Journal of Engineering Science and Technology
Vol. 18, No. 2 (2023) 1164 - 1183
© School of Engineering, Taylor's University

A NEW MANIPULATION DETECTION AND LOCALIZATION SCHEME FOR DIGITAL FACE IMAGES

ZAHRAA AQEEL SALIH[1], RASHA THABIT[2,*], KHAMIS A. ZIDAN[3]

[1]Computer Engineering, Al-Iraqia University, Baghdad, Iraq
[2]Department of Computer Techniques Engineering, Dijlah University College, 10001 Baghdad, Iraq
[3]Vice Rector of Al-Iraqia University for Scientific Affairs, Al-Iraqia University Baghdad, Iraq
*Corresponding Author: rashathabit@yahoo.com

Abstract

Recently, the research community spent a lot of effort to present face image manipulation detection techniques, however, all the available schemes have their own limitations and there is no global detection scheme. To avoid the problems that can face the deep-learning based techniques, we suggest another direction of research to implement a new Face Image Manipulation Detection (FIMD) scheme which is based on face detection and image watermarking techniques. The proposed FIMD scheme at the sender side has two stages where the first stage is applied to detect and select the face area and the second stage is applied to generate and embed the manipulation detection and localization data. The proposed FIMD scheme at the receiver side has also two stages where the first stage is applied to detect and select the face area and the second stage is applied to extract the embedded data and check the authenticity of the received face image. The experiments that have been conducted to check the performance of the proposed FIMD scheme proved its efficiency in detecting different types of face manipulations such as face swap, expression swap, attribute attacks, and retouching attacks. The detection accuracy is 100 % and no false detection results have been recorded, in addition, the scheme obtained promising results in terms of visual quality of the watermarked face images and high embedding capacity. The general comparison with the state-of-the-art detection schemes proved the superiority of the proposed FIMD scheme.

Keywords: DeepFakes reveal, Face image security, Face manipulation detection, Face manipulation localization, Multimedia forensics.

1. Introduction

Recently, the term ‘DeepFakes’ have been widely distributed which refers to fake digital data that are generated using deep-learning algorithms [1-6]. In most cases, the DeepFakes have been used for harmful purposes such as: swapping the faces into porn images and videos, generating fake news, imposing financial fraud, and many others [7, 8]. On the other hand, the digital face manipulation applications are also increasing which can be used for harmful or harmless targets. Therefore, multimedia forensics and digital data security systems have dedicated a lot of effort and researches to detect the manipulations in digital face images. In the last few years, several face manipulation detection algorithms have been developed and used in multimedia forensics [9-14].

Various face image generation and manipulation methods are available such as face synthesis [15], face swap [16, 17], face morphing [18-22], face attribute manipulation or retouching [23-25], face expression swap [26, 27], and others. Some valuable review papers have been presented which contain detailed information about the abovementioned face image manipulation methods and their detection algorithms [28-32].

Most of the available manipulation detection algorithms have been implemented based on the methods of generating the fake or manipulated face images which makes them restricted, and they can reveal only one type of manipulation [33-39]. The deep-learning based manipulation detection algorithms rely on training using specific data sets thus the detection results are good only when the input image is close to the training data [40]. Most of the available deep-learning based techniques have recorded false detection cases which reduces their accuracy [41-44]. On the other hand, many detection methods used supervised algorithms therefore their practical application is time and efforts consuming [45-47]. To improve the performance of the deep-learning detection techniques, large and high-quality datasets are required which are usually not available for free [48]. Some manipulation detection algorithms obtained promising results; however, the production of high-quality fake images is one of the challenges that can face the detection process [14, 49, 50]. Obviously, the available face manipulation detection algorithms have many limitations and to the current date there is no universal detection algorithm which makes inventing new detection algorithms to overcome the limitations an urgent need.

In this paper, we suggest a new Face Image Manipulation Detection (FIMD) scheme based on face detection and image watermarking techniques. The proposed scheme is inspired by the medical image authentication schemes in which the images are divided into two regions called Region-of-Interest (ROI) and Region-of-Non-Interest (RONI). In order to protect the face area in the image and reveal manipulation in this region, the face can be considered as ROI while the remaining part of the image can be considered as RONI.

To implement an efficient FIMD scheme, the face region must be detected accurately using face detection algorithm. Then a successful image authentication technique must be adopted to ensure the integrity of the face region. Several face detection techniques have been presented in the literature and one of the most successful algorithms is the Multi-Task Cascaded CNN-based technique [51] which has been adopted in different face recognition systems [52-60]. In the proposed scheme, the face detection algorithm from [51] is used to detect the face area which can be considered as the first level in the implementation of the proposed FIMD scheme.

On the other hand, several watermarking-based medical image authentication algorithms are available [61-64], however, the most preferred are the robust reversible-based watermarking techniques [65-69] because they can withstand unintentional attacks while preserving the integrity of the medical image [70]. To implement the proposed scheme, the Slantlet transform (SLT) has been applied for transforming the image's blocks and the content-based embedding algorithm has been used to embed the binary bits in SLT coefficients. The choice of SLT and content-based embedding depends on several previous studies such as the methods in [65-67] which proved the efficiency of SLT-based watermarking compared with discrete wavelet transform (DWT)-based watermarking. The SLT-based methods perform better in terms of visual quality, robustness, and execution time. Based on that, the SLT-based medical image authentication algorithms that have been presented in [71] have been modified and adopted in the second level of implementing the proposed FIMD scheme. The algorithms in [71] have been presented for grayscale medical images with manually selected region of interest, therefore, they are not directly applicable in the proposed FIMD scheme. The novelty of the proposed FIMD scheme can be summarized in the use of face detection and selection steps for automatically and correctly localizing the face region and excluding it from the embedding procedure to ensure its integrity, in addition to the use of modified SLT and content-based embedding procedure for embedding and extracting the authentication information into colour face images.

The rest of the paper is organized as follows: section 2 presents the details of the proposed FIMD scheme including the block diagrams and algorithms; section 3 presents the experimental results and discussion; and section 4 illustrates the conclusions of this paper.

2. Proposed FIMD Scheme

The proposed FIMD scheme has two main algorithms that are the embedding algorithm which must be applied at the sender side and the extraction algorithm which must be applied at the receiver side. The following subsections present the proposed embedding and extraction algorithms in details.

2.1. Embedding procedure

The proposed embedding procedure starts by applying the Multi-Task Cascaded CNN-based technique [51] to detect the face box in the input image. The output of the detection algorithm contains fractional numbers refer to the upper left corner of the face box, its width, and its height. With simple calculations, the proposed algorithm can select the pixels of the face area and thus the face image can be divided into two regions ROI (i.e., face area) and RONI (i.e., the remaining area of the face image). The proposed FIMD scheme is inspired by the medical image authentication scheme from [71], however, the algorithms of [71] are not directly applicable in the proposed embedding algorithm. In [71], the ROI is selected manually using a polygon while in the proposed scheme the ROI is selected based on Multi-Task Cascaded CNN technique, in addition, the watermarking algorithms in [71] have been implemented for grayscale medical images while in the proposed scheme the algorithms are implemented for colour face images. Figure 1 presents the block diagram of the proposed FIMD embedding scheme. The detailed algorithms of the proposed FIMD scheme at the sender side are explained in the following subsections.

