When One Sense Fails: Towards Multi-Modal Gait Recognition Framework Bridging Vision and Structural Vibration Sensing

Mainak Chakraborty[1] Bodhibrata Mukhopadhyay[2] Subrat Kar[1]
[1]Indian Institute of Technology Delhi [2]Indian Institute of Technology Roorkee
mainak.chakraborty@iddc.iitd.ac.in subrat@ee.iitd.ac.in bodhibrata@iitr.ac.in

Abstract

Gait recognition from video silhouettes has seen significant progress, yet occlusions and changes in appearance continue to limit its reliability. Structural vibrations induced by footsteps offer a complementary signal that is inherently privacy-preserving, but this modality still lacks the benchmarks and principled fusion strategies necessary for real-world use. In this work, we introduce a multi-modal framework that combines silhouette sequences with floor-vibration measurements for person identification. Our fusion architecture employs intra-modal self-attention to refine each representation independently, bidirectional cross-modal contextualization to exchange information between the two streams, and a learned gating mechanism that adaptively weights each modality's contribution. We evaluate the approach under four experimental protocols and compare it against several alternative fusion strategies. The proposed model achieves approximately 89% rank-1 identification accuracy. Further analysis shows that vibration features provide view-invariant cues that complement the appearance information captured by silhouettes, accounting for much of the gain over either modality alone. To encourage reproducible work, we publicly release our source code, dataset, and evaluation protocols.

1. Introduction

Gait recognition has applications in security surveillance, healthcare monitoring, and human-computer interaction. Over the past two decades, the field has been dominated by vision-based approaches that analyse video sequences to extract characteristic gait signatures [2, 19, 22]. However, camera-based systems face fundamental limitations: they are vulnerable to occlusions, lighting conditions, viewpoint variations, and raise significant privacy concerns [32]. Recently, structural vibration sensing has emerged as a complementary modality for gait recognition [1, 8, 21, 26, 27, 33, 35]. When a person walks across a floor, their footsteps generate vibrations that propagate through the building structure. These vibrations can be captured by geophone sensors and analyzed to identify individuals. Unlike cameras, vibration sensors do not capture appearance information, making them inherently privacy-preserving. They are also immune to visual occlusions and lighting conditions. However, vibration-based methods lack the rich spatial information available in video. The complementary strengths of vision and vibration sensing motivate multi-modal approaches that fuse information from both modalities. As shown in figure 1, progress in this direction has been severely limited by the absence of suitable datasets. Large-scale vision datasets like CASIA-E [32] (1,014 subjects) lack vibration sensing, while the few existing multi-modal datasets [6, 13] are limited to 10 subjects, insufficient for training modern deep learning models.

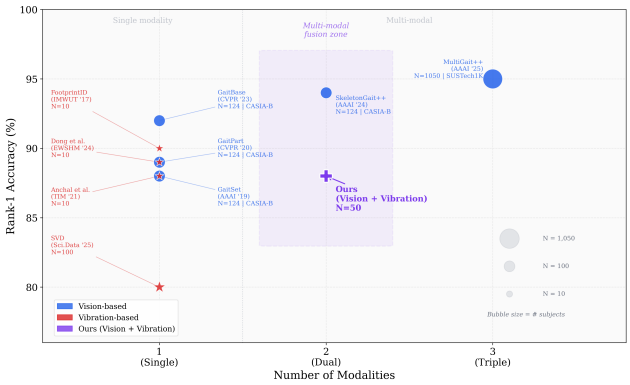

Figure 1. Landscape of gait recognition methods by sensing modality and number of input modalities. Vision-based methods (blue circles) are evaluated on CASIA-B (N=124) or SUSTech1K (N=1,050) using Rank-1 accuracy under the large-sample-test (LT) protocol, averaged across normal walking (NM), bag-carrying (BG), and clothing-change (CL) conditions. Vibration-based methods (red stars) report best Rank-1 accuracy from their respective studies under varied experimental settings with smaller cohorts (N=10–100). Marker size is proportional to the number of subjects.

Table 1. Comparison with Existing Gait Recognition Framework. Our work is combining vision and structural vibration sensing with comprehensive metadata.

Dataset	Year	Subjects	Samples	Modalities	Cameras	Sensors	Public	Biometric
USF HumanID [29]	2005	122	1,870	Vision	✓	×	✓	✓
CASIA-B [38]	2006	124	13,640	Vision	✓	×	✓	✓
OU-MVLP [34]	2018	10,307	1.6M	Vision	✓	×	✓	✓
CASIA-E [32]	2022	1,014	778,752	Vision	✓	×	✓	✓
FootprintID [27]	2017	10	–	Vibration	×	✓	×	✓
USLEET [1]	2020	10	7,750	Vibration	×	✓	×	×
UMGP [31]	2024	50	–	Vibration	×	✓	×	×
GaitVibe+ [13]	2022	20+	-	Vision+Vibration	✓	✓	×	×
VibeGait [6]	2025	10	13,070	Vision+Vibration	✓	✓	×	✓
SVD [5]	2025	15	-	Vision+Vibration	✓	✓	✓	✓
Ours	**2026**	**50**	**29,360**	**Vision+Vibration**	✓	✓	✓	✓

1.1. Contributions

This work makes four primary contributions to understanding multi-modal biometric fusion:

- **C1. Hierarchical Attention Fusion for Gait and Vibration** We integrate and systematically evaluate a Hierarchical Attention Fusion architecture that combines gait silhouette features with structural vibration signals through self-attention, cross-attention, and learned modality gating. Evaluated across 50 subjects with 2 camera views and 2 vibration sensors, the model achieves competitive single-domain verification (EER of 19.90%) and demonstrates stable open-set across increasing unknown ratios.
- **C2. Multi-modal Evaluation Testbed** We curate a multi-modal sensing testbed spanning gait silhouettes from camera setup and CWT spectrograms from dual floor-mounted vibration sensors. As summarized in Table 1, the dataset yields a total of 29,360 samples from vision and vibration sensors. We define four evaluation protocols namely, closed-set identification, single-domain verification, open-set novelty detection, and missing modality robustness, with strict subject-disjoint splits and frame-level data partitioning to prevent temporal leakage.
- **C3. Extensive Baseline Comparison with Leakage-Free Evaluation** We implement and evaluate 19 model configurations: classical GEI+PCA+SVM, three gait encoders (Simple-CNN, GaitSet, GaitPart), 3 vibration encoders (SpectrogramCNN, Footsnet, Resnet50), and 3 fusion strategies. All evaluation code and protocols are publicly released.

