Beyond the Pixel World: A Novel Acoustic-based Face Anti-Spoofing System for Smartphones

Chenqi Kong, Kexin Zheng, Shiqi Wang, *Senior Member, IEEE*,
Anderson Rocha, *Senior Member, IEEE*, and Haoliang Li, *Member, IEEE*

***Abstract*—2D face presentation attacks are one of the most notorious and pervasive face spoofing types, which have caused pressing security issues to facial authentication systems. While RGB-based face anti-spoofing (FAS) models have proven to counter the face spoofing attack effectively, most existing FAS models suffer from the overfitting problem (i.e., lack generalization capability to data collected from an unseen environment). Recently, many models have been devoted to capturing auxiliary information (*e.g.*, depth and infrared maps) to achieve a more robust face liveness detection performance. However, these methods require expensive sensors and cost extra hardware to capture the specific modality information, limiting their applications in practical scenarios. To tackle these problems, we devise a novel and cost-effective FAS system based on the acoustic modality, named Echo-FAS, which employs the crafted acoustic signal as the probe to perform face liveness detection. We first propose to build a large-scale, high-diversity, and acoustic-based FAS database, Echo-Spoof. Then, based upon Echo-Spoof, we propose designing a novel two-branch framework that combines the global and local frequency clues of input signals to distinguish inputs, live vs. spoofing faces accurately. The devised Echo-FAS comprises the following three merits: (1) It only needs one available speaker and microphone as sensors while not requiring any expensive hardware; (2) It can successfully capture the 3D geometrical information of input queries and achieve a remarkable face anti-spoofing performance; and (3) It can be handily allied with other RGB-based FAS models to mitigate the overfitting problem in the RGB modality and make the FAS model more accurate and robust. Our proposed Echo-FAS provides new insights regarding the development of FAS systems for mobile devices.**

***Index Terms*—Face Anti-spoofing, Acoustic signal, Multi-modality, Mobile applications.**

I. Introduction

USER authentication systems have been widely deployed, ranging from phone unlocking to financial payment systems. Generally speaking, ubiquitous user authentication methods on smartphones include PINs, fingerprint, iris, and face authentication. PINs are the most traditional authentication method, but users can easily forget them, and they are vulnerable to shoulder-surfing attacks [1]. Fingerprint and iris-based authentication methods demand the integration of the fingerprint and iris sensor [2], [3] to authentication devices, and they may fail to work properly in some common conditions (e.g., wet finger). Face recognition can provide a more user-friendly authentication mechanism than the techniques above. The face is the dominant biometric trait of a person, a unique FaceID, and a vehicle itself of non-verbal but powerful messages [4]. Therefore, face authentication systems are becoming increasingly pervasive due to their promising recognition performance and user-friendly usage. As a result, we have seen various critical applications using face recognition, such as device unlocking, online shopping, online banking, and account log-in based on smartphones.

Despite the demonstrated success of developed face authentication systems on smartphones, the potential menace of face presentation attacks has raised pressing concerns. This work mainly focuses on 2D face presentation attacks (PA) since 2D face PA can be easily launched by the attacker with the target subject's high-quality face images. In turn, 3D mask face PA demands advanced fabrication systems capturing the 3D geometric and texture information of the target person's face, which requires high costs. Thus, 3D PA is not as pervasive as 2D PA [5]. 2D face presentation attacks, including print attacks and video-replay attacks, have caused severe security issues to face authentication systems. In addition, an attacker without professional skills can easily acquire the target user's face data and deploy it to hack the face authentication system, resulting in critical disconcerting security problems.

Various traditional Face Anti-Spoofing (FAS) technologies have been proposed to defend against presentation attacks. Most of them focus on extracting traditional hand-crafted features including histograms, gradients and texture, such as [5]–[10]. Thanks to recent advances in artificial intelligence and deep learning, many deep learning-based FAS solutions [11]–[18] have been proposed to learn representative features between live and spoofing faces directly from available training data. While some signs of progress have been achieved, most existing models suffer substantial detection performance drops when the training and testing data distributions are misaligned (domain shift problem). Regarding real-world application scenarios with various uncontrolled environmental variables (e.g., acquisition devices and illumination changes), it is non-trivial to develop a FAS model with high generalization capability. Some recently proposed methods [19]–[23] capture depth, infra-red, or thermal information of input faces as auxiliary cues. However, these models might require expensive sensors in practical scenarios and bring extra hardware costs when deployed. In turn, some recent work [24]–[26] have demonstrated

C. Kong, K. Zheng, and S. Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. (email: cqkong2-c@my.cityu.edu.hk; kexizheng3@cityu.edu.hk; shiqwang@cityu.edu.hk).

H. Li is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China. (email: haoliang.li@cityu.edu.hk).

A. Rocha is with the Artificial Intelligence Lab. (Recod.ai) at the University of Campinas, Campinas 13084-851, Brazil (e-mail: anderson.rocha@ic.unicamp.br), URL: http://recod.ai

H. Li is the corresponding author.

that applying adversarial attack techniques on fake input faces can easily fool FAS models, causing severe security concerns. It is only natural to ask: is there a more cost-effective, secure, and robust FAS system that can effectively cope with the issues mentioned above?

Inspired by the recent success of acoustic signals [27]–[31] that can efficiently capture reliable biometric information of users for various mobile-oriented applications, we devise a secure and robust acoustic-based FAS system in this work henceforth referred to as *Echo-FAS*.

As shown in Fig. 1, unlike RGB-based systems leveraging front cameras to capture the input vision data, the devised acoustic-based system uses the available speaker to emit a customized crafted acoustic signal and leverages the microphone to collect the reflected recording that has been modulated by the input live/spoof face. Furthermore, the proposed method extracts feature representatives from the captured recording with rich geometric information to distinguish the input query between genuine and spoofing faces.

Our Echo-FAS only requires one front speaker and microphone as the data collection sensors, which are ubiquitous on most smartphones. Echo-FAS is cost-effective and can be readily deployed on commercial mobile devices in a plug-and-play fashion. Compared with the RGB data that adversarial attack techniques can quickly attack [24]–[26], the designed acoustic signal is emitted by the system itself. Therefore, it is more secure and much more difficult to attack or manipulate even by expert attackers. Our paper shows that the acoustic-based FAS system can achieve an accurate and robust face liveness detection performance as the collected acoustic recording carries much face geometric information, which is largely ignored in the RGB-based FAS system. Last but not least, our experiment further demonstrates that the proposed Echo-FAS system can be flexibly assembled with the RGB-based FAS model to mitigate the RGB domain gaps.

We first propose to collect a large-scale acoustic-based database Echo-Spoof by exploiting different smartphones based on the consideration of diverse environmental variables, including device, distance, ambient noise, and pitch, accommodating applications for practical scenarios. Specifically, our Echo-Spoof database includes more than 250,000 acoustic signal segments collected from 30 volunteers. Echo-Spoof is the largest acoustic-based FAS database so far, to the best of our knowledge. We also design a novel framework for the acoustic-signal-based FAS problem. Extensive experimental results demonstrate the effectiveness and robustness of the proposed Echo-FAS system. The main contributions of this work can be summarized as follows:

- We built a large-scale high-diversity acoustic-based FAS database (Echo-Spoof) from 30 volunteers. Echo-Spoof includes more than 250,000 signal segments with four environmental variables: device, distance, ambient noise, and pitch.
- We propose a novel two-branch framework, smartly fusing the global and local frequency clues, to accurately detect face liveness. Extensive ablation experiments demonstrate the effectiveness of the designed framework.

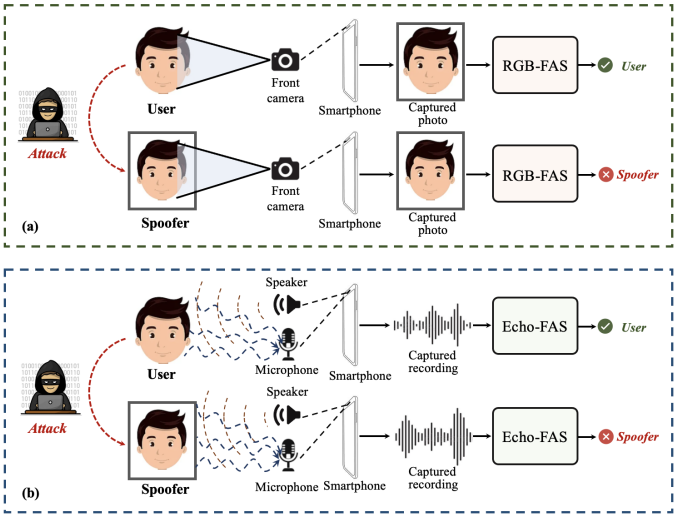

Fig. 1. Illustration of (a). the traditional RGB-based FAS system and (b). the proposed acoustic-based system.

- A benchmark based on the Echo-Spoof database, with extensive experiments showing that the proposed Echo-FAS system achieves almost 99% AUC performance. Echo-FAS consistently achieves the best liveness detection results under various experimental settings, demonstrating its high robustness. Finally, Echo-FAS can be flexibly assembled with RGB-based FAS models to mitigate RGB domain problems, and the use of acoustic data significantly boosts the AUC score from 93.48 to 99.08 in the multi-modality experiment.

In the remainder of this paper, Section II introduces related work of face anti-spoofing and biometric applications of acoustic signals. Section III details the data collection process. Section IV discusses the philosophy of the designed acoustic signal. Section V presents our proposed two-branch framework to the problem. Section VI evaluates comprehensive algorithms on the proposed benchmark database Echo-Spoof and provides rigorous ablation studies. Section VII evaluates the effectiveness of Echo-FAS allied with RGB-based FAS models. Finally, section VIII concludes the paper and presents possible future work.

II. Related Work

A. Face anti-spoofing methods

Existing face anti-spoofing methods can be generally divided into traditional or handcrafted methods and data-driven (deep learning-based) methods. Most traditional approaches capture texture cues to distinguish genuine faces from spoofing ones. Hand-crafted descriptors such as LBP [6], [7], HOG [8], SIFT [5], and SURF [9] are extracted and associated with a classifier to perform FAS. Thanks to the advances in deep learning in recent years, many learning-based approaches have been developed, leading to much more reliable FAS performance. Researchers leverage auxiliary supervisions to improve FAS models and overcome overfitting problems. Jourabloo *et al.* [32] extract spoof noise patterns to differentiate live faces and attacks. Nowara *et al.* [33] and Liu *et al.* [34] explore rPPG

(remote photoplethysmography) information as discriminative clues, while [18] combines both rPPG and depth maps to conduct learning-based anti-spoofing.

