Journal of Software Engineering and Applications, 2023, 16, 1-19
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Advanced Face Mask Detection Model Using Hybrid Dilation Convolution Based Method

Shaohan Wang[1,2,3], Xiangyu Wang[1*], Xin Guo[4]

[1]COSCO SHIPPING Technology Co., Ltd., Shanghai, China
[2]School of Naval Architecture, Ocean & Civil Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]Shanghai Ship and Shipping Research Institute, Shanghai, China
[4]Marine Traffic Safety and Application Laboratory, College of Merchant Marine, Shanghai Maritime University, Shanghai, China
Email: *wang.xiangyu@coscoshipping.com

How to cite this paper: Wang, S.H., Wang, X.Y. and Guo, X. (2023) Advanced Face Mask Detection Model Using Hybrid Dilation Convolution Based Method. *Journal of Software Engineering and Applications*, **16**, 1-19.
https://doi.org/10.4236/jsea.2023.161001

Received: December 19, 2022
Accepted: January 28, 2023
Published: January 31, 2023

Abstract

A face-mask object detection model incorporating hybrid dilation convolutional network termed ResNet Hybrid-dilation-convolution Face-mask-detector (RHF) is proposed in this paper. Furthermore, a lightweight face-mask dataset named Light Masked Face Dataset (LMFD) and a medium-sized face-mask dataset named Masked Face Dataset (MFD) with data augmentation methods applied is also constructed in this paper. The hybrid dilation convolutional network is able to expand the perception of the convolutional kernel without concern about the discontinuity of image information during the convolution process. For the given two datasets being constructed above, the trained models are significantly optimized in terms of detection performance, training time, and other related metrics. By using the MFD dataset of 55,905 images, the RHF model requires roughly 10 hours less training time compared to ResNet50 with better detection results with mAP of 93.45%.

Keywords

Face Mask Detection, Object Detection, Hybrid Dilation Convolution, Computer Vision

1. Introduction

On January 30, 2020, the World Health Organization (WHO) declared the novel coronavirus outbreak as public health emergency, after a few months the outbreak of neo-coronavirus pneumonia constituted a global "pandemic". The WHO has shown that the transmission route of the new coronavirus is mainly through respiratory droplets and close contact, and research has shown that medical masks can effectively minimize the spread of the new coronavirus [1].

It is proved that by correctly wearing masks can effectively lower the spreading of coronavirus by more than 70% [2]. By monitoring people's mask wearing status (*i.e.*, properly wearing, not wearing, or improperly wearing) can reduce the risk of epidemic transmission effectively by shortening or cutting off the transmission chain of the virus, thus preventing the spreading of infections which are usually increased exponentially [3].

Face mask detection, as an important computer vision task to assist people worldwide to counter the epidemic, though sometimes seems to be an easy task, it is facing many challenges in real world scenarios [4]: face angles, varying face size scales, facial expressions, facial occlusions, illumination brightness, mask appearances, and mask positions. For example, side faces images are missing more than half of the facial feature information, a long distance from the camera position leads to fewer feature information, and low illumination brightness or special mask appearance leads to difficulty in obtaining the relative position feature information of face and mask. The above challenges from face mask detection can increase the difficulty of mask detection and reduce the detection accuracy [5] significantly. Current research has shown that the use of supervised learning methods can have significantly better accuracy than self-supervised learning [6].

The performance of supervised learning depends highly on the quality of the dataset being used. The existing dataset contains a great number of mask-wearing image data, but those images are lack variation in scale and illumination. The limitations of the dataset cause the algorithm to fail to detect small targets, severe occlusions, and extreme angles which are very common to see in real world.

There are three main contributions to this paper:

1) A wider perceptual field is adopted to highlight the image features retained by the convolution process of the algorithm, which significantly improves the detection accuracy and efficiency of the algorithm in complex environments.

2) A dataset with more comprehensive category labels, richer data scenes, and more diverse image sizes was constructed. This dataset will be openly accessible to all future research workers for rapid deployment of mask detection subtasks during the New Crown outbreak and in all possible future scenarios.

3) A benchmark model was constructed to facilitate subsequent research workers to test the optimization of their algorithms on the dataset provided in this paper.

