Displays 78 (2023) 102403

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

LightR-YOLOv5: A compact rotating detector for SARS-CoV-2 antigen-detection rapid diagnostic test results☆

Rongsheng Wang [a], Yaofei Duan [a], Menghan Hu [b], Xiaohong Liu [c], Yukun Li [a], Qinquan Gao [d], Tong Tong [d], Tao Tan [a,*]

[a] Faculty of Applied Sciences, Macao Polytechnic University, Rua de Luís Gonzaga Gomes, 999078, Macao Special Administrative Region of China
[b] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200240, China
[c] John Hopcroft Center, Shanghai Jiao Tong University, Shanghai 200240, China
[d] College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China

ARTICLE INFO

Keywords:
YOLOv5
Lightweight
Rotating object detection
RDT detection

ABSTRACT

Nucleic acid testing is currently the golden reference for coronaviruses (SARS-CoV-2) detection, while the SARS-CoV-2 antigen-detection rapid diagnostic tests (RDT) is an important adjunct. RDT can be widely used in the community or regional screening management as self-test tools and may need to be verified by healthcare authorities. However, manual verification of RDT results is a time-consuming task, and existing object detection algorithms usually suffer from high model complexity and computational effort, making them difficult to deploy. We propose LightR-YOLOv5, a compact rotating SARS-CoV-2 antigen-detection RDT results detector. Firstly, we employ an extremely light-weight L-ShuffleNetV2 network as a feature extraction network with a slight reduction in recognition accuracy. Secondly, we combine semantic and texture features in different layers by judiciously combining and employing GSConv, depth-wise convolution, and other modules, and further employ the NAM attention to locate the RDT result detection region. Furthermore, we propose a new data augmentation approach, Single-Copy–Paste, for increasing data samples for the specific task of RDT result detection while achieving a small improvement in model accuracy. Compared with some mainstream rotating object detection networks, the model size of our LightR-YOLOv5 is only 2.03MB, and it is 12.6%, 6.4%, and 7.3% higher in mAP@.5:.95 metrics compared to RetianNet, FCOS, and R^3Det, respectively.

1. Introduction

Accurate identification of nucleic acid test results is particularly important during the COVID-19 pandemic. Currently, there are two non-imaging popular methods for detecting SARS-CoV-2: polymerase chain reaction (PCR)-based assays and the SARS-CoV-2 rapid antigen test (RDT), an instrument-free rapid chromatographic immunoassay designed to qualitatively detect the presence of specific SARS-CoV-2 antigens in nasopharyngeal or mixed nasopharyngeal/oropharyngeal samples [1]. RDT has received increasing attention in public medicine due to the high transmissibility and long incubation period of the Omicron virus [2,3]. Studies have shown that RDT has significant advantages over PCR assays in detecting SARS-CoV-2. For example, RDT is more convenient to operate, and people can perform self-tests without aggregation, thus reducing the risk of cross-infection during PCR testing; since detection time is short, decision-making can be achieved in 15–20 min [4]. During the COVID-19 epidemic, many works on diagnosis and prediction of the disease emerged using medical imaging such as [5–11], but they are medical health issues, and we focus here more on the social security issues raised by SARS-CoV-2. Therefore, the use of RDT for rapid detection of the viral genome of patients in the region can be useful for early triage and rapid management.

RDT results are commonly used to determine whether individuals can enter medical institutions, social service agencies, schools, and take public transportation, etc. Two red lines on both “T” symbol and “C” symbol from the RDT result indicate an infection of SARS-CoV-2, while one red line on the “C” symbol from the result indicates no infection [12]. Individuals are only able to enter authorized places if their RDT results are negative. To facilitate this, a RDT declaration platform has been opened by the public health department to collect images of RDT results submitted by the users. Meanwhile, the fact that the RDT declaration platform cannot evaluate RDT result images. And a large

☆ This paper was recommended for publication by Zhi-Cheng Li.

* Corresponding author.
E-mail address: taotan@mpu.edu.mo (T. Tan).

https://doi.org/10.1016/j.displa.2023.102403
Received 6 January 2023; Received in revised form 7 February 2023; Accepted 19 February 2023
Available online 14 March 2023

number of test images makes it difficult for local health departments to screen them manually on a case-by-case basis. Consequently, there is an urgent need to develop a lightweight model that can be applied to a large number of samples with accurate assessment results to better understand and control the spread of SARS-CoV-2.

2. Related works

In recent years, Convolutional Neural Networks (CNNs) have been widely used in object detection tasks. CNNs have powerful feature extraction capabilities, which can extract semantic information at deeper levels, thus improving the expression ability of features and significantly improving the accuracy of localization and classification compared to traditional methods using manually extracted features. Currently, the commonly used horizontal box object detection includes Faster RCNN [13], SSD [14], YOLO [15–18], and rotation object detection includes RetinaNet [19], FCOS [20], and R^3Det [21]. These algorithms have been widely used in various downstream object detection tasks. Jin et al. [13] proposed a weakly supervised fine segmentation and lightweight Faster-RCNN for the detection of forest fire smoke regions, which achieved 99.6% detection accuracy and a real-time detection speed of only 151 ms. Loey et al. [22] proposed a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection, which solved the problem of mask-wearing detection in Covid-19 and improved the accuracy of detection. Lin et al. [23] proposed a Soft-YOLOX network specifically for garbage-clogging sewers, which solved the urban flooding problem caused by garbage accumulation. Compared with the traditional YOLOX, the proposed method improved by 2.17%. Tran et al. [24] proposed a RetinaNet-based One-stage detector for asphalt pavement crack detection considering pavement distress and surface objects, which achieved a high accuracy detection of 89.1%. Li et al. [25] proposed an optical remote sensing image ship detection method based on R^3Det, which solved the problem of small object detection in a complex environment of remote sensing images and greatly enhanced the accuracy of rotation object detection.

