Deepfake Algorithm Using Multiple Noise Modalities with Two-Branch Prediction Network

Hsuan-Wei Hsu* and Jian-Jiun Ding†
Nation Taiwan University, Taipei, Taiwan
E-mail: r08942035@ntu.edu.tw *, jjding@ntu.edu.tw †, TEL: +886-2-33669652

***Abstract*—In this paper, we propose a facial manipulation detection method based on multiple image noise analysis modalities and a two-branch prediction network to separation different types of forgery artifacts. The proposed architecture reveals whether the input image can be decomposed into a blending of two images from different sources, and checks whether some patches of the input image are generated from a deep learning networks at the same time. We observe that most of the existing forgery detection work] only focuses on finding one of the blending or manipulation artifacts in the input image. As a result, this method provides an effective way for forgery detection by simultaneously checking the manipulation and blending artifacts. In addition, for use with different types of image noise analysis modalities, our method can find more robust detection features in the high-frequency domain compared with traditionally detection in the RGB domain, thereby obtaining better performance. Extensive experiments show that our method outperforms other existing forgery detection methods on detecting synthesized face image, no matter on detecting training dataset or on detecting unseen face manipulation techniques.**

I. INTRODUCTION

In the last few years, advances in computer vision and graphics are very significant. Nowadays AI-based generator [1, 2] can generate realistic synthetic faces, which is challenge for humans and computers alike to distinguish between real and fake. We all agree that faces play a vital role in human interaction, a person's face can represent one's own identity and sometimes conveys a message by the facial expression or behavior. Nevertheless, the rapid development of the AI-synthesized method of the forgery face (commonly known as "*DeepFakes"* for public) threatening the trustworthiness of information transmission. In a Deepfake video, the faces of a target person are replaced by the faces of a source person synthesized by the forgery face generator. Due to the strong association between face and identity, an elaborate Deepfake video can create bogus behaviors of the specific person's activities. These forgeries may cause serious disputes and trust issues no matter to the individuals or to the country and society. As a result, to avoid widespread abuse of Deepfake videos, it is important to develop an effective method for detecting these face-swapping videos.

Due to subtle differences in the generation process, each real image has its own unique mark, which may be caused by differences in software components or hardware settings. Generally, these marks tend to show a similar distribution throughout the whole image. On the other hand, a face-swapping image can be viewed as stitched image patches from different sources together (see Fig. 1), and this procedure would inevitable make some discrepancies located at the blending boundary. Some recent work [3-5] has focused on capturing the intrinsic image divergence across the blending boundaries, instead of relying on the generated facial forgeries for supervision. Compared with relying on manipulation artifacts, using the inconsistency between blending boundaries to measure forgery facial image does help to improve the model's generalization ability.

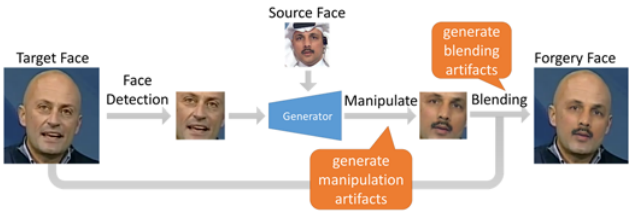

Figure 1. An overview of a typical face operation process. Previous works detect artifacts produced by manipulation methods or blending steps, while our approach focuses on detecting both.

Although these studies have indicated that the use of the difference between the blending boundaries can help improve the generalization ability of the model, little attention has been paid to combining the blending artifacts with the manipulation artifacts together for forgery detection. If we can refer not only manipulation artifacts but also blending artifacts, the detection approach would be more robust intuitively. However, studies on separating blending and manipulation artifacts are still lacking. Therefore, from this perspective as a starting point, we proposed a framework for forgery facial image detection.

The contributions of this paper are as follows. First, our experiments demonstrate that detecting both blending and the manipulation artifacts at the same time certainly improves the generalization ability through a thorough analysis. Second, using multi-modal noise analysis as input, we can extract more robust features to do prediction, compared with using an original RGB image as input. Third, in comparison with many previous face forgery methods, our experimental results show that this framework outperforms the performance in the in-dataset and cross-dataset evaluation.

II. RELATED WORKS

In this section, we will briefly introduce several deepfake detection techniques, from the early research on image forgery detection to the latest work related to our proposed method and some related dataset, as well as some related datasets.

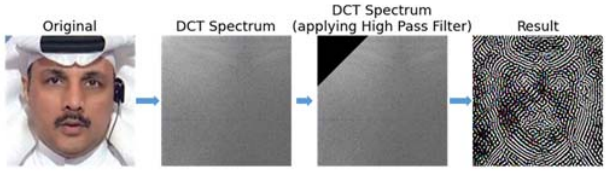

Figure 2. The pipeline of high-pass filter DCT image noise analysis.

