Forensic Science International: Digital Investigation 42 (2022) 301395

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2022 USA - Proceedings of the Twenty-Second Annual DFRWS USA

Deepfake noise investigation and detection

Tianyi Wang [a], Ming Liu [b], Wei Cao [b], Kam Pui Chow [a, *]

[a] *The University of Hong Kong, Pok Fu Lam, Hong Kong, China*
[b] *Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250014, Shandong, China*

ARTICLE INFO

Article history:

Keywords:
Deepfake detection
Digital image forensics
Noise trace extraction and investigation
Siamese network

ABSTRACT

The fast development of Deepfake has brought huge current and potential future negative impacts to our daily lives. As the circulating popular Deepfake videos have become difficult to be distinguished by human eyes, various Deepfake detection approaches have been attempted using deep learning models. However, even though some existing detection methods have achieved reasonable detection performance with respect to the statistical evaluation metrics, the actual underlying Deepfake forensic traces have been barely discussed. In this study, we investigate the special forensic noise traces within Deepfake image frames and propose a noise-based Deepfake detection model approach using a deep neural network. We train a Siamese noise extractor using a novel face-background strategy to investigate different forensic noise traces of a synthesized face area and an unmodified background area. A similarity matrix module is proposed to analyze the forensic noise trace difference between a cropped face square and a cropped background square from a candidate image frame for the Deepfake detection task. As a result, our proposed model trained on the high-quality Celeb-DF dataset has achieved state-of-the-art performance with 99.15% accuracy and 99.92% AUC score on the in-dataset testing set and 88.95% AUC score on the highly difficult unknown-attack Deepfake video dataset. Furthermore, the visualization of the Deepfake forensic noise traces has shown the explicit distinction between synthesized faces and any unmodified area. We believe that the visualized evidence could provide better proof of Deepfake detection results rather than simply the statistical evaluation numbers.

1. Introduction

Have you ever seen the video of a Deepfake synthesized Barack Obama giving a speech insulting Donald Trump in 2018 that is widely spread on YouTube[1]? Without knowing the truth that the face is synthesized using Jordan Peele's, you would possibly get tricked and believe it to be genuine. Deemed to be the most serious artificial intelligence threat in 2020 (ScienceDaily, 2020), Deepfake has become popular and been frequently appeared. Deepfake is firstly introduced by the Reddit user ‘deepfakes’ in 2017, which refers to a facial synthesis technique that completes face-swapping operation and generates hyper-realistic fake videos using deep neural networks (Chawla, 2019; Maras and Alexandrou, 2019). Despite the positive effects of Deepfake that could benefit human lives in various industries such as movies, educational media and digital communications (Westerlund, 2019), huge consequences have gradually appeared in the negative aspects. Potential major victims under Deepfake threats include society, political system and business (Westerlund, 2019), and even anyone can become a target by Deepfake in the future (Melville, 2019; Kietzmann et al., 2020; Tolosana et al., 2020). As the quantity and quality of the online-circulating Deepfake videos become higher, it is hard to manually solve the Deepfake detection task via human eyes. Therefore, preventing high-quality and unseen Deepfake attacks from affecting the human lives is highly desired.

Several Deepfake video datasets have been constructed and released publicly accessible on the Internet within the past few years, and they are categorized into two generations based on their qualities by Tolosana et al. (2020) in a survey. However, due to stark contrasts in visual quality of the existing Deepfake datasets to the actual Deepfake videos circulated on the Internet, prior Deepfake detection algorithms trained on the existing constructed datasets may not perform well against the unknown Deepfake attacks. Li et al. (2020a) released a new high-quality Deepfake dataset with over 5000 videos, namely Celeb-DF, generated using an improved Deepfake synthesis method in 2020. Meanwhile, a set of 518 high-

* Corresponding author.
E-mail addresses: tywang@cs.hku.hk (T. Wang), chow@cs.hku.hk (K.P. Chow).

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

https://doi.org/10.1016/j.fsidi.2022.301395

quality and highly difficult Deepfake videos generated with mysterious techniques was released by Li et al. and has failed many existing state-of-the-art Deepfake detection algorithms. Considering the large quality and difficulty gap between the regular Celeb-DF video dataset and the challenging 518 highly difficult ones, the latter is more proper to play the role of unknown future Deepfake attack, which has been largely used as the benchmark testing set for evaluating Deepfake detection algorithms trained with the existing high-quality datasets.

Various approaches have been attempted for Deepfake detection using deep learning techniques. However, they mostly rely on the computer vision techniques and the magic black box of deep neural networks. To our knowledge, no approach have utilized forensic noise traces for Deepfake detection except the ones utilizing Photo Response Non-Uniformity (PRNU) (Lukas et al., 2006) but failed to achieve satisfied performance. Furthermore, although recent computer vision methods are able to visualize the extracted image feature heatmaps that the Deepfake detection decisions are made based on, the displayed heatmaps appear to be similar and indistinguishable for both real and fake faces. In other words, there is no model up to date that is able to probe the underlying Deepfake traces and perform detection accordingly. Moreover, many methods have ignored the importance of video keyframes for Deepfake detection, leading to huge information loss.

As the hyper-realistic Deepfake synthesized videos become hard to find evidence visually, forensic noise traces are left within the face area whenever there are modifications. Meanwhile, the background area in a Deepfake video is usually unchanged since the goal is face-swapping, and the less the original video is modified, the more authentic it remains. In this paper, we present a novel noise-based Deepfake detection method that mainly focuses on the underlying forensic noise traces of the Deepfake videos. In specific, we utilize the publicly ignored key image frames within the videos and propose a novel face-background strategy that crops the face square and a furthest background square from each video keyframe. We study the different noise trace patterns between fake faces and unchanged real faces while the background squares are always unmodified and authentic. We adopt the Siamese (Bromley et al., 1993) architecture and train the improved DnCNN Denoiser (Zhang et al., 2017) as a noise trace extractor to probe the underlying Deepfake forensic noise traces from video keyframes. We propose a similarity matrix (Huang et al., 2018) to compare and analyze the similarity between noise traces of the face square and a background square from each image frame. A Deepfake manipulated face is expected to have respectively different noise traces from the background square since the background remains unchanged all the time. The ultimate Deepfake detection decision is made based on the similarity matrix value after further refining and projections in the deep neural network. As a result, our proposed approach achieves a frame-level accuracy of 99.15% and an area under the receiver operating characteristic (ROC) curve (AUC) score of 99.92% over the regular Celeb-DF testing dataset and an 88.95% frame-level AUC score over the high-quality and high-difficulty unknown Deepfake attack testing dataset with 518 videos, outperforming the existing state-of-the-art Deepfake detection algorithms in the comparative tests. Furthermore, we have visualized the Deepfake noise traces that have shown strong evidence to support the Deepfake detection results of our proposed model to distinguish fake videos and real videos.