2.1.1. Face detection and selection algorithm

The input of the face detection and selection algorithm is the original face image (I_f) of size $(M \times N \times 3)$. The output of this algorithm is a mask image (I_M) of size $(M \times N)$. The steps of the algorithm are as follows:

1. Read the input face image I_f of size $(M \times N \times 3)$.
2. Apply Multi-Task Cascaded CNN algorithm to detect the face box in (I_f). The output of this step contains fractional numbers that refer to the top left corner of the face box (Cy, Cx) and its width (w) and height (h).
3. Round each number in the outputs $(Cy, Cx, w, \text{and } h)$ to its nearest integer that is greater than the value. The results after rounding are $(RCy, RCx, Rw, \text{and } Rh)$.
4. Calculate the positions of the face region pixels as follows:

   $px1 = RCx\ ,\ px2 = RCx + Rh,\ py1 = RCy, py2 = RCy + Rw.$
5. Define the face area pixels as follows $(px1{:}px2, py1{:}py2, :)$ which contains the face area for the three channels of the input colour image.

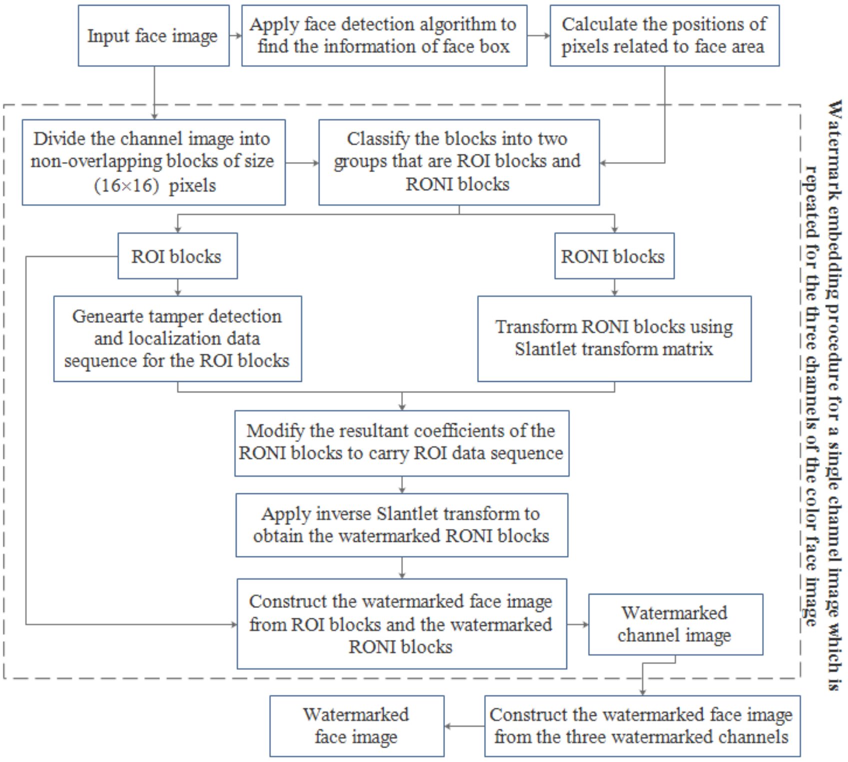

Input face image
Apply face detection algorithm to find the information of face box
Calculate the positions of pixels related to face area
Divide the channel image into non-overlapping blocks of size (16×16) pixels
Classify the blocks into two groups that are ROI blocks and RONI blocks
ROI blocks
RONI blocks
Genearte tamper detection and localization data sequence for the ROI blocks
Transform RONI blocks using Slantlet transform matrix
Modify the resultant coefficients of the RONI blocks to carry ROI data sequence
Apply inverse Slantlet transform to obtain the watermarked RONI blocks
Construct the watermarked face image from ROI blocks and the watermarked RONI blocks
Watermarked channel image
Watermark embedding procedure for a single channel image which is repeated for the three channels of the color face image
Watermarked face image
Construct the watermarked face image from the three watermarked channels

Fig. 1. Block diagram of the proposed FIMD scheme at the sender side.

6. Create a black image (i.e., binary image of zeros) of size $(M \times N)$, then change the pixels at the face positions to white. The resultant black and white image is called mask image I_M.

Figure 2 illustrates an example of face detection and selection results.

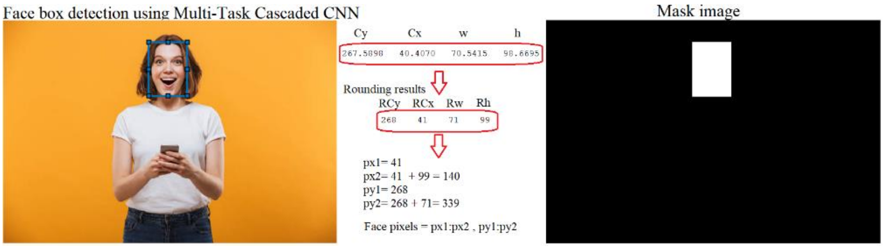


Fig. 2. Example on face detection and selection algorithm results for a face image of size (408×612×3) pixels.

2.1.2. Watermark embedding algorithm for a single channel

The input of this algorithm is one channel from the face image (ch) of size $(M \times N)$ and the mask image (I_M) of size $(M \times N)$ that is generated from the face detection and selection algorithm. The output of this algorithm is the watermarked channel image (W_{ch}) of size $(M \times N)$. The steps of the algorithm are as follows:

1. Read the channel image (ch) and the mask image (I_M)
2. Divide ch and I_M into non-overlapping blocks of size (16×16).
3. Calculate the average of I_M blocks and classify the ch blocks into two groups called ROI blocks and RONI blocks based on the following rule:

$$\text{Average of } I_M \text{ block } \begin{cases} = 0 & ch\ block\ at\ the\ same\ position\ belongs\ to\ RONI \\ \neq 0 & ch\ block\ at\ the\ same\ position\ belongs\ to\ ROI \end{cases}$$

