Fast and Generalized DeepFake Detector Through Feature Space Transformation

Anonymous authors
Paper under double-blind review

Abstract

The current state-of-the-art DeepFake or manipulated image detection algorithms are not generalized against an unseen database, manipulation types, and image degradation due to compression. Existing literature shows different input transformations can boost the detection performance of deepfake detection algorithms. However, these algorithms only transform the spatial pixel values with the hope that the transformation will help in learning a linearly separable decision boundary. The transformation of a 2D volume containing millions of pixel values is computationally complex and on top of that, the amalgamation with the original image further increases the computational complexity. The proposed algorithm utilizes the concept of transformation; however, the transformation of a feature space that is 1-D and compact representation of an image. The transformed representation is then used to calculate the discriminative feature maps used for the binary classification as real or altered images. Extensive experimentation on multiple databases under several unconstrained settings establishes the effectiveness of the proposed algorithm and its desirability in the current era. Under each set, the proposed algorithm achieves state-of-the-art detection performance on Face Forensics++ and Celeb-DF databases. The proposed algorithm is almost 'parameter-free' and achieves its two-fold aim of giving a robust detection algorithm and an energy-saving medium.

1 Introduction

The presence of fake media on 'any' social media platforms has created havoc and made it hard to establish the authenticity of digital multimedia content. Due to this, it is now extremely dangerous to blindly trust the content on social media platforms and react to that. One such reason for such mistrust of social media content is the wide availability of DeepFake videos Li et al. (2023b); Wang et al. (2023); Liang et al. (2022). These deepfake videos can be used for many purposes ranging from harassment to personal gain whether including monitory or achieving any post WION (2022). At the beginning of DeepFake, the face of celebrities was replaced with the face of pornstars in a video Harwell (2018). However, several recent incidents claim the importance of identifying which digital data is real and which is fake because it can severely impact any individual. A few such examples recently came into the picture are (i) the deepfake video of Ukraine's president telling his army to surrender Telegraph (2022), (ii) a candidate in the USA 2020 election claimed to be diagnosed with severe cancer and request the public to not vote for him as he is unfit for the post because of health issues Cook (2020), and (iii) a deepfake audio is used to rob millions of dollars Good (2021). The problem became even more complicated because not only the sophisticated machine learning networks publicly available but simple mobile applications are now available which can easily be used for such malicious manipulations by a novice user Agarwal et al. (2019a); Agarwal et al. (2017); Agarwal et al. (2021b). Therefore, the correct identification of these manipulated videos is not only important to build trust in society but also to make positive progress toward the advancement of artificial intelligence.

By looking at the severity of the problem, several research efforts have started to detect deepfake videos. The detection algorithms can be broadly divided into handcrafted plus machine learning classifiers and deep neural network data-driven algorithms. In one of the early works, Agarwal et al. Agarwal et al. (2017a) proposed a novel feature engineering algorithm to highlight the subtle moiré patterns in the face swap videos.

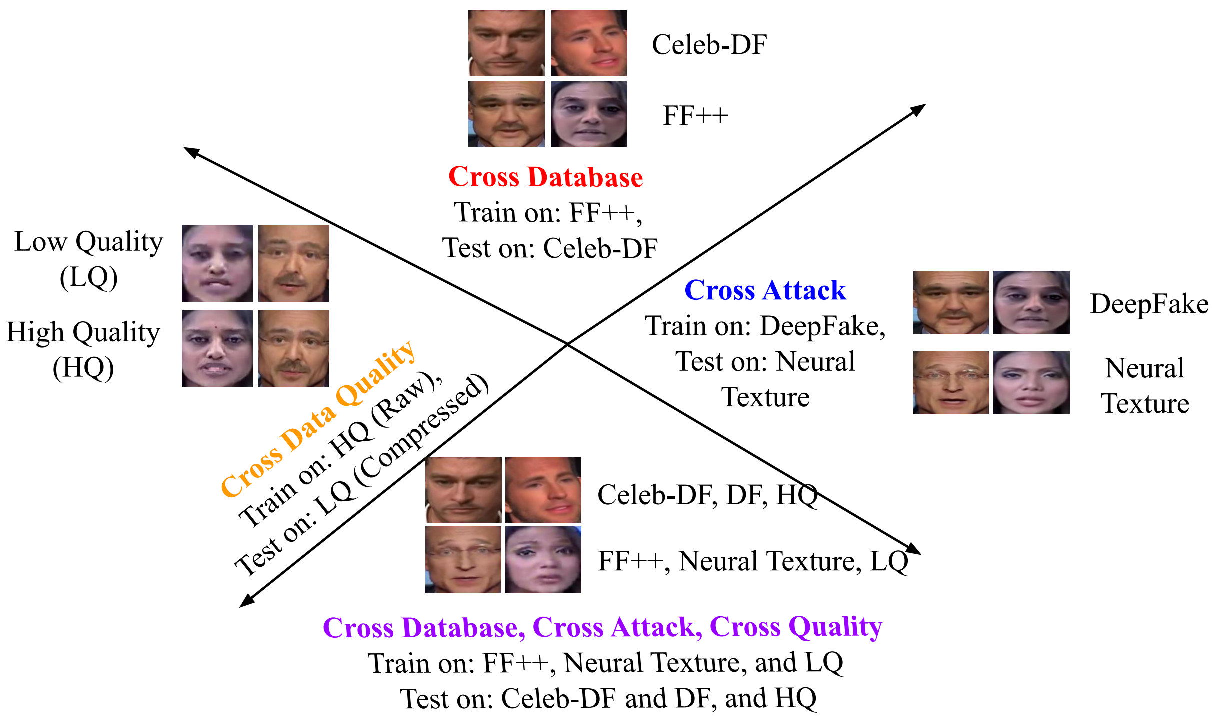


Figure 1: Missing challenges in existing image manipulation detection which we are trying to solve for a practical and robust defense against the serious real-world threat of the time. (i) cross-database, (ii) cross-attack, (iii) cross-data quality, and (iv) cross each possible variation (i.e., database, attack, and data quality).

Other image feature features algorithm used for deepfake detection are: eye color and missing reflections Matern et al. (2019), 3D head poses Yang et al. (2019), facial movements Baltrusaitis et al. (2018), and image artifacts Raghavendra et al. (2017); Zhang et al. (2018). Zhao et al. Zhao et al. (2021c) and Nirkin et al. Nirkin et al. (2021) have proposed source image features and face contextual information extraction networks for deepfake detection. The deepfake detection network proposed by Zhao et al. Zhao et al. (2021a) uses the multi-attention convolution network consisting of spatial attention heads and textural feature enhancement block. Agarwal et al. Agarwal et al. (2021a) have proposed the generalized convolutional network architecture utilizing two branches consisting of raw images and transformed images and introduced the cross-stitch connections to transfer knowledge among layers of two branches Agarwal et al. (2021a). Zhou et al. Zhou & Lim (2021) have utilized both audio and video discrepancies for the detection of deepfake videos. DSP-FWA Li et al. (2020c), Face X-ray Li et al. (2020a), and PCL + I2G Zhao et al. (2021c) proposed the boundaries in the facial regions which possibly exist due to the swapping of two faces. The above category of research algorithms possess two important characteristics: (i) either they are computationally efficient or (ii) yield state-of-the-art detection performance on multiple datasets. However, there is one catch here, the SOTA performance reported is usually obtained in a scenario where the training and testing images/videos belong to the same database and/or same attack type.

Henceforth, Fig. 1 shows the motivation and desirability of the proposed research in the current era of fake examples surfacing over every possible corner of life. The existing defense algorithms are lacking these challenging scenarios while developing robust defense which makes them still not ready for practical deployment. The proposed research tackles all these challenges to build a secure and deployable detection algorithm. To summarize:

- A novel DeepFake detection algorithm is presented by utilizing an amalgamation of various machine learning eras to identify the discriminating cues helpful for classification. The proposed algorithm

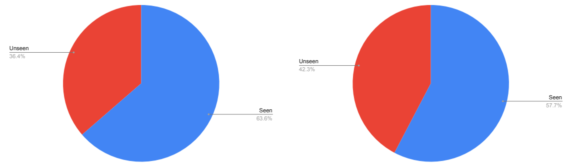


Figure 2: Critical areas where significant attention is needed in the deepfake detection research. The left chart shows the limitations of existing research published in top-tier venues under the cross and same attack setting. The average AUC of three recent algorithms namely Face X-ray Li et al. (2020a), XceptionNet, and High-Frequency Luo et al. (2021) on seen attacks (Deepfake and face swap) is 0.9903 which drops to 0.5656. Similarly, the right pie chart shows the bottleneck of the existing algorithms in terms of seen and unseen dataset training-testing. The average performance of MD-CSDNet Agarwal et al. (2021a), SPSL Liu et al. (2021a), Two branches Masi et al. (2020), and Frequency Qian et al. (2020) in the seen setting is 0.9694 whereas, in the unseen dataset, it shows the true strength with the AUC value of 0.7105.

takes the deep neural network representation of an image and extracts the forensics features to build to binary decision hyperplane separating manipulated images/videos from the clean ones;

- The proposed algorithm is robust against manipulation types, databases, and image degradation, which occurs due to compression. Compression is an important part of social media document upload to save both the transmission bandwidth and memory. Therefore, the defense must be able to handle such natural phenomena;
- An extensive comparison with the existing algorithms and surpassing them with a large margin establishes the superiority of the proposed algorithm.

