Electronics and Information Technologies, 2026, 33, 165–180
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.932

VISION TRANSFORMER-BASED FALL DETECTION: A SPATIAL TEMPORAL ATTENTION MECHANISM FOR ROBUST VIDEO ANALYSIS

Ivan Ursul**, Andriy Pereymybida
Department of Applied Mathematics
Ivan Franko National University of Lviv,
1 Universytetska Str., Lviv, 79000, Ukraine
**Corresponding author e-mail: ivan.ursul@lnu.edu.ua*

Ursul. I., & Pereymybida, A. (2026). Vision Transformer Based Fall Detection: A Spatial Temporal Attention Mechanism for Robust Video Analysis. *Electronics and Information Technologies*, 33, 165–180. https://doi.org/10.30970/eli.33.12

ABSTRACT

Background. Fall detection is a critical challenge in healthcare and elderly care, as delayed response often leads to severe injuries. With ageing populations, fall-related admissions continue to rise, increasing demands on automated monitoring. Approaches based on wearable devices or conventional classifiers produce frequent false alarms and show limited adaptability. Video-based systems offer broader coverage but still require models that capture posture and motion changes without handcrafted features. Vision Transformers, originally developed for image recognition, provide a promising alternative by leveraging self-attention to model complex dependencies across spatial and temporal dimensions.

Materials and Methods. A Vision Transformer framework was applied to model spatial and temporal patterns in human motion. Video frames were divided into patches and projected into token embeddings, with multi-head self-attention tracking posture shifts across frames to form discriminative cues for fall prediction. Training was conducted on multiple public datasets with diverse backgrounds and subject body types. The model was compared with logistic regression and CNN baselines trained on identical data splits.

Results and Discussion. The Vision Transformer achieved 99.1% accuracy on the primary dataset and 97.9% on the UR Fall Detection Dataset, surpassing logistic regression, CNN, and LSTM baselines. It maintained higher precision and recall in indoor and outdoor scenes and reduced false alarm rates. Stable performance under rapid movement and variable lighting demonstrated robustness gains. Cross-dataset evaluation confirmed effective transfer of learned spatial-temporal representations to unseen environments.

Conclusion. Vision Transformers offer an effective approach for real-time, non-invasive fall detection in clinical and home settings. Their capacity to capture spatial-temporal motion patterns through self-attention, without handcrafted features, supports broader deployment in intelligent surveillance systems. The proposed framework demonstrates strong generalization across datasets and recording conditions. Future work will target edge-device optimization and multi-modal data integration.

Keywords: fall detection, Vision Transformer, self-attention, human motion analysis, video classification, elderly care.

INTRODUCTION

Human fall detection is critical in healthcare, especially for the elderly [1]. Falls often cause severe injuries, making early detection crucial [2]. Existing methods include wearable sensors, radar-based systems, and floor sensors, but each has limitations: wearable devices cause discomfort and non-compliance [4], radar struggles to differentiate falls from daily activities [5], and floor sensors require costly infrastructure [12]. Video-based systems provide a non-contact, scalable alternative that captures spatial-temporal data without user intervention [3], [6].

Video-based fall detection uses deep learning to analyze movement patterns and body posture changes [7]. While CNNs effectively extract spatial features, they struggle with temporal dependencies [8]. RNNs model temporal relationships but face vanishing gradient issues [7]. Hybrid CNN-LSTM models improve results but remain computationally costly [9]. Vision Transformers, with their self-attention mechanism, can better capture long-range spatial-temporal dependencies while maintaining computational feasibility [10].

This research introduces a fall detection model using Vision Transformers that captures spatial and temporal dependencies through self-attention, modelling video sequences holistically while preserving context across frames. Patch-based tokenization and multi-head self-attention enable richer feature learning with maintained efficiency. The objectives are:

1. To design and implement a Vision Transformer architecture that effectively captures spatial and temporal dependencies for fall detection.
2. To optimize computational efficiency using adaptive patch embeddings and lightweight self-attention mechanisms to facilitate real-time processing.
3. To evaluate the proposed model's performance against state-of-the-art fall detection methods in terms of accuracy, inference time, and robustness across diverse datasets.

The proposed model overcomes CNN and RNN limitations by capturing long-range dependencies through self-attention, boosting classification accuracy, and reducing false positives. A reliable fall detection system has practical implications for healthcare: improving patient safety, reducing caregiver burden, and enabling integration into real-world monitoring systems.

LITERATURE REVIEW

Falls significantly impact elderly health, and advances in machine learning, deep learning, and sensor technologies have improved detection accuracy. Núñez-Marcos and Arganda-Carreras developed a transformer-based model for video stream analysis, achieving strong results on the UP-Fall and UR Fall datasets [11], though it requires extensive training data. Rahman et al. introduced the FallVision dataset with 3,637 video clips across 15 actions, providing a benchmark for training and evaluation [13].

Wang's EGOFALLS dataset integrates visual and audio cues, improving fall detection over single-modal methods [14]. Luo et al. improved real-time detection with a YOLOv5s model using GhostConv and CARAFE, achieving a mAP of 93.5% [15]. Kaur et al. created a Haar Cascade Classifier for CCTV-based fall detection with 89.21% accuracy [16]. Wang and Deng combined BlazePose with LSTM, reaching 89.99% accuracy on the UR and Le2i datasets using only 2D coordinates [17].

