IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 4, 2021, accepted March 11, 2021, date of publication March 17, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066538

A Novel Detection Framework About Conditions of Wearing Face Mask for Helping Control the Spread of COVID-19

JUN ZHANG[1], FEITENG HAN[1], YUTONG CHUN[1], AND WANG CHEN[2]
[1]School of Management and Engineering, Capital University of Economic and Business, Beijing 100070, China
[2]School of Information, Renmin University of China, Beijing 100872, China

Corresponding authors: Feiteng Han (h_feiteng@163.com) and Yutong Chun (chunyutong@cueb.edu.cn)

This work was supported by the National Social Science Fund of China under Grant 20BGL001.

ABSTRACT Properly wearing a face mask has become an effective way to limit the COVID-19 transmission. In this work, we target at detecting the fine-grained wearing state of face mask: face without mask, face with wrong mask, face with correct mask. This task has two main challenging points: 1) absence of practical datasets, and 2) small intra-class distance and large inter-class distance. For the first challenging point, we introduce a new practical dataset covering various conditions, which contains 8635 faces with different wearing status. For the second challenging point, we propose a novel detection framework about conditions of wearing face mask, named Context-Attention R-CNN, which enlarge the intra-class distance and shorten inter-class distance by extracting distinguishing features. Specifically, we first extract the multiple context feature for region proposals, and use attention module to weight these context feature from channel and spatial levels. And then, we decoupling the classification and localization branches to extract more appropriate feature for these two tasks respectively. Experiments show that the Context-Attention R-CNN achieves 84.1% mAP on our proposed dataset, outperforming Faster R-CNN by 6.8 points. Moreover, Context-Attention R-CNN still exceed some state-of-the-art single-stage detectors.

INDEX TERMS COVID-19, conditions of wearing face mask, detection framework, new dataset.

I. INTRODUCTION

COVID-19 has seriously threatened the safety of human life, and quickly spread to the majority of countries worldwide. Fortunately, the surgical face masks could cut the spread of coronavirus [1]. The World Health Organization (WHO) recommends that people should wear face masks if they have respiratory symptoms, or they are taking care of the people with symptoms [2]. Furthermore, many public service providers ask customers to use the service only if they wear masks [3]. Therefore, detections on conditions of wearing face mask have become a valuable and meaningful computer vision task.

However, most of current researches and mask detection applications target at solving the two-classes detection problem: masked face and not-masked face, which ignores the problem of whether a face mask is worn correctly. The lack of research will lead to virus spreading by people who wear face mask in an incorrect way.

To address the limitations existed in previous works, we target at detecting more fine-grained conditions of wearing face mask: face without mask, face with a wrong mask, face with correct mask. For such a purpose, a practical valuable related dataset is needed, which is another limitation in current works. Therefore, we propose a clean, practical and challenging dataset. Specifically, we collect the most raw images from the MAFA [4], and supplement some mask incorrect images(rare in the real world) by downloading images from the internet. In order to make the dataset more various and practical, we classify the MAFA dataset to several subclasses, and collect the raw images from each subclass. Besides, we label objects in raw images with bounding boxes and clean the dataset by manual inspection.

Detecting conditions of wearing face mask is a challenging task. The reason is that large appearance variations in intra-classes and small apperance variations between different

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

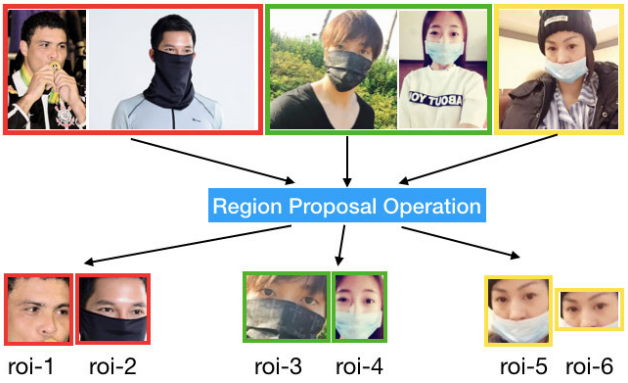

FIGURE 1. **Red:Faces without face masks; Green:Faces with correct face masks; Yellow: Face with an incorrect face mask; The Region-of-Interests generated by region proposal operation show the challenge of this task: different classes are very highly similar, same classes are very different.**

classes, which leads to difficulties in learning distinguishing features for detection(See Fig. 1). From the example images shown in Fig. 1, we can find that the local Region-of-Interests(RoIs) in different classes may have no difference, but have relatively obvious distinction when we consider the surrounding context feature(Fig. 1 roi-2 vs. roi-3). On the contrary, the surrounding context features of some RoIs (Fig. 1 roi-4 vs. roi-5) have no significant difference, but have obvious difference we when consider the inside local spatial region. Based on this observation, there are two problems should be solved: (1) how to extract context features? (2) how to balance these context features from multiple dimensions for generating more distinguishing features? In this paper, we propose the Context-Attention R-CNN, which aims to solve the above problems and further achieve accurate detection on conditions of wearing face mask. Firstly, we introduce a context feature extractor to get the multiple receptive-field context feature for the RoIs. Secondly, we use an attention mechanism to explicitly balance the multiple context features. Specifically, channel attention is used to weight the multiple context features while latter spatial attention is adopted to balance the spatial features for each context feature. Due to the difference between localization and classification task, we decouple the conventional coupled architecture for more accurate performance.

The contributions of this paper are summarized as follows:

1) We propose an accurate conditions of wearing face mask detection framework named Context-Attention R-CNN, which extend the two-classes face mask detection problem into triple-classes detection problem and detect conditions of wearing face mask more fine-grained.
2) We present a challenging, practical, and fine-grained dataset about conditions of wearing face mask, which can be used as a valuable data resource for studying new detectors.
3) An effective practice in controlling spread of the COVID-19. The Context-Attention R-CNN is effective and practical in detecting properly wearing face mask and preventing the transmission of COVID-19 infection.

