EchoFence: Non-Intrusive Forgery Detection in Video Conferencing via Ultrasonic Sensing

Leqi Zhao*, Luxin Shi*, Jianwei Liu*†, Rui Xiao‡§, and Jinsong Han*
*Zhejiang University, †Hangzhou City University, ‡Shanghai University of Finance and Economics,
§MoE Key Laboratory of Interdisciplinary Research of Computation and Economics

***Abstract*—Real-time video conferencing is increasingly prevalent in everyday life, yet it faces growing threats from forgery attacks where synthetic or replayed videos are injected into live calls, potentially leading to serious privacy breaches or financial losses. Unfortunately, existing approaches remain inadequate for detecting such attacks in a reliable, efficient, and user-friendly manner. In light of this, we propose EchoFence, a non-intrusive, lightweight, and robust framework for authenticating live video streams. EchoFence actively emits imperceptible ultrasonic signals during video conferencing, which physically interact with user's natural facial and body movements and are captured by the microphone. Motion-related features are then extracted from both the ultrasonic responses and the video frames, and a training-free cross-modal verification strategy is employed to assess their temporal coherence. Significant misalignment between the two modalities is taken as strong evidence of forgery. Additionally, each ultrasonic signal carries a random credential via frequency modulation, which is validated through template-based matching, preventing tampering attempts involving ultrasound replay or removal. Extensive experiments show that EchoFence effectively detects three representative types of video forgeries with over 94% accuracy, and remains robust under diverse conditions, making it a practical solution for trustworthy video conferencing.**

***Index Terms*—forgery detection, cross-modal consistency, ultrasonic sensing, video conferencing**

I. Introduction

Video conferencing has become an essential part of daily life, enabling remote communication, collaboration, and identity verification. While these systems offer convenience and accessibility, they also introduce serious security concerns. With the rise of social media and advanced deepfake generation techniques [1], adversaries can easily obtain video footage or synthesize highly realistic videos of a target individual. These enable forgery attacks that impersonate the legitimate user during video conferencing, posing significant threats to privacy and financial safety. A recent incident, where the attackers posed as the chief financial officer using deepfake techniques and scammed out of 25 million dollars via a video call [2], has highlighted the real-world risk of forgery attacks.

To defend against such attacks, existing solutions can be broadly categorized into passive and active approaches. Passive methods analyze audio-visual content to detect cues of manipulation, including resolution-related artifacts [3], [4], frequency-domain inconsistencies [5], [6], spatiotemporal anomalies [7], [8], and lip-audio mismatch [9], [10]. However, as manipulation techniques advance, forgeries exhibit fewer detectable artifacts [11]. Meanwhile, lossy compression, common in real-time video conferencing, can further degrade or remove these subtle cues [12]. Finally, passive methods are generally ineffective against replay attacks, as they lack explicit indicators to distinguish live content from pre-recorded footage of legitimate users.

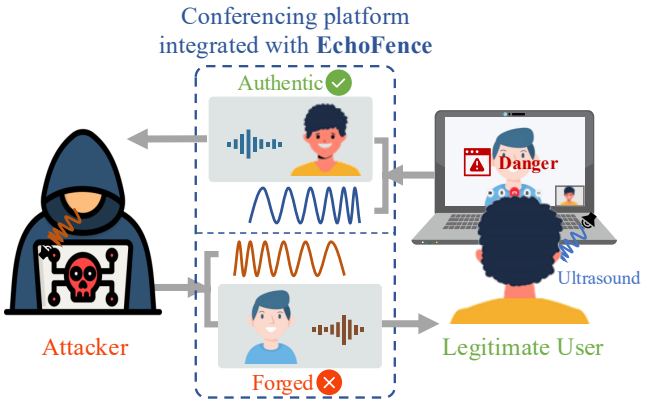


Fig. 1: An application scenario of EchoFence. An attacker attempts to substitute the live video with forged content. The conferencing platform, equipped with EchoFence, emits imperceptible ultrasonic signals and records synchronized audio-visual streams. It performs local verification on the ultrasonic responses and video streams to detect such video forgery attacks and ensure authenticity.

These limitations have motivated the development of active methods, which aim to enhance robustness by introducing controlled stimuli or challenge-response prompts during live interaction. These stimuli introduce verifiable signals that are harder to forge, remain detectable under compression, and can distinguish live users from replayed content. For example, some approaches instruct users to perform specific physical or verbal actions, such as facial expressions or spoken passphrases, and verify them through audio-visual contents or other sensor feedback [13]–[15]. Others project dynamic visual patterns onto the screen and verify them via corneal or facial reflections [16], [17]. While these methods show promise under controlled conditions, they suffer from several key limitations: **1) Intrusiveness:** Active methods rely on explicit user cooperation, such as performing prompted actions, or involve

Jianwei Liu and Rui Xiao are corresponding authors.

displaying full-screen dynamic patterns that obscure the call interface. Such interventions are disruptive to the natural flow of video conferencing and are therefore unsuitable, especially in professional video conferencing settings. **2) Sensitivity to occlusion and behavior variation:** Many methods focus narrowly on facial regions, making them fragile when the face is partially occluded (e.g., by hand gestures) or when the user exhibits non-standard behaviors.

In light of this, we pose the following research question: *Is it possible to design a novel framework that can robustly identify both synthetic and replay-based attacks, without requiring user cooperation or disrupting the natural video conference experience?* To this end, we propose EchoFence, a non-intrusive yet robust forgery detection framework tailored for trustworthy video conferencing. Our key idea is to actively probe the physical world using imperceptible ultrasonic signals to verify the authenticity of the video, with no need for user cooperation. Specifically, in a genuine video stream, user movements, such as head turns or hand postures, are physically performed and result in temporal fluctuations in the reflected ultrasonic signals. These fluctuations remain temporally coherent with the motion captured in the video stream. In contrast, forged content, whether synthesized or replayed, lacks such real-world physical coherence, leading to observable inconsistencies between audio and visual modalities. EchoFence detects such inconsistencies for forgery detection.

EchoFence's design relies solely on built-in microphones, speakers, and cameras. Therefore, it can be seamlessly integrated into mainstream video conferencing platforms (e.g., Zoom or Teams) as a plug-and-play module. As illustrated in Fig. 1, when users communicate through a trusted software equipped with EchoFence, the software emits ultrasonic probes and simultaneously records the reflected audio along with the video stream during the call. Verification is performed locally on the sender side. If a forgery is detected, the platform can immediately alert the remote party, mitigating potential security risks. The inaudible ultrasonic artifacts are removed before forwarding the streams to the other side, preventing potential interference. EchoFence is resilient to both synthetic and replay-based forgeries, while preserving a seamless and natural user experience. Unlike previous active approaches, it introduces no visual distractions and operates transparently in the background, making it suitable for real-world deployments.