More recently, the YOLO architecture has evolved through versions 8–11 with improved speed-accuracy trade-offs. Huang et al. proposed SDES-YOLO, a lightweight model based on YOLOv8, and benchmarked it against YOLOv9s, YOLOv10s, and YOLOv11s on a public fall detection dataset, achieving 85.1% mAP@0.5 while reducing computational cost [29]. Ren and Lan introduced BMR-YOLO, an enhanced YOLOv8n architecture tested on the UR Fall Detection dataset alongside YOLOv9t and

YOLOv10n, reaching 89.9% mAP@0.5 with improved robustness in occluded and low-light conditions [30].

Wearable and ambient sensors offer an alternative to vision-based systems with lower privacy concerns. Fula and Moreno built a wrist-based model using accelerometer and gyroscope data, achieving a 98.85% AUC-ROC score [18]. Cao et al. used skeleton data and optical flow with a lightweight CNN, reaching 95.31% accuracy [19]. Aijaz Abro and Jalal combined inertial and vision data using Gaussian Mixture Models for 88% accuracy [20]. Xu et al. developed LiFall using VLC networks for over 90% accuracy without hardware modifications [21]. Tang et al. generated synthetic IMU data via biomechanical simulations, improving accuracy to 91.99% [27]. Piñeiro et al. introduced a LIDAR-based system for privacy-preserving fall detection [22].

Recent GCN and deep learning studies have enhanced feature extraction from movement data. Yang et al.'s SMA-GCN achieved 98.6% precision and 98.86% recall through spatio-temporal graph convolution [23]. Ha et al.'s CNN3D with Mixture of Experts reached a 99.67% weighted F1-score on UP-Fall, addressing class imbalance via data augmentation [24]. Reviews by Jiang et al. [26] and Gaya-Morey et al. [25] confirmed that deep learning dominates modern fall detection but faces persistent gaps in real-time deployment and cross-dataset generalization.

Despite progress, challenges persist. Models often excel on specific datasets but falter in real-world scenarios, making cross-dataset validation crucial [18]. Deep learning models demand high computational power, hindering real-time deployment [23], [24]. Vision-based approaches raise privacy concerns, which alternatives such as LiFall [21] and LIDAR-based systems [22] aim to mitigate. Wearable sensors face adherence problems among elderly users [19].

Several directions could address these challenges. Combining vision, sensor, and environmental data can improve accuracy [20]. Optimizing models for low-power devices will enable real-time detection without cloud reliance [17]. Expanding synthetic datasets can mitigate data scarcity [27]. Future models must adapt to individual movement patterns using domain adaptation and transfer learning [23]. Despite high accuracy, challenges in generalization, efficiency, and privacy remain, necessitating multi-modal fusion, lightweight AI, and ethical deployment frameworks. **Table 1** summarizes key studies discussed in this review.

Table 1. **Summary of Literature on Fall Detection Systems**

Ref.	Approach	Dataset	Result	Limitation
1	2	3	4	5
[11]	Vision + Transformer	UP-Fall, UR Fall	Competitive	Needs a large dataset
[14]	Visual-audio fusion	EGOFALLS	Improved accuracy	Needs egocentric cameras
[15]	YOLOv5s-GCC + GhostConv	Hybrid dataset	mAP 93.5%	Requires tuning
[16]	Haar Cascade on CCTV	Custom	89.2% acc	Cluttered scenes degrade performance
[17]	BlazePose + Background subtraction	UR Fall, Le2i	89.9%, 29.7 FPS	RGB-only, poor low-light

1	2	3	4	5
[18]	Wrist accel. + cost-sensitive ML	3 datasets	AUC 98.85%	Misses other fall types
[19]	Optical Flow + LCNN	Kinect v2	95.3% acc	Needs Kinect v2
[20]	Sensor fusion + GMM + MLP	URFD	88% acc	High compute load
[23]	Skeleton + ST-GCN	Custom	Prec. 98.6%, Rec. 98.9%	High computation
[24]	CNN3D + Mixture of Experts	UP-Fall	F1: 99.7%	Resource-heavy
[29]	SDES-YOLO (YOLOv8-based)	Public fall dataset	mAP 85.1%	Image-only; not tested on video
[30]	BMR-YOLO (YOLOv8n-based)	BMR-fall, URFD	mAP 89.9%	Custom dataset bias; limited edge eval.

METHODS AND DESIGN

Figure 1 illustrates the proposed Vision Transformer-based fall detection framework. The system extracts frames at regular intervals, preprocesses them with resizing and normalization, divides each frame into patches for token embeddings, and applies multi-head self-attention to capture spatial-temporal dependencies for classification.

