SonarGuard: Ultrasonic Face Liveness Detection on Mobile Devices

Dongheng Zhang, Jia Meng, Jian Zhang, Xinzhe Deng, Shouhong Ding, Man Zhou, Qian Wang, *Fellow, IEEE*, Qi Li, *Senior Member, IEEE*, Yan Chen, *Senior Member, IEEE*

***Abstract*—Liveness detection has been widely applied in face authentication systems to combat malicious attacks. However, existing methods purely depending on visual frames become vulnerable once visual perception is not reliable. The emerging face spoof and forge techniques urge the systems to exploit the defensive potential of non-visual modalities. To tackle this challenge, we introduce SonarGuard, a system combining ultrasonic and visual information to achieve robust liveness detection on mobile devices. More specifically, SonarGuard simultaneously extracts micro-doppler signatures from ultrasound reflections and motion trajectories from video frames both corresponding to the user's lip movement. To further confirm the collected ultrasonic and visual information is not derived from malicious audio/video attacks, we consolidate the system via introducing a cross-modal matching mechanism, which demands the inherent consistency between these two modalities. Extensive experiments on a new dataset collected with existing mobile devices demonstrate that the proposed system could achieve average classification error rate of 0.91% under presentation attacks. This result indicates that SonarGuard can boost the security of face authenfication systems in real world usage without additional hardware modification.**

***Index Terms*—Liveness Detection, Ultrasound Signal Processing, Information Fusion.**

I. Introduction

With the development of machine learning and computer vision techniques, the past decade has witnessed the rapidly growing application of face authentication systems. While these systems have gained widespread popularity due to their convenience, the threat caused by malicious attacks has led to severe security concerns. As a result, liveness detection, which verifies whether the captured data is the actual measurement from live person, has been the most important building block to prevent face authentication systems from malicious attacks. Based on the modality utilized, existing liveness detection methods can be generally categorized into two different categories: single-modal methods and multi-modal methods.

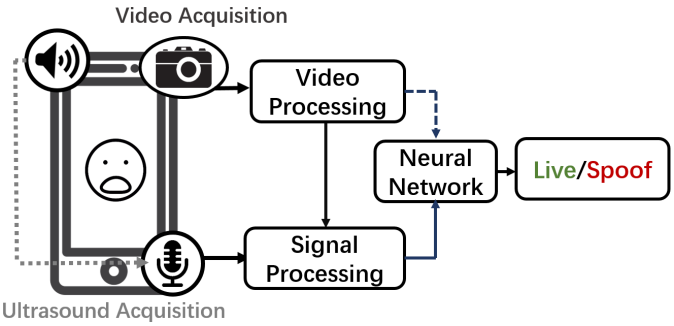


Fig. 1. Overview of SonarGuard. It actively transmits ultrasound signal, simultaneously collect and process ultrasound and video data, and finally outputs the face liveness detection result through the neural network.

Due to its simplicity and efficiency, single-modal methods, which accept video frames recorded by RGB cameras as input, have been widely adopted in current face authentication systems. These methods discover discriminative characteristics unique to attack medium through texture or temporal features. Although these methods achieve accurate liveness detection in some scenarios, they are limited in generalization ability [1]. What is worse, recent white paper figures out the risk the attacker could bypass these liveness detection modules through hardware video injection [2]. More specifically, both texture and temporal based methods are established on the assumption that there are discrepancies between video frames recorded by attack mediums and live users. However, [2] proposes to inject pre-recorded videos directly through the camera serial interface. The injected video frames in this case are identical to the ones recorded by live users, which leads to the failure of all these methods.

Since achieving liveness detection purely depending on the RGB camera videos is vulnerable, a natural solution to enhance the system security is to add other modalities which can provide complementary information. Benefiting from multi-view learning process provided by different modalities, multi-modal methods can achieve impressive performance [3]. How-

This work was supported by National Natural Science Foundation of China under Grant 62201542, fellowship of China Postdoctoral Science Foundation under grant 2022M723069 and the Fundamental Research Funds for the Central Universities.

Dongheng Zhang and Yan Chen are with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China (email: {dongheng, eecyan}@ustc.edu.cn).

Jia Meng, Jian Zhang, Xinzhe Deng and Shouhong Ding are with Tencent Youtu Lab, Shanghai 200235, China. (email: {jeffmeng, timmmyzhang, xinzhedeng, ericshding}@tencent.com). Dongheng Zhang was a research intern at Tencent Youtu Lab.

Man Zhou is with the Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: zhouman@hust.edu.cn).

Qian Wang is with School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (E-mail: qianwang@whu.edu.cn).

Qi Li is with the Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China, and Zhongguancun Laboratory, Beijing 100194, China. (E-mail: qli01@tsinghua.edu.cn).

Shouhong Ding and Yan Chen are corresponding authors.

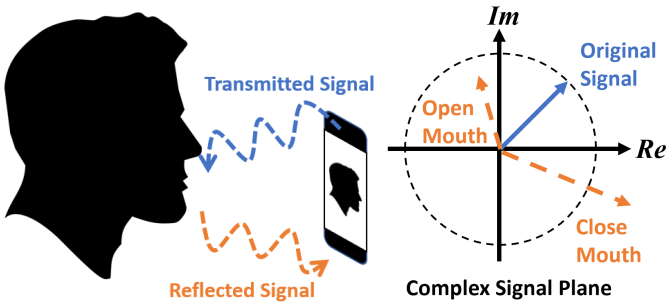
ever, existing multi-modal methods are still limited in vision modalities, i.e, the term "multi-modal" in contemporary works refer to different measurement of a single frame: RGB, depth and infrared image. Since depth and infrared cameras are unavailable in common mobile devices, the applications of existing multi-modal methods are restricted to specialized devices equipped with customized cameras.

To address aforementioned issues, we investigate new modalities for liveness detection on mobile devices. To achieve this, a modality candidate should satisfy the following three requirements:

- **Information complement**: To achieve liveness detection with better performance, the new modality should provide complementary information which may be difficult for existing modality to capture.
- **Hardware reusing**: Designing brand-new hardware for liveness detection is costly, especially for mobile devices. Hence, the new modality should reuse existing hardware to make it possible for large scale real-world deployment.
- **Injection resistance**: Since hardware hacking has been a great challenge for liveness detection, the new modality should be robust when hardware injection into the camera is frequently encountered.

In this paper, we introduce SonarGuard, which exploits the potential of ultrasonic modality for consolidating the security of liveness detection. As shown in Fig. 1, SonarGuard actively transmits ultrasound signal, simultaneously collect and process ultrasound and video data, and finally outputs the face liveness detection result through the neural network. The core of SonarGurad lies in the fact that the propagation of ultrasound signal would be affected by the motion of surrounding reflectors, which is referred as micro-doppler effect [4]. By requiring the user to perform lip movement, the ultrasound reflection would be affected with unique pattern as demonstrated in Fig. 2. On the contrary, the spoof only performs the lip movement on the attack medium, which exerts no effort on the ultrasound signal propagation. The pipeline of SonarGuard is composed of three modules. The ultrasound signal processing module enhances received signal caused by lip movement through a series of beamforming and filtering techniques. The lip motion extraction module recovers motion trajectory from facial landmarks. The trajectory not only facilitates ultrasoic signature segmentation, but also provides a complementary modality to ultrasound. The ultrasonic-visual transformer module aggregates the information from both ultrasound signal and lip landmarks to achieve accurate liveness detection.

Compared with RGB video, ultrasound acquisition is also supported by the majority of mobile devices, which has more promising characteristics. First, ultrasound is much more sensitive to real lip movements, while vision modality might be easily deceived once lip movements are pre-recorded and then presented. As shown in Fig. 2, the lip movement would modulate the ultrasound signal propagation. Hence, we can discriminate still or irregular motion from real lip movement by analyzing the received ultrasound signal. Second, the ultrasound transmitter and receiver are difficult to hack due to the self-verification mechanism. Here, the mechanism denotes



(a) Ultrasound signal propagation model (b) Effect of lip movement on complex signal domain

Fig. 2. Ultrasound propagation model. To acquire ultrasound reflections, we adapt earpiece speaker and microphone to transmit and receive signal as shown in (a). During face authentication, we prompt the user to perform lip movement (open and close mouth). The lip movement would modulate the reflected signal (change the amplitude and phase of the signal) as shown in (b), which can be utilized for subsequent liveness detection.

that the received ultrasound should be in accordance with the transmitted one. Despite all hardware deployed on mobile devices is exposed to the attackers and thus easily becomes unreliable if hacked, we can randomly change the signal modulated frequency or modulation type to perform self-verification, countering hardware injection in system-level.

The contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first attempt towards leveraging the temporal consistency between ultrasonic and visual modality to achieve accurate liveness detection on mobile devices.
- We propose a systematic framework including ultrasound signal acquisition and processing, lip motion extraction and cross-modality information fusion, making face authentication system more secure via an active manner.
- Extensive experiments indicates that SonarGuard could achieve average classification error rate of 0.91% under presentation attacks and 1.43% under hardware video injection attacks, which demonstrates the effectiveness of the proposed framework.

