AI Medicine
https://www.sciltp.com/journals/aim

Article

Faster R-CNN-MobileNetV3 Based Micro Expression Detection for Autism Spectrum Disorder

Hanni Li [1], Yutong Gu [1], Jiarui Han [1], Yimeng Sun [1], Hongwei Lei [1], Chen Li [1,*,†] and Ning Xu [2,*,†]

[1] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110016, China
[2] College of Art and Design, Liaoning Petrochemical University, Fushun 113001, China
* Correspondence: lichen@bmie.neu.edu.cn (C.L.); xuning201096@hotmail.com (N.X.)
† These authors contributed equally to this work.

How To Cite: Li, H.; Gu, Y.; Han, J.; et al. Faster R-CNN-MobileNetV3 Based Micro Expression Detection for Autism Spectrum Disorder. *AI Medicine* **2025**, *2*(1), 2. https://doi.org/10.53941/aim.2025.100002.

Received: 24 December 2024
Revised: 11 February 2025
Accepted: 11 March 2025
Published: 24 March 2025

Abstract: Autism spectrum disorder (ASD) is a neuropathic disease which is characterized by deficits in social interaction and communication. Therefore, the ASD patients have weak ability to express themselves or let others know about their thoughts. As society pays more attention to ASD patients, early intervention programs, behavioral therapy and technological assistance have emerged to help ASD patients improve their quality of lives. This paper aims to propose an improved object detection algorithm based on Faster R-CNN-MobileNetV3 to analyze the micro expressions of ASD patients. The data set includes 1358 face images of ASD patients built from 12 ASD movies with the method of Cinemetrics. Through the training and testing of the ASD data set with the improved model, the overall precision rate has reached 0.9 and mean Average Precision also has significant improvement. As a result, the improved Faster R-CNN-MobileNetV3 model achieves a good performance to recognize micro expressions and emotions of ASD patients.

Keywords: autism spectrum disorder; micro expressions; Cinemetrics; object detection; Faster R-CNN; MobileNetV3

1. Introduction

Autism spectrum disorder (ASD) is a kind of neurodevelopmental disorders disease, mainly for social interaction barriers and repetitive stereotyped behavior or interest. In the first 2 years of a child's life, typical early symptoms of ASD may include not responding to their name when called, minimal or absent use of gestures for communication, and a lack of imagination. ASD includes a series of symptoms that reflect social disorders and restricted repetitive behaviors, ranging from mild to severe. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), a guideline developed by the American Psychiatric Association, patients with autism spectrum disorders usually have the symptoms of difficulty in communicating and interacting with others, and limited interest and repetitive behaviors, that affect their abilities in school, work and other aspects of their lives. The complexity and heterogeneity of ASD are related to genetic developmental factors (such as age and IQ) and environmental factors (such as the availability of support, including personalized educational services and speech, language and behavioral interventions). Several different genes may be the cause of autism spectrum disorders. Some children with autism spectrum disorders may have genetic diseases, such as Rett syndrome or fragile X syndrome. As for the environmental factors, autism spectrum disorders may also be related to factors such as viral infections, drugs, air pollutants, or complications during pregnancy [1–3]. In the United States, about one in 59 school-age youth has autism spectrum disorder (ASD). In the past two decades, the prevalence of ASD has been steadily increasing, and it is currently estimated to be as high as 1/36 of children. Nearly 75% of ASD patients

have comorbid mental illness or symptoms, which may include attention deficit hyperactivity disorder (ADHD), anxiety, bipolar disorder, depression, and Tourette's syndrome. In 2000, the Centers for Disease Control's Autism and Developmental Disorders Monitoring (ADDM) network estimated that the incidence of ASD was one in every 150 children. In 2006, the incidence of ASD increased to one in every 110 children, and in 2008 it increased again to one in every 88 children. In 2012, the ADDM network revised its ASD estimate to one in every 68 children [4,5].

Analyzing micro expression is a useful method to diagnose autism spectrum disorder. In the field of autism spectrum disorder, analyzing micro expressions of patients can help better understand their emotional states and social interactions. Micro expression is a sort of imperceptible facial movements with very short duration, that can reveal underlying true feelings. It is crucial for identifying emotional disorders and improving communication strategies. Typically, people consciously express their emotions through macro-expressions that last from 0.5 to 4 s and are easily perceived by humans. However, psychological research has shown that macro-expression may mislead human emotional recognition [6,7]. Different from macro expression, the duration of micro expression is usually less than half a second, mostly unconscious expression, which can reveal real emotions [8]. Micro expressions, as a key form of nonverbal communication, serve as a valuable tool for understanding genuine human emotions. They can be utilized for non-contact and non-perceptual detection of deception or recognition of abnormal emotions. These micro expressions can disclose an individual's authentic emotional state [9,10]. The analysis of micro expressions in autistic patients is of great help for diagnosis and treatment. Micro expressions can provide subtle cues of emotional state and help doctors to evaluate emotional understanding and social ability of patients more accurately. By identifying specific micro-expression changes, it can indicate emotional distress, provide data support for the formulation of treatment plans, and then guide personalized treatment plans so that the needs of patients will be met more effectively. In addition, the analysis of micro expressions can enhance communication between patients and therapists during treatment, promote more effective interventions, and improve treatment outcomes.

In this paper, an improved algorithm based on Faster R-CNN (Region-based Convolutional Neural Networks) is proposed to recognize micro expressions of ASD patients better. The flowchart of the overall project is shown in Figure 1. Experimental data is derived from ASD movies and expanded to 1358 images of ASD patients, 31% more than the previous study. To pinpoint the faces of ASD patients in the movies, we use Adobe Premiere Pro 2024 software to capture shots of patients with ASD in the movies and then perform video framing with MATLAB (2020b) at 24 interval frame number. Then, the face parts of the framed images are labeled with the types of micro expressions. There are 482 kinds of micro expressions, which expressed different and complex emotions of ASD patients. The basic Faster R-CNN model has been progressed to achieve a higher precision, recall rate and MAP (Mean Average Precision). The original backbone network is VGG16 (Visual Geometry Group), while the improved one is MobileNetV3. More convolution layers are added to enhance feature extraction capability and classification performance. Some hyperparameters are also changed, which are mentioned below, so that the model achieves a better performance.