The main contributions of this study include:

- Besides the detection methods using PRNU that have failed, our study is the first time to raise the idea of Deepfake detection in the perspective of forensic noise traces that achieves good detection performance on the highly difficult Deepfake videos.
- Our proposed inductive approach using the novel face-background strategy and similarity matrix achieves the state-of-the-art performance on both normal training and testing datasets and the high-quality and highly difficult Deepfake video dataset that plays the role of an unknown future attack.
- Different from the existing computer vision Deepfake detection approaches that display indistinguishable heatmaps for both real and fake faces, we successfully visualize the Deepfake forensic noise traces to support the satisfactory detection performance.

The rest of the paper is organized as follows. Section 2 discusses the related work of our study, including the Deepfake video generation process and the existing work that has utilized computer vision techniques and forensic noise traces for Deepfake detection. Section 3 introduces the main methodology and workflow of our proposed deep learning model. Section 4 presents experimental results of the proposed approach and the comparative models and analyzes the results accordingly. Finally, section 5 concludes the paper.

2. Related work

We brief the background work of Deepfake in this section. Specifically, we first introduce the two popular Deepfake generation methods that are widely used. Thereafter, we enumerate the existing Deepfake detection work including the unsuccessful attempts using noise traces.

2.1. Deepfake generation methods

The term 'Deepfake' is raised by the Reddit user 'deepfakes' along with the released source code in 2017. The main architecture of Deepfake is an autoencoder (Kingma et al., 2014) with a shared encoder and two individual decoders composed of convolutional neural network backbones. The shared encoder takes charge of learning the common facial features regardless of facial identity, while the two individual decoders each is trained to construct faces of a unique identity. In specific, in order to swap a source facial identity onto a target face, the well-trained autoencoder model takes in the target face as an input and passes the encoder-learned facial features to the unique decoder that is corresponding to the source facial identity. The decoder generates a face with the identity of the source person while maintaining the facial expression and action of the input target face. Another main architecture that is recently frequently adopted to improve face-swapping quality is the generative adversarial network (GAN) (Goodfellow et al., 2014). GAN is composed of a generator and a discriminator that battle with each other to gradually improve the output quality during the training process. Introducing the discriminator to Deepfake generation makes the synthesized faces more authentic by periodically training the generator to fool the discriminator with the generated faces. The synthesized face is inserted back into the original image frame with further smoothing and blur techniques to clean up the obvious Deepfake traces.

2.2. Deepfake detection approaches

The computer vision Deepfake detection approaches mainly focus on the image features. Early approaches mostly utilize the convolutional neural network (CNN) backbone and rely more on its

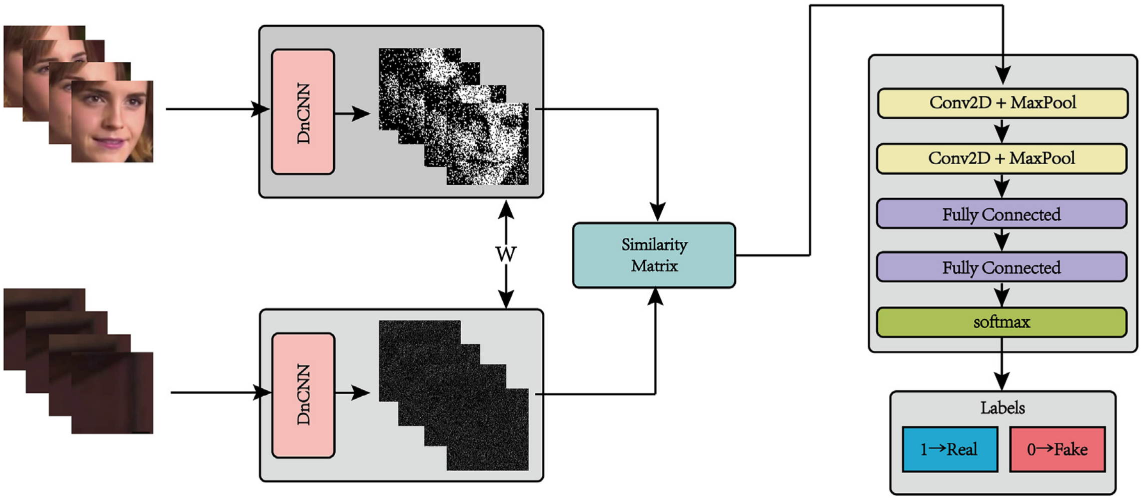

Fig. 1. Workflow of the proposed Deepfake detection model. Cropped keyframe face squares and background squares are passed through the Siamese Network for forensic noise trace extraction, then the two sets of noise traces are computed the similarity matrix. The result is then passed through a stack of convolutional layers with max pooling operations and fully connected layers to further refine the features. In the end, a softmax function is applied to perform Deepfake detection.

image processing ability. Recent computer vision methods tend to include global features for performance enhancement. Two-Stream (Luo et al., 2021) utilizes two streams of Xception (Chollet, 2017) backbones and studies both RGB frames and high-frequency frames, respectively, and further analyzes the cross-modal relations between the two streams. The MAT model (Zhao et al., 2021) adopts EfficientNetB4 as the backbone for Deepfake detection and introduces the idea of attention mechanism to study global features within different local parts of the image frame.