2. Related Work and Gap Analysis

2.1. Vision-Based Gait Recognition

The choice of input representation is a fundamental design decision in vision-based gait recognition. Binary silhouettes, extracted via background subtraction or semantic segmentation, have been the dominant representation since the field's inception. Han and Bhanu [19] proposed the Gait Energy Image (GEI), a spatio-temporal template formed by averaging silhouette frames over a gait cycle. GEI became one of the most widely adopted gait representations due to its compactness and robustness to minor silhouette noise.

Beyond silhouettes and GEI, skeleton-based (model-based) representations encode body joint positions across frames. Early skeleton-based methods suffered from limited pose estimation accuracy, but the advent of modern pose estimators has revived interest in this direction. More recently, human parsing maps [40], optical flow [24], and 3D mesh representations such as SMPL [39] have been explored to provide richer geometric and semantic information for gait recognition.

GaitSet [7] pioneered the treatment of a gait sequence as an unordered set, applying maximum pooling along the temporal dimension to aggregate frame-level spatial features. Its simplicity and effectiveness made it one of the most influential methods in the field. GaitPart [17] extended this line of work by emphasizing local spatial details and modeling temporal dependencies through a focal convolution layer and micro-motion capture module.

Despite these advances, vision-based gait recognition datasets rely on camera sensors (RGB, depth, or LiDAR). Camera-based systems remain sensitive to environmental factors including illumination changes, occlusion, and viewpoint variation [23, 28]. The reliance on visual data raises privacy concerns in deployment contexts where capturing identifiable imagery is undesirable or legally restricted. These limitations motivate the development of multi-modal gait datasets that combine visual and non-visual sensor data to enable more robust, privacy-aware gait recognition research.

2.2. Structural Vibration Sensing

Early work established that footstep vibrations could identify small groups of people in controlled settings. Pan et

al. [26] first demonstrated indoor person identification using footstep-induced floor vibrations captured by geophone sensors, achieving over 83% accuracy with classification at step and trace levels. The key insight is that each person's unique walking pattern (determined by physical characteristics, center of gravity, and foot-ground contact dynamics) produces a distinctive vibration signature in the floor structure. Recent methods achieve higher accuracy and scale to more subjects and diverse environments [4, 9, 26, 27, 37].

Beyond identification, structural vibration sensing has been extended to occupant localization using time-difference-of-arrival (TDOA) across multiple sensors [25], gait health monitoring for conditions such as muscular dystrophy [11], and even gait abnormality detection through footstep contact analysis [15]. Dong et al. [14] introduced a ubiquitous gait analysis framework that estimates standard spatiotemporal gait parameters from floor vibrations in daily living spaces, validating consistent results across wood and concrete floors. The Re-Vibe system [16] addressed the cross-structure person re-identification problem using optimal transport to align vibration feature distributions across different floor structures. Wu et al. [36] has shown that emotion recognition can be achieved using footstep-induced floor vibrations.

Classification methods have evolved from traditional machine learning applied to hand-crafted features [27]. Xu et al. [37] proposed deep domain generalization for pedestrian identification, addressing the distribution shift caused by structural heterogeneity across different floor locations. Most recently, Dong and Noh [10] formulated continual person identification, where the system gradually learns identities online without pre-collected enrollment data, achieving 90% accuracy with a feature transformation that reduces vibration variability by 70%.

Two major challenges continue to hinder progress in this field. First, there is a scarcity of publicly available datasets or test-bed suitable for effectively training robust models [6]. Most existing datasets include limited (10-20) participants [5, 12, 27], limiting generalizability. Although more recent efforts have expanded the scale to approximately 100 subjects, these datasets remain uni-modal [5], restricting their applicability for multi-sensor learning. Second, there is a lack of multi-modal datasets that combine structural vibration signals with complementary data sources such as video, skeletal representations, or inertial measurements. In contrast, vision-based gait recognition has benefited from datasets that provide synchronized silhouettes, skeletons, 3D meshes, and parsing maps [30, 39], enabling effective cross-modal feature learning. No comparable resource currently exists for structural vibration sensing, limiting research on sensor fusion and cross-modal knowledge transfer between privacy-preserving vibration data and information-rich visual modalities. To address these limitations, we propose a unified cross-attention framework that jointly uses structural vibration and visual modalities.

3. Dataset Details and Pre-processing

3.1. Dataset Details

Vision Sensors Two cameras are arranged in a calibrated array at 15°(A1), and 60° (A2) relative to the walking path (3 MP CMOS, 20 fps). Raw frames are processed into Gait Energy Images (GEIs) through standard silhouette extraction and temporal averaging. We recruited 50 individuals (34 males and 16 females) for this study, with ages spanning from 20 to 60 years. For the walking experiments, we instructed each participant to wear comfortable, flat-soled shoes. The participants' physical measurements show considerable diversity. Heights ranged from 1.40 m to 1.90 m, and weights varied between 40 kg and 90 kg. [1]

Vibration Sensors Two geophone sensors (S1, S2) with a sensitivity of 2.88 V/(m/s), sampling rate of 32 kHz are floor-mounted at 0 m and 3 m along the walking path. The sensors have an effective measurement range of 4.5–500 Hz and a coverage radius of approximately 3–5 m at a carpeted floor. Signals are denoised, segmented per footstep event, and converted to Continuous Wavelet Transform (CWT) spectrograms for model input.

