Contribution of residual signals to the detection of face swapping in deepfake videos

Paul Tessé[1,†], Emmanuel Giguet[1,†] and Christophe Charrier[1,*,†]

[1]*Université de Caen Normandie, ENSICAEN, CNRS, GREYC, Normandie Univ, F-14000 Caen, France*

Abstract

The remarkable ascent of Deep Learning, particularly with the emergence of generative adversarial networks, has transformed the landscape of Deepfake technology. These fabrications are evolving into remarkably realistic renditions, posing greater challenges for detection. Verifying the authenticity of video content has become increasingly delicate. Moreover, the widespread availability of forgery tools is a growing concern. Despite numerous proposed detection techniques, determining the efficacy of these methods despite rapid advancements is challenging. Hence, this paper introduces an approach to detect face swapping in videos using residual signal analysis.

Keywords

Deepfake videos, Face swapping, Residual error, Digital forensics, Deep learning

1. Introduction

In today's hyper-connected society, billions of units of data traverse our networks daily. Regrettably, there's a growing uncertainty regarding the reliability and safety of this data. This concern is particularly pronounced when considering the vast dissemination and sharing of videos and images, which account for a substantial portion of this data flow. With around 5.35 billion internet users worldwide, each person can potentially generate approximately 15.87 TB of data daily. From those data, approximately 28.08 billion photos were stored online daily in 2023. Moreover, accessing video content has become exceedingly convenient with the prevalence of mobile devices, streaming services, the internet, and social media platforms. For instance, videos constituted 28% of all internet traffic in 2022, and 625 million videos are viewed on TikTok every internet minute, that's up from 167 million just two years ago [2].

The rapid growth in Deep Learning schemes has led to the emergence of numerous efficient models for generating false images or videos. As a matter of fact, while these models are becoming increasingly powerful, they are also becoming more and more accessible to the public thanks to the internet and social media. We are observing a notable surge in the proliferation of counterfeit multimedia content, notably hyper-realistic videos, commonly referred to as "deepfakes". For instance, AI-powered software applications such as FaceApp [3] and FakeApp [4] have been employed for the creation of convincingly realistic face swaps in both images and videos. This capability enables users to modify facial appearances, hairstyles, genders, ages, and other personal features. The dissemination of such manipulated videos has raised significant concerns and has gained notoriety as Deepfake technology. The main threats are the increasing difficulty to distinguish real from fake, even for informed people, and the inability to trust images and videos even though they are ubiquitous media. Detecting these manipulated videos has become a crucial societal challenge.

As a consequence, many researchers have studied deepfake detection, proposing methods mainly based on the use of Deep Learning [5]. However, although these models presented as state-of-the-art

models show very good performance, the durability of these models and their ability to generalize to any database are lacking. Nevertheless, the main drawback of these models lies in the fact that they are usually designed and used as black boxes. In fact, it is not really possible to provide a justification for the pronounced verdicts, which makes these detectors unusable.

In this paper, an explainable model that takes a video as input and returns a verdict on its authenticity as output is proposed. The problem is therefore characterized as a binary classification problem where the classes are defined as "authentic" and "forged". The secondary objectives considered in this paper are the following:

- the system must be able to be used to support the judicial system, and a particular attention must be paid to the explainability of the results;
- the model must work without any reference to pronounce its diagnosis;
- the model must be as robust and generalizable as much as possible;
- the approach focuses on face swapping detection, and a face detection mechanism is required in order to target the area to be studied;
- the model must work regardless of the length of the video and regardless of the position or the moment where the fake part appears.

Eventually, the aim is to achieve the best possible trade-off between efficiency and explainability. The aim is not to come up with a perfect solution, but rather a proof of concept to determine whether or not the proposed approach is viable.

