Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241160

Enhancing mask detection performance based on YOLOv5 model optimization and attention mechanisms

Guangyuan Yang

The Department of Software engineering, Hubei University of Technology, Wuhan municipality, 430068, China

2011631119@hbut.edu.cn

Abstract. Due to the COVID-19 pandemic, there has been a significant increase in the usage of masks, leading to more complex scenarios for mask detection techniques. This paper focuses on optimizing the performance of mask detection using the You Only Look Once (YOLO) v5 model. In this study, the yolov5 target detection model was employed for training the mask dataset. Diverse model improvement techniques were explored to enhance the model's capability to capture crucial features and differentiate masks from the background in complex scenarios. Finally, the modified model was compared with the earlier original target detection model to identify the most considerable performance gain. The CSPDarknet design with the TensorFlow framework is utilized in this study, and the Attention Mechanism module is implemented through the Keras library. The objective is to optimize the three feature layers between the backbone network and the neck by integrating multiple attention mechanisms. This will enable the model to more quickly and accurately capture important features when dealing with complex scenarios by adjusting the feature map weights. Additionally, in the feature pyramid network, shallow feature maps are fused with deeper feature maps in a certain order to determine the most efficient feature fusion method. Finally, this study identified the optimal combination of attention mechanism and feature fusion through ablation experiments. The results of the experiment demonstrate that the combination of SE block and shallow feature fusion (SE + FF2 model) can greatly enhance category confidence, leading to an improved model performance.

Keywords: Mask Detection, YOLOv5, Attention Mechanism, Feature Fusion.

1. Introduction

The rapid progress in deep learning and computer vision has presented promising prospects for target detection technology. Nonetheless, as artificial intelligence extends its applications, target detection should also adapt to increasingly diverse and intricate usage scenarios. The COVID-19 pandemic has devastated the globe, highlighting the significance of face coverings for daily protection [1]. Although the storm of the epidemic has weakened, wearing masks remains an efficient approach to prevent illnesses. Detecting masks automatically in public or crowded areas may significantly increase adherence to mask requirements, thereby mitigating the transmission of the virus.

As a significant field of target detection technology, You Only Look Once (YOLO)v5 locates and identifies objects by learning their features in images or videos. It excels at detecting small or intricate

DOI: 10.54254/2755-2721/50/20241160

objects with several benefits such as real-time and high-speed performance, accurate multi-scale detection, lightweight design, and intuitive usability and good compatibility with deep learning [2]. In the early YOLO mask detection models [3], the focus was mainly on local features, lacking the fusion of global information. This made it challenging for the model to distinguish the relationship between the mask and other backgrounds accurately, resulting in inaccurate key feature extraction. In [4], Sheng Xu utilized a novel backbone network architecture referred to as Shuffle Coordinate Attention Network (CANet). This structure integrates the ShuffleNetV2 network alongside the coordinate attention mechanism, resulting in a considerable enhancement of the image segmentation performance. Given these findings, I have opted to implement the Cross Stage Partial Darknet (CSPDarknet) as the network structure for the feature fusion approach in the original baseline model. The mask dataset was trained, and experiments revealed that the algorithm encounters occasional challenges, including sudden confidence drops and the inability to capture significant features in the mask detection task. As such, the model's accuracy needs further enhancement. Therefore, this research aims to optimize the YOLOv5 model to improve mask detection performance and obtain higher accuracy [2].

This paper investigates the training of the mask dataset using the yolov5 target detection model. This study explored multiple model modification methods for deeper understanding. In [5], Vaswani introduced a new neural network design employing the attention mechanism, which displayed favorable results in terms of performance, parallelization, and shorter training duration. This paper aims to enhance the original CSPDarknet structure by optimizing the three feature layers situated between the backbone network and the neck, taking into account the research direction of [5]. The first step in this process involves integrating various attention mechanisms such as Convolutional Block Attention Module(CBAM), Squeeze-and-Excitation(CE), and Efficient Channel Attention(ECA) to enhance the model's sensitivity to essential features. Furthermore, through a comparative analysis of experimental results, this study explored various feature fusion techniques and identified the optimal combination of attention mechanisms and feature fusion, which resulted in achieving greater accuracy than the findings presented in [4]. These attention mechanisms effectively aid the model in accurately locating the features associated with the mask, thereby enhancing the model's performance.