Collection and quality control Data were collected from 50 subjects. A three-stage quality pipeline covering automated frame and SNR checks, manual review of a 10% subset, and statistical outlier removal at 3σ vibration amplitude produced a final set of samples (87% retention rate). Multi-session data across different days was not collected. We acknowledge this as a limitation and plan to extend the dataset with multi-session recordings, varied walking speeds, different shoe types, and load-carrying conditions in future work. This study was approved by the Institute Ethics Committee, IIT Delhi (Reference No. P-059), and all participants provided written informed consent permitting public release of anonymized data.

3.2. Data Pre-Processing

Vibration-based Pre-processing The raw seismic audio signals were first pre-processed through an anti-aliasing low-pass filter and downsampled to a target frequency of 500 Hz. These signals were then segmented into discrete temporal chunks for localized analysis. For each chunk, the CWT was computed using the Morlet wavelet ("morl") across a range of 127 widths, which allows for a multi-resolution decomposition of the signal in the time-frequency domain. The resulting coefficients were converted to their absolute magnitude and cropped to a specific frequency band to focus on relevant seismic features. To

[1] Code and Dataset is available at Link

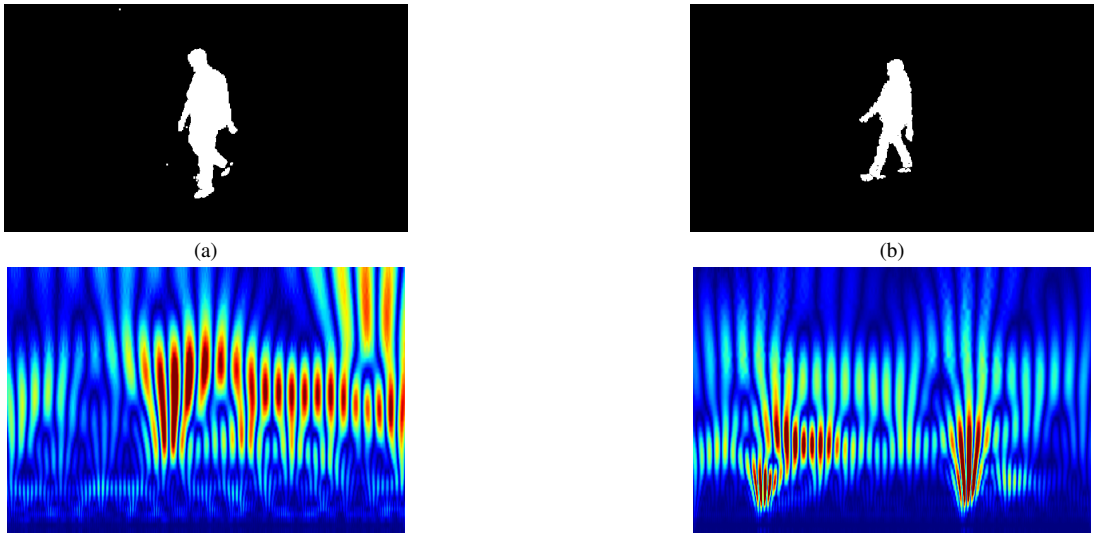

Figure 2. Experimental setup for data collection. (a) Participant (male) walking near the geophone, (b) Participant (female) walking near the geophone, (c) Continuous Wavelet Transform (CWT) of the walking signal of male participant, (d) CWT of the walking signal of female participant.

Table 2. Intra-Subject CWT Consistency Analysis. $S_{\text{align}} = 1 - (\overline{\Delta t} + \overline{\Delta f})$.

Sensor	Avg. Blobs/Step	Mean S_{align}	Mean $\overline{\Delta t}$	Mean $\overline{\Delta f}$
S1	11.3	0.705 ± 0.117	0.1011	0.1940
S2	11.3	0.733 ± 0.068	0.0978	0.1691

ensure consistent visual representation, the magnitude values were normalized to a 0–1 range and clipped at the 99th percentile to prevent outliers from skewing the contrast. Finally, these normalized coefficients were saved as high-resolution .png files. To evaluate the temporal consistency of floor-mounted vibration sensors, we computed the Continuous Wavelet Transform (CWT) for consecutive footstep events and assessed the alignment of energy blobs corresponding to heel-strike and toe-off impulses across successive steps of the same subject. The alignment score is defined as

$$S_{\text{align}} = 1 - (\overline{\Delta t} + \overline{\Delta f}),$$

where $\overline{\Delta t}$ and $\overline{\Delta f}$ denote the mean normalized time and frequency drift between matched blob centroids in adjacent CWT spectrograms. As shown in Table2, both channels exhibit moderate-to-good intra-subject consistency. Channel S2 achieves a slightly higher mean alignment ($S_{\text{align}} = 0.733 \pm 0.068$) compared to Channel S1 (0.705 ± 0.117), with a lower standard deviation indicating more stable signal acquisition. The mean time drift remains below 0.11 for both channels, suggesting consistent gait phase recurrence across steps. The somewhat larger frequency drift ($\overline{\Delta f} \approx$ 0.17–0.19) likely reflects natural stride-to-stride variability in ground-reaction forces rather than sensor noise. Overall, the results confirm that the CWT preprocessing pipeline produces repeatable time–frequency representations suitable for downstream biometric fusion.