2. State-of-the-art methods

Although deepfake technology has the potential for beneficial applications, such as in film-making and virtual reality, its predominant usage remains for malicious intents [6, 7]. Out of all manipulated videos, our focus will be on face swapping. Various method have been proposed to detected forged video. Detection based on general Convolutional Neural Network (CNN) is commonly used in literature, where deepfake detection is considered as a classification task. In those method, face images are extracted from the suspicious video and are used to train the CNN. The learned model is applied to predict whether the video is real or a fake.

In [11] Guëra *et al* present a deep architecture split into a feature extractor part and a classifier part. The feature extractor uses convolutional networks to extract spatial features from images, while a LSTM is used to extract temporal features. Regarding the classifier, the authors use a likelihood estimation in order to predict the probability of a video to be real or fake. The authors argue very good performance with this method, 97.1% of detection. However, these results remain poorly explainable. As a matter of fact, it is difficult to give a sense to the verdict since it is based on features that are not easily explainable. Whatever the proposed scheme based on this strategy, the detection accuracy heavily relies on both the used neural network and the train dataset, without considering anymore the need to exploit specific distinguishable features.

Recently, many methods seek to analyze what are known as residual signals. These are characteristics intrinsic to an image that are generated during the acquisition process and that can be altered during the face swapping process. Some methods based on mid-level manipulation traces are devoted to finding inconsistencies in fake content. In [1], Gao *et al.* have explored the inconsistency of identity information between inner and outer faces. In [8] Wu *et al.* regarded DeepFake detection as a source detection task and utilized the multi-scale spatiotemporal photoplethysmography map from multiple facial regions to capture counterfeit cues. According to [9] emphet al., fake faces typically involve a fusion step. Based on this assumption, they proposed Face X-ray to detect the presence of face fusion boundaries in images and use it as a metric to measure the authenticity.

Despite the above papers reporting good generalization performance, the extracted cues only provide limited global information which, combined with complex learning models, increases computational complexity and training difficulty. To address these challenges more effectively, this paper introduces a
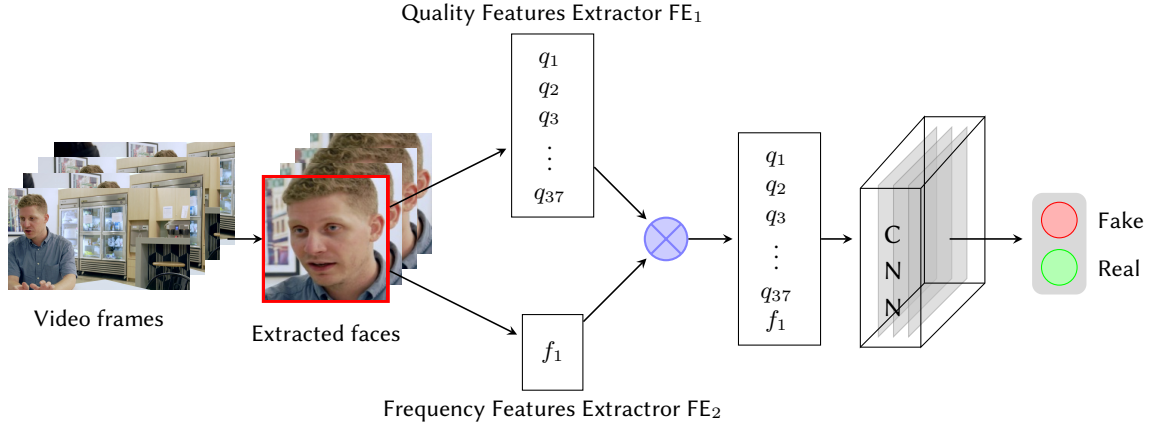
Figure 1: Synposis of the proposed architecture.

straightforward deepfake detection method where used features can be explained, combining thus both explainability and performance

Considering video as a succession of images, we investigate how these residual signals can be analyzed frame by frame in both spatial and frequency domain. Actually, these signals, which are invisible to the naked eye and look like a hidden signature, are varied and can be explained, enabling us to extract features that can be explained and thus, used for prediction.