Designing EchoFence involves three main challenges:

a) Reliable motion estimation under non-cooperative and unconstrained conditions: In contrast to existing active schemes that rely on users cooperatively performing predefined gestures, EchoFence must operate in non-cooperative settings where participants may exhibit only subtle and unconstrained motions, such as slight head tilts or casual hand shifts. Frequent occlusions or motions occurring partially off-frame can further degrade visual tracking, posing a significant challenge for reliable motion sensing. To address this, we leverage frequency-modulated ultrasonic chirps to capture fine-grained upper-body motion, including subtle movements. In parallel, we extract visual motion using optical flow analysis, which covers a broader body region beyond the face and is tolerant to occlusions and partially out-of-frame motion. The combination yields consistent and robust cross-modal motion observations in unconstrained video conferencing scenarios.

b) Cross-modal verification without modality-specific training: Prior detection methods based on cross-modal validation typically require large annotated datasets to learn complex correspondences between high-level cues, such as speech content and lip movements, which is labor-intensive and may not be practical in real-world settings. To address this challenge, EchoFence converts ultrasonic echo fluctuations and visual frame differences into a unified motion-energy signal. This allows verification to be formulated as a lightweight signal-matching task, therefore alleviating the requirement of semantic alignment or modality-specific training.

c) Ensuring replay resilience while maintaining sensing utility: To defend against replay attacks, it is essential to verify that the ultrasonic response is captured in real time, rather than reused from prior recordings. Our key idea is to embed a randomized credential, conceptually similar to a nonce, into the ultrasonic signal. However, introducing randomness must be done carefully, as it can disrupt the signal structure essential for accurate motion sensing. To balance both goals, EchoFence encodes the credential into the sensing waveform by allocating a randomized, time-varying frequency band for frequency-modulated continuous-wave (FMCW) modulation. The system then performs adaptive motion sensing over these structured chirp sequences. This design preserves the temporal and spectral integrity required for robust motion estimation while enabling reliable detection of replayed content.

We implement and evaluate EchoFence on commercial off-the-shelf devices by capturing real-world conversation sessions totaling over 12 hours. We comprehensively evaluate EchoFence's performance under various conditions, including different forgery methods, device placements, and video resolutions. Overall, EchoFence achieves an accuracy of over 94% in detecting both synthetic and replay-based forgeries, significantly outperforming the state-of-the-art systems, while maintaining a non-intrusive and seamless user experience.

In summary, our contributions are as follows:

- We propose a non-intrusive forgery detection framework for video conferencing that operates seamlessly without requiring user cooperation or disruptive visual patterns.
- We extract reliable motion features from ultrasonic responses and visual frames, and design a lightweight strategy for cross-modal consistency verification.
- We embed randomized authentication credentials into ultrasonic signals via frequency modulation, ensuring protection against replay attacks without compromising motion sensing capability.
- We conduct extensive real-world experiments, demonstrating the effectiveness and robustness of our framework in detecting both synthetic and replay-based forgeries.

II. System and Threat Model

A. System Model

EchoFence aims to detect forgery attacks—including synthetic and replay-based ones—in real-time video conferencing by leveraging imperceptible ultrasonic signals. To support practical deployment, the system must satisfy the following requirements: 1) compatibility with commodity devices; 2) unobtrusive operation without disrupting normal conversation; and 3) robustness across diverse users and conditions.

We envision a typical video conferencing setup, where online participants communicate through built-in speakers, microphones, and cameras. The conferencing platform integrates EchoFence, which emits ultrasonic signals and simultaneously records audio-visual streams. By analyzing the ultrasonic responses and visual content, EchoFence detects forgery attacks without degrading the user experience.

B. Threat Model

We consider adversaries who aim to spoof their identity by manipulating the video and audio streams in a video conferencing scenario. Based on the attacker's awareness of the ultrasonic verification mechanism and how they address the ultrasound, we define three representative attack types:

- **Blind Forgery Attack:** The attacker is unaware of the ultrasonic verification mechanism and injects synthetic or replayed audio-visual streams to impersonate a legitimate user. Since the microphone input is fully overridden, no meaningful ultrasonic response is captured.
- **Hybrid Forgery Attack:** The attacker is aware of the ultrasonic verification mechanism and allows live ultrasonic responses to be recorded. Meanwhile, the visual stream is forged through synthetic generation or recording replay, possibly accompanied by synchronized speech to preserve lip-sync consistency.
- **Full Replay Attack:** The attacker replays previously recorded synchronized audio-visual streams, including valid ultrasonic responses, to simulate a legitimate user, preserving high cross-modal consistency.

III. System Overview

EchoFence's design enables non-intrusive forgery detection by leveraging cross-modal physical consistency between acoustic and visual motion cues. Unlike traditional audio-visual watermarking or digital signatures, our system injects randomized ultrasonic signals into the live audio stream to actively probe and verify the motion presented in video. We define a *session* as a specific time window in an ongoing video call, during which EchoFence emits ultrasonic signals and verifies the video content. As illustrated in Fig. 2, our approach leverages the following four modules to secure the conference within each session:

Anti-Replay Ultrasound Modulation. Before each verification session, we modulate the ultrasonic signal to empower it with the capability of replay detection. To this end, we generate a randomized one-time credential sequence, and embed it into ultrasonic chirps via frequency modulation, ensuring each session features an unpredictable chirp sequence. At the start of the session, the modulated ultrasonic signal is emitted into the live conference, physically interacting with the surrounding environment and generating verifiable responses that capture human motion back to the microphone.