Although learning-based methods have achieved high detection rates, performance drops drastically under various cross-domain scenarios, limiting their applications in real-world scenarios. Many recently proposed models seek to use auxiliary modality information to improve the detection models' generalization. For example, depth maps [20]–[23] and thermal infrared images [20], [35]–[37] have been used in FAS models and achieved promising liveness detection performance. However, these models typically demand expensive sensors to capture the specific modality information. Thus, they cannot be ubiquitously equipped on commodity smartphones and other devices. In contrast, the Echo-FAS proposed here only needs one speaker and microphone on commercial-off-the-shelf (COTS) smartphones to collect the acoustic modality data and perform a secure and robust anti-spoofing. In other words, the proposed method leverages existing hardware on typical smartphones and can be flexibly allied with RGB-based models to improve FAS RGB-based detection methods.

B. Biometric applications of acoustic signals

Thanks to the ubiquitous availability of speakers and microphones on COTS smartphones, acoustic-based sensing has been recently adopted in mobile applications to capture users' information for various purposes. The basic idea is employing the speaker to emit a custom-tailored signal and leveraging the microphone to collect the modulated signal, and the captured recording can reflect abundant user's biometric information such as gesture [27], [28], and face structure [29]–[31].

Biometric applications can be generally categorized into dynamic and static applications. Most dynamic-related works investigate phase and frequency shifts to extract modulated patterns that indicate specific movements. For example, Finger IO [38], LLAP [39], Strata [28], and UltraGesture [27] achieve remarkable tracking performance in acoustic-based device-free gesture tracking on COTS devices. Following similar schemes, silentKey [40] and Endophasia [41] recognize mouth commands through ultrasound-sensing. LVID [42] and LiPass [43] detect unique mouth movement patterns for user authentication. Chen *et al.* [44] and EchoLock [45] measure touch gesture and hand geometry to perform user authentication.

Unlike dynamic biometric applications based on unique behavioral biometrics that require user actions, static biometric applications detect the user's biometric structure in a more user-friendly way. Face authentication and face anti-spoofing are two typical static application scenarios. EchoPrint [29] adopted Frequency Modulated Continuous Wave (FMCW) [46] and visual facial features for secure authentication. However, face spoofing detection is largely ignored in Echoprint, leaving a severe security concern in smartphone authentication systems. EchoFace [30] extracts the target signals from two reflected signal segments collected by the earpiece and bottom microphones to perform spoofing attack detection. However, its performance deteriorates when the bottom microphone cannot sense the signal well in many practical scenarios. RFace [31] uses a radio frequency identification tag array to perform face authentication and face anti-spoofing synchronously. Although RFace achieves a desired authentication success rate and a spoofing attack detection performance, its proposed work frequency (920.625MHz) cannot be applied on existing smartphones [47] and other commercial devices due to hardware limitations.

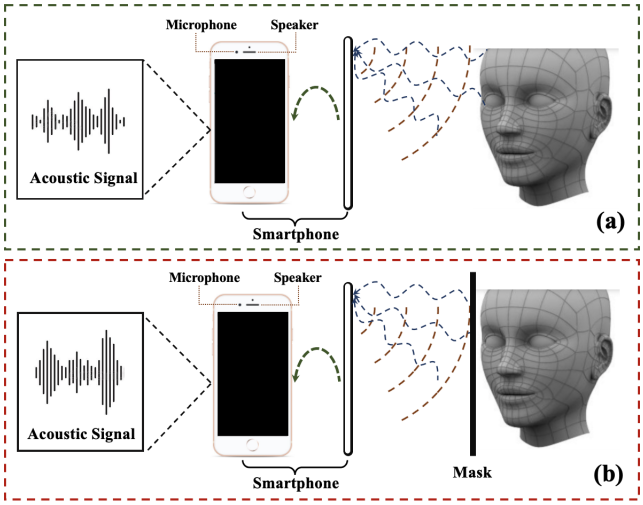

Fig. 2. Data collection processes of (a). genuine/live and (b). spoofing faces. The smartphone emits a custom-tailored inaudible signal via earpiece speaker to illuminate live/spoofing faces, and the reflected acoustic signal is recorded by the microphone.

Differently from existing work in the prior art, we propose Echo-FAS, which designs an acoustic signal as the probe and can be flexibly deployed on most commodity devices in a plug-and-play fashion. In addition, Echo-FAS takes advantage of data-driven methods such as CNN and transformer architectures for conducting more secure and robust face liveness detection.

III. Dataset Acquisition

We first introduce the motivation of an Echo-Spoof dataset collection where the database has the following desired properties:

- Acoustic-based: Unlike most existing FAS databases that collect face images and videos, the proposed Echo-Spoof database collects the acoustic signals that contain rich geometric information of user faces to perform face liveness detection in a more privacy-preserving way;
- Large-scale: The Echo-Spoof database is collected from 30 volunteers (15 females and 15 males), containing more than 250,000 acoustic signal segments;
- High-diversity: To accommodate the unlimited variations of environmental conditions in real-world application scenarios, we set various environmental variables in the data collection process, such as collection distances, devices, ambient noise levels, and pitches.

We illustrate the Echo-FAS dataset collection process in Fig. 2, where Fig. 2 (a) and Fig. 2 (b) respectively, illustrate the live/genuine face and spoofing face data collection processes. The earpiece speaker first emits the designed acoustic

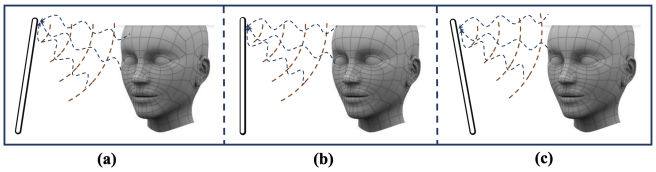

Fig. 3. Illustration of different collection pitches between the device and live/spoofing face. (a). -10 degrees; (b). 0 degree; (c). +10 degrees.

signal; then, the emitted signal will be modulated by the surface of the live/spoof face. The reflected signal carrying rich geometric information of the live/spoof face will be collected by the top microphone of the smartphone. We have obtained the approval of the human ethics application from the Human and Artefacts Ethics Sub-Committee before data collection.

To handle complex attack environments and generalize to real-world scenarios, we build a large-scale and high-diversity acoustic database, Echo-Spoof. We consider the following application variables during the database collection process:

Device. We conduct data collections on four Android smartphone devices: `Samsung s9`, `Samsung s21`, `Samsung edge note`, and `Xiaomi Redmi7`. Generally speaking, different smartphones have different hardware conditions due to the imperfect manufacturing of sensors, including both speakers and microphones [48]. These physical differences will introduce diverse non-uniform noise patterns, and the signals from different smartphones tend to have different data distributions. Thus, it is non-trivial to investigate the robustness of the proposed model on different data collection devices.

Distance. The distance influences FAS performance as the signal-to-noise ratio (SNR) of the received acoustic signal tends to be lower as the distance becomes larger. According to our investigations, the comfortable distance from the user's nose to the phone in people's daily usage is around 25-45cm. We set three data collection distances of 25cm, 35cm, and 45cm during the data collection process.

Ambient Noise. Ambient noise is also a key factor impacting FAS performance as it degrades the SNR of the received acoustic signal. Moreover, ambient noise ubiquitously exists in people's daily usage, so our dataset should consider this aspect. We set three ambient noise levels by controlling another device to play audios (*e.g.*, songs and BBC news) in different volumes. We install a noise detector APP on the data collection smartphone to precisely monitor the ambient noise levels. In this dataset, three noise levels have been set as 40 dB, 60 dB, and 70 dB, corresponding to quiet, little noisy, and very noisy environments in real-world scenarios.

Pitch. To accommodate diverse usage scenarios, we further introduce different *pitches* in our dataset. Pitches affect the collected geometric information because relative positions between face regions and the phone have altered. Relative pitch angles between the smartphone and human face are set as -10 degrees, 0 degrees, and +10 degrees, corresponding to different holding habits of users, as illustrated in Fig. 3.

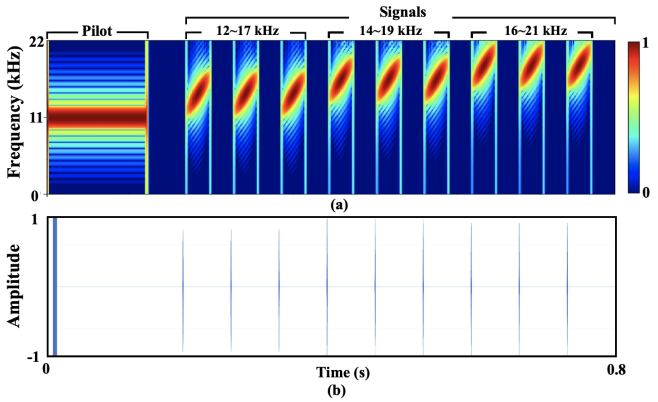

Fig. 4. Illustration of designed signal in (a). frequency and (b) time domain. The pilot is employed for the synchronization between the speaker and microphone. Three chirp signals with different frequency sweep ranges repeat three times, and the overall signal cover the frequency sweep from 12 to 21 kHz. (The frequency spectrogram has been scaled for better visualization.)

IV. Acoustic Signal Design for FAS

A. Signal Designing

Previous acoustic signal designs for biometric applications [29]–[31] suffer various weaknesses such as: requiring extra sensors, low SNR, and poor user experience. Therefore, developing a new acoustic signal is necessary to promise a better user experience and higher liveness detection performance with commodity hardware.

We illustrate our designed signal in Fig. 4, where the whole signal duration is around 0.8 seconds. The sample rate of the designed signal is 44.1 kHz as it is the most commonly supported for Android phones [47]. The highest frequency that smartphones can sense is around 22 kHz. Therefore, we add a 250-sample 11.025 kHz pilot before the signals and set the interval between the pilot and the first chirp as 8,000 samples (*i.e.*, ~0.18s). The following signals, including nine chirps with three different frequency sweep ranges, covers the frequency range from 12 to 21 kHz. The signal chirp at each group covers the corresponding frequency range in a linearly-increasing manner and repeats three times in the final emitted signal. Each chirp consists of 60 samples, and the interval between two chirps is designed as 3000 samples.

B. Signal analysis

The design of the signal considers some fundamental properties. First, the signal should carry rich and distinct geometric information from the target face region for high-quality anti-spoofing. Second, it should be reasonably robust to ambient noise and support accurate localization of target face reflection. Lastly, the emitted signal should produce a minimum annoyance to the human ear to ensure a good user experience. To fulfill the properties above, we design the signal mainly considering the following aspects: signal waveform and frequency, chirp duration and interval, and annoyance control.