2 Related Works

In this section, a comprehensive survey of the existing works done towards both the generation of manipulated videos and the defense against them is presented. SWAPPED Agarwal et al. (2017) is one of the earliest databases prepared using the social media application namely Snapchat. The database contains more than 600 face swap videos and 120 real videos captured in unconstrained environments reflecting real-world applications. Later, Rossler et al. (2019) prepared one of the largest face manipulation databases covering both identity swap and expression manipulation. Four different attack variants are presented in the database and each type contains 1000 videos along with 1000 real videos. Dolhansky et al. (2020) released the competition database on DeepFake detection. The database is prepared by Facebook research to enhance the defense against manipulated videos. The database is prepared from more than 3429 paid actors and consists of more than 100,000 face swap videos. Li et al. (2020c) released the fake videos database of better visual quality that usually surfaces on social media platforms. The database is different from several existing databases which generally consist of visual artifacts helpful in the easy identification of them. The database contains 890 real videos taken from YouTube and high-resolution and color-consistent 5639 DeepFake videos. Apart from these sophisticated databases generated using computationally complex systems, Majumdar et al. (2019) have shown the effect of partial tampering of facial attributes such as blending of an eye and mouth portion on deep face recognition networks. Not only do these facial manipulations utilize pair or more images for manipulation, adversarial noise on face images and presentation attack is also a serious concern Agarwal et al. (2016; 2021c; 2017b); Goswami et al. (2019); Sanghvi et al. (2020). However, this research focuses on the alteration around DeepFake because of its high impact on society.

Due to the adverse effect of these manipulations, several detection strategies are also proposed ranging from traditional handcrafted features to deep neural networks. Agarwal et al. (2017) have proposed a feature extraction algorithm effective for highlighting the artifacts that occurred due to the swapping of faces. These DeepFake videos generally contain facial and voice information and utilizing both these visual and audio features, Chugh et al. (2020) and Mittal et al. (2020) have proposed a DeepFake detection algorithm.

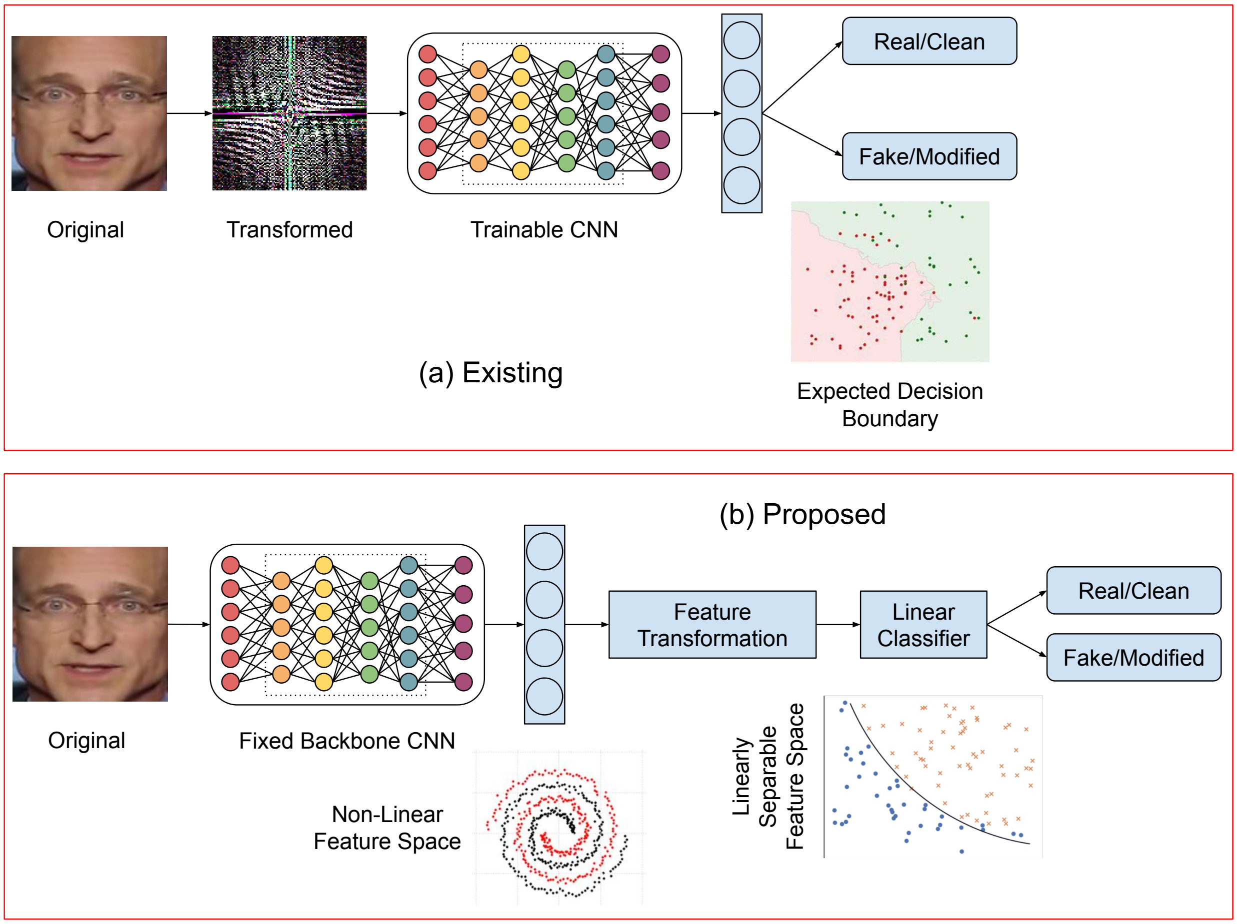


Figure 3: Comparison between the existing image transformation-based approaches vs. the proposed feature transformation-based algorithm. Input transformations that are applied on either 2D input are computationally complex as compared to the proposed feature transform which works on 1D representation.

As discussed earlier, most of the databases exhibit visual artifacts such as head pose and eye blinking, based on these inconsistencies, Agarwal et al. (2019b) and Jung et al. (2020) have proposed deepfake detectors. Rossler et al. (2019) have performed a detailed study with several steganalysis features combined with either traditional classifiers or deep neural networks for manipulation detection. Wang & Dantcheva (2020) have proposed a 3DCNN architecture for the identity swap manipulation. To highlight the minute visual artifacts that occurred due to swapping, Majumdar et al. (2019) have performed high-pass filtering on the input images. The enhanced images combined with the original images are passed through the Siamese type of architecture for the detection of partial swapping of face attributes. Kumar et al. (2020) have proposed the multi-regional deep neural network architecture for the detection of reenactment manipulation. The authors have reported state-of-the-art results ranging from high-quality videos to tightly compressed videos. Tolosana et al. Tolosana et al. (2021) have performed a deepfake detection study by utilizing different components of facial regions. The authors have found that the existing SOTA algorithms are not generalized against recent and complex deepfake datasets. Fernando et al. Fernando et al. (2020) have proposed a memory neural network to counter the deepfake threat. However, as reported through experiments, similar to the existing algorithm, the algorithm is not generalized against unseen attacks. The survey on these face manipulation types and defense against them can also be found in the survey paper for further study Mirsky & Lee (2021); Singh et al. (2020); Tolosana et al. (2020).

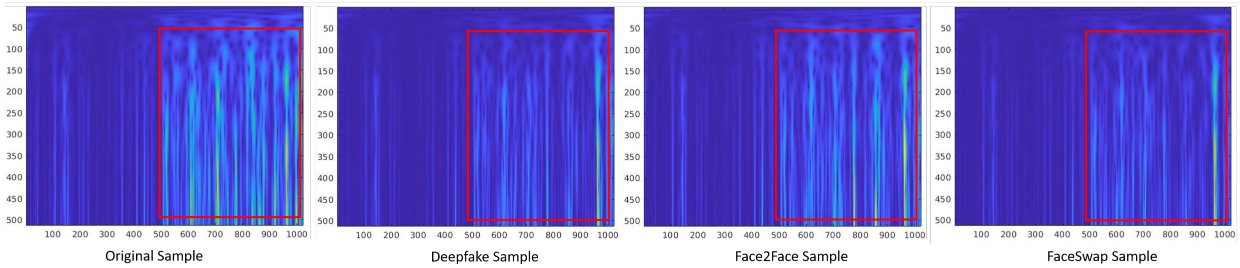


Figure 4: S-transform reflects the potential of differentiating different classes of images including real and several image manipulations. The S-transform heatmap is reported using the images of the FF++ dataset. The heatmap on real images shows a high amount of energy information in the right part whereas, the information is severely suppressed in the manipulated images, especially in identity manipulated images including deepfake and face swap. The red box shows the example of discriminative flow across kinds of data, i.e., real and multiple manipulations. (Best view in color and zoom.)