The rest of this paper is organized as follows: the second section of this paper comprehensively analyses the previous main research, and section 3 summarizes major datasets related to Face Mask and describes in detail the characteristics and labels of the datasets, the dataset enhancement methods, and why these data enhancement methods are used. The fourth section describes the object detection algorithm in detail and the methodology in this work that can be used to optimize the current object detection field. The content of section 5 gives how the experiment was conducted, including model training with MFD and LMFD datasets followed by the validation process with images on the Internet. Section

6 discusses the performance of the experiment and the possibility of using other machine learning methods such as clustering as pre-processing technique to boost the efficiency of the work. Section 7 provides public links to the datasets in this paper, as well as the open source address of the code. At the end of this paper, the eighth section concludes the whole mask detection model along with the possible future work directions.

2. Related Work

In recent years, computer vision has been used in a variety of fields such as transportation, security monitoring, finance, medical, and industrial productions. It has been changing the way people live and work. Thanks to the rapid development of computational power and advancing cameras the use of deep learning models with cameras is becoming possible. The convolutional neural network (CNN) is a classical model in computer vision with four major directions: image classification, target detection, semantic segmentation, and instance segmentation. Those fields have drawn great attention from researchers. And there is still room for exploration of relevant research focusing on mask detection in the context of global epidemics.

In the direction of object detection for practical uses such as the detection of mask wearing status, it is extremely difficult for computers to distinguish the features of correctly worn masks from those of partially incorrectly worn masks as shown in **Figure 1**, which illustrates the difficult samples in the face mask detection tasks. For that, researches are putting effort on either on the contructions of better dataset or the improvement of mask detection network architecture.

Shiming Ge *et al.* [7] proposed the use of LLE-CNNs model and construct the MAFA dataset, which contains 30,811 images and 35,806 face mask labels information. Each face contains three location types and three facial feature types. The model classifies the candidate faces based on the extracted facial features, and the position and scale of the candidate's face refinement are performed. Its performance on the MAFA dataset is 15.6% better than the other six models. But

Figure 1. Challenging samples in face mask detection tasks.

there are fewer number images containing blurred faces and low-resolution faces in this dataset which can affect the generalization of the model's performance.

To solve the incomplete detection categories and difficulties in detecting small targets and blurred faces issues discussed in the above paper, Xinqi Fan *et al.* [8] proposed the RetinaFaceMask model and further constructed the MAFA-FMD (MAsked FAces for Face Mask Detection) dataset. This dataset contains about 56,000 annotated image information, dividing the categories into "no mask wearing", "correctly wearing the mask", and "incorrectly wearing the mask". ResNet50 is used as the backbone network with Feature Pyramid Network (FPN), and the lightweight MobileNetV1 model is used to improve the operational efficiency when the model is being applied on mobile devices, combined with contextual attention module and migration learning to improve accuracy. However, this model has the drawback of detecting small targets due to its insufficient ability to learn enough useful image features.

Khanh-Duy Nguyen *et al.* [9] constructed the NFM dataset by extracting 581,108 facial images from 18,088 video frames to further increase the data labels. This model can only detect masks that are correctly wearing, incorrectly wearing, and "unknown". However, by setting "unknown" labels, the model will be confused during the training process. Therefore, it is essential to set the correct and clear label for the dataset used for training purposes.

3. Datasets

3.1. Major Existing Datasets for Mask Detection

Many datasets have been constructed in the field of face-mask detection problems. Those datasets can be categorized into real-world face mask datasets [7] [8] and simulated generated face mask datasets [9] [10]. By manually analyzing the classes in each dataset and mapping them to "with mask", "poor mask", "none mask", and others to observe the distribution of the number of images within classes in each dataset. The other classes indicate that the class in this dataset cannot be mapped to one of the three classes defined in this paper.

The MAFA [7] dataset contains public web images with a minimum edge length of 80 pixels while removing images containing obscure faces, resulting in 30,811 images, each of which contained at least one face with a mask on.

The AIZOO [11] dataset selected 3894 images from the WIDER FACE [12] dataset and 4064 images from the MAFA dataset for a total of 7959 images. By re-labelling those images, the AIZOO dataset contains a total of 16,636 images with faces in it: including two classes "with mask" and "none mask".