Vision-based Pre-processing The Gait Energy Image (GEI) is generated by condensing a sequence of binary silhouettes into a single grayscale representation that captures both the static body structure and dynamic gait characteristics. Binary silhouettes are first extracted from the video using background subtraction and thresholding, then cropped to their bounding boxes and filtered using an aspect ratio constraint ($0.8 \leq \text{AR} \leq 5.0$) to ensure consistency. Each silhouette $B(x, y, t)$ is subsequently aligned by centering its centroid within a fixed canvas $\mathcal{I}$ of size $H \times W$ (e.g., 224×224), where the centroid $(\bar{x}, \bar{y})$ is computed using image moments $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$. The GEI is then obtained through temporal aggregation by averaging the aligned silhouettes over N frames,

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B(x, y, t),$$

which encodes the spatial–temporal gait pattern in a single image. Finally, the GEI is min–max normalized to the range $[0, 1]$ to improve numerical stability during training,

$$G_{\text{norm}}(x, y) = \frac{G(x, y) - \min(G)}{\max(G) - \min(G) + \epsilon},$$

where $\epsilon = 10^{-8}$ prevents division by zero. We adopt GEI as the primary visual representation for the fusion frame-

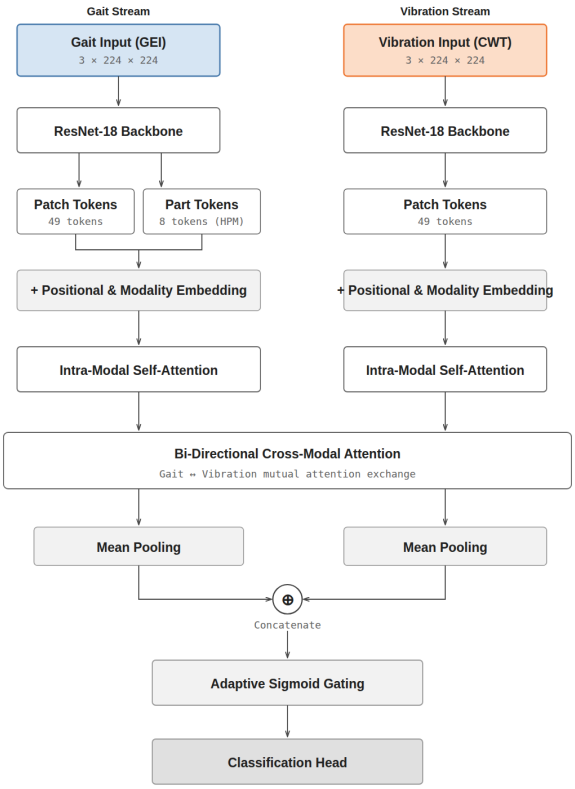

Figure 3. Our overall architecture of the Hierarchical Attention Fusion Network.

work due to its compactness (single image per sequence), low computational cost, and proven robustness to minor silhouette noise. We note that our evaluation already includes sequence-based deep models as baselines in Section 6, which operate on raw silhouette sequences rather than GEI. The fusion framework can be extended to incorporate sequence-level visual encoders, which we leave as future work.

4. Baseline Methods

To demonstrate the utility of our method and establish performance benchmarks, we implement 12 methods spanning classical approaches to modern fusion architectures. Vision-based benchmarks include classical Gait Energy Images (GEI+SVM), set-based learning (GaitSet), GaitPart and GaitBase model [7, 17, 18]. We have implemented these models on our end and tested with our data. Vibration-based baselines comprise CNNs trained on CWT spectrograms and raw waveforms, alongside our own implementations. For multi-modal integration, we examine fusion strategies, concatenation via MLP, early fusion, late fusion, cross-attention, and self-attention. All models were trained for 300 epochs using the AdamW optimizer ($LR = 10^{-4}$, batch size 32) on an NVIDIA P100 GPU to maintain experimental consistency.

As shown in Figure 3, two ResNet-18 backbones encode gait GEI images and vibration CWT strips into $512 \times 7 \times 7$ feature maps. The gait branch reshapes its feature map into 49 spatial tokens and appends 8 horizontal-strip tokens from HPM-style pooling (57 total), while the vibration branch produces 49 spatial tokens. Both token sequences are augmented with learned positional and modality embeddings and processed by Transformer encoder layers for intra-modal self-attention ($d = 128$, 4 heads, FFN dimension $4d$). A bidirectional cross-attention stage enables gait tokens to attend to vibration tokens and vice versa, each followed by residual connections, layer normalization, and a feed-forward sub-layer. The resulting tokens are mean-pooled into d-dimensional vectors and fused using a sigmoid gate, which adaptively weights each modality. The fused representation z is then passed to a lightweight head (LayerNorm $\rightarrow$ Linear $\rightarrow$ SiLU $\rightarrow$ Dropout (0.3) $\rightarrow$ Linear) for classification or verification.

5. Benchmark Protocols and Tasks

We define four evaluation protocols that span diverse recognition scenarios. Each protocol specifies train/gallery/query splits, metrics, and target use cases.

5.1. Protocol 1: Closed-Set Identification

The Protocol 1 (P1) classification process evaluates the closed-set identification performance of the models across 50 subject classes. To ensure statistical robustness, the dataset undergoes an 80/20 train-validation split and is evaluated across three independent experimental repeats.

During training, the models are optimized using a joint metric learning objective:

$$L = L_{CE} + \lambda_1 L_{Center} + \lambda_2 L_{Triplet}.$$

This formulation combines label-smoothed Cross-Entropy (L_{CE}) to establish classification boundaries, Center Loss (L_{Center}) to minimize intra-class variance by pulling embeddings toward learnable class centroids, and Batch-Hard Triplet Loss ($L_{Triplet}$) to maximize inter-class variance by pushing apart the hardest anchor-negative pairs within a batch. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.3$, with a batch-hard triplet margin of 0.5 and label smoothing factor of 0.05. During inference, performance is quantified using standard multi-class metrics (Accuracy and F1-score) alongside Rank-1 and Rank-5 identification accuracies, computed via cosine similarity between the gallery subject centroids and individual probe embeddings.

5.2. Protocol 2: Single-Domain Verification

We measure how well a model learns identity-discriminative embeddings that generalize to entirely

unseen subjects. We adopt a strict subject-level disjoint split where identities in the test set are never observed during training. Each unimodal or multimodal encoder f_θ is trained within a Siamese framework. Given a pair of inputs (x_A, x_B), the shared encoder produces embeddings $\mathbf{e}_A = f_\theta(x_A)$ and $\mathbf{e}_B = f_\theta(x_B)$. A two-layer classifier receives the concatenation of the element-wise absolute difference and Hadamard product, $[|\mathbf{e}_A - \mathbf{e}_B| \,\|\, \mathbf{e}_A \odot \mathbf{e}_B]$, and outputs a similarity logit indicating whether the pair corresponds to the same identity.