The rest of the paper is organized as follows. Section II introduces the related work. Section III illustrates the attack model. Section IV provides the overview of the system. Section V presents the detailed system design. Section VI introduces the dataset to evaluate our system. Extensive experimental results are shown in Section VII. Discussions on the proposed framework are presented in Section VIII. Finally, conclusions are drawn in Section IX.

II. Related Work

A. Single-modal liveness detection

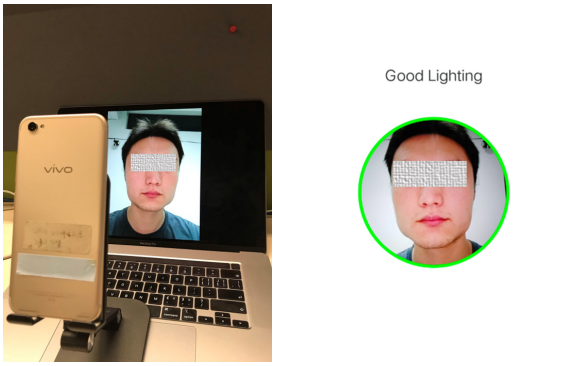
According to the cues leveraged for liveness detection, single-modal methods can be further categorized into texture and temporal based methods. Texture based methods exploit the texture discrimination between live faces and spoof ones. Hand-crafted features such as LBP [5], SIFT [6], HOG [7] are utilized in traditional texture based methods to capture spoofing patterns, while recent progress in CNN makes it possible to achieve face anti-spoofing through a data-driven

TABLE I
NOTATIONS

Notation	Description
f	the ultrasound signal frequency
Δf	the difference among adjacent frequencies
k	the index of ultrasound signal frequency
K	the number of frequency for ultrasound signal
t	the time index of signal
s	the transmitted signal
τ	the Time of Flight of the signal
l	the index of signal corresponds to lip movement
I	the number of paths of irrelevant signals
α	the complex signal attenuation coefficient
r	the demodulated ultrasound signal
$\mathbf{\Phi}$	the phase shift among different frequencies
$\boldsymbol{L}$	the vector of lip landmarks
s	the time index of mouth open
e	the time index of mouth close
W	the time window size of STFT
$\mathbf{X_v}$	the lip motion trajectory
$\mathbf{X_u}$	the segmented ultrasound spectrogram
$\mathbf{T_v}$	the time length of lip motion trajectory
$\mathbf{T_u}$	the time length of ultrasound spectrogram
$\mathbf{E_u}$	the embedding of lip motion trajectory
$\mathbf{E_v}$	the embedding of ultrasound spectrogram

manner [8]–[11], [11]–[17]. However, they both suffer from poor generalization to unseen attacks and complex lighting conditions, especially when RGB sensors are of low resolution or quality. To improve the robustness of face authentication, various face hallucination algorithms have also been proposed to enhance low-resolution facial images [18]–[20]. Temporal based methods concentrate on the temporal difference between live faces and spoof ones. [21] and [22] propose to achieve liveness detection by recognizing spontaneous eyeblinks and remote photoplethysmography (rPPG), respectively. [23] proposes to extract temporal features from visual dynamics to achieve liveness detection. In [24], the authors propose a challenge-response method using face reflections as in-band digital watermarking to address replay attacks for face recognition on consumer devices. Inspired by Captcha, Uzun et al. propose rtCaptcha, a real-time Captcha for face liveness detection in [25]. In [26], Liu et al. propose to utilize skin reflection to distinguish live user and pre-recorded video. Nevertheless, only relying on the captured video would result in a overconfidence judgement once RGB sensors are hacked and certain generated video is injected into the system [2].

To overcome the limitation of vision modality, performing face authentication using ultrasonic signals has been explored [27]. EchoFace proposed in [28] utilizes human face reflections to distinguish live user and media attack. A recent work [29] utilizes lip motion patterns built upon well-designed ultrasonic signals to enhance the security of face liveness detection. However, the security of the system with only ultrasound signal for liveness detection is still limited. To break aforementioned limitations, we introduce a cross-modal matching framework to detect attacks by judging whether the lip movement is performed by live user.



(a) Spoof with laptop screen (b) Captured image on smartphone

Fig. 3. Demonstration of video replay attack. By replaying the user video with required action, video replay attack can spoof the challenge-response protocol.

B. Multi-modal liveness detection

Multi-modal methods achieve liveness detection with impressive accuracy by fusing complementary information from different modalities [30]–[34]. Compared with single modal based methods, multi-modal could leverage the complementary information among different modalities to achieve better accuracy. Leveraging the complementary information among different modalities, multi-modal machine learning could achieve impressive performance [35]–[38]. [39] releases a dataset and benchmark for large scale multi-modal liveness detection. [3] proposes to fuse multi-modal features with a central difference network. Although recent advances demonstrate multi-modal methods are promising solutions to liveness detection, the application of these methods are limited to customized devices due to the fact that these methods require specialized hardware for depth and infrared imaging.

Our work also seeks new modality to strengthen existing systems. However, existing methods are still limited in visual modality, i.e., "multi-modal" in existing works refer to different visual sensors including RGB, depth and infrared cameras. On the contrary, we propose to combine ultrasonic and visual modalities, which raises more challenges since these modalities are totally different. In addition, the proposed framework could be deployed on existing mobile devices, while existing methods require customized hardware for data acquisition.

III. Attack Model

Based on the medium utilized for generating malicious data, existing attacks for facial authentication systems can be generally categorized into presentation attack [40] and hardware injection attack [2].

Presentation attacks can be further categorized into static and dynamic attacks. Static attacks adopt 2D photos or 3D models of a authorized user to spoof authentication systems. To combat these attacks, a natural solution is to design a challenge-response protocol where the user is asked to respond a challenge such as read a word, open mouth or blink. Since it is difficult to utilize static mediums to respond the challenge,

the challenge-response protocol can effectively detect static attacks. To spoof the challenge-response protocol, as shown in Fig. 3, dynamic attacks replay a video of authorized user with required action. With advanced face forgery techniques, the video of user with required action has been easily accessible, which makes the challenge-response protocol be vulnerable.

Hardware injection attack is recently proposed in [2], which injects user video directly through the hardware Camera Serial Interface (CSI). In this case, the injected video frames are identical to the ones recorded by live users, which makes all existing vision based methods fail to detect the attack. As a result, hardware injection attack has been a severe threat for existing face authentication systems

In this paper, we aim to prevent face authentication systems from dynamic presentation attack and hardware injection attack leveraging the consistency between ultrasonic and visual modality. Our design is under the assumption that high-quality user video with lip movement has been accessed by the attackers and utilized for spoofing, which is a common threat for existing face authentication systems.

IV. System Overview

The overview of the proposed system is shown in Fig. 4. Our system is based on a challenge-response protocol, i.e, it requires the user to perform lip movement during liveness detection. Compared with existing methods which only take camera frames as input, the security of our system is guaranteed by enabling the challenge and detect whether the challenge is finished by a live user leveraging the consistency between ultrasound and vision modality. The reason that we choose ultrasound modality is three-fold. First, the lip movement would modulate ultrasound signal, which leads to unique patterns which could be utilized for liveness detection. Second, ultrasound signal acquisition can be achieved with earpiece speaker and micrphone, which is avaliable on most mobile devices. Third, different from RGB cameras which passively capture video frames, the ultrasound signal is actively generated, which provides a self-verification mechanism, i.e., we can verify whether the signal is generated by authorized device or attacker. However, introducing such a modality is non-trivial. Specifically, despite of these advantages, the new ultrasound modality has totally different characteristics compared with vision modality, which poses three main challenges to its application for liveness detection.

- **Ultrasound signal acquisition and processing.** Due to the limited signal power reflected by human lip, extracting signal corresponding to lip movement from ultrasound reflections is non-trivial. To consolidate the system security, the signal acquisition and processing must be effective, secure and inaudible simultaneously.
- **Visual lip motion extraction.** Our design aims to leverage the consistency between visual and ultrasonic modalities to achieve liveness detection. Hence, we need to extract the lip motion information appropriate for cross-modal information fusion from captured videos.
- **Cross-modal information fusion.** The micro-doppler signature and lip motion trajectory come from different modalities, which leads to misalignment problems including inconsistent sampling rate, different sequence lengths, etc. Such misalignment poses challenges to effectively fuse the cross-modal information.

In the following sections, we would introduce how we tackle these challenges in detail.

V. Method

The pipeline of the proposed system is illustrated in Fig. 4, which is composed of three modules to resolve aforementioned challenges. We provide detailed introductions of these modules in the following section.

A. Ultrasound Signal Acquisition and Processing Module

1) Signal Acquisition.: The proposed system relies on transmitting ultrasound signal and receiving its reflections from user or spoof medium presented in front of the mobile device. To achieve this, we adapt earpiece speaker for signal transmitting and microphone for signal receiving. The transmitted signal should satisfy three requirements. First, the signal should be inaudible to avoid annoyance to users. Second, due to the multi-path propagation effect in wireless systems [41], simply transmitting signal with single frequency would suffer from severe multi-path fading, which leads to Signal-to-Noise Ratio (SNR) degradation in the received signal. The signal modulation scheme should be capable of alleviating this problem. Third, similar to vision based replay attacks, an attacker may hack the received ultrasound signal of the authorized user to perform ultrasound replay attack. To avoid such risk, the signal modulation scheme should be self-encrypted.

