Diffusion Based Data Augmentation for Face Anti-spoofing

Tianheng Zhang

{kevin_tianheng2002@outlook.com}

College of Computer and Information Science, Southwest University, Chongqing, 400715, China

Abstract. As the internet continues to evolve, the application of facial recognition technology across different sectors ranging from unlocking smartphones to enhancing airport security and facilitating financial transactions is on the rise. Consequently, the necessity for robust face anti-spoofing (FAS) technology is becoming ever more critical. This article presents a new FAS method that enhances the security of face recognition systems through the optimization of the diffusion model and neural network structure. The diffusion model is used to generate spoof face data similar to the real face, which enriches the original dataset and enhances the robustness of the model. Furthermore, the UNet network structure is enhanced by incorporating the Attention mechanism and optimizing the loss function through a comprehensive consideration of SSIM, Cross-Entropy, and MSE. These modifications aim to improve the ability to distinguish between authentic and fraudulent faces. Experimental results demonstrate that our proposed approach surpasses existing technologies in testing scenarios, showcasing its potential in the field of face anti-spoofing.

Keywords: Diffusion model, computer vision, face anti-spoofing, UNet, self-attention.

1 Introduction

In the realm of security, the implementation of face anti-spoofing techniques within face recognition systems is increasingly capturing the interest of both academic researchers and industry professionals. However, the variety of spoofing methods ranging from print assaults and replay tactics to mask and lighting challenges continues to complicate the task of identifying different types of spoofed faces. Recently, scholars have been striving to develop more comprehensive and differentiated features aimed at thwarting face spoofing, including techniques like LBP, HOG, and LBP-TOP among others [1-3]. Usually, these features are referred to as manual features because they are designed manually. As technology evolves, attackers can trick motion and texture analysis by creating 3D masks, printed photos, or videos that mimic real human movements. Moreover, although conventional CNNs have achieved remarkable success in computer vision, they encounter hurdles when it comes to anti-spoofing, particularly regarding their ability to generalize effectively [4]. In recent years, various methods have emerged that utilize the estimation of supplementary signals obtained from RGB data to reveal intrinsic differentiators between genuine images and their fraudulent duplicates. These techniques include depth maps, remote photoplethysmography (rPPG) [5], color textures, and distortion analyses, all of which have demonstrated promise in enhancing generalization performance [6]. The advent of deep learning has enabled neural networks to autonomously

CONF-MLA 2024, November 21, Adana, Turkey
Copyright © 2025 EAI
DOI 10.4108/eai.21-11-2024.2354616

acquire hierarchical features directly from raw pixel data. Architectures like VGG and ResNet often prioritize high-level semantic representations while overlooking low-level characteristics [7-8]. However, the main network structure used in this paper to differentiate between real and fake faces is the UNet network structure with the incorporation of the Attention mechanism. Furthermore, for dataset optimization in training models, an initial training is conducted using real faces with a diffusion model, and then the augmentation data as a fake face is generated with the help of the back-diffusion process.

2 Literature review

2.1 Traditional FAS techniques

The manual feature extraction-based approach relies on artificially designed features that exhibit different feature patterns in different types of attacks. There are mainly four methods, as follows: By examining the textural characteristics of an image, methods based on texture are capable of discerning between authentic and counterfeit faces. Commonly used texture features include Local Binary Mode (LBP), Gabor filter, and Directional Gradient Histogram (HOG). These methods take advantage of texture differences to detect subtle differences in fake facial images. Second, the motion-based approach uses the dynamic changes of the face in the video sequence to counter spoof. For instance, the identification of genuine faces from static images or videos can be effectively achieved by analyzing subtle movements like blinking and lip motions. Another approach involves utilizing 3D facial structures to detect counterfeit faces, where variations in the three-dimensional shape of a real face are utilized. By employing 3D sensors or stereo vision technology, depth information about the face can be acquired to determine if the input image possesses an authentic three-dimensional structure. Last, methods based on multispectral reflection. The method based on multispectral reflection distinguishes between real and fake faces by the light reflection characteristics of different wavelength bands.

2.2 Deep learning for FAS techniques

As deep learning has evolved, neural network-based feature extraction techniques have become increasingly popular in combating facial fraud.

