

A real-world indoor dynamic skeleton dataset of elderly abnormal gait including crutch-assisted walking

Shaojiang Huang

Zhejiang Sci-Tech University

Manqing Yang

Zhejiang Sci-Tech University

Wei Zhang

zhangweicse@zstu.edu.cn

Zhejiang Sci-Tech University

data-descriptor

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11215944/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A real-world indoor dynamic skeleton dataset of elderly abnormal gait including crutch-assisted walking

Shaojiang Huang¹, Manqing Yang¹, Wei Zhang^{1,*}

¹School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China

*Correspondence: zhangweicse@zstu.edu.cn

Abstract

Falls and gait imbalance are a major health threat to older adults, yet public skeleton datasets that target elderly adults in real indoor care environments, particularly for clinically critical unbalanced gaits such as crutch-assisted walking, remain scarce. Here we present ElderlyPipe, a dynamic skeleton dataset of elderly abnormal gait recorded in real indoor care environments. The release comprises 63,855 quality-assured skeleton frames grouped into 287 continuous segments and 28,735 standardized 24-frame windows, covering four behaviours: standing, walking, crutch-assisted walking (using_crutch), and sitting. The skeletons are completed by hierarchical occlusion compensation and refined by visibility-aware adaptive Kalman smoothing. Benchmarks on the release show that the temporal association of skeleton sequences underpins imbalance analysis: a spatiotemporal graph model outperforms its single-frame counterpart in overall F1 (0.656 vs. 0.605) and better separates crutch-assisted walking, the representative unbalanced state, from everyday behaviours. The anonymized skeleton sequences, metadata, cross-validation splits, documentation, the complete processing pipeline, and the baseline code are openly released under a persistent identifier.

Background

Falls are the second leading cause of unintentional injury death worldwide, and adults older than 65 years suffer the greatest number of fatal falls [1]. Early and continuous monitoring of gait anomalies, especially high-risk unbalanced states such as crutch-assisted walking, is therefore a pressing need in nursing homes, community care centers, and home environments. However, progress of data-driven gait analysis for older adults is bottlenecked not only by algorithms but, more fundamentally, by the scarcity of high-quality datasets that target this population. To address this bottleneck, we present ElderlyPipe (Fig. 1), a real-world indoor skeleton dataset that focuses on elderly residents and covers the clinically critical crutch-assisted walking, together with the quality-assurance pipeline used to construct it and a leakage-free evaluation protocol that accompanies the release.

Current mainstream monitoring technologies for older adults fall into three categories: wearable devices, infrared radar, and visual surveillance. Wearable devices provide continuous physiological indicators, but their reliance on continuous wearing leads to poor long-term compliance [2]. Infrared or radar sensors are not limited by lighting conditions, but their cost and false positive rates in complex environments limit large-scale deployment. Vision-based gait

monitoring is non-contact, operates in real time, and can run continuously, which makes it attractive for 24/7 smart elderly care. Furthermore, various diseases (e.g., Parkinson’s disease and sarcopenia) are closely related to motor disability, so objective motor measurements extracted from skeleton sequences without attaching any clinical diagnosis, can serve as a non-invasive proxy for downstream mobility assessment.

General-purpose benchmarks such as NTU RGB+D [3] and Kinetics [4] offer large-scale action-recording resources, and HuGaDB [5] records gait through wearable inertial sensors, but they target young and middle-aged subjects in controlled capture; clinical collections such as UR-Fall [6] and UP-Fall [7] include older adults but expose only wearable-sensor fall events, not the four behaviour classes targeted here; the most recent elderly-care effort, CarePD [8], focuses on Parkinsonian gait rather than crutch-assisted walking. No public dataset simultaneously satisfies the following conditions: (i) real elderly subjects aged 65 and above; (ii) real indoor care environments rather than scripted laboratory capture; (iii) crutch-assisted walking as an annotated behaviour; and (iv) frame-level skeleton sequences suitable for obtaining spatio-temporal representations of imbalanced walking [9]. ElderlyPipe was constructed to cover exactly this combination, complementing rather than replacing these larger general-purpose

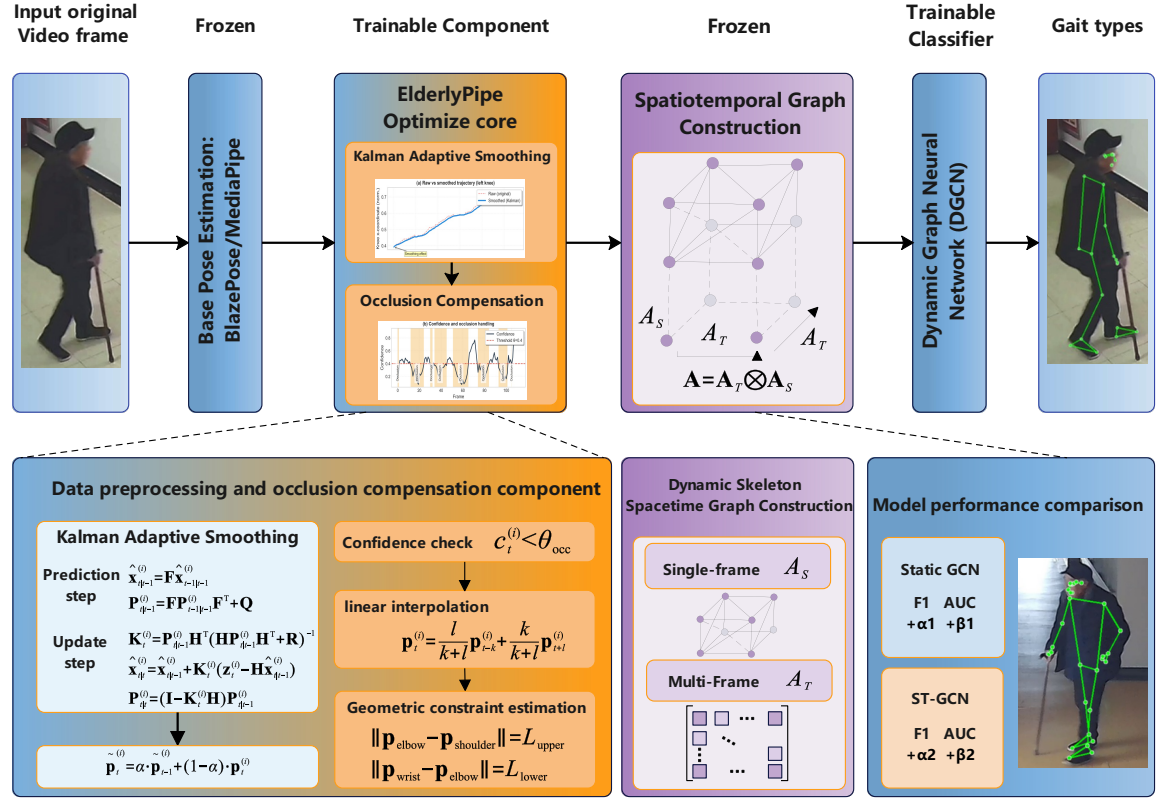


Figure 1. Overview of the ElderlyPipe method. Indoor care-scene videos are processed by BlazePose extraction, hierarchical occlusion compensation, and visibility-aware adaptive Kalman smoothing to obtain quality-assured dynamic skeleton sequences, which are then organized into spatiotemporal representations and evaluated under the leakage-free video-stratified protocol.

resources. The overall scale of the release is summarised in Supplementary Table S1, and a structured comparison with representative skeleton datasets is given in Supplementary Table S2.

