|  |  |
| --- | --- |
| ***E:\IIETA logo.jpg*** | **Traitement du Signal**  Vol., No., Month, Year, pp. \*\*-\*\*  Journal homepage: http://iieta.org/journals/ts |

|  |  |  |  |
| --- | --- | --- | --- |
| **Diffusion and Frequency Enhance Self-Blending Images to Deepfake Detection** | | |  |
| Peng Yanhan1\*, Bbb B. Surname2, Ccc C.C. Surname3 | | |  |
| 1 Affiliation, address  2 Affiliation, address  3 Affiliation, address  Corresponding Author Email: | | | |
| Copyright: ©2024 The author(s). This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/). | | | |
| https://doi.org/10.18280/ts.xxxxxx |  | **ABSTRACT** | |
|  |  |  | |
| **Received:**  **Revised:**  **Accepted:**  **Available online:** |  | Recent research on deepfake detection has achieved significant results when the training and testing datasets are derived from the same source. However, deepfake detection faces a critical generalization challenge: when the distribution of training and testing data is mismatched, performance degrades substantially. A widely accepted explanation is that these detectors tend to overfit specific artifacts of the forgeries, rather than learning features that generalize across different types of forgeries. It has been shown that forged data can assist in deepfake detection, and data augmentation is an effective approach to improve the generalization capability of detectors. In response, this paper proposes a novel method based on diffusion models and frequency domain features to enhance deepfake detectors through self-blended images. First, a central mask diffusion model is used for data augmentation on either the source or target image, followed by image blending. Finally, discrete cosine transform (DCT) is applied to extract low-frequency features from the self-blended image, which are then fused with the original image. Experimental results demonstrate that the proposed method improves the performance of deepfake detection models. | |
| ***Keywords:***  *Deepfake Detection, Diffusion model, key word 3, key word 4, key word 5, key word 6, key word 7, key word 8*  *(no more than 8 keywords)* |  |

# Introduction

With the rapid advancement of digital technology, the generation of fake images and videos has become unprecedentedly simple and fast, a transformation that has not only infused the entertainment industry with unparalleled creative vitality but also inadvertently paved the way for malicious activities. Among these technologies, deepfake technology stands out as particularly noteworthy due to its remarkable capability of creating lifelike fictional scenarios by subtly altering or entirely replacing the facial features of targeted individuals. Unfortunately, these meticulously manipulated data can subsequently be utilized to disseminate misleading information, unjustly harm innocent victims, or manipulate public opinion.

The majority of previous deepfake detectors [11, 12, 4, 6, 17] have demonstrated effectiveness within the confines of their respective datasets, yet they frequently encounter challenges when confronted with the cross-dataset scenario, where disparities in the distribution of training and testing data become apparent. In real-world scenarios, characterized by unpredictability and intricacy, one of the paramount criteria for a dependable and efficient detector is its generalization ability. This capability enables the detector to perform consistently and accurately across diverse data distributions, making it suitable for addressing the uncertainties and complexities inherent in practical applications. However, given that each forgery method typically possesses its specific characteristics, the overfitting to a particular type of forgery may impede the model's ability to generalize effectively to other types. Attempts have been made in recent arts to improve the generalizations. For example, to overcome dataset bias, some studies suggest data augmentation is an effective tool against poor generalization. These methods augment training data by synthesizing new face forgeries with their empirically designed augmentations. However, their augmentations are with a limited choice of strategy [15, 23, 25].

Recently, the diffusion model in generative models has also successfully achieved face forgery [26], making the detection task more difficult. However, diffusion models can also serve as a means of data augmentation. Though diffusion model has made great progress in image generation and editing, there are few research focus on training deepfake detection models with images generated by diffusion models.

In addition, directly using the images generated by diffusion models will lead to a series of problems, for example, making the detection model focus on unnecessary (e.g. non-facial) information, thus causing a performance decline when tested on other face datasets.to address this problem, Chen [27] propose a novel approach that extends the existing deepfake datasets with a diffusion model.

