created using deepfake technology for face-swapping. This incident sparked widespread discussion across various sectors of society, and since then, deepfakes have gradually entered the public's consciousness.

Today, deepfakes are widely present in people's daily lives, providing significant momentum for the development of the entertainment industry. However, with the continuous iteration and development of this technology, several issues have gradually emerged. These include, but are not limited to, fake face-swapping videos and images that may cause social and public opinion issues for the individuals involved, as well as widespread economic damage to citizens caused by imitation and deception using deepfake technology.

On January 15, 2024, a fake advertisement featuring an AI-generated Taylor Swift promoting cookware spread across social media platforms like Facebook. In the ad, Swift claimed to be offering a "free cookware set" to victims. However, when victims were redirected to a fake website, they were

DOI: 10.54254/2755-2721/135/2025.21086

asked to pay a $9.96 shipping fee. Despite the claim of a free giveaway, the cookware was not actually delivered.

It is evident that the information security crisis brought about by the rapid development of deepfakes is expanding. Against this backdrop, both academia and industry have begun exploring detection technologies to combat deepfakes.

With the development of deep learning [1], not only has the technology for forgery advanced, but detection technologies have also emerged in parallel. The most notable technologies in this area are Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [2,3]. These two techniques are complementary, each advancing the development of the other. For instance, Wadajo et al. combined CNNs with Vision Transformers (ViTs) to manage feature learning and input processing, used attention mechanisms for classification, and trained their model using the challenging DFDC dataset, ultimately achieving competitive results[4].

The first section of this paper provides a general introduction to the relevant technologies for detecting facial deepfakes. The second section presents and compares existing deepfake datasets. The third section systematically categorizes facial deepfake detection technologies based on spatial image features and provides a more detailed introduction to each. The fourth section enumerates the significant challenges currently faced in facial deepfake detection and proposes potential directions for future development. Finally, the fifth section concludes with a summary of the work presented in the paper.

2. Overview of facial deepfake

The two most common techniques in deepfake technology are autoencoders and Generative Adversarial Networks (GANs). Based on the classification of facial deepfakes, they can be further divided into four categories: reproduction, editing, replacement, and synthesis [5].

2.1. Reproduction

Reproduction refers to manipulating facial expressions and actions while maintaining the original identity and appearance, thereby generating fabricated dynamic visuals [6]. This technique has many positive applications, such as in film language translation, where the character's lip movements can be adjusted according to the dialogue, as well as in the creation of vivid animated products.

The accuracy of early graphics-based methods for creating images and videos largely depended on the precision of reproducing the facial model during the process. Wu et al. initially proposed a method based on facial key points for reproduction [7]. Kim et al. also introduced a method based on 3D models to reproduce dynamic videos of objects [8]. This approach involved creating a 3D model of the subject, predicting lighting, facial expressions, and other motion parameters, and finally integrating these into a temporal module to generate the complete rendered head image.

However, Thies et al. proposed a delayed neural rendering framework that simultaneously optimizes both the neural texture and rendering network, reducing the misalignment of the model, and achieving more ideal results [9].

2.2. Editing

Facial attribute editing refers to the manipulation of specific facial features, such as adjusting the eyes, nose, age, skin tone, and other attributes. It can be considered a specific application of unpaired image-to-image translation, where multiple images share common underlying features.

The challenge of this technology lies in how to edit specific facial attributes while keeping the unrelated attributes unchanged. This is difficult because facial images have strict graphical constraints,

DOI: 10.54254/2755-2721/135/2025.21086

and there are correlations between different attributes. As a result, editing any single attribute can lead to unintended changes in other attributes.

In 2018, Choi et al. proposed the StarGAN model, which simultaneously edits multiple facial attributes by jointly training and sharing the generator [10]. However, due to the excessive constraints of one-hot encoding, the model struggled with balancing the target attributes and non-target attributes, leading to the loss of texture details.

In 2024, to address the issue of inter-attribute coupling, Gu et al. introduced the GP-GAN method [11]. By repeatedly encoding images, translating the encodings, and having a discriminator judge the results, the method optimizes both the generator and discriminator to produce higher-quality images, achieving more ideal outcomes.

2.3. Replacement

Replacement refers to using deepfake technology to achieve identity swapping on a target face, commonly known as "face-swapping." In the early stages, this algorithm typically used autoencoders to achieve the face-swapping effect. However, due to limitations in the structure of autoencoders and training data, the generated fake images often had low resolution, and the texture details needed improvement.