From the perspective of sample size, the primary dataset is built from 63,855 quality-assured skeleton frames; these are organized into 287 continuous segments (222.5 frames per segment on average) and 28,735 standardized dynamic skeleton windows of 24 frames with stride 2 (see Section Leakage-free evaluation protocol). To protect subject privacy, the original RGB footage is not part of the public release; only the derived, anonymized skeleton sequences are distributed. The original capture conditions were 30 frames per second at a resolution of 1920×1080 , which are documented here solely to characterise the source material before skeleton extraction. Representative capture scenes, spanning the corridor, the lobby, and the dining area, are shown in Fig. 2, where the overlaid skeletons illustrate that ElderlyPipe recognizes complete skeletons even under strongly varying illumination and occlusion.

As for age diversity, individuals are primarily older resi-

dents of nursing homes, together with a small number of family members and caregivers who were naturally present in the care facilities. These younger subjects are deliberately retained: they enlarge the diversity of body shapes and movement styles, faithfully reflect real care environments where caregivers coexist with residents, and cannot inflate reported metrics because the video-stratified protocol confines every clip, regardless of age, to a single fold.

From a clinical biomechanics perspective, crutch-assisted walking exhibits characteristic markers, including reduced gait speed, altered knee and hip range of movement, and increased left-right limb asymmetry. Because this dataset provides frame-by-frame 33-keypoint skeleton sequences, these indicators can be quantified directly from the released data without any additional instrumentation. The release follows one design principle: real elderly subjects in real care environments, annotated for clinically critical unbalanced gaits.

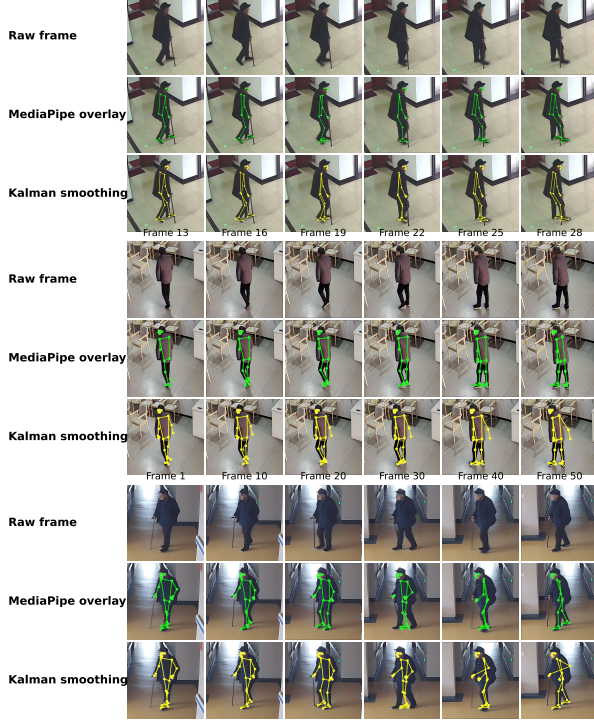


Figure 2. Skeleton recognition by the ElderlyPipe pipeline in three representative scenes. Example frames from the corridor, the lobby, and the dining area, with the skeletons extracted by the ElderlyPipe pipeline overlaid. Across the mixed lighting of the corridor, the backlighting of the lobby, and the dim, heavily occluded dining area, the 33 keypoints are consistently detected and completed, illustrating that ElderlyPipe recognizes reliable skeletons under the realistic conditions of care scenes.

Methods

The ElderlyPipe dataset was produced by a four-stage chain that runs from raw surveillance video to benchmarked skeleton windows (Fig. 1). Walking videos of real elderly subjects were recorded in real indoor care scenes, covering the clinically critical crutch-assisted walking together with standing, walking, and sitting. Each video frame was then converted into a 33-keypoint skeleton by BlazePose, and the raw skeletons were quality-assured by two cascaded stages: hierarchical occlusion compensation restores missing keypoints, and visibility-aware adaptive Kalman smoothing suppresses jitter while bridging unreliable frames. On top of the cleaned sequences, the ElderlyPipe pipeline organizes the data into standardized dynamic windows of 24 frames with stride 2, which serve as the released analysis unit. These windows are benchmarked with representative skeleton-based models under a leakage-free video-stratified 5-fold protocol, so that the reported scores reflect the genuine difficulty of the data rather than identity leakage.

Gait Data Collection

Walking videos were collected in real indoor scenes with various lighting conditions and floor materials, and the recorded indoor exercise modes were divided into four categories: standing, walking, using_crutch, and sitting, with using_crutch being the unbalanced category. Raw walking videos were obtained in five indoor sub-areas: the activity rooms and activity areas, which are well lit and only slightly occluded (6 clips in total); the nursing-home corridor, which mixes natural and artificial light and is almost free of occlusion (217 clips); the nursing-home lobby, which is backlit but unoccluded (23 clips); and the public dining area, which is dimly lit and heavily occluded (3 clips). A scene breakdown is given in Supplementary Table S3. In each scene, subjects perform daily activities naturally, without scripted instructions, so that the recorded gaits reflect the residents’ true motor ability in everyday life.

Fixed surveillance cameras capture raw video at 30 frames per second and a resolution of 1920×1080 . Consecutive frames of the same behaviour are grouped into 287 continuous segments averaging 222.5 frames, from which 28,735 standardized 24-frame windows are extracted at stride 2. The full evaluation protocol is described in Section Leakage-free evaluation protocol. The raw RGB footage is withheld, and only the derived skeleton sequences are distributed.

Skeleton detection and preprocessing

To guarantee data quality, we adopt a semi-automatic annotation pipeline that combines MediaPipe pose estimation with manual verification. The BlazePose framework provides 3D coordinates for 33 keypoints to generate the skeleton sequence; its two-stage detector-regressor architecture is described in Supplementary Note S1. Kalman filtering was applied for skeleton smoothing, and geometric constraints compensate for occlusions. The detected coordinates are normalized into the range $[0, 1]$. For the standard frame-by-frame skeleton sequence, each frame is represented as a 33×4 matrix containing the x, y, z coordinates and a visibility score per keypoint. Behaviour labels are assigned by the annotators at the video level and verified frame by frame against the rendered skeleton overlay. In total, 63,855 skeleton frames survive this verification.