Given that the SBIs method already demonstrates considerable generalization capability, we seek to enhance its performance further through data augmentation and frequency-based enhancement.

To introduce more diversity into the synthetic data generated by SBIs, we adopt a central random mask diffusion model. This model effectively augments the source or target images. In the ablation studies of SBIs, the generalization ability of the detection model is closely tied to the data augmentation applied to both the source and target images.

Low-frequency information, by capturing the global features of the image, helps the model adapt to various scenes and lighting conditions, thereby improving its generalization capacity. In certain cases, low-frequency information may contain more general features, which are more transferable across different datasets, thus reducing overfitting. Consequently, incorporating low-frequency information enhances the generalization ability of the synthetic datasets generated by SBIs.

1. Related works

Deepfake generation can be broadly categorized into face-replacement and entire image synthesis techniques. Face-replacement typically involves identity swapping through autoencoder-based or graphics-based methods, while face-reenactment transfers the expressions from a source video to a target video, preserving the identity of the target individual. In contrast, entire image synthesis bypasses face-swapping processes like blending and directly generates fully synthetic facial images using generative models such as GANs and Diffusion models. While our work specifically focuses on detecting face-swapping forgeries, it also demonstrates potential in identifying fully synthesized images.

* 1. **Deepfake generation**

Deepfake generation typically involves face-replacement [10], face reenactment [5, 1], and entire image synthesis [3]. Face-replacement generally involves the ID swapping utilizing the auto-encoder-based [10] or graphics-based swapping methods, whereas face-reenactment utilizes the reenactment technology to swap the expressions of a source video to a target video while maintaining the identity of the target person. In addition to the face-swapping forgeries above, entire image synthesis utilizes generative models such as GAN [3] and Diffusion models [8, 22] to generate whole synthesis facial images directly without face-swapping operations such as blending. Our work specifically focuses on detecting face-swapping but also shows the potential to detect entire image synthesis.

* 1. **Deepfake detection**

Recent studies have made various attempts for deepfake detection and achieve remarkable success. Several arts discuss low-level differences between pristines and forgeries and suggest using them as classification clues. Li et al. assume blending artifacts exist in all pristines and suggest finding the blending boundary beside the detection task; Qian et al. and Luo et al. use high-frequency details as additional inputs for their models; Liu et al. adopt phase spectrums to capture the up-sampling artifacts of face forgery for the task. Despite their effectiveness in many cases, the low-level artifacts are sensitive to post-processing steps which vary in different datasets, thus limiting their generalization. Some works suggest borrowing features from other tasks for deepfake detection. Such as the features from lips reading, facial image decomposition , and landmark geometric features. Although these features can bring promising improvements, their generalization performance to the deepfake data are questionable. While annotating the deepfake data for these tasks is rather expensive, incorporating these tasks often lead to limited improvements.

The task of deepfake detection grapples profoundly with the issue of generalization. Recent endeavors can be classified into the detection of image forgery and video forgery. The field of detecting image forgery have developed novel solutions from different directions: data augmentation [19, 9, 23, 18], frequency clues [21, 14, 16].

* 1. **Data Augmentation**

One effective approach in deepfake detection is the utilization of data augmentation, which involves training models using synthetic data. For instance, in the early stages, FWA employs a self-blending strategy by applying image transformations (e.g., down-sampling) to the facial region and then warping it back into the original image. This process is designed to learn the wrapping artifacts during the deepfake generation process. Another noteworthy contribution is Face X-ray, which explicitly encourages detectors to learn the blending boundaries of fake images. Similarly, I2G [13] uses a similar method of Face X-ray [9] to generate synthetic data and then employs a pair-wise self-consistency learning technique to detect inconsistencies within fake images. Furthermore, SLADD [20] introduces an adversarial method to dynamically generate the most challenging blending choices for synthesizing data. Rather than swapping faces between two different identities, a recent art, SBI [23], proposes to swap with the same person's identity to reach a high-realistic face-swapping. Hu et al.[24] utilized a masked autoencoder to reconstruct missing regions based on the remaining facial parts. In contrast to these approaches, Chen [27] employed a masked conditional diffusion model for deepfake data augmentation, improving the authenticity of generated images at both the pixel and feature levels.

