sensors

MDPI

Article

Robust 3D Skeletal Joint Fall Detection in Occluded and Rotated Views Using Data Augmentation and Inference–Time Aggregation

Maryem Zobi [1,*], Lorenzo Bolzani [2], Youness Tabii [1] and Rachid Oulad Haj Thami [1]

[1] ADMIR Laboratory, National School of Computer Science and Systems Analysis, Mohammed V University, Rabat 10000, Morocco; youness.tabii@ensias.um5.ac.ma (Y.T.); rachid.ouladhajthami@ensias.um5.ac.ma (R.O.H.T.)
[2] flAIght, Viale dell'Innovazione 13, 20126 Milan, Italy; lorenzo.bolzani@flaight.it
* Correspondence: maryem_zobi@um5.ac.ma

Abstract

Fall detection systems are a critical application of human pose estimation, frequently struggle with achieving real-world robustness due to their reliance on domain-specific datasets and a limited capacity for generalization to novel conditions. Models trained on controlled, canonical camera views often fail when subjects are viewed from new perspectives or are partially occluded, resulting in missed detections or false positives. This study tackles these limitations by proposing the Viewpoint Invariant Robust Aggregation Graph Convolutional Network (VIRA-GCN), an adaptation of the Richly Activated GCN for fall detection. The VIRA-GCN introduces a novel dual-strategy solution: a synthetic viewpoint generation process to augment training data and an efficient inference-time aggregation method to form consensus-based predictions. We demonstrate that augmenting the Le2i dataset with simulated rotations and occlusions allows a standard pose estimation model to achieve a significant increase in its fall detection capabilities. The VIRA-GCN achieved 99.81% accuracy on the Le2i dataset, confirming its enhanced robustness. Furthermore, the model is suitable for low-resource deployment, utilizing only 4.06 M parameters and achieving a real-time inference latency of 7.50 ms. This work presents a practical and efficient solution for developing a single-camera fall detection system robust to viewpoint variations, and introduces a reusable mapping function to convert Kinect data to the MMPose format, ensuring consistent comparison with state-of-the-art models.

Keywords: fall detection; 3D skeletal joints; data augmentation; graph convolutional networks; occlusion; rotated views; MMPose; Kinect; VIRA-GCN

Academic Editors: Yunus Celik and Gill Barry

Received: 27 September 2025
Revised: 26 October 2025
Accepted: 30 October 2025
Published: 6 November 2025

Citation: Zobi, M.; Bolzani, L.; Tabii, Y.; Oulad Haj Thami, R. Robust 3D Skeletal Joint Fall Detection in Occluded and Rotated Views Using Data Augmentation and Inference–Time Aggregation. *Sensors* **2025**, *25*, 6783. https://doi.org/10.3390/s25216783

1. Introduction

The global population is undergoing a significant demographic shift, with the number of individuals aged 65 and over projected to more than double by 2050, increasing from 9.3% in 2020 to approximately 16% of the world's population, according to the United Nations (UN) [1]. This trend is particularly pronounced in countries like Japan, where 29% of the population is expected to be over 65, underscoring the growing public health challenge of falls among the elderly. Falls are a critical concern, as the World Health Organization (WHO) [2] identifies them as the second leading cause of unintentional injury deaths globally, especially among adults over 60. With an estimated 684,000 annual fatalities, a large proportion of which occur in low- and middle-income countries, falls pose a substantial threat to public health. These events, defined as unplanned descents to the

ground, can lead to severe injuries, loss of independence, and even death, especially if assistance is delayed. The high incidence of falls also places a significant economic burden on healthcare systems, with direct medical costs in the United States alone estimated at $50 billion annually. This demonstrates the critical need for effective fall detection systems to ensure timely intervention and improve outcomes for the elderly.

However, the reliability of these systems is often undermined by the challenges of real-world environments [3]. To be truly effective, a fall detection system must function accurately despite the variability of camera positions and the presence of occlusions, where a subject is partially obscured by furniture or other objects. This requires sophisticated systems that can robustly interpret visual data in diverse and cluttered settings.

This challenge is illustrated by the difficulty a standard model has in correctly classifying a fall when the camera is positioned at an unusual height or angle [4–6]. A model trained on a controlled, side-view dataset may fail to recognize a fall when it is viewed from a head-on or elevated perspective. Overcoming this requires either training on a massive, diverse dataset or developing models that can effectively generalize from limited data.

Our work addresses this generalization problem by focusing on improving the robustness of fall detection models to variations in camera viewpoint and occlusions. We present a novel methodology that operates on existing skeletal data, making it applicable to a wide range of pose estimation architectures used for fall detection.

Our key contributions are summarized as follows:

- We introduce a "Synthetic Viewpoint Generation" data augmentation pipeline that efficiently generates realistic occluded and rotated skeletal sequences specifically for training fall detection models, significantly enhancing robustness to camera perspective variation and stability.
- We propose a simple but effective inference-time aggregation method that leverages multiple rotated views of a single input to achieve a consensus-based final fall prediction, further enhancing model performance.
- We propose a per-joint affine transformation to convert the Kinect skeleton format to the MMPose format, enabling the model to be trained on standard Kinect datasets.
- We conduct a comprehensive ablation study to systematically quantify the impact of each of our contributions on the accuracy of fall detection.

The remainder of this paper is organized as follows: Section 2 summarizes existing research and identifies related works. Section 3 details the dataset preparation process and the proposed fall detection system. Section 4 presents our evaluation, including a comparison with a state-of-the-art approach. The discussion and key findings are presented in Section 5, and the paper concludes with a summary and directions for future research in Section 6.

2. Related Work

Human fall detection systems can be categorized into two main types: those based on wearable sensors and those that employ machine vision-based techniques. This section reviews both paradigms, with a specific focus on state-of-the-art machine learning and deep learning methods used for fall detection.

Wearable sensor systems typically use devices such as gyroscopes, accelerometers, pressure sensors, tilt switches, and magnetometers to capture various signals. From these signals, parameters like angles, distances, and statistical measures are directly extracted. Wearable systems for fall detection are inherently data-set-dependent for training and can be deployable across various environments. In an early study [7], researchers developed an accelerometer-based system. It first checks if the signal vector magnitude (SVMA) exceeds a given threshold. If it does, a pulse sensor and trunk angle are used for verification. This

system, which automatically contacts emergency services upon fall confirmation, achieved a high accuracy of 97.5%. Another approach was seen in study [8], where a hip-worn system that used a variety of sensors with a logistic regression classifier achieved 100% accuracy in detecting falls. However, this study did not detail the variability of fall actions. A different approach [9] used a highly optimized machine learning framework with acceleration and angular velocity data for fall detection, achieving 100% accuracy using the QSVM and EBT algorithms.

Accelerometers have proven to be highly efficient in fall detection and were widely used in most systems developed. However, they also have significant disadvantages. A major issue is user compliance, as individuals may find the devices uncomfortable or forget to wear them, which compromises their effectiveness. The sensors themselves also have technical limitations, most notably a tendency for false alarms, where normal activities are mistaken for falls. Furthermore, their drawbacks include limited battery life, which can be particularly problematic for elderly people. Other significant issues involve the high cost of some devices and the associated privacy concerns about the sensitive data they collect.

In contrast, vision-based systems have been applied in various fields, including human fall detection. Vision-based systems offer a non-intrusive alternative that does not require users to wear physical devices. These systems analyze RGB and depth images from real-time video streams. This visual data is then processed using advanced deep learning algorithms to enable accurate, real-time monitoring and event classification.

The core input for many of these systems is the human skeleton, represented by a set of 2D or 3D keypoints. 3D coordinates can be obtained from low-cost depth sensors like Microsoft Kinect (Manufactured by Microsoft Corporation, Headquarters: Redmond, Washington, U.S.), or extracted from 2D images using sophisticated pose estimation algorithms powered by Convolutional Neural Networks (CNNs), such as Openpose [10] (Developed by researchers at Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.) and MMPose [11], an open-source toolbox developed by the OpenMMLab project (Headquarters: China). Early approaches for fall detection used machine learning algorithms with skeleton joint data. For example, this study [12] introduced a feature-based method that first divides the human skeleton into five parts before applying a Support Vector Machine (SVM), achieving 93.56% accuracy on the TST fall detection dataset v2. In another study, a depth camera-based system [13] analyzed the 3D trajectory of a person's head with an SVM classifier to detect falls while checking for post-fall recovery to prevent false alarms, reporting no false positives in tests. Another work [14] used a four-camera setup, applying a median filter for background removal and morphological techniques to isolate the area of an individual before using an SVM classifier with features like head velocity, center of gravity speed and the proportion of the individual to the picture's size. This method reported a sensitivity of 77.8% and a PPV of 36.66%. Deep learning methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also been widely used. It has been shown that using multiple cameras improves performance over a single kinematic camera.