2. Method

2.1. Dataset preparation

The dataset utilized for the experiment is a binomial classification dataset for mask detection, downloaded from China Software Developer Network (CSDN), a professional IT information exchange platform in China. Since the dataset has been preprocessed and improved [6]. The dataset includes a total of 5,508 images, which have been split into a training set and a validation set at a ratio of 9:1. The training set consists of 2,692 masked faces and 9,475 unmasked faces, whereas the validation set has 9,475 unmasked faces. The images are entered into the model in RGB format with a size of 640x640 pixels for training.Figure 1 and Figure 2 depict the example datasets employed in the experiments, featuring images of individuals wearing masks and those without.

Figure 1. Images of faces with masks [6].

DOI: 10.54254/2755-2721/50/20241160

Figure 2. Images of faces without masks [6]

2.2. YOLO v5 model

YOLOv5, the fifth-generation target detection model of the YOLO series, incorporates a backbone, a Feature Pyramid Network (FPN), and a classification regression head (YOLO Head) [2]. The Backbone network employs the CSPDarknet design, consisting of a residual network, CSPnet structure, Focus network structure, Sigmoid Linear Unit(SiLU) activation function, and Spatial Pyramid Pooling(SPP) structure. The feature pyramid incorporates the FPN network to achieve up-sampling and down-sampling, thereby enabling the fusion of various features. The experiment in this study aims to enhance each of the aforementioned aspects and identify the model with the superior performance. In the initial stage of the experiment, the backbone network generates feature maps called feature01, feature0, feature1, feature2, and feature3. To enhance the network's ability to capture critical features, this study introduced the Squeeze-and-Excitation (SE) block According to Figure 3, which adaptively adjusts the weight of the feature maps. Next, this study advances to the initial training phase. Within the feature pyramid segment, the FPN network is adapted to integrate different feature maps (P5 and feature2) from the backbone network. This integration is achieved through convolution and upsampling of the P5 feature map with the feature2 feature map. Similarly, the P4 feature map is fused with the feature1 feature map, whereas the P3 feature map is downsampled and fused with the feature3 feature map, producing more coherent feature maps. The latter half of model training is then conducted. As shown in Figure 4, the attention mechanism and feature fusion techniques introduced in this study undergo ablation experiments to determine the method of pairing that achieves the best performance.

2.3. Implementation details

The experimental setup utilizes a NVIDIA GeForce RTX 3090 graphics card with 24GB of video memory. The TensorFlow [7] and Keras [8] libraries are employed for model selection and implementation of the attention mechanism module [5] which includes SE-Block [9, 10], Channel Attention [11], Spatial Attention[11], CBAM Block[12], ECA Block[13], and CA Block[14]. The optimizer used is Stochastic Gradient Descent (SGD) [15], with the aim of optimizing model performance. The initial learning rate is set to 0.01. To achieve improved convergence during training, a gradual decrease in learning rate based on the number of rounds was employed. To ensure model performance, two loss functions were used: the Complete Intersection over Union (CIoU)[16] loss function for bounding box[17] and the Binary Cross-Entropy (BCE)[18] loss function for target existence and classification. Clear and concise language was used throughout, with causal connections between statements. Technical term abbreviations were explained when first used, allowing for greater understanding by readers. An excellent loss function enables the model to comprehend and record the target's position more fluidly, thereby securing the bounding box's accuracy and category classification. To obtain a more comprehensive and precise comparison of model performance, this study opted for measuring performance from several perspectives, including comparison mAP, AP, Precision, and Recall.

