Striking the Balance: Human Pose Estimation based Optimal Fall Recognition

Aniruddh Balram
University of Maryland
abalram1@umd.edu

Tharun Puthanveettil
University of Maryland
tvpian@umd.edu

Abhijay Singh
University of Maryland
abhijay@umd.edu

Krishna Hundekari
University of Maryland
krishnah@umd.edu *

Abstract

This study addresses the pressing issue of fall detection among the elderly, emphasizing early intervention for improved well-being. Utilizing spatiotemporal transformers, our approach effectively tackles challenges such as illumination variations, noise, and occlusions. Leveraging joint positions in a curated video sequence dataset, our model achieves strong binary classification performance, distinguishing falls from non-falls. Comparative analysis between the transformer-only model and the enhanced GCN+Transformer demonstrates the latter's superior precision, recall, F1 score, and accuracy, particularly excelling in challenging scenarios. Insights from view generalization and occlusion cases underscore the model's adaptability and sensitivity, calling for ongoing research to fortify robustness. The augmented performance of the GCN+Transformer is attributed to the enriched spatial context provided by the GCN.

Acknowledging limitations, including reliance on human pose key points and sensitivity to training data diversity, the proposed methodology stands as a robust solution for efficient and accurate fall detection, especially in scenarios emphasizing privacy preservation. The numerical results affirm the pragmatic applicability of the GCN+Transformer in real-world settings, with ongoing refinement efforts crucial for continual enhancement.

1. Introduction

Falls by elderly people are a significant cause of health issues because they can result in disabling fractures, and dislocations and reduce the sense of independence for the elderly. Aside from this, major falls - though injuries associated with which may be cured - can also result in psychological conditions such as PTSD and disrupt the functioning of the person. According to one survey, falls are one of the leading causes of injury-related deaths in the elderly above the age of 79[1]. Therefore, early detection of falls is important as a precaution.

Current approaches use CNNs to capture spatial information to detect falls in single images or 3D CNNs along with LSTMs to learn spatio-temporal features for a video sequence. The features learned may not be accurate due to illumination differences, noise, or occlusions. Networks detecting falls based on human poses/skeletons alleviate the above-mentioned problems. Most of the research in fall detection using human poses is based on graph neural networks(GCNs). Transformers were a groundbreaking discovery for dealing with sequential data in NLP problems which also shifted to the domain of computer vision after VITs came[13]. Transformers have shown some promise in the research of human action recognition but much of it is unexplored. In our approach, we plan to explore the usage of spatio-temporal transformers for detecting falls.

In our pipeline, we curated a dataset of joint positions in video sequences as input to train a deep network model with a transformer[18] only and a Graph Convolutional Network(GCN) backed Transformer. The models performed binary classification, detecting falls versus no falls. For the dataset, we maintained a good balance between positives (video sequences with falls) and negatives (video sequences with no falls). It is essential to note that the input to the network was joint keypoints rather than RGB frames. This choice resulted in more accurate predictions of human action, even in the presence of occlusions, and improved robustness to noise and illumination differences compared to using RGB images[14]. As the prediction of falls happened solely based on keypoints, our pipeline also preserved the privacy of the person under observation by not storing images but only the location of skeletal keypoints. Additionally, we explored how incorporating other spatial priors us-