One notable approach [15] demonstrated this by combining a 3D CNN with an LSTM-based mechanism to identify crucial motion features, achieving 100% accuracy on a sports dataset and a fall detection benchmark. Similarly, another system utilized RF signals and a state machine-governed CNN, demonstrating strong performance with a recall of 94% and a precision of 92% in tests involving over 140 people [16]. A third study proposed a system that used Mask R-CNN to extract moving objects and an attention-guided Bi-directional LSTM for classification, achieving 96.70% accuracy on the UR-Fall detection dataset [17]. Furthermore, a pose-based system utilized OpenPose for precise pose estimation and a 1D-CNN to classify motion-based features extracted with a short-time Fourier transform (STFT),

achieving high accuracies across multiple datasets: 99% on the UR Fall Dataset, 91% on the MCFD dataset, and 98% on the NTU RGB+D Dataset [18]. Finally, a separate system was developed combining traditional algorithms with a 1DCNN model, demonstrating high accuracy and robustness on the NTU RGB+D dataset [19]. Despite their success, initial deep learning methods for processing skeleton data, such as RNNs and CNNs, have limitations. They require pre-processing steps to re-organize 3D joint coordinates into 1D sequences or 2D pseudo-images. RNNs struggle to model the non-sequential, structural relationships between non-adjacent joints, while CNNs are constrained by their fixed-grid nature and fail to capture the intrinsic, non-Euclidean topology of the human skeleton.These limitations have driven attention toward methods based on Graph Convolutional Networks (GCNs), which are naturally suited for graph-structured data. Recent research has increasingly adopted GCNs for action recognition because these models treat the human body as a natural graph structure. This representation allows GCNs, and their spatio-temporal variants (ST-GCNs), to effectively model and capture the complex spatio-temporal dependencies between all joints simultaneously. This transition has led to state-of-the-art results in applied fields. For instance, an ST-GCN model model in [20] has been successfully employed for fall detection using skeletal data from a Kinect v2 sensor and transfer learning achieved 100% accuracy on the TST Fall v2 dataset and 97.33% on the Fallfree dataset. In related work, a GCN-based model developed by [21] achieved approximately 99% accuracy on the ImViA RU-Fall detection dataset. In another study [22], a novel three-stream system with an ST-GCN model demonstrated superior performance, achieving accuracies of 99.68%, 99.97%, 99.47%, and 98.97% on the ImViA, Fall-UP, FU-Kinect, and UR-Fall datasets, respectively. Deep learning methods are powerful for classification tasks due to their feature extraction capabilities, but they are highly contingent on the availability of large, annotated datasets. They also often exhibit vulnerability to environmental factors like occlusion and are not robust to variations in camera viewpoint. Although multi-camera setups [6,23] have been used to enhance viewpoint robustness, they require complex fusion layers, leading to significant computational demands and complicated deployment logistics [24].

To address these challenges, our work proposes the VIRA-GCN model, an adaptation of the Richly Activated Graph Convolutional Network (RA-GCNv2) for fall detection using skeleton data derived from the MMpose framework. Our approach employs a lightweight, inference-time aggregation method that focuses on robustness to real-world occlusions. Crucially, this system operates on a single input stream by generating synthetic views and aggregating their predictions, thereby eliminating the need for multiple physical cameras. This mechanism results in a fall detection system that is not only highly accurate and efficient but also robust and generalizable.

3. Methodology

This section outlines the entire methodology developed for robust human fall detection. The approach is organized around four critical stages: dataset preparation, feature extraction, model definition, and data transformation. The process commences with the dataset preparation stage, which defines the selected data for the experimentation and evaluation of the proposed method and includes the necessary augmentation strategies. Feature extraction then details the vital process of feature extraction and justifies the selection of the MMPose framework for obtaining robust 3D skeletal data. This is followed by a detailed description of the proposed VIRA-GCN Model Architecture. Finally, the section concludes with an explanation of the specialized Kinect to MMPose Data Transformation pipeline.

3.1. Dataset Preparation

A set of publicly available datasets containing both fall and non-fall actions was first identified for this study. The primary selection criterion was the availability of RGB video to enable the extraction of 3D skeletal data using the MMPose framework [11].

As summarized in Table 1, several datasets meet these requirements. The Le2i Fall Dataset was the first selected and utilized for model training and validation. This selection was motivated by its provision of the essential RGB video sequences required for 3D skeletal data extraction and its representation of complex and realistic fall scenarios, which are crucial for validating the model's robustness to real-world conditions. The NTU RGB+D dataset was then specifically chosen for a technical validation purpose: facilitating a fair comparison between data from Kinect sensors and data processed by MMPose. Since this dataset provides both types of data, it allows for the validation of the transformation pipeline.

Table 1. A Comparative Analysis of Video Modalities and Feature Sets in Fall Detection Datasets.

Dataset	RGB Videos	Kinect Data	Viewpoint Variation	Recording Environment
Le2i Fall Dataset [25]	✓		✓	Indoor Home
UR Fall Detection [26]	✓	✓		Indoor Lab
SDUFall [27]	✓	✓		Indoor Home
SisFall [28]				Indoor Home
NTU RGB+D [24]		✓	✓	Indoor Lab
TST Fall Detection [29]	✓	✓		Indoor Lab
KFall Dataset [30]		✓		Indoor Home
MobiAct [31]				Indoor Lab
FallAllD [32]	✓			Uncontrolled
FARSEEING [33]				Uncontrolled

3.1.1. Le2i Dataset

The Le2i dataset [25] contains video sequences of various fall events and normal daily activities. The dataset was recorded using a single, fixed camera setup across various indoor environments, including a home, coffee shop, lecture hall, and office. The "fall" videos typically include a sequence of consecutive actions: standing, the actual fall, and remaining fallen, often preceded by sitting, while the "normal" category features common actions such as walking, standing, squatting, and sitting. To ensure diversity, actors wore different clothing and simulated a range of activities; however, only one person is displayed in the video. All videos have a resolution of 320 × 240 pixels, a frame rate of 25 frames/s, and feature a fixed, simple background contrasted with complex image textures.

To ensure a consistent and fair evaluation, we meticulously addressed the issue of potential data leakage by first splitting the original Le2i videos into distinct training and testing subsets. The various datasets for our experiments were then created by processing these pre-defined subsets:

1. Baseline Dataset:This dataset was established by extracting the initial 3D joint coordinates from the RGB videos in both the training and testing subsets, utilizing the MMPose framework, as detailed in Section 3.2. This corpus served as the essential performance baseline for subsequent comparative analysis.
2. Occluded Dataset: This dataset was derived from the baseline by applying various types of random occlusions as illustrated in Figure 1. This method simulates realistic scenarios where a subject might be partially obscured before or during a fall. We utilized distinct occlusion modalities, including random black boxes over sequential

frames and temporal occlusions, where all frames are rendered black for a short duration. From a single source video, 10 unique occluded variations were generated. The joint predictions were subsequently re-extracted from these altered frames, thereby yielding intentionally degraded coordinate data for robustness training. The resulting corpus was subjected to a fixed, random split for training and testing evaluation.

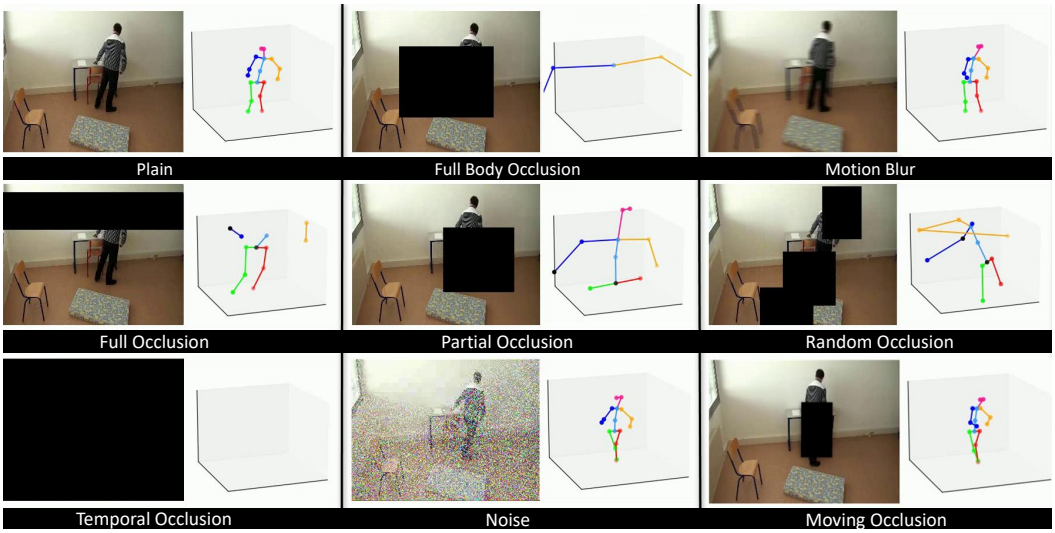

Figure 1. A raw image from the Le2i dataset showing a subject in an office environment. The image illustrates various types of simulated occlusions applied to the dataset to improve model robustness.