4. Calculate the mean of the ROI blocks to be used as manipulation detection and localization data.
5. Convert the mean values to binary sequence each of size 8 bits and concatenate the results to obtain one binary sequence. Check the length of the resultant binary sequence, if the length is dividable by 11 without a remainder, then continue; else extend the binary sequence by zeros to make its length divisible by 11 without a remainder. This process is called extend 1.
6. Apply BCH (15,11,1) coding to the resultant binary sequence to increase its robustness. Save the resultant coded binary sequence as B_{seq}.
7. Check the length of B_{seq}., if the length is dividable by 64 without a remainder, then continue; else extend B_{seq}. by zeros to make its length divisible by 64 without a remainder. This process is called extend 2.
8. Calculate the length of B_{seq} and the capacity of RONI blocks. According to [71], the capacity of each (16×16) block is 64 bits thus the total capacity can be calculated using:

$$Capacity\ of\ RONI\ (bits) = \ total\ number\ of\ RONI\ blocks \times 64$$

9. Compare the length of B_{seq} with the $Capacity\ of\ RONI$; if the capacity is less than the length of B_{seq} (i.e., there is not enough space to hide the information) then stop the execution of the algorithm; else continue to the next step.
10. Divide B_{seq} into sub sequences each of length 64 bits.
11. Transform RONI block using Slantlet transform (SLT) matrix of size (16×16) and divide the resultant matrix of coefficients into 4 subbands called (High High (*HH*) subband, High Low (*HL*) subband, Low High (*LH*), and Low Low (*LL*) subband) as follows:

    $$TB = SLT_N\ B\ SLT_N^T$$

    where, B is the original RONI block, TB is the transformed block, and SLT_N is Slantlet matrix of size $(N \times N)$ [71]. Note that B, TB, and SLT_N have the same size.

    $$LL = TB\left(1{:}\frac{N}{2}, 1{:}\frac{N}{2}\right), \qquad LH = TB\left(\frac{N}{2} + 1{:}N, 1{:}\frac{N}{2}\right),$$
    $$HL = TB\left(1{:}\frac{N}{2}, \frac{N}{2} + 1{:}N\right), \qquad HH = TB\left(\frac{N}{2} + 1{:}N, \frac{N}{2} + 1{:}N\right).$$
12. Hide one binary subsequence from $Bseq$ in the transformed block by modifying *HL* and *LH* coefficients using the rules that have been applied in [71] to embed the binary sequence in RONI blocks.
13. Apply inverse SLT transform to obtain the watermarked RONI block.

    $$B = SLT_N^T\ TB\ SLT_N$$
14. Repeat the steps 11 and 13 until finish hiding all the binary sub sequences of B_{seq}.
15. Construct the watermarked channel image W_{ch} from the ROI blocks and watermarked RONI blocks.

2.1.3. Main embedding algorithm

The input of the main embedding algorithm is the original face image (I_f) while its output is the watermarked face image (WI_f). The steps of the algorithm are as follows:

1- Read the input face image (I_f).

2- Apply face detection and selection algorithm as illustrated in subsection (2.1.1).

3- Select the first channel image (ch) from the input face image.

4- Apply the watermark embedding algorithm for a single channel as illustrated in subsection (2.1.2).

5- Repeat the steps 3 and 4 to the second and third channels of the input face image.

6- Construct the watermarked face image (WI_f) from the three watermarked channels.

2.2. Extraction Procedure

As in the embedding procedure, the proposed extraction procedure starts by applying the face detection and selection algorithm as explained in subsection

(2.1.1). Then the watermark extraction procedure is applied to extract the embedded sequence from the RONI blocks. The algorithm depends on calculating the mean values of the ROI in the received image and comparing them with the extracted values; if the calculated and the extracted mean values for the same block are equal then the block is considered authentic; else the block is considered not authentic, and the localization process is applied to localize the manipulated blocks. Figure 3 presents the block diagram of the proposed FIMD extraction scheme. The detailed algorithms of the proposed FIMD scheme at the receiver side are explained in the following subsections.

2.2.1. Face detection and selection algorithm

The algorithm that has been explained in subsection (2.1.1) (from the embedding algorithms) is applied as the first stage of the extraction procedure which gives the mask image (I_M) at its output.

2.2.2. Watermark extraction procedure for a single channel

The input of this algorithm is one channel from the watermarked face image (W_{ch}) and the mask image (I_M) while its output is the authentication result for the input channel image. The steps of this algorithm are as follows:

1- Read the watermarked channel image (W_{ch}) and the mask image I_M.

2- Divide W_{ch} and I_M into non-overlapping blocks of size (16×16).

3- Calculate the average of I_M blocks and classify the ch blocks into two groups called ROI blocks and RONI blocks based on the following rule:

$$\text{Average of } I_M \text{ block } \begin{cases} = 0 & ch\ block\ at\ the\ same\ position\ belongs\ to\ RONI \\ \neq 0 & ch\ block\ at\ the\ same\ position\ belongs\ to\ ROI \end{cases}$$

4- Calculate the mean of the ROI blocks and save them for later comparison steps.

5- Transform RONI block using SLT matrix of size (16×16) and divide the resultant matrix of coefficients into 4 subbands called (High High (*HH*) subband, High Low (*HL*) subband, Low High (*LH*), and Low Low (*LL*) subband) as explained at the embedding procedure.

6- Extract the hidden binary sub sequences in the transformed RONI blocks using the extraction rules from [71]. Then concatenate the extracted sub sequences to form one binary sequence.

7- Remove the zeros that have been added to the sequence in the ‘extend 2’ process which has been explained in step 7 of subsection (2.1.2) (from the embedding algorithms).

8- Apply BCH (15,11,1) decoding to convert the resultant binary sequence.

9- Remove the zeros that have been added to the sequence in the ‘extend 1’ process which has been explained in step 5 of subsection (2.1.2).

10- Convert the resultant binary sequence to 8 bits sub sequences and convert each subsequence to decimal to recover the original mean values.

11-Compare the extracted mean values from RONI with the calculated mean values for ROI in step 4. When the compared mean values are equal then the face image is authentic, and the execution stops here. If the compared mean values are not equal then the face image is considered not authentic, and the algorithm is continued to the next step.

12-Localize the tampered blocks at which the mean values are not equal by drawing a border on each detected tampered block.

