Available Online: 26 March 2026 DOI: 10.54254/2977-3903/2026.32415

A high-quality localization-aware action recognition algorithm based on YOLOv11

Zheng Han

School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing, China

h978021540@126.com

Abstract. To address the challenges of large-scale variations in human targets, the loss of spatial details, and the inconsistency between prediction confidence and localization quality in complex scenarios, this study proposes a high-quality localization-aware action recognition method based on YOLOv11d. An SPDConv downsampling structure is introduced into the backbone network and the feature fusion stage to enhance the representation capability of small-scale target features. In addition, a localization quality estimation branch is incorporated into the detection head to explicitly model the Intersection over Union (IoU) of bounding boxes, and the confidence score is reweighted by combining the estimated localization quality with class probability. Experimental results demonstrate that the proposed method achieves an mAP@50 of 96.0% and an mAP@50–95 of 72.3%, representing improvements of 0.3% and 2.8%, respectively, compared with YOLOv11.

Keywords: action recognition, YOLOv11, high-quality localization awareness, SPDConv

1. Introduction

With the rapid development of computer vision and deep learning technologies, human action recognition has demonstrated broad application prospects in fields such as intelligent surveillance [1], public security [2], behavior analysis, and human–computer interaction [3]. In recent years, object detection algorithms based on convolutional neural networks have evolved continuously. In particular, the single-stage detection framework represented by the YOLO [4] series, benefiting from its end-to-end architecture and real-time performance, has achieved promising results in practical applications. Nevertheless, human action recognition in complex environments still faces a number of challenges.

First, in complex scenes the scale of human targets varies significantly, and small-scale, distant, or partially occluded targets are difficult to model effectively. Traditional downsampling methods mainly rely on strided convolutions or pooling operations. While these approaches reduce feature resolution, they inevitably lead to the loss of spatial detail information, resulting in insufficient feature representation for small targets. Second, during the object detection process [5], inconsistencies often arise between classification confidence and actual localization accuracy. Existing detection heads typically optimize the classification branch and the regression branch independently. As a result, the predicted confidence score primarily reflects class probability and fails to adequately capture the localization quality of the bounding box. This mismatch affects candidate box

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

ranking and reduces the stability of the filtering process during the Non-Maximum Suppression (NMS) stage, thereby lowering the overall reliability of detection results.

Alzahrani [6] and colleagues proposed the YOLO-Act action detection framework, which introduces multi-frame information fusion and an object tracking mechanism based on YOLOv8. By integrating spatial detection results with temporal sequence features, the framework enables unified spatiotemporal action detection. Elnady [7] and colleagues proposed the YOLO-LSTM action recognition method, which employs YOLOv7 as a spatial feature extractor to obtain human-region features and combines it with a Long Short-Term Memory (LSTM) network to model temporal dynamics, thereby enabling human action recognition in video sequences. Chen [8] and colleagues proposed a joint human localization and action recognition approach based on an improved YOLOv11 model. By employing a dual-output detection head to simultaneously perform object detection and action classification tasks, the method achieves end-to-end training and improves behavior recognition accuracy in complex scenarios.

Although these methods have made progress in spatiotemporal modeling and the integration of detection and classification, they still commonly suffer from insufficient feature representation for small-scale targets and a mismatch between detection confidence and localization quality in complex environments. These limitations restrict the stability and reliability of action recognition results. To address these issues, this paper proposes a human action recognition method based on a high-quality localization-aware mechanism within YOLOv11. During the feature extraction stage, an SPDConv downsampling structure is introduced. By replacing conventional strided convolutions with spatial rearrangement and channel reconstruction, the method reduces feature resolution while effectively preserving spatial structural information, thereby enhancing the model's ability to represent small-scale human targets. In the detection head, a detection quality–aware branch is constructed to explicitly model the IoU-based localization quality of predicted bounding boxes. A confidence reweighting mechanism is then employed to fuse class probability with localization quality predictions, enabling the final confidence score to more accurately reflect the overall quality of the detection box.

