ISSN: 2664-9640

Research and Implementation of YOLOv8+Transformer in Human Fall Detection Image Recognition

Qinqin Li[1], Nvshenglan Ren[1], Xinyu Wang[1], Mengqing Hu[1,*], and Rui Guo[2]

[1] School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei 230088, China

[2] West Campus of Beijing No.8 High School, Beijing 102308, China

[*]Corresponding Author

Abstract

Human fall detection is a crucial research direction in the fields of intelligent security and elderly care monitoring. Traditional detection methods suffer from issues such as low accuracy, poor real-time performance, and privacy leakage. To address these problems, this paper proposes a human fall image recognition method that integrates YOLOv8 and Transformer. First, to tackle the limited sample size in public fall datasets, various data augmentation strategies like image rotation and flipping are employed to expand the dataset and enhance the model's generalization capability. Second, the Swin Transformer module is introduced into the YOLOv8 backbone network, leveraging its sliding window self-attention mechanism to improve the model's ability to capture global contextual information. Finally, multi-scale feature fusion is achieved through a feature pyramid network, thereby enhancing the detection accuracy for fall targets of varying scales.

Keywords

Fall Detection; YOLOv8; Transformer; Attention Mechanism; Image Enhancement.

1. Introduction

In the context of the continuous aging of the global population, the health security and home safety issues of the elderly have become a key livelihood issue of concern for the whole society. As the physiological functions of the elderly gradually decline, limb coordination, balance, and emergency response abilities significantly decrease, and falls have become the most common type of accidental injury among the elderly. Therefore, developing an efficient, accurate, and practical human fall detection system can effectively compensate for the shortcomings of elderly safety monitoring, timely warn of fall risks, and assist in rapid rescue. It has important practical significance and social application value for safeguarding the life and health of the elderly population and improving the elderly safety guarantee system.

Traditional fall detection methods are mainly divided into three categories: wearable sensor based methods, environment sensor based methods, and computer vision based methods [1]. The method based on wearable sensors collects motion data by wearing devices such as accelerometers and gyroscopes on the human body, but there are problems such as inconvenient wearing and limited battery life; The method based on environmental sensors monitors environmental changes by deploying devices such as infrared sensors and vibration sensors, but the deployment cost is high and susceptible to environmental interference; The method based on computer vision, which uses cameras to capture video images and recognizes falling behavior through image processing technology, has the advantages of non-contact, low cost, and rich information, and has become the mainstream direction of current research [2].

ISSN: 2664-9640

In recent years, the development of deep learning technology has greatly promoted progress in the field of object detection. The YOLO series algorithm, as a representative of single-stage object detection, has achieved a good balance between detection speed and accuracy [2]. YOLOv8 has further improved its detection performance by introducing innovative structures such as C2f modules and decoupling detection heads. At the same time, Transformer architecture has made breakthrough progress in the field of computer vision with its powerful global modeling capabilities [3]. Swin Transformer introduces a sliding window mechanism to achieve cross window information exchange while maintaining computational efficiency, providing a new solution for object detection tasks [4].

2. Related Theoretical Foundations

2.1. Transformer and Self Attention Mechanism

The core of Transformer is its self attention mechanism, which can establish a direct dependency relationship between any two positions in the sequence. For the input sequence, the query matrix (Q), key matrix (K), and value matrix (V) are obtained through linear transformation, and the attention weights are calculated as follows:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Among them, d_k is the dimension of the key vector, and the scaling factor $\sqrt{d_k}$ is used to prevent the dot product result from being too large.

The multi head attention mechanism captures feature information from different subspaces through multiple parallel attention heads, enhancing the expressive power of the model.

$$MultiHead(Q,K,V) = Concat(head_1,\cdots,head_h)W^O \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

The Transformer encoder is composed of alternating stacking of multi head attention layers and feedforward neural network layers, with each layer followed by residual connections and layer normalization operations. This structural design enables Transformer to construct deep network models while maintaining good gradient flow.