The existing object detection algorithms mainly belong to two types: the Two-stage object detection network, represented by RCNN, and the One-stage object detection network, represented by YOLO. The main difference between them is whether there is a region proposal process involved in the detection. Although the Two-stage network has a higher detection accuracy, it has a higher training cost and slower detection speed and requires a large dataset due to its deep network. On the other hand, the One-stage network has the advantage of fast detection speed and requires small dataset due to its shallower network and fewer parameters. However, its detection accuracy is slightly lower, and it is not good at detecting small objects. Due to privacy issues, RDT datasets are usually small, YOLOv5 can be a candidate as the basic framework of the network. The RDT result images also differ in angle and position, with large uncertainties such as complex background information and lack of differentiation between RDT result regions. Moreover, a large number of parameters make the model difficult to deploy on low-power hardware. To address these issues, this paper has the following contributions:

1. We are the first to propose a framework to deliver accurate RDT results detection with fast responses. The solution reduces the workload for human detection which could be beneficial to smart medical care.
2. We propose an improved lightweight feature extraction network, L-ShuffleNetV2, which can be widely used for feature extraction in other detection and segmentation tasks. To achieve low computation and low parameterization, we design a lightweight feature fusion module, and introduce the parameter-free NAM attention mechanism, to obtain faster and higher precision detection results.
3. We propose a novel data augmentation method, Single-Copy–Paste. The core idea of this method is to copy and paste the object on one single image, which is different from the traditional Copy–Paste method. Compared with the latter, our proposed method can achieve data augmentation without changing the data distribution.
4. We propose a lightweight rotating box object detection method to address the variable result feedback area in RDT result detection. This method can effectively detect the geometric contour and position information of the result and reduce the redundancy of information included in the RDT results.

3. Methodology

In this section, we will go through the details of the proposed LightR-YOLOv5. In the first part, we offer an overview of the network structure. And we will provide details about the proposed key modifications we made in LightR-YOLOv5. Then we will introduce data augmentation methods. Finally, we will explain the rotating box detection method.

3.1. Network architecture

In this paper, we propose a highly lightweight YOLOv5 rotating object detection network, which improves YOLOv5 horizontal box object detection and transforms it into rotated box object detection, and also conducts a lightweight study. This network shown in Fig. 1 consists of three components: Backbone, Neck, and Head. Specifically, Backbone is a feature extraction module for images and we improved and fused L-ShuffleNetV2 network in it. Neck is a feature fusion module, in which we skillfully combined the GSConv, CARAFE [26], Depth-Wise Convolution, Concat, ADD, and NAM attention. Finally, Head is the prediction output.

CARAFE is a lightweight upsampling operator that improves accuracy while requiring limited parameters and computation. We use GSConv and CARAFE in the upsampling stage to reduce the computation and efficiently fuse the information after upsampling because retaining enough a priori information can improve the upsampling effect. Both ADD and Concat are feature fusion operations, but ADD can enhance the information of features while Concat can increase the number of features. So the combination of Concat and Depth-Wise Convolution (DWConv) is used to increase the features extracted by Backbone. We add NAM parameter-free attention after the layer at the end of downsampling to increase the information weight of effective features even further.

3.1.1. Lightweight structure: L-ShuffleNetV2

Convolution Neural Network (CNN) displays significant advantages in extracting images with distinct local features, which is determined by structure of CNN. CNN is a network structure made up of convolution kernels, with the most notable feature which is its local receptive field. However, the computational load of deep convolutions is huge that will be unfriendly for model deployment. While ShuffleNetV2 [27] is a lightweight network that considers how convolution accumulation affects computational speed.

ShuffleNetV2 can balance the accuracy and speed of the model at the same time with the design of fused parallel standard convolution and the addition of depth-wise convolution, which is very friendly to the end-side deployment of model. ShuffleNetV2 has four different output channels: 0.5×, 1×, 1.5×and 2×. The output channel represents the number of feature maps, the larger the output channel has better accuracy, but also brings greater computational effort and number of parameters. By improving on ShuffleNetV2 1×, L-ShuffleNetV2 maintains the network design of ShuffleNetv2 1×but removes the 1024×1024 convolution layer and the 5×5 MaxPooling layer. The network structure is designed to give ShuffleNetV2 1×higher

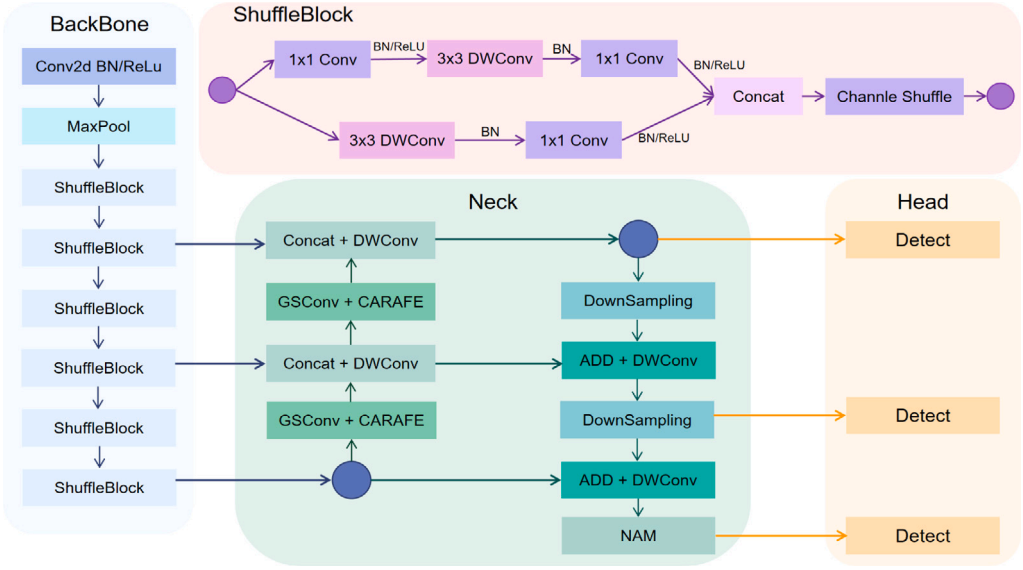

Fig. 1. LightR-YOLOv5 Architecture: Backbone is used for image feature extraction, Neck is used for the fusion of extracted features from different layers, Head is used for localization and prediction output of classification results.

