Noise Based Deepfake Detection via Multi-Head Relative-Interaction

Tianyi Wang, Kam Pui Chow*

Department of Computer Science, The University of Hong Kong, Hong Kong, China
{tywang, chow}@cs.hku.hk

Abstract

Deepfake brings huge and potential negative impacts to our daily lives. As the real-life Deepfake videos circulated on the Internet become more authentic, most existing detection algorithms have failed since few visual differences can be observed between an authentic video and a Deepfake one. However, the forensic traces are always retained within the synthesized videos. In this study, we present a noise-based Deepfake detection model, NoiseDF for short, which focuses on the underlying forensic noise traces left behind the Deepfake videos. In particular, we enhance the RIDNet denoiser to extract noise traces and features from the cropped face and background squares of the video image frames. Meanwhile, we devise a novel Multi-Head Relative-Interaction method to evaluate the degree of interaction between the faces and backgrounds that plays a pivotal role in the Deepfake detection task. Besides outperforming the state-of-the-art models, the visualization of the extracted Deepfake forensic noise traces has further displayed the evidence and proved the robustness of our approach.

Introduction

The video of a synthesized Barack Obama giving a speech insulting the former president of the United States, Donald Trump, is widely spread on YouTube[1]. Without knowing the truth that the face is synthesized using Jordan Peele's, people would possibly get tricked and believe it to be genuine. The video is generated via deep neural networks that perform face identity swap and generate hyper-realistic fake videos appearing authentic, also known as Deepfake (Chawla 2019; Maras and Alexandrou 2019). Deepfake is first introduced by the Reddit user 'deepfakes' in 2017, utilizing deep neural networks to swap a source person's facial identity onto the target one, maintaining the target person's facial expression.

Since the first occurrence of the Deepfake face identity swap technique, methods have been explored to perform Deepfake detection with the help of deep neural networks (Afchar et al. 2018; Nguyen, Yamagishi, and Echizen 2019; Zhao et al. 2021; Zhang et al. 2022; Hu et al. 2022b; Wang et al. 2022; Hu et al. 2022a; Cheng et al. 2022; Wang and Chow 2022). In specific, most of them focus on the video itself and boost their performance through the techniques of computer vision. While the hyper-realistic synthesized faces are hard to find differences from the authentic ones visually, forensic traces are left within the face area regardless of different face identity swap, fine-tuning, or smoothing techniques. On the contrary, the background area in each image frame is usually unmodified as the purpose of Deepfake is face identity swap, and the less the original video is modified, the more authentic the synthesized one is likely to be. However, besides the studies employing the Photo Response Non-Uniformity (PRNU) (Lukas, Fridrich, and Goljan 2006) but failing in the Deepfake detection task, few approaches resort to forensic traces such as noise. Moreover, many existing methods do not utilize video keyframes for Deepfake detection, leading to huge information loss. This is because the keyframes contain the most integrated video information after common video compression.

In this work, we present a novel noise-based Deepfake detection method, NoiseDF for short. In particular, we study the underlying forensic noise traces of the Deepfake videos. We crop the face and a background square from each video keyframe and investigate the different noise patterns between real and fake ones, given the background squares are unmodified. Also, we adopt the Siamese (Bromley et al. 1993) architecture and train the enhanced RIDNet denoiser (Anwar and Barnes 2019) to extract the underlying Deepfake forensic noise traces from the face and background squares. Thereafter, we propose a new Multi-Head Relative-Interaction method to measure the degree of interaction within each face-background pair in multiple views of dimension and perform Deepfake detection accordingly. The specialty of the Siamese design is to share the learnable weights for both branches so that heterogeneous inputs lead to distinct output features. In specific, the level of relative-interaction is higher for a real face-background pair than that for a fake face-background pair since only the fake face contains Deepfake forensic noise traces. We further apply depth-wise separable convolution for the projection of the noise features to overcome the efficiency decay of the convolutional neural network (CNN). Overall, our proposed NoiseDF approach achieves promising Deepfake detection performance against many existing state-of-the-art baseline methods. Furthermore, we have visualized the ex-

*Corresponding author.

