This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

Use Your PIN Securely: Acoustic-based Second-Factor User Authentication with Enhanced Tap Biometrics

Feiyu Han, Zelei Wang, Yuanhao Feng, Jinyang Huang, Meng Li, *Senior Member, IEEE*, Panlong Yang, *Senior Member, IEEE*, Zhangjie Fu*, *Member, IEEE*

***Abstract*—The vulnerability of PIN-based user authentication on mobile phones has been widely criticized. Despite the ongoing research efforts, existing solutions still face significant limitations, including weak anti-interference capabilities, ambiguity/inconsistency of biometrics, and low adaptability over time. In this work, we design an acoustic-based second-factor user authentication mechanism, named TapPass, which derives device-bound and user-specific tap biometrics from hand-reflected ultrasonic signals during PIN input. To overcome the issue of the inherent dynamics of behavioral patterns and the diversity of passcode patterns, we design a joint deep learning biometric representation framework that combines adversarial learning and contrastive learning to extract discriminative and consistent biometric features. Furthermore, based on the insights of unique acoustic patterns of tap behaviors, a lightweight and interpretable frame-wise augmentation strategy is proposed to improve the training efficiency. Lastly, an incremental learning-based long-term adaptation strategy is introduced to improve the stability over time and the scalability among new users. Through extensive experiments with 50 participants, TapPass can yield an averaged authentication accuracy of 93.24%, outperforming the state-of-the-art work by 14.48%. Even in the 8-week longitudinal study, the authentication performance exhibited minimal fluctuation, with a standard deviation of just 1.42%, highlighting its long-term stability.**

***Index Terms*—Second-factor User Authentication, Mobile Authentication, Acoustic Sensing, and Finger Biometrics.**

I. Introduction

As a widely adopted mobile authentication approach, PIN (Personal Identification Number) code remains a cornerstone of device security, offering a straightforward mechanism for verifying user identity. Despite its simplicity, the PIN code is often criticized for its vulnerability to shoulder surfing, brute-force attacks, and smudge tracing [1], [2], leading to weak security. Although several solutions have attempted to enhance PIN entry security by changing the layout features of the keypad, like the position, size, or color of the digits [3], [4], these approaches are still not fundamentally secure. Moreover, they often increase the user's cognitive and operational burden, since users must relearn or adapt to unfamiliar keypad layouts, making authentication less convenient and more stressful.

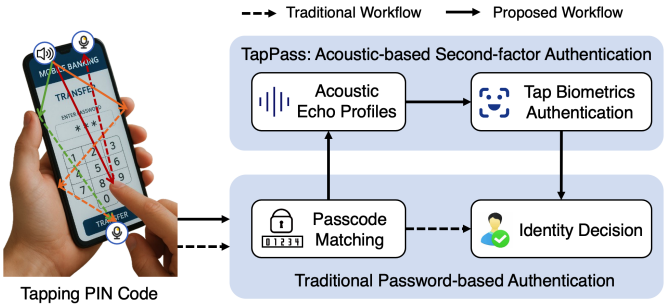


Fig. 1: Overview of TapPass.

Nowadays, second-factor authentication mechanisms are proposed by academic researchers to enhance the security of PIN-based authentication, ensuring that access requires both something the user knows (*i.e.*, the passcode sequences) and something the user has (*i.e.*, biometrics). Prior literature [5]–[7] leverages tapping patterns (like tapping durations, speeds, and touch gestures) to derive behavioral biometrics. Although it is difficult for attackers to mimic the behavior patterns of victims, the inherent instability and variability of behavioral biometrics have hindered the widespread application of this technology [8]. For instance, behavioral patterns may change over time due to changes in mental states or environmental conditions. Therefore, a wide study of physiological biometrics is conducted to extract the stable, consistent, and hard-to-forge characteristics (*e.g.*, hand geometry and fingerprint) underlying the PIN tapping process [9]–[14]. However, in practice, they still face the following limitations. ❶*Additional authentication burden.* Tapping pressure verification is validated as a promising mechanism of enhanced PIN code authentication, which senses the pressure from the user's finger taps [9], [15]. Although this technology does not change the PIN interface and the input mode, it requires users to additionally remember the tapping pressure. ❷*Weak anti-interference capability.* Some solutions [10]–[12], [16] utilize IMU sensors or microphones to collect tap/swipe-induced vibrations or sounds to extract biometric features, extremely susceptible to motion artifacts and ambient noises. In addition, they have specific requirements for the user's tapping gestures and pressure to ensure a high-SNR signal acquisition, limiting the usability in real-world scenarios. ❸*High enrollment efforts.* PIN-based authentication does not require high-effort data enrollment, since it only needs to match the input code pattern with the pre-defined pattern. The second-factor authentication mechanisms designed to enhance

*Zhangjie Fu is the corresponding author.

Feiyu Han, Zelei Wang, and Panlong Yang are with the School of Computer Science and the School of Software, Nanjing University of Information Science and Technology, 210044, Nanjing, China. (E-mail: fyhan@nuist.edu.cn, wangzelei1223@gmail.com, plyang@nuist.edu.cn)

Zhangjie Fu is with the School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, 210044, Nanjing, China. (E-mail: wwwfzj@126.com)

Yuanhao Feng is with University of Electro-Communications, 182-8585, Chofu, Japan. (E-mail: fengyuanhao@uec.ac.jp)

Jinyang Huang and Meng Li are with the School of Computer Science and Information Engineering, Hefei University of Technology, 230601, Hefei, China (E-mail: hjy@hfut.edu.cn, mengli@hfut.edu.cn).

PIN security should avoid imposing extensive enrollment on users, posing a challenge for existing solutions. For example, the enrollment phase in [10] needs 32 samples to generate a user template, which consumes about 1–2 minutes. ❹*Ambiguity of biometrics.* Existing solutions leverage both behavioral biometrics and physiological biometrics underlying the tap behaviors to enable user authentication [13], [14]. However, a significant challenge arises due to the inconsistency of these biometric features induced by dynamic behavior patterns. Additionally, users tend to set different passcodes for different mobile applications, further reducing the consistency of the extracted biometric features. ❺*Long-term instability.* While prior research has achieved considerable progress in improving the accuracy and robustness of biometric authentication, little attention has been paid to ensuring its temporal stability. Most existing methods implicitly assume that biometric features remain consistent over time. Representative studies on PIN-based second-factor authentication are summarized in Tab. I.

To overcome the aforementioned limitations, we introduce an acoustic-based second-factor authentication mechanism, named TAPPASS, which exploits the ubiquitous built-in audio transceivers of mobile phones to sense acoustic channel variations induced by finger movements. As shown in Figure 1, TAPPASS leverages the top speaker to emit continuous ultrasonic waves, and then leverages a combination of top and bottom microphones to comprehensively capture hand-reflected echo profiles, without relying on a dedicated sensor. If a user enters incorrect PIN codes, TAPPASS will directly deny his access. Otherwise, TAPPASS further extract biometric features from hand-reflected ultrasounds for the secondary authentication. Specifically, the hand-reflected echo profiles are fundamentally coupled with not only the user's hand shape but also device-specific factors. From this perspective, the derived biometrics from hand-reflected ultrasounds can be viewed as a joint representation of "the user" and "the user's device". This configuration makes spoofing attacks more difficult because an attacker would need to simultaneously reproduce both the user's interaction characteristics and the target device's acoustic channel properties. Moreover, TAPPASS preserves the conventional PIN input procedure and imposes no additional operations on users, ensuring seamless integration and high usability. While promising, developing such a second-factor authentication solution with ultrasonic signals presents several technical challenges:

- ***Extracting discriminative and consistent biometrics against dynamic behavioral patterns and diverse passcode patterns.*** PIN input often involves a series of dynamic actions such as tapping, pressing, and swiping, which can vary greatly depending on the user's habits, physical state, environment, and even mood. Furthermore, the passcode patterns (combinations of digits) also introduce inconsistency and ambiguity by altering the spatial distribution of the digits. **To address it**, we design a deep learning-based biometric representation framework that incorporates adversarial learning and supervised contrastive learning to decouple behavioral and passcode variations from inherent hand geometry features. Specifically, the contrastive objective enforces identity-level clustering across diverse behavioral conditions. Simultaneously, the adversarial discriminator suppresses passcode-dependent information, encouraging the feature extractor to learn passcode-invariant representations.
- ***Minimizing the amount of enrollment data while maintaining reliable authentication performance.*** Traditional biometric systems often require extensive enrollment periods to collect sufficient data for building accurate and personalized user templates. However, such lengthy enrollment processes can be cumbersome for users, leading to poor adoption rates and user dissatisfaction. **To address it**, based on the physical insights into unique acoustic patterns of tap behaviors, a frame-wise data augmentation is proposed to generate more diverse hand-reflected acoustic responses by simulating the speed and gesture variations. Specifically, the augmentation probability is first adaptively computed at the frame level according to the proportion of effective DCIR components. The augmentation operations are then performed based on this probability, enabling content-aware simulation instead of uniform manipulation as in previous solutions.
- ***Maintaining long-term stability and scalability against temporal variations and new users joining.*** The distribution of biometrics behind tap behaviors changes over time, which may decrease the authentication accuracy [17]–[19]. Additionally, as new users are added to the authentication, the traditional authentication approaches often need to retrain the authentication model from scratch, decreasing the scalability among new users. **To address it**, we introduce a long-term adaptation mechanism coupled with a dual-distance-constraint update strategy against biometric distribution shift. Specifically, the long-term adaptation mechanism determines whether a newly incoming sample should trigger the template update process based on both intra-class distance (quantifying compactness) and inter-class distance (quantifying separability) constraints. This adaptive strategy enables TAPPASS to adapt to biometric distribution shifts while maintaining stable authentication performance and improving scalability for accommodating new users.

Evaluation Summary. We construct diverse datasets in real-world scenarios to evaluate the authentication performance. Among 50 participants, TAPPASS can achieve reliable authentication with an averaged accuracy of 93.24%. Benefiting from the designed long-term adaptation strategy, the authentication accuracy ranges from 89.27% to 93.33% in an 8-week longitudinal study. Furthermore, TAPPASS exhibits a strong attack resistance capability in terms of passcode replay attacks (EER of 2.45%), behavior mimic attacks (EER of 1.33%), artificial finger presentation attacks (EER of 1.43%), and co-located acoustic replay attacks (EER of 0.82%). Based on the evaluation analysis, we provide valuable insights into how external factors influence the authentication performance.

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

TABLE I: Comparison of our approach with previous approaches that are designed for PIN-based second-factor authentication. B: behavioral biometrics. P: Physiological biometrics. If the number of registered samples in the enrollment phase does not exceed 10, it is considered enrollment effortless.

Approach	Modality	Behavior	Biometrics	Long-term Stability	Biometric Consistency	Interference Resistance	Enrollment Effortless	User Scale	Authentication Accuracy
SonicPrint [16]	Friction sounds	Swipe	P	Unknown	Medium	✗	✓	31	98%
FingerPattern [10]	Friction sounds	Swipe	P	4 weeks	High	✗	✗	20	97.5%
PressPIN [9]	Structure-borne sounds	Tap	B	5 weeks	Medium	✓	✓	30	96.7%
TaPIN [11]	IMU+Audio	Tap	B+P	1 week	Low	✓	✗	20	1.9% (EER, sitting) 3.1% (EER, walking)
FingerVib [12]	IMU+Audio	Tap	P	2 weeks	Medium	✓	✓	41	98.53% (sitting) 92.42% (walking)
TouchPrint [14]	Ultrasounds	Swipe+Tap	B+P	Unknown	Low	✓	✗	30	92.8%
SwipePass [13]	Ultrasounds	Swipe	B+P	Unknown	Medium	✓	✓	36	96.8%
TAPPASS (Ours)	Ultrasounds	Tap	P	8 weeks	High	✓	✓	50	93.24%

II. RELATED WORK

A. *Enhanced PIN-based User Authentication*

As a widely used user authentication technology on mobile phones, password-based authentication relies on users memorizing and manually inputting a secret string (numbers, letters, or symbols) to gain access to protected mobile services. Despite its simplicity and ease of deployment, password authentication is vulnerable to attacks like shoulder surfing or smudge tracing [1], [2]. These vulnerabilities have motivated extensive research into enhancing password security. For instance, numerous common studies strengthen PIN authentication by changing the position, size, or color of the digit keypad [3], [4]. Yet, it still faces the risk of PIN leakage and brings additional authentication burden on users. The tapping behavioral pattern varies from person to person, inspiring various solutions to extract behavioral biometrics for authentication [5]–[7]. Although behavioral biometrics cannot be easily imitated, it faces the limitations of motion artifacts and authentication instability. Subsequently, some solutions explore alternative authentication to enhance the security level of PIN input. The pressure applied when tapping the PIN code is validated as an effective authentication approach. For example, Zhou *et al*. [9] analyze the structure-borne sound propagation attenuation to sense the magnitude of tapping pressure and extract the pressure code. However, pressure-based authentication places a memory burden on users; for example, they not only have to remember the PIN code but also the pressure of their fingers when pressing the keys. Other related solutions leverage unobservable channels to tap the PIN code, like eye gazing [20], [21]. Additionally, vibrations induced by finger tapping are related to finger shapes and tapping behavioral habits, underlying sufficient biometrics [22]. It motivates prior literature [11], [12], which leverages the fusion of IMU sensors and microphones to achieve user authentication on mobile phones. The sounds produced by the friction between the fingerprint and the screen contain biometric information about the fingerprint's texture. Rathore *et al*. [16] and Zhou *et al*. [10] leverage microphones to capture sound friction sounds for user authentication. Yet, these solutions still face severe interference induced by motion artifacts and ambient noise.

