Calibrated VideoLevel Scoring of Skeleton Reconstruction Residuals for Campus Safety Monitoring

Zheqi Lyu, Ke Chen*
Zhejiang Technical Institute of Economics, Hangzhou, China
DOI: 10.65285/HSS.V2I4.005

Abstract: Skeletonbased anomaly detection provides a lightweight approach to humanmotion analysis in campus safety monitoring. Campus safety events are often brief and temporally sparse, posing a challenge for conventional temporal averaging. In this paper, we propose a videolevel residual scoring protocol for skeletonbased anomaly detection. Without modifying the underlying reconstruction architecture, the protocol calibrates local residual responses using normalvalidation statistics and aggregates highresponse temporal segments through uppertail selection. Under a skeletonobservable safetyevent evaluation setting on ShanghaiTech, our method improves the ROCAUC of LSTMAE from 0.789 to 0.840 and that of GRUAE from 0.783 to 0.825 when the uppertail ratio is selected automatically using only normal validation data. Ablation studies show that uppertail aggregation provides the main performance improvement by retaining high-response temporal segments, while regional calibration further improves residual comparability. Supplementary experiments on a curated subset of NWPU Campus show similar gains across the evaluated residualgenerating backbones.
Key words: skeletonbased anomaly detection; campus safety monitoring; intelligent alerting; anomaly analysis; reconstructionbased scoring

1. Introduction

Video anomaly detection (VAD) aims to identify events that deviate from normal patterns in video streams [1-2]. It supports applications such as intelligent surveillance and public safety. Campus environments contain diverse and sometimes crowded activity patterns, making timely screening of potential safety incidents desirable. However, safety-related anomalies occur infrequently, may be brief, and exhibit diverse behavioral patterns. These characteristics make it difficult to collect abnormal samples that adequately cover the target event space. Consequently, a normal-only anomaly detection setting, in which models learn exclusively from normal samples, aligns with practical data constraints in campus safety monitoring.

Existing VAD methods primarily learn normal patterns through reconstruction, prediction, or feature-distribution modeling. Among them, skeleton-based approaches provide a compact representation of human motion. By representing the human body as a sequence of keypoints, skeleton data preserves motion information while reducing the influence of appearance variation, illumination changes, and complex backgrounds [3-4]. This makes skeleton representations suitable for campus surveillance settings in which human motion is the primary observable signal.

Recent skeletonbased anomaly detection methods have explored motion representation learning, temporal relationship modeling, and anomalyscore estimation [3, 5-7]. However, skeleton inputs remain susceptible to poseestimation errors, missing keypoints, and temporal jitter, which can produce unstable anomaly responses [4, 7]. While many existing efforts pursue stronger feature extractors or specialized architectures, videolevel score construction under imperfect pose

observations remains comparatively underexplored. Improving these scores without substantially increasing model complexity is therefore an important practical problem.

Additionally, many existing VAD datasets are designed for general anomaly detection, where anomaly annotations are often defined according to deviations from normal visual patterns, including contextual information, object interactions, or semantic cues. Such annotations may not fully correspond to the alerting requirements of campus safety monitoring when only skeleton inputs are available. Therefore, directly adopting the original labels can introduce a mismatch between evaluation objectives and the observable capabilities of skeleton representations. For example, certain daily activities or athletic behaviors may be annotated as anomalies in general VAD benchmarks but do not necessarily indicate safety risks in campus scenarios. To address this issue, we establish a skeletonobservable safetyevent evaluation setting that better aligns anomaly assessment with campus surveillance requirements [8-9].

In summary, our work makes three main contributions for campus security anomaly detection:

1. We propose a videolevel residual scoring protocol for skeletonbased anomaly detection. The protocol operates on the outputs of existing reconstruction models without modifying or retraining the backbone and without requiring abnormal samples for score calibration.
2. We calibrate residual responses using normalreference statistics and preserve concentrated highresponse evidence through proportional uppertail aggregation.
3. We establish a skeletonobservable safetyevent evaluation setting on the ShanghaiTech and NWPU Campus datasets, providing a more appropriate benchmark for evaluating skeletonbased anomaly detection in campus surveillance scenarios.

2. Related work

2.1 Skeletonbased video anomaly detection

Existing skeletonbased video anomaly detection methods can be broadly categorized into reconstruction and prediction, latent representation learning, probability density estimation, and generative future modeling. MPEDRNN employs a recurrent encoder–decoder to model global body displacement and local pose dynamics [5]. Normal Graph uses a spatiotemporal graph convolutional network to predict future skeleton motion [10], while STGCAELSTM integrates LSTM units into a graph convolutional autoencoder to jointly reconstruct observed skeletons and predict future poses [11]. TrajREC learns temporal motion regularities by reconstructing masked continuous segments at different positions within a trajectory [12].

In representation- and distributionbased approaches, GEPC maps pose graphs into a latent space and clusters the resulting representations into action patterns [13], whereas COSKAD constrains normal skeleton representations within a compact latent hypersphere [6]. STGNF employs normalizing flows to estimate the probability density of spatiotemporal pose graph sequences [3], while SeeKer models skeleton sequences through jointlevel autoregressive factorization and conditional Gaussian distributions [7]. Generative approaches explicitly model multiple plausible motion futures. MoCoDAD uses a conditional diffusion model to generate multiple candidate futures [14], and GiCiSAD combines graph attention with a bodyregion jigsaw task to learn regional dependencies and multimodal future motion patterns [15].