Early methods mostly focused on interacting with global features while neglecting the importance of modeling local facial features, such as eyebrows, wrinkles, and other details, which limited the model's ability to maintain identity consistency. To ensure such consistency, Xu et al. proposed the RAFSwap network, which employs a local-global approach to generate high-resolution faces with consistent identity, making the results appear more natural [12].

2.4. Synthesis

Face synthesis, or face generation, is fundamentally different from the previously mentioned techniques. This technology does not rely on real or pre-input faces but instead starts with other attributes to create a completely fabricated face that does not exist in reality and has no prototype.

Since the introduction of GANs in 2014, the development of face synthesis technology has progressed rapidly. However, it was soon realized that early GAN models produced images with low resolution and often generated artifacts. To address this issue, Karras et al. proposed the Progressive GAN (ProGAN) network, which begins with a low-resolution generator and progressively increases the resolution layer by layer, ensuring stable high-resolution generation [13].

However, considering that ProGAN lacked the ability to generate specific attributes and styles, Karras et al. redesigned the normalization method to eliminate tear-drop artifacts. They also replaced the progressive strategy with a large-capacity model, solving issues such as the immobility of areas like the teeth [14].

3. Face Deepfake Data set

In the field of deep learning, datasets have always been indispensable parts, with the function of training, testing, and evaluating model performance. A portion of the common data sets are enumerated here, as shown in Table 1.

Table 1: dataset

data set	Real material	Forge material	distinguishing feature
UADFV	49	49	The data were small and generally of low quality
DF-TINIT	320	—	

DOI: 10.54254/2755-2721/135/2025.21086

Table 1: (continued).

Faceforensics++	1000	5000+	Many kinds, high quality
DFDC	23564	104500	Large scale, more types, but the motion produces artifacts
FFIW10K	10000	10000	Labor costs are low, and often have stability
DeakeAVMiT	—	6480	Contains the audio, which is more realistic
DeepFaceGen	776990	773812	Large scale, much technology used, and a good balance for different races

Early datasets were similar to UADFV [15] and DF-TIMIT [16], where UADFV contains 49 real face videos and 49 fake videos; the DF-TIMIT dataset contains 320 real videos and contains their own HD and low-definition versions. Their data are relatively small and simple. This rough and low-quality dataset is mostly used for the training of the underlying model.

In 2019, the dataset Faceforensics + +, created by Rossler et al., solved the problem of fewer types of forgery and generally low quality of the dataset at that time[17]. It downloaded 1,000 real-life original videos used as material on YouTube, and generated more than 5,000 faces using Five advanced methods of face forgery, FSGAN, DeepFakes, Face2Face, NeuralTextures and FaceShifter.

In 2020, facebook and Microsoft released the DFDC [18] data set, which is a large public data set, containing 23,564 real source videos and 104,500 fake videos, containing a variety of fake face types, and has a diverse background, can be used in a variety of occasions, but has the disadvantage that character movement sometimes produces artifacts.

In 2021, Zhou et al. proposed FFIW10K data set with 10,000 real original videos and equal amount of high-quality face fake videos, containing three faces per frame[19]. The operation process is fully automatic, highly scalable, and the labor cost is lower than previous models. However, because existing data sets usually iterate on them through the development of forgery techniques, and fake videos are usually consistent under normal stability, it is difficult to update the data.

In 2023, Yang et al. proposed DefakeAVMiT, which, compared to other traditional data sets, also contains the falsification of lsah corresponding audio, providing a relatively more realistic environment[20] .

In 2024, Bei et al. proposed the DeepFaceGen face forgery detection dataset, which contains 463,583 real face images and 313,407 real videos, and 350,264 fake images and 424,548 fake videos, using 34 existing mainstream technologies, not only covering a wide range of face data, but also ensures a balance between different races and backgrounds, providing a solid foundation for the evaluation and iterative development of facial forgery detection techniques[21].

4. Detection methods based on the spatial image features

Based on spatial features of image detection method is in many methods more traditional and more direct and effective methods, neural network such as XceptionNet[22], EfficientNet[23] are able to effectively extract the image deep features, but although they are in the face of the existing data set object effect is good, but in the face of unknown fake data set, the network will appear bad effect. Therefore, many experts and scholars have proposed many new detection methods[5] based on image spatial features for this phenomenon.

