DAD-YOLO: EFFICIENT FIGHTING IMAGE RECOGNITION MODEL

Zheng LI[1,2], Dong WANG[1,2,*], Dengcong MU[1,2,*]

In today's society, fight detection technology is vital for preventing violence. To address issues like low accuracy, slow speed, and high complexity in existing video fight detection algorithms, we propose DAD-YOLO, an efficient model. Experiments show it outperforms traditional methods, with average precision increasing from 92.6% to 93.3%, while parameters and computational complexity decrease by 0.1M and 0.1G, respectively. This provides valuable reference for public security monitoring.

Keywords: Fighting Image Recognition, YOLOv8, C2f-DualConv, ADown

1. Introduction

Social harmony and stability are fundamental for the flourishing development of all aspects of society, including politics, economy, and culture. Building a harmonious society can accelerate economic growth and ensure social stability. In recent years, fight behavior recognition has become one of the research hotspots [1]. Fight behavior carries certain harmful consequences and can adversely affect educational environments and social order [2]. As a result, it has garnered increasing attention, becoming a frequent topic in media discourse and one of the major social issues. Addressing public security concerns is therefore extremely urgent. Traditional video surveillance relies on staff observing footage to detect anomalies, but this manual monitoring method falls short of practical demands. Automatically identifying and detecting fight behavior in videos has thus become crucial [3]. Deep learning is increasingly transforming human life [4-6]. Artificial intelligence technology is being increasingly applied in fields such as human-computer interaction, video retrieval, video surveillance, and intelligent security [7-9]. Dong et al. [10] used 2D CNN to extract spatial features and employed LSTM for time-series modeling to achieve violence recognition. Sudhakaran and Lanz [11] adopted a convolutional LSTM architecture to extract spatiotemporal features from

* Corresponding authors: Dong Wang (1422832698@qq.com) and Dengcong Mu (mudc@chnu.edu.cn).
[1] Anhui Province Key Laboratory of Intelligent Computing and Applications, Huaibei Normal University, Huaibei, 235000, China
[2] College of Physics and Electronic Engineering, Huaibei Normal University, Huaibei, Anhui 235000, China

frame differences, replacing fully connected layers with CNN to reduce computational load. Liang et al. [12] utilized attention mechanisms to minimize background interference. Islam et al. [13] not only employed average frame background suppression technology but also proposed separable convolution technology to further decrease computational requirements. Hanson et al. [14] introduced a bidirectional convolutional LSTM neural network architecture, combining forward and backward temporal information via bidirectional LSTM to enhance recognition capability. Asad et al. [15] proposed a frame-level violence recognition method that uses dual VGG16 networks to extract low-level and high-level features between frames, combined with LSTM for temporal modeling to improve overall feature representation. Additionally, Traoré et al. [16] utilized bidirectional gated recurrent units to boost recognition efficiency. Ye et al. [17] introduced a framework for the automatic detection of physical violence by leveraging the feeds from a network of distributed surveillance cameras, extracting key features after obtaining human body position information via lightweight OpenPose and employing SVM for fight behavior classification. These computer vision-based methods have significantly reduced the impact of environmental changes on models but generally suffer from low feature utilization and low model recognition accuracy. Therefore, based on the YOLOv8n algorithm, this paper introduces the neural network building block DBB, the Adown module, and the lightweight C2f-DualConv module, proposing a new detection model named DAD-YOLO. The model achieves a recognition rate of 93.3%, enabling effective detection of fighting image and providing valuable insights for detecting such behavior in public environments.

2. The improved YOLOv8n model

2.1 YOLOv8 algorithm

YOLOv8 is widely regarded as a state-of-the-art algorithm in the field of object detection. YOLOv8 uses the C2f module, which ensures lightweight design while obtaining rich gradient flow information. It adjusts the number of channels based on the model scale, significantly improving model performance. The anchor-free detection method predicts the center point of the target. It also predicts the aspect ratio of the target. This method delivers a boost to both detection speed and precision. Moreover, the decoupled head independently handles the tasks of object categorization and bounding box localization. Two parallel branches extract category and location information, this method improves the model's stability and more effectively defines the connection between spatial position and classification. The YOLOv8 network framework is shown in Fig. 1.

2.2 The DAD-YOLO model

Fig. 2 illustrates the architecture of the proposed DAD-YOLO model, which

builds upon the YOLOv8n baseline.