* 1. **Diffusion model**

Diffusion models, a new family of generative models, have recently achieved state-of-the-art results. Unlike earlier generative models such as GANs and VAEs, which often face instability during training, diffusion models optimize using score matching, typically implemented with a simple mean squared error (MSE) loss. This stable training process allows diffusion models to more effectively capture conditional data distributions. Given their controllability and ability to produce high-fidelity outputs, diffusion models represent a promising direction for further exploration in generative modeling.

* 1. **Discrete Cosine Transform**

The Discrete Cosine Transform is similar to the Discrete Fourier Transform, but the key difference is that it is a transformation in the real domain. Compared with the Discrete Fourier Transform that requires complex operations, the Discrete Cosine Transform can improve the operation speed and can be widely used in real-time scenarios. In digital image processing, it can be used for image compression, transforming signals in the spatial domain into the frequency domain. At the same time, because its process is lossless, the inverse transform can be performed after the Discrete Cosine Transform to restore the original image information. The discrete cosine transform has good decorrelation and can gather important information, so it has a wide range of applications in natural signal processing with high correlation such as images.

The calculation formula for two-dimensional discrete cosine transform is shown in equation (1):

|  |  |
| --- | --- |
|  | (1) |

Among them, is a two-dimensional matrix of size **in the spatial domain, andrepresents the matrix obtained after two-dimensional discrete cosine transform. When processing raw signals with strong correlation, such as images, audio, etc., the coefficient energy will be concentrated in the upper left corner, and most of the remaining coefficients will be almost zero.

1. Proposed method

Our model has two branches in data processing, one using SBI and the other using Mask Diffusion followed by SBI, Finally, concatenate the two data branches into one block; In terms of model training, we use a dual stream network that combines RGB features and frequency features. As shown in the figure.

* 1. **Self-Blended Image**

The forgery method proposed by Face X-ray relies on facial blending operations, utilizing two different faces. When blending, it requires selecting images with similar facial shapes. However, the Self-Blended Images (SBI) method eliminates this requirement by generating both the fake source and target  images from a single original image, followed by a blending operation. To enhance the diversity of forgeries, SBI applies multiple image transformations. The SBI method consists of three key stages that synthesize data from the original image.

In the first stage, the Source Target Generator (STG) generates forged source  and target  images. These images are created by applying random transformations to the original image, aiming to generate two images with mismatched landmarks.

In the second stage, a Mask Generator (MG) enhances the grayscale mask image. The MG uses the Face X-ray method to detect facial landmarks, calculates convex hulls from the landmarks, and then applies operations such as elastic deformation, Gaussian blur, and erosion to produce the mask.

In the third stage, the source and target images are blended using the generated mask. These steps can be summarized as follows:

|  |  |
| --- | --- |
|  | (1) |

where  is the source image,  is the target image, and M is the generated mask.【插一张SBI图】

* 1. **Mask-Conditioned Diffusion**

Diffusion models can generate high-quality images with spatial continuity; however, using diffusion models to enhance forged images may lead to changes throughout the entire image. For instance, enhancing Self-Blended Images (SBI) with diffusion models can result in the degradation of blending details, making it difficult for the classifier to learn these intricate details. Therefore, we choose to apply diffusion-based image enhancement only to either the source or target images of SBI, setting the random number to 0.5. We then adopt a masked diffusion approach, focusing primarily on central mask diffusion within the facial region and generating random masks based on this.