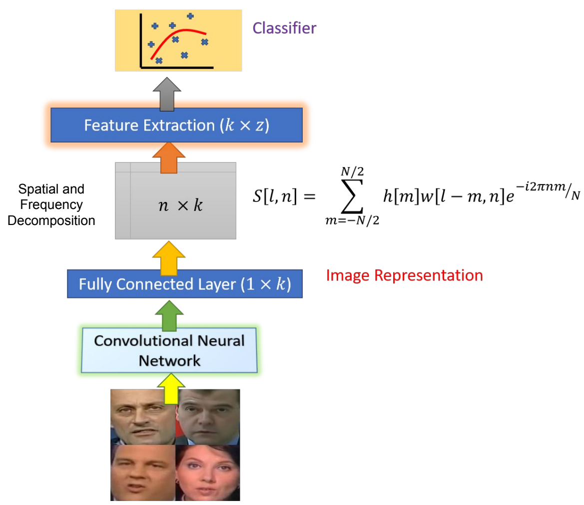


Figure 5: Proposed network for the detection of face-manipulated videos. (Best view in color.)

As discussed, several research algorithms are presented in the literature however there are several critical reasons which make the proposed study impactful. The biggest reason is the generalizability of the existing works against several conditions such as unseen datasets and unseen attacks. Fig. 2 shows the limitations of the recent state-of-the-art algorithms when evaluated in the out-of-distribution samples of unseen attacks or datasets. Another drawback, interestingly with the inception of several high-quality and large-scale datasets, the majority of research is focused on two datasets only namely FF++ Rossler et al. (2019) and Celeb-DF Li et al. (2020c). On top of that, only the Celeb-DF dataset is used for cross-dataset generalization. Very little research focuses on cross-attack generalization. Apart from these two generalizations, we assert several other factors are important as reflected in Fig. 1. This research dealt with these critical challenges such as experimental evaluation on multiple large datasets and maybe reflecting in-the-wild settings and several evaluation generalization settings to make the algorithm real-world ready.

3 Proposed Detection Algorithm to Build Trust on Multimedia Content

The manipulated videos are generated using artificial operations and lack natural frequency and spatial compactness. Some recent input transformations-based defense approaches have articulated the path forward Agarwal et al. (2021c); Chai et al. (2020); Frank et al. (2020); Masi et al. (2020); Qian et al. (2020); Wang

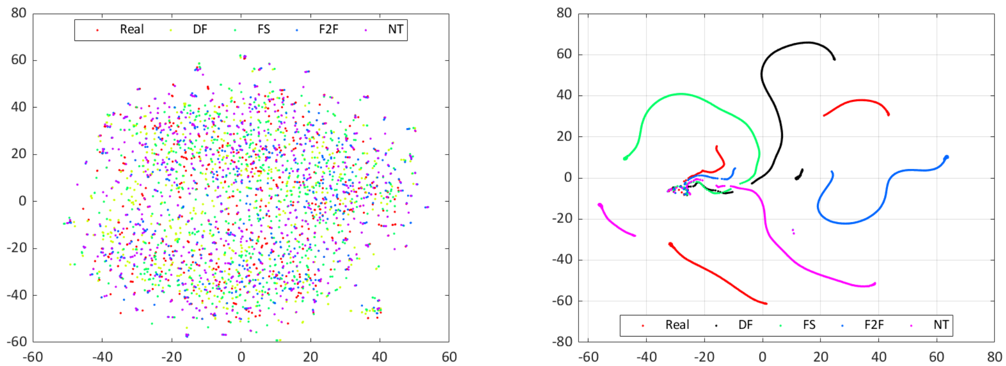


Figure 6: t-SNE Van der Maaten & Hinton (2008) visualization of real and different manipulation videos in the original feature space (left) proposed feature transformation space (right). The proposed feature transformation depicts its strength in separating the classes effectively.

et al. (2020) in detecting the fake media. However, these transformations are restricted to the input pixel space which is high dimensional space due to the high resolution of images. Inspired by existing domain knowledge, we have utilized the concept of transformation but at the feature level, to build an efficient deepfake detector. Fig. 3 shows the comparison between the proposed and existing transformation-based research. The proposed DeepFake detection algorithm consists of four modules: (i) image representation, (ii) spatial-frequency decomposition to enhance the artifacts, (iii) feature extraction, and (iv) classification.

Image Representation: The fully connected feature vector obtained from a convolutional neural network (CNN) is used as an image representation. The idea of utilizing the CNN is its effectiveness in handling the variation such as translation, and minute visual artifacts due to environmental factors such as natural noises in the images. Additionally, the use of a convolutional neural network helps in extracting the discriminative and compact representation of an input which earlier shows advantages in multiple applications ranging from object classification Vo et al. (2019); Xie et al. (2016) to detection Ren et al. (2015). In this research, we have used the ImageNet Deng et al. (2009) pre-trained CNN model for the extraction of a representative vector of an image. It can be defined as follows: $F = CNN(x)$, where, F is a k dimensional 1-D vector representing the input sample(x).

spatial-frequency Decomposition: Recent studies Agarwal et al. (2021c); Chai et al. (2020); Frank et al. (2020); Masi et al. (2020); Qian et al. (2020); Wang et al. (2020) have shown effective use of transformation in the detection of synthetically generated images and deepfake manipulated videos. These studies however mainly utilize the standard frequency transformations such as Fourier or cosine transformations and limit the transformation to the input level only, i.e., only the input images are transformed. These transformed images are then fed into the CNN for end-to-end feature extraction and classification. Fourier transforms Brigham & Morrow (1967); Cooley et al. (1969) converts the data from its original spatial space into the different frequency components. In such a case, the original spatial information is lost which might be required for the task at hand. Therefore, a few researchers have used the transformed information as a separate component for the DeepFake examples detection Majumdar et al. (2019); Masi et al. (2020). These two branch algorithms not only increase the computational complexity of the entire defense algorithm but are also found not generalizable. In this research, we have applied the transformation on an image representation in place of raw pixel space and chose the transform that can decompose the original signal into both spatial and frequency space. For that for the first time in the deepfake detection research, we have utilized the Stockwell Transform (S-transform) Stockwell et al. (1996) which maps the input to the simultaneous space of spatial and frequency. The importance of frequency information has been highlighted in the existing synthetic media detection research Wang et al. (2020); Qian et al. (2020); however, it is obtained from the input space only. The input image space can suffer from several bottlenecks such as the impact of illumination, quality, and compression; while the representation obtained from CNN somewhat mitigates these bottlenecks by learning rich invariant features Wang & Yeung (2013); Nanni et al. (2017). Therefore, obtaining the frequency information at the same time retaining the spatial information which is the image representation can help in building a robust detection algorithm. The s-transform helps in achieving that local frequency information Stockwell (2007a;b). Apart from that, the S-transform combines the properties of both Fourier transform and wavelet decomposition which we believe helps in learning robust features in the transformed space Ventosa et al. (2008); Agarwal et al. (2021c). Therefore, decomposing a signal into its corresponding frequency signal along with that spatial information can highlight those individual artifacts helpful for the discrimination of fake videos.

In the proposed research, the spatial axis is the k dimensional representation of the input image computed from the CNN. The spatial-frequency decomposition through S-transform generates the complex 2D matrix of size $k \times z$ from the 1D representation of an image, where k is the size of the image representation vector and z is the frequency component computed over each value. From the complex values of spatial-frequency decomposition, we have computed the absolute value for further feature extraction and classification.