The rest of this paper is organized as follows: related work is reviewed in Section II. The details of data construction and proposed model are shown in Section III. Experiments are presented in Section IV. Finally, this paper is concluded with a summary in Section V.

II. RELATED WORK

A. FACE DETECTION

Generic face detection. As one of the early most famous face detector,the Viola-Jones proposed the boosted-based cascade architecture with simple yet fast Haar features [5]. The work of [6] introduced an ensemble of decision trees to detect faces. It implemented a fast detection speed through comparing pixel intensities between different nodes. A unified face detector was presented by [7], which combined detection and alignment in a model. In [8], a hierarchical DPM-based framework was developed to achieve face detection and keypoint localization.

In addition to the above conventional face detectors, the convolutional neural network(CNN) based models show a remarkable progress in recent years. In [9], the proposed face detector adopted the feature aggregation model [10], while the features were extracted by CNN. The authors of [11] proposed attribute-related CNN to predict the confidences for candidate windows. Recently, a region-based CNN face detector was proposed in [12], which also took the contextual information into account. The work of [13] developed a novel grid loss to solve the occlusion issues in face detection task. For the same purpose, [4] proposed locally linear embedding module to get a similarity-based descriptor. Combined with dictionaries mechanism, it achieved an accurate performance on occlusion face task. Besides, a novel supervised transformer network was designed to relieve the pose variations problem [14]. [15] design a cascaded framework, which consist of three stage,to progressively improve the face detection performance. The authors of [16] proposed an multiple-branches framework to focus on the accurate detection of small faces. The framework fuses the feature maps of different branches for detecting hard small faces. [17] introduces a receptive field block to extract the robust feature map. Combined with cascaded CNN, it achieves continuous improvements on multiple related dataset.

Masked face detection. The masked face detection refer to the face mask detection task, which is emerged along with the outbreak of COVID-19. Its goal is to check a person whether wear face mask to cut the spread of the COVID-19 virus. [18] proposed RetinaMask detector based on the single-stage generic detector named RetinalNet. The RetinaMask also took the context information into account, and try to extract the robust feature. The work of [19] presented a hybrid detection framework which combined the deep neural network and conventional machine learning algorithms. [20] implement

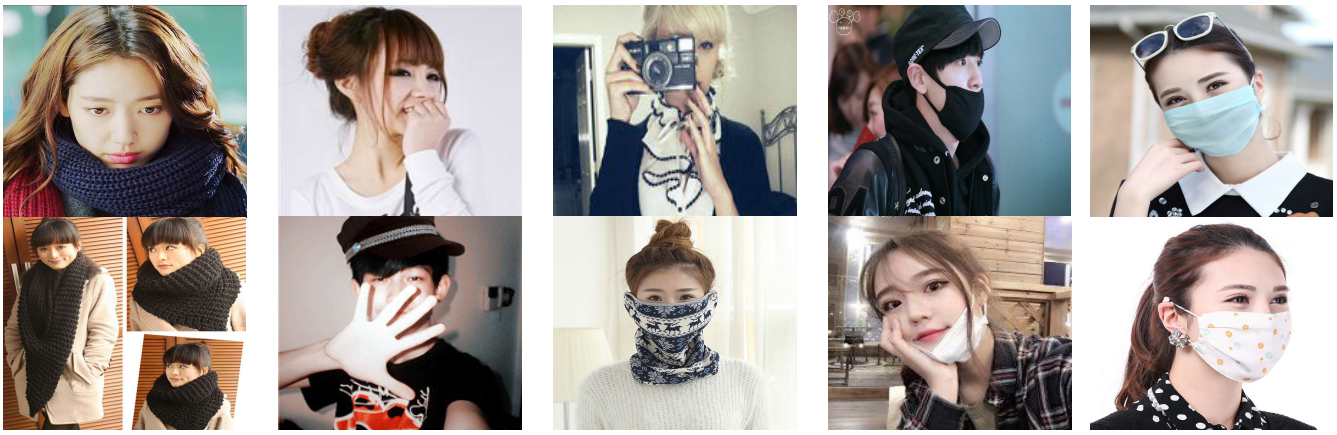

FIGURE 2. **Examples of our dataset.**

TABLE 1. **Comparisons of previous works.**

Method	Judge Whether Wearing Facemask	Judge Whether Wearing Correctly	Advantages	Disadvantages
RetinaMask [18]	YES	NO	1. Strong ability to extract robust feature. 2. Efficient. 3. High precision and recall on specific dataset.	1. More training time. 2. More complicated post-processing and sensitive to the related hyper-parameter.
Hybrid model [19]	YES	NO	1. Ensemble various classical classifiers for better performance. 2. High precision on specific dataset. 3. Simple to deploy.	1.Restricted to the scene that one person per image, and the face covering the most region of image. 2. Training process is relatively complex.
ResNet-YOLOv2 [20]	YES	NO	1.More reasonable number of anchor boxes. 2. High speed.	1. Medium performance. 2. The confidence of the predicted results is relatively low.
RTFR [21]	YES	NO	1. Easy to deploy	1.Non-End-to-End. 2. Strong rely on pre-processing. 3. Only local feature was used, which will lead to errors in some complex cases.
SCRNet [22]	YES	YES	1. Good performance on low quality image.	1. Complicated training process. 2. Tend to make mistakes when local face was cropped.
Ours	YES	YES	1. Strong ability to extract the distinguish feature map. 2. High average precision on the practical dataset. 3. End-to-end pipeline, easy to implement (including training and test) and deploy.	1.More hyper-parameters(i.e. pre-defined anchor boxes).

the medical face mask detection by combining the ResNet-50 and the classical single-stage generic detector YOLOv2. It improves detection performance by using mean IoU to set the proper number of anchor boxes. [21] first divide the input image to several segments.Then use the VGG16 model to predict the final results.

