HF-Detect A Hybrid Detector for Manipulated Face Detection*

Ankit Shakya[1] K. Jenni[2] Murukessan Perumal[3] and M. Srinivas[4]

***Abstract*— The recent advancement of fake face creation and fake face generation motivates the development of an excellent fake face detection method that can effectively detect the difference between fake and real. Various fake detection methods are available with adequate performance, but the limitation of those available methods is they are not performing well with highly compressed images with degraded quality. Manipulation of face images is getting advanced, and becoming difficult to trust the content over the media, and generating and detection should go parallelly to balance society. Therefore we are proposing a novel approach to solve this problem which uses the hybrid model HF-Detect, which combines the advantage of the Xception network along with the F^{3}Net.**

***Keywords*— DeepFake Detection, Computer Vision, Deep Learning, Artificial Intelligence**

There are various methods and techniques available for creating fake faces. Nowadays, machine learning and deep learning methods are used for many different application, like in the case of videos, text and medical [1], [2]. The existing deep learning and machine learning methods generate realistic images that the naked human eye can not identify. These advanced techniques can be deep learning or regular graphical manipulation. They use general photo editing tools to generate fake faces in graphical manipulations. However, deep learning techniques use GANs to generate fake face images using deep learning techniques. These deep learning-based generated faces are almost realistic and can not be easily identified, especially when the images or videos are highly compressed or low-quality. In highly compressed videos and images, it becomes difficult to extract the important, relevant features for the detection of manipulation. When we talk about manipulation, it can be classified into two categories, identity manipulation, and expression manipulation. In identity manipulation, a person's identity in an actual image is changed by putting another person's face over the target image. That is done by using a FaceSwap or using Deepfakes techniques. For expression manipulation, we change the person's expression in the image. For example, if the person seems to have a regular face in the video, we can change his facial expression by making him smile. However, these manipulations can be done using NeuralTextures.

Different manipulation methods are available, which are depicted in Fig. 1. The figure displays example images of Deepfake, FaceSwap, Face2Face, and NeuralTexture. From the image, we can get an idea of different manipulation methods, these methods are highly filtered, so the human eye can not easily detect fake images or videos.

Fig. 1: Images with different methods of manipulation (Deepfakes, Face2Face, FaceSwap and NeuralTextures) sampled from Faceforensics++ dataset

Various methods [3]–[10] have been proposed for fake face detection, which has proven effective for high-quality image detection. When we consider a low-quality or highly compressed image, all these models perform limited because, in compressed images or videos, they are sometimes blurred and sometimes do not have enough highlighting features to distinguish between real and fake images. Therefore, extracting features from highly compressed images is a difficult task.

When some videos and images go viral through different social media platforms, their quality degrades at each forward, making detecting those degraded images difficult. These images are degraded; therefore, extracting essential features from this highly compressed image is complex. We mainly focused on these compressed images so we could detect these images correctly. Many companies are working on this fake face detection problem to stop the fake forwarding of information over social media. Nowadays, there are many platforms to post content, and all people are connected to a vast network; therefore, any information gets viral in a few seconds. We can deploy such detector models over each social media platform to authenticate and then allow the information to post.

Motivated by different state-of-the-art methods, we are developing a hybrid technique HF-Detect to handle this problem. This will extract the underlying hidden feature of highly compressed images by highlighting the manipulated regions. To highlight the manipulated region, some combine additional feature extraction modules such as binary mask extraction and noise map extraction, followed by XceptionNet [11]. Rossler et al. depicts the importance of XceptionNet to develop a good detector model. [12]. The manipulated area is highlighted, and the resultant features are sent to the deep network. Frequency is used as a critical feature to extract the manipulated region to reveal the underlying vital features of highly compressed images. Like a generative model, GAN [13] generates almost realistic images that are very difficult to distinguish between real and fake human eyes. The generative model has many advantages, and they are enhancing more

*This work was not supported by any organization

[1]Ankit Shakya Author is with the Department of Computer Science and Engineering, National Institute of Technology Warangal, India ascs21122@student.nitw.ac.in

