FAMM: Facial Muscle Motions for Detecting Compressed Deepfake Videos over Social Networks

Xin Liao*, Yumei Wang, Tianyi Wang*, Juan Hu, Xiaoshuai Wu

Abstract—**As a face manipulation technique, the misuse of Deepfakes poses potential threats to the state, society, and individuals. Several countermeasures have been proposed to reduce the negative effects produced by Deepfakes. Current detection methods achieve satisfactory performance in dealing with uncompressed videos. However, videos are generally compressed when spread over social networks because of limited bandwidth and storage space, which generates compression artifacts and the detection performance inevitably decreases. Hence, how to effectively identify compressed Deepfake videos over social networks becomes a significant problem in video forensics. In this paper, we propose a facial-muscle-motions-based (FAMM) framework to solve the problem of compressed Deepfake video detection. Specifically, we first locate faces from consecutive frames and extract landmarks from the face images. Then, continuous facial landmarks are utilized to construct facial muscle motion features by modeling the five sensory and face regions. Finally, we fuse the diverse forensic knowledge using Dempster-Shafer theory and provide the final detection results. Furthermore, we demonstrate the effectiveness of FAMM through analyzing mutual information, compression procedure, and facial landmarks for compressed Deepfake videos. Theoretical analyses illustrate that compression does not affect facial muscle motion feature construction and the differences in designed features exist between the real and Deepfake videos. Extensive experimental results conclude that the proposed method outperforms the state-of-the-art methods in detecting compressed Deepfake videos. More importantly, FAMM achieves comparable detection performance on compressed videos that are over real-world social networks.**

Index Terms—**Multimedia forensics, compressed Deepfake videos, facial muscle movements, social networks.**

I. Introduction

DURING the mobile internet period, short video applications such as Youtube [1], Pear Video, Instagram [2], and TikTok have flourished, which increases entertainment options for people and makes it more convenient for them to share their lives by videos. However, when forged videos containing misleading contents are spread over social networks, they will pollute the information ecosystem.

Deepfakes [3] is the product of technological advances in artificial intelligence (AI) and stems from a combination of "deep learning" and "fake". It superimposes the target person's face onto the corresponding position of the original person's face to create a forgery video containing the target person [4]. With the development of artificial intelligence technology, Deepfake techniques have been rapidly improved and most algorithms are open. These techniques have a lot of positive application scenarios. For example, Deepfake techniques have been applied in many industries including film, education, and health care [3]. However, if these techniques are abused, they will severely threaten the credibility and security of countries, society, and individuals. As shown in Fig. 1, the fake video synthesized by Deepfake techniques is indistinguishable to eyes, with the potential to create a crisis of public trust or cause a mist of war when distributed on the Internet [5]. In addition, the spread and dissemination of false political information can lead to political outcry and affect the security of the state and individuals.

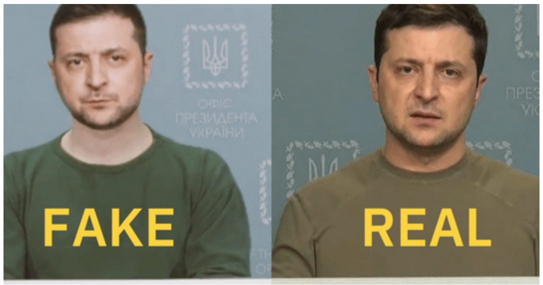


Fig. 1. Illustration of a pristine image (right) and its forgery (left) by using Deepfakes. The fake image is derived from Deepfake video that was spread on social media platforms of Volodymyr Zelensky announcing his surrender.

To cope with potential threats of Deepfake techniques and mitigate the spread of false information, several Deepfake detection approaches [6]–[30] have been developed. These methods have achieved excellent performance in detecting uncompressed videos, which can be roughly divided into two types, i.e., handcraft-based and deep-learning-based methods. The handcraft-based methods [6]–[12] identify forged videos by diagnosing the forgery process and mining heuristic features. With the forgery techniques developing, there are no conspicuous tampering traces in the generated Deepfake videos, which might lead to the progressive ineffectiveness of these methods. Furthermore, deep-learning-based methods [13]–[30] were proposed, which required large-scale real and fake data to drive the models to extract distinguishable features.

In social networks, limited by the bandwidth and storage space, videos are usually compressed when uploading. For

This work is supported by National Natural Science Foundation of China (Grant Nos. U22A2030, 61972142), National Key R&D Program of China (Grant No. 2022YFB3103500).

Xin Liao, Yumei Wang, Juan Hu and Xiaoshuai Wu are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: xinliao@hnu.edu.cn, yumeiwang@hnu.edu.cn, hujuan@hnu.edu.cn, shinewu@hnu.edu.cn).

Tianyi Wang is with the Department of Computer Science, The University of Hong Kong, China (e-mail: tywang@cs.hku.hk).

*Corresponding author: Xin Liao, Tianyi Wang.

instance, the average resolution of the videos has dropped by 0.8 and 0.64 after being transmitted through Facebook and Youtube, respectively [31]. Meanwhile, down-sampling and quantization noise are introduced in the video compression process to reduce temporal and spatial redundancy, which results in compression artifacts such as blocking, blurring, ringing, and jaggies [32] in the compressed videos, making it more challenging for the detection model to extract distinctive features. When the videos are the compressed ones transmitted through social networks, they would mislead the neural network learning, resulting in inferior detection performance. The methods in [28]–[30] investigated Deepfake videos and considered compression scenes. However, they attempted to detect compressed Deepfake videos by using Convolutional Neural Network (CNN) and the detection performance of the models would be weakened by co-existing compression artifacts. Furthermore, they focused on videos that are hard-coded using the H.264 or H.265 codec instead of the ones that are actually shared through current and active social networks. Hence, effectively detecting compressed Deepfake videos and facilitating a deeper understanding of detection methods are the essential issues that remain to be resolved in video forensics.

This paper proposes a method FAMM that detects compressed Deepfake videos by analyzing facial muscle motion features. When people talk or make micro-expressions, the curvature of the facial skin undulates and causes the displacement of facial feature points [33]. Agarwal et al. [6] recognized Deepfake videos by tracking facial expressions and movements to protect world leaders. Specifically, they extracted the intensity of 16 action units to construct features and calculated the Pearson correlation coefficient between these features to characterize the subconscious mind of different characters. Unlike this approach, we analyze Deepfake videos and find that they are distorted in the facial area during the synthesis process, resulting in unnatural muscle movements. We describe facial muscle movements and construct facial muscle motion features from a geometric perspective. Specifically, the lengths and angles of feature vectors that are constructed by facial landmarks jointly represent the facial muscle motions in our paper. FAMM focuses on mining unnatural facial muscle motions by modeling the facial muscle features in the temporal dimension. To capture unnatural facial muscle movements, the feature difference between facial muscle movements of the current frame and the succeeding one is used to catch temporal relationships. In order to minimize the effect of noise for temporal sequence, we construct time-series features by calculating the absolute energy, absolute sum of first-order differences, time-series complexity, kurtosis, and coefficient of variation of the facial muscle motion features. The difference and time-series features are utilized to train two different classifiers and the results are fused using Dempster-Shafer theory. The fusion strategy incorporates multiple forensic knowledge, which provides a more comprehensive insight into the facial muscle motions, resulting in improved compressed Deepfake video detection performance. The main contributions of this paper are as follows:

1) We construct facial muscle motion features from a geometric perspective, which brings a distinct viewpoint for Deepfake detection. Different from previous methods that focus on facial tampering artifacts, we formulate geometric features by utilizing facial landmarks to represent unnatural facial muscle motions due to temporal discontinuity of Deepfake videos.
2) We analyze the video compression process, mutual information, and facial landmarks, which theoretically illustrate the interpretability of FAMM. The intra-frame, inter-frame, and mutual information analyses indicate that facial landmarks are invariant before and after video compression. Furthermore, facial landmark analysis of the whole and central faces indicates the existence of unnatural facial muscle movements in fake videos.
3) We conduct extensive experiments to verify the detection performance on compressed Deepfake videos and the results demonstrate that FAMM achieves promising performance compared to the state-of-the-art methods. More importantly, FAMM performs well on the videos that are over real-world social networks such as Facebook and Youtube.

The rest of the paper is organized as follows. In Section II, we illustrate the related works of Deepfake video generation and detection techniques. In Section III, we perform several statistical analyses to show that compression might affect the quality of Deepfake videos. In Section IV, we present the details about the proposed method. In Section V, we give the mutual information, compression procedure, and facial landmark analyses to show the interpretability of FAMM. In Section VI, we provide extensive experimental results to demonstrate the performance of our method. Finally, the conclusion is drawn in VII.

II. Related Work

Continuous improvements in Deepfake techniques make synthesized images/videos more realistic and bring new challenges to Deepfake video forensics. In this section, we will introduce Deepfake generation techniques and Deepfake detection methods, respectively.

A. Deepfake Generation Techniques

Currently, Deepfake generation techniques can be divided into two categories in accordance with the method of manipulation, i.e., facial replacement and facial attribute editing. Facial replacement mainly replaces the facial region of the target person with the face of original face, which involves a modification of the identity attributes of the target image. The typical representative of face replacement, FaceSwap [34], uses classical 3D graphics techniques [35]. Nevertheless, their complexity and high technical costs make them difficult to be popularized. With the rapid development of deep learning techniques, several researchers apply deep learning to face generation. “Deepfake” [36] is the first representative algorithm to apply deep learning technology to facial swapping tasks, which makes it intelligent and highly realistic. As the understanding of deep forgery continues, more methods are proposed to improve face replacement. For example, FaceShifter [37] is a method that takes face replacement to a new level.

Attribute editing focuses on editing various attributes outside of the target person's face identity information, such as expression migration [38], facial reenactment, etc. Methods such as Face2Face, NeuralTexture, and Head2Head [39]–[41] migrate expressions by exchanging the expression parameters of the original face and the target face. Most of the above methods are already available on the open-source website, which lowers the threshold for the use of deep generation techniques. When these methods are used indiscriminately and maliciously, there are numerous pitfalls.

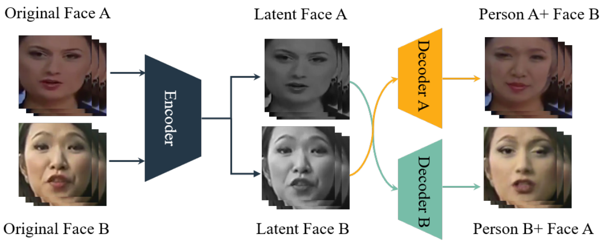


Fig. 2. Illustration of the process for generating fake videos. The figure shows that Deepfake is essentially a face swap of the video frames and then the video frames are composited into the video whose temporal relationships are ignored.