To satisfy aforementioned requirements, we transmit signal with frequency higher than 18kHz, which is inaudible to adults [42]. To alleviate the multi-path fading problem, we transmit signal with Multi Carrier Modulation (MCM), which has been widely adopted in sonar and radar systems [43]–[45] to enhance signal of interest based on the signal Time of Flight (ToF). The transmitted signal is composed of multiple frequency components, which can be expressed as

$$s(t) = \sum_{k=1}^{K} e^{j2\pi f_k t}, \tag{1}$$

where f denotes the signal frequency, K denotes the number of frequencies we adopt for signal transceiving, k, t denote the frequency and time index, respectively. To avoid the risk of ultrasound replay attack, we randomly update f_k for every new authentication. In this case, the received ultrasound signal would be uncoupled with the expected frequency under ultrasound replay attack, which results in further demodulation and processing failure.

2) Signal Processing.: To extract the signal corresponding to lip movement, we design a four-step signal processing pipeline, as shown in the bottom left part of Fig. 4. We first perform complex demodulation on the raw received signal to

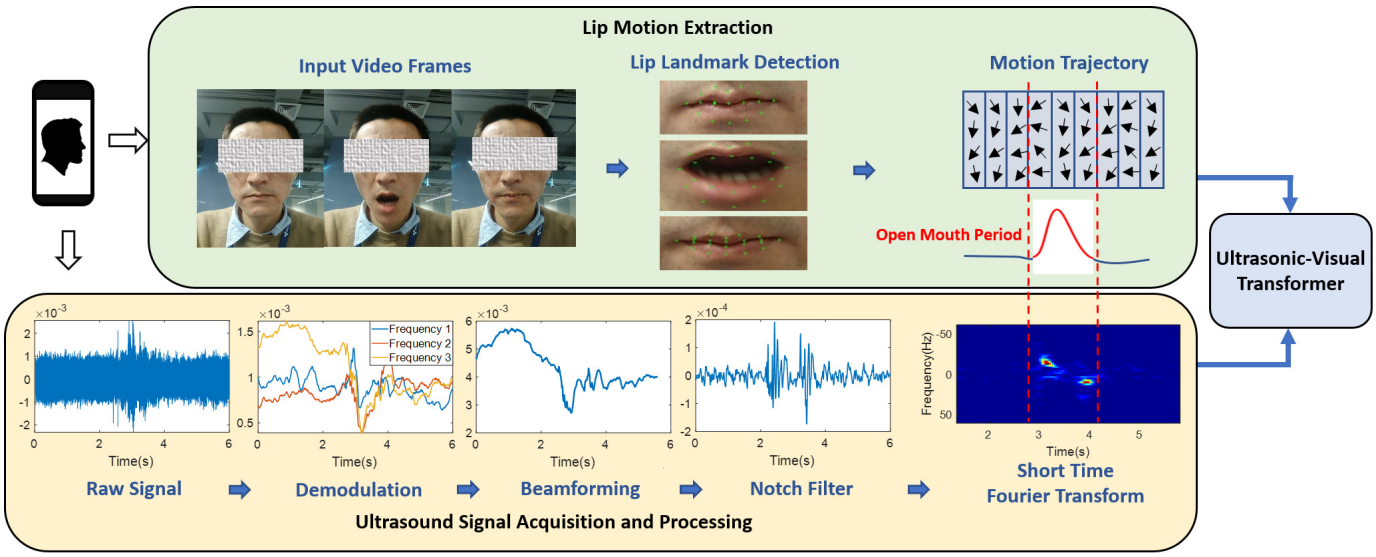


Fig. 4. Pipeline of the proposed system. The proposed system is composed of three modules. The ultrasound signal acquisition and processing module actively transmits and receives signals, extracts the signal correlated to lip movement from raw samples. The lip motion extraction module detects landmarks on the lip to judge the start and end of lip movement and segments ultrasound samples correspondingly. The ultrasonic-visual transformer module accepts the output of ultrasound signal processing and lip motion extraction module to classify whether the input stems from live user or spoof.

extract the complex baseband signal from the carrier wave. The demodulated signal can then be expressed as

$$r_k(t) = \sum_{i=1}^{I} \alpha_i e^{-j2\pi f_k \tau_i} + \alpha_l e^{-j2\pi f_k \tau_l(t)}, \tag{2}$$

where l denotes the index of signal corresponding to lip movement, while i corresponds to index of irrelevant signals. τ, k, α the signal Time of Flight (ToF), the index of carrier frequency and the complex signal attenuation coefficient, respectively. Since $r_k(t)$ is the mixture of lip reflection and irrelevant signals, we need to extract the lip reflection first. Hence, we adopt a conventional beamformer [46] with pre-defined ToF bins to separate the reflected signal, where each ToF bin corresponds to the signal with specific ToFs. More specifically, the relative phase shift on adjacent signal frequencies can be expressed as

$$\Phi(\tau) = e^{-2\pi\Delta f \tau}, \tag{3}$$

where Δf denotes the difference among adjacent frequencies. By compensating the phase shift and adding the signals on different frequencies, the signal from ToF τ would superimpose coherently while the signals from other location would be suppressed as

$$y(\tau) = \mathbf{\Phi}^H(\tau)\boldsymbol{r}, \tag{4}$$

where

$$\boldsymbol{r} = \{r_1, r_2, ..., r_K\}^T, \tag{5}$$

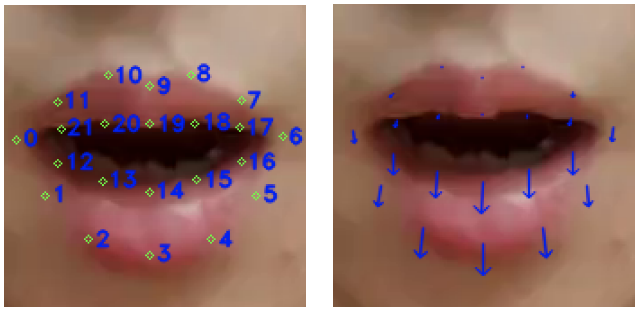
$$\mathbf{\Phi}^H(\tau) = \{1, e^{-2\pi\Delta f\tau}, ..., e^{-2\pi(K-1)\Delta f\tau}\}. \tag{6}$$

To extract the lip signal, we need to determine the ToF corresponding to lip first. To achieve this, we first set a series of possible ToF values and transform raw signals into ToF domain as Eq. 4. Then, we perform difference among consecutive samples. Due to the fact that lip reflections vary with time due to the lip movement while other irrelevant signals are time-invariant, the remained signal only keeps the time-variant components which corresponds to lip movement. Hence, we search the ToF bin with the maximum amplitude as the one corresponding to lip movement and extract the original signal for further processing. Note that after beamforming, although the lip signal has been strengthened, it still contains strong DC components which is caused by the direct signal leakage from speaker and static reflections around lip. To further enhance the signal most related to lip motion, we deliver the signal through a Notch filter with the zero point at zero frequency to suppress low frequency components. As shown in Fig. 4, the signal passed Notch filter is series of time samples containing severe random noise, which could not directly capture fine-grained frequency information according to [47]. Hence, we transform the signal into spectrogram by short time fourier transform (STFT). As a visual representation of the orginal signal, the STFT spectrogram is more suitable to convolutional architectures.

B. Visual Lip Motion Extraction Module

In the proposed system, we extract the lip motion by detecting lip landmarks from the captured videos. The lip motion information, including the period of mouth open and close, can be effectively represented by the location variation of lip landmarks. Note that the ultrasonic STFT spectrogram spans over the whole user interaction period containing irrelevant motion, which would interfere the liveness detection. With the period of lip movement, we can segment the signal-of-interest correspondingly to suppress the interference. In the following, we would introduce how we detect lip landmarks and segment ultrasound signal in detail.

