Mask-wearing Detection Based on YOLOv5 and Knowledge Distillation

Haoran Liu, Meixia Qu*

Abstract—**With the ongoing persistence and global spread of the COVID-19 pandemic, numerous viral mutations have emerged, alongside the transmission of other infectious diseases, such as mpox, primarily via respiratory droplets. As a result, automated systems for mask-wearing detection have become essential for monitoring individuals in public spaces. Currently, deep learning-based object detection technologies dominate the field of mask detection systems. However, these methods often rely on deep network architectures, leading to slow detection speeds. Additionally, conventional large-scale models are unsuitable for lightweight deployment, rendering them inadequate for meeting specific commercial requirements. To address these challenges, we propose a mask detection method that integrates the YOLO object detection model with knowledge distillation. This approach effectively and accurately determines whether individuals are wearing masks while enhancing detection speed and enabling lightweight deployment for targeted commercial applications. By refining the object detection framework, the proposed method improves mask detection accuracy and model performance without increasing model complexity. Experimental results show that the enhanced student network, following the distillation process, achieves a 2.56% higher accuracy compared to the more complex YOLOv5x network. Furthermore, it surpasses state-of-the-art object detection algorithms in both speed and accuracy in real-world applications. These findings underscore the feasibility of the proposed method and its potential for widespread practical use across diverse conditions.**

Index Terms—**Mask-wearing detection, knowledge distillation, performance improvement, model compression**

I. Introduction

TARGET recognition and location are the main issues in computer vision. Image segmentation, target tracking, target behavior analysis, and so on are based on target image detection. Due to advancements in deep learning technology, the target detection algorithm has achieved a notable breakthrough. How to further improve the accuracy of target detection and the overall performance of the model has become an important issue in the current target detection research. Guan et al. [1] suggested that attention should be paid to further improving detection efficiency among existing methods. In the existing studies, in the context of the widespread transmission of COVID-19 and other airborne infectious diseases, there are numerous examples of using mathematical models and computer neural networks to detect and solve problems[2][3][4]. This also provides certain assistance for our research.

The outbreak of the prevalent diseases has brought mask detection to the forefront of research in computer vision. Nevertheless, deploying deep learning-based mask detection systems in industrial settings presents two primary challenges: (1) Large-scale mask datasets, in addition to data augmentation techniques, are essential for effective training. (2) Achieving robust, real-time, and high-precision models remains a formidable task.

Existing mask detection systems often fail to deliver the desired speed and performance, which significantly affects user experience. In recent years, researchers have proposed various solutions for mask detection. For example, Li [5] introduced a mask detection algorithm based on YOLOv3 with the integration of a SENET attention mechanism, while Fan [6] developed a detection system using YOLO5-Face to determine mask usage. Unfortunately, these efforts tend to focus primarily on improving accuracy, neglecting the equally critical need to enhance both accuracy and detection speed simultaneously.

To address these challenges, we propose a mask detection system based on the first-stage object detector YOLOv5, enhanced using a knowledge distillation framework. A key innovation of our approach lies in leveraging the knowledge distillation process, where the teacher network evaluates the complexity of each sample, enabling the student network to engage in adaptive learning tailored to the specific characteristics of the data. Experimental results demonstrate that the student network, optimized through knowledge distillation, significantly improves the average accuracy of the detection model without compromising detection speed or with only minimal accuracy loss. As a result, the network derived from our proposed method outperforms existing solutions, making it well-suited for industrial deployment. The primary contributions of this study encompass:

1. We propose an enhanced knowledge distillation technique that improves mask detection accuracy without adding additional parameters or increasing detection time.
2. We develop a mask detection method based on the YOLO object detection model and our novel knowledge distillation framework. This approach achieves improved average precision (mAP@.5 and mAP@.5:.95), meeting the practical application requirements for mask detection systems.

Manuscript received May 7, 2025; revised July 11, 2025. The study is supported by the Shenzhen Fundamental Research Program (JCYJ20230807094104009).

Haoran Liu is an undergraduate student of Shandong University - Australian National University Joint Science College, Shandong University, Weihai, 264299, China (e-mail: liuhaoran_sdwh@163.com).

Meixia Qu is an associate professor in the School of Mechanical, Electrical and Information Engineering of Shandong University, Shandong University, Weihai, 264299, China (Corresponding author to provide phone: +86-187-6310-0933; e-mail: mxqu@sdu.edu.cn).

II. Related Work

A. Knowledge distillation

The traditional target detection algorithm is typically segmented into three steps: (1) Candidate region selection: the sliding window algorithm delineates potential target regions on the input image by iteratively moving windows of varying sizes to pinpoint the target initially. (2) Feature extraction: use local binary pattern, directional gradient histogram, and other algorithms to extract the features of candidate regions. (3) Classification: Classify the extracted image features through support vector machine (SVM), AdaBoost, and other algorithms.

However, in recent years, target detection networks have grown increasingly complex, with a substantial increase in computational parameters. The excessive number of network parameters leads to significant memory consumption, high computational costs, and considerable inference latency. These challenges make deploying trained models on mobile devices particularly difficult. To mitigate the issues of increased latency and heavy computational demands, model compression techniques, such as knowledge distillation, low-rank decomposition, and network pruning [7] are widely employed. Among these, knowledge distillation, first proposed by Hinton et al. [8] in 2015, has demonstrated strong practical utility and consistently outperforms other model compression approaches in various applications. Knowledge distillation involves extracting insights from the teacher model and subsequently transferring them to a more streamlined, compact model. Through this process, the compact model assimilates the generalization capabilities of the larger model, gradually converging toward the performance levels of the extensive model. Knowledge distillation serves as a potent training technique for transferring model insights, streamlining model deployment, and expediting inference processes.

Currently, knowledge distillation finds applications across diverse domains. In computer vision, many researchers have researched knowledge distillation in image classification. One practical and straightforward approach is to utilize the output probability of the Softmax network layer as the soft target. The schematic diagram is illustrated in Fig.1. Moreover, the activation representations, neurons, or features within intermediate layers can be leveraged as knowledge to steer the learning process of the student network. The interplay among distinct activation representations, neurons, or paired samples encapsulates the wealth of information gleaned by the teacher network from the dataset [6][9][10].

In image classification, knowledge distillation plays a good role by compressing models, transferring the correlation of different labels to student models, integrating heterogeneous models, etc. However, more research needs to be conducted on object detection. In addition, most of the research on knowledge distillation methods has focused on two-stage detectors, whereas only some studies have focused on one-stage object detectors. Responses in object detection tasks can contain logits and bounding box offsets[11]. Choi [12], which built upon the two-stage target detection model Faster RCNN, extracted three parts of dark knowledge in the teacher network, namely, dark knowledge in the intermediate feature layer, dark knowledge in the RPN/RCN classification layer, and dark knowledge in the RPN/RCN regression layer; the research improved the accuracy of small target detection. Meanwhile, Mehta [13] used the knowledge distillation method to enhance the one-stage detection model; more importantly, it optimized the structure of the student network and improved the detection accuracy.

