International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Smoking Behavior Recognition Method Based on Improved YOLOV8

Weixiong zhang[1], Anton Louise De Ocampo*[2]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: Cigarettes, as critical indicators of smoking behavior, often appear as small targets in video surveillance. Due to camera resolution limits and shooting distance, cigarettes may only appear as bars, increasing false detection possibilities. Additionally, complex environments or lighting can cause partial or complete occlusion of cigarette targets, complicating detection. To accurately and quickly identify smoking behavior in such conditions, an improved YOLOV8-based method is proposed to enhance real-time recognition performance. The improved YOLOV8 backbone network uses deformable convolution layers to replace some conventional layers, extracting the smoking behavior feature map. Deformable pooling layers are added to process and extract critical features. In the enhanced neck network, a residual attention module combines channel and spatial attention mechanisms to refine the critical feature map. The detection head network employs scale, spatial, and task awareness attention modules to fuse important features from different scales, locations, and tasks, respectively. This fusion determines the smoking behavior recognition frame and score, optimizing the recognition process. An improved loss function further refines the recognition box, boosting accuracy. Experiments demonstrate that this method effectively extracts critical features of smoking behavior, achieving better feature extraction and accurate identification across varying lighting conditions, resulting in high accuracy.

Keywords: *Improving YOLOV8, Smoking behavior recognition, Deformable colluvium, Attention mechanism, Residual attention, Nonmaximum suppression*

1. Introduction

Smoking not only poses a serious threat to personal health, such as increasing the risk of lung cancer and cardiovascular disease but also adversely affects the surrounding environment and the health of others [1], such as the harm of secondhand smoke. Therefore, accurate and efficient identification of smoking behavior is crucial for promoting public health, disease prevention, and maintaining social order [2]. Traditional smoking behavior identification methods mainly rely on manual observation and judgment, which has many limitations. First, manual observation is inefficient and cannot cover all possible smoking scenarios. Secondly, subjective factors easily influence manual judgment, resulting in low recognition accuracy. In addition, smoking behavior in public places [3] is often difficult to monitor manually in real-time and effectively. Therefore, finding a method to automatically and accurately identify smoking behavior has become a hot and challenging point in current research. The study of smoking behavior recognition technology has crucial academic value and a wide range of application prospects. First, in public places such as shopping malls, stations, airports, etc., smoking detection systems can be set up to monitor and warn against smoking behavior in real time [4] to maintain public order and public health. Secondly, in education, smoking behavior recognition technology can be used in schools, training institutions, and other places to monitor whether students have smoking behavior and to provide timely intervention and correction of illegal behavior. In addition, in the medical field, smoking behavior recognition technology can also be used to assist doctors in diagnosis and treatment to help patients better control smoking behavior [5].

A series of technical challenges must be solved to achieve accurate and efficient smoking behavior recognition. First, smoking behavior is diverse and complex, and different smokers, different smoking environments, and different smoking attitudes will affect the recognition results. Therefore, how to design a deep learning model that can adapt to various situations is the key to smoking behavior recognition technology. Secondly, smoking behavior recognition must process many images and video data, which requires high computing resources. Optimizing the algorithm and model structure to improve calculation efficiency is also a problem that needs to be solved by smoking behavior recognition technology. Given the above challenges, scholars at home and abroad have conducted much research on smoking behavior recognition in recent years. For example, Tran, D. N, and others first introduced the advantages of IoT technology in data acquisition, including real-time capture of human motion data and environmental parameters through various sensors. Then, the human behavior recognition algorithms are elaborated in detail, including key technologies such as feature extraction

1College of Engineering, Batangas State University, Batangas City, 4200, Philippines.
ORCID: 0009-0009-6009-3070
2Digital Transformation Center, Batangas State University, Batangas City, 4200, Philippines
ORCID: 0000-0002-6280-6259
**Corresponding email: antonlouise.deocampo@ieee.org*

and classifier design. These algorithms can extract meaningful features from complex original data and conduct accurate behavior classification through machine learning models [6]. Human behavior recognition algorithms need to balance accuracy and speed. Increasing the computational complexity and feature dimensions may be necessary to improve the algorithm's accuracy. It is necessary to reduce the computational complexity and feature dimensions to improve the algorithm's speed. This trade-off makes the human behavior recognition algorithm perform poorly in complex scenes. Mahalakshmi, V, and others first collected video data containing a variety of human behaviors. Then, the deep learning model was used to extract the feature representation of human behavior from the video, including spatial features (such as human posture, shape, etc.) and temporal features (such as motion trajectory, speed, etc.). Finally, a specific few-shot learning algorithm is used to learn and train the extracted features to recognize new category behavior [7]. In an environment with poor lighting conditions (such as too bright, too dark, shadow, etc.), the quality of video images will significantly decline, affecting the feature representation of human behavior extracted by the depth learning model, resulting in the model being unable to accurately capture the spatial features such as human posture and shape, as well as the temporal features such as motion trajectory and speed, reducing the recognition accuracy. Chen, X and Dinavahi, V map human skeleton points to nodes in the graph structure through the spatiotemporal graph convolution network and define the connection relationship between nodes to build the spatiotemporal graph. The spatial features of human skeleton points are extracted through graph convolution operation, and the changes of these features over time are captured through temporal convolution structure to realize accurate recognition of human behavior patterns [8]. Under complex lighting conditions, the recognition of bone points will be affected, especially when the light is too strong or weak; the extraction of bone points will become inaccurate, resulting in subsequent image convolution operations based on wrong data, thus affecting the accuracy of behavior recognition. Farooq, M. U and others first used motion shape descriptors to capture the dynamic changes of the human body between consecutive frames. This motion-shape information contains the spatiotemporal information of behavior and can effectively represent the subtle differences in human posture and motion. The convolutional neural network (CNN) model is trained to extract deep-level behavior characteristics by taking this motion shape information as input. Secondly, we introduce the Long Short Memory Network (LSTM) model to process the features extracted by CNN and further learn the temporal dynamics of behavior. Finally, CNN and LSTM models are integrated to build an end-to-end behavior recognition framework, automatically learn the motion shape information extracted from the original video, and generate accurate behavior classification [9]. Baek, K, and others first obtained real-time human behavior and environmental status data through multimodal sensors (such as cameras, infrared sensors, accelerometers, etc.). Then, the deep learning model is used to extract and fuse the features of these data to capture the temporal and spatial characteristics of human behavior and the dynamic changes in the environmental state. Finally, a behavior recognition algorithm based on environmental response requirements is designed to automatically adjust the parameters and thresholds of behavior recognition according to the current environmental status and user characteristics to achieve effective recognition and response to different behaviors [10]. In a complex environment, if there are other moving objects or background noise, these interference factors will be confused with the human motion shape descriptor, resulting in the CNN model being unable to distinguish human behavior accurately. In this case, the robustness of the model will be challenged. The deep learning model needs a lot of labeled data for training to ensure its generalization ability under various environments and lighting conditions. However, in real-world applications, collecting data sets covering all possible scenes and lighting conditions is difficult, resulting in a decline in the model's performance in certain specific environments.

