ISSN: 2710-0588

A Lightweight YOLOv5 Real-time Mask-wearing Detection Algorithm for the Post-pandemic Era

Xu Pan[1,2], Xiyin Liang[1,2], Zhen Ma[1,2], Pengfei Deng[1,2,*]

[1] Department of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China

[2] Gansu Province Intelligent Information Technology and Application Engineering Research Center, Lanzhou, Gansu 730070, China

[*]Corresponding author: 1908396445@qq.com

Abstract

Since the outbreak of the COVID-19 pandemic, standardized mask-wearing has become a powerful measure to combat the epidemic. Although the epidemic has been brought under control, vigilance in densely populated areas remains essential. Manual supervision is not only inefficient but also increases the risk of infection among relevant personnel. As a result, this paper proposes a lightweight real-time mask-wearing detection algorithm to monitor mask-wearing in crowds in real time. Built upon the YOLOv5 framework, the proposed algorithm replaces the backbone feature extraction network of the original model with an improved EfficientNetV2, reducing the model's parameter count and enhancing accuracy. The introduction of the ECA module in place of the SE module in the EfficientNetV2 network, coupled with the substitution of DIoU-NMS for the weighted NMS in the original model, further reduces model parameters and improves convergence. Additionally, this approach enhances the detection of occluded objects. Experimental results based on a publicly collected mask dataset demonstrate that the proposed algorithm reduces the model's parameter count by 44.7%, achieves a mAP of 95.3%, and attains an inference speed of 270.3 FPS. The algorithm introduced in this paper effectively identifies whether individuals are wearing masks correctly. Its lightweight nature makes it suitable for deployment on resource-constrained mobile devices, aligning well with post-pandemic epidemic prevention and control efforts in the era to come.

Keywords

COVID-19; Lightweight Mask-wearing Detection; Post-pandemic; YOLOv5.

1. Introduction

Since the outbreak of the COVID-19, it has severely impacted the global healthcare system, presenting unprecedented challenges to human society [1]. Wearing masks, as a simple yet effective protective measure, can significantly reduce the risk of droplet transmission and thereby lessen the potential for the spread of the pandemic. As of now, although domestic outbreaks have been largely controlled, the ever-changing and complex external environment continues to pose formidable challenges to epidemic prevention and control efforts. Particularly in the post-pandemic era [2], monitoring and managing mask-wearing in gatherings of individuals have become increasingly urgent. Traditional manual surveillance methods are evidently inadequate to cope with the intricacies of such a demanding task. Consequently, the application of target detection technology has become pivotal, underpinning the development of a real-time facial mask detection system, which holds paramount significance for epidemic prevention and control.

ISSN: 2710-0588

Target detection technology possesses the capacity to automatically recognize and localize specific objects within images or videos, offering precise data support to regulatory authorities and facilitating more effective monitoring and management of mask-wearing compliance in public spaces [3-5]. Among various target detection techniques, the You Only Look Once (YOLO) algorithm has garnered substantial attention due to its high speed and accuracy. By transforming the target detection task into a regression problem, the YOLO algorithm achieves simultaneous object localization and classification within a single forward pass, significantly enhancing detection efficiency [6].

Addressing the issues of the YOLOv5 backbone feature extraction network's excessive parameter count and insufficient feature extraction, this study proposes an improved lightweight YOLOv5 real-time mask-wearing detection algorithm, catering to the pressing need for mask-wearing monitoring in the post-pandemic context [7]. The main improvements include:

1) Substituting the Efficient NetV2 lightweight network to reduce model parameter count and markedly enhance training speed.

2) Introducing ECA modules to replace the original SE attention modules in Efficient NetV2, further reducing model parameter count and improving convergence.

3) Introducing DIoU-NMS to replace the original YOLOv5 model's weighted NMS approach, enhancing detection of occluded targets and demonstrating superior performance in eliminating redundant prior boxes.

2. Related Study

2.1. Object Detection

Object detection is one of the crucial research directions in the field of computer vision. Its primary task is to identify all relevant objects in an input image based on user-defined target information, and then analyze the objects' categories and spatial positions. It has been widely applied across various domains, including autonomous driving, remote sensing imagery, video surveillance, and medical diagnostics.

Traditional object detection methods leverage machine learning algorithms, using sliding windows of different scales to determine candidate regions for potential objects. Features such as histogram of oriented gradients (HOG) [8], scale invariant features transform (SIFT) [9], local binary pattern (LBP), Haar-like features (Haar) [10] and more are extracted from these regions. These extracted image features are then fed into classifiers like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, and Adaboost for object classification.