Furthermore, Wang [14] solved the problem in which the global distillation algorithm introduces background information; the study used masked feature maps to ensure that the teacher network only transfers knowledge near the actual box, thus resulting in better performance. Hou [15] proposed a self-attention distillation method, which uses the feature map of its convolutional layer as the extraction target of the shallow neural network; this application improved the detection effect. Moreover, Yang [16] introduced a variant of self-distillation termed snapshot distillation, where the initial knowledge of the teacher network is transferred to its subsequent student network stages, aiding in the training progression within the same network. Roheda [17] proposed using GANs to perform cross-modal distillation between missing and available models, which can subsequently improve object detection performance.

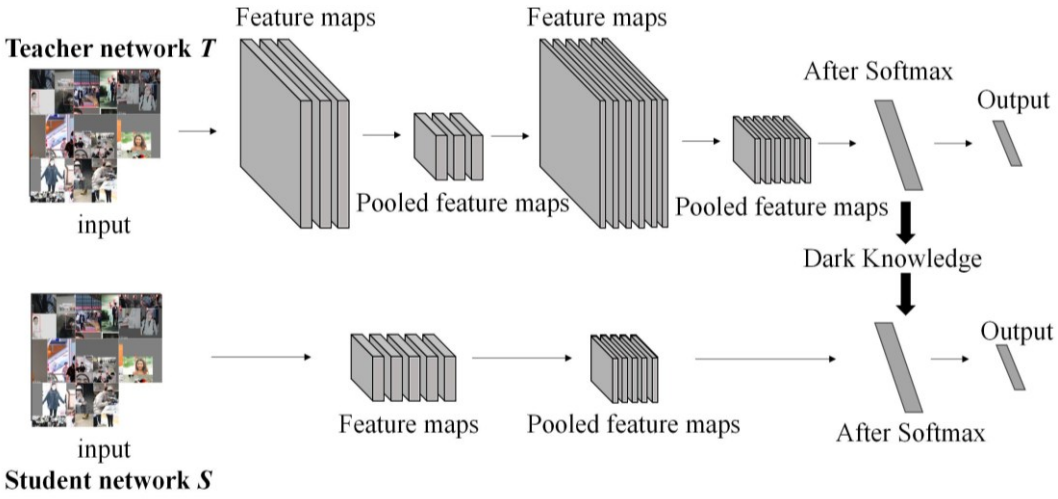

Fig. 1. The transmission process of dark knowledge.

B. Target detection model YOLO

The YOLO object detection model, which can detect objects in real-time, was first proposed in 2015 [18]. YOLO is a "one-stage" idea. The detection object is a regression problem but not a classification problem [19], as shown in Fig.2.

First, the network is divided, and the grid is detected, where the object's center falls into that network. Moreover, the grid is responsible for the target. Each grid needs to predict both the bounding box and the category simultaneously. Compared with the above two-stage target detector, YOLO's ban on candidate frame generation and using a single neural network to complete the detection can greatly improve the detection speed. YOLOv5 significantly reduces the model's size, improves the operation speed, and performs similarly to YOLOv4 [20].

While the two-stage object detector exhibits superior performance, it suffers from significant speed delays, which run counter to the low-latency objectives sought after for edge devices. Meanwhile, regarding space occupation, only YOLO can adapt to hardware with limited storage volume. Therefore, the two-stage detection model still needs to be completed on the mobile terminal [21]. Furthermore, systems characterized by high real-time constraints, such as mask-wearing detection systems, persist in their reliance on the one-stage target detector YOLO. For the above reasons, the YOLOV5 target detection algorithm is used as the baseline and starting point of the mask-wearing detection method in the present study.

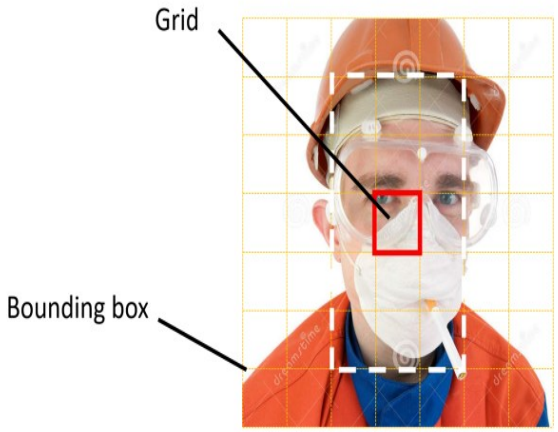

Fig. 2. The idea of the "one-stage" target detection model: Detecting objects based on the central grid.

III. Improved Knowledge Distillation Mask-Wearing Detection Method Based on YOLOv5

A. Traditional Knowledge Distillation Algorithm

Knowledge distillation transfers the dark knowledge from a larger model (teacher network) to a more compact model (student network). Hence, the tiny model's accuracy reaches the large model's level. Thus, the model used for prediction has a faster speed and higher accuracy.

The knowledge distillation method used in this study extracts response-based knowledge [22]. Moreover, it simultaneously extracts localization and classification knowledge. The teacher network predicts the target object's location, category, and confidence after sufficient learning from the dataset [23]. We impart this knowledge to the student network through knowledge distillation.

The distillation loss is formulated by quantifying the disparity between the forecasts of the teacher and the student, delineated in formula (1):

$$\boldsymbol{L_{KD}} = \boldsymbol{L}\big(\boldsymbol{f_t}(\boldsymbol{x}, \boldsymbol{\theta_t}), \boldsymbol{f_s}(\boldsymbol{x}, \boldsymbol{\theta_s})\big) \tag{1}$$

Where $\boldsymbol{f_t}$ denotes the teacher network, $\boldsymbol{f_s}$ denotes the student network, $\boldsymbol{\theta_t}$ denotes the model parameters of the teacher network, $\boldsymbol{\theta_s}$ denotes the model parameters of the student network, $\boldsymbol{x}$ denotes the input of the same data set, and $\boldsymbol{L}(.)$ is the measure of the distance between teachers and students, that is, a specific type of the loss function. Our method, $\boldsymbol{L}(.)$ denotes not only the regression loss for location prediction but also the classification loss between wearing a mask and not wearing a mask. Under distillation and back-propagation, the knowledge encapsulated within the teacher network is swiftly conveyed to the student network. Furthermore, if the student network merely mimics the teacher network during knowledge distillation, its potential is constrained by the knowledge of the teacher network. It cannot exceed the possibility of the teacher network [24]. Therefore, the distillation loss ought to be integrated with the original task-specific loss $\boldsymbol{L_{hard}}$ to facilitate enhanced learning for the student network [25]. The calculation method is shown in formula (2), where the combination of the distillation loss and the loss between the student network and the labeled value $label$ form the comprehensive loss function:

$$\boldsymbol{L_{total}} = \boldsymbol{L_{hard}}(\boldsymbol{f_{s'}}\boldsymbol{label}) + \boldsymbol{\lambda L_{KD}} \tag{2}$$

Where $\boldsymbol{L_{total}}$ denotes the total loss, $\boldsymbol{L_{hard}}$ denotes the loss between the student network and the ground-truth label, while $\boldsymbol{L_{KD}}$ denotes the distillation loss, $\boldsymbol{f_s}$ denotes the student network, and $\boldsymbol{label}$ denotes the ground-truth label. In addition, λ represents the equilibrium parameter in distillation, where λ is usually regarded as a hyperparameter in previous knowledge distillation methods and set as a fixed constant.

