SONICUMOS: An Enhanced Active Face Liveness Detection System via Ultrasonic and Video Signals

Yihao Wu, Peipei Jiang, Jianhao Cheng, Lingchen Zhao, Chao Shen, *Senior Member, IEEE*, Cong Wang, *Fellow, IEEE*, and Qian Wang, *Fellow, IEEE*

***Abstract*—SONICUMOS is an enhanced behavior-based face liveness detection system that combines ultrasonic and video signals to sense the 3D head gestures. As face authentication becomes increasingly prevalent, the need for a reliable liveness detection system is paramount. Traditional behavior-based liveness detection methods (e.g., eye-blinking, nodding, etc.), which are widely deployed in mission-critical scenarios like finance and banking applications today, are prone to advanced media-based facial forgery attacks.**

SONICUMOS aims to incorporate the traditional behavior-based method for active liveness detection without introducing extra user burden. By employing ultrasonic signals, SONICUMOS capitalizes on the head gestures, significantly raising the security bar. Our approach utilizes the frequency-modulated continuous-wave (FMCW) ultrasonic radar for robust 3D gesture recognition compatible with face authentication. We also propose a new dual-feature fusion network that integrates audio and video features at the feature level to increase detection accuracy and resilience against numerous attacks. Our prototype has been tested on seven off-the-shelf Android/iOS smartphones, achieving an overall detection accuracy of 95.83% at an equal error rate (EER) of 4.96% when dealing with 3D impersonation attacks.

***Index Terms*—Face liveness detection, face and gesture recognition, ultrasound, mobile device security.**

I. INTRODUCTION

FACE authentication is arguably one of the most popular biometric authentication methods today, with the market expected to reach $12 billion by 2028 [1]. However, despite being a popular choice, face authentication is vulnerable to many low-cost and high-success rate presentation attacks [2], [3]. By simply using samples reconstructed from the victim's publicly exposed media or candid photos, the attackers can replay them on physical photos or high-resolution digital screens to easily spoof the face authentication systems.

Y. Wu, J. Cheng, L. Zhao and Q. Wang are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: {yihao.wu, polariris, lczhaocs, qianwang}@whu.edu.cn).

P. Jiang is with School of Cyber Science and Engineering, Wuhan University, Hubei 430072, China, and also with the Laboratory for AI-Powered Financial Technologies Ltd., Hong Kong SAR 999077, China. (e-mail: pp.jiang@my.cityu.edu.hk).

C. Shen is with the MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China, and also with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: chaoshen@mail.xjtu.edu.cn).

C. Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR 999077, China (e-mail: congwang@cityu.edu.hk).

(Corresponding authors: Peipei Jiang and Lingchen Zhao).

To combat the vulnerabilities of face-spoofing attacks, researchers have devised liveness detection techniques that analyze unique features of live users during face authentication. These techniques can be implemented in two ways: (1) *actively*, where users are required to participate actively in the authentication process, such as responding to simple movement instructions [4], [5], or (2) *passively* without the user's active engagement. Passive liveness detection methods typically rely on computer vision techniques to evaluate facial texture patterns or reconstruct depth information to distinguish fake 2D photo/video samples [6]–[12] (more related work in §II-A). In the industry, the common choice is to extract facial depth features from a 3D depth-sensing camera (e.g., by Apple, Samsung, etc.). However, the depth information can still be forged from 2D images [13]. Furthermore, an even more significant concern is the lack of awareness in passive systems, as attackers can use covert or candid cameras to perform the detection without the users' consent. For example, even the popular Apple FaceID, equipped with (passive) attention-based detection, can still be fooled by putting glasses to cover the eye area, allowing access to the device while the user was asleep or unconscious [14].

Hence, we advocate that involving users' active engagement in liveness detection is still important in many security-sensitive and mission-critical scenarios. Among the active liveness detection methods, the traditional behavior-based ones [4], [15]–[20] are the most widely-used ones today in many finance/banking applications where confirming the presence and the awareness of the user is a must. This is because such methods are reliable and effective in validating users' awareness, while the interactions are not heavy for users. However, they have the same limitations as traditional passive 2D liveness detection methods. The recorded 2D samples can be easily tampered with through video replay or editing, known as media-based facial forgery (MFF) attacks [21].

In this paper, we explore how to enhance the widely-used 2D behavior-based liveness detection by integrating 3D sensing capabilities to defending against the advanced 2D/3D MFF attacks [13], [21]. We aim to enhance security without increasing the user's burden. By leveraging natural user movements and integrating *ultrasonic* signals, we add an extra layer of protection, significantly improving security while preserving the user experience. Leveraging *ultrasound* to feature the dynamic gestures offers two key benefits: (1) Through a carefully designed approach, the acoustic signals reflected by the perceived facial behaviors, attain unforgeable 3D spatial resolution capabilities, effectively thwarting

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

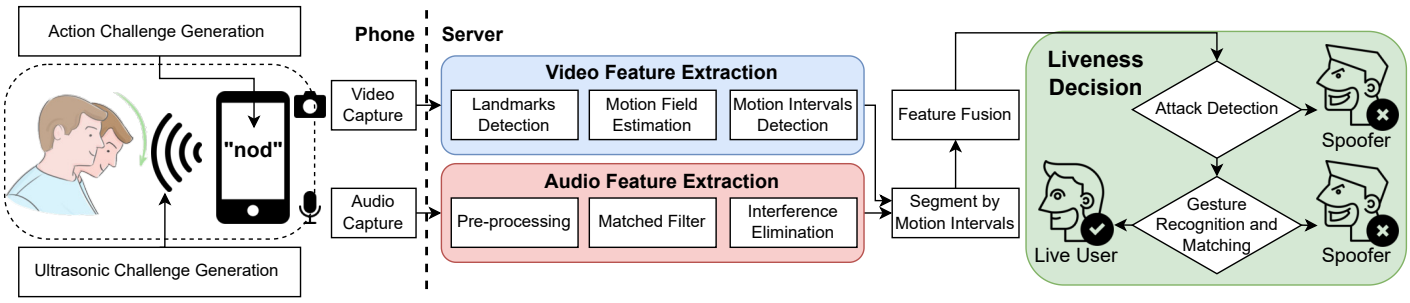


Fig. 1. System overview of SONICUMOS.

advanced MFF attacks. (2) The system can effortlessly expand the randomness space by incorporating challenges into the acoustic signal with random carrier frequencies to mitigate replay attacks. Compared to the traditional methods that use a sequence of behaviors as challenges, it largely eases user burden by requiring only one behavior. Lastly, from a practical standpoint, integrating ultrasound sensing into face liveness detection only further necessitates a microphone and a speaker, which are low-cost components readily available on today's smartphones.

Technical overview. We propose SONICUMOS[1], an active liveness detection system that combines ultrasonic and video signals to provide strong security, high accuracy, resilience to environmental interference, and good compatibility with face authentication. The high-level idea is to let the user respond to an action challenge with a specific head gesture (e.g., nodding) while simultaneously responding to ultrasonic signals embedded with frequency challenges on the speaker (as depicted in the left part of Figure 1). Next, the camera records the video of the head gesture, while the microphone captures the acoustic signals reflected by the user's face, known as the micro-Doppler signal. Finally, through feature extraction and fusion steps, the system determines the user's liveness.

1) Robust 3D gesture recognition via ultrasound. Using ultrasonic signal reflection to sense facial movement is not new in literature [22]–[29]. However, while existing research primarily focuses on lip motion recognition within an effective distance of 10cm between the user and the device, the face authentication requires a 15cm distance to fully capture the user's face. A greater distance implies a broader detection scope for the signals, making the recognition performance more susceptible to dynamic interference (e.g., movements from surrounding objects). As a result, rather than utilizing *single-frequency* continuous-wave (SFCW) as in prior art, we adopt a new technical approach, specifically employing *frequency-modulated* continuous-wave (FMCW) ultrasonic radar, which offers good velocity, range resolution, and multi-path interference resolution capabilities [30]. The key idea is to utilize FMCW radar to extract the motion features within a specific range (with estimation), effectively eliminating dynamic interference originating from illegitimate distances.

[1]SONICUMOS stems from "Sonic" plus the "Lumos" spell from Harry Potter series. It suggests a system that uses acoustic signals to detect and illuminate live faces.

2) Fusion of video and audio features. Different from prior liveness detection system solely via acoustic signals [22]–[26], SONICUMOS utilizes video signals as a second modality alongside ultrasonic sensing for face liveness detection. As SONICUMOS is compatible with face authentication, recording face video becomes an effortless (regarding user experience) and advantageous choice. It naturally provides two additional "guards" for liveness detection: face biometrics and feature consistency, both of which raise the bar for attackers attempting to forge spoofing samples. For instance, the attacker must not only forge both the acoustic feature (to bypass liveness detection) and the video feature of the victim's face (to bypass the face authentication based on face biometric), but also carefully conduct the segmentation and alignment to ensure the consistency between these features. This is a challenging task because all the samples should be generated in real time to respond to the random challenges correctly. For consistency analysis, SONICUMOS employs a dual-modal fusion network to integrate the extracted audio and video features at the higher-dimensional feature level, rather than simply checking their consistency in the time dimension. As a result, the fusion process further bolsters the system's accuracy and resilience against attacks.

Contributions. In summary, this paper contributes:

- A new 3D head gesture recognition approach using FMCW radar, offering robustness to dynamic interferences and compatibility with face authentication.
- A new dual-feature fusion network that integrates audio and video features at the feature level.
- A prototype of SONICUMOS on seven off-the-shelf Android/iOS smartphones, achieving an overall accuracy of 95.83% at an EER of 4.96% when dealing with 3D impersonation attacks.

II. BACKGROUNDS AND RELATED WORK

For security, face authentication often employs liveness detection techniques, with passive methods operating without user interaction and active methods involving a challenge-response mechanism. In this section, we introduce the related literature on passive and active liveness detection, as well as the typical system model for face authentication with behavior-based liveness detection.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

A. Passive Liveness Detection

Passive liveness detection does not require users' interaction and focuses on identifying and evaluating the difference in subtle characteristics present in face images/videos between fake and real samples. Representative characteristics include skin texture, depth information, environmental features, and casual behaviors of users, etc.

In recent years, passive liveness detection has explored various methods to distinguish live faces from fake ones. Early work focused on handcrafted features using extractors like LBP, DoG [7], IDA [31], SIFT [32], and SURF [33] to analyze facial surface properties [7]–[9], [17], [31]–[34]. Later studies shifted to deep learning models like CNN and LSTM [35]–[37]. Some methods also use hardware to obtain 3D information or skin vibrations, such as multispectral cameras [38], RFID [39], infrared cameras [40], and mmWave radars [41]. Other approaches modify the environment to reveal physical phenomena [12], [42]–[45]. Additionally, techniques relying on casual behaviors, like blinking or body movements, have been studied for improving robustness against sophisticated spoofing attacks [46], [47].

Despite being lightweight and user-friendly, passive detection methods remain vulnerable to attacks such as 3D masks [26], [48] and 3D image reconstruction [13]. More critically, attackers can exploit the lack of user interaction for unauthorized access, using phishing techniques or targeting unconscious users [14]. This highlights the need for active techniques in security-sensitive scenarios, like modern banking applications.

B. Active Liveness Detection

Active liveness detection usually involves a challenge-response mechanism, where users are required to engage with the system. Traditional behavior-based methods prompt users to perform a specific action (challenge), such as lip movements [16], [19], eye blink [4], [20], or head rotations [15], [17]–[19], [49]. While these methods resist the 2D presentation attacks, the are vulnerable to simple 3D mask attacks [48], [50].