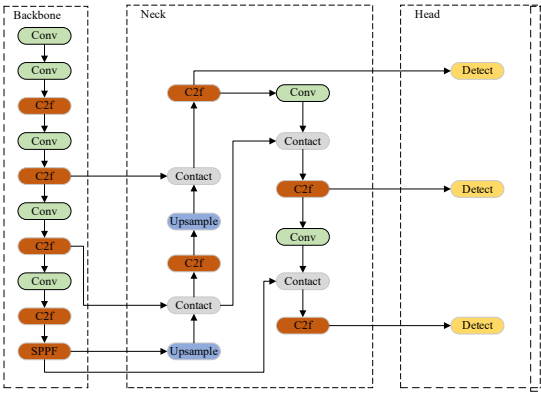


Fig. 1. YOLOv8 network framework

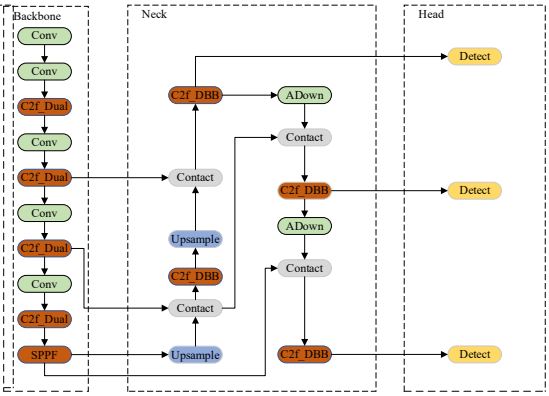


Fig. 2. The DAD-YOLO model

2.3 Introduction of the diverse branch block module

The concept of the Diverse Branch Block (DBB) originates from the seminal work of Ding and colleagues [18]. It leverages a multi-branch architecture to enhance CNN performance. Crucially, through structural re-parameterization, it achieves this without incurring any additional inference time cost. The DBB module leverages two key properties of convolution—homogeneity and additivity—As shown in Fig. 3, the DBB structure is designed to balance lightweight design with high accuracy. During the training phase, the original K×K convolution is enhanced using different combinations such as 1×1, 1×1−K×K, and 1×1−AVG. Through structural re-parameterization, the multi-branch structure used during training is transformed via the six transformation methods shown in Fig. 4. Specifically, batch normalization layers are first eliminated using BN fusion. Then, the 1×1−K×K branch is merged into a K×K convolutional kernel via convolutional concatenation. Next, the 1×1−AVG branch is converted into a K×K convolutional kernel by combining average pooling operations and convolutional concatenation. Subsequently, the 1×1 convolutional branch is expanded into a K×K convolutional kernel through multi-scale convolutional transformation. Finally, the convolutional kernels from all branches are first merged. Similarly, the biases from each branch are also combined. As a final step, these fused parameters form a single, equivalent kernel that is used for computation. Here, homogeneity means that convolving with a scaled kernel is equivalent to convolving with the original kernel and then scaling the resulting feature map. Additivity means that the sum of two convolutions is equivalent to convolving the input with the sum of their respective kernels. This relationship is formally defined by the following equation (1):

$$\begin{aligned} I \otimes (pF) &= p(I \otimes F), \forall_p \in R \\ I \otimes F^{(1)} + I \otimes F^{(2)} &= I \otimes (F^{(1)} + F^{(2)}) \end{aligned} \tag{1}$$

In the formula: The symbol I denotes the feature map that is fed into the convolutional layer. $\otimes$ denotes the convolution operator; F is a K×K convolution kernel; p is an operation parameter; The notations $F(1)$ and $F(2)$ represent two separate convolution filters.

The Diverse Branch Block (DBB) employs structural reparameterization to transform a multi-branch architecture during training into a single-branch, inference-efficient convolution. However, its specific advantages in violence detection are reflected in the following two aspects:

1). Enhanced multi-scale pose representation:

Violence involves rapid, varied body movements that manifest at different spatial scales—from localized fist impacts to full-body grappling. During training, DBB's parallel branches (1×1 conv, K×K conv, 1×1→K×K sequence, and 1×1→AVG) explicitly capture:

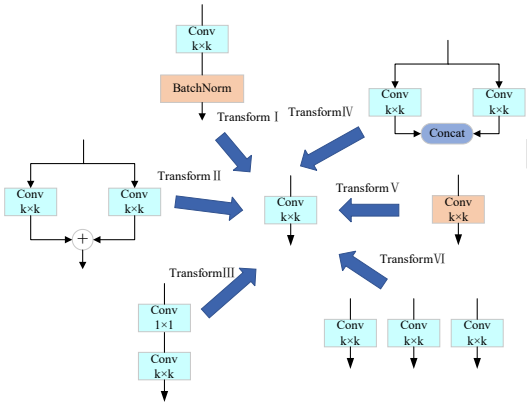


Fig. 3. Diagram of the DBB structure

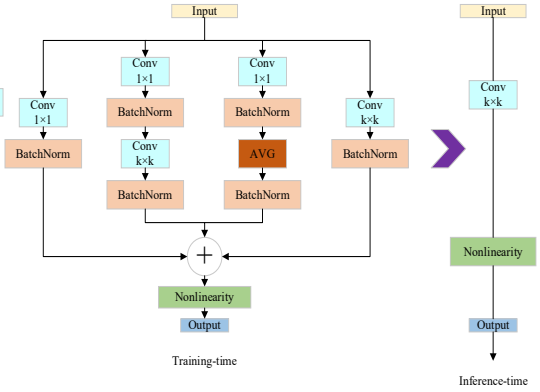


Fig. 4. Diagram of the DBB conversion method

- 1×1 branches: Focus on channel-wise dependencies, useful for recognizing clothing deformation during physical struggle
- K×K branches: Extract spatial patterns crucial for identifying punch/kick trajectories
- Average pooling branches: Preserve motion blur signatures common in rapid violence sequences

During inference, the reparameterization technique losslessly merges these complex parallel structures into a single standard K×K convolutional layer. This mathematical transformation ensures that the multi-scale receptive fields cultivated during training are fully encoded into the final convolution kernel. Consequently, this kernel functions as a pose-sensitive feature extractor that not only integrates the multi-scale knowledge learned for various fighting styles but also maintains high efficiency during inference.