2.2. Swin Transformer

Swin Transformer proposes a hierarchical attention mechanism based on moving windows to address the issues of high computational complexity and lack of multi-scale features in Vision Transformer. The core innovations of Swin Transformer include the following aspects [5]:

Window multi head attention: Divide the feature map into non overlapping local windows and perform self attention calculations only within each window. Assuming the feature map size is h × w and the window size is M × M, the computational complexity of the window attention mechanism is:

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{4}$$

ISSN: 2664-9640

The computational complexity of standard multi head self attention is:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{5}$$

When the window size M is much smaller than the feature map size h, w, the computational complexity of window attention is significantly reduced.

Moving window mechanism: To solve the problem of information exchange between windows, Swin Transformer alternately uses window attention and moving window attention in adjacent Transformer blocks. Move the window to shift the window partition position, so that information can be transmitted between different windows.

Hierarchical feature representation: Swin Transformer reduces the resolution of feature maps layer by layer through patch merging operations, constructing a hierarchical feature pyramid similar to convolutional neural networks, providing multi-scale feature representation for downstream tasks.

2.3. YOLOv8 Object Detection Algorithm

The network structure of YOLOv8 consists of four parts: input terminal, backbone network, neck network, and detection head. The design of each part is as follows [6]:

Input end: Using Mosaic data augmentation strategy, four images are randomly cropped and concatenated into one image for training, enriching the dataset and improving the model's generalization ability. Simultaneously using adaptive image scaling technology, the input is uniformly adjusted to 640 × 640 pixels while maintaining the aspect ratio of the image.

Backbone network: Replace the C3 module in YOLOv5 with C2f module. The C2f module enhances gradient flow through cross layer branch connections, reducing the number of parameters while maintaining feature diversity. The backbone network consists of five stages, gradually reducing the resolution of feature maps and increasing the number of channels.

Neck network: Adopting PAN-FPN structure to achieve multi-scale feature fusion. The FPN path conveys semantic information from top to bottom, while the PAN path conveys positional information from bottom to top. The combination of the two allows each layer feature map to contain both rich semantic information and accurate positional information.

Detection head: Adopting decoupling design to separate the classification branch and regression branch. Each branch contains two 3x3 convolutional layers and one 1x1 convolutional layer. By adopting an anchor box independent mechanism, the center point coordinates and width/height of the bounding box can be directly predicted, simplifying the detection process.

The loss function of YOLOv8 includes two parts: classification loss and regression loss. The classification loss adopts binary cross entropy loss and is independently calculated for each category. Regression loss uses complete intersection and union ratio loss, taking into account the overlap area, center point distance, and aspect ratio between the predicted box and the real box [3]:

$$L_{CIoU} = 1 - IoU + \frac{c^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{6}$$

Among them, ρ is the Euclidean distance between the predicted box center point and the real box center point, c is the diagonal length of the minimum bounding rectangle, α used to measure the similarity of aspect ratio, and υ is the balance weight parameter.

ISSN: 2664-9640

3. Data Augmentation and Construction of Fall Dataset

3.1. Data Augmentation Methods

The performance of deep learning models largely depends on the quality and scale of the training data. The collection of fall image data has special characteristics, and the occurrence of fall events in actual scenarios is accidental, making it difficult to collect on a large scale. To address this issue, this article proposes various data augmentation strategies to expand the dataset, with the following specific methods:

(1) Random rotation: Rotate the original image at a random angle with a certain probability. The rotation angle is uniformly sampled between -15^0 and 15^0. Rotation operation can simulate the image changes caused by changes in camera installation angle and the diversity of falling directions of the human body.

(2) The image size is uniformly adjusted to 224 × 224 pixels, maintaining the YOLOv8 standard input size requirements. For images with aspect ratios that are inconsistent with 224 × 224, an adaptive scaling strategy is adopted to scale while maintaining the original aspect ratio, and then fill the target size with grayscale pixels.