1) Lip motion representation.: Given video frames $\{F_0, F_1, ..., F_T\}$ of length T, we use the face alignment



(a) Lip landmarks (b) Lip motion vectors

Fig. 5. Lip landmarks and motion vectors.

method in [48] to obtain lip landmarks from each video frame. Let $\{\boldsymbol{L}_0, \boldsymbol{L}_1, ..., \boldsymbol{L}_T\}$ denotes the sequence of landmarks, where each vector $\boldsymbol{L}_t = \{p_1, p_2, ..., p_{22}\}$ contains 22 designated landmarks coordinates of one video frame as illustrated in Fig.5(a). The displacement of a landmark between two consecutive frames is then denoted as the *motion vector* of that landmark, as shown in Fig.5(b). To capture the status of mouth action, the aspect ratio of mouth area can be calculated from euclidean distances between selected landmarks as follows:

$$asr_t = \frac{\|p_{20} - p_{13}\| + \|p_{19} - p_{14}\| + \|p_{18} - p_{15}\|}{\|p_{10} - p_2\| + \|p_9 - p_3\| + \|p_8 - p_4\|}. \quad (7)$$

This calculated aspect ratio sequence consists a curve indicating the variation of mouth area along time axis. Thus, the mouth open and close time indices s and e can be found at the points where the curve crosses predefined thresholds upward and downward respectively. The desired lip motion trajectory $\mathbf{X}_v$ is composed of motion vectors of all lip landmarks throughout this interval, which is of length T_v and can be expressed as:

$$\mathbf{X}_v = \{L_{s+1} - L_s, L_{s+2} - L_{s+1}, ..., L_{e-1} - L_{e-2}, L_e - L_{e-1}\}. \quad (8)$$

2) Ultrasound signal segmentation.: We segment ultrasound signal from the entire recorded spectrogram according to the visual lip motion information. On one hand, this design temporally cut out interference from irrelevant motion such as head motion before or after opening mouth. On the other hand, it can detect temporal mismatch to prevent attacks that mimic lip motion on prerecorded videos. Specifically, we pick out spectrogram segment within the same time interval of the above lip motion trajectory as the final ultrasound signal $\mathbf{X}_u$. Notably, the length T_u of ultrasound might differ from the length T_v of lip motion trajectory since the sample rate of ultrasound signal and the video frames are not identical. In the next section, we fuse the information from two modalities to achieve accurate liveness detection.

C. Ultrasonic-Visual Transformer Module

Inspired by the successful application of the Transformer [49] in vision and language tasks [50]–[52], we design a Ultrasonic-Visual Transformer (UVT) model which can effectively fuse the information from ultrasound signal and lip motion trajectory.

Fig. 6 illustrates the detailed architecture of UVT model, which is made up of feature extractors, transformer encoder, transformer decoder, and followed by a light MLP classification head. The feature extractor is a pair of CNNs to extract high dimensional features from ultrasonic spectrogram and visual motion trajectory. The transformer encoder reads ultrasound spectrogram feature and captures long-range context information among spectrogram bins via self-attention mechanism. The transformer decoder's inputs are visual movement feature of each video frames, which are further aggregated by self-attention layers inside the decoder. Furthermore, the fusion of ultrasound and motion information is accomplished by the multi-head attention between encoder output features and decoder intermediate features. The intuition behind our design is that the micro-doppler information in ultrasound spectrogram comes from lip movement. This long-range cross-modal relationship between ultrasonic and visual sequences can be learnt by the transformer architecture. More design details are elaborated below.

1) Feature extractor.: The ultrasound spectrogram $\mathbf{X}_u \in \mathbb{R}^{W \times T_u}$ is composed of T_u frequency coefficient vectors, and each vector contains W coefficients corresponding to the window length of STFT. The visual motion trajectory $\mathbf{X}_v \in \mathbb{R}^{2 \times 22 \times T_v}$ contains a 2D tensor sequence of length T_v. A common and practical issue of multi-modal task is the imperfect alignment between different modalities. Specifically, the length of the two sequences may be different, and sequence element of the two modalities may span different duration. To address this issue, we design the feature extractor as a pair of modified ResNet-18 where all conv3x3 filters are replaced by conv1x1. That is, the feature extractor only captures frequency dimension features from ultrasonic input and spatial dimension features from visual input respectively. Furthermore, the encoder-decoder attention block inside transformer, by its design, can accept queries and keys of different lengths and attend to global contexts across time dimension to handle imperfect alignment issue. Note that, this split spatio-temporal design has demonstrated superior performance than direct 3D convolution for action recognition tasks in [53].

2) Transformer.: Let d be the input embedding dimension of transformer. The ultrasonic feature map from feature extractor is averaged pooled and flattened into ultrasonic embedding $\mathbf{E}_u \in \mathbb{R}^{d \times T_u}$. Similarly, The motion vector embedding $\mathbf{E}_v \in \mathbb{R}^{d \times T_v}$ is obtained through average pooling. We take the advantage of transformer's self-attention mechanism and encoder-decoder architecture to fuse multi-modal features. Specifically, $\mathbf{E}_u$ is fed into transformer encoder and transformed by a stack of self-attention layers to capture long-range temporal information. Intuitively, the noise can be assumed to be stationary stochastic process during the short period. Hence, encoder can focus on useful micro-doppler features while eliminate noise features given enough contextual information. Inside transformer decoder, the visual movement features are aggregated via self-attention layers similarly. Then, the encoder-decoder multi-head attention layer takes aggregated

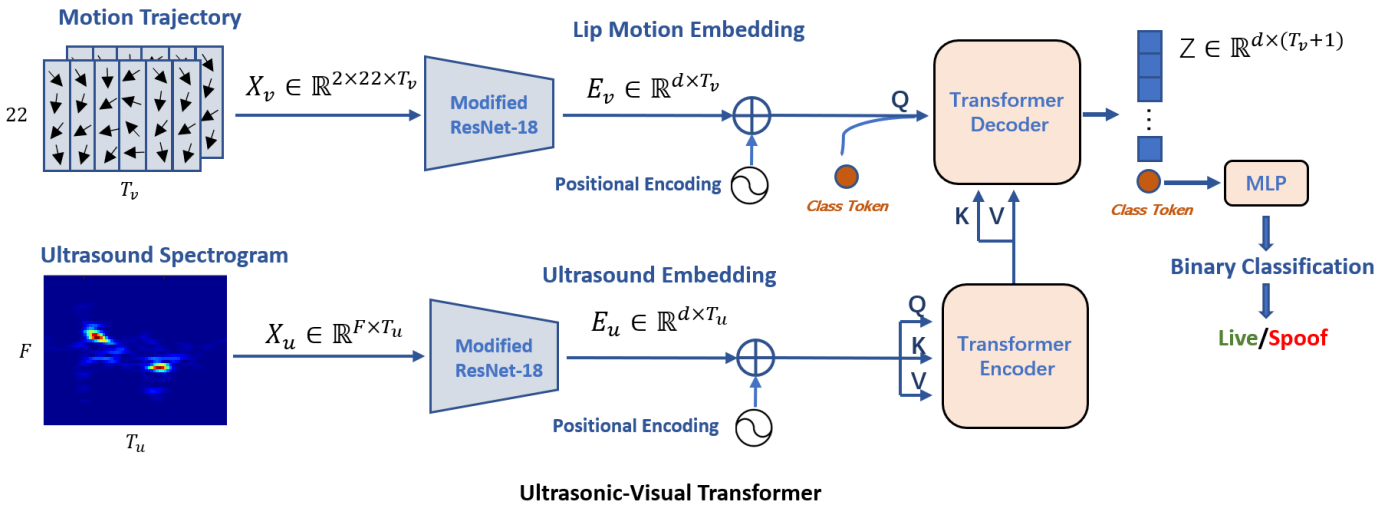


Fig. 6. Detailed architecture of UVT. UVT uses CNNs to extract embeddings from motion trajectory and ultrasound spectrogram, which are added with sinusoid positional encoding to maintain positional information. Transformer encoder makes use of self-attention mechanism to capture long-range information from ultrasound embedding, and feeds its output to transformer decoder as key-value pairs. Cross-modality matching is carried out within transformer decoder. Specifically, for every query slot inside lip motion embedding, it looks over all key vectors to aggregate information from matching value vectors. The aggregated information is carried by class token to classifier.

visual features as query, and audio features inside encoder memory as key and value. The output is the weighted sum of most relevant audio features for each video frame. Formally, the output of the Transformer encoder can be expressed as

$$\begin{aligned}
&\mathbf{h}^0_{enc} = \mathbf{E}_u + \mathbf{Pos}_u, \\
&\tilde{\mathbf{h}}^l_{enc} = \mathrm{LN}(\mathrm{MSA}(\mathbf{h}^{l-1}_{enc}, \mathbf{h}^{l-1}_{enc}, \mathbf{h}^{l-1}_{enc}) + \mathbf{h}^{l-1}_{enc}),\ l = 1...L_{enc} \\
&\mathbf{h}^l_{enc} = \mathrm{LN}(\mathrm{FFN}(\tilde{\mathbf{h}}^l_{enc}) + \tilde{\mathbf{h}}^l_{enc}),\ l = 1...L_{enc}
\end{aligned} \tag{9}$$

where MSA, LN, and FFN denotes Transformer's multi-head self-attention, layer norm, and feed forward network respectively. $\mathbf{Pos}_u$ is the positional encoding. l is the index of the encoder layer and L_{enc} denotes the number of layers for Transformer encoder. Then, the output of the decoder can then be given by

$$\begin{aligned}
&\mathbf{h}^0_{dec} = [\mathbf{E}_{cls}; \mathbf{E}_v] + \mathbf{Pos}_v, \\
&\tilde{\mathbf{h}}^l_{dec} = \mathrm{LN}(\mathrm{MSA}(\mathbf{h}^{l-1}_{dec}, \mathbf{h}^{L_{enc}}_{enc}, \mathbf{h}^{L_{enc}}_{enc}) + \mathbf{h}^{l-1}_{dec}),\ l = 1...L_{dec} \\
&\mathbf{h}^l_{dec} = \mathrm{LN}(\mathrm{FFN}(\tilde{\mathbf{h}}^l_{dec}) + \tilde{\mathbf{h}}^l_{dec}),\ l = 1...L_{dec}
\end{aligned} \tag{10}$$

where $\mathbf{E}_{cls}$ is class token. The output of the decoder in the last layer, i.e., $\mathbf{h}^{L_{dec}}_{dec}$, is denoted as $\mathbf{Z} \in \mathbb{R}^{d\times(T_v+1)}$ for simplicity. The design of Transformer is inspired by the principles of database, where Query and Key are utilized to measure the consistency of the inputs [52]. Similarly, we propose the SonarGuard framework to distinguish live user and malicious attack leveraging the consistency between ultrasound signal and lip motion trajectory. Hence, we adopt ultrasound signal and lip motion trajectories as key value and query, respectively.