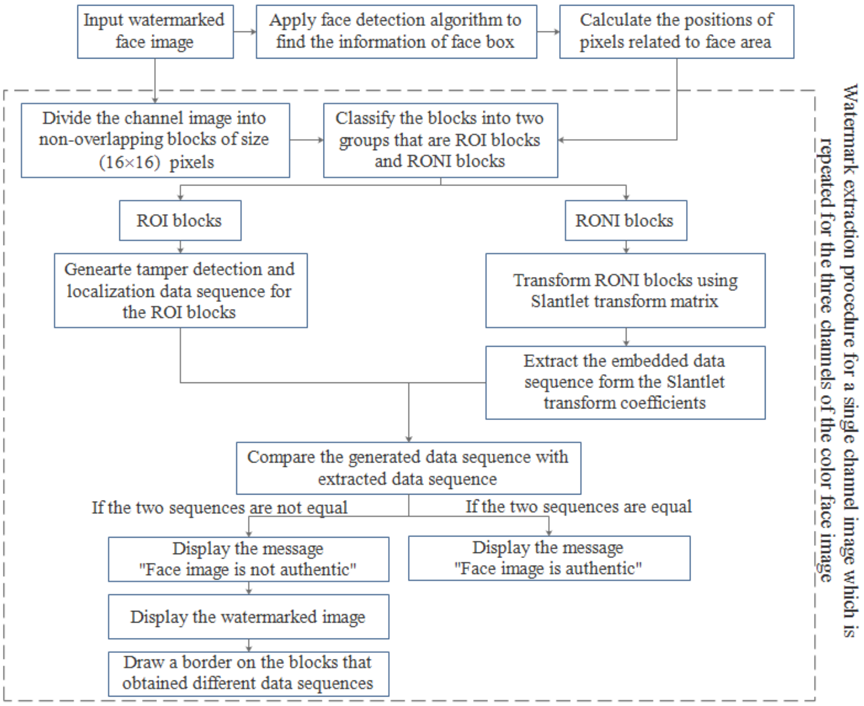

Input watermarked face image
Apply face detection algorithm to find the information of face box
Calculate the positions of pixels related to face area
Divide the channel image into non-overlapping blocks of size (16×16) pixels
Classify the blocks into two groups that are ROI blocks and RONI blocks
ROI blocks
RONI blocks
Genearte tamper detection and localization data sequence for the ROI blocks
Transform RONI blocks using Slantlet transform matrix
Extract the embedded data sequence form the Slantlet transform coefficients
Compare the generated data sequence with extracted data sequence
If the two sequences are not equal
If the two sequences are equal
Display the message "Face image is not authentic"
Display the message "Face image is authentic"
Display the watermarked image
Draw a border on the blocks that obtained different data sequences
Watermark extraction procedure for a single channel image which is repeated for the three channels of the color face image

Fig. 3. Block diagram of the proposed FIMD scheme at the receiver side.

2.2.3. Main extraction algorithm

The input of this algorithm is the watermarked face image (WI_f) while its output is the authentication and manipulation localization results. The steps of this algorithm are as follows:

1- Read the input watermarked face image (WI_f).

2- Apply face detection and selection algorithm as illustrated in subsection (2.2.1).

3- Select the first channel image (W_{ch}) from the input watermarked face image.

4- Apply the watermark extraction algorithm for a single channel as illustrated in section (2.2.2).

5- Repeat the steps 3 and 4 to the second and third channels of the input watermarked face image.

6- Display “Face image is authentic” when the three channels are authentic. Else display “Face image is not authentic” and display the face image after manipulation localization.

3. Experimental Results and Discussion

To test the performance of the proposed FIMD scheme, the experiments have been conducted for colour face images with different sizes which have been collected from various websites such as [72, 73]; samples of these test are shown in Fig. 4. The experimental work includes the face detection and selection test, invisibility test, capacity and payload test, and tamper localization test for different face image manipulation attacks. The final subsection presents a general comparison with the state-of-the-art face image manipulation detection schemes.

Image 1 Image 2 Image 3 Image 4

Image 5 Image 6 Image 7 Image 8

Image 9 Image 10 Image 11 Image 12

Image 13 Image 14 Image 15

Fig. 4. Samples of the face images that have been used in the experiments.

3.1. Face detection and selection test

The face detection and selection algorithm has been tested to ensure the ability of the detection scheme to accurately select the face area. The experimental results proved that the detection scheme is accurate in the detection and selection of the face area and there is no false detection result. Samples of the results of the face detection and selection test are shown in Fig. 5.

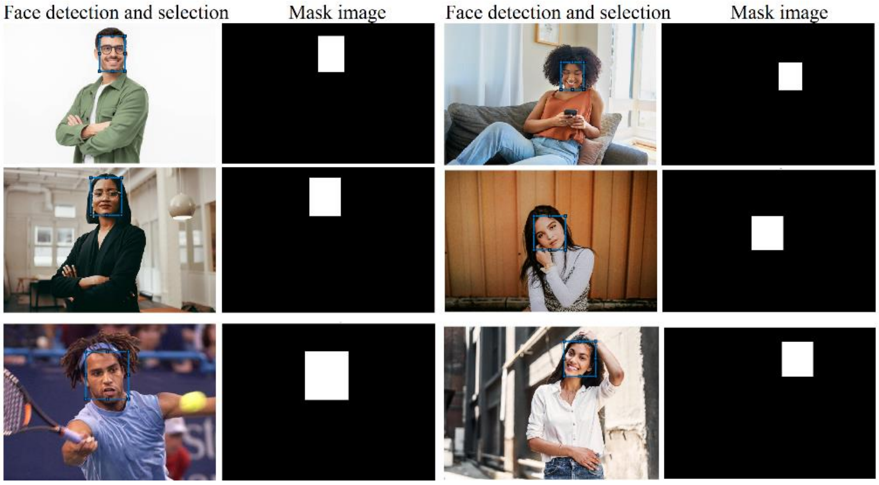


Fig. 5. Samples of face detection and selection test results.

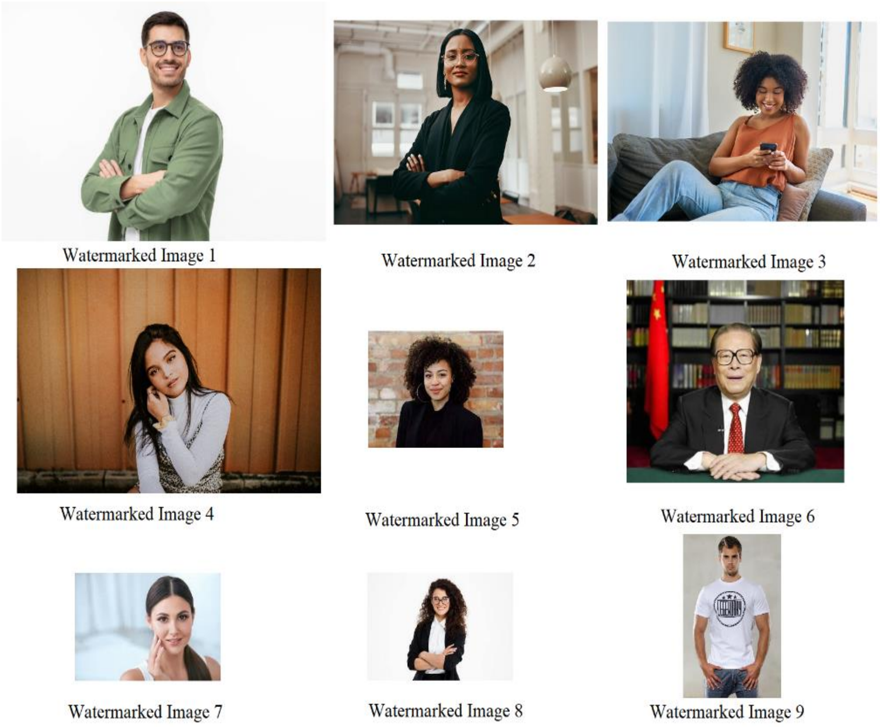
3.2. Invisibility test

To test the invisibility of the watermarking process which is also refers to the visual quality of the watermarked face images, two kinds of experiments have been conducted which are subjective and objective evaluation experiments. For subjective evaluation, the resultant watermarked face images have been displayed to check if any artifacts have been generated in the watermarked face images. The results of subjective evaluation proved the ability to embed the detection and localization information in the RONI area without generating any artifacts in the resultant watermarked image.