Hierarchical Occlusion Compensation

In the traditional MediaPipe framework, it outputs a confidence $c_t^{(i)} \in [0, 1]$ of every keypoint; if $c_t^{(i)} < \theta_{\text{occ}}$, we consider that this point is occluded/unreliable ($\theta_{\text{occ}} = 0.5$). We use two compensation methods depending on how long the occlusion lasts and set $T_{\text{max}} = 15$ frames as its boundary:

1. **Linear interpolation:** if the keypoint is visible (confidence above threshold) in consecutive frames before and after, and the occlusion lasts fewer than $T_{\max} = 15$ frames, linear interpolation is employed to obtain missing values. Let the missing frame be t , the linear interpolation is described by Eq.(1).

$$\mathbf{p}_t^{(i)} = \frac{l}{k+l} \mathbf{p}_{t-k}^{(i)} + \frac{k}{k+l} \mathbf{p}_{t+l}^{(i)}. \quad (1)$$

The visible frames before and after the gap are denoted $t-k$ and $t+l$, respectively.

2. **Geometric constraint estimation:** if a keypoint j remains occluded for $T_{\max} = 15$ consecutive frames or more, its position is estimated from its parent joint j_p and child joint j_c based on learned bone length ratios. For instance, when the elbow is occluded but the shoulder and wrist are visible, we compute the direction from shoulder to wrist and place the elbow along that direction at the pre-trained upper arm length. If only the parent is visible, we extend along an average direction (e.g., vertically downward for lower limbs). The bone length ratios are dynamically updated from visible keypoints in each video.

Formally, for a joint j occluded for $T_{\max}$ frames or more, its position is estimated from the visible parent joint j_p on the same limb according to Eq.(2).

$$\hat{\mathbf{p}}_j = \mathbf{p}_{\text{parent}} + \hat{\mathbf{d}} \cdot l_{\text{std}}, \quad (2)$$

where the unit direction $\hat{\mathbf{d}}$ is determined by the visible child joint j_c when available and falls back to a preset default direction otherwise, defined by Eq.(3).

$$\hat{\mathbf{d}} = \begin{cases} \frac{\mathbf{p}_{\text{child}} - \mathbf{p}_{\text{parent}}}{\|\mathbf{p}_{\text{child}} - \mathbf{p}_{\text{parent}}\|}, & \text{if } j_c \text{ is visible,} \\ \hat{\mathbf{d}}_{\text{default}}, & \text{otherwise,} \end{cases} \quad (3)$$

with $\hat{\mathbf{d}}_{\text{default}}$ taken as the vertically downward unit vector, and the standard bone length l_{std} is the median of the valid length samples collected from the historical frames in which both endpoints satisfy the visibility threshold τ ($\tau = 0.7$):

$$l_{\text{std}} = \text{median} \{ \|\mathbf{p}_{\text{parent}}(t) - \mathbf{p}_j(t)\| \mid v_{\text{parent}}(t) > \tau \wedge v_j(t) > \tau \}. \quad (4)$$

This formulation requires only j_p to be visible, which covers the common cases in which terminal joints (wrists, ankles) are occluded together with their children.

The standard bone lengths required by the geometric constraint estimation are not fixed constants but are updated online. For the eight limb bones (the bilateral upper arms, lower arms, upper legs, and lower legs) we maintain a sliding sample window of at most 300 length samples per bone; a new sample is accepted only when both endpoint

joints of the bone have a confidence above 0.7, and the standard bone length is taken as the median of the window, which is updated frame by frame as each video proceeds. This online median estimation adapts the bone geometry to the body proportions of the current subject while remaining robust to occasional misdetections.

Compensated coordinates carry a confidence below 1.0 so that downstream stages can distinguish them from directly observed detections. Specifically, interpolated points receive 0.6 and geometrically estimated points receive 0.8. The geometric estimate is rated higher: it is anchored to a visible parent joint and to the per-subject bone length, whereas an interpolation can drift when the gap is long. The primary function of occlusion compensation is to restore the completeness of the skeleton sequence, but the coordinates of visible keypoints still suffer from high-frequency jitter introduced by per-frame independent detection. In order to alleviate the jitter problem, we introduce a visibility-aware adaptive Kalman smoother to suppress the jitter while bridging residual unreliable frames.

Visibility-aware adaptive Kalman smoothing

As a linear dynamic system, the indoor movement of each keypoint is smoothed by a Kalman filter [10]. Let the true state of keypoint i at timestamp t be $\mathbf{x}_t^{(i)} \in \mathbb{R}^6$ (including 2D image coordinates (x, y) , depth z , and their first derivatives, i.e., velocities $\dot{x}, \dot{y}, \dot{z}$), and the observed keypoint coordinates $\mathbf{z}_t^{(i)} \in \mathbb{R}^3$ detected by the MediaPipe framework.

When the confidence score of a keypoint falls below a threshold, the update step is skipped for that frame, and only the prediction is retained. This allows the filter to bridge brief occlusions without being disturbed by unreliable detections. The state transition and observation equations are defined by Eq.(5).

$$\begin{aligned} \mathbf{x}_t^{(i)} &= \mathbf{F} \mathbf{x}_{t-1}^{(i)} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathcal{N}(0, \mathbf{Q}), \\ \mathbf{z}_t^{(i)} &= \mathbf{H} \mathbf{x}_t^{(i)} + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathcal{N}(0, \mathbf{R}). \end{aligned} \quad (5)$$

where $\mathbf{F}$ refers to the state transition matrix (constant velocity model) while $\mathbf{H}$ refer to the observation matrix; both $\mathbf{Q}, \mathbf{R}$ denote the process noise and observation noise covariance matrices, respectively. A Kalman filter performs a recursive estimation of an optimal state by using two steps, prediction and update:

• Prediction step:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1}^{(i)} &= \mathbf{F} \hat{\mathbf{x}}_{t-1|t-1}^{(i)}, \\ \mathbf{P}_{t|t-1}^{(i)} &= \mathbf{F} \mathbf{P}_{t-1|t-1}^{(i)} \mathbf{F}^\top + \mathbf{Q}. \end{aligned}$$

• **Update step:**

$$\begin{aligned}\mathbf{K}_t^{(i)} &= \mathbf{P}_{t|t-1}^{(i)} \mathbf{H}^\top \left(\mathbf{H} \mathbf{P}_{t|t-1}^{(i)} \mathbf{H}^\top + \mathbf{R} \right)^{-1}, \\ \hat{\mathbf{x}}_{t|t}^{(i)} &= \hat{\mathbf{x}}_{t|t-1}^{(i)} + \mathbf{K}_t^{(i)} \left(\mathbf{z}_t^{(i)} - \mathbf{H} \hat{\mathbf{x}}_{t|t-1}^{(i)} \right), \\ \mathbf{P}_{t|t}^{(i)} &= \left(\mathbf{I} - \mathbf{K}_t^{(i)} \mathbf{H} \right) \mathbf{P}_{t|t-1}^{(i)}.\end{aligned}$$

Adaptive Kalman filter was implemented with a 6-dimensional state vector $[x, y, z, v_x, v_y, v_z]^\top$ for each keypoint, with base process noise covariance $\mathbf{Q} = 10^{-4}$ and observation noise covariance $\mathbf{R} = 10^{-2}$. Positions are predicted during occlusions and updated when observations are available.