The improved YOLOV8 inherits the advantages of YOLO series algorithms [11] and has a fast detection speed. This real-time feature is essential for smoking behavior recognition because smoking behavior may occur at any time, and the algorithm needs to respond quickly [12]. Unlike simple object detection, smoking behavior recognition must consider spatial and temporal characteristics. This means that in addition to detecting the presence of cigarettes, it is also necessary to analyze the movement patterns of smokers to determine whether they are smoking accurately. The complexity of this behavior is beyond the processing scope of a simple object detection algorithm. Therefore, the research is based on the improved YOLOV8 smoking behavior recognition method. By improving YOLOV8 to deal with objects of different sizes and distances, the algorithm has good adaptability to various sizes and positions of cigarettes in smoking behavior. Even in complex environments or lighting conditions, the algorithm can accurately identify smoking behavior, significantly improving public health and maintaining social order.

2. The general framework of the smoking behavior identification method

The YOLOV8 model mainly includes three parts, namely the backbone network, neck network, and detection head network. In the feature extraction of the ordinary convolution layer in the backbone network, due to the significant changes in the smoker's posture, movement, and the size and shape of smoking tools (such as cigarettes), the conventional convolution layer uses convolution cores of fixed size and shape [13], which will reduce the recognition performance. The deformable convolution layer adds a learnable offset to each sampling point of the convolution kernel so that the shape of the convolution kernel can be adaptively adjusted according to the target object's shape, thus improving the model's ability to simulate the object's deformation. Therefore, the deformable convolution layer is used to replace some conventional convolution layers in the C2f module of the backbone network. The deformable pooling layer is introduced to improve the backbone network so that the YOLOV8 model can better adapt to target objects with different scales and shapes, better distinguish smoking behavior from similar behaviors, and enhance the robustness of the YOLOV8 model.

In the recognition of smoking behavior, it is often necessary to focus on specific parts of smokers (such as mouth, hand, etc.). In the neck network, introducing a residual attention module, combined with channel attention and spatial attention mechanism, can help the YOLOV8 model pay more attention to this crucial information in the feature extraction process and improve recognition accuracy. At the same time, smoking behavior may vary significantly in different scenes, such as lighting conditions, angle changes, occlusion, etc. After introducing the residual attention module to improve the neck network, the YOLOV8 model can also better adapt to these changes and improve the recognition performance in different scenarios.

In the detection head network, the scale-aware attention module, spatial-aware attention module, and task-aware module are introduced to process the smoking behavior feature map, and the improved non-maximum inhibition algorithm is used to determine the smoking behavior recognition frame and the recognition frame score. Improving the detection head network can improve the recognition accuracy and accelerate the recognition speed. Elements involved in smoking behavior (such as cigarettes, mouth, etc.) may have significant changes in scale in different scenes. The introduction of an aware attention module can adaptively adjust the weight according to the scale of the input features so that the model has excellent response-ability to the features of different scales and ensures that smoking behavior can be accurately identified at various scales. The spatial position of the image's critical elements of smoking behavior is not fixed. Introducing the spatial awareness and attention module can teach and emphasize the spatial position information of the essential elements of the input features, which makes the model pay more attention to these critical positions and improves the accuracy of smoking behavior recognition. The recognition of smoking behavior usually involves multiple tasks, such as classification (whether it is a smoking behavior), positioning (the position of cigarettes or mouth), etc. Introducing a task awareness module can help the model allocate attention between different tasks, ensure that each task can be handled appropriately, and thus improve the overall performance.

The overall framework of smoking behavior recognition using improved YOLOV8 is shown in Figure 1.