Recently, with the development of ultrasonic sensing technology on mobile phones, acoustic-based second-factor authentication has attracted wide attention [23]. Recent solutions leverage noise-resistant ultrasounds to capture the finger-reflected echo profiles and extract finger-related biometrics. For example, SwipePass [13] and TouchPrint [14] are designed to leverage ultrasounds emitted by the built-in speakers on mobile phones to extract hand/finger shape traits for mobile user authentication. However, they are mainly designed for continuous hand swipe gestures. Tapping gestures have shorter durations and limited motion dynamics, which can only provide fewer temporal and spatial cues for reliable biometric discrimination. The detailed qualitative comparison between our solution and prior solutions is presented in Tab. I.

B. *Other Mobile User Authentication*

In addition to the above PIN-based authentication, various studies of secure user authentication have been widely conducted. *(i) Touch-based.* When users hold the mobile phones, the vibration responses of their hands vary from person to person. Thus, several solutions [24], [25] leverage the built-in vibration motor to generate active vibrations and leverage the built-in accelerometer to collect hand-related vibration responses for user authentication. *(ii) Face-based.* Traditional CV-based face authentication faces the issues of poor light conditions and privacy concerns. Recently, acoustic-based face authentication has emerged as an alternative or complementary modality to conventional face authentication on mobile devices [26], [27]. To improve the robustness in real-world scenarios, the adaptive face-based authentication techniques are proposed by [28], [29] to address the changes in relative distances and environments. *(iii) Hand-based.* Human hands contain rich and distinctive physiological and behavioral traits [30], like palmprint, hand geometry, and vein patterns. For instance, Wu *et al*. [31], [32] design a multi-modality user

authentication scheme that fuses the camera and the audio transceivers to sense hand geometry. Yang *et al*. [33] design an acoustic-based mobile user authentication that leverages the impact of hands on structure-borne sound propagation. *(iv) Articulatory gesture-based.* Early studies primarily leverage unique dynamics of articulatory gestures as biometrics. For instance, LipPass [34] and VocalLock [35] have explored the authentication effectiveness of leveraging ultrasonic signals to sense the variations of lip motions and vocal tract vibrations, respectively. Additionally, Li *et al*. [36] have tried to derive vocal tract and vocal source biometrics from skin-reflect mmWave signals to enable anti-spoofing user authentication. *(v) Breathing-based.* BreathPrint [37] and BlowPrint [38] leverage built-in microphones on mobile phones to capture breathing or blowing sounds for acoustic-based user authentication. Although they are considered promising first-factor or second-factor authentication solutions on mobile phones, their authentication scenarios are different from our work. Unlike these approaches, which derive biometrics from relatively stable breathing and blowing behaviors, TAPPASS focuses on extracting biometrics from tap behaviors to enhance the security of PIN-based authentication while maintaining seamless integration with existing PIN entry workflows.

III. Attack Model and Biometrics Analysis

A. Attack Model

With the increasing reliance on mobile banking services, users commonly perform account inquiries, fund transfers, or payment confirmations directly on their smartphones rather than visiting offline bank branches. Although this enhances convenience, it also increases the exposure of PIN-based authentication to security risks. Nowadays, PIN codes, typically consisting of 4 or 6 digits, are widely used for authentication in various mobile transaction applications. However, traditional PINs are susceptible to significant security threats in real-world scenarios. In this study, we assume that an attacker may gain unauthorized access to the victim's device through various means, such as theft. Subsequently, the attacker may exploit four distinct attack models, assuming the following prior knowledge:

- *Passcode Replay Attack*. The attacker only knows the passcode of the target mobile phone through shoulder-surfing attacks or other observation-based ways, but has no additional information about the victim's biometric characteristics. The attacker then attempts to authenticate by simply entering the correct passcode.
- *Behavior Mimic Attack*. In addition to knowing the passcode, the attacker observes the victim's interaction behaviors during passcode entry, such as tapping rhythm or finger movement patterns, and attempts to imitate these behaviors during authentication. Furthermore, we consider a stronger adversarial capability in which the attacker can roughly estimate the geometric characteristics of the victim's finger, such as its approximate shape and size, through visual observation.
- *Artificial Finger Presentation Attack*. The attacker attempts to spoof the authentication mechanism by fabricating an artificial finger that approximates the geometric characteristics of the victim's finger. The attacker can measure the shape of a victim's fingers using measurement tools like measuring tape and vernier calipers. Based on measured physical parameters such as finger length, width, or thickness, the attacker creates a fake finger using materials such as medical-grade silicone and uses it to imitate the victim's tapping interaction during authentication.
- *Co-located Acoustic Replay Attack*. The attacker is aware that the authentication mechanism employs ultrasound-based sensing for authentication and secretly records the acoustic signals generated during the victim's legitimate authentication session. The attacker then replays the recorded acoustic signals near the victim's smartphone while entering the passcode to attempt to spoof the authentication mechanism.

B. Biometrics Analysis of Tap Gestures

1) ***Distinctness among Different Users****:* As shown in Fig. 1, on the surface of a mobile phone, the acoustic channel between the speaker and the microphone will be affected by tapping gestures. The variations in these acoustic responses are primarily attributed to differences in users' finger shapes, hand positioning, and tapping behavior. To take a preliminary study, we require three users to tap the same PIN codes with their natural input ways and calculate the channel impulse response (CIR) to represent how the channel is affected by the finger movements. Fig. 2 demonstrates the distinctness in the acoustic responses of different users. As depicted, each user generates unique acoustic profiles when tapping the same PIN, indicating that biometric features derived from these responses can be used for identity authentication.

2) ***Biometrics Inconsistency Caused by Behavioral Patterns****:* The diversity of behavioral patterns may cause the acoustic responses to fluctuate, resulting in instability in the extracted biometrics. As shown in Fig. 3, we can clearly observe that the acoustic responses induced by tapping gestures exhibit significant variability over time. These variations are primarily driven by differences in tapping pressure, tapping speed, hand positioning, and subtle changes in tapping gesture between different attempts. This inconsistency poses a challenge for reliable biometric feature extraction, as stable and repeatable characteristics are essential for accurate authentication. The variability in tapping gestures underscores the difficulty of using tapping-based biometrics for identity authentication, highlighting the necessity to extract consistent biometric features despite these fluctuations. Fig. 4(a) demonstrates the distribution of acoustic responses induced by tap behaviors. We can clearly observe that behavioral patterns significantly increase intra-class distances.

3) ***Biometrics Inconsistency Caused by Passcode Patterns****:* To enhance PIN-based authentication security, users often set different passcodes for different mobile applications. Prior studies mainly focus on mining biometrics from a single passcode pattern. Yet, users prefer to set different passcode patterns for different mobile applications, bringing a barrier to

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

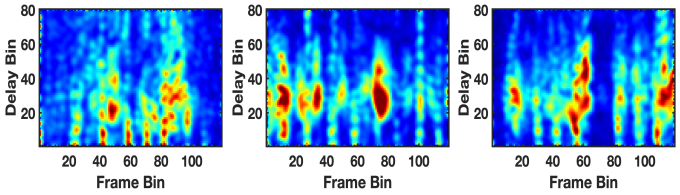


Fig. 2: Distinctness among different users.

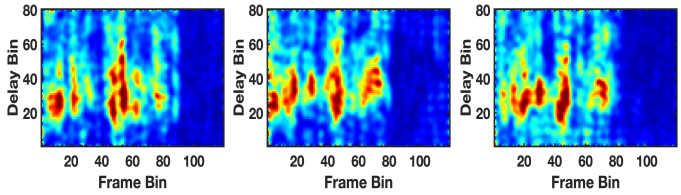


Fig. 3: Inconsistency induced by tap behavior dynamics.

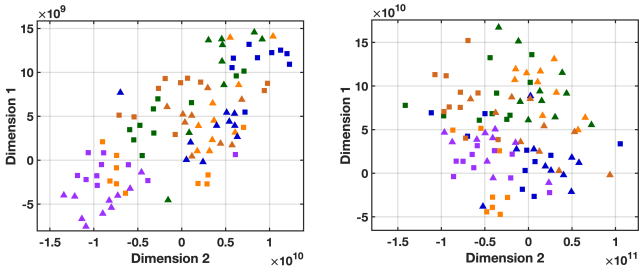


Fig. 4: The distribution of hand-reflected acoustic responses among different sessions. Different colors represent different users, and different shapes represent different passcode patterns (□: "102030"; △: "123456").

existing solutions [13], [14]. As shown in Fig. 4(a), even for the same user, acoustic responses of different passcode patterns still show a large intra-class distance, which may significantly degrade the authentication performance.

*4) **Biometrics Shift over Time**:* Prior studies [17]–[19] have validated that biometrics, while often considered stable, can shift over time due to various factors such as aging, physical condition changes, environmental influences, or habitual behavior variations. These shifts can significantly impact the accuracy of authentication technologies. For instance, factors like changes in a user's finger pressure, typing style, or even skin texture over extended periods may cause deviations in the biometric data that was initially enrolled. Fig. 4 demonstrates acoustic responses of tap behaviors among two sessions. There is a one-week interval between Session 1 and Session 2. Thus, addressing these temporal variations is crucial for ensuring the long-term effectiveness and accuracy of biometric systems, necessitating ongoing adaptations to account for such shifts.

IV. System Design

A. Overview

The overall technical workflow of TapPass is shown in Fig. 5, consisting of the signal preprocessing module, the frame-wise data augmentation module, the enhanced biometrics representation module, and the long-term adaptation module. Specifically, TapPass includes two phases: enrollment phase and authentication phase. When a user performs the PIN entry, TapPass leverages the top speaker to emit ultrasounds and leverages the top and bottom microphones to collect acoustic signals reflected by tap gestures. After the pre-processing, based on the unique acoustic patterns of tap behaviors, the derived high-SNR DCIR profile is augmented to generate more diverse DCIR profiles to save enrollment efforts and improve the generalization capability against dynamic behavioral patterns. It is noted that the data augmentation module is only involved in the enrollment phase. Then, the enhanced biometrics representation module aims to extract consistent biometric features that are robust to passcode and behavior variations for secure authentication. Lastly, TapPass introduces a long-term adaptation mechanism with an adaptive update strategy to guarantee the stability and scalability over time and new users.

B. Acoustic Signal Pre-processing

*1) **Probe Signal Modulation and Reception**:* Benefiting from the characteristics of privacy-preserving, ubiquitousness, and fine-grained granularity, acoustic signals are leveraged as the probe signals to capture fine-grained finger movements. We modulate the acoustic signal with the Frequency Modulated Continuous Wave (FMCW) technology, because it has strong anti-interference ability and accurate measurement capability. Specifically, we set the frequency range of the FMCW to 18-22 KHz (bandwidth size is 4 KHz), because this frequency range is almost inaudible to the human ear and is conducive to non-inductive authentication [39]. The sampling rate of audio signals is set to 48 kHz, and the single-chirp signal duration is set to 0.02 s.