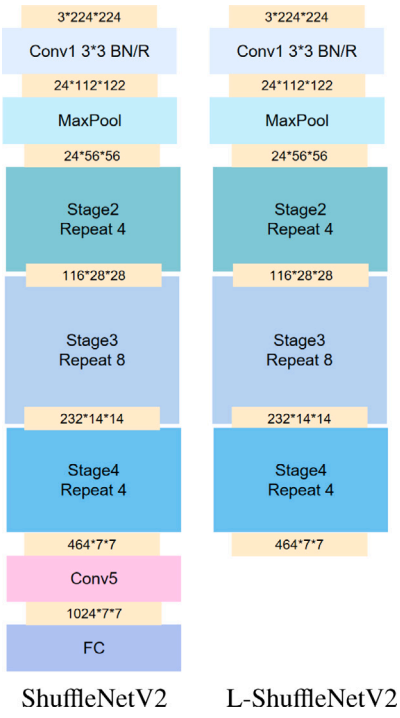

Fig. 2. Network structure of ShuffleNetV2 and L-ShuffleNetV2.

accuracy on the ImageNet. Although it leads to excellent feature extraction, it also greatly increases the computational effort and the number of parameters. We choose to remove this design to keep it lightweight. Network structure of ShuffleNetV2 and L-ShuffleNetV2 are shown in Fig. 2.

3.1.2. Lightweight convolution: Depth-wise convolution

Fig. 3 shows the process of standard convolution (SC). When a 3-channel image is input, according to the size of the output channel, the corresponding size of the filter to convolution operation is set, and then the same number of channels as the output feature map is generated. In the domain of computer vision, such convolutions become

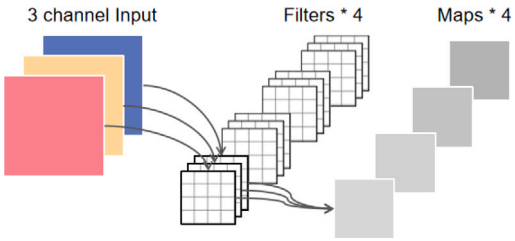

Fig. 3. Process of standard convolution (SC).

popular and has produced excellent results. An important indicator in practice is the convolution neural network-based computation of model and parameter count. Smaller models can be efficiently distributed training, reduce the weight update overhead, reduce the platform size power storage and computational power limitations, and facilitate the deployment of the model on mobile.

There have been a lot of efficient convolutions developed recently. Like depth-wise separable convolution [28], which performs convolution in stages, dilated convolution, which performs convolution with some features lost, and deformable convolution, which performs convolution with arbitrarily shaped convolution kernels. The first step of depth-wise convolution is applied as a separate convolution process with the help of the design of depth-wise separable convolution. In the depth-wise convolution process, one filter is responsible for one channel, and one channel is convolved by only one filter. When a 3-channel image is an input, the depth-wise convolution generates a feature map with the same number of channels as the previous layer, shown in Fig. 4. Compared with standard convolution, using depth-wise convolution will reduce the multiplication operation, reduce the computational complexity, and speed up the network operation.

3.1.3. Feature mixing: GSConv

Although depth-wise convolution can significantly reduce the parameters and computation of the model, it lacks feature fusion between channels because depth-wise convolution is a single-channel convolution, and the number of channels cannot be changed during the operation. To overcome the disadvantage of channel information separation during depth-wise Convolution computation, GSConv [29]

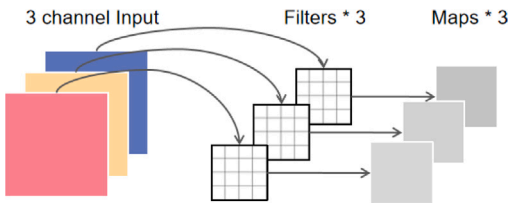

Fig. 4. Process of depth-wise convolution (DWConv).

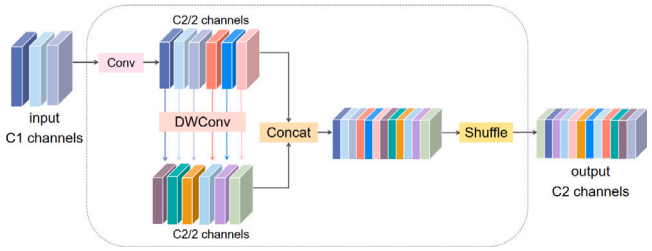

Fig. 5. GSConv: fusion of features from standard convolution and DWConv.

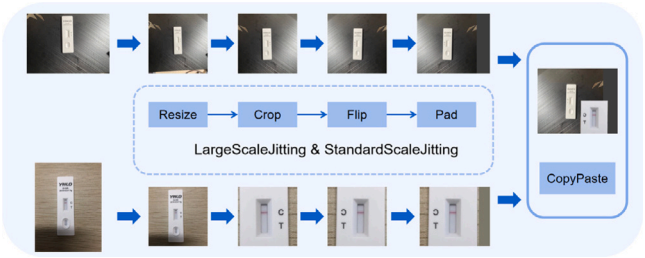

Fig. 6. Process of Copy–Paste data augmentation.

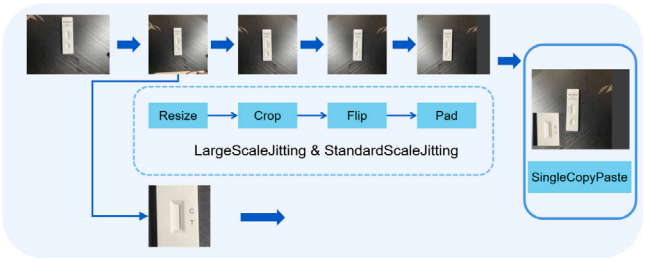

Fig. 7. Process of Single-Copy–Paste data augmentation.

employs shuffle to fuse the feature map obtained from depth-wise convolution and SC, resulting in the output of SC being completely fused into depth-wise convolution. Fig. 5 shows the structure of GSConv. The GSConv method minimizes the negative impact of depth-wise convolution defects on the model and effectively takes advantage of the low computational effort of depth-wise convolution. It has been experimentally demonstrated that GSConv improves the accuracy and speed.