3. The proposed approach

To address both explainability and performance, we propose a light architecture in which the features used can all be explained.

Based on a classical two-part architecture (extractor and classifier), the proposed architecture is depicted in Fig. 1. First, the input video is split into frames from which the faces are extracted. The obtained faces are then analyzed by a set of designed Features Extractors (FE), which generate a vector of explainable features. Finally, these explainable features are concatenated and passed on to the deep classifier, which gives its verdict at frame level.
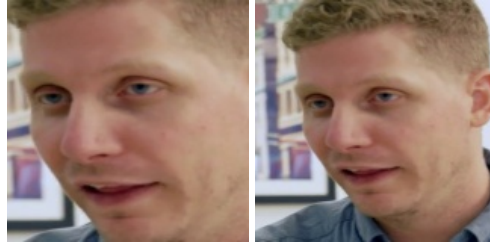
3.1. Face extraction

For each video, a face detector is used to locate the face area in the image. Among all solutions, we used the Multi-task Cascaded Convolutional Neural Networks (MTCNN) introduced by Zhang *et al.* [10].

Face swapping generates non visible artifacts around the area where the face has been swapped. Thus extracting only faces does not guarantee we are able to capture such hidden artifacts, since it focusing only on faces whereas artifacts arise in a larger area around the face. A margin is thus applied when faces are extracted.

3.2. Features FE_i

Face swapping entails the process of seamlessly replacing a face from a source image with another face in a target image, achieving a natural blend between the replacement and the target image. Since both images (source and target image) came from two distinct video, the quality of them is not necessarily the same. Assessing the quality of the face with regards to its neighborhood can provide relevant information. In addition, replacing face into an image may introduce artificial high frequencies into its neighborhood. Capturing such a variation may help to detect face swappping. In this paper, we investigate those alluded above residual signals.

(a) Face extracted without margin (b) Face extracted with margin

Figure 2: Illustration of face extraction with or without margin around the face using MTCNN

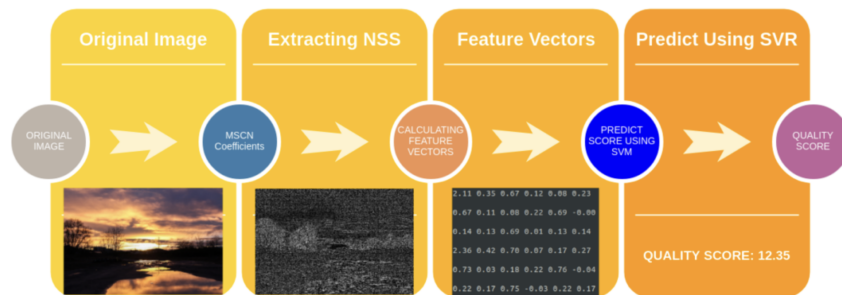### 3.2.1. Image Quality Assessment Features

Figure 3: Quality features extraction process [26]

The first residual signal we focus on is the image quality. Actually, it can be observed that image falsification processes tend to reduce the quality of the original images introducing artifacts. In [13], a detection method based on image quality analysis was presented. This work was then taken up more recently in [14]. The authors greatly increase the number of quality measurements carried out and test their regression model on state-of-the-art databases.

Since we only have access to the video and process it frame by frame to decide if it is a forged or real video, we consider No-Reference Image Quality Assessment (NR-IQA) algorithms. Among all available schemes, we selected the Blind/Referenceless Image Spatial Quality Evaluator also known as BRISQUE [15] due to its high correlation with human judgments. The principle of this method is schematized in Fig. 3.

It achieves this by analyzing spatial domain features within the image, such as local mean and standard deviation. A machine learning model is then trained on a dataset of natural images with mean opinion scores to predict the image quality score. The higher BRISQUE scores, the lower image quality. And lower scores suggest higher image quality. Furthermore, BRISQUE has very low computational complexity and is an agnostic NR-IQA, that is useful for our purpose, since we do not know which kind of distortions may appear when considering deepfake video frames.