Signal waveform and frequency. Signal waveform drives signal processing methods and further affects the final face feature extraction results. Therefore, inspired by EchoPrint [29], we adopt a continuous waveform with a linearly-increasing

frequency based on Frequency-Modulated Continuous Wave (FMCW) [46] technique, which is widely used in the radar prior art for distance measurement.

Since distances between different parts of the face and the smartphone are different, the received signal from the face region is a combination of multiple echoes reflected by different regions of the face (*e.g.*, nose and mouth) with different time delays and phase variations, which represents the unique geometry of the face region.

In FMCW, the echo separating resolution depends on the bandwidth of the chosen signal, which is set as 5 kHz in our design. For a minimum measurable frequency shift δf, the corresponding time interval δT can be represented as $\frac{\delta f}{k}$, where k is the slope of the signal chirp. Thus, the resolution that FMCW can separate mixed echoes can be calculated as follow:

$$R_d = \frac{v \cdot \delta T}{2} = \frac{v \cdot \delta f}{2k} = \frac{v}{2BW} = \frac{343m/s}{2 * 5kHz} = 3.43cm \quad (1)$$

where v is the speed of sound in the air.

Signal frequency affects robustness to ambient noise, annoyance control, and feature extraction quality. We carefully chose our signal frequency range to produce minimum annoyance to the users, considering the audible frequency range of humans. In terms of robustness, the signal frequency should be distinguishable from ambient noise frequency to remove ambient noise using a high-pass filter. In addition, it is essential to ensure commercial smartphones can play the signal with enough energy carried at each frequency for high-quality face feature extraction.

The audible frequency range of human beings is approximately 20 Hz to 20 kHz, and the upper limit for adults is 15-17 kHz on average [49]. According to Chen *et al.*, [29], the ambient noise frequency is usually under 8 kHz. High frequency is desirable for annoyance control and ambient noise removal. However, our preliminary experiments have shown that the emitted signal performs poorly when the frequency range exceeds 20 kHz due to mobile hardware limitations. To balance the trade-off between annoyance control and accurate face anti-spoofing, we set our target frequency range as 12 to 21 kHz and divided it into 12 - 17 kHz, 14 - 19 kHz, and 16 - 21 kHz.

While the previous design on the signal waveform and frequency focused on improving user experience and feature extraction quality, it is challenging to locate target face reflection signals in the captured recording accurately. The reason is that the delays between the microphone and speaker are not consistent across smartphones [50].

To address this issue, we adopt an 11.025 kHz continuous pilot signal at the beginning of signal emission, as Fig. 4 shows. Its frequency is distinguishable from both our feature extraction signal (12-21 kHz) and the environment noise (<8 kHz). By calculating the cross-correlation between the pilot signal and the captured recording, we can locate the beginning of signal emission in the final recording and perform further signal processing steps to extract the target face echoes.

Chirp duration and interval. The chirp duration is the sample length of the chirp, and the chirp interval is the time between the emission of two consecutive chirps. According to the nine-chirp design of our emitted signal, we will also get nine chirps in the captured recording. Intuitively, each chirp contains three components: the direct transmission signal, target face echo, and background echo. The direct transmission signal represents the directly transmitted signal from the speaker to the microphone. The face echo and background echo are signals reflected by face and far-away objects. According to their traveling distances, the arrival times of direct transmission, face echo, and background echo are sorted from early to late. To facilitate extracting the target face echo from the captured mixed recording in the signal processing stage, we designed an appropriate chirp duration and interval to prevent overlap between these three signal components in the time domain. The duration of a chirp is crucial for its signal-to-noise ratio (SNR) [50].

In general, a longer chirp enables more energy to be collected at different frequencies [50]. However, if the chirp duration is too long, the directly transmitted signal from the speaker to the microphone could overlap with echoes from the nearby face region in the recording. For example, for sensing a face that is 30 cm away from the phone, the estimated arrival delay of the first sample point of the chirp is ∼1.7 ms at the speed of sound (∼340 m/s), which is around 77 samples in the recording. Therefore, if the chirp duration is longer than 77 samples, the echoes from the face region will overlap with the direct transmission of the emitted signal in the recording. According to our user study, the comfortable distance from the nose to the smartphone is 25 - 45 cm, which is around ∼1.4 - 2.5 ms at the speed of sound. Hence, we set the chirp duration as 60 samples, corresponding to ∼1.4 ms at the 44.1 kHz sample rate.

For chirp interval, a shorter interval can save the sensing time. However, it might harm the quality of face echoes because echoes from far-away objects could mix with face echoes in the final recording. Following the previous example, when the face is 30 cm away from the phone, if we set the chirp interval as 100 samples, an echo of a chirp from objects $(60+100+77)/44100 \times 34300/2 \approx 92$ cm away from the phone will overlap with the face echoes of the next chirp. Therefore, in the final signal, we adopt a 3000-sample interval for sensing chirps and an 8000-sample interval for the pilot tone to avoid such echo interference while maintaining reasonable sensing time. The total sensing time, including the pilot tone, is around $[(250+8000)+9\times(60+3000)]/44100 \approx 0.8$ seconds, which is short enough to be deployed in real user scenarios.

Annoyance control. As our chosen frequency range is 12-21 kHz, parts of which overlap with the audible frequency range of humans (<15-17 kHz), the final signal might produce audible annoyances to the users. To provide a better user experience, we reduced the annoyance to the minimum level. The audibility of the signal was controlled by both the signal design and the volume of the device. For Echo-FAS, we carefully estimated these factors to reduce the annoyance to the user. We first apply a Hamming window function to the chirps to increase the peak-to-side ratio [51], thereby creating less audible effects in the emitted signal. The hamming window also increases SNR for the chirps, supporting noise removal

in the signal processing stage.

On the other hand, the volume of emitted signals affects the detection quality by controlling the overall energy of the emitted signal. Different mobile phones have different optimal sensing volumes due to their hardware differences. For Echo-FAS, we find optimal sensing volumes of our experiment devices through preliminary studies. To evaluate our annoyance control method, we have collected user feedback during the experiments; more than 90% of volunteers can hardly sense the annoyance. Hence, we can conclude that Echo-FAS can be deployed with a good user experience in real-world scenarios.

C. Signal processing

Echo-FAS uses one microphone (the top one) to collect the signal for further analysis. As we analyzed in the previous subsection, each chirp in the received recording should contain three parts: the direct transmission signal, target face echo, and background echo.

The main idea of signal processing is to eliminate the interference of the direct transmission and background reflection signals and extract the target face region echo.

Fig. 5 illustrates our signal processing pipeline: signal segmentation, direct transmission removal, and target reflection extraction. Signal segmentation locates the beginning of the recording and subsequently segment it into nine signal segments, corresponding to the nine chirps of the emitted signal in Fig. 4; Direct transmission removal aims at removing the direct transmission signal for each clip; Finally, the target reflection extraction process applies the adaptive algorithm to perform the target face region signal extraction.

Signal segmentation. The final signal combines echoes and direct transmissions, which makes directly locating face echoes very challenging. Hence, we segment the recording into nine clips for coarse localization in the signal segmentation stage, which will be further processed for accurately locating face echoes. We first perform a synchronization step to locate the beginning of signal emission in the recording. The signal emission and signal collection are performed simultaneously. However, due to hardware limitations, the microphone and speaker are not perfectly synchronized in actual scenarios. Thus, the synchronization step can find the delay between signal emission and collection in the final recording, which helps us coarsely estimate the location of the face echoes following the signal duration design. First, Echo-FAS synchronizes the recording by locating the pilot signal using cross-correlation. As shown in Fig. 6 (a), the first and most prominent peak represents the beginning of signal emission. After synchronization, we apply a low-pass filter on the residual signal to remove ambient noise under 12 kHz. The signal then can be split into nine short clips.

Each clip contains direct transmission signals, background echoes, and target face echoes and only the face echoes contain the target face information for anti-spoofing. Thus, we need to further process the coarsely segmented clips to segment face echoes from the mixed signals.

Direct transmission removal. Since the direct transmission is the signal that directly transfers from the speaker to the microphone during signal emission, it should be in the same shape and length as the emitted chirp. The direct transmission contains significantly greater energy than face echoes and background echoes because encountered objects like faces and walls do not absorb energy during reflection. In addition, considering the shorter traveling distance, the arrival time of the direct transmission signal is expected to be earlier than the target face reflection signal. As shown in Fig. 5, we adopt a matched filter based on the above analysis, which convolutes the segmented recording clips with the original chirps to detect the location d of the direct transmission in each clip. As shown in Fig. 6 (b), the first and highest peak in the matched filter result indicates the beginning of the direct transmission signal. Echo-FAS removes the sample points before and in the direct transmission signal and keeps the remaining data samples for target face signal extraction.

Target reflection extraction. After the direct transmission removal, the remaining clips only contain face echoes and background echoes from far-away objects. Since face echoes are closer to the phone than surrounding objects, the face echoes have the shortest route among all echoes, making its arrival time the earliest in the recording. Furthermore, as illustrated in Fig. 6 (b), the face reflection is less sensitive to the matched filter than the direct transmission but still distinguishable from the echoes of the surroundings. Therefore, the target reflection should have the first and highest peak in the remaining matched filter results after the direct transmission removal. However, our experiment found that echoes from far-away objects could sometimes create noticeable peaks in the matched filter results, damaging the extraction accuracy. To address this issue, we design an adaptive algorithm to estimate the suitable location of the face echo for all nine clips in one detection. Since the distance from the face to the phone is fixed during detection, the locations of the face echos in all nine clips should be close. Therefore, in this algorithm, we iteratively calculate the nine peak locations of the nine clips and output the mean peak location with the minimum standard deviation representing the beginning of the target face echo location l.

Our adaptive algorithm iteratively processes the nine clips using the matched filter. The target face echo localization process is summarized in Algorithm 1. We calculate the mean of the most prominent peak locations in each clip to estimate the average face echolocation. In each iteration, each clip's 60-sample (same as the chirp duration of the designed signal) segment will be passed through the matched filter, and the most prominent peak locations will be recorded. The algorithm regards the mean of the most prominent peak locations with the minimum standard deviation as the average face echolocation. Besides, our algorithm processes the clips in a near-to-far manner. Considering our detection distance is 25 - 45 cm, we assume the actual distance from the face to phone is 23 - 50 cm during detection, corresponding to 60 - 130 samples transmission time at the sound of speed and 44.1 kHz sample rate.