[1]https://www.youtube.com/watch?v=cQ54GDm1eL0

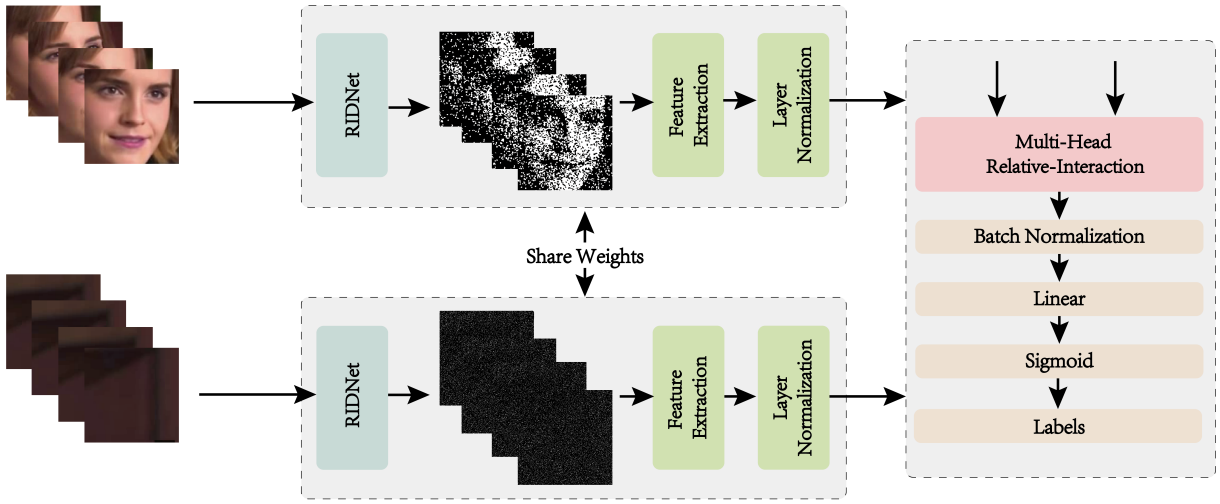

Figure 1: Workflow of the proposed NoiseDF model. Cropped face and background squares are passed through the Siamese architecture for noise feature extraction. Thereafter, the two sets of noise features are computed to obtain the degree of Multi-Head Relative-Interaction to perform Deepfake detection.

tracted Deepfake noise traces to verify the robustness of our proposed model in distinguishing fake and real faces.

The contributions of this work are threefold:

- Significantly distinguished from the traditional approaches relying on computer vision techniques, our study is the first to achieve good Deepfake detection performance in the perspective of digital forensic noise traces. We further visualize the Deepfake forensic noise traces, which have not been achieved by the computer vision based detection models.
- We present a novel idea of extracting the face-background pairs via the Siamese structure towards Deepfake noise features analyses. Meanwhile, we devise a Multi-Head Relative-Interaction method that justifies the level of interaction and similarity between the face and background noise features in multiple perspectives of head dimensions regarding the variants of the original video authenticity.
- We emphasize the keyframe importance in videos and integrate them for the Deepfake detection task while many existing detection algorithms have neglected the potential performance improvements brought by the keyframes. The proposed NoiseDF approach achieves state-of-the-art performance against the comparative baseline methods for both in-dataset and cross-dataset experiments.

Noise Based Deepfake Detection

Although noise trace is not new in the digital forensics domain for various tasks and applications, it has been barely discussed in Deepfake detection. The noise-based Deepfake detection approaches up to date are mostly relying on the Photo Response Non-Uniformity (PRNU), a noise pattern created by small factory defects in the light-sensitive sensors of a digital camera (Lukas, Fridrich, and Goljan 2006). PRNU has shown strong abilities in source device identification (Marra et al. 2017; Saito, Tomioka, and Kitazawa 2017) and source anonymization (Picetti et al. 2022). However, none of the PRNU-based work has shown strong evidence that the PRNU noise can be used for Deepfake detection. Koopman et al. (Koopman, Macarulla Rodriguez, and Geradts 2018) claimed that the mean normalised cross correlation score of PRNU noise per video can be used to distinguish Deepfakes from authentic videos by an experiment with only 10 videos in total given the correct video labels. But the inversion, Deepfake detection, could not be performed without knowing the correct labels. Weever and Wilczek (de Weever and Wilczek 2020) made several experiments calculating the correlation of the PRNU noise and found out that none of the PRNU noise analyses had resulted in a definite proof of real or fake. In summary, the PRNU noise pattern can be a useful tool for device identification studies, but it is not a good forensic noise tracing material for Deepfake detection. Therefore, our study is the first to achieve good performance using a forensic noise trace based Deepfake detection approach.

Methodology

In this section, we illustrate our novel noise-based Deepfake detection model in four parts, namely, the data preprocessing idea on keyframes and face-background pair cropping, the Siamese Noise Feature Extraction module, the Multi-Head Relative-Interaction method, and the final Deepfake detection. The workflow of our approach is shown in Figure 1.

