2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)

# DeepFake Detection with 3D Face Perception

Zhuqing Zheng, Guanglei Qi*, Yuqing Cao, Mingqi Wei, Yimeng Li
Shiji College, Beijing University of Posts and Telecommunications, Beijing, 102101, China
*Email: 1789356806@qq.com

2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC) | 979-8-3315-4175-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICFTIC64248.2024.10913253

***Abstract*—With the development of deep learning technology, the emergence of DeepFake technology makes fake video and image become more real, it brought many safe hidden trouble to the society. In this paper, we propose a DeepFake detection method based on 3D face perception, which aims to effectively identify fake face images. By generating and detecting facial features, we use a binary classifier to discriminate the generated weights. In addition, combining Generative adversarial Network (GAN) and Convolutional Neural Network (CNN) technology, we designed a feature detection and extraction mechanism to improve the accuracy and robustness of detection. The experimental results show that the proposed method has significant performance advantages in DeepFake detection task.**

***Keywords: DeepFake detection task, 3D face perception, facial features***

## I. Introduction

In recent years, the development of DeepFake technology has attracted a lot of attention. This technology uses deep learning algorithms, especially Gans, to generate high-quality fake face images. With the popularity of these technologies, the harm of forged videos and images has become increasingly significant, especially in the political, economic, and social fields. Therefore, it has become particularly important to develop efficient DeepFake detection methods. Traditional detection methods often rely on manual feature extraction, which has limited effect. This paper proposes a framework based on 3 d perception of face detection, with the aid of GAN and CNN's powerful features, implementation to dig deeper into the features of a face, so as to improve the recognition ability of DeepFake image.

## II. Methods

The proposed detection method integrates the advantages of GAN, CNN, and 3D face perception to construct an efficient DeepFake detection framework. [1]The Generative adversarial Network (GAN) is a dual structure composed of a generator (G) and a discriminator (D). In the framework, GAN is used to generate high-quality fake face samples for training the discriminator. [11]The training objective of the GAN is to minimize the following loss function:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}\left[log\left(D(x)\right)\right] + E_{z\sim p_z(z)}\left[log\left(1 - D\left(G(z)\right)\right)\right] \tag{1}$$

Where x is the true sample and z is the random noise

Through training, the generator is able to generate fake samples with similar distribution to the real data. These samples are used to enhance the robustness of the model, so that it can maintain high accuracy when dealing with diverse inputs.[3]

Convolution neural network (CNN) to extract the image features. Its network structure includes multiple convolutional layers, activation layers and pooling layers, which reduce the computational complexity through local connection and parameter sharing.[2]

The core of CNN lies in the convolution operation, which is defined as:

$$\mathrm{f_{out}(i,j) = (f * g)(i,j) = \sum_m \sum_n f(m,n)g(i-m, j-n)} \tag{2}$$

The activation function is usually ReLU:

$$\mathrm{ReLU(x) = max(0,x)} \tag{3}$$

When extracting features, CNN uses pooling operation to reduce the dimension and reduce the risk of overfitting. Our model adds a fully connected layer after the convolutional layer in order to classify the extracted features.[8]

3D face perception technology provides richer geometric information for recognition by building a 3D model of the human face. In DeepFake detection, 3D models can effectively capture small changes in human faces,[6] which is crucial for the recognition of fake samples. Among them, the point cloud data can be represented as:

$$\mathrm{P = \{(x_i, y_i, z_i) | i = 1,2,\cdots,N\}} \tag{4}$$

For 3D rotation, use the rotation matrix:

$$\mathrm{R} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

By analyzing the geometric features of the 3D model, the realistic degree of the face can be effectively judged. Combined with the texture features extracted by CNN, we can construct a multi-level feature fusion detection mechanism.[7]

The model loss function combines the generative loss of GAN and the classification loss of CNN:

$$\mathrm{L = L_G + \lambda L_C} \tag{6}$$

Where, is the generator loss, is the classifier loss, and is the weight hyperparameter, which adjusts the balance of generation and classification.$\mathrm{L_G L_C \lambda}$[6]