2. A high-quality localization-aware action recognition algorithm based on YOLOv11

2.1. Overall network architecture

The overall structure of the proposed high-quality localization-aware action recognition algorithm based on YOLOv11 is illustrated in Figure 1. The method adopts YOLOv11 as the fundamental framework and introduces targeted improvements to both the backbone network and the detection head, aiming to enhance the localization accuracy of human targets and the quality of confidence estimation while maintaining real-time performance. The complete model mainly consists of three components: a feature extraction module, a feature fusion module, and a high-quality localization-aware detection head.

During the feature extraction stage, an SPDConv downsampling structure is introduced to replace part of the conventional strided convolution operations. Through spatial rearrangement and channel reconstruction, this structure compresses feature resolution while minimizing the loss of spatial information, thereby strengthening feature representation capability and improving the perception of small-scale or complex human poses. In the feature fusion stage, a multi-scale feature pyramid structure is employed to effectively integrate human features at different scales, enhancing the model's adaptability to action targets of varying sizes.

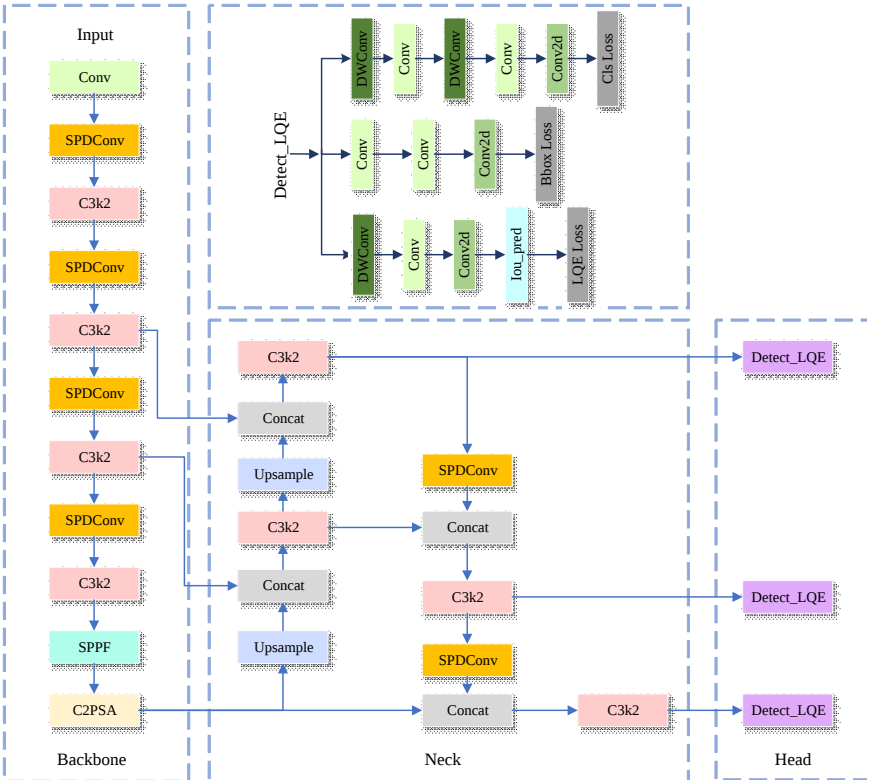

Figure 1. Overall network architecture

In the detection head, a three-branch decoupled structure based on Localization Quality Estimation (LQE) is constructed, including a bounding box regression branch, a classification branch, and a localization quality prediction branch. By integrating classification probability with the predicted localization quality, a localization-aware confidence recalibration mechanism is achieved, enabling the ranking of detection results to better reflect their actual localization accuracy. The model ultimately outputs high-quality human target regions, providing more accurate and stable input information for subsequent action recognition models.

Through the dual optimization of feature representation capability and confidence evaluation mechanisms, the proposed algorithm achieves high-precision localization and high-reliability filtering of human action regions, thereby improving overall action recognition performance.