After obtaining the average face echo location l, we crop 60 samples of each clip after l as the extracted target face echoes. For each input acoustic signal, nine 60-sample segments are

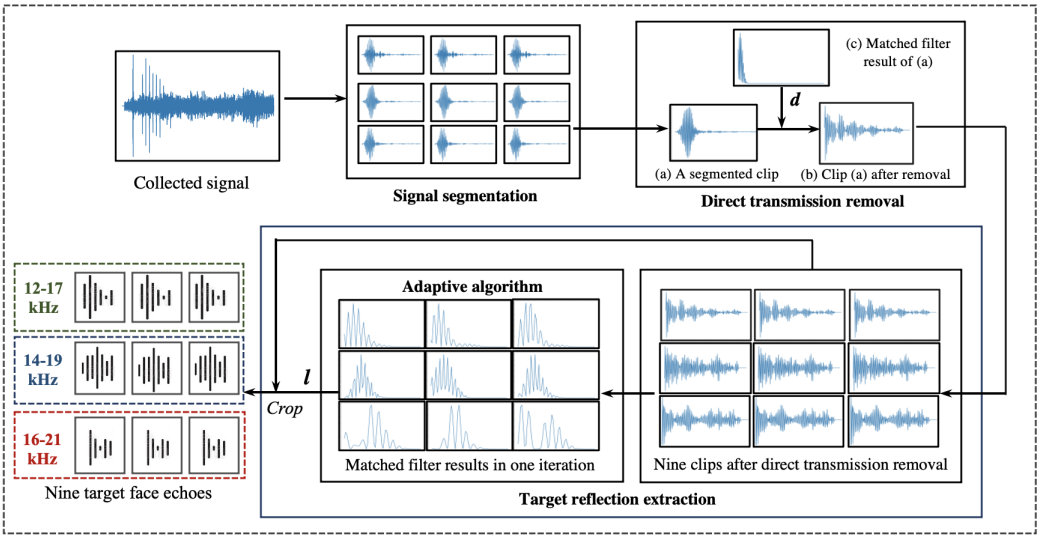

Fig. 5. Overview of the signal processing pipeline. The collected signal is firstly synchronized to locate the beginning of signal emission in the recording in the signal segmentation stage. Then it will be segmented into nine clips following the signal duration design. Each clip comprises our target face echo, direct transmission, and background echoes. In the direct transmission removal stage, we apply a matched filter on each clip to find the location d of direct transmission and remove it from the clip. Then, in the target reflection extraction stage, the remaining clips are fed into our adaptive algorithm to estimate the target face echo location l in the clips. Finally, we crop the nine 60-sample segments according to l as the extracted target face echo signals.

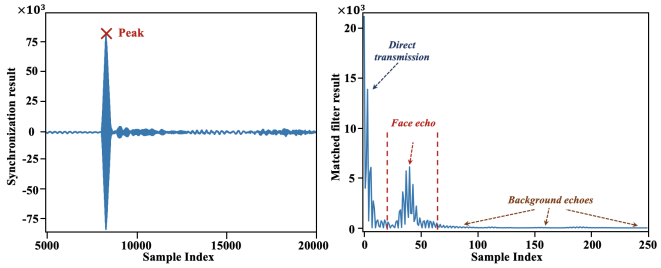

Fig. 6. (a). Synchronization result between microphone and speaker. The peak indicates the pilot position of the received signal. (b). Result of applying match filter to the synchronized signal. The first and largest peak represents the direct transmission signal. The second peak segmented by red dotted lines indicates our target face echo. The background echoes with a higher time delay have a lower amplitude.

extracted to distinguish the input query between live and spoof faces.

D. The superiority of our designed signal

It is worth noting that there exist several works which also utilize acoustic signal for face anti-spoofing [29]–[31], where the signal was designed with the following desired properties: commodity hardware requirement, better user experience, and higher final liveness detection performance. These factors are critical to the system's deployment in real-world applications. Our signal is more advanced than the signal designs in the previous related work.

Unlike RFace [31] using 920.625 MHz as the working frequency, our signal covers a much lower 12-21 kHz range, which does not require any additional hardware and can be handily deployed on commodity smartphones and other devices. Compared with Echoprint [29] that chosen a single 16-22 kHz pulse as an emitted signal, we flexibly divide our acoustic signal into nine chirps. Such design enjoys three benefits: (a). It keeps more energy at each frequency, leading to a higher SNR in the final recording; (b). It reduces the chirp duration, preventing the overlapping between the target face echo and other signals; (c). The emitted signals with diverse frequency components can facilitate Echo-FAS to capture more spatial structural features. Finally, Echoface [30] designed the signal with a 3 kHz bandwidth in each chirp, and its overall duration of the signal is over 1.48 s. However, Echo-FAS designed the signal with a larger bandwidth of 5

Algorithm 1: Target Face Echo Localization

Input : The signal clips after direct transmission signal removal $R_i[n], i \in [1,9]$, where n represents the sample index

Output: The estimated unified face echo location l

Initialize reference standard deviation: $std_r \leftarrow MAXINT$
for $p \in [0, 70]$ **do**
  Initialize empty list of peaks index: K;
  for $i \in [1, 9]$ **do**
    Pass $R_i[p : p+60]$ through the matched filter
    Add the index of first and largest peak k_i to K
  end
  Compute the average of K: $\mu_K \leftarrow \frac{1}{9}\sum_{j=1}^{9} K$
  Compute the standard deviation of K: $\sigma_K \leftarrow \sqrt{\frac{1}{9}\sum_{j=1}^{9}(K_j - \mu_K)^2}$;
  if $\sigma_K < std_r$ **then**
    $std_r \leftarrow \sigma_K$;
    $l \leftarrow \mu_K$;
  end
end

kHz, thus achieving a better echo separating resolution. On the other hand, the overall duration of our signal is 0.8 s, which is much shorter than Echoface.

Echo-FAS can promise a better user experience and higher liveness detection performance with commodity hardware by employing such acoustic signal design, facilitating its deployment in practical scenarios. Based on our preliminary study, we also find that our signal can achieve better FAS performance compared with the design above strategies, especially under complex environment settings.

V. Framework Design for FAS

This section proposes a novel two-branch Echo-FAS framework, which combines the global and local frequency features of the input signals, leading to high-accuracy face liveness detection. As illustrated in Fig. 7, the reflected signal collected by the device is firstly fed to the signal preprocessing model, including noise removal and signal extraction, which have been elaborated in Sec. IV-C). Then, the processed signal will be split into nine segments, corresponding to three different frequency sweep ranges. Echo-FAS processes the input signal in the frequency domain as the received signal recording is uneven over different frequencies, thus presenting significant distinguishable clues for classifying the input query between live and spoof face. This phenomenon is mainly caused by two factors: (a). the target surface's absorption of the signal; (b). the mixed echoes with different phases may be constructive at some frequencies while destructive at other frequencies.

A two-branch framework is designed to complementarily combine the Fast Fourier Transform (FFT) global frequency feature and the Short Time Fourier Transform (STFT) local frequency feature of the input query. The first branch applies Fast Fourier Transform (FFT) to the nine signal segments and converts them to nine corresponding frequency segments. FFT can reflect the global frequency statistics of each signal segment. Thus, the first branch employs a transformer architecture to extract the global frequency feature of the nine input tokens. The cascaded self-attention modules in transformers can effectively capture long-distance feature dependencies while tending to ignore local feature details.

In contrast, numerous works have demonstrated that the convolution operations in CNN are good at extracting local features but experience difficulty in capturing global representations. Therefore, we employ a CNN architecture to complementarily mine more local informative clues in the second branch. We first leverage the Short-Time Fourier Transform (STFT) to convert the processed signals into corresponding spectrograms. According to the frequency sweep rationale of our designed signal, for each chirp signal, the emitted signal's frequency linearly increases over time. Herein, we employ STFT to analyze the frequency content of local windows of the processed signals. The CNN subsequently processes the spectrogram, and then the local frequency feature can be extracted.

Finally, we devise two cross attention modules to model the correlations of the extracted global and local frequency features. Moreover, the two attended features will be combined to determine whether the input query is a live person or a spoofer. The framework is trained in an end-to-end manner and supervised by the cross-entropy loss between the prediction result $\hat{c}_i$ and ground truth label c_i:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(c_i \log \hat{c}_i + (1 - c_i)\log(1 - \hat{c}_i)), \tag{2}$$

VI. Acoustic-based FAS Experiments

A. Implementation Details

The proposed framework is implemented by Pytorch [52]. The model is trained using Adam optimizer [53] with β_1=0.9 and β_2=0.999. We set the learning rate and weight decay as 1e-4 and 1e-5, respectively. The model is trained on 1 Quadro RTX 8000 GPU with batch size 1024. We split the training, validation, and testing sets as 8:1:1. We train our model for 1000 epochs and validate it at the end of every epoch. We pick the checkpoint with the best AUC score on the validation set and test it on the testing set.

Following [29], [30], we directly use the raw signal as the feature for our proposed method as well as other baseline techniques. In our preliminary study, we also experienced using some feature extraction techniques (e.g., Mel-frequency cepstral coefficients (MFCC) [54] and spectral contrast (CONT) [55]) but found the performances are not satisfactory compared with using raw signal data.

B. Evaluation metrics

In this work, we adopt the following evaluation metrics which have been widely used in previous FAS works:

1) Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

2) Area Under Curve (AUC): AUC measures the area under the Receiver Operating Characteristic (ROC) curve. A higher AUC score indicates better FAS performance.

3) Equal Error Rate (EER): EER measures the False Positive Rate (FPR) that equals True Positive Rate (TPR).