However, these traditional object detection methods come with inherent limitations. The sliding window-based region selection algorithm lacks specificity for different feature targets, resulting in high time complexity and excessive window redundancy. Extracted image features are essentially hand-crafted, and their quality significantly impacts detection performance, requiring researchers to tailor features to the specific task. Moreover, machine learning-based object detection methods involve a complex process, from low-level image feature extraction to performance feature extraction, heavily reliant on the design of manual features.

In contrast to traditional machine learning approaches, deep learning-based object detection algorithms utilize deep neural networks to directly perform convolution and pooling operations on images, thereby extracting essential image features. The development of deep learning, particularly the emergence of Convolutional Neural Network (CNN), offers entirely new solutions for image feature extraction. CNN are locally connected networks where small regions of an image are selected as input data for each layer. Information propagates through the network layers via forward propagation, each of which consists of filters designed to

ISSN: 2710-0588

capture prominent features of the observed data. The close interconnections and spatial information between layers make CNN especially suited for image feature extraction tasks [11].

2.2. YOLO

Up to now, in the field of object detection, there are mainly two types of approaches: one-stage algorithms based on regression ideas and two-stage algorithms based on region proposals. Among the two-stage algorithms, Fast-CNN and Faster-CNN stand out as representatives with high detection accuracy. However, their model complexity leads to slower detection speeds. On the other hand, one-stage algorithms represented by the YOLO series and SSD series have significantly improved detection speeds, but their detection accuracy is relatively lower [12].

With the rapid development of deep learning, an increasing number of object detection algorithms are being deployed on embedded devices to achieve object detection. This demands algorithms to have both fast detection speeds and high detection accuracy. The YOLO series is one of the frequently used and mature object detection algorithms. Through extensive optimization and validation by numerous researchers, this algorithm has gradually bridged the gap between increased detection speed and decreased detection accuracy, achieving a balance between detection speed and maintaining relatively high detection accuracy.

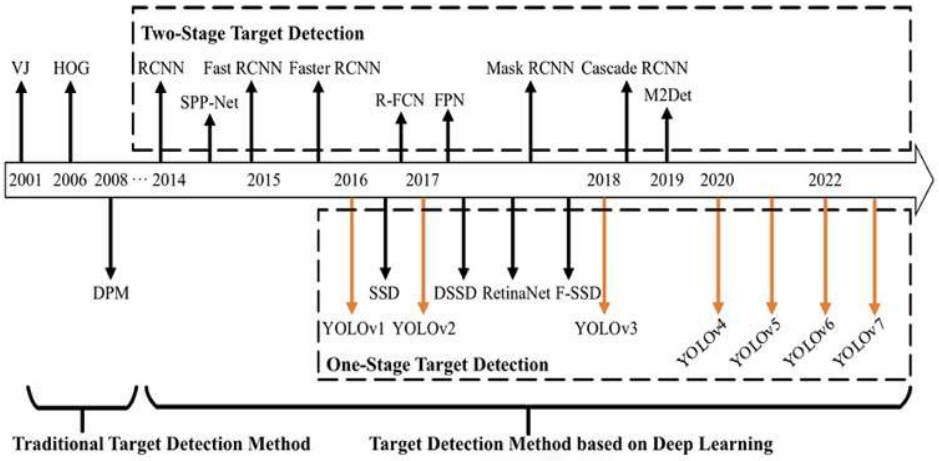

Figure 1. Development history of target detection

As shown in Figure 1, since the introduction of RCNN in 2014, deep learning-based methods have replaced traditional methods and become the mainstream for object detection. Initially, deep learning-based object detection employed a two-stage strategy, explicitly dividing the detection process into candidate region selection and target region determination. While achieving high detection accuracy, this approach had slower detection speeds. In 2016, single-stage object detection methods represented by YOLOv1 emerged. These methods do not require pre-extraction of candidate regions but directly predict the class probabilities and positions of target objects, significantly reducing computational resource consumption and improving detection speed, thus gaining more attention.

Reviewing the development history of single-stage object detection methods, from the inception of the first single-stage method, YOLOv1, to YOLOv7 in 2022, the YOLO series of object detection methods have evolved in tandem with the progress of single-stage object detection, becoming typical representatives of the One-Stage approach. In fact, considering the high detection efficiency, modular framework design, and ease of improvement of YOLO algorithms, not only have the foundational models of YOLOv1 through YOLOv7 appeared in single-stage object detection methods, but also numerous improved models based on YOLO have emerged and been applied across various domains [13].