Crucially, the noise covariances are not kept constant: the filter adapts them online in three ways. First, a Sage-Husa online estimator [11] updates the observation noise covariance according to Eq. (6),

$$\mathbf{R}_k = (1 - \alpha_{\text{eff}}) \mathbf{R}_{k-1} + \alpha_{\text{eff}} \left(\mathbf{e}_k \mathbf{e}_k^\top - \mathbf{H} \mathbf{P}_k^- \mathbf{H}^\top \right) \quad (6)$$

where $\mathbf{e}_k$ denotes the innovation term, the forgetting factor is $\alpha = 0.05$, and α_{eff} ramps up linearly over 10 steps to stabilize initialization. The observation noise was scaled in a visibility-aware manner, according to Eq.(7).

$$\mathbf{R}_{\text{eff}} = \mathbf{R}_{\text{adaptive}} \times (1 + 4 \max(0, 1 - \text{visibility})), \quad (7)$$

In this case, low-confidence detections were assigned low credibility. Meanwhile, the process noise adapts to the indoor movement magnitude, $\mathbf{Q}_k = \mathbf{Q}_{\text{base}} \times (1 + \min(200 \|\mathbf{v}_k - \mathbf{v}_{k-1}\|, 9))$, which enlarges the process noise during abrupt movements. To prevent divergence, the diagonal entries of the noise covariance matrices are clamped to $0.1 \times 50 \times$ (for $\mathbf{R}$) and $0.5 \times 50 \times$ (for $\mathbf{Q}$) of the base values $\mathbf{Q} = 10^{-4}$ and $\mathbf{R} = 10^{-2}$, respectively. For occluded frames, the filter performs only the predicted steps without updating.

Using this Kalman smoothing and occlusion compensation we obtained smooth and continuous skeleton sequences $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T\}$ with each frame $\mathbf{s}_t \in \mathbb{R}^{33 \times 4}$ containing normalized coordinates (x,y,z) and a visibility flag as the data basis of our dynamic graph representation and gait analysis below.

Dynamic Spatiotemporal Representations

Spatiotemporal embeddings of this kind have shown measurable benefits in other health-related learning tasks, where models that integrate temporal and topological views outperform single-view baselines in Parkinson’s disease subtype identification [12] and progression prediction [13], and evolving-graph spatiotemporal representations improve neuropathological-stage assessment [14].

Single-frame skeletons mainly reflect the instantaneous posture of the human body. In this study, the primary purpose of gait analysis lies in capturing the evolution patterns of joints during indoor movement. Consecutive skeleton sequences are therefore organized into dynamic skeleton sequences and formalized as spatiotemporal representations that simultaneously encode the spatial dependencies of human skeletons and the temporal dynamics of indoor movements.

Given a skeleton sequence $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T\}$ containing T frames, where $\mathbf{s}_t \in \mathbb{R}^{N \times F}$ represents the $N = 33$ keypoints and their $F = 4$ features (normalized coordinates x, y, z and visibility c) in frame t , we construct a spatiotemporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

- **Node set** $\mathcal{V} = \{v_{ti} \mid t = 1, \dots, T, i = 1, \dots, N\}$, where node v_{ti} corresponds to the i -th keypoint in frame t . Each node is associated with a feature vector $\mathbf{f}_{ti} \in \mathbb{R}^F$.
- **Edge set** $\mathcal{E}$ consists of two parts:
 - **Spatial edges** $\mathcal{E}_S$: within the same frame, keypoints are connected according to the natural skeletal connections of the human body (e.g., shoulder–elbow–wrist, hip–knee–ankle), following the predefined topology of MediaPipe Pose [15]. Spatial edges encode the spatial constraints of the human skeleton, allowing graph convolutions to aggregate neighborhood information.
 - **Temporal edges** $\mathcal{E}_T$: between adjacent frames, nodes v_{ti} and $v_{(t+1)i}$ of the same keypoint i are connected. Temporal edges capture the temporal evolution of keypoint trajectories, enabling the model to learn the continuity, periodicity, and symmetry of gait.

The adjacency matrix for this spatiotemporal graph can be factorized as: a spatial adjacency matrix $\mathbf{A}_S \in \mathbb{R}^{N \times N}$ shared among all frames, and a temporal adjacency matrix $\mathbf{A}_T \in \mathbb{R}^{T \times T}$ (sliding window connection), which follows the design in ST-GCN [9]. The spatial graph convolutions are applied on each frame using a learnable edge weight adjacency $\mathbf{A}_{\text{adj}} = \mathbf{A}_S \odot \mathbf{W}_e$, where $\mathbf{W}_e \in \mathbb{R}^{N \times N}$ is a trainable parameter matrix and $\odot$ represents an element-wise product; we use a 1D temporal convolution to connect the same keypoint between adjacent frames. Our factorization is mathematically equivalent to our spatiotemporal graph described above but much more memory efficient.

Spatiotemporal representations compress the raw videos by over four orders of magnitude: a 24-frame 1080p clip contains roughly 1.49×10^8 pixel values, whereas the corresponding dynamic skeleton window contains only $24 \times 33 \times 4 = 3,168$ numbers, which is the root cause of the low inference cost quantified in Table 4. Fig. 3 illustrates the construction of the spatiotemporal graph.

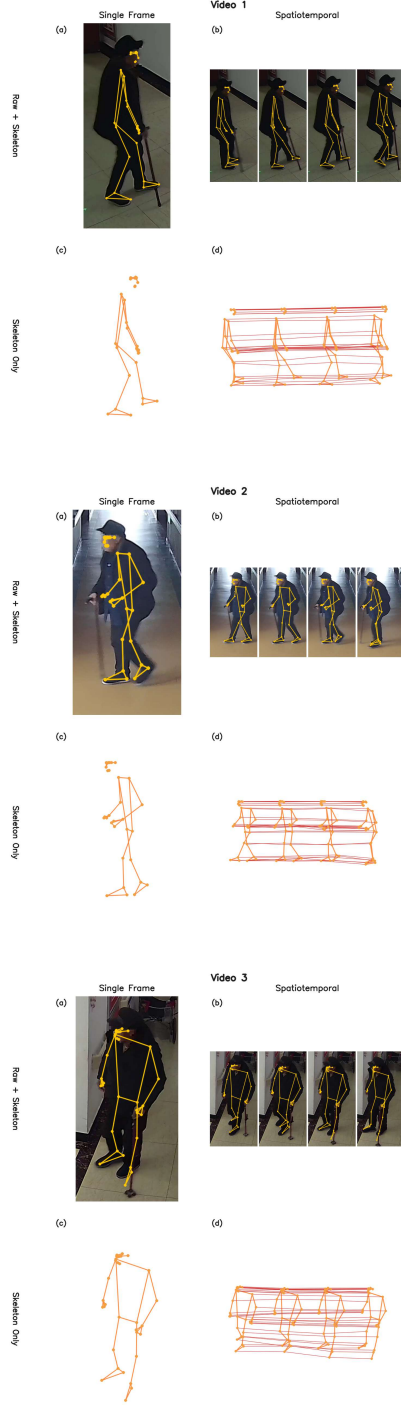


Figure 3. Spatiotemporal representations of gait impairment.