Figure 6 presents samples of the resultant watermarked face images. For objective evaluation, the Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE) have been calculated. Table 1 illustrates the samples of objective evaluation results. The results proved the ability of the proposed scheme to generate high-quality watermarked face images which is good for the security purposes. The results in Table 1 proved that when the size of the face area is small in comparison to the size of the original face image, the visual quality of the watermarked image will be higher because fewer modifications are conducted in the RONI.

3.3. Capacity and payload test

The capacity and payload are different in each image and their values are dependent on the size of the original face image and the size of the face area. Table 2 presents the results of the capacity test for the test face images that have been shown in Fig. 4. The capacity is large when the size of the face area is small in comparison to the size of the original face image. The capacity is calculated by multiplying the total number of RONI blocks by 64 bits because each RONI block can carry 64 bits. The payload is calculated based on the total number of ROI blocks, bits of ‘extend 1’ process, and bits of ‘extend 2’ process. Table 3 illustrates the details of the obtained payload results. The results proved that the larger the number of ROI blocks, the higher the payload and vice versa.



Watermarked Image 1

Watermarked Image 2

Watermarked Image 3

Watermarked Image 4

Watermarked Image 5

Watermarked Image 6

Watermarked Image 7

Watermarked Image 8

Watermarked Image 9

Fig. 6. Samples of face detection and selection test results.

Table 1. Invisibility test results for different face images.

Image name	Image size	Size of face area	MSE	PSNR
Image 1	408×612×3	104×75×3	0.0275	+63.75 dB
Image 2	347×497×3	90×74×3	0.1503	+56.36 dB
Image 3	339×508×3	67×56×3	0.0079	+69.14 dB
Image 4	3054×4581×3	735×684×3	0.0743	+59.42 dB
Image 5	198×255×3	73×62×3	0.2085	+54.94 dB
Image 6	344×410×3	108×89×3	0.2298	+54.52 dB
Image 7	183×275×3	102×88×3	0.1381	+56.73 dB
Image 8	183×275×3	46×38×3	0.1173	+57.44 dB
Image 9	4446×2945×3	726×562×3	0.0100	+68.12 dB
Image 10	2832×4256×3	1368×1201×3	0.0408	+62.02 dB
Image 11	2003×3000×3	855×762×3	0.1245	+57.18 dB
Image 12	2973×4460×3	699×553×3	0.0104	+67.96 dB
Image 13	2000×3000×3	841×662×3	0.0183	+65.51 dB
Image 14	316×410×3	93×84×3	0.0421	+61.89 dB
Image 15	2975×4460×3	683×659×3	0.0573	+60.55 dB

Table 2. Capacity test results for different face images.

Image name	Image size	Size of face area	Total number of RONI blocks	Capacity (bits)
Image 1	408×612×3	104×75×3	910×3	174720
Image 2	347×497×3	90×74×3	609×3	116928
Image 3	339×508×3	67×56×3	631×3	121152

Image 4	3054×4581×3	735×684×3	52272×3	10036224
Image 5	198×255×3	73×62×3	155×3	29760
Image 6	344×410×3	108×89×3	483×3	92736
Image 7	183×275×3	102×88×3	145×3	27840
Image 8	183×275×3	46×38×3	175×3	33600
Image 9	4446×2945×3	726×562×3	49276×3	9460992
Image 10	2832×4256×3	1368×1201×3	40546×3	7784832
Image 11	2003×3000×3	855×762×3	20783×3	3990336
Image 12	2973×4460×3	699×553×3	49890×3	9578880
Image 13	2000×3000×3	841×662×3	21107×3	4052544
Image 14	316×410×3	93×84×3	426×3	81792
Image 15	2975×4460×3	683×659×3	49538×3	9511296

Table 3. Payload test results for different face images.

Image name	**Single channel results**				**Payload for single channel (bits)**	**Total payload (bits)**
	No. of ROI blocks	**No. of bits after extend 1**	**No. of bits after BCH**	**No. of bits after extend 2**		
Image 1	40	330	450	512	512	1536
Image 2	42	341	465	512	512	1536
Image 3	20	165	225	256	256	768
Image 4	2068	16555	22575	22592	22592	67776
Image 5	25	209	285	320	320	960
Image 6	42	341	465	512	512	1536
Image 7	42	341	465	512	512	1536
Image 8	12	99	135	192	192	576
Image 9	1692	13541	18465	18496	18496	55488
Image 10	6536	52294	71310	71360	71360	214080
Image 11	2592	20746	28290	28352	28352	85056
Image 12	1540	12331	16815	16832	16832	50496
Image 13	2268	18150	24750	24768	24768	74304
Image 14	49	396	540	576	576	1728
Image 15	1892	15147	20655	20672	20672	62016

3.4. Tamper localization test

To test the tamper localization performance of the proposed FIMD scheme, different face manipulation attacks have been imposed on the watermarked face images. The proposed scheme obtained perfect results in detecting and localizing the manipulated blocks in the face region. The accuracy of detection is 100 % and there is no false detection for all test images. Samples of the manipulation localization results are shown in Figs. 7 to 10 where different manipulation attacks have been imposed on the watermarked images such as attributes manipulation, retouching attack, expression swap, and entire face swap.

Figure 7 presents the result for ‘Image 1’ where attribute manipulation attack has been imposed on the watermarked image in which the colour of the eyes has been changed. Figure 8 presents the result for ‘Image 5’ where expression swap attack has been imposed on the watermarked image in which the mouth of ‘Image 2’ has been swapped with the mouth of ‘Image 5’. Figure 9 presents the result for ‘Image 1’ where face swap attack has been imposed on the watermarked image in which the face of ‘Image 1’ has been swapped with the face of ‘Image 9’. Figure 10 presents the result for ‘Image 7’ where face retouching attack has been imposed on the watermarked image 7 where the cheeks are brightened, and lipstick has been added.

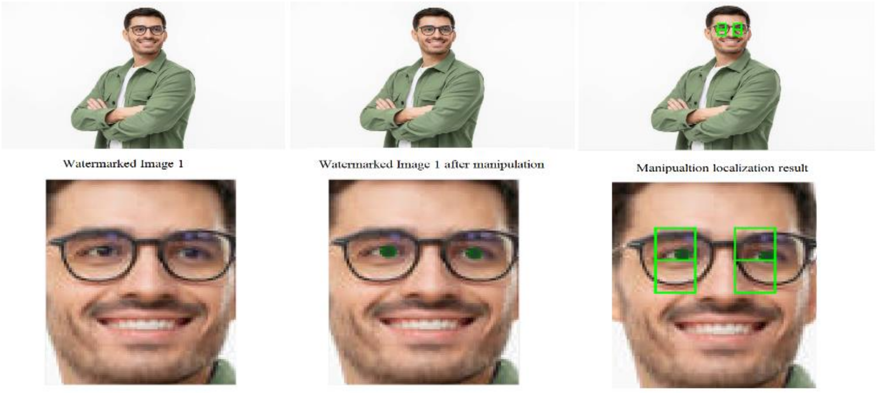

Watermarked Image 1
Watermarked Image 1 after manipulation
Manipualtion localization result

Fig. 7. Manipulation localization result for ‘Image 1’ (attributes attack).