B. Deepfake Videos Forensics

The spread of Deepfake videos accelerates the proliferation of false information, which may damage the reputation of countries and individuals. Therefore, a large number of research works have been proposed to detect Deepfake videos, which include two types, i.e., handcraft-based and deep-learning-based methods.

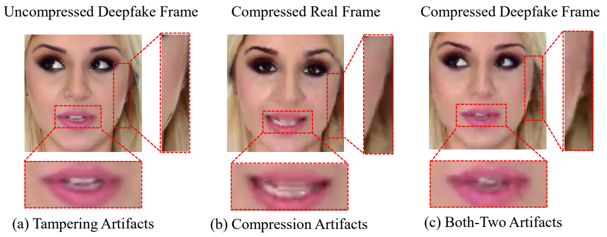
Handcraft-based methods exposed observable features to detect Deepfake videos. Shruti et al. [6] assumed the presence of different expressions and movements when individuals speak. They detected forged videos by tracking facial and head movements to extract the intensity of specific action units. Yang et al. [7] presented a detection method based on 3D head pose estimation by observing errors introduced in the face synthesis process. A novel perspective on detecting forgery videos was adopted [8], which exposed spatial and temporal inconsistencies in biological signals. Li et al. [9] detected Deepfake videos by exposing blinking frequency. A detection method based on visual artifacts was proposed to distinguish Deepfake videos by exposing the texture structure of the face and teeth [10]. Hosler et al. [11] found that there are unnatural emotions in Deepfake videos and proposed a method to detect Deepfake videos using the speaker's faces and voices to predict emotions. LRNet [12] was utilized to detect Deepfake videos through temporal modeling on landmarks.

In recent years, with the rise of deep learning techniques, many researchers proposed deep-learning-based methods to detect Deepfake videos. Due to the presence of unique distortion artifacts in fake videos, FWA [13] compared the generated face areas and their surrounding regions, and a dedicated CNN model was utilized to capture the distinctive artifacts for Deepfake video detection. Afchar et al. [14] proposed a network, namely Mesonet, with few parameters to detect Deepfake videos, which focused on the mesopic features of images. VGG-19 was introduced to extract image features that were then used as input to a capsule network to detect forged videos [15]. Sun et al. [16] first attempted to introduce self-information metrics and designed a novel attention mechanism based on self-information metrics, which can highlight manipulated regions. A multi-scale self-texturing network [17] was presented to track potential texture traces during image generation to detect Deepfake videos from the perspective of image generation, simulating the forgery process based on image generation. RECCE [18] is a forgery detection framework based on reconstruction-classification learning, where the reconstruction learning enhances the learning representation and the classification learning mines the essential differences between real and forged images to facilitate the understanding of forged images. Nevertheless, the single-frame approaches captured the local features of the images adequately but ignored the temporal information. Güera et al. [19], and Sabir et al. [20] proposed the time-aware Deepfake video detection methods using CNN to extract frame-level video features. The optical flow field was adopted to detect Deepfake videos [21]. Masi et al. [22] extracted a multi-domain fusion approach, which fused the spatial, temporal, and frequency-domain information to achieve detection. Deeprhythm [23] monitored the heartbeat rhythms to expose Deepfakes. Based on the analysis of motion, Haliassos et al. [24] realized the detection of Deepfake videos by using the rich semantic features learned from lip reading to represent the high level of inconsistency in semantics to distinguish natural and abnormal oral movements. Wang et al. [25] proposed a novel dual-branch complementary dynamic interaction network to detect Deepfake videos using global and local anomalous dynamic artifacts. A new multi-rate excitation network [26] was utilized to detect Deepfake videos, which can efficiently excite dynamic spatiotemporal inconsistencies from a multi-rate perspective. Yang et al. [27] proposed a masked relation learning framework to address the problem of large redundancy in relational information, which used a masked relation learner to aggregate and propagate relational information to capture irregularities for Deepfake video detection.

In fact, compressed videos are more common in social networks. The above methods suffer from performance degradation when detecting compressed videos because of compression artifacts and noise. FT-two-stream [28] was a two-stream network that analyzed the frame-level and temporality-level features to detect compressed Deepfake videos. In the meantime, they introduced a pruning algorithm to prevent the model from fitting the compression noise. Since the up-sampling operation is common in the forgery process and usually leaves detectable artifacts in the phase spectrum, SPSL [29] used the phase spectroscopy approach to detect forged videos. In addition, they dropped many convolutional layers to reduce the receptive field to focus on local regions that lack high-level semantic information. ADD [30] applied frequency-domain learning and optimal transfer theory to knowledge distillation, improving the detection performance

TABLE I
MEASURING NOISE LEVEL OF HARD-CODED VIDEOS (FACEFORENSICS++ LQ) AND COMPRESSED VIDEOS IN REAL SCENES (FACEFORENSICS++ SOCIAL) USING PSNR, SSIM, VIF, RECO.

Index	PSNR	SSIM	VIF	RECO
FaceForensics++ LQ	32.60	0.84	0.48	0.85
FaceForensics++ Social	37.95	0.92	0.56	1.16



(a) Tampering Artifacts (b) Compression Artifacts (c) Both-Two Artifacts

Fig. 3. Analyses of compression and tampering artifacts. The uncompressed Deepfake frame (a) generated by Face2Face leaves visible tampering artifacts. The compressed real frame (b) leaves compression artifacts and the compressed Deepfake frame (c) generated by Face2Face leaves both compression and tampering artifacts.

of deepfake images. However, these methods benefit from the powerful ability of deep learning, which is not conducive to an intensive understanding of detection methods, and the compressed detection performance of the model is necessary to be improved. Furthermore, they only concentrate on compressed videos with hard-coded implementations, while compressed videos in realistic scenarios are ignored.

III. THE ANALYSIS OF COMPRESSED DEEPFAKE VIDEOS

In this section, we first demonstrate the motivation of designing facial muscle movement features from a geometric perspective by analyzing compressed Deepfake videos. Then, we describe the synthesis process of Deepfake videos to illustrate the presence of unnatural facial movements.

A. Motivation

Compressed videos are widely used in social networks due to the expensive cost of storage and transmission. Compression coding mainly exploits a large amount of redundant data in videos, such as spatial redundancy, temporal redundancy, and visual redundancy. The size and visual quality of the videos are affected by compression, and compression artifacts are generated in these videos. We use Peak Signal to Noise Ratio (PSNR) [42], Structural Similarity (SSIM) [43], Visual Information Fidelity (VIF) [44], and Relative Polar Edge Coherence (RECO) [45] to measure the video quality. PSNR is one of the most commonly used objective criteria for evaluating images and a larger value of PSNR usually means less distortion. SSIM mainly considers the three key features of image brightness, contrast, and structure. A value closer to 1 indicates a stronger perceived similarity between the two images. VIF is an image quality index based on information theory and more significant value demonstrates the better quality of the fused image. RECO can be used to measure the quality of videos. As shown in Table I, the values of PSNR and SSIM between the uncompressed videos and hard-coded compressed ones are 32.60 dB and 0.84, and the values of PSNR and SSIM between the uncompressed videos and the ones of social networks are 37.95 dB and 0.92, which means that compression brings noise information. The values of VIF and RECO indicate that there exists noise in the vision and edge.

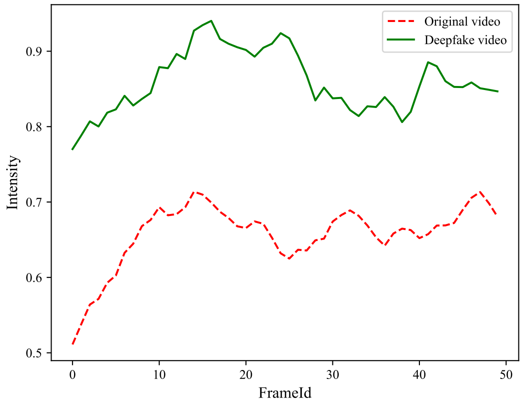


Fig. 4. The statistical intensity of action unit (AU07) is tracked in 100 original videos and 100 Deepfake videos. AU can represent facial muscle movements and there is a significant contrast of the intensity between the real and fake videos.

In addition, we demonstrate some intuitive results. As shown in Fig. 3(a), Deepfake videos compositing process distorts face area and the tampering artifacts are observed. Besides, compression artifacts (Fig. 3(b)) that are similar to tampering artifacts exist in the compressed real videos. More importantly, as shown in Fig. 3(c), the compression artifacts and tampering artifacts co-exist in the compressed Deepfake frames, which makes it challenging to extract distinguishable features from compressed Deepfake videos.

B. The Analysis of Unnatural Facial Movements

The synthesis process of Deepfake videos can be divided into training and testing phases. In the training phase, the encoders are used to learn paired face features, and in the decoding process, the different decoders are used to decode the faces. The testing phase is shown in Fig. 2, such that the synthetic faces are obtained by decoding the latent faces generated by face B (A) using the decoder of face A (B). These faces are utilized to generate Deepfake videos frame by frame and these videos have poor fitting in the temporal dimension. Therefore, unnatural expressions presented in Deepfake videos lead to abnormal facial muscle movements. The movements of facial muscles can be coded using facial action units (AU) [46]. Thus, we select 100 original videos and 100 corresponding fake ones from FaceForensics++ LQ [47] and track the intensity of the action units by utilizing the open-source facial behavior analysis toolkit Openface2 [46]. We normalize the intensity of the lid tightener (AU07) and calculate the corresponding statistical values. The result is

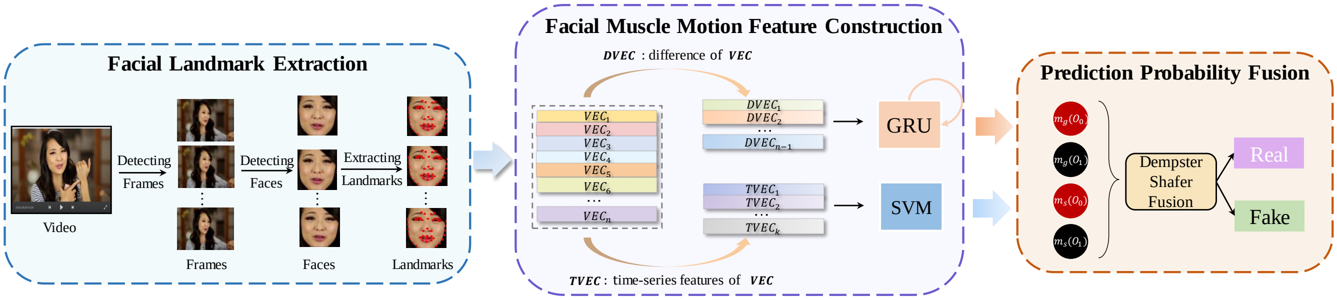


Fig. 5. An overall architecture of FAMM. The facial landmark extraction module obtains the landmarks from videos. The facial muscle motion feature construction module generates the geometric features that input the classifiers. The prediction probability fusion module utilizes Dempster-Shafer theory to integrate the outputs.

shown in Fig. 4, which illustrates that the intensity of real videos is significantly different from those of fake videos. Furthermore, we can observe that the curve of real videos is smoother than that of fake videos. In this paper, considering the robustness of geometric features, we devise the facial muscle movement features from a geometric perspective for the compressed Deepfake video detection case.

