Preventing DeepFake Attacks on Speaker Authentication by Dynamic Lip Movement Analysis

Chen-Zhao Yang, Jun Ma, Shilin Wang, *Senior Member, IEEE*, and Alan Wee-Chung Liew, *Senior Member, IEEE*

***Abstract*— Recent research has demonstrated that lip-based speaker authentication systems can not only achieve good authentication performance but also guarantee liveness. However, with modern DeepFake technology, attackers can produce the talking video of a user without leaving any visually noticeable fake traces. This can seriously compromise traditional face-based or lip-based authentication systems. To defend against sophisticated DeepFake attacks, a new visual speaker authentication scheme based on the deep convolutional neural network (DCNN) is proposed in this paper. The proposed network is composed of two functional parts, namely, the Fundamental Feature Extraction network (FFE-Net) and the Representative lip feature extraction and Classification network (RC-Net). The FFE-Net provides the fundamental information for speaker authentication. As the static lip shape and lip appearance is vulnerable to DeepFake attacks, the dynamic lip movement is emphasized in the FFE-Net. The RC-Net extracts high-level lip features that discriminate against human imposters while capturing the client's talking style. A multi-task learning scheme is designed, and the proposed network is trained end-to-end. Experiments on the GRID and MOBIO datasets have demonstrated that the proposed approach is able to achieve an accurate authentication result against human imposters and is much more robust against DeepFake attacks compared to three state-of-the-art visual speaker authentication algorithms. It is also worth noting that the proposed approach does not require any prior knowledge of the DeepFake spoofing method and thus can be applied to defend against different kinds of DeepFake attacks.**

***Index Terms*— DeepFake spoofs, dynamic lip movement, lip biometrics, liveness detection, multi-task learning.**

I. Introduction

IN THE past few years, user authentication systems based on biometric features have been widely used in many real-life applications [1]–[4] such as mobile payment, smartphone unlock, access control, etc. Human face [5]–[9] is one of the most popular biometric features and it can usually provide a high level of security and convenience compared with password or Personal Identity Number (PIN). Recent research [10]–[22] has shown that human lip alone can be used as a biometric feature to differentiate different speakers. Compared with traditional biometric features such as face, fingerprint and iris, lip feature has the following advantages: i) It contains rich identity-related information that is both static, i.e. lip shape and appearance, and dynamic, i.e. lip movement reflecting the talking habit of a speaker; ii) The capturing device of the lip feature, i.e. a common video camera, is inexpensive and readily available. In many applications, the image capturing device for face authentication can be directly used to capture the lip feature. Hence, lip feature can be integrated in many face-based authentication systems to provide a very high level of security.

Speaker authentication based solely on lip feature (which is also referred to as Visual Speaker Authentication, VSA) has been investigated since 1990s [10], [11]. Sophisticated VSA approaches [19]–[22] can achieve reliable authentication results (with a Half Total Error Rate (HTER) [19]–[22] of less than 1%) under the fixed-password scenario, where the password for each user is assumed to be fixed during training and authentication. Although fixed-password VSA ensures double security, i.e. the password and the lip appearance/talking behavior, a replay attack using a prerecorded video can compromise such a system [21], [23]. To resist replay attacks and guarantee liveness, a random password scheme was first proposed in [23] and a sophisticated VSA system based on this scheme was presented in [21]. A flowchart of the random password based VSA system is given in Fig. 1. The authentication system will randomly generate a prompt text and the system will accept the request if both the identity and the pronounced content are correct. With the assumption that the imposter/attacker cannot prerecord videos of the client pronouncing all the prompt texts, the client's "liveness" can therefore be ensured.

With the development of advanced computer graphics techniques, modern face identity swap/synthesis methods, especially DeepFake techniques, can produce visually realistic fake videos containing facial information [24]–[26]. Recent research has demonstrated that DeepFake methods can deceive traditional face-based user authentication systems [27] and compromise many liveness detection algorithms. With modern DeepFake techniques, even lip-based authentication systems using random password are under threat. To fool a VSA system, an imposter/attacker can record a video of himself/herself speaking the random prompt text on the spot and swap the face in the video with that of the client by DeepFake (an example is

Manuscript received June 30, 2020; revised October 17, 2020 and December 6, 2020; accepted December 7, 2020. Date of publication December 18, 2020; date of current version January 5, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61771310. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vitomir Štruc. *(Corresponding author: Shilin Wang.)*

Chen-Zhao Yang, Jun Ma, and Shilin Wang are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: 568466781@sjtu.edu.cn; madajun@sjtu.edu.cn; wsl@sjtu.edu.cn).

Alan Wee-Chung Liew is with the School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4222, Australia (e-mail: a.liew@griffith.edu.au).

Digital Object Identifier 10.1109/TIFS.2020.3045937

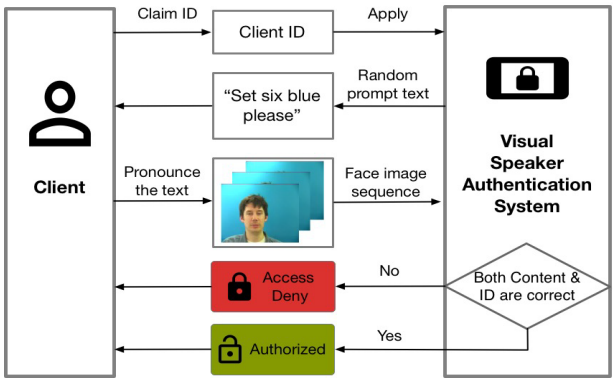


Fig. 1. A visual speaker authentication system under the random password scenario.

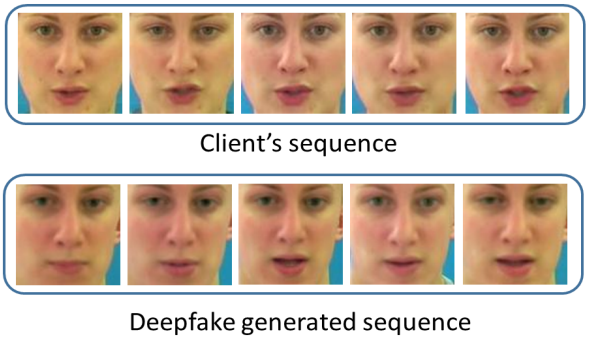


Fig. 2. Top: pristine client's talking sequence; Bottom: computer generated client's talking sequence by DeepFake.

shown in Fig. 2). As the lip appearance generated by DeepFake is very similar to that of the client and the pronounced content is also correct, a VSA system can be easily compromised.

Recently, some pioneering approaches to detect DeepFake forgeries have been proposed [28]–[32]. These methods can achieve high detection accuracy on certain type of DeepFake forgeries. However, in user authentication, the imposter/attacker can produce different fake videos using different kinds of DeepFake methods. In the worst case, there will be no negative (forgery) samples to train the detection classifier when a totally new DeepFake technique is encountered. This will bring great difficulties for existing DeepFake detection approaches.

Facing this challenging problem, we proposed a new deep learning based VSA system, which can effectively detect DeepFake attacks without any prior knowledge about the forgery video or the manipulation method. Considering that the static information is vulnerable to DeepFake attacks, our system employs the dynamic information describing the client's unique talking habit for authentication. The effectiveness of our approach rests on the assumption that the attacker can only obtain very limited information about the client, e.g. a few photos or video recordings of his/her lip movements uttering some specific prompt texts. The major contributions of the proposed work are four-fold: i) A motion-based fundamental feature extraction network is proposed to extract information about talking habit; ii) A comprehensive lip feature representation is proposed. The extracted high-level lip feature has high discriminative power against human imposters as well as having a good representation ability to reconstruct the client's talking styles; iii) A new loss function is designed and an end-to-end, multi-task learning strategy is adopted to integrate all the properties described above in an efficient manner; iv) The proposed authentication scheme can successfully reject human imposters as well as DeepFake videos produced by different manipulation techniques. Moreover, our work also proposed a new solution for DeepFake detection under the user authentication scenario.

This paper is organized as follows. Section II briefly reviews related works on VSA and DeepFake manipulation and detection. Section III presents the proposed deep neural network (DNN) for lip-based speaker authentication against DeepFake attacks. Section IV presents the experiment results of the proposed scheme in comparison with several state-of-the-art VSA and DeepFake detection approaches. Finally, Section V draws the conclusion.

II. Related Works

A. Visual Speaker Authentication

In the past decades, many researchers have proposed various VSA methods [10]–[22], which can be roughly divided into two categories: handcrafted-feature based approach and automatic feature learning-based approach. For the handcrafted-feature based approach, various kinds of feature representations have been proposed to describe the static (lip shape and appearance) and dynamic (talking habits) information of lip biometric. Some widely used lip features include: i) Lip shape descriptors: geometric contour descriptors [12], [15], [16] and contour model parameters obtained by the Active Shape Model (ASM) [10]; ii) Lip texture descriptors: intensity profile along the contour points [10], intensity distribution of the lip region [15], texture model parameters obtained by the Active Appearance Model (AAM) [14] and the visibility of teeth/tongue [12]; iii) Lip movement descriptors: motion vector of the lip contour [14], [15]. Hidden Markov Model (HMM) [10], [15], [16] is the most widely used speaker classification method for handcrafted features. The above works using handcrafted features had demonstrated the feasibility of using lip feature for speaker authentication; however, their performance was not fully comparable to that of recent sophisticated face-based authentication approaches. One of the most successful handcrafted lip features was proposed in 2012 by Chan *et al.* [19]. Their feature was essentially a texture descriptor, which was referred to as the Local Ordinal Contrast Pattern (LOCP). Three Orthogonal Planes (TOP) were employed in their feature to consider both the spatial and temporal information. A low HTER of 0.36% was achieved on the XM2VTS database [33] with about 300 speakers, which demonstrated the high discriminative power of the lip biometric.