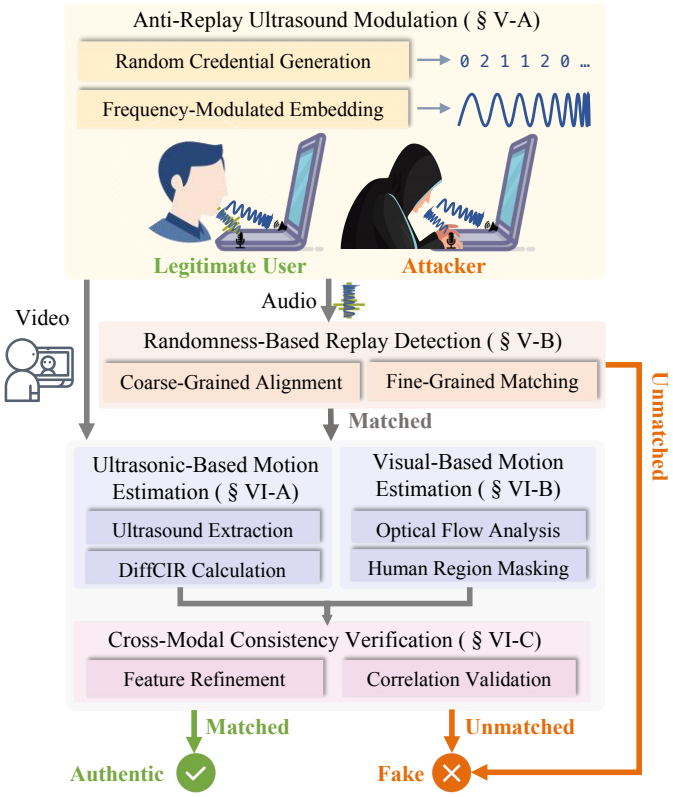


Fig. 2: System overview. EchoFence emits inaudible ultrasonic signals with embedded credentials to capture physical motion, recorded in sync with the video. Replay-based and synthetic forgeries are detected by verifying both the embedded credential and the ultrasound-visual motion consistency.

Randomness-Based Replay Detection. After receiving the acoustic responses, we perform replay detection by first identifying the signal segment containing the modulated ultrasound and then validating it against the pre-embedded random credential. Missing or mismatched signals—commonly observed in blind forgery or full replay—are flagged as attacks before proceeding to dual-modal motion tracing.

Dual-Modal Motion Quantification. For audio-video streams that pass replay detection, the system extracts motion cues from both modalities. On the audio side, we recover the physical reflections of the chirp signals and compute the differential Channel Impulse Response (DiffCIR), which captures changes in backscattered energy induced by human motion. On the video side, we apply dense optical flow estimation to derive frame-level motion magnitude, further enhanced by human-region masking to eliminate background interference. Both modalities independently capture temporal variations

corresponding to the same physical movement.

Cross-Modal Consistency Verification. To determine whether the two modalities match, we evaluate the temporal correlation between audio-derived and video-derived motion energy. In genuine recordings, these sequences exhibit strong synchronization, as they stem from the same physical human motion. In contrast, manipulated videos, whether replayed or synthesized, often disrupt this physical consistency, leading to temporal misalignment or reduced correlation. We employ correlation-based metrics to quantify this alignment, enabling robust detection without the need for training or labeled data.

By embedding an environment-sensitive yet unobtrusive physical probe into live communication and verifying its response through credential check and dual-modality analysis, our system offers a practical and robust method for detecting forgeries in real-world video conferencing scenarios.

IV. Replay-Resistant Modulation and Validation

EchoFence builds on an active probing of the physical world, thus it is crucial to ensure that the physical signals leveraged are themselves robust against replay-based forgery. To this end, we design a randomized ultrasonic modulation scheme that introduces a verifiable and dynamic physical stimulus into the conference. This embeds a one-time random sequence as a security credential into the live audio stream without disrupting normal conversation. By verifying the consistency between the random sequence embedded in received ultrasonic responses and the authentic security credential, we can determine whether the session is a replay.

A. Frequency-Modulated Randomness Embedding

To introduce an additional acoustic signal for trustworthy video conferencing, it is essential to ensure that it provides accurate motion capture and does not interfere with normal live communication. To achieve this, we leverage FMCW technique, and set the frequency range to an inaudible band for humans, i.e. 18–23 kHz. Moreover, it remains within the operating range of commodity 48 kHz audio hardware on laptops and smartphones, allowing for reliable emission and reception of the ultrasonic signals.

To facilitate the introduction of randomness, the ultrasonic signal is structured as a sequence of M units ($M \geq 10$), each lasting a fixed duration. Within each unit, N chirp cycles (e.g., $N = 20$) are transmitted sequentially. As illustrated in Fig. 3, all chirps within the same unit share the same frequency configuration, while that vary across different units based on a random credential sequence:

$$\text{random}[i] \in 0, 1, 2, \quad i = 1, \ldots, M \tag{1}$$

Each chirp starts at 18 kHz and linearly sweeps to:

$$f_{\text{end}}^{(i)} = 21\,\text{kHz} + 1\,\text{kHz} \cdot \text{random}[i] \tag{2}$$

This results in three discrete chirp variants with end frequencies of 21, 22, and 23 kHz, respectively, all with fixed durations. Rather than switching parameters on a per-chirp basis, this structure preserves the coherent temporal pattern required for robust motion sensing.

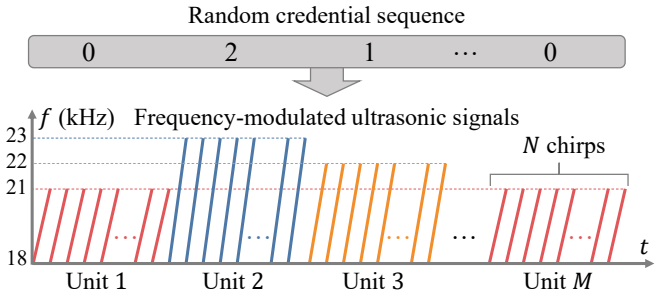


Fig. 3: Illustration of embedding a random credential into ultrasonic signals via frequency modulation.

In this way, the ultrasonic signal within a complete session encodes M independent selections from a three-valued set, resulting in a total encoding space of 3^M for the random credential. Since $M \geq 10$, the probability of an attacker blindly guessing the correct pattern is no greater than $\frac{1}{3^{10}}$.

B. Template-based Credential Verification

To defend against replay-based forgeries, EchoFence verifies whether the received ultrasound matches the originally embedded credential sequence. A natural yet naive approach would be extracting the frequency components of the recorded audio via time-frequency analysis. However, this is inherently unreliable in real-world settings, as it suffers from limited resolution of short-time windows and is highly susceptible to noise and reverberation, often resulting in high false positive or false negative rates. To address this, we propose a template-based credential matching method, which performs accurate replay detection through a two-stage process comprising coarse-grained alignment and fine-grained matching.