Initially, Yang first introduced the Convolutional Neural Network (CNN) for face anti-spoofing detection [9]. This method significantly reduces the error rate and improves the generalization ability of features. However, CNNs mainly rely on local receptive fields for feature extraction, resulting in their limited ability to capture large-scale feature connections and long-distance dependencies. In addition, CNN performs poorly in dealing with complex spatial transformations and non-local dependencies. Then, Gu introduced the self-attention mechanism and auxiliary MLP convolution to further enhance the ability of global feature extraction [10]. Compared with the combination of CNN and MLP, this mechanism has excellent performance in flexibly adjusting the weight of the feature map and capturing global information. However, there is still room for optimization in the performance of multi-scale feature extraction, contextual information fusion, and feature map reconstruction in this literature. In the past few years, there has been significant progress in the field of face anti-spoofing detection technology, particularly with the utilization of Generative Adversarial Networks (GAN). Wu introduced a method based on GAN for generating and discriminating [11]. In the GAN framework, the generator creates synthetic images, whereas the discriminator's role is to identify and differentiate between authentic and fabricated images. By simultaneously generating counterfeit data and training discriminators, this approach can effectively detect intricate patterns in fraudulent activities and enhance the precision of facial anti-spoofing detection. Nevertheless, GAN models are prone to pattern crashes or instability during training, which requires more refined tuning of training strategies.

Overall, Facial anti-spoofing detection technology has steadily evolved from the conventional approach of manual feature extraction to an automated method leveraging deep learning for feature identification. All kinds of methods have their advantages and disadvantages. In particular, there are still some bottlenecks in the fineness of feature extraction and the diversity of datasets. Hence, this study aims to enlarge the current dataset, create additional data through the fusion of the diffusion model, and employ an improved UNet-Attention framework to enhance the efficiency of detecting face spoofing.

3 Methods

In this segment, the writer introduces an innovative approach to facial anti-spoofing, leveraging the cutting-edge diffusion model techniques of the latest generation, as shown in Figure 1.

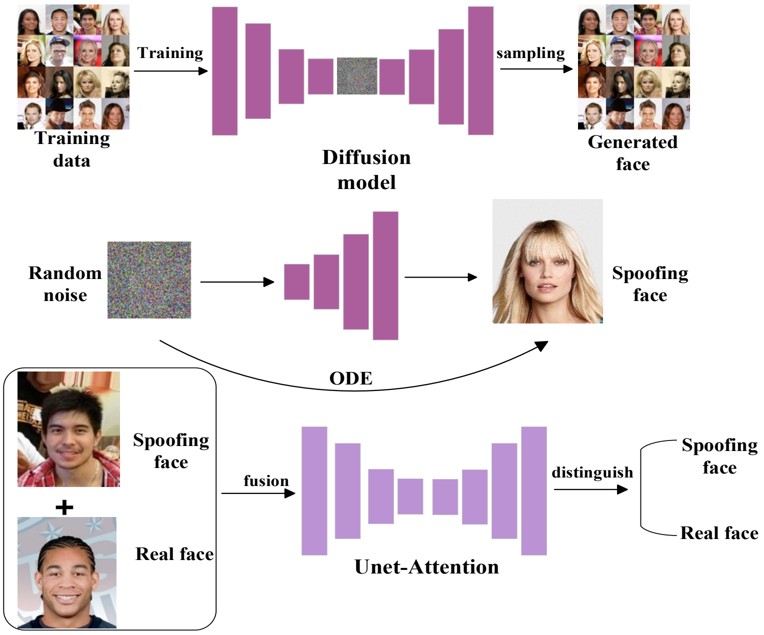

Fig. 1. Flowchart of the scenario in this article.

3.1 Enhanced data generation based on diffusion models

This study presents a novel approach to augmenting data by utilizing the Denoising Diffusion Probabilistic Model (DDPM) [12]. The objective is to improve the robustness and ability of face recognition systems in effectively handling spoofing challenges. Given the diverse and everevolving nature of spoofing attempts ranging from photo to video attacks, it is crucial to develop a training dataset that sufficiently encompasses a broad spectrum of these deceptive methods. However, collecting sufficiently diverse real-world data often faces high costs and data scarcity. Therefore, the dataset's diversity can be effectively enhanced and the model's resilience to intricate deceitful techniques can be strengthened by producing synthetic facial data that closely resembles real faces.

