The Extended Utilization of Lightweight HRNet and VGG16 in Fall Detection

Yuanhang Ren, Ye Tao*, Jinzhen Liao, Xiyan Li, and Wenhua Cui

***Abstract*—With the rapid development of deep learning technology, deep learning-based object detection algorithms have found widespread applications across various fields. In particular, fall detection plays a crucial role in elderly care and safety monitoring. This paper presents a deep learning-based human pose estimation algorithm and applies it to the task of fall detection. The proposed algorithm utilizes the lightweight network Lightweight HRNet to detect human keypoints, ensuring accuracy through high-resolution feature map outputs. Subsequently, a VGG16 network is employed to classify the concatenated images and pose maps, effectively capturing high-level features from the images to identify fall events. Experimental results demonstrate that the proposed algorithm achieves high accuracy in fall detection. Compared with traditional methods, the proposed approach can adaptively learn scene-specific features, thereby enhancing robustness and generalization. These properties meet the practical requirements for real-time and accurate fall detection. Comprehensive experiments validate the effectiveness of the proposed algorithm, providing a feasible solution for applications in elderly care and safety monitoring.**

***Index Terms*—Human Pose Estimation, Lightweight Networks, HRNet, VGG16, Fall Detection**

I. Introduction

With the increasing aging population, fall incidents have become a major healthcare issue for the elderly in contemporary society. Sending an immediate alarm to caregivers when an elderly person falls can prevent further injuries, reduce treatment costs, and enhance recovery opportunities [1]. Approximately 30% of individuals over the age of 65 experience falls annually [2]. Falls are the leading cause of accidental injuries and deaths among the elderly population over 65 years old [3]. Rapid assistance following a fall can reduce the risk of hospitalization by 26% and the risk of death by 80% [4]. Fall detection technology plays a crucial role in identifying falls promptly and enabling timely rescue efforts, thereby reducing the risk of injury for the elderly. Human pose estimation is a key technology in fall detection. By analyzing the changes in human keypoint positions in images or videos, it can determine whether a person is in a fallen state.

In recent years, significant progress has been made in deep learning-based human pose estimation and fall detection methods. Numerous researchers have proposed various complex deep learning models, such as the OpenPose algorithm by Z. Cao et al. [5] and the AlphaPose algorithm by H. Fang et al. [6]. While these models demonstrate excellent accuracy, they are computationally demanding and difficult to deploy in real-time applications. Additionally, some fall detection methods rely on supplementary human object detection tasks, which increase system complexity and latency. To address these challenges, researchers have begun exploring the application of lightweight networks in fall detection.

This paper presents a deep learning-based human pose estimation algorithm and applies it to fall detection. The proposed algorithm utilizes the Lightweight HRNet [7] network for detecting human keypoints, and employs the VGG16 [8] network to classify concatenated images and pose maps for fall status recognition. Experimental results demonstrate that the proposed algorithm achieves high fall detection accuracy while maintaining robust performance, particularly in complex background environments. Compared to traditional methods, the approach based on VGG16 and Lightweight HRNet can adaptively learn scene-specific features, thereby enhancing robustness and generalization ability. This meets the practical requirements for real-time and accurate fall detection.

In recent years, deep learning-based visual fall detection methods have been extensively studied abroad. Kim D. et al. [9] analyzed abnormal behavior of pedestrians in surveillance camera environments, discussing the advantages and disadvantages of various algorithms and how to improve detection performance in real-world applications. They noted that although deep learning methods perform excellently in terms of accuracy, their high computational resource requirements make them difficult to deploy widely in real-time systems. Dubey S. et al. [10] conducted a survey on various 2D and 3D human pose estimation techniques, covering both classical and deep learning approaches. These methods provide solutions to various computer vision problems and also examine different deep learning models

Manuscript received January 5, 2025; revised June 4, 2025.

This work was supported by the National Natural Science Foundation of China (62272093), the Economic and Social Development Research Topics of Liaoning Province (2025-10146-244), the Postgraduate Education and Teaching Reform Research Project of Liaoning Province (LNYJG2024092), and the Undergraduate Innovation Training Program of University of Science and Technology Liaoning.

Yuanhang Ren is an undergraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: 1450920725@qq.com).

Ye Tao is an Associate Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (corresponding author to provide phone: +86-133-0422-4928; e-mail: taibeijack@163.com).

Jinzhen Liao is a graduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China.(e-mail: 2977354475@qq.com).

Xiyan Li is a graduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China.(e-mail: 522053709@qq.com).

Wenhua Cui is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: taibeijack@126.com).

used in pose estimation. Their study highlights the significant application potential of human pose estimation; however, they also emphasize that existing methods still have room for improvement in terms of real-time performance and robustness.

Additionally, Ibrahim M. R. et al. [11] conducted a study on detecting cycling-related accidents using video streams, exploring the combination of video processing techniques and deep learning models, and suggesting future research directions. Their research indicates that real-time video stream processing plays a crucial role in detection; however, it is necessary to address the balance between computational resources and real-time performance. Gao J. et al. [12] proposed several pioneering deep learning models for integrating multimodal big data. Their study found that multimodal data can significantly improve detection accuracy and robustness, although data fusion techniques still require further optimization.

In China, deep learning-based fall detection technology has also received considerable attention. Zhao R. et al. [13] proposed a novel Dynamic Centroid Model (DCM) for detecting abnormal pedestrian behavior in public spaces, particularly focusing on common anomalies such as turning and falling.

Despite significant progress in deep learning-based fall detection both domestically and internationally, several challenges remain. Existing methods often require high computational resources, hindering the feasibility of real-time detection on embedded devices. Although multimodal data fusion techniques can enhance detection performance, the increased complexity of data preprocessing and fusion algorithms complicates system implementation. Additionally, detection performance under varying lighting conditions and complex background environments requires further improvement, as these factors can adversely affect the robustness and generalization ability of the algorithms.

To address the aforementioned challenges, this paper proposes a fall detection method based on the Lightweight HRNet and VGG16 networks. Specifically, the Lightweight HRNet network is utilized for human keypoint detection, ensuring accuracy through high-resolution feature map outputs. The main advantage of Lightweight HRNet lies in its bottom-up approach, which eliminates the need for additional human object detection tasks, thereby offering inherent advantages in detection speed. The VGG16 network is employed to classify the concatenation of images and pose maps, enabling the identification of fall states. The key contribution of VGG16 is its demonstration that increasing network depth can improve performance under certain conditions [14]. Known for its simple architecture and excellent performance, VGG16 effectively extracts high-level features from images, making it a classic model for image classification and object detection tasks. Experimental results show that the proposed algorithm achieves high accuracy in fall detection. This research not only provides technical innovation but also offers a feasible solution for elderly care and safety monitoring. The effectiveness of the algorithm is validated through experiments, and it holds potential for application in various real-world scenarios to enhance the quality of life and safety of the elderly.

II. Related Work

Recent research in fall detection has made significant strides through the use of deep learning models, particularly human pose estimation and image classification. Several approaches have focused on improving the accuracy and efficiency of fall detection systems by combining lightweight networks and human pose estimation techniques.

Human pose estimation is a key component in fall detection as it involves detecting keypoints of the human body to understand its posture. The OpenPose algorithm (Cao et al., 2017) and AlphaPose (Fang et al., 2017) are popular methods in pose estimation, but their high computational costs limit their application in real-time systems. To address this, the use of lightweight networks like Lightweight HRNet has been explored. This network offers efficient pose estimation with reduced computational overhead, making it suitable for real-time applications on embedded platforms.

VGG16, a classic convolutional neural network (CNN), has been widely used for image classification tasks, including fall detection. In recent work, combining VGG16 with pose estimation maps has shown improved performance, as the network can leverage both visual features from images and spatial information from pose maps for more accurate classification. The integration of these two sources of data allows for a more robust fall detection system, especially in complex environments with varying conditions.

Fall detection systems are increasingly being deployed on embedded devices like the NVIDIA Jetson Xavier NX. These platforms offer powerful computational capabilities for real-time processing, which is essential for fall detection applications. Lightweight networks like HRNet and VGG16 are optimized for these systems to ensure that the models can run efficiently without compromising on performance, making real-time detection feasible even in resource-constrained environments.

This paper builds upon these prior works by proposing a novel combination of Lightweight HRNet for pose estimation and VGG16 for fall classification. The proposed system aims to achieve high accuracy while maintaining real-time performance, making it suitable for practical deployment in elderly care and safety monitoring systems.

III. Proposed Method

A. System Overview

The proposed fall detection system is based on deep learning and human pose estimation, leveraging the Lightweight HRNet and VGG16 networks. The overall system architecture is designed to accurately detect falls using input from images and human pose maps, while ensuring efficient real-time performance.

The system operates in the following sequence: first, the input image or video frame is processed using the Lightweight HRNet to detect human keypoints. The Lightweight HRNet is a lightweight network specifically designed for real-time human pose estimation, offering efficient keypoint detection without the need for additional human object detection tasks. This approach significantly reduces the computational complexity compared to traditional methods, making it ideal for embedded systems with limited resources.