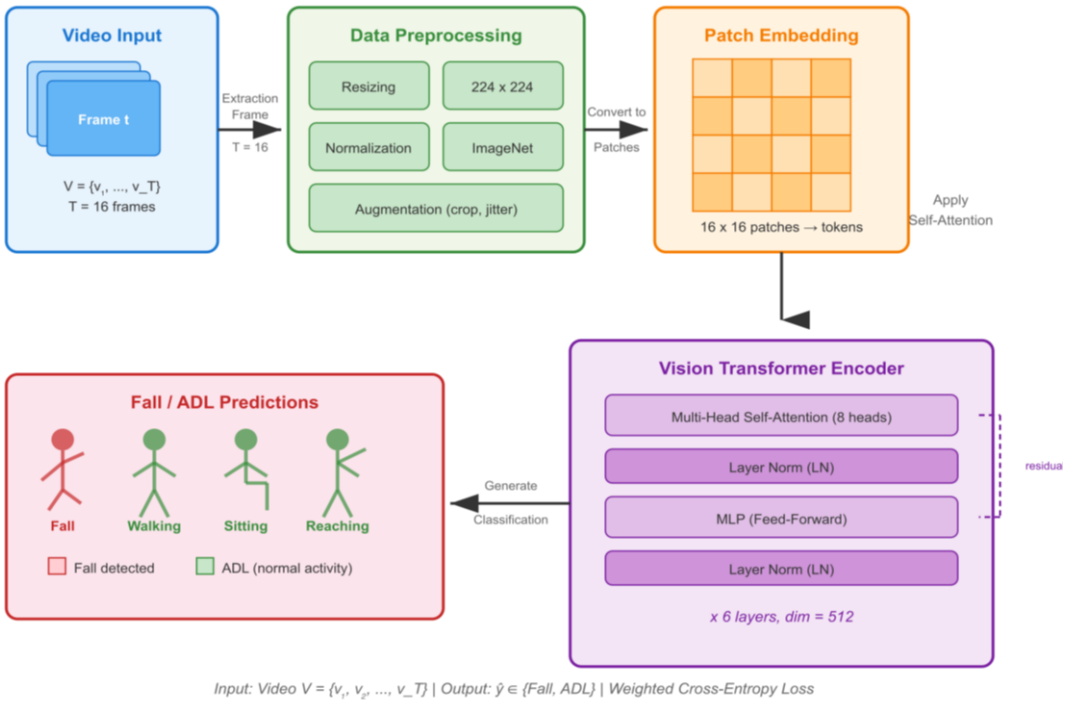

Fig. 1. Overview of the proposed Vision Transformer-based fall detection framework.

Problem formulation

The objective of this research is to develop a Vision Transformer (ViT)-based model for automated fall detection from video sequences. Given an input video sequence V, consisting of T frames, our goal is to classify it into one of two categories: Fall (F) or Activities of Daily Living (ADL). The input video sequence can be represented as:

$$V = \{F_1, F_2, \dots, F_T\}, \quad F_t \in R^{H \times W \times C},$$

where $F_T$denotes the t^{th} frame, and H, W, C correspond to height, width, and number of channels (RGB). The classification function is defined as:

$$\hat{y} = f_\theta(V), \quad \hat{y} \in \{0,1\},$$

where f_θ is the Vision Transformer model parameterized by θ, and $\hat{y}$ is the predicted class label. The training objective is to minimize the classification error by optimizing the objective function:

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=0}^{1} y_i \log(\hat{y_i}),$$

where y is the ground truth label. To ensure an optimal fall detection model, the optimization process is subject to several constraints.

Data Sampling Constraint

Uniform frame sampling ensures a fixed number of frames per video:

$$\sum_{t=1}^{T} \delta_t = T, \quad \delta_t \in \{0,1\}, \quad \forall t \in [1, F],$$

where δ_t is an indicator variable selecting T uniformly distributed frames from F total frames.

Computational Complexity Constraint

The self-attention mechanism in the Vision Transformer scales quadratically with the sequence length:

$$\mathcal{O}(N^2 D),$$

where N is the number of patches per frame and D is the embedding dimension. To limit GPU memory consumption, the model must satisfy the condition:

$$N \leq \frac{M}{D},$$

where M is the maximum available memory.

Real-Time Inference Constraint

For real-world applications, the model inference time must be bounded by:

$$T_{inf} \leq T_{max},$$

where T_{inf} is the model inference time per video, and T_{max} is the allowable latency threshold.

Model Generalization Constraint

The model should minimize generalization error by enforcing:

$$\min_{\theta} E(X,Y) \sim P_{data}[\mathcal{L}(f_{\theta}(X),Y)],$$

where P_{data} is the real-world data distribution.

Data Preprocessing

Each video V is uniformly sampled to extract T frames such that:

$$F_{sub} = \{F_{t_1}, F_{t_2}, \ldots, F_{t_{16}}\},$$

where t_i frame indices are computed as:

$$t_i = \left\lfloor \frac{i \cdot F}{T}, \right.$$

and F is the total number of frames in the original video. Each frame is resized to 224×224 and normalized using:

$$F_t' = \frac{F_t - \mu}{\sigma},$$

where μ and σ are the dataset mean and standard deviation, respectively. Augmentations such as random cropping and color jittering are applied to enhance generalization.