2.2 Anomaly evidence construction and reliability handling

Different modeling paradigms construct anomaly evidence in different forms. Reconstruction- and predictionbased methods quantify anomalies using the discrepancy between model outputs and observed skeleton motion [5, 10-12]. Clustering and latentrepresentation methods measure deviations based on clusterassignment patterns or the geometric relationship between test representations and learned normal representations [6, 13]. Probabilistic methods assess normality using posesequence likelihoods or jointlevel conditional likelihoods [3, 7], whereas diffusionbased methods compare generated motion futures with the corresponding observed future motion [14-15]. Accordingly, anomaly evidence may be defined at different levels, including individual joints, pose sequences, personwindows, and latent representations.

Several studies further incorporate bodyregion structure or pose observation quality into anomaly evidence modeling. GiCiSAD partitions the skeleton into multiple body regions and uses a graphlevel jigsaw task to learn structural dependencies among these regions [15]. SeeKer, by contrast, weights jointlevel conditional negative loglikelihoods using keypoint confidence scores produced by the pose detector, thereby reducing the influence of lowconfidence observations on the skeleton anomaly score [7]. Thus, GiCiSAD models regional structural dependencies, whereas SeeKer explicitly accounts for poseconfidencedependent observation reliability.

2.3 Anomaly score aggregation

Skeletonbased video anomaly detection methods generally need to convert responses from short pose sequences, human trajectories, or individual joints into framelevel anomaly scores. MPEDRNN averages the losses of overlapping trajectory segments and then takes the maximum response among multiple people within the same frame [5]. COSKAD similarly projects overlapping personwindow scores onto their corresponding frames and aggregates responses from multiple people [6]. STGNF converts posesequence likelihoods into framelevel normality scores [3], whereas SeeKer first combines jointlevel conditional likelihoods into a singleperson skeleton score and then aggregates scores across people in multiperson scenes [7]. These operations primarily address temporal overlap and the presence of multiple human instances within the same frame.

Diffusion-based methods additionally need to aggregate multiple candidate futures generated from the same observed history. MoCoDAD and GiCiSAD compare each generated future with the corresponding observed future motion and apply statistical selection or aggregation across the generated samples [14-15]. Overall, existing aggregation mechanisms address different dimensions, including overlapping temporal segments, multiple people, and multiple generated futures.

3. Proposed Method

The proposed method is built upon a skeleton reconstruction model trained solely on normal samples and operates directly on the reconstruction residuals produced by the model. During the scoring phase, the network architecture and parameters of the backbone remain unchanged. As illustrated in Figure 1, the overall pipeline consists of local residual generation, normalreference residual calibration, windowlevel residual scoring, and proportional uppertail aggregation, which together produce the final videolevel anomaly score.

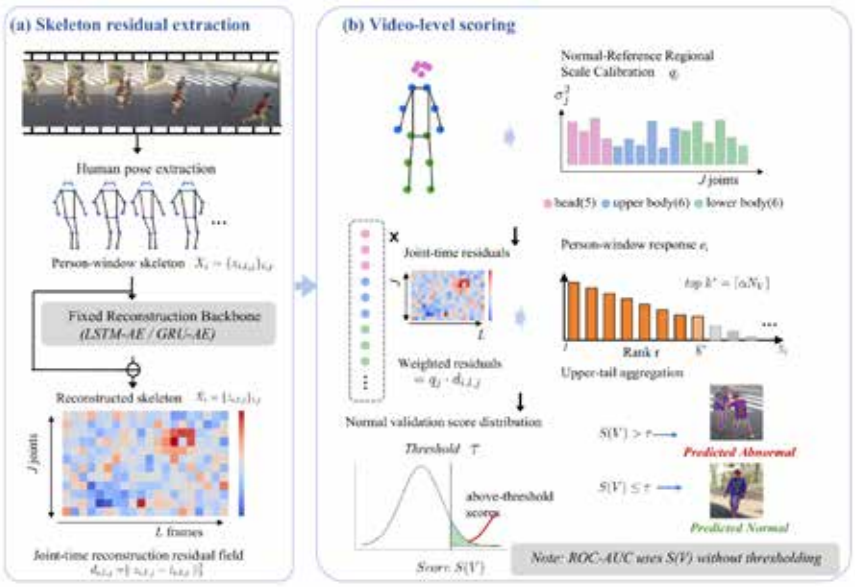

Figure 1. Skeleton residual refinement pipeline. (a) A fixed reconstruction backbone trained on normal data generates joint–time residuals for each person window. (b) Normalreference calibration and uppertail aggregation convert these residuals into a videolevel anomaly score. Thresholding is applied only for binary alarm decisions.