BRISQUE is based on a two-scale approach where 18 features are computed by scale, resulting in a total of 36 quality features.

In our approach, we choose to compute 37 quality features associated to image quality which include the 36 BRISQUE features along with the predicted overall quality score.

3.2.2. Frequency spectrum analysis

Given the advantages of fast operation speed and energy concentration, Discrete Cosine Transform (DCT) is often used for many tasks. Altering the structure of an image when faces are swapped for instance introduces artificial high frequencies. In [16], the authors investigated the impact of the

deepfake generation process on the frequency spectrum of images. Their results, illustrated by the Fig. 4, clearly indicate that deepfakes have a higher high-frequency intensity than real images and that it is possible to use this property to detect deepfakes. The origin of this phenomenon lies in the use of Generative Adversarial Networks (GANs), which are the keystone of modern deepfake models. These models are forced to use upsampling to generate images, which introduces more high frequencies due to interpolation.

We propose to extract features by exploiting the DCT domain to deal with the problem of artificial high frequencies. At this step, the image is divided into equally sized $n \times n$ blocks and a local 2-dimensional DCT transform is computed for each of these blocks. For each frequency block, we compute the ratio between the standard deviation of frequencies F and the average of frequency defined as:

$$r = \frac{\sigma|F|}{\mu|F|} \tag{1}$$

The highest 10th percentile average of the local block ratios across the image is then computed. This pooling result is used as frequency feature.

Finally a comprehensive set of 38 explainable characteristics is computed, encompassing the 37 quality features along with the frequency trait.

3.3. The designed deep classifier

The light designed classifier is mainly based on a succession of linear layers and RELU activations. Fig. 5 provides a detailed description of its characteristics. A batch normalization is applied on the input in
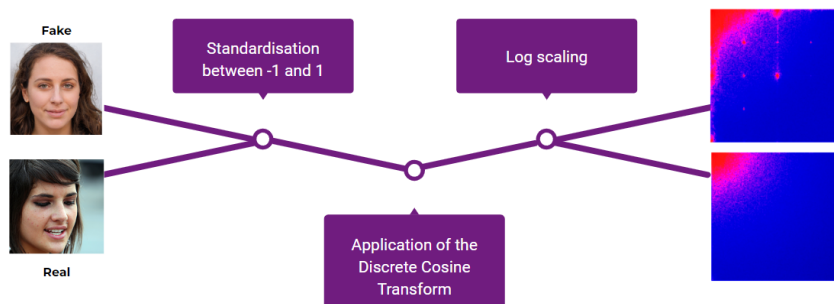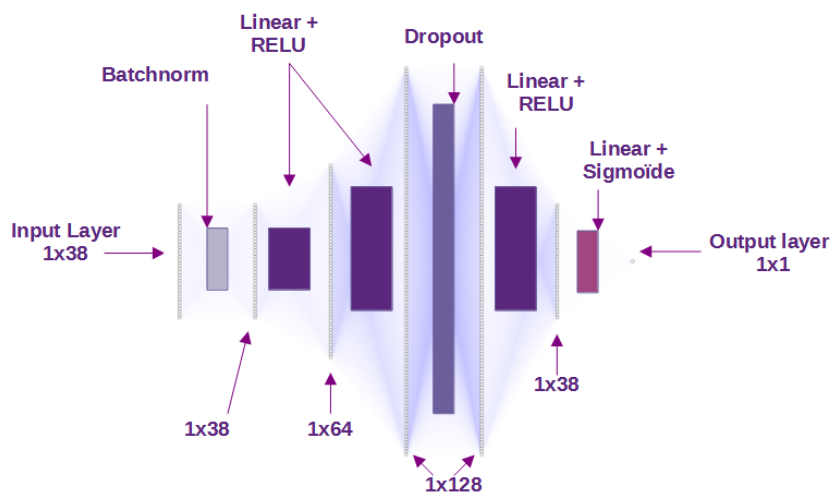

Figure 4: Frequency spectrum analysis

Figure 5: The designed deep classifier

order to re-center and re-scale input features. Then a succession of two linear layer and RELU activation block is defined. A dropout regularization with a rate of 0.33 is then applied to prevent overfitting. A new bock of linear layer and RELU activation function is then applied, followed by a linear layer and a sigmoid activation function. The loss used is the Binary Cross Entropy Logit function considering that we are operating within a binary classification framework with relatively evenly distributed classes.