DOI: 10.54254/2755-2721/50/20241160

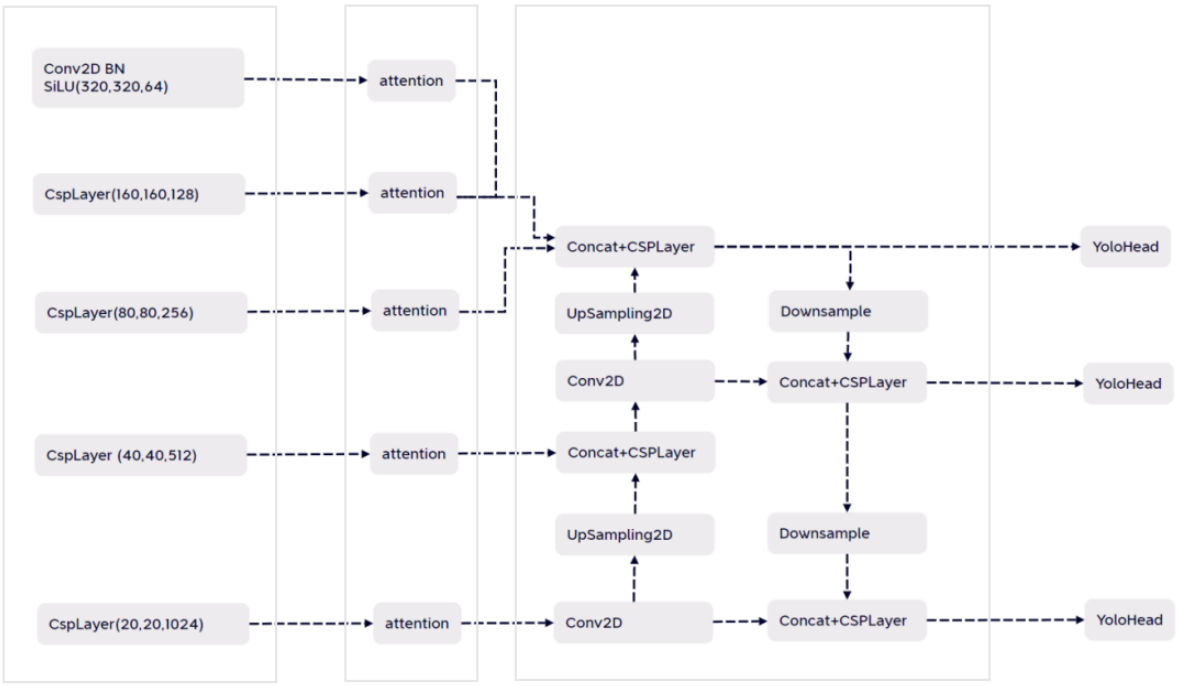

Figure 3. Adding attention mechanism (Photo/Picture credit: Original)

3. Results and discussion

Through attention comparison experiments, it was determined that the SE block [9] has a more prominent enhancing effect on the mask detection model [4], as depicted in Figure 4 (mAP is utilized for initial comparison). Specifically, the introduction of the SE block [9] has a more significant effect on average accuracy (mAP) metrics, where the average accuracy (AP) of both categories (mask-wearing faces and unmasked faces) is improved. In the two feature fusion experiments, the model's feature extraction ability is only significantly improved when shallow features are introduced (as shown in Figure 4). This is due to the model's ability to acquire more contextual information. The results of the experiments are summarized in Figure 5. In the backbone network's convolutional layer, the shallow layer detects small targets while the deep layer detects large targets, leading to improved performance in both face and face mask categories. Results demonstrated that the SE block [9] and shallow feature fusion (SE + FF2 model) resulted in significant performance enhancements of 3.15%, as depicted in Table 1, after conducting two experiments and ablation experiments. To better demonstrate the model's improvement, one can compare its graphs in Table 1, Table 2 and Table 3. A comparison of Figure 6 and Fiuge 7 shows that, although the difference in detection position is minor, the enhanced model's bounding box detection shows significantly improved category confidence. This leads to the conclusion that the model combining feature fusion and attention mechanism is the most effective.