Evaluation uses GroupKFold cross-validation ($k = 5$) with subject identity as the grouping variable to prevent identity leakage. Genuine pairs ($y = 1$) are formed from all $\binom{n_s}{2}$ combinations of gait images belonging to the same subject, while impostor pairs ($y = 0$) are created from cross-subject image combinations. To maintain class balance, impostor pairs are randomly sub-sampled to three times the number of genuine pairs (3:1 ratio).

5.3. Protocol 3: Open-Set Verification

We evaluate open-set performance by simulating a deployment scenario where probe samples may include unseen identities. The gallery contains only enrolled (known) identities, while the probe set includes both known and unknown individuals.

Subjects are randomly split into known and unknown groups using an unknown ratio $u_r \in \{0.2, 0.3, 0.4\}$. Let $\mathcal{S}_k$ and $\mathcal{S}_u$ denote the known and unknown identity sets, respectively, with $|\mathcal{S}_u|/(|\mathcal{S}_k| + |\mathcal{S}_u|) \approx u_r$. The known identities are further divided into training ($\mathcal{S}_k^{\text{train}}$) and evaluation ($\mathcal{S}_k^{\text{eval}}$). For each $s \in \mathcal{S}_k^{\text{eval}}$, a subset of samples forms the gallery $\mathcal{G}_s$, while the remainder constitutes genuine probes.

A model trained only on $\mathcal{S}_k^{\text{train}}$ extracts d-dimensional embeddings $\mathbf{e} = f_\theta(\mathbf{x})$. Gallery embeddings are averaged to form identity templates

$$\bar{\mathbf{e}}^{(s)} = \frac{1}{|\mathcal{G}_s|} \sum_{\mathbf{x}_g \in \mathcal{G}_s} f_\theta(\mathbf{x}_g),$$

and probe embeddings are compared using cosine similarity

$$s(\mathbf{p}, s) = \frac{\mathbf{e}_p^\top \bar{\mathbf{e}}^{(s)}}{\|\mathbf{e}_p\| \, \|\bar{\mathbf{e}}^{(s)}\|}.$$

We report results for two tasks: **known-identity verification**, comparing genuine pairs against impostors within $\mathcal{S}_k^{\text{eval}}$ ($\text{EER}_{\text{known}}$, $\text{AUC}_{\text{known}}$), and **novelty detection**, distinguishing known probes from unknown probes matched to any gallery identity ($\text{EER}_{\text{unknown}}$, $\text{AUC}_{\text{unknown}}$).

5.4. Protocol 4: Ablation and Missing-Modality Robustness

In this protocol we evaluate the robustness of the model in 2 parts. Part-1 corresponds to an architectural ablation study.

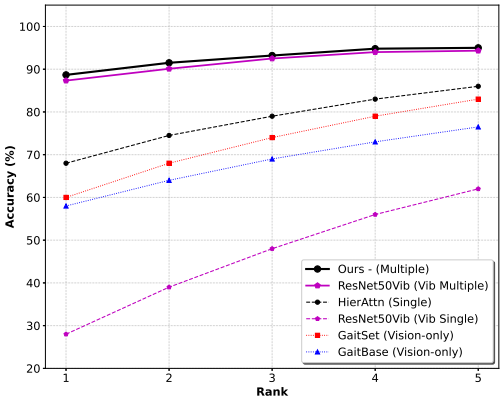

Figure 4. Cumulative Match Characteristic (CMC) curves comparing the proposed Hierarchical Attention (HierAttn) and ResNet50Vibration models against vision-only baselines (GaitSet and GaitBase). Results are presented for both "Single" (Single-footstep) and "Multiple" (Multiple-footstep) scenarios.

Table 3. Single-domain verification results comparing individual modalities Gait and Vibration against the proposed fusion model.

Modality	Model	EER (%) ↓	AUC (%) ↑	d' ↑	TAR @ 1% FAR ↑
Gait	SimpleCNN	39.64 ± 2.15	64.36 ± 2.40	0.531 ± 0.045	0.0027 ± 0.0005
	GEI_SVM	39.36 ± 1.80	64.77 ± 2.10	0.537 ± 0.038	0.0031 ± 0.0004
	GaitSet [7]	36.37 ± 1.55	69.10 ± 1.85	0.699 ± 0.052	0.0040 ± 0.0007
	GaitBase [18]	35.14 ± 2.95	69.75 ± 3.86	0.604 ± 0.124	0.0043 ± 0.0046
	GaitPart [17]	37.59 ± 1.70	67.31 ± 1.95	0.588 ± 0.048	0.0042 ± 0.0008
Vibration	Spectrogram-CNN	24.99 ± 1.25	80.13 ± 1.30	1.103 ± 0.085	0.0373 ± 0.0045
	CNN1D [3]	33.52 ± 2.10	68.81 ± 2.25	0.664 ± 0.060	0.0514 ± 0.0082
	ResNet50 [20]	23.00 ± 1.10	81.87 ± 1.15	1.216 ± 0.075	0.0665 ± 0.0055
Fusion	**ResNet18**	$\mathbf{19.90} \pm 0.85$	$\mathbf{85.22} \pm 0.90$	$\mathbf{1.350} \pm 0.095$	$\mathbf{0.2520} \pm 0.0120$

During part-2 training,the model receives both modalities during optimization. When a modality is missing at test time, we apply an imputation function $I(\cdot)$ to replace the absent input. Specifically, we consider three strategies: **zero** imputation, which replaces the missing modality with an all-zero tensor $\mathbf{0}$; **noise** imputation, which replaces it with i.i.d. Gaussian noise $\mathcal{N}(0, 0.1)$; and **mean** imputation, which uses the per-channel mean computed over the training set. Formally, for an input pair $(\mathbf{g}_a, \mathbf{v}_a)$ and $(\mathbf{g}_b, \mathbf{v}_b)$, we replace the missing modality using $\mathbf{g}_i \leftarrow I(\mathbf{g}_i)$ if gait is unavailable or $\mathbf{v}_i \leftarrow I(\mathbf{v}_i)$ if vibration is unavailable. We report verification performance under four conditions: (i) both modalities present (baseline), (ii) gait missing with zero imputation, (iii) vibration missing with zero imputation, and (iv) gait missing with mean or noise imputation.