The proposed classifier has been trained using a decreasing learning rate starting at 0.005. The number of epochs is 100 for a batch size of 1024 to yield the best correlation scores is applied. Early Stopping technique is used for regularization and model generalization enhancement.

4. Results

4.1. Experimental setup

To evaluate the performance of the proposed method, four databases have been selected:

1. VidTIMIT dataset [17] which comprises of video and corresponding audio recordings of 43 people, reciting short sentences. This database will serve as real samples.
2. DeepFakeTIMIT dataset [14] which contains videos where faces are swapped using the open source GAN-based approach which, in turn, was developed from the original autoencoder-based Deepfake algorithm. A total of 620 total videos with faces swapped is provided.
3. FF++ dataset [18] consisting of 1,000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data has been sourced from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries.
4. Celeb-DF [19] is a large-scale challenging dataset for deepfake forensics. It includes 590 original videos collected from YouTube with subjects of different ages, ethnic groups and genders, and 5,639 corresponding DeepFake videos.

From previously alluded databases, we generate one new database containing 79 385 real and 85 826 fake extracted frames from

- 300 real videos randomly selected from the VidTIMIT database,
- 320 fake videos randomly selected from the DeepFakeTIMIT database,
- 200 real and 600 fake videos both randomly selected from the FF++ dataset,
- 50 real and 50 fake videos both randomly selected from the Celeb-DF database.

From this new database, frames were then randomly separated into 4 different sets used during the learning and evaluation process of the proposed light CNN model, such as:

1. training set contains 31,627 True (real) frames and 34,826 False (fake) frames,
2. validation set containing 13,474 True (real) frames and 15,629 False (fake) frames,
3. test set (13,590 True (real) frames and 14,466 False (fake) frames,
4. generalization set (20,694 True (real) frames and 20,905 False (fake) frames,

The results were computed with a 5-fold process.

4.2. Performance evaluation

In order to evaluate the performance of the proposed scheme, five different measures have been used: 1) Accuracy, 2) Recall, 3) Precision, 4) F1-score and 5) AUC (Area Under the ROC Curve).

T_P, T_N, F_P, and F_N respectively represents true positive, true negative, false positive, and false negative.

Set	Acccuracy	F1	AUC	Precision	Recall
Train	0.96	0.96	0.99	0.97	0.95
Validation	0.82	0.84	0.88	0.83	0.86
Test	0.84	0.85	0.89	0.83	0.87
Generalization	0.49	0.60	0.52	0.50	0.75

Table 1
Performance evaluation of the proposed scheme.

The **Accuracy** is the fraction of predictions correctly identified by the model, ad is defined as:

$$\text{Accuracy} = \frac{T_P + F_N}{T_P + F_P + T_N + F_N} \tag{2}$$

The **Recall** is the percentage of positives well predicted by the model, defined as:

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{3}$$

The higher it is, the more the Machine Learning model maximizes the number of True Positives. When recall is high, this means it won't miss any positives. However, this gives no indication of its predictive quality on negatives.

The **Precision** is the number of positive predictions made defined as:

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{4}$$

The higher the precision, the more the Machine Learning model minimizes the number of False Positives. When precision is high, this means that the majority of the model's positive predictions are well-predicted positives.