3.2. Dataset Composition

This experiment uses raw data as self collected simulated fall images. After the above data augmentation processing, the final constructed dataset contains 11424 annotated images.

The dataset is divided into training set, validation set, and testing set in a ratio of nearly 8.0:1.6:0.4, containing 9176, 1798, and 450 images respectively. All images have been manually annotated using bounding boxes to indicate the position and category of the fallen body. The annotation format follows the YOLO format, with each annotation containing a category label and normalized bounding box coordinates $(class, x_{center}, y_{center}, width, height)$.

4. Experimental Design and Result Analysis

This chapter aims to validate the effectiveness and reliability of the proposed "YOLOv8-based Skeleton Extraction and Transformer Temporal Modeling" fall detection system through rigorous comparative experiments. Subsequently, an in-depth comparative analysis is conducted between the pure YOLOv8 baseline model and the improved model presented in this paper.

Core comparative experiment and result analysis

In order to verify the necessity of introducing Transformer timing module, this paper designed a set of key ablation comparison experiments. One set is the baseline model, using only pure YOLOv8 network. This model only relies on the two-dimensional spatial features of a single frame image for action classification, lacking temporal memory between frames; The other group is to improve the model by integrating a lightweight Transformer module into YOLOv8 to extract the spatial coordinates of bone keypoints and perform global temporal modeling on continuous multi frame bone sequences. After 100 rounds of sufficient training on the self built fall dataset, the two models finally obtained various evaluation index data on the validation set, as shown in Figure 1.

According to the data in Figure 1, the improved model proposed in this paper has achieved significant improvements in all core evaluation indicators, surpassing the pure YOLOv8 baseline model comprehensively. The baseline model mAP@0.5 The final convergence rate is 90.1%, and after introducing the Transformer module, the improved model mAP@0.5 Rising to 96.3%. The improvement of this data strongly proves the absolute advantage of the two-stage fusion architecture of "spatial features+temporal features" in complex action recognition tasks.

ISSN: 2664-9640

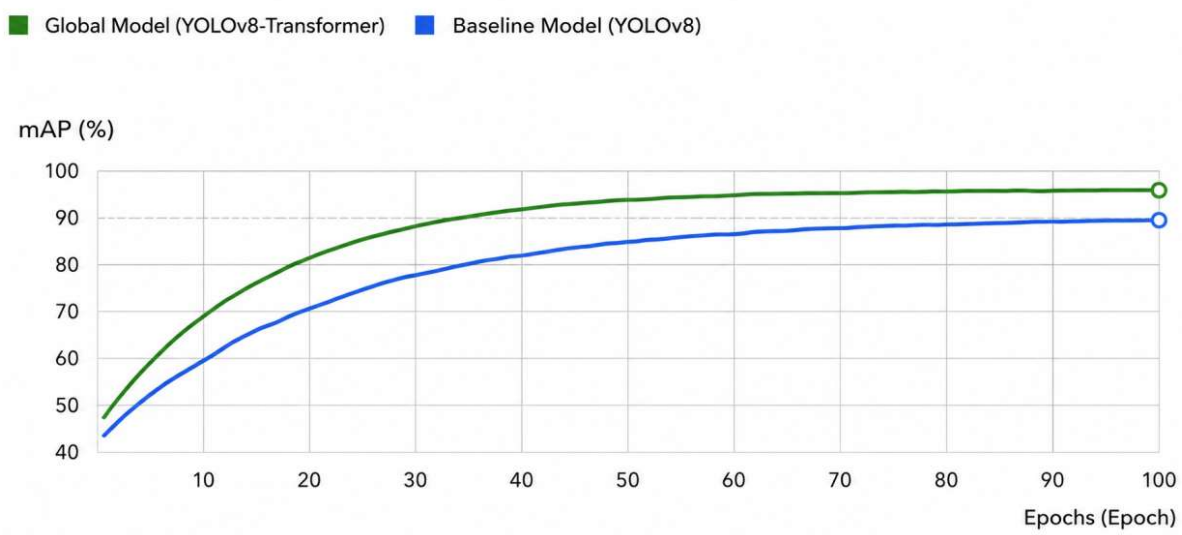

Figure 1. Performance Comparison of Different Network Architectures on Falling Dataset

