Smoking Detection Model Based on Improved YOLOv8-s

Yujun Zhu[1] Canyang Zhou[1] Bi Zeng[1*]

[1](GuangDong University of Technology, *GuangZhou 510006, GuangDong ,China*)

Abstract Target detection algorithms face challenges in smoking detection tasks, particularly in the identification of small targets and the occurrence of misidentification in various scenes. In this paper, we propose Smoking-YOLO based on YOLOv8-s, which adopt ConvNeXtv2 as the backbone that can extract features at four scales, obtaining stronger contextual information to enhance small target detection capability. During the feature fusion stage, we employ a bidirectional three-channel four-scale fusion strategy in the fusion stage to output four-scale prediction maps, strengthening the semantic information focus on smoking details and improving the ability to distinguish pseudo-smoking behaviors. Finally, we adds a slide weighting function to enhance attention to hard negative samples. Experimental results on the self-built Smoking-3k dataset show that our model achieves a detection effect of $AP_{small}(0.5{:}\,0.95)$ 0.31 for small targets, an improvement of 10.6%. The model's precision and recall reach $mAP_{0.5}$ 0.947 and $mAP_{0.5:0.95}$ 0.652, respectively, increasing by 3.1% and 7%, demonstrating the effectiveness of the model improvement. The code is available at https://github.com/TaroPlay/Smoking-YOLO.git

Keywords Smoking Detection Small Target Detection Bidirectional Three-channel Four-scale Fusion Strategy

1 Introduction

With the development of society and the improvement of people's health awareness, the control of smoking behavior in public places has become more and more important. According to the World Health Organization [1], smoking is one of the main causes of various diseases and health problems , such as cardiovascular disease, lung cancer and so on. However, traditional methods of monitoring smoking behavior often require manual intervention, which is costly and inefficient. In this context, smoking behavior detection technology based on computer vision provides a new way to solve this problem.

The smoking detection technology based on wearable smart devices [2] mainly serves people who have the need to quit smoking, and it is difficult to achieve such detection equipment in public places with intensive personnel flow. Zhang [3] proposed a smoking detection model for public places based on YOLOv5, but only realized the identification of cigarette types. Lakatos [4] proposed a multimodal smoking detection method that uses a large language model to fine-tune and train video streams, with high accuracy, but the model is too complex and the dataset quality requirements are high.

In order to solve the problems in the above work, we proposed smoking-YOLO, a multitype smoking detection method. We use YOLOv8-s as the base network. In order to better learn features of small target objects, ConvNeXt-v2, a pure convolutional network with multi-scale features output, is used as a new backbone network. Feature maps with smaller receptive fields can focus more on small targets in images. In the process of feature fusion, it is necessary to ensure that shallow features and deep features are mutually balanced in spatial and semantic terms, and the number of parameters in the model should not be too large. Therefore, we proposed the bidirectional three-channel four-scale fusion method, and carried out lightweight implementation in the fusion module. Finally, because the smoking detection dataset is a small dataset, in order to make full use of difficult samples in model training, we use slide weighting function to increase the weight of difficult samples in the loss function, increasing the model's attention to difficult samples. The code and pre-trained models are released at https://github.com/TaroPlay/Smoking-YOLO.git.

2 Related Work

2.1 Smoke detectors using sensors.

Traditional smoking behavior monitoring methods mainly include video surveillance and sensor technology [5]. The method based on video surveillance usually relies on manual observation and judgment, which has some problems such as limited monitoring range, high cost and low efficiency. Although the sensor technology can realize automatic monitoring, its application scope is limited, and the environmental requirements are high, and it is not suitable for all scenarios.

2.2 Smoking detection using deep learning.

With the development of deep learning technology, the smoking detection method based on neural network has gradually become the mainstream [6]. Deep learning can learn the characteristics of smoking behavior from a large number of data, and has good generalization ability and robustness. Researchers have proposed many deep learning based smoking behavior detection methods, such as convolutional neural network based method, recurrent neural network based method and so on. These methods have achieved good results in smoking behavior detection and provide new ideas and methods for smoking behavior monitoring and control. The current smoking behavior detection methods are mainly to identify the smoking action of ordinary cigarette, but there are still some challenges and problems, such as few detection of rare smoke type, limited adaptability to different smoking posture and environmental light, and more misjudgment behaviors of pseudo-smoking action. Therefore, further research and improvement of smoking behavior detection methods are still of great significance.