3.1.4. Feature weighting: NAM attention

Attention mechanisms can help neural networks to suppress less significant features in channels and space. Many previous studies have focused on how to obtain significant features by attentional operators, such as CBAM attention [30], SE attention [31], etc. However, these works lack the consideration of the contribution factor of the weights which can further suppress the less salient features, and NAM attention [32] uses the contribution factor of weights to enhance the effect of attention. NAM attention is an efficient and lightweight attention mechanism. NAM attention takes the module integration of CBAM attention and redesigns the channel and spatial attention submodules. The NAM attention uses a sparse weight penalty, which allows these weights to be computationally more efficient while maintaining the same performance. And this approach avoids the addition of fully connected and convolutional layers as in SE attention and CBAM attention.

3.2. No impact data augmentation: Single-copy–paste

There are many challenges from datasets in practical object detection tasks, such as small object detection and long-tailed distribution of data, which refers to the fact that only a few categories contain a large number of samples while most categories contain only a small number of samples.

Data augmentation and data resampling are effective strategies to solve these problems. The method of generating new data using Copy–Paste [33] is very simple. First, randomly select two images and apply random scaling dithering and random horizontal flipping. Then, randomly select a subset of objects from one of the images and paste it onto the other image. Finally, adjust Ground Truth accordingly,show in Fig. 6. This method is a useful way to create new training samples and can help to improve the accuracy and generalizability of machine learning models. This approach significantly improves the performance of object detection models trained with supervised learning and semi-supervised learning. In the RDT results dataset, the images are single image with single label. It implies that using the Copy–Paste method may randomly paste a disjoint label into the image, and such a large number of Copy and Paste processes will lead to a shift in the distribution of the data, which results in inconsistent data distribution and leads to poor model training.

To solve this problem, we propose an improved data augmentation method, Single-Copy–Paste in Fig. 7. Specifically, when an object from one image is copied, it will not be pasted randomly to another image, but instead to a random location on the original image. This not only resolves the problem of data imbalance in the dataset but also preserves the original data distribution. As a result, the model will have increased detection accuracy.

3.3. Rotating box detection

The border labeling adopted by the object detection method should be designed according to the shape characteristics of the detected object itself. However, the original YOLOv5 project's application scenario is based on this assumption. Objects in a natural scene, which can be firmly to be included in a Horizontal Bounding Box (HBB). Moreover, due to the random angle at which the photo is taken, the subset of objects in the RDT result detection samples contains a large number of RDT equipments that can be oblique. At this time, the more accurate the labeling method, the less redundant background information is provided to the network training, and the less irrelevant the network need to detect. The benefit is to discipline the training direction of the network and reduce the convergence time.

Usually, for a rotated box we have two expressions: one is based on five parameters, namely the center point (x, y), the width and height (w, h), and the rotation angle a, as shown in Fig. 8(a). The other is based on eight parameters, namely the four vertices of the rotated box (x1, y1), (x2, y2), (x3, y3), and (x4, y4), as shown in Fig. 8(b).

Qian et al. [34] proposed five-parameter regression method performs integration leading to training instability and performance degradation due to the inherent periodicity of angles and the associated abrupt changes in width and height leading to loss discontinuity. This approach leads to inconsistent regressivity between parameters of different measurement units. In contrast, the eight-parameter regression method has parameter consistency, and this method can describe arbitrary quadrilaterals and thus be used in more complex application

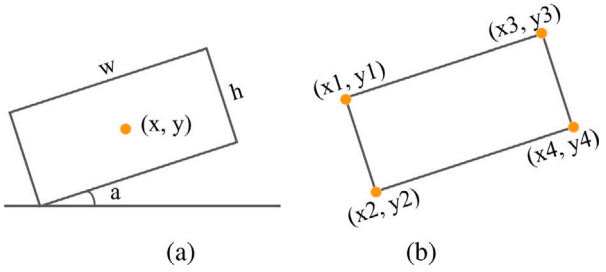

Fig. 8. Two expressions of rotated box.

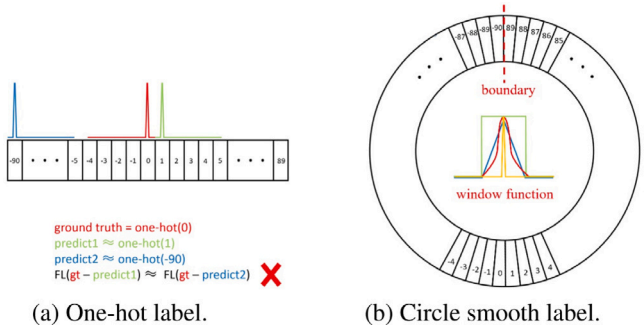

Fig. 9. Two kind of labels for angular classification.

scenarios, but it also inevitably suffers from the inherent periodicity of angles and the associated abrupt changes in width and height.

An effective way to solve this problem is to use the angle classification approach, which is to classify the entire defined range of angles into categories, such as classify one degree into one category proposed by yang et al. [35], as shown in Fig. 9(a) below.

Converting a regression problem to a classification problem is continuous to a discrete problem, and there is a loss of accuracy in this conversion process. For example, in the case of one degree as one class ($w = 1$), we cannot predict a result of 0.5 degree. Therefore, we need to calculate the maximum loss of accuracy and the average loss which follows a uniform distribution to determine how much this loss affects the final result. In order to solve this accuracy loss problem, yang et al. [35] proposed Circular Smooth Label (CSL), which is formulated as follows.

$$\mathrm{CSL}(x) = \begin{cases} g(x), & \theta - r < x < \theta + r \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$G(x)$ is the window function and the radius of the window is controlled by r.

Due to the setting of the window function, the model can measure the angular distance between the predicted label and the ground truth label, i.e., the closer the predicted value is to the true value within a certain range, the smaller the loss value is. And the problem of angular periodicity is solved by introducing periodicity, such as turning the two degrees 89 and −90 into being near neighbors. It should be noted that Circular Smooth Label is equivalent to One-hot label when the window radius of the window function is small (Fig. 9(b)). For this reason we use a combination of eight-parameter regression and CSL to improve the original YOLOv5 to make it a rotating object detector.