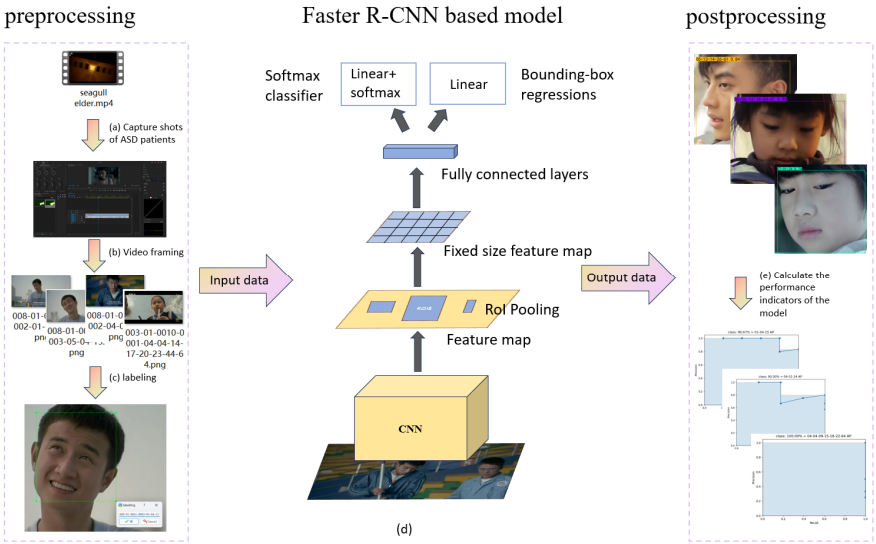


Figure 1. Flowchart of the whole project (step (**a**) capturing the shots of ASD patients from ASD movies, step (**b**) framing the videos captured, step (**c**) labeling the images framed with micro expressions, step (**d**) using Faster R-CNN-based model to train and predict with data set, step (**e**) calculating the performance indicators of the model, such as, precision, MAP).

Compared with the basic Faster R-CNN model experiment we made, we design a MobileNetV3-based Faster R-CNN network. The improved model includes MobileNetV3 network, Region Proposal Network, RoI Pooling (Region of Interest Pooling), fully connected layers and output layers. The Depthwise Separable Convolution of MobileNetV3 network significantly reduces the amount of computation and parameters, which improves computational efficiency. Some additional convolutional layers can further increase the complexity of the model and help improve feature extraction capabilities, resulting in better performance in face detection tasks. In this improved model, the scale of the small object detection anchors is reduced, which increases the face detection rate and recall rate. Smaller anchor frames can cover more potential targets. In the field of application, the Faster R-CNN-MobileNetV3 model can judge the types of facial micro expressions more accurately. This can be applied to hospital departments that diagnose ASD patients or rehabilitation centers that treat ASD patients.

2. Related Works

2.1. Micro Expressions of ASD Patients

To study the micro expressions of ASD patients, we cut off lens from 12 ASD films and frame them getting 1358 images. Then, we label these images depending on the specific labeling criterion to get 1358 labeling files.

The names of data set labels represent the different categories of micro expressions. The categories, which include seven kinds of macro expressions, are shown in Figure 2. The 30 kinds of micro expressions, described in Table 1, are classified according to the Facial Action Coding System (FACS). The Facial Action Coding System evolved from Micro-expressions (ME) phenomenon in 1969 by Ekman et al. [11]. The 30 kinds of micro expressions correspond to 30 action units (AUs), which are components of facial expressions in FACS [12].

AU analysis can effectively resolve the ambiguity issue to represent individual expression and increase Facial Expression Recognition (FER) performance [11]. After recognizing the variation of micro expressions of ASD patients, researchers and clinicians will understand the emotional states of people with autism better and develop personalized treatment plans. In addition, during the rehabilitation process, micro expressions can serve as feedback to help doctors adjust therapying strategies so that achieving better therapeutic effect.

Figure 2. Seven kinds of macro expressions (‘**01**’ corresponds ‘disgust’, ‘**02**’ corresponds ‘anger’, ‘**03**’ corresponds ‘fear’, ‘**04**’ corresponds ‘sadness’, ‘**05**’ corresponds ‘happiness’, ‘06’ corresponds ‘contempt’, ‘**07**’ corresponds ‘surprise’) [13].

Table 1. Micro expressions classification and details [14].

AU	Micro Details	AU	Micro Details	AU	Micro Details
1	Inner Brow Raiser	13	Cheek Puffer	25	Lips part
2	Outer Brow Raiser	14	Dimpler	26	Jaw Drop
4	Brow Lowerer	15	Lip Corner Depressor	27	Mouth Stretch
5	Upper Lid Raiser	16	Lower Lip Depressor	28	Lip Suck
6	Cheek Raiser	17	Chin Raiser	41	Lid droop
7	Lid Tightener	18	Lip Puckerer	42	Slit
9	Nose Wrinkler	20	Lip stretcher	43	Eyes Closed
10	Upper Lip Raiser	21	Lip Funneler	44	Squint
11	Nasolabial Deepener	23	Lip Tightener	45	Blink
12	Lip Corner Puller	24	Lip Pressor	46	Wink

2.2. Cinemetrics

Cinemetrics is a tool which quantitatively analyzes the study of film, with the aim of using data and statistics to explore the structure, style, and narrative aspects of film. It combines statistics, big data, and cloud computing to extend digital humanities in the field of film study [15,16]. Cinemetrics can help determine the style of a film more accurately through the length of each shot, field, angle, scheduling mode, and other formal elements combining quantitative and qualitative comprehensive analysis of authors, text and data [17].

The process of data preparation uses the method of Cinemetrics to collect and quantify data on micro expressions features in ASD films, such as the switching frequency of the lens and the change of facial features. Through analyzing performances of ASD patients in the films quantitatively, we can compare them with the emotional expressions in other types of films to differ the characteristics of micro expressions of ASD patients. This helps understand the difficulty of ASDs social communication. Facial recognition and sentiment analysis technology can be used for more in-depth analysis of micro-expressions in the films. These models perform data mining based on specific micro expressions data to recognize the emotions of ASD patients.