B. Improved Knowledge Distillation Algorithm

The previous knowledge distillation loss function, such as formula (2), has encountered hyperparameter sensitivity and poor ability to detect simple samples [25]. Therefore, the previous loss function is set to a fixed value. Unlike the previous practice of setting hyperparameters, the distillation method in the present study achieves difficulty awareness. It means the network judges the samples with high confidence in the target detection training set as low difficulty, thus indicating that the teacher network has fully learned and extracted this knowledge. Subsequently, the dark knowledge in this scenario is assigned a greater weight and transmitted to the student network through the teacher network. The process of the knowledge distillation algorithm with adaptive difficulty in this study is shown in Fig. 3.

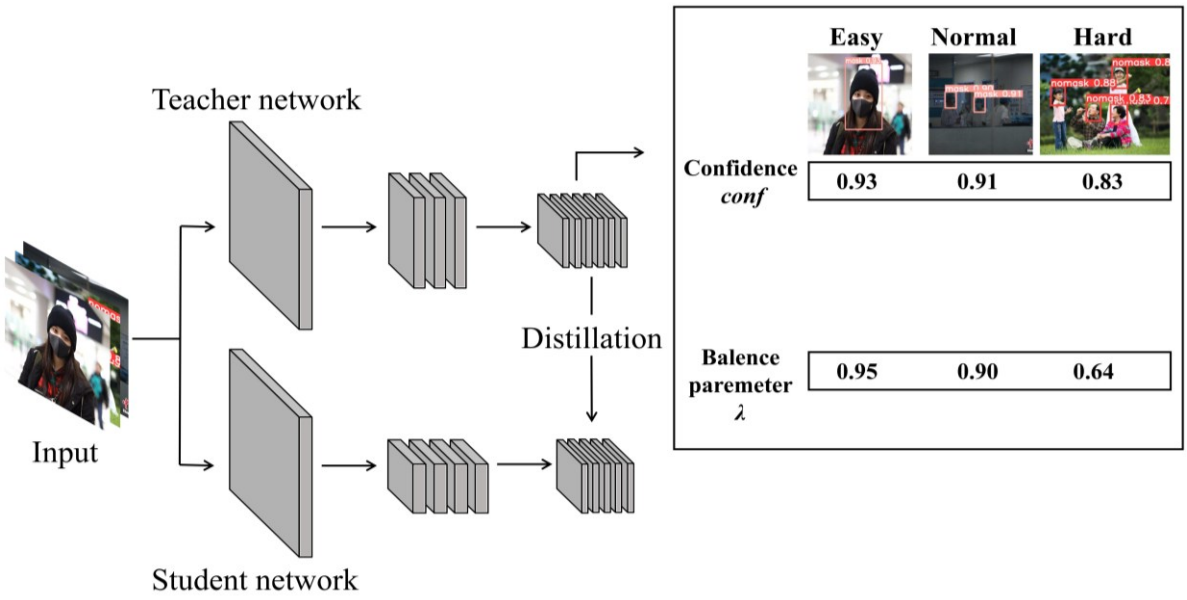

Fig. 3. The overall framework of the improved YOLOv5 algorithm based on knowledge distillation.

Confidence indicates the certainty that the predicted bounding box encompasses the object and the precision of the bounding box prediction. The confidence score for the grid is computed as illustrated in formula (3):

$$\boldsymbol{conf} = \boldsymbol{Pr(object)} \times \boldsymbol{Pr(class_i|object)} \times \boldsymbol{IOU_{predict}^{label}} \quad (3)$$

Where $\boldsymbol{Pr(object)} = \boldsymbol{1}$ if a target falls on the grid; otherwise, $\boldsymbol{Pr(object)} = \boldsymbol{0}$. For the class $\boldsymbol{i}$ predicted by the grid, $\boldsymbol{Pr(class_i|object)}$ indicates the probability that the predicted object belongs to class. $\boldsymbol{IOU_{predict}^{label}}$ represents the intersection-over-union ratio between the ground-truth box and the predicted bounding box.

To realize the adaptive adjustment of dark knowledge weights for samples of different difficulties, this study defines different weights in distillation loss $[\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_n]$ corresponding to different samples, as shown in Equation (4):

$$[\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}, \ldots, \boldsymbol{\lambda_n}] = \boldsymbol{sigmoid}([\boldsymbol{conf_1}, \boldsymbol{conf_2}, \ldots, \boldsymbol{conf_n}] - \boldsymbol{t}) \quad (4)$$

Among them, $[\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_n]$ is the distillation loss balance parameter for each sample of different difficulties in formula (2). The number of samples is $\boldsymbol{n}$. In addition, $[\boldsymbol{conf_1}, \boldsymbol{conf_2}, \boldsymbol{conf_3}, \ldots, \boldsymbol{conf_n}]$ is the confidence of each object predicted by the teacher network, and $\boldsymbol{t}$ is the hyperparameter employed to regulate the extent to which knowledge is assimilated from the teacher.

On the one hand, a hyperparameter is used when calculating the weight $\boldsymbol{\lambda}$ of knowledge passed by the teacher to control how much knowledge is obtained from the teacher model. The hyperparameter $\boldsymbol{t}$ is akin to the "temperature T" hyperparameter in prior knowledge distillation techniques. A higher value of $\boldsymbol{t}$ enables the student network to glean more knowledge from the teacher network, and a lower value of t serves to shield the student network from potential errors in teacher predictions. On the other hand, using the $sigmoid$ function not only remaps the confidence interval to [0, 1], but it also diminishes the range within which the teacher network perceives the difficulty level. Initially, the teacher network gauges the complexity of the samples. Then, it guides the student network to conduct targeted learning with different learning plans for samples of various problems; meanwhile, the student network focuses on the learning, which improves the distillation effect [25] and the performance of the student network in practical applications.

Here is the algorithm description for the framework of improved part calculation:

Algorithm 1: Framework of improved part calculation.