Coarse-grained alignment. Since the length of the received audio signal may exceed that of the emitted ultrasonic signal, we first perform coarse-grained alignment to determine the approximate location of the embedded ultrasonic sequence. Specifically, EchoFence constructs a template waveform of the complete ultrasonic sequence based on the known credential. Cross-correlation is then applied between the recorded audio signal and the template, with the highest peak in the correlation curve indicating the coarse starting point of the ultrasonic sequence. Using this offset, we extract M units from the audio signal, each slightly longer than the nominal unit duration to account for possible temporal misalignment.

Fine-grained matching. For the i-th unit, EchoFence refers to the known credential random$[i]$ and generates a chirp template with corresponding end frequency. A high-resolution correlation is performed to compare each unit with its expected template, and the following three metrics are employed:

- Peak-to-Noise Ratio (PNR) measures peak prominence over background fluctuations:

$$\text{PNR} = \frac{p}{\sigma} \tag{3}$$

where p is the correlation peak amplitude, and σ is the standard deviation of off-peak noise.

- Peak-to-Mean Ratio (PMR) reflects peak dominance over the average correlation baseline:

$$\text{PMR} = \frac{p}{\mu} \tag{4}$$

where μ is the average value of the correlation curve.

- Peak Interval Error (PIE) measures the deviation of the detected peak intervals from the expected chirp duration:

$$\text{PIE} = \frac{|\Delta t - T_c|}{T_c} \tag{5}$$

where Δt denotes the time interval between two adjacent correlation peaks, and T_c is the nominal chirp period.

A unit is considered a match with the authentic template if these three metrics pass corresponding thresholds for at least 90% peaks. The session passes only if all M units are matched.

In real-time video conferencing, the credential is randomly generated and embedded into the ultrasonic signals during the live session, making it impractical for attackers to accurately predict the ultrasound sequences. Any attempt to inject forged ultrasonic patterns—whether by replaying past signals or pre-synthesizing new ones—fails to pass this randomness-based credential verification. Additionally, cases where the ultrasonic signal is absent (i.e. in blind forgery attacks) can be easily detected and rejected at this stage without proceeding to subsequent phases. As a result, this approach serves as a reliable gatekeeper for subsequent cross-modal coherence validation.

V. Cross-Modal Motion Consistency Verification

After passing the replay detection, EchoFence proceeds with further security verification by extracting shared information from both the audio and video streams. During video conferencing, human motion in front of the screen is captured by video frames, while the corresponding acoustic reflections inherently encode the same dynamic behavior. This observation motivates us to extract motion features from both modalities and perform cross-modal consistency analysis.

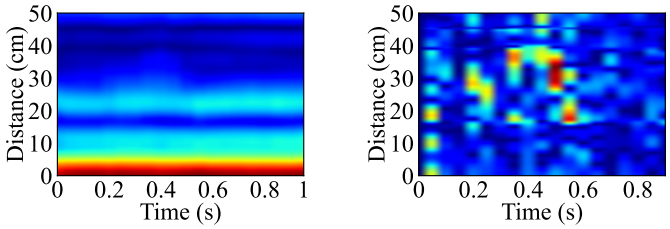
A. Motion Features in Audio Stream

Since the ultrasonic signal employs FMCW technology, which enables accurate channel estimation, we leverage it—rather than audible components—to extract dynamic motion information for verification purposes.

Specifically, during each verification session, EchoFence transmits ultrasonic chirps that reflect off the surrounding environment and are subsequently captured by the microphone. Each received chirp can be modeled as:

$$r(t) = s(t) * h(t), \quad t \in [0, T_{\text{chirp}}] \tag{6}$$

where $s(t)$ denotes the transmitted chirp signal and $h(t)$ represents the Channel Impulse Response (CIR), which characterizes multipath reflections with varying delays and amplitudes. These reflection patterns are highly sensitive to facial and upper-body movements, thereby providing a motion-dependent acoustic signature.



(a) CIR (b) DiffCIR

Fig. 4: Visualization of ultrasound CIR and DiffCIR for a user uttering the word "zero".

To recover CIR, we first apply a high-pass filter to suppress audible-band interference. Channel estimation is then performed through signal mixing and Fast Fourier Transform (FFT). The estimated CIR is discretized into a fixed number of range bins, yielding a 1D profile $h(d)$ for each chirp. The distance resolution of the chirp is given by:

$$\Delta d = \frac{c}{2B} \tag{7}$$

where c is the speed of sound and B is the chirp bandwidth. By stacking the CIRs of successive chirps, we construct a CIR matrix $C_t(d)$, where t indexes the chirp frame and d denotes the physical distance.

To focus on motion-induced changes, we compute the frame-wise difference between consecutive CIRs, forming the differential CIR (DiffCIR) $\Delta C_t(d)$. We retain only the portion within a selected distance range (e.g., 50 cm), ensuring that the magnitude of the DiffCIR is primarily determined by user's upper-body motion. Fig. 4 illustrates the CIR and DiffCIR extracted from ultrasonic signals as a user uttering "zero".

As described in Section IV, we employ chirps with varying end frequencies, which results in different distance resolutions in the DiffCIR according to (7). To obtain a temporally compact representation, we aggregate the DiffCIR magnitude across all selected distance bins at each time frame:

$$E_{\text{aud}}^{(t)} = \sum_d |\Delta C_t(d)|, \quad \mathbf{E}_{\text{aud}} = [E_{\text{aud}}^{(1)}, E_{\text{aud}}^{(2)}, \ldots, E_{\text{aud}}^{(T)}] \tag{8}$$

This produces an audio-domain motion energy sequence $\mathbf{E}_{\text{aud}} \in \mathbb{R}^T$, which captures the cumulative changes in backscattered energy, effectively reflecting the intensity of user movement over time.

B. Motion Features in Video Frames

To extract motion features from video frames, an intuitive approach might be to apply facial landmark detection (e.g., 68-point facial alignment) and compute the displacement of landmarks between consecutive frames. While straightforward, this method has several limitations: 1) it captures only coarse motion at predefined anchor points, thereby missing subtle and non-rigid movements (e.g., in the cheeks or neck); and 2) its accuracy degrades significantly under pose variations, occlusions, or lighting changes, leading to instability in the extracted motion features.

To address the above limitations, we adopt optical flow for dense motion estimation, which computes per-pixel displacements between consecutive frames by analyzing temporal changes in pixel intensity. This approach enables the capture of subtle and non-rigid facial and upper-body movements with improved spatial coverage and robustness. Given two consecutive video frames I_t and I_{t+1}, we employ the pretrained RAFT model [18] to compute a dense optical flow field:

$$F_t(x, y) = (u_t(x, y), v_t(x, y)) \tag{9}$$

where (u_t, v_t) denotes horizontal and vertical displacement.