The MAFA-FMD [8] dataset selected 50% of the images from the AIZOO dataset and 50% of the MAFA images. They re-annotated the selected images into the same three classes of annotations as in this paper.

The NFM [9] dataset extracts 18,088 video frames from 17 street videos and obtains 581,108 faces, which contain many tiny and obscured faces. This dataset has a strong practical significance which can distinguish between the three types

of classes: "masked", "unmasked", and "unknown". In this article, "masked", and "unmasked" map to "with mask" and "none mask".

The MaskedFace-Net (MF-Net) [10] dataset selected images from the Flickr-Faces-HQ3 [13] dataset, which is constructed by combining unmasked faces and taking a generative adversarial network to generate images of faces that wear masks correctly and incorrectly.

The RMFRD [14] dataset contains only 5000 images of faces wearing masks and 90,000 images of faces not wearing masks, which are selected from Internet resources.

Other than using real-world data images, computer generated data are also contributing to this field. The SMFRD [14] dataset generates face images with simulated masks for the LFW [15] and WebFace [16] datasets by using simulated mask-wearing generation software.

Table 1 shows the detailed number of face annotations and the number of images within the dataset. The short-dashed line indicates no data of this category exists in the dataset. If the number of Totals is greater than the sum of the three classes, it means that the dataset has data from other classes. Furthermore, the number of Totals and the number of Images is the same means that all the images in this dataset contain only one face.

3.2. Masked Face Datasets (MFD) and Light Masked Face Datasets (LMFD) Constructions

Up to now, there are fewer datasets with balanced samples containing three cases of wearing masks at the same time. For example, the number of samples with incorrect masks in MAFA-FMD is 20 times smaller from the other two samples, which will make it difficult for the algorithm to learn the feature information of the label sample.

In this paper, researchers construct two multi-category, multi-scene, multi-difficulty mask datasets. LMFD [17] is a raw image dataset containing a total of 8232 images with only mask labels inside each picture provided. MFD [17] is a new dataset composed of LMFD after data enhancement methods, which includes both image data before and after the image enhancement techniques. These datasets allow many researchers to focus their research efforts on algorithm innovation and model deployment, contributing to facilitating the application of

Table 1. Masked face image datasets.

Name	With Mask	Poor Mask	None Mask	Totals	Images
MAFA [7]	31,032	4774	-	35,806	30,811
AIZOO [11]	3993	-	12,643	16,636	7959
NFM [9]	54,766	-	472,500	527,266	18,088
MF-Net [10]	67,193	69,823	-	137,016	137,016
RMFRD [14]	8589	-	90,426	99,015	99,015
SMFRD [14]	500,414	-	6000	506,414	506,414

face mask detection algorithms. The LMFD and MFD datasets contains the seven datasets mentioned above, real-world face mask images publicly available on the Internet, and the simulated generated face mask images.

The resolution of images in the MFD and LMFD datasets is not the same. The resolution of squared images is 320 px × 320 px. Other rectangular images are mainly within 400 px in width and height. 4086 images are used to train the model under LMFD dataset, and 55,905 images are used to train the model under MFD dataset.

3.3. Data Labels Description

3.3.1. With Mask

According to the recommendations of the World Health Organization, we define the proper wearing of a mask as one that at least completely covers and wraps around the nose, mouth, and chin [18]. Also, only disposable medical masks, medical protective masks, dust masks, gas masks, respirators, and other masks are accepted. In addition, because the Centers for Disease Control and Prevention also recommends the use of cloth masks [19], cloth masks are also considered correct and useful masks in this dataset.

3.3.2. Poor Mask

The poor mask is defined as one that does not completely cover and wrap around any part of the nose, mouth, or chin. For example, pulling the mask down to the mouth from a properly worn mask would be considered an improperly worn mask. Exposure of the nasal cavity results in aerosols carrying the new coronavirus remaining in contact with the nasal mucosal tissue [20].

3.3.3. None Mask

For objects that do not block solid or liquid particles suspended in a gaseous medium, such as veils, scarves, etc., the mask is considered not worn. In particular, the act of covering the nose and mouth with the hand is considered as not wearing a mask. Figure 2 shows some sample images of the labels.