*https://github.com/abhijaysingh/Fall-Detection.git

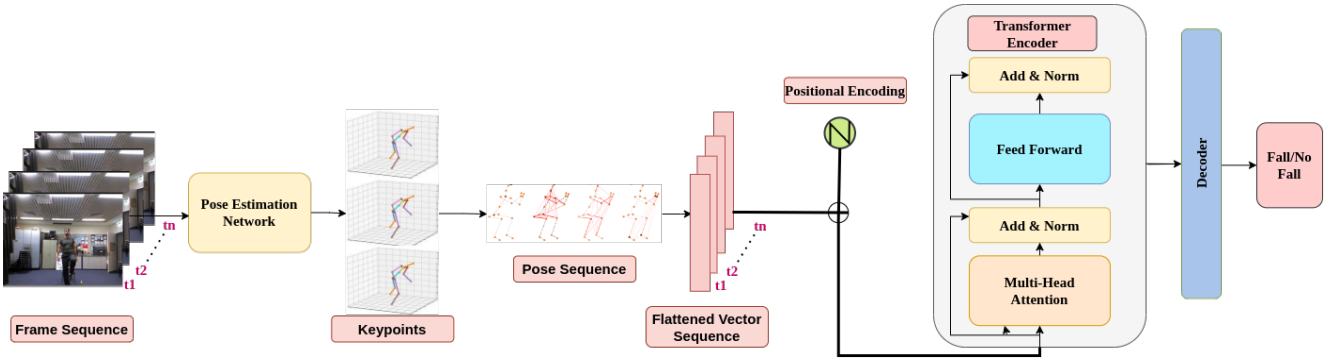


Figure 1. Transformer-only pipeline for Fall Detection

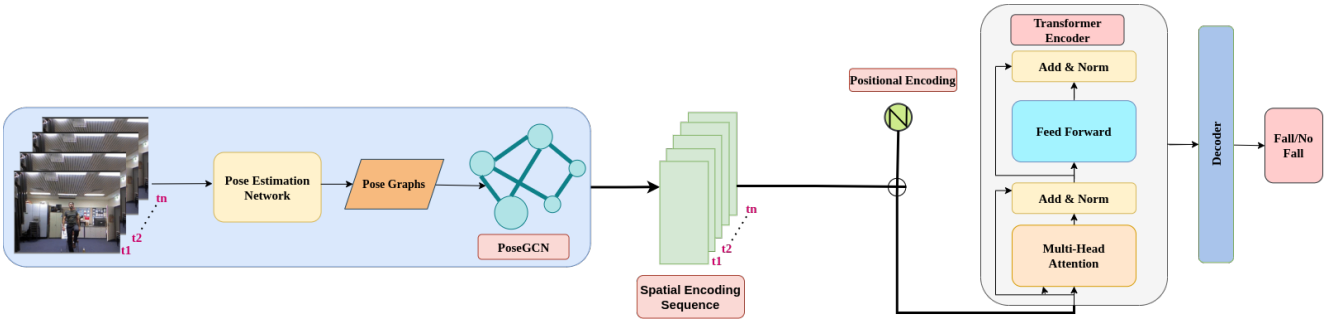


Figure 2. GCN+Transformer pipeline for Fall Detection

ing GCN along with skeletal keypoints, could enhance the model's accuracy.

2. Related Work

Falls represent a significant threat to the health and independence of adults 65 years of age and older. [4].Expensive research is done to design devices which accurately detect falls. In computer vision, fall detection comes under the study of Human action recognition which has been an integral part in the community since before the boom of AI. Early research included designing handcrafted features to train a linear classifier to predict action[12]. With the advent of deep learning, several methods were introduced for the task of action recognition. Initially, deep learning architectures such as CNNs[6] and RNNs[9] were used to classify human actions based on RGB/Depth image sequences or by a meaningful multi-modal fusion[17]. Sequences had to be short due to inherent vanishing and exploding gradient problems in RNNs. With the introduction of Vision Transformer(ViT)[5], transformers could be integrated with CNNs to handle temporal information which overcame all the issues related to RNNs. For action recognition, as there can be changes in body scale, occlusion, viewpoint variation, and fast motion between frames; CNNs fail to capture the important features. Besides, for high-resolution videos; CNNs can be computationally burdening. Skeletal graphs of humans are computationally effi- cient and maintain structural information of humans across frames without an effect from background variables[16]. Skeleton-based action recognition captures meaningful temporal sequences of human actions, outperforming CNN- based methods. Therefore, much of the research in human action recognition is skeleton based. Several techniques involving CNNs for skeletal data have been explored[2, 10].CNNs are popular for encoding spatial information, but since a skeleton is just a graph, CNNs aren't a natural fit for extracting features from a skeleton. Therefore, newer deep learning frameworks like Graph Convolutional Net- works (GCNs) have become fairly popular in tackling ac- tion recognition with skeletal data.[19] utilizes GCNs for encoding both spatial and temporal information achieving better performance than CNN-based methods. For encod- ing temporal information, the state-of-the-art architecture is transformer[15]. Methods fusing transformer architecture to encode temporal information and GCN to encode graph information [20] outperform all the previously explored methods. However, the real bottleneck of skeleton-based methods lies in the accuracy of pose estimation. Therefore, accurate human pose estimation is necessary for high qual- ity action recognition especially in the case of Fall Detection. Errors in pose estimation can occur due to noise, occlusions or high motion speeds. Existing models like PoseNet[7] and OpenPose[3]

can give fairly accurate estimations of 3D poses. Pose Estimation is not a focus of this study and therefore, we assume that the estimated poses from MediaPipe solutions[11] as implemented in the workflow are fairly accurate.

In purview of the importance of fall prediction, our study proposes an end-to-end framework which estimates 3D poses from RGB Images. The poses are then used as an input to a GCN-Transformer based architecture to encode graphical and temporal information. The embedding is used to classify whether a sequence of images results in a fall or not.