3.1 Residual calibration and personwindow scoring

For a video V, pose estimation and trajectory association produce NV person windows $X_i \in \mathbb{R}^{L\times J\times 2}$, where L is the window length and $J = 17$ follows the COCO keypoint layout. Track fragments containing fewer than 15 valid pose observations are excluded during dataset construction. Fragments containing at least 15 but fewer than L observations are zeropadded to the fixed window length, while their valid observation counts are retained for residual scoring and completeness weighting. Let $z_{i,l,j} \in \mathbb{R}^2$ denote keypoint j at time step l, and let $m_{i,l} \in \{0,1\}$ equal one for a valid pose observation and zero for padding. A fixed reconstruction model f_θ trained only on normal samples produces $X_i = f_\theta(X_i)$. The corresponding joint–time residual is

$$d_{i,l,j}=\left\|z_{i,l,j}-\hat{z}_{i,l,j}\right\|_2^2.$$

Because normal residual scales differ across body regions, calibration statistics are estimated from the set of normalvalidation person windows, denoted by W_{val}^{N}. The COCO keypoints are grouped into the head (nose, eyes, and ears), upper body (shoulders, elbows, and wrists), and lower body (hips, knees, and ankles). Let $b(j)$ denote the region containing joint j and $J_{b(j)}$ its joint set. The joint reference scale, regionshared scale, and resulting calibration weight are

$$\hat{\sigma}_j^2=\frac{\sum_{X_i\in W_{val}^N}\sum_{l=1}^{L} m_{i,l}\, d_{i,l,j}}{\sum_{X_i\in W_{val}^N}\sum_{l=1}^{L} m_{i,l}},$$

$$\tilde{\sigma}_j^2=\frac{1}{|J_{b(j)}|}\sum_{k\in J_{b(j)}}\hat{\sigma}_k^2,$$

$$q_j=\frac{1}{\tilde{\sigma}_j^2+\varepsilon}.$$

Here, $\varepsilon = 10^{-8}$ ensures numerical stability. The regionlevel sharing reduces sensitivity to jointspecific validation noise, while inversescale weighting reduces the contribution of regions with larger normal reconstruction variability. For each trained model instance, the q_j values are estimated once from its normalvalidation residuals and then remain fixed during testing.

The calibrated residual for person window X_i is

$$R_{rel}(X_i)=\frac{\sum_{l=1}^{L}\sum_{j=1}^{J} m_{i,l}\, q_j d_{i,l,j}}{\sum_{l=1}^{L}\sum_{j=1}^{J} m_{i,l}\, q_j+\varepsilon}.$$

To reduce unreliable high responses from short or fragmented trajectories containing many padded positions, the score is weighted by the observationcompleteness ratio:

$$\rho_i=\frac{1}{L}\sum_{l=1}^{L} m_{i,l}, e_i=\rho_i R_{rel}(X_i).$$

3.2 Automatic proportional uppertail aggregation

Safetyrelevant evidence may occur in only a small fraction of a video. Mean pooling can dilute such responses, whereas max pooling is sensitive to isolated errors. We therefore sort the personwindow scores in descending order, $e_{(1)} \geq \cdots \geq e_{(NV)}$, and average a proportion $\alpha \in (0,1]$ of the upper tail:

$$k=\max(1,\lceil \alpha N_V\rceil),$$

$$S_\alpha(V)=\frac{1}{k}\sum_{r=1}^{k} e_{(r)}.$$

This empirical uppertail mean is related to conditional valueatrisk applied to high residual responses [16]. When $\alpha = 1$, Eq. (5) reduces to global mean pooling; smaller values retain a more concentrated subset of highresponse windows.

α is selected independently for each trained backbone and random seed using only the set of normalvalidation videos,

denoted by V_{val}^{N}.. The candidate set is

$$A_{tail}=\{0.01,0.03,0.05,0.10,0.15,0.20\}.$$

For each candidate, we compute

$$CV(\alpha)=\frac{\mathrm{Std}_{V\in V_{val}^{N}}[S_{\alpha}(V)]}{\left|\mathrm{Mean}_{V\in V_{val}^{N}}[S_{\alpha}(V)]\right|+\varepsilon}.$$

The tail fraction is selected directly from these candidatewise values:

$$\alpha^{*}=\max\left\{\alpha\in A_{tail}:CV(\alpha)\leq(1+\delta)\min_{\alpha'\in A_{tail}}CV(\alpha')\right\}.$$

Here, δ = 0.005. Selecting the largest candidate within this CV tolerance avoids excessive dependence on only a few windows. The selected α* is fixed before test evaluation and is not adjusted using test results or labels; the final videolevel score is $S(V) = S_{\alpha}^{*}(V)$.

3.3 Alarm decision threshold

When a binary decision is required, the threshold is estimated only from normalvalidation video scores:

$$\tau=\mu_{val}+\beta\sigma_{val},$$

where μ_{val} and σ_{val} are their mean and standard deviation. We use β = 0.25 for all reported binary decisions and classify V as anomalous when $S_{\alpha}^{*}(V) > \tau$. Neither τ nor β is adjusted using test labels.

4. Experimental Setup

4.1 Safetyevent benchmarks

Experiments were conducted on the ShanghaiTech [8] and NWPU Campus [9] datasets. Both benchmarks have been widely adopted in general video anomaly detection (VAD) research. The ShanghaiTech dataset contains 437 video sequences collected from 13 surveillance scenes. It covers diverse environments, including indoor corridors, campus roads and plazas, with variations in camera viewpoints, object scales and crowd densities. Introduced by Cao et al., the NWPU Campus dataset consists of 43 campus scenes, 28 anomaly categories and 16 hours of surveillance videos. Some anomaly categories involve contextual and semantic cues beyond human motion patterns. For example, behaviors such as riding a bicycle across a lawn are annotated as anomalies in the original dataset; however, their anomalous nature is mainly determined by spatial scene constraints rather than observable human motion dynamics.