2). Temporal stability through spatial enrichment:

Unlike generic object detection where spatial features suffice, violence

recognition benefits from motion-consistent spatial representations. DBB's training-time multi-scale features help stabilize per-frame detections by providing richer contextual information.

2.4 Introduction of the ADown module

During the deployment of device models, it is crucial to balance two key factors: model size and detection accuracy. While deep learning models such as ResNet and DarkNet can achieve high detection accuracy, their excessively large model sizes not only complicate deployment but also significantly prolong detection time. To address this issue, a more lightweight and efficient convolutional structure named Adown [19] is proposed, with its network architecture illustrated in Fig. 5. The ADown module begins by applying average pooling to reduce the spatial dimensions of the input feature map, minimizing computational load while retaining critical information, as shown in Equation (2). It then employs a partitioning strategy to efficiently allocate computations for the feature map, reducing redundant parameters and optimizing model inference efficiency. In the left branch, max pooling is applied to highlight salient features. A 1×1 convolution is subsequently applied for channel count reduction and feature fusion, as expressed in Equation (3). The right branch performs downsampling using a 3×3 convolution kernel to enhance the capture of local details, as shown in Equation (4). Finally, the feature maps produced by the two branches are merged via channel-wise concatenation. This processing step produces the final output feature map, Y, which is defined in Equation (5). This approach effectively integrates multiple types of features, thereby improving detection capability.

$$X1, X2 = Split(Avgpool2d(X)), X \in R^{B \times C \times H \times W} \tag{2}$$

$$Y1 = Conv_{1\times1}(Maxpool2d(X1)) \tag{3}$$

$$Y2 = Conv_{3\times3}(X2) \tag{4}$$

$$Y = Concat(Y1, Y2), Y \in R^{B \times 2C \times \frac{H}{2} \times \frac{W}{2}} \tag{5}$$

The design of the Adown module carries clear physical significance and architectural advantages for violence detection tasks. Violent behavior is typically characterized by sudden limb movements and complex interpersonal interactions, where traditional downsampling operations (such as strided convolution or pooling) tend to lose critical motion details and spatial relationships. Adown employs a dual-path heterogeneous downsampling structure: the left path uses max pooling to highlight areas of intense motion (e.g., punching trajectories), while the right path preserves relative positioning and contact information between individuals through 3×3 convolutions. Finally, channel concatenation enables multi-scale feature fusion. This structure essentially establishes a motion-aware feature selection mechanism, allowing the network to adaptively enhance violence-related features (such as rapid displacement and limb overlap) while suppressing static background interference.

As a result, it improves discriminative capability in dynamic conflict scenarios while reducing computational load.

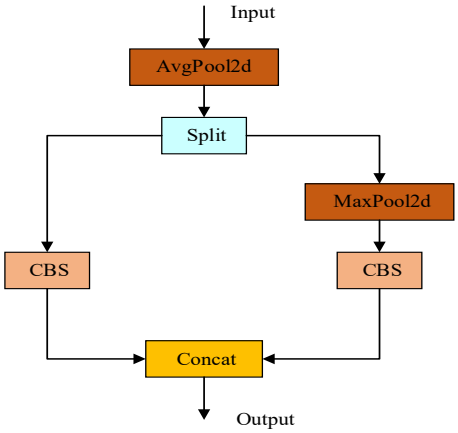


Fig. 5. ADown network structure diagram

2.5 Introduction of C2f-DualConv module

The development of lightweight convolutional modules offers a pathway to substantially lower the computational footprint of neural networks, streamlining their architecture. Fig. 6 depicts the architecture of the DualConv module [20], a lightweight efficiency optimization module, which incorporates parallel convolutional pathways.