To obtain 3D information, a line of work [11], [21], [51] requires users to move their smartphones during the authentication period, and the system analyzes videos captured from different camera positions. With the help of auxiliary sensors, FaceLive [21] measures the consistency between the motion data obtained from the inertial sensors and the head poses captured in the video. FaceFlashing [3] flashed well-crafted images displayed on a screen as the challenge and examined the reflected light while the user should make a facial expression.

Another line of research explores the use of alternative biometrics, such as gaze movement [52], [53], voiceprint [54], pop noise [55], [56] and lip motions [26], [29]. However, they may introduce inconveniences for users (e.g., having to make speak out) or face challenges in proving the uniqueness of biometrics within a large user base, which may negatively impact user trust. Indeed, whether these biometric features can achieve the same effectiveness as well-established methods like facial recognition and fingerprints, which have proven reliable for large-scale client bases and are widely recognized as unique identifiers, still requires further validation and investigation.

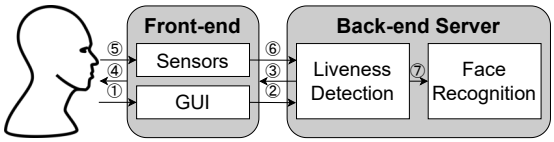


Fig. 2. System model of a typical face authentication system with active liveness detection.

A more related line to our work is to leverage acoustic signals to sense the physical features of a live person, e.g., the articulatory gesture [23], [25], [57]–[60] or the time-difference-of-arrival (TDoA) changes in a sequence of phonemes [22]. However, these acoustic-based methods are also limited by hardware cost and signal type. Besides, since vocal posture movements are subtle, they need to be implemented at a very short distance, which is not suitable for face authentication at a longer distance (detailed discussion in §III). A recent and related work, SonarGuard [29], also utilizes the temporal consistency between ultrasonic and video signals in users' lip movements. However, its method employs SFCW to capture audio features and relies exclusively on lip features as the action challenge, resulting in suboptimal performance. Since the movement consistency check only focuses on the lip area, there are chances for attackers to evade the face authentication [26].

To summarize, SONICUMOS makes advancements by 1) introducing a more robust 3D head motion recognition system, which supports various head gestures leveraging advanced frequency-modulated continuous-wave techniques; 2) ensuring compatibility with existing 2D behavior-based face liveness detection systems while preserving the facial biometric features; and 3) offering a more seamless and user-friendly experience, as it requires no additional effort or knowledge from users compared to commonly used methods, while still enhancing resilience against targeted attacks.

C. Face Authentication with Behavior-based Liveness Detection

A typical face authentication system with active liveness detection consists of a front-end device and a back-end server [3], as illustrated in Figure 2. The front-end device captures the user's face biometrics using sensors (⑤) and displays (④) the authentication results, while providing a user interface for interaction (①). The back-end server is composed of two modules: liveness detection and face recognition. During the authentication process, the back-end server sends challenge parameters to the front-end (③) upon receiving a user-initiated request (②). The user then performs head gestures in response, which are captured by the front-end sensors and sent back to the liveness detection module (⑥). If the user is determined to be live, the recorded video is forwarded to the face recognition module for identification (⑦).

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

III. Revisiting Ultrasonic Sensing for Facial Gesture on Smartphone

Using ultrasonic to sense 3D objects and movement is a hot topic in recent decades, among which utilizing the micro-Doppler effect [61]–[63] is one representative technical route for gesture recognition technologies [23], [30]. The common practice is to treat the smartphone as a radar (primary source) to transmit *ultrasound* and receive the reflected signal from the moving object (secondary source). By analyzing the Doppler frequency shift of the received signal, the system can determine the characteristics of interest for movement.

In this section, we first introduce the micro-Doppler concept and revisit the prior efforts in ultrasonic sensing and illustrate our insight of using frequency continuous wave as ultrasonic radar to capture the micro-Doppler shift.

A. Micro-Doppler Effect

The Doppler effect refers to the change in frequency caused by the relative motion between a signal source and a signal receiver. When a primary source emits a signal toward an object, and the object reflects the signal, the object can be considered a secondary source. The movement of the object relative to the receiver results in a Doppler frequency shift. Assuming the object moves at velocity v relative to the signal receiver, the Doppler frequency shift is given by:

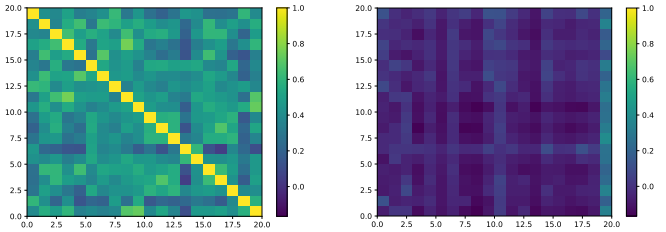
$$\Delta f = \frac{v}{c} f_s, \tag{1}$$

where c is the propagation velocity of the signal in the current medium, and f_s is the frequency of the reflected signal from the object. Besides the Doppler shift caused by the object's coarse-grained motion, any fine-grained micromotion dynamics, such as vibrations or rotations of the object or its components, induce modulations on the returned signal. This phenomenon, known as the micro-Doppler effect [61]–[63], introduces additional frequency components on top of the basic Doppler frequency shift.

B. Single Frequency Continuous Wave

To capture the Doppler frequency shifts, there are many choices for the ultrasound. We find that almost all existing related work on facial gesture sensing [23], [25]–[28] employ single-frequency continuous wave (SFCW) for the ultrasonic choice. However, as mentioned earlier, these methods typically require close proximity (within 10 cm) for effective sensing. Below we experimentally explored the feasibility of extending SFCW sensing to greater distances (15 cm and beyond). The results indicated that SFCW becomes highly sensitive to dynamic interference at these extended distances, limiting its effectiveness.

Our experiments demonstrate that SFCW-based sensing is significantly affected by nearby dynamic interference, reducing the correlation and classification accuracy of motion features. We transmit ultrasonic to sense the lip movement, which is 20 cm away from the smartphone, and a jammer waves his arm 60 cm away from the smartphone as dynamic interferences. The frequency of ultrasonic was set to 20kHz.



(a) No interference (b) With interference

Fig. 3. The matrix of the Pearson correlation coefficient. A value closer to 1 indicates a stronger positive linear correlation.

We collect 40 samples of lip movement that are as similar as possible in terms of the movement pattern, including 20 samples without interference and 20 samples with interference. The microphone of smartphone was configured in MIC mode and speaker in STREAM_VOICE_CALL mode. The audio recording is mono channel and the sampling rate is set at 48 kHz. Finally, audio is saved as a WAV file based on PCM encoding. After demodulation and filtering, we use the Pearson correlation coefficient to measure the correlation between extracted features. The Pearson correlation coefficient matrix of features without interference is shown in Figure 3(a). The average value of this matrix is 0.423, which indicates that the 20 interference-free actions are similar.

C. Frequency Modulated Continuous Wave

To overcome the limitations above, we introduce a method based on Frequency Modulated Continuous Wave (FMCW) radar, offering improved velocity and range resolution [30], [64]. We will provide detailed formula descriptions in Section V-B2. By utilizing ultrasonic radar capable of resolving multi-paths, we can extract motion features within a specific range, making the system more robust to surrounding dynamic interference.

In particular, this method eliminates the need to consider the impact of mobile device frequency response on sensing signal power and produces a strong enough signal to sense motion features within a wide range of scanners. The method based on FMCW radar ensures that the surrounding objects will not introduce unpredictable mixtures.

When designing the FMCW radar for our face authentication scenario, it's essential to consider the propagation paths from the source (i.e., the speaker) to the destination (i.e., the microphone). As illustrated in Figure 5(a), including the structural-borne path (via the body of the mobile device), the line-of-path (LOS) path (via air), the head reflection path (reflected by head gesture), and the environment reflection path (reflected by surrounding objects except for the victim). The power attenuation and phase shift will occur when signals propagate on different paths. The received acoustic signal is a mixture of multiple acoustic signals modulated with different phase shifts and attenuation coefficients of amplitude. Since the arrival time of the structural-borne path and the LOS path is close, it is difficult to distinguish them, we refer to them collectively here as the direct path.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

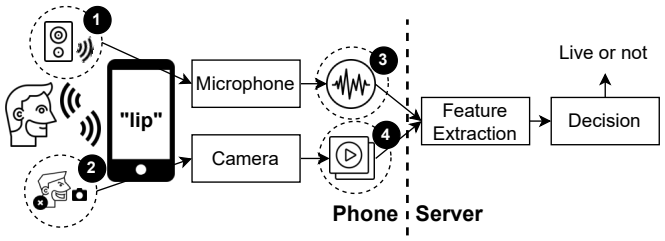
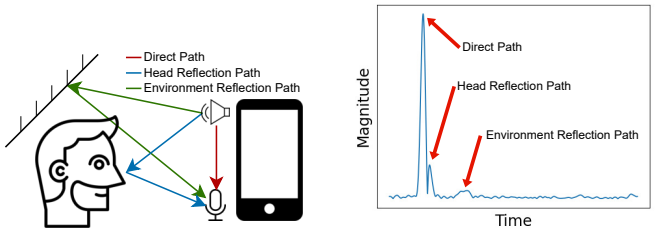


Fig. 4. The multiple vulnerable parts in SONICUMOS pipeline.



(a) Multi propagation paths (b) Correlation estimator

Fig. 5. Multipath and multi-peaks in a frame after the matched filter.

The direct path has the lowest power attenuation, and a peak denoting the direct path will appear on the correlation estimator after processing the mixed multi-path acoustic signal as shown in Figure 5(b). The head reflection path also leads to several peaks, but due to the greater phase shift and power attenuation, the peak will lag behind the peak of the direct path with a lower amplitude. The lag is greater for further interference objects.

To ensure the quality of the captured video, systems often incorporate visual indicators on the screen surface view to limit the distance between the user's face and the mobile device. The phase shift between the direct path and the head reflection path is within a relatively fixed range. Referring to its corresponding correlation estimator, the distance between the peak of the reflection path and the peak of the direct path is relatively fixed.

Since the direct path signal usually has the highest amplitude [42], [65], we can easily locate its peak position. With the peak as the benchmark, we trim these signals where their peaks fall outside a reserved range off the correlation estimator to eliminate the interference of the environmental reflection path with a larger phase shift and more lag, thus extracting the head motion features.

IV. THREAT MODEL

A. Attack Surfaces

A summary of the attack surfaces of SONICUMOS is shown in Figure 4. Firstly, the attacker can hijack/replace video or audio in the communication between the front-end and the back-end. We assume that the attackers can collect the raw and expired audio and video samples of the victims. We also assume that attackers can reconstruct gesture videos from accessible high-resolution photos through social media or public monitors. With the crafted video, the attackers can then select the appropriate response to specific challenge of the system requiring specific action and replace the audio/video during the transmission (❸ and ❹ in Figure 4).

We assume that the attackers can manipulate the inputs (❶ and ❷ in Figure 4) by attempting to conduct attacks using another device. In particular, the ultimate goal for the attacker is to bypass liveness detection and authentication simultaneously, so it is necessary for us to ensure that the input video contains valid face for authentication.

B. Face Spoofing Attacks

We also consider the following face spoofing attacks.

2D video attacks. The attacker attempts to defeat face biometric recognition by presenting non-live biometric data. Specifically, the attacker places a digit screen that replays the victim's facial video according to the challenge to assert a false identity in front of the front camera (❶).

Replay attacks. The attacker illegally collects the input audio and video pair, initiates a request for authentication, and records arbitrary signals through the microphone and speakers. During transmission between smartphone and server, the attacker replays previous sample pairs, replacing the original video with an expired video that contains the victim's face with a corrective action response and replacing the original audio with a corresponding audio that contains the micro-Doppler shift (❸ and ❹).