To quantify instantaneous motion strength, we compute the motion magnitude map as:

$$M_t(x, y) = \sqrt{u_t^2(x, y) + v_t^2(x, y)} \tag{10}$$

This motion field provides a high-resolution, dense representation of all observable movements, overcoming the sparsity and instability issues associated with landmark-based methods.

To ensure that the extracted motion features reflect user activity rather than irrelevant background motion or compression artifacts, we apply spatial masking to isolate the user region. Specifically, a binary mask $B(x, y)$ is generated using a pretrained Mask R-CNN [19] to segment the prominent human subject from the background. The refined motion magnitude map is then given by:

$$\tilde{M}_t(x, y) = M_t(x, y) \cdot B(x, y) \tag{11}$$

This masking operation effectively removes background disturbances , ensuring that the extracted motion features better correspond to genuine user dynamics, as illustrated in Fig. 5.

To align the visual motion features structurally with those derived from the audio modality, we extract a scalar-valued motion energy per frame from motion magnitude maps. This is achieved by spatially aggregating the masked motion field:

$$E_{vid}^{(t)} = \sum_{x,y} \tilde{M}_t(x, y), \quad \mathbf{E_{vid}} = [E_{vid}^{(1)}, E_{vid}^{(2)}, \ldots, E_{vid}^{(T)}] \tag{12}$$

This temporal energy sequence $\mathbf{E_{vid}}$ serves as a compact representation of visual motion and facilitates subsequent matching with audio-based features.

C. Consistency Verification

So far, we have obtained two temporal sequences, $\mathbf{E_{aud}}$ and $\mathbf{E_{vid}}$, which represent motion energy derived from ultrasonic responses and visual frames, respectively. This scalar-level aggregation enhances robustness by mitigating the impact of factors commonly encountered in real-world deployment scenarios, including sensor-view misalignment, spatially localized noise, and device-dependent variability. In genuine, unmanipulated videos, these sequences are expected to exhibit strong temporal alignment, as shown in Fig. 6, since they are governed by the same underlying physical motion.

Instead of relying on a learning-based model that requires extensive supervision, we propose a lightweight and interpretable validation module based on classical signal processing principles. This approach is computationally efficient and well-suited for resource-constrained devices.

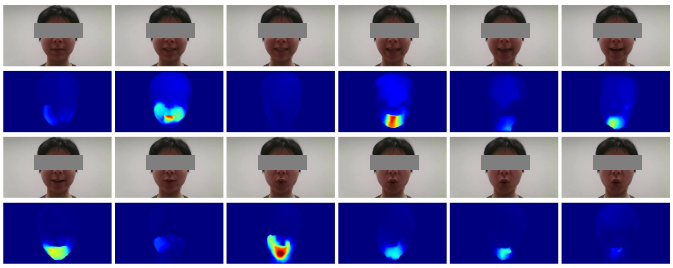
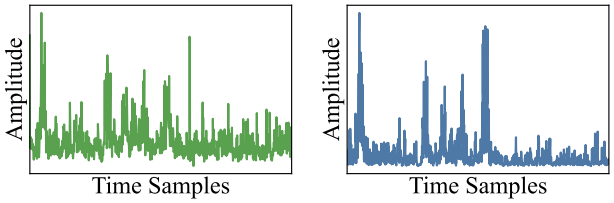


Fig. 5: Visualization of consecutive video frames and corresponding motion magnitude maps from a user uttering "zero".



(a) Acoustic Motion Energy (b) Visual Motion Energy

Fig. 6: Visualization of motion energy features extracted from ultrasonic responses and visual frames.

To assess the coherence between dual-modal motion features, we first normalize both sequences and apply a Savitzky-Golay filter to suppress high-frequency noise while preserving global motion trends. Then we employ a sliding-window strategy to locally evaluate temporal alignment. Specifically, each session is divided into overlapping windows of fixed length L, with stride s. For each window, we compute the full cross-correlation between audio and visual energy sequences:

$$corr(\tau) = \sum_{t} E_{aud}^{(t)} \cdot E_{vid}^{(t+\tau)} \tag{13}$$

where τ denotes the lag index.

Two key metrics are extracted to quantify alignment quality:

- Best Lag (τ^*): The temporal offset that yields the highest correlation value.
- Peak-to-Noise Ratio (PNR): A measure of alignment clarity, as defined in (3).

A window is considered *matched* if both τ^* and PNR exceed empirically determined thresholds. If the number of matched windows surpasses a predefined fraction of the total, the entire session is classified as motion-consistent.

This validation process offers a robust and low-complexity method to assess video authenticity by measuring cross-modal motion coherence, while remaining tolerant to transient noise and minor temporal misalignment.

VI. Evaluation

This section presents a comprehensive evaluation of EchoFence through real-world experiments, demonstrating its effectiveness and robustness in practical scenarios. All the

experiments are conducted with approval from the Institutional Review Board (IRB).

A. Experimental Setup

Platform. We implement EchoFence on a commercial off-the-shelf laptop (HUAWEI MateBook X Pro [20]) equipped with built-in speakers, microphones, and a camera. To simulate real-time video conferencing scenarios, we simultaneously record audio-visual streams using OBS Studio [21]. The video frame rate is set to 20 FPS, and the audio sampling rate is 48 kHz. Each symbol in the randomized credential is mapped to a chirp unit consisting of 20 chirps, with each chirp lasting 0.05 s.

Data Collection. We recruit five volunteers (two males and three females) and collect benign sessions totaling more than 12 hours, with each session lasting approximately 100 seconds. In each benign session, the volunteer sits naturally in front of the laptop and performs spontaneous actions such as speaking, nodding, or shaking their head, mimicking a typical online conversation. During recording, the laptop emits modulated ultrasonic signals embedded with credentials and simultaneously captures synchronized audio and video.

To simulate attacks, we construct samples corresponding to the three attack types defined in the threat model using the following methods:

- Blind Forgery Attack: In this case, the attacker employs forged videos that lack ultrasonic responses. Accordingly, we utilize recorded videos without embedded ultrasound, as well as synthetic videos generated using a face-swapping technique [22].
- Hybrid Forgery Attack: For this case, we combine forged videos (including synthetic and pre-recorded ones) with live-recorded ultrasonic responses. The video contains another person rather than the attacker behind the screen.
- Full Replay Attack: In this case, we reuse previously recorded benign samples, including ultrasonic responses.

