scientific reports

Check for updates

OPEN

Interpretable and lightweight fall detection in a heritage gallery using YOLOv11-SEFA for edge deployment

Siqi Wu[1], Hao Yang[2,3]✉, Yanfeng Hu[4], Xiaoke Ji[5] & Si Cheng[4]

Falls are a critical safety risk in aging societies, causing severe injuries and fatalities, particularly in urban public buildings where elderly visitors frequently gather. Cultural and heritage spaces such as museums and galleries present additional challenges for monitoring due to complex lighting, reflective display cases, and fluctuating visitor densities, underscoring the need for reliable fall detection systems that can be seamlessly deployed without intrusive infrastructure. This study proposes an interpretable, lightweight fall detection and alert system based on the YOLOv11-SEFA architecture. The model integrates P2 feature enhancement and SimAM attention into the YOLOv11n backbone, achieving consistent detection reliability while maintaining low computational cost. A four-layer sensing-to-cloud pipeline is combined with random forest classification of six-dimensional structural features to predict multi-level fall risk, with feature importance analysis verifying aspect ratio, distance to camera, and crowd presence as key predictors aligned with safety logic. The system demonstrates stable performance across confusion matrices, PR curves, and ROC-AUC learning curves, indicating operational feasibility and edge suitability. Practical tests show sub-270 ms latency, low power and bandwidth requirements, and smooth integration into weak-current infrastructures. Pilot validation at Rochfort Gallery, a restored 1920s heritage building in North Sydney, demonstrates feasibility under real-world conditions, supporting future deployment in smart city health and safety applications.

Keywords Deep learning, Edge intelligence, Fall detection, Heritage building, Risk prediction methods

Abbreviations

AI	Artificial Intelligence
SHAP	SHapley Additive exPlanations
YOLO	You Only Look Once
GFLOPs	Giga Floating Point Operations per Second
SimAM	Simple Attention Module
SEFA	Structure-Enhanced Feature Attention
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
2MP PoE	2 Megapixel Power over Ethernet (Camera)
GStreamer	GStreamer Multimedia Framework
EIB	Edge Intelligence Layer
Grad-CAM	Gradient-weighted Class Activation Mapping
TLS-encrypted MQTT channel	Transport Layer Security-encrypted Message Queuing Telemetry Transport channel
RGB	Red Green Blue (Color Model)
HCI	Human–Computer Interaction
ConvLSTM	Convolutional Long Short-Term Memory
NAS	Neural Architecture Search

[1]Faculty of Arts, Design & Architecture, UNSW Sydney, Kensington, Sydney, NSW 2052, Australia. [2]Design Academy, Kashi University, No. 29, Naizibag Road, Kashi 844000, Xinjiang, China. [3]National Research Center of Cultural Industries, Central China Normal University, No. 152, Luoyu Road, Hongshan District, Wuhan City 430079, Hubei Province, China. [4]Faculty of Design and Architecture, Universiti Putra Malaysia, Jalan Universiti, Serdang 43400, Selangor, Malaysia. [5]School of Art and Design, UNSW Sydney, Kensington, Sydney, New South Wales 2052, Australia. ✉email: ksdx_yh@163.com

ST-Transformer	Spatio-Temporal Transformer
GNN-TCN	Graph Neural Network with Temporal Convolutional Network
IoU	Intersection over Union

The world is rapidly entering an aging society. According to the United Nations World Population Prospects 2022, the population aged 60 years and above will reach 2.1 billion by 2050, representing 22% of the global population[1]. Falls are one of the most serious health risks for older adults, causing approximately 684,000 deaths annually and ranking second among unintentional injury deaths after road traffic accidents[2]. Due to age-related declines in physical function and reflexes, older adults are especially vulnerable, highlighting the urgent need for responsive and scalable fall detection systems in heritage gallery governance.

Various detection technologies have been explored, including wearable devices[3], environmental sensing systems[4], and non-contact video-based methods[5]. Early systems primarily employed triaxial accelerometers and gyroscopes[6–8] integrated into Personal Emergency Response Systems (PERS). More recent studies have combined artificial intelligence (AI) with sensor data to enhance detection accuracy[9,10]. While low-cost and easy to deploy, wearable devices suffer from poor user compliance and comfort issues, limiting their widespread adoption among older adults[11].

Environmental sensing solutions such as pressure or infrared sensors offer non-intrusive alternatives but face challenges including high deployment costs, environmental interference, and restricted coverage[12–14]. Vision-based methods using indoor cameras have emerged as effective for behavioral recognition, but their deployment in private areas raises ethical and legal concerns, especially under regulations such as the GDPR and China's Personal Information Protection Law[15]. Consequently, traditional approaches remain constrained in large-scale heritage gallery applications by both technical and societal barriers.

Advances in deep learning and computer vision have revitalized interest in video-based fall detection as a non-contact, real-time solution. Leveraging existing public surveillance infrastructure, these methods are particularly suitable for urban public spaces such as galleries, museums, and community centers[16]. The YOLO (You Only Look Once) family of single-stage detectors has proven especially effective for real-time applications in security and traffic monitoring due to its high frame rate and low latency[17–19]. Recent variants have been tailored for fall detection. For example, Raza et al.[20] adapted YOLOv5 on the UR-Fall dataset, Zhao et al.[21] incorporated GSConv and multi-branch DBB into YOLOv7 for improved efficiency, and Hwuang et al.[11] integrated Transformer modules into YOLOv9 to achieve an mAP@0.5 of 0.982. Overall, existing YOLO-based fall detection studies predominantly emphasize architectural refinements and benchmark-level performance improvements under controlled datasets. While some works have incorporated post-detection logic—such as heuristic rules or confidence-based thresholds for alarm triggering—these mechanisms are typically limited in scope and are not systematically designed to support graded risk assessment or differentiated response strategies in complex public environments. Comparatively less attention has been paid to how detector design choices interact with real-world public-space constraints (e.g., crowding, occlusion, and edge deployment), and how detection outputs can be further structured into interpretable, multi-level risk representations beyond binary alarms.

To address these limitations, this study presents an interpretable and lightweight fall detection and alerting system built upon YOLOv11, explicitly designed for complex public indoor environments such as heritage galleries. Rather than pursuing generic architectural expansion, the proposed system adopts a set of detection and deployment choices motivated by practical challenges observed in such spaces, including occlusion, long-range viewpoints, dense visitor flow, and strict constraints on latency, bandwidth, and privacy. Specifically, a P2 detection branch and a lightweight SimAM attention module are incorporated to improve sensitivity to small or partially occluded human postures frequently encountered in gallery settings. A multi-camera fusion strategy is introduced at the system level to suppress spurious detections arising from reflections and background clutter, while edge-based inference supports low-latency operation and reduced data transmission. Beyond binary fall detection, the system extracts a compact set of six semantic features derived from image-level structural attributes (e.g., body aspect ratio, relative camera distance, and crowd presence), which are subsequently analyzed using a random forest classifier to produce graded fall-risk levels with interpretable feature attributions. Pilot validation has been initiated at Rochfort Gallery, a restored 1920s heritage building in North Sydney, to examine system behavior under challenging lighting conditions and variable visitor densities.

The main contributions of this study can be summarized as follows: (i) A scenario-driven fall detection pipeline is developed for large public indoor spaces, with a particular focus on heritage gallery environments, where occlusion, crowding, and lighting variability pose challenges that are not fully addressed in existing benchmark-oriented studies. (ii) A lightweight detector configuration based on YOLOv11 is presented, in which a P2 detection branch and a minimal attention mechanism are selectively employed to improve robustness to small-scale and partially occluded fall patterns, while maintaining suitability for edge deployment. (iii) A three-tier system framework encompassing camera perception, edge processing, and cloud integration is described, illustrating how fall detection can be integrated into a practical monitoring pipeline under real-time, power, and privacy constraints. (iv) A post-detection risk grading module based on six semantic features and a random forest classifier is introduced to extend fall detection beyond binary alarms, providing interpretable risk-level outputs that support differentiated response strategies in intelligent public environments.

Materials and methods

Rochfort gallery: case study background

Rochfort Gallery, located at 317 Pacific Highway in North Sydney, is housed in a meticulously restored heritage building dating back to the 1920s. Originally constructed as a Masonic Temple, the site has been repurposed into a contemporary exhibition gallery while retaining its architectural heritage features, including high ceilings,

Fig. 1. Location and architectural features of Rochfort Gallery in North Sydney, Australia. Base map data OpenStreetMap contributors. All annotations and graphical elements are created by the authors. Photographs were taken by the authors during on-site deployment.

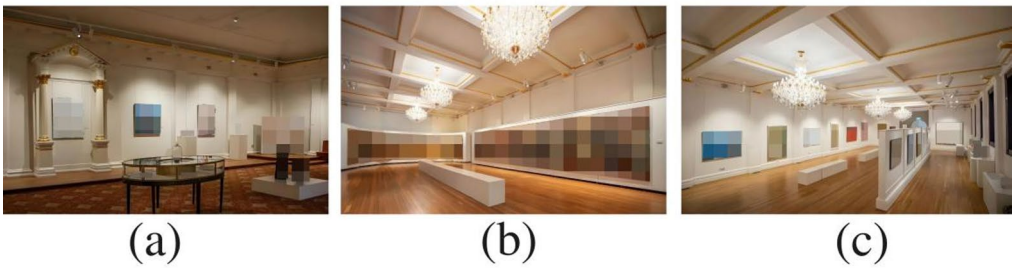

Fig. 2. Interior views of Rochfort Gallery: (**a**) a neoclassical exhibition hall with heritage decorative finishes and glass display furniture; (**b**) a large gallery space with a panoramic mural and chandelier lighting; and (**c**) a contemporary exhibition hall with partition walls and variable artwork displays. These representative settings reflect the architectural and environmental complexity of the heritage-protected site. Photographs taken by the authors.

ornate interior finishes, and reflective display cases, as shown in Fig. 1. These characteristics make the gallery both a culturally significant venue and a challenging environment for intelligent monitoring systems.

From a research perspective, Rochfort Gallery provides a representative case study for fall detection in complex public cultural spaces, as shown in Fig. 2. The venue attracts diverse visitor groups, including elderly audiences, and its spatial layout comprises exhibition halls, corridors, and transitional lobbies. Representative views of the interior are illustrated in panels (a)–(c), showing a neoclassical exhibition hall with decorative finishes, a large-scale mural gallery, and a modernized hall with reflective flooring and chandelier lighting. Such environments exhibit several conditions that complicate video-based monitoring: (i) variable lighting, ranging from natural daylight through windows to directed spotlights on artworks; (ii) fluctuating visitor densities, with crowded conditions during exhibition openings and sparse traffic at other times; and (iii) architectural constraints, since heritage protection regulations limit intrusive modifications such as permanent cabling or drilling for device installation. These factors collectively highlight the necessity of developing a fall detection system that is not only accurate but also lightweight, interpretable, and deployable under constrained infrastructure conditions. The pilot validation at Rochfort Gallery therefore serves two purposes: first, to evaluate the proposed YOLOv11-SEFA model under real-world, heritage-sensitive conditions, and second, to demonstrate the feasibility of integrating intelligent safety monitoring into culturally significant public spaces without compromising privacy, visitor experience, or architectural integrity.

The pilot validation at Rochfort Gallery therefore serves two purposes: first, to evaluate the proposed YOLOv11-SEFA model under real-world, heritage-sensitive conditions, and second, to demonstrate the feasibility of integrating intelligent safety monitoring into culturally significant public spaces without compromising privacy, visitor experience, or architectural integrity. Crucially, a subset of data collected from Rochfort Gallery was primarily used for pilot-level evaluation to examine system behavior under real deployment conditions, rather than for large-scale quantitative benchmarking. Consequently, all reported field performance metrics (false alarm rate, latency under crowd load, nighttime behavior) were computed exclusively from the Rochfort Gallery deployment. It should be noted that the site-specific evaluation was conducted on a limited number of

samples, and results should be interpreted as indicative of operational feasibility rather than exhaustive statistical performance. Due to the exploratory nature of the deployment, field data is reported as aggregated operational metrics (e.g., average latency, hourly alarm rates) rather than as a granularly stratified dataset, paving the way for future structured site-disjoint evaluations.