4) Attack Presentation Classification Error Rate (ACER): ACER represents the mean of Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER), where APCER measures the ratio that spoof faces are misclassified into live faces. At the same time, BPCER denotes the ratio that live faces are misclassified into spoof faces. A lower ACER value indicates better performance.

$$ACER = \frac{APCER + BPCER}{2} \tag{4}$$

5) Half Total Error Rate (HTER): HTER denotes the average of the False Acceptance Rate (FAR) and False Reject Rate (FRR):

$$HTER = \frac{FAR + FRR}{2} = \frac{1}{2}\left(\frac{FP}{TN + FP} + \frac{FN}{TP + FN}\right) \tag{5}$$

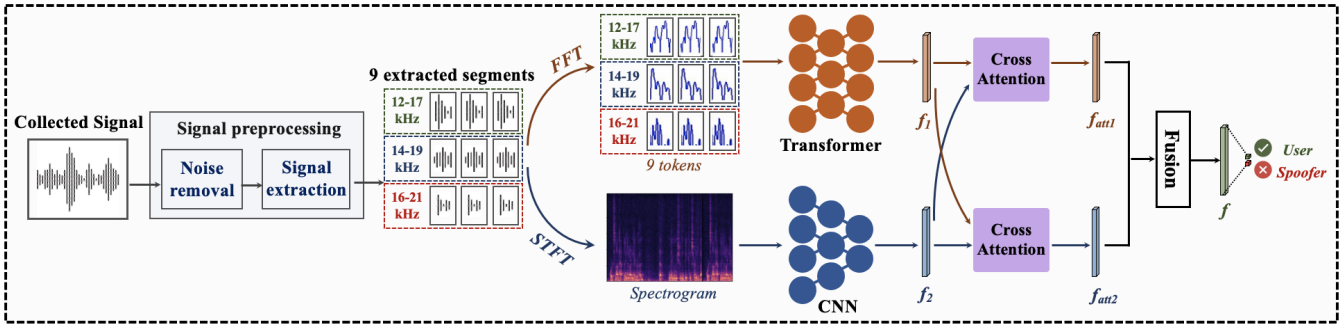

Fig. 7. Overview of the proposed two-branch Echo-FAS pipeline. The input signal is firstly fed forward to the Signal preprocessing module to extract nine signal segments. In the first branch, we apply Fast Fourier Transform (FFT) to obtain nine frequency signals, which are regarded as nine tokens and will be sent to the Transformer to obtain the global frequency feature. In the second branch, we further employ Short Time Fourier Transform (STFT) to obtain the corresponding spectrogram to the processed signal. The CNN is used to extract the local frequency feature of the input signals. The Cross Attention mechanisms model the correlation of the extracted global frequency feature f_1 and local frequency feature f_2. Subsequently, the attended features will be finally combined to determine whether the input query is a live person or a spoofer.

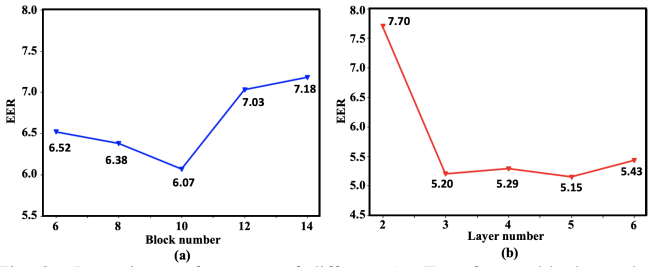

Fig. 8. Detection performance of different (a) Transformer block numbers and (b) CNN layer numbers.

C. Evaluation algorithms

In this section, we adopt the following algorithms to classify the input acoustic signal between live and spoofing faces: Logistic Regression (LR) [56], Linear Discriminant Analysis (LDA) [57], Decision Tree (DT) [58], Naive Bayes (NB) [59], K-Nearest Neighbors (KNN) [60], Support Vector machine (SVM) [61], Multilayer Perceptron (MLP); Convolutional Neural Network (CNN); Transformer (TF), and our proposed system Echo-FAS.

D. Network Architecture

Echo-FAS devises a two-branch architecture and takes the spectrogram and fft tokens as inputs. To better understand the impacts of different CNN and Transformer structures, we conduct an ablation study of the detection performance versus the network depth. Following [62], we depict the curves of EER values versus different Transformer block numbers and different CNN layer numbers in Fig. 8. Finally, we select the 10-block Transformer and 5-layer CNN as the backbones of the two branches. Moreover, we further present the detailed architecture of the two-branch Echo-FAS system in Fig. 9(a). The left branch presents the **CNN** network, while the right one details the **Transformer** structure. We also specified the output feature shapes of each layer. Fig. 9(b)-(d) illustrate the detailed structures of **Multi-Head Attention**, **Feed Forward**, and **Cross Attention**. The source code is available at: https://github.com/ChenqiKONG/EchoFAS.

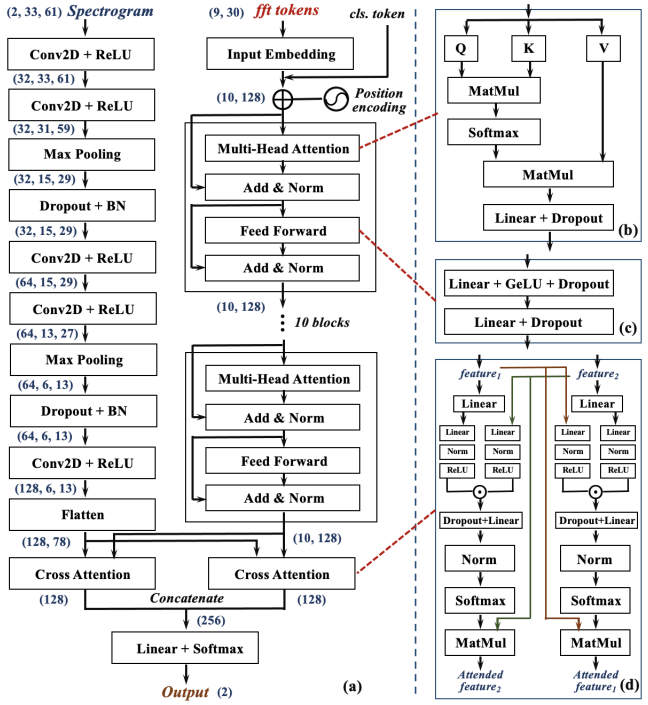

Fig. 9. (a). The detailed architecture of the proposed two-branch Echo-FAS system (**Dropout** probability: 0.5, **BN**: batch normalization). The left branch presents the **CNN** network, while the right one details the **Transformer** structure. We also illustrate the architectures of some transformer modules: (b). **Multi-Head Attention**; (c). **Feed Forward**; and (d). **Cross Attention**. (Q: query; K: key; V: value; MatMul: matrix multiplication; Dropout probability: 0.1.)

E. Annoyance user study.

The frequency of our designed signal is 12-21 kHz, which is slightly audible for users. We perform the annoyance reduction by adding Hamming window to the designed signal and setting lower volume for Andriod smartphones, as introduced in Sec. III-B). Considering that noise annoyance has to be controlled in users' daily uses, we carried out an annoyance user study with 30 volunteers to survey the annoyance level of four data collection devices. As illustrated in Fig. 10, we set three annoyance levels in this user study: no annoyance,

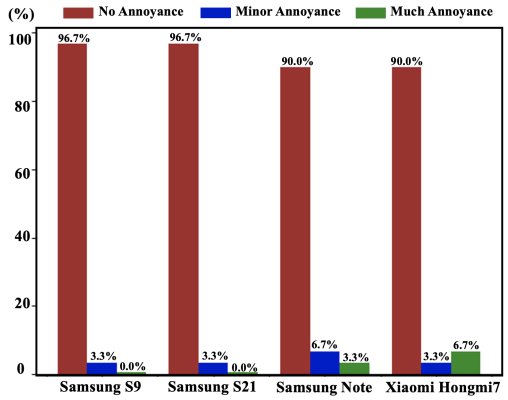

Fig. 10. Annoyance user study of four data collection devices.

TABLE I
FACE LIVENESS DETECTION RESULTS ACROSS DIFFERENT IDENTITIES.

Methods	AUC(%)↑	ACC(%)↑	HTER(%)↓	ACER(%)↓	EER(%)↓
LR	95.20	90.87	8.53	9.14	8.59
LDA	94.82	89.63	9.39	10.38	8.98
DT	81.25	81.25	18.01	18.75	22.10
NB	90.65	84.32	14.16	15.69	13.81
KNN	96.21	91.36	8.26	8.65	7.98
SVM	96.53	90.99	8.73	9.02	8.25
MLP	96.82	92.97	6.82	7.04	7.18
CNN	98.16	94.55	5.45	6.20	5.15
TF	97.69	93.07	6.93	7.16	6.07
Ours	**98.79**	**95.18**	**4.83**	**5.35**	**4.34**

minor annoyance, and much annoyance. It is shown that more than 90% of users can hardly notice the sound, and only around 2.5% of users feel much annoyance. Therefore, we can conclude that the devised Echo-FAS system has successfully addressed the annoyance issues and can be deployed in real-world applications.

F. Face liveness detection results

1) Face liveness detection results cross identities: Considering the FAS model in the real-world application scenarios should be directly deployed to unseen identities, we train the Echo-FAS model on 25 identities and test it on the five unseen identities. The face liveness detection results are listed in Table I. It can be observed that the proposed Echo-FAS system achieves outstanding 98.79 AUC and 95.18 ACC detection scores. On the other hand, other classification algorithms such as MLP, CNN, and TF can also achieve promising detection performances, further demonstrating that the acoustic signal carries large informative clues for identifying the input query between live and spoofing. Moreover, the Echo-FAS system outperforms all the classification algorithms listed in Table I, which shows that the designed Echo-FAS system can conduct a high-accuracy FAS.

2) Face liveness detection results of different devices: In this work, we apply four smartphones (Samsung s9, Samsung s21, Samsung edge note, and Xiaomi Redmi7) to conduct the data collection process. Generally speaking, different devices have different imperfect hardware conditions, and speaker and microphone differences between devices will lead to severe data distribution gaps. Thus, to perform a more accurate and secure attack detection, training the specific model for each device is necessary. Herein, we show the face liveness detection results of the specific devices in Fig. 11. It can be readily observed that the Echo-FAS can achieve high-accuracy face liveness detection performances on all four smartphones and outperforms all the classification algorithms presented in the baseline.

TABLE II
FACE LIVENESS DETECTION RESULTS ACROSS DIFFERENT AMBIENT NOISES

Noise	40dB		60dB		70dB	
Methods	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
LR	86.89	19.98	89.28	15.98	88.45	17.08
LDA	84.56	24.78	88.05	19.46	87.14	19.48
DT	77.92	21.30	78.46	20.70	76.55	22.70
NB	74.56	31.20	74.88	28.43	74.73	25.44
KNN	95.01	10.03	96.10	8.73	94.90	11.52
SVM	96.12	9.39	95.83	9.30	94.97	12.95
MLP	96.31	9.39	96.20	9.21	95.88	10.05
CNN	98.48	6.04	98.70	6.50	98.13	7.67
TF	98.31	6.23	98.30	6.06	98.12	7.07
Ours	**98.93**	**4.86**	**98.94**	**5.20**	**98.57**	**6.49**

TABLE III
FACE LIVENESS DETECTION RESULTS ACROSS DIFFERENT DISTANCES

Distance	25cm		35cm		45cm	
Methods	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
LR	86.81	20.97	87.73	20.34	85.21	22.60
LDA	84.54	23.76	85.23	24.39	82.78	24.32
DT	74.04	24.75	74.93	23.64	72.79	26.79
NB	75.19	29.51	76.27	28.66	73.48	29.81
KNN	94.14	11.94	95.29	10.08	93.54	13.08
SVM	94.46	12.32	95.55	10.57	93.78	12.23
MLP	94.13	12.18	96.01	9.82	93.76	12.85
CNN	97.40	7.76	98.17	6.52	97.26	8.69
TF	97.31	8.68	97.94	6.82	97.15	8.32
Ours	**98.33**	**6.70**	**98.64**	**5.48**	**98.16**	**6.53**

3) Robustness against ambient noise: Considering that ambient noise is inevitable in users' daily usage, we conduct a robustness evaluation against ambient noise in this subsection. To investigate the noise resistance of the Echo-FAS system, we evaluate our model's face liveness detection performance across different background noise levels. In Table II, we list the face liveness detection results of all detection algorithms in 40 dB, 60 dB, and 70 dB noise environments, corresponding to quiet, noisy, and very noisy scenarios. It can be observed that the proposed model again achieves the best detection performance among all detection methods. Besides, the Echo-FAS can consistently achieve outstanding AUC and HTER scores across various ambient noise settings. Finally, it is worth noting that the Echo-FAS can achieve promising 98.57% detection even in a very noisy environment (70dB), demonstrating the high robustness of our system.

