The elderly falling detection based on improved YOLOv8

Zesen Dong

Lanzhou University, Lanzhou, 730000, China

dongzs@lzu.edu.cn

Abstract. Accidental falls are one of the major safety problems that cause the injury and death of the elderly over 65 years old in China. In the face of the lack of accuracy of existing fall detection technologies in dealing with complex environments such as object occlusion and illumination changes, this study based on the improved algorithm of YOLOv8, this study introduces a multi-head attention mechanism to improve the accuracy, robustness, universality and scalability of fall detection. The main goal of this work is to improve the YOLOv8 model's structural adjustments, which include adding a multi-head attention mechanism to improve the model's capacity to identify important characteristics. This method enables the model to learn several features of the input data in separate representational subspaces at the same time, leading to more accurate fall behavior identification and localization. The modified model was tested in several fall scenarios, including ones with varying illumination and shade levels, in the experimental section. This technique outperforms the original YOLOv8 model, achieving 79.4% mAP@0.5 on the dataset, according to a comparison of performance benchmarks. In summary, YOLOv8's multi-head attention method adds to the fall detection algorithm's detection accuracy while simultaneously.

Keywords: Target Detection, YOLOv8, MHSA, Fall Detection.

1. Introduction

With the increasing aging of the global population, the safety of the elderly at home has become an increasingly prominent social issue. Data shows that accidental falls are one of the most common major safety threats among people over 60 years of age. Such accidents not only pose a serious threat to the health of the elderly, but also cause a medical burden arising from long-term medical treatment. Fall events are often caused by a variety of complex factors, including the safety hazards of the living environment and individual behavior patterns. How to detect fall behaviors in time and ensure their safety has become a hot issue. Therefore, fall detection has become one of the most popular research directions.

In the era of rapid development of the Internet, deep learning technology provides a new solution for the safety monitoring of the elderly at home. By combining the computer vision with camera monitoring, the recognition and judgment of human behavior are realized under the operation of behavior recognition algorithm. The application of this method in daily life can not only monitor the behavior of the elderly in real time, but also detect and warn of potential fall risks. Compared to traditional emergency response models, this technology can prevent accidents more proactively and reduce the reliance on human monitoring.

This study mainly studies the deep learning fall detection method based on computer vision, adopts the YOLOv8 object detection algorithm and combines multi-head self-attention mechanism module to design and optimize the fall monitoring network model according to the dynamic monitoring needs of the elderly in the home environment. The multi-head self-attention mechanism is introduced into the backbone network of the YOLOv8 model to obtain more detailed feature information about the human body, so as to better help the network learn important features. Through the training of large-scale data sets, this model can timely and accurately identify the fall event in the complex home environment, thus significantly improving the safety of the elderly's home. The development of this deep learning based monitoring technology can create a safer living environment for the elderly and reduce the risks they may face.

2. Basic Knowledge of Relevant Theories

2.1. Target Detection

Target detection based on computer vision is mainly concentrated in two directions: single-stage (YOLO, SSD) and two-stage (R-CNN). The workflow of the two-stage detection algorithm consists of first generating a series of candidate regions in the image, and then classifying and positioning these candidate regions. This design allows for more precise positioning and classification of targets, which improves detection accuracy. The high processing capacity leads to the slow processing speed and low efficiency of this method. In contrast, the single-stage algorithm can directly predict the position and category of objects on the whole image with only one forward propagation, which has high computational efficiency and processing speed, and can meet the real-time requirements. In the human fall detection of daily life scenes, a model with high real-time performance and both lightweight and detection accuracy is needed to achieve ideal results. Therefore, YOLO and its derivative series detection algorithms have been widely used.

For this study, some people have optimized the YOLO models [1-9]. For example, Wanli Huang [8] successively embedded three self-designed modules in the network structure of YOLOv7: the attention module, the adaptive mean module and the feature enhancement module. By introducing these modules, the network enhanced feature extraction and improved the accuracy of model recognition. Libu Lan [9] made the following optimization for the YOLOv7 algorithm: Introduce the ECA attention mechanism module to make the algorithm more focused on the fall behavior of characters. Then the SIoU loss was replaced by the CIoU loss to reduce the degree of freedom of the loss function and improve the speed of model training.