Diffusion model. Typically, a diffusion model operates through two main phases [13]. The first phase involves gradually transforming the input data x_0 into noise x_t, a process that relies heavily on manual design techniques. The second phase focuses on reversing this gradual noise transformation, ultimately reconstructing the datax_0. Essentially, the diffusion model kicks off with noise and systematically refines it into progressively clearer samples x_{t-1}, x_{t-2}..., until it yields the final product x_0. At each time step t, a distinct noise level is associated, and x_t can be regarded as a combination of the original signal x_0 intertwined with a certain amount of noise, such as Gaussian noise. In order to tailor the DDPM model, $\theta(x_t, t)$ is employed to forecast the noise element present in a corrupted input example x_t .Throughout the training phase of these models, every instance is crafted by randomly drawing from the data x_0 ,selecting a specific time point t, and incorporating noise θ, which together produce the noise sample x_t. Then, optimize the training target by minimizing $\|\varepsilon - \varepsilon_0(x_t, t)\|^2$ [14].

Figure 2 demonstrates the concept of progressive diffusion, wherein noise is gradually incorporated into the original image. This results in the image progressively shedding its original details, ultimately transforming into nothing but Gaussian noise.

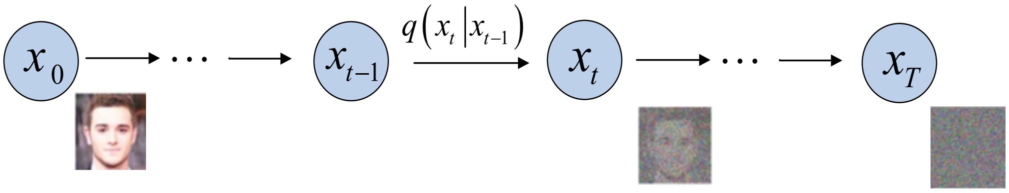

Fig. 2. The forward diffusion process of the diffusion model.

In particular, when you feed a real image $x_0 \sim q(x)$ into the diffusion model, it injects Gaussian noise into the image a total of T times, resulting in the transformed image $x_1, x_2, \ldots, x_T$. The magnitude of each incremental movement is governed by a set of hyperparameters dictating the dispersion of the Gaussian distribution $\{\beta_t \in (0,1)\}_{t=1}^{T}$. At any given point in time, t, the progression solely depends on the state at time t-1, allowing the sequence to be represented as follows (equations 1&2):

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}) \tag{1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

As time progresses, x_t gradually approaches pure noise. Eventually, with $T \to \infty$, what was once an identifiable image becomes indistinguishable from standardized Gaussian noise.

In the depicted reverse diffusion process in figure 3, the neural network model progressively eliminates noise and restores the initial image.

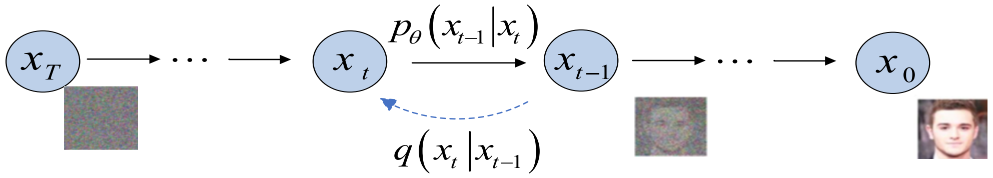

Fig. 3. The reverse diffusion process of the diffusion model.

The goal of the process is to start with pure noise x_T and gradually produce a realistic image. The reverse diffusion of each step can be expressed as (3):

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1};\ \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \tag{3}$$

where μ_θ denote the functions for mean and Σ_θ for variance, while θ signifies the model's parameters. The primary objective of reverse diffusion is to decrease the Evidence Lower Bound (ELBO) shown as (4) to learn the optimal generation process:

$$L_{ELBO} = E_q[\sum_{t=1}^{T} D_{KL}\left(q(x_{t-1}|x_t,x_0)\ ||\ p_\theta(x_{t-1}|x_t)\right)] \tag{4}$$

The KL divergence in this equation acts as a measure to assess the difference between the distribution generated by the model and the real data, represented as (5):

$$D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) = \frac{1}{2\sigma_t^2}||\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)||^2 \tag{5}$$

By minimizing this difference, the model gradually approximates the desired generation effect.