3. Rotated Dataset: This dataset was generated to enhance the model's robustness against camera viewpoint variations. This was achieved through Synthetic Viewpoint Generation, a technique that programmatically augments each skeletal sequence by creating multiple synthetic views via 3D rotations.
Let a skeleton sequence be represented as:

$$\text{Skel} = \{p_{t,v} \in \mathbb{R}^3 | t = 1, ..., T; v = 1, ..., V\}$$

where $p_{t,v} = (x_{t,v}, y_{t,v}, z_{t,v})^\top$ denotes the 3D coordinates of joint v at frame t, with $T = 300$ frames and $V = 17$ joints.
We applied two-axis rotations to each sequence: a vertical (z-axis) rotation and a horizontal (x-axis) tilt. The rotation angles were defined as:

$$\Theta_z = \{0^\circ, 40^\circ, 80^\circ, 120^\circ, 160^\circ, 200^\circ, 240^\circ, 280^\circ, 320^\circ\}, \quad \Theta_x = \{0^\circ, 15^\circ, 25^\circ\}.$$

For each pair of angles $(\theta_x, \theta_z) \in \Theta_x \times \Theta_z$, we constructed a combined rotation matrix:

$$R(\theta_x, \theta_z) = R_x(\theta_x) R_z(\theta_z)$$

where

$$R_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}, \quad R_z(\theta_z) = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Each joint position was then transformed as:

$$p'_{t,v} = R(\theta_x, \theta_z) p_{t,v}.$$

This procedure generated:

$$N_{views} = |\Theta_x| \times |\Theta_z| = 3 \times 9 = 27$$

synthetic viewpoints per original skeleton sequence, thereby enriching the dataset and enhancing the model's ability to generalize across different camera perspectives. A significant challenge we faced was the severe class imbalance within the dataset, as the number of "fall" instances was disproportionately large compared to actual "no fall" events. To address this, we applied a data augmentation strategy to the minority class. By generating additional data for "no fall" events, we created a more balanced training set, which was critical for building a robust model that performs accurately on both fall and non-fall scenarios.

3.1.2. NTU RGB+D Dataset

To enable a direct and equitable comparison of our model with state-of-the-art results from literature that utilize Kinect data, we developed a per-joint affine transformation to convert the Kinect skeleton structure into the MMPose format. We utilized the NTU RGB+D dataset. This dataset is a widely recognized benchmark for human activity recognition, establishing an ideal platform to demonstrate our model's effectiveness beyond the domain-specific constraints of Le2i.

For training and evaluation, we selected the NTU RGB+D dataset's 'falling down' class (labeled 'A043') and a set of 'no fall' actions. The 'no fall' samples were systematically chosen from all daily action classes (A001–A042, A044–A102), including examples such as A001 (drink water), A002 (eat meal), A008 (sit down), and A024 (kicking something). The total number of selected samples was 1028, and the classes were balanced to ensure fair training.

3.2. Skeletal Feature Extraction

Effective feature extraction forms the foundational step of any vision-based action recognition system. For robust fall detection, this process requires the reliable acquisition of 3D skeletal data, as the third dimension (depth, z) is essential for creating viewpoint-invariant metrics such as true body height and vertical drop velocity. This capability is critical for training a model to generalize effectively against the effects of camera angle variation and partial occlusion.

Accurate implementation of this step necessitates the selection of an optimal Human Pose Estimation framework. We utilized MMPose, an open-source toolbox developed by the OpenMMLab project [11], to obtain the raw skeletal data. The justification for this selection, as summarized in the comprehensive analysis presented in Table 2, highlights MMPose's superior alignment with our research requirements. MMPose was chosen because it facilitates the extraction of robust and accurate 3D skeletal representations from the RGB videos. This rich representation mitigates the viewpoint-dependent limitations inherent in 2D data, which is non-negotiable for training the model on synthetically augmented (occluded and rotated) inputs.

Table 2. A Comparative Analysis of State-of-the-Art Human Pose Estimation Frameworks.

Framework	Year	Target Scope	Nb Extracted Joints	Accuracy AP (%)	Cost (GFLOPs)	Efficiency
OpenPose [10]	2017	Multi-person 2D/3D Pose	15–25 up to 135	61.8	160	Slowest (GPU-Intensive)
AlphaPose [34]	2017	High-Accuracy 2D/3D Pose	17 up to 136	73.3	5.91 → 15.99	Fast (Requires GPU)
MediaPipe [35]	2019	Real-time Mobile Pipelines	33	-	29.3 → 35.4	Extremely Fast on Mobile/CPU
MMPose [11]	2020	SOTA Toolkit Benchmark	17 up to 133	75.8	1.93	90+ FPS CPU/ 430+ FPS GPU

Strategically, MMPose provides the optimal balance of performance metrics. While competing frameworks like OpenPose incur a significantly high computational cost (160 GFLOPs), MMPose remains highly efficient, making it suitable for deployment with the lightweight VIRA-GCN architecture.

3.3. Proposed Model Architecture

This study introduces a Viewpoint Invariant Robust Aggregation Graph Convolutional Network (VIRA-GCN). This model is engineered to address the critical gap of real-world robustness by adapting the Richly Activated Graph Convolutional Network (RA-GCNv2) [36]. Our primary objective is to develop a single-camera solution that is robust to perspective variations and partial occlusions.

The VIRA-GCN architecture retains the foundational analytic strengths of its predecessor. Specifically, it leverages the Graph Convolutional Network (GCN) framework to model the human body as a graph for extracting complex spatio-temporal relationships between joints, and it utilizes the recurrent attention mechanism to dynamically weight and emphasize the most temporally relevant frames indicative of a fall event. These core capabilities remain integral to the proposed methodology.

Our adapted pipeline, illustrated in Figure 2, incorporates three principal modifications to the original RA-GCNv2 architecture to enhance its practical utility and focus:

- Input Dimensionality Reduction: We adapted the input layer to process a more concise 17-joint skeleton, which aligns with modern pose estimation frameworks (e.g., MMPose). The original model was designed for the larger 25-joint skeleton from Kinect.
- Realistic Data Augmentation for Robustness: The original study employed an artificial occlusion methodology that is fundamentally limited in its capacity to represent real-world operating scenarios. This method directly altered the predicted joint values, which may not reflect real outputs from a skeletal prediction model for an occluded frame. To overcome this constraint and enhance model robustness, we implemented a comprehensive preprocessing pipeline incorporating two forms of realistic data augmentation. This strategy includes Realistic Occlusion, which simulates the partial visibility of body parts due to environmental factors, and Viewpoint Rotation, which simulates the diversity of camera perspectives encountered in naturalistic settings. As detailed in the preceding section, this deliberate augmentation process ensures the resulting model can accurately identify falls across a representative range of visual conditions, maintaining performance even when full body visibility is compromised.

- Task Re-purposing for Binary Classification: We adapted the base RA-GCNv2 architecture, transforming its original function of general action recognition into a specific binary classification task. This involved re-purposing the final layers and loss functions to distinctly discriminate between fall and non-fall events. This crucial architectural modification compels the model to concentrate its learning on the subtle, critical kinematic indicators of a fall, such as rapid changes in vertical displacement or sudden, precipitous collapses, thereby optimizing its sensitivity for the target safety application.

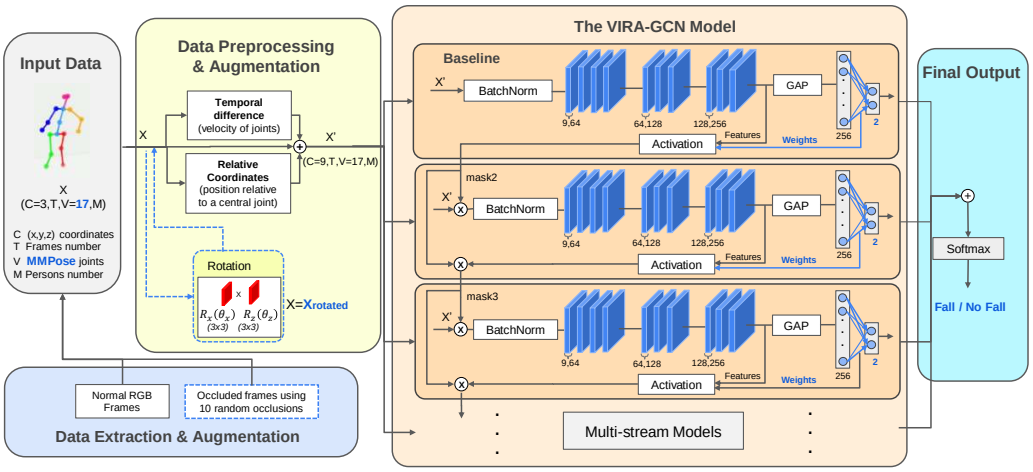

Figure 2. This diagram illustrates our proposed VIRA-GCN architecture for Fall Detection. We enhance its robustness to real-world conditions by introducing a novel preprocessing pipeline that augments the 17-joint input data with realistic occlusion and rotations (adapted from the original RA-GCNv2 work by [36]).