After performing masked blending on the source image , we add random noise *c* to the masked area and utilize a U-Net model to predict this noise. Subsequently, we carry out denoising and restoration of the image to obtain the generated output. The specific formulas are as follows:

|  |  |
| --- | --- |
|  | (2) |

Applying an optimized Markovian noise addition process to

image x0, we obtain an image x\_t with the help of the adaptive

combine module. These processes can be expressed as follows:

|  |  |
| --- | --- |
|  | (2) |

【插一张Hybrid的mask扩散图】【再去看看公式】

* 1. **Frequency Features Generator**

The Discrete Cosine Transform (DCT) is an effective method for converting signals from the spatial domain to the frequency domain. It compresses the local information in an image into a small number of low-frequency components, preserving the main structural features of the image.

The formula for performing the DCT on a two-dimensional image signal  is given by:

|  |  |
| --- | --- |
|  | (2) |

Where and are normalization factors defined as:

|  |  |
| --- | --- |
|  | (2) |

Here, *N* and *M* are the width and height of the image, respectively, and are the coordinates in the frequency domain.

The input image is separated into three channels: Red (R), Green (G), and Blue (B). Each channel image can be represented as , and . The DCT is applied to each channel (where *C* denotes the channel identifier), yielding the frequency domain representation . The low-frequency components are extracted from the DCT results, typically by selecting the top-left submatrix. These low-frequency components contain the primary information of the image. The extracted low-frequency components are scaled, typically by dividing their values by a constant K (e.g., 16), to adjust their influence on the final image:

|  |  |
| --- | --- |
|  | (2) |

The scaled low-frequency components are then subjected to an inverse DCT, resulting in the reconstructed low-frequency image :

|  |  |
| --- | --- |
|  | (2) |

The low-frequency reconstructed image is blended with the original channel image to preserve the details of the original image:

|  |  |
| --- | --- |
|  | (2) |

Finally, the mixed channels and  are merged back into the final RGB image.

【插一张频率转换低频混合图】【可以贴个滤波和DCT变换的公式】

# Experiments

* 1. **Setting**
     1. Dataset

We adopt the widely used benchmark FaceForensics++[12] (FF++) for training, following the convention. It contains 1,000 original videos and 4,000 fake videos forged by four manipulation methods, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). For our cross-dataset evaluation, we use five recent deepfake datasets. Celeb-DF-v2 [10] (CDF) applies a more advanced deepfake technique to celebrity videos downloaded from YouTube. DeepFakeDetection (DFD) provides thousands of deepfake videos generated with consenting actors. DeepFake Detection Challenge Preview [2] (DFDCP) and DeepFake Detection Challenge [7] public test set (DFDC), that are released along with the competition , contain a lot of disturbed videos, e.g., compression, down-sampling, and noise.

* + 1. Evaluation Metrics

We report the video-level area and frames-level area under the receiver operating characteristic curve (AUC) to compare with prior works.

* + 1. Frames baseline

Face X-ray, DSP-FWA, UCF, F3Net, SPSL

* + 1. Videos baseline

LipForensics, FTCN

* 1. **Implementation Details**
     1. Preprocess

We adopt Dlib and RetinaFace to extract facial landmarks and bounding boxes from each video frame, respectively. We use an 81 facial landmarks shape predictor in Dlib.

* + 1. Mask diffusion

【选用pattle 的 hybrid,用celebahq的200轮预训练权重进行训练，训练500个epoch，然后对source 进行推测】

* + 1. Training

We adopt the state-of-the-art convolutional network architecture EfficientNet-b4 (EFNB4) pre-trained on ImageNet as the classifier and train it for 100 epochs with the SAM optimizer. The batch size and learning rate are set to 16 and 0.001, respectively. We sample only eight frames per video for training. If two or more faces are detected in a frame, the face with the largest area of the face bounding box is extracted. Each batch consists of real images and their SBIs, and the same augmentation is applied to each real image and its SBI.

* + 1. Inference Strategy.