2.2. SPDConv module

To alleviate the loss of spatial information caused by conventional convolutional downsampling with a stride of 2 during feature compression, this study introduces the SPDConv (Space-to-Depth Convolution) module into the backbone network to replace the original downsampling convolution structure. The overall structure of SPDConv is shown in Figure 2. By rearranging spatial information into the channel dimension and combining it with convolution operations for feature fusion, SPDConv performs feature map downsampling while preserving as much original spatial detail as possible.

Specifically, SPDConv first divides the input feature map into local regions of size 2×2 along the spatial dimensions. Within each local spatial block, the pixel information is not discarded directly; instead, it is remapped to the channel dimension through a Space-to-Depth (S2D) rearrangement operation, thereby completing feature reorganization. After the S2D operation, the spatial resolution of the feature map decreases

from the original $H \times W$ to $H/2 \times W/2$, while the number of channels correspondingly increases to four times the original.

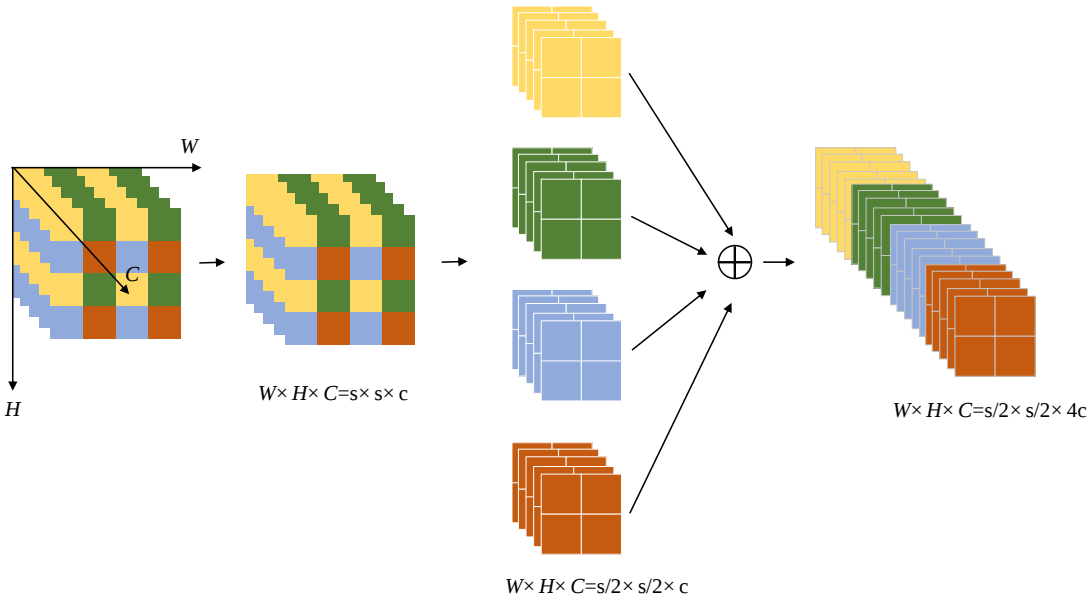

Figure 2. Structure of the SPDConv module

Following the spatial-to-channel rearrangement, SPDConv further introduces a convolution operation with a stride of 1 to fuse and compress the features along the channel dimension. This step enhances the model's capability to capture inter-channel correlations while effectively controlling the scale of the feature dimensionality. Through this process, SPDConv reduces feature resolution while preserving spatial information without introducing additional spatial downsampling loss. As a result, it provides richer and more stable feature representations for subsequent human target localization and action recognition tasks.