ISSN: 2710-0588

2.3. YOLOv5

YOLO is a one-stage algorithm based on regression principles, redefining object detection as a regression problem to enhance detection speed [14]. YOLOv5 is the latest algorithm introduced in the YOLO series. Compared to YOLOv3 and YOLOv4, YOLOv5 maintains accuracy while also having a smaller structure, lighter weight, and being more suitable for deployment on mobile devices. As illustrated in Figure 2, the network architecture of YOLOv5 consists of three parts: Backbone, Neck, and Head [15].

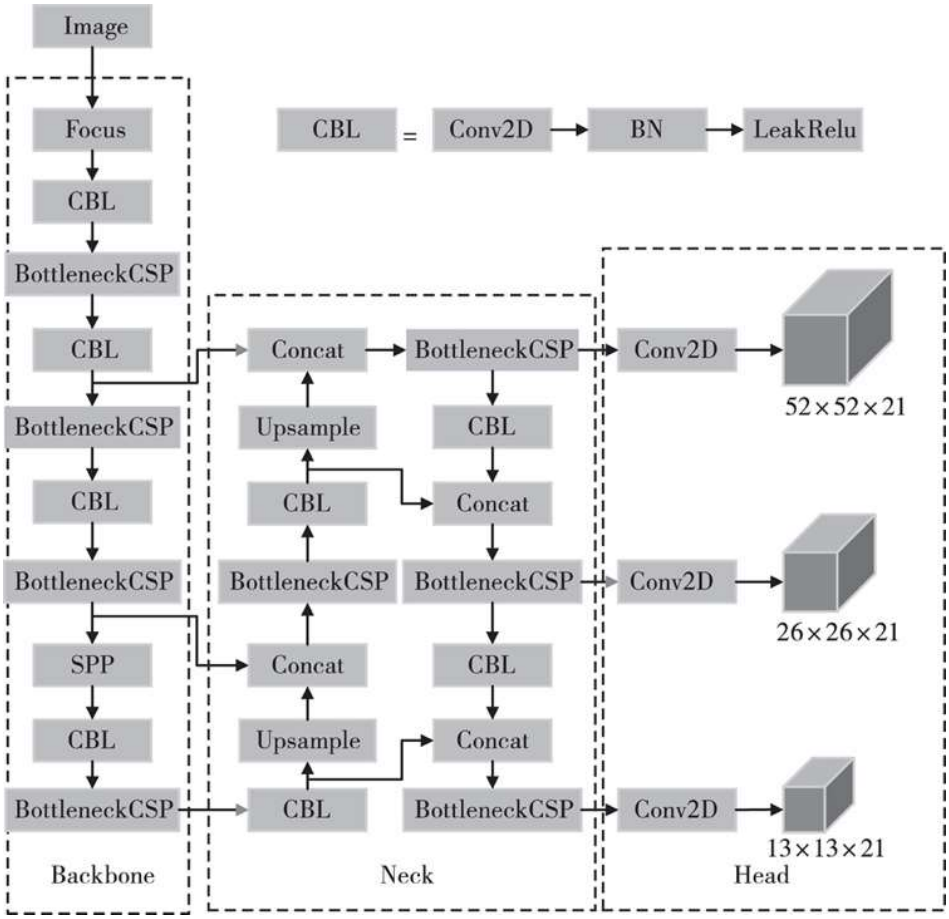

Figure 2. YOLOv5 system framework

In the backbone network, the input image resolution is 416×416×3. Through the Focus structure, it undergoes slicing operations, resulting in a 208×208×12 feature map. Then, it undergoes convolution operations with 32 convolution kernels, ultimately becoming a 208×208×32 feature map. The CBL module comprises Conv2D, Batch Normalization, and LeakyRELU layers. The Bottleneck CSP module primarily focuses on feature extraction from the feature map, extracting rich information from the image. Compared to other large-scale convolutional neural networks, the Bottleneck CSP structure reduces the repetition of gradient information during network optimization. It occupies a significant portion of the network's parameters. By adjusting the width and depth of the Bottleneck CSP module, four models with different parameters can be obtained: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x.The SPP module mainly increases the network's receptive field to capture features of different scales. YOLOv5's Neck section utilizes PANet as its feature fusion network, effectively leveraging detailed feature information to enhance the network's detection capabilities for targets of different scales.