Facemask-wearing conditions detection. Reference [22] divides the task to four steps: preprocessing image, crop face regions, super-resolution operation, predict the final condition. The main novel point is that it apply the super resolution to improve performance for low-quality images. However, it is a non-end-to-end framework, which will make the training process complex.In addition, only the local feature was used when predict the final results.

The above first two face detection tasks target at detecting the clean face or face mask, and it did not care about fine-grained conditions of wearing face mask. Therefore, the related researches can provide some heuristics for our task, but are not suitable to directly detect the conditions of wearing face mask. To show the advantages and disadvantages of our proposed and the most related works(maske face detection and facemask-wearing conditions deteciton), we summary these works in Table. 1.

From the above development of face detectors, we can find that the CNN-based face detectors have outperformed traditional detectors. In addition, some state-of-the-art face detectors are inspired by the generic object detectors. Inspired by the above development, we propose the Context-Attention R-CNN based on the CNN generic object detectors. We will review the generic object detectors in the next subsection.

B. GENERIC OBJECT DETECTION

Two-stage detectors. The two-stage detectors consist of two main stages: the first stage try to provide the candidate region proposals, while the second stage predict the classification scores and localization offsets for these proposals. The R-CNN [23] could be treated as the first most classical region-based detetion framework. The RoIs in R-CNN was proposed by the Selective Search [24]. For each RoI, the R-CNN used the CNN-based extractor to extract the feature map. In order to saved redundant computation, Fast R-CNN [25] and SPPNet [26] extracted the feature for a whole image, and shared the whole feature for all RoIs. For a faster speed, the Faster R-CNN [27] introduced a region proposal network to automatically propose the candidate region proposals,

and shared the feature with the second stage. Through the fully convolutional architecture in RoI-head, the R-FCN [28] further improved the inference speed. In order to implement the high quality object detector, Cascade R-CNN was introduced to predict the classification score and regression offsets with a cascade way [29]. In order to relieve the sensitivity of anchor box hyper-parameter, [30] introduce a dimension-decomposition region proposal network (DeRPN) to independently predict the width and height of object.

Single-stage detectors. The single-stage detectors discard the region proposal stage, and directly predict the classification scores and localization offsets for anchor boxes.YOLO [31] adopted the predefined grids on image as prior regions to predict the classification scores and localization offsets in a single stage. SSD [32] designed a CNN-based architecture to directly predict the final results on different layer feature maps. It still used the prior anchors as the candidate proposals, but discarded the time-consuming operations existed in first stage of two-stage detectors. Later, YOLO9000 [33] was introduced which still used the prior anchors to improve the accuracy of framework. In addition, some training tricks were used to boost better performance. For addressing the imbalance problem existed in one-stage detectors, RetinaNet [34] proposed a novel loss, named focal loss, to automatically adjust the importance of training samples.Similar to two-stage detectors, RefineDet [35] introduced Anchor Refinement Module(ARM) to filter some easy negative samples, and designed Object Detection Module(ODM) to predict the final scores and offsets. Different from two-stage detectors, it did not conduct the time-consuming RoI-wise operation in the first stage. Therefore, it was treated as single-stage detectors.

Many recent face detectors are inspired by the generic object detection algorithms. Though the single-stage detectors were more efficient, two-stage detectors usually show a superior generalization ability when transfer to specific detection task. Therefore, we propose the Context-Attention R-CNN based on the generic two-stage region-based detector.

III. METHODOLOGY

A. BACKGROUND

The purpose of our paper is to detect the fine-grained conditions of wearing face mask, and further make contributions to the limitation of COVID-19. This is a challenging task. Firstly, the practical related dataset is absent. Secondly, there is a fact that subtle appearance variations between different classes and large apperance variations in intra-classes. To address these challenging issues, we first present a clean and practical dataset about conditions of wearing face mask. Based on the proposed dataset, we develop a detection framework, named Context-Attention R-CNN, which can achieve high quality detection performance by capturing the distinguishing feature.

B. DATA CONSTRUCTION

In this section, we will describe details of the related dataset construction. Firstly, we introduce the process of data collection. And then, the dataset annotation and statistics will be described.

1) DATA COLLECTION

Our dataset contains a total of 4672 images, in which 4188 raw images are selected from the MAFA public dataset [4]. The MAFA is a masked face dataset, which contains 30, 811 images and 35, 806 masked faces. Faces in the dataset have various degrees of occlusion and mask types, which includes man-made objects with pure color, man-made objects with complex textures, or face covered by hand etc. Although raw images in the MAFA are diverse, the dataset only focused on detecting masked face, which is not suitable for our target.

We create a new dataset which specificlly contributes to detect conditions of wearing face mask. In order to make the collected images as diverse as possible, we first divide the MAFA raw images into five types: clean face, hand-masked face, non-hand-masked face, masked incorrect face, masked correct face. Examples of the five types are shown in the Fig. 2. Based on our definition, non-hand-masked faces contain a face masked by man-made objects with complex textures like scarf, sunglasses or mobile phone etc.The first three types can be seen the *without_mask* class in our task. Therefore, we ensure that the selected face images cover all of the first three types. The rest 484 images of our dataset are downloaded from the internet, which mainly are *mask_incorrect* instances due to this class is rare in MAFA.

2) DATA ANNOTATION

We use the popular open-source labeling tool (labelImg) [36] to annotate the raw images. For the annotaion format, we choose the PASCAL VOC format [37], which is a popular, classic annotation format for object detection task. It has a series of evaluation tools for evaluating the model performance. As for the label names, we use the *without_mask*, *mask_correct* and *mask_incorrect* to represent the face without a face mask, a face with a face mask in a correct way and a face with a face mask in an incorrect way respectively.