Data prepossessing and dataset construction

High-quality datasets are essential for building reliable and generalizable object detection models. However, most publicly available fall detection datasets are limited by insufficient sample sizes, constrained scene diversity, and inconsistent annotations, which makes them inadequate for modeling real-world safety scenarios in indoor public environments. To address this gap, we extended the FPID dataset[22] by incorporating additional images collected from museums, art galleries, and community centers, thereby enriching scene diversity in terms of viewpoints, postures, and environmental conditions and enhancing suitability for deployment in smart building contexts.

The final dataset contains 3,416 high-resolution RGB images. To ensure a balanced learning target, the dataset maintains a near-balanced binary classification setting: 1,768 samples (51.8%) are labeled as normal actions, and 1,648 samples (48.2%) are labeled as fall states. As illustrated in Fig. 3, panel (a) represents upright standing postures captured from multiple viewpoints (front, back, side, and oblique), whereas panel (b) shows representative fall postures, including supine, prone, side-lying, and curled positions. This categorical separation guided the annotation process and ensured consistency in labeling across all images.

To achieve sample diversity and balance, image acquisition was conducted across a wide range of environmental conditions, as illustrated in Fig. 4. Three different camera viewpoints were considered: (a) top-down views simulating ceiling-mounted surveillance cameras, (b) side views mimicking wall-mounted devices, and (c) front views resembling human-eye level recordings. The dataset maintains a relatively uniform distribution across these perspectives (approximately 35% side views, 35% front views, and 30% top-down views). We explicitly controlled the sampling process to ensure that the ratio of fall to non-fall samples remained approximately 1:1 within each environmental sub-category, preventing class bias in specific scenarios. To support reproducibility and quantitatively characterize the dataset's complexity, we analyzed the statistical distribution across key environmental variables: First is illumination variability: images were collected under (d) low-light conditions, (e) natural daylight, and (f) directed spotlights. Statistically, normal indoor lighting dominates (46.3%), followed by low-light conditions (28.6%) and strong or directed lighting (25.1%), effectively covering the lighting spectrum typical of exhibition areas. While precise lux meters were not feasible for all diverse sources, illumination levels were categorized based on histogram intensity analysis. Second is occlusion levels: to represent realistic obstructions, three scenarios were included: (g) partial occlusion, (h) scenes with no occlusion, and (i) background interference. Quantitatively, 62.7% of samples exhibit no occlusion. However, to challenge the model, 19.4% show partial occlusion and 17.9% involve heavy occlusion or complex background interference, primarily caused by exhibition structures, furniture, or overlapping visitors. Third is crowd density:

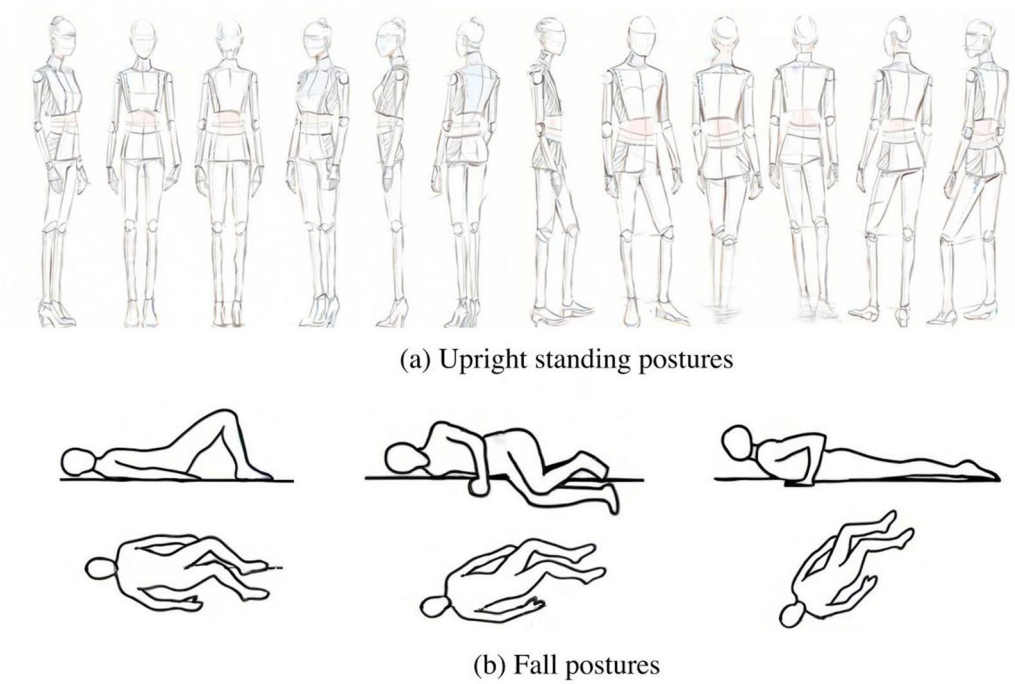

Fig. 3. Representative annotation categories in the extended dataset: (**a**) upright standing postures captured from multiple viewpoints, representing normal actions; (**b**) fall postures, including supine, prone, side-lying, and curled positions, representing fall states. The figure is entirely created by the authors.

Fig. 4. Representative examples of image acquisition conditions in the extended dataset: (**a**) top-down view, (**b**) side view, (**c**) front view, (**d**) low light, (**e**) natural light, (**f**) spotlight, (**g**) partial occlusion, (**h**) no occlusion, and (**i**) background interference. These diverse conditions enhance dataset variability and ensure robustness of fall detection models in realistic public building environments. Photographs taken by the authors.

crowding levels were annotated to reflect background complexity and dynamic obstruction. While 37.2% of the dataset consists of images without additional persons, 41.8% contain sparse crowding (one to two additional persons), and 21.0% exhibit dense crowding (three or more persons). The dense crowding scenario effectively introduces the visual clutter and background interference illustrated in Fig. 4(i). These conditions collectively reproduce realistic challenges encountered in museums, galleries, and community centers, thereby improving dataset representativeness for heritage gallery deployment, with pilot validation conducted at Rochfort Gallery in North Sydney.

Annotation was performed using the LabelImg tool, following strict fall classification standards, and bounding boxes were applied to all target individuals. To enhance reliability, all annotations were conducted by trained professionals and subsequently reviewed by experts in human–computer interaction (HCI) and safety engineering. Specifically, a random subset of 10% of the dataset was independently labeled by two annotators to assess agreement. Discrepancies were resolved through consultation with a senior safety engineer. The inter-rater reliability, measured using the Kappa coefficient, reached 0.89, indicating high annotation consistency. The dataset was divided into training, validation, and testing subsets in a 7:2:1 ratio. Crucially, to prevent data leakage and ensure generalization, the split was performed on a subject-disjoint and scene-disjoint basis. This ensures that images originating from the same video sequence or depicting the same individual do not appear across different subsets.

To improve robustness and mitigate overfitting, five types of data augmentation were applied to simulate real-world challenges: (i) Geometric Transformations: random horizontal/vertical flips, $\pm 90°$ rotations, and constrained cropping; (ii) Noise Simulation: Gaussian noise, salt-and-pepper noise, and motion blur to mimic surveillance distortions; (iii) Lighting Perturbation: random adjustments to brightness, contrast, and saturation to reflect indoor lighting variation; (iv) Synthetic Occlusion: overlays of silhouettes or exhibit items to simulate partial crowding and obstruction; (v) Perspective Correction and Color Normalization: adjustments of scale and tone to account for tilted or uneven illumination. To promote reproducibility, the extended dataset, corresponding annotations, and labeling files are publicly released, with detailed annotation guidelines provided in the online repository.

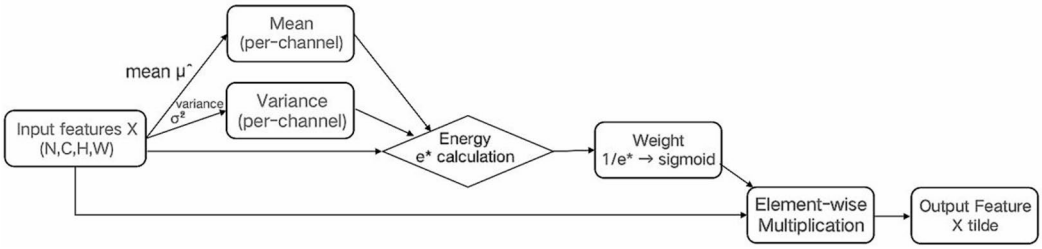

Fig. 5. Structural flow chart of SimAM attention module.Adapted from Yang et al. (2021)[23], with modifications.

Fig. 6. Schematic diagram of the P2 detection head structure. Adapted from Deng et al. (2021)[25], with modifications.

Model comparison and improvement

To validate the effectiveness of the proposed framework, we selected a series of baseline models—YOLOv5n, YOLOv8n, YOLOv10n, and YOLOv11n—which represent the progression of real-time object detection designs. Among these, YOLOv11n was chosen as the backbone for its superior feature extraction capacity and efficiency.

To improve the reliability of fall detection in complex public indoor environments—such as museums, art galleries, and community centers—this study adopts a set of lightweight, scenario-driven design choices within the YOLOv11 framework. These environments are characterized by long viewing distances, frequent partial occlusion, dense backgrounds, and visually cluttered scenes, which have been observed to cause missed detections when standard lightweight detectors are applied without adaptation. In response to these specific challenges, our architectural contribution lies in the tailored integration of two modules to balance sensitivity and efficiency, distinct from generic small-object detection methods that often rely on heavy computational overhead. First, a lightweight SimAM attention mechanism is integrated to enhance the model's sensitivity to fall-relevant spatial cues, such as body deformation patterns and regions of ground contact. Second, a P2 detection head with a stride of 4 is introduced to explicitly capture small-scale or distant human postures that commonly occur in wide-angle indoor surveillance settings. Unlike prior improvements that uniformly increase model depth, these components are selectively employed to strengthen multi-scale perception specifically under cluttered indoor backgrounds. The resulting multi-scale fall detection model, optimized for both accuracy and deployment efficiency, is referred to as YOLO-SEFA (Smart Elderly Fall Alert) in the remainder of this study.

First, a lightweight SimAM (Simple Attention Module) is integrated to enhance sensitivity to fall-relevant spatial cues without introducing additional trainable parameters or computational overhead, as shown in Fig. 5. SimAM is a parameter-free attention mechanism originally proposed in Yang et al.[23]. In our implementation, the SimAM module is inserted after the P5 output of the YOLOv11 backbone, where high-level semantic features are most informative for distinguishing atypical postures and ground-contact regions. Since SimAM is parameter-free, no additional hyperparameters were introduced.

Second, to address missed detections of small-scale or distant fall postures, a P2 detection head (stride = 4)[24] is added to the original YOLOv11 detection hierarchy, as shown in Fig. 6. The P2 branch fuses shallow, high-resolution features from the second stage of the backbone with upsampled features from the P3 head, followed by feature extraction using a C3 module. All other detection head configurations, including channel width, anchor-free design, and loss functions, remain consistent with the original YOLOv11 implementation, ensuring minimal architectural deviation.

Building on the original YOLOv11 framework, this study proposes a multi-scale fall detection network named YOLO-SEFA, as shown in Fig. 7. The model comprises a Backbone, Neck, and four independent detection heads (P2, P3, P4, P5) with strides of 4, 8, 16, and 32, respectively. The input undergoes downsampling through Conv and C3k2 modules, enriched by global semantic information via SPPF and C2PSA modules. The Neck utilizes an FPN + PAN structure to construct cross-scale consistent feature representations, ensuring accurate detection across targets of varying sizes.