2.3. Micro Expression Recognition Detection Techniques

Micro expression recognition has the same basic techniques as face recognition. Face recognition is a biometric technique that identifying the identity of a living individual by analyzing their physiological traits or behaviour patterns with automated methods to find face region [18,19]. However, different from face recognition, micro expression recognition aims to detect subtle emotional changes in the face by paying attention to dynamic changes in facial features, especially subtle muscle movements. In the previous object detection algorithms experiments, Faster R-CNN, SSD (Single Shot MultiBox Detector) and YOLO (You Only Look Once) have been tried to analyse types of micro expressions using the original data set we made and compare the effects of three algorithms.

2.3.1. Faster R-CNN Model

Faster R-CNN, a two-stage object detection method, consists of four modules, which are convolutional layers, region proposal network (RPN), RoI pooling and classification and regression [20]. The conv layers are used to extract features through a set of conv layers, ReLU (Rectified Linear Unit) layers and pooling layers. In the region proposal network, the feature map serves as input and output candidate boxes of objects on the images [21]. RPN uses softmax regression to dichotomy the anchors and correct them to get more accurate proposals [22]. The RoI pooling collects the feature maps from RPN and sends them to fully connected layers to classify and regress. The final part calculates the specific category of each proposal feature maps and position of detection box through bounding box regression. Through the experiment, the Faster R-CNN showed a high accuracy and precision.

2.3.2. SSD Model

Single shot multibox detection (SSD), a single stage object detection method, classifies and locates objects in a single forward propagation. Figure 3 is the architecture of the model. The backbone of SSD contains a basic network VGG16 and multi-scale feature maps. The basic network usually is a pre-trained convolutional neural network to extract feature maps. Then, the model detects on different scales of feature maps through different default boxes. Finally, convolution prediction is responsible for the object category prediction and position prediction to predict the object and position [23,24]. Through the experiment, we found the SSD model had a high

speed of detection and was appropriate for real-time detection and detection on multi-scale objects, but it had a little worse accuracy.

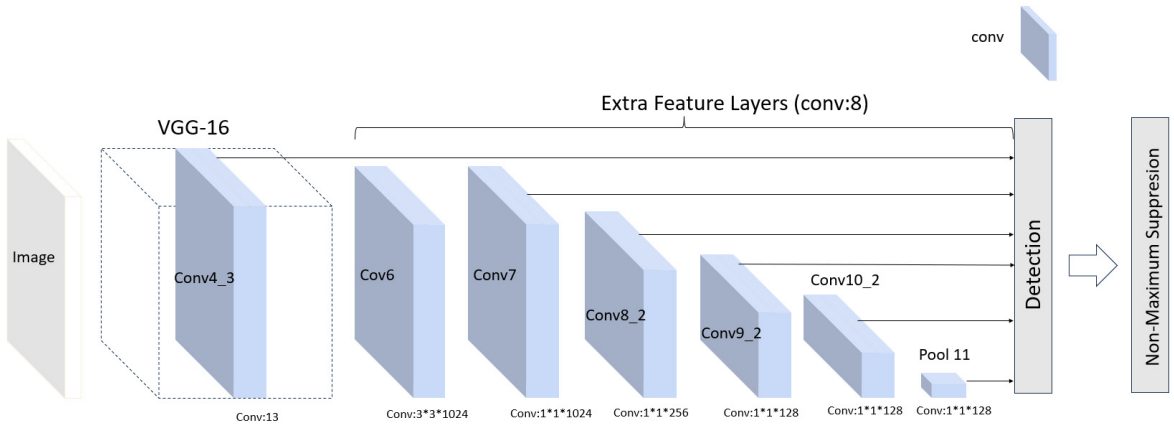


Figure 3. Architecture of SSD model.

2.3.3. YOLO Model

YOLO, standing for You Only Look Once, is a kind of Efficient real-time object detection algorithm. Figure 4 is the architecture of the model. It uses a neural network model to make global predictions directly on the input image, identify multiple targets at the same time, and classify and locate each target. Through the end-to-end unified full convolutional network, YOLO model inputs the whole image to the network and divides grids on the images, and outputs the bounding boxes and detection results directly [25]. After experiments, we found although the detecting speed of the YOLO model is high, the precision is very low probably due to the global prediction. When it is used to detect the micro expressions on smaller face or dense faces of an image, inaccurate and imprecise positioning may occur.

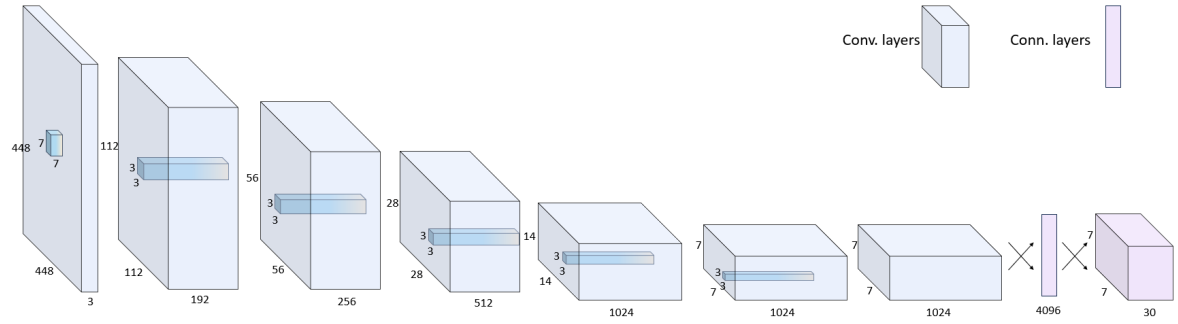


Figure 4. Architecture of YOLO model. (Conv. Layers: 24. Conn. Layers: 2.)

3. Improved Approach of Faster R-CNN Based on MobileNetV3

Through the comparative analysis of the four kinds of algorithms to recognize the micro expressions of ASD patients, Faster R-CNN presents the best training effects and accuracy of prediction. Therefore, we try to improve the Faster R-CNN model to achieve higher accuracy and precision and use a larger ASD data set to test the improving effects.

3.1. Overall Network Structure of Improved Faster R-CNN Based on MobileNetV3

The Faster R-CNN model is built based on Region-Based Convolutional Neural Networks (R-CNN) and Fast R-CNN model, introducing Region Proposal Network (RPN). Traditional feature extraction network, the core part of the model, is mainly VGG16 model and Resnet model, used to extract the features of the input image [26,27]. These feature maps are used to the candidate region generation and target classification in the following steps. The feature extraction network will extract spatial features of images through multilayer convolution and pooling operations and output a high dimensional feature graph, reflecting the feature distribution of different regions in the image [28]. In this paper, we add MobileNetV3 network to replace VGG16 as the backbone. Figure 5 shows the overall structure of improved Faster R-CNN model.