3. Dataset

University of Rzeszow Fall Detection (URFD) Dataset: The URFD dataset, introduced by Kwolek and Kepski in 2014[8], includes 40 daily living activities and 30 falls sequences, captured using two Kinect cameras for depth and RGB images, and includes accelerometer data for contextual information.

NTU RGB60: NTU RGB+D dataset containing 56,880 samples was used. It includes 60 action classes, including daily be- haviors and health-related actions, performed by 40 partici- pants.

4. Methodology

4.1. Overview

In the course of our research endeavors, we adopted a dual-pronged approach, strategically exploring the potential of integrating Human Pose Estimation key points with state-of-the-art sequential data architecture—specifically, Transformers. Our primary objective revolves around leveraging this approach to make nuanced predictions in the context of abnormal cases within Human Action Recognition scenarios, with a specialized focus on Fall detection applications.

In the first approach as shown in Fig.1, we chose to utilize the raw, unprocessed human pose key points as the primary input for the transformer. Our aim was to facilitate the generation of a higher-dimensional embedding through the transformer, enabling a refined binary classification mechanism to distinguish between fall and non-fall events.

The second approach as shown in Fig.2 stems from a thoughtful consideration of the inherent limitations of Transformers, particularly in their sequential nature. Acknowledging potential challenges in extracting spatial contexts, we introduced a sophisticated layer in the form of a Graph Convolutional Network (GCN). This spatial encoding layer actively engages in extracting spatial semantics from the detailed information provided by Human Pose Keypoints. Subsequently, the resulting high-level embedding is carefully channeled into the transformer, intending to craft a temporal embedding as the cornerstone for robust Fall detection capabilities.

This dual-strategy, harmonizing the distinctive attributes of the spatial encoding layer and the transformer, is meticulously designed to enhance the model's discernment capabilities in identifying abnormal events within the domain of Fall Detection.

Offline Pre-processing

In the initial phase (shown in Fig.3), inherent to both methodologies, we perform primary pre-processing. This involves sampling the action sequence video to acquire frames corresponding to the relevant action classes. Subsequently, a Human Pose Estimation algorithm is employed to extract key points specific to each frame. The resulting sequence of key points is stored locally for each view within the action sequence, setting the groundwork for subsequent utilization in the model training pipeline.

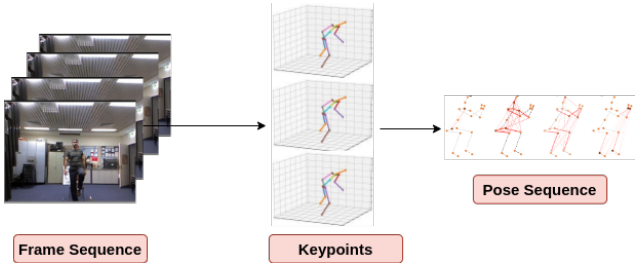


Figure 3. Offline preprocessing pipeline

4.2. Transformer-only Fall Detection

4.2.1 Temporal Encoding using Transformer

Following the pre-processing step, the extracted key points undergo processing through a transformer-based sequential model. Each flattened vector of key points from every frame is positionally encoded with a higher-dimensional vector before being fed into the transformer, generating a single high-dimensional feature embedding for binary classification.

4.2.2 Deep Learning-based Binary Classification

The subsequent stage involves feeding the computed feature embedding into a deep learning-based binary classifier, specifically designed for fall detection applications. The classifier's robustness is paramount for ensuring accurate and timely identification of abnormal events within the given video or image sequences.

4.3. GCN+Transformer Fall Detection

4.3.1 Spatial Encoding using GCN

In the initial phase, subsequent to pre-processing, the extracted key points are directed to a Graph Convolutional

Network (GCN)-based model for spatial encoding. The primary objective is to capture inherent spatial dynamics within the key point sequences and create frame-specific embeddings. The workflow of the PoseGCN for synthesizing frame-specific embeddings is given in Fig.4.

4.3.2 Temporal Encoding using Transformer

Moving to the second stage, the methodology progresses by processing the extracted key points through a transformer-based sequential model. This includes applying higher-dimensional positional encoding to enhance temporal information within the sequence of spatial embeddings. Subsequently, the positionally encoded spatial embeddings traverse through a Multi-head attention-based Transformer encoder layer, chosen for capturing complex temporal dependencies and patterns.