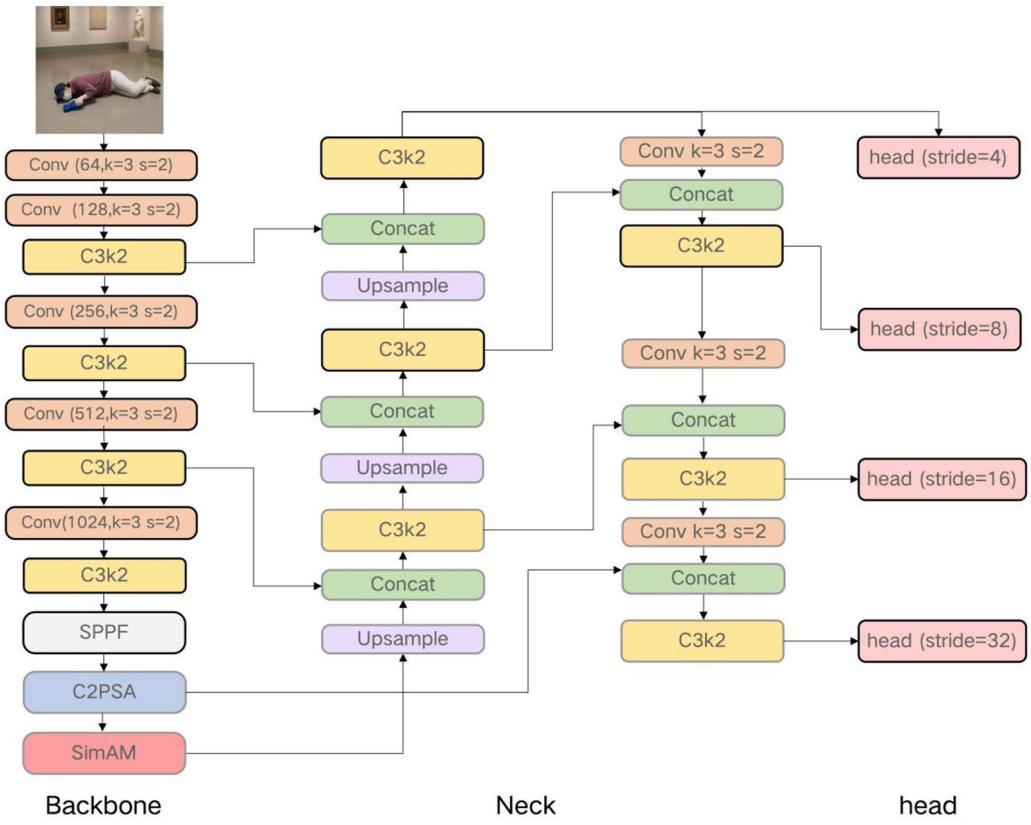

Fig. 7. The framework of YOLO-SEFA. The proposed framework is built upon the YOLO architecture from (Redmon et al., 2016)[17].

Model evaluation metrics

To verify the performance stability and statistical reliability of the YOLOv11-SEFA model, this study conducted 10 independent experimental runs under a consistent train-test split setting, including both the full model and multiple ablation variants. In each run, the network parameters were initialized with a different random seed, and the model's key performance metrics on the test set were recorded, including F1 Score, Precision, Recall, and Mean Average Precision (mAP@0.5).

For each metric, we report the mean (μ) and standard deviation (σ) over 10 runs, along with the 95% confidence interval (95% CI) to assess statistical significance. In addition, we evaluated the model's resource demands by measuring the number of Parameters and Floating-Point Operations (GFLOPs), which reflect deployment cost and edge device adaptability.

The mathematical definitions of the evaluation metrics are as follows:

(i) Precision indicates the proportion of actual falls among the results detected by the model as "falls", as shown in Eq. (1).

$$precision = \frac{True\ Positives\,(TP)}{True\ Positives\,(TP) + False\ Positives\,(FP)} \tag{1}$$

(ii) Recall measures the proportion of all real fall samples that the model successfully identified, as shown in Eq. (2).

$$Recall = \frac{True\ Positives\,(TP)}{True\ Positives\,(TP) + False\ Negatives\,(FN)} \tag{2}$$

(iii) F1-Score is the harmonic average of precision and recall, considering both detection completeness and accuracy, as shown in Eq. (3).

$$F1\ Score = 2 \cdot \frac{Precision \times\ Recall}{Precision + Recall} \tag{3}$$

(iv) Object detection metrics (bounding box evaluation). Since the model performs both classification and object localization, the accuracy of the bounding box must be evaluated using the following metric: Mean Average Precision at Intersection over Union (*IoU*) 0.5 (mAP@50). This metric evaluates the model's object detection performance by measuring the average precision at which the predicted bounding box is at least 50%. A higher mAP@50 score indicates a higher accuracy in identifying and locating fall, as shown in Eq. (4):

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{4}$$

In this study, we adhere to mAP@0.5 as the primary evaluation metric rather than the stricter mAP@[0.5:0.95] for two specific reasons driven by the application scenario. First, the primary objective of safety monitoring is "event recall"—ensuring that every fall incident is detected to trigger an alert. An Intersection over Union (IoU) of 0.5 is sufficient to confirm the correct localization of a fall event for emergency response purposes, whereas higher thresholds (e.g., IoU > 0.75) prioritize pixel-level bounding box alignment, which yields diminishing returns for practical safety operations. Second, in complex surveillance environments characterized by occlusion, wide-angle distortion, and low-resolution inputs, achieving high-precision IoU is often constrained by annotation ambiguity. Adopting mAP@0.5 ensures that the model evaluation focuses on the robustness of behavior recognition rather than sensitivity to minor localization variances.

(v) The mean value ($\bar{x}$) represents the average performance of multiple experimental results, as shown in Eq. (5):

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i \tag{5}$$

(vi) The standard deviation (s) measures the degree of fluctuation of experimental results between multiple runs. The smaller the fluctuation, the more stable the model, as shown in Eq. (6):

$$s = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2} \tag{6}$$

(vii) A 95% confidence interval means that there is a 95% confidence level that the true value falls within the interval, $n = 10$, as shown in Eq. (7):

$$CI = \bar{x} \pm 1.96\frac{s}{\sqrt{10}} \tag{7}$$

The definitions of terms are as follows: TP (True Positive): the number of samples correctly identified as falls by the model. FP (False Positive): the number of normal samples incorrectly identified as falls by the model. FN (False Negative): the number of actual fall samples not identified by the model. TN (True Negative): the number of samples correctly identified as normal by the model (this indicator was not counted in this study because the focus was on detection and recognition effects).

Safety level prediction and SHAP explain ability analysis

To enable graded alarming, we established a quantitative risk assessment model. Unlike binary fall detection, this module evaluates the potential severity of the incident based on the spatial and environmental context of the detected fall.

Six quantitative features were extracted from the YOLOv11 detections to characterize the fall event. To address the need for operational clarity, these features are defined as follows:

(i) Pose Area Ratio (F_1): Defined as the ratio of the subject's bounding box area (A_{bbox}) to the total image frame area (A_{img}), calculated as $F_1 = A_{bbox}?A_{img}$. This feature explicitly quantifies the visual dominance of the subject, serving as a primary indicator of occlusion risk and proximity.

(ii) Tilt Angle (F_2): Calculated as the absolute sine of the angle θ between the subject's major axis (derived from the best-fit ellipse) and the vertical axis (y-axis). Mathematically, $F_2 = |\sin\theta|$. Standing postures yield $F_2 \approx 0$, while fall postures (horizontal alignment) yield $F_2 = 1$, providing a direct geometric metric for abnormal orientation.

(iii) Distance Proxy (F_3): An inverse-depth estimator derived from the bounding box height (hbbox), calculated as $F_3 = 1 - (h_{bbox}/h_{img})$. Note: While F_3 is statistically correlated with F_1 (Pose Area Ratio), it was retained in the feature set because it models depth linearly along the y-axis perspective, whereas F_1 follows a quadratic relationship with distance. The Random Forest model utilizes both to resolve depth ambiguities.

(iv) Aspect Ratio (F_4): Defined as the width-to-height ratio of the bounding box: $F_4 = w_{bbox}/h_{bbox}$. To handle extreme deformations, values are clipped to a normalized range. Upright postures typically exhibit $F_4 \in [0.4,0.6]$, whereas falls result in $F_4 > 1.0$ (flattening) or irregular shapes.
(v) Scene Complexity (F_5): Operationally defined as the Edge Density within the target's context region. It is computed by applying a Canny edge detector to the expanded bounding box and calculating the ratio of edge pixels to total pixels. $F_5 \approx 0$ indicates a clean background, while $F_5 > 0.7$ quantifies high-frequency visual clutter (e.g., texture-heavy artworks or glass reflections).
(vi) Crowd Presence (F_6):A discrete count of other person-class objects (N_p) detected in the same frame, normalized by a maximum crowd threshold (N_{max}=*10*). Calculated as $F_6 = \min(N_p/N_{max}, 1.0)$. This feature explicitly incorporates environmental density into the risk model.

To construct a reliable supervision signal for the Random Forest classifier, we formulated a quantitative risk scoring function. First, the six operationalized features (F_1toF_6) defined in earlier, are normalized to the unit interval [0,1] using Min-Max scaling to ensure dimensional consistency. The comprehensive Risk Score (S) is then modeled as a weighted linear combination of these feature vectors.

Formally, for a given detection sample *x*, the continuous risk score *S(x)* is calculated as shown in Eq. (8):

$$S = \sum_{i=1}^{6} w_i \cdot F_i, \; subject\ to \sum_{i=1}^{6} w_i = 1 \tag{8}$$

Where F_i represents the normalized value of the *i*-th feature, and W_i represents its corresponding importance weight.

To address the challenge of defining "severity" in unlabelled heritage environments, the weight vector $[W = w_1, w_2, w_3, w_4, w_5, w_6, w_7]$ was not assigned arbitrarily but determined through a Delphi method consultation involving five experts in safety engineering and intelligent surveillance. The consensus weights prioritize Tilt Angle (w_2) and Aspect Ratio (w_4) as primary indicators of posture abnormality, while Crowd Presence (w_6) and Scene Complexity (w_5) serve as environmental context modifiers.

The specific instantiation of the scoring function is expressed in Eq. (9):

$$S = 0.15F_1 + 0.25F_2 + 0.10F_3 + 0.20F_4 + 0.20F_5 + 0.10F_6 \tag{9}$$

Here, the variables correspond to: F_1 (Pose Area Ratio), F_2 (Tilt Angle), F_3 (Distance Proxy), F_4 (Aspect Ratio), F_5 (Scene Complexity), and F_6 (Crowd Presence).

To verify that the generated labels are not sensitive to minor fluctuations in these expert-defined weights, a sensitivity analysis was performed. We perturbed each weight w_i by ± 10% and re-calculated the risk levels. The analysis showed a label stability of 96.4%, demonstrating that the scoring rule provides a robust ground truth for training the Random Forest model.

The final continuous score *S* is then discretized into four safety levels (Level 0–3) using the equal-width binning strategy detailed in Table 1. Compared with quantile binning, equal-width binning prioritizes the semantic consistency of the value range, ensuring that the risk levels remain proportional to the linear growth of the expert indicators.

To further validate the reliability of this scoring framework, a preliminary expert consensus experiment was conducted. Specifically, 30 representative fall event images were randomly selected from the dataset, and five senior experts independently scored each image. Inter-rater agreement was assessed using Fleiss' Kappa, yielding a result of $\kappa = 0.79$ ($p < 0.001$), and the correlation between expert ratings and model-generated scores reached $\rho = 0.72$ ($p < 0.001$). These results provide empirical support for the validity of the rating framework as a supervision source.