3D impersonation attacks. This attack happens after the system specifies a challenge. Within a tolerant period of time, the attackers selects the video of required action from a pre-constructed video collection a pre-collected set in replacement in real time [66]. As discussed, these videos can be easily forged by merging user's public photos with their facial features. Normally, the video-only attacks can be effectively resisted, as the system detects liveness by extracting the micro-Doppler feature in the audio response. So a more advanced attacker may try to synchronize with the head motion in the replacing fake video, i.e., the attacker mimics the "physical" action as that occurred in the video. 3D impersonation attacks take place in steps ❸ and ❹.

V. SONICUMOS DESIGN

A. System Overview

Figure 1 (appearing in Section I) shows the system architecture of SONICUMOS, consisting of seven modules, as summarized below.

Challenge Generation aims to generate and transmit a random action challenge and an ultrasonic challenge. The Random action is selected from a set with limited capacity. The ultrasonic challenge is modulation periods (sweeps). In a complete round of the liveness detection process - including challenges, responses, and the decision, the parameters for generating challenges will be delivered to the Feature Extraction module and Detection and Matching module for ensuring that they work properly.

Response Capturing invokes the front camera of the smartphone to record the video response, and the microphone to record the audio response, while ensuring that the video and

audio responses are synchronized, which is a prerequisite for separating features according to the time in the audio response later.

Audio Feature Extraction extracts the feature associated with the micro-Doppler effect induced by head motion from the audio response. It mainly includes three stages: pre-processing, matched filtering and interference elimination. The pre-processing aims to eliminate low-frequency interference and re-organized the signal at time scales. Matched filtering is mainly used to detect reflected signals through multi-path by demodulating. Interference elimination is used to remove interference signals that will degrade the quality of extracted features introduced by dynamic or static surroundings.

Video Feature Extraction performs facial landmarks detection from the video response and estimates the motion field of specified landmarks that can produce the obvious micro-Doppler effect. Eventually, a detection algorithm is performed to distinguish motion intervals and static intervals from the time series according to the motion field estimation.

Segmentation segments the micro-Doppler feature from the time-aligned audio response according to motion intervals to obtain two-modal features that represent motions.

Feature Fusion reconciles two-modal features using a well-designed feature fusion network at the feature level to learn the correlation between the audio and the video feature after segmentation.

Detection and Matching detects attacks and recognizes gestures using the fusion feature. If a user is accepted by the attack detection network and the recognition result of the action is consistent with the challenge transmitted by the system, this module marks the request as a legitimate user.

B. Challenge Generation

When the user proposes a face authentication request, the SONICUMOS server generates an action challenge and an ultrasonic challenge and transmits them to the front-end. Then the smartphone displays the action challenge as a prompt on the screen and invokes the top speaker to play the ultrasonic challenge. The parameters of generated challenges like ultrasonic's frequency and action type will be delivered to the others module for the subsequent processing.

1) Action Challenge: In line with traditional active liveness detection systems, SONICUMOS specifies a randomized instruction from a predefined set of actions, expecting an interactive user to respond accordingly. The selection of actions is carefully chosen to ensure reliable and robust detection. Subtle actions such as blinking or smiling, which generate weak micro-Doppler effects due to slow center-of-mass movement, minimal flexible head motion, or obstruction, are difficult for the system to accurately distinguish. Therefore, SONICUMOS opts for three more distinct and dynamic actions, i.e., lip movements, nodding, and head turning. These actions produce more pronounced micro-Doppler effects, making them easier for the ultrasonic-based system to detect and analyze.

2) Ultrasonic Challenge: In addition to the action challenge, which has relatively limited selectable parameters, SONICUMOS incorporates a more unpredictable challenge into the ultrasonic signals with random frequencies to thwart potential replay attacks. Frequent changes in frequency often cause a ringing effect [67], where high-powered impulses may be noticed by users, potentially affecting the user experience. To mitigate this, we design an ultrasonic challenge that is smoothed by using repeated continuous chirp signals [68], reducing sudden frequency changes and thereby minimizing the ringing effect [69]. This approach also enhances security by randomizing the chirps.

On the premise of ensuring the resolution and discrimination, a single chirp has more selectable parameters, which makes it difficult for the attackers to collect all the audio responses in advance or to predict the challenge during face authentication. A chirp $s(t)$ with period T is defined as:

$$s(t) = \begin{cases} \cos\left(2\pi f_1 t + \mu t^2/2 + \phi\right), & 0 \le t < T/2, \\ \cos\left(2\pi f_2 t - \mu t^2/2 + \phi\right), & T/2 \le t < T, \end{cases} \tag{2}$$

where $[f1, f2]$ is the frequency band of the chirp, μ is the chirp sweep rate in bandwidth $B = (f2 - f1)$, i.e., $\mu = 2\pi B/T$, and ϕ is the initial phase of the signal, which shifts the entire waveform by a constant phase value. $s(t)$ consists of continuous up-and-down symmetric chirp symbols, which avoids producing strident bur that humans can hear [69].

The range resolution of FMCW radar is defined as $\Delta r = \frac{c}{2B}$, where c is the speed of sound in the propagation medium and B represents the chirp's frequency bandwidth. To ensure that the radar has sufficient distance resolution, i.e., $\Delta r < 10$ cm, the system selects the initial frequency f_1 between 18 kHz and 21 kHz, with a bandwidth B greater than 2 kHz, signal duration T ranges from 20–40 ms, initial phase ϕ unconstrained and SNR fixed at 30 dB. Choosing f_1 above 18 kHz helps avoid interference from environmental noise, while frequencies in this range are generally inaudible to most people [70]. Furthermore, due to the Nyquist–Shannon sampling theorem and the hardware limitations of microphones, the maximum frequency of the ultrasonic signal is capped at 24 kHz. Figure 6(a) illustrates the frequency variation of the generated ultrasonic signal.

SONICUMOS generates the ultrasonic challenge by randomly selecting the critical parameters of chirp ultrasonic signals. Based on empirical parameters and experimental validation [71], [72], the minimum distinguishable differences between two chirp signals are defined as $\Delta f = 50\,\text{Hz}$ (frequency), $\Delta B = 50\,\text{Hz}$ (bandwidth), $\Delta T = 0.5\,\text{ms}$ (duration), and $\Delta\phi = 2.56^\circ$ (phase). Assuming parameter independence, the joint probability of an attacker successfully guessing all parameters is:

$$p = \frac{2}{\Delta f} \cdot \frac{2}{\Delta B} \cdot \frac{2}{\Delta T} \cdot \frac{2}{\Delta \phi} \approx 2.9 \times 10^{-6}.$$

This probability ($p \ll 0.1\%$) demonstrates the infeasibility of bypassing detection via random parameter guessing, thereby validating the anti-spoofing robustness of our parameter design.

3) Speaker/Microphone Selection: A typical smartphone is equipped with several audio and camera components, including a top speaker for calls, a bottom speaker, a top

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

microphone for noise cancellation, a bottom main microphone, and both front and rear cameras. For SONICUMOS's face sensing system, we configured the microphone in `MIC` mode and speaker in `STREAM_VOICE_CALL` mode which select the top speaker, top microphone, and front camera as the optimal device group. This selection is based on two key factors: (1) Almost all modern smartphones place the top speaker and front camera in close proximity, which minimizes alignment errors during sensing. (2) During face authentication, users typically focus on the smartphone screen, making the top speaker better positioned to "illuminate" the face with ultrasonic signals. Additionally, the top speaker's proximity to the ultrasonic radar enhances signal reception through the direct path, simplifying the task of identifying the peak. The top speaker is also less affected by variations in the user's posture while holding the smartphone, further improving the robustness of the system.

C. Response Capturing

SONICUMOS invokes the front camera of the smartphone to record the video response and invokes the top microphone to record the audio response. The system prompts the user to adjust their head using a visual indicator on the screen, ensuring a relatively static position before proceeding with the challenge. Once the head is stable, the user is instructed to perform the corresponding action as prompted. After the instruction is issued for a fixed duration, SONICUMOS stops all input devices to complete the recording. The collected data includes the audio response and the video response, which will be delivered to the server for liveness detection after encoding. A raw signal of captured audio response is depicted in Figure 6(b). The video and audio responses are synchronized at time scales, which is a prerequisite for segmentation according to the motion intervals. In addition, we consider the time-of-flight (TOF) of sensing ultrasonic between the face and the smartphone, which is equal to $\Delta t = \frac{2d}{c}$. We add delay Δt to the video response to compensate for the inconsistency of time between them.

D. Audio Feature Extraction

There are multiple propagation paths from the speaker to the microphone, including the direct path, the head reflection path, and the environment reflection path, as shown in Figure 5(a). Assuming that there are N paths, the recorded audio response can be represented as:

$$r(t) = \sum_{k=0}^{N} A_k(t)\{\hat{s}(t - \frac{d_k(t)}{v}) - \theta_k\}, \tag{3}$$

where $\hat{s}$ is transmitted chirp $s(t)$ with a frequency shift $f(t)$, k denotes the k-th path, $A_k(t)$ denotes the amplitude of acoustic signals in k-th path, $\frac{d_k(t)}{v}$ de notes the phase shift caused by delay spread of acoustic channel, and θ_k denotes the phase shift caused by system delay.

Transmitted chirps $s(t)$ can be regarded as the carrier. During multi-path propagation, modulation occurs, leading to both phase and frequency shifts in the recorded audio response.

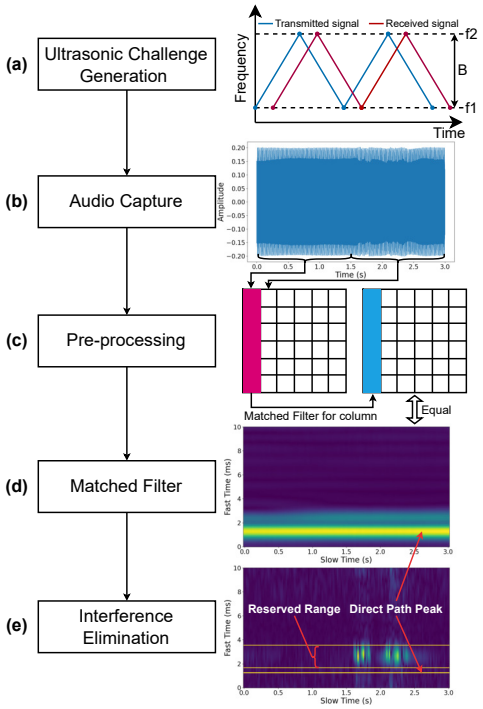


Fig. 6. The pipeline of the audio feature extraction.

For the direct path and reflection paths involving static objects in the environment, the phase shift remains constant, and there is no frequency shift. In contrast, the phase and frequency shifts for paths reflected by the user's head movements and other surrounding dynamic objects are time-variant. The time-varying phase and frequency shifts caused by head gestures are the key features that SONICUMOS aims to extract for accurate liveness detection.

Due to the rotation of rigid bodies or the relative motion of non-rigid bodies, the features are inherently correlated to the transmitted chirps. This correlation makes it possible to extract these features using a matched filter, which can at the same time maximize the signal-to-noise ratio (SNR) in the presence of additive stochastic noise. After obtaining the secret parameters of the playback chirp from the Challenge Generation module, SONICUMOS utilizes the re-generated chirp to detect all reflected periodic chirps that have propagated through the multi-path in the audio response. The detailed steps are introduced below.

1) Pre-processing: Firstly, the audio response needs to pass through a high-pass filter to eliminate low-frequency interferences. We use the Chebyshev filter with a passband cutoff frequency of 17 kHz, a stopband cutoff frequency of 16 kHz, a passband ripple of 0.1, and a stopband attenuation of 50 dB. After filtering, the audio response is divided into multiple fixed-length frames, each matching the length of a single chirp. These frames are then reorganized into a matrix (M^a), as depicted in Figure 6(c), where each column corresponds to a frame of chirp-length duration.