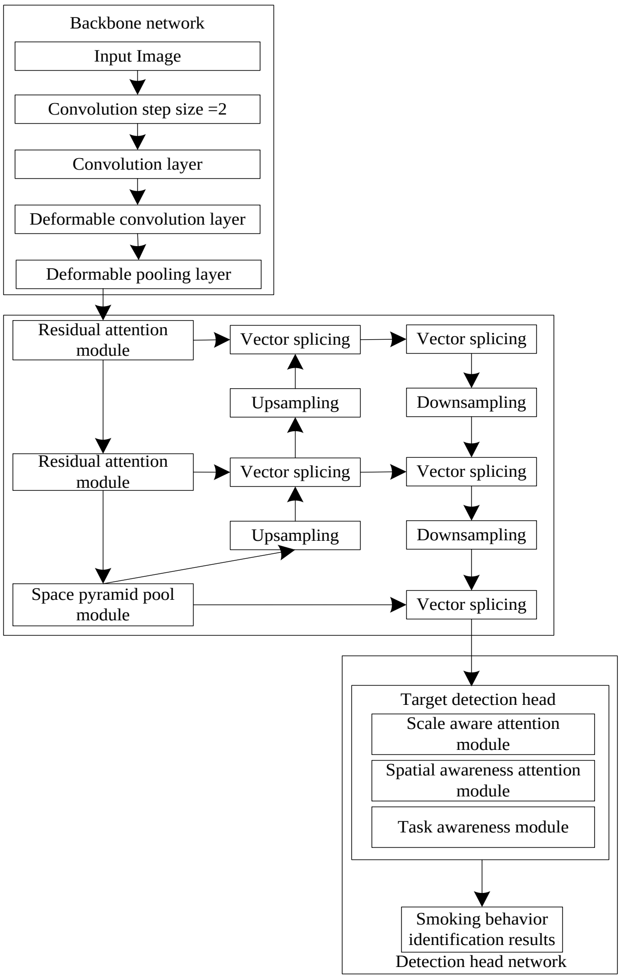


Fig. 1. Overall framework of smoking behavior identification

The specific steps for using improved YOLOV8 to identify smoking behavior are as follows:

Step 1: Input smoking behavior images within the backbone network.

Step 2: Process the input smoking behavior image through the convolutional layer to produce a feature map.

Step 3: Create a deformable convolutional layer, adjust the size and shape of the convolutional kernel according to the need, and adaptively extract the feature maps related to smoking behavior.

Step 4: Use a deformable pooling layer further to optimize the spatial distribution of the feature map.

Step 5: The feature maps processed as described above are passed to the neck network, which is responsible for further processing and integrating the feature maps.

Step 6: In the neck network, the residual attention module combines the channel and spatial attention mechanisms to process the smoking behavior feature maps delivered by the backbone network. This helps the neck network focus on the critical areas in the feature maps, such as the parts where the smoking action occurs, to extract high-quality smoking behavior feature maps.

Step 7: The spatial pyramid pooling module captures the global context information of high-quality smoking behavior feature maps at different scales and converts them into fixed-size outputs.

Step 8: Use the up-sampling module to increase the feature map's fixed size and enhance its resolution.

Step 9: Fuse the enhanced feature maps using the vector stitching module.

Step 10: Use the downsampling module to reduce the size of the fusion feature map, remove redundant information, improve computational efficiency, and help the neck network focus on the higher level of smoking behavior characteristics.

Step 11: The detection head network uses the scale-aware attention module, the spatial-aware attention module, and the task-aware module, respectively, to process the reduced fusion feature map, determine the recognition frame of the smoking behavior as well as the score of the recognition frame, and complete the recognition of the smoking behavior.

Step 12: Optimize the smoking behavior recognition frame by improving the loss function to improve the accuracy of smoking behavior recognition.

3. Smoking Behavior Recognition Methods

3.1. Improve smoking behavior feature extraction of backbone network in YOLOV8

In the overall framework shown in Figure 1, the improved YOLOV8's improved backbone network uses the deformable convolutional layer to extract the feature map related to smoking behavior adaptively.

Make X as a map of smoking behavioral features extracted from the convolutional layer, G_0 Is a position within, the standard convolution operation is:

$$X'(G_0) = \sum_{G_n} w(G_n) \cdot X(G_0 + G_n) \tag{1}$$

Among them, w is the convolution kernel; G_n is the convolution kernel position.

Obtain the deformable convolution operation based on formula (1) as:

$$X'(G_0) = \sum_{P_n} w(G_n) \cdot X(G_0 + G_n + \Delta G_n) \tag{2}$$

Among them, ΔG_n is the offset; X' is a map of features related to smoking behavior extracted by the deformable convolutional layer.

The formula for ΔG_n is as follows:

$$\Delta G_n = \sum_q \psi(q_x, p_x) \cdot \psi^2(q_y, p_y) \tag{3}$$

Among them, p_x and p_y is a random position on the horizontal and vertical coordinates; q_x and q_y is the integral space position; the comparison function is that $\psi(\cdot)$.

The feature map related to smoking behavior is extracted based on formula (2), which can improve the feature extraction effect [14].

X' is processed with the deformable pooling layer to extract key feature maps of smoking behavior. X' is divided into $k \times k$ histograms, and $k \times k$ key features of smoking behavior were simultaneously output. Order $b(i, j)$ as the (i, j)-th histogram, $0 \leq i, j \leq k$, then the corresponding key feature maps for smoking behavior are extracted as:

$$X''(i, j) = \sum_{a \in b(i,j)} \frac{X'(a_0 + a)}{n_{ij}} \tag{4}$$

Among them, a_0 are the coordinates of the upper left corner of the region of interest (the region associated with smoking behavior); a is the pixel point in $b(i, j)$; n_{ij} is the number of pixel points of the pixel $b(i, j)$.

Similar to the process of the deformable convolution operation, an offset Δa_{ij} is introduced into formula (4), a new pixel location is formed, and at this time, the critical feature map of smoking behavior is extracted as follows:

$$X''(i, j) = \sum_{a \in b(i,j)} \frac{X'(a_0 + a + \Delta a_{ij})}{n_{ij}} \tag{5}$$

After the deformable pooling layer processing, the feature extraction effect can be further improved [15], improving the recognition accuracy of smoking behavior.