However, the increase in the number of parameters in the model due to the rise in feature fusion and attention mechanism leads to greater computational demand, which inevitably results in a loss either in space or time, causing a decrease in detection speed index (fps). Nonetheless, this experiment solely focuses on image detection, and thus, does not provide a direct assessment of the impact on fps index. Moving forward, it is necessary to explore a more lightweight algorithmic structure.

DOI: 10.54254/2755-2721/50/20241160

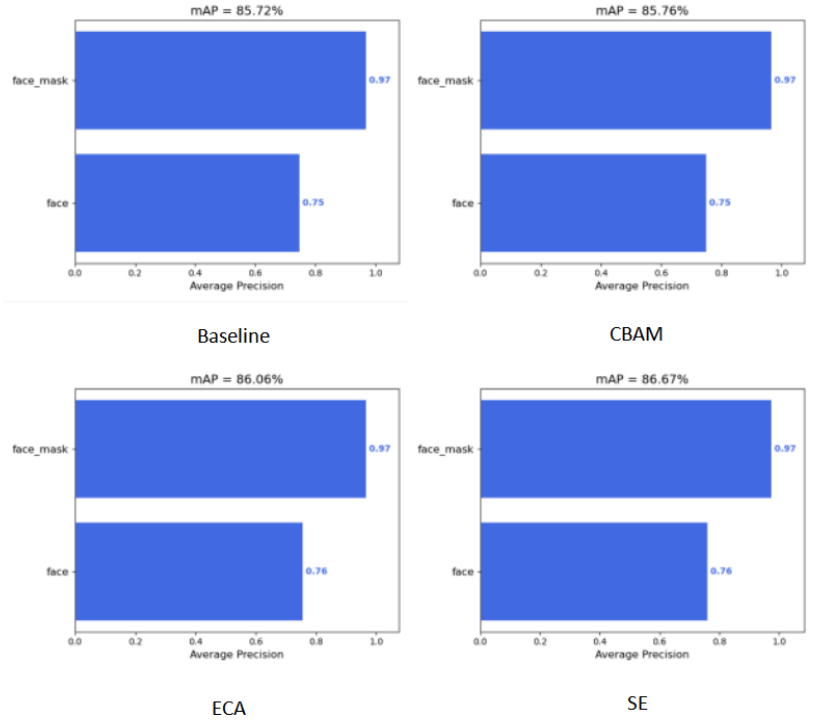

Figure 4. Comparison of training results after adding attention mechanism (Photo/Picture credit : Original)