Although each chirp is brief (i.e., the signal in each column M^a is short in duration), the accumulated data across multiple columns spans a longer time frame. SONICUMOS processes the

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

reflected audio response on fast time scales (i.e., focusing on short time frames) to determine the range of multiple sensed objects, and on slow time scales (i.e., over longer durations) to estimate the speed of these objects.

2) Matched Filter: After pre-processing, a non-coherent demodulator is applied to filter every column in M^a. The specific method to generate an estimator by the matched filter is convolution with the time-reversed signal of the transmitted chirp $s(t)$ as below:

$$e(t) = \mathcal{F}^{-1}\{\mathcal{F}\{w(t) \cdot r(t)\}\mathcal{F}\{w(t) \cdot s(T-t)\}\}, \quad (4)$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote FFT and inverse FFT, respectively. $w(t)$ denotes the hamming window function.

The matched filter achieves coherent superposition of the signal in the time domain by performing phase cancellation, which maximizes the SNR for correlated signals. If the amplitude-frequency characteristics of the matched filter and the audio response are consistent, the matched filter could better amplify the strong frequency points of the signal and remove part of the clutter in the audio response simultaneously. Figure 6(d) illustrates an example of M^a after the matched filter.

3) Interference Elimination: The feature extracted through the above steps still contains all the objects within the sensing range, including static interference caused by the propagation through the environment reflection path and the direct path, as well as dynamic interferences from surrounding moving objects. To deal with the interferences, SONICUMOS employs a two-step process to eliminate these interferences.

Dynamic Interference Elimination. The primary feature we seek to preserve is the signal reflected from the user's head. During the process of face authentication, the distance between user's face and the device can be restrained to a short range using the virtual indicator on the screen, and the gap between the surrounding dynamic objects and the user's head is sufficiently large for separation. Well-parameterized FMCW Radar has a good range resolution, which makes it possible to eliminate dynamic interference by intercepting intervals.

Figure 6(e) shows an example of the interference elimination process. After applying the matched filter to each column of M^a, multiple baseband signals modulated by phase shift and frequency shift will be amplified, resulting in multi-peaks in the frame. Since the attenuation of signal power is the lowest on the direct path, the peak with the highest amplitude in each frame denotes the direct path. Therefore, we regard the top peak as the benchmark.

The phase shift caused by motion and the TOF delay between the radar and the object result in the peaks from other reflected paths arriving later than those from the direct path. These multiple peaks, corresponding to different propagation paths, are shown in Figure 5(b). Our goal is to retain the peaks associated with the head reflection path while discarding those from the environment reflection path. The relative distance between the peak of the direct path and the head reflection path is consistently within a fixed range, while the location between the peak of the head reflection path and the environment reflection path is distinguishable owing to a larger distance. Thus, we keep all peaks in a range

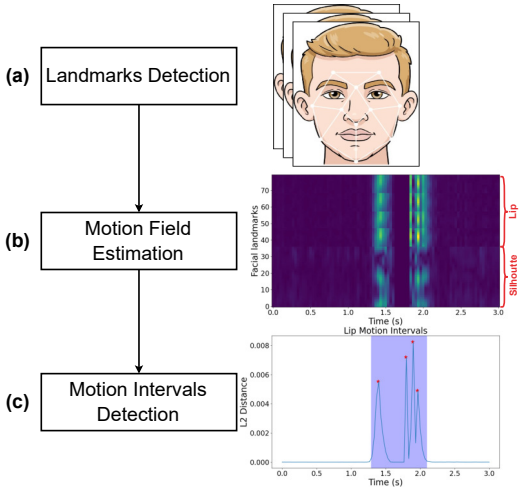


Fig. 7. The pipeline of the video feature extraction.

that lags behind the benchmark, that is, we trim peaks that fall outside the reserved range off the correlation estimator to eliminate the interference of surrounding dynamic objects. Through experiments, we determine the appropriate distance thresholds relative to the direct path for different liveness detection distances to optimize the interference elimination process.

Static Interference Elimination. The recorded audio response contains static interferences from surrounding static objects, including the user's head, leading to constant values in the matrix M^a. Since we discard multiple peaks of dynamic interference by intercepting, the peak corresponding to the head reflection path is still reserved after dynamic interference elimination, which will also produce non-negligible static interference. To address this, we apply a window function within the matched filter to mitigate frequency leakage. However, due to the nature of the up/down chirp sweeps, the matched filter cannot sufficiently reduce the amplitude of sidelobes from all peaks, which leaves non-zero constants in M^a. To eliminate this static interference, SONICUMOS computes the gradient of each row of M^a. Theoretically, if there is no dynamic phase shift and frequency shift, the resulting M^a should consist entirely of zeros. After applying the gradient, as shown in Figure 6(e), the desired features become much more prominent and distinguishable.

E. Video Feature Extraction

Along with the audio, SONICUMOS records the complete video capturing the user's interactions with the system. The video provides richer and more precise motion features than the audio. To extract head motion features, we track the positions of prominent facial landmarks and estimate the head motion field, which serves as the key feature from the video response. In scenarios like 3D impersonation attacks, attackers may attempt to deceive the system using both crafted or pre-recorded video and audio. However, perfectly aligning motion moments across these two modalities is challenging. In order to place more obstacles for the attackers and recognize head

gestures more reliably and precisely, we use the motion field estimation extracted from the video response to accurately detect motion intervals in a fine-grained way. These motion intervals are used to segment both the video and audio responses simultaneously, ensuring synchronized and fine-grained analysis.

1) Landmarks Detection and Motion Field Estimation: For the recorded video response, we first detect facial landmarks and calculate their variations to estimate the motion field. We use a 3D face detector [73] to track 468 facial features of the user's face, as illustrated in Figure 7(a). These features, referred to as vertices, are combined with a predefined triangulated face mesh that has a fixed topology, while depth is predicted using a deep neural network. We then filter out vertices that remain relatively static throughout the duration of the challenge actions (e.g., those around the eyebrows) and focus on estimating the motion field of the head. Specifically, we calculate the $L2$ Euclidean distance between adjacent frames, where this distance reflects the motion velocity of the head in a straightforward manner.

2) Motion Intervals Detection: The velocity of facial landmarks in the video can directly reflect the 3D movement of users in the physical world. The action challenges require the user to revert to the original gesture after completing the action, and the user may have prolonged inaction during the interaction. In order to obtain a complete range from the beginning to the end of drastic actions that can induce the micro-Doppler shift, and eliminate the impact of micro motions during the preparation stage, we develop Algorithm 1 to realize fine-gained motion intervals detection. The algorithm consists of three key steps:

a. Identify peaks that exceed a set threshold ($threshold_1$), where the peak point must be greater than its left and right neighbors.
b. Perform center-proliferation for each peak to define multiple motion intervals.
c. Merge overlapping motion intervals to form continuous segments.

An example of motion interval detection of a head-turn action is shown in Figure 7(c). The highlighted "blue" region represents the detected motion intervals, which will be used to segment features extracted from both the video and audio data.

Algorithm 1 Motion Intervals Detection in Video

Require: Recorded video response X with N frames
Ensure: The motion intervals Y

Description: $L_{\mathbf{k}} = \{l_1, l_2, ...\}$ denotes the $L2$ distance of 3D spatial coordinate of filtered landmarks between k-th frame and $k+1$-th frame;
$A_{\mathbf{k}} \leftarrow$ AVERAGE($L_{\mathbf{k}}$) **for** $k = 1, ..., N-1$
$P = \emptyset$
for $k = 2, ..., N-2$ **do**
  if $A_{\mathbf{k}} \geq threshold_1$ & $A_{\mathbf{k-1}} < A_{\mathbf{k}} < A_{\mathbf{k+1}}$ **then**
    $P = P \cup A_{\mathbf{k}}$
  end if
end for
$Y = \emptyset$
for $i = 1, ...,$ LENGTH(P) **do**
  $left \leftarrow i; right \leftarrow i$
  while $A_{left} \geq threshold_2$ **do**
    $left \leftarrow left - 1$
  end while
  while $A_{right} \geq threshold_2$ **do**
    $right \leftarrow right + 1$
  end while
  $Y = Y \cup [left, right]$
end for
$Y \leftarrow$ MERGEINTERVALS(Y)
return Y

F. Segmentation

In 3D impersonation attacks, the attacker will inject the correct video response forged or collected in advance, and try to imitate the corresponding action in the range of sensing ultrasonic. Imitation is difficult to align completely with the video response on some physical parameters like start and end time, mouth opening amplitude, head-turning velocity, etc. We first utilize start and end time to defend against 3D impersonation attacks. Due to the synchronization between audio and video response, we directly utilize fine-grained motion intervals extracted from the video response to segment the audio. If the audio and video motion intervals are not strictly aligned in time, audio features after segmentation will lose partial motion information, which can be effectively recognized by the detector. The audio feature and video feature is expressed as M_{in}^{a} and M_{in}^{v}, respectively.

G. Multi-Modal Feature Fusion

As mentioned earlier, relying solely on video features fails to effectively detect 2D video attacks, and using only audio features cannot address 3D impersonation attacks. Therefore, we integrate both video and audio features to decide whether to admit a user.

Feature fusion methods are generally divided into four levels: data, feature, score, and decision layers [28]. In SONICUMOS, we adopt feature-level fusion by constructing a dual-modal fusion network. This network extracts temporal features from both audio and video data and subsequently fuses them to fully exploit their temporal correlations. Specifically, the audio features capture the micro-Doppler frequency shifts induced by head posture, while the video features reflect the variations in motion velocity.

Note that the feature maps produced by both modalities have one dimension representing time and another with different physical meanings. To facilitate subsequent fusion, we rearrange the feature maps as follows:

- The audio feature is structured as a T-T matrix, where one dimension corresponds to slow time scales and the other to fast time scales. We regard it as a three-dimensional matrix $M_{in}^{a} \in \mathbb{R}^{1 \times T^{slow} \times T^{fast}}$ with a single channel.
- The video feature is a matrix representing motion field estimations across multiple facial landmarks, reflecting

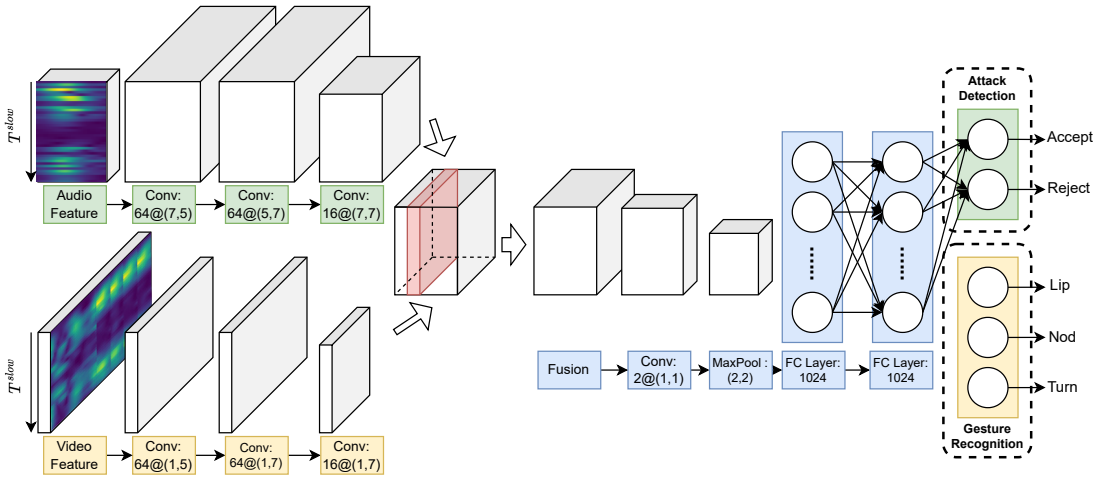


Fig. 8. The architecture of the feature fusion network.

changes in speed. Since the positions of these landmarks are random and adjacent rows do not imply correlation, we treat the dimension representing facial landmarks as the channel dimension. Thus, the video feature is restructured as $M_{in}^{v} \in \mathbb{R}^{C^{fl} \times T^{slow} \times 1}$, where $C^{fl} = 79$ denotes the number of selected facial landmarks used to generate the micro-Doppler frequency shifts.