Unlike general VAD that targets broad behavioral deviations, this work focuses on skeletonbased campus safety monitoring, where anomaly decisions are made from human motion observations. Since the original anomaly annotations were designed for general VAD and may involve visual context beyond human motion, we establish a skeletonobservable safetyevent evaluation setting aligned with campus surveillance objectives. The setting follows two principles:

1.Evaluation is restricted to video sequences from which usable human pose trajectories can be extracted under the predefined pre-processing rules.

2.Event categories are determined according to campus safety relevance. Routine or safetyneutral human activities are regarded as normal samples, while only behaviors with clear safety implications are considered abnormal test samples.

According to these principles, routine human activities, such as walking, running, cycling, jumping, and noncontact safetyneutral gestures, are categorized as normal samples when no additional safetyrelated context is involved. In contrast, safetyrelevant behaviors, including theftrelated actions, fence climbing, fallrelated events, and violent physical conflicts, are included in the abnormal evaluation set. Figure 2 provides representative visual examples of normal activities and safetyrelated anomaly events under this evaluation setting.

Figure 2. Representative ShanghaiTech activities after safetyoriented relabeling. (a) Routine walking and (b) cycling are labeled normal, whereas (c) falling is labeled abnormal.

4.2 Experimental details

Table 1 summarizes the dataset split details for ShanghaiTech and NWPU Campus. For ShanghaiTech, only two sequences were excluded due to unsuccessful skeleton extraction, and no additional manual filtering was performed based on pose confidence. No additional manual videolevel filtering was performed based on pose confidence or anomaly scores. For NWPU Campus, considering its larger scale and longer video duration, the selected samples span 30 of the dataset's 43 scene identifiers. Fixedlength temporal segments, with a duration approximately corresponding to the average duration of safetyrelated abnormal events, were sampled to control the evaluation scale. Segment sampling did not rely on anomaly scores or individual sequencelevel decisions, thereby limiting scoredependent selection bias.

Table 1. Numbers of retained normal and abnormal samples in the ShanghaiTech and NWPU Campus splits.

Dataset	Group	Normal	Abnormal
NWPU Campus	Train	246	0
	Validation	82	0
	Test	83	24
	Total retained	411	24
	Train	64	0
	Validation	22	0
	Test	22	20
	Total retained	108	20

We build a video skeleton analysis pipeline using YOLOv8mpose (Ultralytics 8.4.41) [17] for human keypoint extraction and ByteTrack [18] for crossframe trajectory association. This combination produces the pose trajectories used in the subsequent skeletonbased analysis. Each human target is represented by 17 2D keypoints following the COCO skeleton format [19]. After coordinate normalization, the extracted skeleton sequences are segmented into 60frame clips with 50% overlap using a slidingwindow strategy. Normal sequences are divided into training, validation, and test sets at an approximate ratio of 6:2:2. The training set is used exclusively to train the normal skeleton reconstruction model, while the validation set is used to estimate normal reference statistics and determine alert thresholds. The test set is reserved for final evaluation under the proposed skeletonobservable safetyevent evaluation setting.

Experiments on ShanghaiTech use two reconstruction backbones. The LSTM-AE [20] and GRU-AE [21] use matched single-layer bidirectional encoders and single-layer decoders with 256 hidden units. Two models are trained for 100 epochs using AdamW [22], batch size 64 and weight decay 10−4. Each experiment is repeated with model seeds 42, 43, and 44 on the same fixed split, and all scoring variants reuse the same checkpoint within each run. Performance is

evaluated using video-level ROC-AUC, which measures score-ranking performance across decision thresholds, and macro-averaged F1 for the binary decisions obtained at the fixed normal-validation threshold. False-positive (FP) and false-negative (FN) counts are also reported at this operating point.

5. Results

5.1 Main ShanghaiTech results

Table 2 presents the performance comparison across two backbones under the ShanghaiTech evaluation setup. Relative to the raw temporal averaging baseline, our residual calibration and automatically selected uppertail aggregation better preserve highresponse temporal segments, achieving consistent gains across both backbones. Specifically, ROCAUC increases from 0.789 to 0.840 for LSTMAE and from 0.783 to 0.825 for GRUAE.

Table 2. Videolevel performance of raw and refined residual scoring on ShanghaiTech.

Backbone	Method	ROCAUC ↑	MacroF1 ↑	FP ↓	FN ↓
LSTMAE	Raw	0.789 ± 0.005	0.674 ± 0.018	25.3 ± 0.9	**5.0** ± 0.8
	Ours	**0.840** ± 0.002	**0.697** ± 0.018	**20.3** ± 1.7	6.3 ± 0.5
GRUAE	Raw	0.783 ± 0.005	0.663 ± 0.014	26.7 ± 0.5	**5.0** ± 0.8
	Ours	**0.825** ± 0.010	**0.672** ± 0.010	**23.3** ± 0.5	6.3 ± 0.5

The gains observed for both recurrent backbones are consistent with a scoringstage contribution, because all scoring variants reuse the same trained checkpoint within each run. Under the same validationbased thresholdcalibration strategy, the proposed scoring protocol also reduces false positives. In particular, the FP count decreases from 25.3 to 20.3 for LSTMAE (a 19.7% reduction) and from 26.7 to 23.3 for GRUAE (a 12.5% reduction).

5.2 Ablation study and component analysis

Table 3 presents a componentwise ablation study analyzing the contributions of reliability calibration and uppertail aggregation on videolevel residual scoring. All variants use identical skeleton inputs and model outputs, ensuring a consistent evaluation condition across different scoring configurations.

