Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform

Mohammed Akram Younus
Department of Computer Science
College of Science
University of Diyala
Diyala, Iraq
mohammed.akram@sciences.uodiyala.edu.iq

Taha Mohammed Hasan
Department of Computer Science
College of Science
University of Diyala
Diyala, Iraq
dr.tahamh@sciences.uodiyala.edu.iq

***Abstract*—DeepFake using Generative Adversarial Networks (GANs) tampered videos reveals a new challenge in today's life. With the inception of GANs, generating high-quality fake videos becomes much easier and in a very realistic manner. Therefore, the development of efficient tools that can automatically detect these fake videos is of paramount importance. The proposed DeepFake detection method takes the advantage of the fact that current DeepFake generation algorithms cannot generate face images with varied resolutions, it is only able to generate new faces with a limited size and resolution, a further distortion and blur is needed to match and fit the fake face with the background and surrounding context in the source video. This transformation causes exclusive blur inconsistency between the generated face and its background in the outcome DeepFake videos, in turn, these artifacts can be effectively spotted by examining the edge pixels in the wavelet domain of the faces in each frame compared to the rest of the frame. A blur inconsistency detection scheme relied on the type of edge and the analysis of its sharpness using Haar wavelet transform as shown in this paper, by using this feature, it can determine if the face region in a video has been blurred or not and to what extent it has been blurred. Thus will lead to the detection of DeepFake videos. The effectiveness of the proposed scheme is demonstrated in the experimental results where the "UADFV" dataset has been used for the evaluation, a very successful detection rate with more than 90.5% was gained.**

Keywords— DeepFake, Generative Adversarial Networks GANs, Haar Wavelet, Blur inconsistency.

I. INTRODUCTION

With the advanced techniques of image editing and the emergence of the GAN which is counted as a class of machine learning systems invented in 2014 [1], image generation processes have become devilishly available and made the fake digital videos creation more convenient and acceptable than ever since. The time of producing fraud images and videos has decreased remarkably in the past few years due to the rapid growth of advanced techniques like machine learning and computer vision besides the accessibility to high-throughput computing which overrides the need for manual and human editing steps and traditional applications such as photoshop.

The new methods vein of artificial-based fake video generation named DeepFake has drawn a lot of attention in recent years. A video of a specific individual is taken as an input target and outputs a new video with another individual as a source displaced with the target's faces. Deep neural networks are used in the creation of DeepFake, This is the main idea behind DeepFake, it has been trained to spontaneously to plot the facial expressions of the source face to the face of the target, the outcome videos can create believable forgeries and a high level of authenticity.

The need for a dependable detection method of AI-doctored images is rapidly increasing to distinguish whether a digital image/video is tampered with or not. Testing the content of digital images/videos or distinguishing forged regions would be very useful indeed, for example in the court of law when digital videos are viewed as evidence or the use for malicious purposes.

In this paper, the proposed method illustrates that it can effectively and efficiently recognize DeepFake videos from the original ones. The method relies on the property of the DeepFake videos generation, taking the advantage of the fact that production speed and limitation of resources of computation, the DeepFake algorithm is only able to generate fake faces with specific size and resolution, an affine transformation, and a blur function must be added to the synthesized faces to match and fit the arrangement of the source's face on original videos. This transformation leaves distinguished traces due to the resolution and blur inconsistency between the transformed region "Region Of Interest" (ROI) and the surrounding area. This blur inconsistency is used to detect DeepFake forgery videos. The method detects such inconsistency by comparing the blurred synthesize areas ROI and the surrounding context with a dedicated Haar wavelet transform function by simulating the resolution and blur inconsistency in face transformation directly. Detecting faces in each frame and then extract landmarks to compute the transform matrices to align the faces to a standard configuration is the first step. Some methods can be used to estimate the blur extent based on the presumption that the extracted face image has been blurred.

The proposed DeepFake detection method relies on the Haar wavelet transform. It reveals whether or not a given face image has been blurred by the artifacts traces caused by aligning GAN fake face to an image depending on the analysis of edge type, and it can determine to what extent the ROI has been blurred by the analysis of edge sharpness. This method uses two advantages of the capacity of Haar wavelet transform, first distinguishing different kinds of edges, second retrieving sharpness from the blurred image. Furthermore, it is very effective and fast since the uniform background of faces in the images will have no effect and it does not need to reconstruct the blur matrix function.