Given that the sizes of the feature maps do not match after rearrangement, direct concatenation is not feasible. To overcome this, we design two embedding subnetworks that map the respective features into a shared feature space for further processing.

First, we focus on embedding the audio features. The input M_{in}^{a} is processed by the audio embedding network, as shown in the upper part of Figure 8. Due to rapid and frequent head movements causing several short-duration micro-movements, the network is able to capture the micro-Doppler frequency shifts induced by head posture. We emphasize capturing these correlations locally rather than globally by employing small convolution kernels. Moreover, to extract micro-features at various scales, we utilize convolution kernels of different sizes. After three convolutional layers, the resulting audio feature map is $M_{out}^{a} \in \mathbb{R}^{16 \times T^{slow} \times T^{fast}}$.

For video feature embedding, as illustrated in the lower part of Figure 8, the input is M_{in}^{v}. Considering that the input "image" has dimensions $T^{slow} \times 1$ with C^{fl} channels, we use one-dimensional convolution kernels to focus solely on temporal feature extraction. The output of the video embedding subnetwork is $M_{out}^{v} \in \mathbb{R}^{16 \times T^{slow} \times 1}$.

After feature embedding, we address the alignment of the two feature sets along the slow time dimension. Instead of simply concatenating the two feature maps, we fuse them by performing a dot product across all channels, which yields the fused feature map $M_{in}^{f} \in \mathbb{R}^{16 \times T^{slow} \times T^{fast}}$,

$$F_{\text{fused}} = \sum_{c=1}^{16} M_{out}^{a}(c) \cdot M_{out}^{v}(c).$$

Subsequently, a convolutional layer followed by a max pooling layer is applied to capture inter-channel correlations while reducing the dimensionality as much as possible without sacrificing feature information. Finally, the complete feature map is passed through two fully connected (FC) layers for classification. Each convolutional layer is configured with a stride of 1 and no padding, and is followed by batch normalization (BN) and a ReLU activation function.

H. Liveness Decision

For liveness detection, we use the binary output of the fusion network to detect malicious attacks and the probability output of multi-class to recognize gestures. If the multi-class classifier's prediction matches the action challenge, SONICUMOS marks the user as legitimate.

1) Attack Detection: This module acts as a binary classifier. We add a binary output layer with a sigmoid activation after the fully connected layer of the feature fusion network. Binary cross entropy (BCE) loss and stochastic gradient descent (SGD) optimizers guide the neural network learning. We constructed negative samples from three attacks and trained them alongside positive samples from legitimate users.

2) Gesture Recognition: To ensure user awareness, we design three action challenges. Gesture recognition is performed using fused features, and results are matched with action challenges transmitted by SONICUMOS. A multi-class output layer with softmax is added to the fusion network's last layer, trained using cross-entropy loss and SGD optimizer.

VI. Security Analysis

In this section, we theoretically analyze the security properties of SONICUMOS and evaluate its effectiveness against several common attack vectors.

A. Challenge-response Liveness Detection

The security of SONICUMOS's challenge-response mechanism relies on the consistency between the user's responses to both the *ultrasonic* challenge (chirp selection) and the *action* challenge. To demonstrate the effectiveness of SONICUMOS's liveness detection, the analysis focuses on the challenge space.

Recall that during liveness detection, SONICUMOS randomly generates an ultrasonic chirp and one of three predefined actions as challenges. According to the parameter settings (Section V-B2), An attacker has a probability of approximately 2.9×10^{-6} of generating a signal that is indistinguishable from the original chirp. Therefore, the odds of an attacker successfully guessing both challenges are extremely low. Moreover, the consistency check between the ultrasonic signal and the user's 3D actions makes it highly challenging for attackers to simulate or fake such responses using simple video-based attacks.

B. Resistance to Advanced Attacks

By combining video and ultrasonic signals, SONICUMOS is capable of defending against several common attacks, which can be ranked in increasing order of sophistication. As introduced in our threat model (Section IV), SONICUMOS considers 2D video attacks, replay attacks, and 3D impersonation attacks. Next, we will evaluate how SONICUMOS defends against these attack types.

1) 2D Video Attack: In a 2D video attack, an attacker plays a pre-recorded verification video of the victim on a screen. SONICUMOS defends by requiring signal reflections from the 3D environment during liveness detection, rejecting authentication as 2D videos lack spatial movement data.

2) Replay Attack: In a replay attack, the attacker replays recorded video and ultrasonic audio of the victim. Since SONICUMOS generates a fresh ultrasonic carrier signal for each session, the replayed audio is unlikely to match the correct frequency, preventing this attack from succeeding.

3) 3D Impersonation Attack: In this attack, an adversary mimics the victim's movements while injecting a pre-recorded video. Due to the precise timing of ultrasonic signals, it is highly unlikely the attacker's movements will align with the video. SONICUMOS detects inconsistencies in the reflected signals, thwarting this attack.

VII. PROTOTYPE IMPLEMENTATION

We implement the SONICUMOS prototype using commercial off-the-shelf smartphones to demonstrate its practical feasibility. The implementation covers three key aspects: (1) multi-brand device selection to ensure generalization capability, (2) acoustic hardware configuration considering OS-level constraints, and (3) lightweight signal processing pipelines. Our prototype requires standard smartphone permissions including camera access, microphone recording, local storage, and network communication. This section details the device models, hardware configuration strategies, software implementation choices, and default experimental settings for subsequent performance evaluation.

A. Smartphone Models

To validate cross-device compatibility, we deploy SONICUMOS on seven commercial models spanning three price tiers:

- Entry-level: Redmi Note 10, Redmi K50
- Mid-range: Vivo S15, Oppo Reno7
- Premium: iPhone 15, Samsung S23, Huawei P30 Pro

This selection covers 6 top brands from 2024's global market leaders, representing both Android and iOS ecosystems. Notably, while our implementation primarily utilizes standard smartphone acoustic features, inherent manufacturer-specific optimizations in audio processing pipelines (e.g., noise suppression algorithms and microphone gain calibration) would also create natural performance variations across devices.

B. Acoustic Hardware Configuration

Modern smartphones are equipped with multiple spatially distributed microphones and speakers (e.g., iPhones contain four microphones and two speakers), with varying orientations and power levels. These hardware variations can impact liveness authentication performance.

To configure the microphone and speaker usage scenarios, applications leverage system APIs that are mapped by the operating system to one or more physical devices for audio processing.

Notably, Android and iOS implement distinct API paradigms: Android [74] decouples microphone and speaker selection through separate configuration parameters, whereas iOS [75] employs a unified control framework using three hierarchical elements: *Category*, *Mode*, and *Options* at different levels.

Different acoustic options can lead to different performances of SONICUMOS, which we systematically evaluate in Section VIII-L. Throughout other experiments, unless otherwise specified, Android implementations were configured with `MIC` as the audio source and `STREAM_VOICE_CALL` for speaker output, while iOS implementations utilized the `PlayAndRecord` category paired with `VoiceChat` mode.

C. Software Development

To process audio, we deploy Scipy, a free and open-source Python library that supports FFT, interpolate, and FIR/IIR/Biquad filter on the server. We also deploy MediaPipe, a cross-platform and open-source pipeline framework to build custom machine learning solutions for streaming media to handle video streaming for facial landmarks detection.

D. Evaluation Settings

Default setting. Throughout our experiments, unless otherwise specified, SONICUMOS is evaluated on Oppo Reno7, with the microphone configured in `MIC` mode and speaker in `STREAM_VOICE_CALL` mode. The audio recording is mono channel and the sampling rate is set at 48 kHz. Finally, audio is saved as a WAV file based on PCM encoding. Moreover, the video is stored in MP4 format based on H.264 encoding with the bit rate of $1e6$ and the frame rate of 30. The smartphone is vertically placed on a fixed bracket towards the face at a distance of 20cm which is a comfortable gesture for the user to perform face authentication. Redmi Note10 is chosen as the default smartphone and all experiments are conducted in a quiet conference room. The default frequency band of

the challenge is 18 $\sim$ 20 kHz, and the challenge action type includes all actions, i.e., lip, nod and turn. All classifiers use fusion features network to reconcile two-modal patterns unless claimed otherwise.

Data collection. We conducted two rounds of data collection over a year, each lasting two weeks. In the first round, 20 volunteers (15 male, 5 female) used four smartphones (Redmi Note10, Vivo S15, Redmi K50, Oppo Reno7) to collect positive and negative samples. Volunteers held their phones in a selfie pose and triggered detection by pressing a button, activating the front camera, top speaker, and microphone. They performed instructed head gestures without speaking specific words. Multiple actions per recording were segmented into independent samples. A total of 7,680 samples were collected (4,560 positive, 3,080 negative), covering variations in gestures, distances, ultrasonic frequencies, interference, and mobility. The second round added 13 volunteers (10 male, 3 female) and three devices (iPhone 15, Samsung S23, Huawei P30 Pro) for cross-device compatibility, height, angle, and hardware tests. This yielded 10,179 samples (3,393 positive, 6,786 negative). The final dataset totaled 17,959 samples with original parameters unchanged.

Model Training. We trained on 3,000 samples with a batch size of 64 for 60 epochs, using BCE Loss and an SGD optimizer (momentum = 0.5). The learning rate (0.005) was tuned on the validation set. A single 3090 Ti GPU determined batch size, and early stopping refined the number of epochs.

Metrics. The following metrics are used to evaluate our system: True Positive Rate (TPR) is the likelihood that SONICUMOS correctly detects legitimate users. False Positive Rate (FPR) is the probability that SONICUMOS mistakes an attack as a legitimate user. Receiver Operating Characteristic (ROC) curve describes the relationship between them. Area Under ROC Curve (AUC) is the area under ROC curve used to evaluate a binary classifier's performance, particularly in imbalanced classes. Equal Error Rate (EER) represents the point at which FPR and FNR are equal. Accuracy measures the likelihood that the system accepts legitimate users and rejects attacks and is equal to correctly classified examples divided by the total number of examples.

Attacks. We evaluate SONICUMOS against three attack types: 2D video, replay, and 3D impersonation attacks. For 2D video attacks, a 2K resolution screen displays challenge-specific videos in front of the target smartphone. For replay attacks, we collect audio-video samples from previously accepted users (18–20 kHz range) and inject them into the system. Since the system dynamically selects audio challenge frequencies, attackers cannot predict them. For 3D impersonation attacks, 33 participants mimic a victim's lip, nod, and turn actions by watching videos. They replicate speed, retention time, and other motion details. The victim's video and mimicked actions are then streamed to the server.