In recent years, automatic feature learning approaches such as sparse coding and deep neural networks have outperformed handcrafted-feature based approaches in many computer vision applications. Lai *et al.* [20] employed a sparse coding based feature to describe the lip feature in a spatiotemporal manner. In a recent work of our group [21], an end-to-end deep

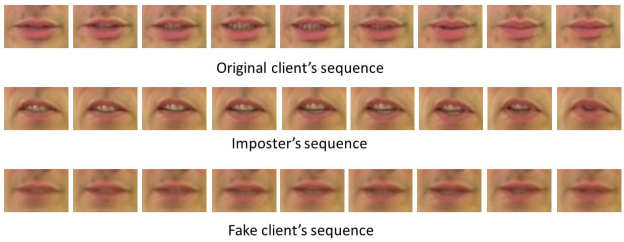


Fig. 3. Top: Client's lip sequence; middle: human imposter's lip sequence; bottom: deepfake lip sequence.

neural network is proposed for visual speaker authentication. Sun *et al.* [22] proposed to pay more attention to the discriminative segment to improve the authentication performance. These methods achieved very good authentication results (zero EER and HTER of 0.46% [20] and 0.6% [21], respectively) on a database with forty speakers. Moreover, the learned-feature based approaches were highly robust against variation caused by different talking poses and positions, compared with the handcrafted-feature based methods [20], [21].

All the above VSA approaches were designed for the fixed password scenario. To resist replay attacks, Liao *et al.* [23] proposed a random/dynamic password scheme for VSA. Recently, a deep neural network-based VSA system with the random password scheme was proposed [21]. This approach is effective and robust against human imposters (by distinguishing different lip shape, appearance and movements) and pre-recorded videos (based on the random password scheme).

B. DeepFake Manipulation and Detection

DeepFake originally refers to a deep learning based technique to produce fake images/videos by swapping the face of a person in the original image/video with the face of another person [34]. Subsequently, the term DeepFake has extended to refer to any AI generated impersonating images or videos [35]. It has caused great public concern because it was first applied to transpose celebrity faces into porn videos [36]. Generally speaking, there are two major kinds of DeepFake manipulation methods, i.e. face-swapping [25], [26], [37] and lip-sync [38]–[40]. Face-swapping manipulation is usually composed of three steps: i) Detect and crop the face region of subject one in the original image/video by face segmentation [41]; ii) Synthesize the face of subject two, based on the original face of subject one using an autoencoder network [25] or a Generative Adversarial Network (GAN) [26]; iii) Insert the synthesized face into the original image/video with some kind of blending postprocessing steps [42], [43]. Lip-sync manipulation modifies a person's mouth region to be consistent with another person's speech [38]–[40]. Compared with the traditional computer graphics based face identity swap techniques [44], [45], DeepFake videos are visually more realistic and difficult to detect [46].

In [27], Korshunov and Marcel demonstrated that DeepFake videos can deceive state-of-the-art face-based authentication methods. Very high false acceptance rates were obtained (85% by VGG [47] and 95% by FaceNet [48]) for DeepFake videos.

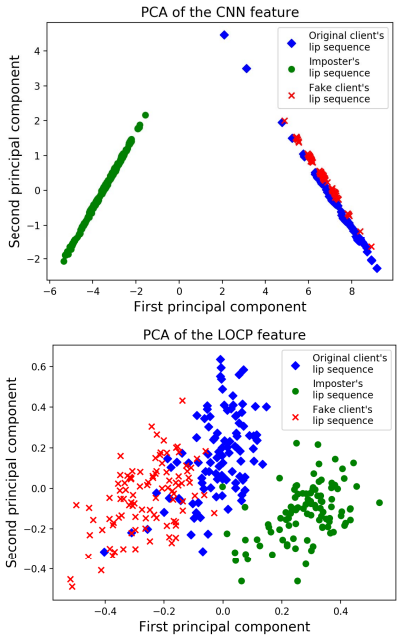


Fig. 4. The feature distribution projected onto the first two principle components for top: CNN features and bottom: LOCP features.

It was also observed that more advanced face authentication systems are more vulnerable to DeepFake attacks [27], [49]. Similarly, traditional VSA systems are also under threat because the static lip shape and appearance are vulnerable to DeepFake attacks. Fig. 3 and Fig. 4 show an example demonstrating this issue. Three kinds of lip image sequences are analyzed in Fig. 4 (each kind contains 100 sequences), including: the original client's sequence pronouncing the prompt text "please", the imposter's sequence pronouncing the same prompt text "please" and the fake client's sequence generated by a DeepFake technique [25] using the imposter's sequence and several photos of the client. Two kinds of lip features, i.e. the automatic CNN feature [21] and the handcrafted LOCP feature [19] are investigated. For ease of visualization, all the features are projected onto the first two principle components by Principle Component Analysis (PCA). In Fig. 4, it is observed that in both feature representations, compared with the feature points of the imposter's sequences (shown in green), the feature points of the fake client's sequences (shown in red) are much closer to the original client's sequences (shown in blue). Such a phenomenon is much more obvious for the traditional CNN features, which indicates that the traditional CNN features rely heavily on the static lip shape and appearance and thus are very vulnerable to DeepFake attacks.

In the past three years, great efforts have been devoted to accurate and reliable DeepFake detection. To facilitate research in this area, many datasets containing various DeepFake videos have been published, including: FaceForensics ++ Dataset [50], Celeb-DF Dataset [51], DFDC Dataset [52], DeeperForensics-1.0 Dataset [53]. Some pioneering works [28]–[32] on DeepFake detection have been proposed. In [28], Li *et al.* observed that eye blinking was not well synthesized in the fake facial videos and designed a corresponding forensics

method based on this observation. In [29, 30], the authors of the same group proposed another DeepFake detection approach, by exploiting the head pose inconsistency [29] and face warping artifacts [30] in the fake videos. Afchar *et al.* proposed a Convolutional Neural Network (CNN), named as MesoNet [31], to detect DeepFake [24] and Face2face [45] manipulations by examining the artifacts left by these autoencoder based face synthesis methods. Sabir *et al.* [32] proposed a CNN+RNN DeepFake detection network. Dense modules are used for feature extraction and bidirectional Recurrent Neural Network (RNN) units are used to analyze the dynamic information of the video. Recently, Rossler *et al.* proposed a new CNN with the Xception structure [50], which achieved state-of-the-art performance in detecting fake videos generated by both the autoencoder and GAN based DeepFake techniques. Agarwal *et al.* [54] observed that most lip-sync manipulation methods cannot synthesize the lip movement of phoneme "M, B, P" correctly and designed a lip-sync DeepFake detection method based on the above observation. Interested readers may refer to [34, 55] for a more comprehensive study. Although prior methods perform well in detecting specific manipulation techniques, the detection performance will degrade significantly for different/unknown manipulation approaches [55]. Especially in the user authentication systems, attackers can select different face manipulation approaches to generate fake facial videos, which calls for a universal detector. In [56], Agarwal *et al.* tried to model an individual's unique speaking pattern by analyzing his/her facial expression and movement. A set of handcrafted features are extracted to characterize an individual's motion signature. The SVM classifier is then used to differentiate the genuine speaker's talking video from that generated by various Deepfake methods. Similar to [56], here we propose a new behavioral feature based approach, which can detect different kinds of DeepFake attacks without any prior information. The proposed method is effective owing to its ability to capture the unique behavioral characteristics of individual speaker and the information asymmetry between the attacker and the authentication system.

III. The Proposed Method

The overall architecture of the proposed authentication system is shown in Fig. 5. In order to guarantee liveness, the random password strategy [23] is adopted. The Dlib [41] detector is adopted to extract the lip region from the face video. The content authentication network of [21], which uses the Connectionist Temporal Classification (CTC) [57] as the decoder, is adopted to verify the pronounced text solely based on the visual information, i.e. the lip region sequence. The prompt text is divided into a series of isolated words based on the CTC output. Note that in our system, the user ID authentication is performed at the word-level in order to reduce the variations caused by pronouncing different content.

After word segmentation, the newly proposed Speaker Authentication network based on Dynamic Talking Habit (SA-DTH-Net in short) will be used to examine whether the lip subsequence complies to the client's talking habit when pronouncing a specific word. The output of the SA-DTH-Net

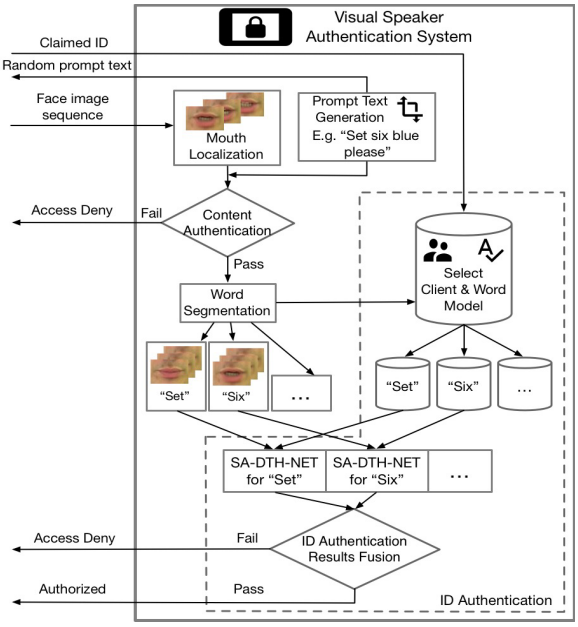


Fig. 5. The overall architecture of the proposed authentication system.

is a probability of the lip subsequence belonging to the client. After thresholding, the word-level authentication result can be obtained. The final decision is obtained by voting over all the word-level authentication results in the sentence. The SA-DTH-Net is the key component of the authentication system that can resist DeepFake attacks and it will be elaborated in this Section.