3.2. Improve the treatment of smoking behavior characteristics of the YOLOV8 internal neck network

In the overall framework shown in Figure 1, YOLOV8 is improved. By improving the key feature map of smoking behavior output from the neck network processing backbone network, a higher level of smoking behavior feature map is obtained [16], and the accuracy of smoking behavior recognition is improved.

The residual attention module is introduced into the neck network to improve the neck network and improve the accuracy of smoking behavior recognition. In the residual attention module, there are two attention mechanisms, namely spatial and channel attention mechanisms, which enable the neck network in the improved YOLOV8 to focus on essential regions in the map X'' of the critical characteristics of smoking behavior to improve the accuracy of smoking behavior recognition. The neck network in the improved YOLOV8 will focus on the characteristic channel most related to smoking behavior [17], further improving the accuracy of smoking behavior recognition. By combining these two attention mechanisms, we can more accurately extract and use the feature information related to smoking behavior [18] and improve the accuracy and efficiency of smoking behavior recognition.

The spatial attention mechanism belongs to the process of encoding two one-dimensional maps of critical features of smoking behavior, sequentially following two spatial directions, aggregating X'', which not only can extract the spatial relationship between cigarettes and smokers' hands and mouths [19] but also can retain precise information about the location of critical elements such as cigarettes, smokers' hands and mouths, to more accurately identify smoking behavior [20]. In the horizontal and vertical directions, through the size of $(H, 1)$ and $(1, W)$ of the pooling kernel, coded to handle each channel, order $Q_c^h(h)$ and $Q_c^{\varpi}(\varpi)$ as the encoding output of the c -th channel in position h and ϖ and is calculated as follows:

$$Q_c^h(h) = \frac{\sum_{\hat{\imath}=1}^{W} X_c''(h,\hat{\imath})}{W} \tag{6}$$

$$Q_c^{\varpi}(\varpi) = \frac{\sum_{\hat{\jmath}=1}^{H} X_c''(\hat{\jmath},\varpi)}{H} \tag{7}$$

Among them, $\hat{\imath}$ and φ respectively are the channel index at the position of h and ϖ.

After completing the decomposition, an intermediate feature tensor φ that encodes spatial information can be obtained, with the following formula:

$$f = \varphi[Z(Q^h, Q^{\varpi})] \tag{8}$$

Among them, Z is a two-dimensional convolution; φ is a nonlinear activation function; $[\cdot,\cdot]$ is a splicing operation.

Splitting φ by convolutional transformation, can obtain attention weights ω^h and ω^{ϖ},. The map of essential features of smoking behavior output by the spatial attention mechanism is as follows:

$$\hat{X}_1 = X_c''(\hat{\imath},\hat{\jmath}) \times \omega_c^h(\hat{\imath}) \times \omega_c^{\varpi}(\hat{\jmath}) \tag{9}$$

In the channel attention mechanism, the global pooling layer is utilized for compression processing X'' to get $Y^{c\times1\times1\times1}$, then the spatial characteristic map of smoking behavior for each channel of compression is as:

$$Y_l = \frac{\sum_{\hat{\jmath}=1}^{H}\sum_{\hat{\imath}=1}^{W} X_l''(\hat{\imath},\hat{\jmath})}{H\times W} \tag{10}$$

Where, $X_l''(\hat{\imath},\hat{\jmath})$ is the l -th element within X''.

Using two full connection layers and the ReLU activation function ϕ, processing $Y^{c\times1\times1\times1}$, we get:

$$Y^* = \varphi\left(\omega_1\left(\phi\left(\omega_2(Y)\right)\right)\right) \tag{11}$$

Among them, ω_1 and ω_2 are the weights of the 2 fully connected layers.

Treat in a multiplicative manner for Y^* and X'', the essential features of smoking behavior that can be accessed by the channel attention mechanism are mapped as follows:

$$\hat{X}_2 = Y^* X'' \tag{12}$$

The map of essential features of smoking behavior extracted by the residual attention module is as follows:

$$\hat{X} = X'' + \hat{X}_1 \otimes \left(X'' + \hat{X}_2\right) \tag{13}$$

3.3. Improve smoking behavior recognition of detection head network in YOLOV8

In the overall framework shown in Figure 1, YOLOV8 is improved by introducing a scale-aware attention module, a spatial-aware attention module, a task-aware module, and an improved non-maximum inhibition algorithm to enhance the detection head network, accurately determine the smoking behavior recognition box and its score, complete the smoking behavior recognition, and improve the accuracy of smoking behavior recognition.

The map $\hat{X}$ of essential features of smoking behavior at different scales through the scale-aware attention module, fusing different scales with the semantic importance of the $\hat{X}$, which helps to improve the detection head network in YOLOV8 to better adapt to varying scales of input smoking behavior important feature maps, capture critical details, and improve the accuracy of smoking behavior recognition. This module integrates different scales $\hat{X}$ with the formula as follows:

$$\tilde{X}_L(\hat{X}) = \delta(\rho(\frac{\hat{X}}{S})) \cdot \hat{X} \tag{14}$$

Among them, ρ is a linear function; $\rho(x) = max\left(0, min\left(1, \frac{x+1}{2}\right)\right)$; x represents $\rho\left(\frac{\sum_S \hat{X}}{S}\right)$; S is the scaling factor, which is used to adjust the map $\hat{X}$ of important characteristics of smoking behavior importance at different scales. The larger the scale S, the more significant the scale's weight.