Input: $\boldsymbol{x_t}, \boldsymbol{y_t}, \boldsymbol{h_t}, \boldsymbol{x_s}, \boldsymbol{y_s}, \boldsymbol{w_s}, \boldsymbol{h_s}$ from the output of the teacher and the output of the student network's previous round, $\boldsymbol{conf}$ given by the teacher network;

Output: $\boldsymbol{L_{total}}$ used for neural network backpropagation;

1: Calculate student network weight λ of knowledge distillation, that is, $[\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}, \ldots, \boldsymbol{\lambda_n}]$.

$$[\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}, \ldots, \boldsymbol{\lambda_n}] = \boldsymbol{sigmoid}([\boldsymbol{conf_1}, \boldsymbol{conf_2}, \boldsymbol{conf_3}, \ldots, \boldsymbol{conf_n}] - \boldsymbol{t});$$

2: Calculate the distillation loss

$$\boldsymbol{L_{KD}} = (\boldsymbol{x_t} - \boldsymbol{x_s})^2 + (\boldsymbol{y_t} - \boldsymbol{y_s})^2 + (\boldsymbol{w_t} - \boldsymbol{w_s})^2 + (\boldsymbol{h_t} - \boldsymbol{h_s})^2 + \boldsymbol{BCELoss}\left(\boldsymbol{soft\,max}\left(\frac{\boldsymbol{D_t}}{\boldsymbol{T}}\right), \boldsymbol{soft\,max}\left(\frac{\boldsymbol{D_t}}{\boldsymbol{T}}\right)\right)$$

3: Calculate the total loss

$$\boldsymbol{L_{total}} = \boldsymbol{L_{hard}}(\boldsymbol{f_{s'}}, \boldsymbol{label}) + \boldsymbol{\lambda L_{KD}}$$

4: Return $\boldsymbol{L_{total}}$

The process of transferring dark knowledge and training the student network is shown in Fig.4.

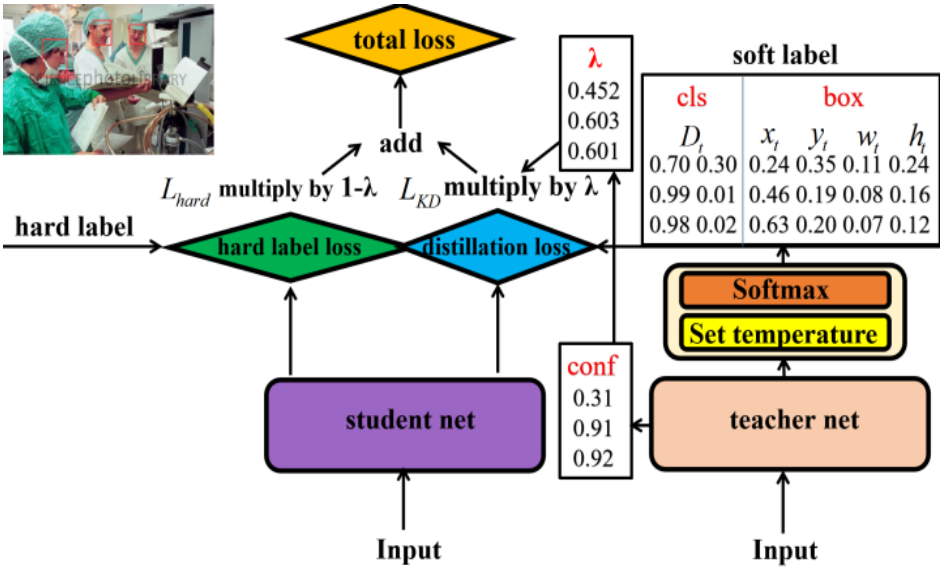

Fig. 4. Dark knowledge transfer and the process of training the student network.

The process involves extracting the forecasted center position from the teacher network, the frame's width and height, and the probabilities associated with each category as a soft label, denoted by the following symbols: $\boldsymbol{x_t, y_t, w_t, h_t}$ and D_t respectively. When specifically guiding students to network learning, the knowledge distillation loss function L_{KD} in this study is defined as follows:

$$\boldsymbol{L_{KD} = (x_t - x_s)^2 + (y_t - y_s)^2 + (w_t - w_s)^2 + (h_t - h_s)^2 + BCELoss\left(softmax\left(\frac{D_t}{T}\right), softmax\left(\frac{D_t}{T}\right)\right)} \quad (5)$$

The distillation loss comprises the squared distance between the teacher and student prediction locations and the BCELoss function applied to the probability distributions of the teacher and student prediction categories. $\boldsymbol{x_t, y_t, w_t, h_t, x_s, y_s, w_s}$ and h_s respectively represent the rectangle box predicted by the teacher and student network, $\boldsymbol{D_t}$ and D_s represent the teacher's and student's probability distribution of the object category predicted by them. For example, in Fig.4, $\boldsymbol{D_t}$ is [0.70, 0.30]. This value indicates that, for the lady in the upper left corner, the probability of belonging to the "without a mask" category is 0.7, while the probability of belonging to the category of "wearing a mask" is 0.3. $\boldsymbol{softmax}(.)$ ensures that the probability distribution is normalized. $\boldsymbol{T}$ represents the distillation temperature, $\boldsymbol{BCELoss}(.)$ represents the *BCELoss* function and the BCELoss function between $\boldsymbol{C_t}$ and the $\boldsymbol{C_s}$ is as follows:

$$\boldsymbol{BCELoss(C_t||C_s) = -1/n \sum_i C_t[i] * logC_s[i] + (1- C_t[i]) * log(1 - C_s[i]))} \quad (6)$$

Where $\boldsymbol{i}$ represents the $\boldsymbol{i_{th}}$ category.

The optimization objective of distillation $\boldsymbol{L_{KD}}$ delineates the variance between the forecasted values of the teacher network and those of the student network. The actual label loss $\boldsymbol{L_{hard}}$ is incorporated to derive the total loss of the student network, as depicted in Equation (2), thereby facilitating the transmission of dark knowledge. The student network optimizes the model parameters through backpropagation to simultaneously fit dark knowledge and actual labels, realize the learning of dark knowledge and actual labels, and finally obtain a student model with higher performance and accuracy.

To sum up, the core of the method in this study is to use the teacher's network to perceive the sample's difficulty in the early training stage. Using the network helps to obtain the coefficient of the teacher's mastery of the sample, which can guide the student network learning with different confidence levels for samples of other difficulties. The student network can effectively grasp straightforward examples with the teacher's assistance. Yet, its capability to comprehend the teacher's knowledge when confronted with intricate samples remains limited. Therefore, for students, the focus of learning from the teacher network should be biased toward simple samples [25]. In the realm of practical application, particularly in mask-wearing detection, the presence of the teacher network notably accelerates the training and comprehension of the student network, especially for uncomplicated instances.

C. Selected teacher network and student network

1) Determine a standard YOLO network

In the experiment, the chosen teacher network is YOLOv5x, comprising 86,180,143 parameters. The chosen baseline models are the YOLOv5m model and the YOLOv5s model, trained separately without distillation. The selected student network1 is YOLOv5s, and the student network2 is YOLOv5m.

2) Determine a lightweight YOLO network

The target detection algorithm emphasizes real-time monitoring. Therefore, to substantiate the efficacy of the proposed algorithm in enhancing lightweight models and to showcase its applicability in practical engineering scenarios, a more lightweight target detection model, YOLO-Lite [26], has been chosen for the experiment. YOLO-Lite is an enhanced real-time target detection network built upon YOLOv5. Compared with YOLOv5, YOLO-Lite is faster, lighter, and easier to deploy.