MobileNetV3 is a lightweight convolutional neural network, designed with depthwise separable convolution, bottleneck and attention mechanism to reduce calculation effectively [29]. Depthwise separable convolution includes depthwise convolution and pointwise convolution. The depthwise convolution convolves each channel and uses a single convolution kernel for each input channel. Pointwise convolution combines the information of each channel from depthwise convolution, applying each convolution kernel to all spatial locations, and generates the output. Therefore, the depthwise separable convolution improves the efficiency and number of parameters of model. The bottleneck structure, also called inverted residual block, adds the number of channels with 1 × 1 convolution to expand the dimensions of feature maps at first [30]. This makes the model catch more features of input. Then, it deals with the spatial information with 3 × 3 convolution. There is a squeeze-and-excitation module (SE) among the MobileNetV3, used to generate attention weights for channels and adjust the response for each channel through adaptive average pooling, full connection layers and activation functions. Finally, it also uses 1 × 1 convolution to generate output [31]. The bottleneck structure also helps the network more efficient. If the dimensions of input and output are the same, there will be a skip connection to enable information skip directly to the output, preventing the disappearing of gradient. Besides the traditional ReLU activation function, Hard Swish is also utilized in most convolution layers [32], which reduces the complexity of calculation and increases the expressiveness and nonlinearity of the model.

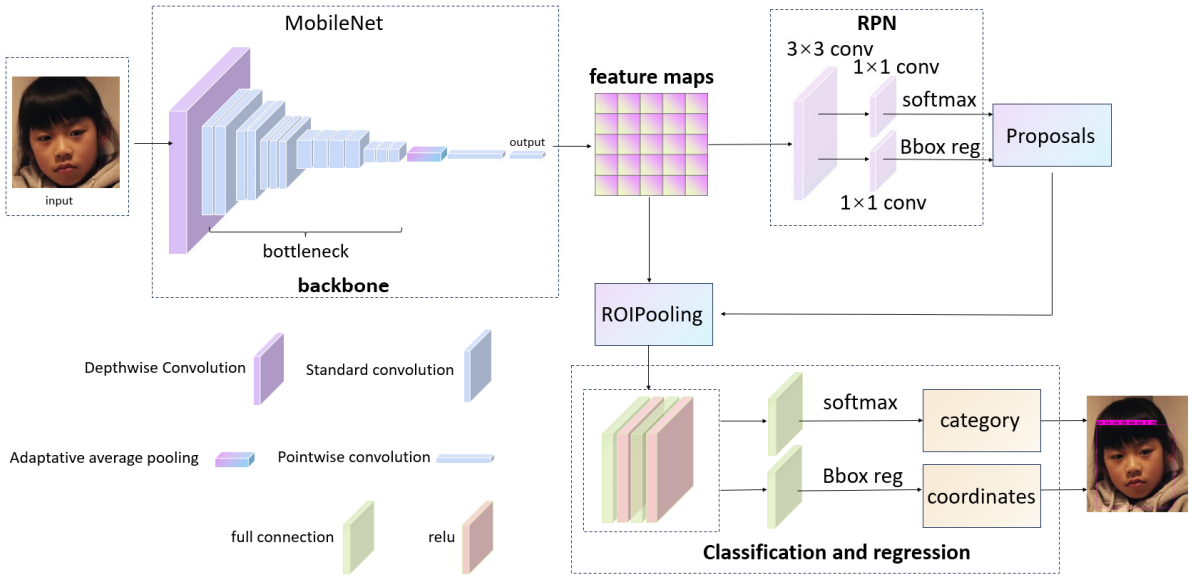


Figure 5. Architecture of Faster R-CNN model based on MobileetV3.

3.2. RPN Module of Faster R-CNN

Region Proposal Network (RPN) is one of the core components of Faster R-CNN model, aiming to generate a series of region proposals from the feature maps. These region proposals are likely to contain the detected objects, which helps reduce the computational amount of subsequent processing [33]. These region proposals are produced by sliding window convolution and anchors [34]. RPN uses a 3 × 3 convolution kernel to traverse the entire feature map through sliding windows and extract local feature of each position. The region proposals are framed by the anchors with different sizes and scales. Each anchor corresponds to a region in the image. The anchors have three kinds of scales and three kinds of aspect ratios, which enables to generate nine different proposals [35]. In this project, we change the scale of small object detection 128 × 128 to 112 × 112, and the scale of medium object detection and big object detection are still 256 × 256 and 512 × 512. The aspect ratios are 1:1, 1:2 and 2:1. This helps improve the detection to small and medium object detection and the accuracy to recognize the micro expressions. Then, these anchors will screen out the areas that might contain the object through classification and regression.

The RPN module includes classification branch and regression branch. The classification branch will determine if each anchor contains objects using softmax function, while the regression branch will adjust the positions and scales of anchors to approach to object bounding boxes, which also outputs center point coordinates and width to height ratio of anchors. Then, the non-maximum suppression (NMS) can reduce the redundant candidate regions produced by the repeated anchors effectively. Finally, a series of region proposals are outputted and enter the RoI Pooling.

3.3. RoI Pooling of Faster R-CNN

Region of Interest Pooling (RoI Pooling) transforms arbitrarily sized region proposals into fixed-size feature maps so that classification and bounding box regression tasks can be handled with consistent input sizes. First, RoI Pooling maps the coordinates of the candidate region from the input image space to the feature map scale. Then, it divides the feature maps into fixed-size grids. At last, RoI Pooling applies the maximum pooling operation to select the maximum activation value in the grid as the output. This helps to capture the most important features from the grid, minimize feature loss, and condense the feature map into a fixed size. The fixed size feature maps are output and enter the classification and regression layers.