The blur inconsistency introduced by the construction pipeline of DeepFake Fig. 3 is a result of the affine transforms of the synthesized face. Detecting the implicit resampling method is widely studied, e.g., [2-4]. However, these

978-1-7281-5249-3/20/$31.00 ©2020 IEEE

algorithms commonly aim to estimate the sampling operation for the whole image, for our DeepFake detection approach, a straightforward settlement obtained by only comparing the area of potentially synthesized faces (ROI) to the remainder of the image (background). Where the background is supposed to have no blur effect or at least less blur effect than (ROI).

In this paper, a new DeepFake detection method is proposed using the Haar wavelet transform. The method takes advantage of the Haar wavelet transformation ability in discriminating different types of edges, also determines the extent of the blurred image, which is based on edge sharpness analysis, thus it will be the indication of manipulation in video frames. The rest of the paper is organized as follows: In section II, the related work is presented and Fig. 1 shows the DeepFake algorithm overview, section III, the method presented in detail and Fig. 2 shows the edge classification diagram, Fig. 3 illustrate the blur inconsistency introduced by the construction pipeline, section IV, reveal the method algorithm steps, Fig. 4 shows how Haar Wavelet is used to detect DeepFake, section V, shows experimental results and Figs. 6 and 7 illustrate the example of the proposed method on one of the DeepFake generated videos from the "UADFV" dataset; finally, we conclude the paper in section VI.

II. Related Works

The most recent new developments of Artificial-based video synthesis methods rely on the new deep learning models such as GANs [1]. The GAN model comprises of two parts which both are deep neural networks trained in conjunction with each other, the first one is the generator network, this part aims to generate face images so close to the real images, while the discriminatory section of the network aims to distinguish between them, that will lead into producing synthesized face images with a very realistic appearance. That is how the DeepFake creation algorithm works. The DeepFake video creation begins with an input video as a target of a specific individual, a new video is produced as an outcome with the target's faces swapped with an individual of the source Fig. 1, that is what GAN has been trained for which is to swap the faces between the target and the source. Zhu et al. [5] recently proposed a scheme to increase the performance of GAN, called Cycle-GAN. Recycle-GAN is proposed by Bansal et al. [6] where stepped further than before, by incorporating the spatial information with temporal cues by conditional generative adversarial networks.

Algorithms such as [7, 8] are used to detect traditional forgeries. Zhou et al. [8] proposed face tampering detection using two-stream CNN. Cozzolino et al. [7] proposed the Noise Print CNN model to detect the trace fingerprints for forgery. The lack of realistic eye blinking in DeepFake videos is observed by Li et al. [9] since training images which are obtained as a source, ordinarily do not have photographs with the individual's closed eyes. The CNN/RNN model is used to detect the lack of eye blinking exposed in DeepFake videos. However, revoking this method of detection can be easily made by purposely combine images in training time with faces of closed eyes. Yang et al. [10] employed the inconsistency of head pose to detect DeepFake videos. Tan et al. [11] utilized the color disparity in non-RGB color spaces between the original images and GAN generated images to classify them. Detecting GAN generated imagery using color cues [12] also applied the color difference method between original images and GAN generated images. However, this method works with the whole image and it is not tested to inspect specific zones as in the DeepFake case. The MesoNet [13] utilized CNNs to classify original faces images and forged faces images produced via the DeepFake pipeline. The DeepFake video detection using recurrent neural networks [14] extended the MesoNet method to the temporal domain by combining RNN with CNN. This method has its drawbacks. In specific, it needs big data for both original and forged images as training data, leading into consuming time and resources.

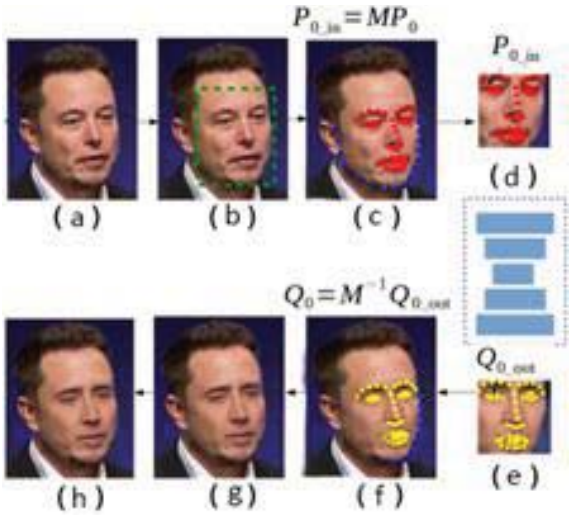

Fig. 1. Overview of DeepFake algorithm. (a) The source face. (b) The face boundary has been detected. (c) Landmarks have been spotted. (d) The face is cropped then warped to a standardized face. (e) The DeepFake face is synthesized. (f) The face is metamorphosed. (g) The synthesized face is fine-tuned based on landmarks in (c). (g,h) The fake face is synthesized into the source image.