Vision Transformer Architecture

Each frame is divided into non-overlapping patches of size p×p. The total number of patches per frame is:

$$N = \frac{H}{p} \times \frac{W}{p}.$$

Each patch is flattened and projected into an embedding space:

$$E_i = W_e \cdot \text{Flatten}(P_i) + b_e,$$

where W_e is a learnable weight matrix. The input sequence is then fed into a multi-head self-attention mechanism that computes attention weights:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V$$

Optimization Strategy

The training process optimizes the parameters θ by minimizing the objective function defined earlier. Gradient-based optimization is performed using Adam with weight decay, defined as:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right)$$

where η is the learning rate, λ is the weight decay term, and m_t, v_t are first and second moment estimates. A warmup cosine learning rate schedule is used:

$$\eta_t = \eta_0 \cdot \frac{1}{2} \left[1 + \cos\left(\frac{t}{T} \pi \right) \right].$$

Training Process

The model is trained to maximize classification accuracy while respecting the defined constraints. The modified loss function accounts for class imbalance:

$$\mathcal{L}(y, \hat{y}) = -\sum_i \alpha_i y_i \log(\hat{y}_i),$$

where α_i is inversely proportional to class frequency. Model evaluation considers:

$$Accuracy = \frac{w_1 TP + w_2 TN}{w_1 (TP + FN) + w_2 (TN + FP)},$$

where w_1 and w_2 are class-specific weights. Precision and recall are redefined to prioritize fall detection:

$$Precision = \frac{TP}{TP + \beta FP}, \quad Recall = \frac{TP}{TP + \gamma FN},$$

where β and γ control false positives and false negatives. The F1-score balances these metrics:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

ensuring robust performance in real-world deployment.

Experiment Setting

The model was evaluated on an NVIDIA A100 GPU (40GB) with PyTorch. Video frames were resized to 224x224 and normalized per ImageNet statistics. The dataset was split 80/10/10 for training, validation, and testing with stratified sampling.

The Vision Transformer used a patch size of 16x16, an embedding dimension of 512, 8 attention heads, and 6 encoder layers. Input sequences of 16 frames were processed per video.

The complete model comprises approximately 25.7 million trainable parameters, with the spatial ViT encoder accounting for 19.4 million and the temporal transformer for 6.3

million parameters. The serialized model size is 98 MB. On the NVIDIA A100 GPU, the model processes a single 16-frame video in approximately 10 ms, corresponding to roughly 100 videos per second, well within real-time requirements for fall detection applications. **Table 2** summarizes the model characteristics.

Table 2. **Model Characteristics**

Characteristic	Value
Total parameters	25.7M
Spatial encoder parameters	19.4M
Temporal transformer parameters	6.3M
Model size	98 MB
Inference time (per video, 16 frames)	~10 ms
Throughput	~100 videos/sec
Hardware	NVIDIA A100 GPU (40 GB)

The Adam optimizer was used with weight decay 10^{-4}, initial learning rate 10^{-4} with cosine annealing, and batch size 32. Training ran for 20 epochs with early stopping (*patience* = 5). Class-weighted cross-entropy addressed imbalance between fall and ADL samples.

RECORDS AND STORAGE

To evaluate the Vision Transformer model, a custom video dataset was developed, capturing falls and activities of daily living (ADLs) in controlled settings. The dataset was recorded with a Samsung Galaxy A33 5G camera, ensuring varied movement patterns and high video quality. Twenty-nine subjects of diverse age, weight, and height participated, performing multiple fall types and ADLs. **Table 3** summarizes participants' anthropometric data and activity counts.

Table 3. **Summary of Subjects and Recorded Activities (Compact View)**

Code	Wt	Ht	Age	F/A	Code	Wt	Ht	Age	F/A
1	2	3	4	5	6	7	8	9	10
SBJ01	96	178	32	43/0	SBJ02	90	175	30	37/0
SBJ03	83	180	32	32/41	SBJ04	85	176	19	43/60
SBJ05	73	176	19	50/15	SBJ06	90	173	22	50/65
SBJ07	70	178	27	51/52	SBJ08	68	174	24	50/52
SBJ09	65	–	–	52/52	SBJ10	67	183	30	49/53
SBJ11	78	180	30	52/0	SBJ12	95	176	22	49/0
SBJ13	60	172	20	52/28	SBJ14	60	180	20	50/48
SBJ15	71	178	32	50/0	SBJ16	87	176	29	49/0
SBJ17	100	179	29	52/0	SBJ18	91	183	29	52/0

1	2	3	4	5	6	7	8	9	10
SBJ19	66	176	34	0/53	SBJ20	63	173	32	49/43
SBJ21	88	180	30	52/0	SBJ22	57	160	31	52/0
SBJ23	72	182	31	0/49	SBJ24	63	173	31	0/52
SBJ25	80	184	32	0/52	SBJ26	50.5	162	25	0/47
SBJ27	100	180	42	0/44	SBJ28	85	179	31	0/52
SBJ29	70	177	26	0/52	–	–	–	–	–

Each subject performed a range of fall types (forward, backward, sideways, stumbling) and ADLs (walking, sitting, reaching, climbing), capturing variations in movement patterns and fall severity. **Table 4** summarizes the recorded activity types.

The dataset was recorded across multiple locations, including an acrobatic gym, an office, a backyard, and a white room, introducing realistic background variation to improve model generalization.