3) Positional encoding.: Positional encoding is important for self-attention modules to maintain positional information. In our case, both ultrasonic and motion information are represented as time sequences, so positional encoding is injected to both transformer encoder and decoders. Regarding the type of positional encoding, we observe that the fixed sinusoid coding works well in our experiments.

4) Liveness detection.: The liveness detection can be modeled as a binary classification problem. To this end, we adopt the class token method following BERT [52]. The class token is a trainable embedding pretended to transformer decoder's input tokens. It keeps aggregating useful classification information, while going through the stack of transformer decoder layers. Let the transformer decoder's output be $\mathbf{Z} \in \mathbb{R}^{d\times(T_v+1)}$. The classification head is formed by connecting a MLP to the first position at encoder's output hidden state $\mathbf{Z^0}$. The training objective of our model is the standard binary cross-entropy (BCE) loss:

$$L_{BCE} = -\frac{1}{M}\sum_{i=1}^{M} y_i log p_i + (1-y_i) log(1-p_i), \tag{11}$$

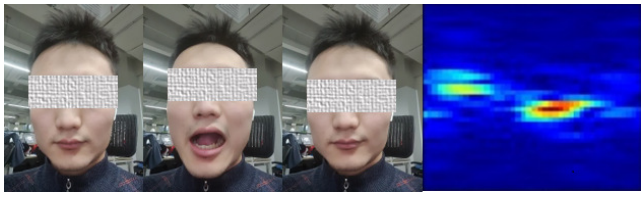
where y_i and p_i denote the label and model prediction for the ith data sample, respectively. M is the number of data samples.

VI. Datasets

A. Datasets for training and validation.

Since this is, to our best knowledge, the first work towards achieving liveness detection leveraging the consistency between ultrasound and vision modality, there is no existing dataset that can be utilized to evaluate our method. Hence, we setup a new multi-modal *Ultrasound-Vision Liveness Detection* dataset, which contains synchronized ultrasound and video samples, to evaluate our method. We implement an Android application to collect data, which records ultrasound signal with 48000Hz sample rate, and live video at 30 frames per second by front camera. We utilize 3 carrier frequencies for ultrasound signal transceiving. The difference among adjacent frequencies is set to be 100Hz.

For live user data collection, the subjects are asked to hold the smartphone, open and close their mouths following user interface instructions shown on the screen. Fig. 7(a) shows an



(a) Data captured by live user, of which the ultrasound spectrogram explicitly shows the unique pattern.

(b) Data captured under video replay attack, of which the ultrasound spectrogram only contains random noise.

(c) Hardware injection attack data by combing live user video and ultrasound spectrogram collected under replay attack.

(d) Attack data in *Challenge* set with finger gestures to mimic the effect of lip movement. The pattern of ultrasound spectrogram under this case is similar to the one under lip movement.

Fig. 7. Illustration of our dataset.

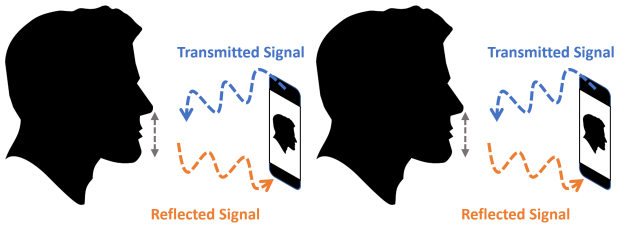
example of data captured by live user, of which the ultrasound spectrogram explicitly shows the unique pattern caused by lip movement.

For attack data, we collect both presentation attack and hardware injection attack data as illustrated in Sec. III. We first collect presentation attack data via replaying live user video data on various screens of phone, laptop and monitors. As shown in Fig. 7(b), there is no signal pattern in ultrasound spectrogram since the lip movement is only replayed on the screen, which does not affect the propagation of ultrasound signal. For hardware injection attack data, the only difference is that the captured user video is directly injected to camera interface rather than replayed on a medium. Hence, we combine the video data of live user and ultrasound data captured under replay attack to build the dataset of hardware injection attack as shown in Fig. 7(c).

The dataset is collected by Asian people with ages ranging from 18 to 40. The numbers of samples captured by male/female are approximately same. Both real face and attack samples are collected in a noisy office room. Totally, our dataset contains over 30000 live and spoof samples from over 200 subjects.

B. Datasets for testing.

We collect two types of test datasets to verify the feasibility of the proposed system and its effectiveness in practical deployment, respectively. The first one is *Basic* testset, which is collected in the same way as the training dataset. For live user data, the subjects hold the smartphone, open and close their mouths following user interface instructions. Similarly, the attack data is also collected in the same way as the training dataset. Another one is *Challenge* testset, which targets to create challenging usage scenarios to evaluate the robustness of our system. The live user data is collected from 20 subjects who are the first time to use our system. In this case, the subjects may not follow the user interface instruction to perform lip movement perfectly, which could demonstrate the performance of the proposed system for new users in practical deployment. On the contrary, for the attack data, we assume that the attacker is familiar with the principles of our system. To this end, we have noted that a professional attacker may perform finger gestures to mimic the effect of lip movement on ultrasound propagation as shown in Fig. 8. Specifically, for live user, the lip movement would modulate the propagation of ultrasound signal, which is utilized to achieve liveness detection. Meanwhile, we can perform finger gestures, i.e., pinch fingers to yield similar effects on ultrasound signal to fool the liveness detection system. Hence, we let the attacker perform finger movement synchronized with the pre-reorded video to better fool the proposed framework. The detailed information of the media for performing presentation attack is summarized in Table. II. In general, the *Basic* dataset containing 19376 positive samples and 16882 negative samples, where the proportion of the training, validation and test samples is 6:2:2. The *Challenge* dataset containing 1120 positive and 1152 negative samples for test.



(a) Ultrasound signal propagation with live user.

(b) Ultrasound signal propagation under replay attack with finger gesture.

Fig. 8. Ultrasound signal propagation for live user and attack.

TABLE II
DEVICES FOR REPLAY ATTACK

Device	Size(mm)	Resolution
Dell P2314	553x312	1920x1080
Thinkpad T470	339x232	1920x1080
Apple iPad	243x190	2048x1536
Oppo R9s	153x74.3	1920x1080

VII. Experiments

A. Implementation Details

The UVT model is trained from scratch by PyTorch [54]. All convolution and fully-connected layers are initialized with normal weight distribution. The model is trained for 100 epochs using SGD optimizer with weight decay of 10^{-4}, and learning rate is adjusted by 1cycle policy [55] with max learning rate of 10^{-2}. The transformer contains two encoder layers and two decoder layers, both of which uses four-head self-attention layers of 256 embedding dimension.

B. Evaluation metrics

To compare the performance of different methods, we report the experiment results with the following metrics: Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) [3]. We utilize positive/negative samples to denote attack/real samples. Then, these metrics can be expressed as

$$\begin{aligned} \mathrm{APCER} &= \mathrm{FN}/(\mathrm{TP}+\mathrm{FN}) \\ \mathrm{BPCER} &= \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) \\ \mathrm{ACER} &= (\mathrm{APCER}+\mathrm{BPCER})/2, \end{aligned} \tag{12}$$

where TP, TN, FP and FN are the abbreviations of True Positive, True Negative, False Positive and False Negative, respectively.

C. Results and Analyais

Besides our full model, we experiment with another two models: ultrasound-only model and late-fusion model. Specifically, the ultrasound-only model applies a single ResNet18 feature extractor on ultrasonic spectrograms. The late-fusion model uses two ResNet18 to extract features from two modalities, and fuses them by simple concatenation. Experiment results below verify the effectiveness of ultrasound on basic liveness detection, as well as the the importance of lip motion trajectory and ultrasonic-visual transformer module on preventing more challenging attacks.

1) Results under presentation attack: Performance comparison result is presented in Tab. III. The ultrasound-only model achieves acceptable result on *Basic* testset, but does not generalize well on *Challenge* testset. By incorporating the lip motion trajectory, the late-fusion method reduces ACER by 1.46 points and 6.96 points on the two testsets respectively. On top of the late-fusion method, our full model gains additional 0.36 points and 0.45 points improvements. We attribute this achievement to transformer's strong information fusion ability. Furthermore, when switching from *Basic* testset to the *Challenge* one, the Ultrasound-only method's ACER increase 8.68 points, while our method maintains more stable performance relatively. To inspect performance comparison at different thresholds, Fig. 9(a) shows the receiver operating characteristic (ROC) curves of *Challenge* testset under presentation attack. We observe that our full model has strongest performance (i.e. largest area under curve), and its advantage is more obvious on the *Challenge* testset due to better generalization ability.