To play the FMCW waveforms, we utilize a single speaker rather than dual speakers to prevent interference between the sound waves emitted by the two sources. Specifically, we selected the top speaker (*i.e.*, the earpiece speaker) over the bottom speaker (*i.e.*, the main speaker), as the latter is primarily used for high-priority playback tasks, which may conflict with our intended acoustic signals. Additionally, users are unable to enter passcodes during phone calls (except when in hands-free mode, which uses the bottom speaker), thus eliminating any potential conflict with the top speaker. Regarding microphones, both the top and bottom microphones are employed to capture reflected sound waves. The top microphone detects signals reflected from the back of the finger, while the bottom microphone captures signals reflected from the side of the finger. This dual-microphone setup enables TapPass to gather information from multiple directions, thereby enhancing the accuracy of the identification process.

*2) **Acoustic Echo Derivation**:* Several candidate metrics can be used to measure signal changes caused by finger movements. Compared with other measurement metrics, such as frequency difference and Doppler frequency shift, CIR

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

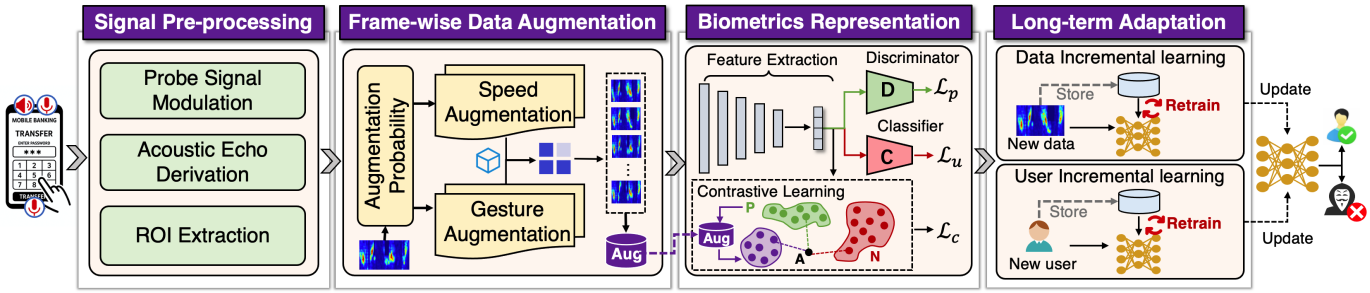


Fig. 5: Technical workflow of TAPPASS, including the signal pre-processing module (§ IV-B), the frame-wise data augmentation module (§ IV-C), the enhanced biometrics representation module (§ IV-D), and the long-term adaptation module (§ IV-E).

(channel impulse responses) can offer fine-grained sensing granularity. For instance, the sensing granularity of the frequency difference can be represented as:

$$d = \frac{c * T * \triangle f}{2B}, \quad \triangle f = \frac{F_s}{N} \tag{1}$$

where $c = 340$ m/s is sound velocity, $\triangle f$ is frequency difference induced by hand movements, $B = 4$ KHz is bandwidth, T is duration of single-chirp signal, F_s is sampling rate, and N is number of FFT points. When N is set to 1920, $\triangle f$ is calculated as 25 Hz, and the value of d is correspondingly calculated as only 2.125 cm, which is insufficient to reflect subtle changes in the fingers. For the Doppler shift, given by $f_c = 20$ KHz, $\triangle v = \frac{c\triangle f}{2f_c} \approx 21.25$ cm/s. It requires users to use an extremely fast speed to input PIN codes unrealistically. Finally, we calculate the CIR to measure subtle finger movements, where the sensing granularity can be denoted by $d = \frac{c}{2F_s} = 0.354$ cm.

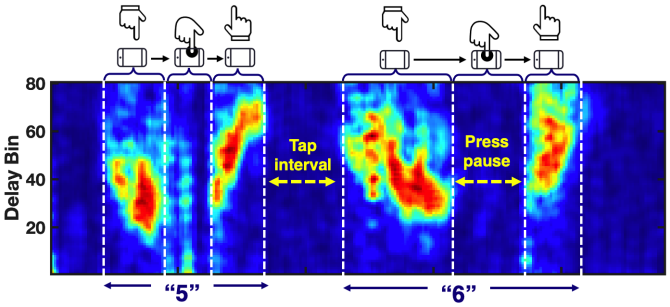


Fig. 6: Acoustic responses of the PIN entry process.

To extract the channel impulse response (CIR), we correlate each received signal frame with the corresponding transmitted signal frame. For the received signals s_r, we first apply a high-pass filter with a 16 KHz to remove the environmental noise interference. Then, s_r is divided into fixed-length frames, *i.e.*, $\{s_r^1, s_r^2, ..., s_r^T\}$, where T denotes the number of frames. The cross-correlation operation between each frame and the transmitted frame is calculated, which is given by:

$$r_{rt}^i(\tau) = \sum_{n=-\infty}^{\infty} s_r^i(n)\, s_t^i(n+\tau) \tag{2}$$

where n and τ denote the samples in the signal frame and the time lag (delay), respectively. The CIR matrix can be expressed as $M_{cir} = \{r_{rt}^1, r_{rt}^2, ..., r_{rt}^T\} \in \mathbb{R}^{D\times T}$, where D and T denotes the number of delay bins and frame bins, respectively. Furthermore, we perform differential calculations on extracted CIR profiles to obtain differential CIR, denoted by $M_{dcir} = \{(r_{rt}^2 - r_{rt}^1), ..., (r_{rt}^T - r_{rt}^{T-1})\}$, eliminating the static interference induced by the chest and the palm.

*3) **ROI Extraction**:* After obtaining the DCIR profiles, we need to extract the Region of Interest (ROI) to focus on the spatial range that is most relevant to tap gestures. In our work, we only preserve the first 80 delay bins of DCIR profiles, which can cover a sensing distance of 30 cm. Then, we leverage a min-max normalization approach to eliminate the negative impact of individual diversity. Most importantly, sometimes we observe that the DCIR presents several areas with particularly high energy, which may be caused by multi-path energy superposition. These high-energy areas may cause other informative areas to be misclassified as noise. Thus, before the min-max normalization, a peak detection and soft thresholding are performed to eliminate abnormal fluctuations. Specifically, we calculate the average value of the entire DCIR matrix (denoted by n_a) and set the soft threshold τ as $5 * n_a$. Then, if the value of the DCIR matrix is greater than the threshold, then it is directly set as τ. Then, we introduce a mean filter with a kernel size of $3 * 3$ to smooth out high-frequency noise while preserving important edge information.

When a user interacts with a mobile phone keypad, they typically perform a series of tap gestures on the touchscreen. Fig. 6 illustrates the acoustic responses (*i.e.*, DCIR) induced by tapping the numbers "5" and "6" during the PIN entry process. When entering each digit of the PIN codes, the movement of the finger can be broken down into three parts: finger falling, finger contact, and finger leaving. Generally, the tap gestures occur in a rhythmic sequence, with small pauses (tap intervals) between each tap gesture. In addition, a brief pause will also occur when the user's finger touches the screen. The tap gesture intervals and press pause intervals vary not only from gesture to gesture, but also from person to person. The variation in the acoustic response highlights the dynamic nature of the tapping gestures, with each key press generating a unique DCIR profile based on factors such as finger movements and contact positions.

C. *Frame-wise Data Augmentation*

The limited scale of the dataset has always been a major issue hindering the user enrollment phase. Prior solutions

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

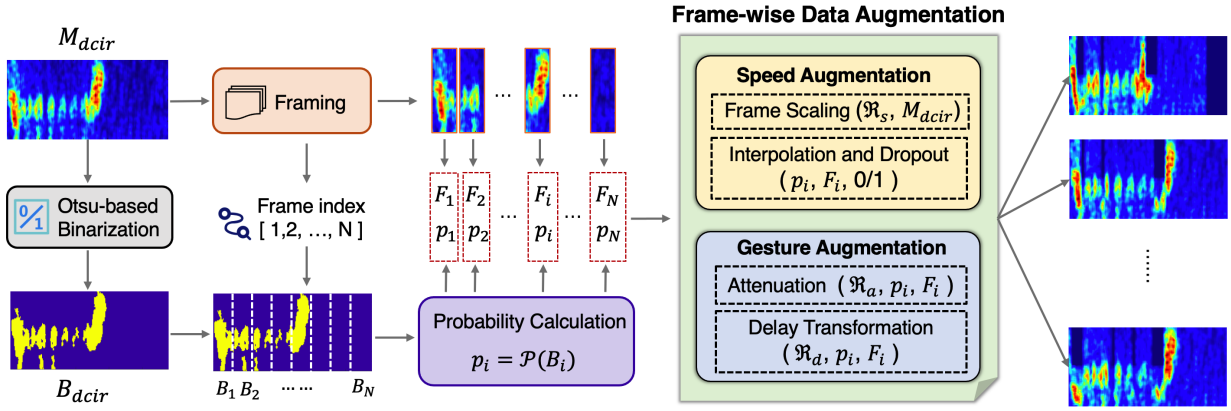


Fig. 7: The workflow of the computing-efficient and interpretable frame-wise data augmentation strategy.

have tried to leverage deep learning to synthesize data or augment data. However, these deep-learning data augmentation approaches (*e.g.*, GAN [40]–[42]) face limitations such as low computational efficiency and large training data requirements [43]. SwipePass [13] designs a lightweight data augmentation approach, which randomly inserts or interpolates frames to simulate variations in speed. However, this design may result in the loss of keyframes. Additionally, the approach mainly focuses on speed variations, overlooking gesture variations, which limits its scalability. In our work, we design a frame-wise DCIR augmentation approach that is based on the unique acoustic patterns and behavioral patterns of tap gestures. The pipeline of data augmentation is shown Fig. 7. First, the augmentation probability is computed adaptively at the frame level based on the proportion of effective DCIR components, enabling content-aware simulation instead of uniform manipulation. Second, the augmentation process explicitly models two distinct sources of variability inherent to tap interactions, *i.e.*, speed variation (via frame-axis scaling and interpolation) and gesture-induced multipath effects (via delay-axis shearing and attenuation modeling).

1) ***Augmentation Probability Calculation****:* Given a DCIR matrix M_{dcir}, we first perform a binarization to identify whether a value of M_{dcir} is the effective component induced by the tap gesture. Specifically, in the binarization process, we introduce the Otsu threshold selection method [44] to determine an optimal threshold, and then construct a binary mask matrix B_{dcir}. The same frame operations are applied on the given DCIR matrix M_{dcir} and the binary mask matrix B_{dcir} to obtain $F = \{F_1, F_2, ..., F_N\}$ and $B = \{B_1, B_2, ..., B_N\}$, respectively. Based on the frames of the binary mask matrix, we calculate the augmentation probability for each frame. For the i-th frame (*i.e.*, B_i), the corresponding augmentation probability is defined as:

$$p_i = \mathcal{P}(B_i) = \frac{\sum_j (B_i(j) == 1)}{\text{size}(B_i, 1) * \text{size}(B_i, 2)} \tag{3}$$

where j represent the element index of B_i. Eq. 3 represents the ratio of effective components in the i frame (*i.e.*, F_i). A higher value for p_i indicates that the F_i contains more tap-related information. Guided by the augmentation probability, adaptive augmentation can be performed based on the ratio of effective components in each frame. After the above steps, the speed augmentation and gesture augmentation are applied on F_i with the augmentation probability of p_i.

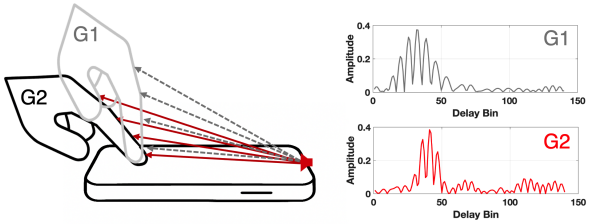


Fig. 8: Delay patterns for different gestures when tapping the same position of the screen.

2) ***Speed Augmentation****:* During the process of PIN entry, the speed of tap gestures is affected by individual states and the external environments, highlighting the dynamic nature of behavioral patterns.