To our knowledge, the existing noise-based Deepfake detection attempts are mainly based on the Photo Response Non-Uniformity (PRNU), a noise pattern created by small factory defects in the light-sensitive sensors of a digital camera (Lukas et al., 2006). PRNU has been frequently adopted in source device identification (Marra et al., 2017; Saito et al., 2017) and source anonymization (Picetti et al., 2020). Unfortunately, no PRNU-based Deepfake detection work has shown strong evidence of the ability of PRNU noise on Deepfake detection. Koopman et al. (2018) evaluated the mean normalised cross correlation score of PRNU noise per video and performed a preliminary experiment to distinguish Deepfake videos from authentic videos with only 10 videos in total given the correct video labels. However, such an attempt is not available without knowing the correct labels in this reported work. Weever and Wilczek (de Weever and Wilczek, 2020) tried several experiments computing the correlation of the PRNU noise and concluded that none of the attempted PRNU noise analyses had led to a definite proof of Deepfake or authenticity. As a conclusion, the PRNU noise pattern has a strong ability for device identification related studies, but it is not a good forensic noise tracing material for Deepfake detection. Therefore, our study is the first to successfully achieve good performance using forensic noise trace based Deepfake detection and visualize the extracted Deepfake forensic noise traces.

3. Methodology

The workflow of our approach as shown in Fig. 1 mainly consists of the following parts: face-background strategy, Siamese noise trace extraction, and noise similarity analysis. The face-background square pairs are firstly extracted from the video key image frames and fed to the Siamese noise extractor. The Siamese noise extractor extracts the Deepfake forensic noise traces from the face and background squares, respectively, and the noise traces are then analyzed via similarity matrix for the Deepfake detection results.

3.1. Face-background strategy

The videos we commonly see are usually under video compression for the purpose of space saving. As a result, three types of image frames are derived, namely, keyframe (I-frame), P-frame, and B-frame. Only the keyframes among all image frames carry complete image information with the largest sizes within a compressed video (Vijayanagar, 2020). To acquire the image frames with high quality and complete information, we extract only the key image frames from the videos using FFmpeg when training our model for optimal Deepfake detection performance.

The major effect of Deepfake is the facial identity swap. Therefore, Deepfake usually only modifies the face area when performing face-swapping, and most of the background area remains unchanged. For each keyframe, we locate the face position using the dlib library[2] and crop the face square and a background square that has the largest Euclidean distance from the face square. This face-background strategy (as shown in Fig. 2) of locating the furthest background area guarantees to crop the background square that is the least likely to be modified by Deepfake even though the background area close to the face may be modified along with the target face. In other words, for each face-background pair, the cropped background square is always unmodified while the face square may be manipulated by Deepfake with noise traces left behind.

3.2. Siamese noise trace extraction

Since being firstly introduced in 1993 for signature verification, the Siamese Network (Bromley et al., 1993) has been frequently utilized for feature comparison (Chopra et al., 2005). We adopt the Siamese design for noise trace extraction from the face squares and background squares where the two branches share the same weights. In particular, for a face-background pair, the face square and the background square are each passed through one branch of the Siamese architecture. In each Siamese branch, a pre-trained

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

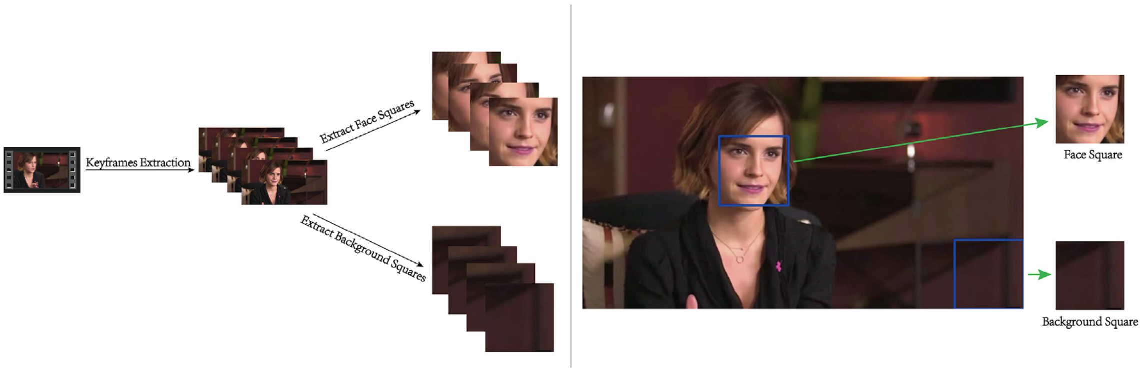

Fig. 2. Keyframes are extracted from the Deepfake video, then the face and background squares are cropped from each keyframe. The face square in a keyframe is detected and cropped using the dlib library, and a background square with the same size and having the furthest Euclidean distance from the detected face within the image frame is found and cropped.

DnCNN denoiser is adopted and improved for Deepfake forensic noise trace extraction.

Classical denoisers share the same drawbacks that they mostly involve a convex optimization problem that is time consuming and cannot handle images with unknown noise levels. Besides, most existing denoisers perform denoising directly without finding out the noise. Zhang et al. (2017) proposed a DnCNN denoiser that is composed of mainly multiple layers of convolutional neural networks and handles denoising over unknown noise levels by extracting the underlying noise. The DnCNN denoiser structure, as shown in Fig. 3, was further utilized by Cozzolino and Verdoliva (2018) for the camera model fingerprint study that can perform source device classification based on the extracted noise. In this study, the face and background squares are passed through the improved DnCNN model to extract the underlying noise traces instead of eliminating them. We exploit the pre-trained weights of the DnCNN that can firstly extract a general level noise. Then, we train it along with further similarity matrix and classification module to enforce additional restrictions for weight updating and achieve Deepfake forensic noise trace extraction that serves for our Deepfake detection purpose.

For a real image frame that is unmodified, it contains the same noise pattern everywhere within the image, in other words, no Deepfake forensic noise trace. On the other hand, a Deepfake image frame has the face area synthesized such that the underlying noise pattern of the face area is different from that of the unmodified background area. Since the two branches of the Siamese architecture share the weights on noise trace extraction, different noise patterns are extracted under the same noise trace extraction process from the face and background squares of the Deepfake synthesized videos, while the same noise patterns can be found within the face and background squares of the real videos. The Siamese architecture is coded as one single network in implementation since both branches share the same set of network weights. The extracted forensic noise traces from the face and background squares are further fed to a similarity matrix design for noise trace pattern comparison and Deepfake detection decision making.