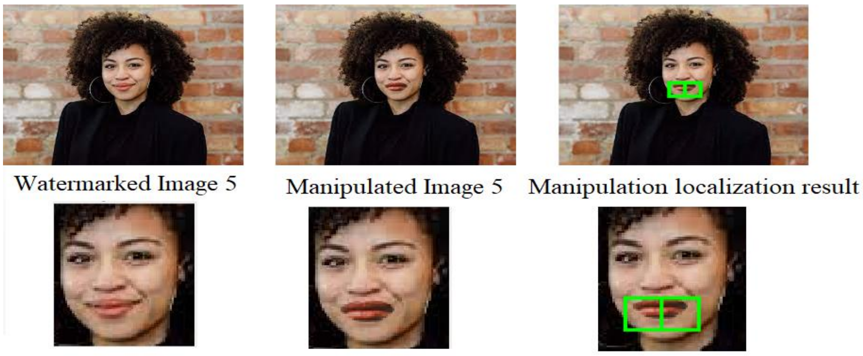


Fig. 8. Manipulation localization result for ‘Image 5’ (expression swap).

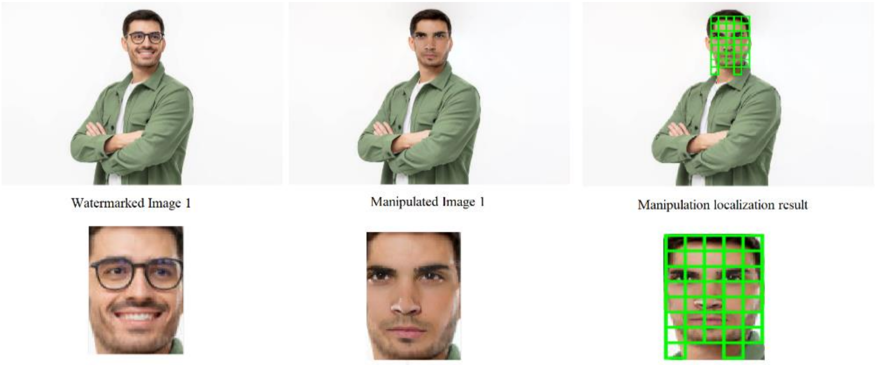
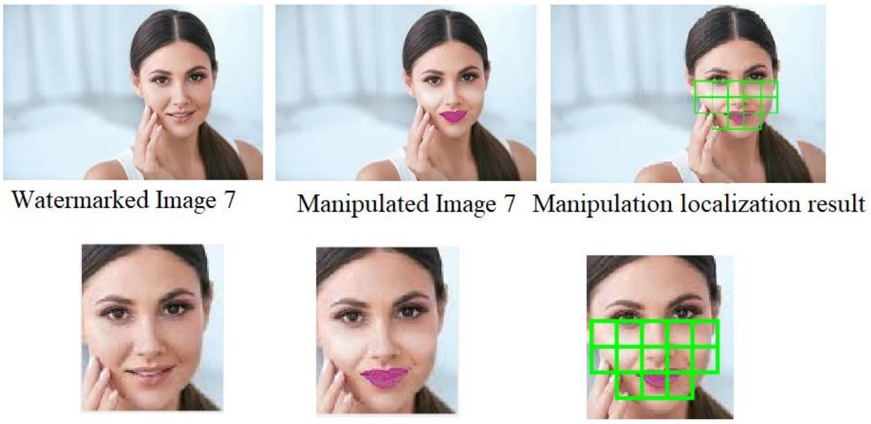


Fig. 9. Manipulation localization result for ‘Image 1’ (face swap attack).



Watermarked Image 7 Manipulated Image 7 Manipulation localization result

Fig. 10. Manipulation localization result for 'Image 7' (face retouching attack).

3.5. Comparison with state-of-the-art schemes

Since the idea of the proposed FIMD scheme is novel, it is not valid to conduct direct comparisons with the state-of-the-art manipulation detection schemes. Generally, the proposed scheme outperforms many face-image manipulation detection techniques. Table 4 presents a general comparison between the proposed FIMD scheme and the state-of-the-art schemes in [33-47] which proved the efficiency and superiority of the proposed scheme.

Table 4. General comparison with state-of-the-art detection schemes.

State-of-the-art detection schemes [33-47]	Proposed FIMD scheme
Most of the available schemes can detect only one type of manipulation because they depended in their implementation on the methods of creating fake or manipulated images [33-39].	The proposed scheme can detect different types of manipulation and it does not need to know the method of creating fake or manipulated images.
The deep-learning based schemes require large datasets for training and the detection performance will be good only when the input image is close to the training set [40-44].	No training is required in the proposed scheme.
Most techniques relied on supervised networks for detection which are time and efforts consuming [45-47].	The proposed scheme is completely automatic therefore it is better in terms of saving time and efforts.
False detection results have been recorded especially when the input image is different from the training dataset [41-44]. For instance, the maximum accuracy values in [41] to [44] are 84.7%, 99.3%, 87.5%, and 81.6%, respectively.	The detection accuracy is 100 % and there is no false detection.

4. Conclusions

The available face manipulation detection schemes have many limitations and the research in this field is very interesting. This paper presents a new manipulation detection scheme for digital face images based on face detection and image watermarking techniques. The proposed FIMD scheme starts by detecting the face box in the image using Multi-Task Cascaded CNN algorithm followed by some calculations to select the face area.

According to the selected face area, a black and white image called mask image is generated. The scheme has been implemented for colour images in order to be more suitable for the practical applications.

In the proposed watermarking stage, one channel is selected and divided into blocks, then the blocks are classified based on the mask image into two types that are ROI blocks and RONI blocks. The manipulation detection and localization data are generated from the ROI blocks and embedded in the RONI blocks. The watermarking procedure is repeated for the three channels in the colour face image. At the receiver side, the manipulation detection and localization data are extracted and used to check the authenticity of the received face image and to detect manipulation if available.

The experimental results that have been conducted to test the performance of the proposed FIMD scheme proved its efficiency in terms of the visual quality of the resultant watermarked face images, the high embedding capacity, and the accuracy of detecting and localizing manipulations.

The accuracy of detection is 100 % and no false detection results have been recorded. The scheme can detect different face image manipulations such as retouching, expression swap, attribute attacks, and face swap.