3) DATA STATISTICS

In this section, we statistic the distribution of bounding box ground truth from the three aspects: Firstly, as shown in Fig. 3(a), statistics represent the distribution of classes based on the number of classes in an image. We notice that the number of images who have three classes are the least. Secondly, as shown in Fig. 3(b), we divide the size of bounding boxes into small, medium and large based on the standard of Microsoft COCO [38]. The number of small size is much less than the other sizes.Thirdly, as shown in Fig. 3(b), the number of *mask_incorrect* is much less than the others classes,which

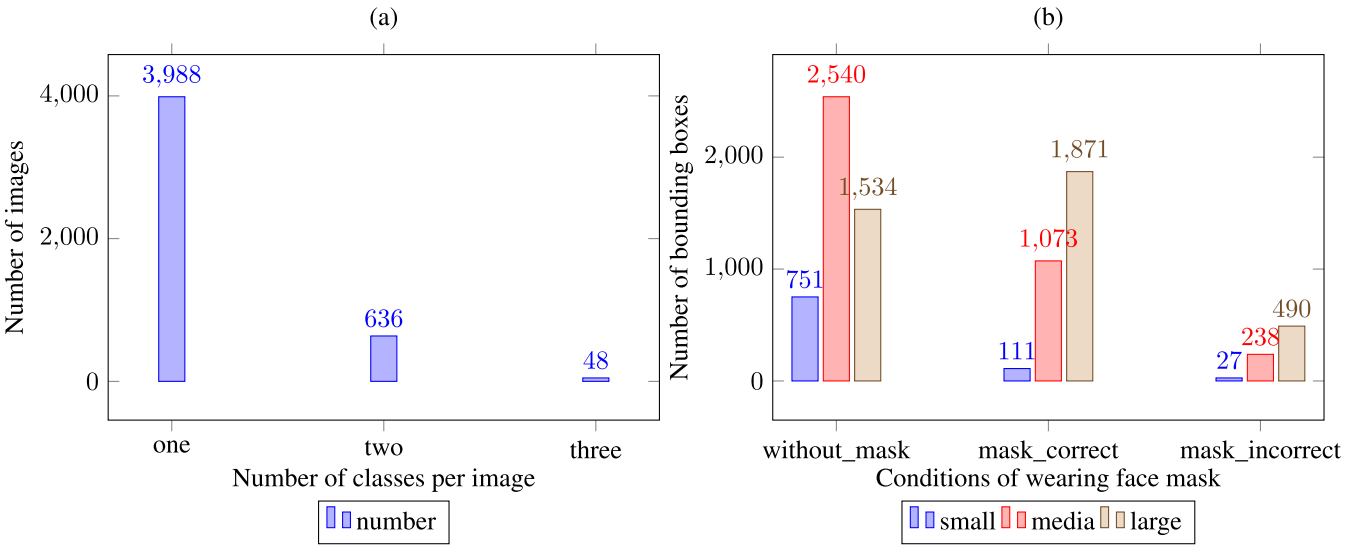

FIGURE 3. **Data statistics of our dataset.**

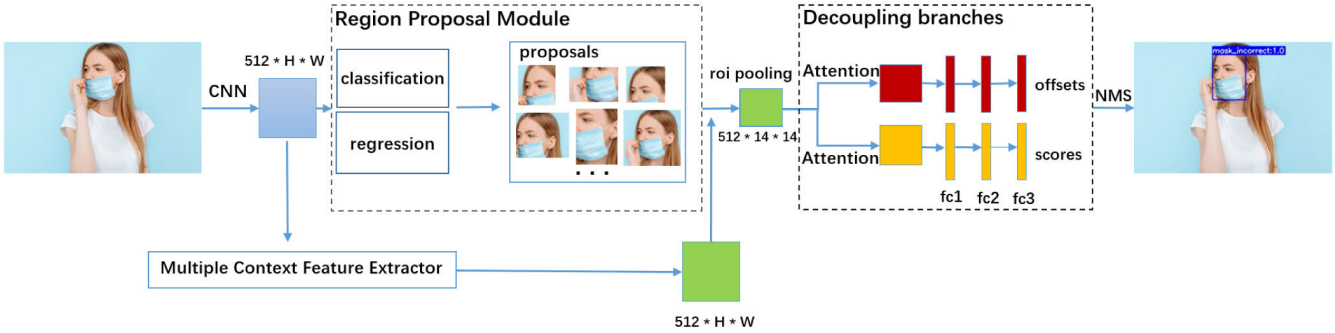

FIGURE 4. **The overall pipeline of Context-Attention R-CNN. Based on the backbone feature extracted by CNN, the region proposal module outputs some class-agnostic region proposals. Simultaneously, the multiple context feature extractor extracts context features. After that, the context features of each region proposal can be generated by the roi-pooling operation. Finally, the decoupling classification-localization branches, which consists of an effective attention module and three fully-connected layers, predict the classification scores and localization offsets. After Non-Maximum Suppression(NMS), the final results can be obtained.**

is a reflection of reality. Based on the Fig. 3, we can know that our dataset is imbalance in the number of class per image, distribution of size and the number of bounding box in an image. It is a trully reflect of reality situation and make more challenge for the detection task.

C. CONTEXT-ATTENTION R-CNN

Based on region-based two-stage detection framework, we proposed a novel detection framework about conditions of wearing face mask, named Context-Attention R-CNN, which contains three novel components. Firstly, a Multiple Context Feature Extractor is used to extract multiple context information for RoIs. Secondly, the decoupling classification-localization branches are proposed to predict the accurate class scores and localization offsets. Thirdly, channel-spatial attention is used as a key component in decoupling branches. The overall pipeline is shown in Fig. 4.