ISSN: 2710-0588

3. Lightweight YOLOv5

3.1. The Improved EfficientNetV2 Network

In 2019, Tan Mingxing and colleagues introduced a straightforward and efficient network model structure called EfficientNet [16]. By employing composite coefficients to achieve scaling of a multi-dimensional mixed model, this approach balances the scaling rates (d, r, w) across the dimensions of network depth, width, and image resolution. This method enhances the network's performance, resulting in improved capabilities [17].

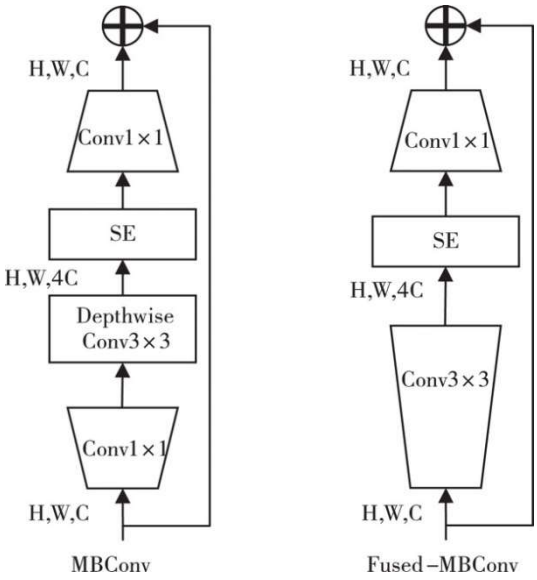

Figure 3. MBConv and Fused-MBConv structure diagram

EfficientNet V2 is a new model proposed by the authors as an improvement upon the EfficientNet architecture. This model effectively addresses three issues encountered during the training process of EfficientNet: slow training speed when working with large-sized training images, slow training speed when utilizing DepthwiseConvolution in shallow networks, and suboptimality in equal amplification of each stage. To tackle these challenges, EfficientNet V2 introduces the Fused-MBConv module and an improved progressive learning method. This method dynamically adjusts regularization techniques based on the size of the training images.Compared to EfficientNet, EfficientNet V2 achieves an 11-fold increase in training speed and reduces the model's parameter count by 6.8 times. Figure 3 illustrates the structural diagrams of the MBConv and Fused-MBConv modules. In the Fused-MBConv module, a 3x3 convolution replaces the 1x1 convolution and Depthwise Convolution 3x3 in the MBConv module. This substitution is made because Depthwise Convolution does not fully leverage the model for acceleration. Incorporating the Fused-MBConv module into the shallow network structure leads to improvements in training speed, accuracy, and floating-point computation.

In this paper, both the MBConv module and the Fused-MBConv module within the EfficientNet V2 network architecture incorporate the SE (Squeeze-and-Excitation) channel attention mechanism. However, within this mechanism, dimension reduction is crucial for learning channel attention, necessitating appropriate cross-channel interactions to maintain performance while reducing model complexity.Therefore, this paper proposes the use of a non-reducing local cross-channel interaction strategy called the ECA (Efficient Channel Attention) module to replace the original SE modules in both the MBConv module and the Fused-MBConv module, as depicted in Figure 4 [18].