2.2. YOLOv8

YOLOv8 [10] is a cutting-edge, state-of-the-art (SOTA) model that builds on the successes of previous YOLO iterations, introducing new features for faster training speeds and improved detection performance. A structural diagram is provided in Figure 1.

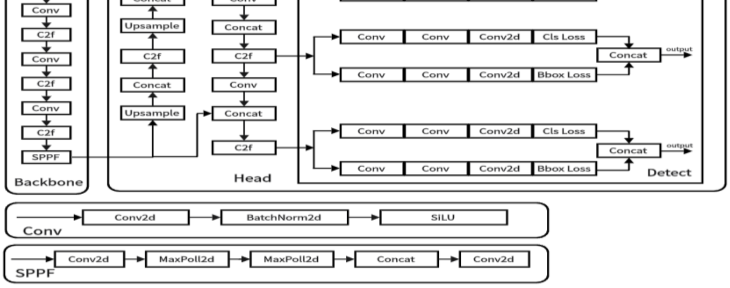

Figure 1. YOLOv8 Structure

The backbone network of YOLOv8 adopts the ELAN design philosophy from YOLOv7, reducing the size of the first convolutional layer's kernel to 3x3 and replacing the C3 structure with the more gradient-rich C2f structure which includes more residual connections, thereby allowing the model to capture more gradient flow information. Additionally, the head part of the model has been revised to use a new decoupled head structure, separating classification and regression tasks. The model has shifted from an Anchor-Based to an Anchor-Free approach, directly predicting the positions and sizes of targets to enhance recognition capabilities. In the data augmentation phase of training, mosaic augmentation is disabled in the final 10 epochs, and a dynamic Task-Aligned Assigner strategy for sample distribution is utilized to improve model accuracy.

The working principle of YOLOv8 can be summarized as follows: Initially, the input image is scaled to a fixed size and fed into a convolutional neural network. The CNN divides the input image into several grids, with each grid cell responsible for detecting targets whose center points fall within the cell. Each cell predicts B bounding boxes along with their confidence levels. The features extracted by the convolutional layers are passed to YOLO’s output layer via a fully connected layer. The output of the YOLO algorithm includes the predicted bounding boxes and their class probabilities for each grid in the image.

2.3. Multi-Head Attention

The multi-head attention [11] mechanism operates by decomposing the attention computation into multiple independent units, known as "attention heads". Initially, the model's Query, Key, and Value parameters are divided into N parts. Each part is then processed by an independent attention head, allowing the model to perform multiple attention computations in parallel. Each attention head applies a distinct linear transformation to its corresponding Query, Key, and Value to prepare them for input into the scaled dot-product attention mechanism, with unique parameters for each head, meaning that the weight matrices for these transformations are not shared. This design enables each head to capture different features and relationships within the input sequence.

After performing these parallel computations, the results from all attention heads are concatenated together and passed through another linear transformation. This step amalgamates the information from all heads to produce the final output of the multi-head attention. The structure diagram is shown in Figure 2.

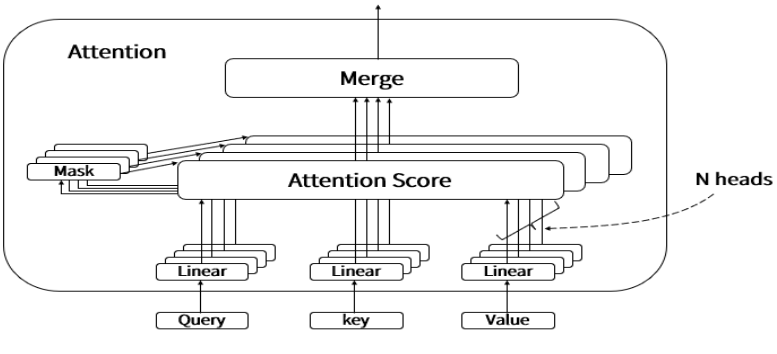

Figure 2. MHSA Structure

Attention computation: Attention (Q, K, V) =softmax $(\frac{QK^T}{\sqrt{dk}})$V, where Q(Query) is the matrix representing the set of queries, K(Key) is the matrix representing the set of keys, V(Value) is the matrix representing the set of values, and dk is the dimensionality of the keys and queries.

Calculation per attention head: headi=Attention (QWiQ ,KWiK,VWiV), where WiQ is the matrix for the query transformation, WiK is the matrix for the key transformation, and WiV is the matrix for the value transformation in head.