Data Preprocessing

Keyframes are first extracted from the videos in the dataset. After that, the face square and a background square are cropped from each keyframe. Details of a data preprocessing example are shown in Figure 2.

Figure 2: An example of data preprocessing including keyframe extraction and face and background squares cropping. The face square in a keyframe is located and cropped, and a background square with the same size and the furthest Euclidean distance from the detected face within the image frame is found.

Keyframe Extraction Image frames within a video are usually under video compression for the purpose of space-saving. Specifically, three types of frames are commonly included in a video clip after video compression, I-frame, P-frame, and B-frame. I-frame, also known as the keyframe or intra-frame, is a complete image with the largest size that contains intact information and plays the most important role in video compression. P-frame, the predicted picture, holds only the variations in the current image frame from the previous image frames within a video. B-frame, the bidirectional predicted picture, records only the changes in the current image frame from both the preceding and following image frames to specify the content and saves even more space than the P-frame (Vijayanagar 2020). In other words, both the B-frame and P-frame are not complete image frames and lack image information as compared to the I-frame (keyframe). The number of keyframes within a video varies depending on the video quality and content motion complexity. In this study, the FFmpeg tool is adopted for keyframe extraction from the videos in the datasets.

Face and Background Extraction The purpose of Deepfake is to accomplish identity swap while maintaining as much frame area unchanged as possible within the video and image frames to guarantee authentic looks. Therefore, a Deepfake video normally has only the face area modified for identity swap, and most of the background area is unchanged. For each extracted image frame, we first perform face detection using dlib[2] library and crop the square area containing the face. It is uncontrollable that the adjacent area of the face square is totally unmodified since some Deepfake techniques make unavoidable changes to the background area by warping back a face square. Thus, we search for the background area with the largest distance from the detected face area to ensure that the extracted background square is unmodified. We crop the background square with the same size as the face square, and the Euclidean distance between the central points of the squares is considered when locating the furthest background square. For each face-background pair from each image frame, the face square is manipulated if it is extracted from a Deepfake synthesized video, and the background one is always unchanged regardless of the authenticity of the video it comes from. The ground-truth label for each face-background pair is determined based on the authenticity of the face in the source video.

Siamese Noise Feature Extraction

We employ the Siamese Network structure for noise feature extraction from the face and background squares, and force both Siamese branches sharing the weights of the noise feature extraction network. Within the Siamese architecture, a pre-trained RIDNet is adopted and improved for Deepfake forensic noise trace extraction, followed by further feature extractions from the extracted noise traces.

RIDNet Noise Extraction Comparing to the existing denoisers (Guo et al. 2019; Zhang et al. 2017), the single-stage RIDNet (Anwar and Barnes 2019) is proven to be more efficient and flexible and can handle both variant and invariant noises with better performance regardless of the noise standard deviation. RIDNet is composed of a single convolutional layer for feature extraction, four cascaded enhancement attention modules (EAM) for feature learning, and another convolutional layer for the reconstruction of a clean output image. In detail, the EAMs follow the residual-on-the-residual architecture and are mainly composed of convolutions while the novel idea is the utilization of channel attention for emphasizing the weights of important features. In this study, the face and background squares are passed through the enhanced RIDNet model to extract the underlying forensic noise traces instead of eliminating them. We exploit the pre-trained weights of the RIDNet that can firstly extract a general level noise n by

$$n = x - \hat{y}, \tag{1}$$

[2]https://pypi.org/project/dlib

a subtraction of the output clean image $\hat{y}$ from the original input noisy image x. Then, we train this noise extractor with additional network architectures and further restrictions for weight updating and finally achieve Deepfake forensic noise trace extraction that serves for our Deepfake detection purpose. The extracted face and background noise traces are passed through convolutional layers for further Deepfake noise feature extraction and dimension adjustment. A layer normalization is followed to restrict the model training direction and ensure training efficiency.

Siamese Network Siamese Network (Bromley et al. 1993) is first introduced in 1993 for signature verification purposes based on cosine similarity of the signature features such as the curvature of the trajectory and the acceleration. The characteristic of this structure is the shared weights for both branches, and corresponding similar or unique output features are derived depending on the inputs to the branches accordingly. A later study has utilized the idea of the Siamese Network for face verification (Chopra, Hadsell, and LeCun 2005) by minimizing a discriminative loss function that makes the similarity metric small for pairs of faces from the same person, and large for pairs from different persons. In this study, the face and background squares from an image frame are respectively passed through branches of the Siamese architecture for noise feature extraction. An authentic video image frame is unmodified and contains the same kind of clean forensic noise pattern everywhere within the image, while on the contrary, a Deepfake video image frame has the face area synthesized and warped so that the Deepfake noise pattern of the face area is different from that of the unchanged background area.