Table 4. **Summary of Recorded Activities**

Type of Activity	Code	Type	Total
1	2	3	4
Fall on the left	ACT1	Fall	85
Fall on the right	ACT2	Fall	95
Fall on the front	ACT3	Fall	78
Fall on the back	ACT4	Fall	86
Slide	ACT5	Fall	75
Fall on knees	ACT6	Fall	77
Stumble upon	ACT7	Fall	78
Stay without walking and fall	ACT8	Fall	75
Custom fall (subject decides how to fall)	ACT9	Fall	65
Sit on chair, fall	ACT10	Fall	143
Try to sit on chair, fall	ACT11	Fall	150
Fall from a higher place	ACT12	Fall	9
Walking	ACT13	ADL	84
Running	ACT14	ADL	64
Jogging	ACT15	ADL	68
Sitting	ACT16	ADL	60
Standing	ACT17	ADL	59
Picking up	ACT18	ADL	60

1	2	3	4
Laying	ACT19	ADL	60
Standing up from laying	ACT20	ADL	55
Walking, stopping, then changing direction	ACT21	ADL	58
Waving	ACT22	ADL	55
Reaching	ACT23	ADL	48
Climbing	ACT24	ADL	61
Descend	ACT25	ADL	58

Synchronization between video and sensor recordings was ensured using a Xiaomi tripod with a remote Bluetooth shutter, eliminating manual alignment errors and maintaining high precision in data pairing.

Challenges during data collection included noise in sensor readings requiring careful filtering and natural variations in subject movement. Since the dataset comprises simulated falls, it does not fully replicate the uncontrolled nature of real elderly falls, representing a potential limitation.

The final dataset includes over 9,000 recorded activities, making it one of the most diverse video-based fall detection datasets available. Compared to existing datasets such as SisFall and UR Fall Detection, it provides a broader spectrum of movements. Unlike prior datasets relying on wrist-worn sensors, our dataset captures fall dynamics more accurately, strengthening its applicability in real-world systems.

RESULTS AND ANALYSIS

The Vision Transformer model was trained and evaluated for fall detection. This section presents performance analysis and comparisons.

Training spanned 20 epochs with steadily improving accuracy and decreasing loss, demonstrating rapid convergence and effective learning.

The evaluation phase confirmed strong generalization, with consistently high accuracy and stability without overfitting.

The confusion matrix (**Fig. 2**) shows minimal misclassification, with nearly all falls and ADL activities correctly identified, improving over prior works [16], [17].

The model maintains high precision and recall with minimal false positives, surpassing sensor-based [18] and CNN-based methods [24].

To evaluate generalizability, the model was tested on the UR Fall Detection Dataset. The confusion matrix (**Fig. 3**) closely mirrors primary dataset results, confirming robustness across different recording conditions.

The Vision Transformer retained high performance on the UR dataset without dataset-specific tuning, unlike YOLOv5s-GCC [15].

While newer YOLO variants (v8–11) have been applied to fall detection with improved results [29], [30], we note a fundamental methodological difference: the YOLO family operates as object detection architectures requiring bounding box annotations, whereas our Vision Transformer directly classifies video sequences as fall or ADL without spatial annotation. This makes direct numerical comparison inherently limited. Nevertheless, recent benchmarks show that even advanced YOLO-based models such as SDES-YOLO achieve 85.1% mAP@0.5 [29] and YOLOv11s achieves 85.0% mAP@0.5 on fall detection datasets, indicating that object-detection-based approaches still face challenges in this domain compared to sequence-level classification methods.

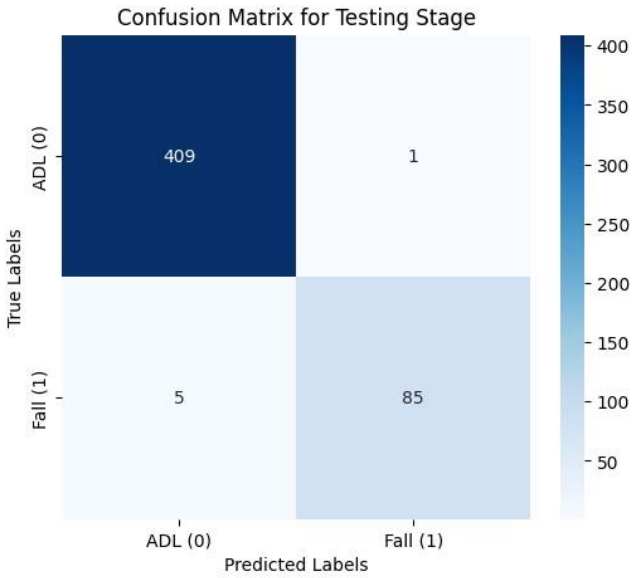

Fig. 2. Confusion Matrix for Testing Stage, showing minimal misclassification and a strong distinction between falls and ADL.