Figure 3. An example of error level analysis of the input images.

Face forensics datasets. The rise and progress of deep learning techniques such as VAE and GAN have made facial image processing of forensic models more challenging, because they not only fake facial attributes, but also retain more and more photo details such as poses, facial expressions, or lighting, etc. To support the research of facial manipulation detection works, several Deepfake related datasets [6-11] have been released and countermeasures have been introduced.

Hand-crafted Deepfake Detection. Early handcraft detection methods [12-14] usually used the intrinsic statistics of forgery facial images for classification. However, for this type of detection method, it is difficult to generate features that is suitable for detecting different types of synthetic methods that are constantly evolving.

Learning Based Face Manipulation Detection. Regarding the threat of image forgery in terms of privacy and trust issues, since 2018, many studies [15-17] have begun to detect face manipulation images. For instance, Xuan *et al.* [18] proposed to use image pre-processing steps, such as Gaussian blur and Gaussian noise, to remove low-level high frequency features. Kumar *et al.* [19] train a triplet network to enhance the feature space distance between the cluster of real and fake videos embedding vectors. Li *et al.* [20] employ the Long-term Recurrent Convolutional Networks (LRCN) model to capture temporal dependencies of human eye blinking. Masi *et al.* proposed a two-branch structure [21]: one branch propagates the original information, while the other branch suppresses the face content yet amplifies multiband frequencies using a Laplacian of Gaussian as a bottleneck. The goal is to isolate manipulated faces by learning to amplify artifacts while suppressing the high-level facial content. For improving the generalization ability, Face X-ray [1] and PCL [2] have also produced their own data generation pipelines, focusing on predicting the blending boundaries in fake video frames.

Our method combines the viewpoints of several works and improves on their shortcomings. First, use a variety of different noise modalities to obtain advanced high-frequency features for detecting, which helps us to obtain more robust facial content during the training phase. Secondly, two-branch prediction involves not only blending artifacts, but also manipulating artifacts to maximize the use of artifacts present in deepfake pictures. To achieve the above points, we constructed a multi-task learning framework for the two prediction branches to predict the blending boundary and manipulation region in turn.

III. Proposed Methods

In our approach, we try to use not only the manipulated artifacts, but also the blending artifacts at the same time for Deepfake detection. In other words, we will figure out the manipulation region and the blending boundary separately. Through this strategy, the model can achieve good generalization ability for some unseen datasets. The overall architecture composed of three parts: (A) Using multiple image noise analysis modalities as the training input, (B) Two branches of multi-task learning predict manipulation artifacts and blending artifacts, (C) Multi-task learning schedule:

A. Multiple Image Noise Analysis Modalities

Since Deepfake detection can be regarded as a binary classification of true and false, in most cases, training will easily face severe overfitting problems. To avoid overfitting on the training dataset and then reduce model generalization, we abandon the regular RGB image as input, and instead train the noise analysis of the input image. This strategy has several advantages. First, it can reduce the influence of the bias of the training dataset, because most of the dataset bias are the low frequency components or the low-level high frequency components in the facial images. Second, most face image synthesis methods evaluate generation quality based on RGB domain, so using the noise analysis of the input image to find the detection features is more able to find the robust features. Third, the noise analysis of the input image can be regarded as high-frequency facial content. Training on these high-frequency components helps to find more high-level forgery artifacts for better performance. In our approach, we choose (I) high-pass filter DCT, (II) error level analysis and (III) photo response non-uniformity as the image noise analysis modalities we adopt.

(I) High-pass filter DCT

The DCT is a transformation that can convert the image from the spatial domain to the frequency domain, and the converted energy can be more concentrated in the low frequency. By doing this transform, we can separate the image into spectral sub-bands of different importance.

image DCT conversion formula:

$$D(i,j) = \frac{1}{\sqrt{2N}} C(i)C(j) \cdot \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} p(x,y) cos\left[\frac{(2x+1)i\pi}{2N}\right] cos\left[\frac{(2y+1)j\pi}{2N}\right] \quad (1)$$

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}}, if\ u = 0 \\ 1, if\ u > 0 \end{cases} \quad (2)$$