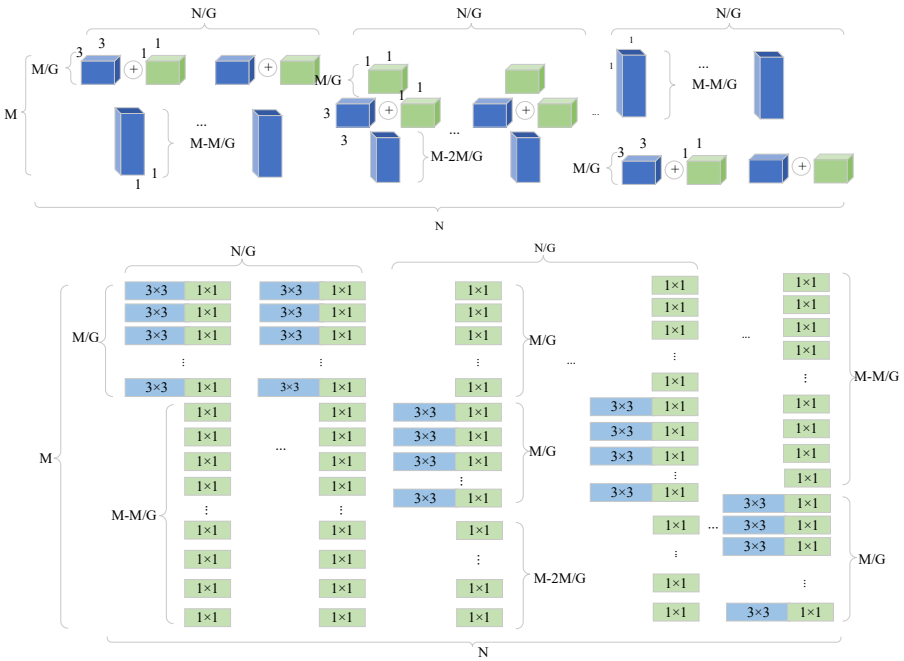


Fig. 6. DualConv architecture diagram

These are composed of 3×3 and 1×1 convolutional filters, which operate in parallel to gather channel-specific information from the input data. It maintains baseline detection accuracy while effectively reducing computational and parameter overhead. Consequently, the target model's recognition accuracy is improved. The input feature map undergoes processing via a 1×1 convolutional operation. This operation helps preserve the original information, facilitating deeper convolutions in better feature extraction and enabling information sharing between convolution operations. In this context, M corresponds to the input channel count, N designates the quantity of convolutional filters, and G signifies the group count for both group and dual convolution operations.

Compared to the traditional single convolution kernel (Conv) architecture, DualConv offers significant advantages in computational cost. As shown in Formula (6), the computational cost (FLSC) of a standard convolution layer is relatively high, while Formula (7) describes the computational cost (FLDC) of DualConv. Formula (8) provides the computational cost ratio of DualConv to a standard convolution layer. Thus, by incorporating group convolution technology, DualConv effectively lowers computational complexity and model size. At the same time, it maintains strong feature extraction performance. This provides a more lightweight and efficient network structure for practical deployment of fighting image recognition.

$$FLSC = D_o^2 \times K^2 \times M \times N \tag{6}$$

$$FLDC = D_o^2 \times M \times (9N/G + 1) \tag{7}$$

$$\frac{FLDC}{FLSC} = (9/G + 1)/K^2 \tag{8}$$

In the formula: D_o^2 represents the width and height of the output feature map, the convolutional kernels possess a size of K×K, and the group/dual convolutions comprise G groups. Meanwhile, the symbols M and N specify the quantity of input and output channels, correspondingly, with N also signifying the total number of kernels.

3. Test results and analysis

3.1 Dataset introduction

This study utilizes data sourced from the Real-Life Violence Situations Dataset [21]. The video data is first converted into images through frame extraction. The final dataset consists of 5,399 images; the dataset was partitioned into training and validation sets following a standard 9:1 split.

3.2 Experimental environment

The experiments run on Windows 11 and utilize the PyTorch framework to

build the network architecture. The specific experimental platform parameters are listed in Table 1. To ensure effective model training and performance optimization, the study adopts a unified parameter configuration: the learning rate is set to 0.001, the batch size is set to 4, the number of training epochs is set to 100, The number of worker threads is configured to 4, while the input image size is maintained at 640 pixels. Other parameters follow the default settings of the YOLOv8n model.

Table 1

Experimental Platform Parameters

Parameter	disposition
Operating System	Windows 11
GPU	NVIDIA RTX 3090
GPU Memory	24 GB
GPU Acceleration Environment	CUDA 11.3
Network Framework	PyTorch 1.11.0
Development Language	Python3.8

3.3 Model experimental results

To verify the performance improvements brought by the Diverse Branch Block mod-ule, ADown module, and C2f-DualConv module to the original YOLOv8n model, the Real Life Violence Situations Dataset was used. Each improvement was incrementally added to the YOLOv8n model: the Diverse Branch Block module denoted as Method A, the ADown module denoted as Method B, and the C2f-DualConv module denoted as Method C. Ablation experiments were designed for comparative validation, Table 2 presents the corresponding experimental outcomes.

Table 2

Ablation Experiment Results

Baseline Model	Improvement Methods	P/%	R/%	mAP@0.5/%	GFLOPS/G	Params/M	Fps/($f \cdot s^{-1}$)
YOLOv8n	——	88.7	87.8	92.6	8.2	3.0	73.3
YOLOv8n	+A	84.8	89.3	93.0	8.8	3.3	65.8
YOLOv8n	+B	84.9	89.0	92.7	8.0	2.8	85.2
YOLOv8n	+C	85.2	88.5	92.6	7.7	2.8	87.3
YOLOv8n	+A+B	86.5	89.1	93.3	8.6	3.1	62.5
YOLOv8n	+A+C	86.4	87.3	93.0	8.3	3.1	80.6
YOLOv8n	+B+C	82.8	88.8	91.6	7.6	2.7	75.6
YOLOv8n	+A+B+C	87.5	86.4	93.3	8.1	2.9	78.4

From the Table 2, after introducing the Diverse Branch Block module, ADown module, and C2f-DualConv module, the mAP@0.5 increases by 0.4%, 0.1%, and 0%, respectively. Among the three individual modules, the Diverse Branch Block module contributes the most to the improvement of model mAP@0.5, with an increase of 0.4 percentage points. The incorporation of the lightweight C2f-DualConv module into the backbone network serves to enhance its efficiency. A lightweight deep neural network is constructed, which efficiently

arranges convolutional filters using group convolution technology, reducing computational costs and the number of parameters. The combination of modules A and B achieves the highest mAP@0.5 of 93.3%. On this basis, further introducing the C2f-DualConv module maintains the mAP@0.5 at 93.3%, while reducing the number of parameters and computational complexity, and increasing the frame rate from 62.5 FPS to 78.4 FPS. The full three-module combination achieves the optimal comprehensive balance between detection accuracy and inference efficiency, which is more suitable for real-time deployment on edge monitoring devices. This indicates that the model presented in this paper is capable of effectively detecting fighting image.