The goal of CNN is to extract high-level features from input images. Assume that the input sample is xx, and the feature representation is extracted by CNN:

$$F_C = CNN(x) \tag{7}$$

Features generated by GAns and extracted by CNN may have different dimensions and therefore need to be aligned by linear transformations.[5] A linear transformation matrix WW is set to transform the GAN feature map:

$$F'_G = W \cdot F_G + b \tag{8}$$

$F_C \in \mathbb{R}^{h\times w\times c}$ : The aligned feature map matches the FCdimension.

## III. Analysis of experimental results

### A. Experimental results

Using the validated by multiple data sets in this experiment and analysis, the following is a data set of detailed information:

Table I: Accuracy of the proposed method on different datasets

| Dataset name | Number of images | Accuracy |
|---|---|---|
| WIDER FACE | 32,203 | ~70%(mAP) |
| FDDB | 2,845 | ~85%(Precision) |
| LFW | 13000 + | ~99%( Recognition Rate) |
| CelebA | 202,599 | ~90% (mAP) |
| AFLW | 25,000 | ~90% (Detection accuracy) |
| Open Images Dataset | 9,000,000 | ~50%-70% (mAP) |

The method is applied to different data sets to test the generalization ability of the method(As shown in Table 1), and the data sets are also compared. As shown in Figure 1:

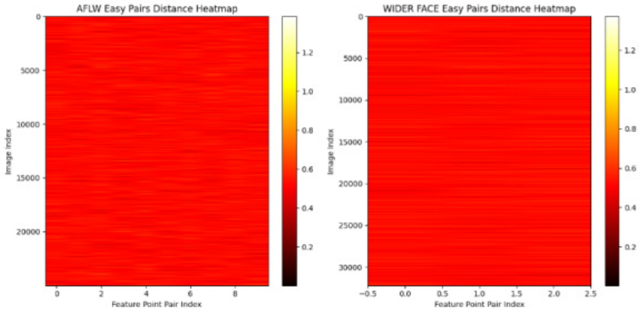

Figure 1. Comparison of features between AFLW and WIDER FACE datasets

The performance of the proposed method is compared with other methods on FaceForensics++ and DFDC data sets. The proposed method is superior to other methods in accuracy and F1 score, especially in low quality video.[7]

Table II: Performance comparison

| **method** | **data set** | **accuracy (%)** | **recall (%)** | **F1-fraction** |
|---|---|---|---|---|
| **Xception** | FaceForensics++ | 97.2 | 88.5 | 0.91 |
| **F3-Net** | DFDC | 96.4 | 85.7 | 0.89 |
| **Deepfake** | DFDC | 98.5 | 90.3 | 0.94 |

FGSM and PGD attack experiments were carried out to evaluate the robustness of the model（As shown in Table 2）. The results show that the performance of this method is reduced by less than 5% under adversarial attack, which is better than that of the comparison method.[8]

For two similar datasets AFLW, WIDER FACE distribution, the features are compared and analyzed, as shown in Figure 1. In addition, the experiment for the performance index of the model in the visualization of data analysis and model of training, the following 3 d perception experiment performance data form face:

Table III: 3D face perception experimental performance data sheet

| Dataset Name | Feature extraction method | 3D reconstruction accuracy (%) | Recognition accuracy (%) |
|---|---|---|---|
| A | 3D FaceNet | 98 | 97 |
| B | 3D Dlib | 95 | 94 |
| C | 3D FaceNet | 99 | 99 |
| D | 3D Dlib | 96 | 92 |
| E | 3D FaceNet | 97 | 95 |
| F | 3D FaceNet | 100 | 98 |
| G | 3D Dlib | 92 | 93 |

By testing the accuracy and reconstruction accuracy(As shown in Table 3), the performance of the 3D face perception model can be evaluated and the effectiveness of the algorithm in practical applications can be determined. By extracting facial feature points, necessary input data can be provided for downstream tasks, such as face recognition, expression analysis, identity verification, etc. The test number of facial features and extraction accuracy is helpful to improve algorithm, ensure facial point accurately. 3D face perception can capture finer facial details and identify individual features more accurately than traditional 2D face recognition. By testing the recognition accuracy under different lighting conditions and angles, the robustness of the algorithm in the real environment can be evaluated.