After thoroughly studying the ablation experiments, the YOLO-Lite network implements the removal of the focus point

layer to avoid multiple slicing operations. Improvements such as preventing the use of high-channel C3 layers at various times have also been made, thus enabling easy deployment on the mobile side without reducing the feature extraction capability of the model. Regarding space occupation, when the YOLO5s model still has 12.4M, the size of the YOLO-Lite model is only 3.2 M. Moreover, the number of parameters of the YOLO-Lite model is 1/5 of that of the YOLO5s model. This network can also be used on the edge device, the Raspberry Pi. After deployment on Raspberry Pi 4B, the detection speed meets the real-time requirements.

In YOLO-Lite-s and YOLO-Lite-g, we choose our experiments' lighter network, YOLO-Lite-s, as the student network. In our enhanced knowledge distillation approach for mask-wearing detection using YOLOv5, the identical network is chosen as the baseline and student networks in the knowledge distillation algorithm. This ensures that the number of student network parameters after knowledge distillation remains aligned with the baseline network, without any addition.

The network structure of YOLO-Lite-s is shown in the following table Ⅰ:

TABLE Ⅰ
THE STRUCTURE OF THE STUDENT NEURAL NETWORK, ALONG WITH THE PARAMETER COUNT IN EACH LAYER.
(AGGREGATE OF ALL WEIGHTS AND BIASES)

Neural network layer name	Parameters of each layer of the YOLO5 Lite-s student network
Maxpool	928
Shuffle_Blocks	8812
Shuffle_Blocks	44588
Shuffle_Blocks	167968
Conv	59648
Upsample	0
Concat	0
C3	104192
Conv	8320
Upsample	0
Concat	0
C3	26496
Conv	36992
Concat	0
C3	74496
Conv	147712
Concat	0
C3	296448
Detect	9471

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Dataset

The implementation of this system requires two classes of training images: mask and no mask. Then, the mask images are crawled to realize the mask detection system. In addition, the dataset we chose came from public datasets and pictures on the network, and the characters and backgrounds in the pictures are closer to the scenarios that require lightweight deployment and commercial use in real life.

Simultaneously, instances where facial features are partially obscured without the utilization of facial masks are incorporated to enhance the network's capacity for generalization and resilience. The curated dataset consists of 577 images, distributed with a partition ratio of 8:1:1 among the training, validation, and test sets.

B. Experimental configuration

The research experiment was conducted within the Windows 11 operating system environment. Moreover, PyTorch is selected as the deep learning framework. In addition, the graphics card NVIDIA GeForce RTX 4060 and the acceleration library CUDA 12.4 is used.

TABLE Ⅱ
The super-parameter setting of the experiment

Super-parameter	Value
batch size	4
image scale	640 x 640
initial learning rate	0.01
mixup	0.1

C. Experimental results

In assessing the efficacy of the target detection model in this study, evaluation metrics such as mAP@0.5, Precision, mAP@0.5:0.95, and Recall are chosen as performance indicators.

(1) Precision and recall. They are fundamental metrics utilized in model evaluation. Precision is the ratio of correctly identified positive samples to all samples identified as positive, while recall is the ratio of correctly identified positive samples to all truly positive samples based on the ground truth labels. These rates are calculated using formula (7):

$$\begin{cases} \boldsymbol{P = TP/(TP + FP)} \\ \boldsymbol{R = TP/(TP + FN)} \end{cases} \tag{7}$$

In this context, TP (True Positives) denotes the count of samples where the predicted object type by the target detection model aligns with the actual object type. Conversely, FP (False Positives) represents the count of samples where the predicted object type by the target detection model differs from the actual object type. Finally, FN is the number of real objects in the sample not detected by the object detection model.

(2) mAP@.5, mAP@.5:.95 (mean Average Precision). The precision-recall curve (P-R curve) is obtained by drawing a dotted line on the (P, R) value of the target detection result. The mAP for each object category is computed by aggregating the area under the P-R curve, and the overall mAP is derived by averaging the mAP values across all object categories. The formula for calculating the mAP is delineated as shown in the formula (8):

$$\boldsymbol{mAP = \sum_{j=1}^{C} \frac{AP_j}{C}} \tag{8}$$

1) Analysis of experimental results on the standard YOLO network

First, the processes outlined in Section 3.2 are adhered to for training the teacher network. The training loss curve of the teacher network is depicted in Fig.5.

Subsequently, the teacher network undergoes distillation, extracting dark knowledge. The subsequent phase encompasses the training of the student network.

The performance of YOLO5s baseline model and student network 1 (average accuracy index map@.5, map@.5:.95) is shown in Fig. 7. Above are performance images of the YOLOV5s baseline network, and below are performance images of student network 1 after distillation.

The performance of the YOLO5m baseline model and student network 2 (average accuracy index chart @. 5, map@. 5:.95) is shown in Fig. 8. Above are performance images of the YOLOV5m baseline network, and below are performance images of student network 2 after distillation.

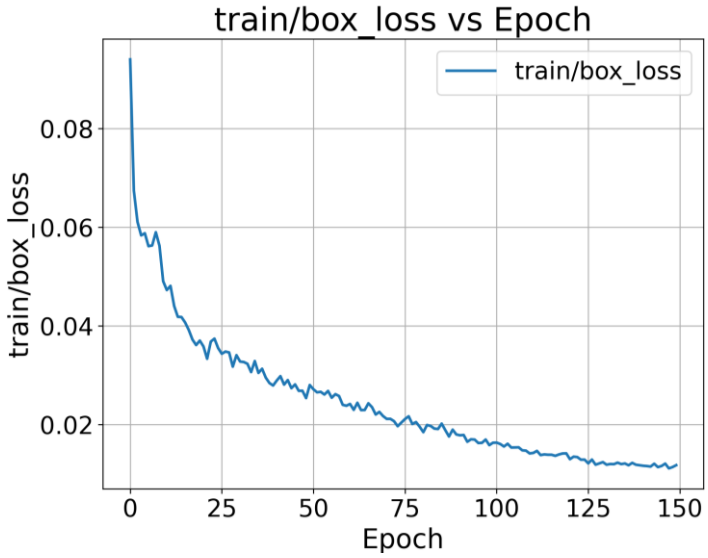

Fig. 5. Loss convergence image of the teacher network.