[Key Observation.] *As shown in Fig. 6, the speed variations are directly reflected in the varying number of frames in the DCIR matrix, such as tap intervals and press pause intervals.* Based on that, in order to simulate the effects of different tap speeds, we perform two-step augmentation operations, including coarse-grained frame scaling and fine-grained interpolation and dropout. First, an affine shear transformation is implemented to scale the entire DCIR matrix M_{dcir} along the frame axis (x), where the shear is defined as $x' = k \cdot x$, with $k = \text{rand}(\mathcal{R}_s)$ controlling the shear factor. Considering the variation in tap speed during PIN entry, $\mathcal{R}_s$ can be set as $[0.8, 1.2]$. Then, a fine-grained frame interpolation and dropout operation is performed to simulate different speeds on a micro scale. Specifically, for the i-th frame (*i.e.*, F_i), it will be randomly inserted with interpolation or dropped out with a probability of p_i.

3) ***Gesture Augmentation****:* In addition to the speed variations, gesture variations also contribute to the behavioral patterns. As shown in Fig. 8, even when tapping the same screen position, the user's tapping gestures are different each time, such as the degree of finger bending and the angle

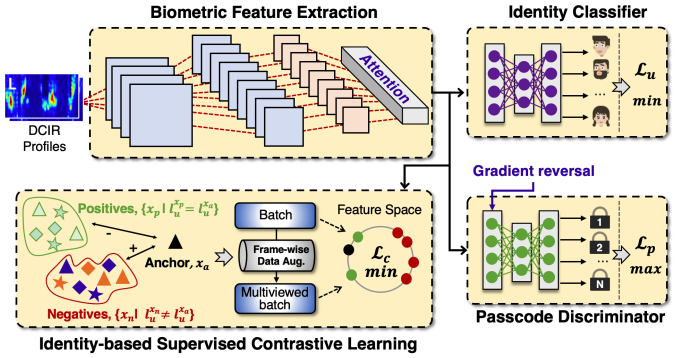


Fig. 9: The structure of the deep learning-based biometric representation network, which fuses the contrastive learning and adversarial learning to extract consistent biometrics for stable authentication.

between the finger and the screen.

[Key Observation.] *Tap gesture variations directly affect the propagation paths of acoustic signals, as demonstrated in Fig. 8, where G1 and G2 present significantly different delay patterns.* Given transmitted signals S_t, the reflected acoustic signals can be expressed as $S_r = \sum_{i=1}^{L} A_i * S_t(t - \tau_i)$, where A_i and τ_i represent the attenuation factor and delay of the i-th signal path, and L denote the numbers of the reflected paths. Therefore, it can be seen that different gestures directly introduce different attenuation and delay patterns as shown in Fig. 8. Based on that, we design a data augmentation strategy from signal attenuation and signal delay to simulate the impact of gesture variations. First, for the i-th frame (F_i), it will be determined whether to be attenuated with the probability of p_i. Specifically, the attenuation factor is defined as $A_i = p_i * \text{rand}(\mathcal{R}_a)$, where $\mathcal{R}_a = [0.6, 1.2]$. Subsequently, we perform the shear transformation on F_i along the delay axis (y) with the augmentation probability of p_i. Specifically, the shear transformation is defined as $y' = k \cdot y$, where $k = \text{rand}(\mathcal{R}_d)$ is the shear factor and $\mathcal{R}_d = [0.7, 1.5]$.

It is noted that the objective of the designed augmentation approach is not to reconstruct real acoustic fields, but to transform delay-temporal structures in DCIR representations to approximate the effects of different tap speeds and gestures. Even with only the limited enrollment samples, the authentication model in TAPPASS can learn consistent user biometrics from augmented DCIR representations and yield accurate and robust authentication performance against realistic behavioral variability.

D. Enhanced Biometrics Representation

To address the challenge of the inconsistency of biometrics caused by dynamic behavioral patterns and passcode patterns, we design a deep learning-based biometrics representation network that combines contrastive learning and adversarial learning. The detailed structure is demonstrated in Fig. 9.

*1) **Common Biometric Feature Extraction**:* The DCIR profiles extracted from hand-reflected acoustic signals contain both behavioral information (*e.g.*, tap speed and duration) and physiological information (*e.g.*, finger shape and hand geometry). The goal of TAPPASS is to extract the consistent physiological features that are more distinctive and more stable, as biometrics. Thus, we design a biometric feature extraction module (*i.e.*, G_f) to map the DCIR profiles to high-level feature representations. The calculation process is denoted by $x = G_f(M_{dcir}, \omega_e)$, where ω_e denotes the parameters of the feature extraction module. The feature extractor is implemented as a lightweight hierarchical convolutional neural network composed of three convolutional blocks. Each block consists of a convolutional layer, a batch normalization layer, a spatial dropout layer, and a max-pooling layer followed by ReLU activation. Specifically, the kernel sizes of three convolutional blocks are 3×3, 5×5, and 7×7. Such a progressive design in kernel size allows the network to learn discriminative features across multiple scales.

*2) **Identity-based Supervised Contrastive Learning**:* Inspired by the success of supervised contrastive learning (SCL) [45], the high-level feature representation x is fed into a carefully-designed supervised contrastive learning network to eliminate the negative impact of dynamic behavioral patterns. Given an anchor x_a with the user's identity label of $l_u^{x_a}$, the positive sample is defined as x_p with the same identity label, and the negative sample is defined as x_p with a different identity label. The supervised contrastive learning network aims to pull together an anchor and all positive samples and push apart the anchor from all negative samples with the user's identity labels. Traditional SCL techniques leverage random augmentations to generate different views of data. These random augmentations are designed for image data and are hard to model channel-sensitive acoustic signals. Thus, the designed data augmentation strategy in Sec. IV-C is adopted to generate the multiviewed batch. After that, the similarity of samples in the multiviewed batch is constrained by the supervised contrastive loss [45]:

$$\mathcal{L}_c = \sum_{i \in I} \mathcal{L}_i^c = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \quad (4)$$

where I is the sample set including the batch and the multiviewed batch, $P(i)$ is the set of positive samples, τ is a scalar temperature parameter, $A(i)$ is the set of all samples except the anchor sample, $\boldsymbol{z}_i$ is the anchor sample, and $\boldsymbol{z}_p$ is the positive sample.

*3) **Passcode Discriminator**:* Inspired by adversarial learning, a passcode discriminator network is introduced to force the network to ignore the passcode variations. The passcode discriminator network is trained adversarially with the biometric feature extraction network in a way that encourages the feature extractor to produce features that are passcode-invariant. The idea is that the feature extraction network should make it difficult for the passcode discriminator network to identify which passcode class the high-level feature representation x belongs to. The passcode discriminator is implemented as a shallow fully connected network that consists of two fully connected layers. The final fully connected layer projects the features into different passcode classes and applies the LogSoftmax function to generate probability outputs. The loss function $\mathcal{L}_p$ is defined as the cross-entropy function of the true passcode pattern labels and the predicted results. Specifically, a gradient inversion layer (GRL) is added before the first fully

connected layer of the passcode discriminator to invert the gradient during backpropagation.

*4) **Identity Classifier**:* To ensure that the model does not lose identity information during adversarial learning, an identity classifier is added. Unlike the prior passcode discriminator network, the identity classifier aims to recognize users' identities as accurately as possible based on high-level feature representations x. The identity classifier consists of three fully connected layers, and the last layer is followed by a LogSoftmax operation to produce numerically stable log-probability scores that characterize the user's identity. The loss function $\mathcal{L}_u$ is also defined as the cross-entropy function of the true labels and the predicted results. The detailed model training strategy is introduced in Sec. V-A2. Then, the output score is compared with a global decision threshold that is shared across all enrolled users to determine whether the user's identity is legitimate. The global threshold is empirically selected based on the similarity score distribution of genuine samples and spoofing samples in a validation dataset. In Sec. V-D, we provide an evaluation analysis of the authentication decision threshold selection.

E. Long-term Adaptation

*1) **Adaptation to Biometrics Distribution Shifts**:* Although the above steps can extract the relatively stable tap biometrics for user authentication, biometric features still yield distribution shifts over time. The dynamic behavioral patterns cause shifts in the underlying data distributions, making it difficult for static authentication models to maintain high performance across different conditions and periods. Therefore, we leverage data incremental learning to cope with biometrics distribution shifts over time. Unlike the conventional incremental learning technique [46] that updates models only based on intra-class distances, TAPPASS triggers model retrain whenever a significant drift is detected using both intra-class and inter-class distances.

We maintain a exemplar set $\mathbb{E} = \{E_1, E_2, ..., E_N\}$, where $E_i = \{e_i^1, e_i^2, ..., e_i^K\}$ represents the enrollment samples of the i-th registered user. For each registered user, the feature centroid C_i is calculated by mean(E_i). The biometric feature shift estimator is designed to continuously monitor the intra-class and inter-class distances between the newly incoming sample x_{in} and the existing class centroids $\mathbb{C} = \{C_1, C_2, ..., C_N\}$, determining whether to retain the authentication model A_{model}. The basic idea of the model update strategy is that after a new legitimate sample is added, it should still maintain a relatively close distance to the updated class centroids and be far away from other class centroids. Inspired by this, the design of the model update strategy is as described in Alg. 1. A newly incoming sample that will trigger the update mechanism must satisfy two conditions. (1) The sample must also fall within the maximum intra-class distance threshold (lines 10 to 13). (2) The distance between the sample and the updated class centroid must remain smaller than the minimum inter-class distance (lines 14 to 20). Only the sample that meets the above conditions will be included in the example set and trigger the model update mechanism. Even if an illegitimate sample is misidentified as a legitimate user, this sample should be relatively far from the class centroid compared to samples in the exemplar set $\mathbb{E}$. Therefore, this sample does not trigger model retraining. This scalability is essential in ensuring that TAPPASS can handle biometrics distribution shifts over time without compromising the authentication process.

Algorithm 1: Model Update Strategy

Input: x_{in}: the newly incoming biometric feature;
$\mathbb{E} = \{E_1, E_2, ..., E_N\}$: an exemplar set including biometric features of N registered users;
$E_i = \{e_i^1, e_i^2, ..., e_i^K\}$: biometric features of the i-th registered user;
A_{model}: the authentication model.
Output: $\hat{A}_{model}$: the updated authentication model..

1 $\mathbb{C} = \{C_1, C_2, ..., C_N\}$; // Class centroids of N registered users.
2 **for** $i \leftarrow 1$ **to** N **do**
3 $C_i = \text{mean}(E_i) = \frac{1}{K}\sum_{k=1}^{K} e_i^k$;
4 **end**
5 **if** $u = A_{model}(x_{in}) \notin N$ **then**
6 **return**; // x_{in} is identified as an illegitimate user.
7 **else**
8 $\hat{C}_u = \text{mean}(E_i \bigcup \{x_{in}\})$; // Update the class centroid of the u-th registered user.
9 $d_w = \text{Distance}(\hat{C}_u, x_{in})$;
10 **for** $i \leftarrow 1$ **to** K **do**
11 $d_{in}^i = \text{Distance}(\hat{C}_u, e_u^i)$; // Calculate the intra-class distances.
12 **end**
13 $d_{in}^{max} = \max(\{d_{in}^1, d_{in}^2, ..., d_{in}^K\})$;
14 **for** $i \leftarrow 1$ **to** N **do**
15 $d_{out}^u = +\infty$;
16 **if** $i \neq u$ **then**
17 $d_{out}^i = \text{Distance}(C_i, x_{in})$; // Calculate the inter-class distances.
18 **end**
19 **end**
20 $d_{out}^{min} = \min(\{d_{out}^1, d_{out}^2, ..., d_{out}^N\})$;
21 **end**
22 **if** $d_w < d_{out}^{min}$ *and* $d_w < d_{in}^{max}$ **then**
23 $m = \text{argmax}(\{d_{in}^1, ..., d_{in}^i, ..., d_{in}^K\})$;
24 $e_u^m = x_{in}$; // Update the exemplar set using x_{in}.
25 $\hat{A}_{model} = \text{RETRAIN}(A_{model}, \mathbf{E})$; // Retrain the authentication model using the updated exemplar set.
26 **else**
27 Do not update.
28 **end**

*2) **Adaptation to New Users**:* Additionally, traditional authentication models may be biased toward the data from a limited set of initial users, leading to reduced accuracy for new enrollment users. Intuitively, the authentication models can be retrained from scratch using new users' data. But this process requires a high computational cost and brings a long authentication waiting time. To address it, we introduce the user incremental learning technique to incorporate new users without retraining the entire authentication model. When a new user is added, TAPPASS collects several samples in the enroll-

ment phase. Then, TAPPASS leverages the designed frame-wise data augmentation approach (Sec. IV-C) to generate more diverse samples. After that, the biometric features are extracted from these original samples and augmented samples and are stored in the exemplar set $\mathbb{E}$.