Feature Extraction: Fig. 4 shows the magnitude spectrum of the S-transform computed from the real and various types of manipulation videos. Each manipulation shows a significant difference in the frequency spectrum (y-axis) across the time-axis (feature vector) from the real videos. Such significant differences in the manipulation cover a range of techniques such as identity swap (FaceSwap Kowalski (2018) and DeepFake deepfakes (2018)) and expression swap (Face2Face Thies et al. (2016)) in the spatial-frequency decomposition makes them perfect for their detection. Therefore, based on such discriminative features of the magnitude values of the spatial-frequency decomposition, we have utilized that for feature extraction. The wavelet energy features at three levels are computed along the frequency axis which acts as the summation of the frequency information at an interval. It leads to a matrix of dimension $k \times 3$, where k is the dimensionality of the image representation and 3 represents the number of energy features. As seen from Fig. 4 some frequency stamps are highly discriminative as compared to others. Therefore, we assert that preserving those spectra can provide better classification performance and further reduce the computational cost. To find the discriminative directions best suitable for classification, principal component analysis (PCA) Pearson (1901); Wold et al. (1987) has been applied to the extracted features from the training set. For the computation of PCA, the feature matrix of size $k \times 3$ is first converted into the vector of size $1 \times (k \times 3)$. Additionally, Fig. 6 also depicts the separation capability of the proposed transformation applied to the feature space.

Classification: Once the features are extracted and only the essential components are preserved, a linear support vector machine (SVM) Cortes & Vapnik (1995) classifier is trained for the detection of manipulated images. The parameters of the classifier are optimized using a grid search on the training or validation set. We assert that the proposed features transformed the images into space where they are linearly separable and restrict the need for complex non-linear classifiers. Therefore, the linear binary classifier is trained to classify the images into either real or fake classes. Fig. 5 shows the overall structure of the proposed algorithm containing multiple steps covering image representation to classification.

3.1 Implementation Details

We want to *highlight* that the proposed face manipulation detection algorithm is majorly a 'parameter-free'[1] architecture. We have used the ImageNet pre-trained DenseNet Huang et al. (2017) CNN network to extract the representation of the input images. The feature representation is of the dimensionality 1×1024. The representation is then converted into the spatial-frequency decomposition. The decomposition used in this research does not yield any parameter and the default parameter of the transform has been applied Dash (2021). The transformation yields the complex-valued matrix of dimension 513×1024, where 513 is the frequency dimension and 1024 is the representation dimension. Later, the magnitude of the spatial-frequency matrix is computed for wavelet energy feature extraction. Again the features extraction step is parameter-free. The only parameters lie in the binary classifier; however, they also do not contain significant parameters.

[1]It represents that the trainable parameters in the proposed architecture are lying with the classifier only. The CNN has the parameters, although they are fixed in our setting.

Table 1: Manipulation-specific detection Accuracy (%) of the existing and proposed algorithm. The results are reported on raw and compressed datasets of all four manipulation methods (DF: DeepFakes, F2F: Face2Face, FS: FaceSwap, and NT: NeuralTextures). The accuracy on each subset also includes the accuracy on the pristine (real) videos, i.e., the reported accuracy is the average performance on both the clean and attack images/videos.

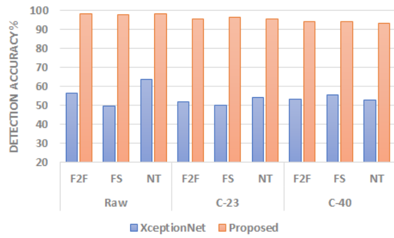
Algorithm	Raw				C-23				C-40			
	DF	F2F	FS	NT	DF	F2F	FS	NT	DF	F2F	FS	NT
Fridrich et al. Fridrich & Kodovsky (2012)	99.03	99.13	98.27	99.88	77.12	74.68	79.51	76.94	65.58	57.55	60.58	60.69
Cozzolino et al. Cozzolino et al. (2017)	98.83	98.56	98.89	99.88	81.78	85.32	85.69	80.60	68.26	59.38	62.08	62.42
Bayar and Stamm Bayar & Stamm (2016)	99.28	98.79	98.98	98.78	90.18	94.93	93.14	86.04	80.95	77.30	76.83	72.38
Rahmouni et al. Rahmouni et al. (2017)	98.03	98.96	98.94	96.06	82.16	93.48	92.51	75.18	73.25	62.33	67.08	62.59
Afchar et al. Afchar et al. (2018)	98.41	97.96	96.07	97.05	95.26	95.84	93.43	85.96	89.52	84.44	83.56	75.74
Chollet et al. Chollet (2017)	99.59	99.61	99.14	99.36	98.85	98.36	98.23	94.50	94.28	91.56	93.70	82.11
Liu et al. Liu et al. (2021b)	–	–	–	–	97.45	98.33	97.20	90.84	–	–	–	–
Wu et al. Wu et al. (2020)	–	–	–	–	–	–	–	–	95.33	90.48	94.09	–
Liu et al. Liu et al. (2021a)	–	–	–	–	–	–	–	–	93.48	86.02	92.26	76.78
Qian et al. Qian et al. (2020)	–	–	–	–	–	–	–	–	97.97	95.32	96.53	83.32
Zhao et al. Zhao et al. (2021b)	–	–	–	–	–	–	–	–	99.73	96.38	98.20	91.79
Trans-FCA Tan et al. (2022)	–	–	–	–	98.10	–	–	–	91.43	–	–	–
Agarwal et al. Agarwal et al. (2021a)	–	–	–	–	98.82	99.19	99.10	94.55	97.34	93.57	94.64	81.78
MRE-Net Pang et al. (2023)	–	–	–	–	99.64	99.64	99.28	97.50	98.21	97.50	96.07	89.64
Zhao et al. Zhao et al. (2023)	–	–	–	–	99.60	99.60	100.00	96.80	98.90	96.10	97.50	92.10
She et al. She et al. (2024)	99.87	99.16	98.91	90.84	–	–	–	–	–	–	–	–
Proposed	**99.90**	**99.99**	**99.97**	**99.98**	**99.97**	**99.95**	**99.15**	**99.50**	**99.90**	**99.65**	**99.00**	**98.70**

The proposed **'parameter-free'** architecture makes it a suitable choice for its real-world deployment for the cost-effective device and advanced computing devices. Face manipulation is not only affecting the celebrity class of society but can be harmful to any common person as well. Therefore, the availability of an algorithm that everyone in society can use without worrying about the heavy computation resources can be impactful. The proposed architecture is *'energy-efficient* which should also not be ignored because of the high impact of ML algorithms on the environment Lu (2019).

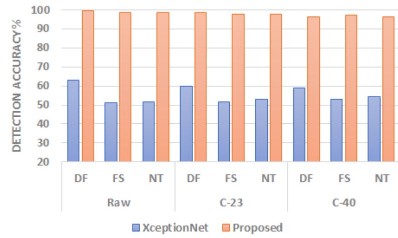
4 Experimental Databases

In this research, we have used two popular face manipulation databases namely FaceForensics++ (FF++) Rossler et al. (2019), Celeb-DF Li et al. (2020c), DFinal Ciftci et al. (2020), and deepfake in-the-wild (DFW) Zi et al. (2020). FF++ is one of the most popular databases because it covers a large spectrum of manipulation concerning identity swaps and expression swaps. We would also like to mention that the videos in the database contain three variations: (i) Raw (original quality), ii) HQ (quantization parameter equal to 23 and referred to as *C-23*), and iii) LQ (quantization parameter equal to 40 and referred to as *C-40*). The compression variations are generated to simulate the real-world processing generally performed on social media uploads. In total, each variant contains 5000 videos, out of which 1000 belongs to each class, i.e., real and four manipulation types.

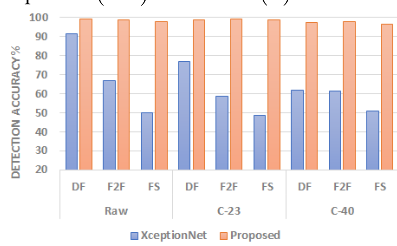
Celeb-DF and DFinal overcome the limitation of several existing low-quality DeepFake databases by generating the high-quality version which was released over several social media platforms CtrlShiftFace (2020). The one probable drawback of the above datasets is that they might not truly reflect the real-world case study of deepfake detection. With this intuition and better support for real-world deepfake detection research, Zi et al. Zi et al. (2020) have collected a deepfake in-the-wild dataset. The experiments on each dataset are performed using the pre-defined protocol mentioned by the authors of the original paper. For example, the FF++ database comes with a pre-defined training, validation, and testing set and we have used these pre-defined settings for the experimentations and fair comparisons. The results in the paper are reported on the dataset containing both real and attack classes, not just one class.



(a) Train on DeepFake (DF)



(b) Train on Face2Face (F2F)



(c) Train on NeuralTextures (NT)

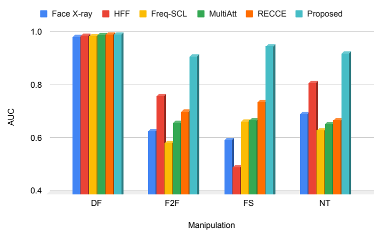
Figure 7: **Attack agnostic** classification accuracy (%) of the best FF++ model, i.e., XceptionNet and proposed algorithm. The detection algorithms are trained on a specific type of manipulation and evaluated on others in the FF++ database. (Best view in color.)