III. DeepFake Detection by Detecting Blur Edges

Two methods can be used to estimate the blur extent, direct and indirect. The Linear blur image can be described as:

$$G = H * F + N \quad (1)$$

Where *G, F*, and *N* matrices, represent the noisy image, while the blur function represented by the *H* matrix.

Indirect methods depend on the blur reconstruction function when the *H* matrix is unknown, which is blurs estimation and blur identification. There are a large number of related studies concerning this matter such as Rahtu, E., et al. [15] propose a local phase quantization for blur-insensitive image analysis; Zhu, X., et al. in [16] resort to Estimating spatially varying defocus blur from a single image; Kerouh, F., et al in [17] proposed a general approach to detect tampered and generated image.

Direct methods can measure the blur function extent by testing some distinctive features in an image. The edge feature is one of these features that can be used. If the blur is present, both the edge sharpness and its type will be changed and that will indicate whether the face image has been manipulated or not.

Different types of edges are present in an image. Generally, there are three classes of edge type: Dirac-Structure, Step-Structure, and Roof-Structure [18]. The Step-Structure type is divided according to the change of intensity whether it is gradual or not into: "A Step-Structure" and "G

Step-Structure" as shown in Fig. 2. [19], [20]. However, every image has all types of edges more or less, most of the G Step-Structure and Roof-Structure are sharp enough. In case it is blurred, the edges lose their sharpness. The sharpness parameter is measured by the sharpness parameter (α). This method detects whether a face image is blurred or not based on Dirac-structure and A Step-Structure. A blur extent is identified by taking sharpness of Roof-Structure and G Step-Structure into account. The sharpness of the edge is indicated by the parameter α ($0<\alpha<\pi/2$), if α is larger, means the edge is sharper.

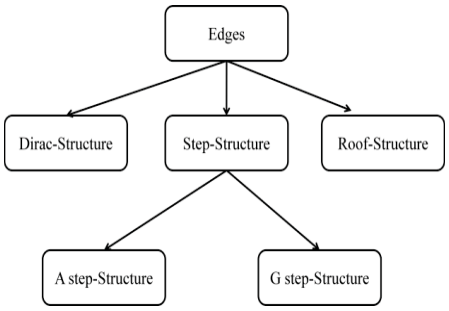

Fig. 2. Edge Classification.

The scheme we used judges if a given image has been blurred or not depending on whether if it has A Step-Structure or Dirac-Structure, the percentage of G Step-Structure and Roof-Structure which are more probably found in a blurred image is used to set the value of the blur extent. By comparing the blur extent of the (ROI) with the blur extent of the rest of the image we can determine if the images (Frames of the video) have been tampered with or not.

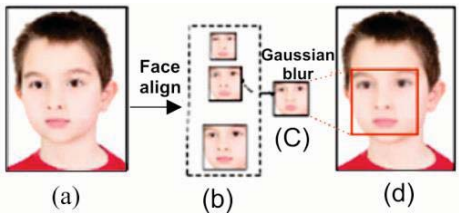

Fig. 3. (a) Orginal image. (b) Different scales of aligned face, (c) Arbitrarily picked scale of the face and applied Gaussian blur to it. (d) Affine warped to source image with a low α value.

IV. DEEPFAKE DETECTION ALGORITHM

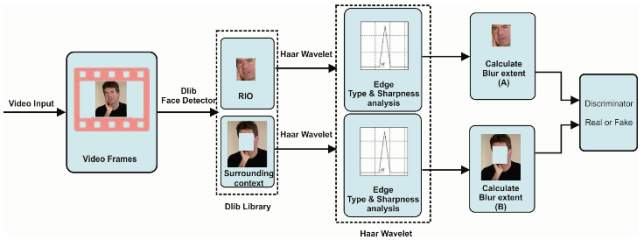

Fig. 4. Work Diagram for the DeepFake detection system

The above diagram shows the DeepFake system. The algorithm used to detect DeepFake Fig. 5 using Haar wavelet transform and edge detection is:

Step 1: Using software package Dlib [21] the faces are detected in the original images and the face region extracted.