4. Experiment and analysis

4.1. Experimental environment and experimental setup

In this paper, the algorithm was trained on a 64-bit operating system, Ubuntu-22.04. The 4-card parallel training was conducted on Intel (R) Core (TM) i9-10980XE CPU @ 3.00 GHz and NVIDIA RTX A4000 high-performance GPU, each with a computer with 16 GB memory. The model was built and trained with Pytorch 1.10.0, CUDA 11.3, and CUDNN 8.5. Pretraining was conducted on the MS COCO dataset [36], with the initial momentum and initial learning rate set to 0.937 and 0.01, respectively. The optimization used was the stochastic gradient descent (SGD), with a batch size of 32. The model converged at the 250th epoch. Before formal training, three generations of preheating learning were carried out, with a preheating learning momentum of 0.8 and a Warmup learning rate of 0.1, to make the model gradually stabilize before formal training.

Table 1
Dataset splitting and object statistics.

Type(sheet)	Training set	Validation set	Test set	Total
	1140	253	250	1643
Positive(A)	102	15	19	136
Negative(A)	508	264	156	928
Invalid(A)	696	143	86	925

4.2. Evaluation metrics

Four evaluation metrics are used to complete the comparison experiment, including Precision, Recall, and mean Average Precision (mAP): mAP@.5 and mAP@.5:.95. The mAP@0.5 denotes the average AP calculated for all images under the three categories (negative, positive and invalid) in the RDT dataset when the ***IoU*** is set to 0.5. The mAP@.5:.95 denotes the average AP over different ***IoU*** thresholds (0.5 to 0.95, step size 0.05). ***IoU*** is used to measure the degree of overlap between the predicted box and the real box in object detection. Commonly used concepts in evaluation metrics are expressed as follows:

(1) True Positive (TP): the number of positive classes predicted to be positive classes.
(2) True Negative (TN): the number of negative classes predicted as negative classes.
(3) False Positive (FP): the number of negative classes predicted as positive classes, which is the number of detection errors.
(4) False Negative (FN): the number of positive classes predicted as negative classes, which is the number of missed detections.

Hence, Precision is defined as

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is defined as

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

mAP is defined as

$$mAP = \int_0^1 P_r(R_e) d R_e \tag{4}$$

where $P_r(R_e)$ is the curve constructed from the detection Precision and Recall at different ***IoU*** thresholds.

4.3. Dataset

In this paper, we used a self-collected dataset for training, validation and testing. As shown in Table 1, the final dataset we obtained contains 1643 RDT result images divided into three categories: negative, positive, and invalid. We used 1140 of these images as the training set, 53 images as the validation set, and the remaining 250 images as the test set. The samples from our dataset contains single background or complex-background information, and object sizes have different image scales.

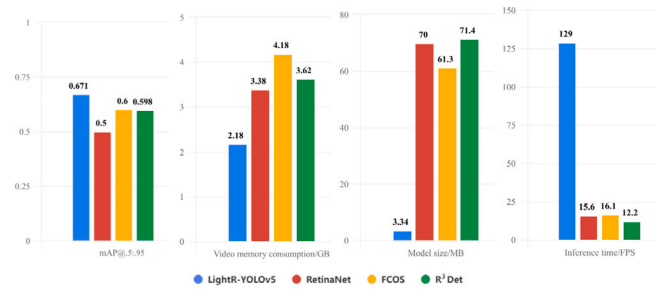

Fig. 10. LightR-YOLOv5 compared to RetinaNet, FCOS, and R^{3}Det in terms of different performance.

Table 2
Performance of different lightweight backbone.

Backbone	Params(M)	FLOPs(G)	FPS
MobleNetV3	1.93	3.5	98.2
ShuffleNetV2	0.71	1.0	126.3
GhostNet	1.39	3.3	116.4
PP-LCNet	0.95	2.0	112.5

4.4. Experimental result

To validate the improved model, we carried out a series of comparison experiments on our dataset. We evaluate the experimental results based on benchmarks of mean Average Precision (mAP), Model size, Video memory consumption, and Inference time. The model accuracy metric (mAP@.5 and mAP@.5:.95) is a common measure of object detection accuracy. The speed of model forward inference is measured by inference time. The number of parameters and the amount of computation are also compared in experiments.

We compare LightR-YOLOv5 to the RetinaNet, FCOS, and R^{3}Det models. As shown in Fig. 10, our improved YOLOv5 model has achieved remarkable improvements in terms of model size, model complexity, and inference speed. The model size and complexity of LightR-YOLOv5 are about one-twentieth of the other models, and the inference speed is 8.5 times faster. Additionally, the video memory consumption of the LightR-YOLOv5 model is almost half that of the other models, while the mAP is significantly higher than that of the other three models. These results demonstrate that LightR-YOLOv5 is capable of effectively extracting features from RDT result images.

In addition, we recorded the variations in Precision, Recall, mAP@0.5, and mAP@.5:.95 for each iteration of the model training process. The blue, red, yellow, and green curves in the following Fig. 11 illustrate the accuracy comparison of our proposed LightR-YOLOv5, RetinaNet, FCOS, and R^{3}Det, respectively. Finally, our model achieves significantly faster convergence and has higher Precision, Recall, and mAP values.

To verify the rationality and necessity of each part of the improved model, ablation experiments were conducted. We separately evaluated each part of the model, including the L-ShuffleNetV2, NAM attention, and Single-Copy–Paste method. And gradually added these modules to experimentally evaluate the YOLOv5 model.In CNN-based network design, there are many lightweight network models designed. MobileNetV3 [37] uses complementary search techniques to reduce model computation. GhostNet [38] reduces computational cost by generating more feature maps using Ghost module operations, and PPLCNet [39] is an improved approach based on CPU. To compare the performance of different networks in our method, we designed comparison experiments with different models as Backbone, and the experimental results are shown in Table 2.

The results of the ablation experiments are shown in Table 3. Although the addition of L-ShuffleNetV2 loses certain mAP values, the addition enables the network algorithm to perform effective feature extraction with less computation and number of parameters. To compensate for the loss of mAP values, we introduced the NAM attention, which suppressed the less significant weights and finally improved the mAP@.5:.95 by 2.5%. Additionally, based on the characteristics of the dataset, we propose an improved Copy–Paste data augmentation method called Single-Copy–Paste. The mAP@.5:.95 metric of the model is improved by 3.3% by the data augmentation method. Therefore, we merge L-ShuffleNetV2, NAM attention, and Single-Copy–Paste together into the YOLOv5 model.