Typically, to simplify optimization, the objective function of DDPM can be refactored to directly predict the noise of each step as (6):

$$L_{simple} = E_{t,x_0,\varepsilon}[||\varepsilon - \varepsilon_0(x_t,t)||^2] \tag{6}$$

where ε represents the noise added during the forward diffusion process, and $\varepsilon_\theta(x_t,t)$ is the noise predicted by the neural network. In the backward diffusion method, the algorithm trains itself to strip away the noise and progressively restore an image that closely mirrors the actual data.

Leveraging the diffusion model to generate augmented data. The diffusion model data augmentation method proposed in this study is as follows:

First, the original real face dataset should be prepared. The researcher gathers an extensive collection of authentic facial images, which serve as the foundational dataset for training the generative model. This dataset contains a variety of face images with different angles, lighting, expressions, and backgrounds to ensure that it is broadly representative, and the researcher used the celeb A dataset for training in this study. Then, using the real face dataset, train a generator based on the diffusion model. The training process enables the model to learn how to gradually denoise random noise through forward and reverse diffusion mechanisms to generate realistic face images. In the training procedure, a single face image is randomly chosen from the dataset at each iteration, and incremental Gaussian noise is applied to produce a sequence of noisy image samples. Through backpropagation, the model learns how to remove different levels of noise and ultimately restores a clear image that closely resembles a real human face. After the training is completed, generate a batch of dummy faces that are highly similar to real faces by inputting completely random Gaussian noise into the model and going through the reverse diffusion process. The resulting dummy faces are not only visually high-quality, but also reflect similar features to those in the original dataset. Finally, the generated dummy face data is added to the original dataset to form an extended augmented dataset. By utilizing this method, the author can effectively expand the range of our data and improve the diversity of our datasets. As a result, we provide more comprehensive and valuable data to strengthen the training of future facial anti-spoofing models.

3.2 Neural network structure design: UNet-Attention

This research developed a novel neural network architecture, as illustrated in figure 4, which combines the UNet and Attention mechanisms. This integration enables the processing of a combined input comprising augmented data and raw data generated by the diffusion model, thereby enhancing the precision of face anti-spoofing detection. In order to optimize the neural network training process, the author has also updated the loss function to regularize and optimize the model from multiple perspectives. The following sections of the network are explained in detail, the design of the loss function, and the role of the Attention mechanism.

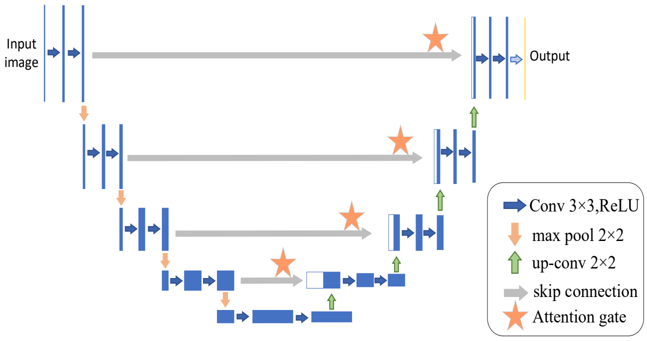

Fig. 4. UNet-Attention network architecture.

UNet structure analysis. The UNet architecture is known for its comprehensive convolutional neural network structure, which includes pathways for both subsampling and upsampled information. Its main advantage lies in its ability to preserve spatial resolution details through skip connections, leading to improved performance in tasks like image segmentation and feature extraction. In our face anti-spoofing task, UNet plays a crucial role in identifying real and fake faces by extracting multi-level features from the input image. The subsampled path consists of a series of convolutional and pooling layers that effectively capture both fine-grained and overall characteristics of the image. Each layer utilizes adaptable kernels during the convolution process, leveraging the ReLU activation mechanism to handle complex nonlinear attributes. Additionally, the pooling layer reduces the size of the feature map while retaining essential representative features through dimensionality reduction operations.