Step 2: Perform Haar Wavelet Transform to the (ROI) found in step 1 and the original image, the level for the decomposition would be three.

Step 3: The edge map constructs in each scale:

$$E\ map_i(K,l) = \sqrt{LH_i^2 + HL_i^2 + HH_i^2}\,(i = 1,2,3) \qquad (2)$$

Step 4: Find the local maxima in each window by partitioning the edge maps. The size in the highest window scale is (2×2), the next window scale is (4×4), and the last one is (8×8). The outcome is denoted as

$$E\ max_i(i = 1,2,3)$$

Where $E\ \max_i$ shows the intensity of the edge, the higher the value of $E\ \max_i$, the stronger intense the edge is.

Step 5: For a specific threshold value, if $E\ max_i(k,l) > threshold, (k,l)$ is categorized as an edge point; else, it is a non-edge point. Let N_{edge} be the total number of them.

Step 6: Find all edge points in the image (the Dirac-Structure and the A Step-Structure). For each edge point (k,l), if the $E\ max_1(k,l) > E\ max_2(k,l) > E\ max_3(k,l)$, (k,l) is considered either Roof-Structure or G Step-Structure. Let N_{da} be the total number of them.

Step 7: Find all Roof-Structure points, if $E\ max_2(k,l) > E\ max_1(k,l)$ and $E\ max_2(k,l) > E\ max_3(k,l)$, (k,l) is considered as a Roof-Structure Let N_{rg} be the total number of them.

Step 8: Find all G Step-Structure and Roof-Structure edge points which have lost their sharpness. For each edge point (k,l), if $E\ max_i(k,l) < threshold, (k,l)$ is more probable to be in a blurred image. Let N_{brg} be the total number of them.

Step 9: Calculate the ratio of all the edges (Dirac-Structure and A Step-Structure), $per = \frac{N_{da}}{N_{edge}}$, if Per < MZero, consider the image as a blurred image else, it is not blurred, where MZero is a positive number close to zero.

Step 10: Calculate BlurExtent by $BlurExtent = \frac{N_{Brg}}{N_{rg}}$. Which represents the image blur confident coefficient.

Step 11: If the blur is found in (ROI), compare the BlurExtent of (ROI) with the rest of the image.

Step 12: Depending on step 11, judge if the image frame is real or fake.

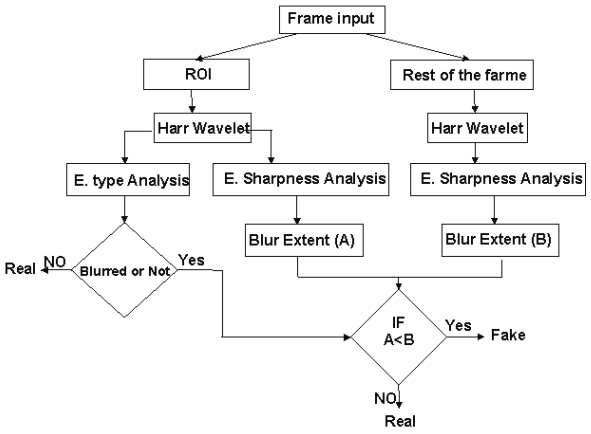

Fig. 5. DeepFake detection Algorithm using Haar Wavelet.

A. The Effect of Blur on Different Edges

The noise factor *N* in "(1)" can be neglected since there is a small noise ratio in photos usually acquired from digital

cameras. The main blur functions H is a convolution operation that affects the equation and will change the edge property. Take into consideration that there will be no Dirac-Structure or A Step-Structure in the blurred image. On the other hand, both Roof-Structure and G Step-Structure will tend to lose their sharpness (less α value).

B. Sharpness Detection and Edge Type

The best multi-resolution analysis ability is found in the Wavelet Transform. The capability to detect the area of irregular structures by the local maxima of a Wavelet Transform is proved in [22], Tallavó, F., et al. [23] utilize the Modulus-Angle Separated Wavelet (MASW) for the detection of both edge, Dirac-Structure and Step-Structure. After all, (MASW) is weighed as time-consuming. Instead, the Haar Wavelet Transform (HWT) is used in this work.