[2] Dr. K Jenni is with the Computer Science Department, King Khalid University, Abha, Saudi Arabia jenni.k.507@gmail.com

[3]Murukessan Perumal is with the Department of Computer Science and Engineering, National Institute of Technology Warangal, India muruap87@student.nitw.ac.in

[4]Dr. M Srinivas is with the Department of Computer Science and Engineering, National Institute of Technology Warangal, India msv@nitw.ac.in

rapidly. Still, to maintain the balance of generation and detection, we must develop an enhanced detector to detect all the fake faces and videos generated by these advanced generative models. In the event of protecting against the spread of fake viral images and videos, we can use these detector models to validate the authenticity of the content. Many fake news videos are getting viral nowadays, which can create social unrest. We are mainly focused on news videos and mainly on the face part. Because in news videos, the most crucial part is the face of the person delivering the news, we cropped the images over the face to focus on the critical part of the video.

What people do with news videos they manipulate the face of the person in the video. However, they can swap that person's face with the other one, or they can change the expression of that person in the video. This can cause social unrest and loss of trust in social media.

I. Related work

There are various traditional and old methods available for fake face detection. Rossler et al. developed the standard dataset FaceForensics++ and proposed XceptionNet [12], resulting well-defined model for fake face classification. Then as the advancement of the methods grew, the other researcher came up with different solutions that focused on extracting additional features, like noise maps and binary masks, as done in [11]. Nicolo Bonettini et al. [14] came up with an implementation of the attention model with the EfficientNetB4, where they can learn the critical feature of fake faces using the advantages of the attention module. The weight-sharing mechanism uses the Xception Network [15], which uses contrastive learning. As the popularity of Transformer come into the picture, Zhengbo Luo et al. [16] come up with a new approach to deal with this fake face detection problem. They used the Transformer and XceptionNet as DNN extractors with a dual path network.

All these advanced methods are enhanced and result in good accuracy with the FaceForensics++ dataset, and some authors also generalized the model for various datasets; there are different datasets available for forgery face detection. The problem with the FaceForensics++ dataset is its very huge dataset. After extracting the frame, it takes almost 2.2.TB space. When the dataset is enormous, the problem becomes automatically complex because handling the vast dataset and tuning the parameter take time. One of the problems is the manipulated videos are not available. We have to create manipulated videos by using different methods. All the available methods try to develop a suitable detector, but all these methods have limitations when we deal with highly compressed images or videos. These models can not extract the critical hidden features lying in low-quality videos. To understand the low-quality or highly compressed images, many researchers have developed different models that focus on detecting fake images and highlighting the manipulated region by using different additional techniques.

Peng Chen et al. [17] proposed a novel approach that detects the fake and real class of videos or images and localizes the manipulated area in the images. They used separate modules for each functionality, one for the face detection branch, one manipulated classification branch, and one for the manipulated localization branch. In [18], they proposed a multi-modal multi-scale transformer for Deepfake detection. They use multi-scale transformers and the advantage of frequency features and classify them as real or fake. "Detect and Locate: Exposing Face Manipulation by Semantic- and Noise-Level Telltale" [11] introduce an efficient approach in which they used a noise map prediction module along with a semantic map prediction module which helps to extract the underlying critical feature in highly compressed images or videos, They also compared various method, but F^3Net [19]is having highest AUC (Area Under Curve). Inspired by all these methods, we proposed a highly efficient hybrid model HF-Detect that uses the advantage of the Xception Network and the Cat-feature extraction Module by considering the frequency information.

Adding important information to the learning model results in an excellent, sensitive detector model. Regular CNN (Convolutional Neural Network) extracts the local features by using regular convolutional operation; some images have different features to learn that got suppressed by other highlighted features, so to extract that hidden features, we implemented Cat-Feature extraction which takes the frequency information extracted using Discrete Cosine Transform (DCT).