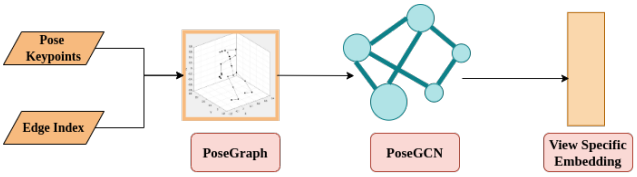


Figure 4. Synthesis of frame-specific embedding using PoseGCN

Deep Learning-based Binary Classification

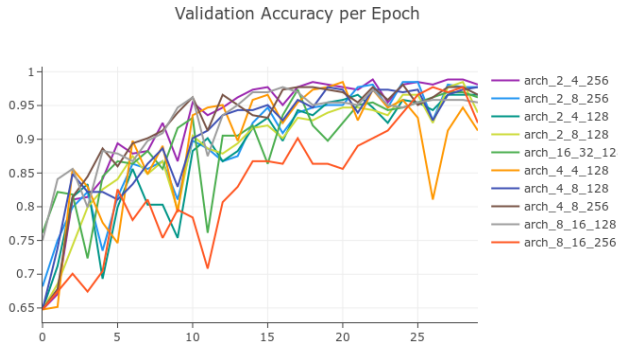
Both methodologies converge at this stage, where the computed higher-dimensional temporal embedding is fed into a deep learning-based binary classifier. This classifier, tailored for fall detection applications, plays a crucial role in predicting abnormal cases, ensuring robustness for accurate and timely identification of abnormal events within the given video or image sequences.

5. Experiments

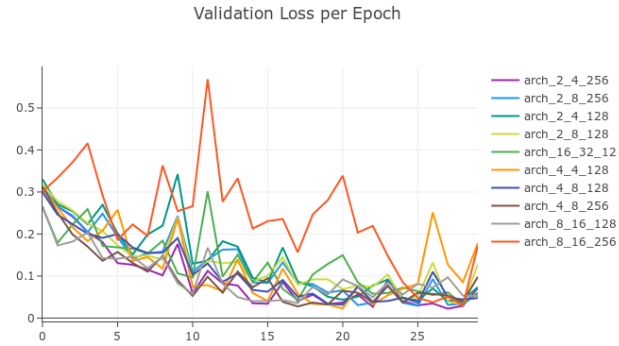
5.1. Experimental Setup

In our experimental setup, we focused on two core datasets: UR and NTU, boasting 90 and 938 data samples, respectively. The UR dataset, designed specifically for fall detection, comprises two classes: Fall and ADL. From the available video sequences captured from Front and Top views, we exclusively utilized the Front view for model training. On the other hand, the NTU dataset spans 60 action classes, and for our binary classification approach, we selected six classes, designating one as Fall and the rest as non-fall.

We adhered to a conventional 60-25-15 training-validation-testing split for both datasets. The training phase involved two distinct model architectures: Transformer and GCN+Transformer. Across both datasets, training parameters included 30 epochs, a batch size of 32, and an initial



(a) Validation accuracy per epoch



(b) Validation loss per epoch

Figure 5. Validation Accuracy and Loss Curves of Transformer Model

learning rate of 0.0001. Employing the Reduce-On-Plateau scheduler facilitated adaptive learning rate adjustments during training. Notably, for the Transformer model, we explored various architectural configurations, adjusting hyperparameters such as the number of layers, heads, and feed-forward dimensions. The final chosen configuration served as the architecture for the Transformer component within the GCN+Transformer model. The various transformer architectures that were tested are summarized in Table 1.

Our ablation study aimed to assess model robustness under different scenarios. To simulate occlusion, we systematically masked keypoints corresponding to three distinct scenarios: obstruction due to indoor furniture, bending over to fetch items, and working on a laptop or carrying items. Additionally, for evaluating model generalization across views, we leveraged multi-view data from the NTU dataset. Performance analysis was conducted for each view, with View 1 representing the front view, akin to the perspective of a social robot for medical assistance. In an attempt to obtain optimal inference efficiency, we introduced

a skip-frame parameter to determine the number of frames skipped between those used for inference in a video sequence. Lastly, we benchmarked the performance of our proposed GCN+Transformer model against other prominent spatial and sequential models, all trained on the UR dataset.