Multi-head attention computation: MultiHead(Q,K,V)=Concat (head1,…,headh)WO, where headi is the output from the attention head representing the specific focus and information extracted by the head, h is the number of attention heads, and WO is the output weight matrix which linearly transforms the concatenated output of all heads into the final output space.

3. YOLOv8 with Multi-Head Self-Attention

3.1. Data Preparation

This study uses a data set specifically for the detection of falls in the elderly. The data set consists of images and corresponding label files. The images are mainly scenes of daily activities and falls among the elderly, and the label files record the location of the targets in the images in detail. The dataset has a total of 1443 images and contains diverse fall scenes to improve the practicality and robustness of the detection algorithm. The images in the dataset cover different environments, different lighting conditions, and multiple fall positions to simulate a variety of situations that might occur in real life.

3.2. Model Construction

A Multi-Head Self-Attention Mechanism (MHSA) has been added to the YOLOv8 backbone network after the Spatial Pyramid Pooling layer (SPPF) in order to improve the model's recognition performance of complicated situations. Thanks to this method, the network may concentrate more on important details in the picture, including the way the human body is positioned after a fall. The MHSA increases the expressive capability of the model by computing several attention heads in simultaneously, each concentrating on distinct subsets of features. The model structure is shown in Figure 3.

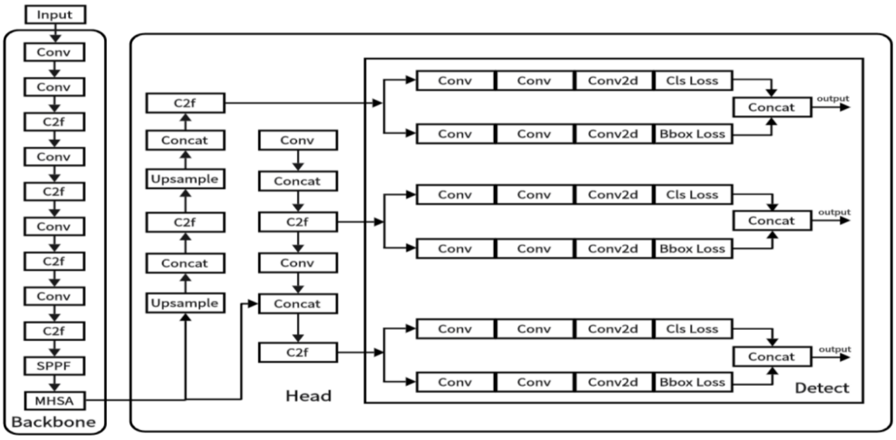

Figure 3. Improved YOLOv8 Structure

3.3. Model Training

The improved YOLOv8 remains a target detection algorithm, employing a single-stage detector approach for object recognition. The training process can be divided into the following steps:

(1) Data Preparation: Prepare training data, including training images and annotation files, with images sized at 640x640. In this step, the training dataset is divided into training, validation, and testing sets in proportions of 80%, 10%, and 10% respectively.

(2) Model Configuration: Utilize a pre-trained YOLOv8 model, adjusting model parameters to meet the specific needs of elderly fall detection.

(3) Data Processing: To increase data randomness and enhance model generalization, an automated script is used to partition the data.

(4) Model Training: Iteratively train the model using the training set, optimizing model parameters through the calculation of cross-entropy loss and bounding box regression loss. At the end of each training cycle, evaluate the model's performance using the validation set, and adjust the learning rate and other hyperparameters as needed.

(5) Model Testing: Conduct a performance evaluation on the test set to verify the model's ability to recognize unseen data.

Model parameters are shown in Table 1.

Table 1. Model Parameters

Parameters	**Settings**
Epochs	200
Batch	32
Weight Decay	0.0005

Epochs (200): It refers to one complete cycle through the full training dataset. Setting the epochs to 200 means that the entire dataset will be passed forward and backward through the neural network 200 times.

Batch (32): Its size refers to the number of training samples used to train the model in one forward/backward pass. Setting the batch size to 32 means that 32 samples from the training dataset are used to compute the model loss and subsequently update the weights of the model.