Aggregating maps $\hat{X}$ of essential features of smoking behavior at the exact spatial location using spatial perceptual attention module, focused attention spatial layout ,and contextual relationships of $\hat{X}$, capturing the interdependence between regions in $\hat{X}$ which helps improve the detection head network in YOLOV8 to understand and locate regions of interest more accurately The formula for the module consolidating the same space position $\hat{X}$ is as follows:

$$\tilde{X}_C(\hat{X}) = \frac{\sum_{m=1}^{M}\sum_{u=1}^{U} \omega_{m,u} \cdot \hat{X}(m; p_u + \Delta p_u; c) \cdot \Delta\beta_u}{M} \tag{15}$$

Among them, $p_u + \Delta p_u$ is the location of the movement in the region of interest where the self-learning spatial offsets are centered M is the number of layers in the spatially aware attention module; U is the number of sparse sampling locations; $\omega_{m,u}$ is the weight of the u -th sampling bit m set of the first layer. c is the channel index. $\Delta\beta_u$ is the self-learning fundamental scalar in position p_u.

The task-aware attention module is responsible for dynamically adjusting the attention distribution according to the current task goal, making the model more focused on the critical feature map of smoking behavior related to the task, thus improving the generalization ability of the detector network in YOLOV8. The formula for this module handles $\hat{X}$ is as follows:

$$\tilde{X}_\lambda(\hat{X}) = max\left(\gamma^1(\hat{X}) \cdot \hat{X} + \varepsilon^1(\hat{X}), \gamma^2(\hat{X}) \cdot \hat{X} + \varepsilon^2(\hat{X})\right)$$
(16)

Among them, γ^1, γ^2, ε^1 and ε^2 are hyperfunction that controls the activation threshold.

$\tilde{X}_L(\hat{X})$, $\tilde{X}_C(\hat{X})$ and $\tilde{X}_\lambda(\hat{X})$ are processed through 2 fully connected layers with a normalization layer, generate a smoking behavior recognition box, output the score of the smoking behavior recognition box through the Sigmoid function, and complete the smoking behavior recognition. The score of each smoking behavior identification box reflects the confidence in smoking behavior. The higher the score, the more likely the box contains smoking behavior.

3.4. Improve the optimization of YOLOV8 smoking behavior recognition

By improving the loss function, optimizing and enhancing the smoking behavior recognition box output by YOLOV8, and improving the accuracy of smoking behavior recognition. Order $Z = [x, y, w, h]$ as the smoking behavior identification box output from the previous subsection; x is the horizontal coordinate of the center of the smoking behavior identification box. y is the vertical coordinate of the center of the smoking behavior identification frame. w is the width of the smoking behavior recognition box. h is the height of the smoking behavior identification box.

YOLOV8 takes CIoU Loss as the loss function, and CIoU Loss mainly focuses on the IoU (cross merge ratio) between the recognition frame and the actual frame. Still, the IoU cannot fully capture the relative position and size differences between the targets, leading to the model's inability to identify smoking behavior accurately. WIoU Loss considers targets' position and size information by introducing weight coefficients to more accurately measure the performance of smoking behavior recognition and enable the model to more comprehensively assess the relative position and size differences between targets during training. Therefore, WIoU Loss is taken as the loss function, and the calculation formula is as follows:

$$L_{WIoU} = \frac{\xi e^{\frac{(x-x')^2+(y-y')^2}{w'^2+h'^2}} L_{IoU}}{\tau r^{\xi-r}}$$
(17)

Among them, $Z' = [x', y', w', h']$ is the real box. ξ is the extent of the anomaly for Z. r is the equilibrium parameter. τ is the learning parameter; L_{IoU} is the normal mass anchor frame loss function.

To further improve the optimization effect of the smoking behavior recognition frame, the normalized Wasserstein distance position regression loss function is introduced into the loss function, and the formula is as follows:

$$L_D = exp\left(-\frac{\sqrt{\left\|\left(x,y,\frac{w}{2},\frac{h}{2}\right),\left(x',y',\frac{w'}{2},\frac{h'}{2}\right)\right\|_2^2}}{2}\right) \qquad (18)$$

The final loss function optimized for the smoking behavior recognition box is:

$$L = \kappa \cdot L_{WIoU} + (1-\kappa) \cdot L_D$$
(19)

Among them, κ is the balancing factor.

4. Experimental analysis

4.1. Experimental setup

The experimental shooting environment is an indoor room with an area of about 120m2. Digital cameras are installed in each room corner to collect images of real-time smoking behavior. The distance between each digital camera and the ground is 3m. Different backgrounds and lighting conditions are set artificially during the shooting to analyze different environments and lighting conditions. The recognition effect of smoking behavior of this method. Table 1 shows the relevant introduction of the collected smoking behavior image data set.

In addition to data set preparation, the parameters of the proposed improved YOLOV8 are set as shown in Table 2, and the training process is shown in Figure 2.