Table 3. ROCAUC ablation of residual calibration and temporal aggregation on ShanghaiTech.

Variant / Strategy	GRUAE	LSTMAE
Component Contribution		
Raw residual baseline	0.783 ± 0.005	0.789 ± 0.005
With region calibration	0.791 ± 0.007	0.794 ± 0.002
With tail summarization	0.820 ± 0.005	0.828 ± 0.006
Ours	**0.825** ± 0.010	**0.840** ± 0.002
Calibration Design		
Perjoint calibration	0.774 ± 0.009	0.777 ± 0.002
Regionshared calibration	0.791 ± 0.007	0.794 ± 0.002
Randomized scale factors (+ summarization)	0.813 ± 0.008	0.823 ± 0.011
Ours	**0.825** ± 0.010	**0.840** ± 0.002
Aggregation Strategy with Calibration Fixed		
Mean pooling	0.791 ± 0.007	0.794 ± 0.002
Fixed top5 pooling	0.779 ± 0.009	0.780 ± 0.005
Softmax pooling (Tsm = 1)	0.827 ± 0.007	0.829 ± 0.006
GeM pooling (p = 2)	0.817 ± 0.007	0.819 ± 0.002
GeM pooling (p = 4)	0.832 ± 0.008	0.832 ± 0.006
Max pooling	0.820 ± 0.010	0.822 ± 0.004
Ours (auto α)	**0.825** ± 0.010	**0.840** ± 0.002

Experimental results indicate that regional residual calibration and uppertail aggregation address different aspects of anomaly scoring: residual comparability and videolevel evidence aggregation, respectively. Compared with raw temporal averaging, automatically selected uppertail aggregation provides the primary gain, increasing the ROCAUC of GRUAE from 0.783 to 0.820 and that of LSTMAE from 0.789 to 0.828. Building upon this improvement, regional residual calibration increases the ROCAUC to 0.825 and 0.840, respectively, demonstrating that their combination provides a further correction under the same automaticselection protocol.

We then examine how calibration granularity affects scoring stability. Directly applying independent perjoint calibration performs worse than the raw baseline on both backbone networks, achieving ROCAUC values of 0.774 for GRUAE and 0.777 for LSTMAE. This degradation may reflect the limited validation sample size, because independently estimating scales for all 17 keypoints makes the weighting scheme sensitive to keypointlevel noise. In contrast, regionshared calibration groups keypoints into coarse body regions, such as the head, upper body, and lower body, and estimates shared scales at the regional level. This reduces the number of independently estimated scale parameters while preserving differences across body regions. Substituting statistical calibration with random scale factors yields lower ROCAUC values of 0.813 and 0.823. This result is consistent with normalsamplebased scale estimation contributing more useful weighting information than arbitrary scale adjustments.

The aggregation analysis further investigates proportional selection of highresponse windows for shortduration anomaly events. Fixed top5 pooling performs poorly (with ROCAUC values of 0.779 and 0.780), because retaining the same number of windows does not adapt to variation in sequence length and extractedwindow count across clips. Unlike fixednumber pooling, proportional uppertail aggregation retains an automatically selected fraction of the available windows. This design is intended to preserve informative highresponse evidence without relying on a single maximum response. The automatic uppertail rule achieves the highest ROCAUC on LSTMAE. On GRUAE, it remains above

mean, fixed top5, GeM (p = 2), and max pooling, but is slightly below softmax pooling and GeM with p = 4.

5.3 Sensitivity and robustness analysis

We assess scoring robustness across two dimensions: temporal window length and pose observation perturbations. All experiments in this section use the LSTMAE backbone, and Table 4 summarizes the results under different experimental conditions.

Table 4. ROCAUC sensitivity to temporal segmentation and pose perturbations on ShanghaiTech using LSTMAE.

Evaluation Configuration	Raw Baseline	Ours
Temporal Window & Stride Sensitivity		
L = 20 / stride 10	0.760 ± 0.005	**0.817** ± 0.008
L = 40 / stride 20	0.774 ± 0.005	**0.813** ± 0.018
L = 60 / stride 30 (Default)	0.789 ± 0.005	**0.840** ± 0.002
L = 80 / stride 40	0.807 ± 0.005	**0.831** ± 0.012
Resilience to InferenceTime Perturbations		
Clean (No perturbation)	0.789 ± 0.005	**0.840** ± 0.002
Gaussian noise (σ = 0.04)	0.793 ± 0.005	**0.839** ± 0.003
Joint mask (10%)	0.580 ± 0.016	**0.636** ± 0.025
Joint mask (20%)	0.565 ± 0.047	**0.616** ± 0.010
Joint mask (30%)	0.551 ± 0.036	**0.593** ± 0.051

Across window lengths ranging from 20 to 80 frames, our scoring protocol outperforms the raw residual mean baseline under all evaluated settings. In particular, the default 60frame window (stride 30) reaches the highest ROCAUC of 0.840. Even with a short 20frame window, ROCAUC reaches 0.817 (vs. 0.760 for Raw), showing that the scoring strategy remains above the raw baseline under a shorter temporal span. The observed pattern is consistent with short windows providing limited temporal context and longer windows including more normal motion that can dilute concentrated responses.