A. Overview of the SA-DTH-Net

The network structure of the proposed SA-DTH-Net is given in Fig. 6. Our network is composed of two parts, i.e. the low-level Fundamental lip Feature Extraction subnet (FFE-Net in short) and the high-level Representative lip feature extraction and Classification subnet (RC-Net in short). The FFE-Net aims to emphasize the lip motion characteristics and to reduce the influence of the static lip shape and appearance during authentication. This design is based on the fact that static information is vulnerable to Deepfake attacks [27]. The RC-Net is designed to extract high-level representation lip features for authentication. The high-level features not only provide a high discriminative power against human imposters, but also provide good representation to describe the client's talking habit. Based on the high-level features, a two-class (i.e. acceptance or rejection) classification network is then applied in the RC-Net. There are two distinguishing features in the proposed SA-DTH-Net, which makes it more robust against DeepFake attacks than traditional methods: i) A motion-only lip feature extraction subnet that extracts the dynamic information of the lip movements describing the user's talking habit but removes the static information of the lip region (i.e. the lip shape and appearance), which is vulnerable to DeepFake attacks; ii) A multi-task learning scheme that enables the final features for classification to have a high discriminative power against imposters while providing a comprehensive representation of the client's talking style.

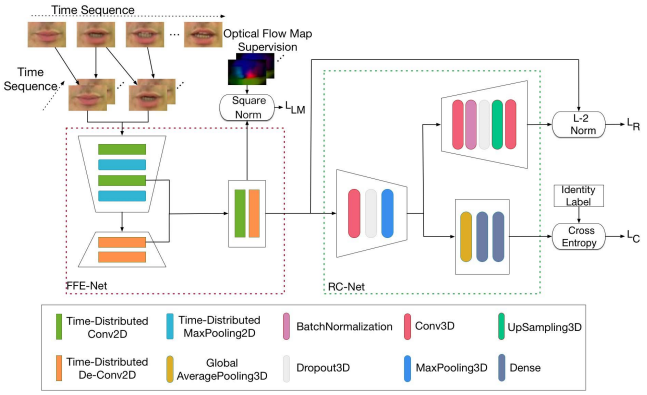


Fig. 6. The network structure of the proposed SA-DTH-Net.

The details of the FFE-Net and the RC-Net will be elaborated next.

B. FFE-Net (Fundamental Lip Feature Extraction Subnet)

As static lip shape and appearance can be easily obtained in the DeepFake generated videos, the FFE-Net attempts to remove the vulnerable static information and to emphasize the dynamic lip motion information. The optical flow [58]–[61] between two successive frames is used as a reference for describing the dynamic information as it has high discriminative power in visual speaker authentication [15]. Inspired by the flowNet [62], the proposed FFE-Net adopts a fully convolutional network structure as illustrated in Fig. 7.

As shown in Fig. 7 (a), two successive lip images with the size of W (width) by H (height) are stacked along the channel dimension and fed into the network. Assume that the lip images are in RGB format, the network input tensor is of the size ($H \times W \times 6$). The network output contains the motion information of every pixel in the lip image in both the horizontal and vertical dimensions and is of the size ($H \times W \times 2$). The optical flow map between the two frames is used as a reference to guarantee that the network only extracts motion information. To better fit the optical flow map, a hierarchical structure is employed and the FFE-Net can be divided into two stages, i.e. the contraction stage and the expansion stage. In the contraction stage, four convolutional layers are used to extract useful features at different resolution. Following each of the first three convolutional layer, a max-pooling layer with a stride of 2 by 2 in the spatial domain are employed to downsample the size of the feature map by half in both dimensions. In the expansion stage, three deconvolutional layers (the deconvolutional operation is illustrated in Fig. 7 (b)) are used to generate the expanded feature map with the spatial size doubled in both dimensions.

The motion maps at various scales are generated by the output network. At the lowest resolution, i.e. $W/8$ by $H/8$, a convolutional layer is used to transform the feature map F4 into the motion map at the corresponding scale. Then, a deconvolutional layer is applied to construct a feature map with double size in both dimensions and to transmit the motion information from a lower scale to a higher scale. By integrating the feature maps at the corresponding scales in both the contraction and expansion networks by the addition

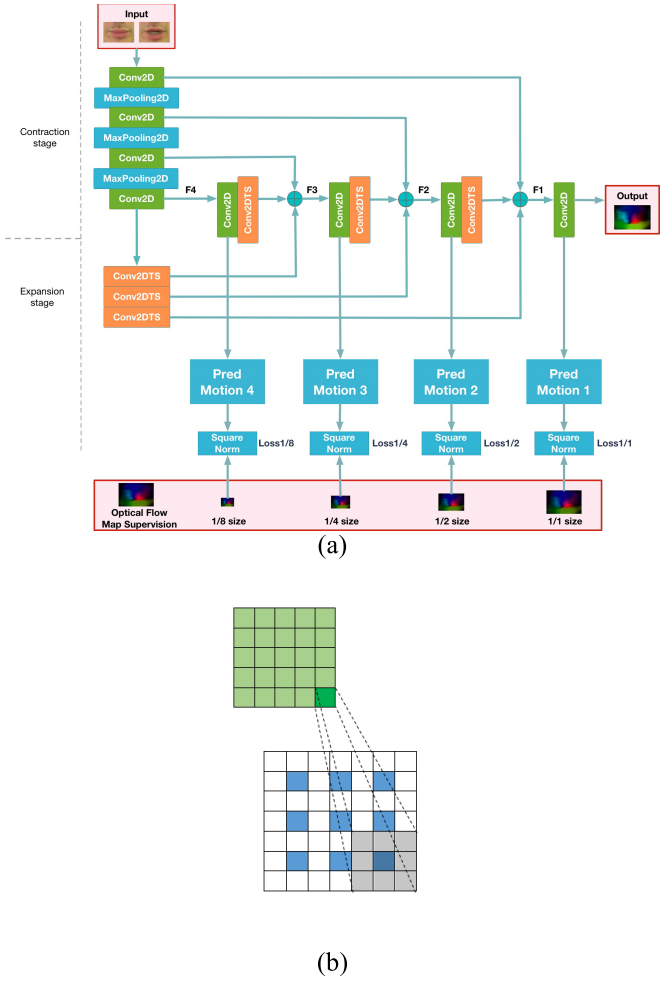


Fig. 7. (a) Flowchart of the FFE-Net; and (b) Illustration of the deconvolution (Transposed convolution) operation with kernel size of 3×3 and stride of 2×2.

operation, the feature map F3 with a higher resolution of $W/4$ by $H/4$ is generated. The motion map in this scale can also be obtained by applying a convolutional layer. Similar operations are repeated three times and the predicted motion features at each scale can be obtained. Finally, the motion map at the original scale is adopted as the network output. As suggested in [62], the reference optical flow map at various scales, i.e. (1/8, 1/4, 1/2, 1) in both dimensions, are used as the supervision information for the predicted motion maps at each scale accordingly.

It should be noted that FFE-Net performs dynamic feature extraction on each pair of successive lip images and thus the time-distributed modules are adopted in this subnet. The final input and output of the FFE-Net are the lip image sequence pronouncing a specific word and the dynamic feature map sequence, respectively.

C. RC-Net (Representative Lip Feature Extraction and Classification subnet)

The RC-Net extracts high-level representation features and verifies the speaker's identity. The network structure of the RC-Net is given in Fig. 8. Three key components, i.e. the feature extraction module, the reconstruction branch and the classification branch, are included in

the RC-Net. In the feature extraction module, a hierarchical structure in the spatial domain is adopted to describe the characteristics of the lip motion at various scales. Four convolutional layers are used to extract features at different resolution and three max-pooling layers with a stride of 2 by 2 in the spatial domain is used for down-sampling. Considering that the network input is the lip feature map sequence, 3D network elements including 3D convolutional layers, 3D max pooling layers, etc., are employed to depict the spatiotemporal lip dynamics.

As shown in the example in Subsection IIB, DeepFake videos are much more similar to the client's videos compared to the human imposter's videos. To enhance the authentication performance against DeepFake attacks, the extracted representation features need to have a high discriminative power in authentication and a good representation ability to depict the client's talking habit. Hence, in the RC-Net, two branches, i.e. the reconstruction and the classification branches, are designed to endow the extracted feature with the above abilities. The reconstruction branch aims to reconstruct the motion map from the extracted features and its structure corresponds to that of the feature extraction part, i.e. with four convolutional layers and three upsampling layers. The classification branch performs a two-class classification with the aid of two fully connected layers and one global average pooling (GAP) layer. Two kinds of supervision information are applied to the RC-Net, including the lip motion map which guides the reconstruction branch and the label information which guides the classification branch.

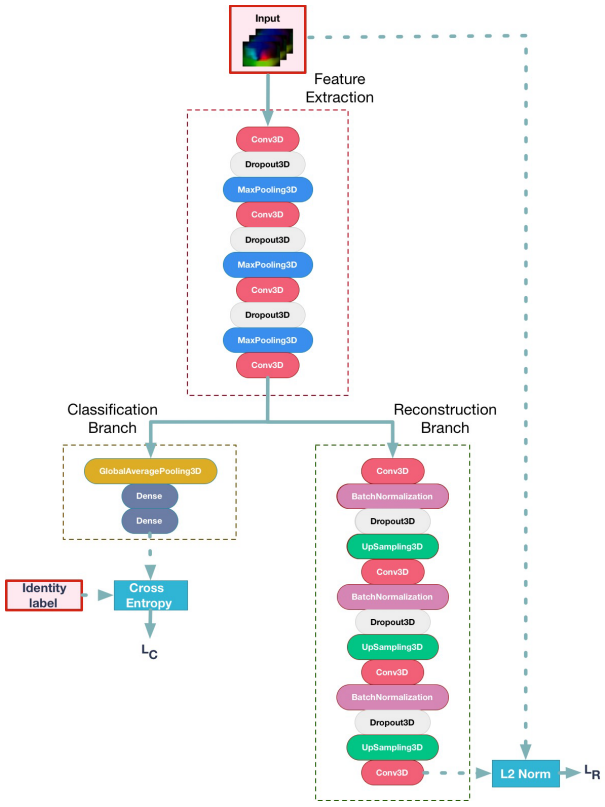


Fig. 8. Flowchart of the RC-Net.