The **F1-score** is an harmonic mean and provides a relatively accurate assessment of our model's performance. It is defined as

$$\text{F1-score} = 2 \times \frac{\text{Recall.Accuracy}}{\text{Recall} + \text{Accuracy}} \tag{5}$$

The higher the F1 Score, the better the model's performance.

The **AUC** computes the area under the ROC curve when plotting the precision versus the recall value. It represents the overall performance of the model, *i.e.*, the probability that a randomly chosen positive sample will be ranked higher by the model than a randomly chosen negative sample. A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5.

4.3. Results

Table 1 displays the obtained results. Whatever the considered measure, we obtained very good performance both in validation and testing, although, with repesct to the training performance, one notes a drop in performance, which is symptomatic of Deep Learning. The results for the generalization base (Celeb-DF) are much lower, which shows that our model is not yet sufficiently robust. These results are nevertheless satisfactory since Celeb-DF uses the latest deepfakes models, *i.e.*, unseen data, for which the quality is superior to those used in the three other sets. There is therefore a significant gap, which may explain this drop. To face this drawback, the amount of data for the training process is increased by adding samples that are more difficult to diagnose from the DFDC database [20].

To evaluate the performance of the proposed strategy against state-of-the-art approaches, we selected four methods. The first method, Xception [23], is an architecture based on Inception, but it replaces the traditional modules with depthwise separable convolutions. The second method, DSP-FWA [24],

	Xception	DSP-FWA	EfficientNetB4	EfficientNetB4ATTST	Ours
Accuracy	0.80	0.70	0.98	0.69	0.84
F1	0.85	0.83	0.89	0.72	0.85
AUC	0.87	0.88	0.92	0.82	0.89

Table 2
Performance evaluation obtained from the test set.

	Xception	DSP-FWA	EfficientNetB4	EfficientNetB4ATTST	Ours
Size	22,855,952	25,636,712	19,341,616	21,995,642	16,730

Table 3
Number of trainable parameters of each trial scheme.

uses convolutional neural networks (CNNs) to capture artifacts from the warping process needed to adapt the new face to the source image. The third method, EfficientNetB4 [25], is currently one of the leading networks in deepfake detection, as identified in a broader study. Lastly, EfficientNetB4ATTST [25] builds on EfficientNetB4 by integrating an attention mechanism and siamese training to enhance the model's generalization capability.

The performance of the proposed method is competitive with the analyzed state-of-the-art techniques. Although it does not surpass the best performer, EfficientNetB4, the proposed strategy outperforms the remaining three methods across the three performance measures used (Table 2). Overall, the performance is generally strong on face swap generation techniques.

Table 3 displays the number of trainable parameters for each trail state-of-the-art schemes and the proposed one. One may note that the introduced lightCNN clearly has significantly fewer parameters compared to state-of-the-art methods. Although this number of parameters is limited, the results achieved are competitive with those of state-of-the-art networks. Furthermore, the proposed scheme can be trained in just a few minutes (less than 5 minutes on a Dell Laptop SP15 with a NVIDIA GeForce RTX 4070, 8 Go GDDR6).

4.4. Discussion

To conclude, the performance of our model is very encouraging even if it does not outperform the trial state-of-the-art schemes.

Potential improvements include incorporating more advanced feature extraction techniques, such as multi-scale feature aggregation or hybrid models that combine convolutional layers with transformers. These techniques can capture fine-grained details crucial for detecting subtle manipulations in face-swapping.

Another approach is the use of perceptual loss functions to enhance the model's ability to detect subtle artifacts introduced by face-swapping. By focusing on high-level features from both authentic and manipulated faces, the model can better distinguish real faces from swapped ones.

Additionally, it is essential to assess and address potential biases in the model to ensure it performs well across different demographic groups. Implementing fairness-aware training strategies can help achieve this goal.