Once the human keypoints are detected, the pose map is generated. This pose map is then concatenated with the original input image, resulting in a combined feature map. The concatenated image and pose map are passed through the VGG16 network for classification. VGG16, a well-established deep learning model known for its simple and effective architecture, classifies the combined feature map to determine whether the person is in a fall state or standing.

The system’ s main advantage lies in its combination of a lightweight pose estimation model with a classic deep learning classifier. This hybrid approach allows the system to efficiently detect falls while maintaining high accuracy in diverse real-world scenarios. Additionally, the use of a lightweight network such as Lightweight HRNet ensures that the system can run in real-time on embedded devices like the NVIDIA Jetson Xavier NX, which is used in this study for practical deployment.

In summary, the system consists of three primary components: (1) the Lightweight HRNet for pose estimation, (2) the VGG16 network for fall classification, and (3) a real-time detection pipeline running on the Jetson Xavier NX platform. The system is designed to be both computationally efficient and accurate, meeting the needs of real-time applications in elderly care and safety monitoring.

B. Lightweight HRNet for Pose Estimation

In recent years, the rapid advancement of deep learning has led to significant improvements in human pose estimation technology. Despite these developments, traditional pose estimation networks generally require considerable computational resources and large amounts of storage, which limits their applicability in real-time environments where efficiency is crucial. To overcome these challenges, researchers have introduced lightweight pose estimation networks. These networks are specifically designed to minimize computational load and storage requirements, while striving to maintain high levels of accuracy in pose estimation tasks. By using such networks, it becomes feasible to implement pose estimation models on devices with limited resources. The architecture of the Lightweight HRNet network, which embodies these advancements, is illustrated in Figure 1.

The Lightweight HRNet, a lightweight network based on the HRNet architecture, has gained widespread attention due to its excellent performance and efficiency. Proposed by Jinzhen Liao et al. in 2024, the Lightweight HRNet is an efficient network designed specifically for bottom-up multi-person human pose estimation tasks. By employing a multi-resolution, multi-branch parallel network architecture, Lightweight HRNet enables real-time human pose estimation. Its main advantage lies in the bottom-up approach, which eliminates the need for additional human object detection tasks [7], thereby providing an inherent advantage in detection speed. The Lightweight HRNet shows great potential for real-time application scenarios, such as fall detection.

The key advantage of the Lightweight HRNet is its ability to detect human keypoints directly from the input image, without the need for additional human object detection tasks. This simplifies the system by eliminating the need for separate object detection algorithms, which are commonly used in other pose estimation models. By directly estimating keypoints, the network speeds up the process and reduces the computational overhead. The architecture of the Lightweight HRNet consists of multiple parallel branches operating at different resolutions, which are then fused to produce a final pose map. This multi-branch, multi-resolution design is essential for capturing both fine-grained details and global context, which is critical in detecting human falls, especially in crowded or complex environments.

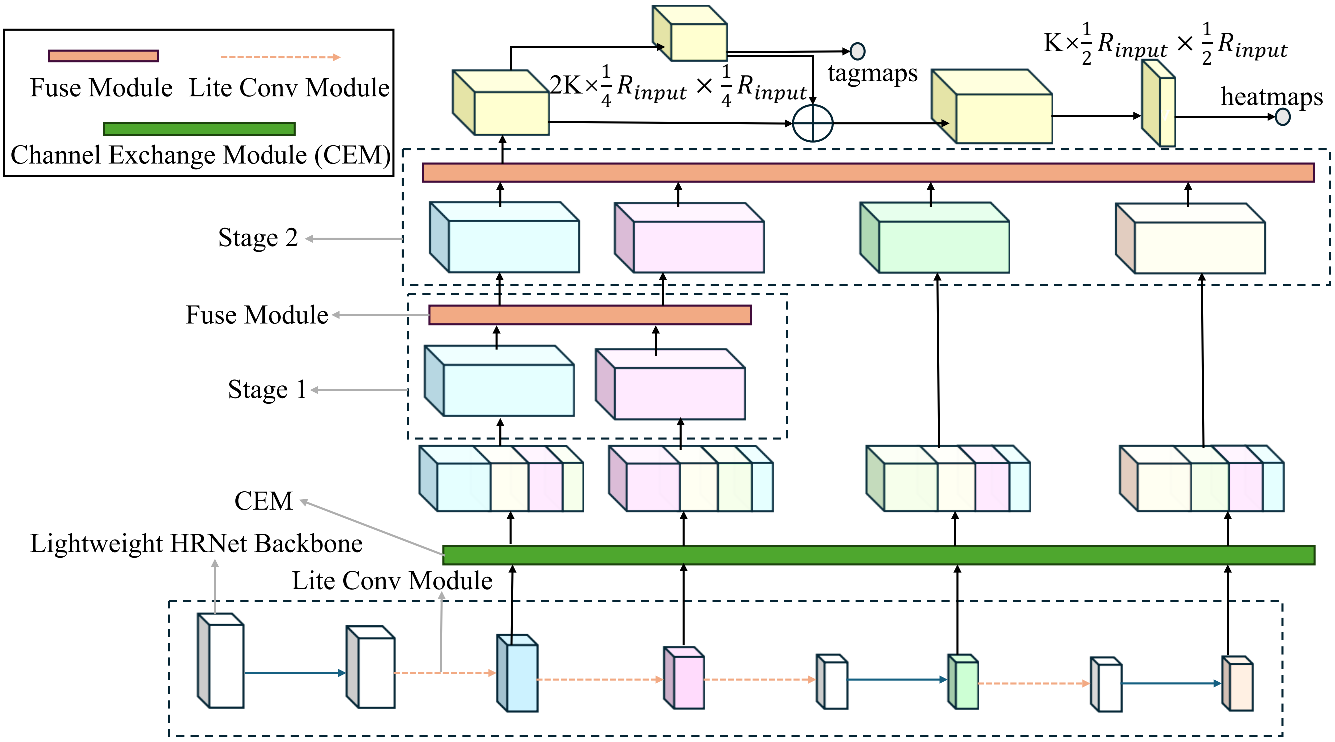

Fuse Module
Lite Conv Module
Channel Exchange Module (CEM)
$2K\times\frac{1}{4}R_{input}\times\frac{1}{4}R_{input}$
tagmaps
$K\times\frac{1}{2}R_{input}\times\frac{1}{2}R_{input}$
heatmaps
Stage 2
Fuse Module
Stage 1
CEM
Lightweight HRNet Backbone
Lite Conv Module

Fig. 1. Lightweight HRNet architecture

In our system, the pose map generated by the Lightweight HRNet represents the human body's keypoints and their respective locations in the image. These keypoints are used to identify the human pose, and any significant changes in the pose, such as falling, can be easily detected. By leveraging the efficient design of the Lightweight HRNet, the system can estimate human poses in real-time, which is essential for timely fall detection.

The Lightweight HRNet offers a remarkable balance between strong performance and enhanced computational efficiency, making it ideal for use in fall detection systems deployed on embedded devices with limited resources. Its ability to detect human keypoints quickly and accurately, importantly without the need for any extraneous additional detection tasks, makes it a key and indispensable component of our proposed method.

C. VGG16 Model

The VGG network employs the technique of stacking multiple 3 × 3 convolution filters to replace larger filters, which not only increases the depth of the network but also helps in reducing the total number of parameters, thus improving the computational efficiency without sacrificing performance [8]. The VGG16 model, specifically, is composed of five convolutional blocks, three fully connected layers, and one final softmax layer. The network is organized into five distinct stages, each containing several convolutional layers followed by a 2 × 2 max-pooling layer. More specifically, the first two stages each have two convolutional layers, whereas the remaining three stages contain three convolutional layers. This structure ensures a gradual extraction of more abstract and complex features from the input images, starting from basic low-level features such as edges and textures and progressing to more high-level object-related features.

The architecture of VGG16 is designed with the aim to capture increasingly complex patterns through its layered convolution process. After the convolutional layers and pooling operations, the model utilizes three fully connected layers for image classification. In the final layer, the softmax function is applied, which generates a probability distribution for the possible classes, with each node representing the probability of a particular class being the correct classification. Despite the fact that the model contains a substantial number of parameters — approximately 138 million—and requires significant computational resources for training, VGG16 has proven to be highly effective in various tasks such as image classification and object detection. It has become a foundational model for those starting in deep learning, often serving as a benchmark for comparison in numerous computer vision tasks.

The depth and design of VGG16 allow it to capture intricate details from images and perform exceptionally well on large-scale image recognition challenges. Although the model's complexity may require powerful hardware to train efficiently, its ability to generalize well across diverse datasets has made it an invaluable resource in the field of computer vision. Moreover, VGG16's architecture and performance have influenced the development of subsequent models, demonstrating its lasting impact on the field.