6. Experimental Results and Analysis

6.1. Closed-Set Identification

The closed-set classification and identification evaluation (Protocol 1) demonstrates the strong performance of both the standalone non-intrusive sensor approach and the cross-modal fusion strategy across 50 subjects. The proposed

Table 4. Cross-View Evaluation for the Fusion Model. Results indicate the model's ability to generalize across different viewing angles.

Source View	Target View	Scenario	EER (%)	AUC
A1	A2	S1	30.01 ± 1.45	0.7756 ± 0.0125
A2	A1	S2	34.66 ± 1.80	0.7492 ± 0.0150
A2	A1	S1	28.87 ± 1.50	0.7371 ± 0.0185
A2	A1	S2	26.67 ± 1.20	0.8062 ± 0.0095

identification model, combining both vision and vibration features achieves the highest overall identification performance. The hierarchical attention mechanism reached a peak Rank-1 accuracy of 88.67%. We note that the absolute Rank-1 gain from fusion over vibration-only is modest in closed-set identification. However, the benefit of multimodal fusion is more pronounced in verification, cross-view generalization (Table 4), and missing-modality robustness (Table 6), where the complementary gait modality provides improvements beyond what single modality alone achieves. The transition from single-footstep evaluation (Single) to multi-footstep (Multiple), (five consecutive footstep) yielded a substantial performance leap, improving the Rank-1 rate of the fusion model from 68.0% to 88.67%.

6.2. Verification Under Subject-Disjoint Splits

Table 3 reports the 1-vs-1 verification performance of all evaluated methods under the single-modality setting, where only one input modality is available during inference. Among the gait-only approaches, sequence-based deep models outperform both the simple frame-wise CNN and the classical GEI+SVM baseline. Specifically, GaitSet achieves the best gait-only performance with an EER of 36.37%, AUC of 69.10%, and $d' = 0.699$, followed by GaitPart with an EER of 37.59% and AUC of 67.31%. All gait-only methods exhibit relatively high error rates (EER $> 36\%$), indicating the overall performance for this dataset. In contrast, the vibration modality provides substantially stronger discriminative capability. The CNN1D baseline [3] achieves an EER of 33.52% and AUC of 68.81%, while spectrogram-based models further improve performance. In particular, the ResNet50 model trained on vibration spectrograms achieves the best single-modality result with an EER of 23.00%. Finally, the proposed fusion model achieves the best overall verification performance with an EER of 19.90%.

Additionally, we explore the effects of cross-view and cross-sensor evaluation.

6.2.1. Cross-View Evaluation

As shown in Table 4, the fusion model achieves its best cross-view performance when transferring from the profile view (A2) to the frontal view (A1) under vibration scenario S2, achieving the lowest EER of 26.67% and the highest AUC of 0.8062. For the same transfer direction (A2→A1), scenario S1 yields a comparable EER of 28.87% but a lower AUC of 0.7371, suggesting that the dynamic footstep responses captured by S2 better complement visual feature translation across viewpoints.

In the reverse direction (A1→A2), the model with S1 records an EER of 30.01% and an AUC of 0.7756, while the A2→A1 pairing with S2 in the first evaluation group shows the weakest performance (EER: 34.66%, AUC: 0.7492). Overall, these results indicate that structural vibration signals can help stabilize visual representations under viewpoint shifts.

Table 5. Cross-Sensor Performance Metrics for our Model(S) (mean values)

Config	Train → Test	EER ↓	AUC ↑	D-prime ↑	TAR@1% FAR
A1	S1 → S2	23.47	0.8541	0.9923	0.0130
	S2 → S1	21.57	0.8436	0.9487	0.0046
A2	S1 → S2	24.89	0.8236	0.7752	0.0000
	S2 → S1	22.63	0.8335	0.7462	0.0021

6.2.2. Cross-Sensor Performance

Table 5 presents the cross-sensor evaluation of the proposed fusion model. The visual configuration plays a notable role in cross-sensor generalization. Under camera angle A1, the model achieves higher d' values (0.9923 for S1→S2 and 0.9487 for S2→S1) and stronger AUC scores (0.8541 and 0.8436) compared to A2, which yields d' values of 0.7752 and 0.7462 with AUCs of 0.8236 and 0.8335.

A consistent directional pattern also emerges across both views: training on S2 and testing on S1 produces the lowest EERs—21.57% under A1 and 22.63% under A2. In contrast, the reverse direction (S1→S2) attains slightly higher AUC and d' values, suggesting that S1 captures a broader set of generalizable vibration features, potentially due to its spatial placement or structural coupling characteristics.

The uniformly low TAR@1% FAR values across all configurations (ranging from 0.0000 to 0.0130) highlight the difficulty of strict biometric verification under cross-sensor conditions, indicating that the system is better suited for continuous multimodal behavioral sensing rather than high-security access control.