Architecture	Precision	Recall	F1	Accuracy	Geometric Mean
(2, 4, 128)	0.983	0.9821	0.9822	0.9821	0.9858
(4, 4, 128)	0.9602	0.9554	0.9558	0.9554	0.9641
(4, 8, 128)	0.9913	0.9911	0.9911	0.9911	0.9929
(8, 16, 128)	0.9913	0.9911	0.9911	0.9911	0.9929
(2, 8, 128)	0.9913	0.9911	0.9911	0.9911	0.9929
(2, 8, 256)	0.9913	0.9911	0.9911	0.9911	0.9929
(2, 4, 256)	**0.9913**	**0.9911**	**0.9911**	**0.9911**	**0.9929**
(4, 8, 256)	0.983	0.9821	0.9822	0.9821	0.9858
(16, 32, 128)	0.9913	0.9911	0.9911	0.9911	0.9929
(8, 16, 256)	0.9506	0.9464	0.9454	0.9464	0.9239

Table 1. Performance comparison of different Transformer Architectures based on metrics

6. Results

6.1. Evaluation Metrics

In assessing the effectiveness of the proposed fall detection framework, which is treated as a binary classification problem, several key performance metrics were employed to evaluate the model's accuracy, robustness, and generalization capabilities in detecting fall events.

- **Accuracy and Precision:** Accuracy is the main statistic for assessing the classification performance of the model. It displays how accurate action predictions have been overall across all classes. In order to evaluate the model's capability to accurately detect positive instances among the anticipated samples, precision for each action class will also be computed.

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

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

- **Recall:** Recall gauges the model's capability to correctly identify all positive instances, capturing the proportion of actual positives that were predicted correctly.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- **Confusion Matrix:** To provide a thorough examination of the model's predictions, a confusion matrix will be created. It will highlight true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each action type. This makes it easier to pinpoint specific areas where the model may struggle or perform well.
- **Training and Testing Time:** In real-world applications, computational efficiency is essential. To evaluate the model's effectiveness, the time required for both training the model and making predictions during testing will be recorded.
- **F-Score:** The F-score provides a balanced assessment of both precision and recall, offering a single metric to gauge the overall effectiveness of an action recognition model.

$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

- **Geometric Mean:** The geometric mean is particularly useful when dealing with imbalanced datasets, as it helps prevent one metric from dominating the evaluation.

$$Geometric\_Mean = \sqrt{Precision \times Recall} \tag{5}$$

6.2. Discussion

The comparative analysis underscores GCN+Transformer's superiority in fall detection, outperforming the Transformer in precision (92.5%) and F1 score (90.3%) in the UR dataset. While the Transformer achieves high metrics (99.1% across the board) in the NTU dataset, it faces a minor dip compared to GCN+Transformer, emphasizing the latter's prowess in diverse action scenarios. The training and loss curves for the GCN+Transformer in Fig.5a and Fig.5b show that the models are consistently getting better at learning the patterns.

The excellence of GCN lies in capturing spatial dependencies, modeling relationships between keypoints, and encoding contextual information crucial for understanding the spatial arrangement of body parts. This spatial awareness complements the sequential information encoded by Transformers, providing a holistic representation of video data.

6.3. Ablation Study

In the context of this research, the following ablation studies were performed to comprehensively assess various facets of the proposed pipeline:

1. **Camera Angle Variability:** In evaluating the influence of varied camera angles on our pipeline's performance, our study seeks to quantify potential reductions in efficacy when faced with angles divergent from the training phase. Additionally, we examine adaptability under an ego-centric view for the potential deployment of social robots.
   As can be seen in Table 3 the GCN+Transformer model is evaluated for its ability to generalize across various perspectives. It achieves perfect precision, recall, F1 score, accuracy, and geometric mean when trained and tested on NTU View 1. However, when tested on NTU View 2 and View 3, the model shows slight declines in