Performance Metrics. To quantify the system's performance toward different cases, we compute the recognition accuracy individually for each case as follows:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{all}}} \tag{14}$$

where N_{correct} denotes the number of correctly recognized samples (accepted in benign cases and rejected in attack cases), and N_{all} represents the total number of test samples for that specific case.

B. Overall Performance

We evaluate the overall performance of EchoFence on the laptop across benign cases and three attack cases. In the default setting, the user sits 25 cm away from the device.

To establish a comparative baseline, we benchmark EchoFence against both a representative detection approach MesoNet [4] and a state-of-the-art method SONICUMOS [15]:

- MesoNet is visual-only deep learning model designed for detecting facial video forgeries.
- SONICUMOS is a liveness detection system that also utilizes ultrasonic chirps to probe user motion; however, it requires users to respond to prompted action challenges, imposing explicit constraints in a natural conversation.

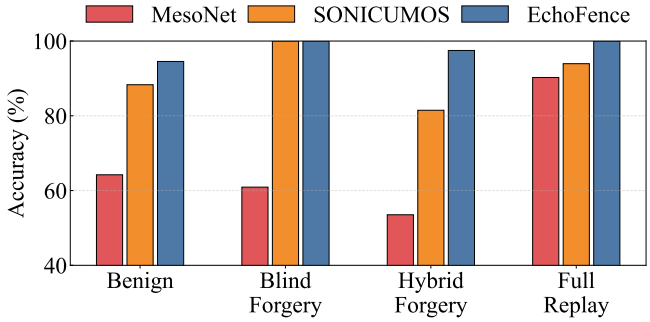


Fig. 7: Overall performance of EchoFence compared to two baselines: MesoNet and SONICUMOS.

In SONICUMOS, the original method segments audio-visual samples according to distinguished motion intervals. This approach is effective in its original setting, where only three action types are considered and users perform them on demand. However, it struggles in our non-intrusive, unconstrained scenario. Therefore, we adapt SONICUMOS by removing its motion segmentation module while retaining its feature extraction process and feature fusion network.

The performance of EchoFence and the two baseline methods is presented in Fig. 7. Since MesoNet relies solely on visual frames, it performs poorly in our setting, which includes replay-based forgeries. Although it appears to reject 90% of full replay samples, the high false rejection rate for benign cases highlights its limitations. SONICUMOS demonstrates high accuracy in detecting blind forgery and full replay attacks by leveraging randomized ultrasound. However, its effectiveness declines against hybrid forgery attacks. This is due to its original design for constrained user actions and reliance on specific facial landmarks, which proves inadequate in our non-intrusive scenario involving diverse subtle motions.

In contrast, EchoFence consistently achieves over 94% recognition accuracy across both benign and attack cases, outperforming both baselines. Notably, all samples from blind forgery and full replay attacks are successfully rejected through randomized credential validation alone, without requiring additional cross-modal verification.

C. Robustness Study

To evaluate the robustness of EchoFence under real-world variability, we assess its performance across different conditions, including changes in chirp frequency band, user-to-device distance, physical disturbances, and video resolution. Since blind forgery attacks contain no ultrasonic signals and are consistently rejected regardless of these factors, we exclude this case from the following analysis.

a) Impact of Chirp Frequency Band: As introduced in Section IV, we adopt three different frequency bands for credential embedding: 18–21 kHz, 18–22 kHz, and 18–23 kHz.

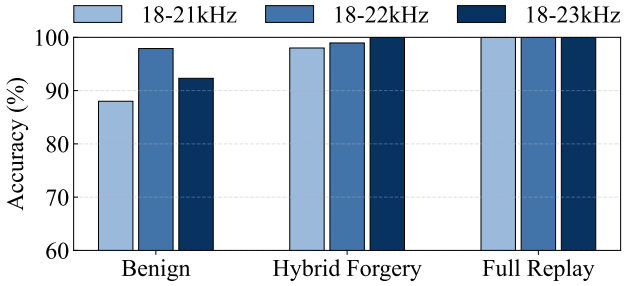


Fig. 8: Performance under different chirp frequency bands.

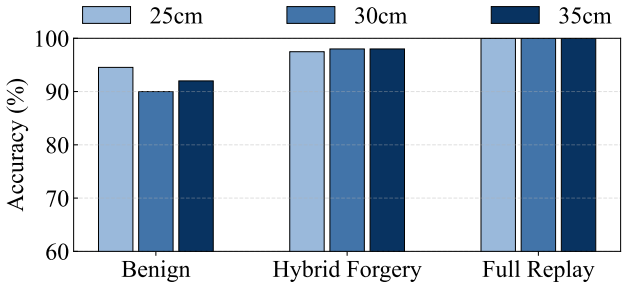


Fig. 9: Performance across different user-to-device distances.

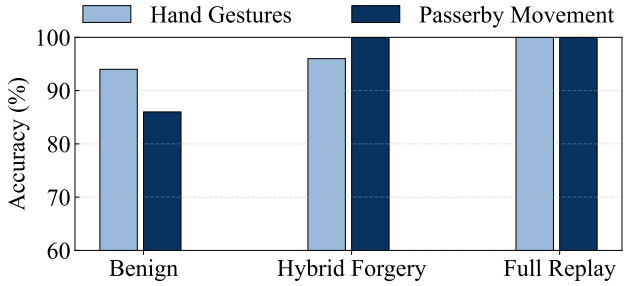


Fig. 10: Performance under physical disturbances.

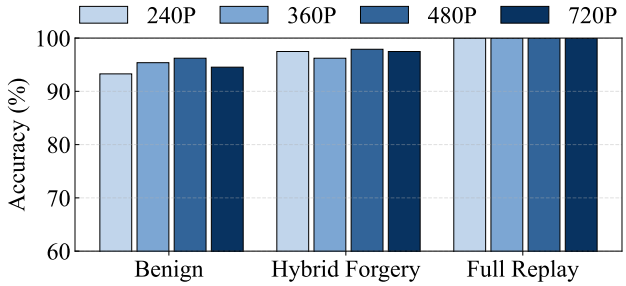


Fig. 11: Performance under different video resolutions.

The choice of frequency band can influence the resolution of ultrasonic sensing and the stability of signal transmission, both of which may affect detection performance. To assess this impact, we conduct controlled experiments using each fixed frequency band and evaluate the resulting detection accuracy.