2.3 YOLOv8

Ultralytics YOLOv8 [24] is a cutting-edge, state-of-the-art (SOTA) model that builds upon the success of previous YOLO versions and introduces new features and improvements to further boost performance and flexibility. YOLOv8 is designed to be fast, accurate, and easy to use, making it an excellent choice for a wide range of object detection and tracking, instance segmentation, image classification and pose estimation tasks.

2.4 Contribution of smoking detection work in this paper.

Smoking detection is the next important sub-task in the field of computer vision, but its standardization work is less, and the smoking task is only transformed into the human behavior recognition task based on simple cigarette labeling. However, it did not take into account the actual factors in the picture, such as the incomplete human body, the diversity of cigarettes, the wide range of smoking scenes, and the false judgment caused by smoke and white gas. Therefore, according to some shortcomings of previous work, this paper has made improvements. The main contributions of this paper are as follows:

i). For the first time, a high-quality detection-based smoking recognition classification dataset was produced. The images were collected from network pictures and personal photos. Smoking behaviors in various scenes were included as positive samples, and fake smoking behaviors were also marked to increase anti-interference ability during detection.

ii). ConvNeXt-v2 [7] was used as the backbone network to strengthen the ability to capture small targets, and four-scale feature map was used as the main body of the network feature fusion process, emphasizing the exchange of high-level semantic information and low-level semantic information. Finally, slide weighted function was used to make full use of difficult samples and improve the efficiency of the use of Smoking-3k.

iii). A pixel-level detection model is proposed, which has the ability to recognize various types of smoke with fast reasoning speed and can recognize all kinds of smoking actions effectively.

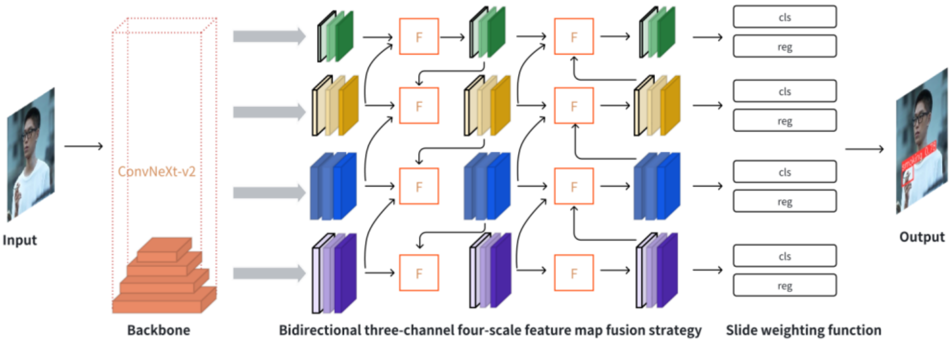
Input
ConvNeXt-v2
Backbone
F
cls
reg
Bidirectional three-channel four-scale feature map fusion strategy
Slide weighting function
Output

Figure 1. Overview of smoking-YOLO. Four scale features are extracted from the backbone network, and a bidirectional three-channel four-scale feature fusion strategy is used in the feature fusion stage. And the Slide weighting function is additionally referenced during the training of the prediction head.

3 Methods

3.1 Model Overview.

At present, the basic framework of detection tasks is as follows: the backbone network extracts multi-scale feature information, the neck network fuses the feature information to obtain higher dimensional semantic information, and finally uses the header network to reduce the dimension of the feature map and output the prediction box information to meet the needs of different downstream tasks. Smoking-YOLO also follows the above framework. First, the image is taken as the network input with the pixel size of 640×640. The pure convolutional network ConvNeXt-v2 is used to extract multi-scale features and output four kinds of scale feature maps. Then, the neck network uses the bidirectional feature pyramid path to fuse the information of the four scales, and the fusion feature map of the four scales is also obtained. Finally, the four types of prediction graphs input the same structure but do not share the weight of the head network to predict the presence of smoking behavior in the picture. An overview of the model is shown in Fig. 1.