(a) Single-frame spatial graph; (b) multi-frame skeleton sequence; (c) spatiotemporal graph.

In Fig. 3, solid edges denote spatial connections within a single frame, whereas dashed edges denote temporal connections across adjacent frames.

Baseline skeleton models

Static GCN. Static graphs use only the spatial graph of a single frame, ignoring temporal information. Each sample is a single-frame skeleton graph, classified by a vanilla graph convolutional network (GCN).

ST-GCN. The spatio-temporal graph convolutional network (ST-GCN) backbone [9] projects the $F = 4$ input channels to 64 hidden channels, then stacks three spatiotemporal convolution units. Each unit applies (i) a spatial graph convolution with the learnable edge-weight adjacency $\mathbf{A}_{\text{adj}} = \mathbf{A}_S \odot \mathbf{W}_e$, where $\mathbf{W}_e$ is initialized to an all-ones matrix and learned end-to-end so that the model can autonomously strengthen the upper-lower limb coordination patterns characteristic of using crutch, followed by (ii) a one-dimensional temporal convolution with kernel size 3, batch normalization, and ReLU function.

The spatial adjacency includes self-connections and is symmetrically normalized, $\hat{\mathbf{A}}_S = \mathbf{D}^{-1/2} \mathbf{A}_S \mathbf{D}^{-1/2}$. After the stacked units, global average pooling over all nodes and frames produces a 64-dimensional embedding, which is classified by a two-layer fully connected head ($64 \rightarrow 32 \rightarrow 4$) with dropout 0.5. Skeleton-based gait classification models were trained with the Focal Loss [16]. Focal loss down-weights the abundant easy samples of the majority classes and focuses learning on the rare abnormal gaits, formulated as follows.

$$\mathcal{L}_{\text{focal}} = -\frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C y_{i,c} (1 - p_{i,c})^\gamma \log p_{i,c}, \quad (8)$$

where B denotes the batch size, $C = 4$ is the number of gait categories, $p_{i,c}$ the predicted probability of sample i belonging to the class c , and $\gamma = 2$ is the focusing parameter.

Why only skeleton baselines. We benchmark only skeleton-based models. The released data are skeleton sequences; the raw RGB clips are not part of this deposit, so any video-based baseline would be unreproducible for users of the dataset. Skeleton-based recognition has since moved from graph convolutional networks to transformer architectures such as SkateFormer [17], and pose-based fall-detection continues to rely on UR-Fall and UP-Fall [18]; the two baselines here are kept deliberately simple so that every reported number can be reproduced from the released code without external data.

Data augmentation

To address sample imbalance and improve model generalization, several augmentations are applied to skeleton sequences during training (see `config.py` in the released code for the parameters):

Spatial augmentation: random translation, rotation, and scaling of keypoint coordinates to simulate different camera angles and distances ($\pm 5\%$ translation, $\pm 5^\circ$ rotation, $\pm 5\%$ scaling, plus Gaussian coordinate noise with $\sigma = 0.01$).

Temporal augmentation: random cropping (retaining 70–100% of the frames) followed by linear interpolation back to the original length to simulate variations in movement speed.

Joint noise: Gaussian noise added to keypoint coordinates to improve robustness against detection errors.

Occlusion augmentation (probability 0.1): a skin-colored rectangular patch covering 15–35% of the frame is simulated by setting the visibility of the enclosed keypoints to 0.2 and perturbing their coordinates, mimicking the detector failures caused by real occluders.

Sitting-specific augmentation: random knee bending (probability 0.3, simulating knee angle changes in sitting) and lowering of the center-of-mass height to increase the diversity of sitting samples.

During training we additionally apply a high dropout rate (DROPOUT=0.5) together with the occlusion augmentation above, forcing the model to learn robust feature representations.

Leakage-free evaluation protocol

Skeleton-based gait classification models are evaluated under an identical video-stratified K -fold cross-validation protocol ($K = 5$, random seed 42). The 247 videos that contain at least one 24-frame window are partitioned into K class-stratified folds; the remaining two released clips (14 and 17 frames) are shorter than a single window and are therefore released in `skeletons/` but not part of the benchmark (see Supplementary Note S2). Each fold serves once as the test set, while a class-stratified 15% of the remaining videos is held out as a validation set for early stopping and model selection. Every window extracted from a video is confined to the same partition as that video, so the model never sees any frame of a test video during training; the out-of-fold (OOF) predictions from all folds are then pooled to compute the reported metrics.

To counteract class imbalance, the training indices of each fold are oversampled with replacement up to class balance, capped at a $5\times$ amplification per class, while the validation and test distributions are kept intact. Windows are extracted strictly within contiguous frame segments, and each window inherits the majority label of its 24 frames. If the smallest class has fewer videos than K , the number of folds is automatically reduced to that count so that every fold still contains every class.

The key training hyperparameters of the skeleton-based gait classification models are listed in Supplementary Table S4.

Validation experiments are conducted on a workstation equipped with an Intel Core i9-14900HX CPU, an NVIDIA GeForce RTX 4060 GPU (8 GB), and 32 GB RAM, running Windows 11 with Python 3.11 and PyTorch 2.5.1 (CUDA 12.1).

Ethics statement

The visual data were recorded in the public spaces (corridor, lobby, activity and dining areas) of a residential elderly care facility, with the written authorization of the facility for a non-commercial preliminary study on fall-risk monitoring. The collection involved no clinical intervention, no biological samples, and no personal identification of individuals, and therefore fell outside the scope of biomedical ethics review at our institution; no ethics approval number is applicable. Recording was carried out under the supervision of care-facility staff. Only anonymized skeleton sequences (numerical keypoint coordinates) are released publicly; no facial imagery, names, or direct identifiers are included in the release. Raw RGB video frames are withheld from this release; only anonymised skeleton sequences are distributed. Any future release of de-identified imagery would require renewed written authorization from the care facility.

Data Records

The primary release of ElderlyPipe is deposited at Zenodo [19], under DOI 10.5281/zenodo.23028481. All skeleton data are anonymized numerical sequences; no raw facial imagery is distributed. The 249 released clips were captured across five indoor sub-areas (activity room, activity area, corridor, lobby and dining area); three of them are illustrated in Fig. 2. Lighting, occlusion and floor material vary widely across scenes, mirroring real elderly care environments. The release is organized as follows:

```
elderlypipe/
  README.md           # format and usage
  metadata.csv         # per-clip metadata
  labels.csv           # behaviour labels
  keypoint_index.csv   # BlazePose mapping
  skeletons/           # one file per clip
    <clip_id>.npz      # T x 33 x 4 float32
  splits/              # 5-fold CV splits
    fold{1..5}_train.txt
    fold{1..5}_val.txt
    fold{1..5}_test.txt
  seed.txt             # split seeds
  windows/             # dynamic windows
    windows.npz        # shape (N,24,33,4)
    windows_index.csv
```

Skeleton files. Each clip is stored as a $T \times 33 \times 4$ float32 array (one file per clip in `skeletons/`); the four channels per keypoint are (x, y, z , visibility) in that order.