We sample 32 frames per video for inference. If two or more faces are detected in a frame, the classifier is applied to all faces and the highest fakeness confidence is used as the predicted confidence for the frame. Once the predictions for all frames are obtained, we average them to get the prediction for the video. For fair comparison, we use all videos of all test sets for evaluation by setting the confidences to 0.5 for the videos where no face is detected in all frames.

* 1. **Results**

All our experiments follow a commonly adopted generalization evaluation protocol by training the models on the FF++ c23 and then evaluating on other previously untrained/unseen datasets (e.g., CDF and DFDC).

* + 1. Cross-dataset evaluation

To show the generality of our method, we conduct a cross-dataset evaluation where models are trained on FF++ and evaluated on other datasets.

* + 1. Cross-manipulation evaluation

we evaluate our model on four manipulation methods of FF++, i.e., DF, F2F, FS, and NT. We use the raw version for evaluation as well as the competitors.

* 1. **Ablation**

**Effect of Each Augmentation Process**

Diff+Freq

* 1. **Visualization**

4.5.1 T-SNE

We then apply t-SNE visualization to feature vectors from the last layers of the models. We emphasize again that it is easy for the baseline to recognize the forged faces because they are seen in its training, and that our goal is to separate real faces from others, not to classify types of manipulations.

4.5.2 Grad-cam

To visualize where the models are paying their attention on the forged faces, we apply Grad-CAM to the models on manipulated frames of FF++, i.e., DF, F2F, FS, and NT, as shown in Fig. 5. It can be observed that our method encourages the model to make its attentions sparser than the baseline. This is because our model detects minor artifacts independent of manipulations, e.g., blending boundaries, while the baseline captures method-specific pixel distributions that are widely spread in the forged faces.

1. **CONCLUSIONS**

It is mandatory to have conclusions in your paper. This section should include the main conclusions of the research and a comprehensible explanation of their significance and relevance. The limitations of the work and future research directions may also be mentioned. Please do not make another abstract.

**REFERENCES**

[1] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387-2395.

[2] B. J. a. p. a. Dolhansky, "The dee pfake detection challenge (DFDC) pre view dataset," 2019.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401-4410.

[4] E. Sabir, J. Cheng, A. Jaiswal, W. Abdalmageed, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," 2019.

[5] J. Thies, M. Zollhöfer, and M. J. A. T. o. G. Nießner, "Deferred neural rendering: Image synthesis using neural textures," vol. 38, no. 4, pp. 1-12, 2019.

[6] X. Yang, Y. Li, and S. J. I. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," 2019.

[7] B. Dolhansky *et al.*, "The deepfake detection challenge (dfdc) dataset," 2020.

[8] J. Ho, A. Jain, and P. J. A. i. n. i. p. s. Abbeel, "Denoising diffusion probabilistic models," vol. 33, pp. 6840-6851, 2020.

[9] L. Li *et al.*, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001-5010.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207-3216.

[11] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues," 2020.

[12] A. Rssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[13] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning Self-Consistency for Deepfake Detection," 2020.

[14] H. Liu *et al.*, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772-781.

[15] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing Face Forgery Detection with High-frequency Features," 2021.

[16] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16317-16326.

[17] H. Zhao, W. Zhou, D. Chen, T. Wei, and N. Yu, "Multi-attentional Deepfake Detection," 2021.

[18] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15023-15033.

[19] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18710-18719.

[20] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection," 2022.

[21] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 1, pp. 735-743.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684-10695.

[23] K. Shiohara and T. J. a. e.-p. Yamasaki, "Detecting Deepfakes with Self-Blended Images," 2022.

[24] J. Hu *et al.*, "Mover: Mask and recovery based facial part consistency aware method for deepfake video detection," 2023.

[25] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22412-22423.

[26] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568-8577.

[27] T. Chen *et al.*, "Masked conditional diffusion model for enhancing deepfake detection," 2024.