Discrete Haar functions can be described as functions determined by sampling the Haar functions at 2^n points. A matrix form can represent the function in a convenient means. Every row in the matrix H(n), have a discrete Haar sequence *Haar(w, t)* each alone, where the index (w) represents the number of the Haar function, the discrete point of the function determination interval is identified by index (t). By the following recurrence relation [24], any dimension of the Haar matrix can be gained.

$$H(n) = \begin{bmatrix} H(n-1) & \otimes [1 \quad 1] \\ 2^{\frac{n-1}{2}} I(n-1) & \otimes [1 \quad 1] \end{bmatrix}, H(0) = 1$$

Where $H(n)$ - the discrete Haar functions of degree 2^n matrix, $I(n)$ - identity matrix of degree 2^n.

V. Results

The DeepFake video dataset "UADFV" [10], is experimented with the proposed method. The "UADFV" dataset comprises 98 videos in total, 49 unmanipulated videos and 49 videos with the faces has been changed. Each video lasts approximately 12 seconds (total frames around 35280). The experiment showed that this model has a powerful generalization strength to detect DeepFake videos generated by GANs. This means that images/videos that are generated by various GANs model can be distinguished in fast and acceptable accuracy. TABLE I shows the Area Under Curve (AUC) performance comparison between the proposed method and some other methods [25].

A test sample is shown in Fig. 6, the video is one of the "UADFV" datasets, with a length of 292 frames (approximately 10 sec) which prove the strength of the proposed method in the detection of fake synthesized face. The results show the difference between the sharpness edges of (ROI) area by the blur extent value (the large value the more sharpness the edges are).

TABLE I. A Comparison of DeepFake Methods Performance

Methods	**"UADFV" dataset**
Two-stream NN [12]	85.1
Meso-4 [17]	84.3
MesoInception-4	82.1
Head Pose [14]	89.0
Proposed Model	90.5

The two frames in Fig. 7, revealed the difference in the blur extent between the ROI and the rest of the two frames image leading to the distinction of the manipulation which is the result of GAN (the fake image), the left image shows a value an approximate blur extent value of 22.7 for both regions, where it shows a large difference in blur extent value between ROI and the rest of the image in the right frame.

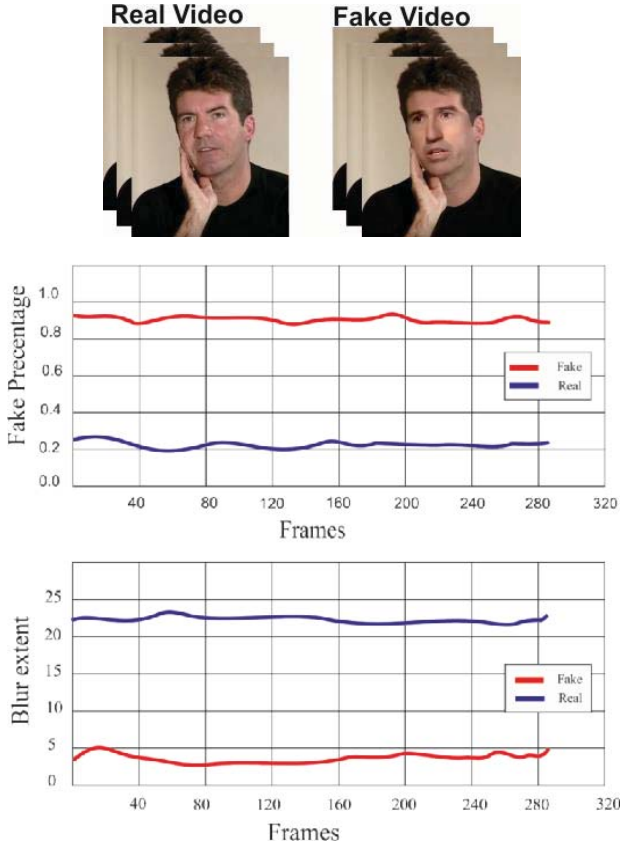

Fig. 6. Example of the proposed method on one of the DeepFake generated videos from the "UADFV" dataset.

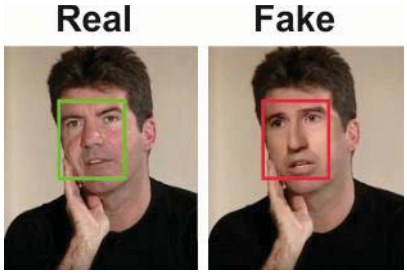

Fig. 7. The discrimination between the real and fake faces.