In the subsampled path, each layer consists of two convolution operations and one pooling operation. Multiple convolution kernels are used to extract features, defined as (7):

$$F_l = \sigma(W_l * F_{l-1} + b_l) \tag{7}$$

where F_{l-1} is the input feature map of layer l-1, W_l is the convolutional kernel weight, b_l is the bias, and σ is the ReLU activation function. The feature map obtained after the convolution operation will be normalized in batches to stabilize the training process.

By applying a 2×2 pooling operation, the feature map's spatial resolution is reduced by half while retaining crucial feature information. This process can be mathematically represented as (8).:

$$P_l = max(\{F_l[i,j]\}) \tag{8}$$

Here is the pooled feature map, which represents the pixel values of the lth position i and j. The path that is subsampled obtains the features of the image at different levels, and as more layers are added to the network, it becomes capable of capturing increasingly abstract characteristics.

The upsampled path restores the high-level semantic features obtained during the subsampled process to the spatial resolution of the original image through a deconvolution layer. The operation of deconvolution is to perform an inverse operation on the convolution and gradually zoom in on the feature map. The deconvolution process can be expressed as (9):

$$F_l^{'} = \sigma\left(W_l^{'} * F_{l-1} + b_l^{'}\right) \tag{9}$$

The features from the subsampled path in UNet are directly transmitted to the corresponding layer in the upsampled path through a skip connection. Through this structure, the model can use the detailed features extracted earlier when upsampled to avoid information loss. The skip connection adeptly integrates the feature map from the encoding phase with its counterpart in the decoding phase by means of concatenation.

Attention mechanism. In the conventional UNet structure, the primary emphasis of the convolutional layer lies in extracting local characteristics from the image. Moreover, by increasing the size of the convolutional kernel, a greater amount of global information is captured by the model. However, this design can lead to the neglect of some important detail features when dealing with complex tasks. Therefore, a self-attention mechanism was added to each layer of the upsampled path of the UNet architecture to increase the focus on important features. The Self-Attention mechanism has the potential to elevate the quality of feature representations by determining the relationship between every location on an input feature map with all other locations within that map. The formula is as follows (10):

$$A(Q,K,V) = Soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the scaling factor. The Attention mechanism adjusts the importance weights of the feature map by calculating the correlation between each location and other positions. The introduction of the Attention module after sampling each layer of the convolution of the path on the UNet ensures that the network can focus more on the key regions in the spoof image in the process of recovering spatial information. This mechanism is especially important for anti-spoofing tasks, because faked features tend to have special patterns only in certain local areas (e.g., light anomalies, detail fakes, etc.), and Attention can capture these subtle but crucial differences.

Design of the loss function. The primary objective of the model's training is to enhance its proficiency in differentiating between genuine and counterfeit facial images. The researcher designs a composite loss function that combines multiple loss terms, and its formula is as follows (11):

$$L_{overall} = L_{SSIM} + \alpha L_{constructive} + \beta L_{cross-entropy} \tag{11}$$

L_{SSIM}:It is used to maintain the perceptual similarity between the generated image and the real image, and the Structural Similarity Index (SSIM) is used. The formula is (12):

$$L_{SSIM} = 1 - SSIM(x, y) = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{12}$$

$L_{constructive}$:Mean square error (MSE) is used to measure the pixel-level difference between the generated image and the real image. The formula is (13):

$$L_{constructive} = MSE(x, y) = \|x - y\|^2 \tag{13}$$

$L_{cross-entropy}$:Classical classification loss, which is used to supervise classification accuracy. The formula is (14):

$$L_{cross-entropy} = -\sum_{i=0}^{1} p_i \log(1 - p_i) \tag{14}$$

The design of this comprehensive loss function helps the model to optimize the classification accuracy and pixel-level similarity while ensuring perceived similarity. The model's generalization capability can be enhanced by receiving diverse informational feedback from each loss item.