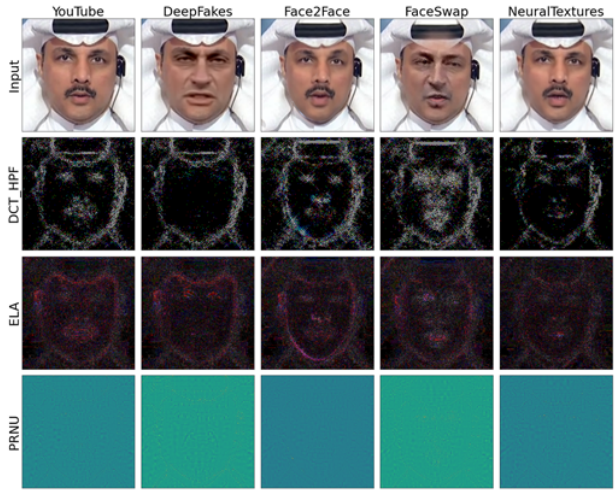

Figure 4. Image noise analysis example (high pass filter DCT, error level analysis, photo response non-uniformly) of a real image (YouTube) and fake images (Deepfakes, Face2Face, FaceSwap, NeuralTextures) from Faceforensics++.

The high pass filter design:

$$H(m,n) = \begin{cases} 0, if\,(m+n) < T \\ 1,\ otherwise \end{cases} \quad (3)$$

where T is threshold, m, n are the corresponding coordinate axes in the input spectrum.

(II) Error-Level-Analysis

This method can identify portions of an image with different compression levels. For any non-fake JPEG image, the entire photo should be roughly at same compression level. If an image has a significantly different error level, this may indicate that there might exist some digital modifications in the photo. The steps of this algorithm as following:

1) Use the same image as the target image and pass uniform 90% quality JPEG compression.
2) The difference between the two images shows a variation of the artifacts in JPEG compression.

(III) Photo Response Non-uniformity

When a uniform light falls on the sensor array in a digital camera, the array would output slightly differently voltage due to the variety of factors, including small variations in cell size, material or interference with the local circuitry. This difference in response to a uniform light source is called *Photo Response Non-Uniformity* or *PRNU* for short. PRNU is one source of pattern noise in digital cameras. It is the pixel variation under illumination. The more detail information of PRNU extraction algorithm can see in [22].

B. Two Branches of Multi-task Learning Predict Manipulation Artifacts and Blending Artifacts

As shown in Figure 1, due to the different formation stage, the property of the manipulation artifacts and the blending artifacts are distinct. The former is generated during the process of synthesizing face images because of the imperfections of the GAN-related or VAE-related techniques, and then the latter one is generated in the progress of fusing the synthesis face to the target real face due to the intrinsic image discrepancies in the blending boundary. Both types of artifacts will be present in almost every facial swapping image. We know that both two types of artifacts can help us distinguish fake images. Nevertheless, if we don't separate these two forgery clues in training stage, then the model might be limited in performance on learning to recognize fake images. It might not be guarantee what will be learned more during the training steps. We don't know that the trained model will more rely on manipulation artifacts or on blending artifact to do deepfake detection. Then this will lead to the generalization ability of the model may be limited.

To consider both manipulation and blending artifacts, we use two branches multi-task architecture for prediction. One branch is used to predict the manipulation region, and the other branch is used to predict the blending boundary. Each branch also combines a classifier for binary classification. The multi-task learning in our approach refers to the combination of semantic segmentation (region prediction) and binary classification (label prediction), also means to detection the manipulation region and the blending boundary. With the two-branch prediction network, we can detect the manipulation artifacts and the blending artifacts simultaneously in our work. The overall proposed architecture is shown in Figure 5.

C. Multi-task Learning Schedule

Since the output of our method has two branches, each branch predicts its own classification label and semantic segmentation mask, we must use multi-task learning to arrange our training phase. Our multi-task learning schedule is as follows.

During the training process, in the first half of each epoch, we will train our two branch outputs in turn, that is, each iteration only backpropagates one branch output. In the second half of each epoch, we will train these two branches jointly. In each iteration, the respective loss functions of these two branches are calculated, and the average value is taken for back propagation.

IV. IMPLEMENTATION

A. Overall Procedure

Inspired by [23], our solution based on three types of image noise analysis transforms: high-pass filtering performed by discrete cosine transform, error level analysis and photo response non-uniformity, combined in end-to-end pipeline for deepfake detection. We have introduced a fusion architecture with ResNet18 backbone as the feature extractor. Each noise analysis modality is processed by an individual ResNet18 backbone, which returns latent vector of size d = 256. Then, connect these latent vectors to form a tensor of shape 3 × d. After that, apply Max, Avg and Min pooling in the first dimension, and concatenate the final feature vector to get a 3×d tensor as the feature embedding. Based on the ideas provided by [23], given different inputs, the importance of features of