D. Network Optimization by Multi-Task Learning

Multi-Task Learning (MTL) [63] is a widely used machine learning technique where a series of tasks with certain commonalities and differences are learned together. In the proposed SA-DTH-Net, there are three tasks to be fulfilled, including: i) to derive fundamental feature maps that only contain the lip motion information; to extract high-level representation features which can not only ii) depict the client's talking style but iii) be highly discriminative against human imposters. According to the above three tasks, three kinds of loss functions are designed, i.e. the lip motion loss (L_{LM}) in the FFE-Net, the reconstruction loss (L_R) and the classification loss (L_c) in the RC-Net. L_{LM} is the average end-point error between the ground truth optical flow vector map, i.e. $G = \{G_X(t,i,j),\ G_Y(t,i,j)\}, 1 \le t \le T-1, 1 \le i \le H, 1 \le j \le W$, and the predicted fundamental feature map, i.e. $P = \{P_X(t,i,j), P_Y(t,i,j)\}, 1 \le t \le T-1, 1 \le i \le H, 1 \le j \le W$ as shown in (1), at the bottom of the page, where T denotes the length of the original input sequences and (H, W) is the size of every frame. The subscripts X and Y denote the feature map in the horizontal and vertical dimensions, respectively, hereinafter. L_R is the mean squared error, describing the difference between the original predicted fundamental feature map P and the reconstructed feature map $R = \{R_X(t,i,j), R_Y(t,i,j)\}, 1 \le t \le T-1, 1 \le i \le H, 1 \le j \le W$ as shown in (2), at the bottom of the page. L_c describes the classification error in term of the cross-entropy loss function as shown in (3), at the bottom of the page, where I_g and I_p denote the identity vector of the ground truth and the prediction, respectively. The overall loss function that is used to optimize the SA-DTH-Net is the weighted sum of the three loss functions as shown in (4), at the bottom of the page, where α, β and γ are the weights balancing the importance of each loss.

$$L_{LM} = \frac{1}{(T-1)\times H\times W}\sum_{t=1}^{T-1}\sum_{i=1}^{H}\sum_{j=1}^{W}\sqrt{(P_X(t,i,j)-G_X(t,i,j))^2+(P_Y(t,i,j)-G_Y(t,i,j))^2} \tag{1}$$

$$L_R = \frac{1}{2\times(T-1)\times H\times W}\sum_{t=1}^{T-1}\sum_{i=1}^{H}\sum_{j=1}^{W}[(R_X(t,i,j)-P_X(t,i,j))^2+(R_Y(t,i,j)-P_Y(t,i,j))^2] \tag{2}$$

$$L_C = -\sum\nolimits_k I_g(k)\log I_p(k) \tag{3}$$

$$L = \alpha L_{LM} + \beta L_R + \gamma L_C, \quad s.t.\ \alpha+\beta+\gamma = 1 \tag{4}$$

Then the partial derivative of the loss function with respect to the network weights can be derived as:

$$\frac{\partial L}{\partial W_{FFE-Net}} = \alpha \frac{\partial L_{LM}}{\partial W_{FFE-Net}} + \beta \frac{\partial L_R}{\partial W_{FFE-Net}} + \gamma \frac{\partial L_C}{\partial W_{FFE-Net}} \quad (5)$$

$$\frac{\partial L}{\partial W_{F-ext}} = \beta \frac{\partial L_R}{\partial W_{F-ext}} + \gamma \frac{\partial L_C}{\partial W_{F-ext}} \quad (6)$$

$$\frac{\partial L}{\partial W_{Re}} = \beta \frac{\partial L_R}{\partial W_{Re}} \quad (7)$$

$$\frac{\partial L}{\partial W_{Cl}} = \gamma \frac{\partial L_C}{\partial W_{Cl}} \quad (8)$$

where $W_{FFE-Net}$, W_{F-ext}, W_{Re} and W_{Cl} denote the weights of FFE-Net, Feature-Extraction subnet, Reconstruction-branch, and Classification-branch in RC-Net, respectively.

In the optimization procedure, the Adam algorithm [64] is adopted to reduce the overall MTL loss function.

E. Implementation

As shown in Fig. 5, the input of the SA-DTH-Net is the user's lip image sequence when pronouncing a specific word and the output after thresholding is a binary decision indicating whether the user matches the claimed identity. The implementation of the proposed SA-DTH-Net involves three stages, i.e. the training stage, the evaluation stage and the test stage. Details of the above stages are elaborated as follows.

Training Stage: For each client pronouncing a specific word, $N_{T,c}$ client's lip image sequences are used as the positive samples and $N_{T,i}$ lip image sequences from the other speakers are used as the negative samples. To reduce the computational costs, time-distributed convolution, de-convolution and pooling modules are adopted in the FFE-Net to process every two successive image pairs in the lip image sequence simultaneously. The optical flow maps are pre-computed for each training sample and are used to calculate the lip motion loss (L_{LM}). Both the reconstruction and classification branches are activated and the corresponding reconstruction loss (L_R) and classification loss (L_c) are calculated. Then, the Adam algorithm is employed to optimize the network based on the overall loss function. Note that in order to speed up the training process, the SA-DTH-Net for the first word is learned from scratch with random initialization and the networks for the other words are trained by fine-tuning that of the first word.

Evaluation stage: For each client model, $N_{E,c}$ client's lip image sequences are used as the positive samples and $N_{E,i}$ lip image sequences from the other speakers are used as the negative samples. Then the sequences are directly fed into the FFE-Net to generate the fundamental lip motion feature maps without computing the optical flow maps. In the RC-Net, the reconstruction branch is deactivated and the corresponding outputs are analyzed. A threshold θ is then obtained when the equal error rate (EER) is reached.

Test Stage: For an unknown lip image sequence pronouncing a specific word, the corresponding SA-DTH-Net is selected according to the claimed identity and the prompt word. Similar to that in the evaluation stage, the reconstruction branch is not activated and the authentication result is determined by the output of the classification branch. When the output is greater/smaller than the threshold θ, the sequence is recognized as a client/imposter sample.

Based on the word-level authentication results, a final sentence-level authentication decision can be made by voting. If the acceptance votes are dominant, i.e. the number of acceptance votes is great than half of the total votes, accept the request, or reject it otherwise. Details of the proposed network is given in Appendix and we also provide the source code in https://github.com/chenzhao-yang/lip-based-anti-spoofing.

TABLE I
ALL THE WORDS FOR ANALYSIS IN THE GRID DATASET

Category	Words
command	bin, set, lay, place
color	red, white, blue, green
number	zero, one, two, three, four, five, six, seven, eight, nine
adverb	again, please, soon, now

IV. EXPERIMENTS AND DISCUSSIONS

A. Experiment Setup

Dataset: To evaluate the authentication performance of the proposed method, the widely used GRID [65] dataset and MOBIO [66] dataset were used. The GRID dataset is composed of thirty-three speakers and every speaker is asked to speak 1,000 sentences in the following format: "command + color + preposition + letter + number + adverb", e.g., "Place blue at F 9 now". Considering that pronouncing the prepositions and letters usually lasts for very few frames (less than 10), these words cannot provide sufficient information to describe the client's talking style and thus they are excluded from the vocabulary to be analyzed, i.e. four words are analyzed in each sentence. Table I lists all the words under investigation.

The MOBIO dataset consists of bimodal AV data taken from 152 people. Compared with the GRID dataset, it has a much larger vocabulary; however, most words in the vocabulary were pronounced very few times, which cannot provide sufficient information to learn the speaker's talking style. Hence, a subset of MOBIO dataset is investigated in our experiment, which contains six digits, i.e. "two", "four", "five", "seven", "eight", "nine", and 59 speakers who pronounced all of the above digits for at least twenty times. Note that only the word-level authentication experiments are performed on the MOBIO dataset and for both the GRID and MOBIO datasets, only the visual information, i.e. the lip image sequence, is used for speaker authentication.

DeepFake manipulation methods: Three kinds of widely used face-swapping manipulation methods were considered, including *Faceswap* (*FS*) [25], *Deepfacelab-Quick96* (*DFL*) [37] and *Faceswap-GAN* (*FS-GAN*) [26]. All these methods adopted the autoencoder architecture with different implementation and in [26], an adversarial loss was incorporated to improve the quality of the generated face images. In addition,

a recent lip-sync (*LS*) manipulation method [40] was also investigated.

Experiment settings: All samples from 75% of the speakers (24 in GRID and 44 in MOBIO) were randomly selected as the user set and those of the remaining 25% speakers (8 in GRID and 15 in MOBIO) were used as the attacker set. Based on each face-swapping manipulation method, the corresponding face-swapping models for a specific user-attacker pair were trained using three randomly selected videos (containing 200–300 face images in total) for both the user and the attacker. Hence, for each manipulation method, there are overall $24 \times 8 = 192$ models and $44 \times 15 = 660$ models trained for GRID and MOBIO, respectively. For the lip-sync manipulation method, the pretrain model in [40] was used directly. Similarly, three videos (containing 200–300 face images in total) for the user were randomly selected from the training set. Using these user's facial images, the lip-sync model [40] can generate the user's fake videos by synthesizing the mouth region to match the attacker's speech.