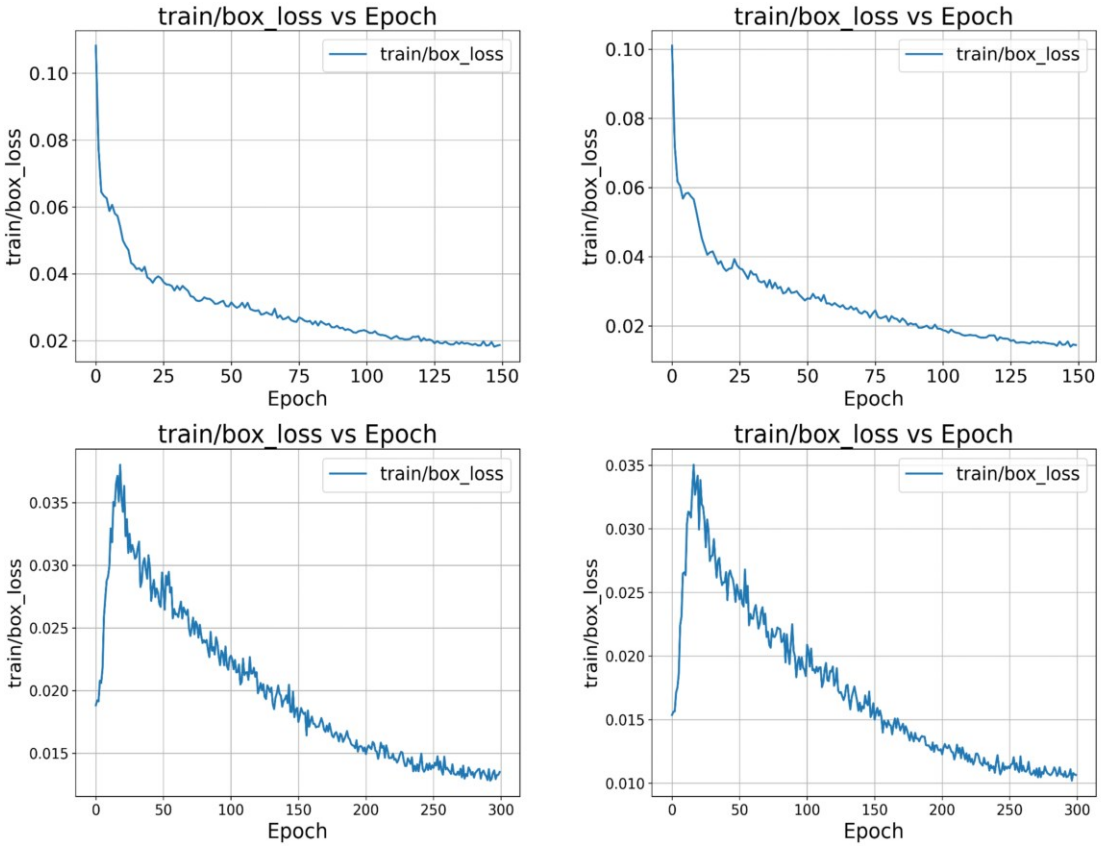

Fig. 6. The loss convergence images, top left YOLOV5s baseline network, top right YOLOV5m baseline network, bottom left student network 1, and bottom right student network 2.

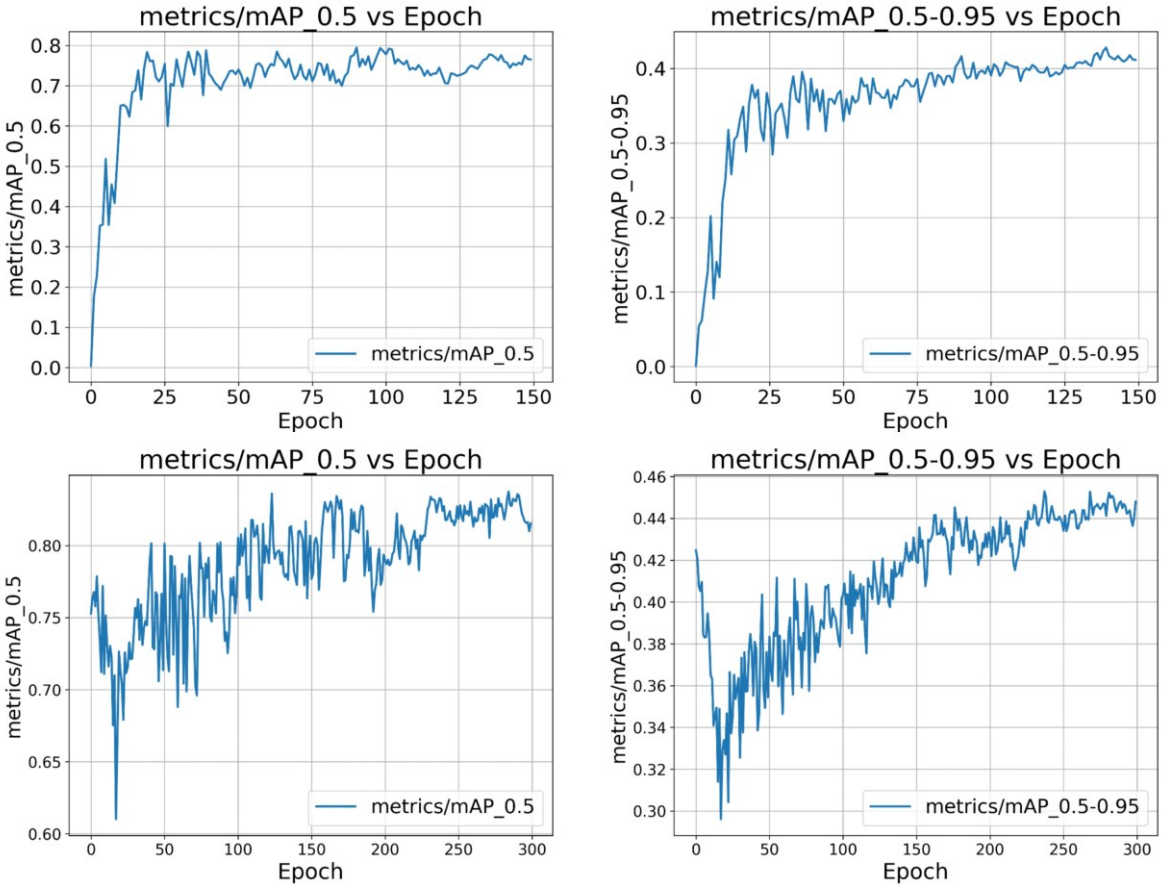

Fig. 7. Experimental results of the distillation experiment on YOLOV5s, map@.5, and map@.5:.95 of YOLO5s baseline network and YOLO5s student network.