ISSN: 2710-0588

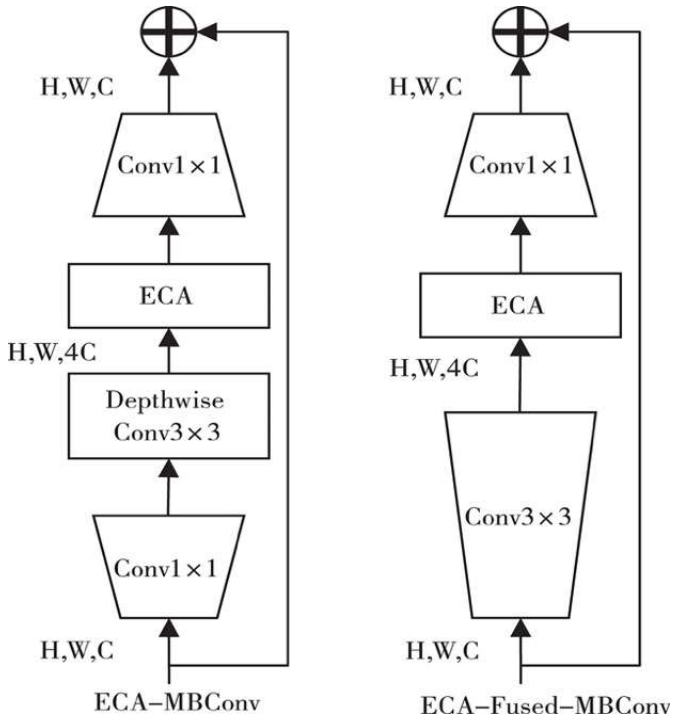

Figure 4. Improved MBConv and Fusion-MBCONV structure diagrams

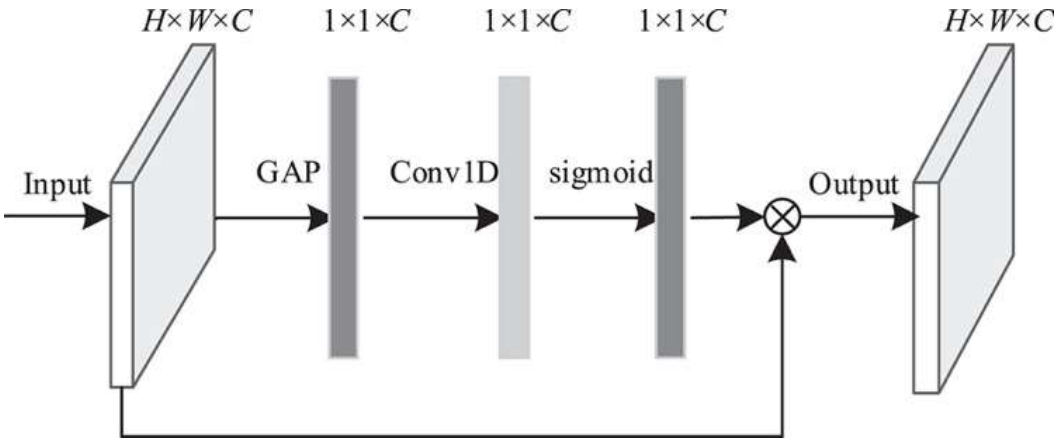

Figure 5. ECA structure diagram

The ECA module enables per-channel global average pooling without reducing the channel dimensionality [19]. It captures local cross-channel interaction information by considering each channel and its neighboring K channels, preventing features from becoming entirely independent. The structure of the ECA module is illustrated in Figure 5.

3.2. Improved YOLOv5 Architecture

This paper proposes an improved network architecture for YOLOv5, specifically designed for lightweight applications, as illustrated in Figure 6. The input image size is 416x416x3. The backbone of this structure employs the enhanced EfficientNet V2 network as the feature extraction network. In the neck section, the paper replaces the original Bottle Neck CSP unit with a C3 unit [20]. The C3 unit is a deeper residual structure composed of CBL layers, as depicted in Figure 7 [21]. The Concat and Add layers are used for tensor concatenation of CBL modules and upsampling layers, enhancing feature reuse.