4.1. Detection method based on the attention network

In deep learning, the attention mechanism has always been a component technique that enables the trained model to focus on a specified part. Stehouwer On the basis of XceptionNet and VGG

DOI: 10.54254/2755-2721/135/2025.21086

16introduced the attention network mechanism to process and improve the feature map of the classification task[24] . In Figure 1, the attention map specifically shows the information area, and was used to further improve the judgment of true and false faces, which significantly improved the detection effect.

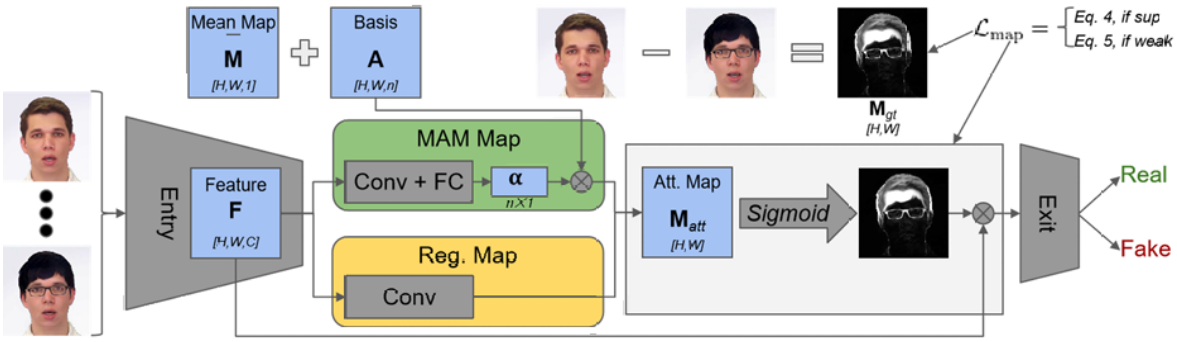

Figure 1: The method flow proposed by Stehouwer et al[24]

Duan proposed a dual-stream extraction and multi-scale enhancement multi-feature fusion network [25], designed a dual-stream feature extraction module to extract more feature information, then put forward the multi-scale feature enhancement module, let the network model can analyze the current information from different angles, and finally use the attention network training detection module to analyze the input image and the original image overlap, so as to determine the tampered area. Because the method learning is not limited to the general use of a specific feature, very high detection accuracy and performance are achieved. The attention network makes the model have the ability to acquire more subtle features for specific parts, which makes the model have more generalization.

4.2. Detection method based on the autoencoder

The autoencoder is divided into two parts, the encoder and the decoder, as shown in figure 2. The encoder has the function of squeezing the input data to extract the hidden features in the data, while the decoder and the encoder can reconstruct the extracted hidden features to restore the data. The principle of tampering based on the autoencoder is to reorganize through two sets of autoencoders, yes, their encoder and decoder, so that the way of mutual exchange and reconstruction finally generates the tampered face [26].

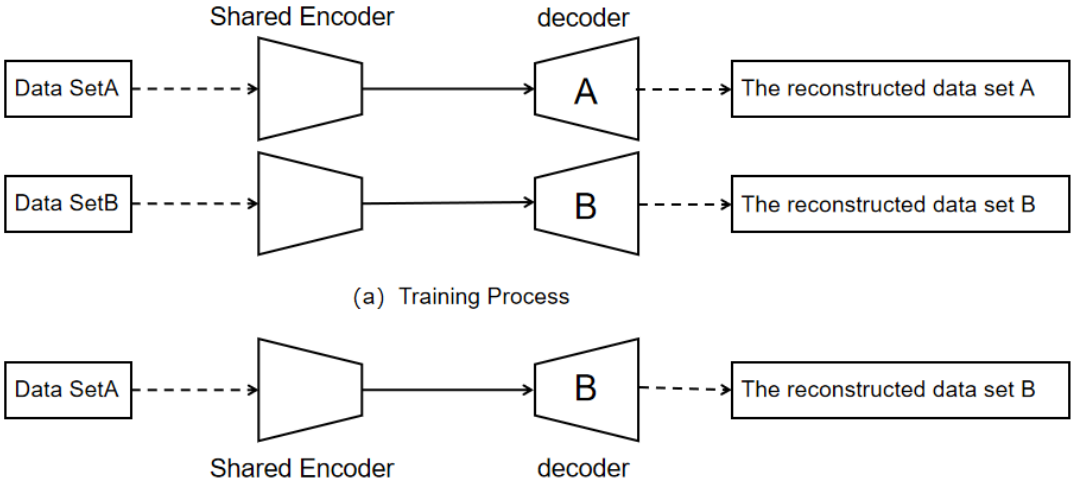

Figure 2: Auto-encoder tamper-up process Figure [26]