The STM32 series microcontrollers, which are based on the ARM Cortex-M core, offer robust support for the deployment of neural networks, making them suitable for applications such as fall detection. In this context, the fall detection task is formulated as a multi-label classification problem, where the VGG16 neural network is utilized to classify input images. The architecture of the VGG16 network, as depicted in Figure 2, consists of multiple layers designed to extract increasingly complex features from the input image. The input image is standardized to a fixed size of 224×224 pixels, ensuring consistency across all processed data. During the convolutional process, the network exclusively employs 3 × 3 convolution kernels that are stacked together to form the layers. This design choice allows for the construction of a deeper network while minimizing the increase in the number of parameters and computational complexity. This approach helps strike a balance between the network's depth and its efficiency. The depth of the network plays a significant role in improving performance, as a deeper network has the capacity to learn more intricate and abstract representations of the input data.

The 3 × 3 kernel size is particularly advantageous for capturing local features in the image, such as edges, textures, and other low-level features. These features are essential in many image processing tasks, as they help the model identify key elements of the image. The use of this kernel size is not only common but also proven to be highly effective in a variety of deep learning networks. Furthermore, the network incorporates pooling layers to reduce the spatial resolution of the feature maps, as opposed to using convolution with a stride of 2. This design choice helps to minimize the network's computational load and reduces the total number of parameters, making the network more efficient. Pooling layers, particularly max-pooling, are known to be effective in extracting the most relevant features from the feature maps by selecting the maximum value within local regions. This process serves to retain the most important information while discarding irrelevant details and noise. As a result, the network can focus on the most salient features in the image, enhancing its ability to make accurate predictions. In addition, pooling layers increase the translational invariance of the learned features. This means that slight variations in the position of features within the input image have minimal impact on the network's output. Such translational invariance is crucial in tasks like object recognition, where objects may appear at different locations in the image.

The VGG16 network's terminal phase, for classification results, employs several fully connected layers as depicted in Figure 2. These layers adeptly process high-level features from earlier convolutional stages. This procedure generates an output feature map, a vector of size n, where n denotes the distinct image categories. The VGG16 architecture's recognized simplicity, its computational efficiency, and ease of implementation have established it as a foundational, reliable model in deep learning. Consequently, it has been widely applied in tasks like image recognition and classification, owing to its robust performance and user-friendliness. Given these merits and its proven visual understanding, employing VGG16 for accurate fall detection using static image analysis presents an appropriate and scientifically validated effective solution.

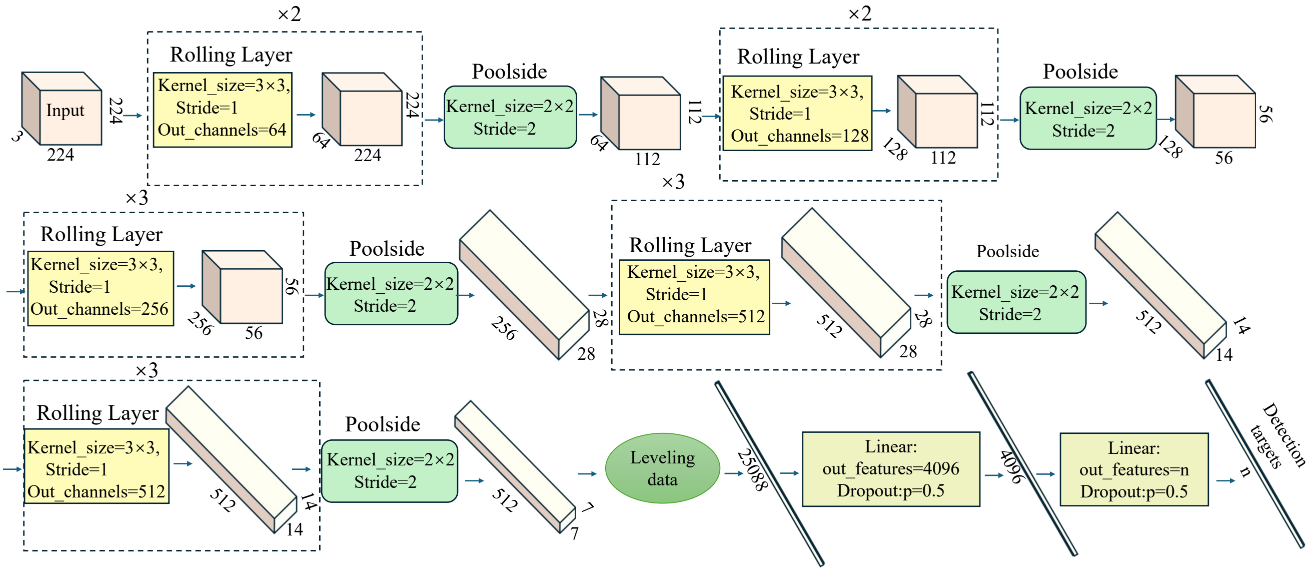


Fig. 2. Network architecture of VGG16

D. Jetson Xavier NX

The NVIDIA Jetson series encompasses low-power embedded platforms designed for GPU-accelerated computing [15]. These systems offer powerful computational performance and are specifically engineered to support applications in deep learning, computer vision, and artificial intelligence (AI). They feature independent CPU, GPU, PMIC, and flash memory, making them ideal for real-time detection tasks on local devices. The Jetson Xavier NX, one of the platforms in the Jetson series, is equipped with the NVIDIA Volta GPU and a six-core ARM Cortex-A57 CPU, providing substantial computational power within a compact module. It is designed for deep learning and AI applications on edge devices and is suitable for applications requiring high-performance deep learning inference. This platform can be applied in lightweight human behavior recognition tasks and can also be used for object detection tasks. As shown in Figure 3.

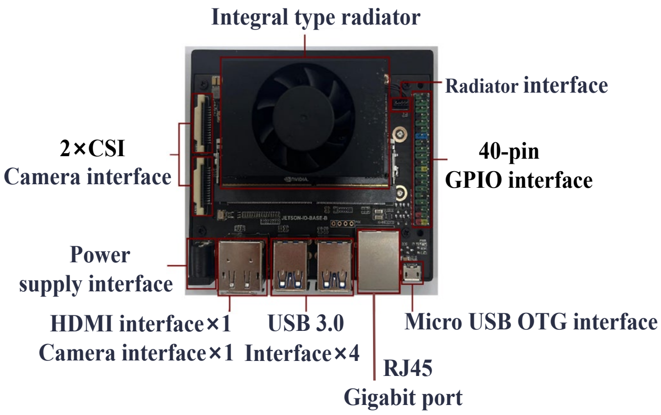


Fig. 3. Jetson Xavier NX

The Jetson Xavier NX features two CSI camera interfaces, ensuring the completion of real-time computer vision and image processing tasks. It also includes an HDMI interface, enabling the device to connect to a display for real-time monitoring of program execution. Additionally, the system is equipped with four USB interfaces, allowing users to connect devices such as USB drives, keyboards, and mice for easy operation within the system. The RJ45 gigabit Ethernet port facilitates network connectivity, while a Micro USB OTG interface is included to support a wider range of device connections.

The proposed fall detection system is implemented on the NVIDIA Jetson Xavier NX, a powerful embedded platform designed for high-performance computing and real-time processing. This platform is equipped with a Volta GPU, a six-core ARM Cortex-A57 CPU, and 8GB of LPDDR4x memory, which makes it suitable for running deep learning models with significant computational requirements. The Jetson Xavier NX is particularly well-suited for applications in edge computing, as it enables efficient deployment of AI models without the need for cloud-based processing.

The fall detection pipeline leverages the powerful Jetson Xavier NX for robust real-time performance, skillfully integrating Lightweight HRNet for human pose estimation (detecting crucial body keypoints) and VGG16 for fall classification ("fall" or "normal"). The entire streamlined pipeline, from initial image capture to final classification, executes on the Jetson Xavier NX, ensuring consistently low-latency detection.

Input images/frames are captured via a connected camera; Lightweight HRNet then processes these to extract keypoints and generate pose maps, crucial for understanding posture and dynamic movement.

The pose map is then strategically concatenated with the original image, creating a combined feature map enriched with visual and pose information. This map is passed to VGG16 for definitive classification; its output probability distribution for the two classes determines the final prediction.

Real-time operation is ensured by several key optimizations like efficient models (e.g., Lightweight HRNet, minimizing overhead while maintaining accuracy) and Jetson Xavier NX's GPU acceleration for both pose estimation and fall classification, enabling quick real-time response to incidents.

This Jetson Xavier NX pipeline effectively integrates these models for accurate, real-time fall detection, enabling efficient edge deployment in elderly care, healthcare monitoring, and other similar safety-critical environments requiring immediate alerts.