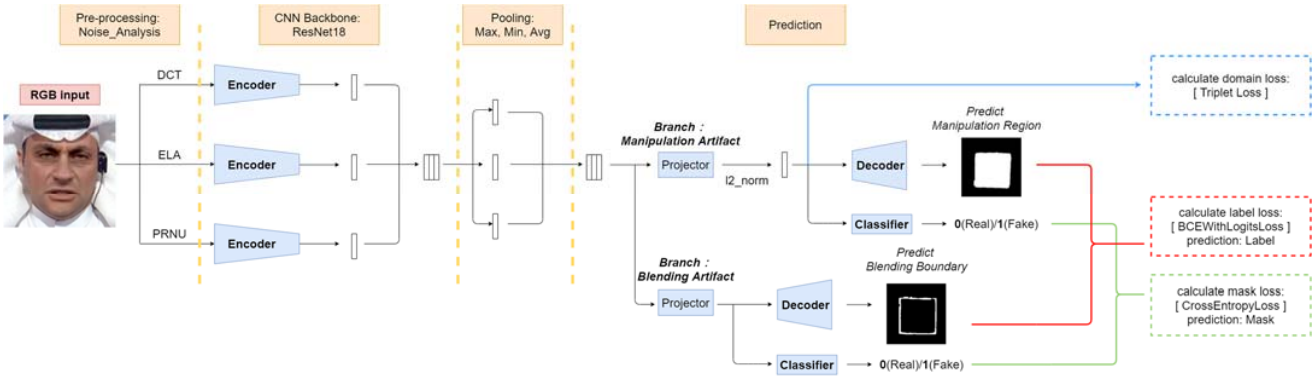

Figure 5. Our frame-based face operation detection architecture. The input image is processed by three types of noise analysis methods: high-pass filter DCT, ELA, PRNU, and then through the ResNet18 backbone to do feature extraction. After that, the three feature embeddings are fused, so that the following modules can learn richer representations. A two-branch neural network is applied in the prediction stage. One branch is used to predict the manipulation region, and the other branch is used to predict the blending boundary.

Figure 6. The pipeline of the pre-processing steps

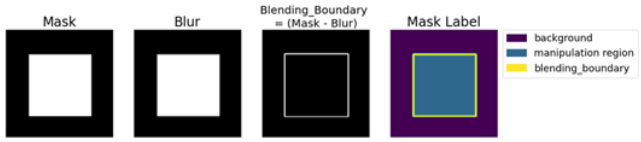

Figure 7. The ground truth mask definition for region prediction. The forgery mask is provided by the FF++ data set. We first apply Gaussian blur to the original fake mask, and then subtract the blur one from the original one to obtain the blended boundary area. The area outside the blending boundary is the background, and the area inside the boundary is the manipulation region.

different modalities may be different. Therefore, various pooling strategies are better than just average pooling. Next, we flatten the embedding vector and pass it to the two-branch neural network for doing multi-task prediction. Each branch is composed with one projector, one decoder and one classifier. The projector is used to take out the manipulation or the blending artifacts part in the embedding vector. The decoder is used to prediction the manipulation region or the blending boundary of the input image. And the classifier is used to classify whether the noise analysis of the input image exists the manipulation artifacts or the blending artifacts. Finally, use the prediction scores of classifiers from the two-branch output to do average, we could get the final soft decision score of the deepfake detection results.

In the proposed network, the number of epochs is 25, the batch size is 25, the optimizer is Adam, the learning rate is reduced from 1e-2 to 1e-6, and the scheduler is OneCycleLR.

B. Pre-processing

Our solution is based on frame-by-frame classification approach. The flow chart of the pre-processing step is shown in Fig 6. For a training data video, we first use OpenCV as the frame extractor, and then use MTCNN as the face extractor for each frame to get the target face position. After obtaining the target face position, we crop these faces and save and resize them to shape 255x255 as the input facial image for our training.

C. Data Augmentation

To achieve better generalization, we use heavy augmentation such as Image Compression, Gaussian Blur, Gaussian Noise, Random Crop, Flip, Rotation, etc. In the light of [24], adding some Cutout-like augmentations would also help to achieve better generalization ability. That is, we deleted some parts of the input image, hoping to improve the robustness of the model. The cutout-like augmentation we used is like the data augmentation techniques of Cutout [40] and Random Erasing [41]. The biggest difference is that when selecting the area to be discarded, we have added additional conditions to choose. For example, delete part of the input face based on facial landmarks or based on ground truth mask area. Combined with the other heavy data augmentation skills we mentioned in the paper, these enhancements can help our method achieve better generalization abilities.

D. Loss Function

Our proposed architecture contains two branch prediction networks, one for detecting manipulation artifacts in the input image, and the other for detecting blending artifacts. For each branch, the output will predict about the classification score and the region of the artifact location. In other words, this is a combination of binary classification tasks and semantic segmentation tasks. In the former task we choose the binary cross entropy as the loss function, and in the latter task we choose the cross entropy as the loss function. The definition of the ground truth masks for semantic segmentation tasks is shown in Figure 7.