While the linear scoring rule defines the risk standards, a Random Forest (RF) classifier was employed as the inference engine. The rationale for training an RF model rather than using the raw linear equation for deployment is twofold: (1) Non-linear Robustness: Detection inputs from YOLO in real-world scenarios contain noise; RF ensembles can model complex decision boundaries that smooth out feature jitter more effectively than a rigid linear threshold; (2) Interpretability: RF enables the use of SHAP to provide global and local feature attribution, transforming the risk score into actionable safety insights.

To prevent data leakage and ensure rigorous evaluation, we employed a subject-disjoint split strategy. The dataset was divided into training (70%), validation (10%), and testing (20%) sets, ensuring that images of the same individual or specific video sequence did not appear across subsets.

The training was implemented using the TreeBagger algorithm in MATLAB. To optimize generalization, a 5-fold cross-validation was performed strictly within the training set to tune hyperparameters via grid search (as shown in Table 2). The final model performance reported in Sect. 3.4 is based exclusively on the held-out test set.

Level	Value range	Explanation
Level 0	$0.38 \leq x_7 \leq 0.50$	Safety: normal activities, natural posture
Level 1	$0.51 \leq x_7 \leq 0.56$	Low risk: minor falls or atypical postures
Level 2	$0.57 \leq x_7 \leq 0.61$	Medium-high risk: posture distortion or occlusion
Level 3	$0.62 \leq x_7 \leq 0.74$	High risk: obvious falls, accompanied by occlusion and crowded environment

Table 1. Classification criteria of safety level.

To ensure the Random Forest classifier functions transparently, we implemented a SHAP (SHapley Additive exPlanations) analysis. It is important to clarify that since the ground truth risk labels were generated via the weighted scoring rule (Eq. 10), the primary role of SHAP in this study is not to discover new causal relationships, but to serve as a verification mechanism. It audits whether the machine learning model has faithfully learned the expert-defined logic rather than relying on spurious correlations (e.g., background noise).

We utilized the TreeExplainer method, which is specifically optimized for tree-based ensemble models. For the multi-class Random Forest output (Levels 0–3), SHAP values were computed for each class separately. Let $f(x)$ be the model prediction probability for a specific risk level; the SHAP value $\varnothing_i$ for feature *i* represents its marginal contribution to the deviation of the prediction from the baseline expectation:

$$f(x) = E[f(x)] + \sum_{i=1}^{M} \varnothing_i \tag{10}$$

This additive property allows us to decompose the model's decision path into quantifiable feature attributions, enabling two levels of analysis: (i) Local Explanation: Using beeswarm plots to visualize how feature magnitudes (e.g., high Tilt Angle) shift prediction probabilities; and (ii) Global Alignment Check: Comparing the mean absolute SHAP values against the predefined expert weights (w_i) to confirm model consistency.

Fall detection and alert system for the elderly in heritage gallery

This section describes the proposed system architecture for fall detection in heritage gallery environments and distinguishes it from the components that were empirically evaluated through field deployment.

To address real-time responsiveness, privacy constraints, and scalability requirements in public indoor spaces such as museums and art galleries, we propose a multi-layer fall detection and alert system architecture, as illustrated in Fig. 8. The system follows a modular four-layer design: perception, edge intelligence, transmission, and cloud service layers. This architecture is intended to support low-latency operation and privacy-aware deployment; however, not all components were exhaustively validated in long-term operational settings.

(i) Perception Layer (Video Acquisition). Multiple 2MP PoE network cameras (25 fps) are deployed in a 4–6 m grid to provide multi-angle coverage of public areas. Video streams are transmitted via RTSP to edge nodes located on the same floor. To reduce computational load and limit privacy exposure, adaptive frame-rate reduction (8–15 fps) and region-of-interest cropping are applied prior to inference.

(ii) Edge Intelligence Layer (EIB). Low-power edge servers equipped with Jetson AGX Orin or Intel Arc A770 GPUs (≤ 60 W) execute the proposed YOLOv11-SEFA model using a GStreamer-based inference pipeline. The system performs object-detection–based fall recognition, rather than full pose estimation. When a fall is detected with confidence exceeding 0.70, a local alert is generated, containing event ID, bounding box coordinates, confidence score, camera ID, and timestamp. For interpretability analysis, Grad-CAM heatmaps are generated locally and transmitted together with alert metadata.

(iii) Transmission and Cloud Layers. Alert messages are transmitted via TLS-encrypted MQTT channels to a centralized server. Raw video streams are not uploaded by default; short historical clips are retrieved only when post-event tracing is required. The cloud layer performs alert deduplication, confidence-weighted fusion across cameras, and scheduling integration. These cloud-side mechanisms are part of the proposed system design and were evaluated at a pilot level.

(iv) Privacy Considerations. The system is designed to support compliance with the Personal Information Protection Law of China[25] and GDPR[1] by limiting data transmission and retaining only compressed heatmaps and minimal metadata. However, no formal legal certification or data protection impact assessment was conducted as part of this study. Heatmaps and metadata may still constitute personal data depending on retention duration and contextual use; therefore, privacy claims in this work should be interpreted as design intentions rather than regulatory guarantees.

To evaluate operational behavior beyond dataset benchmarking, a 72-hour continuous pilot deployment was conducted at Rochfort Gallery (North Sydney), covering both public opening hours and nighttime closed periods. The reported metrics reflect average observed values observations under pilot-scale conditions rather than full operational validation.

Hyperparameter	Description	Value range
Number of Trees (numTrees)	Number of decision trees in the random forest	{10, 50, 100, 150}
Maximum Depth (maxDepth)	Controls the complexity of individual decision trees	{2, 5, 10, 20}
Maximum Features (maxFeatures)	Number of features considered at each split	$\left(\sqrt{d}, \frac{d}{3}\right)$ (where d is the total number of features)
Minimum Samples per Leaf (minSamplesLeaf)	Minimum number of samples required to be at a leaf node	{1, 5, 10, 20, 50, 100}

Table 2. Hyperparameter settings for random forest model in fall risk classification. The above parameter adjustment strategy aims to achieve a balance between capturing fine-grained data regularities and preventing overfitting, thereby improving the robustness and generalization ability of the model.

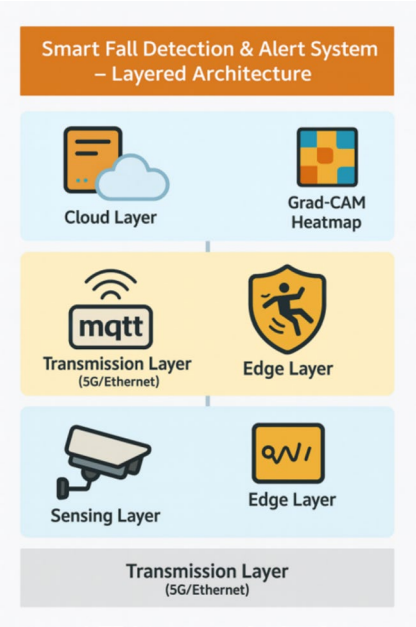

Fig. 8. Layered architecture of the smart fall detection and alert system. The figure is entirely created by the authors.

Metric	Scenario/condition	Performance	Note
False Alarm Rate (FAR)	Peak Hours (Crowded, > 50 visitors/hr)	0.4 alarms/hour	Mainly caused by squatting for photos
	Off-Peak/Nighttime (Low light)	0.05 alarms/hour	Rare shadows or cleaning staff
End-to-End Latency	Sparse Scene (0–2 persons)	265 ms ± 15 ms	Consistent with lab tests
	Dense Scene (Crowded, > 5 persons)	312 ms ± 28 ms	Slight increase due to NMS processing
Frame Rate (FPS)	Edge Node (Jetson AGX Orin)	14.8 FPS (Avg)	Stable under continuous load
Miss Rate (Pilot Verification)	Staged Mock Falls (Closed-hours, controlled)	0 missed/20 staged trials*	Preliminary functional check; sample size not statistically powered (see Limitations)

Table 3. Quantitative field performance metrics of the proposed system during pilot deployment at Rochfort Gallery. *Sample size not statistically powered.

First is False Alarm Behavior. During peak visiting hours (10:00–14:00), the false alarm rate was approximately 0.4 alarms per hour, primarily caused by visitors adopting squatting or kneeling postures for photography. These events were typically classified as low-risk (Level 1) by the post-detection risk grading module, reducing unnecessary escalation. During nighttime low-light conditions, the false alarm rate decreased to 0.05 alarms per hour, with occasional triggers from moving shadows.

Second is Latency under Load. End-to-end latency remained stable across scene densities. In sparse scenes (0–2 persons), average latency was 265 ms (± 15 ms). Under crowd load conditions (> 5 persons per frame), latency increased to 312 ms (± 28 ms), mainly due to additional non-maximum suppression processing. These results remain within commonly accepted response-time expectations for assisted safety monitoring systems.

Third is Miss Rate Verification. A functional check involving 20 staged mock falls was conducted during closed hours. While zero misses were recorded, this serves as a preliminary verification of system availability rather than a statistically powered recall metric.

A summary of field performance metrics is provided in Table 3. Overall, the pilot results provide preliminary evidence of deployment feasibility in heritage gallery environments, while acknowledging that longer-term and larger-scale evaluations are required for comprehensive validation.

Results

To systematically evaluate the performance of the proposed YOLO-SEFA architecture, this section presents both ablation studies and comparative model experiments. In the ablation study, we incrementally integrate the structural enhancement module (P2) and the attention mechanism (SimAM) to quantify their individual and combined contributions to model performance. In the comparative study, we benchmark several mainstream lightweight object detection models to compare key metrics such as detection accuracy, computational complexity, and model size, thereby validating the multidimensional advantages of our proposed model.

Model variation	P2	SimAM	F1 score	*P*(%)	*R*(%)	mAP@50 (%)	GFLOPs	Params (MB)
Baseline			82.74 ± 0.45	87.52 ± 0.52	78.4 ± 0.61	86.67 ± 0.48	6.3	2.58
+ P2 Head	√		82.45 ± 0.31	85.50 ± 0.42	79.60 ± 0.38	87.30 ± 0.33	6.6	2.67
+ SimAM		√	82.46 ± 0.50	84.70 ± 0.47	80.3 ± 0.61	87 ± 0.44	6.5	2.58
YOLO-SEFA	√	√	83.99 ± 0.47	88.50 ± 0.49	80.00 ± 0.63	88.60 ± 0.36	6.6	2.67

Table 4. Experimental results of improvement YOLOv11 model. $n = 10$.

Model	F1 score	*P* (%)	*R* (%)	mAP@50 (%)	GFLOPs	Params (MB)
YOLOv5n	78.89	85.65	73.08	83.27	4.1	1.76
YOLOv8n	82.89	87.1	79	87.3	8.1	3
YOLOv10n	81.04	83.7	78.4	86.1	6.5	2.27
YOLOv11n	83.04	87.8	78.9	87	6.3	2.58
YOLOv11-SEFA	83.99	88.50	80.00	88.60	6.6	2.67

Table 5. Comparison of experimental results of different models.

Ablation study

To investigate the optimization potential of the YOLOv11 baseline in fall detection tasks, we conducted an ablation study focusing on two enhancements: structural reinforcement via the P2 module and semantic enhancement via the SimAM attention module. Table 4 presents the comparative results, reported as mean ± standard deviation (SD) across 10 independent runs to ensure statistical reliability.

As shown in Table 4, the baseline YOLOv11n model established a performance benchmark with an F1 score of 82.74 ± 0.45 and a Recall of 78.40 ± 0.61%. Introducing the individual modules revealed a strategic trade-off between sensitivity and precision. The P2 module alone yielded an F1 score of 82.45 ± 0.31. Although the aggregate F1 score remained statistically comparable to the baseline, the Recall metric improved notably to 79.60 ± 0.38%. This indicates that the P2 layer successfully captured more fine-grained fall cues and reduced missed detections, albeit with a slight increase in false positives (Precision drop). Similarly, incorporating only the SimAM attention module resulted in an F1 score of 82.46 ± 0.50, with Recall further boosting to 80.30 ± 0.61%. This suggests that while individual modules prioritize enhanced sensitivity to abnormal postures—a critical safety requirement—they require joint optimization to recover precision. Consequently, the combined YOLO-SEFA model achieved the best overall performance (F1: 83.99 ± 0.47), effectively harmonizing high Recall (80.00%) with superior Precision (88.50%).