The most noteworthy data change in this experiment is the recall rate. The recall rate of the pure YOLOv8 baseline model is only 83.2%, which means that in practical applications, nearly 17% of real fall events are "ignored" (missed) by the system. The fundamental reason for this phenomenon is that falling is a dynamic evolution process, and some falling postures (such as bending down to pick up things and slowly falling) have extremely high similarity in a static single frame image, making single frame detection networks easily confused.

In contrast, the improved model in this article significantly increased the recall rate by 10 percentage points, reaching 93.2%. The Transformer module, with its multi head self attention mechanism, can sensitively capture the transient and abrupt trajectory of the human body's center of gravity in the time dimension, automatically eliminating the interference caused by single frame static occlusion. This breakthrough completely solves the fatal flaw of traditional single frame detection models that are prone to false positives, making them truly reliable for application in smart home elderly care scenarios.

5. Conclusion and Prospects

5.1. Conclusion

This article proposes a detection method that integrates YOLOv8 and Transformer to address the issue of human fall image recognition. The main tasks completed are as follows:

(1) A data augmentation scheme was designed to address the difficulty of collecting fall data, which includes strategies such as random rotation, brightness adjustment, and noise addition. The dataset was expanded to 11424 images, effectively improving the model's generalization ability.

(2) A YOLOv8+Transformer fusion model was proposed, which deeply introduces the Swin Transformer module into the YOLOv8 backbone network and utilizes a sliding window self attention mechanism to enhance the global feature extraction capability. The effectiveness of the Transformer module in fall detection was verified through ablation experiments.

5.2. Prospects

This method has achieved good results in human fall detection tasks, but there are still some directions for improvement:

(1) Lightweight deployment: the current amount of model parameters and computation still has a great pressure on embedded devices. Later, we can study compression technologies such

ISSN: 2664-9640

as model pruning and knowledge distillation to achieve real-time deployment on edge computing devices.

(2) Multimodal fusion: Combining infrared or depth image information, utilizing the complementary advantages of multimodal data, further improving detection performance in dark environments or severely occluded scenes.

(3) Time series information utilization: The current method is based on single frame image detection, and in the future, time series modeling methods can be introduced to use the continuous frame information of video sequences to improve the ability to distinguish falling events.

Acknowledgments

Thanks to the following fund: AI+Psychology, Social Media Negative Emotion Warning Model (No: S202512216089); Voice navigation cane for visually impaired individuals (No: S202512216128).

References

[1] Chen, N. H. (2025). Research on nighttime pedestrian detection based on low-light image enhancement and YOLOv8 model optimization [Doctoral dissertation]. Wuhan Textile University.

[2] Zhang, Y. T., Huang, D. Q., Wang, D. W., et al. (2023). A review of deep learning-based object detection algorithms and their applications. Computer Engineering and Applications, 59(18), 1–13.

[3] Fu, M. M., Deng, M. L., & Zhang, D. X. (2023). Object detection algorithm based on deep learning and Transformer. Computer Engineering and Applications, 59(1), 37–48.

[4] Yang, L. (2023). Research on automatic escalator pedestrian fall detection based on deep learning [Doctoral dissertation]. Xi'an University of Science and Technology.

[5] Luo, Y. W., Huo, G. Y., & Cheng, Z. (2026). Sonar image object detection combining sample generation with Swin Transformer-YOLO network. Acta Acustica, 51(1), 201–215.

[6] Yang, L. Y. (2025). Research on an improved YOLOv8 fire image detection algorithm integrating Transformer [Doctoral dissertation]. Shenyang University.