3.4. Classification and Regression of Faster R-CNN

The fixed size feature maps are inputted into fully connected layers at first, generating one-dimensional vectors to extract features. Then, the vectors enter a fully connected layers used to classify [34]. It uses softmax activation function to calculate probability distributions for objects and background categories, which are the probability distributions of region proposals. Therefore, all categories of data set will add an 'environment'. The bounding box regression branch is responsible to adjust the bounding boxes of proposals precisely [34]. The output is the offsets of positions and width and height of proposals. The offsets are used to correct the proposals of RPN to bound the objects more closely. Faster R-CNN also uses total loss function to optimize the accuracy of classification and bounding box regression. Classification loss calculates the error between real categories and predicted categories through cross entropy loss function. Regression loss calculates the gap between the predicted boundary of the candidate box and the true boundary box through smooth L1 loss function [36]. The smooth L1 loss adopts L2 loss for smaller errors and L1 loss for larger errors, making the bounding box regression more stable and robust. The final total loss function is the weighted sum of classification loss and regression loss, which improves the whole detection effects.

4. Experiment with ASD Data Set

4.1. Experimental Environment Configuration

The computer for experiments is Windows11, equipped with an AMD Ryzen7 5800H with Radeon Graphics processor with 16 cores and 16 threads (AMD, Inc. Santa Clara, CA, USA). The main frequency is 3.2 GHz and maximum acceleration frequency is 5.2 GHz, achieving excellent performance. The graphics card uses NVIDIA GeForce RTX 3050 and has 2 GB of video memory, with 16.0 GB of RAM (Nvidia Corporation. Santa Clara, CA, USA). Therefore, the computer can run multiple applications smoothly. Besides, the computer uses 135 W, 20 V power supply to ensure stable operation.

The Adobe Premiere Pro 2024 software is used to capture videos from ASD movies that contain ASD patients. It supports many video formats and has high speed of clip and export, which enables flexible and convenient operations. The videos are flamed by MATLAB software (2020b Version). The MATLAB also supports many video formats. It uses 'VidioReader' class to read video files by frame and then obtains segmented images with specific frame intervals. The labeling software we used to label the framed images is Labelimg software (v1.8.6). Users can add rectangular bounding boxes on images with categories. The output files are Pascal VOC (XML) format and YOLO (txt) format, which can be applied in the object detection models. The environment for deep learning is built with Python 3.7.1 and Pytorch 1.7.1, supported by CUDA 11.0 and cuDNN 8.0.4, which ensures GPU acceleration for deep learning tasks. The package management tool is Conda 24.3.0.

4.2. Evaluation Index

We use four kinds of evaluation indexes to evaluate the training effects and overall performance of the improved Faster R-CNN model in the detection of micro expressions project.

Precision is an index to measure the accuracy of the positive sample prediction in the model prediction results. It represents the proportion of real positive samples in all samples predicted by the model to be positive. The higher precision, the fewer false results in the predicted positive samples. IOU (Intersection over Union) is a parameter to determine if predicted boxes match the real boxes [37]. In this project, if IOU is bigger than 0.6, we consider the predicted boxes are TP. Therefore, it can be expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

TP (True Positive): the number of detected objects correctly.

FP (False Positive): the number of targets detected by errors.

Recall is the index to measure the coverage of positive samples in the model prediction results. It represents the proportion of positive samples to be predicted in all real positive samples. The higher recall, the fewer samples missed to predict. It can be expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

FN (False Negative): the number of real positive samples missed to be predicted positive.

F1 Score is the index to evaluate accuracy and recall rate of model comprehensively. It is a harmonic mean to weigh the predictive accuracy and comprehensiveness of the model. The higher F1 Score, the better performance of model. It can be expressed as:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

MAP, standing for Mean Average Precision, is the average of all average precision of each category, used to measure the detection accuracy and positioning accuracy for all categories of the model. The Average Precision is the area of PR curve of a category, which takes Recall as the horizontal axis and Precision as the vertical axis. It can be measured by:

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

N is the number of all categories.

AP_i is the Average Precision value of category i.

4.3. Experimental Results and Discussion

With the ASD data set of 1358 face images of ASD patients and tag files, we use the improved Faster R-CNN-MobileNetV3 model to experiment. The UnFreeze_Epoch is set to 100. Through the epoch loss figure (Figure 6), we can observe the change of the train loss and validation loss (val loss) with the change of epoch and overall trend by smoothing the curve.

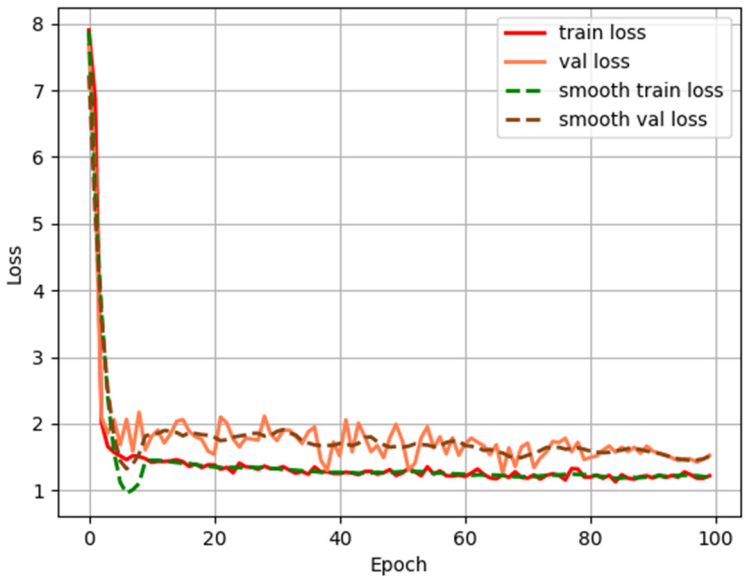


Figure 6. Epoch loss figure of the Faster R-CNN-MobileNetV3 model.

The train loss and validation loss decreased significantly in the first 10 epochs shows model has learnt to extract features and optimize predictions gradually. The trend of training loss and validation loss after smoothing is similar, which illustrates the training process of model is stable. After 20 epochs, the changes in training and validation losses level off, indicating that the optimization of the model gradually converges. The overall fluctuation range of validation loss is not large and the verification loss after smoothing is close to the training loss, indicating that the model is not significantly overfitting. In the final stage, validation loss and training loss

are almost the same, indicating that the performance of the model on the training set and the verification set is close, and the generalization ability is good. Therefore, the improved Faster R-CNN-MobileNetV3 model achieves a good performance.