VIII. EVALUATION

A. Overview of Evaluations

The goal of SONICUMOS is to develop a behavior-based liveness detection system that ensures both security and user experience. We designed a multidimensional evaluation to assess its **effectiveness**, **generality**, and **robustness**:

- **Effectiveness against attacks.** To counter 2D/3D spoofing, we conducted three experiments (Sections *VIII-B*, *VIII-C*, *VIII-D*, and *VIII-E*) to validate the effectiveness in resisting advanced attacks. Results show that dual-modality fusion enhances detection accuracy and resilience.
- **Robustness in real-world scenarios.** Users interact with varying distances, heights, postures, and facial orientations, while environmental factors like noise and hardware diversity further impact performance. We evaluated these variables through interaction distance (Section VIII-G), height (Section VIII-H), facial orientation (SectionVIII-I), interaction posture (Section VIII-J, and environment interference (Section VIII-F), ensuring stability in real-world conditions.
- **Generality for practical deployment.** Through hardware (Section VIII-L) and smartphone (Section VIII-K) tests, we verified compatibility with off-the-shelf devices, confirming SONICUMOS's adaptability to diverse environments and feasibility for large-scale deployment.

This evaluation framework ensures SONICUMOS's reliable performance under adversarial and real-world constraints while maintaining user experience.

B. Overall Performance

We first present the overall performance of SONICUMOS against three attacks: 2D video attacks, replay attacks, and 3D impersonation attacks. Figure 9 shows the ROC curves of the liveness detection system using fusion features. We observe that our system archives high performance in detecting live users and rejecting different attacks. SONICUMOS attains an accuracy of 99.39% and an EER of 0.38% under 2D video attacks, an accuracy of 99.78% and an EER of 0.45% under replay attacks. The performance against 3D impersonation attacks is slightly worse than the other two attacks and results in 95.83% accuracy with 4.96% EER. Only a user accepted by both the attack detection classifier and the gesture recognition classifier will be marked as legitimate. The results show that gesture recognition attains 100% accuracy, which means SONICUMOS can distinguish three gestures perfectly with fusion features.

C. Impact of Feature Fusion

We next take a closer look at the impact of the feature fusion method under three attacks. SONICUMOS capture multimodal biometrics including the user's facial video and the audio containing a micro-Doppler signature. We evaluate the effectiveness of video-only, audio-only, and feature fusion (audiovisual) methods, respectively. Figure 10 illustrates the accuracy and EER of different methods with three attacks. We observe that the video-only method against 2D video attacks produces an accuracy of 56.42% with an EER of 36.36%, which indicates the vulnerability of the traditional visual-based active liveness detection system to 2D video attacks. Due to replay attacks

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

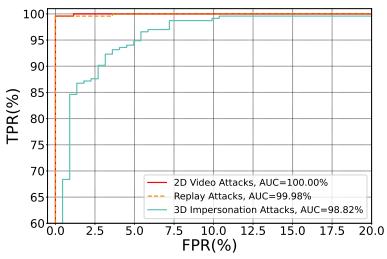


Fig. 9. Overall ROC curves with AUCs under different attacks.

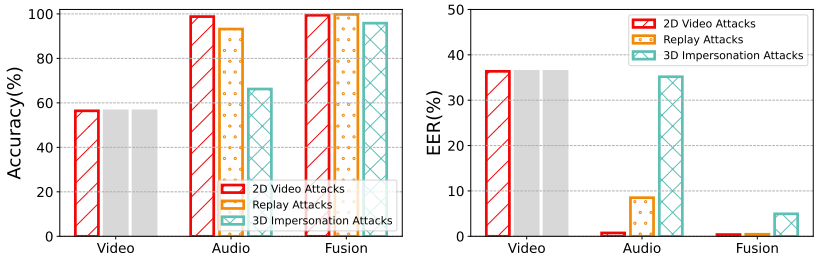


Fig. 10. The effect of feature fusion, showing improved accuracy and reduced EER across all attack types, particularly for 3D impersonation attacks.

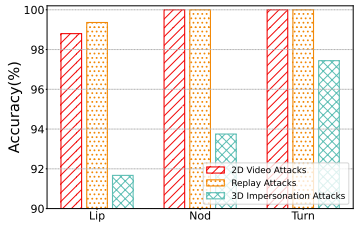


Fig. 11. Accuracy under different challenge actions.

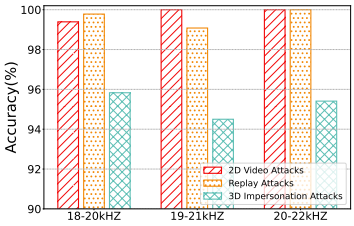


Fig. 12. Accuracy under different frequency bands of ultrasonic challenge.

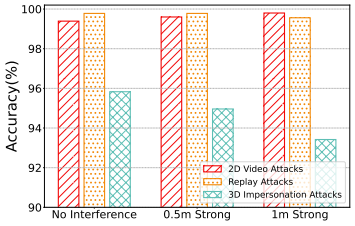


Fig. 13. Accuracy under different interferences.

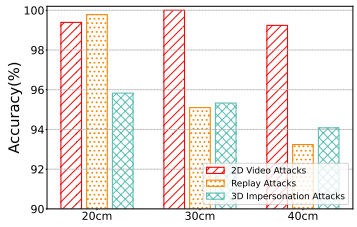


Fig. 14. Accuracy under different interaction distances.

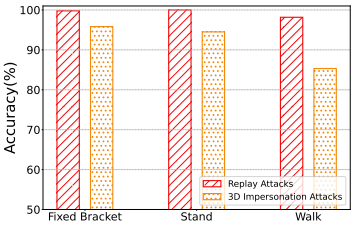


Fig. 15. Accuracy under different interaction postures.

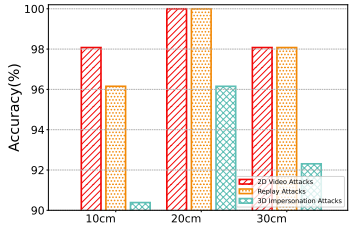


Fig. 16. Accuracy under different heights.

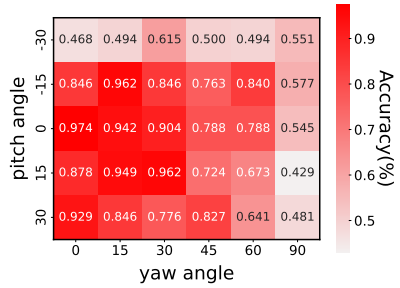


Fig. 17. Accuracy under different angles for action Turn.

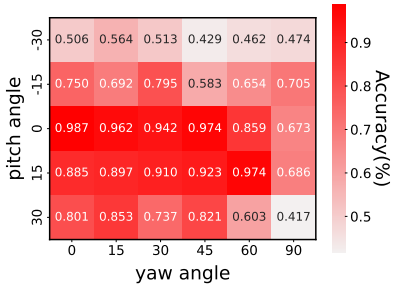


Fig. 18. Accuracy under different angles for action Nod.

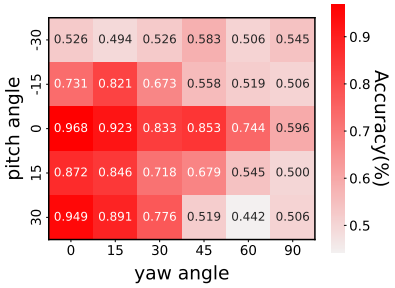


Fig. 19. Accuracy under different angles for action Lip.

and 3D impersonation attacks being conducted by injecting the video containing the victim's face, we assume the system has no ability to distinguish attacks by the 2D video-only method and didn't design correlated evaluation. In the figure, values of accuracy and EER are represented as gray bars. The audio-only method provides much better performance than the video-only method. The accuracy/EER under 2D video attacks and replay attacks, which are 98.79%/0.74% and 93.19%/8.51%, respectively. Although the audio-only method can effectively detect above two attacks, it has poor performance against 3D impersonation attacks with 66.23% accuracy and 35.17% EER. It indicates that audio features after segmentation relying on motion intervals at time scales are not sufficiently effective to distinguish 3D impersonation attacks. We further look at how the feature fusion mechanism performs under three attacks. We observe that the feature fusion method improves performance under all three attacks, especially under 3D impersonation attacks. The accuracy is 99.39%, 99.78%, and 95.83% respectively under 2D video attacks, replay attacks, and 3D impersonation attacks. Associated EER is 0.38%, 0.45%, and 4.95%. The above results show that the feature fusion method is effective since it takes advantage of the consistency between

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

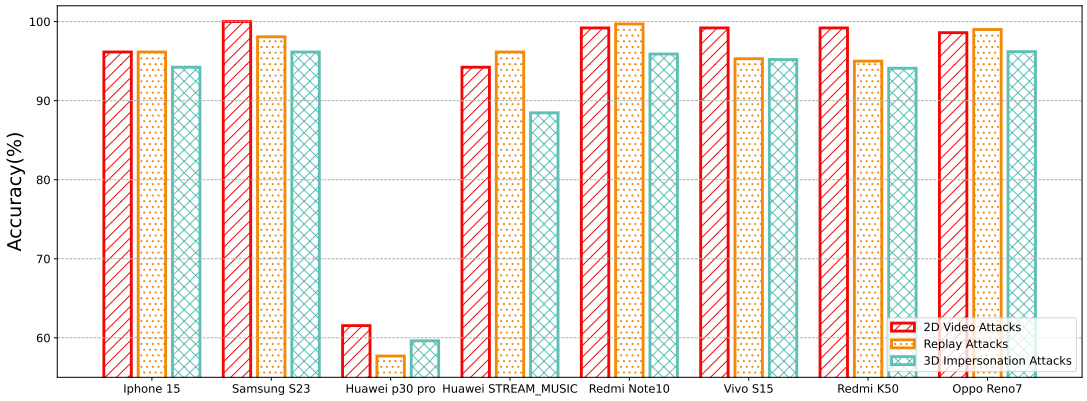


Fig. 20. Accuracy under different smartphone models.

the audio and the video at higher dimensions instead of the consistency in time.

D. Impact of Action Type of Challenge

To ensure that the authentication or detection is initiated with the user's awareness, we retain the random action challenge mechanism. We choose three head gestures (lip movement, nod, head turn) that can be sensed by ultrasonic to evaluate the performance. Figure 12 illustrates the accuracy of different actions with three attacks. We observe that the head turn achieves the best accuracy on three attacks which is 100% under 2D video attacks and replay attacks, and 97.44% under 3D impersonation attacks. The accuracy of the nod under 2D video attacks and replay attacks also produces an accuracy of 100% and the accuracy under 3D impersonation attacks is slightly lower than the turn which results in 93.75%. The lip-only method provides the worst accuracy, which is 98.8%, 99.36%, and 91.67% under 2D video attacks, replay attacks, and 3D impersonation attacks. It is because lip movement contains more tiny flexible relative movements than the other two actions, which are not easily sensed by ultrasonic while the action is not intense enough. Overall, the result shows that our system is capable of detecting three attacks using different action types of challenge.

E. Impact of Sensing Ultrasonic Frequency of Challenge

SONICUMOS improves security by randomly transmitting ultrasonic challenges to expand the selectable parameter space. We study the impact of sensing ultrasonic with different frequencies on system accuracy. In particular, we choose three frequency bands with the same bandwidth B and different start frequencies $f1$ as the sensing ultrasonic, i.e., 18∼20 kHz, 19∼21 kHz, 20∼22 kHz. Note that this is not all the selectable parameters. We just selected several representative samples that can cover the inaudible frequency bands between 18 kHz and 22 kHz that acoustic hardware on most smartphones supports. Figure 12 shows the accuracy of SONICUMOS using three frequency bands under three different attacks. We observe that our system works effectively under all of these sensing frequency bands. For 2D video attacks and replay attacks, the accuracy is over 99%. Minor fluctuations occur in the accuracy under 3D impersonation attacks but all of these are over 94%. These results show that our system can work with different sensing ultrasonic frequencies as a challenge to defend against three attacks and is insensitive to the frequency response of hardware.