Each `.npz` file additionally stores the frame indices and the per-frame behaviour label alongside the landmark array (see Supplementary Note S2). The 33-landmark convention follows MediaPipe BlazePose [20], with the full index-to-landmark mapping released as `keypoint_index.csv`. Compensated keypoints carry visibility 0.6 (interpolated) or 0.8 (geometric), and the released sequences contain no facial imagery. Detailed field descriptions are given in Supplementary Note S2.

The per-window index fields (`window_id`, source `clip_id`, segment identifier, start frame, and majority label) are documented in Supplementary Note S2.

The full field list of `metadata.csv` is documented in Supplementary Note S3.

Split files. The files under `splits/` implement the video-stratified 5-fold protocol described in Section Leakage-free evaluation protocol: each file lists the `clip_ids` assigned to the training, validation, or test partition of one fold, guaranteeing that all windows of a clip remain in a single partition. The split is deterministic with random state 42 (see `splits/seed.txt`); when a class contains fewer clips than the number of folds, the number of folds is automatically reduced so that every fold still contains at least one clip per class.

README. `README.md` documents the file formats, the loading API, the licence, and the changelog of the release.

Data Overview

The ElderlyPipe release contains four behaviour classes with a long-tailed distribution: `walking` dominates, `using_crutch` is the clinically critical minority class, and `sitting` and `standing` are the smallest categories. The class counts at three granularities: clip, frame, and dynamic window are reported in Table 1. The release is intended as a benchmark for skeleton-based recognition of clinically relevant unbalanced gaits; the imbalance shown here is part of the dataset’s ecological validity rather than a limitation.

Technical Validation

The technical quality of the ElderlyPipe release is validated along four axes, without claiming downstream scientific findings. Readers are referred to Section Data Overview for the class distribution and to Section Leakage-free evaluation protocol for the evaluation protocol. First, the benchmark results in Table 2 show that the released data support model training and expose the genuine difficulty of the task: the minority `using_crutch` class, which is the clinically critical unbalanced state of this dataset, is examined separately in Table 3. Second, the error analysis of the

Table 1. Class distributions at three granularities. Per-class video counts sum to more than the 249 distinct clips because a single clip may contain segments of multiple behaviours; the frame and window counts are mutually exclusive. Numbers correspond to release v1.1.

Class	Videos	Frames	Windows
walking	173	37,224	16,600
using_crutch	42	16,236	7,623
sitting	29	6,033	2,649
standing	26	4,362	1,863

predictions is inspected to confirm that misclassifications reflect genuine behavioural ambiguity rather than annotation noise. Third, the effect of the quality-enhancement pipeline is verified visually and by human verification of every released frame. Fourth, the computational footprint of the released models is measured to confirm that the representation supports real-time processing on consumer-grade hardware, while validation on dedicated edge devices is left to future work.

Baseline benchmarks under the leakage-free protocol

To demonstrate that ElderlyPipe supports model training and to characterize the intrinsic difficulty of the dataset, two representative skeleton-based baselines are trained and evaluated under the video-stratified 5-fold out-of-fold (OOF) protocol of Section Leakage-free evaluation protocol: Static GCN, which takes single frames as input and is therefore referred to as frame-level, and ST-GCN, which takes 24-frame windows as input and is referred to as window-level. The overall metrics are reported in Table 2. Static GCN operates on released single frames, whereas ST-GCN operates on the released 24-frame windows; both are evaluated under the same video-stratified OOF protocol. Each reported number is the mean and range (max–min) across five independent training runs (seeds 42, 49, 56, 63, and 70), each rerunning the full 5-fold OOF training with its own fold split, model initialization, and augmentation randomness. The released split files correspond to the seed-42 run; the remaining four seeds serve only to quantify run-to-run variance and are not part of the deposit. The full run protocol and hardware configuration are given in Supplementary Note S4. Per-class one-vs-rest receiver operating characteristic (ROC) curves of the two baselines are shown in Fig. 4c,d.

ST-GCN outperforms Static GCN on every overall metric (F1 0.656 vs. 0.605), confirming that temporal dependence between frames is essential for capturing gait dynamics: a single-frame graph sees only the instantaneous pos-

ture, whereas the spatiotemporal graph learns the continuity, periodicity, and left–right coordination characteristic of gait. The two models share the same training objective (Focal Loss with $\gamma = 2$) and the same minority-oversampling protocol, so the gap reflects the contribution of temporal information rather than a difference in class-imbalance handling.

Validation on the minority crutch-assisted walking class

The clinically critical minority class `using_crutch` is examined separately; per-class results for this class are reported in Table 3. Within the skeleton-based family, the five-run results put the two baselines at `using_crutch` recall of 0.506 ± 0.118 (ST-GCN) and 0.524 ± 0.173 (Static GCN); the ranges overlap heavily, so no recall gain can be attributed to temporal information (see Supplementary Note S5 for why AUC and recall can diverge on an imbalanced minority class). ST-GCN does lead on `using_crutch` precision (0.492 vs. 0.431), F1 (0.498 vs. 0.470), and AUC (0.706 vs. 0.684). For fall monitoring, where missing a `using_crutch` user is far costlier than a false alarm, recall is the relevant quality indicator, and improving it on this minority class remains an open challenge for both baselines.

Error analysis

The OOF confusion matrices of the two baselines (Fig. 4a,b) reveal three error patterns, described below for the ST-GCN baseline (Fig. 4b). The single largest is standing being predicted as walking: on average only about 20% of standing windows are correctly recognised, with roughly 65% absorbed by the majority walking class and the remainder mostly by `using_crutch`. This is not an annotation error; the two classes share nearly identical postures in a single 24-frame window, and short standing segments that border walking segments within the same clip are inherently transitional. Second, `using_crutch` and walking are confused in both directions: roughly 39% of `using_crutch` windows are predicted as walking and 20% of walking windows as `using_crutch`, reflecting the graded nature of crutch-assisted walking, in which mild support is visually close to unassisted walking, particularly when the crutch is partially occluded. Third, `sitting` is reliably separated from all other classes (recall 0.85, precision 0.99); `standing` and `sitting` produce almost no mutual confusion, indicating that the annotation protocol resolves sit-to-stand and stand-to-sit transitions cleanly even though the two classes share a posture in a single frame. Static single-frame graphs cannot resolve the borderline cases: under the 5-fold OOF protocol, Static GCN lags ST-GCN by 5.1 points of overall F1 (0.605 vs. 0.656), confirm-

ing that the temporal information encoded in the released sequences carries discriminative signal, although the two baselines are evaluated on their native sample units (single frames vs. 24-frame windows), and on `using_crutch` recall specifically the two baselines remain comparable (0.524 vs. 0.506).