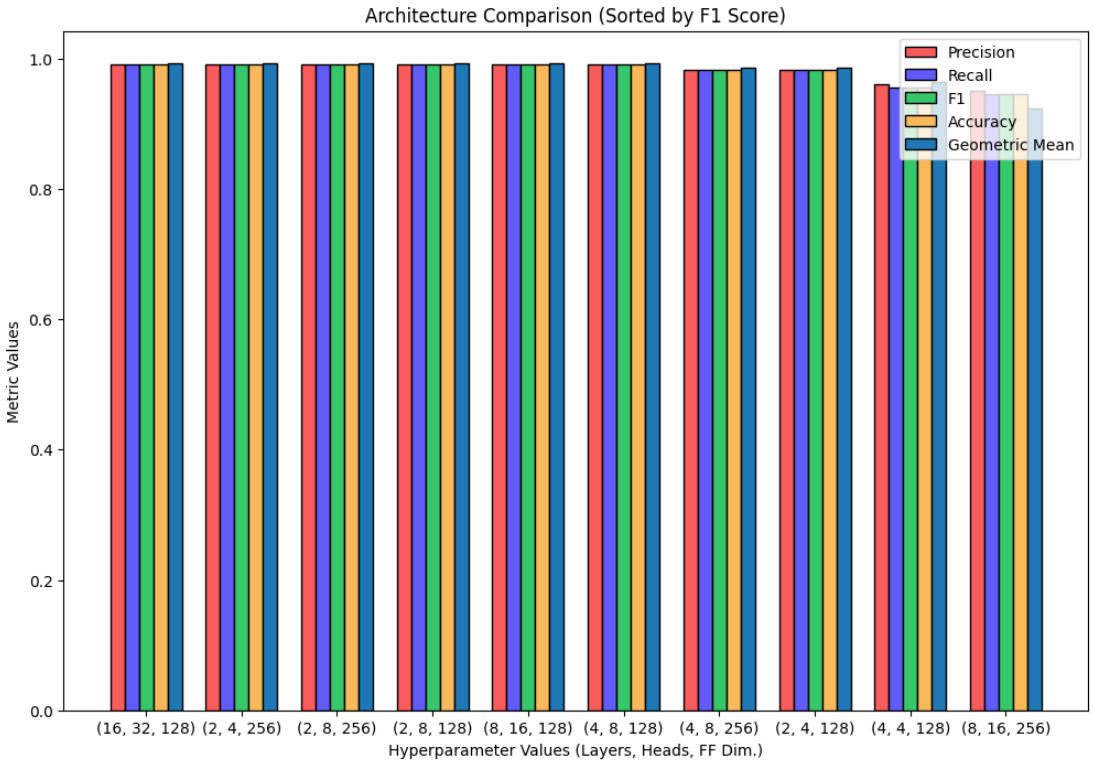


Figure 6. Comparison of performance based on Transformer Architectures

Model	UR					NTU				
	Precision	Recall	F1	Accuracy	Geometric Mean	Precision	Recall	F1	Accuracy	Geometric Mean
Transformer	0.912	0.900	0.893	0.900	0.816	0.991	0.991	0.991	0.991	0.993
GCN+Transformer	0.925	0.900	0.903	0.900	0.925	1.000	1.00	1.00	1.00	1.00

Table 2. Transformer & GCN+Transformer performance on NTU & UR Dataset

Views	Precision	Recall	F1	Accuracy	Geometric Mean
View 2	0.9735	0.9732	0.9733	0.9732	0.9737
View 3	0.9202	0.9196	0.9198	0.9196	0.9159

Table 3. Performance of GCN+Transformer based on views

performance. Despite this, the model maintains high accuracy and effectiveness across different views, indicating a high degree of generalization. Overall, the GCN+Transformer model shows promising adaptability and robustness in handling diverse views within the NTU dataset.

2. **Partial Occlusion Analysis:** A primary focus of our ablation study is the analysis of how partial occlusions impact the pipeline's performance. This investigation aims to understand how the presence of obstructions or partial occlusions within video frames may affect the accuracy of the anomaly detection system, thereby enhancing the overall reliability of our proposed methodology.

The assessment of the GCN+Transformer model across diverse occlusion scenarios exposes distinctive performance trends as can be seen in Table 4. In scenarios marked by Type 1 occlusion, where indoor furniture obstructs lower-body keypoints, the model exhibits resilience, showcasing high precision, recall, F1 score, and accuracy. This robustness is attributed to the sustained visibility of crucial joints essential for effective fall detection. Conversely, Type 2 occlusion, involving torso obstruction during activities like bending over, results in a noticeable performance decline. The compromised visibility of key torso-related keypoints impedes accurate fall-related action recognition, leading to diminished precision, recall, F1 score, and accuracy. In contrast, Type 3 occlusion, affecting upper limbs during activities like working on a laptop or carrying items, yields commendable model performance, maintaining high metrics across all parameters. The preserved visibility of upper limb keypoints enables the model to proficiently iden-