2.3. LQE module

The prediction network is primarily responsible for performing regression and classification on the multi-scale features produced by the neck network. In the original YOLOv11 architecture, the prediction network completes bounding box regression and category confidence prediction through shared feature branches. However, the characterization of localization quality for predicted bounding boxes mainly relies on the regression loss function, lacking explicit modeling of localization reliability. To address this limitation, this study introduces a Localization Quality Estimation (LQE) module into the original YOLOv11 prediction network and correspondingly improves the detection head into Detect_LQE. The LQE module jointly utilizes the geometric information produced by the bounding box regression branch and the high-level semantic features extracted from the classification branch to evaluate the localization quality of predicted bounding boxes. In this way, the reliability of target localization can be explicitly modeled. The overall structure of Detect_LQE is illustrated in Figure 3.

The improved prediction network consists of three components: a bounding box regression branch, a classification feature branch, and the Localization Quality Estimation (LQE) module. The input feature map is first fed separately into the regression branch and the classification feature branch. The regression branch predicts the bounding box parameters of the target through a multi-layer convolutional structure, while the classification feature branch extracts discriminative category semantic features through depthwise separable

convolutions and standard convolution operations. Subsequently, the predicted bounding box information generated by the regression branch is encoded and fused with the classification features within the LQE module to estimate the localization quality score of the target. The localization quality prediction generated by the LQE module is then used to adaptively modulate the classification confidence score. As a result, the final prediction score reflects both the category confidence and the localization accuracy of the target, thereby effectively improving the reliability and stability of the detection results.

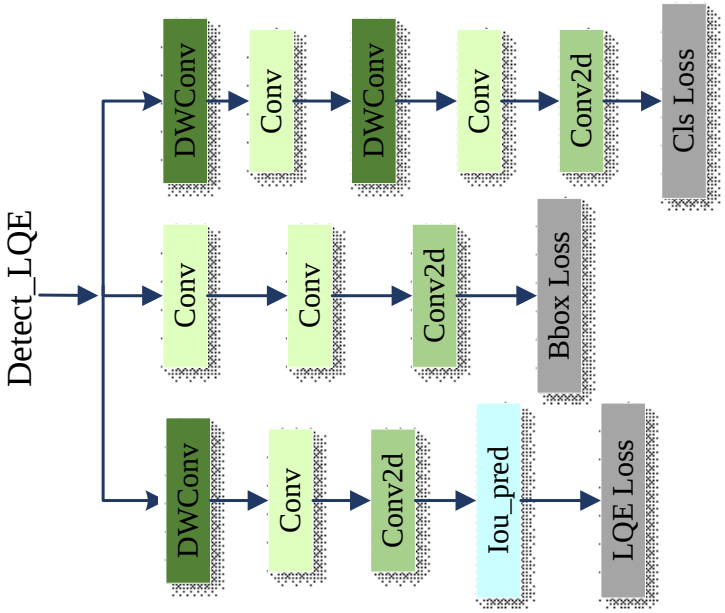

Figure 3. Structure of Detect_LQE

2.4. Improved loss function module

In YOLOv11, the bounding box regression stage typically employs an IoU-based loss function to constrain the overlap relationship between the predicted bounding box and the ground-truth box. However, in complex scenarios, the edge regions of bounding boxes are easily affected by background interference and annotation noise, which limits the localization accuracy of the model under high IoU thresholds. To address this issue, this study introduces a CIoU loss function based on Inner-IoU into the YOLOv11 bounding box regression loss, as shown in Equations (1) to (4).

$$IoU_{inner} = \frac{B_p \cap B_{gt}}{min(|B_p|,|B_{gt}|)} \tag{1}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w_p}{h_p}\right)^2 \tag{2}$$

$$\alpha = \frac{v}{(1-IoU_{inner})+v} \tag{3}$$

$$L_{box} = 1 - IoU_{inner} + \frac{\rho(b_p,b_{gt})}{c^2} + \alpha v \tag{4}$$

In these equations, B_P denotes the predicted bounding box, and B_{gt} denotes the ground-truth bounding box. w_{gt} and h_{gt} represent the width and height of the ground-truth box, while w_p and h_p denote the width and height of the predicted box. ρ represents the Euclidean distance between the center points of the predicted box and the ground-truth box. c denotes the diagonal length of the smallest enclosing rectangle that covers both the predicted box and the ground-truth box. v represents the aspect ratio consistency constraint

term, and α is a balancing coefficient. b_p denotes the center coordinate of the predicted bounding box, and b_{gt} denotes the center coordinate of the ground-truth bounding box.