3.3. Noise similarity analysis

Considering the forensic noise traces for the unmodified area are always clean, the Deepfake manipulated faces are expected to have complicated noise traces. We propose the similarity matrix idea on the extracted face and background noise traces for the noise similarity analysis. In specific, the similarity matrix is implemented as the inner product of the two noise representations to find out the correspondence between every two vector entries. Following the convention, multiplication is more powerful to find relations between deep neural network feature matrices than summations. When performing matrix summation, only entries in the same position can be summed up and thus the correspondence is weaker than conducting a product operation. The similarity between column *i* of the face noise trace matrix and column *j* of the background

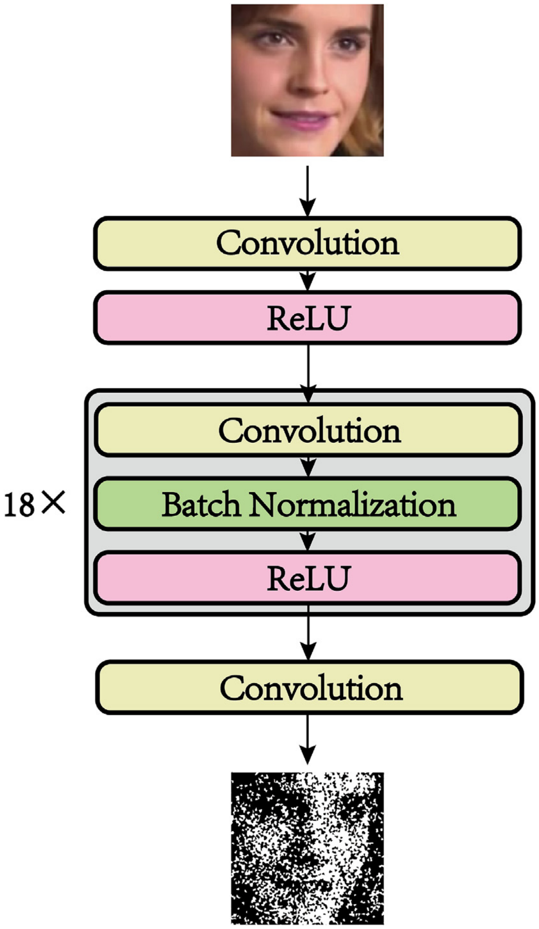

Fig. 3. The architecture of the improved DnCNN denoiser for Deepfake forensic noise trace extraction. The input face or background is passed through a combination of a convolutional layer and a ReLU activation layer, followed by 18 repeated blocks of convolution, batch normalization, and ReLU activation layers, and a convolutional layer in the end to generate the output noise.

noise trace matrix, namely entry **S***i*,*j*, follows

$$\mathbf{S}_{i,j} = \frac{\mathbf{F}_i \times \mathbf{B}_j^{\mathrm{T}}}{\|\mathbf{F}_i\|_2^2 \times \|\mathbf{B}_j\|_2^2}, \tag{1}$$

where **F** and **B** are the extracted noise traces for the face and background squares, respectively.

We make the final Deepfake detection decision based on the similarity matrix result. Therefore, we apply 2D convolutions with fully connected layers to the proposed model to refine the dominant similarity entries, and compute the probability that the input face is Deepfake synthesized and tune the model by

$$L_{CE} = -\Sigma_{i=1}^{2} t_i \log p_i, \tag{2}$$

where t_i is the ground truth value and p_i is the Softmax prediction for class i upon the final output from the last fully connected layer.

4. Experiments

In this section, we first introduce the dataset utilized in the experiment. Then, we describe the experiment settings. Thereafter, we list and briefly introduce the state-of-the-art models employed for comparative tests. Thenceforth, we evaluate the models on the testing datasets following the identical experiment settings and analyze the experiment results. In the end, we visualize the extracted Deepfake forensic noise traces using our well-trained model.

4.1. Dataset

The purpose of the Deepfake detection task is to prevent future unknown Deepfake attacks from affecting human lives with the help of the existing Deepfake video datasets. As the actual Deepfake videos circulated on the Internet have become more authentic and harder to distinguish, it is significant to adopt a Deepfake dataset with good quality for deep detection models training. Tolosana et al. (2020) categorized the existing Deepfake video datasets into two generations with respect to the qualities. Typical datasets in the first one are UADFV (Li and Lyu, 2019), Deepfake TIMIT (Korshunov and Marcel, 2018), and FaceForensics++ (FF++) (R ö ssler et al., 2019) datasets. By contrast, DeepFakeDetection (DFD) (Dufour and Gully, 2019), Celeb-DF (Li et al., 2020a), and Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2019) datasets are in the second generation. To be more specific, the Celeb-DF dataset, constructed by an improved Deepfake synthesis method, is of the highest quality while the other ones all contain a substantial amount of obvious fake videos that can be easily discovered by human eyes. We thus selected the high-quality Celeb-DF dataset to train our proposed model in this study.

The Celeb-DF dataset contains 712 real videos and 5299 fake ones that have good qualities and a certain degree of difficulty to distinguish. Besides, there is a set of 518 high-quality and highly difficult videos (178 real and 340 fake) that contains mysterious synthesis tricks and has challenged all state-of-the-art Deepfake detection models with suboptimal performance. Thereby, it can be regarded as a potential future threat for cross-dataset detection model testing. In order to overcome the possible unknown Deepfake threats in the future, we evaluated our proposed noise-based Deepfake detection model over the Celeb-DF normal dataset after sampling balanced datasets for both training and testing, and further tested the performance over the unbalanced set of 518 high-quality and highly difficult unknown attack Deepfake videos.

4.2. Experiment settings

The proposed model is trained on a balanced preprocessed keyframe dataset as described in Section 4.1 for real and fake, at a ratio of 8:1:1 for training, validation, and testing set. In detail, a total of 45,820 face-background pairs are utilized in the training process. The lengths for the input face and background squares are resized to 64 for consistency in the training process. When refining the similarity matrix, each 2D convolution is set with kernel size of 3 and stride of 1 and each max pooling operation is set with kernel size of 2 and stride of 2. Each combination of convolutional layer and max pooling operation decreases the feature dimension by half. The fully connected layers each decreases the feature dimension fourfold so that the softmax function can be properly applied. While achieving a considerably high performance over the normal testing dataset, we further validated our well-trained Deepfake detection model on the unknown future attack dataset of 518 videos that has high quality and high difficulty. We evaluated the overall Deepfake detection performance using accuracy and AUC score at frame level for the balanced testing datasets. Depending on the testing set distribution, a high value of the accuracy can cover up potential problems within the model. The AUC score is calculated by the area under the ROC curve by adjusting all threshold values throughout the whole range from 0 to 1, plotting over the true positive rate and false positive rate values, which could further prove the robustness of our model. Therefore, we used only the AUC scores for model evaluation upon the unbalanced unknown attack challenging testing set.