1) MULTIPLE CONTEXT FEATURE EXTRACTOR

We implement the multiple context feature extractor(MCFE) by adopting the atrous convolution [39] with different dilation rates. Comparing to convolution with different kernel size, the atrous convolution have fewer parameters and still can extract the feature map with multiple Receptive-Field context information by adopting multiple different dilation rates. With a proper padding size, we can make the feature map outputted by convolution kernel with different dilation rates have the same size. Then we combine the feature maps from different atrous convolutional layers by cross-channel concatenation. If the size of input feature map is $C \times H \times W$, and the number of different dilation rates is K, we set $\frac{C}{K}$ convolution kernels for each dilation convolution. Thus the output of the MCFE is $K \times \frac{C}{K} \times H \times W$. For example, we using the VGG16 [40] as the backbone to extract the basis feature(i.e.,*conv5-3*) in this paper, the size of basis feature is $512 \times H \times W$. With the dilation rates are set to 1, 3, 5,

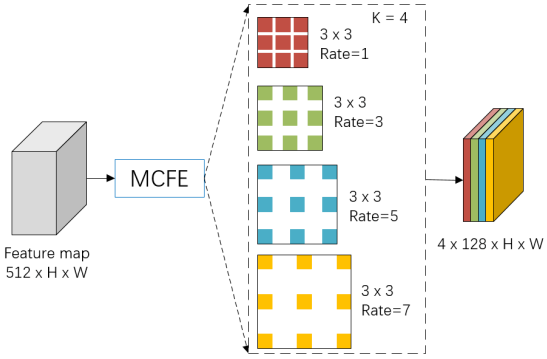

FIGURE 5. **Multiple context feature extractor.**

$7(K = 4)$, the size of context feature map extracted by MCFE is $4 \times 128 \times H \times W$. (see Fig. 5)

2) DECOUPLING CLASSIFICATION-LOCALIZATION BRANCHES

Many existed two-stage detectors shared the feature for localization and classification branches. However, the localization mainly focus on location information while the classification branch try to capture the high level semantic representation. The shared feature is intrinsic not optimal for this two specific tasks.

In this paper, we decoupling the classification and localization branches, share the same architecture but do not share the parameter. For each branch architectures, an attention module first capture the effective context feature from the roi-pooling feature map. Then a series of fully-connected layers is used to predict the final offsets or class scores. Next, we will describe the detail of branch architecture.

Multiple Context Feature Attention. The Multiple Context Feature Attention (MCFA) consists of channel-attention(CA) and spatial-attention(SA), whose target is to capture the effective context feature and depress some noise context features from two aspects of semantic and spatial.

Channel attention. The channel-wise attention module is used to weight the channel features for balancing the different Receptive-Field(RF) context features and different channel features inside each RF feature. Let roi-pooling context feature maps as $\mathbf{F_c} \in \mathbb{R}^{C\times R\times R}$. $\mathbf{F_c} = [\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_k}]$, $\mathbf{f_i} \in \mathbb{R}^{(C/K)\times R\times R}$ represent a context feature block with specific dilation rate, where C, K, R refers to the number of channel, dilation rates and roi-pooling size respectively. Then, the weighted context feature can be formulated as:

$$\mathbf{F_{CA}} = \mathbf{CA}(\mathbf{F_c}) \otimes \mathbf{F_c} \tag{1}$$

where **CA** denotes channel-wise attention module which output weights for all channels, and $\otimes$ represents channel-wise multiplication. For details of **CA**, we adopt the architecture used in CBAM [41],

$$\mathbf{CA}(\mathbf{F_c}) = \sigma(fc2(relu(fc1(AveragePooling(\mathbf{F_c}))) + fc2(relu(fc1(MaxPooling(\mathbf{F_c})))) \tag{2}$$

where σ denotes sigmoid function, and fc operation is implemented by a convolutional layer with 1×1 kernel size.

Spatial attention. The spatial attention aims to assign more attention to important spatial feature and depress the noise feature. Instead of applying a parameter-shared spatial attention for all channels of CA feature, we adopt K spatial attention modules for K channel feature blocks respectively. Let multiple RF context feature weighted by CA as $\mathbf{F_{CA}} \in \mathbb{R}^{C\times R\times R}$. $\mathbf{F_{CA}} = [\mathbf{f'_1}, \mathbf{f'_2}, \ldots, \mathbf{f'_k}]$, $\mathbf{f'_i} \in \mathbb{R}^{(C/K)\times R\times R}$ represents a weighted context feature block with specific dilation rate. Then the CA context feature weighted by SA can be formulated as:

$$\mathbf{F_{SA}} = \mathbf{Concatenate}(\mathbf{SA_1}(\mathbf{f'_1}) \otimes \mathbf{f'_1}, \mathbf{SA_2}(\mathbf{f'_2}) \otimes \mathbf{f'_2}, \ldots, \mathbf{SA_k}(\mathbf{f'_k}) \otimes \mathbf{f'_k}) \tag{3}$$

where $\mathbf{SA_i}$ denotes spatial-wise attention module which output weights for $\mathbf{f'_i}$. $\otimes$ represents spatial-wise multiplication and **Concatenate** denotes channel-wise concatenate. For details of **SA**, we adopt the architecture used in CBAM [41],

$$\mathbf{SA_i}(\mathbf{f'_1}) = \sigma(Conv(Concatenate(Mean(\mathbf{f'_i}), Max(\mathbf{f'_i})))) \tag{4}$$

where σ denotes sigmoid function. *Mean* and *Max* means operation that extracts pooling features across the channel.

Note that our work explores the usage of attention mechanism on this task, and adopts the existing spatial and channel attention proposed in CBAM [41]. Other attention formulas may have a better potential, but are not the study scope of this work.