3.3.1. Training Pipeline Optimization

We identified a major efficiency bottleneck within the initial training framework, evidenced by GPU utilization persistently restricted to approximately 30%. This computational inefficiency was attributable to the sequential execution of the data augmentation operations, specifically skeleton rotation, which occurred synchronously within the main training loop via micro-batches. Our optimization strategy resolved this limitation by refactoring the augmentation logic to be executed entirely within the dataloader. This architectural modification enabled the rotation process to be performed in parallel and decoupled from the primary training thread, ensuring the model received full batches of pre-transformed samples. This procedural change yielded a significant enhancement in throughput, substantially increasing GPU utilization, allowing for the stable use of a larger batch size of 64, and consequently accelerating the training process while maintaining a modest memory footprint of only 8 Gb. Furthermore, we integrated 'torch.autocast' to leverage mixed 16/32-bit precision, providing additional optimization of memory consumption and calculation speed.

3.3.2. Inference-Time Aggregation

To enhance the robustness of the trained model without incurring a costly retraining process, we developed a consensus-based Inference-time Aggregation method. This approach leverages the model's pre-trained viewpoint invariance to derive a more reliable final classification from a single input stream.

The implementation follows a defined sequential process. Firstly, multiple synthetic viewpoints of the input skeleton sequence are generated by applying a set of fixed rotations (e.g., 15°, 25°). Subsequently, each of these rotated sequences is independently processed

by the trained VIRA-GCN model. The model's prior training with rotated skeletons from multiple angles ensures it is able to handle these transformed inputs effectively. Following prediction, the model outputs an individual probability for a fall for each sequence. Finally, the probabilities derived from all rotated sequences are combined using a simple probability averaging schema to produce the final, statistically more robust fall classification.

A transformation from kinect to MMPose was proposed to ensure that the input skeleton to our model is consistent in term of joint topology and spatial alignmenet, regardless of the original data source. The core of our methodology is a per-joint affine transformation, which aligns the Kinect skeleton to the MMPose structure.

Unlike a single transformation for the entire body, our approach calculates a unique transformation matrix and translation vector for each joint. This is a more general approach that accounts for rotation, scaling, and shearing, as well as translation and is able to correct for the different joints schema.

3.4. Kinect to MMPose Data Transformation

The transformation for a single joint j, in 3D Euclidiean space, can be formally expressed as:

$$P_{transformed,j} = P_{kinect,j} \cdot A_j^\top + b_j$$

Here, $P_{transformed,j}$ is the 3D coordinate vector of the transformed Kinect joint, and $P_{kinect,j}$ represents the original Kinect joint's position. Where A_j is the unique 3×3 transformation matrix for joint j and b_j is the unique 3×1 translation vector for joint j. The alignment achieved by this per-joint affine transformation is visually represented in Figure 3.

The transformation parameters (A_j and b_j) are derived from a training dataset of 30 randomly selected frames. These parameters are calculated for each joint by solving a least-squares problem, which finds the optimal transformation that minimizes the distance between the training points of the Kinect and MMPose data. The performance of our model is quantitatively evaluated using the Mean Squared Error (MSE). The MSE for a single joint j is calculated as the average squared Euclidean distance between the transformed and target joints across the training set.

The Euclidean distance, along with its squared form, the Mean Squared Error (MSE), plays a central role in the transformation methodology as it provides both a physically meaningful error measure and the mathematical properties necessary for optimization. For 3D pose data, the Euclidean distance is the most direct and geometrically intuitive metric, quantifying the actual shortest physical displacement between the transformed joint and the ground-truth target joint. Crucially, the squared distance results in a smooth and differentiable MSE loss function, which is essential for the effective application of least-squares optimization to analytically determine the optimal transformation parameters (A_j and b_j). Moreover, the MSE metric is the established standard metric in the field, often referred to as the Mean Per Joint Position Error (MPJPE), ensuring that the quantitative evaluation of the model's positional accuracy is comparable with existing research.

$$MSE_j = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \|P_{mmpose,j,i} - P_{transformed,j,i}\|_2^2$$

The overall model performance is then measured by the average MSE across all joints:

$$\text{Overall MSE} = \frac{1}{N_{joints}} \sum_{j=1}^{N_{joints}} MSE_j$$

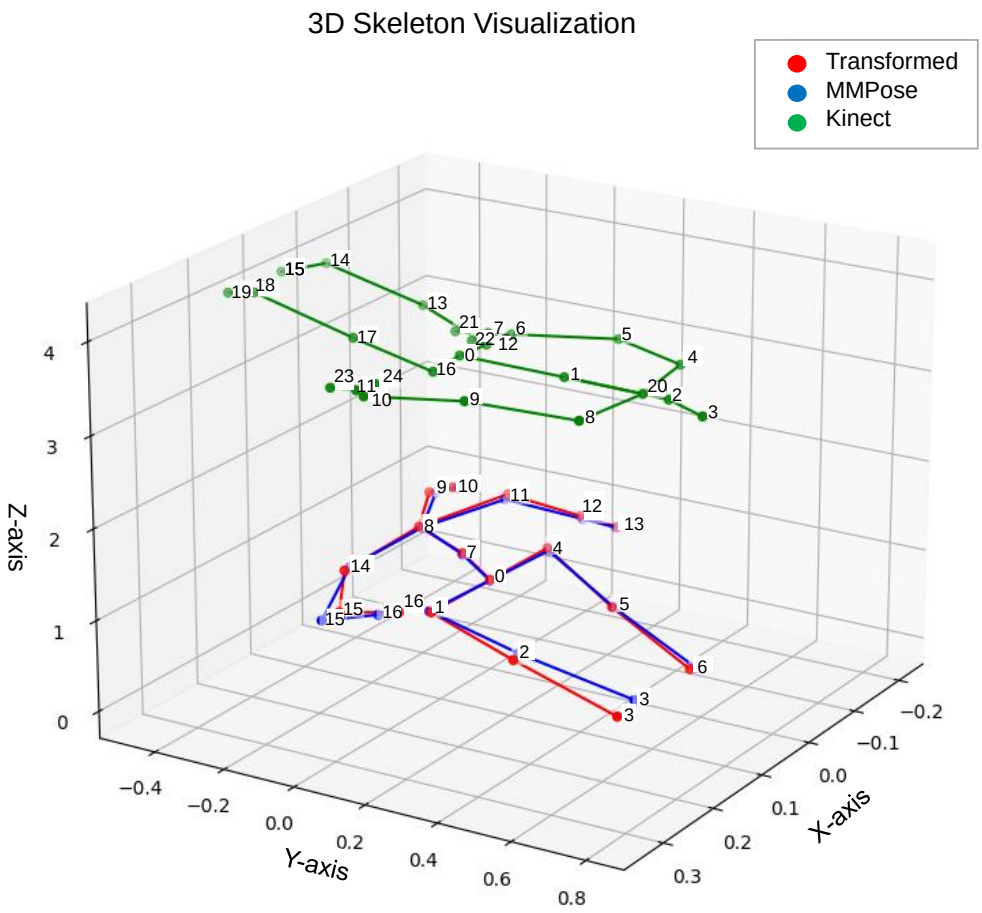

Figure 3. 3D skeleton comparison showing the alignment achieved by the per-joint affine transformation. The original Kinect skeleton (green) is transformed to align with the target MMPose skeleton (blue), resulting in the transformed skeleton (red). The numbers displayed next to the joints represent the specific index (ID) of the joint within the respective Kinect (0-24) or MMPose (0-16) skeleton format. The mapping between these indices is detailed in the X-axis labels of Figure 4.