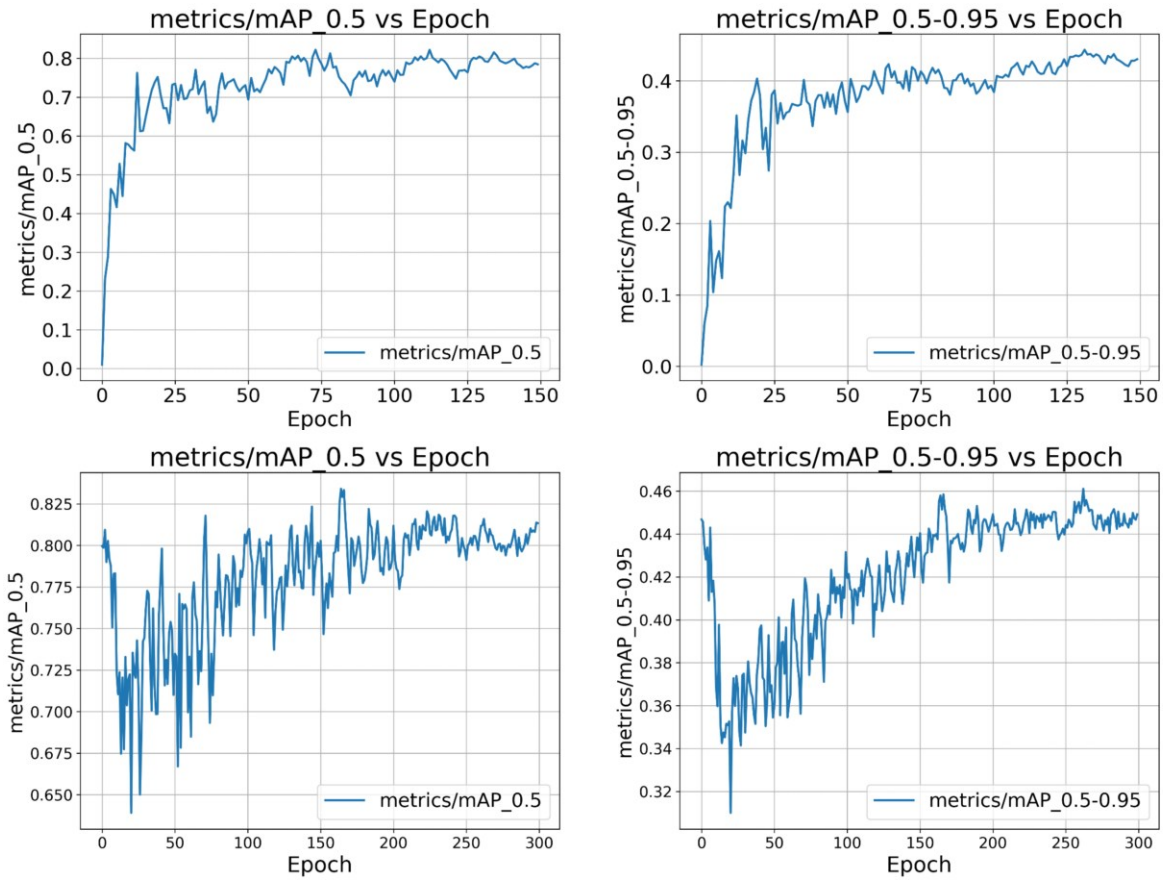

Fig. 8. Experimental results of the distillation experiment on YOLO5m, map@.5, and map@.5:.95 for YOLO5m baseline network and YOLO5m student network.

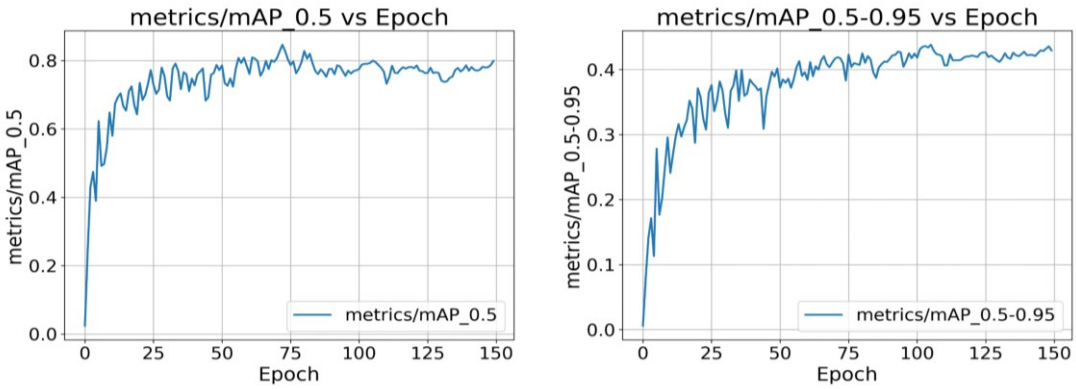

Fig. 9. map@.5 and map@.5:.95 of teacher network (YOLOV5x).

The performance on each evaluation metric is shown in Table III :

TABLE III

The performance of the teacher network, baseline networks, and student networks is assessed across each target detection evaluation metric.

models	map@.5	precision	map@.5:.95	recall
YOLOv5x (teacher network)	82.9	86.0	44.7	76.5
YOLOv5s before distillation (baseline network1)	77.6	84.2	42.8	67.9
YOLOv5s after distillation (student network1)	83.3	86.6	45.3	76.6
YOLOv5m before distillation (baseline network2)	79.2	81.7	44.3	72.5
YOLOv5m after distillation (student network2)	81.6	88.2	46.1	67.8

The results of this experiment are analyzed as follows:

(1) Table III reveals a notable enhancement in the performance of each evaluation metric for Student Network 1 when juxtaposed with the baseline model. Fig.7, 8, and 9 illustrate that through the transmission of dark knowledge from the teacher network, during the initial training stages, the student network progressively diminishes the disparity with the teacher network. In the later stage of training, the student network is infinitely close to the teacher network in each evaluation index, surpassing the teacher network in some evaluation metrics.

Fig. 10. a Teacher network.

Fig. 10. b Student network1.

Fig. 10. c Student network2.

Fig. 10. Results of teacher network, student network1, and student network2 predicting the examples.

(2) Fig.10 demonstrates the detection results of our well-trained teacher network, student network 1, and student network 2, respectively. The student network, after distillation, performed better in detecting mask-wearing. Following knowledge distillation, the performance of the student network surpasses that of the preceding two networks in terms of mAP@0.5 and mAP@0.5:0.95. The experimental findings substantiate the efficacy of the teacher network in imparting knowledge to the student network through the knowledge distillation approach in this study.

(3) In comparing the outcomes of two sets of experiments involving the YOLO5s model and the YOLO5m model, it is observed that while there is a minimal disparity in performance between the YOLO5s model and the YOLO5m model before knowledge distillation, the YOLO5m student network exhibits significantly enhanced performance post knowledge distillation in contrast to the YOLO5s student network. This outcome also confirms the conclusion of Mirzadeh [27] and Huang [28] et al., who observed that distilling between networks with smaller capability gaps might yield better performance. The above experiments are improved based on the standard YOLO [29] and a more lightweight target detection model, YOLO-Lite [26].

2) Analysis of experimental results on the lightweight YOLO network

The experimental outcomes of the lightweight YOLO network are depicted in Table IV:

TABLE IV

Performance of baseline network and student network in each target detection evaluation metric.

models	map@.5	precision	map@.5:.95	recall
YOLO-Lite-s baseline	77.6	84.2	42.8	67.9
YOLO-Lite-s student network	**83.3**	**86.6**	**45.3**	**76.6**

The comparison of some output images of the YOLO-Lite baseline network and the YOLO-Lite student network is shown in Fig. 11:

Fig. 11. a Example of YOLO-Lite-s baseline.