Validation of the occlusion compensation and smoothing pipeline

The two-stage enhancement is applied to every released frame: hierarchical occlusion compensation (linear interpolation for gaps shorter than $T_{\max} = 15$ frames, Eq. (1), and geometric constraint estimation for longer gaps, Eqs. (2)–(4)), followed by visibility-aware adaptive Kalman smoothing (Eqs. (5)–(7)). Across the 63,855 released frames (2,107,215 keypoint values), 2.29% of keypoints were restored by linear interpolation (confidence 0.6) and 2.82% by geometric estimation (confidence 0.8); the remaining 94.9% are directly observed detections (see Supplementary Note S6 for the full statistics and for how the assigned confidence values can be verified from the released files). Segments in which the detector lost a keypoint are completed by the compensation stage, and the high-frequency jitter of per-frame detection is suppressed by the visibility-aware adaptive filter, while the underlying gait rhythm is preserved. During dataset construction, every frame that survives the pipeline is additionally verified by human annotators against the rendered skeleton overlay (see the annotation procedure in the Methods section); frames failing verification are excluded, leaving 63,855 quality-assured frames.

Computational footprint of the released models

Table 4 reports the parameter counts, computational cost (FLOPs), and storage size of each baseline, measured on the released implementations. The complete ElderlyPipe chain—BlazePose extraction, hierarchical occlusion compensation, visibility-aware adaptive Kalman smoothing, and ST-GCN inference—sustains 28 frames per second end-to-end for 1080p video on a consumer-grade laptop CPU (Intel Core i9-14900HX), without requiring a GPU; the smoothing module alone costs less than 2 ms per frame for all 33 keypoints. The released skeleton representation can therefore be processed in real time on consumer-grade hardware, suggesting feasibility for on-site deployment in care facilities; validation on dedicated edge devices is left to future work.

Table 2. Baseline performance under the video-stratified 5-fold OOF protocol (skeleton-based gait classification models: weighted average (Static GCN: frame-level; ST-GCN: window-level), macro AUC). Values are mean $\pm$ range (max–min) across five independent training runs.

Method	Acc	Prec	Rec	F1	AUC
ST-GCN	0.653 $\pm$ 0.029	0.662 $\pm$ 0.019	0.653 $\pm$ 0.029	0.656 $\pm$ 0.024	0.763 $\pm$ 0.031
Static GCN	0.582 $\pm$ 0.016	0.650 $\pm$ 0.062	0.582 $\pm$ 0.016	0.605 $\pm$ 0.014	0.743 $\pm$ 0.011

Table 3. Baseline performance on the minority using_crutch class under the 5-fold OOF protocol (Precision/Recall/F1/AUC). Values are mean $\pm$ range (max–min) across five independent training runs (seeds 42, 49, 56, 63, 70).

Method	Precision	Recall	F1	AUC
ST-GCN (window-based)	0.492 $\pm$ 0.073	0.506 $\pm$ 0.118	0.498 $\pm$ 0.077	0.706 $\pm$ 0.036
Static GCN (frame-based)	0.431 $\pm$ 0.026	0.524 $\pm$ 0.173	0.470 $\pm$ 0.070	0.684 $\pm$ 0.016

Table 4. Model complexity comparison (measured on the released implementations)

Method	Params (M)	FLOPs (M)	Model size (MB)
Static GCN	0.092	6.851	0.364
ST-GCN	0.455	88.303	1.762

Usage Notes

ElderlyPipe is released as a complete, leakage-controlled package of real-world indoor elderly gait data, so that other researchers can directly train, evaluate, and deploy models for the crutch-assisted and everyday behaviours of elderly care.

The window array uses the layout $(N, 24, 33, 4)$ with channels ordered $(x, y, z, \text{visibility})$.

Recommended evaluation protocol. Users should evaluate new models with the released video-stratified 5-fold splits and report out-of-fold aggregated metrics. Random window-level splitting is strongly discouraged: because consecutive windows of one clip are highly correlated and subjects repeat across windows, window-level splits leak subject identity and inflate performance. If the smallest class constrains the fold count, the protocol reduces K automatically (Section Leakage-free evaluation protocol).

At inference time, the released code mirrors the training-side hygiene: the filter and compensator states are reset whenever pose detection fails for more than one consecutive frame, so that a window is never assembled across a temporal gap; the displayed decision is smoothed over the most recent 15 window predictions for stability; and the five fold-models can be ensembled by probability averaging to reduce single-fold variance on this small-scale dataset.

Users should be aware of the following limitations. First, the standing class remains the smallest category with

only 26 videos, which limits generalization on this class—the ST-GCN baseline attains a mean F1 of only 0.18 on standing; we report this value openly rather than hiding it behind weighted averages, and minority-class results should be interpreted with corresponding caution. Second, the released dynamic-graph construction uses temporal connections between adjacent frames only and does not encode cross-period temporal dependencies (e.g., gait-cycle-level relations). Third, the released baselines are representative but not exhaustive and no paired significance tests between models have been performed; such same-protocol comparisons and statistical tests are important future work. Fourth, the current release does not distinguish crutch types or unilateral versus bilateral crutch use, and the release tracks a single person per frame and does not include multi-person interaction scenarios; these metadata and scenario extensions are left to future releases. Planned extensions include enriching the standing class, advanced-age subgroups, and multi-person interaction samples. With regards to gait speed, double-support duration, and left-right asymmetry, subsequent works could incorporate gait biomechanical features as explicit physical inputs. Weakly supervised labeling strategies could be investigated to reduce annotation cost.

Data Availability

The ElderlyPipe dataset (anonymized skeleton sequences, behaviour labels, metadata, and the video-stratified cross-validation splits described in Data