To evaluate the impact of pose observation perturbations on scoring stability, we simulate two common types of skeleton input errors: Gaussian coordinate noise and random keypoint missingness. After adding Gaussian coordinate noise with a standard deviation of σ = 0.04, the ROCAUC remains nearly unchanged (0.839 vs. 0.840), showing that this perturbation has limited influence under the evaluated setting. By contrast, keypoint missingness degrades performance for both the raw baseline and the scoring approach, indicating that incomplete pose observations have a stronger impact than the tested coordinate noise. Across the evaluated keypointmissing ratios, the method remains above the raw scoring baseline. For instance, at a 20% keypoint drop rate, the ROCAUC is 0.616 compared with 0.565 for the raw baseline, although both methods degrade substantially under incomplete pose inputs.

5.4 Supplementary evaluation across residualgenerator types

As a supplementary evaluation on a second dataset and additional residualgenerator types, we conduct experiments on a skeletonobservable subset of NWPU Campus as shown in Table 5. Consistent with the ShanghaiTech safetyevent evaluation setting, data partitioning is performed according to skeleton input availability and campus safety semantics. In addition to LSTMAE and GRUAE, we include two simplified inhouse residual generators. These are a sequencebottleneck recurrent autoencoder and a skeleton graphconvolutional autoencoder inspired by spatiotemporal skeleton graph convolution [23]. All listed residual generators provide reconstruction responses to the same scoring, aggregation, and evaluation pipeline. All four residualgenerator types show improved ROCAUC on the curated NWPU Campus subset. These findings provide supplementary evidence that the scoring procedure can be applied to several

reconstruction architectures.

Table 5. ROCAUC comparison across residualgenerator types on the curated NWPU Campus subset.

Residual Generator	Baseline	+ Ours	ΔROCAUC
LSTM-AE	0.633 ± 0.026	**0.745** ± 0.022	+0.112
GRUAE	0.645 ± 0.006	**0.730** ± 0.012	+0.085
Sequencebottleneck RNNAE	0.605 ± 0.011	**0.717** ± 0.010	+0.112
Graphconvolutional AE	0.500 ± 0.009	**0.646** ± 0.009	+0.146

6. Discussion

6.1 Case study: local residual peaks in a normal multiperson walking video

Figure 3 presents a falsealarm correction example from a normal indoor corridor video in ShanghaiTech. Multiple pedestrians pass through the scene, but no safetyrelated event occurs. The extracted skeleton sequences mainly correspond to the more stably tracked foreground pedestrians in the lowerright region; the distant pedestrian in the upperleft area is not extracted. For this video, the raw temporalmean score is 0.01363 and exceeds its independently fitted normalvalidation threshold of 0.01069. The calibrated uppertail score is 0.01359 and remains below its separately fitted threshold of 0.01469. The case therefore illustrates the combined effect of residual calibration, uppertail aggregation, and validationbased operatingpoint estimation; it does not isolate the contribution of any single step.

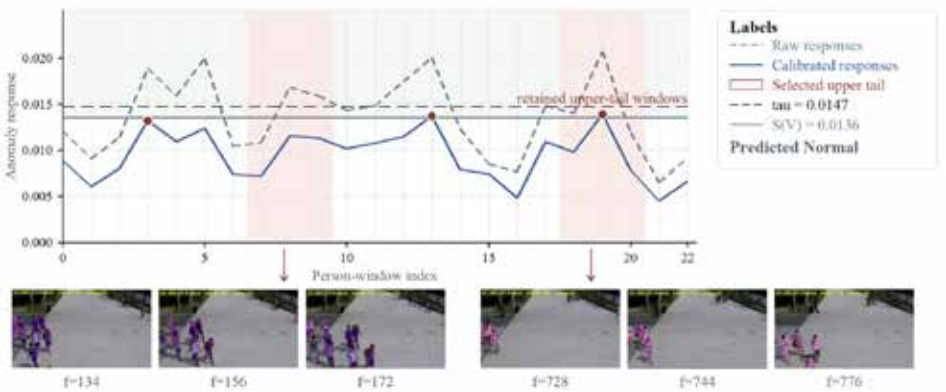

Figure 3. Raw and calibrated personwindow responses for a normal ShanghaiTech walking video. The selected uppertail windows yield $S(V) = 0.0136 < \tau = 0.0147$, resulting in a normal prediction (LSTMAE, seed 44; $\alpha = 0.15$ selected on normalvalidation data).

This case illustrates that localized peaks in skeleton reconstruction residuals do not always correspond to safetyrelevant events. As reflected in the lowresponse segment on the left of Figure 3, calibrated residuals remain low even in crowded scenes when tracked targets have complete skeleton extraction, continuous trajectories, and stable pose observations. Conversely, the highresponse segment on the right corresponds to a local residual peak where fewer pedestrians are present, yet a single target trajectory exhibits a high response. Framebyframe analysis shows that this peak is concentrated primarily in lowerlimb keypoints, particularly the left and right ankles. During oblique or lateral walking, relative ankle displacement and localized trajectory fluctuations across consecutive windows coincide with increased lowerlimb reconstruction residuals. The localized high response is therefore consistent with pose observation instability rather than a safetyrelevant event. This viewpoint sensitivity suggests that orientationaware calibration or viewpoint normalization may help reduce pose uncertainty under challenging surveillance angles. Since the current scoring scheme relies on post hoc statistical calibration without explicitly modeling viewpoint variation, such extensions remain

promising directions for future work.