Moreover, to train the Faceforensics++ dataset, since the dataset contains 4 different face manipulation methods, we pull in the triplet loss as a clustering loss, which is used for clustering manipulation artifact features in each different types of face manipulation method. The cluster loss only applies to the manipulation branch, and is not calculated in the blending branch. That's because we assume that for different face

TABLE I: In-dataset evaluation results on FF++. Our method performs better on all manipulation types compared with other works.

Methods	Backbone	Train Set	Test Set (AUC (%))				
			DF	F2F	FS	NT	FF++
XceptionNet[30]	Xception	FF++	99.08	93.77	97.42	84.23	93.63
SPSL [31]	Xception	FF++	98.50	94.62	98.10	80.49	96.91
MIL [27]	Xception	FF++	99.51	98.59	94.86	97.96	97.73
Fakespotter [28]	ResNet-50	FF++, CD2, DFDC	-	-	-	-	98.50
XN-avg [6]	Xception	FF++	99.38	99.53	99.36	97.29	98.89
Face X-ray [1]	HRNet	FF++	99.12	99.31	99.06	97.27	99.20
S-MIL-T [30]	Xception	FF++	99.84	99.34	99.61	98.85	99.41
Ours	ResNet-18	FF++	100.00	99.32	99.99	99.20	99.63

TABLE II: In-dataset evaluation results on Celeb-DF.

Method	Backbone	Train Set	Test Set (AUC%)
			CD2
Fakespotter [28]	ResNet-50	CD2	66.80
Tolosana et al. [32]	Xception	CD2	83.60
S-MIL-T [29]	Xception	CD2	98.84
Ours	ResNet-18	CD2	99.82

TABLE III: Cross-dataset evaluation results on Faceforensics to Celeb-DF.

Method	[Training] FF++	[Testing] Celeb-DF
Two-stream [33]	70.10	53.80
Meso4 [15]	84.70	54.80
MesoInception4 [15]	83.00	53.60
HeadPose [35]	47.30	54.60
FWA [34]	80.10	56.90
VA-LogReg [36]	78.00	55.10
Xception-raw [6]	99.7	48.2
Xception-c23	99.7	65.3
Xception-c40	95.50	65.50
Multi-task [39]	76.30	54.30
Capsule [38]	96.60	57.50
DSP-FWA [34]	93.00	64.60
SMIL [29]	96.80	56.30
Two-branch [37]	93.20	73.40
SPSL(Xception) [31]	96.91	76.88
Face X-ray [1]	99.20	80.58
Ours	99.63	81.31

manipulation algorithms, although the ways to generate virtual faces are different, post-processing always contains similar steps, such as blending, blurring, and color correction. Therefore, we believe that for different forgery methods, the manipulation artifact features are different, and the blending artifact features are similar to each other.

$$L_{label} = Binary\ Cross\ Entropy = -\sum_{\{I,c \in D\}} (c\ log\ \hat{c} + (1-c)\ log(1-\hat{c})) \quad (4)$$

$$L_{\text{mask}} = \text{Cross Entropy} = -\sum_{\{I,B \in D\}} \frac{1}{N} \sum_{i,j} \left(M_{i,j}\ log\ \widehat{M}_{i,j} + \left(1 - M_{i,j}\right) log(1 - \widehat{M}_{i,j})\right)$$

$$L_{cluster} = \text{Triplet Loss} = \sum_{i}^{N} \left[\left\| f(\chi_i^a) - f\left(x_i^p\right)\right\|_2^2 - \| f(x_i^a) - f(x_i^n)\|_2^2 + a\right] \quad (5)$$

where I represents the input image, M denotes the corresponding ground truth mask of the manipulation region or the blending boundary, and c is a binary label specifying whether the image is real or fake.

$$Total\ Loss = \lambda_{label} L_{label} + \lambda_{mask} L_{mask} + \lambda_{cluster} L_{cluster} \quad (6)$$

The loss weight balancing parameter we adopt here is $\{\lambda_{label}, \lambda_{mask}, \lambda_{cluster}\} = \{1.0,\ 25.0,\ 5.0\}$.

V. EXPERIMENTS

Several experiments are conducted to show the performance of the proposed defake algorithm. We use the most commonly used metrics in the literature to evaluate Deepfake detection results, including **area under the ROC curve** (**AUC**) and **average accuracy** (**AP**). The higher the AUC or AP value, the better the performance. The evaluation result in the experiment is at the video level, which is calculated by averaging the classification score of the video frame.