IV. FAMM: Facial-Mulscle-Motions-based Detection Framework

In this section, we first provide an overview of FAMM. Then, the details of the facial landmark extraction, facial muscle motion feature construction, and prediction probability fusion are introduced, respectively.

A. Overview of the proposed FAMM

The primary purpose of this work is to detect Deepfake videos over social networks. Fig. 5 illustrates the proposed facial-muscle-motions-based framework. FAMM consists of three modules: facial landmark extraction, facial muscle motion feature construction, and prediction probability fusion. First, we extract continuous frames from the videos and detect faces from frames. Since the precision of the landmarks is crucial for feature extraction, face alignment is performed before retrieving the 68-point facial landmarks. Second, we select the regions of interest (ROI) to construct facial muscle motion features. In order to effectively capture unnatural facial motions, we conduct feature transformation from two different perspectives. On the one hand, the relationship between adjacent frames is represented by changes in distance and angle features to expose unnatural facial muscle motions. The difference features between facial muscle motion features of the current and succeeding frames are utilized to train Gate Recurrent Unit (GRU), and then the predicted probability values of the videos are obtained. On the other hand, The presence of compression noise in compressed videos makes it more difficult to model the temporal dimension. Therefore, the time-series features of the facial muscle motion features are calculated to enlarge the unnatural facial muscle motions. Finally, Dempster-Shafer theory is utilized to fuse the outputs of GRU and SVM classifiers. The difference features are utilized to train GRU and the motion information of facial muscle motion features could be captured. As for SVM, the temporal information is obtained by learning the time-series features.

B. Facial Landmark Extraction

The facial landmark extraction module is utilized to extract facial landmarks from videos for feature construction and it consists of four main steps: frame extraction, face detection, face alignment, and landmark extraction. Face detection and face alignment are utilized to extract precise landmarks from videos in this module. In the video detection task, the primary means of determining the authenticity of a video is identifying the real or fake frames. Since FAMM focuses on temporal relationships, it is necessary to process the videos as continuous frames. We focus on the face regions because the tampered part of the Deepfake videos is the face area. The classical face detector, `Dlib` [48], is applied to extract the faces from consecutive frames. Since the designed facial muscle motion features rely on facial landmarks, we need to retrieve landmarks from the detected faces. Moreover, we adopt the face alignment operation in the face detection procedure, aiming at extracting the facial landmarks from the aligned face images. We utilize the human eyes for face alignment. The center points of the left and right eyes are calculated separately and the angle between the central coordinate line and horizontal direction is computed. Then, the affine matrix is calculated and the affine transformation is performed to align the faces. The face alignment process improves the accuracy of extracting landmarks, which facilitates the computation of subsequent features. Finally, the continuous facial landmarks $\boldsymbol{L}$ are extracted.

C. Facial Muscle Motion Feature Construction

Talking or making expressions can cause facial movements, especially in five sensory areas. Traditional methods usually encode the facial action units (AU) and then extract the signal to represent the movement of facial muscles. However, these methods are susceptible to compression. In fact, when people make expressions or speak, they will cause facial muscle movements. When the facial muscles move, the feature points of the face will follow the facial muscles. In this paper, the process of movements generated by talking or making

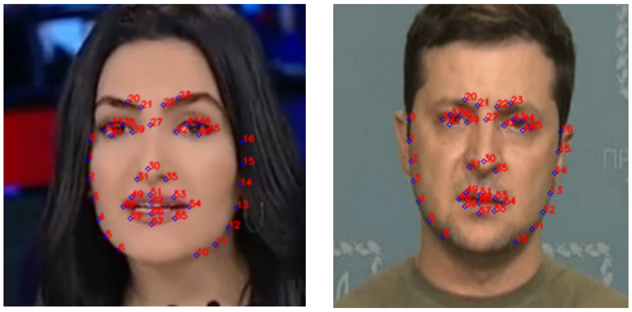
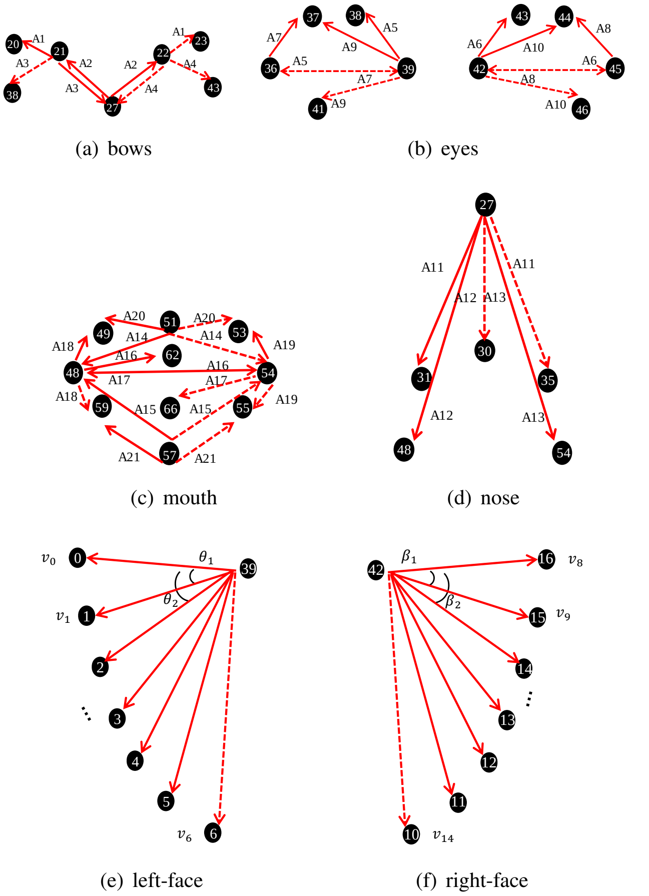


Fig. 6. Illustration of the selected feature points. The 42 feature points were extracted to construct ROI regions, i.e., five sensory regions (brows, eyes, mouth, and nose) and face regions (left-face and right-face).

expressions is defined as the movements of facial muscles and we construct facial muscle motion features from a geometric perspective to improve the robustness of our method for detecting compressed videos over social networks. The five sensory areas and the face areas are selected as ROI (as Fig. 6 shows). Then, the facial landmarks of ROI are selected and the facial muscle motion features are constructed. For modeling of the five sensory areas, inspired by [49], the brows, eyes, nose, and mouth regions are selected to construct geometric facial muscle motion features. The distance and angle features are calculated separately, which characterize the lengths and directions of the feature vectors, respectively. The feature vectors designed for the five sensory regions are shown in Fig. 7, where (a), (b), (c), and (d) are the features designed for the brows, eyes, mouth, and nose regions, respectively. Vectors with the same number construct an angle. Meanwhile, we calculate the lengths of the solid line vectors (as Fig. 6 shows (A2)) to characterize the distance features. For brow area, it is primarily related to the movement between the brows and eyes. The eye area focuses on eye movement processes, including the eyes opening and closing movements. The designed features of the left eye are symmetrical to those of the right eye and their axis of symmetry is a line perpendicular to the line formed by landmark 39 and landmark 42. In Fig. 6(b), for the left eye, taking landmark 39 at the corner of the eye as an example, it constructs a feature vector (A5) with landmark 38 inside the eyebrow and we calculate the length of the feature vector. At the same time, a new feature vector (A5) is constructed by landmark 39 and landmark 36 at the end of the eye and the angle between the two feature vectors is calculated. For the right eye, the features are constructed in the same way as the left eye and we take landmark 42 at the corner of the right eye as an example. We construct a feature vector (A6) with landmark 43 inside the eyebrow and calculate the length of the feature vector. Similarly, landmark 42 and landmark 45 at the end of the right eye construct a feature vector (A6) and calculate the angle between the two feature vectors. The rest of the features are constructed in the same way as the features above. For nose area, we can represent the movements by the variation at the ends of the nose. When a person talks or makes a facial expression, the movements around the lip are more pronounced. We use the lip shape and the process of opening and closing the lip to describe the movements of the muscles surrounding the lip.



(a) bows

(b) eyes

(c) mouth

(d) nose

(e) left-face

(f) right-face

Fig. 7. Schematic diagram of the features of the five sensory areas and face regions design. For five sensory areas, where those with the same number indicate that the two vectors form a vector angle, and the vector of solid lines is chosen to calculate its length. For the face regions, this is illustrated by the left-face region. The vector v_0 is used as the base vector with other feature vectors to construct the features. The right-face region features are constructed in the same way as the left face region. (a) bows, (b) eyes, (c) mouth, (d) nose, (e) left-face, (f) right-face.

The construction of feature vectors for the face regions, as shown in Fig. 7(e) and Fig. 7(f), is different from that of the five sensory areas. For the left-face region, the feature vector v_0 constructed from landmarks 0 and 39 is selected as the base vector to construct angles with the other vectors and lengths of the first 6 feature vectors are calculated. Following the same procedure, the vector of landmarks 16 and 42 in the right-face region is adopted as the base vector to compose the angles with other vectors. Therefore, the facial muscle motion features can be computed by

$$\boldsymbol{VEC} = \begin{bmatrix} l_{1,1} & \theta_{1,1} & l_{1,2} & \theta_{1,2} & \cdots & l_{1,m} & \theta_{1,m} \\ l_{2,1} & \theta_{2,1} & l_{2,2} & \theta_{2,2} & \cdots & l_{2,m} & \theta_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{n,1} & \theta_{n,1} & l_{n,2} & \theta_{n,2} & \cdots & l_{n,m} & \theta_{n,m} \end{bmatrix}, \quad (1)$$

where $m = 33$ is the number of the designed distance or angle features, which is the sum of the number of features in the five sensory and face regions (as Fig. 7 shows). In this paper, we design distance features and angle features to jointly represent

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3278310

facial muscle movements. Therefore, for a certain frame of the video, the feature dimension is 66. $l_{i,j}$ is the j-th distance feature of the i-th frame and $\theta_{i,j}$ is the j-th angle feature of the i-th frame. n is the number of input frames of a video and it might influence the modeling of the temporal dimension. We also analyze the impact of n in Section V.