In the original experiments, we compared and analysed six algorithms, which were Faster-RCNN, SSD, RetinaNet, YOLOv3, YOLOv5 and YOLOv8. We found the experimental results of Faster-RCNN are the highest, with 0.89 precision, 0.82 recall, 0.85 F1-Score and 84% mAP, shown in Table 2. In this study, we repeated the Faster R-CNN-MobileNetV3 experiments three times and found the best effect was improved to 0.90 precision, 0.86 recall, 0.88 F1-Score and 89% mAP, shown in Table 2, which is better than the model before the improvement. This shows that the overall performance of the model in the object detection task is fully optimized.

Table 2. Experimental results of the Faster R-CNN model and Faster R-CNN-MobileNetV3 model with ASD data set.

Times	Precision	Recall	F1-Score	mAP
Original	0.89	0.82	0.85	84%
1	0.88	0.85	0.86	86%
2	0.90	0.86	0.88	89%
3	0.88	0.84	0.86	85%
Average	0.89	0.85	0.87	87%

The comparison graph of original results and results of three times experiments is shown in Figure 7. We can see the recall rate has increased dramatically, which means the number of missed data has decreased. The higher mAP also illustrates the accuracy of object localization and classification are enhanced.

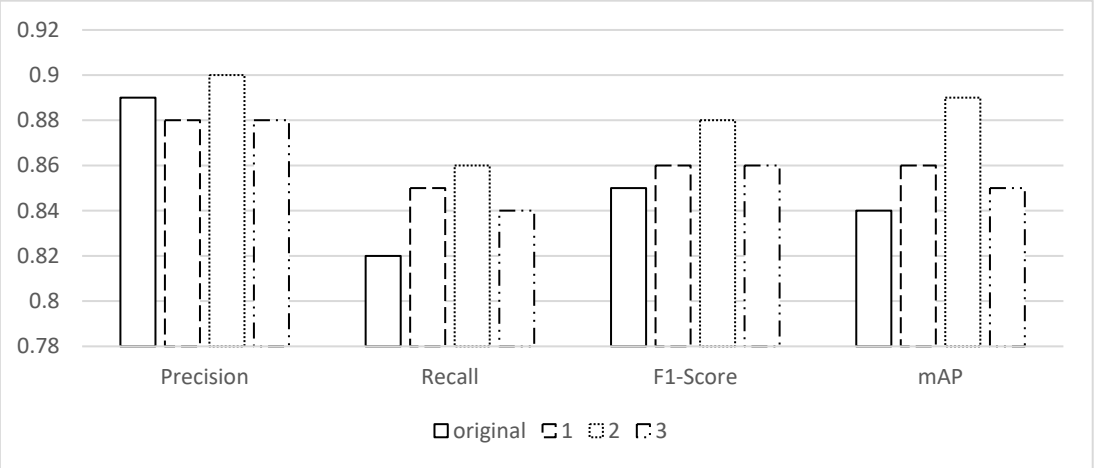
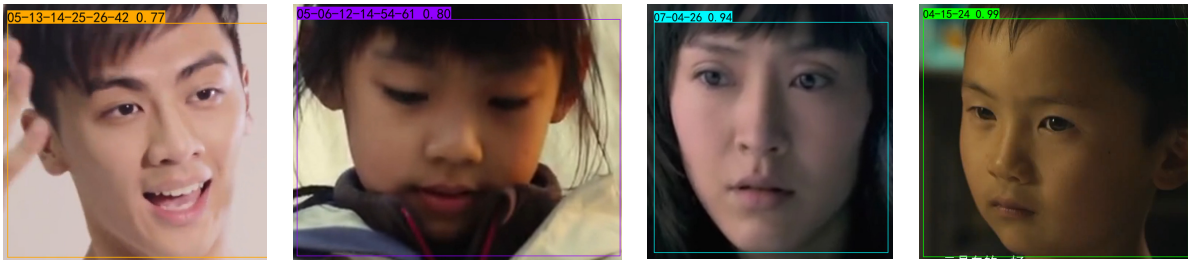


Figure 7. The longitudinal comparison between the experimental results of the original model and the experimental results for three times of the improved model.

The output results examples of Faster R-CNN-MobileNetV3 model are shown in Figure 8. The input images are face images of ASD patients obtained from ASD movies. Through the prediction of the model, images are bounded with boxes on the whole faces and the categories of micro expressions and confidence score are also shown. The confidence score is the degree to which the model is confident that the target in the box belongs to the prediction category. The closer the score is to one, the more confident the model is about the predicted target.



(**a**) (**b**) (**c**) (**d**)

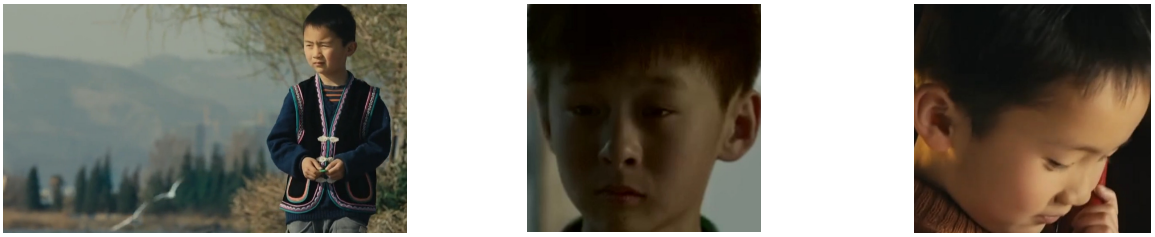
Figure 8. Examples of output results through Faster R-CNN-MobileNetV3 ((**a**) from "*On the road*", (**b**) from "*Me and my partner*", (**c**) from "*Don't say sorry. Don't say goodbye*", (**d**) from "*Old seagull*").

4.4. Limitations

There are also some unsuccessful output images through the prediction of the improved model. Through the analysis, we divided the failed outputs into three categories below.

4.4.1. Missed Detection

Among the output results, a few images are not detected, shown in Figure 9. From the perspective of the images, the object in Figure 9a is too small and the background is too large, which causes difficulty to classify. In Figure 9b, the light is too dim to judge the category clearly. In Figure 9c, the facial expression is incomplete. Therefore, in the actual detection of micro-expression categories of ASD patients, the light should be sufficient, and the patient's face should be facing the camera and as close as possible to the camera, so that clear and complete micro-expression images can be collected. Besides, there are also some problems with the parameters of the model. It may be caused by failed match of the anchor settings and small amount of data in this category. The model generates region proposals depending on the sizes and scales of anchors. If the anchor does not fit the size or shape of the object, the target area cannot be captured correctly. Besides, if the amount of a specific category of micro expression is not enough, the variety of scenes, pose, size, lighting covered by the sample may be very limited. Therefore, models cannot learn the general features of the objects.