3.2 Multi-scale Feature Extraction

Inspired by ConvNeXt-v2, this paper also uses a full convolution-based network as the backbone, and because the number of categories for smoking detection is small, we use lightweight version named atto. Its internal construction of full convolutional mask autoencoder and global response normalization is very suitable for smoking detection task. The experimental results show that ConvNeXt-v2 is better than YOLOv8-s in detecting small objects.

The full convolutional mask autoencoder framework [8] is a self-supervision method based on convolutional neural networks. The idea is to block some regions on the initial image according to probability, and then let the model restore the covered part. For smoking detection tasks with small data sets, generalization is easier to learn in terms of training methods. This mask operation can make the model learn the relationship between the global and local features [9] of the image, so as to improve its generalization ability. In the task of smoking detection, this generalization is very necessary. When there is a lighter in the picture, it often means that the smoker will light a cigarette, and the local area where the lighter appears will have a attention effect on the local area where the human mouth is located. For another example, when a person plays the flute, it may be mistaken for the smoking behavior of a long-pipe smoker at some angles. When the model has a strong local feature learning ability, it can learn that the long-pipe smoke will generate a lot of smoke when it is smoked, and when playing the flute, it should learn the hands placement action of the player. At the same time, compared with ordinary mask autoencoders [10], the advantage of using full convolutional mask autoencoder is that multi-scale mask strategy is used instead of fixed-size mask, which increases the perception ability of the model to different scale information [11], and enough semantic information can be obtained in the smoking detection pictures of the long view and the smoking detection pictures of the near view. In terms of computation and parameter number, it is also less than that of mask generation and image reconstruction using full connection layer.

Global Response Normalization (GRN) [12] is designed to solve the feature collapse problem. Compared with the general batch normalization, GRN has the advantage that no additional parameters are required, because it only normalizes the feature map, and its normalization mode can handle any batch size, while batch normalization requires dynamic adjustment of parameters according to the batch size, and the calculation is large. The implementation of the GRN layer is also very simple and is divided into 3 steps: global feature aggregation, feature normalization and feature calibration. First, the feature graphs on each channel are aggregated using L2-norm to get the aggregated vector. Then, in the feature normalization step, the standard normalization function is used to normalize the aggregated vector again. Finally, in the feature calibration step, the normalized vector is used to calibrate the original feature map. The following is a formulaic implementation process in three steps:

$$F(X) := X \in R^{H*W*C} \rightarrow fx \in R^{C} \qquad (1)$$

$$G(\|X_i\|) := \|X_i\| \in R \rightarrow \frac{\|X_i\|}{\sum_{j=1,\ldots,C}\|X_j\|} \in R \qquad (2)$$

$$X_i = X_i * G(F(X)_i) \in R^{H*W} \qquad (3)$$

$X_i$is the feature graph matrix of the i'th channel after global corresponding normalization. GRN changes the normalization mode from channel normalization to feature graph normalization, which enhances the feature competition among channels, helps channels learn information of different feature domains, and avoids the problem of feature collapse and feature redundancy. This is helpful for the model to learn more diverse feature content and enhance the model extraction performance during the smoking detection. After the backbone network, the feature maps of $P_2(\frac{1}{4}P_0), P_3(\frac{1}{8}P_0), P_4(\frac{1}{16}P_0), P_5(\frac{1}{32}P_0)$ are input into the feature fusion module, and P_0 is the initial feature map size.

3.3 Bidirectional three-channel four-scale feature map fusion strategy

The feature maps of four scales are extracted from the backbone network, and the different receptive fields mean that the feature maps have learned different levels of semantic information respectively. In this paper, the most concerned features are smoke features, cigarette shape features, hand movement features and mouth movement features. For low-level semantic information, such as smoke, tobacco products and hand movements to be recognized, which involve more specific and fine-grained visual information, we choose to use feature graphs P_2, P_3 with small receptive fields to learn. For high-level semantic information, these features involve more abstract and semantic concepts, which usually require a higher level of analysis to understand. High-level semantic information includes the overall action of smoking behavior and the context of the background environment. In this paper, feature figures P_4, P_5 with larger receptive fields are selected for learning. P_2, P_3, P_4, P_5 as the initial feature map of the four semantic information.