DOI: 10.54254/2755-2721/135/2025.21086

Lee in the detection method of development work introduced a kind of highly practical digital forensics tool [27], can detect the depth of different types of forgery at the same time, and put forward the residual autocoder (TAR), based on migration learning with a small number of samples can be in the real environment to detect all kinds of depth forged video.

However, Zhang Ya et al. proposed to add Gaussian filter for preprocessing, and then use the automatic encoder to perform feature extraction, the final accuracy of the detection effect can reach 97.1%, and compared with InceptionV3, ResNet50 and Xception, the three comparison methods obviously have higher generalization [26].

Compared with traditional models, autoencoders can mine more image features, that is, it will be more generalized than traditional methods, but such results still need to sacrifice a part of precision.

4.3. Based on the Vision Transformer method

VIT is an image classification model that has emerged in recent years. Different from the traditional classification block model, VIT is better at the perception and analysis of global information.

Heo et al. proposed a detection method combining VIT on the traditional basis. Figure 3, not only through the traditional CNN image features, but also combines local features with global features, and then uses the deit knowledge distillation model for deep forgery detection. Such a model generalization improves the working performance of the detection[28].

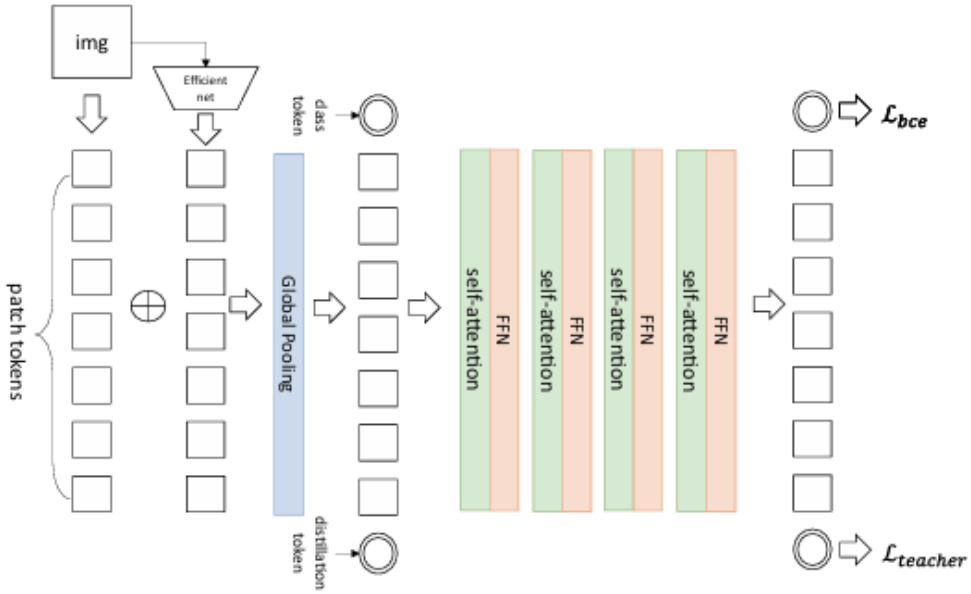

Figure 3: Visual converter-based forgery detection processas proposed by Heo [28]

Lin et al. also proposed a deep forgery detection method based on CNN multiscale convolution and visual converter, which has the ability of global information classification, making it have better detection results on both high-quality and low-quality datasets, and the method has better generalization[29].

Although the ViT-based detection method has extraordinary accuracy, its complex network structure and each detection require special changes and optimization for specific scenarios. As a result, this method is not highly practical.

4.4. Detection methods based on data augmentation

Data enhancement is a diversified method for image processing, while the detection method based on data enhancement is a method to put the center on the modeling.