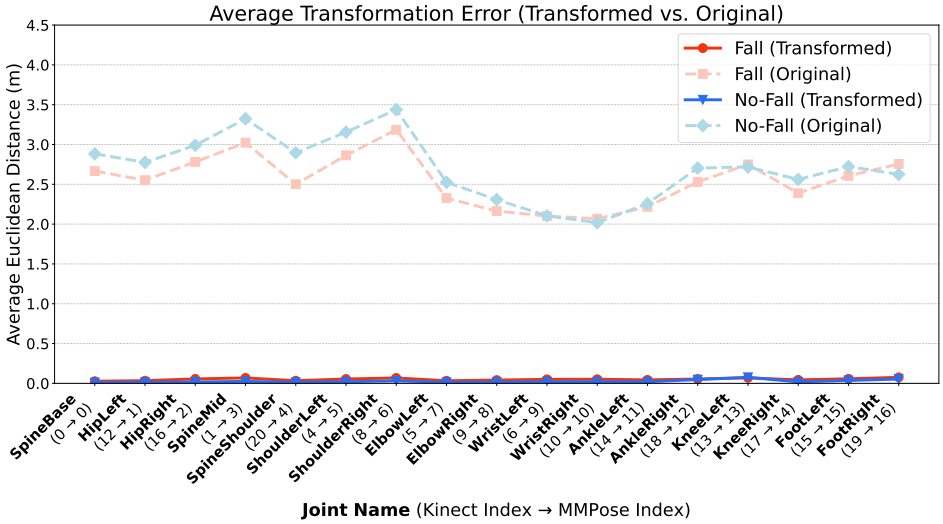

Figure 4. Average Euclidean Distance Error per Joint. The bar chart compares the mean Euclidean distance error for each of the 17 joints between the ground-truth MMPose skeleton and the original Kinect skeleton versus the transformed Kinect skeleton.

4. Experiments and Results

This section systematically presents the empirical evaluation and quantitative results of the proposed fall detection framework. Our evaluation was rigorously structured as a comprehensive ablation study, designed to systematically quantify the isolated impact of data augmentation and, critically, to compare the generalization capabilities of our models when trained on single-domain versus combined-domain data.

4.1. Experimental Design

The rigorous assessment of the VIRA-GCN model's generalization capabilities and robustness necessitated a comprehensive experimental design, centred on controlling for both data composition and stochastic variability in model initialization. To ensure the statistical reliability of the performance metrics, all model configurations underwent a 5-fold cross-validation procedure, which was independently repeated across 5 distinct random seeds. This methodology yielded an aggregate of 25 independent training and evaluation trials for each configuration, allowing for a robust estimation of mean performance and variance.

4.1.1. Evaluation Datasets

The performance and cross-domain generalization capabilities of the models were rigorously assessed against a comprehensive suite of five distinct evaluation datasets, each serving a specific analytical purpose. All evaluations were conducted on the dedicated, non-overlapping held-out test subsets of the respective data sources. This methodical approach allowed for the isolation of factors such as data augmentation efficacy, robustness to corruption, and domain shift resilience. The datasets employed are detailed as follows:

- Plain (Le2i): This dataset comprises the raw, unaugmented 3D skeleton joint sequences extracted directly from the Le2i video corpus using the MMPose. Evaluation on this set established the baseline performance of the models on the native, clean data of the source domain.
- Occluded (Le2i): Consisting of skeleton joints derived from Le2i videos that were intentionally subjected to realistic visual occlusions, this set utilizes the MMPose format. Performance here directly quantifies the model's inherent robustness against common real-world data corruptions, specifically in the form of partial data visibility.
- Rotated (Le2i): This set is composed of 3D skeleton sequences from the Le2i test set that were synthetically transformed via viewpoint rotations. Adhering to the MMPose format, this evaluation measures the model's invariance to synthetic changes in camera perspective within the source domain.
- Mntuplain (NTU MMPose): This dataset features 3D skeleton joints extracted from the original NTU RGB+D videos using the MMPose framework. Evaluation on this set provides a direct measure of the model's generalization performance when confronted with the standard skeleton format of the target domain.
- Kntuplain (NTU Kinect Transformed): This is a critical cross-domain evaluation set comprising 3D skeleton joints that were transformed from the original NTU RGB+D Kinect sensor data into the MMPose Transformed format using the proposed affine transformation pipeline. Performance on this dataset is essential for assessing the efficacy of the proposed transformation methodology in aligning the characteristics of the NTU Kinect data with the model's training space, providing insight into robustness against sensor heterogeneity.

4.1.2. Training Model Variants

The training phase was structured around four distinct model configurations, carefully named to reflect the specific scope of their training data and designed to isolate the impact of progressive data augmentation and domain expansion:

- VIRA-GCN$_{\text{Base}}$ (Model Le2i-Plain): This preliminary model was trained exclusively on the Plain subset of the Le2i skeleton data. It serves as a fundamental reference point, establishing the performance achievable without the introduction of any data augmentation.
- VIRA-GCN$_{\text{Rob}}$ (Model Le2i-Occluded): This second preliminary model expanded the training set to include both the Plain and the Occluded sequences from the Le2i dataset. Its purpose was to quantify the incremental robustness gained solely from simulating occlusions.
- VIRA-GCN$_{\text{Spec}}$ (Model Le2i-FullAug): This primary model was trained exclusively on a highly augmented version of the Le2i dataset, explicitly combining the raw skeleton data Plain, sequences subjected to realistic visual occlusions Occluded, and sequences augmented with synthetic viewpoint rotations Rotated. This configuration used the Le2i data augmented across all defined transformation categories. The primary purpose of this model was to establish the maximum performance and robustness achievable solely within the well-controlled Le2i domain and to serve as a critical baseline to quantify the generalization failure, or domain shift, when evaluated against the distinct NTU RGB+D dataset.
- VIRA-GCN$_{\text{Gen}}$ (Model Le2i-NTU-DomainAug): This model was introduced to specifically address the identified challenge of cross-dataset generalization. It significantly expands the training regime by incorporating augmented data from both the Le2i and the NTU RGB+D domains. It retained the comprehensive Le2i augmented training set and further included the Kinect-transformed data Kntuplain and its synthetic viewpoint rotations Rotated Kntuplain derived from the NTU dataset. By explicitly exposing the model to the characteristics of both datasets, this model was engineered to achieve robust generalization across heterogeneous data sources, aiming to bridge the performance gap observed in cross-domain evaluation.

4.2. Transformation Results

We evaluated the effectiveness of our per-joint affine transformation from the Kinect to the MMPose skeleton format. Figure 3 shows a visual comparison of the transformed skeleton, demonstrating an impressive alignment with the target MMPose structure. Quantitatively, the Mean Squared Error (MSE) chart, as confirmed by the visualization in Figure 4, demonstrates that our transformation method significantly reduces the Euclidean distance error for each joint, bringing the Kinect data much closer to the MMPose format.

4.3. Comparative Model Performance

The models were evaluated using four standard performance metrics: accuracy, sensitivity, specificity, and F1-score, all calculated from the confusion matrix elements: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Fall sequences were consistently labeled as positive samples, and normal daily activities as negative samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

The decisive comparison between the models incorporating different degrees of domain information is visually synthesized in Figure 5, which plots the F1-score and accuracy across all five test datasets. The results demonstrate the critical impact of integrating target domain data into the training regimen.

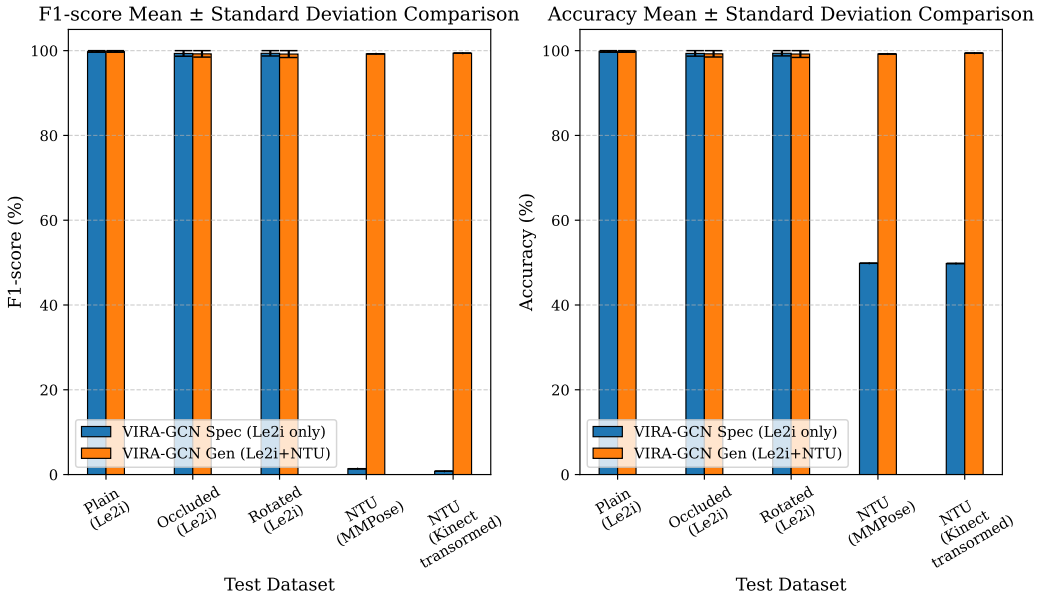

Figure 5. Comparative Model Performance Across Different Test Datasets. The figure shows the F1-score and accuracy (Mean ± Standard Deviation across 25 runs) comparing VIRA-GCN$_{\text{Spec}}$ and VIRA-GCN$_{\text{Gen}}$ when evaluated on five distinct test subsets: Plain, Occluded, Rotated (Le2i domain), and Mntuplain, Kntuplain (NTU domain).