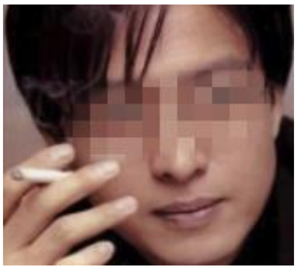
Table 1 Parameters of smoking behavior image dataset

Environmental condition	Light condition	Number of pictures/piece	Resolution
Indoor daytime	Natural light	300	1920×1080
Indoor night	Indoor lighting	200	1280×720
Indoor overcast	Less natural light	150	1600×900
Indoor dusk	Oblique sunlight	100	1024×768
Interior closed doors and Windows	Low room light	50	800×600

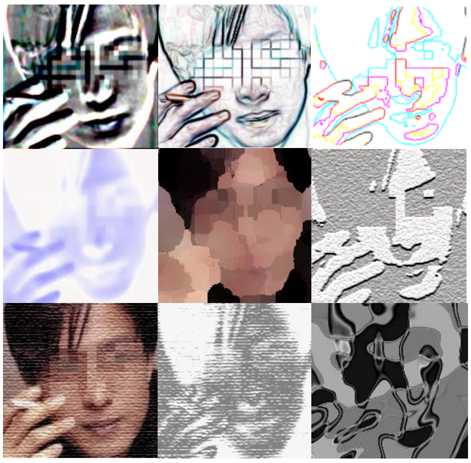
Table 2 Improved YOLOV8 algorithm parameters

The name of the parameter	Values
Learning rates	0.001
Batch size	32
Number of training rounds/rounds	100
Input image size/pixel	416×416
Momentum	0.9
Weight decay	0.0005
Deformable convolution kernel size/pixel	3×3
Number of residual attention modules	4
Scale-aware attention to the number of module layers that	3
Spatially aware attention to the number of module layers	2
Number of layers of the task-aware module	1
The non-maximal inhibition threshold, the	0.5
The loss function equilibrium factor	0.75

An image is randomly selected from the smoking behavior image dataset, and the key features of this image's smoking behavior are extracted using the method of this paper. The results of the feature extraction are shown in Fig. 3.



(a) The original image



(b) Feature extraction results

Fig 3 Key characteristics of smoking behavior

Analysis of Figure 3 (a) and Figure 3 (b) shows that the method in this paper successfully extracts critical features related to smoking behavior in the original image. The feature extraction process preserves the visual information closely related to smoking actions, such as gestures. Moreover, the features extracted by the method in this paper have a clear visual representation, providing a more abstract but recognizable contour and shape. The features obtained by this method can provide reliable data support for subsequent smoking behavior recognition. The reason is that these features not only contain direct evidence of smoking behavior but also consider indirect clues such as hand posture and cigarette position, which is conducive to improving the accuracy and robustness of recognition.

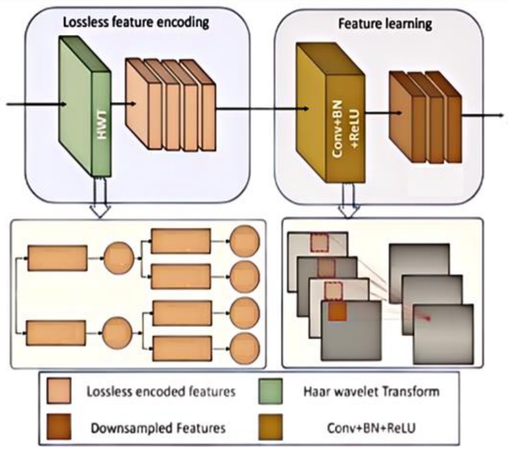
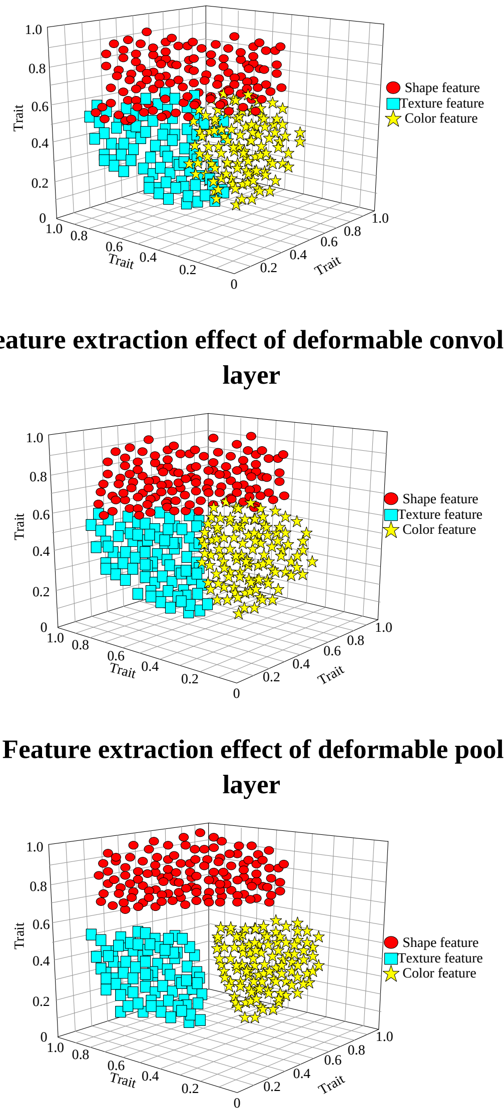


Fig. 2. Improve the YOLOV8 training process

4.2. Analysis of results

In visualization, we analyze the smoking behavior feature extraction effect of this paper's method, using shape features, texture features, and color features as examples. The analysis results are shown in Fig. 4.



(a) Feature extraction effect of deformable convolution layer

(b) Feature extraction effect of deformable pooling layer

(c) Feature extraction effect of residual attention module

Fig 4 Visual analysis results of the feature extraction effect of the proposed method

From the analysis of Figure 4 (a), it can be seen that the deformable convolution layer can effectively extract the characteristics of smoking behavior. Still, the boundary between shape features, texture features, and colour features is not obvious, and there is serious confusion. This is because of the complexity of smoking behavior and the limitations of the deformable convolution layer in processing these features. Smoking behavior includes various shapes, textures, and colour changes; however, deformable colluvium cannot effectively deal with these other characteristics at the same time. Analysis of Figure 4 (b) shows that after the deformable pooling layer processing, the boundary between shape features, texture features and colour features is not obvious. Still, the confusion is alleviated compared with Figure 4 (a). The overlapping area is significantly reduced, which indicates that the feature extraction effect is better than that of the deformable convolution layer at this time, which suggests that the deformable pooling layer can improve the feature extraction effect to a certain extent, Especially when dealing with smoking behavior with significant shape changes. Analysis of Figure 4 (c) shows that after the residual attention module processing, the boundary between shape features, texture features and colour features is obvious, without any confusion, indicating that the residual attention module can effectively extract the characteristics of smoking behavior and clearly distinguish different feature categories, which is mainly due to the role of attention mechanism, It can make the model pay more attention to the key features related to smoking behavior, and ignore those features unrelated to the task.