Crucially, the synergistic integration of both modules (YOLO-SEFA) achieved the best overall performance. When both P2 and SimAM were combined, the model yielded an F1 score of 83.99 ± 0.47, achieving the highest Precision of 88.50 ± 0.49% while maintaining a robust Recall of 80.00 ± 0.63%. This represents a measurable improvement over the baseline (approx. +0.95% in F1) while keeping computational requirements within acceptable bounds (6.6 GFLOPs, 2.67 MB). This demonstrates that the structural (P2) and semantic (SimAM) enhancements are highly complementary, effectively correcting the precision drop observed in individual modules. In conclusion, the ablation experiments confirm that while single-module enhancements prioritize recall, their combination is necessary to achieve the optimal stability required for real-world deployment.

Model comparison experiments

To validate the architectural suitability of the proposed method for edge-based fall detection, we conducted a systematic comparison between the YOLOv11-SEFA model and representative lightweight object detectors from the YOLO family. All models were trained and tested under identical experimental conditions with a unified input resolution of 640 × 640. The comparison focuses on identifying the optimal trade-off between detection capability (F1 score, mAP) and deployment efficiency (GFLOPs, Parameters), as detailed in Table 5.

Early iterations, such as YOLOv5n, offer minimal computational cost (4.1 GFLOPs) but demonstrate limited safety reliability, evidenced by a comparatively low Recall of 73.08%. This high miss rate makes YOLOv5n unsuitable for critical safety monitoring where false negatives are unacceptable. Conversely, YOLOv8n achieves a competitive F1 score of 82.89 but at a substantially higher computational cost (8.1 GFLOPs), which challenges the thermal and power constraints of passive edge devices.

Recent architectures, specifically YOLOv10n and YOLOv11n, effectively bridge this gap. YOLOv11n, selected as our baseline, achieves a robust F1 score of 83.04 with a moderate computational load of 6.3 GFLOPs. It provides a more balanced foundation than its predecessors, offering adequate feature extraction capabilities without the overhead of the v8 series.

Building upon the YOLOv11n architecture, the proposed YOLOv11-SEFA integrates P2 structural reinforcement and SimAM attention. As shown in Table 5, this integration yields the highest overall performance across all tested metrics. Specifically, it achieves an F1 score of 83.99 and mAP@50 of 88.60%, surpassing the standard YOLOv11n baseline. More importantly, it secures the highest Precision (88.50%) and Recall (80.00%) among the group, ensuring reliable event detection with fewer false alarms.

Fig. 9. Comparison of fall detection and attention visualization between YOLOv11n and YOLOv11-SEFA in real-world scenarios.

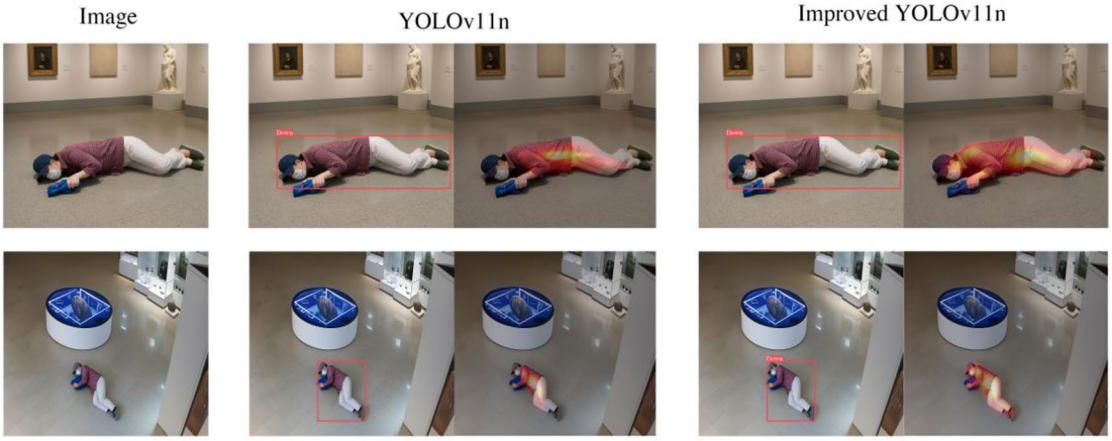

Fig. 10. Comparison of fall detection and attention visualization between YOLOv11n and YOLOv-SEFA in museum indoor environments.

Fig. 11. Comparison of fall detection and attention visualization between YOLOv11n and YOLO-SEFA in occlusion and optical interference scenes.

Crucially, these improvements incur only a marginal increase in computational resources (6.6 GFLOPs vs. 6.3 GFLOPs for the baseline) and parameter size (2.67 MB vs. 2.58 MB). This confirms that the YOLOv11-SEFA configuration offers the most favorable performance-to-efficiency ratio for this study's specific application scenario—heritage gallery monitoring—where both high accuracy and strict resource constraints must be satisfied.

Scenario testing and visualization

To qualitatively examine how the proposed model responds to specific visual challenges commonly encountered in public indoor environments, we present representative inference results in Figs. 9, 10 and 11. These visualizations are intended to complement the quantitative evaluations reported elsewhere and to illustrate model behavior under controlled scenario conditions. Due to privacy protection requirements and intellectual property restrictions associated with exhibited artworks, raw surveillance imagery from the Rochfort Gallery pilot deployment cannot be visually reproduced in this manuscript. Consequently, all visual examples shown in Figs. 9, 10 and 11 are selected from a publicly available, held-out test set that is disjoint from the training data. These samples are used exclusively for qualitative illustration and do not overlap with any data used for training or field-level evaluation.

Figure 9 illustrates representative inference results under low-light conditions. In these examples, the YOLOv11n baseline exhibits false activations on background regions with color and texture similarity to human silhouettes, indicating sensitivity to illumination noise. In contrast, YOLO-SEFA shows more localized activations on human body regions and maintains correct fall detection in these scenarios. This comparison qualitatively reflects the model's ability to suppress background interference under reduced lighting.

Figure 10 focuses on complex indoor environments with reflective materials and heterogeneous lighting, such as glass display cases commonly found in museums. While YOLOv11n detects the fallen target, its Grad-CAM activations are more spatially dispersed and partially influenced by reflective surfaces. YOLO-SEFA, by comparison, exhibits more concentrated attention on the torso and upper-body regions, suggesting improved semantic focus in visually cluttered scenes.

Figure 11 presents examples involving partial occlusion and crowd interference. Under identical confidence thresholds, YOLOv11n fails to detect the fallen individual when significant occlusion is present, whereas YOLO-SEFA successfully identifies the target and allocates attention to key body regions such as the head and shoulders. These examples qualitatively illustrate the model's behavior under occlusion-heavy conditions.

It is important to emphasize that Figs. 9, 10 and 11 are qualitative illustrations rather than exhaustive performance evidence. To address potential concerns regarding cherry-picking, the operational robustness of the system was quantitatively evaluated using field data from the Rochfort Gallery pilot, stratified by illumination, crowd density, and occlusion conditions. These results—including detection rates, false alarm rates, and latency under load—are reported in Table 6 and are computed exclusively from the deployment site.

Together, the qualitative visualizations and the scenario-stratified quantitative metrics provide complementary perspectives on model behavior, with the former illustrating attention patterns and failure modes, and the latter supporting claims regarding deployment feasibility under real-world constraints.

Result of prediction of the safety levels

Model interpretability and logic verification

To validate that the Random Forest model correctly encoded the risk assessment rules defined in Sect. 2.4, we performed a post-hoc attribution analysis using SHAP.

Figure 12 visualizes the directional impact of features on the prediction probability for each risk level. The results confirm that the model's learned decision boundaries align with the intended physical definitions of falls:

In Level 0 (Safety) predictions, The Tilt Angle exhibits the strongest negative SHAP values (blue/purple points pushing to the left) when the angle is low. This indicates that a "vertical posture" is the primary inhibitor of false alarms, actively suppressing risk probabilities.

In Level 3 (High risk) discrimination, Tilt Angle and Pose Area Ratio show strong positive contributions. Notably, high feature values (red points) for these variables significantly push the model output toward the Level 3 class. This attribution pattern is consistent with the definition of a "severe fall" (lying horizontally and close to the camera).

For Level 1 and Level 2, the contributions are more distributed, with Scene Complexity and Crowd Presence playing auxiliary roles. This suggests the model successfully utilizes contextual features to distinguish ambiguous mid-risk states, aligning with the weight distribution assigned in the expert scoring rule.

To further verify the consistency between the trained model and the expert system, we compared the global feature importance (Mean |SHAP|) against the initial design weights. As shown in the stacked bar chart (Fig. 13), Tilt Angle and Pose Area Ratio rank among the top contributors, which directly mirrors their high weights ($W_2 = 0.25$, $W_1 = 0.15$) in the ground truth equation.

While BBox Aspect Ratio shows a higher-than-expected influence in Level 3 detection compared to its linear weight, this likely reflects the non-linear decision capability of the Random Forest. The model effectively identified that "flattened bounding boxes" are a highly distinct visual signature of falls, amplifying this feature's utility beyond its initial linear assignment. Relative Distance and Crowd Presence show moderate contributions, confirming that environmental context is factored into the decision as intended, without overpowering the primary postural cues.

Environmental condition	Scenario definition	Samples/Events*	Detection recall (%)	False alarm rate (per hour)	Avg. latency (ms)	Notes
Illumination	Normal/Spotlight (10:00–16:00)	12 staged falls	98.5	0.40	268 ± 18	False alarms mainly caused by squatting for photography
	Low-light/Night (18:00–06:00)	8 staged falls	96.0	0.05	255 ± 14	Occasional triggers from moving shadows
Crowd Density	Sparse (0–2 per	Continuous monitoring	–	0.18	265 ± 15	Baseline operational load
	Crowded (> 5 persons)	Continuous monitoring	–	0.42	312 ± 28	Latency increase mainly due to NMS overhead
Occlusion Level	None/Minor	10 staged falls	100.0	–	270 ± 16	Clear body contours
	Partial occlusion (exhibit stands, visitors)	10 staged falls	90.0	–	298 ± 22	Missed cases mainly due to head–torso occlusion

Table 6. Quantitative field performance metrics stratified by environmental conditions during pilot deployment at Rochfort Gallery. *Staged falls were performed by trained volunteers during closed hours under supervision, following safety protocols. Continuous monitoring statistics were collected during regular public operation.