II. Proposed Model

In this paper, we proposed an effective approach to solve the fake face detection problem, which used a frequency feature, cat-feature extraction, and combined XceptionNet. The high-level view of the proposed approach is depicted in Fig. 2. The model takes an image as input which is

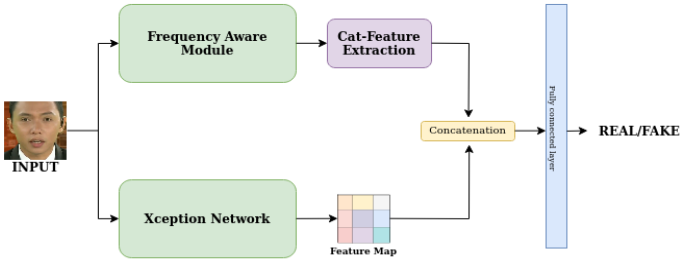

Fig. 2: High-Level architecture of Hybrid Detector HF-Detect

face cropped by using dlib library; the input image is passed to Frequency Aware Module (FAM) that extracts the underlying frequency information from the input image. This Frequency information is extracted and further processed by the Cat-Feature extraction module. The resultant of the Cat-feature extraction module is concatenated with the feature map extracted by passing the input image to Xception Network(XceptionNet) and finally connected with a fully connected layer, which detects the input image as fake or real. Let's discuss each module in detail.

A. Frequency Aware Module

Frequency Aware Module (FAM) extracts the frequency information by decomposing the input image by adaptively dividing the image into learning frequency bands. FAM learns the forgery pattern and highlights the manipulated region in the image by using the Discrete Fourier Transform (DFT) as used in [19]. It applies Discrete Cosine Transform

(DCT) to input image that results in decomposed frequency components. We use Inverse Discrete Cosine Transform (IDCT) to convert these resulting decomposed frequency components into the spatial domain resulting in a series of frequency-aware image components.

B. Cat-Feature Extraction

This Cat-Feature Filtration Module has convolutional layers that extract the frequency component feature map. The internal architecture of the cat-feature filtration module is depicted in Fig. 3, where Conv1 is a 3x3 convolutional layer, "C" represents the concatenation operation between the frequency feature maps, and "-" denotes the subtraction operation. We pass the decomposed frequency components as input to the Cat-Feature Filtration (Cat-Feature Filtration Module), which passes the input to the first Conv1 layer with kernel size 3, the output of which results in feature map 1. The resultant feature map 1 is subtracted from the input to result in F_1, passed to the next layer, a combination of two consecutive Convo layers, resulting in F_2 feature map.

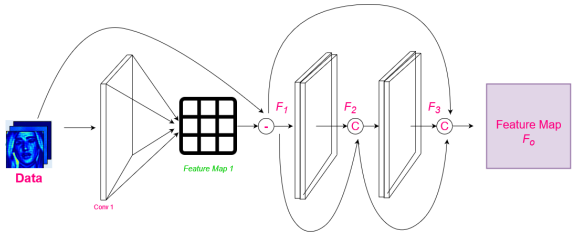

Fig. 3: Overview of Cat-Feature Filtration Module, where Conv1 denotes a 3x3 convolution layer, "C" denotes a concatenation operation, and "-" represents a subtraction operation.

The F_2 feature map is concatenated with the F_1 feature and passed to the next consecutive Convo layer. After getting F_3, we again perform a concatenation operation between F_1, F_2, and F_3, and this results in the high-level feature map F_o. It can be easily understood by using mathematical equations. Let's suppose X is an input to the Cat-Feature Filtration module and Feature Map 1 is represented by W. And F_1 that we get after the subtraction operation is $(X - W)$ so the following equation gives the resultant Feature Map F_o.

$$F_1 = X - W \tag{1}$$

$$L_{1I} = X - W \tag{2}$$

$$L_{2I} = F_1 \otimes F_2 \tag{3}$$

$$F_o = F_1 \otimes F_2 \otimes F_3 \tag{4}$$

$\otimes$ represents the concatenation operation between the feature maps, and L_{1I} is the input to the first consecutive Convo layers at the first level. The layer results in feature map F_2. L_{2I} is input to the second consecutive Convo layer; therefore, the resultant Feature Map F_o is given by concatenating all the intermediate frequency feature maps.