Facial muscle motions are activities with continuity between adjacent frames. Deepfake videos are synthesized frame by frame while ignoring temporal coherence in the synthesis process. Therefore, temporal continuity is disrupted due to the shortcomings of synthetic technologies, resulting in Deepfake videos with unnatural facial muscle motions. To capture unnatural facial motions, we subtract the facial muscle motion features of the current frame from those of the next frame, which can be computed by

$$\boldsymbol{DVEC} = \begin{bmatrix} \Delta l_{1,1} & \Delta\theta_{1,1} & \Delta l_{1,2} & \Delta\theta_{1,2} & \cdots & \Delta l_{1,m} & \Delta\theta_{1,m} \\ \Delta l_{2,1} & \Delta\theta_{2,1} & \Delta l_{2,2} & \Delta\theta_{2,2} & \cdots & \Delta l_{2,m} & \Delta\theta_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta l_{k,1} & \Delta\theta_{k,1} & \Delta l_{k,2} & \Delta\theta_{k,2} & \cdots & \Delta l_{k,m} & \Delta\theta_{k,m} \end{bmatrix}, \quad (2)$$

where $k = n - 1$, $\Delta l_{i,j} = l_{i+1,j} - l_{i,j}$ and $\Delta\theta_{i,j} = \theta_{i+1,j} - \theta_{i,j}$, for $1 \leq i < n$ and $1 \leq j \leq m$. $\boldsymbol{DVEC}$ indicates the relationship between adjacent frames and is used to represent the movement patterns of facial muscle motion features.

Unfortunately, noise is introduced into videos during compression, which could interfere with the modeling of the temporal dimension. Therefore, we enhance unnatural facial muscle motions by utilizing discontinuity in the temporal sequence of facial muscle features. The input frame sequences are regarded as a coherent whole. For the j-th column of the facial muscle motion features, it can be denoted as $\boldsymbol{VEC}_{:,j} = [VEC_{0,j}, VEC_{1,j}, ..., VEC_{n,j}]$ and $VEC_{i,j}$ is the j-th feature of the i-th frame. In particular, we calculate five time-series features, including the absolute energy, absolute sum of first-order differences, time-series complexity, kurtosis, and coefficient of variation of the facial muscle motion features.

- The absolute energy indicates the squared fluctuation of the time-series data away from the origin. It can be computed by

$$E(\boldsymbol{VEC}_{:,j}) = \sum_{i=1}^{n} \boldsymbol{VEC}_{i,j}^{2}. \quad (3)$$

- In order to describe the absolute fluctuations between adjacent observations of time-series data, we calculate the absolute sum of first-order differences by

$$A(\boldsymbol{VEC}_{:,j}) = \sum_{i=1}^{n-1} |\boldsymbol{VEC}_{i+1,j} - \boldsymbol{VEC}_{i,j}|. \quad (4)$$

- The time-series complexity is used to evaluate the complexity of a time-series. It can be generated by

$$O(\boldsymbol{VEC}_{:,j}) = \sqrt{\sum_{i=1}^{n-1} (\boldsymbol{VEC}_{i+1,j} - \boldsymbol{VEC}_{i,j})^2}. \quad (5)$$

Algorithm 1: The Procedure of Facial Muscle Motion Features Construction

```
input :
        The facial landmarks $\boldsymbol{L}$
        The number of input frames of a video $n$
output:
        The difference features $\boldsymbol{DVEC}$
        The time-series features $\boldsymbol{TVEC}$
for $i = 1 \to n$ do
    for $j = 1 \to 33$ do
        Calculate distance feature $l_{i,j}$
        Calculate angle feature $\theta_{i,j}$
        $\boldsymbol{VEC} \leftarrow (l_{i,j}, \theta_{i,j})$
    end
end
for $i = 1 \to n - 1$ do
    for $j = 1 \to 66$ do
        Compute difference between adjacent frames of
          distance features $\Delta l_{i,j}$
        Compute difference between adjacent frames of
          angle features $\Delta\theta_{i,j}$
        $\boldsymbol{DVEC} \leftarrow (\Delta l_{i,j}, \Delta\theta_{i,j})$
    end
end
for $j = 1 \to 66$ do
    Construct continuous vector:
      $\boldsymbol{VEC}_{:,j}$=[$\boldsymbol{VEC}_{0,j}, \boldsymbol{VEC}_{1,j}, ..., \boldsymbol{VEC}_{n,j}$]
    $E(\boldsymbol{VEC}_{:,j}) \leftarrow \boldsymbol{VEC}_{:,j}$ away from the origin
    $A(\boldsymbol{VEC}_{:,j}) \leftarrow$ absolute fluctuations of $\boldsymbol{VEC}_{:,j}$
    $K(\boldsymbol{VEC}_{:,j}) \leftarrow$ complexity of $\boldsymbol{VEC}_{:,j}$
    $O(\boldsymbol{VEC}_{:,j}) \leftarrow$ fourth-order standard moment of
      $\boldsymbol{VEC}_{:,j}$
    $C(\boldsymbol{VEC}_{:,j}) \leftarrow$ ratio of the standard deviation to the
      mean of $\boldsymbol{VEC}_{:,j}$
    Obtain time-series features $\boldsymbol{TVEC}$ by cascading
      $\boldsymbol{VEC}_{:,j}$), $A(\boldsymbol{VEC}_{:,j})$, $K(\boldsymbol{VEC}_{:,j})$, $O(\boldsymbol{VEC}_{:,j})$,
      $C(\boldsymbol{VEC}_{:,j})$
end
Return: $\boldsymbol{DVEC}$, $\boldsymbol{TVEC}$
```

- The kurtosis is the fourth-order standard moment. The kurtosis can be denoted as

$$K(\boldsymbol{VEC}_{:,j}) = E[(\frac{\boldsymbol{VEC}_{:,j} - \mu_j}{\sigma_j})^4], \quad (6)$$

  where σ_j is the standard deviation of the time-series and μ_j is the mean of the time-series.
- The coefficient of variation is the ratio of the standard deviation to the mean. Namely, the coefficient of variation is obtained by

$$C(\boldsymbol{VEC}_{:,j}) = \frac{\sigma_j}{\mu_j}. \quad (7)$$

Then, the time-series features $\boldsymbol{TVEC}$ can be obtained by cascading $E(\boldsymbol{VEC}_{:,j})$, $A(\boldsymbol{VEC}_{:,j})$, $O(\boldsymbol{VEC}_{:,j})$, $K(\boldsymbol{VEC}_{:,j})$, $C(\boldsymbol{VEC}_{:,j})$.

We derive and resolve facial muscle motion features from the videos to characterize the unnatural movement patterns of the facial muscles. A detailed description of facial muscle movement features construction is given in Algorithm 1. ***DVEC*** can be seen as the speed pattern of ***VEC***, which is used to capture time discontinuity by training GRU. Time-series features are applied to enhance unnatural facial movements by training SVM.

D. Prediction Probability Fusion

In the fusion module, Dempster-Shafer theory is utilized to effectively fuse the corresponding prediction probability values of GRU and SVM classifiers. Since GRU could process sequence data, we utilize difference features to train GRU and capture temporal relationships. Time-series features are utilized to train the linear classifier SVM for enhancing unnatural facial muscle movements. For our classification task, let O_0 and O_1 indicate the original video and Deepfake video. GRU outputs the predicted probability values $m_g(O_0)$ and $m_g(O_1)$, which have the following characteristics,

$$\begin{cases} m_g(O_0), m_g(O_1) \geq 0, \\ m_g(O_0) + m_g(O_1) = 1, \end{cases} \tag{8}$$

where $m_g(O_0)$ and $m_g(O_1)$ are the probability values that GRU predicts to be the original and Deepfake videos, respectively. $m_s(O_0)$ and $m_s(O_1)$ are the predicted probability values of the original and Deepfake videos for SVM. For our forensics scenario, $m_s(O_0)$ and $m_s(O_1)$ are the nonnegative numbers, and their sum is equal to 1, i.e.,

$$\begin{cases} m_s(O_0), m_s(O_1) \geq 0, \\ m_s(O_0) + m_s(O_1) = 1. \end{cases} \tag{9}$$

The process of prediction probability fusion of GRU and SVM classifiers contains two steps. First, we calculate the normality factor K for initialization and it can be obtained by

$$K = m_g(O_0) \cdot m_s(O_1) + m_g(O_1) \cdot m_s(O_0). \tag{10}$$

Then, we fuse the results of dual classifiers. For the original videos, the fusion results of the dual classifiers can be expressed as

$$m(O_0) = m_g(O_0) \oplus m_s(O_0) = \frac{m_g(O_0) \cdot m_s(O_0)}{K}, \tag{11}$$

where $m(O_0)$ is the final probability value of the original video and the final probability value of the Deepfake video $m(O_1)$ is fused in the same way as $m(O_0)$.

Through the above fusion process, the prediction probability values of GRU and SVM are integrated. Since Deepfake detection is a binary classification task, the final outputs of the model are two probability values. Furthermore, in Section VII, the ablation study shows the effectiveness of Dempster-Shafer theory.

V. Interpretability of FAMM

We provide the interpretative analyses of FAMM. Theoretical analyses of mutual information and video compression reveal that landmark errors are not introduced in the compression process, and theoretical analysis of facial landmarks indicates discrepancies in real and fake videos.

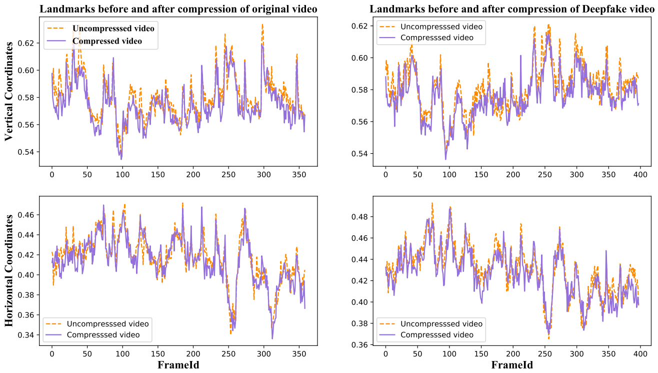


Fig. 8. Difference between horizontal and vertical coordinates before and after compression.