Feature Pyramid Networks [11] are essential for object detection as they aggregate features of varying resolutions extracted from the backbone. While traditional FPNs [13] use a top-down path to fuse multi-scale features, they can only pass feature information in one direction. To address this limitation, PAFPN [14] adds a bottom-up path aggregation network. However, this comes with increased calculation cost and parameter count. BiFPN [15] improves upon PAFPN by removing nodes with only one input edge and adding skip connections from the original input at the same level, enhancing efficiency. Despite these advancements, convolution-based cross-scale feature fusion in GFPN [16] remains inefficient, especially when fusing three-scale feature maps for real-time detection models. For smoking detection tasks, the interaction between high-level semantic information and low-level spatial information is crucial, demanding real-time performance in computation and parameter quantity. Inspired by DAMO-YOLO [17], this paper proposes a

Feature Graph Fusion Semantic Module with several key enhancements : 1). 4-Scale Feature Fusion: Improved from the original 3-scale feature fusion, this strategy employs two-way information transmission to prevent information loss. Different channel numbers are used for different scale features, allowing flexible control over the expression ability of high-level and low-level features within lightweight computation constraints. 2). Optimized Up-sampling Operation: The additional up-sampling operation in the 3-channel feature graph fusion process is optimized, significantly reducing model inference delay with minimal precision reduction. 3). Simply Rep 3×3 Module: A simplified version is used in the fusion process, doubling channels during training and optimizing to single-channel equivalent during inference, thereby improving reasoning speed, shown in Fig. 2. These advancements are particularly effective for detecting small and medium smoking targets in challenging and diverse smoking scenes, capturing various smoking actions and distances effectively.

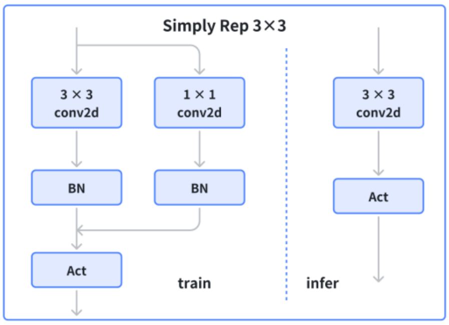

Figure 2. The Simply Rep 3×3 Module saves the inference time of the smoking-YOLO.

Multi-scale feature fusion aims to aggregate output features from different stages of the backbone network, enhancing their expressive power and improving overall model performance. In this paper, we utilize four scale feature maps for feature fusion, including the newly added feature map $P_2(\frac{1}{4}P_0)$, which has a smaller receptive field. This enables better distinction between objects in the image at the shallow feature stage, facilitating the learning of features distinguishing smoke from white gas. Unlike other FPNs that typically use only three high-level semantic feature maps (P_3, P_4, P_5) with larger receptive fields for feature fusion, our approach considers the importance of representational information, such as smoke features and shape features. This consideration is particularly relevant for smoking detection tasks. In Fig. 3, we visualize the attention distribution map [18] of the model after multi-scale feature fusion, demonstrating its strong ability to capture low-level semantic information in the image. After passing through the neck network, four kinds of scale prediction maps are still output, and the size remains unchanged.

3.4 Detection Head And Loss Function

In the head network section, this paper adopts the current mainstream decoupling head structure, separating the classification head and the detection head to improve convergence speed. Additionally, the Anchor-Free mode is employed, which does not rely on prior knowledge in the dataset, reducing a large number of invalid calculations. This network exhibits enhanced capability in expressing the "shape of the object" and demonstrates greater generalization potential. It improves detection of moving objects and objects of different sizes, and offers more flexible detection of blocked objects. The decoupling of the classification branch and regression branch is the mainstream method for detection tasks, which can reduce interference between the two tasks, ensuring that detection and classification can reach their optimal states. The four types of feature maps are trained with their own predictive head weights, ensuring the independence of the four types of key information in the prediction. However, the decoupling head structure network often faces the problem of misalignment between classification and regression. This means that the cells in the feature map and the Ground Truth perform IOU calculation to allocate the cells used for prediction, but the optimal cells for the classification task and regression task are often inconsistent. Frequent misidentification of smoking behavior in smoking detection can lead to loss of significance in the task. To solve the problem of misalignment, this paper uses the Task-Aligned Assigner positive sample allocation strategy to assign labels to the anchor frame of the ground truth feature map constructed by calculating the Loss. The classification branch uses the BCE Loss function, while the regression branch uses the Distribution Focal Loss function [19] and the Complete IoU Loss function [20] of the integral representation. The three Loss functions are weighted by a weight ratio to obtain the joint loss function. The formula is expressed as follows:

$$L_{BCE} = \frac{1}{N}\sum_{i} -[y_i - \log(p_i)\,(1 - y_i) \cdot \log(1 - p_i)] \tag{4}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{5}$$

$$L_{DFL}(S_i, S_{i+1}) = -\big((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})\big) \tag{6}$$

y_i represents the ground truth and p_i represents the predicted value. b, b^{gt} represent the center point of the rectangular box, ρ represents the Euclidean distance between the two rectangular boxes, c is a constant, α is the weight coefficient, and ν is a measure of the consistency of the relative proportions of the two rectangular boxes. DFL is a supplementary loss function to match the anchor-free mechanism and enhance the generalization when blocking and moving objects. DFL optimizes the probability of the two positions closest to the Ground Truth in the form of cross-entropy, so that the network can focus on the target position and the distribution of adjacent regions more quickly. That is, the final learned distribution is theoretically near the real floating point coordinates, and the weights are obtained according to the linear interpolation method.

In the training process, we observed a problem of sample imbalance in the smoking detection dataset, where difficult samples were relatively sparse, while the number of easy samples was large. To address this issue, this paper introduces the slide weighting function [slide]. The method for distinguishing easily separable samples from difficult samples is to predict the IoU size of the box and the Ground Truth box. To avoid setting a hyperparameter threshold, the average value of all IoUs is used as the threshold θ. Samples with an IoU less than θ are considered negative, while those with an IoU greater than θ are considered positive. Negative samples located near θ, called difficult negative samples, are given higher weights to ensure the network is effectively trained using these samples. The slide weighting function can be expressed as follows:

$$\mathrm{f(x)} = \begin{cases} 1 & x \leq \theta - 0.1 \\ e^{1-\theta} & \theta - 0.1 \leq x \leq \theta \\ e^{x} & x \geq \theta \end{cases} \qquad (7)$$

The final loss function consists of three parts: the binary cross-entropy loss function for the classification task, and the Distribution Focal Loss and Complete Intersection over Union loss functions for the bounding box regression task.

$$Loss = L_{BCE} + L_{CIoU} + L_{DFL} \qquad (8)$$

The training strategy adopted in this paper aligns with the baseline network YOLOv8-s. The SGD optimizer is utilized with an initial learning rate of 0.01. The training and testing are conducted on the RTX3090 platform, with a batch size of 32.

4 Experiments

4.1 Smoking-3k Dataset.

In our investigation of relevant smoking detection datasets, we found no openly available dataset for training. Both Paddle and Ali’s DAMO utilize self-built datasets and provide API interfaces for users, but they do not disclose these datasets for experiments. Considering that smoking detection involves small targets and is prone to misidentification, and also has significant relevance to monitoring equipment in public places, we proposed the Smoking-3K Dataset. The Smoking-3K Dataset was created using the LabelMe annotation tool, resulting in a total of 3,059 smoking detection datasets in YOLO format, collected through methods such as web scraping and camera shots. This dataset includes 2,728 training datasets. The label distribution is shown in Fig. 3.