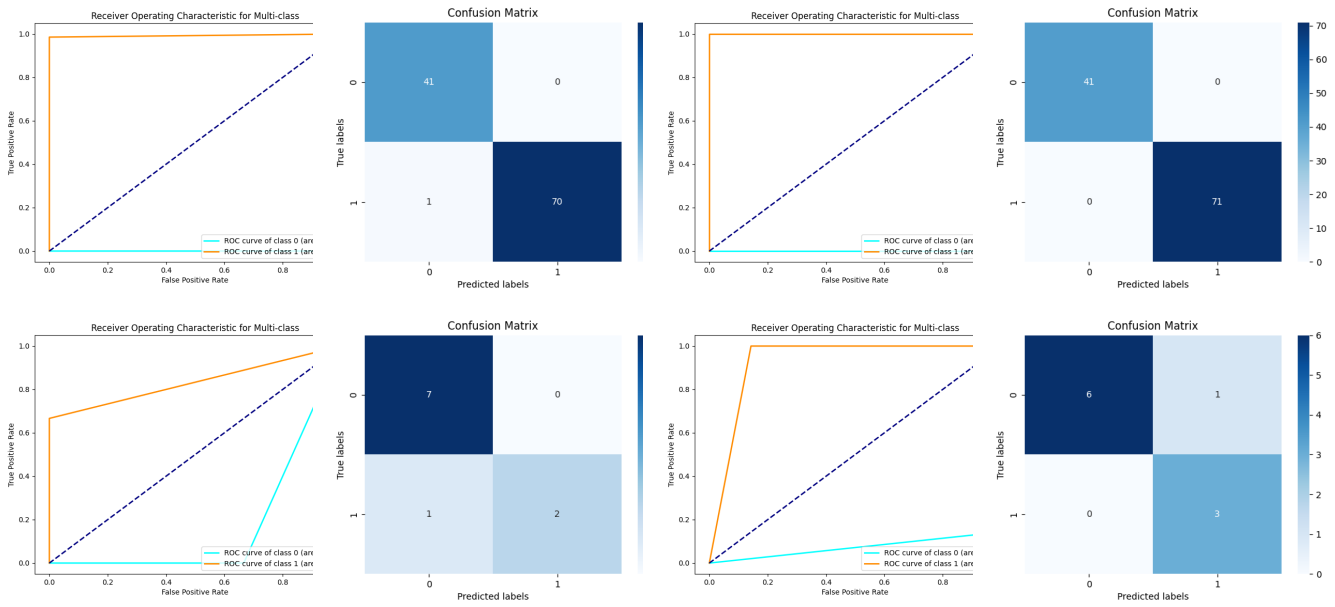


Figure 7. Comparison between ROC Curves and AUC Curves between NTU (first row) and UR (second row) datasets. The ROC Curves and the confusion matrices in the left column are the Transformer-only architecture, while the right column shows the results for the GCN+Transformer architecture.

tify fall-indicative movements. In summary, the study underscores the importance of unobstructed visibility of specific body parts for precise fall detection, highlighting the nuanced impact of occlusion patterns on model effectiveness.

Occlusion Case	Precision	Recall	F1	Accuracy	Geometric Mean
Type 1	0.9398	0.9375	0.9365	0.9375	0.9174
Type 2	0.6558	0.6696	0.6559	0.6696	0.5820
Type 3	0.8976	0.8929	0.8899	0.8929	0.8572

Table 4. Performance of GCN+Transformer based on different occlusion scenarios

3. **Optimal Frame Quantity:** The ablation study includes an exploration of the optimal number of frames necessary for achieving peak performance within the pipeline. This inquiry is designed to discern the most effective frame quantity, ensuring precision in anomaly detection. Systematically varying the number of frames considered helps establish an empirical understanding of the relationship between frame quantity and the overall efficacy of our proposed anomaly detection system.
The analysis of model performance across varying skip rates provides insightful observations regarding the model's sensitivity to frame subsampling as observed in Table 5. A skip rate of 1 demonstrates robust precision, recall, F1 score, accuracy, and geometric mean, suggesting that processing consecutive frames yields optimal results. However, with an increased skip rate, particularly at 7 and 11, there is a noticeable decline in model performance. The reduction in precision, recall, and F1 score indicates that skipping frames between inferences negatively impacts the model's ability to accurately detect anomalies. Notably, at a skip rate of 11, the model encounters a substantial decrease in performance, reaching zero for the geometric mean. This underscores the critical importance of judicious skip rate selection for maintaining optimal anomaly detection capabilities in real-world applications.

Skip Frame Rate	Precision	Recall	F1	Accuracy	Geometric Mean
1	0.925	0.9	0.9033	0.9	0.9258
7	0.9167	0.9	0.899	0.9	0.8944
11	0.6222	0.7	0.6588	0.7	0

Table 5. Performance comparison based on Optimal frame quantity

7. Limitations

1. **Dependence on Human Pose Key Points:** The model's performance is heavily reliant on the quality and availability of Human Pose Estimation (HPE) key points. While this reliance mitigates biases introduced by RGB-based techniques, it also means that contexts captured by RGB information, which could be valuable for making fine-grained decisions, are not utilized.