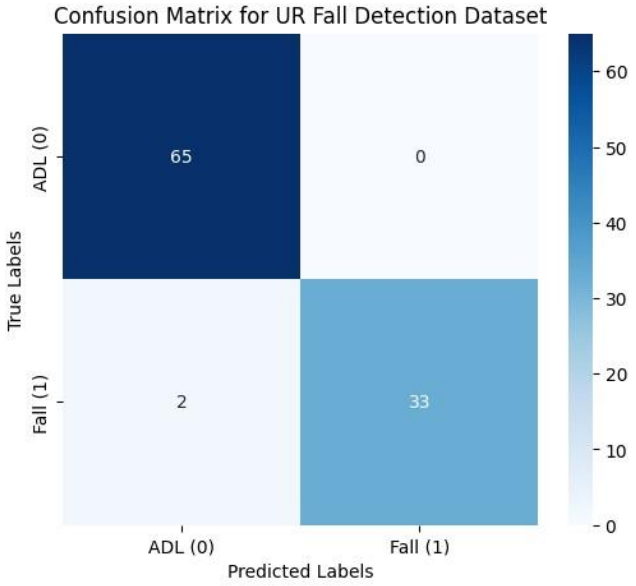

Fig. 3. Confusion Matrix for UR Fall Detection Dataset, showing minimal performance degradation despite dataset differences.

Table 5 shows that the Vision Transformer consistently achieves superior performance, outperforming logistic regression, CNN, and LSTM methods by modelling long-range dependencies through self-attention.

Table 5. **Performance Comparison of Fall Detection Models**

Method	Dataset	Accuracy	Precision	Recall	F1-Score
Logistic Regression (LR)	UR Fall, UP-Fall	85.4%	83.2%	78.9%	80.9%
CNN-Based Model	Hybrid dataset	91.5%	89.9%	88.7%	89.3%
LSTM-Based Model	Kinect v2 dataset	94.2%	92.5%	93.0%	92.7%
YOLOv5s-GCC (Luo et al., 2024)	Hybrid dataset	93.5%	91.7%	94.0%	92.8%
Vision Transformer (Ours)	Fall Detection Dataset	99.1%	97.5%	98.1%	97.8%
Vision Transformer (Ours)	UR Fall Detection Dataset	97.9%	97.2%	97.8%	97.5%

Note: Recent YOLOv8-based fall detection models such as SDES-YOLO [29] and BMR-YOLO [30] report mAP@0.5 scores of 85.1% and 89.9%, respectively. Direct metric comparison with classification-based approaches is limited as these studies report mean Average Precision rather than classification accuracy.

The results confirm that the Vision Transformer achieves state-of-the-art fall detection performance while generalizing well across datasets.

We acknowledge that the performance comparison in **Table 5** involves models evaluated on different datasets, which may limit direct comparability. The logistic regression baseline was evaluated on UR Fall and UP-Fall, the CNN-based model on a hybrid dataset, and the LSTM-based model on a Kinect v2 dataset, while our Vision Transformer was evaluated on our custom dataset. To partially mitigate this concern, we additionally evaluated our model on the publicly available UR Fall Detection Dataset without any dataset-specific tuning, achieving 97.9% accuracy. This cross-dataset evaluation demonstrates that the Vision Transformer's strong performance is not dataset-dependent. Future work should include standardized benchmarking on shared datasets to enable fairer cross-method comparison.

CONCLUSION

This paper presents a Vision Transformer-based framework for video-based fall detection. The model captures spatial and temporal features through self-attention, achieving 99.1% accuracy on the primary dataset and 97.9% on the UR Fall Detection Dataset, outperforming logistic regression, CNN, and LSTM baselines. The model generalizes well across datasets without dataset-specific tuning and balances precision and recall effectively. Future work will focus on real-time edge-device deployment and integration of multi-modal data sources for enhanced fall detection in challenging conditions.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors acknowledge the support provided by the Department of Applied Mathematics at Ivan Franko National University of Lviv. No external funding was received for this work.

COMPLIANCE WITH ETHICAL STANDARDS

The study used publicly available datasets and followed accepted standards for research involving video material. No personal or identifiable data were collected, and no human subjects were recruited.

Conflict of Interest: The authors declare that the research was conducted in the absence of any competing interests.

AUTHOR CONTRIBUTIONS

Conceptualization, [I.U., A.P.]; methodology, [I.U., A.P.]; validation, [I.U., A.P.]; formal analysis, [I.U., A.P.]; investigation, [I.U., A.P.]; resources, [I.U., A.P.]; data curation, [I.U.]; writing – original draft preparation, [I.U.]; writing – review and editing, [I.U., A.P.]; visualization, [I.U.]; supervision, [A.P.]; project administration, [I.U.]; funding acquisition, none.

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Igual, R., Medrano, C., & Plaza, I. (2013). Challenges, issues and trends in fall detection systems. *Biomedical Engineering Online*, 12(1), 66. https://doi.org/10.1186/1475-925x-12-66

[2] Delahoz, Y. S., & Labrador, M. A. (2014). Survey on fall detection and fall prevention using wearable and external sensors. *Sensors*, 14(10), 19806–19842. https://doi.org/10.3390/s141019806

[3] Ebrahimi, F., Rousseau, J., & Meunier, J. (2025). Mobility anomaly detection with intelligent video surveillance. In *L. Deligiannidis, F. G. Mohammadi, F. Shenavarmasouleh, S. Amirian, & H. R. Arabnia (Eds.), Image processing, computer vision, and pattern recognition and information and knowledge engineering* (CCIS vol. 2262, pp. 189–202). Springer. https://doi.org/10.1007/978-3-031-85933-5_13