6.2 Failure mode analysis

We analyze failure cases of the LSTMAE backbone on the ShanghaiTech test set to characterize the error distribution. On this benchmark, our method yields 21 false positives (FPs) and 7 false negatives (FNs). Figure 4 summarizes the error counts and the corresponding regional and jointlevel contributions.

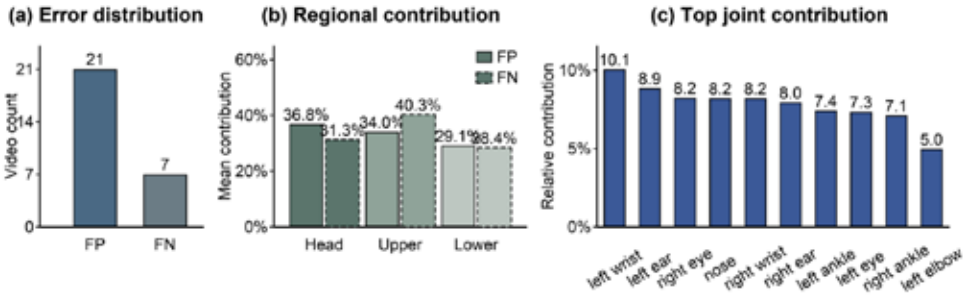

Figure 4. Error profile on the ShanghaiTech test set using LSTMAE (seed 44; $\alpha = 0.15$). (a) Falsepositive (FP) and falsenegative (FN) counts. (b) Mean calibrated residual contributions by body region. (c) Largest jointlevel contributions among thresholded errors.

False positive analysis

Falsepositive (FP) samples are analyzed from two perspectives: trajectory quality and regional residual contributions. FP samples exhibit lower average trajectory coverage than truenegative (TN) samples (0.81 vs. 0.92) and are more likely to occur during track initiation or termination phases (trackboundary flag: 0.89 vs. 0.65). These statistics are consistent with incomplete trajectories and transient observation fluctuations at track boundaries contributing to elevated residual responses and biased videolevel scores.

Examining regional contributions within falsealarm windows reveals that the calibrated head region accounts for the largest proportion (36.85%), followed by the upper body (34.02%) and lower body (29.13%). Note that these proportions reflect postcalibration residual contributions rather than raw residual magnitudes. Because head keypoints exhibit smaller residual scales on the normal validation set, inversescale calibration assigns them higher response weights. Consequently, minor head keypoint variations caused by perspective shifts, profile poses, or distant localization errors can contribute disproportionately during scoring, thereby increasing falsealarm risks.

False negative analysis

Among the seven false negatives (FNs) generated by the system, the upper body region yields the highest residual contribution (40.28%), followed by the head (31.31%) and lower body (28.41%). Manual review showed that all seven missed videos involved multiperson interactions, including theft followed by pursuit, pushing, fighting, and pursuit followed by confrontation. These missed detections appear to reflect limited separability between the residual patterns of certain safetyrelevant interactions and those of normal activities, rather than a complete absence of detected human motion.

For instance, brief physical contact or pursuit sequences involve upperbody and hand movements; however, their joint dynamics in 2D singleperson pose representations may resemble routine arm swinging or turning, which limits the anomaly response after videolevel aggregation.

These cases reveal a limitation of current singleperson skeleton representations: although they effectively capture individual motion, they remain insufficient for modeling multiperson interactions, behavioral intent, and scene context.

Future work could incorporate multiperson interaction modeling and contextual cues to improve discrimination in complex campus safetymonitoring scenarios.

6.3 Scope and deployment considerations

Because campus safety anomalies occur infrequently, the limited scale of the test data may affect the precision of performance estimates. We therefore conduct a paired videolevel bootstrap analysis [24]. Based on 2,000 paired videolevel resampling iterations, the estimated ROCAUC of our method is 0.841 with a 95% confidence interval (CI) of [0.758, 0.917]. Compared with the raw baseline, the estimated ROCAUC improvement is 0.050, although its 95% CI of [-0.026, 0.132] crosses zero. Thus, the current ShanghaiTech safetyevent evaluation setting does not establish a statistically conclusive improvement over the raw baseline. Further validation on larger, multiscene campus safety benchmarks is needed to assess the stability of the estimated gain.

From a practical deployment standpoint, our goal is to improve anomaly scoring without increasing the complexity of the reconstruction backbone. Operating directly on reconstruction residuals, the scoring scheme introduces no additional trainable parameters and requires no abnormal samples for score calibration. It adds only residual calibration, completeness weighting, and uppertail aggregation. Because these operations add no trainable parameters, the protocol has low structural complexity. Endtoend runtime and memory measurements are still needed before making stronger claims about realtime or resourceconstrained deployment.

7. Conclusion

This paper presents a videolevel residual scoring protocol for skeletonbased video anomaly detection. Without altering the underlying skeleton reconstruction architecture, the protocol calibrates residual responses using normalvalidation statistics and retains concentrated highresponse evidence through uppertail aggregation. Experiments under the curated ShanghaiTech and NWPU Campus safetyevent evaluation settings show gains across the evaluated residualgenerating backbones without adding trainable parameters. The failure analysis illustrates how localized pose observation instability can affect videolevel scores and highlights the limitations of singleperson skeleton representations for modeling complex multiperson interactions.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 15:1–15:58, 2009, doi: 10.1145/1541880.1541882.