(**a**) (**b**) (**c**)

Figure 9. Examples of missed detection outputs through Faster R-CNN-MobileNetV3 ((**a**) from "*Ocean Heaven*", (**b**) from "*The boy from the stars*", (**c**) from "*Ocean Heaven*").

4.4.2. Multiple Categories Outputs of Single Object

Some examples about repeated detection and multiple categories outputs of single object are shown in Figure 10. The problem of repeated detection and output of multiple overlapping boxes may have three reasons. The first one is the IOU threshold of NMS, which is used to remove overly overlapping boxes. If the value is set too high, some overlapping enclosures may fail to be filtered, resulting in multiple enclosures. Secondly, if the output box confidence threshold is set too low, the model is not confident enough about these predictions, which may retain many low confidence prediction boxes. Another one is overlapping features. The characteristics of the detection micro expressions are relatively complex, which may lead to multiple similar candidate boxes generated by the RPN module, and therefore, multiple overlapping boxes may be output after classification.

(**a**) (**b**) (**c**)

Figure 10. Examples of multiple categories outputs of single object through Faster R-CNN-MobileNetV3 ((**a**) from "*Ocean Heaven*", (**b**) from "*Don't say sorry. Don't say goodbye*", (**c**) from "*A singing fish*").

4.4.3. Off-Object Detection Boxes

In some output results, the detection boxes deviate from the faces, and the facial micro-expressions are not completely detected, as shown in Figure 11. This phenomenon will make users unable to fully understand the emotional state of ASD patients or even misunderstand them. In future study, we will adjust the size and proportion

of the anchor box to make it more suitable for the characteristics of human face and introduce multi-scale feature pyramid network (FPN) to enhance the detection ability of multi-scale targets. Insufficient learning of the bounding box regressors will also lead to excessive offset between the predicted box and the real box. The regression loss function will be corrected to optimize the positioning accuracy of the bounding box. Besides, increase the training rounds to ensure full convergence of the bounding box regressors.

(**a**) (**b**) (**c**)

Figure 11. Examples of Off-object detection boxes through Faster R-CNN-MobileNetV3 ((**a**) from "*Me and my partner*", (**b**) from "*Don't say sorry. Don't say goodbye*", (**c**) from "*Ocean Heaven*").

4.4.4. Data Security and Privacy Issue of ASD Patients

As the rapid development of the Internet and social media, some medical data or research datasets are available online under authorized conditions. However, it also leads to some security problems, such as the leakage of medical data and invasion of patient privacy for the purpose of illegal profit [38]. Therefore, we will encrypt the face images of ASD patients in future study with the use of some deep learning models. This not only protects the portrait privacy of ASD patients, which avoids unfair treatment from society and complies with the ethical and legal requirements but also ensures the security of data transmission [39,40]. From the perspective of scientific research, accurate medical data will also prompt the improvement of experimental results and help the study about facial micro expressions of ASD patients move forward.

5. Conclusion and Future Work

In this paper, we first expand the original ASD data set to enrich the categories of micro expressions of ASD patients. Then, the Faster R-CNN model is improved by replacing the backbone network with MobileNetV3 and adjusting some hyperparameters. Through the training and testing of the ASD data set with the improved model, the highest precision rate is 0.9, which increases the ability to recognize the micro expressions and emotions of ASD patients. Besides, the experimental results of the three repeated experiments show that the improved model has a stable object detection effect. However, this model still exists the phenomenon of repeated detection and missing detection. Therefore, in the future, we will continue to supplement the data volume and try data enhancement to ensure that the model can learn the diversity of the object characteristics and enhance the robustness of the model for different scenarios. We also consider optimizing model parameters and introducing stronger backbone networks to extract more accurate features of the face objects.

Author Contributions

All authors contributed to the study's conception and design. Material preparation and data collection were performed by Y.S. and J.H. The first draft of the manuscript was written by H.L. All authors have read and agreed to the published version of the manuscript.

Funding

The work is supported by the "National Natural Science Foundation of China" (No.1282220108007).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data used in this paper are from an open dataset for free use.

Acknowledgments

We thank Zixian Li and Guoxian Li for their important discussion.

Conflicts of Interest

The authors have no relevant financial or non-financial interests to disclose.