A. In Dataset Evaluation Datasets

To evaluate our experimental results, we selected some state-of-the-art datasets (FaceForensics++, Celeb-DF) for verification, and compared the results with some works on the same topic.

FaceForensics++ [6] is by far the most famous and popular dataset in the Deepfake detection research community. This dataset was proposed in 2019 as an extended version of the FaceForensics [26] dataset. It consists of 1000 real video sequences extracted from YouTube and processed using four automatic face manipulation methods: "Deepfakes", "Face2Face", "FaceSwap" and "NeuralTextures". The evaluation result is shown in Table I.

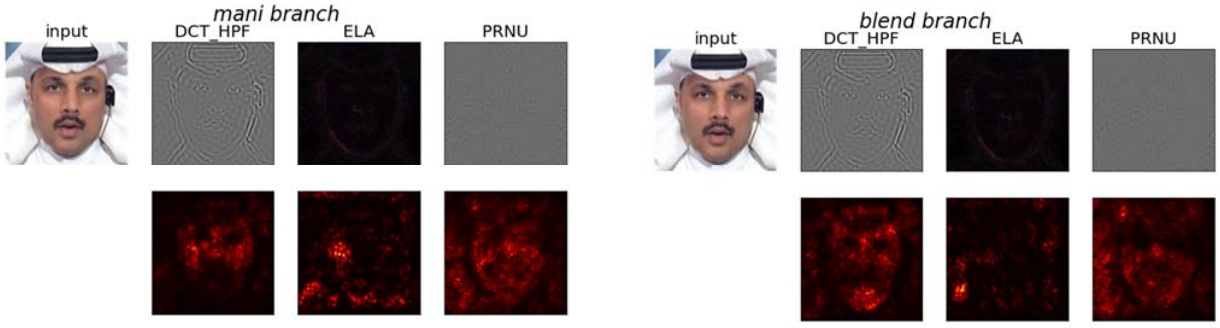

Figure 8. The saliency map of two prediction branch (manipulation & blending). The left half is the saliency map of the manipulation branch, and the right half is the saliency map of the blending branch. We could see that the saliency map in manipulation branch is more concentrated on the middle patch of the input facial noise analysis image. For the saliency map of the blending branch, we see that it is close to forming a complete facial boundary. We can interpret it as the facial boundary is more important for the forgery detection of the blending branch.

TABLE IV: *Ablation study on the impact of cross-dataset performance.* We trained our detection model on the FF++ dataset and tested it on CD2 to see the effect of different experimental settings on the generalization ability of the model.

Method	Test Set (AUC%)
	Celeb-DF-v2
Our approach	0.813
without cutout-like augmentation	0.765
with cutout-like augmentation + ELA to 90%	0.775
with cutout-like augmentation + without cluster loss	0.765
with cutout augmentation + DCT High pass filter Threshold = 100	0.731

Compared with the FaceForensics++ dataset, Celeb-DF-v2 [27] aims to provide fake videos of high visual quality, which is closer to the Deepfake videos circulating on the Internet. This dataset consists of 590 real videos extracted from YouTube by 59 celebrities and 5639 fake videos. The fake videos are created by a refined version of the public Deepfake generation algorithm, which improves the synthesis of faces such as low resolution and color mismatches, inaccurate face masks, etc. The performance of our work trained in this dataset is shown Table II.

B. Cross-Dataset Evaluation

A big problem of deepfake detection is the evaluation of the generalization ability of each detection algorithm. Due to the binary classification property of this task, it does not seem to be a difficult problem for researchers to obtain excellent performance in the in-dataset. However, these excellent performances evaluated in the in-dataset often means overfitting with the training data, which cannot be maintained well when encountering unseen synthetic methods. For real-world scenarios, most of our detection methods need to face unseen forgery synthesis methods instead of known forgery methods. Therefore, maintaining good generalization ability is also one of the key points to evaluate deepfake detection methods.

Here, we designed a cross-validation experiment to test our method in aspect of the model robustness. Our model is first trained on the FaceForensics++ dataset, and then tested on the Celeb-DF-v2 dataset to make a scene to detect facial manipulation images on the unseen dataset. Table III shows the cross-validation performance results of our method in terms of AUC. We observe that even if the performance is significantly degraded in cross-evaluation, the performance of our method is still better than other referred face manipulation detection methods.

C. Ablation Study

We studied the effect of different experimental settings on the generalization ability of the model. Table IV shows that in our approach, cutout-like data augmentation plays a vital role in training model. Without cutout-like data augmentation in our work, the performance will not be much different from other previous forgery detection works. In addition, the hyper-parameters of the input image noise analysis need to be carefully chosen. If our choice is not suitable enough, some forgery features may be discarded in the process of noise analysis, which will seriously reduce the performance of the model.