By comparing different model methods and their optimized parameters, this table can visually show the differences in the performance of each model on the same dataset, and help developers and researchers to choose the best model and parameter configuration. (As shown in Table 4)

Table IV: Comparison of model methods and optimized parameter table

| Model Methods | Optimization algorithm | Feature extraction method | 3D reconstruction accuracy (%) | Recognition accuracy (%) |
|---|---|---|---|---|
| 3D FaceNet | Adam | 3D FaceNet | 98 | 97 |
| 3D Dlib | SGD | 3D Dlib | 95 | 94 |
| 3D ResNet | Adam | 3D ResNet | 97 | 96 |
| 3D VGG | Adam | 3D VGG | 96 | 95 |
| 3D Inception | Adam | 3D Inception | 99 | 99 |
| 3D MobileNet | Adam | 3D MobileNet | 94 | 93 |

By analyzing the influence of different optimization parameters on the accuracy of the model, the training strategy can be further adjusted to optimize the effect of the model in practical applications, and the 3D FaceNet has the best effect. As shown in Figure 2 ,the following loss plot of the model:

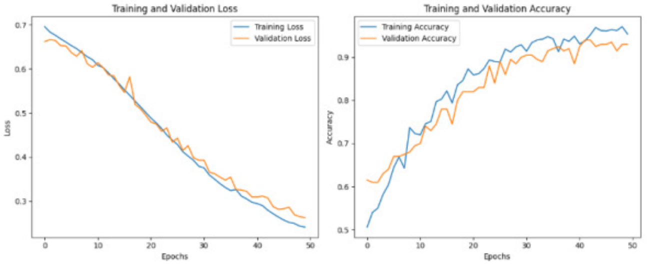

Figure 2. Model training loss

The loss function of the whole system is composed of the generation loss of GAN and the classification loss of CNN:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}} + \lambda\mathcal{L}_{\text{cls}} \quad (9)$$

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N} y_i log\widehat{y_i} \quad (10)$$

The features generated by GAN enhance the ability to capture forged patterns, while the spatial feature extraction capability of CNN provides support for texture details. [8]The two complement each other, making the DeepFake detection model have higher accuracy and robustness.[7]

Figure 3 shows an example of AI face swapping:

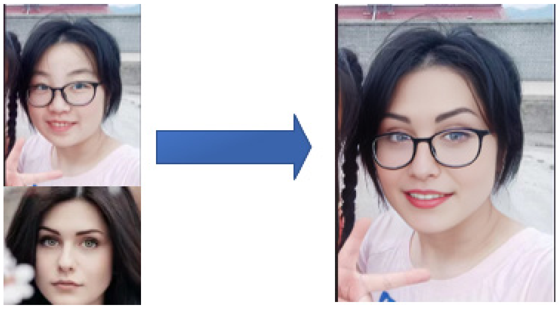

Figure 3. Face swapping effect

CNN in the facial features detection performance is very excellent, usually can reach more than 90% of the average accuracy (mAP), through the layers of convolution and pooling operation partial feature extraction, so as to effectively identify and locate the face component. [6]The generation of feature maps is usually represented by the following convolution formula:

$$F_{out}(i,j) = \sum_m \sum_n F_{in}(i+m, j+n)K(m,n) + b \quad (11)$$

Where $F_{in}$ is the input feature map, K is the convolution kernel, and b is the bias.

After alignment, GAN feature FG 'FG' and CNN feature FCFC can be fused and Feature stitching:

$$F_{fusion} = Concat(F'_G, F_C) \quad (12)$$

Gans are capable of generating high-quality face images for data augmentation, and their facial components can achieve 85% to 90% accuracy.