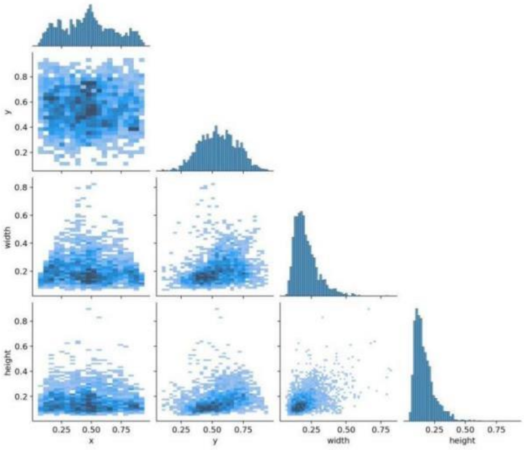

Figure 3. Distribution of smoking-3k detection boxes

The first and second distributions in the distribution map represent the positions of the horizontal and vertical axes of the center coordinates of the detection box within the entire image dataset. It is evident that the center points of the actual boxes are evenly distributed throughout the images. The third and fourth distributions in the map represent the proportions of the true frame width and height of the dataset within the entire image. It can be observed that the width and height of the actual boxes are predominantly less than a quarter of the image size, indicating that small target objects are primarily used as detection boxes.

4.2 Metrics

AP (Average Precision) is an indicator of the accuracy of the prediction box in target detection. In this paper, $AP_{0.5}$ and $AP_{0.5:0.95}$ are used as the evaluation indexes of model accuracy, and AP_{Small} is used as the evaluation index of small object recognition rate. Frames Per Second (FPS) refers to the number of frames per second, which is used to measure the reasoning speed of a model per second.

4.3 Analysis of Smoking-3k experimental results

This paper compared multiple target detection networks on self-made datasets, including PP-YOLOE [21], rt-detr [22], YOLOv7 [23], YOLOv8 [24], Cascade R-CNN [25], Efficientvit [26]. In terms of accuracy index, the model $AP_{0.5}$ and

$AP_{0.5:0.95}$ respectively ranked second only to efficient-Vit and rt-detr, respectively, and belonged to the first echelon of smoking detection accuracy index. Especially in the comparison of $AP_{0.5:0.95}$ index, it is significantly ahead of rt-detr with the third effect, indicating that under the condition of higher confidence as the threshold, our method can still effectively identify the behavior of smoking detection, and the model detection has high stability. In terms of network size, the parameter number of this model is the lowest among all models, only 7.09M. At the same time, the calculation amount of the model is 28.6Gflops, which belongs to the low complexity of the model, far lower than the best precision rt-detr. In terms of reasoning speed, our model's FPS can reach 155.2, which is the highest result among all models.

Table1. The detection algorithm is based on Smoking-3k comparison.

Model	$AP_{0.5:0.95}$	$AP_{0.5}$	FPS	Params	Gflops
PP-YOLOE	0.442	0.894	26.6	7.61	16.3
rt-detr	0.456	**0.953**	25.9	42.78	135.1
YOLOv7	0.452	0.876	29.8	37.20	105.1
YOLOv8-s	0.496	0.916	76.4	11.13	28.6
Cascade R-CNN	0.390	0.939	8.0	69.17	159.2
Efficientvit	**0.567**	0.928	27.8	8.38	**20.4**
Our model	0.566	0.947	**155.2**	**7.03**	28.7

4.4 Ablation experiment.

In this paper, the improvement of YOLOv8-s mainly includes the replacement of backbone network, the number of feature map fusion, and the addition of slide weighting function.

For the backbone network, we conclude that using 4-scale feature map to extract features greatly enhances the ability of the model to detect small target objects. This shows that ConvNeXt-v2 has better feature extraction ability and information retention ability for small target objects. However, the original backbone network of YOLOv8-s cannot generate effective attention to small targets, and the detection accuracy of small targets can reach 0.319 after replacing the backbone, which is 10.6% higher than that of YOLOv8-s. The results of ablation experiment are as follows:

Table2. Comparison of small target detection in

backbone network.