4.3. Comparative models

We considered the existing state-of-the-art Deepfake detection algorithms for comparisons to our exhibited approach. We chose the Deepfake detection algorithms that have source code published and reproducible for training and testing with the same experiment settings as ours while maintaining their optimal parameter settings whenever applicable. For the rest selected detection algorithms that either have no source code published or the source code are unreproducible in experiments, we directly employed the published model checkpoints or the reported evaluation results if applicable. A summary of the methods information considered for comparative tests is as listed in Table 1, including the model name, released date, source code availability, and how the model is evaluated (trained, checkpoint evaluation, or reported result adoption) in comparative tests for each selected method. The comparative test models are briefly introduced in the following paragraphs.

MesoNet (Afchar et al., 2018) focuses on the mesoscopic properties of image frames using CNN based architectures. The proposed model is trained and tested on 175 Internet-collected rushes of Deepfake videos at frame-level using keyframes. We trained and evaluated the MesoNet model on our dataset, which is based on complicated Inception modules (Szegedy et al., 2015) and claims to achieve the best performance on Deepfake videos as reported.

Capsule (Nguyen et al., 2019) model adopts capsule structures (Sabour et al., 2017) and utilizes much fewer VGG19 (Simonyan et al., 2015) based network parameters than traditional CNNs with similar performance. The model is trained on the FF++ dataset (R ö ssler et al., 2019) in the published work. An updated version with better performance is published and we trained and tested the updated Capsule model on our dataset for a fair comparison.

DSP-FWA (Yang et al., 2019), an improved method based on the spatial pyramid pooling module (He et al., 2015) with CNN based ResNet (——, 2016) as the backbone, can handle Deepfake videos

Table 1
Summary of the selected Deepfake detection approaches for comparison tests. Information includes comparative model name, released date, source code availability, and how the model is evaluated in comparative tests. In general, models with source code are trained and tested on our dataset, while models with no source code are either tested on the given checkpoints or adopted the reported performance.

Models for Comparison	Released Date	Source Code Availability	Evaluation Method
MesoNet (Afchar et al., 2018)	Sept. 2018	Published source code available.	Trained and tested on our dataset.
Capsule (Nguyen et al., 2019)	Oct. 2019	Published source code available.	Trained and tested on our dataset.
DSP-FWA (Yang et al., 2019)	Nov. 2019	No published source code but has checkpoint released.	Adopt the experiment result reported on the Celeb-DF test dataset.
Ensemble (Bonettini et al., 2021)	Apr. 2020	Published source code available.	Train and test on our dataset.
DFT-MF (Jafar et al., 2020)	Apr. 2020	No published source code and no checkpoint released.	Adopt the experiment result reported on the Celeb-DF test dataset.
FFD (Dang et al., 2020)	June 2020	Published source code available.	Trained and tested on our dataset.
Face X-ray (Li et al., 2020b)	June 2020	No published source code and no checkpoint released.	Adopt the experiment result reported on the Celeb-DF test dataset.
Multi-Attention (Zhao et al., 2021)	Mar. 2021	No published source code and no checkpoint released.	Adopt the experiment result reported on the Celeb-DF test dataset.
Two-Stream (Luo et al., 2021)	Mar. 2021	Published source code available.	Trained and tested on our dataset.
TAR (Lee et al., 2021)	May 2021	Published source code available.	Trained and tested on our dataset.

with different resolution qualities as reported. The model is trained with self-generated Deepfake dataset and only a checkpoint for the pre-trained model weights is released. This approach is evaluated by Li et al. on the high-quality and highly difficult unknown attack Deepfake testing dataset, and we adopted the reported experiment performance result for the comparative test.

Ensemble (Bonettini et al., 2021) stands for the ensemble of different pre-trained CNN based models. The work takes CNN based EfficientNetB4 (Tan et al., 2019) as the backbone and proposes the attention mechanism for performance improvements. The model utilizes triplet Siamese Network architecture and takes groups of 3 inputs for training, namely, an anchor, a positive sample with the same label as the anchor, and a negative sample with the opposite label to the anchor. The source code for the Ensemble model is publicly available, and we trained and tested its performance on our dataset for performance comparison.

DFT-MF (Jafar et al., 2020) published in 2020 focuses specifically on the mouth features and uses CNN based model to detect by isolating, analyzing, and verifying lip and mouth movements. The authenticity classification is based on a number of fake frames in a video with respect to words per sentence, speech rate, and frame rate. The model is trained and evaluated on Celeb-DF and Deepfake-TIMIT datasets separately. No source code and checkpoint for DFT-MF is published online, and we directly adopted its reported results as it is trained and tested on the Celeb-DF dataset.

FFD (Dang et al., 2020) refers to a CNN based Deepfake detection network that utilizes the attention mechanism to process the feature maps and adopts the previous state-of-the-art Xception (R ö ssler et al., 2019) architecture as the backbone, claiming to outperform the backbone. Xception was introduced for Deepfake detection and has achieved considerable performance when the FF++ dataset was released in 2019. It refers to a detection method based on the XceptionNet model (Chollet, 2017). We trained and tested the state-of-the-art FFD model on the same dataset as ours for performance comparison while maintaining its default parameter settings for optimal performance.

Face X-ray (Li et al., 2020b) simulates a medical x-ray examination by revealing whether the input image can be decomposed into the blending of two images from different sources. The blending boundary appears for a forged image while no blending is detected for a real image. The approach is trained on FF++. No source code or pre-trained model weight checkpoint is published up to the time when this paper is written, and we directly employed the performance results on the same testing dataset reported by the authors for comparison with our approach.