DOI: 10.54254/2755-2721/135/2025.21086

As shown in figure 4, Guo put forward the adaptive operation tracking extraction network (AMTEN), used to suppress the image content and highlight operation tracking preprocessing, through AMTEN and CNN inheritance to build a face detector, proved its accuracy is quite high, AMTENnet1 average accuracy is as high as 98.52%, is better than the best works at the time [30]. In fact, even when detecting face images with unknown postprocessing operations, it reached an average of 95.17%.

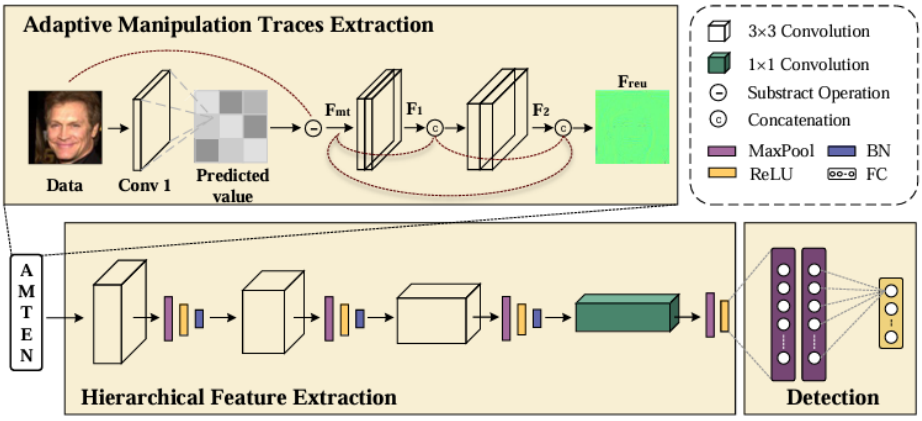

Figure 4: an adaptive tracking extraction network flow proposed by Guo et al[30]

Although data enhancement improves the utilization efficiency of data, it also improves the difference between data and real data, and reduces the detection performance of the model applied in real scenarios.

5. Challenges and prospects

5.1. Challenges for the future

Currently, facial deepfake technology has made significant progress, but there are still several issues that need to be addressed:

1)Poor quality of forged videos: Most existing techniques focus heavily on improving resolution and enhancing the realism of images, often neglecting temporal factors between frames. As a result, there are noticeable differences in lighting, texture, and other details between consecutive frames, leading to unnatural appearances in the video.

2)High hardware requirements: The field typically requires the use of complex models (such as GANs), which have high computational complexity and demand powerful hardware. This creates substantial limitations for research and development in this area.

3)Inconsistent quality of existing datasets: Datasets play an indispensable role in model training and are directly related to the results of the training process. However, existing datasets often rely on similar methods or even a single approach for data generation, leading to discrepancies with real-world conditions. This results in models that struggle to produce optimal results when dealing with real-world situations.

4)Poor generalization of existing models: As deepfake technology continues to evolve, existing models need to possess strong generalization capabilities to ensure detection accuracy. However, most detection models are designed for a specific method of deepfake creation. When applied to cross-dataset detection, their accuracy drops significantly. In real-world applications, this would leave the model ineffective against new, emerging deepfake techniques.

DOI: 10.54254/2755-2721/135/2025.21086

5.2. The future development direction

In the field of deepfake detection, there are still many exploration directions. Amid ongoing evolution and challenges, the following development directions are crucial:

1)Improving the generalization of detection technologies: Detection technologies that can identify common flaws across different data, scenarios, and techniques used to create forged videos or images will be a key area of future research. Possible approaches include: proposing datasets that are universally representative, developing real-time incremental learning models that can quickly adapt to new cases appearing on the internet, and enabling models to learn effectively from a small number of examples.

2)Enhancing the robustness of detection technologies: With the development of deepfake technology and GANs, detection technologies are increasingly facing adversarial attacks. There is an urgent need to improve the robustness of models against adversarial samples and develop countermeasures that can enhance the stability of detection models.

3)Legal and policy aspects: No matter how advanced the detection methods are, they are merely tools for implementation. To prevent malicious events, it is essential to optimize and enforce relevant laws and regulations promptly, strengthen supervision, and increase public awareness. Additionally, widespread education is necessary to establish a "firewall" in the minds of the public, enabling individuals to be fully aware and vigilant against deepfake-related threats.