[2] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, “A survey of singlescene video anomaly detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2293–2312, 2022, doi: 10.1109/TPAMI.2020.3040591.

[3] O. Hirschorn and S. Avidan, “Normalizing flows for human pose anomaly detection,” in ICCV, 2023, pp. 13545–13554. doi: 10.1109/ICCV51070.2023.01246.

[4] G. Alinezhad Noghre, A. Danesh Pazho, V. A. Katariya, and H. Tabkhi, “Understanding the challenges and opportunities of posebased anomaly detection,” in Proceedings of the 8th international workshop on sensorbased activity recognition and artificial intelligence, 2023, pp. 1–9. doi: 10.1145/3615834.3615844.

[5] R. Morais, V. Le, T. Tran, B. Saha, M. R. Mansour, and S. Venkatesh, “Learning regularity in skeleton trajectories for anomaly detection in videos,” in CVPR, 2019, pp. 11996–12004. doi: 10.1109/CVPR.2019.01227.

[6] A. Flaborea, G. M. D’Amely di Melendugno, S. D’Arrigo, M. A. Sterpa, A. Sampieri, and F. Galasso, “Contracting skeletal kinematics for humanrelated video anomaly detection,” Pattern Recognition, vol. 156, p. 110817, 2024, doi: 10.1016/j.patcog.2024.110817.

[7] A. Delić, M. Grčić, and S. Šegvić, “Sequential keypoint density estimator: An overlooked baseline of skeletonbased video anomaly detection,” in ICCV, 2025, pp. 11579–11589. doi: 10.1109/ICCV51701.2025.01077.

[8] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection – a new baseline,” in CVPR, 2018, pp. 6536–6545. doi: 10.1109/CVPR.2018.00684.

[9] C. Cao, Y. Lu, P. Wang, and Y. Zhang, “A new comprehensive benchmark for semisupervised video anomaly detection and anticipation,” in CVPR, 2023, pp. 20392–20401. doi: 10.1109/CVPR52729.2023.01953.

[10] W. Luo, W. Liu, and S. Gao, “Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection,” Neurocomputing, vol. 444, pp. 332–337, 2021, doi: 10.1016/j.neucom.2019.12.148.

[11] N. Li, F. Chang, and C. Liu, “Humanrelated anomalous event detection via spatialtemporal graph convolutional autoencoder with embedded long shortterm memory network,” Neurocomputing, vol. 490, pp. 482–494, 2022, doi: 10.1016/j.neucom.2021.12.023.

[12] A. Stergiou, B. De Weerdt, and N. Deligiannis, “Holistic representation learning for multitask trajectory anomaly detection,” in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2024, pp. 6729–6739. doi: 10.1109/WACV57701.2024.00659.

[13] A. Markovitz, G. Sharir, I. Friedman, L. ZelnikManor, and S. Avidan, “Graph embedded pose clustering for anomaly detection,” in CVPR, 2020, pp. 10539–10547. doi: 10.1109/CVPR42600.2020.01055.

[14] A. Flaborea, L. Collorone, G. M. D. di Melendugno, S. D’Arrigo, B. Prenkaj, and F. Galasso, “Multimodal motion conditioned diffusion model for skeletonbased video anomaly detection,” in ICCV, 2023, pp. 10318–10329. doi: 10.1109/ICCV51070.2023.00947.

[15] A. Karami, T. K. K. Ho, and N. Armanfard, “Graphjigsaw conditioned diffusion model for skeletonbased video anomaly detection,” in WACV, 2025, pp. 4237–4247. doi: 10.1109/WACV61041.2025.00416.

[16] R. T. Rockafellar and S. Uryasev, “Optimization of conditional valueatrisk,” The Journal of Risk, vol. 2, no. 3, pp. 21–41, 2000, doi: 10.21314/JOR.2000.038.

[17] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8.” https://github.com/ultralytics/ultralytics; Ultralytics, 2023.

[18] Y. Zhang et al., “ByteTrack: Multiobject tracking by associating every detection box,” in ECCV, 2022, pp. 1–21. doi: 10.1007/9783031200472_1.

[19] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in ECCV, 2014, pp. 740–755. doi: 10.1007/9783319106021_48.

[20] S. Hochreiter and J. Schmidhuber, “Long shortterm memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[21] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in Proceedings of the 2014 conference on empirical methods in natural language processing, Association for Computational Linguistics, 2014, pp. 1724–1734. doi: 10.3115/v1/D141179.

[22] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in ICLR, 2019. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[23] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeletonbased action recognition,” in AAAI, 2018, pp. 7444–7452. doi: 10.1609/aaai.v32i1.12328.

[24] B. Efron, “Bootstrap methods: Another look at the jackknife,” The Annals of Statistics, vol. 7, no. 1, pp. 1–26, 1979, doi: 10.1214/aos/1176344552.

Author Biography:

Zheqi Lyu (1995.01–), female, Han Chinese, Quzhou, Zhejiang Province, Lecturer at Zhejiang Technical Institute of Economics, data security and artificial intelligence.

Ke Chen* (1980.03—), female, Han Chinese, Yongkang, Zhejiang Province. Associate professor, Zhejiang Technical Institute of Economics, Student Management and Software Engineering.

Fund Project:

This work was supported by the Humanities and Social Science Fund of Ministry of Education of China (Grant No. 24JDSZ3007).