ISSN: 2710-0588

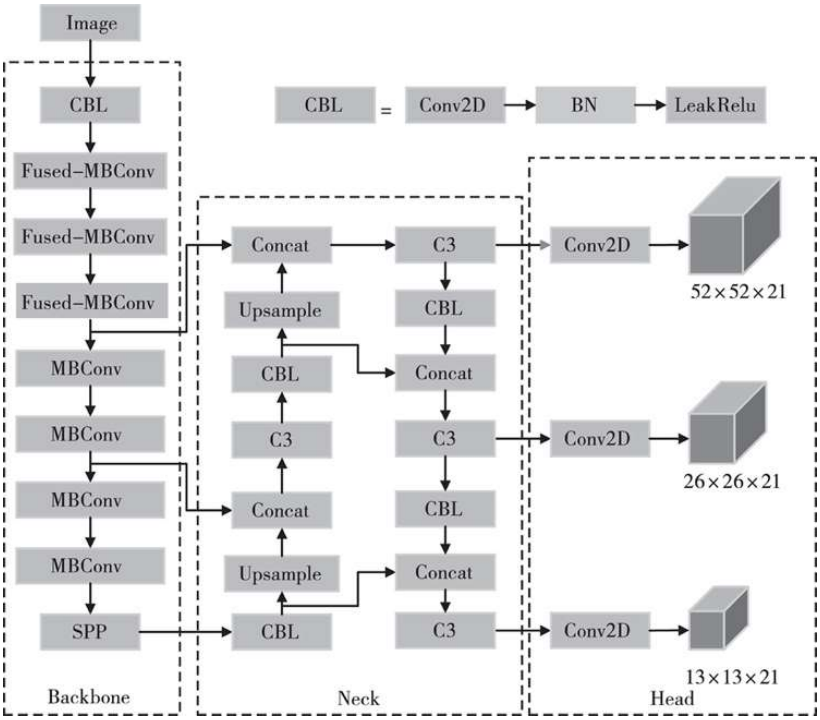

Figure 6. Improved YOLOv5 network structure diagram

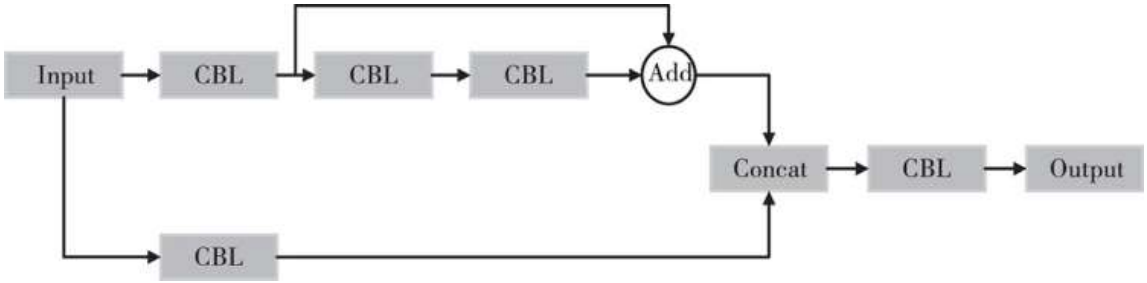

Figure 7. C3 cell structure diagram

4. Experimental Result

4.1. Data Set

Data acquisition methods mainly include: smartphone photography, web photo collection, web open-source datasets, and creating a homemade dataset for face masks. An open-source dataset provided by GitHub-AIZOOTech is also used to supplement the data.

Regarding data preparation, since YOLOv5 uses Mosaic data augmentation at its input, four images from the dataset can be randomly scaled and stitched together, greatly enriching the existing dataset. On the other hand, as the plan involves training the target model using a pre-trained model, the demand for a high quantity of self-made dataset is not high. The final dataset contains a total of 2000 original images [22].

In terms of data format, to facilitate the uniformity of images from different sources, even though the YOLOv5 model is used in the algorithm design, the homemade original images were annotated in VOC format for ease of data format standardization. The annotation scope covers the entire head of the person, including a small portion of the upper shoulders. Data format conversion is handled within the program code. The use of VOC format also allows this dataset to be potentially used with other models in the future.

Regarding data preprocessing, adjustments were made to image sizes. In most object detection algorithms, original images need to be resized, padded, and standardized to a specific size

ISSN: 2710-0588

before being input to the network. This process introduces a certain amount of information redundancy and increases training time. In theory, YOLOv5 is less constrained by the input image size. However, different input image sizes lead to variations in network width, which in turn affects the size of the model file. Since the plan involves using pre-trained model parameter files, the input image size is set to 640x640 in the program. This size is the same as the input image size used by the YOLOv5 authors for their pre-trained model on the COCO dataset.

4.2. Model Construction and Training

Due to hardware limitations and in order to shorten training time while enhancing model generalization, this model underwent secondary training based on a pre-trained model. The official YOLOv5 code provides four network models, ranging from smallest to largest: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Considering practical task requirements and hardware environment conditions, the downloaded official YOLOv5s code (weight files) was used as the pre-trained model, upon which training was performed using a self-made dataset. The model training employed a self-made dataset comprising a total of 2,000 samples, divided into training, testing, and validation sets. Among these, 1,200 samples were allocated to the training set, 400 samples to the testing set, and 400 samples to the validation set. Training was conducted over 100 iterations on an image processing workstation, with ACC and Cost intuitively displayed after each training iteration. The final trained model was saved. The results of the model training are depicted in Figure 8.