The two branches of the Siamese Network share the same weights. Considering the background squares stay authentic and unchanged, the model is tuned such that different noise patterns are extracted from the faces and backgrounds of the Deepfake synthesized videos, and the same noise patterns are extracted from that of authentic ones. The two branches of the Siamese Network are coded as one single branch in implementation since both branches share the same architecture and the same set of network weights. The noise patterns are further analyzed with a novel Multi-Head Relative-Interaction method to perform Deepfake detection.

Multi-Head Relative-Interaction Using Depth-Wise Separable Convolutions

The traditional cosine similarity method used in the early Siamese Network related research has achieved good efficiencies. However, a one-step dimension reduction to make two high-dimensional Deepfake noise trace features become one single cosine similarity numerical value can cause enormous information loss and deteriorate the performance of Deepfake detection. Therefore, we devise a novel Multi-Head Relative-Interaction approach to study the interaction and similarity between the face noise features and background noise features. Furthermore, a depth-wise separable convolutional design is applied to the projection of Multi-Head Relative-Interaction to speed up and compensate for the possible efficiency damping. The architecture detail of the Multi-Head Relative-Interaction is shown in Figure 3.

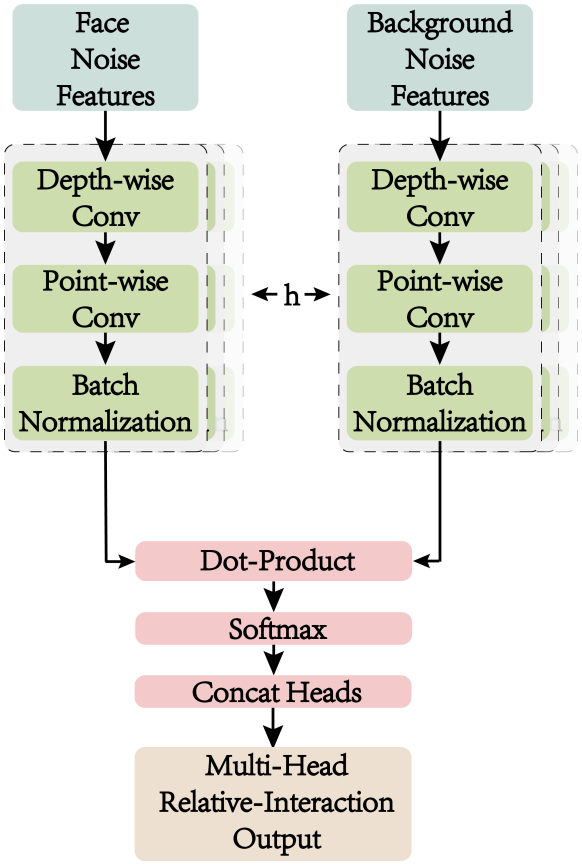

Figure 3: Illustration of the Multi-Head Relative-Interaction module.

Multi-Head Relative-Interaction We propose a novel Multi-Head Relative-Interaction approach to analyze the features of face and background noise traces in multiple dimensions of view. The Relative-Interaction can be described as mapping the face and the background noise trace features each to an output, where the mapped face noise trace feature F_f and background noise trace feature F_b are generated through learnable projections W_f and W_b with CNNs on the noise trace features N_f and N_b extracted by the Siamese architecture by

$$F_f, F_b = W_f(N_f), W_b(N_b). \tag{2}$$

We compute the dot-product of the projected noise trace features F_f and F_b, divide the result by $\sqrt{d_F}$, the square root of the projected noise trace feature dimension, and apply a softmax function to obtain the weights on the interaction and similarity between the noise trace features F_f and F_b. The overall Relative-Interaction output is computed by

$$\text{RelativeInteraction}(F_f, F_b) = \text{softmax}(\frac{F_f F_b^T}{\sqrt{d_F}}). \tag{3}$$

The purpose of the division by $\sqrt{d_F}$ is to prevent value explosion after the dot-product of the projected noise trace features F_f and F_b.