TABLE III
Evaluation results under presentation attack (%).

Testset	Method	APCER	BPCER	ACER
Basic	Ultrasound-only	2.74	2.71	2.73
	Late-fusion	0.49	2.05	1.27
	UVT	0.26	1.56	**0.91**
Challenge	Ultrasound-only	15.23	7.59	11.41
	Late-fusion	1.40	7.50	4.45
	UVT	0.95	7.05	**4.00**

2) Results under hardware injection attack: Although the vision cue under hardware injection attack has been unreliable, our method still can exploit the mismatching between two modalities to achieve accurate liveness detection as shown in Tab. IV. All three methods can resist hardware injection attack. Specifically, the ACER of our methods only increase 0.52 points and 2.75 points respectively on the two testsets. Note that, pure vision methods would be subjected to total failure under such attack. Since the live user data does not change under difference attacks, the BPCER in Tab. IV is the same with the BPCER in Tab. III. The ROC curves of *Challenge* testset in Fig. 9(b) show that our full model is the most robust against hardware injection attack.

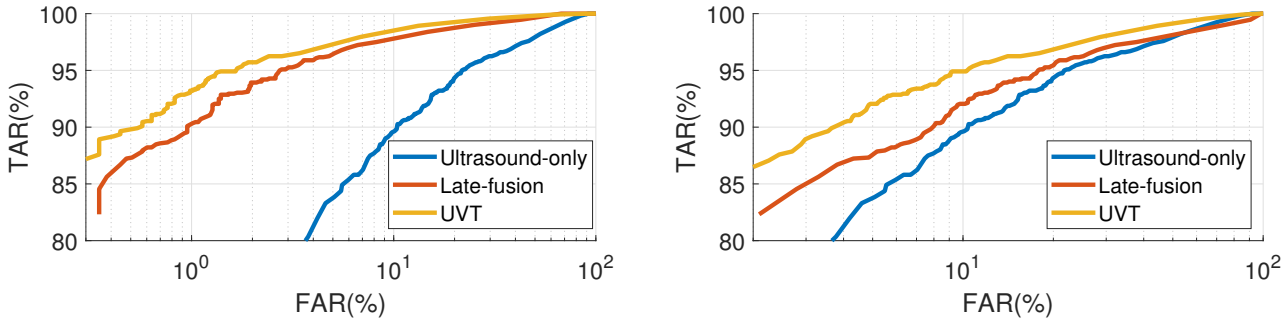
TABLE IV
Evaluation results under hardware injection (%).

Testset	Method	APCER	BPCER	ACER
Basic	Ultrasound-only	2.74	2.71	2.73
	Late-fusion	1.78	2.05	1.92
	UVT	1.29	1.56	**1.43**
Challenge	Ultrasound-only	15.23	7.59	11.41
	Late-fusion	10.66	7.50	9.08
	UVT	6.44	7.05	**6.75**

3) Ablation study on signal processing: Our signal processing module is composed of several steps. Fig. 10(a) shows the ultrasonic spectrogram obtained through the entire signal processing module. To demonstrate the effectiveness of these steps, we perform an ablation study by removing the beamformer and notch filter. The spectrogram without beamformer is shown in Fig. 10(b), which contains severe random noise. This illustrates beamformer's functionality of aggregating information from different carrier frequencies to increase the signal-to-noise ratio. Without notch filter, it would be hard to identify the lip motion trajectory, as shown in Fig.10(c). This is because the ultrasound signal contains strong time-invariant components, which has much higher amplitude compared to lip motion trajectory.

4) Model scaling experiment: We also perform a model scaling experiment by varying transformer's embedding size, and number of layers. The result is reported in Tab. V. Regarding our task, increasing number of layers from 1 to 2 achieves more improvements than enlarging embedding size from 128 to 256. However, stacking additional layers hurts performance in our experiment instead, which suggests that larger transformer requires more training samples.

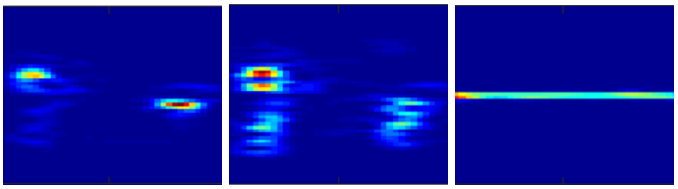
5) Attention visualization: Attention mechanism is a key components of Transformer, so we provide visualizations to understand how it contributes to effective information fusion.



(a) ROC under presentation attack

(b) ROC under hardware injection attack

Fig. 9. ROC curves under *Challenge* testset. The proposed UVT module achieves the best performance under both presentation and hardware inejction attack.



(a) Original (b) Without Beamformer (c) Without Notch Filter

Fig. 10. Ablation study of signal processing module. (a) is the signal extracted by the entire signal processing module. (b) is extracted without beamformer, which contains severe random noise. (c) is extracted without notch filter, which is hard to distinguish lip movement period.

TABLE V
DETAILED RESULTS OF MODEL SCALING EXPERIMENT. (%)

Testset	Embedding size	Layers	ACER
Basic	256	1	1.42
	128	2	1.04
	256	2	**0.91**
	256	4	1.59
Challenge	256	1	6.22
	128	2	4.88
	256	2	**4.00**
	256	4	4.68

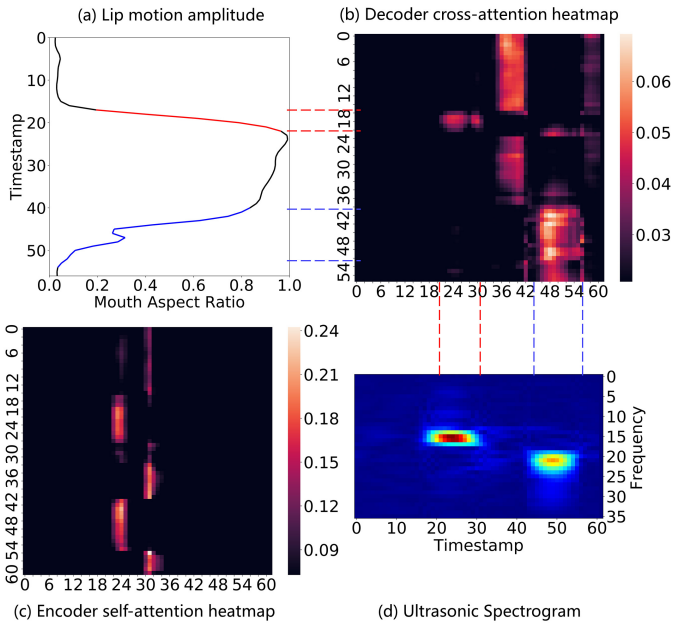


Fig. 11. Visualization of encoder self-attention in ultrasonic modality and decoder cross-attention in both modalities. The encoder-decoder cross-attention plays an key role on effective information fusion and the encoder self-attention provides noise suppression on ultrasonic spectrogram.

Fig. 11(a) shows the curve of mouth aspect ratios during data acquisition, where mouth open and close occur in duration 18˜22 and 42˜54 respectively. The corresponding ultrasonic spectrogram is shown in Fig. 11(d), which shows the mouth open and close signatures in duration 18˜32 and 42˜56. We visualize the attention map of one cross-attention head in Fig. 11(b). The vertical and horizontal axis correspond to the query and key, i.e., visual embeddings $\mathbf{E}_v$, and ultrasound embeddings $\mathbf{E}_u$ accordingly. The visual embeddings within 18˜22 attend to the ultrasonic embeddings within 21˜30, which has highest amplitude in ultrasonic spectrogram shown in red. Similar observation for lip close can be noticed. In other words, the cross-attention block is able to aggregate ultrasonic embeddings most relevant to lip motion. Furthermore, we visualize the attention map of Transformer encoder's self-attention blocks in Fig. 11(c), the vertical axis for query and horizontal axis for key are both related to ultrasonic embeddings. We notice that embeddings associated with mouth open 18˜32 and close 42˜56 both attend to the 24th embedding. Moreover, the attention of other embeddings is located around the 31st embedding, all of which corresponds to noise area in the spectrogram. This attention patterns reveals the model's ability to separate lip motion trajectory and noise. In conclusion, these results demonstrate that the encoder-decoder cross-attention plays an important role on effective information fusion and the encoder self-attention provides noise suppression on ultrasonic spectrogram.

D. Robustness Evaluation

1) Robustness among different devices: Our method is designed for existing mobile devices without using additional hardware. So a consistent performance on different devices is a key consideration for future deployment in real world. To evaluate the the robustness of the proposed system among different devices, we collect data on four smartphones: OnePlus 7Pro, Samsung Galaxy Note10, Google Pixel 2 XL and Vivo X9, respectively. From Tab. VI, we can find that our full model performs consistently among different devices.

TABLE VI
ACER(%) MEASUREMENT AMONG DIFFERENT DEVICES.