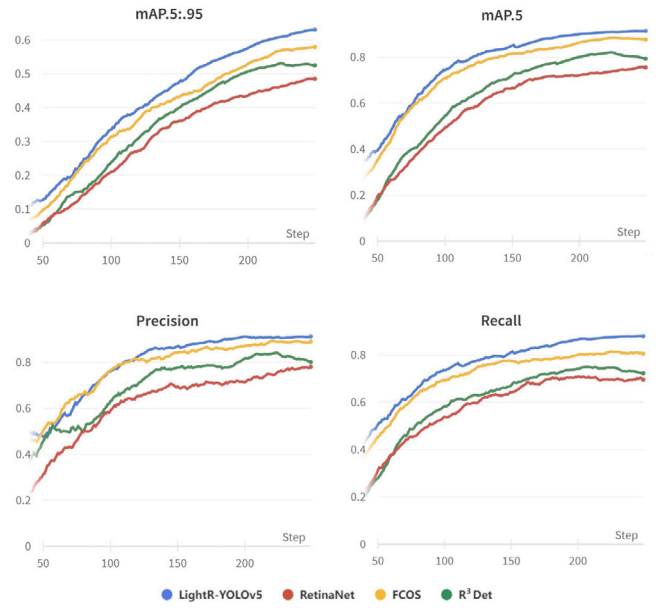

Fig. 11. LightR-YOLOv5 compared to RetinaNet, FCOS, and R^{3}Det of metrics performance on training 250th epochs.

4.5. Visualization of detection basis

Since the deep learning framework is more a black-box to determine health status, understanding machine learning models is essential in improving model credibility and providing transparency in scrutinizing prediction results. To verify whether the LightR-YOLOv5 model can effectively locate and classify objects, we visualize the key focused regions extracted by LightR-YOLOv5 using Grad-weighted class activation mapping (Grad-CAM [40]).

As shown in Fig. 12, LightR-YOLOv5 locates and classifies the ROI in the region where the object is located in the fourteenth and seventeenth layers of the feature extraction network in a relatively scattered manner. While in the twenty-first layer, LightR-YOLOv5 incorporates the NAM attention, so the localization and classification of the object will be more focused on one region, thus achieving good localization and classification results. These visualization results demonstrate the effectiveness and reasonableness of our improved network for RDT image feature extraction.

5. Discussion

To pursue a fast response of automated detection and classification of the SARS-CoV-2 antigen-detection rapid diagnostic tests, we propose a lighter design while maintaining the accuracy. We adopted L-ShufleNetV2 as the Backbone of the model and the extracted effective features are not greatly affected by less model parameters. The combination of lightweight Depth-Wise Convolution and deep fusion features of GSConv can also show comparable results with SC with less computation. Ultimately, L-ShuffleNetV2 and the lightweight Neck

Table 3
Ablation experiments of LightR-YOLOv5 on different methods.

Model	L-ShuffleNetV2	NAM	Single-CP	map@.5:.95	Model size(MB)	Inference time(ms)
YOLOv5(original)				69%	14.6	27.2
YOLOv5(L-ShuffleNetV2)	✓			61.3%	2.03	10.2
YOLOv5(NAM)	✓	✓		63.8%	2.03	10.8
$YOLOv5^{ours}$(**Single-CP**)	✓	✓	✓	67.1%	2.03	11.0

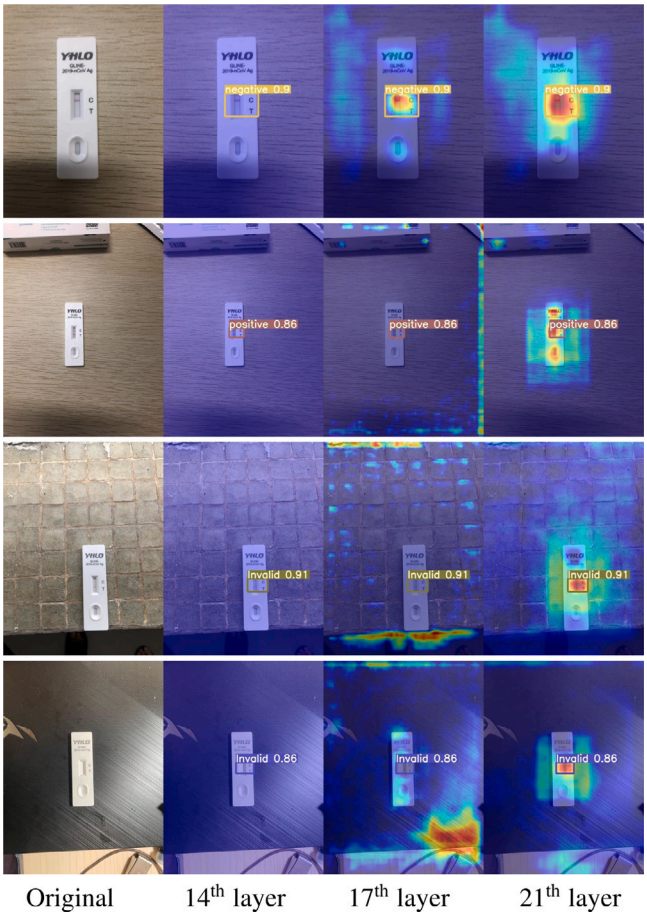

Fig. 12. Visualization of feature map weights extracted by different feature extraction layers.

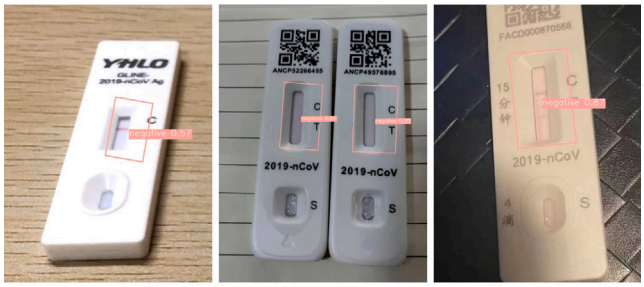

Fig. 13. Model recognition error results.

design can significantly reduce the network redundancy with lower the computational complexity (the model size is reduced to one-seventh of the original size and the inference speed is increased to the double of the original). To improve the accuracy of the lightweighted model, we use transfer learning on MS COCO and add NAM attention mechanism.