5 Experimental Results and Analysis

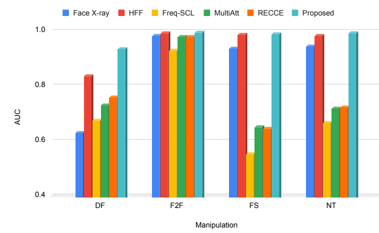
In this research, we have conducted the most rigorous experiments and utilized every potential aspect of the databases. To perform the experiments and report the results, the videos in both the databases are preprocessed such as face regions are cropped and normalized to the fixed resolution of dimension $128\times128\times3$. The performance is reported in terms of video-based classification: where an entire video is classified as real or manipulated.

5.1 Seen Attack: Traditional Settings

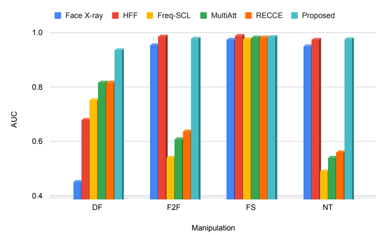
When the algorithms are trained on each specific manipulation set and tested on that particular manipulation set, the results of the proposed algorithm along with existing algorithms are reported in Table 1. The existing algorithms used by Rossler et al. Rossler et al. (2019) range from hand-crafted features with a traditional classifier to deep neural networks. From the results, it is clear that the steganalysis features yield comparable or better performance than several deep neural networks on 'raw' quality data. Another advantage of such an algorithm is its computational efficiency; however, the algorithm suffers severely on compressed quality images and the detection performance degrades at least 17% and 23% on light-compressed and heavily compressed videos, respectively. The performance of the fine-tuned XceptionNet is the best among all the existing algorithms used in the paper Rossler et al. (2019). The architecture yields approximately perfect detection performance on 'raw' quality videos. While on mild compression, the XceptionNet is found robust; the architecture shows a significant drop on heavily compressed videos. Another critical disadvantage of the architecture is the robustness against a variety of manipulation types. For example, the architecture was found robust against 'identity swap' manipulation, i.e., DF and FS, but yields significantly lower performance on 'expression swap' manipulation especially 'NeuralTexture'. As shown in Fig. 4 and Fig. 6, the proposed algorithm depicts clear differences among the different data types including real and manipulated, which gets reflected in the detection performance. The proposed algorithm shows either **'near-perfect'** or state-of-the-art (SOTA) detection accuracy on each manipulation type and quality type. Such universal detection performance which is agnostic to manipulation types and their quality makes the proposed approach an ideal choice for its real-world deployment.



(a) Train on DeepFake (DF)



(b) Train on Face2Face (F2F)



(c) Train on Face swap (FS)

Figure 8: Cross attack evaluation and comparison with the recent state-of-the-art algorithms. The existing algorithms are: Multi-Att Zhao et al. (2021a), Face X-ray Li et al. (2020b), Freq-SCL Li et al. (2021), HFF Luo et al. (2021), and RECCE Cao et al. (2022).

5.2 Cross Attack and Data-Quality

To fully utilize the potential of the FF++ database which is missing from the existing literature, we have performed an evaluation study where one type of manipulation has been used in the training and other manipulations are used for testing. In these experiments, only one *variable* (manipulation type) has changed while another variable (data quality) remains fixed. In other words, when training the detector on the 'raw' quality of one manipulation it is only tested against the 'raw' quality data of other manipulations. As mentioned earlier this type of study is missing from the literature, therefore, we have run the experiments using the best-performing architecture, i.e., XceptionNet, and finetuned it for cross-attack evaluation.

The XceptionNet architecture which shows high detection performance on seen attack images/videos, failed significantly when the attack types were not seen at the time of training. Interestingly we have seen that when the XceptionNet has seen the attack images at the time of training it yields at least 99%, 94.5%, and 82.11% detection accuracy on 'raw', C-23, and C-40 quality images, respectively. Whereas, when the network is tested against unseen attack images where at a time one attack is used for training, the XceptionNet yields only 53.06%, 50.24%, 52.08% average detection accuracy on 'raw', C-23, and C-40 quality images, respectively Majumdar et al. (2021). A similar low performance against unseen attacks is observed by Cao and Gong Cao & Gong (2021). We have also made a comparison with the recently reported results by Liu et al. Liu et al. (2021b). When the MesoNet, XceptionNet, LAE Du et al. (2020) models are trained on the F2F set and tested on the FS attack, they yield 47.32%, 49.94%, 63.15%, respectively. The ADD method proposed by Liu et al. Liu et al. (2021b) also shows a severe drop in the detection performance and yields only 67.02% accuracy. The huge performance drops show the limitation of the existing state-of-the-art network and demand an effective attack agnostic detection algorithm. The proposed research fulfills that gap by producing an attack and data quality agnostic algorithm. The proposed algorithm is not only effective in the case of a seen attack and yields the SOTA detection rate but performs similarly even under unseen attack settings as well. The results of this finding are shown in Fig. 7.

Fig. 8 shows the comparison of the proposed algorithm with two recently proposed generalized deepfake detection algorithms namely Multi-Att Zhao et al. (2021a), Face X-ray Li et al. (2020b), Freq-SCL Li et al. (2021), HFF Luo et al. (2021), and RECCE Cao et al. (2022). When the existing algorithms are trained on the DF subset and tested on the remaining, they are found highly ineffective as compared to when

Table 2: Attack and quality agnostic performance of the proposed algorithm on the FF++ database. The proposed algorithm is duly agnostic against both the variables in the database, i.e., attack types and image quality.

Training		Agnostic Variables		Average
Attack	Quality	Attack	Quality	Accuracy
DF	Raw	FS, F2F, and NT	C-23 and C-40	**98.61 ± 0.43**
	C-23		Raw and C-40	
	C-40		Raw and C-23	
FS	Raw	DF, F2F, and NT	C-23 and C-40	
	C-23		Raw and C-40	
	C-40		Raw and C-23	
F2F	Raw	DF, FS, and NT	C-23 and C-40	
	C-23		Raw and C-40	
	C-40		Raw and C-23	
NT	Raw	DF, FS, and F2F	C-23 and C-40	
	C-23		Raw and C-40	
	C-40		Raw and C-23	

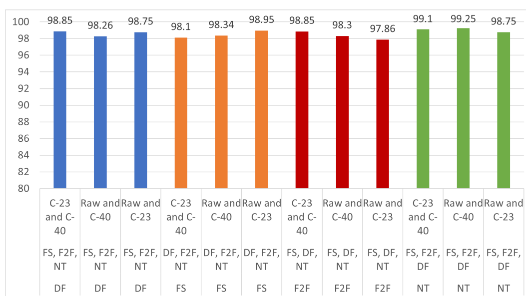


Figure 9: Attack and quality agnostic performance of the proposed algorithm on the FF++ database. For example, the proposed algorithm is trained on DF (first three bars) and tested on remaining attacks of different testing image qualities (leftmost bar: trained on Raw quality and tested on unseen image qualities C23 and C40). The lower variation among the accuracies showcases that the proposed is robust and not biased towards any particular attack or data quality.

they are tested on the DF subset. Whereas, the proposed algorithm yields an attack-agnostic nature and yields significantly better performance than the existing algorithms. In an interesting observation, when the existing algorithms are tested on identity manipulation attacks such as DF and FS, their performance degrades significantly under unseen attack training-testing settings. In contrast, on the expression evaluation manipulations such as F2F and NT, their performance is significantly higher. In both cases, the proposed algorithm yields higher AUC and establishes its attack agnostic strength.

If the proposed algorithm encounters both variables, i.e., attack types and data quality, it is found 'duly' agnostic and yields more than 96% detection performance as reported in Table 2. The prime advantage of such dual robustness is that in the real world, an attacker can come up with a new attack and perform some different degradation that might not be seen at the time of training. Hence, the detection algorithm must tackle all such unwanted testing conditions that are still missing in the literature and future research must tackle all such extensive evaluation settings. The accuracy of individual experiments about cross quality and attack are reported in Fig. 9 showcase that the proposed algorithm is unbiased in handling different attacks and data quality.

Table 3: Cross-dataset evaluation (AUC) and comparison with complex state-of-the-art algorithms on Celeb-DF Li et al. (2020c). The results on FF++ are reported under the same database training-testing and most of the algorithms achieve almost perfect performance. However, each algorithm suffers a drastic drop in performance when an unseen database comes except the proposed algorithm.