Multi-Attention (Zhao et al., 2021) uses the CNN based EfficientNetB4 network as the backbone and applies multiple head attention to analyze and focus on different local parts of the input image frame and zoom in the artifacts in shallow features for an enhancement in Deepfake detection performance. The model is trained on FF++ and no reproducible source code or pre-trained weight is released up to date on the Internet. We directly utilized the experiment performance result on the same testing dataset for comparison with our approach.

The Two-Stream (Luo et al., 2021) approach constructs two streams of Xception backbones and mainly analyzes the RGB frames and high-frequency frames, respectively. The cross-modal relations between the two streams are studied and utilized for Deepfake detection task. The model is trained on FF++ and source code is publicly available. Thus, we trained and tested the Two-Stream model on our utilized dataset with identical experiment settings.

TAR (Lee et al., 2021) is one of the latest Deepfake detection model reporting significantly higher performance over the state-of-the-art methods. The model applies transfer learning based on autoencoders with residual blocks. A Facilitator module is utilized to force and divide the latent space between the real and fake embeddings. The model is trained on the FF++ dataset and self-collected datasets. Source code of TAR is publicly available on the Internet, so we trained and tested the TAR model on our dataset while maintaining its default optimal training setting parameters.

4.4. Evaluation results

We first trained our proposed noise-based Deepfake detection model with our balanced training dataset and evaluated the performance on the testing dataset. Then, we performed Deepfake detection using the well-trained model on the high-quality and highly difficult unknown attack dataset with 518 videos. As a result, our approach achieves high frame-level accuracy of 99.15% and AUC score of 99.92% on the normal testing set. Models with source codes published for comparative tests are trained and tested on the same dataset as ours while maintaining their original optimal experiment settings when possible, and the evaluation results on the normal testing dataset are as shown in Table 2. As a result, our Deepfake detection approach outperforms all adopted open-source state-of-the-art models after training and testing on the same dataset.

Besides the normal testing dataset, the high-quality and highly difficult set of 518 unknown attack videos have failed many well-known Deepfake detection algorithms (Li et al., 2020a). We

Table 2
Frame-level comparative tests accuracy (%) and AUC scores (%) on the normal testing dataset.

Model Names	Accuracy	AUC Score (%)
MesoNet (Afchar et al., 2018)	92.36	97.82
Capsule (Nguyen et al., 2019)	99.02	99.83
FFD (Dang et al., 2020)	98.69	99.92
Ensemble (Bonettini et al., 2021)	70.77	78.15
Two-Stream (Luo et al., 2021)	92.27	98.19
TAR (Lee et al., 2021)	50.00	50.00
Our Approach	**99.15**	**99.92**

Table 3
Frame-level comparative tests AUC scores (%) on the unknown attack challenging Deepfake dataset.

Model Name	AUC Score (%)
MesoNet (Afchar et al., 2018)	70.57
Capsule (Nguyen et al., 2019)	77.55
DSP-FWA (Yang et al., 2019)	64.60
Ensemble (Bonettini et al., 2021)	80.83
DFT-MF (Jafar et al., 2020)	71.25
FFD (Dang et al., 2020)	81.05
Face X-ray (Li et al., 2020b)	80.58
Multi-Attention (Zhao et al., 2021)	67.44
Two-Stream (Luo et al., 2021)	84.19
TAR (Lee et al., 2021)	50.00
Our Approach	**88.95**

further evaluated our well-trained model and it has achieved a state-of-the-art frame-level AUC score of 88.95% on the unknown attack Deepfake dataset with 518 videos. The comparative models trained on the same dataset as listed in Table 1 are also evaluated on the 518 videos for further comparison with our approach. For the rest methods as listed in Table 1 without source code published, we either directly utilized the reported performance on the 518 videos or evaluated the performance using the provided pre-trained checkpoints if applicable. As Table 3 shown, the proposed noise-based Deepfake detection method outperforms all state-of-the-art models on the challenging unknown attack Deepfake dataset with 518 videos. A more vivid look of the AUC score comparisons among all Deepfake detection models on the high-quality and highly difficult unknown attack testing dataset is shown in Fig. 4. The ROC curves of our model when tested on the normal testing dataset and on the unknown attack challenging dataset are shown in Fig. 5. Both curves have shown reasonable behaviors, which further prove the robustness of our proposed model.

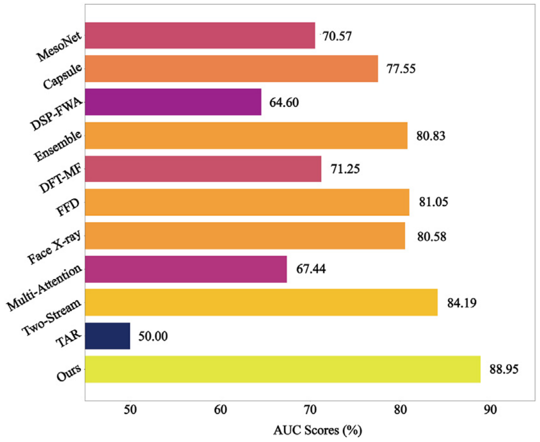

Fig. 4. Bar plot AUC score (%) performance comparisons of the comparative models. The proposed model achieves the state-of-the-art performance over all other models.

As listed in Tables 2 and 3, most models have performed well on the normal testing dataset after training on the corresponding training dataset except the Ensemble method, while our approach still achieves a slightly better performance than all of them. In addition, Ensemble (Bonettini et al., 2021), FFD (Dang et al., 2020), Face X-ray (Li et al., 2020b), Two-Stream (Luo et al., 2021) and our proposed model have achieved the AUC scores over 80% on the challenging testing set, while ours reaches the highest over 85%. The fact that the Ensemble method doesn't perform well on the normal testing dataset might be because of an overfitting by the model on the detection of fake videos while neglecting the work on the real videos, and therefore fails on a balanced normal testing dataset but still achieves good performance on the unbalanced challenging testing dataset. The TAR (Lee et al., 2021) is proved to be very time-consuming at a day-level training process while all other models' training processes are at hours level. Moreover, the TAR model after training with default settings labels all videos to be authentic, which causes the bad performance as shown in Tables 2 and 3. As a matter of fact, since all Deepfake faces contain the underlying forensic noise traces regardless of different manipulation techniques, our proposed approach is more generalizable on the unknown future Deepfake attacks while the existing methods are somehow under higher potential risks of overfitting on the training dataset.