Backbone	$AP_{small}(0.5:0.95)$
YOLOv8-s	0.213
ConvNeXt-v2（our model）	**0.319**

In the neck network, the feature information of four different scales is used in this paper to improve the accuracy of detection, compared with the general feature maps of three scales entering the neck network for the purpose of capturing target information at different scales. The advantage of this approach is that the feature information of different scales in the task of smoking detection can be captured more comprehensively, thus improving the accuracy of detection. This additional scale change can better deal with the ratio change of small and large targets, and more effectively deal with the feature extraction of medium-scale targets. And because the fusion module is optimized, the number of parameters and calculation amount of the model remain lightweight, but the accuracy is not improved much mainly because: There are few training pictures containing white gas, which can be easily confused with smoke. The improvement of this section is mainly to solve the problem of correctly identifying the presence of smoking in pictures where smoke and white gas coexist.

Table3. Comparison of quantitative fusion effect of feature maps.

Feature map Number	$AP_{0.5}$
3	0.932
4（our model）	**0.947**

Use of the slide weighting function. The newly added slide function makes the model pay more attention to difficult to identify samples in the training process, and improves the utilization rate of the model to difficult samples when the number of data sets is too small. The improvement of accuracy results demonstrates the effectiveness of the use of the loss function, which is 0.3% higher than before. The following experiments show the changes in the accuracy of the model after using the loss function:

Table4. Usage of Slide Function.

Setup	$AP_{0.5:0.95}$
W/O slide function	0.487
W slide function	**0.490**

We also show through the heatmap in Fig. 4 that the focus on smoking behavior in the learning process of the model in this paper has generated a corresponding strong correlation. When holding a cigarette, the two fingers holding the cigarette will have higher attention. When smoking, the attention scores of the cigarette object and its surroundings were significantly higher than those of the surroundings. When lighting a cigarette, the flame generated by the lighter will attract the model's attention more easily. When the smoking action is not around the mouth, the model will still notice the mouth area of the smoker.

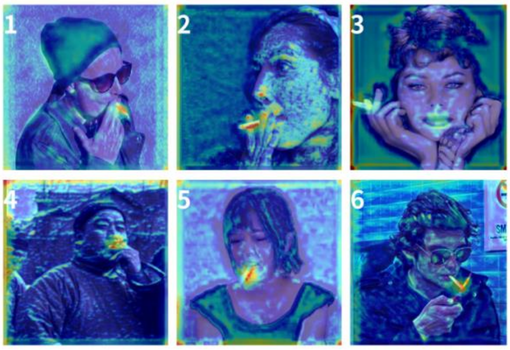

Figure 4. Heatmap of model attention scores.

5 Detection results and visualization

We use the model after training convergence to predict the smoking behavior with severe occlusion. It is found that the smoking behavior of the tested person can still be recognized when only a small amount of cigarette butts are exposed, which indicates that the model not only learns the cigarette object, but also reacts to the action behavior of smoking and the smoke generated by smoking in the process of learning smoking behavior. At the same time, in addition to ordinary RGB images, monochrome black and white photos are also included in the detected images, which can be detected by the model, indicating that the model is robust to the range of color changes. Furthermore, the act of smoking is a continuous action, which means that smoking may occur in the mouth, it may occur in the hand, and it may occur in the light of the fire, because this type of action already indicates that the tested person has a dynamic action of smoking. The model can recognize the smoking action in a variety of situations, not only the cigarette butt located in the mouth will be recognized as smoking. Finally, due to the limitation of existing data sets, smoking detection is generally only performed on ordinary cigarettes. In this paper, other types of cigarettes are added to the smoke-3K dataset, so that the model can maintain stability for uncommon types of cigarettes when learning smoking actions. All the above advantages are given in Fig. 5.

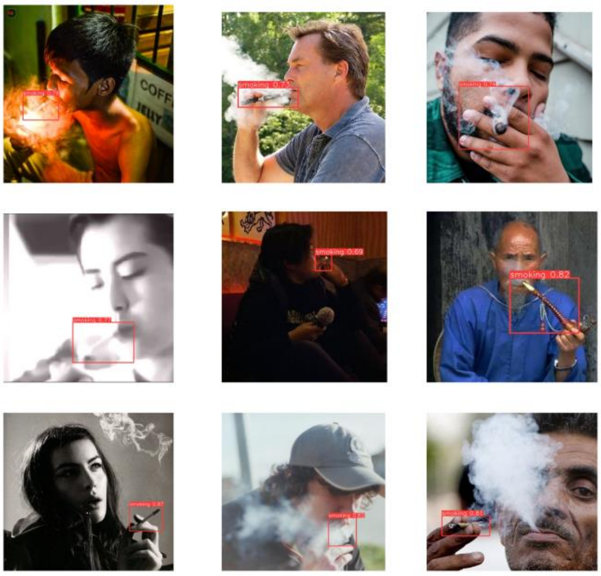

Figure 5. Results of smoking detection in multiple scenes.