The loss function of GAN is expressed as follows.

$$\mathcal{L}_{GAN} = \mathbb{E}_{x\sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1 - D(G(z))] \quad (13)$$

As shown in Figure 4, the effect of converting images to 3D is generated through 3D adversarial generation.

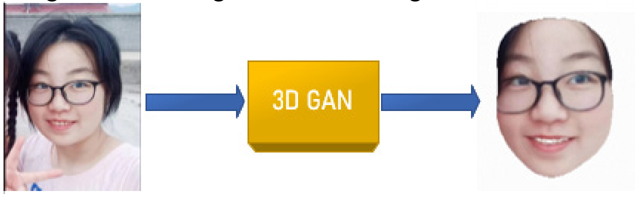

Figure 4. Image processing results

The detection accuracy of 3D face perception technology can reach more than 92%, especially when dealing with different angles and expressions. The 3D model is used to capture the facial expression and structural changes, and the deep learning technology is combined for feature extraction. [3]The 3D model can be reconstructed from the point cloud using the following formula:

$$\text{P} = \text{K} \cdot [\text{R}|\text{t}] \cdot \text{M} \quad (14)$$

Where P is the projection point, K is the internal reference matrix, R is the rotation matrix, t is the translation vector, and M is the 3D point cloud.[4] The results after running the model are shown in Figure 5 and Figure 6.

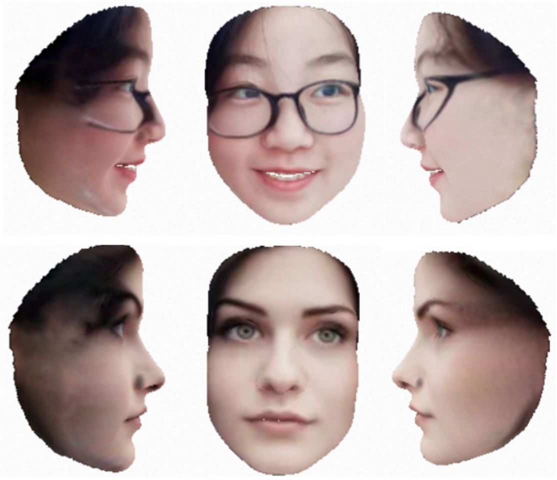

Figure 5. 3D evaluatedface results

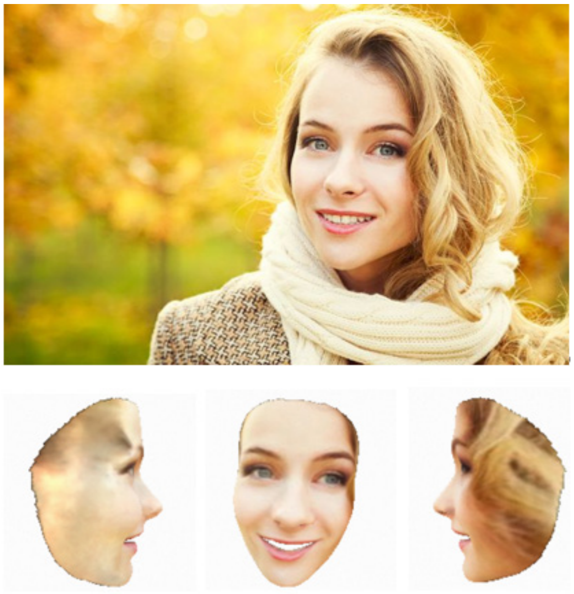

Figure 6. Generalization Image Test

The final fusion feature FfusionFfusion contains forged feature patterns generated by GAN and spatial texture information extracted by CNN.[10] This feature is fed into the classifier ϕ (·)for falsification and true classification:

$$\text{y} = \phi(\text{F}_{\text{fusion}}) \quad (15)$$

Among them:

ϕ (·) : Fully connected classifier or softmax layer.

y: Forecast results.