In the word-level authentication experiments, for each speaker in the user set, a client model for the speaker was trained from the training samples and a corresponding threshold T was calculated from the evaluation samples. The training/evaluation/test data division is illustrated in Fig. 9. In the training stage, $N_{T,c}$ and $N_{T,i}$ were set as 60% samples of the client and imposters, respectively. In the evaluation stage, $N_{E,c}$ and $N_{E,i}$ were set as half (20%) of the remaining samples of client and the remaining 40% samples of imposters, respectively. In the test stage, the remaining 20% samples of the client were adopted as client samples and two kinds of negative samples were collected, including: i) human imposter samples: all the speech videos pronouncing the same word by the speakers in the attacker set were taken as human imposter samples; and ii) DeepFake samples: for face-swapping methods, based on the client-attacker model, generate the face-swapping DeepFake samples from the human imposter samples by swapping the face of the speaker/attacker with that of the client. For the lip-sync method, we generate the lip-sync Deepfake samples from the audio information of the attack's samples. Hence, in the test set, the number of human imposter samples is equivalent to that of the DeepFake samples using each manipulation method.

Evaluation Metric: In our experiments, the Lausanne protocol [67] was employed. Based on the client model obtained in the training stage and the threshold θ determined in the evaluation stage, the false accept rate (*FAR*) representing the ratio of negative samples (human imposter or DeepFake) being falsely authenticated and the false rejection rate (*FRR*) representing the ratio of client samples being falsely denied, were calculated in the test set. Then the half total error rate (HTER, denoted as $\hat{H}$), which was calculated as $\hat{H} = (FRR+FAR)/2$, was used to evaluate the authentication performance.

It should be noted that to avoid the impact of the selection of training samples on authentication performance, all the experiments were repeated for five rounds with different training/test division and the average value of $\hat{H}$ was recorded.

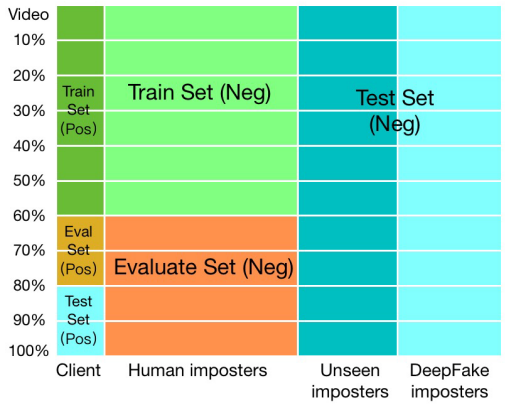


Fig. 9. Division of samples for a specific word.

B. Parameter Selections in Multi-Task Learning

There are three preset parameters, i.e., α, β and γ, in the proposed loss function in (4), which aim to balance the influence among the lip motion loss, the reconstruction loss and the classification loss. To achieve the optimal authentication performance, different parameter selections in MTL are investigated on GRID and their corresponding authentication results in the form of the average $\hat{H}$ of all the words investigated in the GRID dataset are given in Table II. The subscripts *hm*, and *fake* denote the negative samples for human imposters and DeepFake samples, respectively. Note that γ is not listed in the heatmap because when α and β are selected, γ is determined as $1 - \alpha - \beta$.

From the heatmaps in Table II, it is observed that i) when α is set too small (e.g. 0.1), the system performs well defending against human imposters (i.e. $\hat{H}_{hm}$ is small) while is more vulnerable to DeepFake attacks (i.e. $\hat{H}_{fake}$ are large). It is because in this case, the motion map supervision is weak and the authentication network tends to exploit static information which can better differentiate the client and the human imposters; ii) when $\alpha \in \hat{\mathrm{E}}\ [0.3, 0.5]$ and $\beta \in \hat{\mathrm{E}}\ [0.2, 0.3]$, acceptable authentication results ($\hat{H}_{hm} < 4$ and $\hat{H}_{fake} < 6$) can be achieved, which has shown that the proposed approach is not that sensitive to the exact parameter setting. The optimal α, β and γ are therefore set to 0.5, 0.25 and 0.25, respectively, for the best authentication performance.

C. Robustness Tests

To avoid the authentication performance been affected by irrelevant factors, i.e. image resolution, compression ratio of the image, etc., the following experiment has been performed. Each test sample is manipulated by three kinds of post-processing steps, including: i) random compression (with a random compression ratio ranging from 0.5 to 1); ii) random rescaling (the image is rescaled by a random factor ranging from 0.8 to 1.2 and then resized to its original size) and iii) random resolution reduction (the resolution of the image is randomly resized to 1/2 in both the horizontal and the vertical dimensions) Note that only the test samples are manipulated, and no modification has been done on the trained model. The authentication performance for both the original samples and the noisy samples after post-processing is given in Table III.

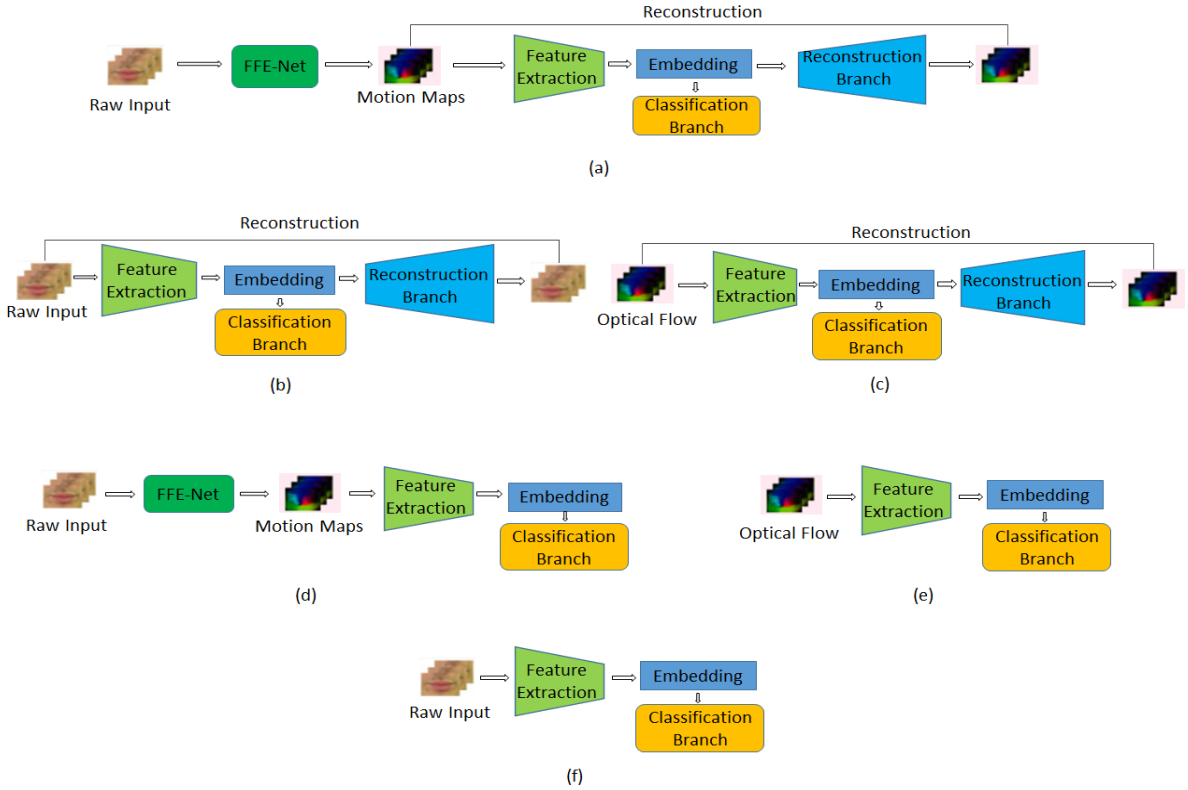
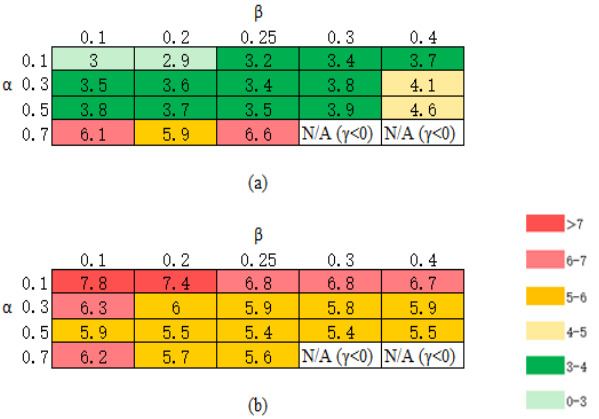


Fig. 10. Network structures for models in Table IV, where the structures for model i to vi are shown in (a) to (f), respectively.

TABLE II

AVERAGE WORD-LEVEL (A) $\hat{H}_{hm}$ (B) $\hat{H}_{fake}$ USING VARIOUS PARAMETER SELECTIONS ON GRID



The subscripts *hm*, *fs*, *dfl*, *gan*, *ls* denote the negative samples for human imposters, DeepFake samples using *FS*, *DFL FS-GAN* and *LS*, respectively. From the table, it is observed that for noisy samples, the detection performance will be degraded by about 1%-2% in $\hat{H}$. Considering the information loss caused by post-processing, the proposed network has been shown to be insensitive to the visual quality of the talking video. Moreover, the performance degradation observed in the GRID dataset captured under laboratory condition is more pronounced than that in the MOBIO dataset because the video samples in the GRID dataset are more similar in quality, making the model trained on the GRID dataset less able to generalize.