F. Impact of Environment Interference

In the real scenario, there will be dynamic surroundings like passers-by around a user who is doing face authentication, which may affect the effectiveness of detecting attacks. To verify the robustness of SONICUMOS in the complicated and dynamic environment interferences, we experiment with three simulated environments: a quiet conference, an indoor market, and a crowded street. Specifically, a quiet conference doesn't contain any dynamic interferences. For simulating the indoor market and crowded street environments, we invited 2 volunteers to walk quickly accompanied by arm waving at around 1m and 0.5m in front of the smartphone. Figure 14 depict the accuracy of three different degrees of interference under three attacks. With the introduction of interference at 1m away from the smartphone, the accuracy of detecting 3D impersonation attacks decreases from 95.83% to 94.96%, and the accuracy is further reduced to 93.42% with a closer interference at 0.5m. Strong dynamic interference has little impact on the detection accuracy of the other two attacks, but all of them still remain over 99%. The above results confirm our system is robust to the interference caused by dynamic surroundings with the observation that even strong interference in a short distance cannot significantly affect the ability to identify spoofing attacks.

G. Impact of Interaction Distance

Generally, acoustic signals decay as the propagation distance increases. Thus the increase of interaction distance may take influence on the quality of micro-Doppler features captured by the ultrasonic carrier and further leads to performance degradation. We evaluated the effect of the interaction distance between the smartphone and the user's face under three attacks. We consider three distance gradients (20cm,

30cm, 40cm), where the user's face fronts straight to the phone's screen. Figure 14 reveals the accuracy of three attacks when the interaction distance varies from 20cm to 40cm. We observe the accuracy degradation while the interaction distance increases. In particular, an interaction distance of 30cm under replace attack results in an accuracy of 95.1%, and the value is 93.24% for 40cm. The accuracy degradation under replay attacks is more oblivious than the other two attacks, where they have a decline of less than 2%. However, even in the farthest interaction distance, the accuracy of our system against three attacks exceeds 93%. The longer interaction distance is difficult to apply in practical scenarios due to the limitation of arm-span. Results show that our system works well for different interaction distances even at a long distance.

H. Impact of Height

The height from the desktop affects the intensity of ultrasonic multipath effects, which may influence the experimental results. To verify the robustness of the solution, we conducted height experiments. During the experiments, volunteers were required to face the phone directly, and only mouth-opening actions were collected. A total of 78 test data points were collected, with an equal number of attack data points for each type of attack. The results are shown in Figure 16.

The experimental results show that at a height of 10cm, the system achieved accuracies of 98.08% and 96.15% under 2D video attacks and replay attacks, respectively, while the accuracy under 3D impersonation attacks was 90.38%. When the height was increased to 20cm, the system achieved 100% accuracy under both 2D video attacks and replay attacks, and the accuracy under 3D impersonation attacks also improved to 96.15%. At a height of 30cm, the system's accuracy under 2D video attacks and replay attacks was 98.08% and 98.08%, respectively, while the accuracy under 3D impersonation attacks was 92.31%.

I. Impact of Facial Orientation

The experiment tested 6 yaw (horizontal) and 5 pitch (vertical) angles. Volunteers first faced the device, then followed prompts to gaze at marked points and perform actions at fixed angles. Each combination yielded 52 test samples, totaling 2,340. To facilitate result presentation, attack results (2D video, replay, and 3D impersonation) were merged, as shown in Figures 17, 18, and 19. In these figures, positive pitch angles indicate upward tilts, negative indicate downward, and yaw represents head rotation. Due to head symmetry, left and right turns were not distinguished.

From the angle analysis results, tilting downward (negative yaw angle) significantly reduces accuracy due to obscured facial landmarks, while upward tilts have minimal impact. The influence of pitch angle is lower than yaw angle, and within ±45°, accuracy remains above 80%. Among actions, turn is less affected by pitch, nod by yaw, while lip movements perform worst across all angles. This is due to inherent action characteristics—turn and nod involve horizontal and vertical motion, partially offsetting facial orientation changes.

J. Impact of Interaction Postures

Users may adopt different postures for face authentication. We evaluated the accuracy of three common postures under two attacks. 2D video attacks are generally organized by placing a fixed screen, independent of the user's posture. We do not consider this attack here. Three detection postures include the selfie posture using the mobile phone placed on the fixed bracket, the standing posture, and holding the smartphone while walking slowly. Figure 15 shows the accuracy of different interaction postures under two attacks. We observed that sitting posture still has high detection accuracy under replay attacks and 3D impersonation attacks, which are 100% and 94.5%. However, the accuracy decreases to 85.32% rapidly under 3D impersonation attacks. The reason is that our system relies on the motion field extracted from the video to segment audio features. When the user starts to walk, the sharp shake of the camera will seriously affect the fine-grained detection of the motion intervals. Therefore, we recommend users use this system while keeping static. The above results show that our system is capable of detecting live user cases and rejecting attacks while using stable interaction postures.

K. Impact of Smartphone Model

Microphones in different smartphones exhibit distinct frequency response characteristics [76]–[78]. Figure 20 presents the accuracy rates across three attack scenarios and different smartphone models. In this evaluation, the dataset comprises 796 valid samples collected under controlled conditions where volunteers maintained 20cm distance from devices while performing lip-synchronized actions. All Android devices except Huawei P30 Pro demonstrated strong defense against 2D video attacks with accuracy exceeding 98%. Vivo S15 and Redmi K50 achieved approximately 95% accuracy in detecting replay attacks. Oppo Reno7 exhibited superior performance against sophisticated 3D imitation attacks, reaching 96.26% accuracy. The Huawei P30 Pro showed reduced effectiveness due to its screen sound technology replacing conventional earpiece speakers, which demonstrates inferior ultrasonic performance. When reconfigured to use the `STREAM_MUSIC` mode with bottom-firing speakers, Huawei's recognition rate improved moderately but remained below conventional earpiece performance, consistent with findings discussed in Section VIII-L. The sole iOS device (iPhone 15) showed slightly lower accuracy across all attack types, attributable to iOS's automatic multiplexing of 4 microphones and 2 speakers – a hardware selection constraint analyzed in Section VIII-L.

L. Impact of Hardware

In our experiment, we evaluated five microphone modes and three speaker modes on Android, collecting 52 samples per configuration. Results (Figure 21) show that the "CAMCORDER" microphone mode with the "STREAM_VOICE_CALL" speaker mode performed best, utilizing the front earpiece speaker and front microphone to efficiently receive reflected ultrasonic waves. In contrast, noise-suppressing modes (e.g., VOICE_RECOGNITION,

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3565689

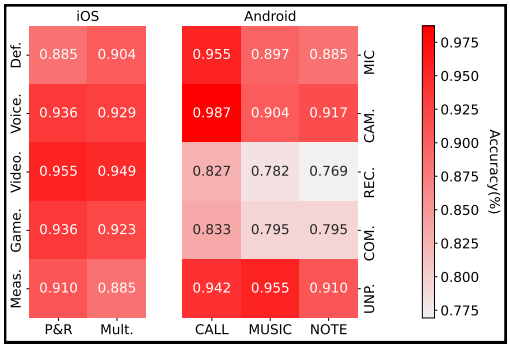


Fig. 21. Accuracy under different hardware selections.

VOICE_COMMUNICATION) performed worse due to system-level noise filtering.

For iOS, we tested two Category options and five Mode configurations, also collecting 52 samples each. The results (Figure 21) showed that "VideoChat" mode yielded the best performance, with noise suppression not significantly affecting the outcomes. These findings highlight the robustness of the proposed approach across different hardware, ensuring reliable performance with well-performing hardware selections.

IX. Discussions

Wider device compatibility. We conducted tests on 7 mobile phone models across 6 brands, which together account for over 70% of the market share, thereby demonstrating the generalizability of Sonicumos on mobile devices. However, for non-mobile devices such as tablets and PCs, the diverse configurations of microphones and speakers prevent us from guaranteeing similarly high accuracy. Moreover, as Sonicumos employs a client-server architecture, development constraints have so far precluded deploying the model directly on mobile devices. We anticipate that future commercial applications will allow for more extensive testing and resolve local deployment issues.

Defense against more powerful attacks. Our method leverages multimodal consistency for liveness detection, effectively countering spoofing. Feature fusion and deep learning help capture variations in action amplitude and texture, suggesting potential defense against 3D mask attacks. However, due to experimental constraints, relevant data collection and model training remain future work.

More efficient utilization of multimodal information. Sonicumos relies on handcrafted feature extraction, leading to information loss. It follows a traditional liveness detection-authentication model, but Transformers offer a promising end-to-end alternative for multimodal fusion, potentially enhancing both identification and liveness detection.

X. Conclusions

This paper has presented Sonicumos, an enhanced behavior-based face liveness detection system that leverages ultrasonic and video signals for sensing 3D head gestures. Our approach addresses the limitations of traditional behavior-based liveness detection methods. By employing frequency-modulated continuous-wave (FMCW) ultrasonic radar, Sonicumos offers a robust 3D gesture recognition solution that is compatible with face authentication and does not introduce extra user burden. We have also proposed a new dual-feature fusion network that integrates audio and video features at the feature level, increasing the system's detection accuracy and resilience against numerous attacks. Our prototype demonstrates promising robustness to various impacts including enviormental interference, distances, and interaction postures.

References

[1] "Facial recognition market - growth, trends, covid-19 impact, and forecasts (2022 - 2027)," 2022.
[2] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *Proc. of IEEE International Joint Conference on Biometrics*, pp. 1–7, 2011.
[3] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang, "Face flashing: a secure liveness detection protocol based on light reflections," in *Proc. of NDSS*, 2018.
[4] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," pp. 1–8, 2007.
[5] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," vol. 9967, pp. 611–619, 2016.
[6] K. Kollreider, H. Fronthaler, and J. Bigün, "Non-intrusive liveness detection by face images," *Image Vis. Comput.*, vol. 27, no. 3, pp. 233–244, 2009.
[7] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proc. of ECCV*, pp. 504–517, 2010.
[8] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. of IJCB*, pp. 1–7, 2011.
[9] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park, and J. Kim, "Face liveness detection based on texture and frequency analyses," in *Proc. of ICB*, pp. 67–72, 2012.
[10] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee, "Face liveness detection using variable focusing," in *Proc. of ICB*, pp. 1–6, 2013.
[11] S. Chen, A. Pande, and P. Mohapatra, "Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones," in *Proc. of ACM MobiSys*, pp. 109–122, 2014.
[12] H. Farrukh, R. M. Aburas, S. Cao, and H. Wang, "Facerevelio: A face liveness detection system for smartphones with a single front camera," in *Proc. of ACM MobiCom*, pp. 1–13, 2020.
[13] Z. Wu, Y. Cheng, J. Yang, X. Ji, and W. Xu, "Depthfake: Spoofing 3d face authentication with a 2d photo," in *Proc. of IEEE S&P*, pp. 1710–1726, 2023.
[14] Y. Chen and H.-T. Ma, "Biometric authentication under threat: Liveness detection hacking," in *Proc. of BlackHat*, 2019.
[15] R. Frischholz and A. Werner, "Avoiding replay-attacks in a face recognition systenm using head-pose estimation," in *Proc. of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 234–235, 2003.
[16] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigün, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Trans. Inf. Forensics Secur.*, vol. 2, no. 3-2, pp. 548–558, 2007.
[17] M. D. Marsico, M. Nappi, D. Riccio, and J. Dugelay, "Moving face spoofing detection via 3d projective invariants," in *Proc. of IAPR*, pp. 73–78, 2012.
[18] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3d structure recovered from a single camera," in *Proc. of ICB*, pp. 1–6, 2013.
[19] K. Kollreider, H. Fronthaler, and J. Bigün, "Evaluating liveness by face images and the structure tensor," in *Proc. of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pp. 75–80, 2005.
[20] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Z. Li, "Face liveness detection by exploring multiple scenic clues," in *Proc. of ICARCV*, pp. 188–193, 2012.
[21] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng, "Seeing your face is not enough: An inertial sensor-based liveness detection for face authentication," in *Proc. of ACM SIGSAC*, pp. 1558–1569, 2015.