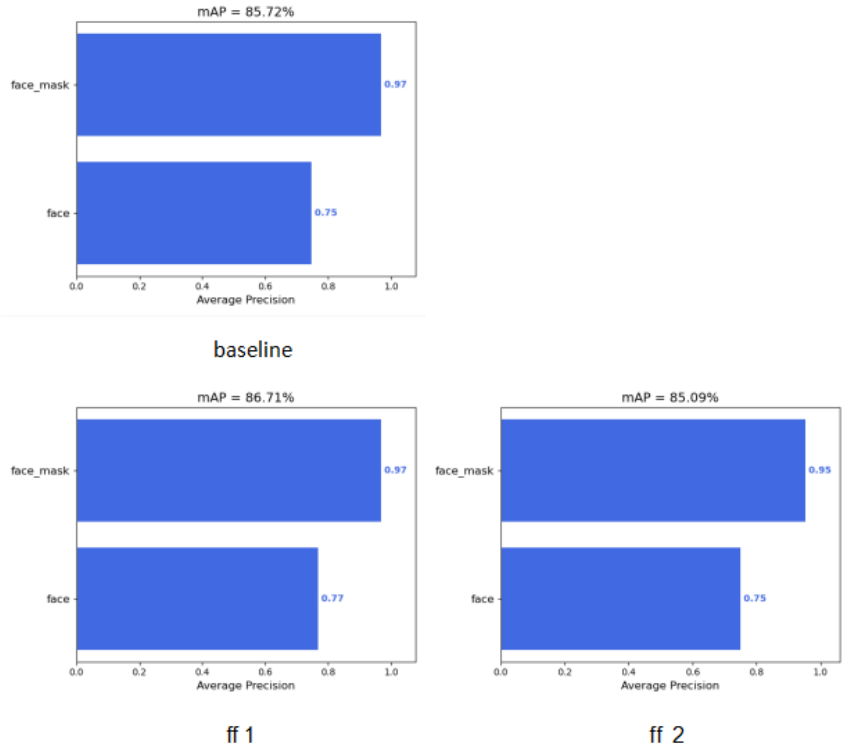

Figure 5. Comparison of training results after feature fusion (Photo/Picture credit: Original)

DOI: 10.54254/2755-2721/50/20241160

Table 1. Comparative experiments on attention mechanisms.

Method	class	mAP	AP	Precision	Recall
baseline	face	85.72	74.69	93.99	64.69
	face_mask		96.74	97.05	89.76
cbam	face	85.76	75.01	93.28	64.87
	face_mask		96.52	97.01	88.74
eca	face	86.06	75.57	94.89	65.05
	face_mask		96.55	96.31	89.08
se	face	**86.67**	76.01	93.97	64.42
	face_mask		97.32	97.04	89.42

Table 2. Comparative experiments on feature fusion

method	class	mAP	AP	Precision	Recall
baseline	face	85.72	74.69	93.99	64.69
	face_mask		96.74	97.05	89.76
ff1	face	85.09	74.94	93.38	64.6
	face_mask		95.24	95.26	89.08
ff2	face	**86.71**	76.7	94.02	64.96
	face_mask		96.71	96.67	89.08

Table 3. Ablation Experiments on Attention Mechanism and Feature Fusion

method	class	mAP	AP	Precision	Recall
baseline	face	85.72	74.69	93.99	64.69
	face_mask		96.74	97.05	89.76
+se	face	86.67	76.01	93.97	64.42
	face_mask		97.32	97.04	89.42
+ff2	face	86.71	76.7	94.02	64.96
	face_mask		96.71	96.67	89.08
+se+ff2	face	**88.87**	80.39	93.77	67.57
	face_mask		97.35	97.79	90.44

DOI: 10.54254/2755-2721/50/20241160

Figure 6. Detection box confidence for the baseline (Photo/Picture credit: Original)

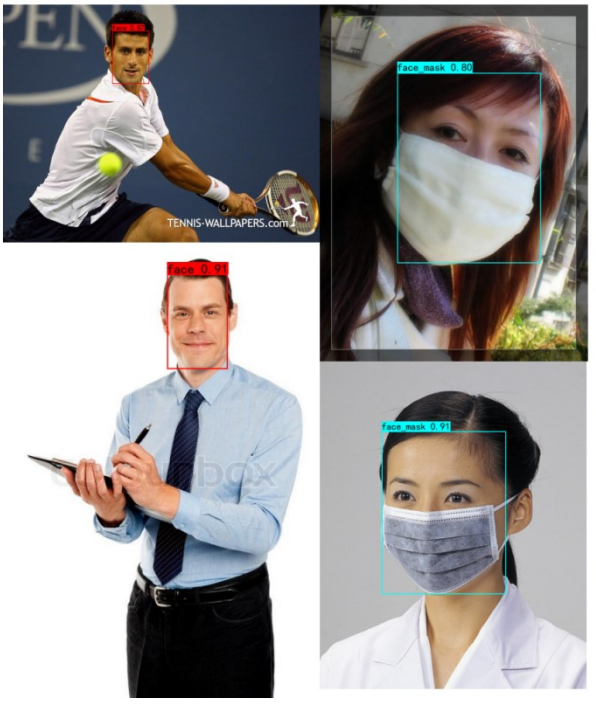

Figure 7. Detection box confidence for SE+FF2 model (Photo/Picture credit: Original)