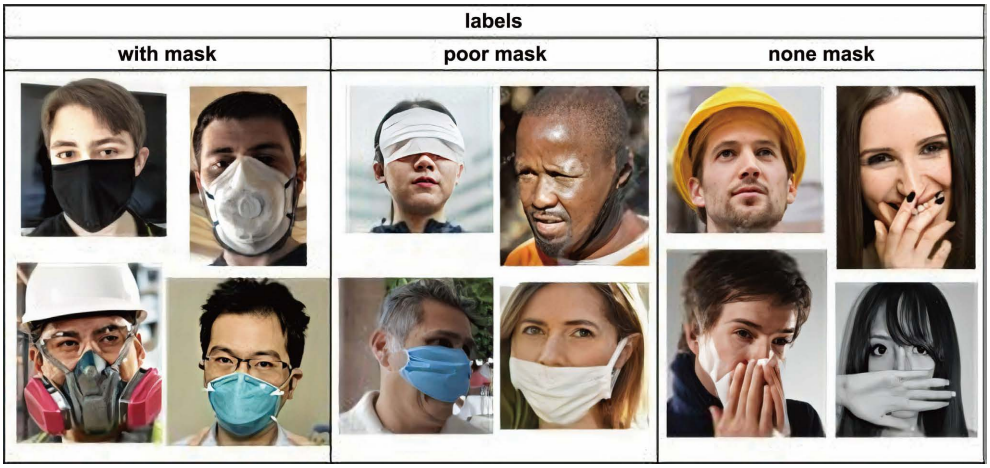

Figure 2. Label category demo images.

3.4. Data Augmentation

Images have properties including but not limited to width and height, resolution, color channels, contrast, brightness, light source properties, saturation, white balance, sharpening, exposure level, etc. Convolutional neural networks are more concerned with information about the absolute properties of a feature in a picture and are not sensitive to the relative relationship with the background or other features. Artificial image enhancement of images, such as rotating the angle, zooming in and out, panning and flipping, imitating perspective, changing saturation, contrast, brightness, sharpening, exposure, etc., can be effective in increasing the number of images and feature information [21].

In this paper, when training the model, in order to make the model easily and comfortably detect pictures with different scenes, different clarity, different facial and mask features, five kinds of data enhancement methods are taken, and each picture is expanded 3 times by taking five data enhancement methods in turn. Finally, each image can be expanded 15 times, and the method significantly reduces the difference in data volume between samples. The experiments in Section 5 show that the method has a more obvious positive effect on the improvement of accuracy.

The ColorJitter [22] method randomly changes the brightness, contrast, saturation, and hue of the image to simulate different ambient light, the special color of the mask, etc. The GaussianBlur [23] method randomly selects a Gaussian blur to blur the image to simulate camera shake during shooting or high-speed movement of the subject. The Sharpness [24] method randomly adjusts the sharpness of the image to simulate the image taken under different sharpness conditions. The Posterize [25] method post-processes the image by reducing the number of bits in each color channel. The Solarize [26] method exposes the image by inverting all pixel values above a threshold value.

Examples of data augmentation results are shown in **Figure 3** for images containing incorrect labels for wearing masks. After performing data enhancement, the dataset can handle the problem of sample imbalance.

Table 2 shows the number of labels for the LMFD and the MFD dataset which we construct for future researchers. When aiming for fast training and deployment, LMFD would be a better choice.

Experiments (Section 5) show that using LMFD can significantly improve the

Table 2. Comparison the number of labels between LMFD and MFD.

Label	LMFD	MFD
With Mask	3229	53,039
Poor Mask	2813	47,203
None Mask	2190	31,180
Totals	8232	131,422
Images	4086	55,905