IV. Experiments and Results

A. Dataset construction

During the initial development of our fall detection method, a static image dataset was constructed as a foundational step to facilitate neural network training, offering a controlled environment for early model experimentation. The original dataset was sourced from the PaddlePaddle website, but it required significant and careful refinement before it could be effectively utilized for robust model development. This raw dataset, although extensive, was unfortunately highly disorganized, primarily because it was aggregated from diverse online sources, leading to a lack of cohesion. The data annotations within the original dataset were often inconsistent, with bounding boxes marking individual locations and labels indicating fallen or standing states sometimes varying in accuracy or format. To further improve its quality and comprehensiveness, a small portion of the dataset was meticulously supplemented by manually adding carefully selected images sourced from other existing video-based fall detection datasets.

The images in these datasets depict a truly wide variety of highly complex scenes, with each image potentially featuring multiple individuals simultaneously at once, sometimes in rapid motion. These images were captured in two distinct primary contexts: athletes falling during various physically demanding sports activities and pedestrians experiencing falls in often unpredictable everyday urban environments. The majority of the images in the dataset were obtained from dynamic news coverage and often low-resolution surveillance footage, which greatly contributed to the sheer variety of the scenes captured. However, the dataset presented several notable challenges, as it included images with crowded scenes, often severe occlusions, and significant variations in the apparent size of individuals, making the task of recognizing falls particularly difficult indeed. Given these inherent complications, it was evident that the original dataset could not be directly applied to the model training without very substantial and meticulous preprocessing.

TABLE I
THE DISTRIBUTION OF THE DATASET

Category	Number of Images (Count)
Training Set	2,839
Validation Set	1,217
Total	4,056

As a result, a series of preprocessing steps were undertaken to prepare the dataset. Initially, individuals were cropped from the images according to the bounding box annotations. Subsequently, the images were manually reviewed to ensure that only clear, unobstructed images were retained, depicting either a fallen person or a person in a normal, upright position. After this selection process, the cropped and cleaned images were systematically renamed using a format that combined the state (either "fall" or "normal") with a serial number (e.g., "fall.00001" for the first image of a person in a fallen state). Following this extensive preprocessing, a total of 4,056 images depicting falls were obtained.

Due to the limited size of the dataset, no separate test set was created. Instead, the data was divided into a training set and a validation set at a ratio of 7:3, yielding 2,839 images for training and 1,217 images for validation. The distribution of the dataset is summarized in Table I.

Some manually processed fall images from the dataset are shown in Figure 4.

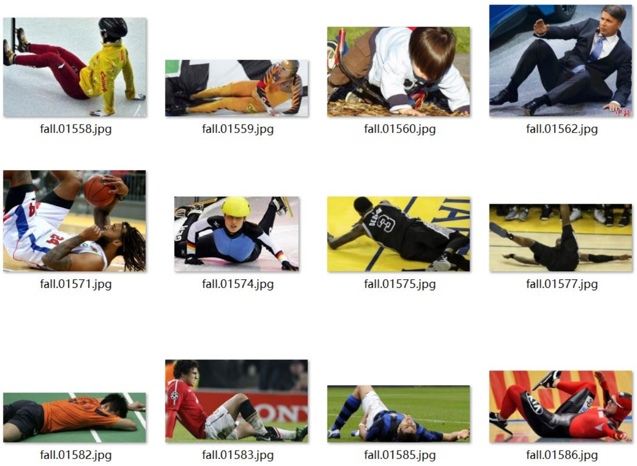


Fig. 4. Illustration of the processed partial fall dataset

B. The training of the network

The model was trained on an NVIDIA GeForce RTX 3060, based on the computing power of the device. The number of samples selected for the training set was 32, and the learning rate was set as 0.00001. Due to the smaller dataset, the model was trained to prevent overfitting by setting the number of training epochs to 30. The loss function selected was the mean squared error, which is commonly used in classification tasks. It is assumed that the network outputs an image with two classes (fall, normal), and it performs classification based on each input image. The network outputs a two-dimensional vector of predicted values, where in Figure 5, n = 2, and each dimension represents the output of the respective class predicted by the network. The ground truth bounding box for each image is represented as a two-dimensional vector, where the corresponding class is set to 1, and the other class is set to 0. If the image belongs to a particular class, the ground truth label for that class is set to 1.

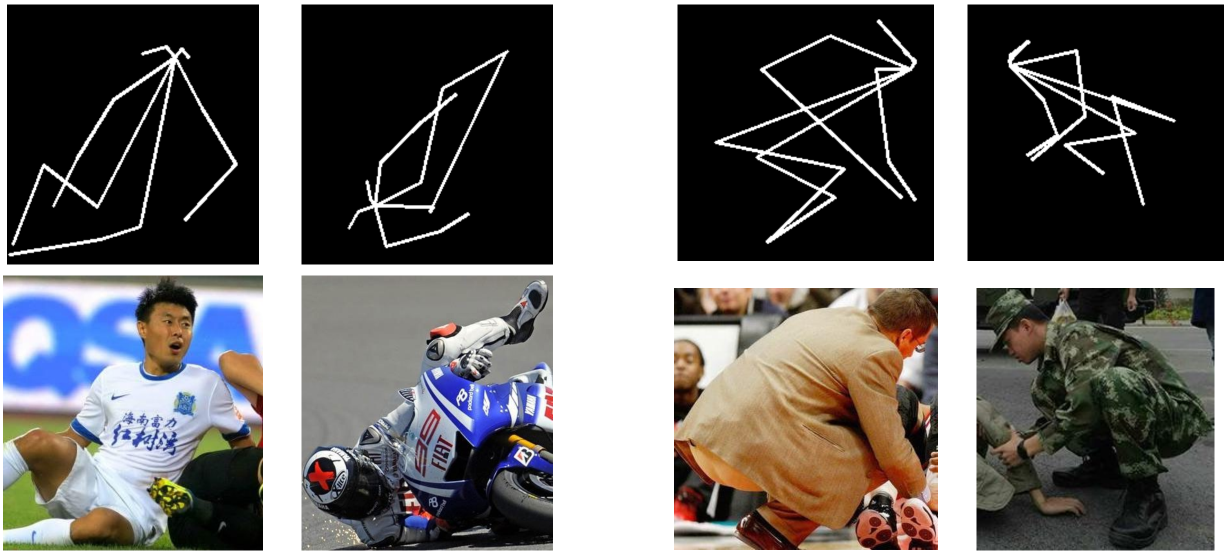
the experimental setup used in this study is summarized in Table II.

TABLE II
THE EXPERIMENTAL SETUP

Parameter Name	Parameter Settings
epochs	30
batch	32
workers	0
imgsz	224
optimizer	Adam
learn_rate	0.00001

The loss function is calculated using the square of the Euclidean distance between the predicted label and the ground truth label, as explicitly detailed in equation (1). This specific choice, the squared error loss function, was strategically employed to more significantly reduce large discrepancies between predictions and actual values, thus improving the overall precision and reliability of the classification task.

$$\mathrm{Loss}(x) = [g(x) - f(x)]^2 \qquad (1)$$



a) Illustration of the fall state dataset

b) Graphical representation of the non-fall state dataset

Fig. 5. Illustration of posture graphs corresponding to some data sets

As mentioned earlier, during the experiment, not only were images used as input in the VGG16 network for the experiment, but also the corresponding pose maps of the images were input into the network for the experiment. Additionally, the pose maps corresponding to the images were input together with the images into the network for the experiment. In the training set, the pose maps of the characters were generated using HRNet-W48, and some of the dataset corresponding pose maps are shown in Figure 5. The pose maps were set as one-dimensional, so if the pose map corresponding to the image was input together with the image into the network, the pose map would be concatenated with the network, and finally a 4×224×224-sized feature map would be formed and input into the network for training.

C. Evaluation Metrics

Accuracy represents the overall proportion of correctly classified instances out of the total instances in the dataset. It is a widely used metric but can be less informative in the case of imbalanced datasets, where one class may dominate. In the context of fall detection, accuracy measures how often the model correctly identifies both falls and normal instances. As shown in equation (2).