A. Interpretation for Detecting Compressed Videos

In this paper, we mainly use the extracted facial landmarks to construct feature vectors that characterize the movements of facial muscles. Hence, we demonstrate the robustness of the designed features since landmark extraction is not affected by compression. In information theory, mutual information is utilized to measure the information shared between two variables. The more redundant information between two variables, the greater the mutual information. We use mutual information to measure the dependence of facial landmarks between uncompressed videos $\boldsymbol{L^u}$ and compressed videos $\boldsymbol{L^c}$, which can be calculated as

$$MI(\boldsymbol{L^u};\boldsymbol{L^c}) = H(\boldsymbol{L^u}) + H(\boldsymbol{L^c}) - H(\boldsymbol{L^u},\boldsymbol{L^c}), \tag{12}$$

where $H(\boldsymbol{L^u})$ is the information entropy of uncompressed video landmarks, $H(\boldsymbol{L^c})$ is the information entropy of compressed video landmarks, and $H(\boldsymbol{L^u},\boldsymbol{L^c})$ is joint entropy of landmarks between uncompressed and compressed videos.

FaceForensics++ Social [31] is a compressed video dataset of the real social network obtained by uploading part of the FaceForensics++ videos to Facebook and Youtube and re-downloading them. For the mutual information of video landmarks in FaceForensics++ Social, we randomly select 100 videos from FaceForensics++ Social and select the corresponding 100 videos from FaceForensics++ raw [47]. According to Eq. (12), the mutual information value of facial landmarks is obtained. The calculation method of the mutual information of FaceForensics++ LQ [47] compressed video is the same as that of FaceForensics++ Social. We also randomly select 100 videos from FaceForensics++ LQ and select the corresponding 100 videos from FaceForensics++ raw. Then, the mutual information value of facial landmarks is calculated. The normalized $MI(\boldsymbol{L^u};\boldsymbol{L^c})$ of FaceForensics++ LQ and FaceForensics++ Social are 0.996 and 0.9958, which demonstrate that mutual information values of landmarks between compressed and uncompressed videos are close to 1.

Besides, we also analyze the compression process to illustrate that facial landmarks are not affected by compression. The video compression process consists of chunking, intra-frame prediction, inter-frame prediction, quantization, and encoding. We analyze the intra-frame and inter-frame predictions to demonstrate that the compression process does not impact facial landmarks.

For intra-frame prediction, let $f_i^{intra}(x,y)$, $\widehat{f_i^{intra}}(x,y)$ be the i-th frame intensity value at pixel (x,y) and the i-th frame predicted value obtained by the reference pixels, respectively. The prediction error at pixel (x,y) of i-th frame is the difference between intensity value and predicted intensity value, which can be calculated as

$$e_i^{intra}(x,y) = f_i^{intra}(x,y) - \widehat{f_i^{intra}}(x,y). \tag{13}$$

For inter-frame prediction, let (v_x, v_y) be the motion vector of the current image block, which is utilized for motion compensation to achieve prediction. Let $f_i^{inter}(x,y)$ be the i-th frame intensity value at pixel (x,y). The intensity value of the previous image block at $(x-v_x, y-v_y)$ is used to predict the current frame. The prediction value can be calculated as

$$\widehat{f_i^{inter}}(x,y) = f_{i-1}^{inter}(x-v_x, y-v_y). \tag{14}$$

The prediction error at pixel (x,y) of i-th frame is the difference between intensity value and predicted intensity value, which can be expressed as

$$e_i^{inter}(x,y) = f_i^{inter}(x,y) - f_{i-1}^{inter}(x-v_x, y-v_y). \tag{15}$$

The prediction errors are quantized and coded to obtain the compressed data. The above process shows that the compression process operates on the pixel values of the image frames (chroma, luminance) without modifying landmarks. Consequently, the landmarks are not affected by compression.

To further illustrate the above views, we select an original video and a corresponding fake video from FaceForensics++ [47] to draw the landmarks before and after compression. The results in Fig. 8 show that the distribution of the landmarks before and after compression is almost identical, which indicates that compression process would not affect the landmarks. Therefore, FAMM has significant advantages for compressed Deepfake video detection.

B. Interpretation for Facial Landmarks between Real Videos and Fake Videos

Estimated head poses of 2D landmarks are utilized to reveal errors introduced by the synthesis procedure. Let $\boldsymbol{R_a}$ and $\boldsymbol{R_c}$ be the rotation matrices estimated using facial landmarks from the whole face and the central region. The 3D vectors in the head direction corresponding to the two rotation matrices can be expressed as

$$\overrightarrow{\boldsymbol{v_a}} = \boldsymbol{R_a}^T \overrightarrow{\boldsymbol{z}}, \overrightarrow{\boldsymbol{v_c}} = \boldsymbol{R_c}^T \overrightarrow{\boldsymbol{z}}, \tag{16}$$

where $\overrightarrow{\boldsymbol{z}} = [0,0,1]^T$. The cosine distance between two vectors can be expressed as

$$cos(\overrightarrow{\boldsymbol{v_a}}, \overrightarrow{\boldsymbol{v_c}}) = 1 - \frac{\overrightarrow{\boldsymbol{v_a}} \cdot \overrightarrow{\boldsymbol{v_c}}}{||\overrightarrow{\boldsymbol{v_a}}||\ ||\overrightarrow{\boldsymbol{v_c}}||}, \tag{17}$$

where $cos(\overrightarrow{\boldsymbol{v_a}}, \overrightarrow{\boldsymbol{v_c}}) \in [0,2]$.

According to the study [7], the cosine distances $cos(\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{r}}, \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{r}})$ between two vectors of real videos are mostly in the range of $0 \sim 0.02$. For Deepfake videos, the majority of $cos(\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{f}}, \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{f}})$ are distributed in the range of $0.02 \sim 0.08$, i.e.,

$$cos(\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{r}}, \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{r}}) < cos(\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{f}}, \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{f}}). \tag{18}$$

According to Eq. (18), we can know that the cosine distances of original videos are not equal to the cosine distances of Deepfake videos in most cases. Combining Eq. (17) and Eq. (18), the cosine distances of original and Deepfake videos can be expressed as

$$\frac{\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{r}} \cdot \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{r}}}{||\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{r}}||\ ||\overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{r}}||} \neq \frac{\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{f}} \cdot \overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{f}}}{||\overrightarrow{\boldsymbol{v_a}}^{\boldsymbol{f}}||\ ||\overrightarrow{\boldsymbol{v_c}}^{\boldsymbol{f}}||}. \tag{19}$$

According to Eq. (16), Eq. (19) can be expressed as

$$\frac{\boldsymbol{R_a^r} \cdot \boldsymbol{R_c^r}}{||\boldsymbol{R_a^r}||\ ||\boldsymbol{R_c^r}||} \neq \frac{\boldsymbol{R_a^f} \cdot \boldsymbol{R_c^f}}{||\boldsymbol{R_a^f}||\ ||\boldsymbol{R_c^f}||}. \tag{20}$$

$\boldsymbol{R_a}$ and $\boldsymbol{R_c}$ are the rotation matrices estimated using facial landmarks from the whole face and central region, and they are inclusion relations. According to Eq. (20), we could derive the relationship between rotation matrices in real videos $\boldsymbol{R_a^r}$, $\boldsymbol{R_c^r}$ and fake videos $\boldsymbol{R_a^f}$, $\boldsymbol{R_c^f}$. Namely, $\boldsymbol{R_a^r} \neq \boldsymbol{R_a^f}$ or $\boldsymbol{R_c^r} \neq \boldsymbol{R_c^f}$. Since the estimation matrices are estimated by landmarks, the landmarks of real and fake videos are not equal. The above analysis demonstrates that landmark errors are introduced in the forged video, which might cause unnatural facial muscle movements. Hence, FAMM can effectively detect Deepfake videos.

VI. Experiments

In this section, we first describe the experiment settings. Secondly, we conduct experiments with different numbers of frames per video to obtain optimal performance. Then, ablation experiments are conducted to demonstrate the effectiveness of Dempster-Shafer theory and the combination of distance and angle features. In addition, we provide the comparisons between FAMM and some state-of-the-art methods for Deepfake video detection. Moreover, a cross-manipulation performance comparison is presented to illustrate the robustness of FAMM. Finally, we visualize and analyze the designed features to illustrate the effectiveness.

A. Experiment Settings

Datasets. Following previous Deepfake detection approaches, we conduct our experiments on FaceForensics++ [47] (FF++) dataset. It is the most widely used dataset that contains 1000 original videos from Youtube. Furthermore, the FF++ dataset comprises five types of forgery videos namely DeepFakes [36], Face2Face [39], FaceSwap [34], NeuralTexures [40] and FaceShifter [37]. These videos are compressed into two versions: medium compression (HQ) and high compression (LQ), using the H.264 codec with a constant rate quantization parameter of 23 and 40, respectively. The compressed videos of FF++ are implemented by hard-coding, but the videos from social networks are compressed by diverse coding methods. Thus, we further evaluate the performance of FAMM on FaceForensics++ Social [31] dataset for compressed

TABLE III
ABLATION STUDY-THE DETECTION ACC (%) OF DEMPSTER SHAFER INTEGRATION.

Datasets	***DVEC***	***TVEC***	FAMM
DeepFakes	80.54	83.69	**90.00**
FaceSwap	79.60	79.00	**92.75**
Face2Face	77.20	72.60	**91.00**
NeuralTextures	81.50	80.50	**85.50**
FaceShifter	73.26	74.05	**87.50**

TABLE IV
ABLATION STUDY-THE DETECTION ACC (%) OF THE COMBINATION DISTANCE AND ANGLE FEATURES.

Datasets	w/ distance features	w/ angle features	FAMM
DeepFakes	86.25	87.50	**90.00**
FaceSwap	66.50	88.50	**92.75**
Face2Face	78.00	89.50	**91.00**
NeuralTextures	68.50	79.00	**85.50**
FaceShifter	70.50	79.00	**87.50**

TABLE V
ABLATION STUDY-THE DETECTION ACC (%) FOR TRAINING ***DVEC*** USING GRU AND TRANSFORMER.

Datasets	GRU	Transformer
DeepFakes	**80.54**	66.52
FaceSwap	**79.60**	62.20
Face2Face	**77.20**	72.60
NeuralTextures	**81.50**	60.40
FaceShifter	**73.26**	61.51

videos in real-life scenarios. FaceForensics++ Social dataset contains 3360 compressed videos that are distributed by two social networks, namely Facebook and Youtube.