TABLE III

ROBUSTNESS TEST: THE AUTHENTICATION RESULTS FOR THE ORIGINAL TEST SAMPLES AND THE TEST SAMPLES AFTER THE POST PROCESSING STEPS

Dataset	Test Samples	*FRR*	$\hat{H}_{hm}$	$\hat{H}_{fs}$	$\hat{H}_{dfl}$	$\hat{H}_{gan}$	$\hat{H}_{ls}$
GRID	Original	0.4	0.6	3.3	3.2	3.5	2.9
	After post processing	1.4	1.7	5.8	5.4	5.5	4.1
MOBIO	Original	2.8	2.4	5.0	5.2	5.3	7.4
	After post processing	3.7	2.8	6.6	5.9	6.4	8.9

D. Ablation Study

In order to comprehensively investigate the authentication performance of the proposed network structure, two ablation studies have been conducted on the GRID dataset. The effectiveness of the FFE-Net and the RC-Net are analyzed separately as follows.

1) Effectiveness of FFE-Net: To examine the effectiveness of the FFE-Net, six variants of network structures have been analyzed, including: i) the proposed model with the FFE-Net and the Reconstruction Branch (RB) as the baseline; ii) the FFE-Net is removed and the lip image sequence is directly fed into the RC-Net; and iii) the FFE-Net is removed and the optical vector flow map sequence calculated from the lip image sequence is used as the input of the RC-Net; iv) the RB is removed and v) the FFE-Net and the RB are both removed as the raw model and the optical vector flow map sequence is used as the input and vi) the raw model. The authentication results using these networks (with the

TABLE IV

AVERAGE WORD-LEVEL AUTHENTICATION PERFORMANCE ON GRID USING VARIOUS NETWORK STRUCTURES

Model	*FRR*	$\hat{H}_{hm}$	$\hat{H}_{fs}$	$\hat{H}_{dfl}$	$\hat{H}_{gan}$	$\hat{H}_{ls}$
i: Proposed Model	2.7	3.5	**5.9**	**5.8**	**6.0**	**3.9**
ii: w/o FFE-Net, raw input	**1.9**	**3.1**	8.9	9.3	9.5	5.2
iii: w/o FFE-Net, optical flow input	4.1	4.4	7.7	7.6	7.4	5.2
iv: w/o RB	2.8	3.8	7.6	7.9	7.9	4.6
v: Raw Model, optical flow input	4.3	4.9	8.1	7.9	8.3	5.3
vi: Raw Model	2.2	3.5	13.5	10.9	12.9	6.7

structures shown in Fig. 10) are given in Table IV. From the table, it is observed that directly feeding the original lip image sequence into the RC-Net (model ii) achieves the lowest *FRR* and $\hat{H}_{hm}$ in the test samples. It is mainly because in order to differentiate the client and the human imposter, the static lip shape and appearance play an important role and when such information is removed (e.g. in model i and iii), the authentication performance is degraded to some extent. However, when confronted with DeepFake attacks, model ii will falsely accept many DeepFake samples and thus has large $\hat{H}$ values for all the three kinds of attacks because the static information in DeepFake samples is much closer to that of the client. For the models without RB (model iv and vi), similar observations can be made in the table, which demonstrates that using the FFE-Net alone can help prevent Deepfake attacks. Comparing the two models that use the dynamic information and RB, i.e. model i and iii, the proposed model with FFE-Net allows end-to-end training and has the following advantages: i) it greatly reduces the processing time in computing the dynamic features and ii) the multi-task learning strategy allows the classification and the reconstruction tasks to influence the weights in the FFE-Net, which helps to improve the discriminative power of the dynamic features extracted by the FFE-Net (an $\hat{\text{H}}$ gain of 0.9% against human imposters and about 1.3%-1.8% against DeepFake attacks). Note that even without the Reconstruction Branch (RB), the FFE-Net can extract more robust and discriminative feature than the optical flow map (comparing model iv and v).

2) Effectiveness of Reconstruction Branch: To evaluate the effectiveness of the reconstruction branch, the authentication results by three pairs of models, i.e. model i vs model iv (with FFE-Net), model ii vs model vi (without FFE-Net, raw input), model iii vs model v (without FFE-Net, optical flow map as input), are analyzed. From Table IV, it is observed that the models with the reconstruction branch (i.e. models i, ii and iii) always perform better than their counterparts without the reconstruction branch (models iv, vi and v) in both authentication tests. Furthermore, the reconstruction branch is more useful at preventing DeepFake attacks than against human imposters. It is mainly because with the reconstruction branch, the extracted high-level lip features will have both the discriminative ability to differentiate between the client and the human imposters and the representation ability to describe the talking style of the client. Hence, the reconstruction branch enhances the robustness of the model against computer-generated videos which are more similar to the client's videos than the human imposter's videos.

TABLE V

AVERAGE WORD-LEVEL /SENTENCE-LEVEL(HIGHLIGHTED) AUTHENTICATION PERFORMANCE ON GRID BY THE PROPOSED MODEL TRAINED WITH VARIOUS NUMBER OF CLIENT SAMPLES. THE DEFAULT SETTING IS TO USE 60% OF THE TOTAL NUMBER OF CLIENT

No. of samples ($N_{T,c}$)	$\hat{H}_{hm}$	$\hat{H}_{fs}$	$\hat{H}_{dfl}$	$\hat{H}_{gan}$	$\hat{H}_{ls}$
10	7.4/**2.4**	12.2/**9.1**	12.4/**9.4**	12.6/**9.5**	8.7/**7.6**
20	4.7/**1.3**	9.7/**6.7**	9.4/**6.7**	9.8/**7.2**	7.0/**5.8**
40	3.7/**0.8**	7.0/**4.0**	6.9/**3.9**	7.3/**4.1**	4.4/**3.2**
Default (about 60)	3.5/**0.6**	5.9/**3.3**	5.8/**3.2**	6.0/**3.5**	3.9/**2.9**

E. Performance Comparisons Vs Number of Training Samples

For most VSA systems, the client has to prerecord a number of samples to train the client model. In Table V, the word/sentence-level authentication results on the GRID dataset by the proposed SA-DTH-Net with different number of client training samples, i.e. $N_{T,c}$, were listed. From the table, it can be observed that the client's behavioral lip feature, i.e. his/her talking style, is better described by having more training samples, and this improve the authentication performance. In addition, compared with the human imposters, the DeepFake samples are more difficult to detect under limited number of training samples in sentence-level authentication.

F. Comparisons With State-of-the-Art VSA Approaches

To evaluate the authentication performance of the proposed method, three state-of-the-art approaches, i.e. Zhang's [19] (handcraft LOCP feature based), Cheng's [21] (CNN based) and Sun's [22] (CNN-based) approaches, are compared. Note that for a fair comparison, these approaches are implemented at word-level and the sentence-level results are derived by majority voting.

The authentication results on both the GRID and MOBIO datasets are listed in Table VI. From the table, it is observed that the proposed method can obtain comparable authentication results against human imposters and achieve much better results against DeepFake attacks. Moreover, it is observed from the table that compared with the existing VSA methods, the proposed approach achieved a comparable false rejection rate (FRR) for the client and the reduction of $\hat{H} = (FAR+FRR)/2$ against DeepFake attacks is mainly due to the large reduction in *FAR* value. The above observation has demonstrated that the proposed method can effectively prevent DeepFake attack, which is a great threat to existing VSA systems. Note that the authentication performance by Zhang's method on the MOBIO dataset degrades greatly due to the great variations in talking style, capturing device, etc., which agrees with the observations in [20].

G. Comparisons With State-of-the-Art DeepFake Detection Approaches

Here, we compare the DeepFake detection performance of the proposed approach with three state-of-the-art DeepFake

TABLE VI

AUTHENTICATION PERFORMANCE BY VARIOUS VSA METHODS

Dataset	Method	*FRR*	$\hat{H}_{hm}$	$\hat{H}_{fs}$	$\hat{H}_{dfl}$	$\hat{H}_{gan}$	$\hat{H}_{ls}$
GRID	Zhang's	**0.2**	0.4	12.8	10.6	22.5	14.4
	Cheng's	0.2	**0.4**	13.5	14.3	12.9	6.7
	Sun's	0.5	0.9	14.2	10.7	11.7	6.3
	Proposed	0.4	0.6	**3.3**	**3.2**	**3.5**	**2.9**
MOBIO	Zhang's	12.9	14.3	27.3	23.0	33.0	22.2
	Cheng's	**2.6**	2.5	12.1	11.5	10.8	9.7
	Sun's	4.1	3.0	13.2	10.4	14.4	10.6
	Proposed	2.8	**2.4**	**5.0**	**5.2**	**5.3**	**7.4**

detection methods, including: two CNN-based methods, i.e. the MesoNet [31] and the XceptionNet [50], and one biometric feature based method, i.e. Agarwal's [56]. Note that the CNN-based DeepFake detection methods are image-based and take all the face region as input. The sentence-level (for GRID) and word-level (for MOBIO) detection results are obtained by majority voting over all the face images in the video.

The dataset division, training & evaluation strategy and cross-validation scheme introduced in Subsection IVA is used. Specifically, for the CNN-based approaches, the images from all the samples in the training set were adopted as the pristine samples and their corresponding fakes images were generated by *FS* [25], *DFL* [37], *FS-GAN* [26] and *LS*[40]. To evaluate the generalization and transferability of the detection models, models trained with various selections of negative training samples (fake images) were investigated. After training, the images from all the samples in the evaluation set and their DeepFake counterparts were used to obtain the optimal threshold θ to compute the *EER*. Finally, the detection results for the entire sentence (GRID) or word (MOBIO) were derived by voting from all the frames. For Agarwal's method [56], all the experiment settings are exactly the same as our approach. Note that the images in the GRID dataset are of similar quality, and thus to avoid information leakage due to image resolution, the data augmentation procedure in [52], i.e. randomly reduce the resolution of the video to 1/4 of its original size, was applied on all the samples in the training/evaluation/test sets for all the four DeepFake detection approaches investigated (MesoNet, XceptionNet, Agarwal's and ours).