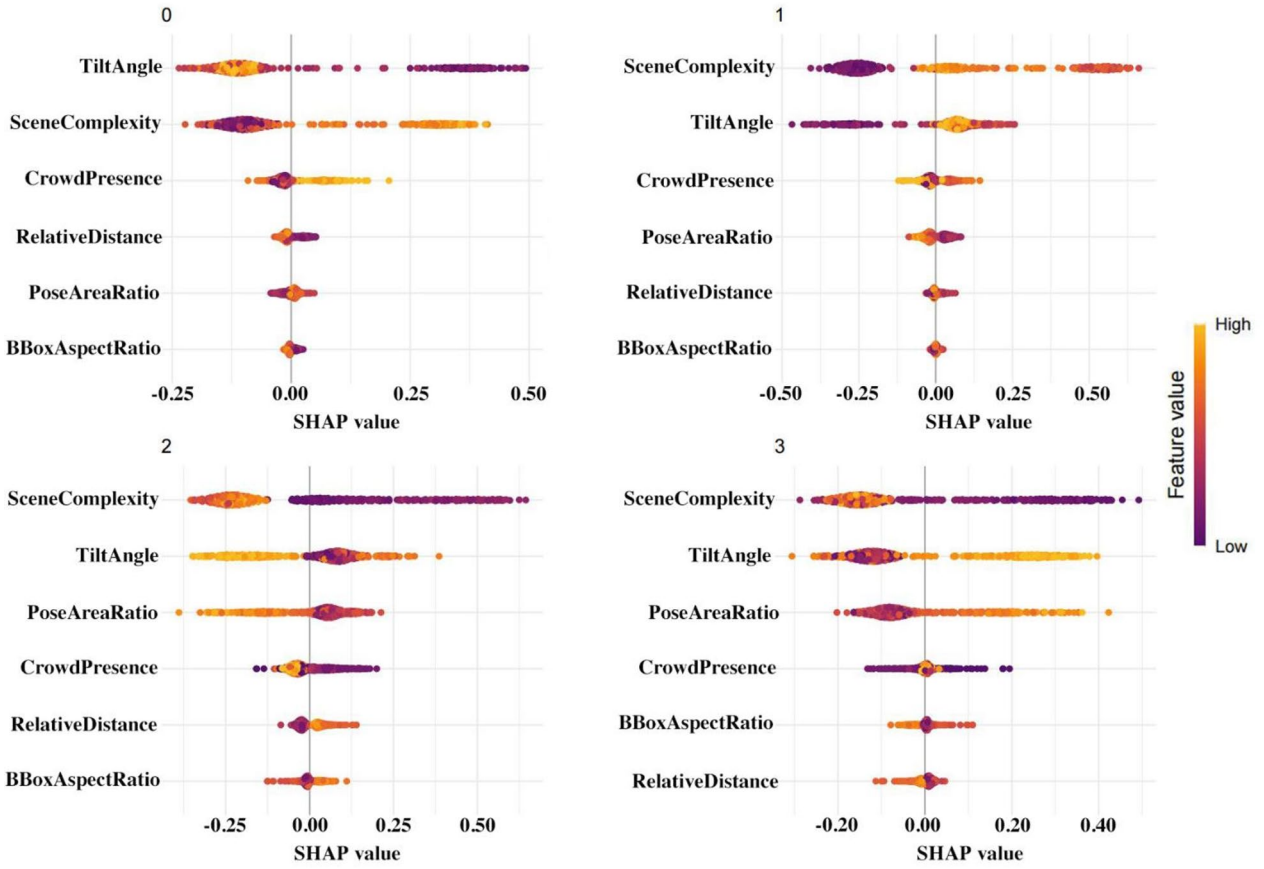

Fig. 12. SHAP beeswarm Plot of feature contributions across fall risk levels (Level 0–3).

In summary, the SHAP analysis serves as a rule verification step. It demonstrates that the YOLOv11-SEFA-RF model has not simply memorized the data but has robustly captured the hierarchical importance of the expert-defined risk factors, providing a transparent basis for its deployment in safety-critical environments.

Generalization and performance stability

While SHAP verifies the model's logical consistency, quantitative evaluation on the test set is crucial to assess deployment readiness. It is important to note that due to the pilot nature of this study, the evaluation relies on a limited hold-out set ($n \approx 400$), which may introduce variance in point estimates. Therefore, the following analysis focuses on structural error patterns rather than absolute precision claims.

Figure 14 presents the confusion matrix on the independent test set. The model exhibits a stable diagonal concentration, particularly at the extremes: Level 0 (Safety) achieved 96.67% accuracy, and Level 3 (High Risk) achieved 89.36%. This high separability is expected, as these classes possess distinct visual signatures (e.g., upright standing vs. horizontal lying) that strongly correlate with the expert-defined risk features.

However, the matrix reveals "boundary ambiguity" in intermediate classes. For Level 2 (Moderate Risk), the recall drops to 73.91%, with 13.49% of samples misclassified as Level 1. Rather than random error, this reflects the inherent semantic overlap in the risk definition—a "moderate" fall often shares visual characteristics with a "mild" fall (Level 1). The model adopts a conservative strategy here: while it misses some Level 2 cases (lower recall), it maintains high precision (100% in this specific split) by only flagging distinct events, which minimizes false escalations in real-world operations.

Figure 15 presents the one-vs-rest precision–recall (PR) curves obtained via 5-fold stratified cross-validation. The curves exhibit a pronounced step-like pattern with relatively few breakpoints, which is an expected artifact of the limited number of positive samples available in each fold.

Specifically, the dataset contains 500 samples evenly distributed across four safety levels (125 samples per class). Under 5-fold stratified splitting, each fold includes 25 positive samples per class, resulting in discretized recall increments and consequently staircase-shaped PR curves.

Despite this sparsity, the AUPRC trends remain informative and consistent across folds. Level 0 and Level 3 achieve high and stable AUPRC values (typically > 0.85), indicating robust discrimination for clearly defined safety and danger states. Level 1 exhibits intermediate performance, with AUPRC values generally above the random baseline and moderate variability across folds, suggesting that early warning states are reasonably distinguishable from the remaining classes, albeit with less confidence than the extreme categories. In contrast, Level 2 shows substantially greater volatility (AUPRC ≈ 0.45–0.60), corroborating the confusion matrix results

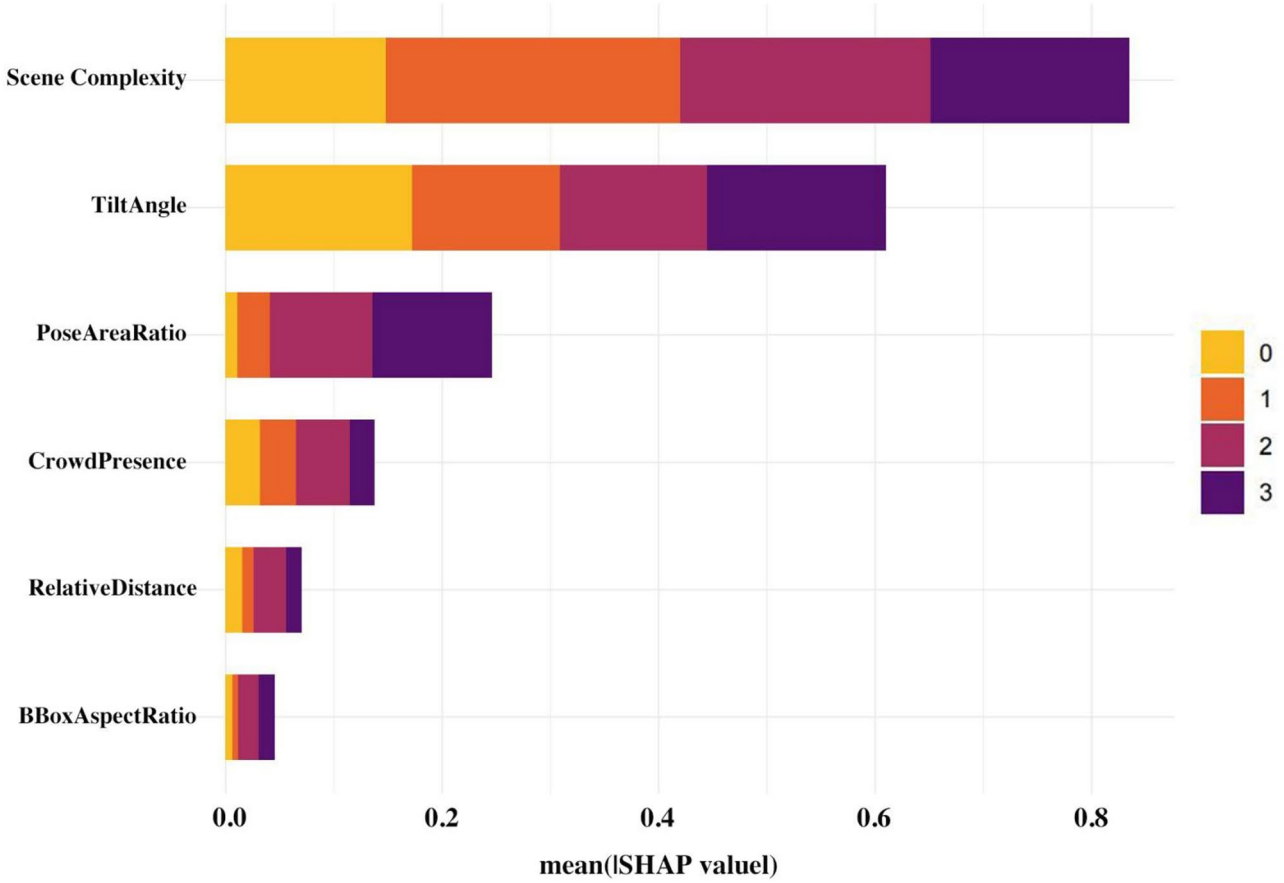

Fig. 13. SHAP global importance ranking, and classification composition bar stacked chart.

and highlighting the inherent difficulty in separating intermediate risk levels that share overlapping visual and contextual characteristics.

Importantly, all classes perform well above the random classifier baseline defined by class prevalence, confirming that the model retains meaningful predictive power even in the most challenging category.

Finally, Fig. 16 presents the ROC-AUC learning curves. Both training and testing AUCs remain consistently high (> 0.98). We clarify that this near-saturation is not indicative of data leakage, but rather a result of the "Rule-Learning" nature of this specific module. Since the ground truth labels are generated deterministically from the input features (via the expert scoring equation), the Random Forest is essentially tasked with approximating a known mathematical function rather than generalizing from noisy, subjective human labels. The high convergence simply confirms that the Random Forest has successfully approximated the expert scoring rule. The slight gap between train and test curves suggests that the model generalizes well to unseen feature combinations without significant overfitting.

Discussion

This study addresses three core challenges currently faced in fall monitoring for the elderly in public space: insufficient real-time detection, limited recognition accuracy, and high difficulty in edge deployment. To this end, we propose a fall detection and early warning system that integrates a lightweight deep learning model with a coordinated "perception–edge–transmission–cloud" architecture. Through multidimensional experiments and visual analysis, we systematically evaluate the system's accuracy, interpretability, and deployability. The results presented in this study are intended to characterize system behavior and feasibility under controlled experimental and pilot deployment conditions, rather than to claim definitive large-scale generalization.

By incorporating the P2 structural enhancement module and the SimAM attention mechanism into the YOLOv11n baseline, the proposed YOLOv11-SEFA model achieves a consistent improvement in F1 score and mAP@50 relative to the YOLOv11n baseline, while maintaining low computational cost (6.6 GFLOPs) and lightweight parameters (2.67 MB). Compared with other lightweight detectors such as YOLOv5n and YOLOv8n, the proposed configuration shows competitive performance under the evaluated settings. These results suggest that targeted architectural adjustments can improve sensitivity to posture variations and detection stability in complex indoor environments, without introducing substantial computational overhead. Such a balance is particularly relevant for edge-based deployment scenarios that require low latency and constrained resources.