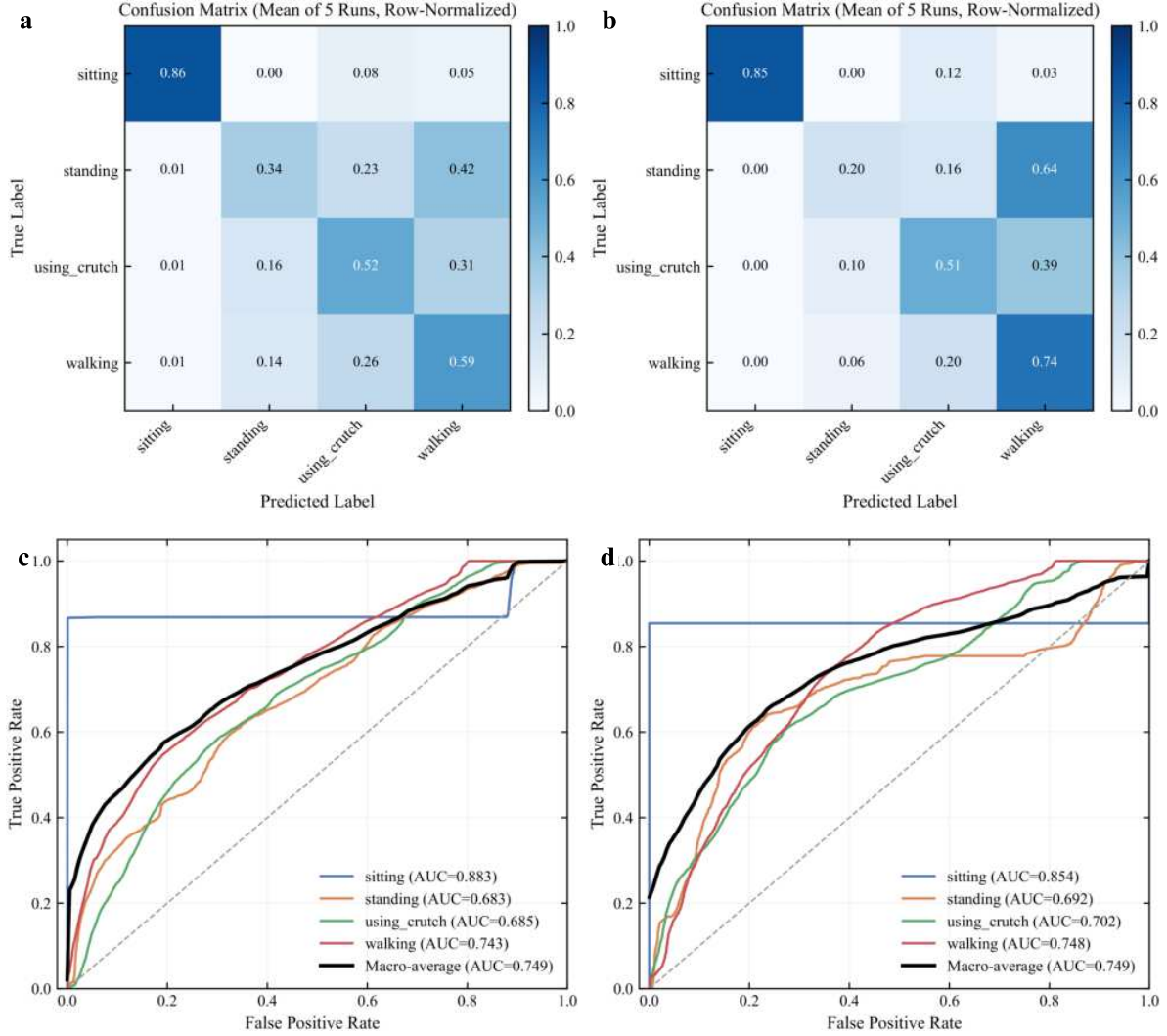


Figure 4. Baseline recognition results under the video-stratified five-fold out-of-fold protocol. **a**, Confusion matrix of the Static GCN baseline, averaged over five independent runs. **b**, Confusion matrix of the ST-GCN baseline, averaged over five independent runs. **c**, Per-class one-vs-rest ROC curves of the Static GCN baseline (one representative run of five). **d**, Per-class one-vs-rest ROC curves of the ST-GCN baseline (one representative run of five).

Records) is deposited in the Zenodo repository and is openly available at <https://doi.org/10.5281/zenodo.23028481> under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. De-identified RGB video frames are not included in this record; only anonymised skeleton sequences are released, and no part of the dataset is available only upon request.

Code Availability

The ElderlyPipe processing pipeline (BlazePose extraction, hierarchical occlusion compensation, visibility-aware adaptive Kalman smoothing), the dynamic-window construction, and the complete training and evalua-

tion code reproducing all baseline results and figures are available at <https://github.com/TurkeyJackie/ElderlyPipe>. The release corresponding to this paper is tagged as v1.0 in the same repository. The code requires Python 3.11, PyTorch 2.5.1, and MediaPipe; exact dependencies are pinned in the repository. The code is released under the MIT licence.

References

References

- [1] World Health Organization. Falls. Fact sheet, 2021.

- [2] Yao Xu. Application research of wearable devices in exercise risk management for the elderly. *Sports World*, (01):76–78, 2025. (in Chinese).
- [3] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.
- [4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset, 2017.
- [5] Roman Chereshnev and Attila Kertész-Farkas. Hugadb: Human gait database for activity recognition from wearable inertial sensor networks. In *Analysis of Images, Social Networks and Texts (AIST 2017), Lecture Notes in Computer Science*, volume 10716, pages 131–141. Springer, 2018.
- [6] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3):489–501, 2014.
- [7] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafoort-Asturiano. UP-Fall detection dataset: A multimodal approach. *Sensors*, 19(9):1988, 2019.
- [8] Vida Adeli, Ivan Klabucar, Javad Rajabi, et al. CARE-PD: A multi-site anonymized clinical dataset for parkinson’s disease gait assessment. In *Advances in Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track*, volume 38. Curran Associates, Inc., 2025. arXiv:2510.04312.
- [9] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, pages 7444–7452, 2018.
- [10] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions of the ASME*, 82:35–45, 1960.
- [11] Andrew P. Sage and G. W. Husa. Adaptive filtering with unknown prior statistics. In *Proceedings of the 1969 Joint Automatic Control Conference*, pages 760–769, Boulder, CO, USA, 1969.
- [12] Wei Zhang, Zeqi Xu, Ruochen Yu, Mingfeng Jiang, and Qi Dai. DualGCN-GE: integration of spatiotemporal representations from whole-blood expression data with dual-view graph convolution network to identify Parkinson’s disease subtypes. *BMC Bioinformatics*, 26, 2025.
- [13] Wei Zhang, Zeqi Xu, Chenjun Wu, and Qi Dai. Multi-view multi-scale dynamic graph attention network (GAT) based prediction of Parkinson’s disease progression using whole-blood RNA sequencing data. *Scientific Reports*, 16(1):11777, 2026.
- [14] Wei Zhang, Ruochen Yu, Chengjie Ding, Mingfeng Jiang, and Qi Dai. STAD-CoAtt: integration of evolving gene graphs in the assessment of neuropathological stages using spatiotemporal representations of brain transcriptomics data. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(6):2882–2894, 2025.
- [15] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines, 2019.
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [17] Jeonghyeok Do and Munchurl Kim. SkateFormer: Skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision (ECCV)*, volume 15099 of *Lecture Notes in Computer Science*, pages 401–420. Springer, 2024. arXiv:2403.09508.
- [18] Xiaoqun Yu, Chenfeng Wang, Wenyu Wu, and Shuping Xiong. A real-time skeleton-based fall detection algorithm based on temporal convolutional networks and transformer encoder. *Pervasive and Mobile Computing*, 107, 2025.
- [19] Shaojiang Huang, Manqing Yang, and Wei Zhang. ElderlyPipe: A real-world indoor skeleton dataset of elderly abnormal gait including crutch-assisted walking, 2026.
- [20] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

Acknowledgements

We sincerely thank the participating elderly care facility, its staff, and the residents and their families for their support and cooperation during data collection. The facility prefers to remain anonymous.

Author Contributions

S.H.: conceptualization, methodology, software, validation, investigation, data curation, writing – original draft, writing – review & editing, visualization. M.Y.: data curation, investigation, validation, writing – review & editing. W.Z.: conceptualization, resources, supervision, project administration, writing – review & editing. All authors reviewed and approved the final manuscript.

Funding

This work is partly sponsored by the National Natural Science Foundation of China (62302455), and the Natural Science Foundation of Zhejiang Province, China (LQ24F030007).

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)