Fig. 11. b Example of YOLO-Lite-s student network.

Fig. 11 Results of YOLO-Lite-s baseline and student network predicting the examples.

Experiments show that the lightweight student network obtains better feature extraction ability under the teaching of the powerful teacher network. As shown in Fig.11. a and Fig.11. b, in cases where many targets appear in the picture simultaneously, the distilled student network can make more accurate predictions.

V. Conclusion

Complex models can significantly enhance the learning performance of deep learning tasks. However, the extensive network parameters associated with these models substantially increase computational demands and lead to delays in practical applications. Such delays pose serious challenges to the deployment and utilization of neural network models on edge devices. To address these limitations, knowledge distillation methods have emerged as a promising solution to balance time efficiency and hardware constraints in industrial applications.

This study presents a novel approach to effectively distill complex tacit knowledge, enabling the efficient transfer of "dark knowledge" from a sophisticated teacher network to a lightweight student network. Our proposed method achieves optimized detection times, enhanced detection accuracy, and improved overall model performance without increasing the number of parameters. Experimental results demonstrate that, after applying knowledge distillation, the student network achieves notable improvements in accuracy and overall performance, even surpassing the teacher network. Furthermore, the distilled student network exhibits substantial performance gains when compared to the baseline network. Most importantly, the proposed method demonstrates broad applicability in both industrial scenarios and commercial applications.

Nevertheless, there remain areas for improvement in our approach. For instance, in real-world scenarios involving dense crowds or significant obstructions between individuals, the model's performance may decline, limiting its effectiveness. Addressing these challenges will be a central focus of our future work. Our goal is to further enhance detection accuracy and model performance while ensuring seamless integration of target detection systems into real-world applications. We are confident that continued research will uncover more effective solutions to these challenges in the future.

References

[1] Guan, Y., Aamir, M., Hu, Z., Dayo, Z. A., Rahman, Z., Abro, W. A., & Soothar, P. "An Object Detection Framework Based on Deep Features and High-Quality Object Locations," *Traitement du Signal*, vol. 38, no. 3, 2021.

[2] Siti Nurmaini, Alexander Edo Tondas, Radiyati Umi Partan, Muhammad Naufal Rachmatullah, Annisa Darmawahyuni, Firdaus Firdaus, Bambang Tutuko, Rachmat Hidayat, & Ade Iriani Sapitri. "Automated Detection of COVID-19 Infected Lesion on Computed Tomography Images Using Faster-RCNNs," *Engineering Letters*, 2020, 28(4), pp. 1295–1301.

[3] Kewalee Suebyat, Pravitra Oyijnda, Sureerat A. Konglok, & Nopparat Pochai. "A Mathematical Model for the Risk Analysis of Airborne Infectious Disease in an Outpatient Room with Personal Classification Factor," *Engineering Letters*, 2020, 28(4), pp. 1331–1337.

[4] Ahmed N. K. Alfarra & Ahmed Hagag. "COVID-19 Crisis: Forecasting the Arab World Economy Performance Using the ARIMA Model," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2022*, 2022, 6–8 July, London, U.K., pp. 48–53.

[5] Li, G. J., & Rong, Y. "Face Mask Detection Algorithm Based on DCN-SERES-YOLOv3," *Journal of Computer and Modernization*, vol. 48, no. 9, pp. 12–20, 30, 2021.

[6] Fan, Q. R., & Li, D. "Mask Face Detection Based on YOLO5Face," *Electronic Production*, vol. 48, no. 19, pp. 61–63, 8, 2021.

[7] Meng, X. F., Liu, F., Li, G., & Huang, M. M. "Overview of Knowledge Distillation Technology in Convolutional Neural Network Compression," *Journal of Computer Science and Exploration*, vol. 15, no. 10, pp. 1812–1829, 2021.

[8] Hinton, G., Vinyals, O., & Dean, J. "Distilling the Knowledge in a Neural Network," 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, 2014, pp. 38–39.

[9] Ahn, S., Hu, S. X., Damianou, A., et al. "Variational Information Distillation for Knowledge Transfer," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020.

[10] Zagoruyko, S., & Komodakis, N. "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer," 2016.

[11] Wang, L., & Yoon, K. J. "Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks," 2020.

[12] Choi, W., Chandraker, M., Chen, G., et al. "Learning Efficient Object Detection Models With Knowledge Distillation: US20180268292," [Patent], 2018.

[13] Mehta, R., & Ozturk, C. "Object Detection at 200 Frames Per Second," *Proceedings of European Conference on Computer Vision*, 2018.

[14] Wang, T., Yuan, L., Zhang, X., et al. "Distilling Object Detectors With Fine-Grained Feature Imitation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020.

[15] Hou, Y., Ma, Z., Liu, C., et al. "Learning Lightweight Lane Detection CNNs by Self Attention Distillation," 2019.

[16] Yang, C., Xie, L., Su, C., et al. "Snapshot Distillation: Teacher-Student Optimization in One Generation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019.

[17] Roheda, S., Riggan, B. S., Krim, H., et al. "Cross-Modality Distillation: A Case for Conditional Generative Adversarial Networks," *ICASSP: IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2018.

[18] Redmon, J., Divvala, S., Girshick, R., et al. "You Only Look Once: Unified, Real-Time Object Detection," IEEE, 2016.

[19] Ming-Hsuan, Y., Kriegman, et al. "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, no. 1, p. 34, 2002.

[20] Bochkovskiy, A., Wang, C. Y., & Liao, H. "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020.

[21] Li, J. N., Wu, X., Liu, J. S., & Wang, H. G. "Research on YOLOv3 Algorithm Based on Knowledge Distillation," *Journal of Computer Engineering and Applications*, [Online], available: http://kns.cnki.net/kcms/detail/11.2127.tp.20210414.0857.002.html, 2021.

[22] Gou, J., Yu, B., Maybank, S. J., et al. "Knowledge Distillation: A Survey," 2020.
[23] Bochkovskiy, A., Wang, C. Y., & Liao, H. "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020.
[24] Clark, K., Luong, M. T., Khandelwal, U., et al. "BAM! Born-Again Multi-Task Networks for Natural Language Understanding," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
[25] Zhang, Y., Lan, Z., Dai, Y., et al. "Prime-Aware Adaptive Distillation," 2020.
[26] Chen. "YOLOv5-Lite (v1.0)," [Computer software], Zenodo, 2021, available: https://doi.org/10.5281/ZENODO.5241425.
[27] Mirzadeh, S. I., Farajtabar, M., Li, A., et al. "Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher," 2019.
[28] Huang, Z., & Wang, N. "Like What You Like: Knowledge Distill via Neuron Selectivity Transfer," 2017.
[29] Jocher, G., Ayush Chaurasia, Stoken, A., et al. "ultralytics/yolov5: v6.1—TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (v6.1)," [Computer software], Zenodo, 2022, available: https://doi.org/10.5281/ZENODO.6222936.