As shown in Fig. 8, EchoFence maintains over 88% accuracy across all three frequency configurations, indicating that each is suitable for randomness-based modulation to ensure replay resistance. Among them, the 18–22 kHz band provides the most reliable performance, achieving a low false rejection rate of 2%. This may be because it strikes an optimal balance between sensing resolution and hardware capability: narrower bands reduce resolution, while wider bands may exceed the effective operating range of typical speakers and microphones, causing signal distortion and instability.

b) Impact of User-to-Device Distance: EchoFence performs verification using ultrasonic signals, which naturally attenuate as the distance between the user and the device increases. To evaluate this effect, we assess the system's performance at varying user-to-device distances, ranging from 25 cm to 35 cm.

As shown in Fig. 9, EchoFence maintains reliable performance across this range, achieving over 90% accuracy even at 35 cm. These results suggest that EchoFence is well-suited for typical video call scenarios, where the user is positioned within arm's length of the device.

c) Impact of Physical Disturbances: Since our method estimates motion from both acoustic and visual streams, physical disturbances may affect the system's robustness. We evaluate EchoFence's performance under two types of common disturbances:

- Hand gestures: In addition to head movements, users perform various hand gestures, such as spreading or waving their hands. These actions can influence both ultrasonic and visual signals.
- Passerby movement: Another person walks behind the user during the video session, introducing potential noise into both modalities.

As shown in Fig. 10, EchoFence maintains acceptable performance under both disturbance types. For hand gestures, since EchoFence captures human body motion rather than focusing solely on the head, it can still track consistent motion patterns across both ultrasonic responses and visual frames. In the case of passerby movement, EchoFence mitigates interference by limiting the DiffCIR extraction range and masking the optical flow of the nearest foreground individual. These strategies ensure that common, unintended physical disturbances do not significantly compromise recognition accuracy.

d) Impact of Video Resolution: Given the variability of video quality in real-world video conferencing, we evaluate EchoFence under four different video resolutions: 720P (1280×720), 480P (640×480), 360P (480×360), and 240P (320×240). As shown in Fig. 11, EchoFence maintains an accuracy of 93% even at 240P. These results suggest that EchoFence can reliably extract motion patterns from low-resolution frames, demonstrating its robustness in resource-constrained scenarios.

D. Device Scalability

EchoFence is designed to be hardware-agnostic, relying solely on off-the-shelf microphones, speakers, and cameras, without requiring device-specific training. This makes it inherently portable across a wide range of consumer devices.

To examine the device scalability, we implement an Android application to emit randomized ultrasonic chirps while record-

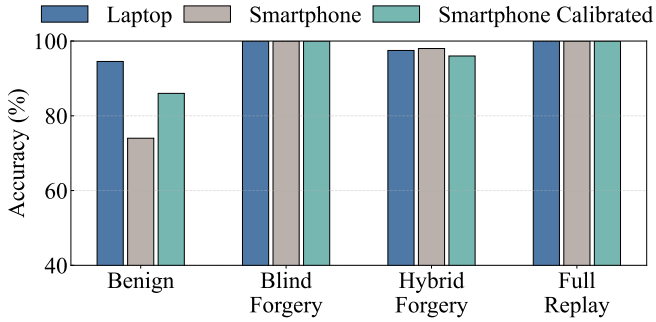


Fig. 12: Performance on a smartphone.

ing video and audio. We deploy it on a commodity smartphone (Samsung Galaxy A53 [23]) and collect session samples.

When using the same metric thresholds as in the laptop setting, EchoFence achieves 98% accuracy on forged samples and exhibits a lower true acceptance rate of 74% for benign ones, as shown in Fig. 12. Further analysis reveals that this drop primarily stems from audio-visual synchronization issues in our Android implementation. To mitigate this, we relax the acceptable range of the best cross-correlation lag τ^* by 500 ms in the cross-modal matching stage, which raises the benign acceptance rate to 86% while not compromising the forgery rejection rate. This suggests that EchoFence has potential for cross-device deployment, with only minor threshold calibration or synchronization compensation needed.

VII. Discussion

We now discuss EchoFence's deployment consideration. EchoFence currently assumes that users are on video conferencing with built-in speakers and microphones enabled. We now consider a common alternative audio setting: users may choose to use earphones or headsets for privacy or improved audio quality. In such cases, ultrasonic signals emitted from the earphones may not effectively propagate into the environment.

Fortunately, modern devices support simultaneous use of built-in speakers and earphones. This allows EchoFence to emit imperceptible ultrasonic chirps through the speakers while routing audible audio through the earphones. We validate this setup on a commodity laptop using audio routing software (e.g., VoiceMeeter [24]), confirming that EchoFence can remain effective in earphone settings without disrupting the user experience.

The EchoFence pipeline operates entirely on the local device, and thus factors related to network transmission do not directly affect the proposed detection mechanism. Looking ahead, we envision collaborating with video conferencing platforms to integrate EchoFence as a background verification module, ensuring robust forgery detection across diverse real-world scenarios.

VIII. Related Work

A. Video Forgery Detection

Existing video forgery detection approaches can be broadly categorized into passive and active ones. Passive methods analyze naturally captured content to detect manipulation. Visual-only approaches detect spatio-temporal artifacts [7], [8], [25], frequency-domain anomalies [5], [6], [26], and inconsistencies in facial features [3], [4], [27], [28]. Others leverage multi-modal cues, such as lip-audio synchronization [9], [10]. While effective in controlled settings, these methods mainly target synthetic artifacts and may fail against replay attacks.

Active methods enhance robustness by introducing controlled prompts or engineered stimuli. Some prompt users to perform specific actions, verifying liveness via audio-visual streams or sensor feedback [13]–[15]. Others exploit corneal or facial reflections by projecting dynamic screen patterns [16], [17], [29]. While effective, such techniques often require user cooperation or introduce visual distractions, reducing their practicality in video conferencing. EchoFence falls under the category of active detection, but distinguishes itself through its non-intrusive design and seamless integration into video conferencing platforms.

B. Ultrasound-Based Sensing

Ultrasound has been widely adopted in human sensing tasks due to its fine-grained motion sensitivity and compatibility with commodity hardware. Prior works have utilized ultrasonic signals for gesture recognition [30]–[34], respiration monitoring [35], [36], and face or lip tracking [37]–[41]. These efforts have demonstrated the sensing potential of ultrasound in benign scenarios. However, they primarily aim to extract accurate motion features under varying environmental conditions, not considering adversarial scenarios where the signals may be forged or replayed. In contrast, our work extends ultrasound from a sensing modality to a verification mechanism, leveraging signal randomization and cross-modal motion consistency to detect tampering and resist spoofing.