While preserving the core design principles of the CIoU loss—including overlap ratio, center distance, and aspect ratio constraints—the proposed method computes the IoU constraint based on an inner region obtained by proportionally cropping the predicted and ground-truth boxes. In this study, the cropping ratio of the inner region is set to 0.7, allowing the loss function to focus more on the spatial overlap within the central region of the target. This design effectively reduces the interference caused by noise in boundary areas during loss computation and guides the model to focus on the precise alignment of the main body of the target during training.

3. Experimental results and discussion

3.1. Experimental environment and training settings

The hardware and software configurations used in the experiments are presented in Table 1. The hardware setup includes a high-performance CPU, GPU, and sufficient memory to ensure efficient model training and inference. The software configuration covers the programming language, development environment, and deep learning framework, ensuring the reproducibility of the experiments.

Table 1. Experimental environment configuration

Name	Version/Model
GPU	NVIDIA GeForce RTX 4060 Ti
CPU	AMD EPYC 7601 (32C/64T, 256GB)
Programming Language	python3.8.18
Development Environment	Pycharm2023
Deep Learning Framework	Pytorch(2.0.0+cu118)

The parameter settings used during model training are shown in Table 2.

Table 2. Experimental parameter settings

Configuration	Parameter
Input Image Size	640×640
Batch Size	32
Training Epochs	100
Initial Learning Rate	0.01
Weight Decay	0.0005
Optimizer	Stochastic Gradient Descent

3.2. Experimental results

3.2.1.1. Ablation study

To verify the effectiveness of the proposed improvements, ablation experiments were conducted on three aspects: the SPDConv module, Detect_LQE, and the improved loss function. The experimental results are shown in Table 3. The first row presents the detection performance of the baseline YOLOv11 model on a

subset of the Fall Detection Dataset, where the mAP@50 and mAP@50–95 values are 95.7% and 69.5%, respectively. These results serve as the baseline for comparison with the subsequent improved models.

Table 3. Ablation results on a subset of the fall detection dataset

Methods	Detect_LQE	SPDConv	Loss	mAP@50(%)	mAP@50-95(%)
YOLOv11	×	×	×	95.7	69.5
A	√	×	×	95.4	71.2
B	×	√	×	95.6	71.1
C	×	×	√	95.7	71.0
D	√	√	×	95.6	71.6
E	×	√	√	95.8	71.2
F	√	×	√	95.6	71.3
G (Proposed)	√	√	√	96.0	72.3

Methods A, B, and C introduce the Detect_LQE module, SPDConv module, and improved loss function, respectively, into the baseline model. The results show that all three individual improvements lead to a certain degree of enhancement in the mAP@50–95 metric, indicating that each module contributes positively to the overall detection performance of the model. Specifically, after introducing the Detect_LQE module, the model demonstrates a more noticeable improvement in detection accuracy under higher IoU thresholds, indicating that the module enhances the model's ability to characterize prediction quality. With the SPDConv module, the model's ability to represent structural information of targets is strengthened, leading to improvements in localization accuracy. The improved loss function also provides a stable improvement in detection performance, confirming its effectiveness in optimizing bounding box regression.

On this basis, Methods D, E, and F further evaluate different combinations of these modules. The results show that when multiple modules are used together, the model performance generally improves compared with single-module modifications, although the degree of improvement varies. This indicates that the modules have certain complementary effects, though their synergistic advantages are not yet fully realized. Finally, Experiment G (the proposed method) simultaneously integrates Detect_LQE, SPDConv, and the improved loss function into the baseline model. Under this configuration, the model achieves the best detection performance, with mAP@50 reaching 96.0% and mAP@50–95 reaching 72.3%. Compared with the original YOLOv11 model, the mAP@50–95 metric shows a significant improvement, which fully validates the effectiveness and rationality of the proposed improvements in the fall detection task.