References

1. Hirota, T.; King, B.H. Autism Spectrum Disorder: A Review. *J. Am. Med. Assoc.* **2023**, *329*, 157–168. https://doi.org/10.1001/jama.2022.23661.
2. National Institute of Mental Health. Autism Spectrum Disorder—National Institute of Mental Health (NIMH). Available online: https://www.nih.gov/ (accessed on 12 February 2024).
3. Mayo Clinic. Autism Spectrum Disorder—Symptoms and Causes—Mayo Clinic. Available online: https://www.mayoclinic.org/ (accessed on 6 January 2018).
4. Wood, J.J.; Kendall, P.C.; Wood, K.S.; et al. Cognitive Behavioral Treatments for Anxiety in Children with Autism Spectrum Disorder: A Randomized Clinical Trial. *JAMA Psychiatry* **2019**, *77*, 474–483. https://doi.org/10.1001/jamapsychiatry.
5. Sharma, S.R.; Gonda, X.; Tarazi, F.I. Autism Spectrum Disorder: Classification, diagnosis and therapy. *Pharmacol. Ther.* **2018**, *190*, 91–104. https://doi.org/10.1016/j.pharmthera.2018.05.007.
6. Porter, S.; Ten Brinke, L. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychol. Sci.* **2008**, *19*, 508–514.
7. Ekman, P.; Martin, C.W. Lie catching and microexpressions. In *The Philosophy of Deception*; Oxford University Press: Oxford, UK, 2009; pp. 118–121.
8. Wang, C.; Peng, M.; Bi, T.; et al. Micro-attention for micro-expression recognition. *Neurocomputing* **2020**, *410*, 354–362.
9. Li, J.; Yap, M.H.; Cheng, W.H.; et al. FME'22: 2nd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis. In Proceedings of the 30th ACM International Conference on Multimedia, New York, NY, USA, 10–14 October 2022; pp. 7397–7399.
10. Guerdelli, H.; Ferrari, C.; Barhoumi, W.; et al. Macro-and micro-expressions facial datasets: A survey. *Sensors* **2022**, *22*, 1524.
11. Li, Y.; Wei, J.; Liu, Y.; et al. Deep learning for micro-expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2028–2046.
12. Sayette, M.A.; Cohn, J.F.; Wertz, J.M.; et al. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *J. Nonverbal Behav.* **2001**, *25*, 167–185.
13. Jeffrey, F. Cohn, Takeo Kanade. Cohn-Kanade Dataset (CK), Available online: https://www.kaggle.com/datasets/davilsena/ckdataset (accessed on 20 September 2024).
14. Zarins Uldis. *Anatomy of Facial Expression*. Anatomy Next, Inc.: Beacon, NY 12508, USA, 2019.
15. Fan, B.; Guo, L. Stylistic Transfer between Intergenerational Directors: a Metrological Study of the Fifth and Sixth-Generation Chinese Directors. *J. Guizhou Univ. Art Ed.* **2021**, *2021*(3), 72–85. (In Chinese).
16. Li, W.; Wang, Z. Visualization Analysis on Intellectual Structures and Research Fronts of "Cinemetrics": Quantitative Film Studies. In Proceedings of the 2023 8th International Conference on Information and Education Innovations, Manchester, UK, 13–15 April 2023; pp.206–210.
17. Qiao, J.Q., A glimpse into the style of early Chinese cinema (1922–1937): A quantitative film studies perspective. *Movie Review* **2022**, *2022*(22). Doi: 10.16583/j.cnki.52-1014/j.2021.22.025. (In Chinese).
18. Tolba, A.S.; El-Baz, A.H.; El-Harby, A.A. Face recognition: A literature review. *Int. J. Signal Process* **2006**, *2*, 88–103.
19. Hassaballah, M.; Aly, S. Face recognition: Challenges, achievements and future directions. *IET Comput. Vis.* **2015**, *9*, 614–626.
20. Zhu, Z.; Liu, L.; Free, R.C.; et al. OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks. *Inf. Sci.* **2024**, *680*, 121141. https://doi.org/10.1016/j.ins.2024.121141.
21. Ren, S.; He, K.; Girshick, R.; et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149.
22. Li, W. Analysis of object detection performance based on Faster R-CNN. *J. Phys. Conf. Ser.* **2021**, *1827*, 012085. https://doi.org/10.1088/1742–6596/1827/1/012085.
23. Shuai, Q.; Wu, X. Object detection system based on SSD algorithm. In Proceedings of the 2020 International Conference

on Culture-oriented Science & Technology (ICCST), Beijing, China, 28–31 October 2020; pp. 141–144. https://doi.org/10.1109/ICCST50977.2020.00033.

24. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. *arXiv* **2017**, arXiv:1705.09587.
25. Zhu, Z.; Wang, S.; Zhang, Y. A Survey of Convolutional Neural Network in Breast Cancer. *Comput. Model. Eng. Sci.* **2023**, *136*, 2127–2172. https://doi.org/10.32604/cmes.2023.025484.
26. Ahmad, T.; Ma, Y.; Yahya, M.; et al. Object detection through modified YOLO neural network. *Sci. Program.* **2020**, *2020*, 8403262. https://doi.org/10.1155/2020/8403262.
27. Zhang, C.; Xu, X.; Tu, D. Face detection using improved faster rcnn. *arXiv* **2018**, arXiv:1802.02142.
28. Cheng, B.; Wei, Y.; Shi, H.; et al. Revisiting rcnn: On awakening the classification power of faster rcnn. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 453–468.
29. Zhu, Z.; Wang, S.; Zhang, Y. ROENet: A ResNet-Based Output Ensemble for Malaria Parasite Classification. *Electronics* **2022**, *11*, 2040. https://doi.org/10.3390/electronics11132040.
30. Qian, S.; Ning, C.; Hu, Y. MobileNetV3 for Image Classification. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 26–28 March 2021; pp. 490–497. https://doi.org/10.1109/ICBAIE52039.2021.9389905.
31. Koonce, B.; Koonce, B. MobileNetV3. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Apress: Berkeley, CA, USA, 2021; pp.125–144.
32. Zhao, L.; Wang, L. A new lightweight network based on MobileNetV3. *KSII Trans. Internet Inf. Syst. Korean Soc. Internet Inf. (KSII)* **2022**, *16*, 1–15.
33. Ren; Yun; Zhu, C.; Xiao, S. Object detection based on fast/faster RCNN employing fully convolutional architectures. *Math. Probl. Eng.* **2018**, *2018*, 3598316. https://doi.org/10.1155/2018/3598316.
34. Liu; Bin; Zhao, W.; Sun, Q. Study of object detection based on Faster R-CNN. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 6233–6236.
35. Jiang, H.; Learned-Miller, E. Face Detection with the Faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 650–657. https://doi.org/10.1109/FG.2017.82.
36. Kong; Xiaohong; Li, X.; Zhu, X.; et al. Detection model based on improved faster-RCNN in apple orchard environment. *Intell. Syst. Appl.* **2024**, *21*, 200325.
37. Cao, C.; Wang, B.; Zhang, W.; et al. An Improved Faster R-CNN for Small Object Detection. *IEEE Access* **2019**, *7*, 106838–106846. https://doi.org/10.1109/ACCESS.2019.2932731.
38. Rehman, M.U.; Shafique, A.; Ghadi, Y.Y.; et al. A Novel Chaos-Based Privacy-Preserving Deep Learning Model for Cancer Diagnosis. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 4322–4337.
39. Rehman, M.U.; Shafique, A.; Khan, I.U.; et al. An efficient deep learning model for brain tumour detection with privacy preservation. *CAAI Trans. Intell. Technol.* **2023**, 1–16. https://doi.org/10.1049/cit2.12254.
40. Rehman, M.U.; Shafique, A.; Khan, M.S.; et al. A novel medical image data protection scheme for smart healthcare system. *CAAI Trans. Intell. Technol.* **2023**, *9*, 821–836.