In the ablation experiments, we observed that the combination of ADown and C2f-DualConv (YOLOv8n + B + C) failed to achieve performance improvement, and its mAP@0.5 decreased by 1.0 percentage points compared with the baseline, which is the largest performance drop among all ablation combinations. The analysis suggests that the superposition of two lightweight downsampling and convolution structures will cause obvious loss of feature representation ability, resulting in a decline in detection accuracy. Analysis suggests that while ADown preserves motion information during downsampling, it may also lead to a simplification of spatial and channel information in feature maps, which in turn affects the C2f-DualConv module's ability to extract interactive features. However, when the DBB module was introduced, the combination of all three (A + B + C) achieved significant performance gains. Through its multi-branch structure and structural reparameterization, DBB provides richer feature representations, effectively compensating for the potential information loss caused by ADown and offering more discriminative input features for C2f-DualConv. This indicates that the synergy among modules requires careful consideration of the continuity and complementarity of the feature flow. While combining individual modules may not directly lead to improvements, appropriately designing and arranging them within the overall architecture can still leverage their respective advantages to achieve superior detection performance.

3.4 Comparative experiments

3.4.1 Comparison of different convolutional detection modules

This study aims to validate the superiority and effectiveness of two key improvements. We introduce the Diverse Branch Block (DBB) and ADown modules into the original network, comparisons were made with the partial convolution module [22], the NTB convolution module, and SPD-Conv module [23]. The corresponding experimental results are presented in Table 3.

Table 3

Results of Different Small Object Detection Networks

Baseline Model	P/%	R/%	mAP@0.5/%	GFLOPS/G	Params/M
+Diverse Branch Block	84.8	89.3	93.0	8.8	3.3

+ADown	84.9	89.0	92.7	8.0	2.8
+partial	84.0	90.1	92.6	8.7	3.6
+NTB	86.4	88.8	92.8	8.6	3.5
+SPD-Conv	83.9	87.6	91.6	7.5	2.7

As shown in Table 3, compared to the partial and NTB convolution modules, the Diverse Branch Block (DBB) convolution module achieves higher average precision, with an improvement of 0.4% in average precision after introducing the DBB module. Additionally, the integration of the ADown module yields a dual benefit: it elevates average precision while simultaneously lowering both the parameter count and computational demands. Although the parameter count is lower after introducing SPD-Conv, the average precision decreases by 1.4 percentage points. In conclusion, incorporating the Diverse Branch Block and ADown modules is more beneficial for improving model performance.

3.4.2 Comparison of different C2f modules

To better demonstrate the superiority and effectiveness of introducing the C2f-DualConv module into the original network, comparisons were made with the C2f_MSBlock module, C2f_repghost module, and the Large Separable Kernel Attention module. Table 4 compares the performance metrics obtained from the experiments.

Table 4

Comparison of Different Improvements in the C2f Module

Baseline Model	P/%	R/%	mAP@0.5/%	GFLOPS/G	Params/M
+C2f-DualConv	85.2	88.5	92.6	7.7	2.8
+C2f_MSBlock	82.4	90.7	92.1	7.9	2.8
+C2f_repghost	86.1	85.5	91.6	7.0	2.6
+C2f_LSKA_Attention	83.7	88.9	91.8	7.4	2.6

According to the data in Table 4, C2f-DualConv significantly reduces the computational load and parameter count compared to C2f_MSBlock without sacrificing the average accuracy; and it has a clear advantage in terms of accuracy compared to C2f_repghost and C2f_LSKA_Attention. This module can effectively reduce the model's parameters and computational cost while maintaining the average accuracy.

3.4.3 Comparison of experimental results across different models

We compared the improved model against several established detectors. This comparative analysis provides a clear visualization of its performance. Such as YOLOv8n, YOLOv5n, YOLOv5s, YOLOv5m, and YOLOv8s [24]. As indicated in Table 5, the DAD-YOLO model surpasses its counterparts in average precision, achieving a superior level of accuracy, surpassing YOLOv8n, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv8s and YOLOv11n by 0.7, 3.2, 2.8, 1.3, 0.1 and 2 percentage points, respectively. Additionally, the parameter count of our model is relatively small. While YOLOv5s and YOLOv5m have larger computational loads

and parameter counts, their precision is low-er. While YOLOv5n and YOLOv11n exhibit lower computational costs and parameter counts compared to our algorithm, their accuracy decreases by 3.2 and 2.0 percentage points, respectively. The parameter counts, computational loads, and precision levels of YOLOv5s, YOLOv5m, and YOLOv8s are all inferior to those of our algorithm. At the same time, in order to verify the superiority of the model, a new comparative experiment was conducted using the fight dataset. The experimental results are shown in Table 6.