Prediction of offsets and scores After MCFA, we exploit a series of fully-connected layers to predict the final localization offsets and classification scores. Let $\mathbf{F^r_{SA}}$, $\mathbf{F^c_{SA}}$ as the context feature are weighted by MCFA for localization and classification branch respectively. Then, the final prediction of localization offsets and classification scores can be formulated as:

$$offsets = fc3(relu(fc2(relu(fc1(\mathbf{F^r_{SA}}))))) \tag{5}$$

$$scores = softmax(fc3(relu(fc2(relu(fc1(\mathbf{F^c_{SA}})))))) \tag{6}$$

where the fc_i in different branches are not shared. Following Faster R-CNN, we use the same transformed offsets expression. Let us denote region proposal to be $[x_p, y_p, h_p, w_p]$ and final predicted box to be $[x, y, h, w]$. The offsets equal to $[t_x, t_y, t_h, t_w]$, where

$$t_x = (x - x_p)/w_p \tag{7}$$

$$t_y = (y - y_p)/h_p \tag{8}$$

$$t_w = \log(w/w_p) \tag{9}$$

$$t_h = \log(h/h_p) \tag{10}$$

IV. EXPERIMENTS

We will describe the detail of our experiments in this section. Firstly, we introduce the dataset and evaluation criterion in Section IV-A. Then the implement details of Context-Attention R-CNN and training process are shown in Section IV-B. The Section IV-C reports the performance

TABLE 2. Comparisons with state-of-the-art detectors.

Detector	Backbone	without_mask	mask_correct	mask_incorrect	mAP
Faster R-CNN [27]	ResNet-50	0.738	0.858	0.722	0.773
Faster R-CNN [27]	ResNet-50-FPN	0.789	0.889	0.807	0.828
Cascade R-CNN [29]	ResNet-50	0.809	0.893	0.794	0.832
RetinaNet [34]	ResNet-50-FPN	0.780	0.862	0.800	0.814
SSD300 [32]	VGG16	0.749	0.880	0.809	0.812
SSD512 [32]	VGG16	0.771	0.882	0.823	0.826
Context-Attention R-CNN(ours)	VGG16	0.785	0.887	0.852	**0.841**

TABLE 3. Ablation studies of three core components in our Context-Attention R-CNN. The check mark means which components an experiment has.

Multiple Context Feature Extractor	Attention Module	Decoupling Branches	without_mask	mask_correct	mask_incorrect	mAP
✓	✓	✓	**0.785**	**0.887**	**0.852**	**0.841**
	✓	✓	0.775	0.884	0.830	0.830
✓		✓	0.778	0.888	0.797	0.821
✓	✓		0.779	0.890	0.842	0.837

about our method and state-of-the-art works. Finally, the ablation study is described in Section IV-D.

A. DATASET AND EVALUATION

Basic Dataset. All experiments in this paper are conducted on the proposed dataset. As mentioned above, this dataset contains 4672 images which includes three types of object: *without_mask*, *mask_correct*, *mask_incorrect*.We randomly split it into two subsets: training set(3504 images) and test set(1168 images).

Evaluation. For evaluation, we choose the standard mean average precision (mAP) [37], which is widely used for evaluating the performance of object detectors. In this paper, we set the IoU (Intersection-over-Union) thresh as 0.5.

B. IMPLEMENTATION DETAILS

Network architecture. We use the VGG16 [40] pre-trained on ImageNet [42] as the backbone network. In order to better display the advantages of spatial attention, we set the size of roi-pooling as 14×14. In decoupling branches, we set the dimension of fully-connected layers as 2048. In addition, the kernel size of spatial attention is 3×3.

Image Preprocessing. In train and test phrase, the image is first resized to 600×1000. Then pixel values of the resized image are normalized to [-1,1]. Additionally, the horizontally-flip operation is used to augment the training samples.

Training. We train Context-Attention R-CNN with batch size 1 for 11 epochs. A stochastic gradient descent (SGD) optimizer with 0.9 momentum was used. The start learning rate is 0.001, and is decreased by a factor 10 after 10 epoch.

C. COMPARISON WITH STATE-OF-THE-ART DETECTORS

In this section, we compare our method with the state-of-the-art generic object detectors. In order to make a comprehensive comparison, we report the results of single-stage detectors and two-stage detectors. The results are shown in Table. 2. Without bells and whistles, our Context-Attention R-CNN outperforms these classical generic object detectors. Specifically, our Context-Attention R-CNN achieves 84.1% mAP with VGG16 backbone. Compared with two-stage detectors, our Context-Attention R-CNN outperforms the ResNet-50 Faster R-CNN by 6.8 points. As a two-stage detectors, ResNet-50 Cascade R-CNN can be seen a strong baseline. Context-Attention R-CNN still achieves 0.9 points higher than ResNet-50 Cascade R-CNN. As for single-stage detectors, SSD512 shows the best performance, but it is still 1.5 points lower than Context-Attention R-CNN. Some visual examples generated by Context-Attention R-CNN can be seen in Fig. 6.

Some phenomena should be noticed. Specifically, the performance rank in the generic detection task (such as COCO, VOC) is changed when transfer into our task. Therefore, the task-related design is important. In addition, compared with single-stage detectors, the two-stage detectors still show more robust ability. It may be due to the imbalance issues existed in single-stage detectors.

D. ABLATION EXPERIMENTS

To understand the effectiveness of each component in Context-Attention R-CNN, we conduct the ablation studies whose results are shown in Table. 3. Specifically, we study three components: multiple context feature extractor, attention mechanism and decoupling branches.

Multiple Context Feature Extractor. Without MCFE, the explicit context information cannot be extracted, and the latter module will act on the raw roi-pooling feature. We can find that it decreases the mAP from 84.1% to 83.0%. Therefore, even though the roi-pooling implicitly contain some context information and the attention mechanism can