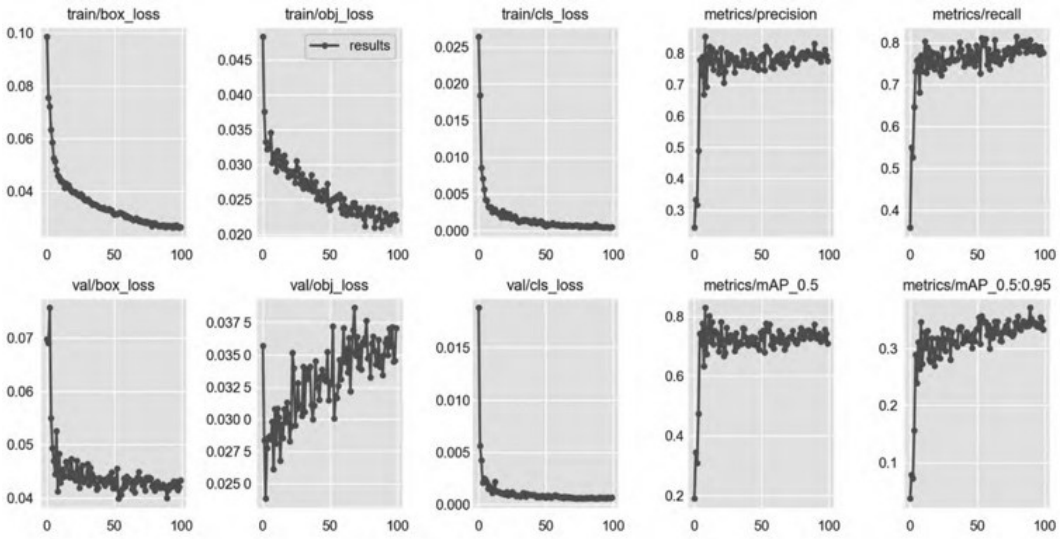

Figure 8. Model training results

4.3. Model Evaluation

As shown in the above graph, with an increase in the number of training iterations, the loss steadily decreases and stabilizes, reaching the desired values. Precision [23] and recall [24] also increase continuously as the iterations progress, stabilizing around 0.8 and 0.79 respectively. The fluctuation of their values at different thresholds is relatively stable. This suggests that the model possesses good precision and stability.

To further analyze the model's performance, testing was conducted on a validation set containing 400 samples. The model performed well on the validation set. Analyzing the source graphs and the confusion matrix in Figure 9, it can be inferred that the model demonstrates strong face detection capabilities, but there is a tendency to misclassify masked faces as background factors. To some extent, this might be due to factors such as low image pixel quality, small size of masked faces within an image, or improper mask wearing.

ISSN: 2710-0588

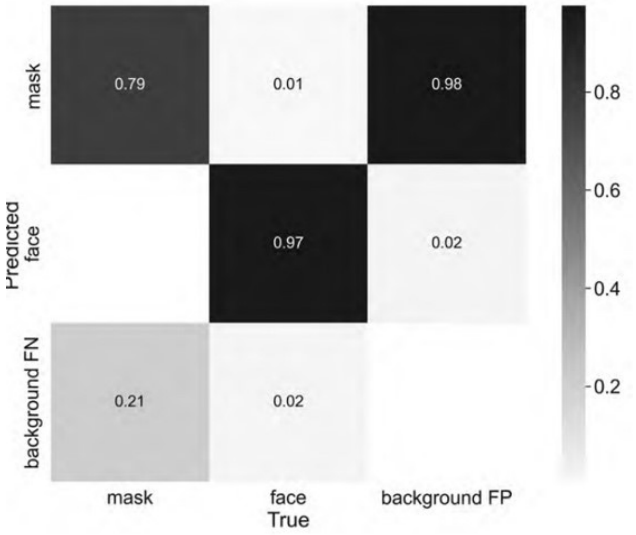

Figure 9. Confusion matrix

5. Conclusion

This paper proposes a real-time mask detection algorithm based on the lightweight YOLOv5 network. The approach involves replacing the original YOLOv5 backbone feature extraction network with an improved EfficientNetV2, substituting the SE channel attention module with the ECA channel attention module, and replacing the weighted NMS method in the original YOLOv5 model with DIoU-NMS. These modifications reduce the network model's parameter count, enhance model convergence, improve accuracy, and enhance the detection of occluded targets.The experimental results, conducted on both publicly available and natural scene mask datasets, demonstrate that the proposed lightweight mask detection algorithm achieves a mAP of 95.3% and an inference speed of 270.3 FPS while reducing the parameter count by 45.7%. Compared to other lightweight networks, the algorithm introduced in this paper maintains high accuracy while reducing parameter count, making it well-suited for deployment on resource-constrained mobile devices. This algorithm effectively identifies proper mask usage, contributing to effective post-pandemic mask-wearing surveillance. In future research, there is potential to integrate this approach with relevant hardware for intelligent detection in various applications such as access control systems and production security.