UNet with attention. In the inference phase of the final model, a hybrid dataset (consisting of augmented data and a raw dataset) passes through the UNet with Attention network as input. This network combines UNet and Attention, can fully capture the fine-grained and global features in the image, especially the fake faces generated by the diffusion model, whose subtle distinctions in the features might not mirror the genuine faces, and the Attention mechanism boosts these disparities, aiding the model in more effectively discerning the authentic from the forged.

4 Experiment

The researcher utilized the Celeb A dataset, which contains over 200,000 celebrity facial images with various facial attribute labels such as gender, hairstyle, and glasses. The dataset is divided into three subsets: training set, validation set, and test set for different applications including face attribute analysis, identity recognition based on facial features and attempts to decect deception in face recognition systems.

During the initial data preprocessing stage, only real human face data was selected to train the diffusion model to ensure that the generated augmented data is very similar to real human faces. Additionally, each image was adjusted to a size of 128×128 pixels to meet model input requirements. All images were normalized within a pixel value range of 0–1 to reduce noise impact on model performance.

In configuring diffusion model parameters, T was set at 1000 steps gradually adding noise and performing denoising training. DDIM algorithm was used during sampling process with 50 backpropagation steps ensuring high-quality fake faces are generated. To further enhance the realism of generated data, small noise adjustments were made during each sampling iteration.

UNet-Attention network was trained using Adam optimizer with an initial learning rate of 1e-4 and applying learning rate decay strategy by reducing it by a factor of 0.1 every 200 iterations. Maximum training rounds were set at 300 while batch size of 32 optimized GPU performance and prevented overfitting.

Classification accuracy, Structural Similarity Index (SSIM), and Normalized Mean Square Error (NMSE) were chosen as evaluation metrics for comprehensive assessment of model performance. In experimental process, the performance of each model was tested on test sets recording their accuracy and similarity in recognizing real and fake faces.

5 Results

This study compares the proposed Anti-diffusion Net with three commonly used deep learning models (CNN, DNN, GAN) in terms of their performance on real and fake faces. The evaluation metrics employed include classification accuracy, structural similarity index (SSIM), and normalized mean square error (NMSE). Table 1 presents the specific results obtained for each model.

Based on the findings, it can be concluded that the proposed Anti-diffusion Net outperforms other models in various performance indicators. Firstly, the accuracy of Anti-diffusion Net reached 95% in detecting real faces, which was improved by 3%, 2% and 4%, respectively, compared to CNN, DNN and GAN. The model also achieved a detection accuracy of 90% for fake faces, exceeding CNN, DNN, and GAN by 5%, 2%, and 3%, respectively.

Table 1. Performance Metrics for Face Detection Methods.

Method	Accuracy (Real Face)	Accuracy (Spoofing Face)	SSIM	NMSE
Anti-diffusion Net (Proposed)	95%	90%	0.98	0.02
CNN	92%	85%	0.95	0.05
DNN	93%	88%	0.96	0.04
GAN	91%	87%	0.94	0.06

In addition to accuracy, Anti-diffusion Net also performs well in image quality evaluation indicators. The SSIM value was 0.98, which was higher than that of CNN (0.95), DNN (0.96) and GAN (0.94), indicating that the proposed model had more advantages in image similarity preservation. At the same time, NMSE value was 0.02, which was significantly lower than those of CNN (0.05), DNN (0.04), and GAN (0.06), indicating that Anti-diffusion Net has a lower error rate and higher reconstruction accuracy in terms of reconstruction error.

In summary, the proposed Anti-diffusion Net not only has higher accuracy in detecting real and fake faces but also performs well in maintaining image quality while controlling reconstruction errors. This implies that this approach has potential for better distinguishing between authentic facial images from counterfeit ones, making it well-suited for practical solutions within facial anti-spoofing field.