Table 5

Comparison of Experimental Results of Different Models on the Real Life Violence Situations Dataset

Baseline Model	P/%	R/%	mAP@0.5/%	GFLOPS/G	Params/M
YOLOv8n	88.7	87.8	92.6	8.2	3.0
YOLOv5n	80.1	91.3	90.1	4.2	1.7
YOLOv5s	83.1	88.7	89.5	16.0	7.0
YOLOv5m	86.2	87.0	92.0	48.2	20.9
YOLOv8s	87.9	88.5	93.2	28.6	11.1
YOLOv11n	85.5	84.6	91.3	6.4	2.6
Ours	87.5	86.4	93.3	8.1	2.9

Table 6

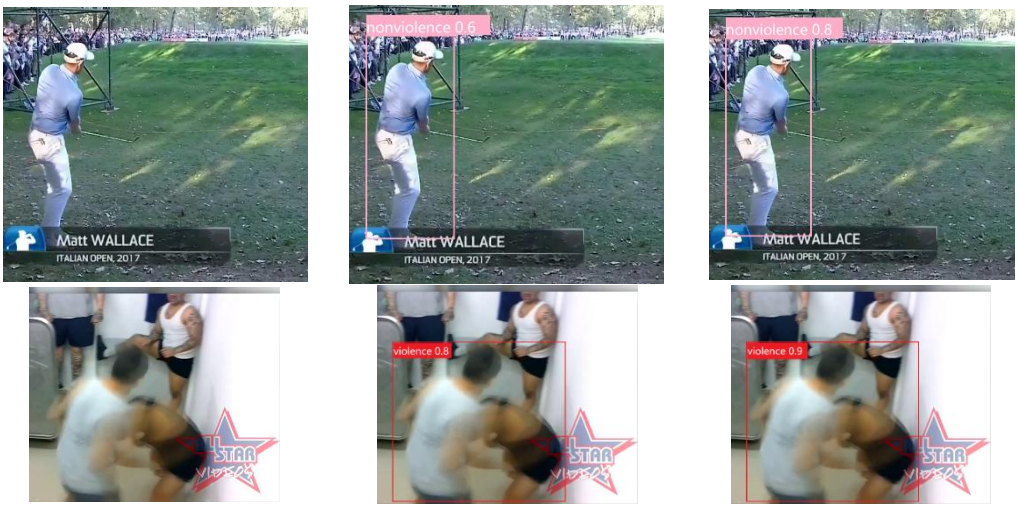
Comparison of Experimental Results of Different Models on the fight Dataset

Baseline Model	P/%	R/%	mAP@0.5/%	GFLOPS/G	Params/M
YOLOv8n	88.6	86.6	91.9	8.2	3.0
YOLOv5n	87.3	75.7	87.8	4.2	1.7
YOLOv5s	84.3	76.7	86.8	16.0	7.0
YOLOv5m	79.1	78.6	85.7	48.2	20.9
YOLOv8s	86.2	78.8	87.0	28.6	11.1
YOLOv11n	86.8	74.8	87.0	6.4	2.6
Ours	93.3	86.4	92.5	8.1	2.9

The DAD-YOLO model outperforms other similar models in terms of average accuracy. Its accuracy rate has reached a higher level, which is 0.6, 4.7, 5.7, 6.8, 5.5 and 5.5 percentage points higher than YOLOv8n, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv8s and YOLOv11n respectively. In addition, the number of parameters in our model is relatively smaller. While YOLOv5s and YOLOv5m have larger computational costs and parameter quantities, their accuracy is lower. Compared with our algorithm, YOLOv5n and YOLOv11n have lower computational costs and parameter quantities, but their accuracy has decreased by 4.7 and 5.5 percentage points respectively. The parameter quantity, computational cost and accuracy level of YOLOv5s, YOLOv5m and YOLOv8s are all lower than our algorithm. After comparing different models, the DAD-YOLO model demonstrates superior overall performance and better detection effectiveness for fighting image.

3.5 Visualization analysis

This article takes two scenarios in the public dataset as examples for detection and comparison. By comparing YOLOv8n with the improved algorithm model, the superiority of the enhanced model becomes more evident, as shown in the detection comparison results in Fig. 7.



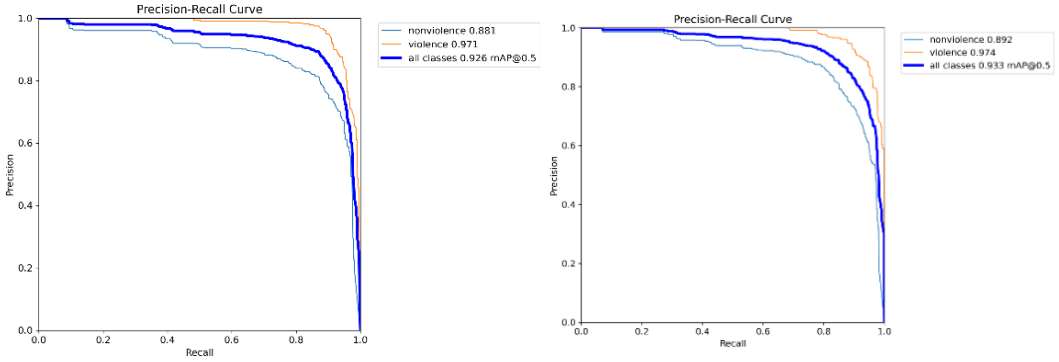
(a) Original article (b)YOLOv8n (c) Algorithm in this paper

Fig. 7. Comparison of detection results before and after algorithm improvement

The magnitude of the confidence score directly corresponds to the probabilistic assurance that a target object is present within the predicted bounding box. and also demonstrates that the model has learned more detailed target information. Experimental results confirm the effectiveness of our improved YOLOv8n. It significantly enhances the accuracy of fighting image detection. Fig. 8 compares the precision-recall relationships of the baseline YOLOv8n model and our proposed algorithm. The improvements lead to a significant increase in the algorithm's average precision.