The general comparison with the state-of-the-art schemes proved the superiority of the proposed scheme. The work in this paper opens the door for a new direction of research in this thriving research field and it can be applied to ensure the intactness and safety of the digital face images in different practical applications such as online face recognition systems, online access control based on face recognition, and others.

References

1. Cote, J. (2022). Deepfakes and fake news pose a growing threat to democracy, experts warn. *Report in News at Northeast*. https://news.northeastern.edu/2022/04/01/deepfakes-fake-news-threat-democracy/.
2. Patel, L. (2020). The Rise of Deepfakes and What That Means for Identity Fraud. Retrieved December 26, 2022 from DarkReading Authentication: https://www.darkreading.com/authentication/the-rise-of-deepfakes-and-what-that-means-for-identity-fraud.
3. Citron, D. (2019). How DeepFake undermine truth and threaten democracy. Retrieved December 26, 2022 from TED official website: https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy.
4. Silva, S.H.; Bethany, M.; Votto, A.M.; Scarff, I.H.; Beebe, N.; and Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble

of weakly supervised models. *Forensic Science International: Synergy*, 4, 100217.

5. Zi, B.; Chang, M.; Chen, J.; Ma. X.; and Jiang, Y.-G. (2020). WildDeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the* 28*th ACM International Conference on Multimedia*, *NY*, *USA: Association for Computing Machinery*, *New York*, 2382-2390.
6. Gray, C. (2020). Add to Cart: Why deepfakes are good for retail. Retrieved December 26, 2022, from AdNews Newsletter. https://www.adnews.com.au/news/add-to-cart-why-deepfakes-are-good-for-retail#:~:text=For%20a%20 small%20retailer%20with,their%20very%20own%20custom%20model.
7. Kolagati, S.; Priyadharshini, T.; and Rajam, V.M.A. (2022). Exposing deepfakes using a deep multilayer perceptron - convolutional neural network model. *International Journal of Information Management Data Insights*, 2(1), 100054.
8. Wang, G.; Jiang, Q.; Jin, X.; and Cui, X. (2022). FFR_FD: Effective and fast detection of DeepFakes via feature point defects. *Information Sciences*, 596, 472-488.
9. Korus, P. (2017). Digital image integrity - a survey of protection and verification techniques. *Digital Signal Processing*, 71, 1-26.
10. Amerini, I.; Baldini, G.; and Leotta, F. (2021). Image and video forensics. *Journal of Imaging*, 7(11), 1-3.
11. Kumar, J.H.; and Devi, T.K. (2022). Fingerprinting of Image Files Based on Metadata and Statistical Analysis. *Proceedings of International Conference on Deep Learning*, *Computing and Intelligence*, vol 1396. Springer, Singapore, 105-118.
12. Berthet, A.; and Dugelay, J.-L. (2020). A review of data preprocessing modules in digital image forensics methods using deep learning. *Proceedings of the* 2020 *IEEE International Conference on Visual Communications and Image Processing* (*VCIP*), Macau, China, 281-284.
13. Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; and Verdoliva, L. (2021). ID-Reveal: Identity-aware DeepFake Video Detection. *Proceedings of the* 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, QC, Canada, 15088-15097.
14. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, Korea (South), 1-11.
15. Papastratis, I. (2020). Deepfakes: Face synthesis with GANs and Autoencoders. Retrieved December 26, 2022, from AI Summer: https://theaisummer.com/deepfakes/
16. Kowalski, M. (2021). FaceSwap. Retrieved December 26, 2022, from GetHub official website https://github.com/MarekKowalski/FaceSwap
17. Store, A. (2019). ZAO. Retrieved December 26, 2022, from Changsha Shenduronghe Network Technology Co., Ltd: https://apps.apple.com/cn/app/id1465199127.
18. Wolberg, G. (1998). Image morphing: a survey. *The Visual Computer*, 14(8), 360-372.

19. Gomez-Barrero, M.; Rathgeb, C.; Scherhag, U.; and Busch, C. (2017). Is your biometric system robust to morphing attacks?. *Proceedings of the* 2017 5*th International Workshop on Biometrics and Forensics* (*IWBF*), Coventry, UK, 1-6.
20. Venkatesh, S.; Ramachandra, R.; Raja, K.; and Busch, C. (2021). Face Morphing Attack Generation and Detection: A Comprehensive Survey. *IEEE Transactions on Technology and Society*, 2(3), 128-145.
21. Scherhag, U.; Rathgeb, C.; Merkle, J.; and Busch, C. (2020). Deep face representations for differential morphing attack detection. *IEEE Transactions on Information Forensics and Security*, 15, 3625-3639.
22. Zhang, H.; Venkatesh , S.; Ramachandra, R.; Raja , K.; Damer, N.; and Busch , C. (2021). MIPGAN - Generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Transactions on Biometrics*, *Behavior*, *and Identity Science*, 3(3), 365-383.
23. Agarwal, A.; Singh, R.; Vatsa, M.; and Noore, A. (2017). SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern. *Proceedings of the* 2017 *IEEE International Joint Conference on Biometrics* (*IJCB*), Denver, CO, USA, 659-665.
24. Snap, I. (2022). Snapchat. Retrieved December 26, 2022, from App. Store Preview: https://www.snapchat.com/.
25. He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. (2017). Arbitrary facial attribute editing: only change what you want. Retrieved December 26, 2022 from ArXiv: https://www.researchgate.net/publication/321374783_Arbitrary_Facial_Attribute_Editing_Only_Change_What_You_Want.
26. Gonzalez-Sosa, E.; Fierrez, J.; Vera-Rodriguez, R.; and Alonso-Fernandez, F. (2018). Facial soft biometrics for recognition in the wild: recent works, annotation, and COTS evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8), 2001-2014.
27. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 8789-8797.
28. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148.
29. Passos, L.A. et al. (2022). A review of deep learning-based approaches for deepfake content detection. Retrieved December 26, 2022, from ArXiv: http://arxiv.org/abs/2202.06095.
30. Pashine, S.; Mandiya, S.; Gupta, P.; and Sheikh, R. (2021). Deep fake detection: Survey of facial manipulation detection solutions. *International Research Journal of Engineering and Technology* (*IRJET*), 8(5), 4441-4449.
31. Almars, A.M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications*, 9(5), 20-35.
32. Ju, X. (2020). An Overview of face manipulation detection. *Journal of Cyber Security*, 2(4), 197-207.