F. Expanding to Dual-hand Enrollment and Authentication

While users tend to hold their phones with their dominant hand in most daily usage scenarios, there are certain situations where they may use their non-dominant hand. Obviously, there are also distinct differences between the left and right hands of the same person. To address this limitation, we introduce a dual-hand enrollment strategy inspired by multiple-fingerprint enrollment.

During the enrollment phase, each user provides tap samples from both hands. For each hand, the user enters several repetitions of their chosen PIN sequences, following the standard tapping procedure. It ensures that TAPPASS captures the hand-specific acoustic channel responses for both the left and right hands. The collected acoustic signals from each hand are processed using the existing technical pipeline as introduced in Sec. IV, including signal pre-processing, data augmentation, and biometric representation. Then, two distinct templates are created per user, *i.e.*, one for the right hand and one for the left hand. Assuming that there are N users who have completed the enrollment phase, the user template set is denoted by $\Gamma_u = \{\Gamma_1^r, \Gamma_1^l, ..., \Gamma_N^r, \Gamma_N^l\}$, where Γ_i^r and Γ_i^l represent the right-hand template and left-hand template of the i-th user. During user authentication, the incoming acoustic signals are processed by signal pre-processing, data augmentation, and biometric representation. Then, the generated high-level feature representation is matched against these templates in Γ_u. A successful authentication occurs if the similarity score exceeds the pre-designed threshold for either template.

V. PERFORMANCE EVALUATION

A. Experimental Setup

*1) **Data Collection**:* To facilitate data collection, we developed an Android application [1]. capable of emitting ultrasounds through the top speaker while simultaneously receiving echo signals through two microphones. Totally, 50 participants (18 female and 32 male, aged between 20 and 43 years) are involved in the data collection process. Before data collection, all participants are required to sign an informed consent form outlining the research objectives, data collection methods, and privacy protection measures. All collected data is anonymized and stored to ensure compliance with privacy regulations. To evaluate the effectiveness, robustness, and security of TAPPASS, we construct three types of datasets.

♢ *Dataset-A.* A total of 50 participants are involved in data collection in an office room with the noise level of 41.13 dB. This basic dataset is leveraged to evaluate the overall performance of TAPPASS. Each participant is asked to use their dominant hand to enter three basic PIN code sequences, *i.e.*, "123456" (P1), "202383" (P2), and "102030" (P3), on a mobile phone (Redmi Note12 Turbo). These sequences were chosen to vary in length and spatial structure to assess the system's adaptability to different passcode patterns. Each participant performs 30 attempts with a specific passcode sequence. To minimize fatigue and ensure natural input behaviors, a 5-minute break was provided after every 10 attempts.

♢ *Dataset-B.* To evaluate the robustness of TAPPASS under different conditions, we select a range of diverse scenarios for data collection, including a living room with the noise level of 32.37 dB, an outdoor street with the noise level of 65.13 dB, and a cafeteria with the noise level of 56.75 dB. Specifically, *Dataset-B* is constructed with 15 participants who participated in *Dataset-A*. In each scenario, 5 participants are selected to be involved in the data collection process. Each participant is asked to enter three different passcode patterns mentioned above with their habits, and each pattern is entered 20 times. In addition, factors such as body postures, device types, handheld postures, and motion states are also present to further test the robustness and usability.

♢ *Dataset-C.* To assess the stability and consistency of the collected biometric data over time, we conducted a long-term data collection process involving 8 participants. Data was gathered from these participants over a period of eight consecutive weeks, with weekly sessions designed to capture variations in user behavior and biometric features. During each session, participants were asked to enter three different PIN codes, with each code repeated multiple times.

*2) **Model Training and Evaluation**:* The proposed deep learning model in Sec. IV-D is implemented using the PyTorch framework. We combine $\mathcal{L}_u$, $\mathcal{L}_p$, and $\mathcal{L}_c$ for model training. The entire model training process is completed on a high-performance server equipped with 144 Intel(R) Xeon(R) Platinum 8352V@2.10GHz CPUs and 3 NVIDIA RTX A6000 GPUs. Specifically, the model is trained using the Adam optimizer with an initial learning rate of 1.0×10^{-3} and a batch size of 128. The learning rate is adjusted using a step-wise schedule, where it is multiplied by a factor of 0.8 every 10 epochs. The training process consists of approximately 100 epochs. During the training process, the best-performing model is selected based on validation loss and averaged authentication accuracy calculated on a held-out validation subset. In this study, we leverage the Authentication Accuracy (ACC), the False Accept Rate (FAR), the False Reject Rate (FRR), the Receiver Operator Characteristic (ROC) curve, the Equal Error Rate (EER), and the Area Under Curve (AUC) as evaluation metrics.

B. Overall Performance

We use *Dataset-A* involving 50 participants to give an overall performance analysis. 45 participants are selected to be legitimate users, and 5 participants are selected to be attackers. The attackers are automatically categorized into one class. For each passcode pattern, 10 samples are used for training, and the remaining 20 samples are used for evaluation. As shown in Fig. 10, TAPPASS can achieve an averaged accuracy of 93.24%. While a small number of participants exhibit slightly

[1]We release the APK (Android Package) on https://anonymous.4open.science/r/TapPass_data_collection_APK-2D8B/

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

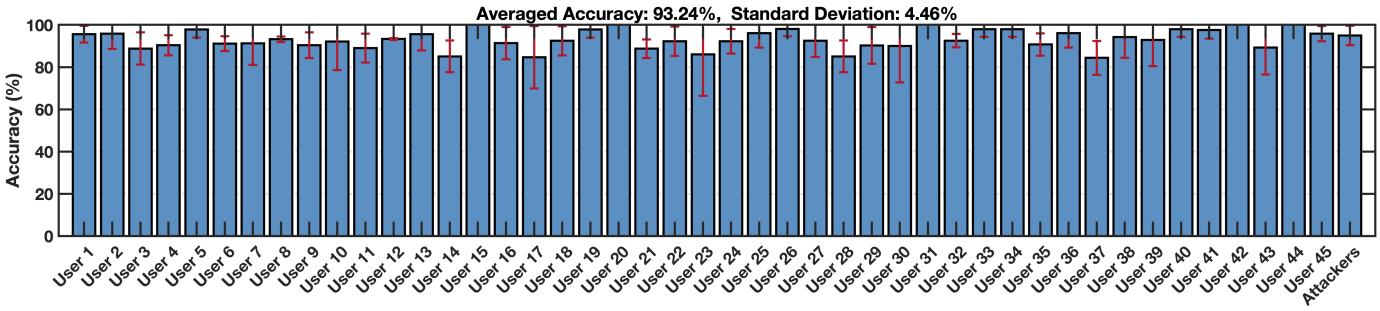
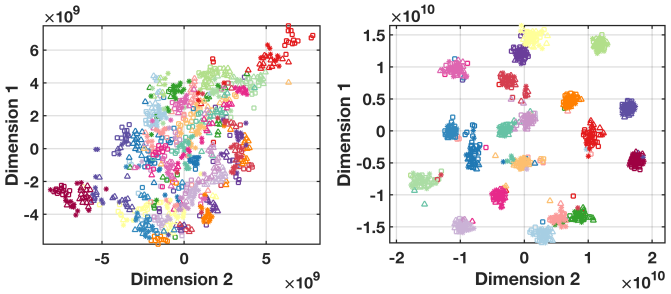


Fig. 10: Overall authentication performance among 50 participants, where 45 participants are considered as legitimate users and the remaining 5 participants are considered as attackers.



(a) Original biometrics. (b) Enhanced biometrics.

Fig. 11: The t-SNE visualization of (a) original tap biometrics and (b) enhanced tap biometrics, where different colors represent different users, and different shapes (*e.g.*, □, △, ∗) represent different passcode patterns.

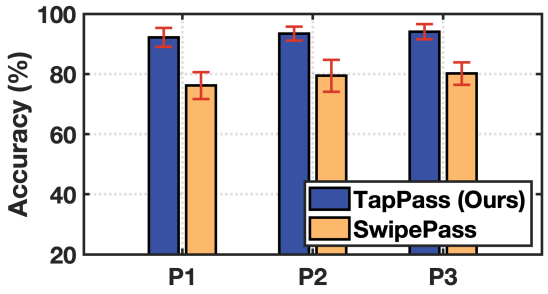


Fig. 12: Baseline comparison between TAPPASS and SwipePass among different passcode patterns.

lower accuracy and larger variance, TAPPASS still achieves consistently high accuracy for most users, with the majority exceeding 90%. These results confirm that TAPPASS can effectively authenticate a wide range of users while maintaining reliable and balanced performance at the individual level. Furthermore, to give a clear demonstration, we use the t-SNE approach to visualize the extracted tap biometric features. As shown in Fig. 11, our solution can effectively extract discriminative and consistent biometric features. For instance, the original feature points of different users in Fig. 11(a) show severe overlap and uneven distribution. Yet, in Fig. 11(b), after the processing of TAPPASS, these feature points show a concentrated distribution.

C. Baseline Comparison

Since our evaluation involves a more diverse set of users and testing conditions, only qualitative comparison in Tab. I is not entirely fair. Thus, we reproduce a representative prior work closely related to our work (*i.e.*, SwipePass [13]) and evaluated it using our dataset to give a quantitative baseline comparison. It is worth noting that the prior work also employs data augmentation techniques to enhance generalization and robustness. Therefore, to ensure a fair baseline comparison, we disable data augmentation for both approaches and compare the performance of authentication models under the same dataset and evaluation protocol. As shown in Fig. 12, among three passcode patterns, TAPPASS significantly outperforms SwipePass by 16.14%, 14.06%, and 13.23% in terms of authentication accuracy, respectively. That is because TAPPASS can effectively eliminate the ambiguity of biometric features induced by passcode and behavior variations, thereby extracting more consistent and more distinct tap biometrics for reliable authentication across a wider user group.

D. Authentication Decision Selection Study

In the authentication phase, the identity classifier (Sec. IV-D4) outputs the similarity score between the extracted biometric representation and the enrolled template. If the similarity score is larger than the pre-defined global threshold, the input acoustic sample is accepted; otherwise, it is rejected. The relationship between FAR and FRR as the similarity threshold changes is shown in Fig. 13. As the threshold increases, more and more spoofing samples will be correctly identified, but consequently, more and more genuine samples will also be incorrectly rejected. Thus, the decision threshold is selected based on the operating point that balances the trade-off between FAR and FRR, *i.e.*, 0.59. In this threshold condition, TAPPASS can achieve reliable authentication performance with FAR of 3.07% and FRR of 3.81%.

In addition to the ML-based authentication decision, DTW (Dynamic Time Warping) is a commonly used quantitative feature similarity technique. Next, we additionally leverage the DTW approach to give a feature-level similarity measure and report authentication performance with different similarity thresholds. As shown in Fig. 14, we can find that when the optimal global DTW threshold is set to 0.16, a trade-off can be achieved between FAR and FRR. In this threshold condition,

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

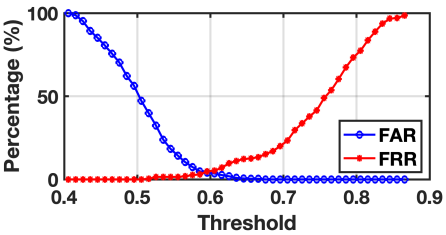


Fig. 13: The impact of threshold on ML-based authentication decisions.

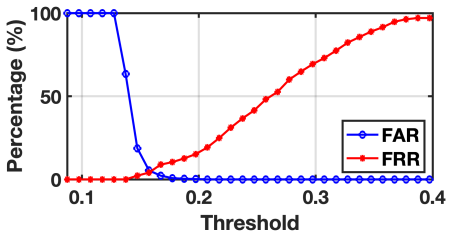


Fig. 14: The impact of threshold on DTW-based authentication decisions.