Table VII shows the detection results of the four approaches. To compare the detection results, three DeepFake detection scenarios are defined as follows:

Detection ofa specific attack: In the training, evaluation and test stage, the negative/fake samples are generated by only one particular DeepFake manipulation technique. The authentication results are displayed in light green in the table. It is observed that both MesoNet and XceptionNet is able to achieve the best detection performance. The fake samples by FS-GAN are the most difficult to be detected.

Detection of multiple attacks: The detection models were trained and evaluated based on the negative/fake samples generated by all four kinds of DeepFake methods. The test samples were generated by one of these attacks and the authentication results are displayed in blue in the table. Compared with those in the previous scenario, the detection performance

TABLE VII

DETECTION PERFORMANCE ON THE GRID AND MOBIO DATASETS, WHERE *a/b* DENOTES THE DETECTION RESULTS FROM THE GRID AND MOBIO DATASET, RESPECTIVELY. *DFS, DDFL* AND *DGAN* DENOTE THE FAKE SAMPLES GENERATE BY *FS, DFL* AND *FS-GAN*, RESPECTIVELY. THE AUTHENTICATION RESULTS IN THE "*Detection of a Specific Attack*", "*Detection of Multiple Attacks*" AND "*Detection of Unknown Attacks*" SCENARIOS ARE DISPLAYED IN LIGHT GREEN, BLUE AND DARK BLACK, RESPECTIVELY

Detection Method	Negative train data	$\hat{H}_{fs}$	$\hat{H}_{dfl}$	$\hat{H}_{gan}$	$\hat{H}_{ls}$
MesoNet	D_{fs}	3.5/2.1	8.2/23.9	39.1/34.9	22.3/17.5
XceptionNet		1.9/1.5	6.7/25.8	41.2/37.1	19.9/24.7
MesoNet	D_{dfl}	7.9/5.1	1.8/2.1	36.8/47.6	15.6/14.3
XceptionNet		6.2/6.5	1.6/2.9	42.4/44.2	16.0/13.6
MesoNet	D_{gan}	21.9/10.1	31.6/23.9	4.3/5.1	11.2/15.3
XceptionNet		21.5/15.5	31.1/15.9	5.3/4.9	20.5/10.2
MesoNet	D_{ls}	19.8/14.7	12.1/9.5	32.8/26.2	2.3/5.1
XceptionNet		24.4/13.3	10.9/17.7	30.1/26.7	1.8/1.4
MesoNet	$D_{fs}+D_{dfl}$ $+D_{gan}+D_{ls}$	7.0/3.9	12.6/10.1	4.6/11.0	8.9/10.3
XceptionNet		6.6/4.5	8.2/8.7	15.7/8.3	7.4/7.7
Agarwal's	*None*	8.1/11.2	7.6/10.5	7.4/12.8	10.3/15.1
SA-DTH-Net	*None*	5.7/5.0	5.5/5.2	5.7/5.3	3.6/7.4

drops greatly, e.g. for the *FS-GAN* detection (with an average $\hat{H}_{GAN}$ increase of about 7.5%).

Detection of unknown attacks: The negative/fake data generation method in the test set is different from that in the training and evaluation set, i.e. the attack approach is unknown at the defender side. The authentication results are displayed in dark black in the table and the following observation can be made: i) For MesoNet and XceptionNet, the detection errors further increased. Particularly, the models trained from *FS* and *DFL* can detect the samples generated by each other to some extent because *FS* and *DFL* adopt similar network structures (autoencoder without adversarial loss); however, they can hardly detect *FS-GAN* fake samples (very high $\hat{H}_{gan}$ of about 30%-40%, note that 50% means random guess). Meanwhile, models trained from *FS-GAN* cannot accurately detect the traces left by DeepFake manipulations using *FS* and *DFL*. Similar observations can be made between face-swapping and lip-sync Deepfake manipulation methods. The above observation shows that the current DeepFake detection approaches have limited ability to detect different/unknown manipulation, which accords with the opinion in [55]; ii) The two biometric feature based approaches, i.e. the proposed method and the Agarwal's method, usually perform well in this scenario and our approach always achieves the best detection performance, especially in detecting fake videos generated by *FS-GAN*. It is mainly because our network can better model the speaker's talking habit than the handcrafted, statistical feature extracted in [56]. Our approach does not produce large differences in detection error between different DeepFake methods. As the proposed approach checks whether the talking style in the test video is consistent to that of the client, it is hardly affected by the type of DeepFake manipulation method and the visual fidelity of the fake videos. Note that data augmentation in the GRID dataset has increased the detection error by about 2.3% for the proposed approach because it is more difficult to learn the speaker's talking style from videos with low resolution; iii) Even when comparing with

TABLE VIII
THE STRUCTURE OF FFE-NET (TIME_DISTRIBUTION IS APPLIED ON EACH LAYER)

layer name	kernel_size / nodes / strides	output size
Conv2D_1	(3,3) / 64 / 1	15×40×64×64
MaxPooing2D_1	(2,2) / - / 2	15×20×32×64
Conv2D_2	(3,3) / 96/ 1	15×20×32×96
MaxPooing2D_2	(2,2) / - / 2	15×10×16×96
Conv2D_3	(3,3) / 128 / 1	15×10×16×128
MaxPooing2D_3	(2,2) / - / 2	15×5×8×128
Conv2D_4	(3,3) / 256 / 1	15×5×8×256
Deconv2D_3	(5,5) / 128 / 2	15×10×16×128
Deconv2D_2	(5,5) / 96 / 2	15×20×32×96
Deconv2D_1	(5,5) / 64 / 2	15×40×64×64
Flow_deconv2D_3	(5,5) / 2 / 2	15×10×16×2
Flow_deconv2D_2	(5,5) / 2 / 2	15×20×32×2
Flow_deconv2D_1	(5,5) / 2 / 2	15×40×64×2
Pred_conv2D_4	(3,3) / 2 / 1	15×5×8×2
Pred_conv2D_3	(3,3) / 2 / 1	15×10×16×2
Pred_conv2D_2	(3,3) / 2 / 1	15×20×32×2
Pred_conv2D_1	(3,3) / 2 / 1	15×40×64×2

TABLE IX
THE STRUCTURE OF RC-NET

layer name	kernel_size / nodes / strides	output size
Conv3D_1	(3,3,3) / 32 / 1	15×40×64×32
MaxPooling3D_1	(1,2,2) / - / 2	15×20×32×32
Conv3D_2	(3,3,3) / 64 / 1	15×20×32×64
MaxPooling3D_2	(1,2,2) / - / 2	15×10×16×64
Conv3D_3	(3,3,3) / 96 / 1	15×10×16×96
MaxPooling3D_3	(1,2,2) / - / 2	15×5×8×96
Conv3D_en	(3,3,3) / 128 / 1	15×5×8×128
Deconv3D_3	(3,3,3) / 96 / 1	15×5×8×96
Upsampling3D	(1,2,2) / - / -	15×10×16×96
Deconv3D_2	(3,3,3) / 64 / 1	15×10×16×64
Upsampling3D	(1,2,2) / - / -	15×20×32×64
Deconv3D_1	(3,3,3) / 32 / 1	15×20×32×32
Upsampling3D	(1,2,2) / - / -	15×40×64×32
Conv3D_de	(3,3,3) / 2 / 1	15×40×64×2
GAP3D	- / - / -	128
Dense1	- / 128 / -	128
Dense2	- / 2 / -	2

the results under the "Detection of multiple attacks" scenario, our method still outperforms MesoNet and XceptionNet in detecting sophisticated *DFL*, *FS-GAN* and *LS* manipulations even though it was not trained using any DeepFake samples.

V. CONCLUSION AND FUTURE WORKS

In this paper, a lip-based visual speaker authentication method is proposed to defend against both human imposters and DeepFake attacks. The proposed approach can differentiate DeepFake attacks without any prior knowledge of the video-generation method based on the assumption that the attackers only have limited information about the client (a small number of photos, a few talking videos, etc.) and cannot exactly reproduce the client's talking habit when uttering random prompt texts. A new deep neural network, called SA-DTH-Net, is designed to extract information about the client's unique talking habit in order to differentiate the client's lip image sequence against human imposters and DeepFake forgeries. The final authentication result for a speaker uttering a random prompt text can be obtained by integrating all word-level authentication results derived from SA-DTH-Net. Experimental results demonstrated that the proposed approach can successfully reject most DeepFake attacks produced using different manipulation methods. Our approach can therefore provide a feasible solution for universal DeepFake detection in VSA systems.

In our future works, we will try to address the following problems:

i) In the user registration stage, the client is asked to prerecord a number of videos uttering different prompt texts for training the network. However, the client will lose patience if he/she is asked to record too many videos. How to extract highly representative feature to describe the speaker's unique talking habit from very limited training samples is an important topic in our future research.

ii) How to extend the idea from user authentication to a more general DeepFake video detection scenario is another important topic. The authenticity of a video with talking face can be examined based on analyzing the speaker's talking habit. A larger vocabulary will bring more variation in speech content and also cause difficulties in recognizing speech content by lipreading. In addition, compared with those in cooperative authentication scenario with frontal faces, greater variations in head poses make representative talking style feature extraction non-trivial.