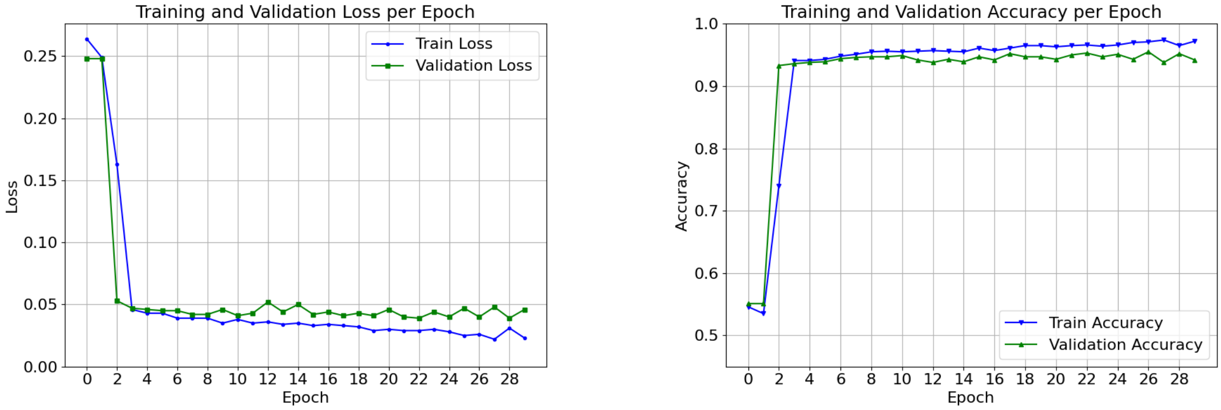
$$\text{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Instances}} \tag{2}$$

Loss is a measure of the error between the model's predicted outputs and the actual class labels. In this study, the cross-entropy loss function was used to calculate the discrepancy between the predicted probability distribution and the true class labels. The loss value is minimized during training, with a lower loss indicating better model performance. As shown in equation (3).

$$\text{Loss} = -\sum \text{True Labels} \times \log\text{(Predicted Probabilities)} \tag{3}$$

D. Input Modality Comparison

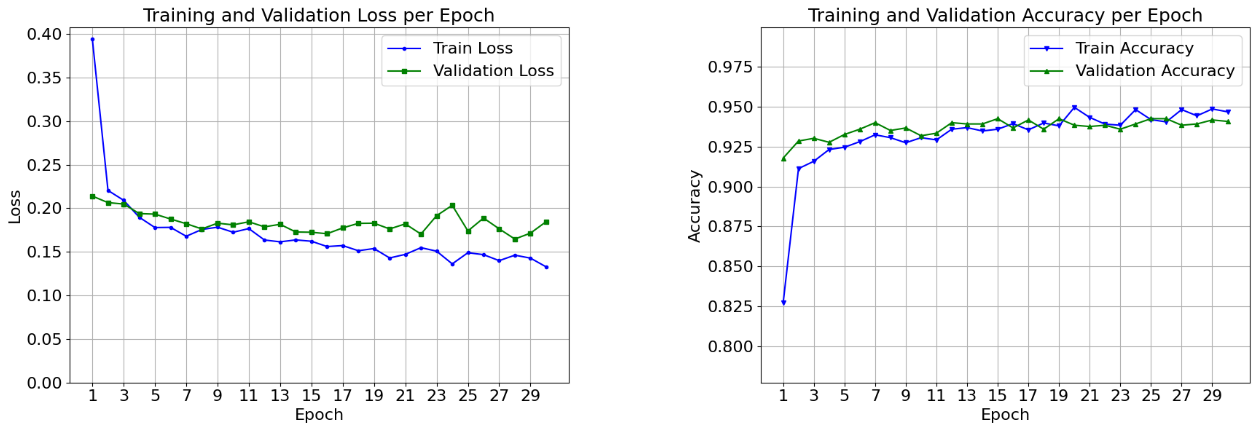
This section presents a detailed comparative analysis of the fall detection system's performance utilizing distinct input modalities, to rigorously assess their individual and combined effectiveness. The primary objective is to quantitatively and qualitatively evaluate the precise influence of different input data representations—specifically, raw RGB images, which capture rich visual context, derived pose maps, providing crucial skeletal abstraction, and their strategic concatenation—on the accuracy and operational robustness of fall detection across diverse challenging scenarios. This investigation provides valuable insights into the relative contributions of visual appearance cues and skeletal structure information to the classification task, thereby informing feature engineering and subsequent model optimization. The underlying architecture employs the robust Higher-HRNet for human pose estimation and a classification network based on an EfficientNet-B0 backbone, as detailed within the experimental setup and methodology section.



a) Loss Curve (RGB Image Input Only)

b) Accuracy Curve (RGB Image Input Only)

Fig. 6. Training Dynamics for Baseline Classification using RGB Image Input.



a) Loss Curve (Pose Map Input Only)

b) Accuracy Curve (Pose Map Input Only)

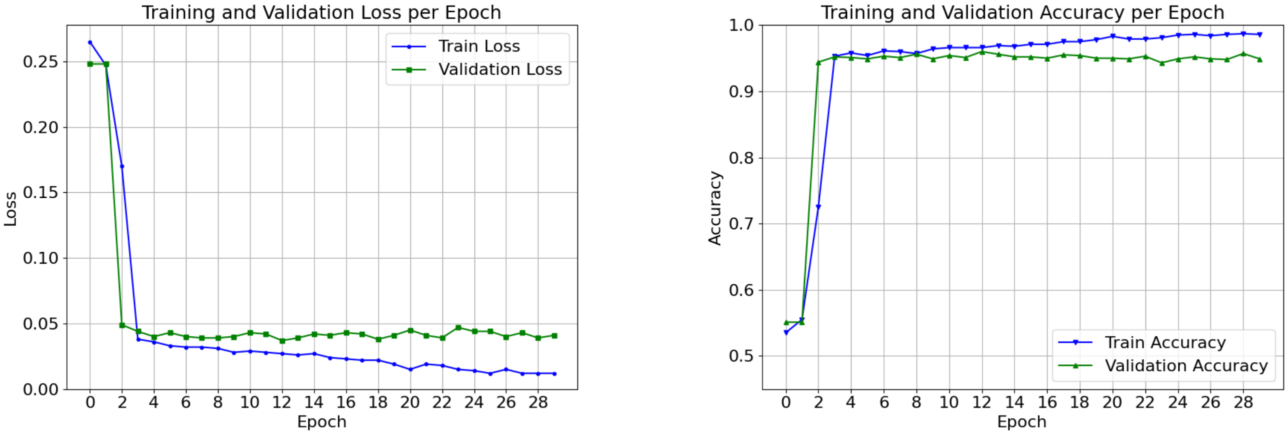
Fig. 7. Training Dynamics for Classification using Pose Map Input Only

In the first experimental configuration, only raw RGB images serve as input to the classification network. While human pose features are implicitly extracted by the Higher-HRNet module during preprocessing (as it's part of the overall pipeline design), the explicit pose map is not provided to the classification stage in this modality. This setup establishes a baseline performance assessment relying solely on visual features learned directly from the pixel data by the classifier. The model must discern fall states based on learned representations of shape, texture, context, and implicit postural cues within the image. The training dynamics and evaluation results for this modality are depicted in Figure 6 a) and Figure 6 b). While capable of capturing visual indicators, this approach may face challenges in scenarios with visual ambiguity, complex backgrounds, or significant variations in subject appearance, where explicit pose information could be beneficial.

The second configuration utilizes only the pose map, generated by Higher-HRNet, as the direct input to the classification network. This modality isolates the contribution of skeletal keypoint information, representing the subject's posture and spatial configuration. By focusing exclusively on the relative positions and connections of body joints, the model can potentially achieve robustness against variations in clothing, illumination, and background clutter that might affect the image-only approach. The performance characteristics under this condition are illustrated in Figure 7 a) and Figure 7 b). A potential limitation, however, is the exclusion of potentially valuable visual context and fine-grained details present in the raw image that might aid in differentiating ambiguous poses.

The third configuration investigates a hybrid approach wherein the raw RGB image and its corresponding pose map are concatenated, typically channel-wise, to form a combined input tensor for the classification network. This strategy aims to synergistically leverage both the rich visual details inherent in the image and the explicit structural information provided by the pose map. The hypothesis is that this fusion allows the model to develop a more comprehensive understanding of the scene, integrating appearance cues with explicit postural geometry for enhanced fall detection accuracy. The training behavior and evaluation metrics for this combined modality are presented in Figure 8 a) and Figure 8 b).

The experiment evaluated three distinct input modalities: firstly, individual raw images; secondly, pose maps from the HRNet-w48 network; and thirdly, a synergistic combination of images with their pose maps. Final validation results from the dataset are presented in Table III. Findings demonstrate that concatenating images and pose maps yielded optimal fall detection performance, achieving 96.0% accuracy on the validation set. These results align with prior analytical work, robustly confirming that pose maps significantly aid the network in more accurately and reliably detecting individual fall states.



a) Loss Curve (Combined RGB + Pose Map Input)

b) Accuracy Curve (Combined RGB + Pose Map Input)

Fig. 8. Training Dynamics for Classification using Combined RGB Image and Pose Map Inputs

TABLE III
EVALUATING THE IMPACT OF DIFFERENT INPUTS ON FALL DETECTION

Input	Enter the picture dimensions	Detection accuracy
Figure diagram	224	95.2%
Pose chart	224	94.0%
Person map + pose map	224	96.0%

These findings empirically validate the hypothesis that combining visual and explicit pose information yields the most effective fall detection system within the tested framework. The pose map itself provides substantial discriminative power, outperforming the image-only approach, confirming the critical role of skeletal structure analysis in this task. The synergistic effect observed in the concatenated modality suggests that the classifier effectively utilizes both sources of information to achieve higher accuracy and robustness.