Implementation Detail. We sequentially extract consecutive frames by utilizing `OpenCV` [50] and `Dlib` [48] is adopted to extract the faces from the frames. For GRU, following the default settings [12], the batch size is set as 1024 and the learning rate is set as 1×10^{-3}. The `Adam` optimizer [51] is utilized to optimize the model. The GRU has an output dimension of 64 and two fully connected layers with dimensions of 64 and 2. A dropout layer with dr = 0.25 is set up between the input and the RNN. Two additional drop layers are added to segment the remaining layers. For SVM, a linear kernel function with $\gamma = 10^{-6}$ and penalty factor $C = 0.2$. In our experiments, we adopt an 8:2 dataset split, which means 80% of the videos are used for training the model and 20% for testing. Specifically, as for DeepFakes, we select 800 original videos and 800 fake videos for training, and 200 original videos and 200 fake videos for testing. The settings of the remaining types of videos are consistent with DeepFakes.

Baselines. We compare FAMM with the baseline methods. **FWA** [13], **Mesonet** [14], **Capsule** [15], **RECCE** [18], **SPSL** [29], **ADD** [30] are representatives of the frame-level based methods, **LRNet** [12], **Re-network** [19], **FT-two-stream** [28] are representatives of the video-level based methods. The detailed parameters of the baselines are described as follows.

- **LRNet** [12]. This paper utilized facial landmarks for temporal modeling to detect Deepfake videos. For the training of the model, we extract the coordinates of the facial landmarks from the videos and calibrate them to train the model.
- **FWA** [13]. This paper applied a CNN (ResNet50) model to capture tampering artifacts to detect deep forged videos. For the training of the model, 10 frames of a video are selected as input and them are resized to 224*224.
- **Mesonet** [14]. This paper utilized a CNN network with few parameters to capture tampering artifacts for Deepfake video detection. For the training of the model, we capture random frames from the videos and resize frame to 256*256 to train the model.
- **Capsule** [15]. VGG-19 was used to extract image features and used these features as input to the capsule network for Deepfake video detection. In the experiments, for each video, the first 10 frames are resized to 256*256 and selected as input to the model.
- **RECCE** [18]. The method was based on the reconstruction idea of mining common features of original faces to detect Deepfake videos. For each video, they randomly selected 12 frames as input and resized them to 299*299.
- **Re-network** [19]. The method is a time-aware detection method that used CNN to extract features for Deepfake video detection. Re-network concerned temporal features and the LSTM model is trained by selecting 40 frames as input to the model to extract features. For every frame, we resize them to 299*299.
- **FT-two-stream** [28]. The method focused on compressed Deepfake video detection. FT-two-stream utilized frame-level and temporality-level streams to extract image features and temporal inconsistency features to detect Deepfake videos. In the course of our experiments, we use `FFmpeg` to extract key frames as input to frame-level. For temporality-level stream, we utilize the video directly as input.
- **SPSL** [29]. The method used the phase spectroscopy approach to detect compressed Deepfake videos. For the model, the final spatial form of the phase spectrum is the IDFT of the absolute value of the pristine phase spectrum.
- **ADD** [30]. The method applied frequency-domain learning and optimal transfer theory to knowledge distillation. They utilized the `Dlib` to detect the largest face in every single frame and resize them to a square image of 128*128.

Following the above setup, we acquire video frames using `Opencv`. Then, we detect faces using `Dlib` from frames and perform data cleaning. For frames where the faces cannot be detected, we clean the data and reprocess them to obtain the final data. Finally, we reproduce **LRNet** [12], **FWA** [13], **Mesonet** [14], **Capsule** [15], **Re-network** [19], and **FT-two-stream** [28]. In the experimental process, the parameters of

TABLE II
THE DETECTION ACC (%) AT DIFFERENT FRAME LENGTHS ON DEEPFAKES, FACESWAP, FACE2FACE, NEURALTEXTURES, AND FACESHIFTER DATASETS.

Datasets	20	30	40	50	60	70	80
DeepFakes	88.75	90.00	89.25	88.25	**90.00**	89.50	88.25
FaceSwap	90.00	92.00	91.75	90.75	**92.75**	90.75	90.25
Face2Face	87.25	89.25	90.75	91.00	91.00	**91.25**	90.75
NeuralTextures	79.25	80.75	81.00	81.75	**85.50**	84.25	84.00
FaceShifter	82.00	85.00	83.00	86.25	**87.50**	83.75	84.50

TABLE VI
THE COMPARISON OF THE DETECTION ACC (%) WITH THE STATE-OF-THE-ART METHODS ON DEEPFAKES, FACESWAP, FACE2FACE, NEURALTEXTURES, AND FACESHIFTER DATASETS OF LQ. THE RESULTS OF SPSL ARE CITED DIRECTLY FROM LITERATURE [29] AND [30].

Datasets	FWA [13]	Mesonet [14]	Capsule [15]	Re-network [19]	FT-two-stream [28]	SPSL [29]	ADD [30]	FAMM
DeepFakes	78.50	82.49	87.50	76.44	94.64	93.48	**95.50**	90.00
FaceSwap	67.25	75.25	80.50	63.32	85.27	92.26	92.49	**92.75**
Face2Face	66.75	73.25	81.25	67.09	86.68	86.02	85.42	**91.00**
NeuralTextures	64.75	63.75	70.50	62.06	80.50	76.78	68.53	**85.50**
FaceShifter	66.33	81.75	85.12	74.00	77.31	-	**91.64**	87.50

TABLE VII
THE COMPARISON OF THE DETECTION AUC (%) WITH THE STATE-OF-THE-ART METHODS ON DEEPFAKES, FACESWAP, FACE2FACE, NEURALTEXTURES, AND FACESHIFTER DATASETS. THE RESULTS OF SPSL [29] ARE CITED DIRECTLY FROM LITERATURE [29].

Datasets	FWA [13]	Mesonet [14]	Capsule [15]	Re-network [19]	FT-two-stream [28]	SPSL [29]	FAMM
DeepFakes	86.00	92.90	92.00	84.82	98.00	**98.50**	95.81
FaceSwap	68.00	83.31	78.00	67.08	94.00	**98.10**	96.98
Face2Face	70.00	72.00	86.00	72.65	94.00	94.62	**95.67**
NeuralTextures	67.00	67.00	78.00	68.70	90.00	80.49	**91.45**
FaceShifter	69.32	92.41	92.29	82.09	77.49	-	**92.90**

these models such as learning rates and optimisers follow the authors' settings and the dataset split between training and testing data follows the setup in "Experimental Details" section.

B. Impacts of the Number for Input Frames

In FAMM, since we focus on the features of continuous frames, the number of input frames might affect the detection performance. A video can be intercepted into multiple clips that are fused to identify it as real or fake. The higher number of frames in a video clip, the fewer can be intercepted. The time clips need to be appropriate for the model to capture temporal features effectively. Excessive video clips might lead to over-fit in the temporal dimension, resulting in ineffective detection performance. The number of frames per video clip is evaluated to obtain the best performance for the model. To improve the detection accuracy (ACC), we change the number of frames per video clip. We set each video clip to contain $n \in \{20, 30, 40, 50, 60, 70, 80, 90\}$ frames and conduct extensive experiments. Table II shows the detection performance, which demonstrates that FAMM achieves the highest detection accuracy when choosing $n = 60$. Therefore, in the following experiments, we set n as 60.

C. Ablation Study

To effectively explore unnatural facial motions, we perform ablation experiments on FAMM to evaluate our fusion approach. Specifically, we compare the detection performance of the individual classifier and their fusion, and the specific results are shown in Table III. The average ACC of FAMM that uses Dempster-Shafer is 89.35%, which is 10.93% and 11.38% higher than the average ACC without fusion. These results indicate that the fusion strategy can effectively integrate forensic knowledge and improve compressed Deepfake video detection performance.

In this paper, distance and angle features are joined to indicate the movements of the facial muscles. We also perform the ablation study on FAMM to evaluate the combination of distance and angle features. Specifically, we use only distance features or angle features to represent facial muscle motions and show the results in the second and third columns of Table IV, respectively. The experimental results demonstrate that the detection performance of combined distance and angle features is superior to that using only distance features or angle features. The reason might be that their combination provides an excellent characterization of facial muscle motions.

To illustrate the effectiveness of the GRU, we compare the

TABLE VIII
COMPARISON OF THE COMPRESSED DEEPFAKE VIDEO EVALUATION (ACC (%) AND AUC (%)) WITH STATE-OF-THE-ART METHODS ON DEEPFAKES, FACESWAP, FACE2FACE, NEURALTEXTS, AND FACESHIFTER DATASETS OF FACEFORENSICS++ SOCIAL.

Methods	Social Networks	DeepFakes		FaeSwap		Face2Face		NeuralTextures		FaceShifter	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
FWA [13]	Facebook	65.18	71.75	50.00	51.37	55.60	59.06	51.63	61.06	65.63	71.71
	Youtube	64.83	70.76	51.40	52.92	55.75	58.02	57.45	59.44	64.68	70.52
Mesonet [14]	Facebook	86.22	93.83	82.50	89.61	84.50	91.75	70.01	83.25	81.91	91.32
	Youtube	85.25	93.67	77.00	84.32	79.95	88.66	66.17	81.32	83.75	92.17
Capsule [15]	Facebook	92.20	**98.18**	93.30	98.58	88.02	95.49	72.28	80.99	70.70	82.11
	Youtube	92.74	97.71	82.78	97.53	87.22	94.71	70.70	82.11	**92.88**	**97.17**
Re-network [19]	Facebook	74.44	83.81	77.75	84.86	76.75	84.63	63.75	69.91	72.25	83.36
	Youtube	74.44	84.01	68.17	73.00	73.25	79.39	67.75	74.58	78.00	88.10
FT-two-stream [28]	Facebook	**92.50**	97.55	90.25	95.98	82.50	89.69	73.00	81.90	89.25	95.08
	Youtube	**93.48**	**97.82**	79.25	91.76	80.50	86.80	77.00	80.85	90.00	96.01
FAMM	Facebook	91.00	97.35	**94.50**	**98.72**	**93.73**	**96.93**	**89.00**	**94.97**	**89.25**	**95.00**
	Youtube	90.75	96.17	**95.00**	**98.82**	**94.25**	**97.97**	**88.25**	**92.44**	88.00	94.53

GRU model with the Transformer. The experimental results are shown in the Table V. The results demonstrate that the detection performance of adopting Transformer is not as well as that of adopting GRU model. The Transformer is based on a self-attentive mechanism that captures the relationship of the temporal sequence by computing the relationship of each position to the other time periods in the sequence. In contrast, GRU controls the flow of information and forgetting by introducing a gating mechanism, which can better to the designed features. Besides, Transformer has a more complex structure and requires a higher amount of data for training and inference. For a video, the difference features designed in our proposed method have a low dimension of 60*59. In the case of small data volume, Transformer is difficult to train and extract effective features. On the contrary, the GRU model has a more detectable structure and is more advantageous in processing time-series data with lower dimensionality.