Model Name	Test Datasets							
	FF++		DFDC		Celeb-DF		DF-1.0	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MesoNet (Afchar et al. 2018)	61.03	58.13	50.02	50.16	36.73	50.01	50.05	50.21
Capsule (Nguyen, Yamagishi, and Echizen 2019)	76.40	83.44	51.30	56.16	61.96	59.93	59.29	61.46
FFD (Dang et al. 2020)	82.29	82.48	59.44	59.47	46.19	55.86	53.69	53.81
CViT (Wodajo and Atnafu 2021)	84.36	93.10	54.36	59.23	46.51	54.33	50.75	51.76
MAT (Zhao et al. 2021)	76.07	86.42	56.99	60.84	62.94	64.99	**69.34**	69.47
Two-Stream (Luo et al. 2021)	80.69	89.19	54.87	57.13	58.30	66.95	60.86	63.60
TAR (Lee et al. 2021)	53.48	50.00	49.85	50.00	63.29	50.00	49.95	50.00
NoiseDF (Ours)	**84.36**	**93.99**	**59.87**	**63.89**	**70.10**	**75.89**	67.49	**70.88**

Table 1: Frame-level comparative tests accuracy (%) and AUC scores (%) on the testing datasets after trained on FF++.

A Multi-Head Relative-Interaction expands the Relative-Interaction to multiple perspectives and dimensions of view when analyzing the Deepfake forensic noise trace features. As the learnable convolutional projection weights are initialized randomly and differently for each head, a concatenation of outputs from h heads by

$$\text{head}_i = \text{RelativeInteraction}(W_{fi}(N_f), W_{bi}(N_b)), \quad (4)$$

and

$$\text{MultiHead}(F_f, F_b) = \text{Concat}(\text{head}_1, ..., \text{head}_h), \quad (5)$$

brings a broader view of the noise features than a single-head Relative-Interaction on N_f and N_b, where the i-th head head_i is computed with face and background projections W_{fi} and W_{bi} on N_f and N_b, respectively.

Depth-wise Separable Convolution The depth-wise separable convolution idea is introduced by Chollet (Chollet 2017) to avoid the possible overfitting and time-consuming problem of CNNs during model training. The depth-wise separable convolution operates the same procedure as a regular convolution does but utilizes much fewer parameters and strengthens the efficiency. In our NoiseDF model, we propose depth-wise separable convolutions instead of standard convolutional layers for the learnable convolutional projections upon the face and background noise trace features and all other convolutions in the model. As a result, a more efficient model is obtained with fewer parameters and fewer chances of overfitting with the help of the depth-wise separable convolution.

Deepfake Detection

Following the innovative Multi-Head Relative-Interaction approach, we apply batch normalization (Ioffe and Szegedy 2015) along with refinement operations to adjust the output shape and gradually reduce the dimension using fully connected layers for a final sigmoid calculation by

$$\sigma(x) = \frac{1}{1 + \exp(-x)}, \quad (6)$$

where x is the output of the final fully connected layer, to generate the ultimate prediction results upon the candidate videos for the Deepfake detection. During the training process, the entire network is tuned according to the loss values of each training epoch. Weights of the improved pre-trained Siamese Noise Extraction Network module are updated to extract the Deepfake noise traces and gradually satisfy our demand.

Experiments

Datasets

FaceForensics++ (FF++) (Rössler et al. 2019) is currently the most widely adopted training dataset in the existing Deepfake detection work. It contains 1,000 real videos and four fake video subsets each contains 1,000 fake videos synthesized from the 1,000 real ones using FaceSwap (FS), Deepfakes (DF), Face2Face (F2F) (Thies et al. 2016), and NeuralTextures (NT) (Thies, Zollhöfer, and Nießner 2019) manipulation techniques, respectively. Three qualities have been released, namely, Raw, HQ (c23), and LQ (c40), where the latter two are compressed using the H.264 codec with different compression levels. We chose the HQ (c23) dataset as our training dataset because it is similar to the real-life Deepfake video quality, and followed the official split ratio to extract a balanced image frame dataset for real and fake.

The purpose of the Deepfake detection task is to prevent Deepfake attacks with unknown manipulation techniques from affecting human lives with the help of the existing Deepfake video datasets. As the real-life Deepfake videos have become harder to distinguish, we also chose the datasets with better qualities for cross-dataset evaluation on the proposed model. In specific, we tested the well-trained model performance on Deepfake Detection Challenge (DFDC) (Dolhansky et al. 2019), Celeb-DF (Li et al. 2020), and DeeperForensics-1.0 (DF-1.0) (Jiang et al. 2020) datasets.