We are getting another feature map from XceptionNet generated at another network branch. In this branch, it takes the cropped input image as input and passes it to the XceptionNet model [12]. This model generates the feature map by learning the spatial features of the input image. This feature map is combined with the output of Cat-feature extraction by using the concatenation operation.

III. EXPERIMENTAL RESULT

A. Datasets:

We are using the FaceForensics++ dataset, which has 1000 videos, especially news videos; we manipulated these videos by using four methods Deepfake, Face2Face, FaceSwap, and NeuralTextures. These manipulation methods are different; FaceSwap and Deepfake is identity manipulation technique, and Face2Face and NeuralTexture is expression manipulation method. We got 5000 videos, of which 1000 are real, and 4000 are manipulated; we know the video is a collection of frames. Therefore, we extracted frames from the videos. The dataset has three different quality videos that are Raw (C0), high-quality (C23), and low-quality (C40). The C0, C23, and C40 is the quantization parameter's compression level. We mainly focus on low-quality videos that are highly compressed and difficult to handle. We take 720 videos for training and 140 for validation, and 140 for testing. We consider 270 frames from each video in the training part and 100 frames for each video for testing and validation. After splitting the dataset, we have the following number of images as in Table I.

TABLE I: Dataset Details (FaceForensics++)

	TRAIN	TEST	VALIDATE
Videos	3600	700	700
Frames	972000	70000	70000

The dataset we are working on is unbalanced because the real class has 1000 videos, and the fake class has 4000 videos, which imbalanced the dataset. We balanced the dataset by applying augmentation techniques to make the model learn properly.

We also tested the model with CelebDF $V1$ dataset, which also contains videos, and we extracted the frames in the same way as with the FaceForensics++ dataset. We consider 270 frames from each video for training and 100 frames for each video for testing and validation. Table II lists the number of videos and frames.

TABLE II: Dataset Details (CelebDF $V1$)

	TRAIN	TEST	VALIDATE
Videos	884	100	219
Frames	238680	10000	21900

B. Evaluation metric

All the fake face detection methods consider accuracy a critical evaluation metric, and we also consider accuracy an evaluation metric along with AUC (Area Under Curve). AUC provides information about the model and its performance for both classes. More the AUC better the model. We use the DGX A100 system with a GPU memory of 80GB, which is very high-end. We run the model for 20 epochs having a batch size of 120.

C. Results

Chenqi Kong et al. [11] have compared the available methods regarding AUC (Area Under Curve) and Accuracy. F^{3}Net model results in the highest AUC among all the available models. Table III lists the comparison between the different models. In it, F^3Net and the proposed method HF-Detect are with locally experimented results.

TABLE III: Accuracy and AUC Comparisons of the proposed model HF-Detect with the state-of-the-art models for the FaceForensics++ dataset

METHODS	ACCURACY	AUC
Steg. Features [20]	55.98	-
Cozzolino et al. [21]	58.69	-
Bayar and Stamm [22]	66.84	-
Rahomouni et al. [23]	61.84	-
MesoNet [24]	58.69	-
Xception [12]	83.61	84.38
Face X-ray [25]	-	61.60
Two-Branch [26]	-	86.59
SPSL [27]	81.57	82.82
Nirkin et al. [28]	80.18	-
Multi-Att [29]	84.49	85.52
Detect and locate [11]	84.84	87.10
F^{3}Net [19]	<u>86.56</u>	<u>88.36</u>
Proposed Model HF-Detect	**86.714**	**88.80**

In the above Table, the highlighted values mean maximum accuracy and AUC, and the underline values refer to the second top values. The proposed model HF-Detect has the topmost values in Accuracy and AUC, followed by F^{3}Net. In it, F^3Net and the proposed method HF-Detect are with locally experimented results.