6.3. Open-Set Novelty Detection

We evaluate four fusion architectures under the open-set novelty detection protocol across three unknown ratios ($\alpha \in \{0.2, 0.3, 0.4\}$), reporting both known verification and unknown detection EER. As shown in Fig. 5, CrossAttn achieves the lowest EER across most individual settings, with a particularly strong result at $\alpha = 0.3$ (Known EER = 0.03, Unknown EER = 0.03), demonstrating its ability to learn discriminative joint representations that generalize

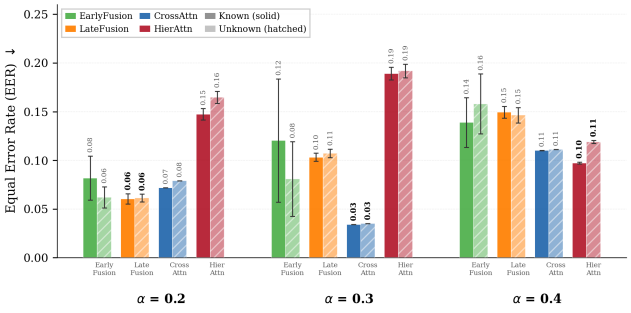

Figure 5. Open-set Novelty Detection performance across models and Unknown Ratios (UR) using Multiple footsteps. Lower EER is better. Our method demonstrates better stability across increasing Unknown Ratios.

Table 6. Performance trade-offs and modality robustness of the framework.

Model / Mode	Condition	EER (%) ↓	AUC ↑	Params / Infer(ms)
Part 1: Architecture Ablation				
Concat Only	Both Present	20.96	0.7936	22.7M / 5.3ms
+ Self-Attn	Both Present	20.60	**0.8626**	23.3M / 6.4ms
+ Cross-Attn	Both Present	23.28	0.8560	23.7M / 5.6ms
+ Gate	Both Present	**19.94**	0.8525	26.1M / 6.7ms
Part 2: Missing Modality				
HierAttn	Both Present	**26.39**	**0.8000**	–
	No Gait (Zero)	26.71	0.7659	–
	No Gait (Noise)	26.62	0.7737	–
	No Gait (Mean)	26.71	0.7650	–
	No Vib. (Zero)	35.98	0.6924	–
	No Vib. (Noise)	36.15	0.6899	–
	No Vib. (Mean)	35.88	0.6929	–

well to both seen and unseen identities. LateFusion performs competitively at lower unknown ratios ($\alpha = 0.2$), achieving Known and Unknown EER of 0.06, but degrades as the proportion of unknown subjects increases.

HierAttn, while exhibiting higher error rates at $\alpha = 0.2$ and $\alpha = 0.3$, recovers at $\alpha = 0.4$ (Known EER = 0.10, Unknown EER = 0.11), suggesting greater robustness under larger open-set conditions, which better reflects real-world deployment scenarios. *EarlyFusion* shows a steady degradation trend across all ratios, indicating limited capacity to handle novel identities. Overall, CrossAttn achieves the best absolute EER at lower unknown ratios, while HierAttn demonstrates the most stable performance degradation as the unknown fraction increases. We select HierAttn as the primary architecture for this reason, though we acknowledge CrossAttn outperforms it at specific operating points.

6.4. Architectural Ablation and Modality Analysis

As shown in Table 6, the architectural ablation (Part 1) confirms that progressively richer fusion strategies improve verification accuracy over simple feature concatenation. The baseline *Concat Only* model achieves an EER of 20.96% and an AUC of 0.7936. Adding self-attention raises the AUC to a peak of 0.8626 with only a modest increase in parameters (23.3M) and inference time (6.4 ms). The full + *Gate* configuration reduces the EER to its lowest value of 19.94% while maintaining a competitive AUC of 0.8525, providing the best balance between error-rate reduction and computational overhead at 26.1M parameters and 6.7 ms inference.

The missing-modality analysis (Part 2) highlights a clear asymmetry in modality contributions. With both modalities present, the HierAttn framework achieves an EER of 26.39% and an AUC of 0.8000. Removing the gait modality—via zero, noise, or mean imputation—results in only minor degradation (EER $\approx$ 26.7%), indicating that the model primarily relies on vibration features. In contrast, removing the vibration modality causes a substantial performance drop (EER $\approx$ 36%) across all imputation strategies. These results suggest that structural vibration acts as the dominant modality in the fusion framework, while gait provides complementary information that refines decision boundaries rather than driving them.

7. Limitations, and Future Directions

This work analyzes gait–vibration fusion for multimodal biometrics by evaluating 19 model configurations, across diverse operational scenarios. This study has several limitations. First, data were collected in a controlled single-walker laboratory setting; vibration signals from multiple simultaneous pedestrians would superimpose, and separating individual gait signatures from mixed signals remains an open challenge. Second, the dataset does not include covariates such as varied walking speeds, shoe types, carried loads, different floor materials or clothing changes for the same subject. Third, multi-session data across different days was not collected, limiting assessment of temporal stability. Future work will address these gaps by expanding the dataset with covariates and multi-session recordings, exploring disentangled representations for view-invariant gait extraction, investigating vibration source separation for multi-pedestrian scenarios, and evaluating cross-floor generalization.

Acknowledgment

We would like to thank Shivang Gaur, Chandan, and especially the participants at IIT-Delhi for providing us with an appropriate environment and their unstinting support while we deployed sensors and recorded data. We also acknowledge the support of the Bharti School (BSTTM), SeNSE, IRD, and Department of Electrical Engineering at IIT Delhi.

References

[1] S. Anchal, B. Mukhopadhyay, and S. Kar. Person identification and imposter detection using footstep generated seismic signals. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2020. 1, 2

[2] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition using gait entropy image. In *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, pages 1–6. IET, 2009. 1

[3] M. Chakraborty and S. Kar. Enhancing person identification through data augmentation of footstep-based seismic signals. *IEEE Signal Process. Lett.*, 30:1642–1646, 2023. 6, 7

[4] Mainak Chakraborty, A Srinivasan, Srinivasa Reddy, Sanjib Kumar Mandal, and Subhasis Bhaumik. Human action classification using seismic sensor and machine learning techniques. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–6. IEEE, 2021. 3

[5] Mainak Chakraborty, Chandan, Sahil Anchal, Bodhibrata Mukhopadhyay, and Subrat Kar. A structural vibration-based dataset for human gait recognition. *Scientific Data*, 12(1): 1617, 2025. 2, 3