[4] Gupta, R., Valencia, X. P. B., Goyal, L. M., & Kumar, J. (2025). *Ambient assisted living (AAL) technologies: Transitioning from healthcare 4.0 to healthcare 5.0*. CRC Press. https://doi.org/10.1201/9781003520184

[5] Wagner, J., Mazurek, P., & Morawski, R. Z. (2022). *Non-invasive monitoring of elderly persons: Systems based on impulse-radar sensors and depth sensors*. Springer. https://doi.org/10.1007/978-3-030-96009-4

[6] Ahmad, I., Asghar, Z., Kumar, T., Li, G., Manzoor, A., Mikhaylov, K., Shah, S. A., Höyhtyä, M., Reponen, J., & Huusko, J. (2022). Emerging technologies for next generation remote health care and assisted living. *IEEE Access*, 10, 56094–56132. https://doi.org/10.36227/techrxiv.19382876

[7] Islam, M. M., Tayan, O., Islam, M. R., Islam, M. S., Nooruddin, S., Kabir, M. N., & Islam, M. R. (2020). Deep learning based systems developed for fall detection: A review. *IEEE Access*, 8, 166117–166137. https://doi.org/10.1109/access.2020.3021943

[8] Roy, D., Komini, V., & Girdzijauskas, S. (2023). Classifying falls using out-of-distribution detection in human activity recognition. *AI Communications*, 36(4), 251–267. https://doi.org/10.3233/aic-220205

[9] Mulo, J., Liang, H., Qian, M., Biswas, M., Rawal, B., Guo, Y., & Yu, W. (2025). Navigating challenges and harnessing opportunities: Deep learning applications in internet of medical things. *Future Internet*, 17(3), 107. https://doi.org/10.3390/fi17030107

[10] Fernandez-Bermejo, J., Martinez-Del-Rincon, J., Dorado, J., Toro, X. D., Santofimia, M. J., & Lopez, J. C. (2024). Edge computing transformers for fall detection in older adults. *International Journal of Neural Systems*, 34(05), 2450026. https://doi.org/10.1142/s0129065724500266
[11] Núñez-Marcos, A., & Arganda-Carreras, I. (2024). Transformer-based fall detection in videos. *Engineering Applications of Artificial Intelligence*, 32(2), 101–115. https://doi.org/10.1016/j.engappai.2024.107937
[12] Wang, X., Ellul, J., & Azzopardi, G. (2020). Elderly fall detection systems: A literature survey. *Frontiers in Robotics and AI*, 7, 71. https://doi.org/10.3389/frobt.2020.00071
[13] Rahman, N. N., Mahi, A. B. S., Mistry, D., Al Masud, S. M. R., Saha, A. K., Rahman, R., & Islam, M. R. (2025). FallVision: A benchmark video dataset for fall detection. *Data in Brief*, 59, 111440. https://doi.org/10.1016/j.dib.2025.111440
[14] Wang, X. (2024). EGOFALLS: A visual-audio dataset and benchmark for fall detection using egocentric cameras. In *Pattern recognition. ICPR 2024. Lecture notes in computer science*. Springer. https://doi.org/10.1007/978-3-031-78166-7_16
[15] Luo, Z., Jia, S., Niu, H., Zhao, Y., Zeng, X., & Dong, G. (2024). Elderly fall detection algorithm based on improved YOLOv5s. *Information Technology and Control*, 53(2), 601–618. https://doi.org/10.5755/j01.itc.53.2.36336
[16] Kaur, N., Rani, S., & Kaur, S. (2024). Real-time video surveillance-based human fall detection system using hybrid haar cascade classifier. *Multimedia Tools and Applications*, 83, 71599–71617. https://doi.org/10.1007/s11042-024-18305-w
[17] Wang, Y., & Deng, T. (2024). Enhancing elderly care: Efficient and reliable real-time fall detection algorithm. *Digital Health*, 10, 1–11. https://doi.org/10.1177/20552076241233690
[18] Fula, V., & Moreno, P. (2024). Wrist-based fall detection: Towards generalization across datasets. *Sensors*, 24, 1679. https://doi.org/10.3390/s24051679
[19] Cao, Y., Guo, M., Sun, J., Chen, X., & Qiu, J. (2024). Fall detection based on LCNN and fusion model of weights using human skeleton and optical flow. *Signal, Image and Video Processing*, 18, 833–841. https://doi.org/10.1007/s11760-023-02776-9
[20] Abro, I. A., & Jalal, A. (2024). Multi-modal sensors fusion for fall detection and action recognition in indoor environment. In *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering* (ETECTE) (pp. 1–6). IEEE. https://doi.org/10.1109/etecte63967.2024.10823705
[21] Xu, Z., Liang, C., Wang, J., Ruan, L., Li, J., Dong, Y., Ding, W., & Song, J. (2024). LiFall: Passive indoor fall detection system based on illumination and visible light communication networks. In *Photonics & Electromagnetics Research Symposium* (pp. 1–10). https://doi.org/10.1109/piers62282.2024.10618192
[22] Piñeiro, M., Araya, D., Ruete, D., & Taramasco, C. (2024). Low-cost LIDAR-based monitoring system for fall detection. *IEEE Access*, 12, 72051–72065. https://doi.org/10.1109/access.2024.3401651
[23] Yang, X., Zhang, S., Ji, W., Song, Y., He, L., & Xue, H. (2024). SMA-GCN: A fall detection method based on spatio-temporal relationship. *Multimedia Systems*, 30, 90–105. https://doi.org/10.1007/s00530-024-01293-0
[24] Ha, T. V., Nguyen, H. M., Thanh, S. H., & Nguyen, B. T. (2024). Fall detection using mixtures of convolutional neural networks. *Multimedia Tools and Applications*, 83, 18091–18118. https://doi.org/10.1007/s11042-023-16214-y