Method	FF++	Celeb-DF
Meslnception4 Afchar et al. (2018)	0.83	0.53
Xception-raw Rossler et al. (2019)	0.99	0.48
Xception-c23 Rossler et al. (2019)	0.99	0.65
Xception-c40 Rossler et al. (2019)	0.95	0.65
Multi-task Nguyen et al. (2019a)	0.76	0.54
Capsule Nguyen et al. (2019b)	0.96	0.57
DSP-FWA Li & Lyu (2019)	0.93	0.64
Face-XRay Li et al. (2020b)	0.99	0.74
F^3-Net Qian et al. (2020)	0.97	0.65
EfficientNet-B4 Tan & Le (2019)	0.99	0.64
MD-CSDNetwork Agarwal et al. (2021a)	0.99	0.68
Nirkin et al. Nirkin et al. (2021)	0.99	0.66
ProtoPNet Chen et al. (2018)	0.98	0.69
DPNet Trinh et al. (2021)	0.99	0.68
DPNet - c40 Trinh et al. (2021)	0.90	0.72
M2TR Wang et al. (2021)	0.99	0.66
Multi-Attention Zhao et al. (2021a)	0.99	0.67
MFFNet Zhao et al. (2021b)	0.99	0.75
ADD Liu et al. (2021b)	0.97	0.66
STIL Gu et al. (2021)	0.97	0.75
PCL + I2G Zhao et al. (2021c)	0.99	0.81
CORE Ni et al. (2022)	0.99	0.79
RECCE Cao et al. (2022)	0.99	0.69
UIA-ViT Zhuang et al. (2022)	0.99	<u>0.82</u>
SCL-KD Lin et al. (2022)	0.99	0.69
Trans-FCA Tan et al. (2022)	0.99	0.78
Forensics Symmetry Li et al. (2023a)	0.99	0.58
ISTVT Zhao et al. (2023)	0.99	<u>0.84</u>
She et al. She et al. (2024)	0.99	0.92
Yan et al. Yan et al. (2024)	–	0.83
Proposed*	**0.99**	**0.98**

*The proposed algorithm yields more than 94.77% detection accuracy.

5.3 Cross Database Evaluation

Li et al. Li et al. (2020c) have performed a comparison of the detection performance of multiple databases using multiple existing DeepFake detection algorithms. The authors claim that the detection of DeepFake images of the Celeb-DF database is most challenging as compared to several other counterpart databases including DFDC Dolhansky et al. (2019) and DFD Dufour & Gully (2020). Therefore, to further showcase the strength of the proposed algorithm, we have utilized the Celeb-DF database, and the database is used for evaluation under cross-database settings only. Each attack and quality image of FF++ databases is used one at a time for training and evaluated on the Celeb-DF database. The FF++ database contains twelve variants, i.e., four attacks and three data qualities. The results reported in Table 3 show the average performance of all twelve variants, where each variant yields at least 98% detection AUC performance on the Celeb-DF database. Along with AUC, the proposed algorithm archives 92.76% real image detection accuracy and 96.78% deepfake image detection accuracy on the Celeb-DF dataset. The average detection accuracy is 94.77%. The performance of the proposed algorithm is compared against several existing state-of-the-art

Table 4: The results on the individual quality images of the Celeb-DF database show that the proposed algorithm is 'robust' against image quality.

Algorithm	Original	C-23	C-40
FWA	56.9	54.6	52.2
Xception-C23	65.3	65.5	52.5
Xception-C40	65.5	65.4	59.4
DSP-FWA	64.6	57.7	47.2
Proposed	**99.84**	**98.25**	**97.89**

Table 5: Attack, database (DB), and quality agnostic performance of the proposed algorithm through the FF++ and Celeb-DF databases. The proposed algorithm is **'tri'** agnostic against each possible variable in the domain, i.e., attack types, database, and image quality.

Training variables			Test Agnostic Variables			Average
DB	Attack	Quality	DB	Attack	Quality	Accuracy
FF++	DF	Raw	Celeb-DF	DF	C-23 and C-40	**98.46 ± 1.24**
		C-23			Raw and C-40	
		C-40			Raw and C-23	
	FS	Raw			C-23 and C-40	
		C-23			Raw and C-40	
		C-40			Raw and C-23	
	F2F	Raw			C-23 and C-40	
		C-23			Raw and C-40	
		C-40			Raw and C-23	
	NT	Raw			C-23 and C-40	
		C-23			Raw and C-40	
		C-40			Raw and C-23	
Celeb-DF	DF	Raw	FF++	DF, FS, F2F, NT	C-23 and C-40	
		C-23			Raw and C-40	
		C-40			Raw and C-23	

algorithms namely (i) Two-stream Zhou et al. (2017), (ii) Meso4 Afchar et al. (2018), (iii) MesoInception4 Afchar et al. (2018), (iv) HeadPose Yang et al. (2019), (v) FWA Yang et al. (2019), (vi) VA-MLP Matern et al. (2019), (vii) VA-LogReg Matern et al. (2019), (viii) Xception-raw Rossler et al. (2019), (ix) Xception-C23 Rossler et al. (2019), (x) Xception-C40 Rossler et al. (2019), (xi) Multi-task Nguyen et al. (2019a), (xii) Capsule Nguyen et al. (2019b), (xiii) DSP-FWA which is an improved variant of FWA. Apart from these algorithms, the comparison with recent complex algorithms is also performed including MD-CSDNetwork Agarwal et al. (2021a), Nirkin et al. Nirkin et al. (2021), ProtoPNet Chen et al. (2018) DPNet Trinh et al. (2021), DPNet - c40 Trinh et al. (2021), M2TR Wang et al. (2021), Multi-Attention Zhao et al. (2021a), MFFNet Zhao et al. (2021b), ADD Liu et al. (2021b), STIL Gu et al. (2021), and PCL + I2G Zhao et al. (2021c). Another desired property of a strong defense is the effectiveness against unseen databases and the performance of the proposed algorithm in that agnostic direction further claims that it is the best possible option for manipulation detection.

Similar to the FF++ database, the Celeb-DF also contains the videos in three image qualities: (i) raw, (ii) C-23, and (iii) C-40. It can be seen from Table 4 that the performance of the existing algorithms significantly degrades with the quality of the images/videos. Whereas, the performance of the proposed algorithm remains the same across each data quality. We want to mention that the proposed algorithm is trained only on FF++ (not seen Celeb-DF images) and tested on individual quality images of Celeb-DF.

5.4 Cross Quality, Attack, and Database Variations

The strong manipulation detection algorithm must have considered all possible variables while evaluating the performance. The three variables that might be possible in the detection databases are (i) attacks (identity swap or expression swap), (ii) quality (raw or compressed), and (iii) database. We have earlier shown the effectiveness of the proposed algorithm under two variables using the FF++ database and the third using Celeb-DF testing. However, now we have to consider all three parameters simultaneously and evaluate the

Table 6: DFinal database Ciftci et al. (2020) results. The proposed algorithm yields state-of-the-art deepfake detection performance on such a challenging high-quality dataset.

Algorithm	Face ↑	Video ↑
Simple CNN	54.56	48.88
InceptionV3 Szegedy et al. (2016)	60.96	68.88
Xception Chollet (2017)	56.11	75.55
ConvLSTM Xingjian et al. (2015)	44.82	48.83
V1 Tariq et al. (2018)	–	82.22
V3 Tariq et al. (2018)	–	73.33
Emsemble Tariq et al. (2018)	–	80.00
Fake Catcher Ciftci et al. (2020)	87.62	91.07
Proposed	**94.65**	**97.80**

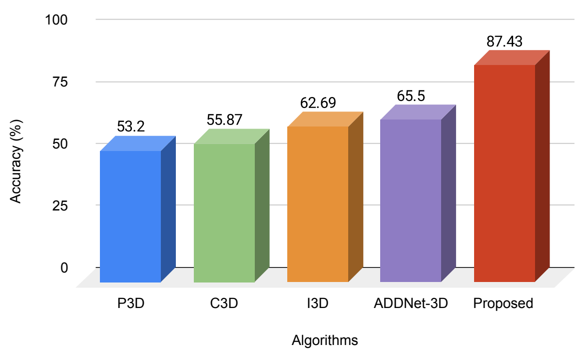


Figure 10: Deepfake detection accuracy (%) on the DFW dataset. Our performance is compared with I3D Carreira & Zisserman (2017), P3D Qiu et al. (2017), C3D Tran et al. (2014), ADDNet-3D Zi et al. (2020).

agnostic nature of the proposed algorithm against them. The finding showcased in Table 5 establishes the desired effect of the proposed algorithm which is missing in the existing algorithms.