4.5. *Forensic noise trace visualization*

Although some state-of-the-art Deepfake detection models have exhibited reasonably good performance upon the testing datasets. To our knowledge, none of them have shown the ability to extract and visualize the Deepfake forensic traces as a straightforward evidence. In specific, recent work is able to display the feature heatmap of which the Deepfake detection decision depends on. However, the heatmaps are for the purpose of feature localization, and they tend to be always similar and indistinguishable for both real and fake faces.

In this study, we trained a Deepfake noise trace extractor and truly extracted the underlying traces within our proposed model. Besides outperforming the state-of-the-art approaches statistically, we further visualized the extracted Deepfake forensic noise traces from a sample testing dataset of face-background pairs. We froze the weights of the noise trace extractor within the Siamese architecture and displayed the extracted Deepfake forensic noise traces of each face background pairs in Fig. 6.

As Fig. 6 shown, the top two rows display the extracted Deepfake forensic noise traces of real faces and backgrounds, while the bottom two rows exhibit that of the fake ones. Every two columns contain a pair of cropped face and background squares along with the corresponding noise traces on the row below them. The fake face squares demonstrate obvious Deepfake forensic noise traces with complicated traces displayed while the real ones have nearly no forensic noise traces on the contrary. The background squares, as expected, have shown clean figures with no Deepfake noise trace. In conclusion, the more complicated the noise traces of a face is displayed, the more likely the face is from a Deepfake video. With the help of the visualized noise traces, it is convincing to point out the particular positions on the faces that are synthesized by Deepfake, which further supports the satisfactory statistical Deepfake detection performance of our proposed model.

5. Conclusion

Deepfake has drawn considerable public attention as its potential security risk is gradually recognized by the society. In this

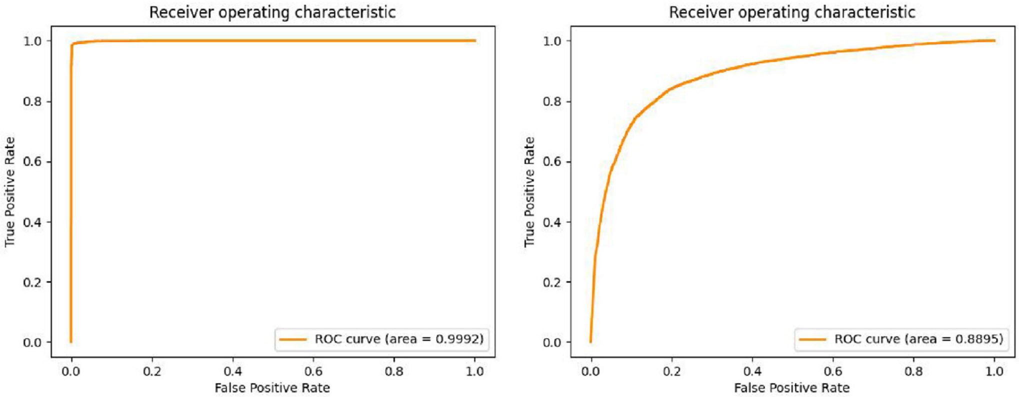

Fig. 5. ROC curves of the proposed model when tested on the normal testing dataset (left) and on the unknown attack challenging dataset (right).

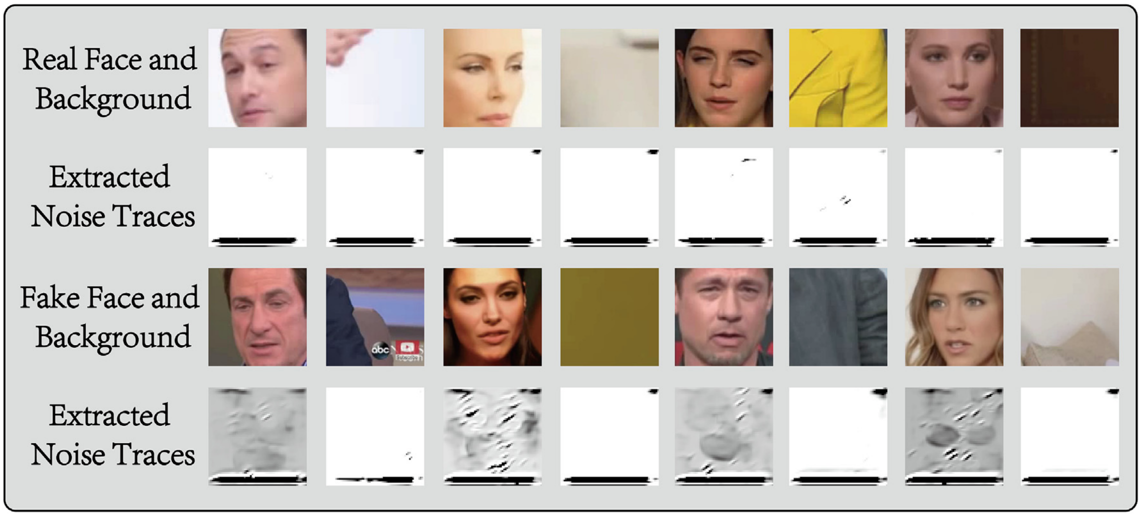

Fig. 6. Visualization of the extracted Deepfake forensic noise traces. The two rows on the top display the extracted Deepfake forensic noise traces of real faces and backgrounds, and the bottom two rows display the extracted Deepfake forensic noise traces of fake faces and backgrounds. Every two columns are a pair of cropped face and background squares along with the corresponding noise traces on the row below them. The more colorful and more complicated the noise traces of a face is displayed, the more likely the face is from a Deepfake video.

study, we present a state-of-the-art noise-based Deepfake detection model that investigates the underlying forensic noise traces of Deepfake. The visualized noise traces have further shown promising evidence for the robustness of the proposed approach. Future work will focus on improving the quality of the extracted forensic noise traces and the Deepfake detection generalization performance on diverse testing sets. Besides, due to computation limit, input images are resized to 64 in this study. We plan to explore the influence of different input image sizes in future work. We will also work on defending our model against potential attacks with perturbations and distortions since our approach is noise-based. Lastly, we wish to elaborate the noise-based model to video-level Deepfake detection in the future.