4) Detection results under various user habits: Echo-FAS aims at performing face liveness detection in a more user-friendly fashion. To cater to various users' habits, we further introduce two factors of distance and pitch in the collected database, corresponding to different phone holding distances and positions. In this subsection, we evaluate our model across various distances and pitches and report the detection results in Table III and Table IV, respectively. Unsurprisingly, the proposed method again outperforms all the listed models and achieves stable and promising detection performance regardless of different distances and pitches, further demonstrating

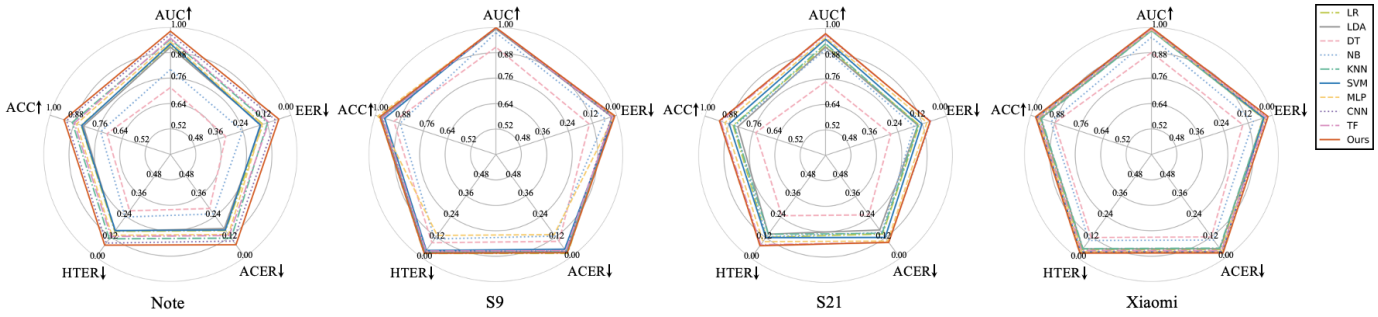

Fig. 11. The face liveness detection results of four different devices. The ↑ indicates the higher the better, while the ↓ indicates the lower the better.

Fig. 12. Face samples of four new presentation attack species. (a). Paper cut; (b). Curved Mask; (c). Half Mask; (d). Cropped Mask. The top and bottom rows respectively present the collected face pictures under normal and dark illumination conditions.

TABLE IV
FACE LIVENESS DETECTION RESULTS ACROSS DIFFERENT PITCHES

Pitch	+10 degree		0 degree		-10 degree	
Methods	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
LR	87.60	20.24	87.87	19.57	83.28	23.70
LDA	85.89	23.36	84.96	23.72	79.92	27.41
DT	76.60	23.21	76.54	22.62	71.85	26.53
NB	80.13	25.07	77.01	28.22	74.17	30.47
KNN	94.95	11.32	95.22	9.93	94.63	11.98
SVM	94.16	12.45	96.13	9.95	94.27	13.11
MLP	94.68	12.06	96.46	9.79	94.76	12.20
CNN	97.32	8.26	98.35	6.52	97.44	9.09
TF	97.00	8.80	97.63	7.17	97.27	9.19
Ours	**97.97**	**6.94**	**98.40**	**5.71**	**98.17**	**8.25**

the effectiveness and high robustness of the designed Echo-FAS system.

5) Generalization capability to unseen PA species: To study the robustness of Echo-FAS against new PA, we further collect four unseen PA species under two illumination conditions, which are illustrated in Fig. 12: (a). Paper Cut: the face and body regions are cropped out to make the attack looks more realistic; (b). Curved Mask: unlike the 2D media only has a flat surface, we bend the printed photographs to fool the Echo-FAS system; (c). Half Mask: we cut half of the printed photography and carefully aligned the edges, leaving half of the 3D face region exposed to the Echo-FAS system; (d). Cropped Mask: we cut the mouth and eye regions to make this database more challenging. The top and bottom rows present the attack attempts under normal and dark illumination conditions.

We train our model on the data of print and replay attacks and directly test it on the four unseen PA species under dark and normal illumination conditions. Note that the 25 training person ids and the 5 testing person ids have no overlap. We report the PA detection performance in Table V. It can be seen that the Echo-FAS again achieves outstanding detection performances for all four challenging PA species. Besides, it is unsurprising that the Echo-FAS can work well under both illumination conditions. Compared with the intra-PA detection results in Table I, the average detection performance on unseen attack types only dropped slightly, from (98.79% AUC, 4.83% HTER) → (97.59% AUC, 6.56% HTER), further demonstrating the generalization capability of Echo-FAS from another point of view.

TABLE V
PRESENTATION ATTACK DETECTION PERFORMANCES ON FOUR UNSEEN ATTACK SPECIES UNDER TWO ILLUMINATION CONDITIONS.

	Dark		Normal		AVG	
Test PA	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
Cropped Mask	98.43	4.76	98.68	5.26	98.56	5.01
Curved Mask	98.13	5.86	99.86	3.36	99.00	4.61
Half Mask	96.58	6.26	95.73	7.76	96.16	7.01
Paper Cut	99.11	4.76	94.13	14.46	96.62	9.61
AVG	98.06	5.41	97.10	7.71	97.59	6.56

TABLE VI
COMPARISON WITH PRIOR ARTS.

	AUC(%)↑	ACC(%)↑	HTER(%)↓	EER(%)↓
Echoface	94.50	90.90	8.90	9.38
Echo-FAS	98.79	95.18	4.83	4.34

6) Comparison with prior arts: The proposed Echo-FAS system differs from prior acoustic-based FAS methods in the following two aspects: (1) signal configurations and (2) presentation attack detection algorithms. We have analyzed the superiority of the proposed signal configuration to Rface, Echoprint, and Echoface in Sec. IV-D). In this part, we further compare the performance of Echo-FAS with prior arts quantitatively. As Rface cannot be built into commodity smartphones due to the working frequency limit, we conduct the FAS performance comparison with the most recently proposed method Echoface. We firstly use four smartphones to emit the acoustic signal in Echoface to acquire data samples; then, we reimplement the signal processing and classification algorithms of Echoface, based on the collected acoustic recordings. During the data acquisition process, even though we keep the person identities, devices, and environmental variables the same as our data collection settings, it is not an ideally

TABLE VII
FAS RESULTS USING DIFFERENT FREQUENCY GROUPS.

12-17kHz	14-19kHz	16-21kHz	AUC(%) ↑	HTER(%) ↓
✓	-	-	96.16	10.77
-	✓	-	95.55	11.72
-	-	✓	93.28	12.14
✓	✓	-	98.32	5.65
-	✓	✓	96.55	8.78
✓	-	✓	98.09	6.52
✓	✓	✓	**98.79**	**4.83**

TABLE VIII
EFFECTIVENESS OF THE DESIGNED TWO-BRANCH ARCHITECTURE.

Methods	AUC(%)↑	ACC(%)↑	HTER(%)↓	ACER(%)↓	EER(%)↓
Local	98.16	94.55	5.45	6.20	5.15
Global	97.69	93.07	6.93	7.16	6.07
Fusion	**98.79**	**95.18**	**4.83**	**5.35**	**4.34**

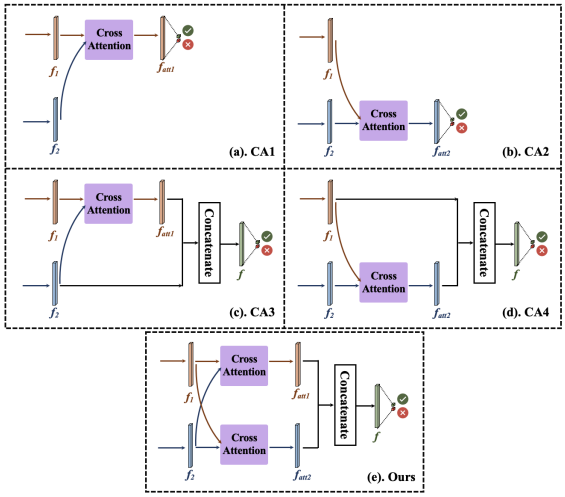

Fig. 13. Illustration of different feature fusion strategies: (a). CA1, (b). CA2, (c). CA3, (d). CA4, and (e). Ours.

TABLE IX
EFFECTIVENESS OF THE TWO CROSS ATTENTION MODULE.

	AUC(%)↑	ACC(%)↑	HTER(%)↓	EER(%)↓
BL	98.15	93.48	6.52	5.87
CA1	98.76	94.89	5.11	**4.03**
CA2	98.44	93.80	6.20	5.66
CA3	98.67	94.55	5.45	4.97
CA4	98.16	93.65	6.36	6.95
Ours	**98.79**	**95.18**	**4.83**	4.34

fair comparison since both signal setups and classification algorithms of Echoface and Echo-FAS (Ours) are distinctive. From Table VI, it can be observed that Echo-FAS achieves a better detection performance, mainly benefiting from our signal design and the proposed two-branch learning scheme.

7) Discussion: We quantitatively evaluate the proposed Echo-FAS system under various experimental settings in this subsection. Extensive experimental results demonstrate that our model can achieve high accuracy and robust face liveness detection performance in a more user-friendly fashion, indicating that the Echo-FAS can be handily deployed in real-world application scenarios.