6. Conclusion

In the ongoing battle between deepfake technology and detection techniques, forgery technology will continue to improve, and the forged content will become increasingly realistic. This will present significant challenges for detection technologies. While there are already many deepfake detection technologies on the market, as mentioned earlier, most current methods are still focused on detecting specific targets or objects. Achieving the generalization of detection technology is likely to require much more time and effort.

Therefore, this paper aims to clarify the current state of research on deepfake detection based on spatial image features and outline the future development directions. To achieve this, the paper have organized and introduced several mainstream detection methods and reviewed many influential research results. At the same time, the paper highlights the challenges currently faced by deepfake detection, outlines the general future directions for development, and seeks to provide reference and assistance for the progress and breakthroughs in deepfake detection.

References

[1] *Heaton, J. (2018). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. Genetic programming and evolvable machines, 19(1), 305-307.*

[2] *Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.*

[3] *Zou X. F., Zhu, D. Z. (2019). Review on Generative Adversarial Network. Journal of Computer Applications (11), 1-9. doi:10.15888/j.cnki.csa.007156.*

[4] *Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.*

[5] *Yang, H. Y., Li, X. H., Hu, Z. (2024). A survey of deepfake face generation and detection technologies Journal of Huazhong University of Science and Technology(Natural Science Edition)1-19. doi:10.13245/j.hust.250021.*

[6] *Peng, X. K., Sun, G. Q., Shao, C. L., Lian, Z. C. (2024). Audio Driven Face Reenactment Method Integrating Facial Deep-perception. Journal of Command and Control (03), 365-371.*

[7] *Wu, W., Zhang, Y., Li, C., Qian, C., & Loy, C. C. (2018). Reenactgan: Learning to reenact faces via boundary transfer. In Proceedings of the European conference on computer vision (ECCV) (pp. 603-619).*

DOI: 10.54254/2755-2721/135/2025.21086

[8] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., ... & Theobalt, C. (2018). Deep video portraits. ACM transactions on graphics (TOG), 37(4), 1-14.

[9] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., & Nießner, M. (2020). Neural voice puppetry: Audio-driven facial reenactment. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16 (pp. 716-731). Springer International Publishing.

[10] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

[11] Gu, G. H., Liu, C., Sun, W. X., Dou, Y. Y. (2024). Facial attribute editing based on global perception. Journal of Huazhong University of Science and Technology(Natural Science Edition) (11), 117-124. doi:10.13245/j.hust.240172.

[12] Xu, C., Zhang, J., Hua, M., He, Q., Yi, Z., & Liu, Y. (2022). Region-aware face swapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7632-7641).

[13] Karras, T. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv preprint arXiv:1710.10196.

[14] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

[15] Yang, X., Li, Y., & Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8261-8265). IEEE.

[16] Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

[17] Rorshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

[18] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

[19] Zhou, T., Wang, W., Liang, Z., & Shen, J. (2021). Face forensics in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5778-5788).

[20] Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18, 2015-2029.

[21] Bang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18, 2015-2029.

[22] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[23] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[24] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition (pp. 5781-5790).

[25] Duan, H., Jiang, Q., Jin, X., Wozniak, M., Zhao, Y., Wu, L., ... & Zhou, W. (2025). Mf-net: multi-feature fusion network based on two-stream extraction and multi-scale enhancement for face forgery detection. Complex & Intelligent Systems, 11(1), 11.

[26] Zhang, Y., Jin, X., Jiang, Q., Lee, S., Dong, Y. Y., Yao, S. W. (2021). Deepfake image detection method based on autoencoder. Journal of Computer Applications (10), 2985-2990.

[27] Lee, S., Tariq, S., Kim, J., & Woo, S. S. (2021, June). Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In IFIP International conference on ICT systems security and privacy protection (pp. 351-366). Cham: Springer International Publishing.

[28] Heo, Y. J., Yeo, W. H., & Kim, B. G. (2023). Deepfake detection algorithm based on improved vision transformer. Applied Intelligence, 53(7), 7512-7527.

[29] Lin, H., Huang, W., Luo, W., & Lu, W. (2023). DeepFake detection with multi-scale convolution and vision transformer. Digital Signal Processing, 134, 103895.

[30] Guo, Z., Yang, G., Chen, J., & Sun, X. (2021). Fake face detection via adaptive manipulation traces extraction network. Computer Vision and Image Understanding, 204, 103170.