Using Grad-CAM visualization, the study reveals how the model's attention mechanism adapts under complex visual conditions (e.g., low light, reflections, occlusion). While YOLOv11n often misactivates non-

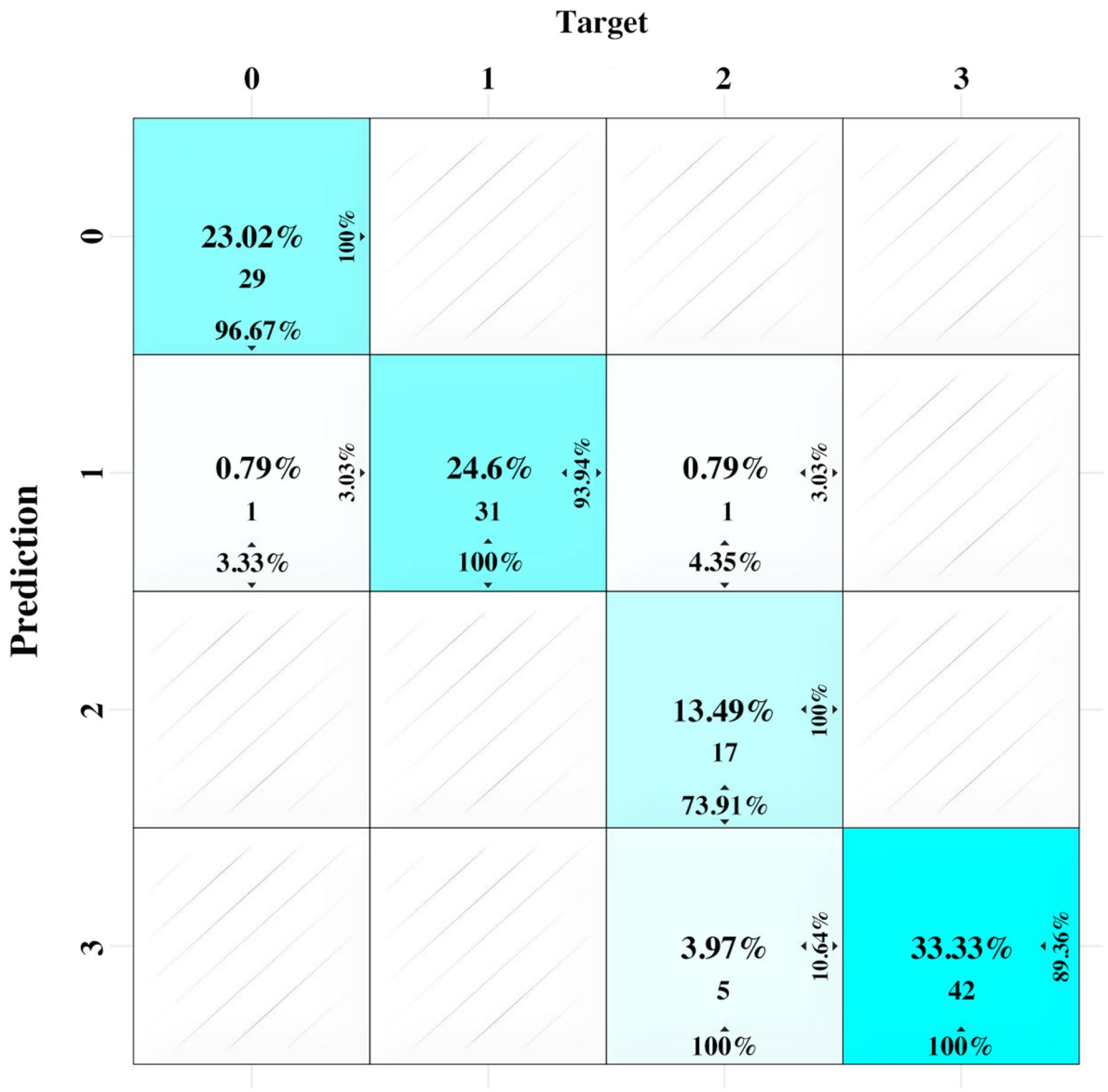

Fig. 14. Confusion matrix of the fall risk level classification on the test set.

human regions in strong lighting or complex backgrounds, YOLOv11-SEFA consistently focuses on key semantic areas such as the head, shoulders, and joints, exhibiting stronger semantic focus and structural perception. These observations indicate that the proposed model tends to allocate attention more consistently to fall-relevant body regions under challenging visual conditions. Such visualization results provide qualitative insight into model behavior… but they should be interpreted as illustrative rather than as exhaustive evidence of robustness across all real-world scenarios.

Following fall detection, the YOLOv11-SEFA model semantically encodes target behaviors using a six-dimensional image structural feature vector, which is then input into a grid search–optimized random forest classifier to determine four levels of fall risk (Levels 0–3). SHAP value analysis indicates that features such as BBoxAspectRatio, RelativeDistance, and CrowdPresence contribute strongly to the model's decision patterns across different risk levels, reflecting posture-related, spatial, and environmental characteristics. PoseAreaRatio and TiltAngle appear to provide additional discriminative cues for intermediate risk levels, while SceneComplexity contributes primarily in lower-risk categories. It should be noted that the current risk-level annotations are derived from the same visual feature space used for prediction, which introduces a degree of label dependency. Accordingly, SHAP-based explanations primarily reflect internal consistency within the constructed risk grading framework rather than independently validated clinical severity, and should not be interpreted as causal explanations of fall risk.

The confusion matrix suggests a structured classification behavior across risk levels, with higher accuracy observed for Level 0 and Level 3 in the evaluated test set. While these results indicate promising separation between low- and high-risk categories, performance on intermediate levels remains more variable. The PR curves

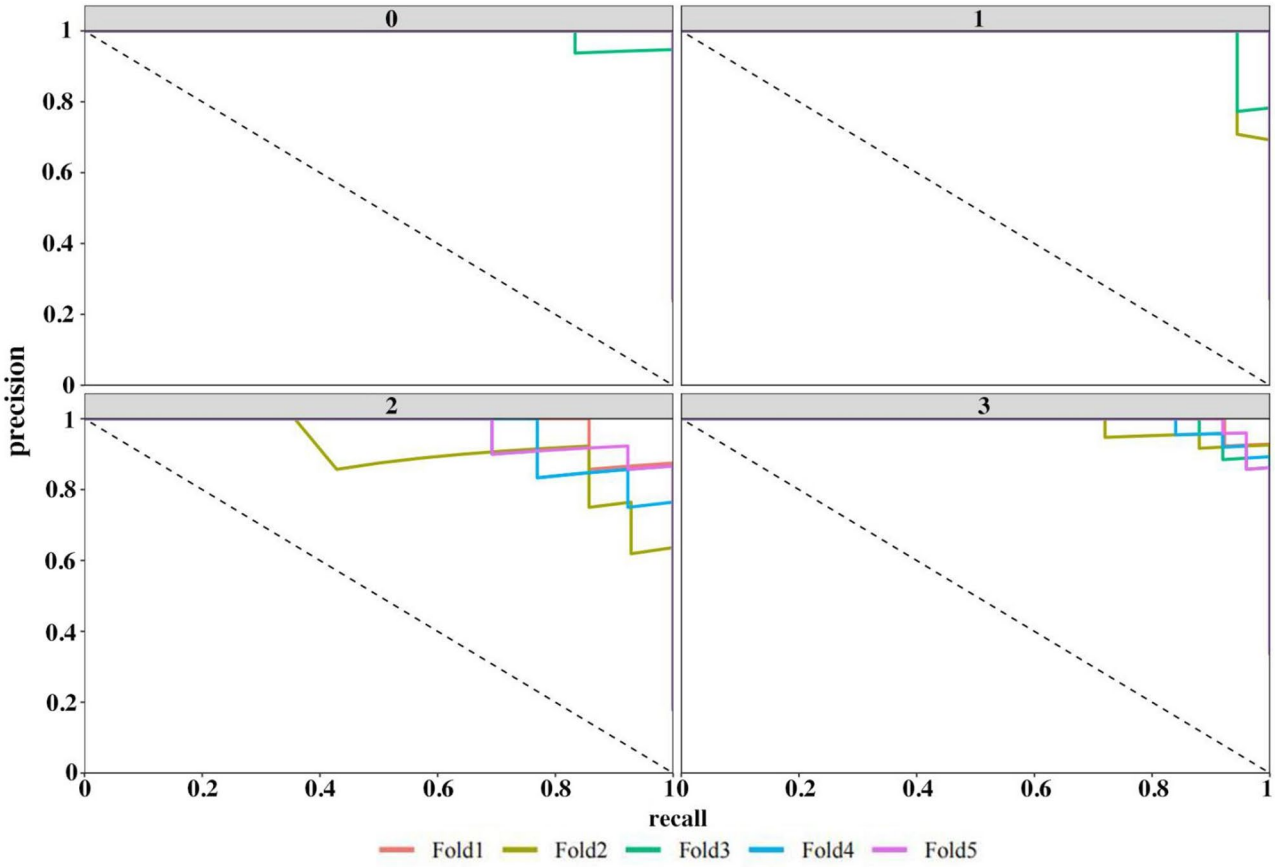

Fig. 15. PR curve comparison across five-fold cross-validation for multi-level fall risk classification. One-vs-rest precision–recall (PR) curves for the four safety levels obtained from 5-fold stratified cross-validation. Each subplot corresponds to one class, and PR curves are computed from predicted class probabilities. Due to the limited number of positive samples per fold (25 positives per class, given 125 samples per class in total), the PR curves exhibit a step-like appearance with discrete recall levels. The horizontal dashed line denotes the random classifier baseline, defined by the positive class prevalence (0.25). Colored curves indicate results from individual folds, and AUPRC trends are consistent with the class-wise confusion matrix analysis.

and cross-validation results provide preliminary evidence of stability under the tested conditions. However, these findings do not yet constitute strong evidence of cross-site or city-scale generalization, and further evaluation on larger, more diverse, and site-disjoint datasets is required to comprehensively assess robustness beyond the studied environments.

Under the experimental setup described in this study, the system exhibited an end-to-end response latency below 270 ms, average edge power consumption of approximately 55 W, and video transmission bandwidth under 1 Mbps. These measurements were obtained via system logging during continuous operation in the specific pilot deployment setting; they represent average observed values under typical load rather than rigorous worst-case performance bounds. The reported metrics suggest the feasibility of integrating the proposed pipeline into existing indoor monitoring infrastructures under controlled deployment conditions. At this stage, the results should be interpreted as indicative of deployment feasibility rather than as a comprehensive validation of large-scale operational performance.

Despite the system's strong performance, several limitations remain. First, limited scene diversity: While the proposed model demonstrates high reliability in heritage gallery environments, direct benchmarking against generic public datasets was limited due to the specific need for multi-level risk annotations and the unique optical challenges of museum settings (e.g., glass reflections and spotlights). Future work will explore cross-domain validation on broader public datasets as they evolve to include complex public scenes. Second, lack of temporal modeling: The current system performs fall inference based on static single-frame images, which limits its ability to capture motion continuity and progressive fall patterns. Although state-of-the-art temporal detectors (e.g., VideoMAE, SlowFast) can achieve higher accuracy by modeling temporal dynamics, this study intentionally prioritizes single-frame efficiency due to both data and deployment constraints. Specifically, the constructed dataset consists of discrete key frames rather than continuous sequences, and temporal models typically incur higher computational cost and latency that are less suitable for low-power edge deployment. Future work will explore the integration of lightweight temporal modules when sequential data and deployment resources become available. Third, limited robustness in extreme environments: The model may still miss falls