5.5 In-the-wild Detection

While the previous datasets are popular benchmark datasets in the deepfake detection study, the recent advancement in deepfake research has introduced several high-quality videos that reflect real-world conditions. Henceforth, to study the effectiveness of our proposed algorithm in deepfake detection, we have utilized two datasets namely DFinal Ciftci et al. (2020) and DFW Zi et al. (2020). The DFinal dataset is claimed to be a high-quality deepfake video dataset and is challenging as compared to other datasets. The detection results of the proposed algorithm along with state-of-the-art algorithms are reported in Table 6. The detection results on another challenging dataset namely DFW are also reported in Fig. 10. The proposed algorithm surpasses the existing algorithms by at least 21.93%. The effectiveness of the proposed algorithm on such challenging datasets and surpassing several state-of-the-art algorithms reflects its superiority in identifying the deepfake threat. The extensive experimental evaluation helps in establishing our aim of developing a unified deepfake detection algorithm that is agnostic to several challenging dimensions such as dataset, manipulation, and image quality.

5.6 Analysis

We have also studied the current strength and bottleneck of the proposed algorithm which can potentially help in building an advanced version of the algorithm and further detecting the deepfake samples. Fig. 11 shows some correctly and incorrectly classified samples of both classes. The prime reasons for misclassification

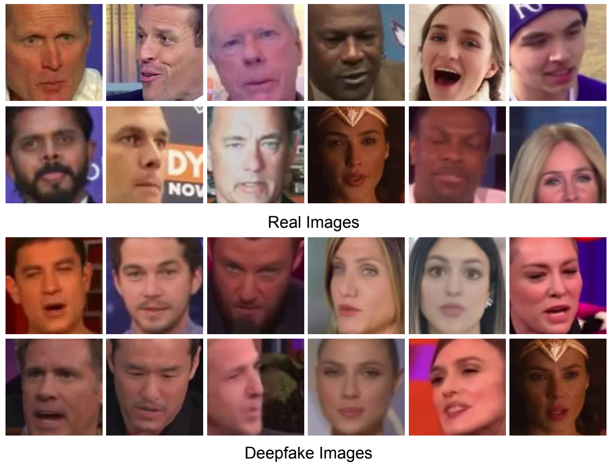


Figure 11: Correct and incorrect prediction samples to study what are the current bottlenecks and strengths of the proposed algorithm. The first row in each category is correctly classified samples and the second row is the incorrect predictions. Interestingly, the majority of the incorrect predictions are somewhat blurry, low quality, and contain a low amount of face region.

we can see from the images are low quality, blurriness of the images, and low facial region. Although, we want to highlight it is not observed that each image with these characteristics gets misclassified; however, we assert that these might be the potential reasons for incorrect prediction. We want to highlight the interesting strength of the proposed algorithm: it shows good resilience in handling poses, expression, contrast, and quality of the images. On top of that, we have observed there is no bias towards gender modality and the algorithm is fair in detecting both gender images.

We have also analyzed the performance of the proposed algorithm concerning individual classes, i.e., real and attack. The prime reason for such evaluation is that the existing algorithms are found highly biased towards the attack class and yield poor performance of the real class Nirkin et al. (2021); Rossler et al. (2019). As mentioned earlier the reported results correspond to both real and fake classes. The proposed deepfake detection algorithm is not found biased towards any particular class and yields approximately similar ($\pm 0.56\%$) accuracy in both classes. For example, if the proposed algorithm yields 98.70% NT examples detection accuracy (Table 1), then the detection accuracy on real images is 98.64% and detection accuracy of NT attack is 98.76%.

5.7 Discussion

S-transform helps in capturing the local frequency information. Research works show that synthetic media has either suppressed frequency information or has them in excess Durall et al. (2020). S-transform is effective in detecting these frequency bursts. Frequency information has proven effective in the detection of various synthetic media and adversarial noise detection and is one of the primary differential components between real and fake images Agarwal et al. (2021a;c); Frank et al. (2020). On top of that, the proposed algorithm also utilizes the wavelet energy features from transformed features. The wavelet energy features contain the information related to both approximation (low frequency) and high-frequency content of an image) Akbarizadeh (2012). Fig. 4 also shows that the energy information in the real images is highly preserved and suppressed in the fake images of different types. Therefore, The proposed combination of S-transformation which attenuates the frequency artifacts, and acquisition of these artifacts in the form of energy features through wavelet decomposition pave the way for an effective deepfake detection algorithm. The extensive experimental evaluation backs our understanding and yields state-of-the-art and generalized deepfake detection performance. However, as shown in Fig. 11, the proposed algorithm is not perfect and we should not be confused that the field of deepfake detection is entirely solved. We believe further advancement is needed to tackle several real-world variations that might be exhibited in future advanced deepfake videos.

5.8 Low Carbon Emission or Green Computing

The proposed research not only aims to build a robust solution but also to reduce carbon emissions. All the experiments are performed using standard datasets and benchmark protocols. As mentioned earlier, the steps involved in the proposed approach are parameter-free excluding the parameter optimization of the SVM classifier. We want to highlight that the linear SVM classifiers are also not high enough to yield significant computational costs. On the FF+ dataset, the proposed algorithm with resolution took approximately 600 seconds. The computational time is computed on the RTX 2080 single GPU system. As we can see the computational time of the proposed algorithm is significantly lower than the existing algorithms that either take multiple hours/days for training or are not generalized against unseen settings.

6 Conclusion and Impact

Manipulated images/videos including DeepFakes have created havoc among society, research personnel, and various governments. The manipulation can be broadly classified into two groups: (i) identity swap: where an entire face of a person is swapped with the face of another person, and (ii) expression swap: where only certain part(s) of a face are manipulated for the desirable effects. By looking at the seriousness of the issue, in this research, we have proposed a state-of-the-art detection algorithm based on simultaneous spatial-frequency signal transformation. The algorithm is evaluated on two large-scale databases that constitute various manipulation types. The experiments are conducted on several challenging scenarios which are still missing from the literature on defense. The proposed algorithm is found state-of-the-art reaching an almost perfect rate at each desirable real-world condition including unseen attack, unseen image degradation due to compression, and unseen database. On one of the challenging high-quality databases namely Celeb-DF, the proposed algorithm is at least 33% better than the state-of-the algorithms. Another significant advantage of the proposed algorithm is its almost *'parameter-free'* nature and hence does not need heavy computational resources and saves energy significantly. The said advantages make it a viable option for anyone around the globe to use without worrying about computational requirements.

References

D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE WIFS*, pp. 1–7, 2018. doi: 10.1109/WIFS.2018.8630761.

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE WIFS*, pp. 1–7, 2018.

A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE IJCB*, pp. 659–665, 2017.

Aayushi Agarwal, Akshay Agarwal, Sayan Sinha, Mayank Vatsa, and Richa Singh. MD-CSDNetwork: Multi-domain cross stitched network for deepfake detection. In *IEEE F&G*, 2021a.

Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *IEEE BTAS*, pp. 1–6, 2016.

Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE IJCB*, pp. 659–665, 2017a.

Akshay Agarwal, Daksha Yadav, Naman Kohli, Richa Singh, Mayank Vatsa, and Afzel Noore. Face presentation attack with latex masks in multispectral videos. In *IEEE CVPRW*, pp. 81–89, 2017b.

Akshay Agarwal, Akarsha Sehwag, Mayank Vatsa, and Richa Singh. Deceiving the protector: Fooling face presentation attack detection algorithms. In *IEEE ICB*, pp. 1–6, 2019a.

Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. Magnet: Detecting digital presentation attacks on face recognition. *Frontiers in Artificial Intelligence*, 4, 2021b.

Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Image transformation-based defense against adversarial perturbation on deep learning models. *IEEE TDSC*, 18(5):2106–2121, 2021c. doi: 10.1109/TDSC.2020.3027183.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE CVPRW*, pp. 38–45, 2019b.

Gholamreza Akbarizadeh. A new statistical-based kurtosis wavelet energy feature for texture recognition of sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4358–4368, 2012. doi: 10.1109/TGRS.2012.2194787.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE F&G*, pp. 59–66, 2018.

Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM WIHMS*, pp. 5–10, 2016.

E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE spectrum*, 4(12):63–70, 1967.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4122, 2022.

Xiaoyu Cao and Neil Zhenqiang Gong. Understanding the security of deepfake detection. *arXiv preprint arXiv:2107.02045*, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, pp. 103–120, 2020.

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE CVPR*, pp. 1251–1258, 2017.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *ACM MM*, pp. 439–447, 2020.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Jesselyn Cook. Deepfake videos and the threat of not knowing what's real. `https://www.huffpost.com/entry/deepfake-videos-and-the-threat-of-not-\knowing-whats-real$_$n$_$5cf97068e4b0b08cf7eb2278`, 2020. Accessed in Feb 2021.

James W Cooley, Peter AW Lewis, and Peter D Welch. The fast fourier transform and its applications. *IEEE ToE*, 12(1):27–34, 1969.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *ACM WIHMS*, pp. 159–164, 2017.

CtrlShiftFace. Ctrl shift face. `https://www.youtube.com/channel/UCKpH0\CKltc73e4wh0_pgL3g`, 2020. Accessed in Feb 2021.