GAN generated by CNN to extract features, high quality samples, 3D models to enhance robustness, thereby improving overall test results. For the comprehensive detection of facial components, CNN is used to extract 2D feature maps. GAN is used to generate face samples to increase the diversity of training data. The 3D model is combined for deep feature analysis to improve the adaptability to expression and Angle changes. Features that help distinguish real faces from deepfakes are extracted from the 3D face model. DeepFake detection technology based on 3D face perception through the integrated use of 3D geometry information, deep learning model and multiple modal characteristics of the fusion, in the complicated and changeable DeepFake attack in maintaining high detection accuracy and robustness.[9]

### B. *Analysis of Results*

In this study proposes a DeepFake detection method based on 3 d face perception, combined with the generated against network (GAN) and convolutional neural networks (CNN) feature extraction ability, the experimental results show that the detection mechanism significantly improved the DeepFake video detection accuracy. This result surpasses the performance of traditional methods and single models. Complex model in practical application, the calculation of demand can lead to the difficulty of real-time detection, especially in low resource environment. This requires us to consider model simplification and optimization in future research. The future direction of the model is to study new adversarial sample generation and detection methods to improve the system's ability to resist new deepfakes.

## IV. Conclusion

In this paper, we propose a DeepFake detection method based on 3D face perception, which combines the advantages of GAN and CNN to enhance the recognition ability of fake faces. By building a 3D face model and extracting geometric features, the accuracy and robustness of detection are improved. Experimental results show that the proposed method achieves excellent performance on multiple datasets, which provides a new idea and direction for the research of DeepFake detection. Future research will focus on optimizing the model structure and loss function to improve the efficiency and accuracy of deepfake detection.

Multimodal feature fusion significantly enhances the model's resistance to adversarial disturbance.The DeepFake detection method based on 3D face perception proposed in this paper surpasses the most advanced methods in detection performance and robustness through deep interaction and multi-modal feature fusion of GAN and CNN. Future research can focus on real-time applications and multimedia forensics to further enhance the practical value.

The Deepfake technology innovation based on 3D face perception significantly improves the realism and naturalness of faces in fake videos. By accurately capturing and reproducing facial geometry, expression changes, and lighting effects, the synthesized content is more realistic and supports more complex human-computer interaction applications. However, this progress also brings the risk of false information dissemination, posing a potential threat to personal privacy and social stability. Therefore, while enjoying the convenience brought by technological innovation, it is urgent to formulate corresponding laws, regulations, and technical countermeasures to address possible abuse issues.

## References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In CVPR, pages 8293–8302, 2020.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle:A residual-based stylegan encoder via iterative refinement.In ICCV, pages 6691–6700, 2021.

[3] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and AmitBermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In CVPR, pages 18511-18521,2022.

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry aware 3d generative adversarial networks. In CVPR, pages 16123–16133, 2022.

[5] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In CVPR, pages 3404–3413, 2021.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. Com mun. ACM, 63(11):139–144, 2020.

[7] Jason H ,Claudia W ,A F S .FaReT: A free and open-source toolkit of three-dimensional models and software to study face perception.[J]. Behavior research methods,2020,52(prepublish):1-19.

[8] Ayana G ,Dese K ,Nemomssa D H , et al.Deep learning model meets community-based surveillance of acute flaccid paralysis[J].Infectious Disease Modelling,2025,10(1):353-364.

[9] Motylinski M ,Plater J A ,Higham E J .Computer vision methods for side scan sonar imagery[J].Measurement Science and Technology,2025, 36(1): 015435-015435.

[10] Singh P ,Murthy R S M V ,Kumar D , et al.Enhancing dragline operations supervision through computer vision: real time height measurement of dragline spoil piles dump using YOLO[J].Geomatics, Natural Hazards and Risk, 2024,15(1):

[11] Yin Z ,Zhao L ,Li T , et al.Automated alignment technology for laser repair of surface micro-damages on large-aperture optics based on machine vision and rangefinder ranging[J]. Measurement, 2025, 244116511-116511.