4.3.1. Initial Augmentation Impact (Le2i-Plain Dataset)

To isolate the contribution of augmentation techniques, the initial model variants were evaluated. Table 3 details the comparative performance of the models when tested exclusively on the unaugmented (Plain) Le2i dataset.

Table 3. Comparative Performance of Preliminary Model Variants Evaluated on the Plain (Unaugmented) Le2i Test Dataset.

Model Name	Test Type	F1-Score Mean (%)	Accuracy Mean (%)	Sensitivity Mean (%)	Specificity Mean (%)
VIRA-GCN$_{\text{Base}}$	Plain	94.52	94.25	90.34	99.09
VIRA-GCN$_{\text{Rob}}$	Plain	99.50	99.69	99.25	100.00
VIRA-GCN$_{\text{Spec}}$	Plain	99.81	99.81	99.62	100.00

The data demonstrates a critical dependency on augmentation strategy, even for canonical views. The strictly VIRA-GCN$_{Base}$ exhibited a measurable deficiency, achieving a 94.52% F1-score, 94.25% accuracy, and a low mean sensitivity of 90.34%. In contrast, both augmentation strategies led to substantial performance gains. VIRA-GCN$_{Rob}$ achieved a 99.50% F1-score, 99.69% accuracy, and 99.25% sensitivity, while VIRA-GCN$_{Spec}$ attained the highest metrics overall, with a 99.81% F1-score and 99.81% accuracy. Furthermore, the perfect mean specificity (100%) achieved by both augmented models VIRA-GCN$_{Rob}$ and VIRA-GCN$_{Spec}$, confirms that the feature learning derived from augmentation successfully eliminates false positives in this controlled test scenario, underscoring its general benefit to feature robustness.

4.3.2. Performance on Le2i Domain Datasets

On the Le2i domain test sets (Plain, Occluded, Rotated), both primary models: the VIRA-GCN$_{Spec}$ and the VIRA-GCN$_{Gen}$ achieved exceptionally high performance, with metrics consistently approximating 100%. This sustained, near-perfect performance confirms the efficacy of the initial Le2i data augmentation strategy for achieving feature robustness within the source domain.

Specifically, VIRA-GCN$_{Spec}$ confirmed the effectiveness of the initial Le2i-only augmentation strategy for viewpoint robustness, achieving 99.38% accuracy and 99.37% F1-score on the Rotated test set. Crucially, the VIRA-GCN$_{Gen}$ maintained this high performance across all Le2i subsets, demonstrating that the inclusion of auxiliary NTU training data did not negatively impact the model's specialized performance or feature discrimination within the original source domain.

For the subsequent comprehensive cross-domain evaluation, we focus on the two highest-performing configurations: the VIRA-GCN$_{Spec}$ (Model Le2i-FullAug) and the VIRA-GCN$_{Gen}$ (Model Le2i-NTU-DomainAug), which were evaluated against all five test subsets, as detailed in Table 4.

Table 4. Comprehensive Comparative Performance of Primary Model Variants Grouped by Evaluation Domain.

Model Name	Test Type	F1-Score Mean (%)	Accuracy Mean (%)	Sensitivity Mean (%)	Specificity Mean (%)
VIRA-GCN$_{Spec}$	Plain	99.81	99.81	99.62	100.00
	Occluded	99.34	99.35	98.70	100.00
	Rotated	99.37	99.38	98.75	100.00
	Mntuplain	1.38	49.84	0.70	99.00
	Kntuplain	0.79	49.79	0.40	99.00
VIRA-GCN$_{Gen}$	Plain	99.81	99.81	99.62	100.00
	Occluded	99.24	99.25	98.50	100.00
	Rotated	99.19	99.20	98.40	100.00
	Mntuplain	99.25	99.22	99.00	99.50
	Kntuplain	99.40	99.42	99.20	99.60

4.3.3. Performance on NTU Domain Datasets

The decisive divergence in model efficacy is rigorously documented on the NTU domain test sets, specifically Mntuplain and Kntuplain. The VIRA-GCN$_{Spec}$, trained exclusively on the Le2i dataset, exhibited a pronounced inability to generalize to the NTU domain. Consistent with initial hypotheses, this configuration registered F1-scores and accuracy metrics approximating 50%. This result unequivocally substantiates the existence of a severe domain shift between the proprietary Le2i data and the publicly sourced NTU dataset characteristics.

Conversely, the VIRA-GCN$_{Gen}$ successfully mitigated this domain generalization deficit. This model achieved near-optimal performance across both NTU evaluation subsets:

- Mntuplain (NTU MMPose): F1-score approximated 99.25% and accuracy approximated 99.22%.
- Kntuplain (NTU Kinect Transformed): F1-score approximated 99.40% and accuracy approximated 99.42%.

Furthermore, the perfect mean specificity (100%) attained by the VIRA-GCN$_{Gen}$ on the NTU test sets confirms that the model successfully acquired the discriminatory features necessary to differentiate between the specific fall dynamics inherent to the NTU domain and general non-fall activities, thereby demonstrating robust cross-domain generalization.

4.4. Computational Analysis and Hardware Configuration

To substantiate the claim that the VIRA-GCN model is suitable for low-resource deployment, we provide a rigorous quantitative analysis of both its complexity and hardware performance. The comprehensive metrics for model size, computational demand, and inference latency are detailed in Table 5.

Table 5. VIRA-GCN Computational Complexity and Performance Metrics.

Reference Hardware	Model Parameters	Computational Complexity	Inference Latency
NVIDIA RTX 3090	4.06 M	7.85 GFLOPs	7.50 ms/sample

The system exhibits high computational efficiency, utilizing 4.06 M parameters and requiring 7.85 GFLOPs. Inference speed, a critical metric for real-time suitability, was measured on an NVIDIA RTX 3090 GPU (Manufactured by NVIDIA Corporation, Headquarters: Santa Clara, California, U.S., 24GB VRAM, 32 GB RAM). The average latency of 7.50 ms per sample is notably well below the typical frame interval of video streams ($\sim$ 33 ms for 30 FPS). These quantitative results affirm that the model is lightweight and fully capable of real-time inference, supporting its suitability for deployment in resource-constrained settings.

A key factor contributing to this high overall efficiency was the optimization of the training pipeline. Initially, the pipeline was highly inefficient, leading to GPU utilization restricted to approximately 30%. We resolved this bottleneck by refactoring the data augmentation logic to be executed entirely within the dataloader, enabling the process to run in parallel. This crucial architectural modification drastically improved GPU utilization and accelerated the training process.

5. Discussion

The empirical findings from the comparative study between the training configurations validate that utilizing a combined, heterogeneous, and augmented dataset is the most effective strategy for building a truly robust and generalizable fall detection model. While Model VIRA-GCN$_{Spec}$ demonstrated exceptional performance within its source domain, achieving near-perfect scores on the Plain, Occluded, and Rotated Le2i test sets, its severe performance drop to approximately 50% on the NTU domain test sets (Mntuplain, Kntuplain) unequivocally confirmed the critical problem of domain shift.

5.1. The Critical Challenge of Domain Shift

The initial evaluation, highlighted by the performance disparity of VIRA-GCN$_{Spec}$, revealed a significant cross-dataset generalization deficit that defines a critical challenge for skeleton-based fall detection systems. This outcome confirms that performance degradation

is attributed to a severe domain shift and inherent differences in the characteristics of the fall events. The main distinctions contributing to this deficit are summarized as follows:

- Fall Action Heterogeneity: As qualitatively illustrated in Figure 6, the NTU dataset generally features slower, more controlled descents where subjects often remain centered. Conversely, the Le2i dataset captures more realistic and complex scenarios, including falls initiated during activities such as walking, and subjects remain prone for the duration of the sequence. The VIRA-GCN$_{\text{Spec}}$ overfit to the complex dynamics of Le2i and could not reliably recognize the subtler, stylized falls of the NTU domain.