Denoisers	Test Datasets							
	FF++		DFDC		Celeb-DF		DF-1.0	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
CBDNet (Guo et al. 2019)	73.29	81.46	**61.94**	**64.04**	65.23	62.80	**69.84**	**73.05**
DnCNN (Zhang et al. 2017)	84.13	92.41	58.34	61.62	61.55	71.63	56.58	54.82
RIDNet (Anwar and Barnes 2019)	**84.36**	**93.99**	59.87	63.89	**70.10**	**75.89**	67.49	70.88

Table 2: Ablation study on adopting different denoisers in our detection model.

Experimental Settings

We adopted $h = 8$ heads with head dimension 64 for the Multi-Head Relative-Interaction method and resized the lengths for all input face and background squares to 64 for consistency during training. The proposed model is trained with a batch size of 16 and a learning rate of $5e-5$. Experiments are conducted on a Tesla V100 GPU. We evaluated the Deepfake detection performance using the overall correctness accuracy and the area under the receiver operating characteristic (ROC) curve (AUC) score at the frame level for testing.

Results

Model Evaluation We only considered the state-of-the-art Deepfake detection algorithms that have source code published and reproducible in the comparative test for a fair game. We maintained the same experimental settings as ours while keeping their optimal parameter settings whenever applicable for training and testing.

We first trained the proposed NoiseDF Deepfake detection model with the FF++ training set and evaluated the model performance on the FF++ testing set. Then, we ran Deepfake detection using the well-trained model on the other high-quality datasets for cross-dataset evaluation. The state-of-the-art baseline models with source codes published for comparative tests are trained and tested on the same datasets as ours while maintaining their original optimal experiment settings when applicable. As a result, our novel NoiseDF model outperforms the state-of-the-art Deepfake detection baselines for both in-dataset and cross-dataset experiments with respect to accuracy and AUC scores. Details are shown in Table 1.

As listed in Table 1, most models perform well on FF++ for the in-dataset experiment except the TAR method and a few early approaches. Meanwhile, our proposed NoiseDF model still slightly outperforms all of them on FF++. The performance damping has occurred for all models in the cross-dataset evaluation. The DFDC dataset is observed to be the hardest dataset because it contains Deepfake videos with 8 different facial manipulation techniques that the well-trained models have not seen. Celeb-DF, although brings huge challenges once released, is becoming easier to be overcome by the recent computer vision approaches such as Two-Stream and MAT by extracting plenty of image features due to its high resolution.

DF-1.0, adopted for the purpose of attacker countermove testing, contains a testing set with different levels of artificial noise trace perturbations and distortions added to the videos. In Table 1, our model is the only one to achieve the highest AUC score over 70% on DF-1.0 against all other comparative baseline methods although it ranks number two in the accuracy evaluation. A high AUC score demonstrates the probability that a random positive sample scores higher than a random negative sample from the testing set, in other words, the ability of our well-trained classifier to distinguish between real and fake faces. The reason for a relatively lower accuracy than the AUC score is that the threshold to classify real and fake is always fixed at 0.5 for the accuracy evaluation upon the sigmoid output scores, while the real threshold for the optimal model performance is usually different from 0.5. Therefore, despite a classification with the threshold value set to 0.5 does not perform well, the proposed model is still able to distinguish between real and fake on DF-1.0 according to the achieved highest AUC score. In conclusion, although introducing random levels of noise trace perturbations and distortions, DF-1.0 does not bring significant challenges to our proposed detection model that relies on the Deepfake forensic noise traces.

In addition, earlier approaches with pure CNN backbones have shown poor transferability in cross-dataset evaluation. The reason might be that the pure CNN architecture is not robust for the generalization on other datasets. On the contrary, recent detection models such as Two-Stream and MAT with attention mechanisms added consider both local and global image features, illustrating considerably high performance in both in-dataset and cross-dataset experiments. The TAR model is proved to be very time-consuming and is observed that it incorrectly labels all input faces to be authentic, which causes bad performance on all testing datasets as shown in Table 1.

Deepfake Forensic Noise Trace Visualization Recent Deepfake detection models using computer vision techniques have attempted to visualize the dominant image features that determine the final predictions of the models. However, although the visualized heatmaps are able to locate the determining parts within the face area, they have exhibited similar feature patterns for both real and fake. In this section, we sampled some real and fake faces from the testing sets and visualized the heatmaps of the image features extracted by the recent Two-Stream and MAT computer vision models. As Figure 4(a) shows, the top three rows display the images and the heatmaps of the correspond-