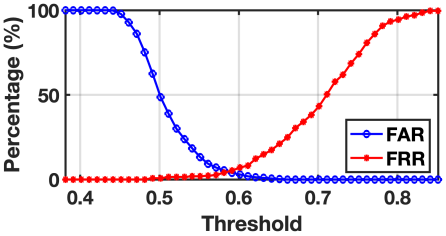


Fig. 15: The dual-hand enrollment strategy with ML-based thresholds.

TABLE II: Ablation study of the biometric representation network designed in Sec. IV-D.

	ACC(%)	FAR(%)	FRR(%)
BFE + IC	74.54 (3.13)	19.04 (1.69)	19.83 (1.46)
BFE + IC + PD	82.16 (2.85)	9.65 (1.34)	10.51 (1.37)
BFE + IC + SCL	86.27 (2.26)	5.63 (1.18)	6.70 (0.95)
BFE + IC + SCL + PD	94.26 (2.32)	2.78 (0.89)	3.35 (0.79)

TAPPASS can achieve authentication performance with FAR of 3.71% and FRR of 5.57%. Compared with the ML-based authentication decision that achieves a FAR of 3.07% and a FRR of 3.81%, the DTW-based authentication decision yields slight performance degradation. It indicates that the ML-based authentication decision is more suitable for our solution. In addition, we also observe that different similarity metrics lead to different decision thresholds, resulting in distinct optimal global thresholds for authentication.

E. Generalization of Dual-hand Authentication Scenarios

During daily usage, users may switch hands to tap PIN codes. Next, we conduct a generalization of dual-hand authentication scenarios to validate the effectiveness of the dual-hand enrollment strategy introduced in Sec. IV-F. There are 23 participants who simultaneously register both their left and right hands. Fig. 15 demonstrates the variation of FAR and FRR with ML-based authentication thresholds. At the threshold of 0.59, TAPPASS can yield a FAR of 3.15%, a FRR of 4.51%, and an EER of 3.62%. By comparing Fig. 13 with Fig. 15, we can clearly observe that TAPPASS can achieve comparable performance in single-hand authentication scenarios and dual-hand authentication scenarios. Essentially, dual-hand authentication can be regarded as an extension of single-hand authentication, where the right and left hands of the same user are modeled as two different authentication templates.

F. Ablation Study

A key step in TAPPASS is to leverage a deep learning-based biometric representation network designed in Sec. IV-D to extract discriminative and consistent features for user authentication. Next, we conduct an ablation study to validate the effectiveness of key components/modules (BFE: biometric feature extraction; SCL: supervised contrastive learning; IC: identity classifier; PD: passcode discriminator) in the biometric representation network. A total of three user partitions are involved in the ablation study. In each round of user partitioning, 10 participants are randomly selected as legitimate users, and the remaining 40 participants are selected as attackers. This setup in each user partition better reflects practical authentication scenarios, where a relatively small set of legitimate users must be protected against a much larger population of potential attackers. The mean (standard deviation) of accuracy, FAR, and FRR among three user partitions are shown in Tab. II. We can clearly observe that each module in the biometric representation network contributes to the final authentication performance. Most importantly, although SCL or PD can mitigate the issue of biometric inconsistency induced by behavior and passcode variations, they still struggle to achieve significant improvements. Although using either the SCL module or the PD module individually can mitigate the issue of biometric inconsistency induced by behavior and passcode variations, they still struggle to achieve significant improvements. Combining these two modules can further improve the authentication performance.

G. Effectiveness of Data Augmentation

Next, we give a quantitative evaluation of data augmentation effectiveness by comparing authentication performance with and without data augmentation across different enrollment sizes. The amount of data collected during the enrollment phase directly impacts user experience and authentication accuracy. Intuitively, the more enrollment data collected, the more accurately the user template can be generated, potentially leading to higher authentication accuracy. However, extensive enrollment data collection can degrade user experience. Fig. 16 illustrates the authentication accuracy of TAPPASS under different enrollment sizes, with and without the proposed data augmentation. Overall, the authentication accuracy consistently improves as the enrollment size increases, indicating that more enrollment data provides richer biometric representations. Notably, the authentication accuracy with augmentation significantly outperforms the one without augmentation, especially when the enrollment size is small. Even with the enrollment size of 5, TAPPASS can achieve a high authentication accuracy of about 91%. It demonstrates that the designed data augmentation effectively compensates for data scarcity and substantially boosts accuracy.

H. Longitudinal Study

Long-term stability is crucial for TAPPASS because it directly relates to reliability and availability in real-world

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

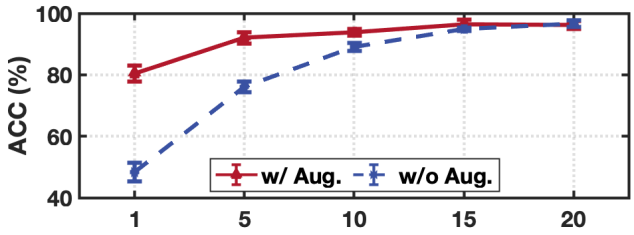


Fig. 16: Data augmentation validation and impact of different enrollment sizes.

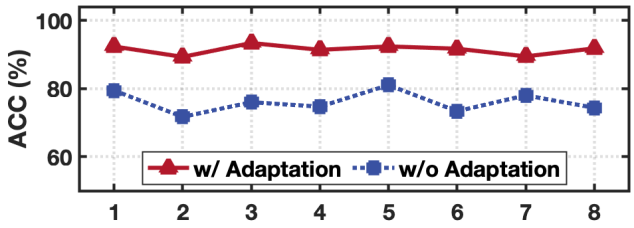


Fig. 17: Long-term adaptation validation with a longitudinal study over 8 weeks.

scenarios. In *Dataset-C*, we collect data from 8 participants over 8 consecutive weeks. The authentication model is trained with the data collected in the first week and evaluated using data collected in subsequent weeks. Fig. 17 presents the authentication performance of TAPPASS over an extended period, comparing performance with and without the proposed long-term adaptation mechanism (Sec. IV-E). With the assistance of the long-term adaptation mechanism, TAPPASS maintains a more stable performance over time, where the authentication accuracy ranges from 89.27% to 93.33% with a standard deviation of 1.42%. Such a small accuracy fluctuation indicates that the designed long-term adaptation mechanism can effectively mitigate the negative impact of biometric drift. Without the adaptation mechanism, the averaged authentication accuracy is only 76.04% with a standard deviation of 3.18% over 8 consecutive weeks, exhibiting significant fluctuations. These results demonstrate that the proposed adaptation strategy plays a crucial role in sustaining reliable authentication performance under temporal variations.

I. Robustness and Usability Study

We use the *Dataset-B* to conduct a robustness and usability study. Since the participants in *Dataset-B* are also involved in *Dataset-A*, we use data in *Dataset-A* for training and data in *Dataset-B* for evaluation.

*1) **Impact of Passcode Patterns**:* Passcode patterns (*e.g.*, length and complexity) are crucial for the security level of user authentication. In addition to the three basic passcode patterns (P1: "123456", P2: "202383", and P3: "102030") in *Dataset-A*, two other passcode patterns (*i.e.*, P4: "555555", P5: "123456789") are involved in evaluation. As shown in Fig. 18, compared with P1, P2, P3, and P5, P4 shows a higher FAR and FRR. Regarding this phenomenon, we have made the following speculation. Repeated tap patterns significantly reduce the spatial diversity present in the acoustic reflections. When users repeatedly tap the same key position, the resulting ultrasonic reflections originate from highly similar spatial configurations. Consequently, the DCIR patterns across different users become more structurally similar compared to passcodes that involve multiple key positions. In contrast, passcodes containing taps on different screen locations introduce richer spatial perturbations in the acoustic propagation paths. These variations produce more diverse multipath reflection patterns that better capture user-specific hand geometry, thereby improving inter-user separability. Inspired by the above results, we recommend that users choose digits with relatively large spatial distances to form the passcode sequence during the enrollment phase. Generally, the longer the passcode sequence, the higher the security.

TABLE III: Cross-volume evaluation on the same device.

	Vol=100%	Vol=75%	Vol=50%	Vol=25%
Redmi Note12 (Basic)	95.84%	95.65%	95.77%	95.14%
HUAWEI P60	96.27%	96.31%	96.53%	96.48%
HUAWEI Mate40 Pro	98.42%	98.37%	98.33%	98.24%
Redmi K50 Pro	95.12%	95.33%	95.02%	94.87%

TABLE IV: Cross-device evaluation among different volume levels.

	Vol=100%	Vol=75%	Vol=50%	Vol=25%
HUAWEI P60	58.74%	56.67%	43.33%	34.48%
HUAWEI Mate40 Pro	6.78%	9.87%	0	0
Redmi K50 Pro	7.42%	6.67%	3.33%	16.67%

*2) **Impact of Diverse Scenarios**: Dataset-B* includes the acoustic samples collected in diverse scenarios with different noise levels, *i.e.*, a cafeteria (Env. 1), a living room (Env. 2), and a street (Env. 3). As shown in Fig. 19, the authentication accuracy in different scenarios only shows slight differences with a standard deviation of 0.08%. It demonstrates that our solution has a strong noise resistance capability, since it works in the high-frequency range. Then, we asked the participants to walk normally in the environment while entering the PIN codes. As shown in Fig. 19, we can find that TAPPASS can effectively mitigate interference induced by slow walking. Yet, for the brisk walking, the authentication performance drops significantly. That is because rapid movements may cause acoustic channel fluctuations and trigger more severe multi-path effects, hindering the authentication performance. In addition, during brisk walking, the user's hand positioning may change, causing TAPPASS to extract features from different areas of the hand, which can lead to inconsistencies in the biometrics.

*3) **Impact of Playback Volumes**:* Since the volume of the ultrasonic waves played by the mobile phone directly affects the signal intensity, it is necessary to investigate the impact of different volume levels on the authentication performance. We first conduct the cross-volume evaluation with four mobile phones (*i.e.*, Redmi Note12 Turbo, HUAWEI P60, HUAWEI Mate40 Pro, and Redmi K50 Pro). In particular, we leverage each phone to collect eight participants with four volume levels to construct a data subset. As for each mobile phone,

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

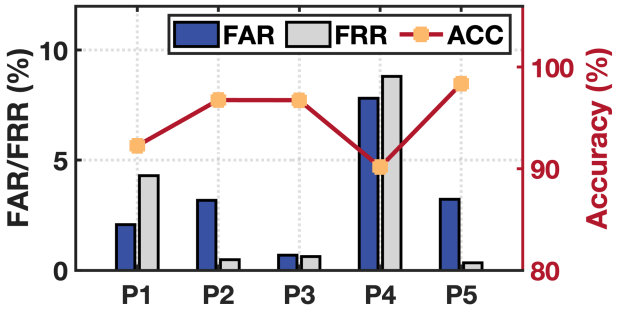


Fig. 18: Impact of passcode patterns on authentication performance.

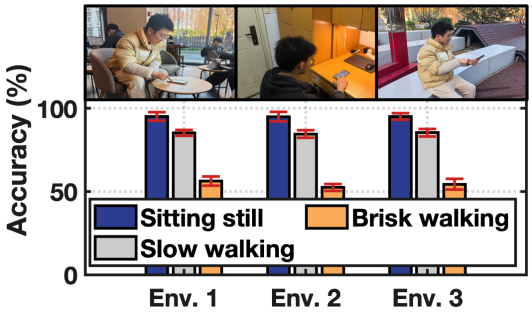


Fig. 19: Authentication accuracy of different scenarios and motion states.

the authentication model is trained using acoustic samples collected at a 50% volume level, and evaluated using acoustic samples collected at other volume levels. As shown in Tab. III, we can find that different volume levels on the same device have a negligible impact on authentication stability. Even with 25% of the volume, TAPPASS still achieves reliable authentication accuracy. This is because the screen interaction area of a mobile phone is limited, so even at low volumes, the acoustic sensing scope can completely cover the area.