IX. Conclusion

We present EchoFence, a novel non-intrusive forgery detection framework tailored for video conferencing, which is practical for seamless integration into existing platforms. By embedding random credentials into ultrasonic signals and verifying cross-modal motion consistency between the ultrasonic responses and visual frames, EchoFence effectively distinguishes authentic live videos from replay-based or synthetic forgeries. Our design is fully compatible with off-the-shelf hardware and does not disrupt user experience or communication quality. Extensive evaluations demonstrate its effectiveness and robustness across various scenarios.

Acknowledgment

We sincerely thank our anonymous reviewers for their insightful comments. This paper is supported by the Natural Science Foundation of China (62372400), the National Key R&D Program of China (2023YFA1009500), the China Postdoctoral Science Foundation (2025M781520), the Postdoctoral Fellowship Program of CPSF (GZC20241488), the Postdoctoral Research Excellence Funding Project of Zhejiang Province (ZJ2025024), and the Fundamental Research Funds for the Central Universities (2025110528-0).

REFERENCES

[1] G. Pei, J. Zhang, M. Hu, G. Zhai, C. Wang, Z. Zhang, J. Yang, C. Shen, and D. Tao, "Deepfake generation and detection: A benchmark and survey," *CoRR*, vol. abs/2403.17881, 2024.

[2] H. Chen and C. Kathleen Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk, 2024.

[3] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 46–52.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[5] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 3247–3258.

[6] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5052–5060.

[7] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *Proceedings of the ACM Multimedia Conference*, 2021, pp. 3473–3481.

[8] X. Ding, W. Zhu, and D. Zhang, "Deepfake videos detection via spatiotemporal inconsistency learning and interactive fusion," in *Proceedings of the IEEE International Conference on Sensing, Communication, and Networking*, 2022, pp. 425–433.

[9] Y. Zhou and S. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 780–14 789.

[10] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, "A robust approach to multimodal deepfake detection," *Journal of Imaging*, vol. 9, no. 6, p. 122, 2023.

[11] Y. Huang, F. Juefei-Xu, R. Wang, Q. Guo, L. Ma, X. Xie, J. Li, W. Miao, Y. Liu, and G. Pu, "Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 1217–1226.

[12] H. Xu, Y. Wang, Z. Wang, Z. Ba, W. Liu, L. Jin, H. Weng, T. Wei, and K. Ren, "Profake: Detecting deepfakes in the wild against quality degradation with progressive quality-adaptive learning," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, p. 2207–2221.

[13] G. Mittal, C. Hegde, and N. D. Memon, "Gotcha: Real-time video deepfake detection via challenge-response," in *Proceedings of the IEEE European Symposium on Security and Privacy*, 2024, pp. 1–20.

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

[15] Y. Wu, P. Jiang, J. Cheng, L. Zhao, C. Shen, C. Wang, and Q. Wang, "Sonicumos: An enhanced active face liveness detection system via ultrasonic and video signals," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2025.

[16] C. R. Gerstner and H. Farid, "Detecting real-time deep-fake videos using active illumination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 53–60.

[17] H. Guo, X. Wang, and S. Lyu, "Detection of real-time deepfakes in video conferencing with active probing and corneal reflection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[18] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020, pp. 402–419.

[19] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[20] Huawei, "HUAWEI MateBook X Pro," https://consumer.huawei.com/en/laptops/matebook-x-pro-ultra-premium-edition/, 2025.

[21] OBS-Studio, "Free and open source software for video recording and live streaming," https://obsproject.com/, 2025.

[22] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *Pattern Recognition*, vol. 141, p. 109628, 2023.

[23] SAMSUNG, "Galaxy A53 5G," https://www.samsung.com/latin_en/smartphones/galaxy-a/galaxy-a53-5g-awesome-black-128gb-sm-a536ezkdtpa/, 2023.

[24] VB-AUDIO, "Voicemeeter virtual audio mixer," https://vb-audio.com/Voicemeeter/, 2025.

[25] Y. Chen, N. Akhtar, N. A. H. Haldar, and A. Mian, "Deepfake detection with spatio-temporal consistency and attention," *CoRR*, vol. abs/2502.08216, 2025.

[26] M. Qiao, R. Tian, and Y. Wang, "Towards generalizable deepfake detection with spatial-frequency collaborative learning and hierarchical cross-modal fusion," *CoRR*, vol. abs/2504.17223, 2025.

[27] H. K. Verma and I. Kumar, "Enhancing deepfake detection with a hybrid cnn-bilstm approach," in *Proceedings of the International Conference on Soft Computing and Machine Intelligence*, 2024, pp. 351–356.

[28] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8261–8265.

[29] H. Liu, Z. Li, Y. Xie, R. Jiang, Y. Wang, X. Guo, and Y. Chen, "Livescreen: Video chat liveness detection leveraging skin reflection," in *Proceedings of the IEEE Conference on Computer Communications*, 2020, pp. 1083–1092.

[30] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "Audiogest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 474–485.

[31] J. McIntosh, A. Marzo, M. Fraser, and C. Phillips, "Echoflex: Hand gesture recognition using ultrasound imaging," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1923–1934.

[32] A. Das, I. Tashev, and S. Mohammed, "Ultrasound based gesture recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 406–410.

[33] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "Ultragesture: Fine-grained gesture sensing and recognition," in *Proceedings of the IEEE International Conference on Sensing, Communication, and Networking*, 2018, pp. 28–36.

[34] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," *IEEE Transactions on Mobile Computing*, vol. 21, no. 5, pp. 1798–1811, 2022.

[35] L. Ge, J. Zhang, and J. Wei, "Single-frequency ultrasound-based respiration rate estimation with smartphones," *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 3 675 974:1–3 675 974:8, 2018.

[36] T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou, "Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2959–2973, 2019.

[37] P. Kar, S. Singh, A. Mandal, S. Chattopadhyay, and S. Chakraborty, "Expressense: Exploring a standalone smartphone to sense engagement of users from facial expressions using acoustic sensing," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2023, pp. 265:1–265:18.

[38] W. Cheng, M. Pang, H. Wan, S. Dong, D. Liu, and W. Wang, "Usee: Ultrasound-based device-free eye movement sensing," in *Proceedings of the IEEE International Conference on Sensing, Communication, and Networking*, 2024, pp. 1–9.

[39] J. Tan, C. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *Proceedings of the IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[40] 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," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–26, 2020.

[41] Y. Zhang, Y. Chen, H. Wang, and X. Jin, "CELIP: Ultrasonic-based lip reading with channel estimation approach for virtual reality systems," in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers*, 2021, pp. 580–585.