Device	Method	*Basic* Set	*Challenge* Set
OnePlus 7Pro	Ultrasound-only	2.70	11.97
	Late-fusion	1.52	4.28
	UVT	**1.20**	**3.18**
Samsung Galaxy Note 10	Ultrasound-only	2.39	14.79
	Late-fusion	1.02	6.00
	UVT	**0.85**	**4.19**
Google Pixel 2 XL	Ultrasound-only	15.71	11.06
	Late-fusion	5.00	4.55
	UVT	**2.14**	**3.26**
Vivo X9	Ultrasound-only	2.71	3.35
	Late-fusion	1.52	2.41
	UVT	**0.82**	**2.22**

TABLE VII
ACER VARIATION UNDER BACKGROUND NOISE(%)

Noise Level(dB)	40	50	60	70	80	90
Basic	-0.2	-0.2	-0.2	-0.2	-0.17	-0.2
Challenge	-0.13	+0.06	-0.23	-0.03	-0.03	-0.13

2) Robustness under background noise.: Since our system relies on transmitting and receiving ultrasound signal on mobile devices, a natural concern is whether the system performance would be affected by environment noise. To address this concern, we evaluate robustness of the proposed system under background noise with the level varies from 40dB to 90dB. The ACER variation of the proposed system versus noise level is shown in Tab. VII, which indicates that the background noise has tiny effect on the system performance. This is because the background noise mainly exists on the frequency under 18000Hz, while our system adopts signal with higher frequency. Note that the environment with noise level higher than 90dB would be harmful to human health, which is not considered for the application of the proposed system.

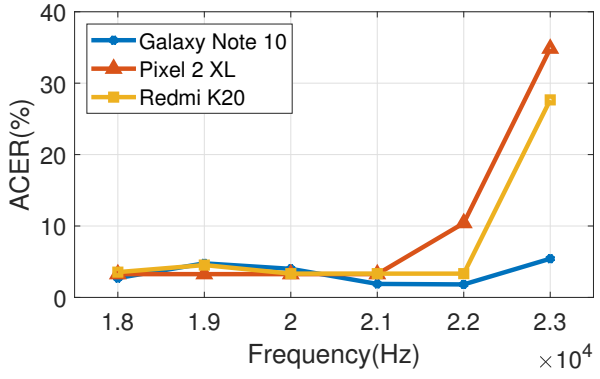


Fig. 12. ACER under different carrier frequencies

3) Robustness under different signal frequencies: In this experiment, we investigate the impact of the signal frequency selected for signal transmitting and receiving. As shown in Fig. 12, the ACER varies slightly when the frequency is lower than 22000Hz, while it increases rapidly with higher frequency. This is due to the fact that signal transceiving with higher frequency could not be perfectly supported by the speaker and microphone on smartphones, which leads to the performance degradation.

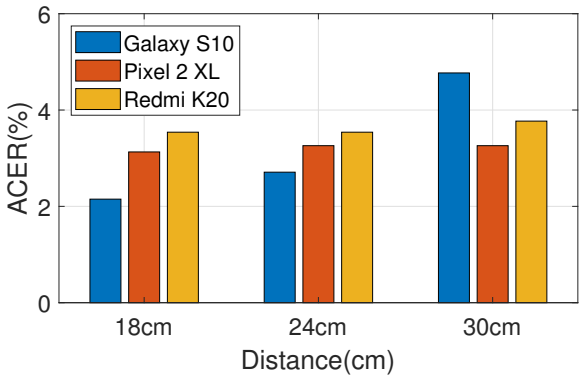
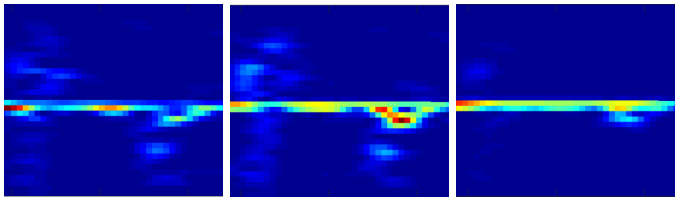


Fig. 13. ACER under different distances.

4) Robustness under different distances: In this experiment, We investigate the impact of the distance between the user and the device. The distance, which is measured from the center of the phone screen to the user lips, varies from 18cm to 30cm. As shown in Fig. 13, the ACER gradually increases when the distance increases. This is because ultrasound signals decay rapidly when the propagation increases, which makes it extremely difficult to capture lip motion information and leads to accuracy degradation. However, it is common for the user to hold the smartphone within 30cm, which indicates that the proposed system can be deployed in practical scenario.



(a) One carrier frequency shifts 3Hz (b) Two carrier frequencies shift 2Hz (c) Three carrier frequencies shift 2Hz

Fig. 14. Robustness of our system under ultrasound replay attack. With the wrong carrier frequency, the signal has been far from the expected one as shown in Fig. 10(a), which leads to the failure of attack.

5) Robustness under ultrasound replay attack.: Similar to visual replay attacks, the ultrasound data of live user may be hacked by attackers to achieve ultrasound replay attack, which would introduce security risk. However, different from vision based methods which passively record RGB videos of the user, our system actively transmits and receives ultrasound signal. As illustrated in Sec. V-A, we randomly change the carrier frequency of the signal in Eq. 1 for every new authentication process. Fig. 14 shows the processed signal when an attacker perform replay attack while the signal frequencies has changed. Due to the wrong signal frequency, the output signal has been far from the expected one, which leads to the failure of the attack.

VIII. DISCUSSIONS

In the past decade, face authentication systems have gained widespread popularity. However, these systems have been vul-

nerable under various kinds of malicious attacks. This is due to the fact that optical cameras themselves are easily to be fooled or hacked, which motivates us to seek for new modalities to achieve liveness detection. While achieving liveness detection based on ultrasound reflection has been investigated, the system security with only ultrasound modality is still limited. To this end, we propose to leverage the consistency between visual and ultrasonic modality to achieve liveness detection, which is more accurate through multi-modal learning. All in all, the proposed technique could enhance the security of face authentication systems on mobile devices and inspire more investigations on multi-modal learning.

For ultrasound signal, since the signal variation caused by lip movement is too tiny to capture, we design the ultrasound signal processing module to enhance the signal-of-interest with elaborated designed beamformer and filters. For visual information, the lip landmarks could represent the lip motion information more efficiently compared with raw frames, which encourages us to deign the visual lip extraction module. To judge the consistency of two modalities, we have noted that Transformer is inspired by the principles of database, where the Query and Key are utilized to measure the consistency of the inputs. Hence, we design the Ultrasonic-Visual Transformer module to achieve final information fusion for accurate liveness detection. While all existing approaches fail under hardware injection attack, the proposed framework could detect it through cross-modality matching. For presentation attacks, the proposed framework could also improve the overall accuracy for liveness detection. The main disadvantage of the proposed framework lies in the fact that it requires synchronized collection of ultrasound signal and video frame, which may lead to additional overhead for data collection.

IX. Conclusion

In this paper, we proposed SonarGuard, a liveness detection system for face authentication on mobile devices by combing ultrasonic and visual information. SonarGuard extracts ultrasound signal related to lip movement with a signal processing module, obtains the lip motion trajectory and segmented ultrasound signal with a motion extraction module and finally fuses the information from ultrasound signal and motion vector with a information fusion module. Extensive experiments on newly collected dataset demonstrates the efficiency of the proposed system in real world usage. To our best knowledge, this is the first attempt towards ultrasonic liveness detection on mobile devices.

References

[1] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8484–8493.

[2] Y. Chen and B. Ma, "Biometric authentication under threat: Liveness detection hacking," in *White Paper For Black Hat USA 2019*, Mandalay Bay, Las Vegas, August 2019.

[3] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[4] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.

[5] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*, 2015, pp. 2636–2640.

[6] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, June 2016.

[7] J. Komulainen, A. Hadid, and M. Pietik¨ainen, "Context based face anti-spoofing," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[8] W. Yan, Y. Zeng, and H. Hu, "Domain adversarial disentanglement network with cross-domain synthesis for generalized face anti-spoofing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 4084–4095, 2022.

[9] H. Wu, D. Zeng, Y. Hu, H. Shi, and T. Mei, "Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4626–4638, 2022.

[10] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *Proceedings of 2017 IEEE International Joint Conference on Biometrics (IJCB)*, October 2017, pp. 319–328.

[11] S. R. Arashloo, "Unseen face presentation attack detection using sparse multiple kernel fisher null-space," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4084–4095, 2021.

[12] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3d face anti-spoofing with factorized bilinear coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4031–4045, 2021.

[13] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 641–657.

[15] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, "Face spoofing detection based on local ternary label supervision in fully convolutional networks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3181–3196, 2020.

[16] X. Zhu, S. Li, X. Zhang, H. Li, and A. C. Kot, "Detection of spoofing medium contours for face anti-spoofing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2039–2045, 2021.

[17] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3507–3516.

[18] R. Wang, M. Jian, H. Yu, L. Wang, and B. Yang, "Face hallucination using multisource references and cross-scale dual residual fusion mechanism," *International Journal of Intelligent Systems*, 2022.