5. Conclusion

This work has enabled us to experiment with a hybrid approach between traditional forensic methods and state-of-the-art deepfake detection methods based on Deep Learning. Our architecture is more explainable and therefore viable in an integration context. What's more, its good performance with just

38 features and little data seems to indicate that by digging deeper in this direction it would be possible to achieve very good results. Finally, our architecture also has the advantage of being lightweight and quick to train and use, which is increasingly rare with the development of Deep Learning. It also makes it easy to integrate new feature extraction modules, which guarantees its durability. Future works will investigate how new residual signals computed in spatial, frequency and color spaces may help to increase the performance of the approach.

References

[1] Gao, J. and Concas, S. and Orrù, G.and Feng, X. and Marcialis, G.L.and Roli, F. Generalized Deepfake Detection Algorithm Based on Inconsistency Between Inner and Outer Faces. In Image Analysis and Processing - ICIAP 2023 Workshops. Vol 14365. Springer (2023)

[2] Datareportal, https://datareportal.com/reports/digital-2024-global-overview-report (2024)

[3] FaceApp, https://www.faceapp.com/ (2016)

[4] FakeApp, https://www.faceapp.com/ (2018)

[5] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung. Deepfake Detection: A Systematic Literature Review, in IEEE Access, vol. 10, pp. 25494-25513 (2022), doi: 10.1109/ACCESS.2022.3154404.

[6] Delfino, R.: Pornographic Deepfake–Revenge Porn's Next Tragic Act–The Case for Federal Criminalization. 887, 88 Fordham L. Rev (2019). SSRN33415933.

[7] Dixon, H.B., Jr.: Deepfakes: More frightening than photoshop on steroids. Judges J. 58(3), 35–37 (2019)

[8] Wu Jiahui and Zhu Yu and Jiang Xiaoben ans Liu Yatong and Lin Jiajun. Local attention and long-distance interaction of rPPG for deepfake detection, The Visual Computer, 40 (2) (2024), pp. 1083-1094

[9] Li Lingzhi and Bao Jianmin and Zhang Ting and Yang Haoand Chen Dong and Wen Fang, Face x-ray for more general face forgery detection, In Proceedings of the conference on computer vision and pattern recognition (CVPR), IEEE (2020), pp. 5001-5010

[10] Zhang, K. and Zhang, Z. and Li, Z. and Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters 23(10), 1499–1503 (Oct 2016).

[11] David Güera and Edward J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1-6, 2018

[12] Verdoliva, Luisa. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing. PP. 1-1. 10.1109/JSTSP.2020.3002101.

[13] J. Galbally and S. Marcel, "Face Anti-spoofing Based on General Image Quality Assessment," 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 2014, pp. 1173-1178, doi: 10.1109/ICPR.2014.211.

[14] Korshunov, Pavel and Marcel, Sébastien. (2018). DeepFakes: a New Threat to Face Recognition? Assessment and Detection.

[15] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

[16] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 304, 3247–3258.

[17] Sanderson, Conrad and Lovell, Brian. (2009). Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. LNCS. 5558. 10.1007/978-3-642-01793-3-21.

[18] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.

[19] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[20] Dolhansky, Brian and Bitton, Joanna and Pflaum, Ben and Lu, Jikuo and Howes, Russ and Wang, Menglin and Ferrer, Cristian. (2020). The DeepFake Detection Challenge (DFDC) Dataset. (arXiv:2006.07397)

[21] Li, L., Bao, J., Yang, H., Chen, D., and Wen, F. (2020). Advancing high fidelity identity swapping for forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5074-5083).

[22] Ding, X., Raziei, Z., Larson, E. C., Olinick, E. V., Krueger, P., and Hahsler, M. (2020). Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security, 2020(1), 1-12.

[23] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)

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

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

[26] Chanda, K., Ahmed, W., Banik, S. (2024). Deepfake Image Detection for Low and High Quality Images for Biometric Face Recognition. In: Nayak, R., Mittal, N., Kumar, M., Polkowski, Z., Khunteta, A. (eds) Recent Advancements in Artificial Intelligence . ICRAAI 2023. Innovations in Sustainable Technologies and Computing. Springer, Singapore.