[25] Gaya-Morey, F. X., Manresa-Yee, C., & Buades-Rubio, J. M. (2024). Deep learning for computer vision-based activity recognition and fall detection of the elderly: A systematic review. *Applied Intelligence*, 54, 8983–9000. https://doi.org/10.1007/s10489-024-05645-1
[26] Jiang, Z., Al-Qaness, M. A. A., Al-Alimi, D., Ewees, A. A., Abd Elaziz, M., Dahou, A., & Helmi, A. M. (2024). Fall detection systems for internet of medical things based on wearable sensors: A review. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2024.3421336
[27] Tang, J., He, B., Xu, J., Tan, T., Wang, Z., Zhou, Y., & Jiang, S. (2024). Synthetic IMU datasets and protocols can simplify fall detection experiments and optimize sensor configuration. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32, 1233–1245. https://doi.org/10.1109/tnsre.2024.3370396
[28] Ursul, I., & Pereymybida, A. (2025). Source code for sensor-based fall detection dataset processing and analysis [Software]. GitHub. https://github.com/ivanursul/fall-detection-phd
[29] Huang, X., Li, X., Yuan, L., Jiang, Z., Jin, H., Wu, W., Cai, R., Zheng, M., & Bai, H. (2025). SDES-YOLO: A high-precision and lightweight model for fall detection in complex environments. *Scientific Reports*, 15, 2026. https://doi.org/10.1038/s41598-025-86593-9
[30] Ren, H., & Lan, P. (2025). BMR-YOLO: A deep learning approach for fall detection in complex environments. *PLOS One*, 20(11), e0335992. https://doi.org/10.1371/journal.pone.0335992

ВИЯВЛЕННЯ ПАДІНЬ НА ОСНОВІ ЗОРОВОГО ТРАНСФОРМЕРА: ПРОСТОРОВО-ЧАСОВИЙ МЕХАНІЗМ УВАГИ ДЛЯ НАДІЙНОГО АНАЛІЗУ ВІДЕО

Іван Урсул*, Андрій Переймибіда
Кафедра прикладної математики
Львівський національний університет імені Івана Франка,
вул. Університетська, 1, Львів, 79000, Україна
**Відповідальний автор e-mail: ivan.ursul@lnu.edu.ua*

АНОТАЦІЯ

Вступ. Виявлення падінь є критично важливим завданням у сфері охорони здоров'я та догляду за людьми похилого віку, оскільки затримка реагування часто призводить до серйозних травм. Зі старінням населення кількість госпіталізацій, пов'язаних із падіннями, зростає, підвищуючи вимоги до автоматизованого моніторингу. Підходи на основі переносних пристроїв або традиційних класифікаторів генерують хибні спрацювання та демонструють обмежену адаптивність. Системи на основі відео забезпечують ширше охоплення, але потребують моделей, здатних фіксувати зміни пози та руху без ручного створення ознак.

Матеріали та методи. Для моделювання просторово-часових закономірностей рухів людини було застосовано архітектуру зорового трансформера. Відеокадри розділялися на фрагменти та проєктувалися у вбудовані токени, що дозволило механізму багатоканальної самоуваги відстежувати зміни пози між кадрами для

формування ознак з метою прогнозування падінь. Навчання проводилося на кількох відкритих наборах даних із різноманітним фоном та різними типами статури суб'єктів. Модель порівнювалася з базовими методами логістичної регресії та згорткових нейронних мереж, навченими на ідентичних вибірках даних.

Результати. Зоровий трансформер досяг точності 99,1% на основному наборі даних та 97,9% на наборі UR Fall Detection, перевершивши базові методи логістичної регресії, згорткової нейронної мережі та довгої короткочасної пам’яті. Модель зберігала вищі показники влучності та повноти в приміщенні та на відкритому повітрі, зменшивши частоту хибних спрацювань. Стабільну роботу спостерігали за складних умов, включаючи швидкий рух та змінне освітлення, що підтвердило переваги у стійкості. Перехресна оцінка наборів даних показала ефективне перенесення просторово-часових представлень на невідомі умови запису.

Висновки. Зорові трансформери пропонують ефективний підхід до виявлення падінь у реальному часі в клінічних та домашніх умовах. Здатність фіксувати просторово-часові закономірності руху за допомогою самоуваги, без ручного створення ознак, сприяє ширшому впровадженню в інтелектуальні системи відеоспостереження. Подальша робота буде зосереджена на оптимізації периферійних пристроїв та інтеграції мультимодальних даних.

Ключові слова: виявлення падінь, зоровий трансформер, самоувага, аналіз руху людини, класифікація відео, догляд за людьми похилого віку.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
23 February, 2026	16 March, 2026	23 March, 2026	30 March, 2026