Figure 5 shows the recognition results of this paper's method for recognising the smoking behavior of this image.

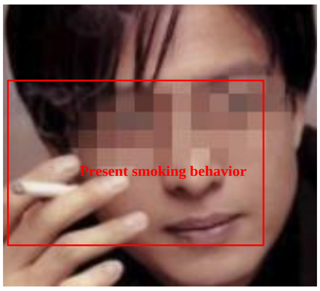


Fig 5 Results of smoking behavior identification

Figure 5 shows that the method of this paper can effectively determine the identification frame of smoking behavior and complete the identification of smoking behavior.

Using the number of detected frames per second (FPS) as the evaluation index of this method, the efficiency of this method's recognition of smoking behavior is analysed through FPS. The larger the value, the higher the recognition efficiency of this method. Based on YOLOV8, the application effect of this method is investigated through an ablation experiment. The analysis results are shown in Table 3.

Table 3 Ablation experiments of the method presented in this paper

Methods and indicators	Improve or not									
YOLOV8	×	×	×	×	×	×	×	×	×	×

Improved backbone network	Deformable convolution layer	√	√	√	√	√	√	√	√	√	√
	Deformable pooling layer	×	√	√	√	√	√	√	√	√	√
Improved neck network	Channel attention mechanism	×	×	√	√	√	√	√	√	√	√
	Spatial attention mechanism	×	×	×	√	√	√	√	√	√	√
	Residual attention module	×	×	×	×	√	√	√	√	√	√
Improve the detection head network	Scale-aware attention module	×	×	×	×	×	√	√	√	√	√
	Spatial awareness attention module	×	×	×	×	×	×	√	√	√	√
	Task awareness module	×	×	×	×	×	×	×	√	√	√
Improved loss function	WIoU Loss	×	×	×	×	×	×	×	×	√	√
	Normalised Wasserstein distance position regression loss function	×	×	×	×	×	×	×	×	×	√
FPS		101	108	111	114	116	125	133	136	140	154

Analysis of Table 3 shows that before improvement, YOLOV8 had the lowest FPS value when it conducted smoking behavior recognition, indicating that YOLOV8 had the slowest efficiency in smoking behavior recognition. After the deformable convolution layer and deformable pooling layer are added to the backbone network, the FPS value increases significantly, indicating that the efficiency of smoking behavior recognition can be accelerated at this time. In the neck network, after adding the channel and spatial attention mechanism, as well as the residual attention module, the FPS value is further improved, indicating that the efficiency of smoking behavior recognition can be further accelerated at this time. In the enhanced detection head network, after adding an aware attention module, spatial-aware attention module and task-aware module, the FPS value has also been improved, further accelerating the efficiency of smoking behavior recognition. After enhancing the loss function, the FPS value of the method in this paper reaches the maximum, indicating that the efficiency of smoking behavior recognition is the fastest. Comprehensive analysis shows that applying improved YOLOV8 in this method can significantly improve the efficiency of smoking behavior recognition.

The AUC area measures the recognition effect of this method's smoking behavior under different lighting conditions. The closer the AUC area is to 1, the higher the recognition accuracy of this method's smoking behavior. The analysis results are shown in Figure 6.

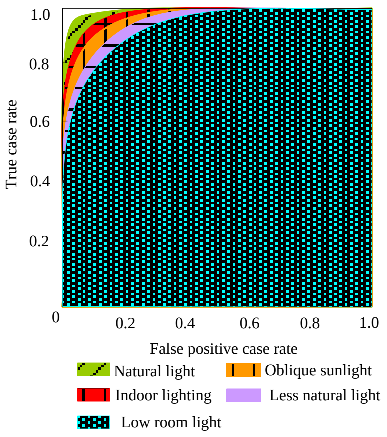


Fig 6 Smoking behavior recognition effect under different lighting conditions

It can be seen from the analysis of Figure 6 that under different lighting conditions, the method in this paper can effectively complete the recognition of smoking behavior. Among them, under natural light conditions, the AUC value of smoking behavior recognition is the largest, which is the closest to 1, indicating that the recognition accuracy of smoking behavior is the highest at this time. Under weak indoor light conditions, the AUC value is the smallest, which is also close to 1, indicating that the recognition accuracy of smoking behavior is also high at this time. Experiments show that under different lighting conditions, the AUC value of the method in this paper is high; that is, the recognition accuracy is high.

5. Conclusion

With artificial intelligence technology's continuous development and application, target detection and recognition technology based on deep learning has shown great potential and value in many fields. Aiming at the specific task of smoking behavior recognition, this paper proposes a method based on improved YOLOV8. By enhancing the backbone network, neck network and detection head network, we can more accurately capture the critical features of smoking behavior, improve the accuracy of smoking behavior recognition, and accelerate the efficiency of smoking behavior recognition. The research of this paper can not only provide a new technical means for smoking behavior recognition but also bring possible solutions for practical application scenarios such as intelligent monitoring and public health management. In addition, the work of this paper also provides some valuable references and enlightenment for researchers in related fields, especially in model optimisation, feature extraction, and real-time monitoring.