D. Comparisons of Compressed Deepfake Video Detection Performance on FaceForensics++ LQ

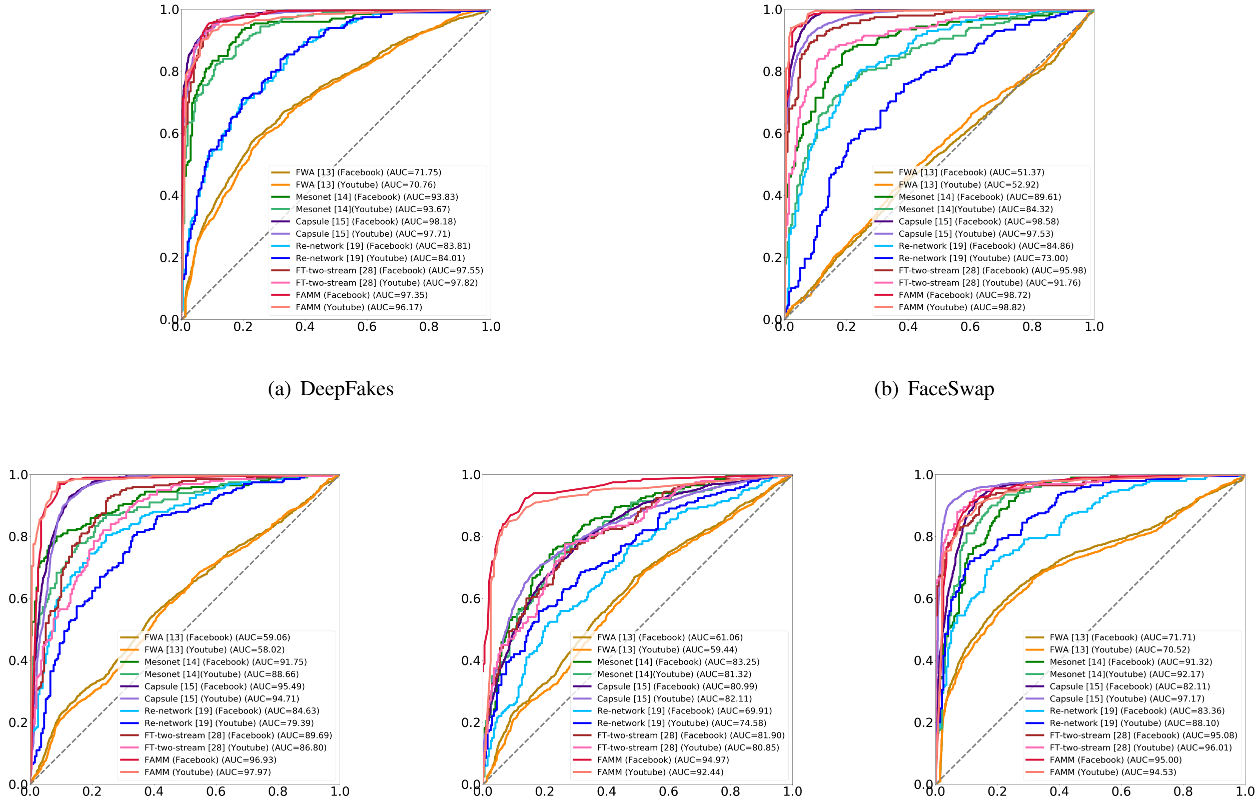
In this section, we use ACC and Area under the curve (AUC) to evaluate the detection performance on the compressed Deepfake datasets. We evaluate FAMM on the FF++ LQ and compare it with some state-of-the-art methods. When detecting compressed videos, the results in Table VI demonstrate that FAMM achieves state-of-the-art performance on Face2Face, NeuralTextures, and FaceShifter. **FWA** [13], **Mesonet** [14] and, **Capsule** [15] focused on a single-frame image and detected Deepfake videos based on facial tampering artifacts. The presence of compression artifacts in compressed videos might impact detection performance, thus these methods are unable to detect compressed Deepfake videos very well. **FT-two-stream** [28] combined frame-level and temporality-level features and **SPSL** [29] based on the phase spectrum to detect compressed videos, which achieved impressive detection performance. **ADD** [30] applied frequency-domain learning to detect compressed images and achieved the best performance on DeepFakes and FaceShifter. FAMM achieves the highest scores on the FaceSwap, Face2Face, NeuralTextures. Therefore, FAMM can effectively detect compressed Deepfake videos.

For compressed Deepfake video detection, we can notice that the detection performance of FAMM is higher than **ADD** [30] on FaceSwap, Face2Face, and NeuralTextures. For ACC metric, our method is outperforming **ADD** [30] by 0.26% on FaceSwap, 5.58% on Face2Face, 16.97% on NeuralTextures. The reason is that in the presence of heavy compression, the videos would lose more information making it difficult to extract discriminatory features. However, FAMM can effectively decouple compression artifacts and tampering artifacts improving the detection performance of compressed Deepfake videos.

E. Comparisons of Compressed Deepfake Video Detection Performance on FaceForensics++ Social

As the compressed dataset of FF++ is implemented by hard-coding, we evaluate the detection performance on FF++ Social, which is the realistic scenes dataset. Since FF++ Social only improves the data for validation and testing, we use FF++ high-quality (HQ) datasets for training and FF++ Social for testing to guarantee the performance of the model. The specific experimental results in Table VIII demonstrate that FAMM works equally well for compressed videos in real scenes and outperforms the FF++ LQ.

We also evaluate the detection performance on FF++ Social of FAMM using Receiver Operating Characteristic (ROC) curve, where the horizontal and vertical coordinates represent the False Positive Rate (FPR) and True Positive Rate (TPR),



(a) DeepFakes

(b) FaceSwap

(c) Face2Face

(d) NeuralTextures

(e) FaceShifter

Fig. 9. ROC (receiver operating characteristic) curves for the state-of-the-art compressed Deepfake videos detection methods on different public datasets: (a) DeepFakes, (b) FaceSwap, (c) Face2Face, (d) NeuralTextures, (e) FaceShifter.

respectively. As shown in Fig. 9, the curve for our method is closer to the upper left corner on FaceSwap, Face2Face, NeuralTextures, and FaceShifter, which means that our method achieves superior performance.

F. Cross-Manipulation Detection Performance Comparisons

In addition, we conduct cross-manipulation experiments on FaceForensics++ to evaluate the robustness of FAMM without introducing confounders such as variations in pose or illumination. Specifically, the fake videos that were created from the same sources. In this paper, we focus on the research of compressed videos, therefore, the classifiers are trained on one tampered database of FF++ LQ but tested on the remaining four databases. The results of the experiment are provided in Table IX, which demonstrate that FAMM achieves comparable robustness performance than most of the baseline methods, except for training on FaceSwap and testing on FaceShifter. When training on DeepFakes, FAMM is outperforming the SOTA methods by 6.83% (ACC) and 7.66% (AUC) (i.e. **LR-Net** [12]), 12.72% and 14.14% (i.e. **Mesonet** [14]), 15.47% (AUC) (i.e. **RECCE** [18]), 7.13% (ACC) and 9.25% (AUC) (i.e., **Re-network** [19]). Our approach also achieves state-of-the-art performance for models trained by other manipulation methods. Therefore, compared with baseline methods, FAMM achieves excellent generalization to different forgeries.

It is noticed that when we use DeepFakes for testing, Face2Face, NeuralTextures or FaceShifetr for training, the model achieves an adequate performance. In other cases, the performance of the model is significantly affected. When we train the model on FaceSwap and test it on Face2Face or NeuralTextures, the performance of cross-manipulation is inferior. The possible reason is that FaceSwap uses classical graphical techniques to implement facial replacement, causing obvious forged boundaries that are easily detected. In contrast, Face2Face and NeuralTextures techniques are expression migration techniques and the synthesized videos fit better. In addition, we perform the experiment by training the model on FaceShifter while testing it on the remaining datasets and we could observe poor performance of the model. This is because FaceShifter generates swapped faces with high fidelity by thoroughly and adaptively exploiting and integrating target attributes.

G. Visualization and Analysis

In this section, we select an original video and a corresponding fake video to visualize and analyze the difference features ***DVEC*** and time-series features ***TVEC*** to illustrate the effectiveness of the designed features. First, we draw the 3D figure of difference features ***DVEC*** to demonstrate the distribution difference between the real video and synthesized

TABLE IX
DETECTION ACC (%) AND AUC (%) FOR CROSS-MANIPULATION EVALUATION WITHIN THE FACEFORENSICS++ OF LOW-QUALITY: DEEPFAKES, FACESWAP, FACE2FACE, NEURALTEXTURES, AND FACESHIFTER. THE RESULTS OF RECCE [18] ARE CITED DIRECTLY FROM LITERATURE [18].