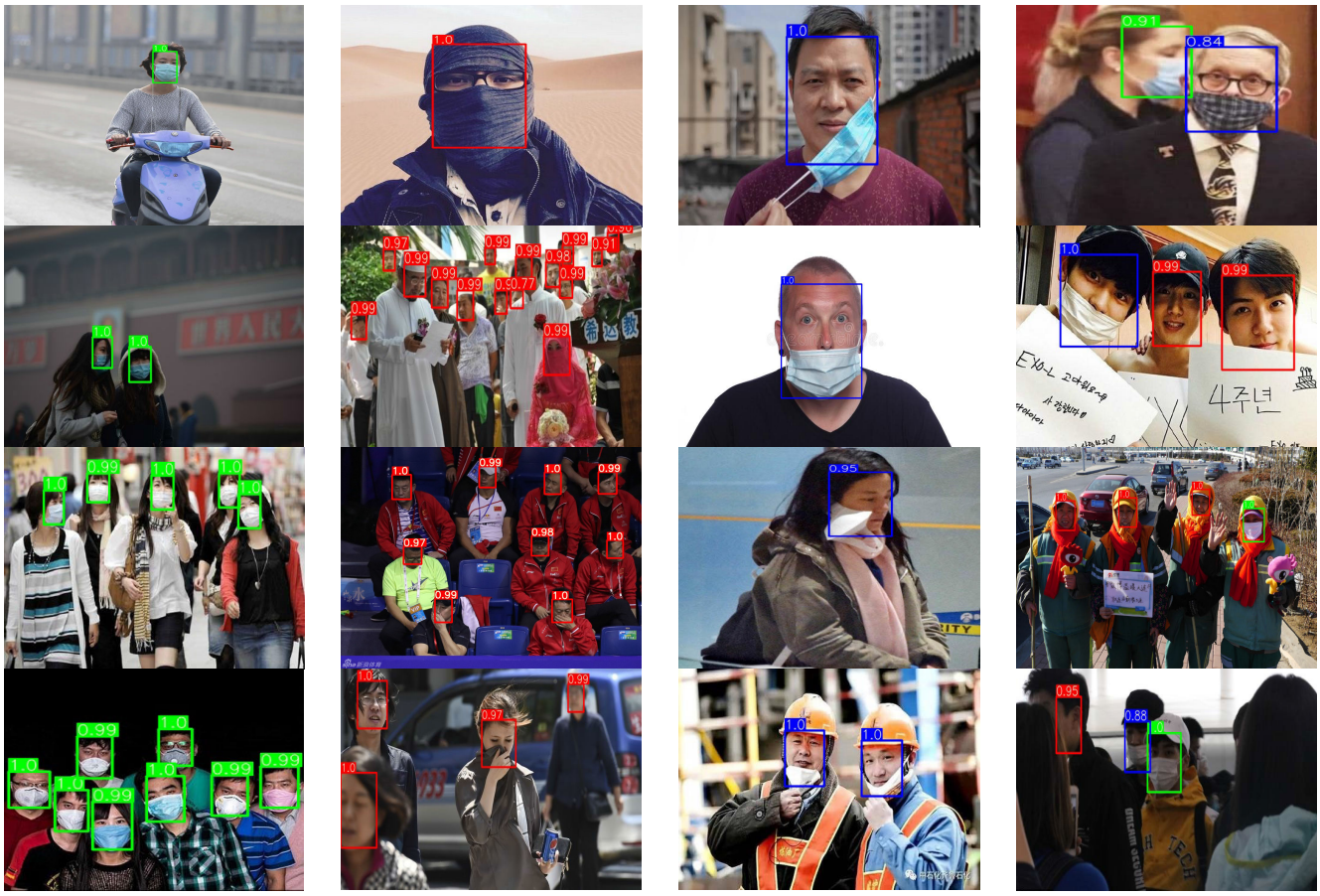

FIGURE 6. Some visual examples generated by Context-Attention R-CNN. The Green bounding boxes represent faces with correct face masks; The Red bounding boxes represent faces without face masks; The Blue bounding boxes represent faces with an incorrect face masks.

further strengthen the representation, the performance is still decreased without the explicit context features.

Attention Mechanism. Attention module can bring 2.0 points higher mAP than the architecture without attention. For *mask_incorrect* class, the attention module brings 5.5 points improvement. The *mask_incorrect* class sometimes is highly similar to *mask_correct*.It suggests that the attention module is a very important component and plays an important role in extracting the distinguishing feature.

Decoupling Branches. Compared with the conventional shared branches, the decoupling branches can improve the mAP from 83.7% to 84.1%. This indicates that it is better to extract the different feature map for different task branch. In this paper, we decouple the branches but share the same network architecture. It is a noticed direction that explore the different architectures for localization-classification branches.

V. CONCLUSION

In this paper, we first propose a practical and challenging dataset, which aims to reflect the conditions of wearing face mask in the era of COVID-19. Then we analyze the main challenging points in this task. Based on these analyses, we further develop Context-Attention R-CNN, a framework to detect conditions of wearing face mask, which contains three novel points: multiple context feature extractor, decoupling branches and attention module. With these components, the Context-Attention R-CNN brings significant improvements for region-based detector. The extensive experiments show that Context-Attention R-CNN outperforms many state-of-the-art detectors, including two-stage detectors and single-stage detectors. We believe that this dataset and Context-Attention R-CNN can make contributions for preventing the COVID-19 virus from spreading. Besides the coronavirus, our work is applicable to protect against other infectious diseases which can be spread by such things as coughing, sneezing, or even speaking at close range. In the future work, we will explore the imbalance problems and a better attention architecture for more accurate detection on conditions of wearing face mask. In addition, there was always one problem with CNN-based detectors: sensitive to hyper parameter.Therefore, the hyper parameter optimization(e.g. anchor box) is another important research direction, which can relieve some issues araised by hyper parameter.

APPENDICES

See Table 4 and Algorithm 1.

TABLE 4. Experimental Setup.