Weight Decay (0.0005): It is a regularization technique used to prevent the model from overfitting, which can happen when the model learns too well to fit every detail of the training data. By adding a weight decay of 0.0005, the training algorithm modifies the learning process to keep the weights of the network small.

4. Experimental Results and Analysis

4.1. Experimental Results

In order to ensure the effectiveness and accuracy of the detection system, this study adopted an improved algorithm based on YOLOv8 to identify fall events.

In the experiment part, we visually demonstrated the detection effect of the improved YOLOv8 algorithm in different fall scenarios. These scenes include indoor and outdoor environments under different lighting conditions, scenes with occlusions, and public spaces with complex backgrounds. This makes it intuitive to see how well the algorithm performs in real-world applications and how well it can identify fall behavior. A number of sample results that illustrate the behavior of the algorithm in the aforementioned conditions are shown in Figure 4.

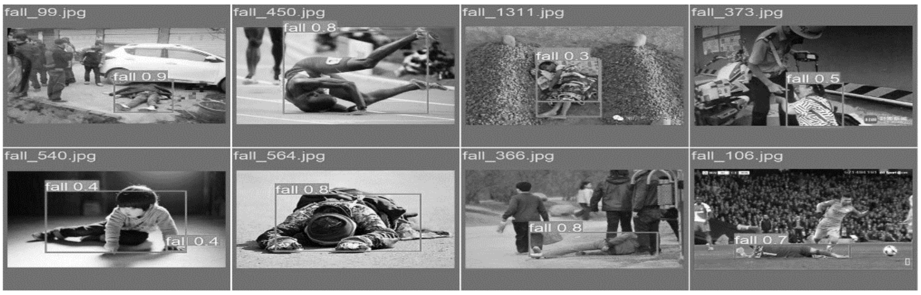

Figure 4. Results

4.2. Analysis of Experimental Results

To validate the effectiveness of the improved algorithm, this study compared the original YOLOv8 model with the enhanced algorithm across several key performance metrics to demonstrate the advantages of the new approach. These metrics include precision(P), recall(R), and average precision (AP).

Precision refers to the proportion of true positive samples in all detected outcomes, calculated using the formula: Precious=$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$, where TP is the number of correctly detected targets, and FP is the number of false positives.

Recall refers to the proportion of detected targets out of all actual targets, calculated using the formula: Recall=$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where FN is the number of missed targets.

Average Precision is a comprehensive metric that considers precision at different levels of recall, obtained by calculating the area under the precision-recall curve. The higher the AP, the better the performance of the algorithm. It is calculated using the formula: AP=$\int_0^1 \mathrm{P(r)dr}$, where P(r) represents the precision at a given recall rate.

Mean Average Precision is the mean of the APs across all categories, calculated as: mAP= $\frac{\sum_{\mathrm{i=1}}^{\mathrm{C}} \mathrm{APi}}{\mathrm{C}}$, where C is the total number of categories.

In object detection, mAP@0.5, which is the mean average precision at a confidence threshold of 0.5, is commonly used. It reflects the algorithm’s ability to distinguish between positive and negative samples at higher confidence levels, measuring the accuracy and stability of the algorithm.

The comparison of the two models is shown in Table 2, Figure 5 and Figure 6.

Table 2. Effect Comparison

Models	P	R	mAP50
YOLOv8_MHSA	0.77073	0.74522	0.79448
YOLOv8	0.75556	0.70064	0.757