*4) **Impact of Device Types**:* Furthermore, we also conduct a cross-device generalization evaluation to assess the impact of device diversity on authentication stability. Specifically, we use acoustic samples collected from Redmi Note12 to train the authentication model and use acoustic samples collected from other phones to assess cross-device generalization performance. As shown in Tab. IV, the authentication performance degrades significantly in the cross-device setting. This performance drop is primarily caused by hardware heterogeneity across devices, including differences in speaker and microphone placement, acoustic emission characteristics, and internal chassis structures. These factors influence the acoustic signal generation, propagation, and sensing processes, leading to distribution shifts in the extracted DCIR features across devices. Most importantly, this phenomenon is not unique to our work. Domain shifts caused by hardware heterogeneity have been widely recognized as a fundamental challenge in the acoustic sensing field and remain an open problem in the research community [39].

However, Tab. III demonstrates that TAPPASS can achieve stable authentication performance even among different volume levels with a fixed hardware configuration. In practical usage scenarios, users typically perform authentication on the same device they enrolled with, and biometric templates are not expected to transfer across different hardware platforms. Even if users change devices, they only need to re-enroll on the new device to provide a small number of enrollment samples, which is similar to widely deployed biometric solutions such as fingerprint or facial authentication on smartphones.

J. Attack Resistance Study

Based on the attack model described in Sec. III-A, a comprehensive experiment is conducted to study the attack resistance capability of TAPPASS in terms of passcode replay attack, behavior mimic attack, and audio replay attack.

*1) **Resistance to Passcode Replay Attack**:* In this attack scenario, we assume that the attacker has the knowledge of the victim's passcode sequence and enters it on the victim's smartphone to carry out the passcode replay attack. Specifically, we randomly select 10 participants as victims and the remaining 40 participants as attackers. As shown in Fig. 20, TAPPASS achieves EER of 2.78%, 2.53%, 2.04% across three passcode patterns (*i.e.*, P1, P2, and P3), respectively. Furthermore, we can clearly observe that the third passcode pattern (*i.e.*, "102030") has a lower EER. This is mainly because the spatial distance between the digits in the third passcode pattern is relatively large. Regardless, the low EER and high AUC values indicate that TAPPASS shows a strong resistance capability against passcode replay attacks.

*2) **Resistance to Behavior Mimic Attack**:* To conduct the behavior mimic attack, we still select 5 participants as legitimate users to register in TAPPASS. We select 10 participants with finger/hand sizes similar to those of the victims to act as attackers. To improve the success rate of attacks, we require these legitimate users to enter their passcodes multiple times and simultaneously record the video of their entry process. The attackers carefully watch recorded videos and imitate the tap behaviors of victims. The ROC curves are shown in Fig. 21. TAPPASS can achieve EER of 1.69%, 1.05%, and 1.25% across three passcode patterns, respectively. These results demonstrate that TAPPASS can effectively defend against mimic attacks.

*3) **Resistance to Artificial Finger Presentation Attacks**:* We assume that the attacker not only obtains the knowledge of the victim's passcode patterns and tap behavioral patterns, but also the physical parameters of the victim's dominant finger, including finger length, finger width profile, finger circumference, and finger thickness. Based on these parameters, the attacker can replicate the victim's finger using medical-grade silicone materials. This is because medical-grade silicone exhibits mechanical properties similar to those of human soft tissue and is capable of preserving fine surface details, making it suitable for replicating skin texture and contact characteristics. To conduct artificial finger presentation attacks, we select four participants as victims and use a vernier caliper and a measuring tape to measure the parameters of their index fingers. Then, the artificial fingers are made by the silicone molding technique and are leveraged for bypassing the authentication mechanism, as shown in Fig. 23. The ROC

This article has been accepted for publication in IEEE Transactions on Dependable and Secure 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/TDSC.2026.3710224

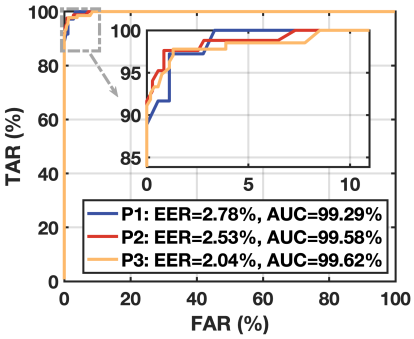


Fig. 20: ROC curves of passcode replay attacks.

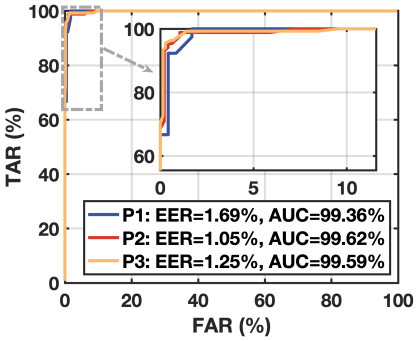


Fig. 21: ROC curves of behavior mimic attacks.

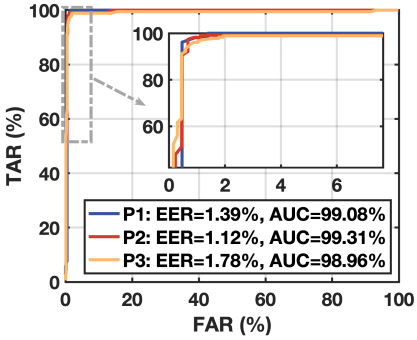


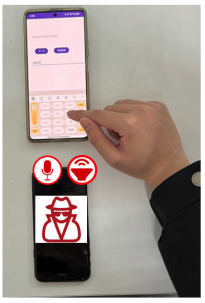
Fig. 22: ROC curves of presentation attacks.

No.	Joint Width			Joint Thickness			Joint Circumference			Phalange Length		
	MCP	PIP	DIP	MCP	PIP	DIP	MCP	PIP	DIP	Proximal	Middle	Distal
1	19.88	16.84	14.06	23.64	14.70	11.45	61.6	56.8	47.3	27.68	20.04	25.64
2	20.73	17.03	15.28	24.70	16.13	12.14	61.8	60.9	51.7	20.44	21.30	28.62
3	19.14	16.20	13.68	24.05	14.90	10.73	59.1	58.8	48.7	22.97	18.13	25.36
4	20.52	18.55	14.06	22.11	15.28	11.40	61.9	59.1	50.4	22.99	22.32	23.81

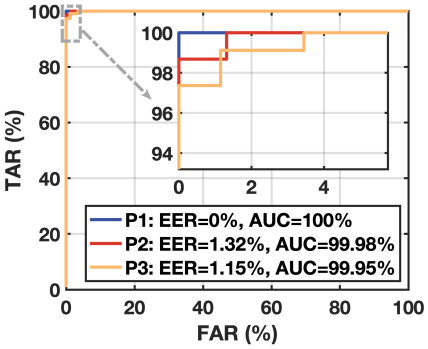
Fig. 23: The artificial finger is made based on the measurement parameters (millimeters) of the victim's index finger. MCP, PIP, and DIP denote the metacarpophalangeal, proximal interphalangeal, and distal interphalangeal joints, respectively. The proximal, middle, and distal phalanx lengths correspond to the distances from MCP to PIP, PIP to DIP, and DIP to the fingertip.

curves of presentation attacks are shown in Fig. 22. The EERs of the three passcode patterns (P1, P2, and P3) are 1.39%, 1.12%, and 1.78%, respectively, while the corresponding AUC values exceed 98.9% in all cases. The results demonstrate that TAPPASS can also effectively discriminate these attack attempts.

*4) **Resistance to Co-located Acoustic Replay Attacks**:* Lastly, we simulate co-located acoustic replay attacks, where the attacker attempts to bypass the authentication system by replaying spoofing acoustic samples. Specifically, the attacker first records acoustic signals generated during a genuine authentication session by placing a recording device (a smartphone) approximately 3 cm away from the victim's smartphone. Although such a short distance will make the victim be alert in reality, we still choose it for the attack resistance capability study. Then, the playback device (a smartphone) is also placed as close as possible to the victim's smartphone to ensure that the spoofing acoustic samples can be directly received by the smartphone's microphone with less attenuation, as shown in Fig. 24(a). The ROC curves of acoustics replay attacks are presented in Fig. 24(b). Surprisingly, TAPPASS can effectively identify replay acoustic attack attempts, with the EER close to 0 and the AUC close to 100%. These results indicate that TAPPASS presents an excellent attack detection performance. In practice, if the attacker gets very close to the victim as we set up in the attack scenarios, the victim will become alert. If the attacker and victim are further apart, the attack attempts will completely fail because ultrasounds attenuate very quickly in space.



(a) Attack setup.

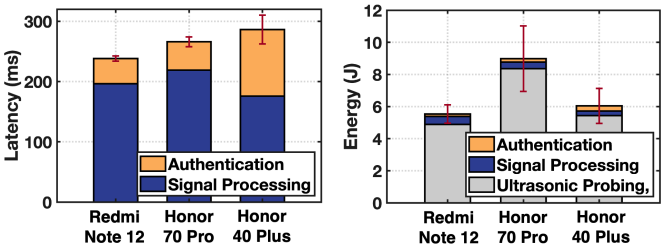


(b) ROC curves.

Fig. 24: Co-located acoustic replay attack setup and corresponding ROC curves.

K. *Latency and Energy Consumption Analysis*

As for COTS mobile phones, model complexity and energy consumption are key factors related to the practicality of the proposed authentication mechanism. Next, we perform runtime latency and energy consumption analysis on mainstream mobile phones to present the deployability of TAPPASS. It is noted that the frame-wise data augmentation approach designed in Sec. IV-C is used only during the enrollment phase and model training phase, and does not introduce runtime latency during the authentication process. We only focus on runtime latency induced by signal processing and authentication model inference. In practical usage, TAPPASS is triggered simultaneously when the PIN code is entered. The duration taken to enter the PIN code varies from person to person, but it generally does not exceed 3 seconds. Thus, we input a 3-second audio clip into TAPPASS to calculate the runtime latency. Fig. 25(a) demonstrates the runtime latency results of three mainstream mobile phones, *i.e.*, Redmi Note



(a) Runtime latency. (b) Energy consumption.

Fig. 25: Runtime latency and energy consumption evaluation on three mobile phones.

12 Turbo, Honor 70 Pro, and Honor Play 40 Plus. We can find that even the maximum latency does not exceed 300 ms. Real-time response indicates that TAPPASS can be seamlessly integrated into PIN code input as a secondary authentication mechanism to further improve security. In addition to response time, energy consumption is also a critical factor directly affecting the deployability of TAPPASS on mobile phones. The energy consumption of TAPPASS mainly comes from ultrasonic probing, signal processing, and authentication inference. Specifically, we measure the energy consumption with the help of *BatteryManager* Android API that can provide detailed information about a device's battery status. As shown in Fig. 25(b), we can find that a large portion of the energy consumption is induced by ultrasonic probing, while the energy consumption for signal processing and authentication inference is only a few hundred millijoules (mJ). Since ultrasonic probing is performed only during PIN entry, the ultrasound emission is transient rather than continuous. Combined with lightweight inference, the overall energy consumption remains low. For example, considering a fully charged Honor Play 40 Plus smartphone equipped with a 6000 mAh battery, the device can theoretically support over $\frac{3.7\ V*6000\ mAh*3600\ s}{1000*6\ J} = 13320$ authentication sessions on a single charge.

VI. Discussion and Future Work

While TAPPASS demonstrates encouraging results for mobile authentication, several limitations remain that warrant further investigation.

(1) Threat Evaluation of Advanced Learning-based Attacks. We have validated that TAPPASS can effectively resist four physical attacks in Sec. V-J. With the development of deep learning technology, more advanced attack techniques have been studied, like signal synthesis attacks and generative replay attacks. Yet, we believe that launching these learning attacks is highly non-trivial in practice. Because acoustic echo profiles contain rich multi-path information induced by hand reflection, accurately synthesizing such signals using deep learning models requires a large amount of victim-specific training data in order to learn these complex acoustic reflection patterns. In practice, obtaining such a dataset is difficult for attackers. Even if an attacker attempts to access the microphone of the victim's smartphone, existing mobile operating systems impose strict permission and privacy controls, and it is difficult to collect large amounts of high-quality acoustic echo data without being noticed by the victim.

(2) Towards Transferable Biometrics. Ideally, biometric features should remain stable across different usage conditions, such as varying handheld postures and heterogeneous device hardware. The biometrics in TAPPASS are jointly determined by the user's hand characteristics and the device-specific acoustic channel characteristics. While such device-bound biometrics can enhance the security of the authentication mechanism, they limit cross-device generalization and reduce portability in multi-device usage scenarios. This highlights the need for research on transferable biometrics, which aims to extract user-specific features that remain reliable across different devices and usage scenarios, improving both usability and scalability in practical deployments. Future work could focus on learning more transferable biometric representations that remain discriminative across different grips and devices, further enhancing the flexibility and scalability in real-world scenarios.