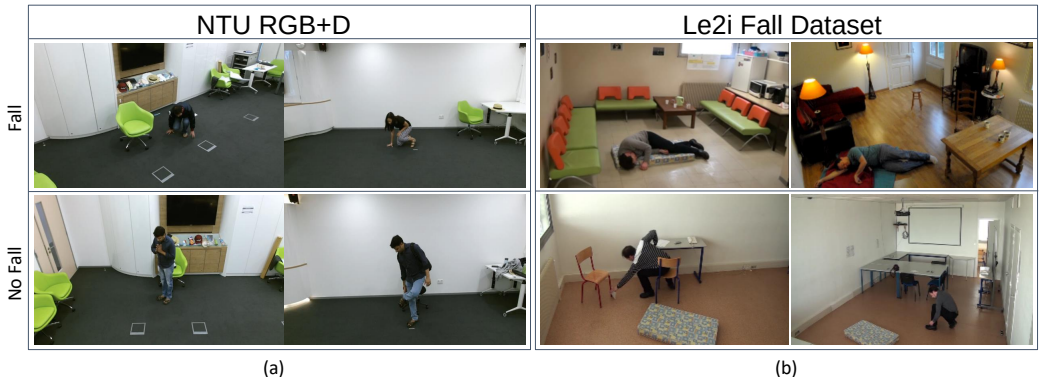

Figure 6. Comparative analysis of fall and non-fall actions in the (**a**) NTU RGB+D and (**b**) Le2i datasets. The Le2i dataset presents greater complexity by including non-fall actions that are visually similar to falls in NTU RGB+D, posing a significant challenge for detection models. The NTU dataset, by contrast, contains more structured and controlled fall events.

- Data Transformation Efficacy: Although our per-joint affine transformation successfully established a consistent mapping function from the Kinect to the MMPose format (Figure 4), the resultant poor cross-domain performance of VIRA-GCN$_{\text{Spec}}$ confirmed that failure was solely attributable to the intrinsic domain differences in fall dynamics, rather than mere data format misalignment. The transformation was necessary, but not sufficient, to bridge the domain gap.
- Implication for Generalization: These results suggest that standard models are prone to overfitting to specific recording conditions and sensor modalities. The successful strategy implemented in the following section demonstrates that the solution must involve explicit domain exposure.

5.2. Overcoming the Generalization Deficit

VIRA-GCN$_{\text{Gen}}$ successfully overcame this generalization deficit. By incorporating the transformed Kinect data (Kntuplain) and its rotated synthetic views into the training pipeline, the model achieved scores of nearly 100% on both the Le2i and the NTU test subsets. This outcome supports two fundamental conclusions:

- Domain Adaptation through Augmentation: The inclusion of the augmented NTU data successfully taught the model to recognize the specific fall characteristics and environmental conditions inherent in the NTU dataset (e.g., slower falls, post-fall recovery actions). This demonstrates that model performance degradation is not an inherent flaw in the model architecture (VIRA-GCN), but rather a failure to generalize due to domain-specific feature scarcity in the original training data.
- Efficacy of the Transformation Pipeline: The fact that VIRA-GCN$_{\text{Gen}}$ performed equally well on both the MMPose-extracted NTU data (Mntuplain) and the Kinect-transformed NTU data (Kntuplain) provides strong validation of our per-joint affine transformation method. It confirms that the transformation creates a consistently

aligned skeleton format, enabling seamless aggregation of different data sources (Kinect vs. MMPose) for robust cross-domain training.

5.3. State-of-the-Art Performance

As summarized in Table 6, the proposed VIRA-GCN$_{Gen}$ achieves superior performance compared to previously established state-of-the-art approaches on the Le2i dataset.

Table 6. Performance evaluation of various state-of-the-art methods on the Le2i dataset.

Model Methodology	Year	Accuracy/%	Sensitivity/%	Specificity/%
Lightweight ST-GCN [37]	2021	96.10	92.50	95.70
V2V-PoseNet [38]	2021	93.67	100.00	87.00
YOLOv7-DeepSORT [39]	2023	94.50	98.60	97.00
Alphapose ST-GCN [40]	2023	98.70	97.78	100.00
3D Multi-Stream CNNs [41]	2023	99.44	99.12	99.12
Three-Stream ST-GCN [42]	2024	99.64	97.25	-
VIRA-GCN (Ours)	2025	99.81	99.62	100.00

Specifically, the model demonstrates dominant accuracy, reaching 99.81%, surpassing all comparable methods. Furthermore, the perfect 100.00% specificity indicates that the model exhibits zero false positives on the evaluation data, a crucial factor for reducing unnecessary alerts in real-world application systems. This state-of-the-art result validates the architectural benefits of VIRA-GCN when coupled with our comprehensive augmentation and domain adaptation strategy.

In conclusion, the VIRA-GCN model, when trained with the comprehensive augmentation strategy of Model VIRA-GCN$_{Gen}$, not only maintains its real-time efficiency and robustness to occlusions and viewpoint changes within the primary domain (Le2i) but also successfully generalizes to the challenging, distinct domain of the NTU RGB+D dataset. This robust generalization capability is achieved through a practical and efficient pipeline that seamlessly integrates different sensor data types via our novel per-joint affine transformation.

6. Conclusions

In this study, we presented a comprehensive methodological framework to significantly enhance the robustness and generalization of vision-based fall detection models against two major real-world limitations: occlusions and variations in camera viewpoint. The core contribution is the development of the Viewpoint Invariant Robust Aggregation Graph Convolutional Network (VIRA-GCN), which utilizes a combination of a carefully designed data augmentation strategy, namely synthetic viewpoint generation, and an efficient inference-time aggregation method, which is consensus-based prediction, to achieve substantial performance improvements without necessitating resource-heavy multi-camera setups.

The efficacy of our approach was quantitatively confirmed on the challenging Le2i dataset, where the VIRA-GCN achieved superior performance metrics, including 99.81% accuracy, 99.62% sensitivity, and 100.00% specificity. Furthermore, the model is inherently lightweight, utilizing only 4.06 M parameters and achieving an average inference latency of 7.50 ms per sample on an NVIDIA RTX 3090 GPU, confirming its suitability for real-time, low-resource deployment.The model's efficiency was further established by the successful optimization of the training pipeline, which eliminated the initial computational bottleneck of restricted GPU utilization.

Crucially, the VIRA-GCN$_{Gen}$ successfully demonstrated robust cross-domain generalization. By leveraging the Kinect to MMPose transformation pipeline and incorporating

augmented NTU data, the model not only maintained its high performance on the Le2i domain but also successfully generalized to the challenging, distinct characteristics of the NTU RGB+D dataset, achieving scores of nearly 100% on all NTU evaluation subsets. This finding, while confirming that model degradation is initially driven by domain shift stemming from inherent differences in fall characteristics (e.g., fast versus slow falls), validates that the problem can be mitigated through our proposed architecture and comprehensive domain adaptation strategy. In summary, this research successfully provides a valuable and computationally practical framework for developing more reliable fall detection systems that are robust to variations in camera perspective, thereby advancing their readiness for deployment in real-world applications.

Author Contributions: Conceptualization, M.Z. and Y.T.; methodology, M.Z. and L.B.; software M.Z. and L.B.; validation, M.Z. and L.B.; formal analysis, M.Z. and L.B.; investigation, M.Z.; data management, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z. and L.B.; visualization, M.Z.; supervision, Y.T. and R.O.H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research utilized publicly available, pre-existing datasets (Le2i Fall Dataset and NTU RGB+D Dataset) which contain anonymized data and were collected with prior ethical approval or waiver by the original authors. The study did not involve new collection of human or animal data

Informed Consent Statement: Not applicable. The study utilized publicly available, pre-existing datasets and did not involve the collection of new human subject data. The original authors of the datasets were responsible for obtaining informed consent from all participants.

Data Availability Statement: The data that support the findings of this study are available from the original sources: Le2i Fall Detection Dataset and the NTU RGB+D Dataset. The core datasets generated and analyzed during the current study, including all model performance metrics and transformation results, are contained within this published article (Tables 1–5 and Figures 1–6).