VI. Conclusions

Our method proposes a new technique to detect artificial-generated fake face images or videos known as the DeepFakes. Based on the fact that the DeepFake method can only produce face images of limited resolutions and fixed size, which are then needed to be further blurred and transformed to match the faces to be swapped in the original video. This additive blur and transformations on the ROI leave special artifacts in the resulting DeepFake videos, which can be effectively captured by detecting the differences between ROI and the rest of the image using Haar Wavelet transformation. We tested our method on a set of available DeepFake videos which showed its effectiveness in practice.

However, till now no full technique or method can equip a perfectly universal solution. There is a need for some techniques that have particular robustness to current image processing such as blurring, scaling, and rotation. Regrettably, some techniques can be accurate but have very high computational complexity and resource consumption.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[2] N. Dalgaard, C. Mosquera, and F. Pérez-González, "On the role of differentiation for resampling detection," in *2010 IEEE International Conference on Image Processing*, Hong Kong, 2010, pp. 1753-1756: IEEE.

[3] M. Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue," in *Proceedings of the 10th ACM workshop on Multimedia and security*, United Kingdom, 2008, pp. 11-20: ACM.

[4] M. Kirchner, R. J. I. T. o. I. F. Bohme, and Security, "Hiding traces of resampling in digital images," vol. 3, no. 4, pp. 582-592, 2008.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.

[6] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Washington, 2018, pp. 119-135.

[7] D. Cozzolino, L. J. I. T. o. I. F. Verdoliva, and Security, "Noiseprint: a CNN-based camera model fingerprint," vol. 15, pp. 144-159, 2019.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 1831-1839: IEEE.

[9] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, 2018, pp. 1-7: IEEE.

[10] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 8261-8265: IEEE.

[11] H. Li, B. Li, S. Tan, and J. Huang, "Detection of Deep Network Generated Images Using Disparities in Color Components," *arXiv preprint arXiv:07276,* 2018.

[12] S. McCloskey and M. Albright, "Detecting GAN-generated Imagery using Color Cues," *arXiv preprint arXiv:08247,* 2018.

[13] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, 2018, pp. 1-7: IEEE.

[14] D. Güera and E. J. Delp, "DeepFake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, 2018, pp. 1-6: IEEE.

[15] E. Rahtu, J. Heikkilä, V. Ojansivu, T. J. I. Ahonen, and V. Computing, "Local phase quantization for blur-insensitive image analysis," vol. 30, no. 8, pp. 501-512, 2012.

[16] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating spatially varying defocus blur from a single image," *IEEE Transactions on image processing,* vol. 22, no. 12, pp. 4879-4891, 2013.

[17] F. Kerouh and A. Serir, "A no reference quality metric for measuring image blur in wavelet domain," *International journal of digital information and wireless communication,* vol. 1, no. 4, pp. 767-776, 2012.

[18] Y.-y. Zheng, J.-l. Rao, and L. Wu, "Edge detection methods in digital image processing," in *2010 5th International Conference on Computer Science & Education*, Hefei, 2010, pp. 471-473: IEEE.

[19] J.-H. Lee, Y.-S. Ho, and I. Representation, "High-quality non-blind image deconvolution with adaptive regularization," *Journal of Visual Communication and Image Representation,* vol. 22, no. 7, pp. 653-663, 2011.

[20] R. Liu and J. Jia, "Reducing boundary artifacts in image deconvolution," in *2008 15th IEEE International Conference on Image Processing*, San Diego, CA, 2008, pp. 505-508: IEEE.

[21] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research,* vol. 10, no. Jul, pp. 1755-1758, 2009.

[22] M.-Y. Shih, D.-C. Tseng, and V. Computing, "A wavelet-based multiresolution edge detection and tracking," *Image,* vol. 23, no. 4, pp. 441-451, 2005.

[23] F. Tallavó, G. Cascante, and M. Pandey, "Experimental and numerical analysis of MASW tests for detection of buried timber trestles," *Soil Dynamics and Earthquake Engineering,* vol. 29, no. 1, pp. 91-102, 2009.

[24] P. Porwik, A. Lisowska, and vision, "The Haar-wavelet transform in digital image processing: its status and achievements," *Machine graphics,* vol. 13, no. 1/2, pp. 79-98, 2004.

[25] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:.00656,* 2018.