E. Baseline and Ablation Experiments

To rigorously and systematically evaluate the multifaceted performance and operational efficiency of the advanced proposed fall detection system, which distinctively integrates the Lightweight HRNet for robust human pose estimation and the VGG16 network for subsequent, accurate fall classification, a comprehensive and meticulously designed set of baseline comparisons was diligently conducted. The primary baseline system, specifically selected for this in-depth comparative analysis, strategically employs a standard, yet computationally more intensive, pose estimation network, namely the HRNet-W48, in direct conjunction with the identical VGG16 classifier architecture. This carefully chosen experimental configuration facilitates a direct, unambiguous, and highly focused assessment of the tangible computational and memory-related benefits derived specifically and exclusively from the strategic utilization of the lightweight pose estimation component within the proposed framework. The key quantitative metrics employed for this critical comparison, which are meticulously detailed and presented in Table IV, encompass the total number of parameters (Params), the requisite floating-point operations (FLOPs), and the overall model size (FP32). These metrics are considered critically indicative and essential for determining the practical deployment feasibility of the system on resource-constrained edge devices.

The comparative analysis, quantitatively detailed in Table IV, elucidates the substantial architectural advantages of the proposed fall detection system, particularly regarding its enhanced computational efficiency and optimized resource utilization when contrasted with a baseline that employs a heavier pose estimation model. The cornerstone of these significant efficiency gains is the strategic incorporation of Lightweight HRNet for the crucial initial human pose estimation stage. This lightweight network demonstrates a remarkable reduction in parametric complexity, requiring only 5.9M parameters versus approximately 63.6M for HRNet-W48. Correspondingly, its computational load is significantly diminished, at 5.5G FLOPs compared to an estimated 63.7G FLOPs for HRNet-W48 when processing comparable 448x448 inputs. Consequently, the model size for the pose estimation component is also drastically reduced, from about 254.4MB for HRNet-W48 to a mere 23.6MB for Lightweight HRNet. These pronounced reductions in both parameters and FLOPs directly translate to a considerably higher potential inference throughput for the pose estimation component of the proposed system.

While both the proposed and baseline systems utilize the VGG16 network for the final fall classification stage—a network architecturally characterized by its considerable parameter count of approximately 138.36M and a moderate computational requirement of 15.5G FLOPs for 224x224 input—the pivotal differences in the underlying preceding pose estimation stage translate directly to notable distinctions in overall system metrics and operational efficiency. The proposed system, representing an efficient synergistic combination of Lightweight HRNet and VGG16, totals 144.26M parameters and 21.0G FLOPs, resulting in an FP32 model size of approximately 577.0MB. In stark contrast, the baseline system (HRNet-W48 + VGG16) aggregates to a significantly larger 201.96M parameters and a considerably more demanding 79.2G FLOPs, with an estimated model size of 807.8MB, highlighting its greater resource requirements. These figures unequivocally demonstrate a substantial reduction of approximately 28.6% in total parameters and a highly significant, practically beneficial 73.5% decrease in total FLOPs for the proposed system, a critical advancement for applicability. Such considerable reductions in computational complexity and model footprint not only facilitate faster end-to-end inference times but also markedly enhance the system's suitability for deployment on resource-sensitive edge computing platforms, such as the Jetson Xavier NX, where computational capabilities and power are inherently constrained. The successful implementation and documented real-time functionality on this specific edge platform further substantiate this inference regarding its practical viability. Moreover, ablation insights from varying input modalities consistently reinforce the intrinsic value of high-fidelity pose information extracted by Lightweight HRNet, as the combined image and pose map input yielded optimal classification performance, thereby robustly validating the superior efficacy of the chosen lightweight pose estimator within the integrated system.

TABLE IV
. COMPARATIVE ANALYSIS OF FALL DETECTION SYSTEM COMPONENTS AND OVERALL PERFORMANCE

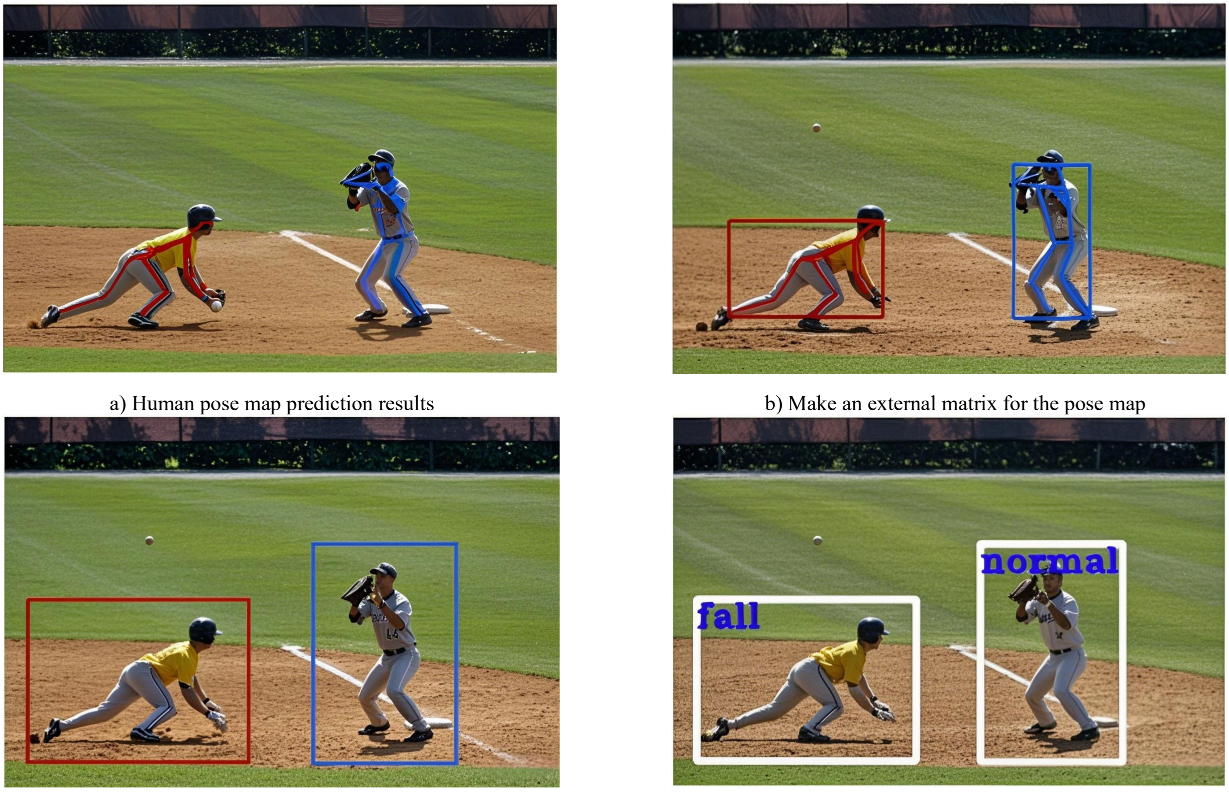
COMPONENT/SYSTEM	NETWORK ARCHITECTURE	INPUT SIZE (POSE/CLASS.)	PARAMS (M)	FLOPs (G)	MODEL SIZE (MB, FP32)
POSE ESTIMATION	LIGHTWEIGHT HRNET (R=-2)	448X448	5.9	5.5	23.6
POSE ESTIMATION	HRNET-W48	448X448	63.6	63.7	254.4
CLASSIFICATION	VGG16	224X224	138.36	15.5	553.4
CLASSIFICATION	VGG16	224X224	138.36	15.5	553.4
OVERALL SYSTEM	LWHRNET + VGG16	448X448 / 224X224	144.26	21	577
OVERALL SYSTEM	HRNET-W48 + VGG16	448X448 / 224X224	201.96	79.2	807.8

F. The implementation of fall detection

Based on the experimental results above, the input of pose maps has significantly assisted the network in performing the fall detection task. Therefore, for the fall detection implementation, Lightweight HRNet is employed to detect the pose maps of individuals. These pose maps are then concatenated with the images and input into the network for the final detection of fall events, leveraging both structural and visual cues. Since Lightweight HRNet utilizes the AE algorithm for human pose detection, issues related to the duplication of keypoint detection may arise. To mitigate this, preprocessing techniques are applied to the detected keypoints, improving the visualization quality of the generated pose maps. This step ensures that the fall detection system remains robust, efficient, and accurate under varying environmental conditions. Furthermore, the AE algorithm plays a critical role in detecting the keypoints of the human body in the input image, which are then classified to successfully accomplish the human pose detection task. Notably, this algorithm does not automatically generate human detection boxes, focusing solely on the accurate identification of key body landmarks. In contrast, the trained VGG16 network specifically addresses the detection of fall statuses in single-person images, offering a more specialized approach to fall recognition. To enhance the model's versatility and accuracy in complex scenarios, the system has been designed to detect and locate all humans present in the input image. This is achieved by utilizing the pose map, which is generated by the network. As illustrated in Figure 9, the system is able to integrate multiple human detections, significantly improving the robustness of the fall detection process in environments with multiple individuals.