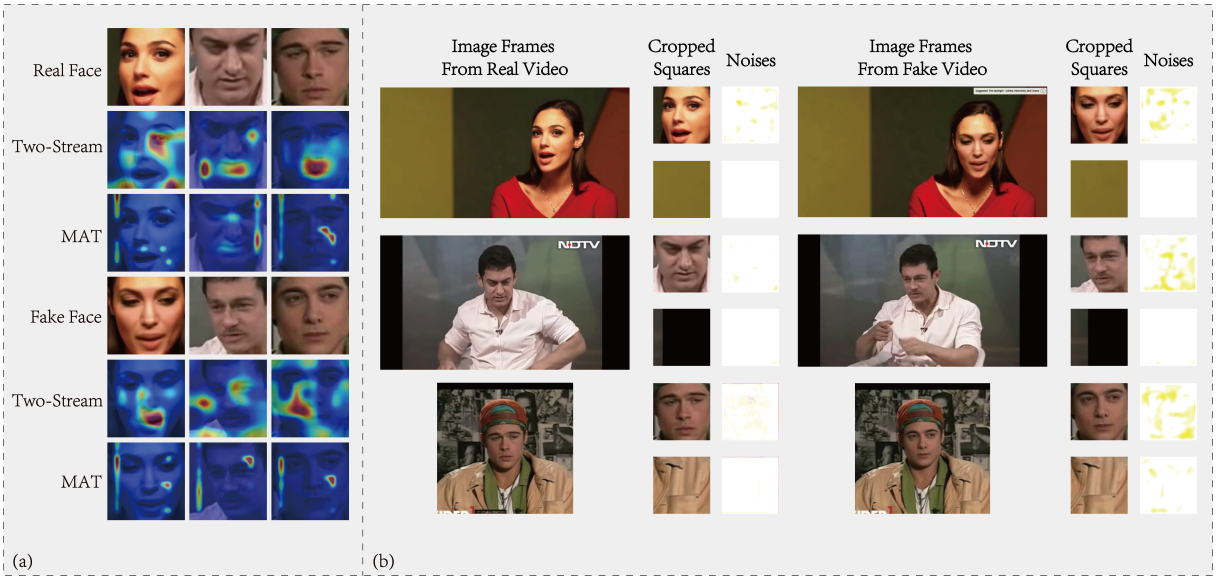

Figure 4: (a) Heatmap visualization of the extracted image features by MAT and Two-Stream Deepfake detection models. The hotter (red color) a position is, the more features are learned by the models. (b) Visualization of the Deepfake forensic noise traces extracted by the NoiseDF model. Obvious face shapes can be observed from the fake face squares.

ing real faces, while the bottom three rows display that of the fake faces. For a certain part in the heatmap, the hotter it is with the red color, the more features were learned by the model at there. As a result, it is observed that although the image features are successfully extracted and visualized, the real and fake faces are indistinguishable simply based on the heatmaps.

To compare with our model, we further visualized the extracted Deepfake forensic noise traces from a sample testing dataset of face-background pairs. In specific, we froze the weights of the noise extraction network within the Siamese Noise Feature Extraction module and displayed the extracted Deepfake forensic noise traces of each face-background pair by our model. In Figure 4(b), we displayed some sample results of the noise traces extracted from face-background pairs for both real and fake image frames using our well-trained NoiseDF model. As Figure 4(b) shows, the fake image frames with cropped face and background squares and the extracted noise traces are displayed on the left, while to the right are the synthesized image frames along with the corresponding noise traces. The fake face squares exhibit obvious Deepfake forensic noise traces with colorful and complex noise traces displayed with obvious face shapes while all other image squares have shallow or nearly no forensic noise trace as expected.

Ablation Study In this study, the purpose of the adopted denoiser is to be enhanced and trained for complete Deepfake forensic noise trace extraction. Therefore, a blind denoiser with better noise extraction performance is preferred. We conducted an ablation study on different state-of-the-art blind-denoising models for comparison to confirm the ability of the chosen denoiser in noise trace extraction. Specifically, we switched the forensic noise trace extractors in our NoiseDF model in different training sessions and tested the performance of CBDNet (Guo et al. 2019), DnCNN (Zhang et al. 2017), and RIDNet for both in-dataset and cross-dataset evaluations. As Table 2 shows, CBDNet performs well in some cross-dataset tests while failing on FF++; on the contrary, the DnCNN model achieves relatively satisfied performance for the in-dataset experiment on FF++ but does not fit well in the cross-dataset experiment. As a result, the proposed Deepfake detection model with RIDNet has shown dominant performance against that with CBDNet and DnCNN for both in-dataset and cross-dataset evaluations.