(3) Suboptimal Performance under Challenging Conditions. The behavioral patterns and passcode patterns shift the distribution of tap biometrics. Achieving reliable authentication under such challenging conditions is fundamentally difficult. Despite these challenges, TAPPASS substantially improves the authentication performance compared with prior acoustic-based PIN authentication systems. Nevertheless, we fully recognize that the current system performance still leaves room for improvement before large-scale deployment in highly usability-sensitive scenarios. The current design prioritizes robustness against dynamic behavioral patterns as well as passcode patterns and resistance against multiple attack models. In future work, we plan to further optimize the trade-off between security and usability through adaptive threshold selection, confidence-aware multi-attempt verification, and more generalized biometric representation learning.

VII. Conclusion

This paper presents TAPPASS, an acoustic-based second-factor authentication solution that enhances PIN security on commodity smartphones by exploiting hand-reflected ultrasonic signals during tap interactions. A complete pipeline is designed to extract the discriminant and consistent biometrics for user authentication, effectively addressing the limitations of prior studies. Extensive experiments are conducted to validate the effectiveness, robustness, and usability. Beyond technical design, our findings also provide practical insights into secure PIN selection, highlighting TAPPASS as a deployable solution for real-world mobile authentication.

References

[1] W. Melicher, D. Kurilova, S. M. Segreti, P. Kalvani, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek, "Usability and security of text passwords on mobile devices," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 527–539.

[2] W. Meng, D. S. Wong, S. Furnell, and J. Zhou, "Surveying the development of biometric user authentication on mobile phones," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1268–1293, 2014.

[3] Q. Yan, J. Han, Y. Li, J. Zhou, and R. H. Deng, "Designing leakage-resilient password entry on touchscreen mobile devices," in *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, 2013, pp. 37–48.

[4] T. Kwon and J. Hong, "Analysis and improvement of a pin-entry method resilient to shoulder-surfing and recording attacks," *Ieee transactions on information forensics and security*, vol. 10, no. 2, pp. 278–292, 2014.

[5] A. Garbuz, A. Epishkina, and K. Kogos, "Continuous authentication of smartphone users via swipes and taps analysis," in *2019 European Intelligence and Security Informatics Conference (EISIC)*. IEEE Computer Society, 2019, pp. 48–53.

[6] Y. Chen, G. Liu, L. Yu, H. Kang, L. Meng, and T. Wang, "Tbauth: A continuous authentication framework based on tap behavior for smartphones," *Expert Systems with Applications*, vol. 264, p. 125811, 2025.

[7] T. Van Nguyen, N. Sae-Bae, and N. Memon, "Draw-a-pin: Authentication using finger-drawn pin on touch devices," *computers & security*, vol. 66, pp. 115–128, 2017.

[8] S. Khan, C. Devlen, M. Manno, and D. Hou, "Mouse dynamics behavioral biometrics: A survey," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–33, 2024.

[9] M. Zhou, Q. Wang, X. Lin, Y. Zhao, P. Jiang, Q. Li, C. Shen, and C. Wang, "Presspin: Enabling secure pin authentication on mobile devices via structure-borne sounds," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 2, pp. 1228–1242, 2022.

[10] M. Zhou, Y. Zhou, S. Su, Q. Wang, Q. Li, S. Hu, C. Yu, and Z. Li, "Fingerpattern: Securing pattern lock via fingerprint-dependent friction sound," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 7210–7224, 2023.

[11] J. Lee, I. Kim, S. Oh, and H. Kim, "Tapin: Reinforcing pin authentication on smartphones with tap biometrics," *IEEE Transactions on Mobile Computing*, 2024.

[12] Y. Wu, S. Bai, R. Lv, X. Gong, X. Liu, L. Ding, and Y. Chen, "Fingervib: Fortifying acoustic-based authentication with finger vibration biometric on smartphone," *IEEE Transactions on Information Forensics and Security*, 2025.

[13] Y. Chen, T. Ni, W. Xu, and T. Gu, "Swipepass: Acoustic-based second-factor user authentication for smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–25, 2022.

[14] H. Chen, F. Li, W. Du, S. Yang, M. Conn, and Y. Wang, "Listen to your fingers: User authentication based on geometry biometrics of touch gesture," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–23, 2020.

[15] K. Krombholz, T. Hupperich, and T. Holz, "Use the force: Evaluating {Force-Sensitive} authentication for mobile devices," in *Twelfth symposium on usable privacy and security (SOUPS 2016)*, 2016, pp. 207–219.

[16] A. S. Rathore, W. Zhu, A. Daiyan, C. Xu, K. Wang, F. Lin, K. Ren, and W. Xu, "Sonicprint: A generally adoptable and secure fingerprint biometrics in smart devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 121–134.

[17] J. Chauhan, Y. D. Kwon, P. Hui, and C. Mascolo, "Contauth: Continual learning framework for behavioral-based user authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–23, 2020.

[18] F. Han, P. Yang, S. Yan, H. Du, and Y. Feng, "Breathsign: Transparent and continuous in-ear authentication using bone-conducted breathing biometrics," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[19] F. Han, P. Yang, Y. Feng, H. Du, and X.-Y. Li, "Exploring earable-based passive user authentication via interpretable in-ear breathing biometrics," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 15 238–15 255, 2024.

[20] M. Khamis, M. Hassib, E. v. Zezschwitz, A. Bulling, and F. Alt, "Gazetouchpin: protecting sensitive data on mobile devices using secure multimodal authentication," in *Proceedings of the 19th acm international conference on multimodal interaction*, 2017, pp. 446–450.

[21] Z. Li, M. Li, P. Mohapatra, J. Han, and S. Chen, "itype: Using eye gaze to enhance typing privacy," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[22] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 73–87.

[23] D. Liu, Q. Wang, M. Zhou, P. Jiang, Q. Li, C. Shen, and C. Wang, "Soundid: Securing mobile two-factor authentication via acoustic signals," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 2, pp. 1687–1701, 2022.

[24] X. Xu, J. Yu, Y. Chen, Q. Hua, Y. Zhu, Y.-C. Chen, and M. Li, "Touchpass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.

[25] Y. Wu, W. Song, C. T. Chou, J. Hu, and W. Hu, "Handid: Towards unobtrusive gesture-independent user authentication on smartphones using vibration-based hand biometrics," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 2, pp. 1–27, 2025.

[26] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 321–336.

[27] C. Kong, K. Zheng, Y. Liu, S. Wang, A. Rocha, and H. Li, "$M^3$3fas: An accurate and robust multimodal mobile face anti-spoofing system," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 6, pp. 5650–5666, 2024.

[28] H. Chen, C. Gu, L. Xu, R. Tan, S. He, and J. Chen, "Listen to your face: a face authentication scheme based on acoustic signals," *ACM Transactions on Sensor Networks*, vol. 21, no. 1, pp. 1–23, 2025.

[29] Z. Xu, T. Liu, R. Jiang, P. Hu, Z. Guo, and C. Liu, "Aface: Range-flexible anti-spoofing face authentication via smartphone acoustic sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–33, 2024.

[30] J. Huang, Y. Feng, F.-Q. Cui, X. Zhang, Z. Liu, X. Liu, J. Liu, F. Zhang, and M. Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–15, 2026.

[31] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "Echohand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 2022, pp. 2931–2945.

[32] C. Wu, K. He, J. Chen, R. Du, R. Yan, and Z. Zhao, "High accuracy and presentation attack resistant hand authentication via acoustic sensing for commodity mobile devices," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 5, pp. 5231–5247, 2025.

[33] Y. Yang, Y. Wang, Y. Chen, and C. Wang, "Echolock: Towards low-effort mobile user identification leveraging structure-borne echos," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, 2020, pp. 772–783.

[34] L. Lu, J. Yu, Y. Chen, and Y. Wang, "Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–24, 2020.

[35] 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*. IEEE, 2018, pp. 1466–1474.

[36] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: A mmwave-based unmediated vocal sensing system for secure authentication," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 589–606, 2021.

[37] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 278–291.

[38] H. Halim, E. G. Chekole, D. Reijsbergen, and J. Zhou, "Blowprint: Blow-based multi-factor biometrics for smartphone user authentication," in *European Symposium on Research in Computer Security*. Springer, 2025, pp. 169–189.

[39] C. Cai, R. Zheng, and J. Luo, "Ubiquitous acoustic sensing on commodity iot devices: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 432–454, 2022.

[40] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4570–4580.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

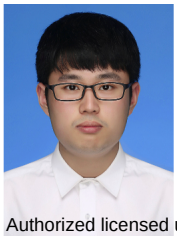
[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[43] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
[44] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems Man Cybernetics-Systems*, vol. 9, no. 1, pp. 62–66, 1979.
[45] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
[46] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

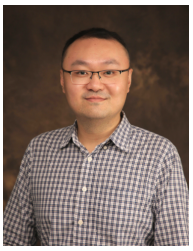
Feiyu Han is an associate professor at School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology. Before that, he received his Ph.D. degree at School of Computer Science and Technology, University of Science and Technology of China, in 2024. He received his B.S. degree at School of Computer Science and Engineering from Nanjing University of Science and Technology, in 2019. His research interests include wireless sensing and mobile computing.

Zelei Wang is currently an undergraduate student at School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, where he is pursuing a Bachelor's degree. His research interests include IoT security, wireless sensing, and mobile computing.

Yuanhao Feng is now a Postdoctoral Fellow in the Department of Computing, The Hong Kong Polytechnic University. He received his Ph.D. degree from University of Science and Technology of China in 2023. He is an outstanding graduate of the School of Computer Science in USTC 2023. He has published many top journal and conference papers, such as MobiCom, INFOCOM, and TMC. His research interest is AIoT, which includes wireless sensing and communication.



Jinyang Huang is a lecturer at the School of Computer Science and Information Engineering, Hefei University of Technology and the Secretary-General of the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligence Machine. He obtained his Ph.D. in Computer Science and Technology from the School of Cyberspace Security, University of Science and Technology of China in 2022. His research interests include multimodal perception, human-computer interaction, wireless security, and signal processing.



Meng Li (Senior Member, IEEE) is a Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He obtained his Ph.D. in Computer Science and Technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He was a Post-Doc Researcher at the University of Padua, Italy, where he worked with Prof. Mauro Conti at the SPRITZ research group. He also conducted Post-Doc research under the supervision of Prof. Fabio Martinelli at CNR, Italy, and was a Joint Ph.D. student supervised by Prof. Xiaodong Lin at the University of Waterloo and Wilfrid Laurier University, Canada. His research interests include security, privacy, applied cryptography, blockchain, TEE, and Internet of Vehicles. He is a Senior Member of IEEE, CIE, CIC, and CCF. He is an Associate Editor for TIFS, TDSC, and TNSM. He has served as a TPC member for conferences, including ICDCS, Inscrypt, ICICS, and TrustCom. He is the recipient of 2024 IEEE HITC Award for Excellence (Early Career Researcher) and 2025 IEEE TCSVC Rising Star Award. He was selected into the IEEE Computer Society Computing's Top 30 Early Career Professionals for 2025.

Panlong Yang (Senior Member, IEEE) is now a professor at School of Computer Science and Engineering from Nanjing University of Information Science and Technology. Before that, he received his B.S. degree, M.S. degree, and Ph.D. degree in communication and information system from Nanjing Institute of Communication Engineering, P.R. China, in 1999, 2002, and 2005, respectively. His research interests include wireless mesh networks, wireless sensor networks, and cognitive radio networks. He is a senior member of the IEEE Computer Society.

Zhangjie Fu (Member, IEEE) is currently a Professor and the Executive Dean with the School of Computer and Cyberspace Security, Nanjing University of Information Science and Technology, Nanjing, China, where he is also the Executive Director of the Engineer Research Center of Digital Forensic, Ministry of Education. From 2015 to 2016, he was a Visiting Scholar with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA. He received the Ph.D. degree in computer science from the College of Computer, Hunan University, Changsha, China, in 2012. His research interests include cloud and outsourcing security, digital forensics, and network and information security.