3.6 Module adaptability analysis and synergistic mechanisms

Violence detection, as a specialized task in action recognition, poses unique challenges beyond mere object localization, requiring the capture of dynamic interactions, pose variations, and temporal continuity. The proposed DAD-YOLO model integrates three modules—DBB, ADown, and C2f-DualConv—into the YOLOv8n architecture. Their combination is deliberately designed to address these challenges through targeted and synergistic mechanisms, as analyzed below:



(a) YOLOv8n (b) Algorithm in this paper

Fig. 8. Comparison of P-R curves between YOLOv8n algorithm and our improved algorithm

(1) DBB module: enhancing feature representation for diverse poses

Violent image often involves varied and exaggerated body poses. The DBB module enriches feature diversity during training through its multi-branch structure (e.g., 1×1 convolution, K×K convolution, average pooling branches) and maintains inference efficiency via structural re-parameterization. This design enables the model to learn richer spatial representations, particularly for key movements such as punching, kicking, and grappling. As shown in Table 2, the inclusion of DBB improves mAP@0.5 by 0.4%, validating its effectiveness in enhancing discriminative capability for violence-related poses.

(2) ADown module: preserving motion information to handle dynamic blur

Traditional downsampling operations (e.g., strided convolutions) often lose fine-grained motion information, which is critical in violence videos characterized by rapid movement and motion blur. The ADown module employs an average pooling + dual-branch convolution design to retain spatial details and motion cues while reducing dimensionality. Its left branch highlights salient features via max pooling, while the right branch captures local details via 3×3 convolution. The final channel-wise concatenation integrates multi-scale information, helping the model perceive limb movement directions and interaction patterns even in lower-resolution feature maps, thereby improving detection stability in dynamic scenes.

(3) C2f-DualConv module: enhancing interaction-aware perception with low computation

Violence typically involves multi-person interactions, requiring the model to jointly represent individual poses and their relationships. C2f-DualConv employs a parallel dual-kernel structure (3×3 convolutions for local interaction features, 1×1 convolutions for channel fusion) combined with group convolution, reducing parameters while strengthening the focus on interaction regions (e.g., hand-torso overlap areas). This module directs network attention toward contact zones between individuals, improving the discriminative accuracy of fighting image recognition.

As shown in Table 4, C2f-DualConv maintains mAP while significantly reducing GFLOPS and parameters.

(4) Synergistic effects among modules

The three modules form a complementary and synergistic system for violence detection:

DBB and ADown synergy: DBB enhances feature diversity, while ADown preserves motion information—together they improve the model's ability to represent dynamic poses.

ADown and C2f-DualConv synergy: ADown provides multi-scale features, and C2f-DualConv performs lightweight interaction modeling on top, enabling efficient recognition of multi-person interaction scenarios.

Overall synergy: DBB increases feature capacity, ADown optimizes feature propagation, and C2f-DualConv strengthens interaction awareness, resulting in a lightweight, efficient, and task-specialized architecture for violence detection.

4. Conclusion

To mitigate missed and false detections in fighting image recognition, we introduce DAD-YOLO, a novel model engineered for balanced accuracy and deployment efficiency. Through the architectural integration of the Diverse Branch Block, ADown, and the C2f-DualConv lightweight efficiency optimization module, the model achieves an accuracy of 93.3% on the Real-Life Violence Situations Dataset while reducing the number of parameters. Additionally, computational load and complexity are significantly decreased, enabling effective application in real-time fighting behavior detection. The scale of the dataset will be expanded in the future, incorporating more diverse scenarios of fighting behaviors, and further improving the model's generalization capability to enhance its adaptability in complex environments.

Funding

This research was funded by Anhui Province Key Laboratory of Intelligent Computing and Applications (No.AFZNJS2024KF03); This research was funded by Provincial Quality Engineering Projects for Education in the New Era of Anhui Province, grant number 2023lhpysfjd044, 2024yjsxxsfkc030; This research was funded by Quality Engineering Project for Degree and Graduate Education at Huaibei Normal University, grant number 2024xxsfkc002. This research was funded by top-tier Online Courses(2023ylkc015 , 2023xskc028).

REFERENCES

[1] *Li Jincheng, Yan Ruiao, Dai Xuejing.* Violence Detection Based on Improved Attention Mechanism and VGG-Bi LSTM. Modern Electronics Technique, 2024, 47(21): 131-138.