Conclusion

In this study, we present a novel noise-based Deepfake detection model from the perspective of digital forensics view and introduce the efficient and novel Multi-Head Relative-Interaction with depth-wise separable convolutions to boost the detection performance. Meanwhile, we introduce the creative face-background cropping strategy to distinguish the Deepfake forensic noise patterns between real and fake videos with the help of Siamese architecture. Our approach derives both the state-of-the-art performance on all baseline datasets and promising visualization evidence of the forensic noise traces. Admittedly, one unavoidable limitation of this work is that we are unable to deal with the case when the face covers a large region of an image, that is, the cropped face and background squares are overlapped. In this work, this kind of image data is omitted in the experiment. Future work will cover this limitation and further improve the model performance.

Acknowledgments

The authors would like to thank Ming Liu and Wei Cao for their help with running and monitoring some of the model-tuning experiments on the remote device.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*.

Anwar, S.; and Barnes, N. 2019. Real Image Denoising with Feature Attention. *IEEE International Conference on Computer Vision (ICCV-Oral)*.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, 737–744.

Chawla, R. 2019. Deepfakes: How a pervert shook the world. *International Journal for Advance Research and Development*, 4: 4–8.

Cheng, H.; Guo, Y.; Wang, T.; Li, Q.; Chang, X.; and Nie, L. 2022. Voice-Face Homogeneity Tells Deepfake. *arXiv preprint arXiv:2203.02195*.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the Detection of Digital Face Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

de Weever, C.; and Wilczek, S. 2020. Deepfake detection through PRNU and logistic regression analyses. Technical report, University of Amsterdam.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:1910.08854.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward convolutional blind denoising of real photographs. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, J.; Liao, X.; Liang, J.; Zhou, W.; and Qin, Z. 2022a. FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 951–959.

Hu, J.; Liao, X.; Wang, W.; and Qin, Z. 2022b. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1089–1102.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*.

Jiang, L.; Li, R.; Wu, W.; Qian, C.; and Loy, C. C. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *CVPR*, 2889–2898.

Koopman, M.; Macarulla Rodriguez, A.; and Geradts, Z. 2018. Detection of Deepfake Video Manipulation. In *Proceedings of the 20th Irish Machine Vision and Image Processing conference*, 133–136.

Lee, S.; Tariq, S.; Kim, J.; and Woo, S. S. 2021. TAR: Generalized Forensic Framework to Detect Deepfakes Using Weakly Supervised Learning. In *ICT Systems Security and Privacy Protection*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3204–3213.

Lukas, J.; Fridrich, J.; and Goljan, M. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2): 205–214.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing Face Forgery Detection With High-Frequency Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16317–16326.

Maras, M.-H.; and Alexandrou, A. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, 23(3): 255–262.

Marra, F.; Poggi, G.; Sansone, C.; and Verdoliva, L. 2017. Blind PRNU-Based Image Clustering for Source Identification. *IEEE Transactions on Information Forensics and Security*, 12(9): 2197–2211.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Use of a Capsule Network to Detect Fake Images and Videos. arXiv:1910.12467.

Picetti, F.; Mandelli, S.; Bestagini, P.; Lipari, V.; and Tubaro, S. 2022. DIPPAS: a deep image prior PRNU anonymization scheme. *EURASIP Journal on Information Security*, 2022.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11.

Saito, S.; Tomioka, Y.; and Kitazawa, H. 2017. A Theoretical Framework for Estimating False Acceptance Rate of PRNU-Based Camera Identification. *IEEE Transactions on Information Forensics and Security*, 12(9): 2026–2035.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.*, 38(4).

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Niessner, M. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2387–2395.

Vijayanagar, K. R. 2020. I, P, and B-frames – Differences and Use Cases Made Easy. https://bit.ly/34OArtI. Accessed: 2021-05-01.

Wang, T.; Cheng, H.; Chow, K. P.; and Nie, L. 2022. Deep Convolutional Pooling Transformer for Deepfake Detection. *arXiv preprint arXiv:2209.05299*.

Wang, T.; and Chow, K. P. 2022. A Lightweight Reliably Quantified Deepfake Detection Approach. In *Annual ADFSL Conference on Digital Forensics, Security and Law*.

Wodajo, D.; and Atnafu, S. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv preprint arXiv:2102.11126*.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155.

Zhang, Y.; Wang, T.; Shu, M.; and Wang, Y. 2022. A Robust Lightweight Deepfake Detection Network Using Transformers. In *PRICAI 2022: Trends in Artificial Intelligence*, 275–288. ISBN 978-3-031-20862-1.

Zhao, H.; Wei, T.; Zhou, W.; Zhang, W.; Chen, D.; and Yu, N. 2021. Multi-attentional Deepfake Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2185–2194.