After preprocessing the keypoints, non-overlapping and accurately predicted single-person poses are selected, as shown in Figure 9 a). Next, the pose map is used for human localization by constructing a bounding box around the keypoints, as shown in Figure 9 b).

The bounding box is determined based on the keypoints, effectively encapsulating all the detected keypoints. However, in some cases, certain bounding boxes may fail to cover the entire human body. To address this issue, the bounding boxes are expanded as necessary to ensure that all detected humans are fully encompassed, as depicted in Figure 9 c). Once the bounding box is appropriately adjusted, individual human frames and their corresponding pose maps are extracted and concatenated. These processed data are then input into the VGG16 network for fall detection, and the corresponding results are displayed in Figure 9 d).



a) Human pose map prediction results

b) Make an external matrix for the pose map

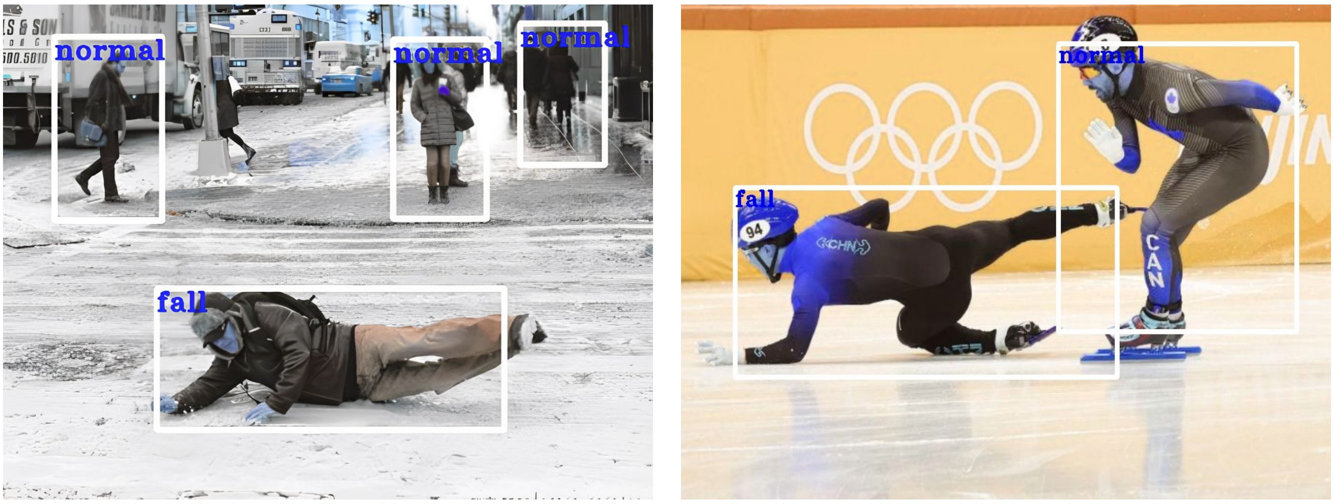
c) Extend the external matrix of the attitude map

d) Fall detection results

Fig. 9. Realization process of fall detection

G. Visual Analysis and Failure Cases

This section provides an in-depth qualitative visual analysis of the proposed fall detection system's operational efficacy, critically evaluating its real-world performance and practical utility through meticulous visual evidence. It examines diverse successful detection instances from particularly challenging scenarios (e.g., varied lighting, occlusions) and representative failures. This granular visual inspection serves to develop a nuanced understanding of system behavior under varying conditions, clearly showing its strengths, inherent limitations, and specific actionable paths for future algorithmic refinement and system enhancement. The findings from this visual audit contribute directly to the ongoing iterative effort to build more robust, adaptable, and generalizable intelligent surveillance systems.



a) Fall Detection in a Complex Multi-Person Street Scene b) Fall Detection in a Specialized Sporting Domain (Ice Skating)

Fig. 10. Examples of Fall Detection and Normal Action Classification in Complex and Specialized Application Scenarios

In baseline scenarios characterized by optimal viewing conditions—specifically, a clear, unobstructed perspective of the subject with adequate illumination and minimal background clutter—the system reliably detects fall incidents. Under these favorable circumstances, the Higher-HRNet component consistently generates accurate and complete pose maps, capturing the subject's keypoints with high fidelity. The subsequent classification stage, leveraging the combined input of the raw RGB image and its corresponding high-quality pose map, then confidently and correctly identifies the fall state, establishing a strong benchmark for the system's core performance.

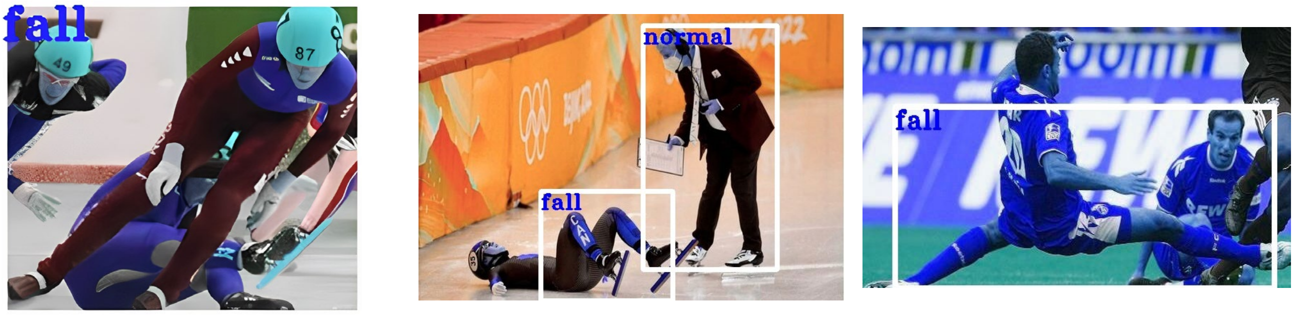
Complex Multi-Person Environments: The system exhibits robustness in complex scenes containing multiple individuals. As illustrated in Figure 10 a), depicting a street scene with several pedestrians, the system successfully isolates and detects the fall event of one individual while correctly classifying others as 'normal'. The pose map distinctly represents the keypoints of all individuals, enabling the classifier to differentiate the fallen person based on the combined image and pose features. This highlights the discriminative power gained from incorporating pose context.

Specialized Domains and Non-Fall Actions: The system's applicability extends to specialized domains, such as sports. Figure 10 b) demonstrates successful fall detection in an ice rink setting. Furthermore, the system accurately distinguishes dynamic, non-fall actions, such as athletic movements in soccer, classifying them correctly as 'normal'. This capability underscores the model's ability to interpret pose dynamics beyond simple static postures.

Despite the generally robust performance, the system exhibits limitations under certain challenging conditions. Analyzing these failure cases is crucial for targeted improvements. Such critical scrutiny not only reveals the current boundaries of the system's capabilities but also highlights specific interactive elements that contribute to errors.

Occlusion: Performance degradation occurs in environments where the subject is partially or significantly occluded by other individuals or objects. Occlusion impedes accurate keypoint localization by the Higher-HRNet module, potentially leading to incomplete or erroneous pose estimations or potentially the background figures in Figure 11 a). This compromised pose information can subsequently lead the classifier to produce false negatives (missed falls) or misclassifications. The interaction scene in Figure 11 b) also presents occlusion challenges, potentially affecting pose accuracy for both individuals.

Rapid or Kinematically Ambiguous Motion: Falls involving very rapid motion, rolling, or significant self-overlap can challenge the pose estimation network. The temporal dynamics or the spatial superposition of limbs may exceed the model's capacity for accurate keypoint tracking and association within a single frame or short sequence. This kinematic ambiguity, potentially exemplified by the dense keypoints in Figure 11 c), can result in misinterpretation of the body's configuration and, consequently, misclassification of the fall event.



a) Occlusion: Background Interference b) Occlusion: Interaction Impact c) Kinematic Ambiguity: Rapid Motion

Fig. 11. Analysis of Failure Cases for the Fall Detection System under Challenging Conditions

The system's performance can be notably compromised when confronted with atypical fall postures—those unusual or out-of-distribution configurations, such as falls occurring at extreme or oblique angles, which were insufficiently represented during the model's training phase. When encountering such novel scenarios, the generated pose map, which is crucial for understanding body mechanics, may become distorted or fail to capture sufficient discriminative features truly representative of a fall event. This deficiency can subsequently mislead the classification stage, causing it to misinterpret the event as a 'normal' state, simply because the pose representation is unfamiliar, incomplete, or lacks the tell-tale signs of a typical fall. While specific images perfectly matching the "extreme angle" description aren't provided, scenarios like those hinted at in Figure 11 c), potentially involving less common body positions during an interaction, could exemplify these challenging situations where the system's learned patterns are less applicable.