In small and unbalanced datasets, applying data augmentation methods can effectively increase the number of training samples and thus improve the generalization ability of the model. The RDT result dataset is a special dataset, which is usually a single image with a single category. Using copy-paste does increase the data samples for RDT training, When RDT image is pasted it will randomly cause a new subset of the pasted object subset to be disjoint from the original image class, eventually causing a shift in the distribution of the entire dataset, which can be effectively mitigated by using Single-Copy–Paste. Single-Copy–Paste belongs to pixel-level data augmentation method, which is usually considered as a more advanced way of data augmentation because it operate on the whole object subset without changing the whole image. In RDT results detection, we also use Mosaic [18] and Cutout [41], which also increase the data samples for training but do not take into account the size relationship between the detection region and the whole image according to RDT results.

During our testing, there are some samples that were not effectively identified and localized. Fig. 13(a) shows that the model is not very compact in locating the entire RDT result detection area under extremely distorted shooting angle. Fig. 13(b) shows that the test sample is incorrectly identified as negative due to a large data category imbalance in the data, resulting in a small bias in the identification of model. Fig. 13(c) shows that the sample is not effectively identified in the weak positive sample. For Fig. 13(a) we believe that the output can be effectively improved by adding some correction methods or adding suppression detection boxes. For Fig. 13(b) and Fig. 13(c) adding more training samples and balancing the number gap between different categories is a fundamental solution.

Currently, the dataset we used contains only 1643 images, of which the number of positive images only accounts for 8.28% of the full dataset. It is far from enough for further improvement of model accuracy. Moreover, there are many RDT kits produced by different manufacturers in the market, and the kits produced by different manufacturers will bring about image differences in the object detection task. In future work, we will include more manufacturers' RDT result images, so that LightR-YOLOv5 can be widely used for most RDT result detection while expanding the number of samples.

SARS-CoV-2 antigen-detection rapid diagnostic test (RDT) plays an important role in the management and surveillance of large-scale novel coronavirus outbreak. RDT is a fast, sensitive, and accurate detection method that can help individuals understand if they have been infected with the SARS-CoV-2 virus, allowing them to take effective preventive or therapeutic measures to further protect public health and safety. Furthermore, RDT can also aid researchers in understanding the transmission mechanisms of the virus and its relationship with the population, thus helping to better control the development of the epidemic. Since traditional manual evaluation methods are difficult to accurately and quickly evaluate large-scale RDT results data, it is important to apply lightweight deep learning models to detect RDT results. In a wider and more far-reaching way, our proposed LightR-YOLOv5 can be applied not only in RDT result detection, but also in computer vision related tasks with fast migration.

6. Conclusions

In this work, we propose a lightweight RDT results rotation detector, which adopts the new L-ShuffleNetV2 feature extraction network and

feature fusion technique with integrated NAM attention, which can effectively reduce the redundant information in the result detection region and improve the detection accuracy, while reducing the number of model parameters and computational complexity. For the special RDT result detection dataset, we also introduce a new data augmentation method Single-Copy–Paste, which can effectively increase the data samples and further improve the model accuracy without changing the data distribution. It is shown that LightR-YOLOv5 outperforms RetinaNet, FCOS and R^3Det networks on this dataset. We demonstrate the impact of each of the proposed contributions in ablation experiments and visually verify the effectiveness of these contributions on the network focus using Grad-weighted class activation mapping.

CRediT authorship contribution statement

Rongsheng Wang: Methodology, Coding, Writing – original draft. **Yaofei Duan:** Data curation, Writing – review & editing. **Menghan Hu:** Investigation, Concept. **Xiaohong Liu:** Investigation, Concept. **Yukun Li:** Writing – review & editing. **Qinquan Gao:** Data curation. **Tong Tong:** Data curation. **Tao Tan:** Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

We thank all the anonymous reviewers for their valuable comments and constructive suggestions. This work is supported by Macao Polytechnic University Grant (P106/DEI/2022).

References

[1] S.N. Kales, A. Fernández-Montero, J. Argemi, J.A. Rodriguez, J.A. Rodríguez, A.H. Ariño, A.H. Ariño, L. Moreno-Galarraga, Validation of a rapid antigen test as a screening tool for SARS-CoV-2 infection in asymptomatic populations. Sensitivity, specificity and predictive values, Soc. Sci. Res. Netw. (2021) http://dx.doi.org/10.2139/ssrn.3814781.

[2] R. Kubina, A. Dziedzic, Molecular and serological tests for COVID-19. A comparative review of SARS-CoV-2 coronavirus laboratory and point-of-care diagnostics, Diagnostics (ISSN: 2075-4418) 10 (6) (2020) http://dx.doi.org/10.3390/diagnostics10060434, URL https://www.mdpi.com/2075-4418/10/6/434.

[3] G. Lippi, B.M. Henry, M. Plebani, An overview of the most important pre-analytical factors influencing the clinical performance of SARS-CoV-2 antigen rapid diagnostic tests (Ag-RDTs), Clin. Chem. Lab. Med. (CCLM) (2022) http://dx.doi.org/10.1515/cclm-2022-1058.

[4] M. Wölfl-Duchek, F. Bergmann, A. Jorda, M. Weber, M. Müller, T. Seitz, A. Zoufaly, R. Strassl, M. Zeitlinger, H. Herkner, H. Schnidar, K. Anderle, U. Derhaschnig, Sensitivity and specificity of SARS-CoV-2 rapid antigen detection tests using oral, anterior nasal, and nasopharyngeal swabs: a diagnostic accuracy study, Microbiol. Spectrum 10 (1) (2022) http://dx.doi.org/10.1128/spectrum.02029-21, URL https://journals.asm.org/doi/abs/10.1128/spectrum.02029-21.

[5] T. Tan, B. Das, R. Soni, M. Fejes, H. Yang, S. Ranjan, D.A. Szabo, V. Melapudi, K. Shriram, U. Agrawal, L. Rusko, Z. Herczeg, B. Darazs, P. Tegzes, L. Ferenczi, R. Mullick, G. Avinash, Multi-modal trained artificial intelligence solution to triage chest X-ray for COVID-19 using pristine ground-truth, versus radiologists, Neurocomputing (ISSN: 0925-2312) 485 (2022) 36–46.