Acknowledgments

This work was supported in part by a grant from Education technology innovation project of Gansu Province (2021CYZC-22).

References

[1] Pastor L, Vorsatz B.Mutual Fund Performance and Flows During the COVID-19 Crisis[J]. Social Science Electronic Publishing [2023-08-10].

[2] Javakhishvili I.Covid-19-Pandemic Measures in Conflict Zones in 2020 and 2021—The Case of the OSCE and South Ossetia in Georgia[M]. 2021.

[3] Sauter,Marian,Liesefeld,et al.Region-based shielding of visual search from salient distractors: Target detection is impaired with same- but not different-dimension distractors[J].Attention Perception & Psychophysics, 2018.

[4] Wu,Zhihuan,Chen,et al.Rapid Target Detection in High Resolution Remote Sensing Images Using YOLO Model[C]//2018.

ISSN: 2710-0588

[5] [1] Lin C , Chen S Y , Chen C C ,et al.Detecting Newly Grown Tree Leaves from Unmanned-Aerial-Vehicle Images using Hyperspectral Target Detection Techniques[J].ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 142(AUG.):174-189.

[6] Dlamini S , Kao C Y , Su S L ,et al.Development of a real-time machine vision system for functional textile fabric defect detection using a deep YOLOv4 model[J].Textile research journal, 2022(5/6):92.

[7] Majeed F , Khan F Z , Nazir M ,et al.Investigating the efficiency of deep learning based security system in a real-time environment using YOLOv5[J].Sustainable Energy Technologies and Assessments, 2022.

[8] Hu L , Liu W , Li B ,et al.Robust motion detection using histogram of oriented gradients for illumination variations[C]//International Conference on Industrial Mechatronics & Automation. IEEE, 2010.

[9] Pérez, Diego Sebastián, Bromberg F , Diaz C A .Image classification for detection of winter grapevine buds in natural conditions using scale-invariant features transform, bag of features and support vector machines[J].Computers and Electronics in Agriculture, 2017, 135:81-95.

[10] Mulyawan I T .Two wheeled Vehicles detection Using Haar-like Features and Support Vector Machine on CCTV Traffic Image[J]. 2019.

[11] Seddati O , Dupont S , Mahmoudi S ,et al.Towards Good Practices for Image Retrieval Based on CNN Features[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[12] Calakli F , Taubin G .SSD: Smooth Signed Distance Surface Reconstruction[J].Computer Graphics Forum, 2011, 30(7):1993-2002.

[13] A S S R , B S K V , C Y K .Review on recent development in infrared small target detection algorithms[J].Procedia Computer Science, 2020, 167:2496-2505.

[14] Envelope J C A , A E B , A J P .Classification and Positioning of Circuit Board Components Based on Improved YOLOv5[J].Procedia Computer Science, 2022, 208:613-626.

[15] Darapaneni N , Kumar S , Krishnan S ,et al.Implementing a Real-Time, YOLOv5 based Social Distancing Measuring System for Covid-19[J]. 2022.

[16] Tan M , Le Q V .EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[J]. 2019.

[17] Tan M , Le Q V .EfficientNetV2: Smaller Models and Faster Training[J]. 2021.

[18] Yang Q , Wei M , Zhu R ,et al.Facial expression recognition algorithm based on efficient channel attention[J].Journal of electronic imaging, 2022.

[19] Hartweg B , Fisher K , Niverty S ,et al.Analysis of electrically conductive adhesives in shingled solar modules by X-ray imaging techniques[J].Microelectronics and reliability, 2022.

[20] Karthik R , Vaichole T S , Kulkarni S K ,et al.Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification[J].Biomedical Signal Processing and Control, 2022, 73:103406-.

[21] Cailing Z , Shihua L , Bo H ,et al.Relationship between the Convective Boundary Layer and Residual Layer over Badain Jaran Desert in Summer[J].Plateau Meteorology, 2016.

[22] [1] Fang S , Mu L , Jia S ,et al.Research on sunken & submerged oil detection and its behavior process under the action of breaking waves based on YOLO v4 algorithm[J].Marine pollution bulletin, 2022(Jun.):179.

[23] Wang G , Ding H , Li B ,et al.Trident-YOLO: Improving the precision and speed of mobile device object detection[J].IET image processing, 2022(1):16.

[24] Yang F .A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5[J].Remote Sensing, 2021, 13.