VI. DISCUSSION

The saliency map in Figure 8 shows that the degree of separation of the two artifacts has reached the result we expected. Although the saliency maps of the two branches still have a certain similarity, for the manipulation branch, it is still more concentrated in the middle patch of the input noise. Then for blending branch, it relies more on entire facial boundary instead of small feature blocks for prediction. Of course, we still have room for improvement the separation degree of these two types of forgery features.

We also checked the t-SNE result of the two-branch feature embedding vector in Faceforensics++. The visualization result is shown in Figs. 9 and 10. Due to the use of cluster loss, the feature space distribution of the manipulated branch is the same as we expected, as in Fig. 9. Each forgery method and real video is in group and distinguishable from others. This is fully

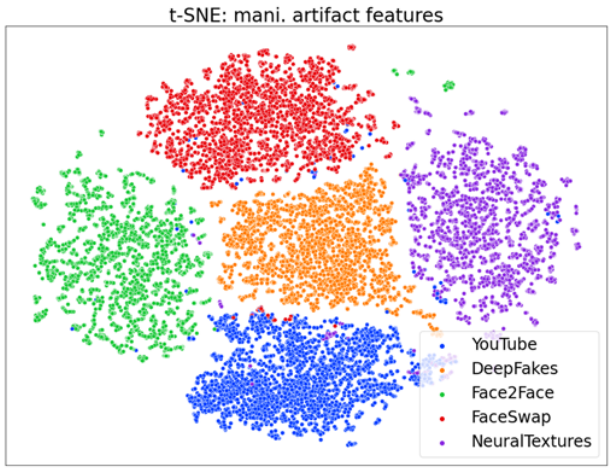

Figure 9. the visualization t-SNE result of the manipulation branch features in Faceforensics++ dataset.

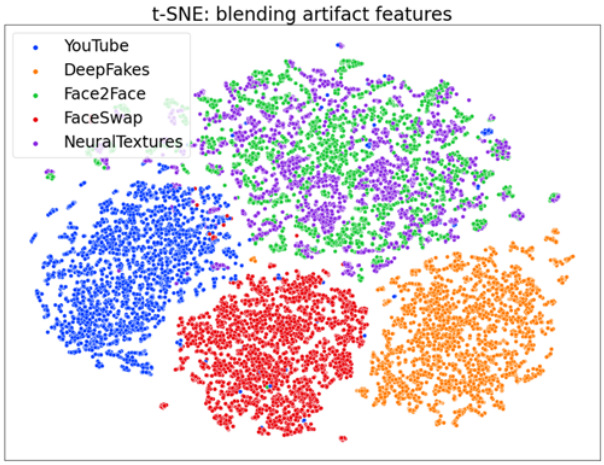

Figure 10. the visualization t-SNE result of the blending branch features in Faceforensics++ dataset.

in line with the nature of our previous assumptions: for each different forgery method, the extracted features of the manipulated artifacts should be different, while the features extracted of the blending artifacts should be similar. The distribution of the blending features in shown in Figure 9. We see that some forgery methods still have obvious clustering in blending features. This is a point that our method should be improved. It is not enough to use heavy data augmentation to separate manipulation and blending features. We would think how to define more constrain and tips for separate these two artifacts much more in future work.

VII. Conclusions

We propose a face manipulation detection method with multiple noise analysis modalities, including high pass filter DCT, error level analysis and photo response non-uniformity. In addition, combined with the two-branch prediction network to separate detect the manipulation artifacts and the blending artifacts for forgery detection. The detection output of our method can not only give true and false soft decision scores, but also provide the location of the manipulation region and the blending boundary. Our experiments show that, compared with recent work, our method can obtain excellent performance and maintain certain performance when encountering cross-evaluation. We show that it is not a bad idea to detect both manipulation artifacts and blending artifacts at the same time, but to maximize the detection ability of the two artifacts, we must have a good strategy to separate the two artifacts in a forged image. Hope our work can be a good reference for researchers who will study the same subject in the future.

Acknowledgment

The authors thank for the support of Ministry of Science and Technology, Taiwan.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in: *Advances in Neural Information Processing* Systems, article 27, pp. 1-9, 2014

[2] D. P. Kingma and M. Welling, “Autoencoding variational Bayes,” *arXiv preprint arXiv*: 1312.6114, 2013.

[3] L. Li, J., Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 5001-5010, 2020.

[4] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, “Learning to recognize patch-wise consistency for deepfake detection,” *arXiv preprint arXiv*: 2012.09311, 2020.

[5] D. K. Kim, D. H. Kim, and K. Kim. "Facial manipulation detection based on the color distribution analysis in edge region," *arXiv preprint arXiv*:*2102.01381*, 2021.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to detect manipulated facial images,” in *IEEE/CVF Int. Conf. Computer Vision*, pp. 1-11, 2019.