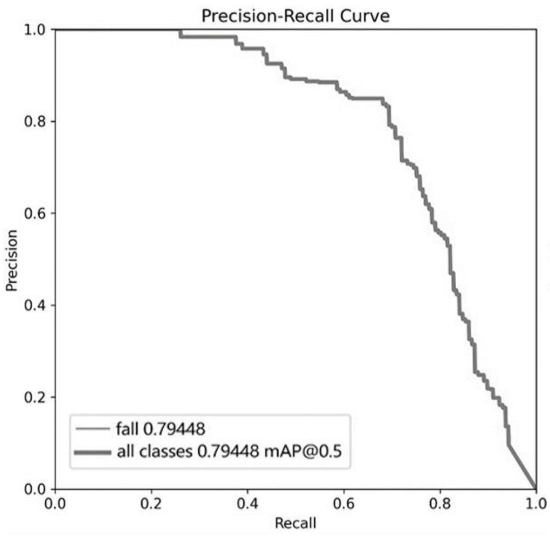

Figure 5. Average Accuracy of Improved YOLOv8

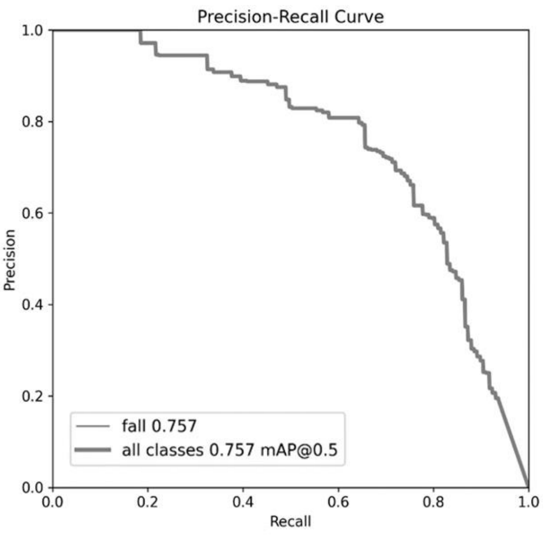

Figure 6. Average Accuracy of YOLOv8

By comparing above Figure 5 and Figure 6, it can be seen that the area of the improved pr curve is larger than that of the original version. At the same time, according to Table 2, it shows that the accuracy of the model increased by 1.5%, the recall increased by 4.5%, and the mean precision increased by 4%.

5. Conclusion

This study introduced an improved YOLOv8 detection algorithm that integrates the concept of transformers, embedding a multi-head self-attention mechanism within the YOLOv8 model. The algorithm demonstrated strong performance on test datasets, with enhanced accuracy in detection, reflecting the effectiveness of combining deep learning with self-attention mechanisms. However, this paper also recognizes that there is significant room for improvement in the model. For instance, integrating an LSTM network to incorporate time series analysis could enable the model to predict subsequent actions based on the elderly's movements over time, allowing for early prediction and a faster response to potential falls.

The findings suggest that deep learning-based technology for in-home elderly fall detection holds substantial potential for practical application, providing robust technical support for the safety of elderly individuals at home. If this technology continues to be developed and applied, it could significantly improve the sense of security among the elderly.

References

[1] Yangyang Jiang. The design and implementation of YOLOv3-based fall detection system for the elderly in home [D]. Nanjing University of Posts and Telecommunications, 2022. DOI:10.27251/d.cnki.gnjdc.2021.001540.

[2] Yeping Wang. Video fall detection algorithm based on YOLO [J]. Computer programming skills and maintenance, 2019(11):137-139.DOI:10.16184/j.cnki.comprg.2019.11.049.

[3] Jiao Li. The algorithm and implementation of fall detection for the elderly based on deep learning [D]. Hebei University of Technology, 2023. DOI:10.27105/d.cnki.ghbgu.2021.001162.

[4] Zhangping Zhong. Research on home-based fall behavior decetion based on improved Yolov5s [D]. Nanchang University, 2024. DOI:10.27232/d.cnki.gnchu.2023.003973.

[5] Yanming He. Research and Application of Fall Behavior Detection Based on Improved YOLOv5 [D]. Anhui Jianzhu University, 2024. DOI:10.27784/d.cnki.gahjz.2023.000544.

[6] Zhuowen Li, jun Xiao. Design of a human fall detection algorithm based on deep learning [J]. Communication and Information Technology, 2024(01):101-105.

[7] Shuangshuang Ma. Research on Fall Detection Algorithm Based on Deep Learning [D]. Beijing Institute of Graphic Communication,2024.DOI:10.26968/d.cnki.gbjyc.2023.000103.

[8] Wanli Huang. Fall Detection Algorithm Research Based on Improved YOLOv7 [D]. Jianghan University, 2024. DOI:10.27800/d.cnki.gjhdx.2023.000335.

[9] Libu Lan. Research on Fall Detection Algorithm based on Improved YOLOv7 [D]. Ningxia University, 2024. DOI:10.27257/d.cnki.gnxhc.2023.001505.

[10] Jocher, G., Chaurasia, A., & Qiu, J. Ultralytics YOLO, Version 8.0.0, 2023. https://github.com/ultralytics/ultralytics

[11] Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need [J] arXiv, 2017. DOI:10.48550/arXiv.1706.03762.