33. Matern, F.; Riess, C.; and Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. *Proceedings of the* 2019 *IEEE Winter Applications of Computer Vision Workshops* (*WACVW*), Waikoloa, HI, USA, 83-92.
34. Hu, S.; Li, Y.; and Lyu, S. (2021). Exposing GAN-Generated Faces Using Inconsistent Corneal Specular Highlights. *Proceedings of the ICASSP* 2021 - 2021 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Toronto, ON, Canada, 2500-2504.
35. Han, X.; Ji, Z.; and Wang, W. (2020). Low Resolution Facial Manipulation Detection. *Proceedings of the* 2020 *IEEE International Conference on Visual Communications and Image Processing* (*VCIP*), Macau, China, 431-434.
36. Yang, X.; Li, Y.; Qi, H.; and Lyu, S. (2019). Exposing GAN-synthesized Faces Using Landmark Locations. *IH&MMSec'*19*: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 113-118.
37. McCloskey, S.; and Albright, M. (2019). Detecting GAN-Generated Imagery Using Saturation Cues. *Proceedings of the* 2019 *IEEE International Conference on Image Processing* (*ICIP*), Taipei, Taiwan, 4584-4588.
38. Li, H.; Li, B.; Tan, S.; and Huang, J. (2018). Detection of deep network generated images using disparities in colour components. Retrieved December 26, 2022, from ArXiv: https://arxiv.org/pdf/1808.07276.pdf.
39. Frank, J.; Eisenhofer, T.; Schonherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. *ICML'*20*: Proceedings of the* 37*th International Conference on Machine Learning*, PMLR, 119(304), 3247-3258.
40. Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5), 6259-6276.
41. Wang, R.; Juefei-Xu, F.; Ma, L.; Xei, X.; Huang, Y.; Wang, J.; and Liu, Y. (2019). FakeSpotter: A Simple baseline for spotting AI-synthesized fake faces. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (*IJCAI*-20), 3444-3451.
42. Marra, C.S.; Boato, G.; and Verdoliva, L. (2019). Incremental learning for the detection and classification of gan-generated images. *Proceedings of the* 2019 *IEEE International Workshop on Information Forensics and Security* (*WIFS*), Delft, Netherlands, 1-6.
43. Jung, T.; Kim, S.; and Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8, 83144-83154.
44. Amerini, I.; Galteri, L.; Caldelli, R.; and Bimbo, A.D. (2019). Deepfake video detection through optical flow based CNN. *Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, Korea (South), 1205-1207.
45. Kong, C.; Chen, B.; Li, H.; Wang, S.; Rocha, A.; and Kwong, S. (2021) Detect and locate: A face anti-manipulation approach with semantic and noise-level supervision. *ArXiv*, arXiv:2107.05821.
46. Cao, L.; Sheng, W.; Zhang, F.; Du, K.; Fu, C.; and Song, P. (2021). Face manipulation detection based on supervised multi-feature fusion attention network. *Sensors*, 21(24), 8181.

47. Bharati, A.; Singh, R.; Vatsa, M.; and Bowyer, K.W. (2016). Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9), 1903-1913.
48. Tolosana, R. et al. (2022). *Future trends in digital face manipulation and detection BT - Handbook of digital face manipulation and detection: From DeepFakes to morphing attacks*. C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, Eds. Cham: Springer International Publishing, 463-482.
49. Jain, A.; Singh, R.; and Vatsa, M. (2018). On Detecting GANs and Retouching based Synthetic Alterations. *Proceedings of the* 2018 *IEEE* 9*th International Conference on Biometrics Theory, Applications and Systems* (*BTAS*), Redondo Beach, CA, USA, 1-7.
50. Majumdar, P.; Agarwal, A.; Vatsa, M.; and Singh, R. (2022). *Facial retouching and alteration detection BT - Handbook of digital face manipulation and detection: from deepfakes to morphing attacks*. C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, Eds. Cham: Springer International Publishing, 367-387.
51. Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing letters*, 23(10), 1499-1503.
52. Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. (2016). *A discriminative feature learning approach for deep face recognition*. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, 9911. Springer, Cham.
53. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; and Wu, X. (2019). Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212-3232.
54. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; and Zisserman, A. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. *Proceedings of the* 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, China, 67-74.
55. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; and Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128, 261-318.
56. Khan, A.; Sohail, A.; Zahoora, U.; and Qureshi, A.S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8), 5455-5516.
57. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; and Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *Proceedings of the* 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, China, 59-66.
58. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. (2018). CosFace: Large margin cosine loss for deep face recognition. *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 5265-5274.
59. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. (2017). SphereFace: Deep Hypersphere Embedding for Face Recognition. *Proceedings of the* 2017

IEEE Conference on Computer Vision and Pattern Recognition (*CVPR*), Honolulu, HI, USA, 6738-6746.

60. Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. (2019). ArcFace: additive angular margin loss for deep face recognition. *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 4685-4694.
61. Alshanbari, H.S. (2021). Medical image watermarking for ownership & tamper detection. *Multimedia Tools and Applications*, 80, 16549-16564.
62. Su, G.-D.; Chang, C.-C.; and Lin, C.-C. (2020). Effective Self-Recovery and Tampering Localization Fragile Watermarking for Medical Images. *IEEE Access*, 8, 160840-160857.
63. Allaf, A.H., Kbir, M.A. (2019). *A review of digital watermarking applications for medical image exchange security*. In: Ben Ahmed, M., Boudhir, A., Younes, A. (eds) Innovations in Smart Cities Applications Edition 2. SCA 2018. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham.
64. Sinha, S.; Singh, A.; Gupta, R.; and Singh, S. (2018). Authentication and tamper detection in tele-medicine using zero watermarking. *Procedia Computer Science*, 132, 557-562.
65. Thabit, R.; and Khoo, B.E. (2015). A new robust lossless data hiding scheme and its application to colour medical images. *Digital Signal Processing*, 38, 77-94.
66. Mohammed, R.T.; and Khoo, B.E. (2013). Robust reversible watermarking scheme based on wavelet-like transform. *Proceedings of the* 2013 *IEEE International Conference on Signal and Image Processing Applications*, Melaka, Malaysia, 354-359.
67. Thabit, R.; and Khoo, B.E. (2014). Robust reversible watermarking scheme using Slantlet transform matrix. *Journal of Systems and Software*, 88, 74-86.
68. Shehab, A.; Elhoseny, M.; Muhammad, K.; Sangaiah, A.K.; Yang, P.; Huang, H.; and Hou, G. (2018). Secure and robust fragile watermarking scheme for medical images. *IEEE Access*, 6, 10269-10278.
69. Rahman, A.U. et al. (2018). Robust and fragile medical image watermarking: a joint venture of coding and chaos theories. *Journal of Healthcare Engineering*, Volume 2018, Article ID 8137436.
70. Thabit, R. (2021). Review of medical image authentication techniques and their recent trends. *Multimedia Tools and Applications*, 80(9), 13439-13473.
71. Thabit, R.; and Khoo, B.E. (2017). Medical image authentication using SLT and IWT schemes. *Multimedia Tools and Applications*, 76(1), 309-332.
72. Bainbridge, W.A.; Isola, P.; and Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General*, 142(4), 1323-1334.
73. iMerit, P. (2021). 5 Million faces - Top 14 free image datasets for facial recognition. Retrieved December 26, 2022, from https://imerit.net/blog/5-million-faces-top-14-free-image-datasets-for-facial-recognition-all-pbm/