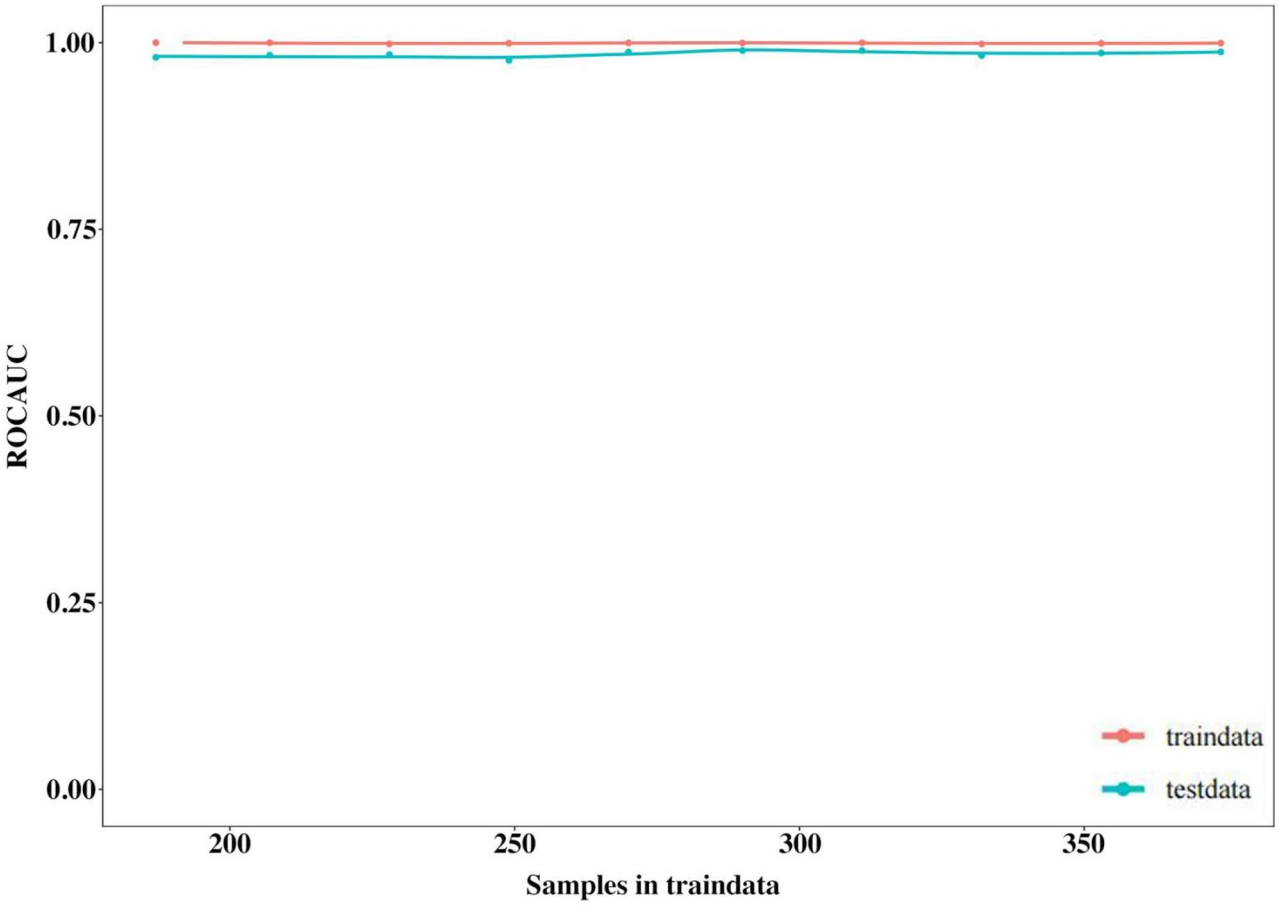

Fig. 16. ROC-AUC comparison between training and testing sets.

in scenarios such as nighttime darkness or sudden occlusions. Introducing multimodal sensors (e.g., infrared, depth, audio) could support the development of a robust fusion system for improved anomaly detection. Fourth, dependency of risk labels: As noted in the discussion, the risk scoring is currently rule-based. The lack of independent clinical validation for risk severity means the current labels reflect algorithm-defined consistency rather than medical ground truth. In particular, the observed variability in intermediate risk-level classification and sensitivity to partial occlusion highlight concrete failure modes that will guide future architectural and data-collection improvements. Finally, regarding the field performance reported in Table 3, it is important to contextualize the reported "zero miss rate." These results were obtained from 20 staged trials conducted under controlled, closed-hour conditions to verify system functionality and end-to-end connectivity. This sample size is not statistically powered to estimate the true recall rate of rare events in the wild. Therefore, the '0/20' result should be interpreted as a preliminary validation of the system's operational readiness, rather than a definitive measure of its statistical robustness against all possible real-world fall variances. These limitations do not undermine the validity of the proposed system in the targeted heritage gallery scenarios, but rather define the current scope of applicability and inform future extensions toward broader and more heterogeneous public environments.

Conclusions

In conclusion, the fall detection and early warning system developed in this study offers a feasible solution for fall monitoring in heritage galleries and similar public indoor spaces. The results demonstrate the practical viability of the proposed approach under controlled experimental settings and pilot deployment conditions, particularly in terms of real-time responsiveness and edge-based operation. Compared to traditional methods that rely on human monitoring and backend analysis, the proposed system leverages edge intelligence for automatic detection and real-time alerts, while balancing computational efficiency, privacy protection, and multi-scenario adaptability. At this stage, the system should be regarded as a validated prototype rather than a fully generalized city-scale solution.

Future work can enhance system performance and expand its applicability across city contexts in three directions, each of which is directly informed by the limitations observed in the current study. Firstly, at the architectural level, with the advancement of Neural Architecture Search (NAS) and edge-aware pruning techniques, fall detection models are expected to evolve toward adaptive light weighting. By incorporating automated architecture search and computation-aware pruning, models can dynamically adjust network depth

and width based on deployment environments, enabling optimal resource allocation and energy efficiency, particularly in response to the observed trade-off between detection accuracy and latency under crowded conditions. Secondly, in terms of behavioral understanding, the integration of temporally aware networks (e.g., ST-Transformer, GNN-TCN) will enhance the ability to parse cross-frame semantic relationships, enabling recognition of progressive falls, sudden behavioral shifts, and persistent instability, which are currently difficult to capture using single-frame inference and were identified as a key source of ambiguity in intermediate risk-level classification. Thirdly, regarding environmental robustness and privacy, future systems should pursue multimodal sensing (infrared, RGB, audio, depth) and privacy-preserving computation (Federated Learning, Differential Privacy). These strategies will improve performance under low-light, heavy occlusion, and noisy backgrounds, addressing failure cases observed during pilot testing, while ensuring user data protection. Together, these directions provide a grounded pathway toward broader urban deployment, subject to further site-disjoint validation, independently labeled risk benchmarks, and systematic evaluation of long-term operational reliability.

Data availability

The data supporting the findings of this study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.29645927.

Received: 6 October 2025; Accepted: 5 February 2026
Published online: 08 February 2026

References

1. European Union. General Data Protection Regulation (GDPR). GDPR. Available online: (2018). https://gdpr-info.eu/ (accessed on 25 July 2025).
2. WHO & Falls May. Available online: (2025). http://www.who.int/en/news-room/fact-sheets/detail/falls (accessed on 6.
3. Montesinos, L., Castaldo, R. & Pecchia, L. Wearable inertial sensors for fall risk as-sessment and prediction in older adults: A systematic review and meta-analysis. *IEEE Trans. Neural Syst. Rehabil Eng.* **26** (3), 573–582. https://doi.org/10.1109/TNSRE.2017.2771383 (2018).
4. Songthap, A., Suphunnakul, P. & Rakprasit, J. Factors affecting home environmental safety management for fall prevention for older adults in Northern Thailand. *BMC Geriatr.* **23**, 704. https://doi.org/10.1186/s12877-023-04419-7 (2023).
5. Sadreazami, H., Bolic, M. & Rajan, S. Contactless fall detection using time-frequency analysis and convolutional neural networks. *IEEE Trans. Ind. Inf.* **17** (10), 6842–6851 (2021).
6. Ajerla, D., Mahfuz, S. & Zulkernine, F. A real-time patient monitoring framework for fall detection. *Wirel. Commun. Mob. Comput.* **2019** (1), 9507938. https://doi.org/10.1155/2019/9507938 (2019).
7. Bourke, A. K., O'Brien, J. V. & Lyons, G. M. Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. *Gait Posture*. **26** (2), 194–199. https://doi.org/10.1016/j.gaitpost.2006.09.012 (2007).
8. Chen, G. C., Huang, C. N., Chiang, C. Y., Hsieh, C. J. & Chan, C. T. June. A reliable fall detection system based on wearable sensor and signal magnitude area for elderly resi-dents. In Proceedings of the Aging Friendly Technology for Health and Independence: 8th In-ternational Conference on Smart Homes and Health Telematics (ICOST 2010), Seoul, Republic of Korea, 22–24 ; Volume 8, pp. 267–270. (2010).
9. He, Y., Li, Y. & Bao, S. D. Fall detection by built-in tri-accelerometer of smartphone. In Proceedings of the 2012 IEEE-EMBS International Conference on Biomedical and Health In-formatics, Hong Kong, China, 5–7 January ; pp. 184–187. (2012).
10. Khan, S. S. & Hoey, J. Review of fall detection techniques: A data availability perspective. *Med. Eng. Phys.* **39**, 12–22 (2017).
11. Hwang, H., Kim, D., Kim, H. & FD-YOLO: A YOLO network optimized for fall de-tection. *Appl. Sci.* **15** (1), 453. https://doi.org/10.3390/app15010453 (2025).
12. Qu, Z., Huang, T., Ji, Y. & Li, Y. *Phys. Sens. Based Deep Learn. Fall Detect. Syst. ArXiv* (2024). arXiv:2403.06994.
13. Song, Z. et al. Fall risk assessment for the elderly based on weak foot features of wearable plantar pressure. *IEEE Trans. Neural Syst. Rehabil Eng.* **30**, 1060–1070 (2022).
14. He, C. et al. A noncontact fall detection method for bedside application with a MEMS infrared sensor and a radar sensor. *IEEE Internet Things J.* **10** (14), 12577–12589 (2023).
15. Hawley-Hague, H., Boulton, E., Hall, A., Pfeiffer, K. & Todd, C. Older adults' per-ceptions of technologies aimed at falls prevention, detection or monitoring: a systematic review. *Int. J. Med. Inf.* **83** (6), 416–426. https://doi.org/10.1016/j.ijmedinf.2014.03.002 (2014).
16. Wang, H., Xu, S., Chen, Y., Su, C. & LFD-YOLO: A lightweight fall detection network with enhanced feature extraction and fusion. *Sci. Rep.* **15** (1), 5069. https://doi.org/10.1038/s41598-025-89214-7 (2025).
17. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June ; pp. 779–788. (2016).
18. Nguyen, H. H. et al. Yolo based real-time human detection for smart video surveillance at the edge. In Proceedings of the 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), Phu Quoc Island, Vietnam, 13–15 January ; pp. 439–444. IEEE. (2021).
19. Wang, X., Kong, L., Zhang, Z., Wang, H. & Lu, X. Keypoint regression strategy and angle loss based YOLO for object detection. *Sci. Rep.* **13**, 20117. https://doi.org/10.1038/s41598-023-47218-2 (2023).
20. Raza, A., Yousaf, M. H. & Velastin, S. A. Human fall detection using YOLO: A re-al-time and AI-on-the-edge perspective. In Proceedings of the 2022 12th International Con-ference on Pattern Recognition Systems (ICPRS), Lille, France, 22–23 June ; pp. 1–6. IEEE. (2022).
21. Zhao, D., Song, T., Gao, J., Li, D. & Niu, Y. YOLO-fall: A novel convolutional neural network model for fall detection in open spaces. *IEEE Access.* **12**, 26137–26149. https://doi.org/10.1109/ACCESS.2024.3383822 (2024).
22. Tianchi Falling posture image dataset. Available online: https://tianchi.aliyun.com/dataset/89276 (accessed on 25 July 2025).
23. Yang, L., Zhang, R-Y., Li, L. & Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. *Proc. Int. Conf. Mach. Learn.* **139**, 11863–11874 (2021).
24. Deng, C., Wang, M., Liu, L., Liu, Y. & Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimedia*. **24**, 1968–1979 (2021).
25. The Personal Information Protection Law of the People's Republic of China (PCPD). Data Privacy Law in Mainland China. Available online: (2021). https://www.pcpd.org.hk/sc_chi/data_privacy_law/mainland_law/mainland_law.html (accessed on 25 July 2025).

Acknowledgements
The authors have no acknowledgments to declare.

Author contributions
Conceptualization, S. W and Y. H; methodology, S. C and X.J; software, H. Y; validation, H. Y, Y.H and S. W; formal analysis, S. W; investigation, S. W and S. C; resources, H. Y and X.J; data curation, H. Y; writing—original draft preparation, Y.H and S. W; writing—review and editing, H. Y and S. C; visualization, H. Y; supervision, H. Y; project administration, S. C. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.