3.2.2. Comparative experiments

To further validate the effectiveness of the proposed method, comparative experiments were conducted on a subset of the Fall Detection Dataset, where YOLOv5, YOLOv8, and YOLOv11 were selected as baseline models. The purpose was to analyze the detection performance differences among various models under the same dataset conditions. The experimental results are presented in Table 4.

Table 4. Comparative experimental results on a subset of the fall detection dataset

Methods	mAP@50	mAP@50-95
YOLOv5	93.6	64.3
YOLOv8	95.2	69
YOLOv11	95.7	69.5
Proposed Method	96.0	72.3

As shown in the table, with the continuous evolution of model architectures, the overall detection accuracy of the YOLO series has exhibited a gradual improvement in the fall detection task. Among them, YOLOv11 outperforms YOLOv5 and YOLOv8 in both evaluation metrics, achieving 95.7% for mAP@50 and 69.5% for mAP@50–95. These results indicate that YOLOv11 possesses certain advantages in terms of feature representation capability and target localization accuracy. In comparison, the method proposed in this paper further improves the performance, achieving 96.0% in mAP@50 and 72.3% in mAP@50–95. These results surpass those of YOLOv5, YOLOv8, and YOLOv11, demonstrating the effectiveness of the proposed approach for the fall detection task.

4. Conclusion

To address the challenges of large variations in human target scale, the loss of spatial detail information, and the inconsistency between detection confidence and actual localization quality in complex scenarios, this paper proposes an improved YOLOv11-based human action recognition method incorporating a high-quality localization-aware mechanism. By introducing the SPDConv downsampling structure into the backbone network and feature fusion stages, the proposed method effectively mitigates the spatial information loss caused by conventional strided convolutions, thereby enhancing the feature representation capability for small-scale human targets. Meanwhile, a Localization Quality Estimation (LQE) branch is incorporated into the detection head to explicitly model the IoU of predicted bounding boxes. Through a confidence reweighting strategy that integrates classification probability with localization quality, the approach improves the rationality of candidate box ranking and enhances the stability of the Non-Maximum Suppression (NMS) stage.

Experimental results demonstrate that the proposed method outperforms the baseline YOLOv11 model on both mAP@50 and mAP@50–95, confirming the effectiveness of the structural improvements and the localization quality–aware mechanism. While maintaining real-time detection capability, the proposed approach improves both the accuracy and reliability of human action recognition in complex environments.

Future work will further explore lightweight architectural designs and more efficient spatiotemporal feature fusion strategies in order to enhance the adaptability and generalization performance of the model in practical deployment scenarios.

References

[1] Wang, Z. (2009). *Motion target detection and abnormal behavior recognition in real-time video surveillance systems* [Master's thesis, Xidian University].

[2] An, Q. (2024). *Research on unsafe behavior detection of personnel in public waters based on pose estimation* [Master's thesis, Chongqing University of Science and Technology].

[3] Li, S. (2013). *Human pose action recognition and imitation algorithms for robots* [Doctoral dissertation, Shanghai Jiao Tong University].

[4] Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction, 5*(4), 1680–1716.

[5] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE, 111*(3), 257–276. https: //doi.org/10.1109/JPROC.2023.3238524

[6] Alzahrani, N., Bchir, O., & Ismail, M. M. B. (2025). YOLO-Act: Unified spatiotemporal detection of human actions across multi-frame sequences. *IEEE Access, 25*(10), 3013.

[7] Elnady, M., & Abdelmunim, H. E. (2025). A novel YOLO–LSTM approach for enhanced human action recognition in video sequences. *Scientific Reports, 15*(1), 17036.

[8] Chen, S., Liu, Y., Zhang, H., & Cai, Y. (2025). A human location and action recognition method based on improved Yolov11 model. *Discover Artificial Intelligence, 5*, Article 232. https: //doi.org/10.1007/s44163-025-00492-6