G. Ablation Study

1) Effectiveness of different frequency groups: In the proposed system, the well-designed signal consists of the following three different frequency groups: 12-17kHz, 14-19kHz, and 16-21kHz. Such signal design philosophy ensures two compelling properties: high detection accuracy and less annoyance. To investigate the contribution of each frequency group to the final liveness detection performance, we conduct an ablation experiment under the cross-identity setting in Table VII. From the results reported in Table VII, we can conclude that: (a). the emitted low-frequency range signal can capture more informative clues than the high-frequency signal; (b). the combination of different frequency groups indeed improves the liveness detection accuracy.

2) Effectiveness of two-branch architecture: In this work, we design a novel two-branch Echo-FAS framework to synchronously combine the global and local information in the frequency domain. To study the effectiveness of each branch, we report the detection performance of the single local branch, the single global branch, and the fusion of the two branches in Table VIII. Compared with the single branch, the detection accuracy using the two-branch framework improves, demonstrating that the Echo-FAS system successfully fuses the local and global frequency features of input acoustic signals. The results also show the effectiveness of the designed two-branch Echo-FAS architecture.

3) Effectiveness of cross-attention (CA) module: It is intuitive that the global frequency feature and local frequency feature of the identical input signal are highly-related. Thus, how to model the relationship between them and learn more refined features is a non-trivial problem. We design two cross-attention (CA) modules for learning the interaction between the global and local frequency features in this work.

To better demonstrate the effectiveness of the proposed cross attention mechanism, we compare our scheme with four feature fusion strategies, all using one single cross attention module. We depict our scheme and other four feature fusion strategies in Fig. 13: (a). **CA1**: f_2 attends to f_1; (b). **CA2**: f_1 attends to f_2; (c). **CA3**: f_2 first attends to f_1, then f_{att1} is concatenated with f_2; (d). **CA4**: f_1 first attends to f_2, then f_{att2} is fused with f_1; (e). **Ours**: two attended features f_{att1} and f_{att2} are concatenated for the final decision-making.

We report the detection results of all cross-attention schemes in Table IX. BL indicates the baseline method that directly concatenates f_1 and f_2. It can be observed that the overall detection performances of CA1-CA4 are superior to BL, and the proposed scheme using two cross-attention modules leads to further performance improvement. Thus, it can be concluded that the usage of a single cross-attention module can indeed improve the detection performance. And the proposed dual cross-attention design boosts the feature interactions on each side, thereby achieving the best detection performance.

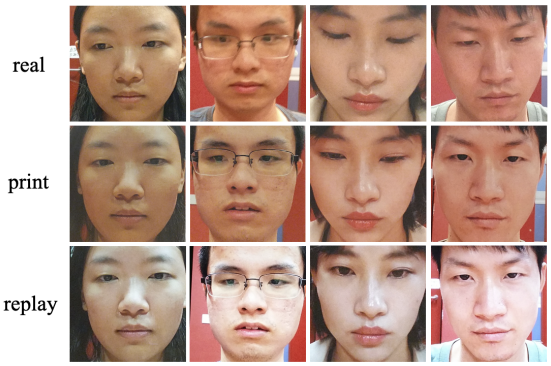

Fig. 14. The collected face samples. The top row presents the live faces, while the middle and bottom rows show the print and replay attack examples.

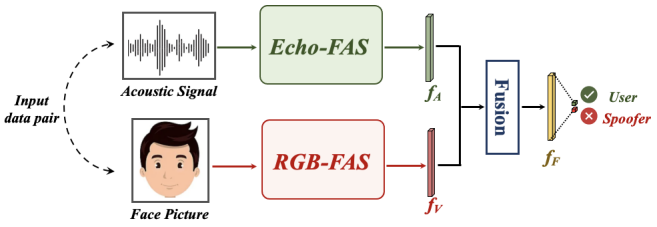

Fig. 15. Overview of the vision-acoustic multi-modality fusion framework. The input data pair includes one acoustic signal segment and one corresponding face image. We employ the Echo-FAS backbone to extract the acoustic feature f_A and leverage the widely-used ResNet18 backbone to extract the visual feature f_V. The feature fusion module fuses these two features and sequentially makes the final decision.

4) Discussion: Herein, we conduct extensive ablation studies to demonstrate the effectiveness of the dedicated learning scheme, the designed framework architecture, and the proposed signal configuration.

VII. Multi-modality FAS Experiments

A. Multi-modality face liveness detection cross to unseen device

Under the uncontrollable environment conditions in the real-world application scenarios, the RGB modality-based FAS models are prone to suffer significant detection performance drops due to the domain gaps between the training and testing data sample distributions. On the other hand, Echo-FAS can capture the surface geometric information from the input query. Moreover, such geometric information can reflect much depth, which can hardly be learned from the RGB inputs. Thus, it is reasonable to flexibly assemble the proposed Echo-FAS system with the RGB modality to conduct a more generalized face liveness detection and improve the robustness of the RGB-based model.

TABLE X
Multi-modality face liveness detection results.

	Vision		Acoustic		Fusion	
Test device	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
Note	89.50	21.30	68.11	38.83	**96.33**	**14.38**
S9	99.91	2.90	95.39	24.48	**100.00**	**0.00**
S21	94.53	23.89	82.08	19.79	**100.00**	**0.00**
Mi	89.99	24.95	89.98	29.21	**99.99**	**2.99**
AVG	93.48	18.26	83.89	28.08	**99.08**	**4.34**

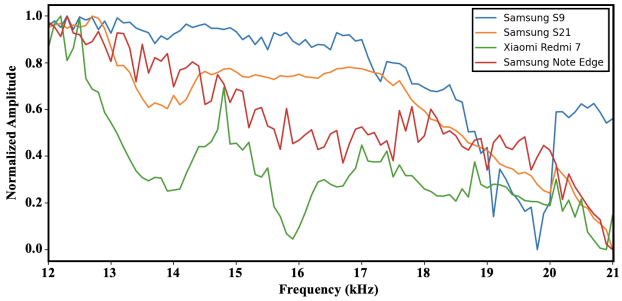

Fig. 16. Frequency response curves of the four smartphones. The amplitude has been normalized to [0, 1].

1) Face liveness detection results: To verify this point, we further control the front camera to capture the face picture of ten volunteers during the data collection process. It is worth noting that the acoustic signal and face image are synchronously captured by the microphones and front cameras of the smartphones, which means that each acoustic signal segment is paired with one corresponding face image. We show some face examples in Fig. 14. The top row presents the live face images, while the middle and bottom rows show the print and replay attack examples. It is challenging for naked eyes to discriminate the live face pictures from the spoofing ones.

We illustrate the overview of the vision-acoustic multi-modality fusion framework in Fig. 15. The input data pair includes one acoustic signal segment and one corresponding face image. We employ the Echo-FAS backbone to extract the acoustic feature f_A and leverage the pervasive ResNet18 backbone to extract the visual feature f_V. The feature fusion module combines these two features and makes the final decision sequentially. The ResNet18 is pretrained on the ImageNet dataset [63], and the whole framework is trained in an end-to-end manner. Four smartphones collect the vision and acoustic data: Samsung edge note, Samsung s9, Samsung s21, and Xiaomi Redmi7, which can be regarded as four domains due to the discrepancy between their sensors. We train the model on three devices and cross it on the other one. The face liveness detection results of the visual modality, auditory modality, and multi-modality are reported in Table X. Thanks to the usage of acoustic modality data, the face liveness detection performance of the multi-modality framework gained a significant improvement compared with the single RGB modality. Therefore, it can be concluded that acoustic data can be regarded as auxiliary information to mitigate the RGB domain gaps effectively.

2) Analysis and discussions: From Table X, we observe that the performance of Echo-FAS somewhat drops when applied to unseen devices, even though it has demonstrated outstanding generalization capability across different distances, noise levels, and user habits. This limitation is mainly caused by the hardware discrepancies among different smartphones. That is, the speaker and microphone in each phone have unique mechanical and electronic features due to the imperfect manufacturing process [64]. Prior works [48], [65], [66] leverage such unique features as fingerprints to identify different devices. To be more specific, we present the acoustic

TABLE XI
FACE LIVENESS DETECTION RESULTS ON UNSEEN PA SPECIES UNDER THE NORMAL ILLUMINATION CONDITION.

	Vision		Acoustic		Fusion	
Test PA	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
Cropped Mask	100.00	5.84	98.68	5.26	100.00	0.00
Curved Mask	100.00	5.64	99.86	3.36	100.00	0.00
Half Mask	96.92	34.10	95.73	7.76	100.00	10.10
Paper Cut	99.86	5.64	94.13	14.46	100.00	2.70
AVG	99.20	12.81	97.10	7.71	**100.00**	**3.20**

TABLE XII
FACE LIVENESS DETECTION RESULTS ON UNSEEN PA SPECIES UNDER THE DARK ILLUMINATION CONDITION.

	Vision		Acoustic		Fusion	
Test PA	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓
Cropped Mask	77.11	29.24	98.43	4.76	100.00	0.10
Curved Mask	77.36	25.94	98.13	5.86	100.00	0.00
Half Mask	75.42	42.02	96.58	6.26	100.00	0.20
Paper Cut	77.03	25.74	99.11	4.76	100.00	12.30
AVG	76.73	30.74	98.06	5.41	**100.00**	**3.15**

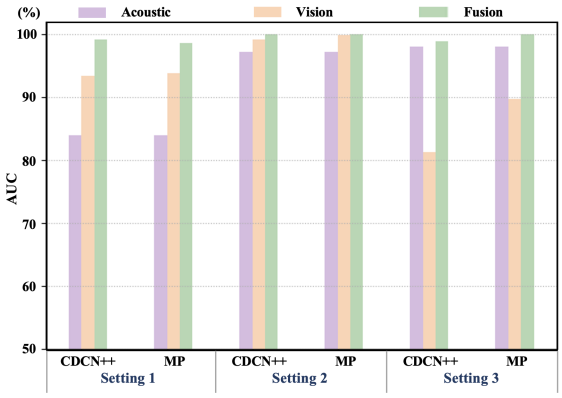

Fig. 17. Multi-modality face liveness detection results with SOTA RGB-based methods under three experimental settings: **Setting 1**: Cross-device; **Setting 2**: Cross-PA; **Setting 3**: Cross-PA & illumination. CDCN++ and MP indicate two SOTA RGB-based FAS methods.

attenuation formula below:

$$R(f,x) = L_S(f)L_M(f)R_0(f)e^{-\alpha(f)x} + noise \qquad (6)$$

where $R_0(f)$ and $R(f,x)$ represent the signal power transmitted and received at frequency f. $L_S(f)$ and $L_M(f)$ denote the energy loss caused by the speaker and microphone imperfection at f. $e^{-\alpha(f)x}$ is the attenuation factor related to the propagation distance x. $noise$ represents the ambient noise. Basically, $L_S(f)L_M(f)$ can be regarded as the fingerprint of the device.