[19] M. Jian, C. Cui, X. Nie, H. Zhang, L. Nie, and Y. Yin, "Multi-view face hallucination using svd and a mapping model," *Information Sciences*, vol. 488, pp. 181–189, 2019.

[20] M. Jian and K.-M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1761–1772, 2015.

[21] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[22] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 4244–4249.

[23] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. Ho, "Detection of face spoofing using visual dynamics," *IEEE Transactions on information forensics and security*, vol. 10, no. 4, pp. 762–777, 2015.

[24] D. F. Smith, A. Wiliem, and B. C. Lovell, "Face recognition on consumer devices: Reflections on replay attacks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 736–745, 2015.

[25] E. Uzun, S. P. H. Chung, I. Essa, and W. Lee, "rtcaptcha: A real-time captcha based liveness detection system." in *NDSS*, 2018.

[26] H. Liu, Z. Li, Y. Xie, R. Jiang, Y. Wang, X. Guo, and Y. Chen, "Livescreen: Video chat liveness detection leveraging skin reflection," in *IEEE INFOCOM 2020*, 2020, pp. 1083–1092.

[27] B. Zhou, Z. Xie, Y. Zhang, J. Lohokare, R. Gao, and F. Ye, "Robust human face authentication leveraging acoustic sensing on smartphones," *IEEE Transactions on Mobile Computing*, 2021.

[28] H. Chen, W. Wang, J. Zhang, and Q. Zhang, "Echoface: Acoustic sensor-based media attack detection for face authentication," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2152–2159, 2019.

[29] M. Zhou, Q. Wang, Q. Li, P. Jiang, J. Yang, C. Shen, C. Wang, and S. Ding, "Securing face liveness detection using unforgeable lip motion patterns," *arXiv preprint arXiv:2106.08013*, 2021.

[30] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.

[31] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "Pipenet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 644–645.

[32] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-modal face presentation attack detection via spatial and channel attentions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[33] T. Shen, Y. Huang, and Z. Tong, "Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[34] H. Kuang, R. Ji, H. Liu, S. Zhang, X. Sun, F. Huang, and B. Zhang, "Multi-modal multi-layer fusion network with average binary center loss for face anti-spoofing," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 48–56.

[35] L.-P. Morency, P. P. Liang, and A. Zadeh, "Tutorial on multimodal machine learning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, 2022, pp. 33–38.

[36] L. Zhang, J. Shen, J. Zhang, J. Xu, Z. Li, Y. Yao, and L. Yu, "Multimodal marketing intent analysis for effective targeted advertising," *IEEE Transactions on Multimedia*, vol. 24, pp. 1830–1843, 2021.

[37] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *International Conference on Learning Representations (ICLR)*, 2018.

[38] Z. Zeng, Y. Hu, G. I. Roisman, Z. Wen, Y. Fu, and T. S. Huang, "Audio-visual spontaneous emotion recognition," in *Artifical intelligence for human computing*. Springer, 2007, pp. 72–90.

[39] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] G. Wang, H. Han, S. Shan, and X. Chen, "Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 56–69, 2020.

[41] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.

[42] P. Dale, J. George, F. David, C. H. William, L. Anthony-Samuel, O. M. James, and W. S. Mark, *Neuroscience*. Sinauer Associates, Inc, 2004.

[43] D. Zhang, Y. Hu, and Y. Chen, "Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3904–3914, 2020.

[44] Z. Wu, D. Zhang, C. Xie, C. Yu, J. Chen, Y. Hu, and Y. Chen, "Rfmask: A simple baseline for human silhouette segmentation with radio signals," *IEEE Transactions on Multimedia*, 2022.

[45] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Breathtrack: Tracking indoor human breath status via commodity wifi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.

[46] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.

[47] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[48] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen, "Towards highly accurate and stable face alignment for high-resolution videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8893–8900.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, 2017.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[51] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[53] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[55] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, pp. 369-386, 2019.

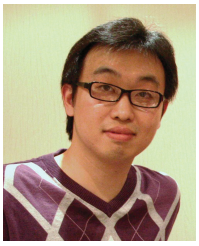
Dongheng Zhang received the B.S. and Ph.D degrees from University of Electronic Science and Technology of China, Chengdu, China, in 2017 and 2021, respectively. He is currently a Postdoctoral Researcher at School of Cyber Science and Technology, University of Science and Technology of China, Hefei, China. His research interests are in signal processing, wireless communications and networking.

Jia Meng received the B.S. and M.E. degrees in signal and information processing from Southeast University, Nanjing, China, in 2003 and 2006, respectively. He is currently a researcher at Tencent YouTu Lab. His research interests include face anti-spoofing, computer vision, and signal processing.

Jian Zhang received the bachelor and master degree at Shanghai Jiao Tong University, Shanghai, China, in 2017 and 2020, respectively. He is currently a computer vision researcher at Tencent YouTu Lab. His research interests include face anti-spoofing, transfer learning and few-shot learning.

Xinzhe Deng received the B.S. from Huazhong University of Science and Technology and the M.S. from University of Illinois Urbana-Champaign, in 2015 and 2019, respectively. He is currently a researcher at Youtu Lab, Tencent. His research interests lie primarily in deep learning, machine learning and computer vision.

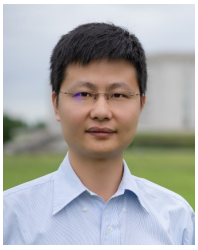
Man Zhou is currently an associate professor with the School of Cyber Science and Engineering, Huazhong University of Science and Technology. He received his Ph.D. degree in Cyberspace Security in 2021, and his B.E. degree in Information Security in 2016, from Wuhan University, China. His research interests include mobile security, authentication security, AI system security, and sensing security. He was the recipient of "National scholarship for graduate students, China" in 2016-2018 and 2020.



Qian Wang is a Professor in the School of Cyber Science and Engineering at Wuhan University, China. He was selected into the National High-level Young Talents Program of China, and listed among the World's Top 2% Scientists by Stanford University. He also received the National Science Fund for Excellent Young Scholars of China in 2018. He has long been engaged in the research of cyberspace security, with focus on AI security, data outsourcing security and privacy, wireless systems security, and applied cryptography. He was a recipient of the 2018 IEEE TCSC Award for Excellence in Scalable Computing (early career researcher) and the 2016 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He has published 200+ papers, with 120+ publications in top-tier international conferences, including USENIX NSDI, ACM CCS, USENIX Security, NDSS, ACM MobiCom, ICML, etc., with 20000+ Google Scholar citations. He is also a co-recipient of 8 Best Paper and Best Student Paper Awards from prestigious conferences, including ICDCS, IEEE ICNP, etc. He serves as Associate Editors for IEEE Transactions on Dependable and Secure Computing (TDSC) and IEEE Transactions on Information Forensics and Security (TIFS). He is a fellow of the IEEE, and a member of the ACM.

Qi Li received his Ph.D. degree from Tsinghua University. He is currently an Associate Professor with the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include network and system security, particularly Internet and cloud security, IoT security, and machine learning security. He is currently an editorial board member of IEEE TDSC and ACM DRTAP.

Shouhong Ding received his Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China, in 2016. Now he is a senior researcher of Youtu Lab in Tencent, China. His current research interests include image/video processing, computer vision, pattern recognition and multimedia technology.



Yan Chen (SM'14) received the bachelor degree from the University of Science and Technology of China in 2004, the M.Phil. degree from the Hong Kong University of Science and Technology in 2007, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2011. He was with Origin Wireless Inc. as a Founding Principal Technologist. From Sept. 2015 to Feb. 2020, he was a Professor with the School of Information and Communication Engineering at the University of Electronic Science and Technology of China. He is currently a Professor with the School of Cyber Science and Technology at the University of Science and Technology of China.

Dr. Chen's research interests include multimodal sensing and imaging, multimedia signal processing, and wireless multimedia. He is a co-author of "Reciprocity, Evolution, and Decision Games in Network and Data Science" (Cambridge University Press, 2021) and "Behavior and Evolutionary Dynamics in Crowd Networks: An Evolutionary Game Approach" (Springer, 2020), as well as co-author of over 200 technical papers including more than 100 IEEE journal papers. He is the Associate Editor for IEEE Transactions on Network Science and Engineering (TNSE) and IEEE Transactions on Signal and Information Processing over Networks (TSIPN). He is the Chair for APSIPA Signal and Information Processing Theory and Methods (SIPTM) Technical Committee, a Distinguished Lecturer for APSIPA, the Secretary-General for the CES Young Scientist Network Multimedia Technical Committee. He is an Organizing Co-Chair of PCM 2017, a Special Session Co-Chair of APSIPA ASC 2017, the 10K Best Paper Award Committee Member of ICME 2017, the Multimedia Communications Symposium Lead Chair of WCSP 2019, an Area Chair for ACM Multimedia 2021, a TPC Co-Chair of APSIPA ASC 2021. He was the recipient of multiple honors and awards, including an Excellent Editor for IEEE TNSE in 2021, the best paper award at the APSIPA ASC in 2020, the best student paper award at the PCM in 2017, the best student paper award at the IEEE ICASSP in 2016, the best paper award at the IEEE GLOBECOM in 2013.