2. **Additional Step for Real-time Inference:** Due to its reliance on HPE, the model introduces an additional step before inference, potentially affecting real-time implementation. This could impact the model's suitability for applications where low latency is crucial.
3. **Impact of Training Data Diversity:** The model's effectiveness is contingent on the diversity and representativeness of the training data. In scenarios where the dataset lacks comprehensive coverage, the model may struggle to generalize to new, unseen situations as can be seen in the ablation study pertaining to view generalization.
4. **Sensitivity to Occlusions:** As observed, occlusions significantly impact the model's performance. Further exploration and enhancement of the model's robustness to various occlusion scenarios could be beneficial.
5. **Training Challenges for Transformers:** Transformers are known for their implicit heaviness, making them challenging to train. This characteristic could contribute to training difficulties, and careful consideration of computational resources and training strategies is essential.

8. Conclusion

This study addresses the urgent concern of detecting falls among the elderly, recognizing the significant health and psychological repercussions linked to such incidents. Our motivation arises from the crucial necessity for early detection to prevent injuries and enhance the well-being of older individuals.

We introduced an inventive approach employing spatio-temporal transformers for fall detection, surmounting challenges posed by existing methods such as illumination differences, noise, and occlusions. Through the curation of a dataset comprising joint positions in video sequences and harnessing transformers, our model exhibited compelling performance in binary classification, effectively discerning falls from non-falls.

The comparative examination between the solitary transformer model and the enhanced GCN+Transformer model produced noteworthy outcomes. The GCN+Transformer demonstrated superior precision (100.0%), recall (100.0%), F1 score (100.0%), and accuracy (100.0%) compared to the transformer-only model. Notably, the GCN+Transformer exhibited resilience in complex scenarios, including occlusions and diverse viewing angles.

Our investigation into view generalization and occlusion cases offered valuable insights, showcasing the model's adaptability across various perspectives and sensitivity to occlusions. In occlusion scenarios, a nuanced balance between precision and recall emerged, underscoring the necessity for further research to bolster the model's robustness.

However, it is essential to recognize limitations, encompassing the model's dependence on human pose key points, an additional step for real-time inference, and sensitivity to training data diversity. Despite these challenges, the proposed approach presents a promising solution for efficient and accurate fall detection, particularly in contexts emphasizing privacy preservation.

Numerical outcomes underscore the GCN+Transformer's capabilities, highlighting its potential for pragmatic implementation in real-world scenarios.

9. Above & Beyond

We initially proposed utilizing transformers alone for training with Human Pose Data. However, as we progressed, we explored and integrated Graph Convolutional Networks (GCN) that have enhanced spatial encoding abilities to enhance the embedding process. This addition proved to be highly effective, and the resulting GCN + Transformer model emerged as our best-performing model during the project development phase.

References

[1] Ekram Alam, Abu Sufian, Paramartha Dutta, and Marco Leo. Vision-based human fall detection systems using deep learning: A review. *Computers in biology and medicine*, 146: 105626, 2022. 1

[2] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A. dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. *CoRR*, abs/1907.13025, 2019. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. 2

[4] Shomir Chaudhuri, Hilaire Thompson, and George Demiris. Fall detection devices and their use with older adults: a systematic review. *J Geriatr Phys Ther*, 37(4):178–196, 2014. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016. 2

[7] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Convolutional networks for real-time 6-dof camera relocalization. *CoRR*, abs/1505.07427, 2015. 2

[8] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3):489–501, 2014. 3

[9] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based

action recognition using temporal sliding lstm networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1012–1020, 2017. 2

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *CoRR*, abs/1804.06055, 2018. 2

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 3

[12] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. In *arXiv*, 2015. 2

[13] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 1

[14] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021. 1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[16] Cailing Wang and Jingjing Yan. A comprehensive survey of rgb-based and skeleton-based human action recognition. *IEEE Access*, 11:53880–53898, 2023. 2

[17] Haoran Wei and Nasser Kehtarnavaz. Simultaneous utilization of inertial and video sensing for action detection and recognition in continuous action streams. *IEEE Sensors Journal*, 20(11):6055–6063, 2020. 2

[18] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 2023. 1

[19] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *CoRR*, abs/1801.07455, 2018. 2

[20] Qipeng Zhang, Tian Wang, Mengyi Zhang, Kexin Liu, Peng Shi, and Hichem Snoussi. Spatial-temporal transformer for skeleton-based action recognition. In *2021 China Automation Congress (CAC)*, pages 7029–7034, 2021. 2