Development Environment	Details
CPU	Intel(R) Xeon(R) CPU E5-1603 0 @ 2.80GHz
GPU	GeForce GTX TITAN X Memory: 12G
Nvidia Driver Version	418.56
CUDA Version	10.1.105
Pytorch	1.6.0
Python	3.7.9
Linux Version	Ubuntu 16.04

Algorithm 1 Context-Attention R-CNN

Input: An image I

Output: A set of predicted object coordinate B, A set of predicted object classification scores C

Generate backbone feature b_f by the CNN f_{cnn} (such as VGG, ResNet etc.): $b_f = f_{cnn}(I)$;
Generate proposals P and objectness scores S by the region proposal module RPM: $P, S = RPM(b_f)$;
Get N proposals P_N and corresponding scores S_N processed by Non-Maximum Suppression NMS operation: $P_N, S_N = NMS(P, S)$
Get the multiple context feature mc_f by multiple context feature extractor $MCFE$: $mc_f = MCFE(b_f)$;
Initialize $B = \varnothing$, $C = \varnothing$
for each $p_n \in P_N$ **do**
    Get roi pooling feature rp_f by roi pooling operation RPO: $rp_f = RPO(p_n, mc_f)$;
    if branch is classification **then**
        Apply channel-attention CA and spatial-attention SA:$c_f = SA(CA(rp_f))$;
        Get the final classification score c by fully-connected layers fcs : $c = softmax(fcs(c_f))$;
        $C = C.append(c)$
    end if
    if branch is localization **then**
        Apply channel-attention CA and spatial-attention SA:$l_f = SA(CA(rp_f))$;
        Get the final offsets o by fully-connected layers fcs : $o = fcs(l_f)$;
        Get the final coordinate b by the offsets transform funciton $trans\_fun$:$b = trans\_fun(p_n, o)$;
        $B = B.append(b)$;
    end if
end for
Get the final predicted B and C: $B, C = NMS(B, C)$;
return B, C

REFERENCES

[1] N. Leung, D. Chu, E. Shiu, K.-H. Chan, J. Mcdevitt, B. Hau, H.-L. Yen, Y. Li, D. Ip, J. S. Peiris, W.-H. Seto, G. Leung, D. Milton, and B. Cowling, "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature Med.*, vol. 26, pp. 676–680, May 2020, doi: 10.1038/s41591-020-0843-2.

[2] S. Feng, C. Shen, N. Xia, W. Song, M. Fan, and B. J. Cowling, "Rational use of face masks in the COVID-19 pandemic," *Lancet Respiratory Med.*, vol. 8, no. 5, pp. 434–436, May 2020.

[3] Y. Fang, Y. Nie, and M. Penny, "Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis," *J. Med. Virology*, vol. 92, no. 6, pp. 645–659, Jun. 2020.

[4] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 426–434.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Feb. 2001, p. I-511.

[6] N. Markuš, M. Frljak, I. Pandžić, J. Ahlberg, and R. Forchheimer, "A method for object detection based on pixel intensity comparisons," in *Proc. 2nd Croatian Comput. Vis. Workshop*. Zagreb, Croatia: Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, 2013. [Online]. Available: https://arxiv.org/abs/1305.4537

[7] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8694, Jul. 2014, pp. 109–122.

[8] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.

[9] B. Yang, J. Yan, Z. Lei, and S. Li, "Convolutional channel features," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 82–90.

[10] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[11] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3676–3684.

[12] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Proc. Comput. Vis. Pattern Recognit.*, Aug. 2017, pp. 57–79.

[13] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischof, "Grid loss: Detecting occluded faces," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9907, Oct. 2016, pp. 386–402.

[14] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9909, Oct. 2016, pp. 122–138.

[15] R. Qi, R.-S. Jia, Q.-C. Mao, H.-M. Sun, and L.-Q. Zuo, "Face detection method based on cascaded convolutional networks," *IEEE Access*, vol. 7, pp. 110740–110748, 2019.

[16] S. Luo, X. Li, R. Zhu, and X. Zhang, "SFA: Small faces attention face detector," *IEEE Access*, vol. 7, pp. 171609–171620, 2019.

[17] X. Li, Z. Yang, and H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," *IEEE Access*, vol. 8, pp. 174922–174930, 2020.

[18] M. Jiang and X. Fan, "RetinaMask: A face mask detector," 2020, *arXiv:2005.03950*. [Online]. Available: https://arxiv.org/abs/2005.03950

[19] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, Jan. 2021, Art. no. 108288.

[20] M. Loey, G. Manogaran, M. H. N. Taha, and 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," *Sustain. Cities Soc.*, vol. 65, Feb. 2021, Art. no. 102600.

[21] S. V. Militante and N. V. Dionisio, "Real-time facemask recognition with alarm system using deep learning," in *Proc. 11th IEEE Control Syst. Graduate Res. Colloq. (ICSGRC)*, Aug. 2020, pp. 106–110.

[22] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19," *Sensors*, vol. 20, no. 18, p. 5236, Sep. 2020.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014.

[24] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[25] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[28] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*. [Online]. Available: http://arxiv.org/abs/1605.06409
[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[30] L. Xie, Y. Liu, L. Jin, and Z. Xie, "DeRPN: Taking a further step toward more general object detection," in *Proc. AAAI Conf. Artif. Intell.*, Nov. 2018, pp. 9046–9053.
[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[32] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9905, Oct. 2016, pp. 21–37.
[33] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[35] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[36] Tzutalin. (2015). *LabelImg*. [Online]. Available: https://github.com/tzutalin/labelImg
[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Zürich, Switzerland: Springer, 2014, pp. 740–755.
[39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Nov. 2016, pp. 1–13.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556
[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2018, pp. 3–19.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

JUN ZHANG received the Ph.D. degree in computer science from Université Paris Dauphine, Paris, France, in 1990. He is currently a Professor of Management Science and Engineering with the Capital University of Economics and Business. His current research interest includes artificial intelligence and its applications.

FEITENG HAN received the M.S. degree from Capital University of Economics and Business, China, in 2017, where he is currently pursuing the Ph.D. degree. His current research interests include object detection, image classification, and semantic segmentation.

YUTONG CHUN was born in Beijing, China, in 1994. He is currently pursuing the Ph.D. degree with the School of Management and Engineering, Capital University of Economics and Business, Beijing. His research interests include machine learning and big data analytics.

WANG CHEN was born in Hebei, China, in 1992. He received the B.S. and M.S. degrees from the Capital University of Economics and Business, China, in 2015 and 2017, respectively. He is an Engineer with the School of Information, Renmin University of China. His research interests include machine learning and computer software and theory.