APPENDIX

The implementation details of the SA-DTH-Net are as follows: For a lip image sequence representing the user pronouncing a specific word, the following preprocessing steps are employed to normalize the network input of SA-DTH-Net. It is observed that in GRID and MOBIO, the maximum duration when pronouncing a single word does not exceed 16 frames. Hence, a maximum time duration is set to 16 frames and any sequence less than 16 frames is extended to 16 frames by padding. The spatial resolution of each frame is then normalized as $W = 64$ and $H = 40$. Owing to the time-distributed modules, all image pairs between neighbouring frames can be fed into SA-DTH-Net at the same time. The shape of the input is therefore $((16\text{-}1)\times 64 \times 40 \times (3 \times 2)) = (15 \times 64 \times 40 \times 6)$. Details of the FFE-Net and RC-Net are shown in Table VIII and Table IX.

Note that a Batch Normalization [68] layer and a ReLU [69] layer are employed following each convolutional layer, and a dropout rate of 0.5 is applied as well.

REFERENCES

[1] J. Liu-Jimenez, R. Sanchez-Reillo, and C. Sanchez-Avila, "Biometric co-processor for an authentication system using iris biometrics," in *Proc. 38th Annu. Int. Carnahan Conf. Secur. Technol.*, Oct. 2004, pp. 131–135.

[2] V. Conti, C. Militello, F. Sorbello, and S. Vitabile, "A multimodal technique for an embedded fingerprint recognizer in mobile payment systems," *Mobile Inf. Syst.*, vol. 5, no. 2, pp. 105–124, 2009.

[3] Y.-H. Jo, S.-Y. Jeon, J.-H. Im, and M.-K. Lee, "Security analysis and improvement of fingerprint authentication for smartphones," *Mobile Inf. Syst.*, vol. 2016, pp. 1–11, Mar. 2016.

[4] Q. Zhang, H. Li, Z. Sun, and T. Tan, "Deep feature fusion for iris and periocular biometrics on mobile devices," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2897–2912, Nov. 2018.

[5] S. Srisuk, M. Petrou, W. Kurutach, and A. Kadyrov, "A face authentication system using the trace transform," *Pattern Anal. Appl.*, vol. 8, nos. 1–2, pp. 50–61, Sep. 2005.

[6] X. L. Meng, Z. W. Song, and X. Y. Li, "RFID-based security authentication system based on a novel face-recognition structure," in *Proc. Wase Int. Conf. Inf. Eng.*, vol. 1, Aug. 2010, pp. 97–100.

[7] S. R. Arashloo, J. Kittler, and W. Christmas, "Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2396–2407, Nov. 2015.

[8] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, p. 1794—1809, Feb. 2018.

[9] P. Perera and V. M. Patel, "Face-based multiple user active authentication on mobile devices," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1240–1250, May 2019.

[10] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," in *Proc. 4th Int. Conf. Spoken Lang. Processing. ICSLP*, Oct. 1996, pp. 62–65.

[11] T. Wark, D. Thambiratnam, and S. Sridharan, "Person authentication using lip information," in *Proc. TENCON Brisbane - Australia. IEEE TENCON Region 10 Annu. Conf., Speech Image Technol. Comput. Telecommun.*, Dec. 1997, pp. 153–156.

[12] T. Wark, S. Sridharan, and V. Chandran, "The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing.*, Jun. 2000, pp. 2389–2392.

[13] C. C. Broun, X. Zhang, R. M. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2002, pp. 685–688.

[14] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2879–2891, Oct. 2006.

[15] S.-L. Wang and A. W.-C. Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognit.*, vol. 45, no. 9, pp. 3328–3335, Sep. 2012.

[16] X. Liu and Y.-M. Cheung, "Learning multi-boosted HMMs for lip-password based speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 2, pp. 233–246, Feb. 2014.

[17] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Aug. 2002.

[18] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.

[19] C. H. Chan, B. Goswami, J. Kittler, and W. Christmas, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 602–612, Apr. 2012.

[20] J.-Y. Lai, S.-L. Wang, A. W.-C. Liew, and X.-J. Shi, "Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling," *Inf. Sci.*, vol. 373, pp. 219–232, Dec. 2016.

[21] F. Cheng, S.-L. Wang, and A. W.-C. Liew, "Visual speaker authentication with random prompt texts by a dual-task CNN framework," *Pattern Recognit.*, vol. 83, pp. 340–352, Nov. 2018.

[22] J. Sun, S. Wang, and Q. Zhang, "Visual speaker authentication by a CNN-based scheme with discriminative segment analysis," in *Proc. ICONIP*. Cham, Switzerland: Springer, 2019, pp. 159–167.

[23] C.-W. Liao, W.-Y. Lin, and C.-W. Lin, "Video-based person authetication with random passwords," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2008, pp. 581–584.

[24] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3685.

[25] *Faceswap*. Accessed: Apr. 7, 2020. [Online]. Available: https://github.com/deepfakes/faceswap

[26] *Faceswap-GAN*. Accessed: Apr. 14, 2020. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN

[27] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*. [Online]. Available: http://arxiv.org/abs/1812.08685

[28] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[29] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[30] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*. [Online]. Available: http://arxiv.org/abs/1811.00656

[31] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[32] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE CVPR Workshop*, Sep. 2019, pp. 80–87.

[33] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, 1999, pp. 964–966.

[34] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," 2020, *arXiv:2001.00179*. [Online]. Available: http://arxiv.org/abs/2001.00179

[35] S. Lyu, "Deepfake detection: Current challenges and next steps," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[36] B. Bitesize. (2019). *Deepfakes: What are They and Why Would i Make One*. [Online]. Available: https://www.bbc.co.uk/bitesize/articles/zfkwcqt

[37] *Deepfacelab*. Accessed: May 3, 2020. [Online]. Available: https://github.com/iperov/DeepFaceLab

[38] O. Fried, A. Tewari, and M. ZollhoferOhad, "Text-based editing of talking-head video," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 68:1–68:14, 2019.

[39] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, p. 95, Jul. 2017.

[40] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. ACM Multimedia*, 2020, pp. 484–492.

[41] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[42] Z. Xingjie, J. Song, and J.-I. Park, "The image blending method for face swapping," in *Proc. 4th IEEE Int. Conf. Netw. Infrastruct. Digit. Content*, Sep. 2014, pp. 95–98.

[43] S. Mahajan, L.-J. Chen, and T.-C. Tsai, "SwapItUp: A face swap application for privacy protection," in *Proc. IEEE 31st Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Mar. 2017, pp. 46–50.

[44] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, p. 39, 2008.

[45] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[46] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 659–665.

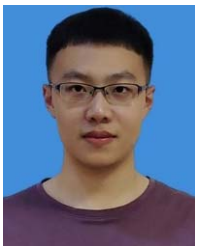
[47] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[49] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Deeply vulnerable: A study of the robustness of face recognition to presentation attacks," *IET Biometrics*, vol. 7, no. 1, pp. 15–26, Jan. 2018.

[50] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.
[51] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," 2019, *arXiv:1909.12962*. [Online]. Available: http://arxiv.org/abs/1909.12962
[52] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*. [Online]. Available: http://arxiv.org/abs/1910.08854
[53] L. Jiang, R. Li, W. Wu, C. Qian, and C. Change Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," 2020, *arXiv:2001.03024*. [Online]. Available: http://arxiv.org/abs/2001.03024
[54] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deepfake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2814–2822.
[55] Z. Akhtar, D. Dasgupta, and B. Banerjee, "Face authenticity: An overview of face manipulation generation, detection and recognition," in *Proc. Int. Conf. Commun. Inf. Process. (ICCIP)*, 2019, pp. 1–9.
[56] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 38–45.
[57] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. ICML*, 2006, pp. 369–376.
[58] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[59] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scandin. Conf. Image Anal.*, 2003, pp. 363–370.
[60] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV*, vol. 3024, 2004, pp. 25–36.
[61] A. Wedel, D. Cremers, T. Pock, and H. Bischof, "Structure- and motion-adaptive regularization for high accuracy optic flow," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1663–1668.
[62] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[63] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn. ICML*, 2008, pp. 160–167.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[65] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustic Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.
[66] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2012, pp. 635–640.
[67] J. Luettin and G. Maitre, "Evaluation protocol for the XM2FDB database (Lausanne Protocol)," in *Communication 98–05*. Martigny, Switzerland: IDIAP, 1998.
[68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
[69] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

Chen-Zhao Yang received the B.Eng. degree in information security from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently pursuing the master's degree with the School of Electric Information and Electronic Engineering. His research interests include computer vision and pattern recognition.



Jun Ma received the B.Eng. degree in information security from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently pursuing the master's degree with the School of Electric Information and Electronic Engineering. His research interests include computer vision and pattern recognition.

Shilin Wang (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001, and the Ph.D. degree from the Department of Computer Engineering and Information Technology, City University of Hong Kong, in 2004. Since 2004, he has been with the School of Electric Information and Electronic Engineering, Shanghai Jiao Tong University, where he is currently a Professor. His research interests include image processing and pattern recognition.

Alan Wee-Chung Liew (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical and electronic engineering from the University of Auckland, New Zealand, in 1993, and the Ph.D. degree in electronic engineering from the University of Tasmania, Australia, in 1997. He worked as a Research Fellow and later as a Senior Research Fellow with the Department of Electronic Engineering, City University of Hong Kong. From 2004 to 2007, he was with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, as an Assistant Professor. In 2007, he joined the School of Information and Communication Technology, Griffith University, as a Senior Lecturer, and currently as an Associate Professor. His current research interests include machine learning, pattern recognition, computer vision, medical imaging, and bioinformatics. He serves on the Organizing Committee or the Technical Program Committee for many international conferences. He also serves as an Associate Editor for several journals, such as the IEEE TRANSACTIONS ON FUZZY SYSTEMS.