I want to express my deepest gratitude to my supervisor for his invaluable guidance, unwavering support, and constructive feedback throughout the research process. His expertise and encouragement have been instrumental in shaping the direction of this study.

Author contributions

SecondAuthor: Research Direction Guidance, Methodological Support, Experimental Supervision and Guidance, Thesis Review and Revision, Academic Resources and Support.

FirstAuthor: Literature Review and Data Collection, Constructing a dataset, Experimental Design and Implementation, Thesis Writing and Revision.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Kazutaka, O. , Kentaro, T. , Ayumi, K. , Shunsuke, S. & Toshiki, S. (2022). Common brain cortical abnormality in smoking behavior and bipolar disorder: discriminant analysis using cortical thickness and surface area. Cerebral Cortex, 32(20),4386-4396.

[2] Tahseen, S. & Danti, A. (2022). Relevance of psychophysiological and emotional features for the analysis of human behavior survey. ECS transactions, 107(1),607-626.

[3] Jed E., R. , Suzanne, F. , David, C. , Alfred, S. , Susan, C. & James M., D. (2023). Smoking reduction using electronic nicotine delivery systems in combination with nicotine skin patches. Psychopharmacology, 240(9),1901-1909.

[4] Lin, S. C. , Gathua, N. , Thompson, C. , Sripipatana, A. & Makaroff, L. (2022). Disparities in smoking prevalence and associations with mental health and substance use disorders in underserved communities across the United States. Cancer,128(9),1826-1831.

[5] Hoon, J. S. , Jin, J. H. , Junhyun, K. & Eun-Cheol, P. (2021). Association between smoking behavior and insulin resistance using triglyceride-glucose index among South Korean adults. The Journal of Clinical Endocrinology & Metabolism,106(11),e4531-e4541.

[6] Tran, D. N. , Nguyen, T. N. , Phi-Khanh, P. C. & Tran, D. T. (2021). An IoT-based design using accelerometers in animal behavior recognition systems. IEEE Sensors Journal, 22(18), 17515-17528.

[7] Mahalakshmi, V. , Sandhu, M. , Shabaz, M. , Keshta, I. , Prasad, K. D. V. & Kuzieva, N. , et al. (2024). Few-shot learning-based human behavior recognition model. Computers in human behavior, 151(Feb.),1.1-1.12.

[8] Chen, X. & Dinavahi, V. (2021). Group behavior pattern recognition algorithm based on spatio-temporal graph convolutional networks. Scientific Programming,2021(Pt.4),2934943.1-2934943.8.

[9] Farooq, M. U. , Saad, M. N. M. & Khan, S. D. (2022). Motion-shape-based deep learning approach for divergence behavior detection in the high-density crowd. The visual computer, 38(5),1553-1577.

[10] De Ocampo, A. L. P., & Dadios, E. (2019). Radial greed algorithm with rectified chromaticity for anchorless region proposal applied in aerial surveillance. International Journal of Advances in Intelligent Informatics, 5(3), 193-205.

[11] Pham, T. N. , Nguyen, V. H. & Huh, J. H. (2023). Integration of improved yolov5 for face mask detector and auto-labeling to generate dataset for fighting against COVID-19. The Journal of Supercomputing,79(8),8966-8992.

[12] Manssor, S. A. F. , Sun, S. , Abdalmajed, M. & Ali, S. (2022). Real-time human detection in thermal infrared imaging at night using the enhanced tiny-yolov3 network. Journal of Real-Time Image Processing, 19(2), 261-274.

[13] Tan, G. X., Cen, M. W. & Su, R. J. (2023). Design of Vehicle and Pedestrian Detection Network Based on YOLOv4.Computer Simulation,40(4):128-133.

[14] Joris Guérin, Stéphane Thiery, Nyiri, E. , Gibaru, O. & Boots, B. (2021). Combining pre-trained cnn feature extractors to enhance the clustering of complex natural images. Neurocomputing, 423(Jan.29), 551-571.

[15] A. L. P. De Ocampo, "Normalized Difference Vegetation Index (NDVI) Estimation based on Filter Augmented Imaging," 2023 International Electrical

Engineering Congress (iEECON), Krabi, Thailand,2023,pp.84-88, doi: 10.1109 IEECON56657.2023.10126616.

[16] Sreelakshmi, D. & Inthiyaz, S. (2021). Fast and denoise feature extraction based admf–cnn with a global framework for MRI brain image. International Journal of Speech Technology, 24(3), 1-16.

[17] Sabry, E. S., Elagooz, S., El-Samie, F. E. A., El-Bahnasawy, N. A. & El-Banby, G. (2024). Assessment of various feature extraction methods for object discrimination in different scenarios. Journal of Optics, 53(1),49-69.

[18] Schleider, L. , Pasiliao, E. L. , Qiang, Z. & Zheng, Q. P. (2022). A study of feature representation via neural network feature extraction and weighted distance for clustering. Journal of Combinatorial Optimization, 44(4),3083-3105.

[19] Lahoti, G. , Ranjan, C. , Chen, J. , Yan, H. & Zhang, C. (2023). Convolutional neural network-assisted adaptive sampling for sparse feature detection in image and video data. IEEE Intelligent Systems,38(1),45-57.

[20] Damaneh, M. M., Mohanna, F. & Jafari, P. (2022). Static hand gesture recognition in sign language based on a convolutional neural network with feature extraction method using orb descriptor and Gabor filter. Expert Syst. Appl, 211(Jan.), 118559.1-118559.13.