Conflicts of Interest: Author Lorenzo Bolzani was employed by the company flAIght. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. The United Nations Department of Economic and Social Affairs (UN DESA). *World Population Ageing 2020 Highlights: Living Arrangements of Older Persons*; (ST/ESA/SER.A/451); United Nations: New York, NY, USA , 2020.
2. WHO. *WHO Global Report on Falls Prevention in Older Age*; WHO: Geneva, Switzerland, 2021.
3. Nazar, M.; Ahsan, A.; Khan, M.I. A review of vision-based fall detection systems for elderly people: A performance analysis. *J. King Saud-Univ. Comput. Inf. Sci.* **2020**, *32*, 808–819.
4. Shaha, I.; Islam, M.; Hasan, S.; Zahan, M.L.N.; Islam, T.; Kamal, A.M.H.; Chowdhury, M.M.I.; Hasan, M.A.; Khan, F.M.H. Exploring Human Pose Estimation and the Usage of Synthetic Data for Elderly Fall Detection in Real-World Surveillance. *IEEE Access* **2022**, *10*, 94249–94261 . [CrossRef]
5. Zhang, A.; Feng, X.; Su, X.; Li, F. Robust Self-Adaptation Fall-Detection System Based on Camera Height. *Sensors* **2019**, *19*, 2390.
6. Sannino, M.M.; De Masi, L.; D'Angelo, V.; Sannino, L.; D'Agnese, M.; Sannino, F. Fall Detection with Multiple Cameras: An Occlusion-Resistant Method Based on 3-D Silhouette Vertical Distribution. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 262–273.
7. Wang, J.; Zhang, Z.; Li, B.; Lee, S.; Sherratt, R.S. An enhanced fall detection system for elderly person monitoring using consumer home networks. *IEEE Trans. Consum. Electron.* **2014**, *60*, 23–29. [CrossRef]
8. Desai, K.; Mane, P.; Dsilva, M.; Zare, A.; Shingala, P.; Ambawade, D. A novel machine learning based wearable belt for fall detection. In Proceedings of the 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Online, 2–4 October 2020; pp. 502–505.
9. Chelli, A.; Pätzold, M. A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition. *IEEE Access* **2019**, *7*, 38670–38687. [CrossRef]

10. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 150–166. [CrossRef]
11. MMPose Contributors. MMPose: An OpenMMLab Pose Estimation Toolbox and Benchmark. 2020. Available online : https://github.com/open-mmlab/mmpose (accessed on 19 October 2025).
12. Yao, L.; Yang, W.; Huang, W. An improved feature-based method for fall detection. Teh. Vjesn. **2019**, *26*, 1363–1368.
13. Bian, Z.-P.; Hou, J.; Chau, L.-P.; Magnenat-Thalmann, N. Fall Detection Based on Body Part Tracking Using a Depth Camera. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 430–439. [CrossRef]
14. Debard, G.; Karsmakers, P.; Deschodt, M.; Vlaeyen, E.; Bergh, J.; Dejaeger, E.; Milisen, K.; Goedemé, T.; Tuytelaars, T.; Vanrumste, B. Camera Based Fall Detection Using Multiple Features Validated with Real Life Video. In Proceedings of the 7th International Conference on Intelligent Environments, Nottingham, UK, 25–28 July 2011; Volume 10.
15. Lu, N.; Wu, Y.; Feng, L.; Song, J. Deep Learning for Fall Detection: Three-Dimensional CNN Combined with LSTM on Video Kinematic Data. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 314–323. [CrossRef]
16. Tian, Y.; Lee, G.-H.; He, H.; Hsu, C.-Y.; Katabi, D. Rf-based fall monitoring using convolutional neural networks. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Singapore, 8–12 October 2018; Volume 2, pp. 1–24.
17. Chen, Y.; Li, W.; Wang, L.; Hu, J.; Ye, M. Vision-based fall event detection in complex background using attention guided bi-directional lstm. *IEEE Access* **2020**, *8*, 161337–161348. [CrossRef]
18. Dutt, M.; Gupta, A.; Goodwin, M.; Omlin, C.W. An interpretable modular deep learning framework for video-based fall detection. *Appl. Sci.* **2024**, *14*, 4722.. [CrossRef]
19. Tsai, T.H.; Hsu, C.W. Implementation of fall detection system based on 3D skeleton for deep learning technique. *IEEE Access* **2019**, *7*, 153049–153059. [CrossRef]
20. Keskes, O.; Noumeir, R. Vision-Based Fall Detection Using ST-GCN. *IEEE Access* **2021**, *9*, 28224–28236. [CrossRef]
21. Egawa, R.; Miah, A.S.M.; Hirooka, K.; Tomioka, Y.; Shin, J. Dynamic fall detection using graph-based spatial temporal convolution and attention network. *Electronics* **2023**, *12*, 3234. [CrossRef]
22. Shin, J.; Miah, A.S.M.; Egawa, R.; Hirooka, K.; Hasan, M.A.M.; Tomioka, Y.; Hwang, Y.S. Fall recognition using a three stream spatio temporal GCN model with adaptive feature aggregation. *Sci. Rep.* **2025**, *15*, 10635. [CrossRef]
23. Kepski, M.; Kwolek, B. A New Multimodal Approach to Fall Detection. *PLoS ONE* **2014**, *9*, e92886.
24. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
25. Charfi, I.; Mit'eran, J.; Dubois, J.; Atri, M.; Tourki, R. Optimised spatio-temporal descriptors for real-time fall detection: Comparison of SVM and Adaboost based classification. *J. Electron. Imaging* **2013**, *22*, 043009. [CrossRef]
26. Kepski, M.; Kwolek, B. A New Fall Detection Dataset and the Comparison of Fall Detection Methods. In Proceedings of the 2019 International Conference on Computer Vision and Graphics, Warsaw, Poland, 16–18 September 2019; Springer: Cham, Switzerland, 2019; pp. 11–22.
27. Ma, X.; Wang, H.; Xue, B.; Zhou, M.; Ji, B.; Li, Y. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE J. Biomed. Health Inform.* **2014**, 18, 1915–1922. [CrossRef]
28. Soto-Valero, L.G.; Medina-Merodio, M. SisFall: A Fall and Movement Dataset. *Sensors* **2017**, *17*, 198. [CrossRef]
29. TST Fall Detection Dataset v2. IEEE Dataport, 2016. Available online: https://ieee-dataport.org/documents/tst-fall-detection-dataset-v2 (accessed on 19 October 2025).
30. Yu, X.; Jang, J.; Xiong, S. KFall: A Comprehensive Motion Dataset to Detect Pre-impact Fall for the Elderly Based on Wearable Inertial Sensors. *Front. Aging Neurosci.* **2021**, *13*, 692865. [CrossRef]
31. Vavoulas, G.; Chatzaki, C.; Malliotakis, T.; Pediaditis, M. The MobiAct Dataset: Recognition of Activities of Daily Living using Smartphones. In Proceedings of the 2nd International Conference on Information and Communication Technologies for Ageing Well and e-Health, Rome, Italy, 21–22 April 2016; pp. 143–151.
32. Saleh, M.; Abbas, M.; Le Jeannès, R.B. FallAllD: An Open Dataset of Human Falls and Activities of Daily Living for Classical and Deep Learning Applications. *IEEE Sens. J.* **2021**, *21*, 5110–5120. [CrossRef]
33. Klenk, J.; Schwickert, L.; Jäkel, A.; Krenn, M.; Bauer, J.M.; Becker, C. The FARSEEING real-world fall repository: A large-scale collaborative database to collect and share sensor signals from real-world falls. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1339–1346. [CrossRef]
34. Fang, H.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; Lu, C. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7157–7173. [CrossRef] [PubMed]
35. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172. [CrossRef]
36. Song, Y.-F.; Zhang, Z.; Shan, C.; Wang, L. Richly Activated Graph Convolutional Network for Robust Skeleton-based Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1915–1925. [CrossRef]

37. Li, W.; Wang, Y.; Wang, Y.; Li, B.; Zhang, H.; Wu, Y. Lightweight Spatio-Temporal Graph Convolutional Networks for Skeleton-based Action Recognition. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–12 June 2021; pp. 1605–1609.
38. Alaoui, A.; El-Yaagoubi, M.; Chehri, A. V2V-PoseNet: A View-Invariant Pose Estimation Network for Action Recognition. In Proceedings of the International Conference on Advanced Communication Technologies and Information Systems (ICACTIS), Coimbatore, India, 19–20 March 2021; pp. 1–6.
39. Zi, Q.; Liu, Y.; Zhang, S.; Yang, T. YOLOv7-DeepSORT: An Improved Human Fall Detection Method. *Sensors* **2023**, *23*, 789.
40. Li, H.; Wei, S.; Xu, Y.; Yang, J.; Zhang, Z. A Novel Fall Detection Method Based on AlphaPose and Spatio-Temporal Graph Convolutional Networks. *Sensors* **2023**, *23*, 5565.
41. Alanazi, S.; Alshahrani, S.; Alqahtani, A.; Alghamdi, A. 3D Multi-Stream CNNs for Real-Time Fall Detection. *IEEE Access* **2023**, *11*, 15993–16004.
42. Shin, J.; Kim, S.; Kim, D. Three-Stream Spatio-Temporal Graph Convolutional Networks for Human Fall Detection. In Proceedings of the 2024 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 16–18 October 2024; pp. 1205–1210.