6 Conclusion

The author introduces an innovative anti-spoofing technique for facial recognition that leverages a diffusion model combined with UNet-Attention enhancements. Through the diffusion model's ability to craft convincing fake faces that mimic genuine ones during the reverse diffusion phase, the diversity of the initial dataset is bolstered, thereby enriching the training material even further. The augmented data generated helps the model better capture complex patterns in fraud attacks compared to existing methods. Subsequently, a UNet network structure combined with attention mechanism was designed, which could extract multi-scale features more finely and effectively focus on important regions in the image. The findings show that the proposed methods are superior to the existing methods after testing. The findings underscore the scheme's viability in the realm of facial anti-spoofing, underscoring the necessity for bolstering the security and dependability of facial recognition tech. Despite the considerable advancements achieved in this research, there are lingering limitations to address. Firstly, despite the enhanced diversity of training data achieved through the diffusion model's generation of counterfeit images, there remains scope for enhancing the quality and genuineness of these fabricated images to more accurately replicate intricate real-world instances of fraudulent attacks. Secondly, existing celeb A datasets have certain limitations in the diversity and extensiveness of real attack types. Future studies should consider incorporating a wider range of datasets from various regions and devices in real-life situations to enhance the model's ability to adapt to different environments.

Future research directions can include the following aspects. First, in terms of datasets, it is necessary to further expand the datasets used for training and evaluation, especially cross-device, cross-cultural, and multimodal data. Secondly, the combination of multi-modal data, such as infrared and depth information, can further enhance the fraud prevention ability and robustness of the model in different environments.

Acknowledgments

Thanks for the advice and help from the instructor.

References

[1] J. Maatta, A. Hadid, and M. Pietikainen. (2011). Face spoofing detection from single images using micro-texture analysis. 2011 International Joint Conference on Biometrics (IJCB), 1 –7.

[2] J. Komulainen, A. Hadid, and M. Pietikainen. (2013). Context based face anti-spoofing. In Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference, 1–8.

[3] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. (2012). Lbp-top based countermeasure against face spoofing attacks. In Computer Vision-ACCV 2012 Workshops, Springer, 121–132.

[4] R. Cai, Z. Li, R. Wan, H. Li, Y. Hu, and A. C. Kot. (2022). Learning meta pattern for face anti-spoofing. IEEE Transactions on Information Forensics and Security, 17:1201–13.

[5] Kim, Seung-Hyun, Su-Min Jeon, and Eui Chul Lee. 2022. "Face Biometric Spoof Detection Method Using a Remote Photoplethysmography Signal" Sensors 22, no. 8: 3070.

[6] Y. Liu, A. Jourabloo, and X. Liu. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In Proceedings of the IEEE conference on computer vision and pattern recognition, 389–398.

[7] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo. (2017). Transfer learning using convolutional neural networks for face anti-spoofing. In Image Analysis and Recognition: 14th International Conference, 14: 27–34.

[8] A. Pinto, S. Goldenstein, A. Ferreira, T. Carvalho, H. Pedrini, and A. Rocha. (2020). Leveraging shape, reflectance and albedo from shading for face presentation attack detection. IEEE Transactions on Information Forensics and Security,15: 3347–58.

[9] J. Yang, Z. Lei, and S. Z. Li. (2014). "Learn Convolutional Neural Network for Face Anti-Spoofing," arXiv preprint arXiv:1408.5601.

[10] H. Gu, J. Chen, F. Xiao, Y.-J. Zhang, and Z.-M. Lu. (2023). "Self-Attention and MLP Auxiliary Convolution for Face Anti-Spoofing," IEEE, 11: 131152-167.

[11] Y. Wu, D. Tao, Y. Luo, J. Cheng, and X. Li. (2022). "Covered Style Mining via Generative Adversarial Networks for Face Anti-spoofing," Pattern Recognition, 132: 108957.

[12] Yang, Ruihan, Prakhar Srivastava and Stephan Mandt. “Diffusion Probabilistic Modeling for Video Generation.” Entropy 25 (2022): n. pag.

[13] B. Zhang, X. Zhu, X. Zhang and Z. Lei, "Modeling Spoof Noise by De-spoofing Diffusion and its Application in Face Anti-spoofing," 2023 IEEE International Joint Conference on Biometrics (IJCB), Ljubljana, Slovenia, 2023, pp. 1-10,

[14] Zhang, B., Zhu, X., Zhang, X., and Lei, Z., "Modeling Spoof Noise by De-spoofing Diffusion and its Application in Face Anti-spoofing," arXiv, 2024.