Furthermore, suboptimal image quality, characterized by issues like low resolution, significant visual noise, or severely poor illumination (even though successful low-light detection was noted in less extreme cases, severe degradation remains a formidable challenge), adversely impacts overall system performance. Such poor visual clarity directly degrades the richness and reliability of features extracted by the EfficientNet-B0 backbone. Concurrently, it significantly hinders the precision of keypoint detection by the Higher-HRNet module, making accurate skeletal tracking difficult. Although the derived pose maps are designed to offer a degree of resilience to minor image imperfections, the combined loss of crucial visual detail can ultimately compromise the final classification accuracy, particularly when the system needs to differentiate genuine falls from ambiguous non-fall states that may share superficial similarities.

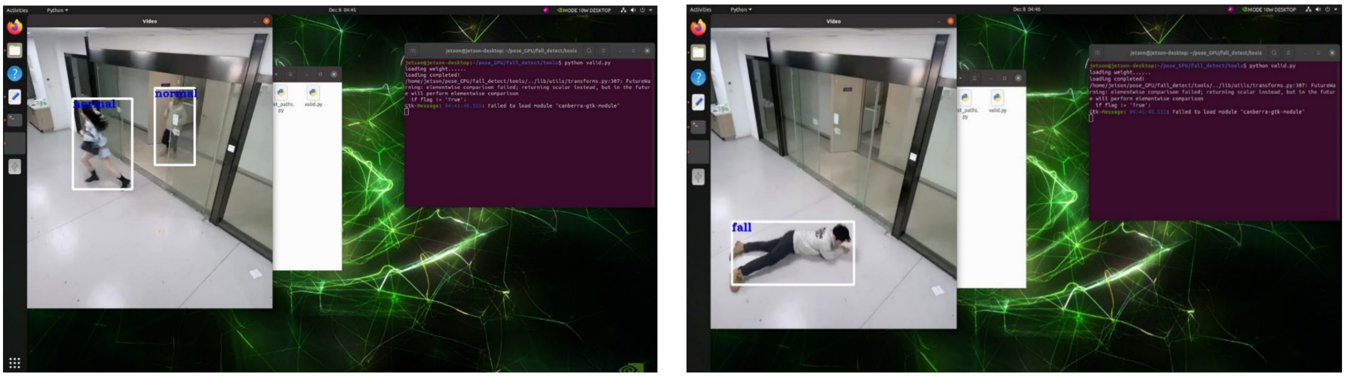
Overall, the visual analysis underscores the system's general effectiveness in standard and moderately complex scenarios, effectively leveraging the synergistic combination of visual and pose-based features. Key strengths undeniably include its capability to handle multiple persons within a scene and reliably distinguish falls from a diverse range of normal daily activities. However, the principal failure modes consistently emerge from challenges such as occlusion, where parts of the body are hidden; rapid or highly complex motion dynamics that lead to inaccuracies in pose estimation; the aforementioned atypical fall configurations that defy learned patterns; and significant image quality degradation that starves the algorithms of necessary information. Addressing these identified limitations is paramount. Future research should therefore prioritize avenues such as enhancing the robustness of pose estimation, perhaps by more explicitly incorporating temporal information across frames to predict and correct body part locations. Additionally, developing sophisticated data augmentation strategies, specifically targeting these edge cases and underrepresented scenarios, and investigating advanced techniques for effectively handling low-quality inputs will be critical for bolstering the system's reliability and broadening its generalization capabilities in real-world deployments.

H. Real-time Performance on Jetson NX

In the previous section, the process of constructing the Jetson Xavier NX was described in detail. This section focuses on the implementation of fall detection on the device. The application leverages the lightweight human pose estimation network, as discussed earlier, to generate human pose maps, and employs the VGG16 network to classify the fall status of individuals.

The implementation of fall detection begins by importing the pre-trained weights of the lightweight HRNet, which is used for pose estimation, and the VGG16 network, which has been trained for fall detection. Once the input image is received, it is processed through the lightweight HRNet to produce a human pose map. This pose map is then subjected to a preprocessing technique designed to enhance the visibility of the pose detection results. In order to improve the quality of human pose detection, the pose map is transformed into a one-dimensional image, which is then concatenated with the original input image. This combined image is subsequently passed through the VGG16 network. The network analyzes the image and determines the fall status of the individual depicted.

Fall detection outcomes are subsequently displayed on the device's interface, a representation of which is shown in Figure 12, allowing real-time viewing of results. This procedure ensures accurate system detection and classification of fall events, offering valuable user feedback and facilitating real-time monitoring in varied settings. The effective integration of these neural networks and preprocessing steps is crucial for optimizing model performance on the Jetson Xavier NX, thereby providing efficient and dependable fall detection capabilities.



```
jetson@jetson-desktop:~/pose_GPU/fall_detect/tools$ python valid.py
loading weight......
loading completed!
/home/jetson/pose_GPU/fall_detect/tools/../lib/utils/transforms.py:307: FutureWa
rning: elementwise comparison failed; returning scalar instead, but in the futur
e will perform elementwise comparison
  if flag != 'True':
Gtk-Message: 04:45:40.555: Failed to load module "canberra-gtk-module"
```

a) Fall Detection Interface 1

```
jetson@jetson-desktop:~/pose_GPU/fall_detect/tools$ python valid.py
loading weight......
loading completed!
/home/jetson/pose_GPU/fall_detect/tools/../lib/utils/transforms.py:307: FutureWa
rning: elementwise comparison failed; returning scalar instead, but in the futur
e will perform elementwise comparison
  if flag != 'True':
Gtk-Message: 04:45:40.555: Failed to load module "canberra-gtk-module"
```

b) Fall Detection Interface 2

Fig. 12. Fall detection screen

V. Conclusion

This study introduces a novel deep learning-based human pose estimation algorithm for fall detection, utilizing the Lightweight HRNet and VGG16 networks. The proposed method effectively detects human keypoints with HRNet's high-resolution output, facilitating accurate pose recognition. Subsequently, VGG16 processes the concatenated images and pose maps to classify fall events, offering robust performance even in complex environments. Experimental evaluations demonstrate that the algorithm achieves high accuracy, with significant improvements in both detection precision and generalization. Moreover, the system maintains low computational overhead, making it ideal for real-time deployment in elderly care and safety surveillance applications. Despite its promising performance, challenges remain in scaling the algorithm to handle larger datasets efficiently. Future research will focus on optimizing the computational efficiency of the system, enhancing its applicability in diverse real-world scenarios, and addressing the limitations in large-scale dataset processing.

References

[1] L. Chen, R. Li, H. Zhang, L. Tian, and N. Chen, "Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch," *Measurement*, vol. 140, pp. 215-226, 2019.

[2] T. L. Christiansen, S. Lipsitz, M. Scanlan, S. P. Yu, M. E. Lindros, W. Y. Leung, J. Adelman, D. W. Bates, and P. C. Dykes, "Patient activation related to fall prevention: a multisite study," The Joint Commission Journal on Quality and Patient Safety, vol. 46, no. 3, pp. 129-135, 2020.

[3] C. Y. Hsieh, K. C. Liu, C. N. Huang, W. C. Chu, and C. T. Chan, "Novel hierarchical fall detection algorithm using a multiphase fall model," *Sensors*, vol. 17, no. 2, 307, 2017.

[4] J. A. Stevens, "Falls among older adults—risk factors and prevention strategies," *Journal of Safety Research*, vol. 36, no. 4, pp. 409-411, 2005.

[5] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291-7299, 2017.

[6] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2334-2343, 2017.

[7] J. Liao, W. Cui, Y. Tao, T. Shi, and L. Shen, " Lightweight HRNet: A Lightweight Network for Bottom-Up Human Pose Estimation," *Engineering Letters*, vol. 32, no. 3, pp. 661-670, 2024.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] D. Kim, H. Kim, Y. Mok, and J. Paik, "Real-time surveillance system for analyzing abnormal behavior of pedestrians," *Applied Sciences*, vol. 11, no. 13, 6153, 2021.

[10] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimedia Systems*, vol. 29, no. 1, pp. 167-195, 2023.

[11] M. R. Ibrahim, J. Haworth, N. Christie, and T. Cheng, "CyclingNet: Detecting cycling near misses from video streams in complex urban scenes with deep learning," *IET Intelligent Transport Systems*, vol. 15, no. 10, pp. 1331-1344, 2021.

[12] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829-864, 2020.

[13] R. Zhao, Y. Wang, P. Jia, W. Zhu, C. Li, Y. Ma, and M. Li, "Abnormal behavior detection based on dynamic pedestrian centroid model: Case study on U-turn and fall-down," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8066-8078, 2023.

[14] H. Yang, J. Ni, J. Gao, Z. Han, and T. Luan, "A novel method for peanut variety identification and classification by Improved VGG16," *Scientific Reports*, vol. 11, 15756, 2021.

[15] B. Jabłoński, D. Makowski, P. Perek, P. Nowak vel Nowakowski, A. P. Sitjes, M. Jakubowski, Y. Gao, A. Winter, and The W-X Team, "Evaluation of NVIDIA Xavier NX Platform for Real-Time Image Processing for Plasma Diagnostics," *Energies*, vol. 15, no. 6, 2088, 2022.