We also evaluated the model with CelebDF $V1$ dataset under the same experimental setup. The results of an experiment with the CelebDF $V1$ dataset are listed in Table IV. Tables III,IV represent the accuracy and AUC comparison between the latest methods, which has the highest AUC, with the proposed approach HF-Detect which uses the hybrid technique to learn the features.

TABLE IV: ACCURACY AND AUC COMPARISONS OF THE PROPOSED MODEL HF-DETECT WITH THE STATE-OF-THE-ART MODELS FOR THE CelebDF $V1$ DATASET

MODELS	ACCURACY	AUC
FWA [30]	-	52.2
Xception-c23 [30]	-	52.5
Xception-c40 [30]	-	59.4
DSP-FWA [30]	-	47.2
F^{3}Net [19]	<u>86.56</u>	<u>88.36</u>
Proposed Model HF-Detect	**92.21**	**96.40**

IV. Conclusion

In this paper, we proposed a novel method performing better than the state-of-the-art methods available. We are focusing on highly compressed videos (C40), and our experiment shows the model performance, which is improved. We evaluate the model with the FaceForensics++ dataset and the CelebDF $V1$ dataset, and both experiments' results effectively, as shown in Table III and Table IV. In the future, we will try to improve the model's performance more densely to be more accurate and generalized. Adding more hand-crafted features to the learning model can improve the performance and develops a good detector model which can effectively distinguish between real and fake images or videos.

References

[1] M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (covid-19) using machine learning methods," *Chaos, Solitons & Fractals*, vol. 139, p. 110050, 2020.

[2] M. Perumal, A. Nayak, R. P. Sree, and M. Srinivas, "Inasnet: Automatic identification of coronavirus disease (covid-19) based on chest x-ray using deep neural network," *ISA transactions*, vol. 124, pp. 82–89, 2022.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[4] H. Jeon, Y. Bang, and S. S. Woo, "Fdftnet: Facing off fake images using fake detection fine-tuning network," in *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings 35*. Springer, 2020, pp. 416–430.

[5] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *2018 international symposium on computer, consumer and control (IS3C)*. IEEE, 2018, pp. 388–391.

[6] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2018, pp. 384–389.

[7] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," *arXiv preprint arXiv:1910.12467*, 2019.

[9] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7556–7566.

[10] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1831–1839.

[11] C. Kong, B. Chen, H. Li, S. Wang, A. Rocha, and S. Kwong, "Detect and locate: Exposing face manipulation by semantic-and noise-level telltales," *IEEE Transactions on Information Forensics and Security*, 2022.

[12] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[13] M. Mahiuddin, M. Khaliluzzaman, M. S. A. Chowdhury, and M. N. Arefin, "Fake face generator: Generating fake human faces using gan," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[14] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 5012–5019.

[15] Z. Ma, P. Xu, X. Mei, and J. Shen, "A novel facial manipulation detection method based on contrastive learning," in *2022 IEEE 5th International Conference on Electronics Technology (ICET)*, 2022, pp. 1163–1167.

[16] Z. Luo, S.-I. Kamata, and Z. Sun, "Transformer and node-compressed dnn based dual-path system for manipulated face detection," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3882–3886.

[17] P. Chen, J. Liu, T. Liang, C. Yu, S. Zou, J. Dai, and J. Han, "Dlfmnet: End-to-end detection and localization of face manipulation using multi-domain features," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[18] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 615–623.

[19] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," *CoRR*, vol. abs/2007.09355, 2020. [Online]. Available: https://arxiv.org/abs/2007.09355

[20] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[21] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," in *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, 2017, pp. 159–164.

[22] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.

[23] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *2017 IEEE workshop on information forensics and security (WIFS)*. IEEE, 2017, pp. 1–6.

[24] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[25] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "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.

[26] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 667–684.

[27] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[28] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6111–6121, 2021.

[29] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[30] 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.