From the analysis above, we believe that the frequency responses of the built-in microphones and speakers in different phones should be distinctive. To verify this point, we further measure the frequency response curves of the four Android smartphones in Fig. 16. It can be observed that the frequency responses of different devices differ dramatically at the working frequency range (12-21 kHz), thus causing performance drops when the training and testing acoustic data is from different phones.

It cannot be denied that cross-device detection is a limitation of Echo-FAS. However, the customized acoustic signal takes advantage of capturing 3D geometric information of input query, thus can be regarded as a complementary modality to RGB modality and can further enhance the generalization capability. Moreover, it is feasible to train a specific model for each phone in practical applications to address this device gap issue. And it is also interesting to design algorithms to erase device fingerprints in Eqn. (6) and extract more general features from the recorded acoustic signals, to further improve the generalization capability. We prefer to investigate these tasks in our future works.

TABLE XIII
COMPUTATIONAL COSTS AND MODEL SIZES.

	MACs	#param.
ResNet18	1.82 G	11.69 M
CDCN++	50.97 G	2.26 M
Meta-Pattern	13.28 G	62.13 M
Echo-FAS (Ours)	0.069 G	3.73 M

B. Multi-modality face liveness detection cross to unseen attack species

As discussed above, the acoustic signal can serve as a complementary modality to mitigate the overfitting problem in the RGB modality. In this section, we investigate the effectiveness of Echo-FAS defending unseen PA species in two-modality fusion experiments. Specifically, we train the models on the data of print and replay attacks under normal illumination and test them on four unseen PA types with normal and dark illuminations. From Table XI, we can observe that the usage of acoustic data improves the detection performance from (99.20% AUC, 12.81% HTER) → (100.00% AUC, 3.20% HTER) under the cross PA setting. Table XII reports the face liveness detection results on four unseen PA species under the dark illumination condition. The RGB-based models suffer significant performance drops while the acoustic modality still works well in poor ambient illumination environments. Finally, the model fusing vision and acoustic modalities takes advantage of both RGB texture features and acoustic-based facial geometric features and achieves the best performance, demonstrating that the acoustic modality plays an auxiliary role in mitigating the domain gaps in the RGB modality.

C. Multi-modality face liveness detection with SOTA RGB-based models

To verify the complementary efficiency of Echo-FAS when allied with SOTA RGB-based models, we present the FAS performances of two-modality fusion under three experimental settings: **Setting 1**: Cross-device, **Setting 2**: Cross-PA, and **Setting 3**: Cross-PA & illumination. We conduct two RGB-based SOTA methodologies: CDCN++ [67] and Meta-Pattern (MP) [68], on our database. As shown in Fig. 17, the acoustic modality plays an auxiliary role to the RGB modality, and it can consistently boost the multi-modality fusion FAS performances under all experimental settings. Thus, in real-world application scenarios, the proposed Echo-FAS system can be solely deployed or flexibly allied with other SOTA RGB-based FAS systems.

Besides, resource consumption is a vital factor for model deployment in practical scenarios, especially for mobile applications. For this reason, we measure the computational

cost and model size of Echo-FAS and SOTA RGB-based methods. As shown in Table XIII, Echo-FAS has only 0.069G MACs (multiply–accumulate operations) and 3.73M parameters, which is much more lightweight than other RGB-based FAS models. Therefore, Echo-FAS tends to be a better choice than RGB-based models under the ultra low-power mode in users' daily mobile usage.

D. Discussions

In this section, we explored the effectiveness of Echo-FAS when allied with RGB-based FAS models. In the two-modality FAS experiments, Echo-FAS played a complementary role to RGB data, and it could consistently boost the face liveness detection performance.

VIII. Conclusions and Future Work

This paper presented a novel acoustic-based framework to tackle the face anti-spoofing problem. We first collected a large-scale, high-diversity, and acoustic-based face anti-spoofing database, Echo-Spoof. Based upon this database, we devised a tow-branch Echo-FAS architecture to synchronously capture global and local frequency clues from input acoustic signals. Extensive experimental results showed that the designed framework performs face liveness detection under diverse experimental settings. Our dataset will be freely available (under a license agreement) on Zenodo upon acceptance.

The proposed acoustic-based FAS system can be handily allied with RGB-based FAS systems to conduct a more secure and robust face anti-spoofing. Acoustic data can effectively reflect much depth information about input faces, largely ignored in the RGB modality. Extensive multi-modality experimental results demonstrated that using the auditory modality could effectively mitigate domain gaps in the RGB modality and improve the final FAS performance.

An ablation study further demonstrated that different frequency groups indeed contributed to the final FAS accuracy. The two-branch architecture effectively learned global and local joint representations of input signals, which further improved the final classification. Another ablation study showed that the scheme of the cross-attention module indeed improves the framework's face liveness detection performance.

The proposed Echo-FAS can provide new insights regarding developing FAS systems for mobile devices, which takes advantage of built-in sensors in the device and bring to bear algorithms capable of harnessing all their capabilities toward more secure and transparent authentication systems.

While our proposed method is effective for the face anti-spoofing (FAS) task, we limit our scope herein to 2D presentation attacks. Adapting our system to 3D face presentation attacks is worth investigating in future work. We envision that the acoustic signal could also capture informative clues between the 3D masks and live person, as mask materials and skin reflectance characteristics should be distinctive. In turn, it is also interesting to further apply domain generalization algorithms to improve the Echo-FAS system's generalization capability.

IX. Acknowledgement

A. Rocha thanks the support of São Paulo Research Foundation (Fapesp) under grant #2017/12646-3.

References

[1] F. Tari, A. A. Ozok, and S. H. Holden, "A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords," in *Proceedings of the second symposium on Usable privacy and security*, 2006, pp. 56–66.

[2] C. Stein, C. Nickel, and C. Busch, "Fingerphoto recognition with smartphone cameras," in *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE, 2012, pp. 1–12.

[3] J. G. Daugman, "Biometric personal identification system based on iris analysis," Mar. 1 1994, uS Patent 5,291,560.

[4] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[5] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE transactions on information forensics and security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[6] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 international joint conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–7.

[7] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp- top based countermeasure against face spoofing attacks," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 121–132.

[8] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.

[9] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[10] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3557–3560.

[11] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

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

[13] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

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

[15] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "Drl-fas: a novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 937–951, 2020.

[16] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 974–11 981.

[17] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 650–651.

[18] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[19] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2019.

[20] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

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

[22] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5042–5051.

[23] A. Pinto, S. Goldenstein, A. Ferreira, T. Carvalho, H. Pedrini, and A. Rocha, "Leveraging shape, reflectance and albedo from shading for face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3347–3358, 2020.

[24] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability," *Computer Vision and Image Understanding*, vol. 197, p. 102988, 2020.

[25] A. Agarwal, A. Sehwag, M. Vatsa, and R. Singh, "Deceiving the protector: Fooling face presentation attack detection algorithms," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–6.

[26] A. Agarwal, A. Sehwag, R. Singh, and M. Vatsa, "Deceiving face presentation attack detection via image transforms," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 373–382.

[27] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "Ultragesture: Fine-grained gesture sensing and recognition," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1—9.

[28] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, 2017, pp. 15—28.

[29] B. Zhou, Z. Xie, Y. Zhang, J. Lohokare, R. Gao, and F. Ye, "Robust human face authentication leveraging acoustic sensing on smartphones," *IEEE Transactions on Mobile Computing*, pp. 1–16, 2021.

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

[31] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: Anti-spoofing facial authentication using cots rfid," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[32] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 290–306.

[33] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan, "Ppgsecure: Biometric presentation attack detection using photopletysmograms," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 56–62.

[34] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3d mask face anti-spoofing with remote photoplethysmography," in *European Conference on Computer Vision*. Springer, 2016, pp. 85–100.

[35] W. Liu, X. Wei, T. Lei, X. Wang, H. Meng, and A. K. Nandi, "Data fusion based two-stage cascade framework for multi-modality face anti-spoofing," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[36] J. Seo and I.-J. Chung, "Face liveness detection using thermal face-cnn with external knowledge," *Symmetry*, vol. 11, no. 3, p. 360, 2019.

[37] O. Nikisins, A. George, and S. Marcel, "Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[38] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2016, pp. 1515—1525.

[39] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, 2016, pp. 82–94.

[40] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," pp. 1–18, 2018.

[41] Y. Zhang, W.-H. Huang, C.-Y. Yang, W.-P. Wang, Y.-C. Chen, C.-W. You, D.-Y. Huang, G. Xue, and J. Yu, "Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands," pp. 1–26, 2020.

[42] L. Wu, J. Yang, M. Zhou, Y. Chen, and Q. Wang, "Lvid: A multimodal biometrics authentication system on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1572–1585, 2019.

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

[44] 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.

[45] Y. Yang, C. Wang, Y. Chen, and Y. Wang, "Echolock: Towards low effort mobile user identification," *arXiv preprint arXiv:2003.09061*, 2020.

[46] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F-Radar and Signal Processing*, vol. 139, no. 5. IET, 1992, pp. 343–350.

[47] "Guideline of android platforms. [online]," http://developer.android.com/reference/android/media/AudioRecord.html.

[48] A. Das, N. Borisov, and M. Caesar, "Do you hear what i hear? fingerprinting smart devices through embedded acoustic components," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 441–452.

[49] D. Purves and S. M. Williams, *Neuroscience*. Sinauer Associates, 2001.

[50] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 525–536.

[51] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library." in *NeurIPS*, 2019.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[54] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *In International Symposium on Music Information Retrieval*. Citeseer, 2000.

[55] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1. IEEE, 2002, pp. 113–116.

[56] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[57] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[58] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[59] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[60] N. S. Altman, "An introduction to kernel and nearest-neighbor non-parametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[61] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[62] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2639–2652, 2018.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[64] D. Han, Y. Chen, T. Li, R. Zhang, Y. Zhang, and T. Hedgpeth, "Proximity-proof: Secure and usable mobile two-factor authentication," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 401–415.

[65] D. Chen, N. Zhang, Z. Qin, X. Mao, Z. Qin, X. Shen, and X.-Y. Li, "S2m: A lightweight acoustic fingerprints-based wireless device authentication protocol," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 88–100, 2016.

[66] Z. Zhou, W. Diao, X. Liu, and K. Zhang, "Acoustic fingerprinting revisited: Generate stable device id stealthily with inaudible sound," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 429–440.

[67] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[68] R. Cai, Z. Li, R. Wan, H. Li, Y. Hu, and A. C. Kot, "Learning meta pattern for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1201–1213, 2022.