[22] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. of ACM SIGSAC*, pp. 1080–1091, 2016.

[23] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. of ACM CCS*, pp. 57–71, 2017.

[24] L. Zhang, S. Tan, Y. Chen, and J. Yang, "A continuous articulatory-gesture-based liveness detection for voice authentication on smart devices," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23320–23331, 2022.

[25] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pp. 1466–1474, 2018.

[26] M. Zhou, Q. Wang, Q. Li, W. Zhou, J. Yang, and C. Shen, "Securing face liveness detection on mobile devices using unforgeable lip motion patterns," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9772–9788, 2024.

[27] L. Wu, J. Yang, M. Zhou, Y. Chen, and Q. Wang, "Lvid: A multimodal biometrics authentication system on smartphones," *IEEE Transactions on Information Forensics & Security*, vol. 15, pp. 1572–1585, 2020.

[28] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *Proc. of ACM MobiCom*, pp. 160–173, 2021.

[29] D. Zhang, J. Meng, J. Zhang, X. Deng, S. Ding, M. Zhou, Q. Wang, Q. Li, and Y. Chen, "Sonarguard: Ultrasonic face liveness detection on mobile devices," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4401–4414, 2023.

[30] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range fmcw monopulse radar for hand-gesture sensing," in *IEEE Radar Conference*, pp. 1491–1496, 2015.

[31] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 746–761, 2015.

[32] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 10, pp. 2268–2283, 2016.

[33] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, 2017.

[34] T. de Freitas Pereira, A. Anjos, J. M. D. Martino, and S. Marcel, "LBP - TOP based countermeasure against face spoofing attacks," in *Proc. of ACCV*, vol. 7728, pp. 121–132, 2012.

[35] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *CoRR*, vol. abs/1408.5601, 2014.

[36] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," in *Proc. of ACPR*, pp. 141–145, 2015.

[37] 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 *Proc. of IEEE CVPR*, pp. 3507–3516, 2019.

[38] R. Raghavendra, K. B. Raja, S. Venkatesh, F. A. Cheikh, and C. Busch, "On the vulnerability of extended multispectral face recognition systems towards presentation attacks," in *Proc. of IEEE ISBA*, pp. 1–8, 2017.

[39] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: Anti-spoofing facial authentication using COTS RFID," in *Proc. of IEEE INFOCOM*, pp. 1–10, 2021.

[40] "About face id advanced technology," 2017.

[41] W. Xu, W. Song, J. Liu, Y. Liu, X. Cui, Y. Zheng, J. Han, X. Wang, and K. Ren, "Mask does not matter: anti-spoofing face authentication using mmwave without on-site registration," in *Proc. of ACM MobiCom*, pp. 310–323, 2022.

[42] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proc. of ACM MobiCom*, pp. 321–336, 2018.

[43] C. Kong, K. Zheng, Y. Liu, S. Wang, A. Rocha, and H. Li, "M3FAS: an accurate and robust multimodal mobile face anti-spoofing system," *CoRR*, vol. abs/2301.12831, 2023.

[44] M. Mohzary, K. J. Almalki, B. Choi, and S. Song, "Apple in my eyes (AIME): liveness detection for mobile security using corneal specular reflections," in *Proc. of ACM MobiSys* (S. Banerjee, L. Mottola, and X. Zhou, eds.), pp. 489–490, 2021.

[45] C. R. Gerstner and H. Farid, "Detecting real-time deep-fake videos using active illumination," in *Proc. of IEEE CVPR*, pp. 53–60, 2022.

[46] H.-K. Jee, S.-U. Jung, and J.-H. Yoo, "Liveness detection for embedded face recognition system," *International Journal of Biological and Medical Sciences*, vol. 1, no. 4, pp. 235–238, 2006.

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

[48] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[49] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Proc. of International Conference on Image Analysis and Signal Processing*, pp. 233–236, 2009.

[50] I. Muslukhov, "How to bypass android liveness check." https://www.youtube.com/watch?v=zYxphDK6s3I, 2012.

[51] Y. Li, Z. Wang, Y. Li, R. H. Deng, B. Chen, W. Meng, and H. Li, "A closer look tells more: A facial distortion based liveness detection for face authentication," in *Proc. of ACM Asia CCS*, pp. 241–246, 2019.

[52] A. Ali, F. Deravi, and S. Hoque, "Liveness detection using gaze collinearity," in *Third International Conference on Emerging Security Technologies*, pp. 62–65, 2012.

[53] Z. Zheng, Q. Wang, C. Wang, M. Zhou, Y. Zhao, Q. Li, and C. Shen, "Where are the dots: Hardening face authentication on smartphones with unforgeable eye movement patterns," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1295–1308, 2023.

[54] E. Uzun, S. P. H. Chung, I. Essa, and W. Lee, "rtCaptcha: A real-time CAPTCHA based liveness detection system," in *Proc. of NDSS*, 2018.

[55] P. Jiang, Q. Wang, X. Lin, M. Zhou, W. Ding, C. Wang, C. Shen, and Q. Li, "Securing liveness detection for voice authentication via pop noises," *IEEE Trans. Dependable Secur. Comput.*, vol. 20, no. 2, pp. 1702–1718, 2023.

[56] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. of IEEE INFOCOM*, pp. 2062–2070, 2019.

[57] J. Tan, C.-T. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *Proc. of IEEE INFOCOM*, pp. 1–9, 2017.

[58] J. Tan, X. Wang, C. Nguyen, and Y. Shi, "SilentKey: A new authentication framework through ultrasonic-based lip reading," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 36:1–36:18, 2018.

[59] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 447–460, 2019.

[60] A. B. Wong, Z. Huang, and K. Wu, "Leveraging speech and ultrasonic signals toward articulation-based smartphone user authentication," in *Proc. of ACM MobiSys*, pp. 547–548, 2022.

[61] T. P. Gill, *The Doppler effect : an introduction to the theory of the effect*. 1965.

[62] M. S. Zediker, R. R. Rice, and J. H. Hollister, "Method for extending range and sensitivity of a fiber optic micro-doppler ladar system and apparatus therefor," 1998.

[63] 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 & Electronic Systems*, vol. 42, no. 1, pp. 2–21, 2006.

[64] S. Zhang, Q. Wang, M. Gan, Z. Cao, and H. Zeng, "Radsee: See your handwriting through walls using fmcw radar," in *Proc. of NDSS*, 2025.

[65] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proc. of ACM MobiSys*, pp. 42–55, 2017.

[66] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang, G. Guo, Z. Lei, S. Z. Li, and D. Zhang, "Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2497–2507, 2022.

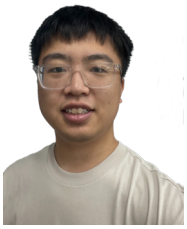
[67] N. Roy, H. Hassanieh, and R. R. Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proc. of ACM MobiSys*, pp. 2–14, 2017.

[68] C. E. Cook, "Linear fm signal formats for beacon and communication systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-10, no. 4, pp. 471–478, 1974.

[69] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *Proc. of IEEE INFOCOM*, pp. 2407–2415, 2015.

[70] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages behind the sound: real-time hidden acoustic signal capture with smartphones," in *Proc. of ACM MobiCom*, pp. 29–41, 2016.

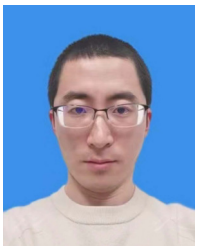
[71] X.-G. Xia, "Discrete chirp-fourier transform and its application to chirp rate estimation," *IEEE Transactions on Signal processing*, vol. 48, no. 11, pp. 3122–3133, 2000.
[72] P. M. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2118–2126, 1990.
[73] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *CoRR*, vol. abs/1907.06724, 2019.
[74] A. Developers, "Audiomanager." https://developer.android.com/reference/android/media/AudioManager, n.d. Accessed: 2025-03-03.
[75] A. Inc., "Avaudiosession." https://developer.apple.com/documentation/avfaudio/avaudiosession, n.d. Accessed: 2025-03-03.
[76] Z. Zhou, W. Diao, X. Liu, and K. Zhang, "Acoustic fingerprinting revisited: Generate stable device ID stealthily with inaudible sound," in *Proc. of ACM CCS*, pp. 429–440, 2014.
[77] A. Das, N. Borisov, and M. Caesar, "Fingerprinting smart devices through embedded acoustic components," in *Proc. of ACM CCS*, pp. 441–452, 2014.
[78] M. Zhou, Q. Wang, T. Lei, Z. Wang, and K. Ren, "Enabling online robust barcode-based visible light communication with realtime feedback," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 12, pp. 8063–8076, 2018.



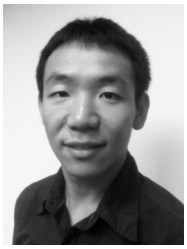
Yihao Wu is currently pursuing his graduate studies at Wuhan University. Prior to this, he obtained his Bachelor's degree in Information Security from Wuhan University. His research interests include AI security and privacy-preserving technologies.

Peipei Jiang currently works as a researcher at the Laboratory for AI-Powered Financial Technologies Ltd., Hong Kong. She received a joint Ph.D. degree from the School of Cyber Science and Engineering, Wuhan University and the Department of Computer Science, City University of Hong Kong. Before that, she received the B.E. degree in Information Security from Wuhan University, China. Her research interests include privacy preservation technologies, network security and AI security.

Jianhao Cheng received the Master's degree from the School of Cyber Science and Engineering, Wuhan University, following the completion of the Bachelor's degree in Aircraft Design and Engineering at Zhejiang University. He currently works as a Software Engineer at Tencent. His research interests include system security and network security.



Lingchen Zhao is currently an Associate Professor with the School of Cyber Science and Engineering, Wuhan University, China. He received his Ph.D. degree in Cyberspace Security in 2021, from Wuhan University, China, and his B.E. degree in Information Security in 2016, from Central South University, China. He was a Postdoctoral Researcher with the City University of Hong Kong, Hong Kong, from 2021 to 2022. His research interests include data security and AI security.



Chao Shen is currently a professor in the School of Electronic and Information Engineering, Xi'an Jiaotong University of China. He is also with the Ministriy of Education Key Lab for Intelligent Networks and Network Security. He was a research scholar in Carnegie Mellon University from 2011 to 2013. His research interests include system and software security, human computer interaction, AI security, and behavioral biometrics.

Cong Wang (Fellow, IEEE) is a Professor and Head of the Department of Computer Science at the College of Computing, City University of Hong Kong. His research encompasses data security and privacy, AI systems and security, and blockchain with decentralized applications. He has made prolific contributions to these fields, witnessed by 30,000+ citations on Google Scholar and multiple best paper awards, including the 2020 IEEE INFOCOM Test of Time Paper Award. He is an IEEE Fellow, an HK RGC Research Fellow, and a Founding Member of the Young Academy of Sciences of Hong Kong. He has served as the Editor-in-Chief for the IEEE Transactions on Dependable and Secure Computing (late 2022-early 2025), a premier security journal within the IEEE Computer Society.

Qian Wang (Fellow, IEEE) is a Professor in the School of Cyber Science and Engineering at Wuhan University, China. He was selected into the National Highlevel 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. In 2021, his PhD student was selected under Huawei's "Top Minds" Recruitment Program. He serves as Associate Editors for IEEE Transactions on Dependable and Secure Computing (TDSC) and IEEE Transactions on Information Forensics and Security (TIFS).