4. Conclusion

This paper primarily aims to enhance mask detection performance using the You Only Look Once (YOLO) v5 model. The yolov5 target detection model was selected for training the mask dataset, and various model improvement techniques were explored to improve its ability to differentiate masks from complex backgrounds and capture essential features. The study also compared the modified

DOI: 10.54254/2755-2721/50/20241160

model with the original target detection model to determine the most significant performance improvements. In this research, the CSPDarknet design within the TensorFlow framework was adopted, and the Attention Mechanism module was integrated using the Keras library. The main objective was to optimize the three feature layers between the backbone network and the neck by incorporating multiple attention mechanisms. This adjustment enables the model to more swiftly and accurately identify crucial features in complex scenarios by adapting feature map weights. Additionally, the study investigated the most efficient feature fusion method by fusing shallow and deep feature maps in a specific order within the feature pyramid network. A series of ablation experiments helped identify the optimal combination of attention mechanisms and feature fusion. The results revealed that the combination of the SE block and shallow feature fusion (SE + FF2 model) significantly improved category confidence and overall model performance.

References

[1] Andrejko K L 2022 COVID-19 Case-Control Study Team Effectiveness of Face Mask or Respirator Use in Indoor Public Settings for Prevention of SARS-CoV-2 Infection - California, February-December 2021. MMWR. Morbidity and mortality weekly report, 71(6), 212–216. https://doi.org/10.15585/mmwr.mm7106e1

[2] Pei W et al 2023 Small target detection with remote sensing images based on an improved YOLOv5 algorithm. Frontiers in neurorobotics, 16, 1074862. https://doi.org/10.3389/fnbot.2022.1074862

[3] Susanto S et al 2020 The Face Mask Detection For Preventing the Spread of COVID-19 at Politeknik Negeri Batam 2020 3rd International Conference on Applied Engineering (ICAE), Batam, Indonesia pp. 1-5, doi: 10.1109/ICAE50557.2020.9350556.

[4] Xu S 2022 An Improved Lightweight YOLOv5 Model Based on Attention Mechanism for Face Mask Detection. International Conference on Artificial Neural Networks.

[5] Vaswani A et al 2017 Attention is All you Need. NIPS.

[6] Pratama Y et al 2021 Application of YOLO (You Only Look Once) V. 4 with Preprocessing Image and Network Experiment. The IJICS (International Journal of Informatics and Computer Science), 5(3), 280-286.

[7] Abadi M et al 2016 TensorFlow: A system for large-scale machine learning. USENIX Symposium on Operating Systems Design and Implementation.

[8] Gong H et al 2022 Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images. Remote Sensing, 14(12), 2861.

[9] Xia W et al 2022 A high-precision lightweight smoke detection model based on SE attention mechanism. In 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 941-944). IEEE.

[10] Qiu Y et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323.

[11] Shi G et al 2021 Combined Channel and Spatial Attention for YOLOv5 during Target Detection. In 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML) (pp. 78-85). IEEE.

[12] Yang S Q et al 2022 Student in-class behaviors detection and analysis system based on CBAM-YOLOv5. In 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 440-443). IEEE.

[13] Linlin Z et al 2021 Improving YOLOv5 with Attention Mechanism for Detecting Boulders from Planetary Images Remote Sensing, 13(18), 3776-3776.

[14] Yang W et al 2023 ST-CA YOLOv5: Improved YOLOv5 Based on Swin Transformer and Coordinate Attention for Surface Defect Detection. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[15] Bottou L 2010 Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France,

DOI: 10.54254/2755-2721/50/20241160

August 22-27, 2010 Keynote, Invited and Contributed Papers (pp. 177-186). Physica-Verlag HD.

[16] Zheng Z et al 2020 Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 12993-13000).

[17] Ren S et al 2015 Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[18] Guo C et al 2022 Multi-Stage Attentive Network for Motion Deblurring via Binary Cross-Entropy Loss. Entropy, 24(10), 1414.