[6] Mainak Chakraborty, Bodhibrata Mukhopadhyay, Sahil Anchal, Subrat Kar, et al. Vibegait: Enhancing structural-vibration based gait recognition using vision. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1, 2, 3

[7] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019. 2, 5, 6

[8] J. Clemente, F. Li, M. Valero, and W. Song. Smart seismic sensing for indoor fall detection, location, and notification. *IEEE J. Biomed. Health Inform.*, 24(2):524–532, 2019. 1

[9] Yiwen Dong and Hae Young Noh. Structure-agnostic gait cycle segmentation for in-home gait health monitoring through footstep-induced structural vibrations. In *Society for Experimental Mechanics Annual Conference and Exposition*, pages 65–74. Springer, 2023. 3

[10] Yiwen Dong and Hae Young Noh. Continual person identification using footstep-induced floor vibrations on heterogeneous floor structures. *arXiv preprint arXiv:2502.15632*, 2025. 3

[11] Yiwen Dong, Joanna Jiaqi Zou, Jingxiao Liu, Jonathon Fagert, Mostafa Mirshekari, Linda Lowes, Megan Iammarino, Pei Zhang, and Hae Young Noh. Md-vibe: Physics-informed analysis of patient-induced structural vibration data for monitoring gait health in individuals with muscular dystrophy. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*, pages 525–531, 2020. 3

[12] Y. Dong, J. Fagert, P. Zhang, and H. Y. Noh. Stranger detection and occupant identification using structural vibrations. In *Eur. Workshop Struct. Health Monit.*, pages 905–914. Springer, 2022. 3

[13] Yiwen Dong, Jingxiao Liu, and Hae Young Noh. Gaitvibe+ enhancing structural vibration-based footstep localization using temporary cameras for in-home gait analysis. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 1168–1174, 2022. 1, 2

[14] Yiwen Dong, Mario Iammarino, Jiarui Liu, Jordan Codling, Jonathon Fagert, Mostafa Mirshekari, Linda Lowes, Pei Zhang, and Hae Young Noh. Ubiquitous gait analysis through footstep-induced floor vibrations. *Sensors*, 24(8): 2511, 2024. 3

[15] Yiwen Dong, Susu Kim, Katharine Schadl, Peter Huang, Wenyu Ding, John Rose, and Hae Young Noh. In-home gait abnormality detection through footstep-induced floor vibration sensing and person-invariant contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7054–7067, 2024. 3

[16] Yiwen Dong et al. Re-Vibe: Vibration-based indoor person re-identification through cross-structure optimal transport. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*, pages 348–352. ACM, 2022. 3

[17] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. GaitPart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14233, 2020. 2, 5, 6

[18] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 5, 6

[19] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006. 1, 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 6

[21] S. Kitic, G. Puy, P. Pérez, and P. Gilberton. Scattering features for multimodal gait recognition. In *Proc. IEEE Glob. Conf. Signal Inf. Process. (GlobalSIP)*, pages 843–847. IEEE, 2017. 1

[22] Toby HW Lam, King Hong Cheung, and James NK Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern recognition*, 44(4):973–987, 2011. 1

[23] Tianhao Li, Weizhi Ma, Yujia Zheng, and Zhengping Li. A survey on gait recognition against occlusion: Taxonomy, dataset and methodology. *PeerJ Computer Science*, 10: e2602, 2024. 2

[24] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. GaitEdge: Beyond plain end-to-end gait recognition for better practicality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 375–390. Springer, 2022. 2

[25] Mostafa Mirshekari, Shijia Pan, Jonathon Fagert, Eve M Schooler, Pei Zhang, and Hae Young Noh. Occupant localization using footstep-induced structural vibration. *Mechanical Systems and Signal Processing*, 112:77–97, 2018. 3

[26] Shijia Pan, Ningning Wang, Yuqiu Qian, Irem Velibeyoglu, Hae Young Noh, and Pei Zhang. Indoor person identification through footstep induced structural vibration. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (HotMobile)*, pages 81–86. ACM, 2015. 1, 3

[27] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–31, 2017. 1, 2, 3

[28] Amandeep Parashar, Rajesh Singh Shekhawat, Weiping Ding, and Imad Rida. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 2

[29] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005. 2

[30] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1054–1063, 2023. 3

[31] M. Single, L. C. Bruhin, A. C. Naef, P. Krack, T. Nef, and S. M. Gerber. Unobtrusive measurement of gait parameters using seismographs: An observational study. *Scientific Report*, 14(1):14487, 2024. 2

[32] Chunfeng Song, Yongzhen Huang, Weining Wang, and Liang Wang. Casia-e: A large comprehensive dataset for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2801–2815, 2022. 1, 2

[33] A. Sundaresan, A. Subramanian, P. K. Varshney, and T. Damarla. A copula-based semi-parametric approach for footstep detection using seismic sensor networks. In *Proc. SPIE Multisensor, Multisource Inf. Fusion: Archit., Algorithms, Appl.*, pages 103–114. SPIE, 2010. 1

[34] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10(1):4, 2018. 2

[35] Divya Venkatraman, Vinod V Reddy, Andy WH Khong, and BP Ng. Polarization-cum-energy metric for footstep detection using vector-sensor. In *2011 IEEE International Conference on Technologies for Homeland Security (HST)*, pages 196–201. IEEE, 2011. 1

[36] Yuyan Wu, Yiwen Dong, Sumer Vaid, Gabriella M Harari, and Hae Young Noh. Emotionvibe: Human emotion recognition through footstep-induced floor vibrations. *IEEE Transactions on Affective Computing*, 2026. 3

[37] X. Xu, R. Deng, G. Zhao, B. Zhang, and C. Liu. Deep domain generalization-based indoor pedestrian identification using footstep-induced vibrations. *IEEE Transactions on Instrumentation and Measurement*, 73:1–8, 2024. 3

[38] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition (ICPR'06)*, pages 441–444. IEEE, 2006. 2

[39] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3D representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20228–20237, 2022. 2, 3

[40] Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 3040–3049, 2023. 2