Methods	Train	DeepFakes		FaceSwap		Face2Face		NeuralTextures		FaceShifter	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
LRNet [12]	DeepFakes	-	-	60.00	65.66	59.00	65.84	60.19	64.09	58.00	63.00
Mesonet [14]		-	-	55.10	62.17	51.52	52.27	51.40	53.53	55.61	64.71
RECCE [18]		-	-	-	74.29	-	70.66	-	67.34	-	-
Re-network [19]		-	-	54.77	58.63	58.04	62.43	61.55	64.18	61.75	67.01
FAMM		-	-	**62.75**	**69.79**	**73.25**	**79.14**	**65.50**	**70.64**	**63.00**	**69.67**
LRNet [12]	FaceSwap	58.25	73.84	-	-	**59.00**	**68.54**	50.07	49.26	50.75	56.94
Mesonet [14]		56.46	63.12	-	-	48.61	48.74	48.98	47.74	53.32	54.70
RECCE [18]		-	**82.39**	-	-	-	64.43	-	56.70	-	-
Re-network [19]		52.88	54.18	-	-	48.24	46.67	44.22	43.28	**59.25**	**63.17**
FAMM		**67.75**	76.83	-	-	55.00	64.37	**52.75**	**58.82**	51.50	58.21
LRNet [12]	Face2Face	74.50	**81.93**	56.00	62.66	-	-	70.00	**79.53**	56.25	**60.17**
Mesonet [14]		49.74	54.54	49.87	52.26	-	-	50.13	54.53	49.62	59.83
RECCE [18]		-	75.99	-	64.53	-	-	-	73.32	-	-
Re-network [19]		57.14	61.11	52.01	53.37	-	-	59.05	60.56	49.50	49.15
FAMM		**75.25**	81.29	**60.75**	**65.65**	-	-	**71.00**	78.00	**58.00**	59.76
LRNet [12]	NeuralTextures	71.00	80.46	54.50	61.57	81.25	**89.39**	-	-	56.50	65.75
Mesonet [14]		51.52	55.94	46.29	43.00	52.67	55.98	-	-	54.22	57.78
RECCE [18]		-	78.83	-	**63.70**	-	80.89	-	-	-	-
Re-network [19]		61.65	67.91	49.00	50.13	59.55	63.80	-	-	57.74	61.08
FAMM		**74.50**	**81.40**	**59.00**	62.38	**81.50**	87.33	-	-	**63.00**	**68.08**
LRNet [12]	FaceShifter	75.75	82.05	**63.75**	67.11	65.75	67.18	61.25	62.10	-	-
Mesonet [14]		60.86	69.37	52.55	53.19	47.34	49.11	47.45	48.85	-	-
RECCE [18]		-	-	-	-	-	-	-	-	-	-
Re-network [19]		64.41	70.58	50.25	52.72	52.02	55.82	57.29	61.94	-	-
FAMM		**77.00**	**82.72**	62.50	**68.04**	**66.00**	**71.33**	**62.00**	**66.24**	-	-

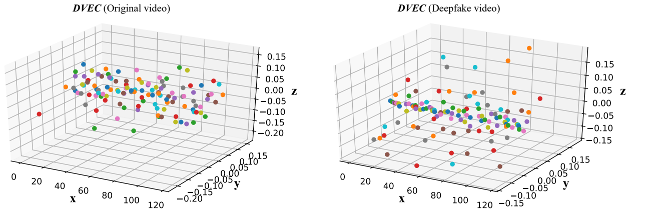


Fig. 10. The difference features ***DVEC*** visualization of original (left) and Deepfake (right) videos. The distribution of original videos is more concentrated, while the distribution of Deepfake videos is relatively discrete.

video. As shown in Fig. 10 (left), ***DVEC*** of the original video is concentrated in a certain area. However, the Deepfake video has discrete distribution regarding ***DVEC*** (as Fig. 10 (right) shows), implying deficiencies in the temporal sequence.

Then, the curve chart of time-series features ***TVEC*** is illustrated in Fig. 11, which shows that there is a visible difference between the original video and Deepfake one. For

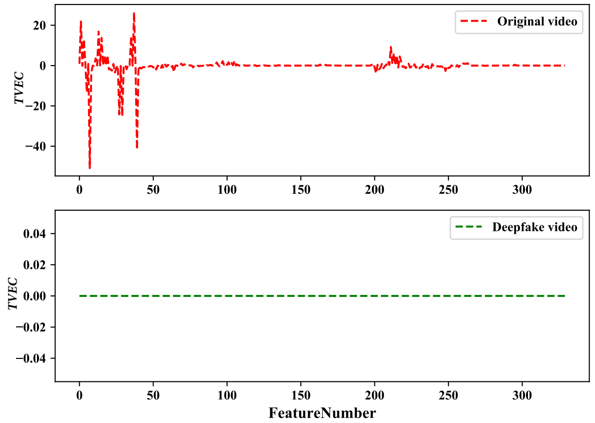


Fig. 11. The time-series features ***TVEC*** visualization of original (red) and Deepfake (green) videos. The distribution of original videos is random, while the distribution of Deepfake videos is mainly concentrated around 0.

the original video, the feature distribution is random because of the variety of expressions, while the distribution of Deepfake video is mainly concentrated around 0. The aforementioned visualization results demonstrate that the proposed FAMM can capture the difference in the designed features between the original and Deepfake videos and would be capable to obtain excellent detection performance.

VII. Conclusion

In this work, we design facial muscle motion features from a novel perspective to detect compressed Deepfake videos. The compressed videos contain compression artifacts, which are detrimental to feature extraction and network learning. We investigate geometric features, which construct features to mine and enhance unnatural facial muscle motions. Besides, we provide interpretive analyses, including compression procedure, mutual information, and facial landmark analyses, which demonstrate the effectiveness of FAMM theoretically. We carry out FAMM on the FaceForensics++ LQ and FaceForensics++ Social. Furthermore, we conduct cross-manipulation experiments without introducing confounding factors. The extensive experimental results demonstrate that FAMM achieves excellent performance in detecting compressed Deepfake videos. In addition, the possibility of mining more effective facial features would also be further studied.

References

[1] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A subjective and objective study of stalling events in mobile streaming videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 183–197, 2019.

[2] T. Leaver, T. Highfield, and C. Abidin, *Instagram: Visual social media cultures*, 2020.

[3] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–49, 2019.

[4] P. Zhang, L. Yang, X. Xie, and J. Lai, "Lightweight texture correlation network for pose guided person image generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4584–4598, 2022.

[5] https://www.bitdefender.com/blog/hotforsecurity/.

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[7] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8261–8265.

[8] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, DOI:10.1109/TPAMI.2020.3009287.

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

[10] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings of IEEE Winter Applications of Computer Vision Workshops*, 2019, pp. 83–92.

[11] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm, "Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1022.

[12] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3609–3618.

[13] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[14] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *Proceedings of IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[15] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2307–2311.

[16] K. Sun, H. Liu, T. Yao, X. Sun, S. Chen, S. Ding, and R. Ji, "An information theoretic approach for attention-driven face forgery detection," in *Proceedings of European Conference Computer Vision*, 2022, pp. 111–127.

[17] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, "Msta-net: Forgery detection by generating manipulation trace based on multi-scale self-texture attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4854–4866, 2022.

[18] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.

[19] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of IEEE International Conference on Advanced Video and Signal based Surveillance*, 2018, pp. 1–6.

[20] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces*, vol. 3, no. 1, pp. 80–87, 2019.

[21] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019, pp. 1–3.

[22] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 667–684.

[23] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms," in *Proceedings of ACM International Conference on Multimedia*, 2020, pp. 4318–4327.

[24] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

[25] H. Wang, Z. Liu, and S. Wang, "Exploiting complementary dynamic incoherence for deepfake video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, DOI:10.1109/TCSVT.2023.3238517.

[26] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "Mre-net: Multi-rate excitation network for deepfake video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, DOI:10.1109/TCSVT.2023.3239607.

[27] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1696–1708, 2023.

[28] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089–1102, 2021.

[29] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 772–781.

[30] S. Woo *et al.*, "Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 122–130.

[31] F. Marcon, C. Pasquini, and G. Boato, "Detection of manipulated face videos over social networks: A large-scale study," *Journal of Imaging*, vol. 7, no. 10, pp. 193–208, 2021.

[32] X. Zhang and X. Wu, "Multi-modality deep restoration of extremely compressed face videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, DOI:10.1109/TPAMI.2022.3157388.

[33] F. Saxen, P. Werner, and A. Al-Hamadi, "Real vs. fake emotion challenge: Learning to rank authenticity from facial activity descriptors," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3073–3078.

[34] Faceswap github - non official project based on original deepfakes thread. Available at. https://github.com/MarekKowalski/FaceSwap/. 2018.

[35] J. D. Foley, F. D. Van, A. Van Dam, S. K. Feiner, and J. F. Hughes, *Computer graphics: principles and practice*, 1996.

[36] Deepfakes github - non official project based on original deepfakes thread. Available at. https://github.com/deepfakes/faceswap/. 2018.

[37] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5074–5083.

[38] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1831–1845, 2015.

[39] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[40] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–12, 2019.

[41] M. C. Doukas, M. R. Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou, "Head2head++: Deep facial attributes re-targeting," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 31–43, 2021.

[42] M. A. Baig, A. A. Moinuddin, and E. Khan, "Psnr of highest distortion region: an effective image quality assessment method," in *Proceedings of IEEE International Conference on Electrical, Electronics and Computer Engineering*, 2019, pp. 1–4.

[43] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[44] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[45] V. Baroncini, L. Capodiferro, E. D. Di Claudio, and G. Jacovitti, "The polar edge coherence: a quasi blind metric for video quality assessment," in *Proceedings of European Signal Processing Conference*, 2009, pp. 564–568.

[46] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[47] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 1–11.

[48] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[49] K. Wu, M. Zhou, G. Li, and X. Chen, Zengzhaoand He, "Facial expression recognition based on geometrical features of angles," *Computer Applications and Software*, vol. 37, no. 7, pp. 120–124, 2020.

[50] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to opencv," in *Proceedings of International Convention MIPRO*, 2012, pp. 1725–1730.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations*, 2015.

Xin Liao (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from Beijing University of Posts and Telecommunications in 2007 and 2012, respectively. He is currently a Professor and a Doctoral Supervisor with Hunan University, China. He worked as a Post-Doctoral Fellow with the Institute of Software, Chinese Academy of Sciences, and also a Research Associate with The University of Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, USA. His current research interests include multimedia forensics, steganography, and watermarking. He is a member of Technical Committee (TC) on Multimedia Security and Forensics of AsiaCPacific Signal and Information Processing Association, TC on Computer Forensics of Chinese Institute of Electronics, and TC on Digital Forensics and Security of China Society of Image and Graphics. He is serving as an Associate Editor for the IEEE Signal Processing Magazine. He is a senior member of the IEEE.

Yumei Wang received the B.E. degree from College of Computer Science and Technology, Guizhou University, Guiyang, China, in 2020. She is currently pursuing the M.S. degree with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her current research interests include multimedia forensics.

Tianyi Wang (Student Member, IEEE) received the B.Sc. degrees in Computer Science and Applied and Computational Mathematical Sciences from the University of Washington, Seattle, USA, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Hong Kong. His major research interests include multimedia forensics, face forgery detection, and artificial intelligence.

Juan Hu is a Ph.D. candidate in the College of Computer Science and Electronic Engineering, Hunan University, China, where she received her M.S. degree in 2019. She received the B.S. degree in the College of Electronic Information Science and Technology from Nanjing Agricultural University, China, in 2017. She is a visiting Ph.D. student at the National University of Singapore from March 2022 to March 2023. Her current research interests include multimedia forensics and artificial intelligence.

Xiaoshuai Wu received his B.S. degree in 2019 from Nanyang Institute of Technology and a M.S. degree in 2022 from Hangzhou Dianzi University. He is currently pursuing the Ph.D. degree at Hunan University. His research interests include data hiding and AI security.