6 Conclusion

This method is improved on YOLOv8-s, and the ConvNeXt-v2 with powerful feature extraction ability is used as the backbone network to improve the detection ability of small objects. In the feature fusion stage, feature maps of four scales are used for fusion. The relationship between high-level semantic information and low-level semantic information is strengthened, and the stability of detection is improved. In the loss optimization strategy, a higher weight is used to focus on the difficult negative samples, and this part of samples is emphasized. Experimental results show that the network has an $mAP_{0.5}$ of 0.947 and$mAP_{0.5:0.95}$ of 0.652 on the Smoking-3K dataset, which are increased by 3.1%and 7% respectively. In addition, the inference speed is much faster than other models, which indicates that it can have good robustness in occlusion scenes and complex environments.

At present, this method mainly focuses on the detection of smoking, and it is unable to do anything about other behaviors that need attention in public places. The focus of subsequent work will be to improve the detection of multi-category abnormal behaviors.

References

[1] Tripathi O, Parada Jr H, Shi Y, et al. Perception of harm is strongly associated with complete ban on in-home cannabis smoking: a cross-sectional study[J]. BMC Public Health, 2024, 24(1): 669.

[2] Ortis, Alessandro, et al. "A report on smoking detection and quitting technologies." International journal of environmental research and public health 17.7 (2020): 2614.

[3] Zhang, Zhen, et al. "Research on smoking detection based on deep learning." Journal of Physics: Conference Series. Vol. 2024. No. 1. IOP Publishing, 2021.

[4] Lakatos, Róbert, et al. "A multimodal deep learning architecture for smoking detection with a small data approach." Frontiers in Artificial Intelligence 7 (2024): 1326050.

[5] Stone C J, Essery R, Matthews J, et al. Evaluating the feasibility and acceptability of a smartwatch-based smoking relapse intervention ('StopWatch')[J]. 2024.

[6] Senyurek V Y, Imtiaz M H, Belsare P, et al. A CNN-LSTM neural network for recognition of puffing in smoking episodes using wearable sensors[J]. Biomedical Engineering Letters, 2020, 10: 195-203.

[7] Woo S, Debnath S, Hu R, et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16133-16142.

[8] Lei K, Tan Z, Wang X, et al. Semi-Symmetrical, Fully Convolutional Masked Autoencoder for TBM Muck Image Segmentation[J]. Symmetry, 2024, 16(2): 222.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[10] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.

[11] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[12] Pacal I. Enhancing crop productivity and sustainability through disease

identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model[J]. Expert Systems with Applications, 2024, 238: 122099.

[13] Ghiasi G, Lin T Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7036-7045.

[14] Wang K, Liew J H, Zou Y, et al. Panet: Few-shot image semantic segmentation with prototype alignment[C]//proceedings of the IEEE/CVF international conference on computer vision. 2019: 9197-9206.

[15] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[16] Jiang Y, Tan Z, Wang J, et al. Giraffedet: A heavy-neck paradigm for object detection[J]. arXiv preprint arXiv:2202.04256, 2022.

[17] Xu X, Jiang Y, Chen W, et al. Damo-yolo: A report on real-time object detection design[J]. arXiv preprint arXiv:2211.15444, 2022.

[18] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[19] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[20] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.

[21] Xu S, Wang X, Lv W, et al. PP-YOLOE: An evolved version of YOLO[J]. arXiv preprint arXiv:2203.16250, 2022.

[22] Lv W, Xu S, Zhao Y, et al. Detrs beat yolos on real-time object detection[J]. arXiv preprint arXiv:2304.08069, 2023.

[23] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[24] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

[25] Cai Z, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(5): 1483-1498.

[26] Liu X, Peng H, Zheng N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430.