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., Dec 2018. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). https://doi.org/10.1109/WIFS.2018.8630761 [Online]. Available:

Bonettini, N., Cannas, E.D., Mandelli, S., Bondi, L., Bestagini, P., Tubaro, S., 2021. Video face manipulation detection through ensemble of cnns. In: 2020 25th International Conference on Pattern Recognition. ICPR), pp. 5012–5019.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993. Signature verification using a "siamese" time delay neural network. In: *Proceedings Of the 6th International Conference On Neural Information Processing Systems*, Ser. NIPS'93. Plus 0.5em Minus 0. Morgan Kaufmann Publishers Inc., 4emSan Francisco, CA, USA, pp. 737–744.

Chawla, R., 2019. Deepfakes: how a pervert shook the world. Int. J. Adv. Res. Develp. 4, 4–8.

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, pp. 539–546, 1.

Cozzolino, D., Verdoliva, L., 2018. Noiseprint: a cnn-based camera model fingerprint. abs/1808.08396, CoRR [Online]. Available: http://arxiv.org/abs/1808.08396.

Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K., June 2020. On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

de Weever, C., Wilczek, S., 2020. Deepfake Detection through Prnu and Logistic Regression Analyses.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C., 2019. The Deepfake Detection Challenge (Dfdc) Preview Dataset.

Dufour, N., Gully, A., 2019. Contributing Data to Deepfake Detection Research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. accessed: 2021-05-01.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: *Proceedings Of the 27th International Conference On Neural Information Processing Systems - Volume 2*, Ser. NIPS'14. Plus 0.5em Minus 0. MIT Press, 4emCambridge, MA, USA, pp. 2672–2680.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.

Huang, Y.-S., Chou, S.-Y., Yang, Y.-H., 2018. Generating music medleys via playing music puzzle games. In: *Proceedings Of the Thirty-Second AAAI Conference On Artificial Intelligence And Thirtieth Innovative Applications Of Artificial Intelligence Conference And Eighth AAAI Symposium On Educational Advances In Artificial Intelligence*, Ser. AAAI'18/IAAI'18/EAAI'18. Plus 0.5em Minus 0. 4emAAAI Press.

Jafar, M.T., Ababneh, M., Al-Zoube, M., Elhassan, A., 2020. Forensics and analysis of deepfake videos. In: 2020 11th International Conference on Information and Communication Systems. ICICS), pp. 53–58.

Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C., 2020. Deepfakes: trick or treat? aRTIFICIAL intelligence and machine learning. Bus. Horiz. 63 (2), 135–146 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681319301600.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings [Online]. Available: http://arxiv.org/abs/1312.6114.

Koopman, M., Macarulla Rodriguez, A., Geradts, Z., 2018. Detection of deepfake video manipulation. In: Proceedings of the 20th Irish Machine Vision and Image Processing conference, 8, pp. 133–136.

Korshunov, P., Marcel, S., 2018. Deepfakes: a New Threat to Face Recognition? Assessment and Detection.

Lee, S., Tariq, S., Kim, J., Woo, S.S., 2021. Tar: Generalized Forensic Framework to Detect Deepfakes Using Weakly Supervised Learning.

Li, Y., Lyu, S., 2019. Exposing deepfake videos by detecting face warping artifacts. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW).

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020a. Celeb-df: a large-scale challenging dataset for deepfake forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B., 2020b. Face x-ray for more general face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR), pp. 5000–5009.

Lukas, J., Fridrich, J., Goljan, M., 2006. Digital camera identification from sensor pattern noise. IEEE Trans. Inf. Forensics Secur. 1 (2), 205–214.

Luo, Y., Zhang, Y., Yan, J., Liu, W., June 2021. Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16, pp. 317–416 326.

Maras, M.-H., Alexandrou, A., 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. Int. J. Evid. Proof 23 (3), 255–262. https://doi.org/10.1177/1365712718807226 [Online]. Available:

Marra, F., Poggi, G., Sansone, C., Verdoliva, L., 2017. Blind prnu-based image clustering for source identification. IEEE Trans. Inf. Forensics Secur. 12 (9), 2197–2211.

Melville, K., 2019. The Insidious Rise of Deepfake Porn Videos and One Woman Who Won't Be Silenced. https://www.abc.net.au/news/2019-08-30/11437774. accessed: 2021-05-01.

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

Picetti, F., Mandelli, S., Bestagini, P., Lipari, V., Tubaro, S., 2020. Dippas: A Deep Image Prior Prnu Anonymization Scheme.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M., 2019. Faceforensics++: learning to detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. In: *Proceedings Of the 31st International Conference On Neural Information Processing Systems*, Ser. NIPS'17. Plus 0.5em Minus 0, 4emRed Hook. Curran Associates Inc., NY, USA, pp. 3859–3869.

Saito, S., Tomioka, Y., Kitazawa, H., 2017. A theoretical framework for estimating false acceptance rate of prnu-based camera identification. IEEE Trans. Inf. Forensics Secur. 12 (9), 2026–2035.

ScienceDaily, 2020. 'deepfakes' ranked as most serious ai crime threat. https://www.sciencedaily.com/releases/2020/08/200804085908.htm accessed: 2021-05-01.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. May 7-9, 2015, Conference Track Proceedings. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations. ICLR 2015, San Diego, CA, USA [Online]. Available: http://arxiv.org/abs/1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., June 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tan, M., Le, Q.V., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), *Proceedings Of the 36th International Conference On Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Ser. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114 plus 0.5em minus 0.4emPMLR. http://proceedings.mlr.press/v97/tan19a.html.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J., 2020. Deepfakes and beyond: a survey of face manipulation and fake detection. Inf. Fusion 64, 131–148 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303110.

Vijayanagar, K.R., 2020. I, p, and b-frames – differences and use cases made easy. https://bit.ly/34OArtI accessed: 2021-05-01.

Westerlund, M., 11/2019 2019. The emergence of deepfake technology: a review. Tech. Innovation. Manage. Rev. 9, 40–53 [Online]. Available: timreview.ca/article/1282.

Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP), pp. 8261–8265.

Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2017. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. 26 (7), 3142–3155.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N., 2021. Multi-attentional Deepfake Detection.

——, 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.