[6] R.M. Pereira, D. Bertolini, L.O. Teixeira, C.N. Silla Jr., Y.M. Costa, COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios, Comput. Methods Programs Biomed. 194 (2020) 105532.

[7] R. Soni, T. Tan, G.B. Avinash, D. Pati, H. Krupakar, V.R. Saripalli, Synthetic training data generation for improved machine learning model generalizability, 2022, US Patent App. 16/999, 665.

[8] T. Tan, G.B. Avinash, M. Fejes, R. Soni, D.A. Szabó, R. Mullick, V. Melapudi, K.S. Shriram, S.R. Ranjan, B. Das, et al., Multimodality image processing techniques for training image data generation and usage thereof for developing mono-modality image inferencing models, 2022, US Patent App. 17/093, 960.

[9] T. Tan, B. Das, R. Soni, M. Fejes, S. Ranjan, D.A. Szabo, V. Melapudi, K.S. Shriram, U. Agrawal, L. Rusko, Z. Herczeg, B. Darazs, P. Tegzes, L. Ferenczi, R. Mullick, G. Avinash, Pristine annotations-based multi-modal trained artificial intelligence solution to triage chest X-ray for COVID-19, 2020.

[10] V. Ramesh, B. Rister, D.L. Rubin, COVID-19 lung lesion segmentation using a sparsely supervised mask R-CNN on chest X-rays automatically computed from volumetric CTs, 2021.

[11] S. Hu, Y. You, S. Zhang, J. Tang, C. Chen, W. Wen, C. Wang, Y. Cheng, M. Zhou, Z. Feng, et al., Multidrug-resistant infection in COVID-19 patients: A meta-analysis, J. Infection 86 (1) (2023) 66–117.

[12] G.L. Salvagno, B.M. Henry, G. Bongiovanni, S.D. Nitto, L. Pighi, G. Lippi, Positivization time of a COVID-19 rapid antigen self-test predicts SARS-CoV-2 viral load: a proof of concept, Clin. Chem. Lab. Med. (CCLM) (2022) http://dx.doi.org/10.1515/cclm-2022-0873.

[13] J. Pan, X. Ou, L. Xu, A collaborative region detection and grading framework for forest fire smoke using weakly supervised fine segmentation and lightweight faster-RCNN, Forests (ISSN: 1999-4907) 12 (6) (2021) http://dx.doi.org/10.3390/f12060768, URL https://www.mdpi.com/1999-4907/12/6/768.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[15] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.

[16] M. Loey, G. Manogaran, M.H.N. Taha, N.E.M. Khalifa, Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection, Sustainable Cities Soc. (ISSN: 2210-6707) 65 (2021) 102600, http://dx.doi.org/10.1016/j.scs.2020.102600, URL https://www.sciencedirect.com/science/article/pii/S2210670720308179.

[17] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[18] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[20] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[21] X. Yang, J. Yan, Z. Feng, T. He, R3det: Refined single-stage detector with feature refinement for rotating object, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35 no. 4, 2021, pp. 3163–3171.

[22] M. Loey, G. Manogaran, M.H.N. Taha, M. Taha, N.E.M. Khalifa, Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection, Sustainable Cities Soc. (2020) http://dx.doi.org/10.1016/j.scs.2020.102600.

[23] J. Lin, C. Yang, Y. Lu, Y. Cai, H. Zhan, Z. Zhang, An improved soft-YOLOX for garbage quantity identification, Mathematics (ISSN: 2227-7390) 10 (15) (2022) http://dx.doi.org/10.3390/math10152650, URL https://www.mdpi.com/2227-7390/10/15/2650.

[24] V.P. Tran, T.S. Tran, J.-J. Lee, H.J. Lee, K.D. Kim, J. Baek, T.T. Nguyen, T.T. Nguyen, T.T. Nguyen, One stage detector (RetinaNet)-based crack detection for asphalt pavements considering pavement distresses and surface objects, J. Civ. Struct. Health Monit. (2020) http://dx.doi.org/10.1007/s13349-020-00447-8.

[25] J. Li, Z. Li, M. Chen, Y. Wang, Q. Luo, A new ship detection algorithm in optical remote sensing images based on improved R3Det, Remote Sens. (ISSN: 2072-4292) 14 (19) (2022) http://dx.doi.org/10.3390/rs14195048, URL https://www.mdpi.com/2072-4292/14/19/5048.

[26] J. Wang, K. Chen, R. Xu, Z. Liu, C.C. Loy, D. Lin, Carafe: Content-aware reassembly of features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3007–3016.

[27] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.

[28] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[29] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, Q. Ren, Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles, 2022, arXiv preprint arXiv:2206.02424.

[30] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[31] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[32] Y. Liu, Z. Shao, Y. Teng, N. Hoffmann, NAM: Normalization-based attention module, 2021, arXiv preprint arXiv:2111.12419.

[33] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E.D. Cubuk, Q.V. Le, B. Zoph, Simple copy-paste is a strong data augmentation method for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2918–2928.

[34] W. Qian, X. Yang, X. Yang, X. Yang, S. Peng, Y. Guo, Y. Guo, C. Yan, J. Yan, Learning modulated loss for rotated object detection, 2019, http://dx.doi.org/10.1609/aaai.v35i3.16347, arXiv: Computer Vision and Pattern Recognition.

[35] X. Yang, J. Yan, Arbitrary-oriented object detection with circular smooth label, 2020, arXiv: Computer Vision and Pattern Recognition.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft COCO: Common objects in context, Null (2014) http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[37] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[38] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.

[39] C. Cui, T. Gao, S. Wei, Y. Du, R. Guo, S. Dong, B. Lu, Y. Zhou, X. Lv, Q. Liu, et al., PP-LCNet: A lightweight CPU convolutional neural network, 2021, arXiv preprint arXiv:2109.15099.

[40] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization, IEEE Int. Conf. Comput. Vis. (2016) http://dx.doi.org/10.1007/s11263-019-01228-7.

[41] T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout., 2017, arXiv: Computer Vision and Pattern Recognition.