Baba Dash. Stockwell transform (s-transform). `https://www.mathworks.com/matlabcentral/\fileexchange/45848-stockwell-transform-s-transform`, 2021. Accessed in Feb 2021.

deepfakes. Faceswap. `https://github.com/deepfakes/faceswap`, 2018. Accessed in Feb 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pp. 248–255, 2009.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. In *CIKM*, 2020.

Nick Dufour and Andrew Gully. Deepfakes detection dataset by google & jigsaw. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`, 2020.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.

Tharindu Fernando, Clinton Fookes, Simon Denman, and Sridha Sridharan. Detection of fake and fraudulent faces via neural memory networks. *IEEE Transactions on Information Forensics and Security*, 16:1973–1988, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pp. 3247–3258, 2020.

Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE TIFS*, 7(3):868–882, 2012.

Jamie Good. Deepfake voice, ceo voice deepfake blamed for scam that stole 243 000. `https://www.ign.com/articles/2019/09/05/240000-stolen-in-worlds-first-artificial-intelligence-heist`, 2021.

Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *IJCV*, 127(6):719–742, 2019.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM MM*, pp. 3473–3481, 2021.

Drew Harwell. Scarlett johansson on fake ai-generated sex videos. `https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex\-videos-nothing-can-stop-someone-cutting-\pasting-my-image/?noredirect=on`, 2018. Accessed in Feb 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE CVPR*, pp. 4700–4708, 2017.

Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.

Marek Kowalski. Face swap. `https://github.com/MarekKowalski/FaceSwap`, 2018. Accessed in Feb 2021.

Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *IEEE/CVF WACV*, pp. 2589–2597, 2020.

Gen Li, Xianfeng Zhao, and Yun Cao. Forensic symmetry for deepfakes. *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2023a. doi: 10.1109/TIFS.2023.3235579.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 6458–6467, 2021.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE CVPR*, 2020a.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020b.

Qilei Li, Mingliang Gao, Guisheng Zhang, and Wenzhe Zhai. Defending deepfakes by saliency-aware attack. *IEEE Transactions on Computational Social Systems*, pp. 1–8, 2023b. doi: 10.1109/TCSS.2023.3271121.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE CVPRW*, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF CVPR*, pp. 3207–3216, 2020c.

Lingyu Liang, Lianwen Jin, and Yong Xu. Pde learning of filtering and propagation for task-aware facial intrinsic image analysis. *IEEE Transactions on Cybernetics*, 52(2):1021–1034, 2022. doi: 10.1109/TCYB.2020.2989610.

Yuzhen Lin, Han Chen, Bin Li, and Junqiang Wu. Towards generalizable deepfake face forgery detection with semi-supervised learning and knowledge distillation. In *IEEE International Conference on Image Processing (ICIP)*, pp. 576–580, 2022. doi: 10.1109/ICIP46576.2022.9897792.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *IEEE/CVF CVPR*, pp. 772–781, 2021a.

Ping Liu, Yuewei Lin, Yang He, Yunchao Wei, Liangli Zhen, Joey Tianyi Zhou, Rick Siow Mong Goh, and Jingen Liu. *arXiv preprint arXiv:2106.10705*, 2021b.

Donna Lu. Creating an ai can be five times worse for the planet than a car. `https://www.newscientist.com/article/2205779-creating-an-ai-can-be-five-times-\worse-for-the-planet-than-a-car/`, 2019. Accessed in Feb 2021.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16317–16326, 2021.

Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Evading face recognition via partial tampering of faces. In *IEEE/CVF CVPRW*, pp. 0–0, 2019.

Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Facial retouching and alteration detection. *Handbook of Digital Face Manipulation and Detection From DeepFakes to Morphing Attacks*, 2021.

Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pp. 667–684, 2020.

Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE WACVW*, pp. 83–92, 2019.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1): 1–41, 2021.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM MM*, pp. 2823–2832, 2020.

Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71:158–172, 2017.

Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE BTAS*, pp. 1–8, 2019a. doi: 10.1109/BTAS46853.2019.9185974.

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE ICASSP*, pp. 2307–2311, 2019b. doi: 10.1109/ICASSP.2019.8682602.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12–21, 2022.

Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE TPAMI*, 2021.

Guilin Pang, Baopeng Zhang, Zhu Teng, Zige Qi, and Jianping Fan. Mre-net: Multi-rate excitation network for deepfake video detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023. doi: 10.1109/TCSVT.2023.3239607.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pp. 86–103, 2020.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, 2017.

Ramachandra Raghavendra, KiranB Raja, Sushma Venkatesh, and Christoph Busch. Face morphing versus face averaging: Vulnerability and detection. In *2017 IEEE IJCB*, pp. 555–563, 2017.

Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *IEEE WIFS*, pp. 1–6, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF ICCV*, pp. 1–11, 2019.

Nilay Sanghvi, Sushant Kumar Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Mixnet for generalized face presentation attack detection. *ICPR*, 2020.

Huimin She, Yongjian Hu, Beibei Liu, Jicheng Li, and Chang-Tsun Li. Using graph neural networks to improve generalization capability of the models for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2024.

Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, volume 34, pp. 13583–13589, 2020.

RG Stockwell. Why use the s-transform. *Pseudo-differential operators: partial differential equations and time-frequency analysis*, 52:279–309, 2007a.

Robert Glenn Stockwell. A basis for efficient representation of the s-transform. *Digital Signal Processing*, 17(1): 371–393, 2007b.

Robert Glenn Stockwell, Lalu Mansinha, and RP Lowe. Localization of the complex spectrum: the s transform. *IEEE TSP*, 44(4):998–1001, 1996.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *ICML*, volume 97, pp. 6105–6114, 2019. URL `http://proceedings.mlr.press/v97/tan19a.html`.

Zichang Tan, Zhichao Yang, Changtao Miao, and Guodong Guo. Transformer-based feature compensation and aggregation for deepfake detection. *IEEE Signal Processing Letters*, pp. 1–5, 2022. doi: 10.1109/LSP.2022.3214768.

Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *international workshop on multimedia privacy and security*, pp. 81–87, 2018.

The Telegraph. Deepfake video of volodymyr zelensky surrendering surfaces on social media. `https://www.youtube.com/watch?v=X17yrEV5sl4&t=4s`, 2022.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE CVPR*, pp. 2387–2395, 2016.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance. In *International Conference on Pattern Recognition*, pp. 442–456. Springer, 2021.

Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2(7):8, 2014.

Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable and trustworthy deepfake detection via dynamic prototypes. In *IEEE/CVF WACV*, pp. 1973–1983, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Sergi Ventosa, Carine Simon, Martin Schimmel, Juan Jose Dañobeitia, and Antoni Mànuel. The s-transform from a wavelet point of view. *IEEE Transactions on Signal Processing*, 56(7):2771–2780, 2008.

Anh H Vo, Minh Thanh Vo, Tuong Le, et al. A novel framework for trash classification using deep transfer learning. *IEEE Access*, 7:178631–178639, 2019.

Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2tr: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770*, 2021.

Naiyan Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. *Advances in neural information processing systems*, 26, 2013.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *IEEE/CVF CVPR*, pp. 8695–8704, 2020.

Yaohui Wang and Antitza Dantcheva. A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In *IEEE F&G*, 2020.

Zihan Wang, Olivia Byrnes, Hu Wang, Ruoxi Sun, Congbo Ma, Huaming Chen, Qi Wu, and Minhui Xue. Data hiding with deep learning: A survey unifying digital watermarking and steganography. *IEEE Transactions on Computational Social Systems*, 10(6):2985–2999, 2023. doi: 10.1109/TCSS.2023.3268950.

WION. Deepfake of south korea's presidential candidate al yoon ahead of election. `https://www.youtube.com/watch?v=yIUTvPOXkk8`, 2022.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *IEEE ICASSP*, pp. 2952–2956, 2020.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *AAAI*, volume 30, 2016.

SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pp. 802–810, 2015.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE ICASSP*, pp. 8261–8265, 2019.

Le-Bing Zhang, Fei Peng, and Min Long. Face morphing detection using fourier spectrum of sensor pattern noise. In *IEEE ICME*, pp. 1–6, 2018.

Cairong Zhao, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. Istvt: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2023.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pp. 2185–2194, 2021a.

Lei Zhao, Mingcheng Zhang, Hongwei Ding, and Xiaohui Cui. Mff-net: Deepfake detection network based on multi-feature fusion. *Entropy*, 23(12):1692, 2021b.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF ICCV*, pp. 15023–15033, 2021c.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *IEEE CVPRW*, pp. 1831–1839, 2017. doi: 10.1109/CVPRW.2017.229.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *IEEE/CVF ICCV*, pp. 14800–14809, 2021.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision (ECCV)*, 2022.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020.