DOI: 10.16652/j.issn.1004-373x.2024.21.021.
[2] *Wang Haoyu.* New Trends and Implications of School Violence Prevention in South Korean Primary and Secondary Schools. Journal of Shanghai Educational Research, 2024, (12): 25-31. DOI: 10.16194/j.cnki.31-1059/g4.2024.12.001.
[3] *Tan Dengtai, Zhao Jinlong, Wang Yiqun, et al.* Violence Recognition Model for Inter-class Mutual Learning. Journal of People's Public Security University of China (Science and Technology), 2023, 29(04): 75-83. DOI: 10.3969/j.issn.1007-1784.2023.04.010
[4] *He Gongshan, Zhao Chuanlei, Jiang Jinhu, et al.* A Survey of Data Storage Technologies for Deep Learning. Chinese Journal of Computers, 1-56 [2025-02-17]. DOI: 10.11897/SP.J.1016.2025.01013
[5] *Li Shuhui, Cai Wei, Wang Xin, et al.* Review of Infrared and Visible Image Fusion Methods in Deep Learning Frameworks. Computer Engineering and Applications, 1-19 [2025-02-17]. DOI: 10.3778/j.issn.1002-8331.2410-0012
[6] *Li Tongwei, Qiu Dawei, Liu Jing, et al.* Review of Human Behavior Recognition Based on RGB and Skeletal Data. Computer Engineering and Applications, 1-26 [2025-02-17]. DOI: 10.3778/j.issn.1002-8331.2407-0456
[7] *Sheng Qincheng, Tang Wei, Qin Hao, et al.* A review on AI-driven face robot for human-robot expression interaction. Scientia Sinica (Technologica), 2025, 55(10):1603-1637. DOI: 10.1360/SST-2025-0217
[8] *Huang Haoyu.* A Media Study on Hidden Roles in Video Surveillance. Modern Publishing, 2025, (05):84-92. DOI: 10.3969/j.issn.2095-0330.2025.05.009
[9] *Lin Jun 'an, Bao Cuizhu, Dong Jianfeng, et al.* Multilingual Text-Video Cross-Modal Retrieval Model via Multilingual-Visual Common Space Learning. Chinese Journal of Computers, 2024, 47(09):2195-2210. DOI: 10.11897/SP.J.1016.2024.02195
[10] *Dong Z, Qin J, Wang Y.* Multi-stream deep networks for person to person violence detection in videos // Chinese Conference on Pattern Recognition, November 5-7, 2016, Chengdu, China. Singapore: Springer Singapore, 2016:517-531. DOI: 10.1007/978-981-10-3002-4_43
[11] *Sudhakaran S, Lanz O.* Learning to detect violent videos using convolutional long short-term memory//201714thIEEEInternationalConferenceon Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017:1-6. DOI: 10.1109/AVSS.2017.8078468
[12] *Liang Q, Li Y, Yang K, et al.* Long-term recurrent convolutional network violent behaviour recognition with attention mechanism. MATEC Web of Conferences, 2021,336(01):5013. DOI: 10.1051/matecconf/202133605013
[13] *Slam Z, Rukonuzzaman M, Ahmed R, et al.* Efficient two-stream network for violence detection using separable convolutional lstm //2021International Joint Conference on Neural Networks (IJCNN). IEEE, 2021:1-8. DOI: 10.48550/arXiv.2102.10590
[14] *Hanson A, Pnvr K, Krishnagopal S, et al.* Bidirectional Convolutional LSTM for the Detection of Violence in Videos//European Conference on Computer Vision. Springer, Cham, 2019. DOI: 10.1007/978-3-030-11012-3_24.
[15] *Asad M, Yang J, He J, et al.* Multi-frame feature-fusion-based model for violence detection. The Visual Computer, 2021, 37(6):1415-1431. DOI: 10.1007/s00371-020-01878-6.
[16] *Traoré, Abdarahmane, Akhloufi M A.* 2D bidirectional gated recurrent unit convolutional Neural networks for end-to-end violence detection In videos. 2024.DOI: 10.1007/978-3-030-50347-5_14.
[17] *Ye L, Yan S, Zhen J, et al.* Physical Violence Detection Based on Distributed Surveillance Cameras. Mobile Networks and Applications, 2022, 27(4):1688-1699. DOI: 10.1007/s11036-021-01865-8.
[18] *Ding X, Zhang X, Han J, et al.* Diverse Branch Block: Building a Convolution as an Inception-like Unit. 2021. DOI: 10.48550/arXiv.2103.13425.
[19] *Wang C Y, Yeh I H, Mark Liao H Y.* YOLOv9: learning what you want to learn using

programmable gradient information//Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland, 2024: 1-21. http://arxiv.org/abs/2402.13616

[20] *Zhong J, Chen J, Mian A.* DualConv: Dual Convolutional Kernels for Lightweight Deep Neural Networks. 2022. DOI: 10.48550/arXiv.2202.07481.

[21] *Soliman M M, Kamal M H, Nashed E M, et al.* Violence Recognition from Videos using Deep Learning Techniques//2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS).2019. DOI: 10.1109/ICICIS46948.2019.9014714.

[22] *Chen J, Kao S H, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[J]. ArXiv, 2023, abs/2303.03667.*

[23] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459. DOI:10.1007/978-3-031-26409-2_27

[24] *Wang J, Han Q, Ge K, et al.* SPPF-CGA: Marine Garbage Detection and Image Enhancement in Turbid and High-Dynamic Underwater Environments. Journal of Ocean University of China. JOUC, 2025, 24(5). DOI: 10.1007/s11802-025-6087-5.