[7] N. Dufour and A. Gully, “Contributing data to deepfake detection research,” 2019, available from https://ai.googleblog.com/2019/09/contributingdata-to-deepfake-detection.html.

[8] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3207–3216, 2020.

[9] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge dataset,” *arXiv e-prints arXiv*: 2006.07397, 2020.

[10] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2886–2895, 2020.

[11] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, “WildDeepfake: A challenging real-world dataset for deepfake detection,” in *ACM Int. Conf. Multimedia*, pp. 2382–2390, 2020.

[12] X. Pan, X. Zhang, and S. Lyu, “Exposing image splicing with inconsistent local noise variances,” in *IEEE Int. Conf. Computational Photography*, pp. 1-10, 2012.

[13] J. Fridrich and J. Kodovsky, “Rich models for steganalysis of digital images,” *IEEE Trans. Inf. Forensics Secur.*, vol. 7, issue 3, pp. 868–882, 2012.

[14] M. Goljan and J. Fridrich, “CFA-aware features for steganalysis of color images,” in *Media Watermarking, Security, and Forensics*, vol. 9409, pp. 94090V, 2015.
[15] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," *IEEE Int. Workshop on Information Forensics and Security*, pp. 1-7, 2018.
[16] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-generated fake images over social networks,” in *IEEE Conf. Multimedia Information Processing and Retrieval*, pp. 384-389, 2018.
[17] W. Quan, K. Wang, D. M. Yan, and X. Zhang, “Distinguishing between natural and computer-generated images using convolutional neural networks,” *IEEE Trans. Information Forensics and Security*, vol. 13, issue 11, pp. 2772-2787, 2018.
[18] X. Xuan, B. Peng, W. Wang, and J. Dong, “On the generalization of GAN image forensics,” in *Chinese Conf. Biometric Recognition*, pp. 134-141, 2019.
[19] A. Kumar and A. Bhavsar, “Detecting deepfakes with metric learning,” in *IEEE Int. Workshop on Biometrics and Forensics*, pp. 1-6, 2020.
[20] Y. Li, M. C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv*: 1806.02877, 2018.
[21] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in *Euro. Conf. Computer Vision*, pp. 667-684, 2020.
[22] J. Lukas, J. Fridrich, and M. Goljan. "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Secur.*, vol. 1, issue 2, pp. 205-214, 2006.
[23] A. Parkin and O. Grinchuk, "Creating artificial modalities to solve RGB liveness," *arXiv preprint arXiv*: 2006.16028, 2020.
[24] Dfdc_deepfake_challenge, 2020, available from https://github.com/selimsef/dfdc_deepfake_challenge.
[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” *arXiv preprint arXiv*: 1803.09179, 2018.
[26] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, S. “Celeb-DF (v2): A new dataset for deepfake forensics,” *arXiv preprint arXiv*: 1909.12962, 2019.
[27] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu. “Revisiting multiple instance neural networks,” *Pattern Recognition*, vol. 74, pp. 15-24, 2018.
[28] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, “FakeSpotter: A simple yet robust baseline for spotting ai-synthesized fake faces,” *arXiv preprint arXiv*: 1909.06122, 2019.
[29] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, “Sharp multiple instance learning for deepfake video detection,” in *ACM Int. Conf, Multimedia*, pp. 1864-1872, 2020.
[30] F. Chollet, “Xception: “Deep learning with depthwise separable convolutions,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1251-1258, 2017.
[31] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 772-781, 2021.
[32] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, “DeepFakes evolution: Analysis of facial regions and fake detection performance,” *arXiv preprint arXiv*: 2004.07532, 2020
[33] P. Zhou, X. Han, V. Morariu, and L. Davis, “Two-stream neural networks for tampered face detection,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 1831-1839, 2017.
[34] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 46-52, 2019.
[35] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 8261–8265, 2019.
[36] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *IEEE Winter Applications of Computer Vision Workshops*, pp. 83-92, 2019.
[37] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in *Euro. Conf. Computer Vision*, pp. 667-684, 2020.
[38] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsuleforensics: Using capsule networks to detect forged images and videos,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 2307-2311, 2019.
[39] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *IEEE Int. Conf. Biometrics: Theory, Applications, and Systems*, pp. 1-8, 2020.
[40] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv*: 1708.04552, 2017.
[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. "Random erasing data augmentation," *arXiv preprint arXiv*: *1708.04896*, 2017.
[42] P. Chen, S. Liu, H. Zhao, and J. Jia, "Gridmask data augmentation," *arXiv preprint arXiv*: 2001.04086, 2020.