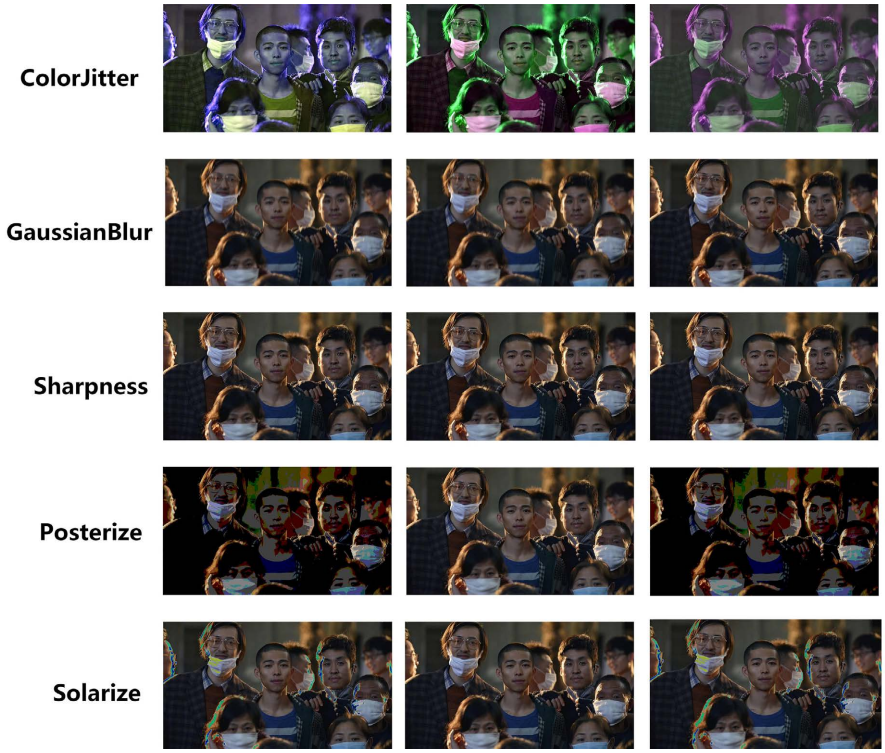

Figure 3. Data augmentation effect and quantity.

detection accuracy of the algorithm and obtain an excellent model with a high yield at the cost of low consumption. And by using MFD, detection performance is good under the condition of blurred faces, dark environments, etc. Besides, both datasets have a balanced distribution of labels which can affect the stability of the model performance.

4. Methodology for Object Detection

4.1. Expanding Convolution Kernel Perceptual Field

The convolutional kernel in the convolutional layer determines the perceptual field size of the feature map. The following problems exist in the traditional deep convolutional neural network:

1) The representation feature space obtained from the up-sampling and pooling layers does not change with different viewpoints, thus the model is not robust enough for handling situations with different viewing angles.

2) The internal data structure is lost; the spatial hierarchical information is lost.

3) It is hard to learn the inconspicuous features from smaller objects. Therefore, the detection accuracy of people far away from the cameras is low.

In order to solve the above issues, we use the Hybrid dilation convolution method in the mask detection network. The Hybrid dilation convolution is usually applied to the field of semantic segmentation after its initial proposal, and in this paper, we apply hybrid dilation convolution [27] to the muzzle detec-

tion subtask in target detection, which is found to be better than traditional convolution through the experiments discussed in section 5. By using hybrid dilation convolution, small object's features can be more easily learned with the use of dilation factor. When the dilation factor value is 1, 2, and 3, the convolution kernel shape is shown in Figure 4, and the red points represent pixels that are involved in the convolution calculation. It is worth mentioning that the dilation factor r is equal to 1 in the case of traditional convolution.

When r is equal to 3, and the number of pixels involved in the calculation will be 9, the number of pixels ignored due to the expansion will increase to 40, and the information loss rate is 40/(40 + 9) = 82%. It is obvious that the larger the expansion factor, the larger the information loss rate. Therefore, it is more balanced and reasonable to use a mixture of convolution kernels with different dilation factors.

Besides, it is important to design a reasonable expansion rate by assuming that we have N convolution layers, the size of the convolution kernel is $K \times K$, and the hybrid expansion rate is $[r_1, \cdots, r_i, \cdots, r_n]$. The goal of the dilation convolution process is to make a series of convolution operations with continuous perceptual fields and no missing image information, and we define the "maximum distance between two non-zero values" as M_i. The design requirements for the inflation rate need to satisfy condition (1).

$$M_i = \max\left[M_{i+1} - 2r_i, M_{i+1} - 2\left(M_{i+1} - r_i\right), r_i\right] \tag{1}$$

where:

$$M_2 \le K \tag{2}$$

$$M_n = r_n \tag{3}$$

r_i is the expansion factor of layer i, and M_i is the maximum expansion rate of layer i. The design principle is that the conventional number of the expansion rate of each layer cannot be greater than 1.

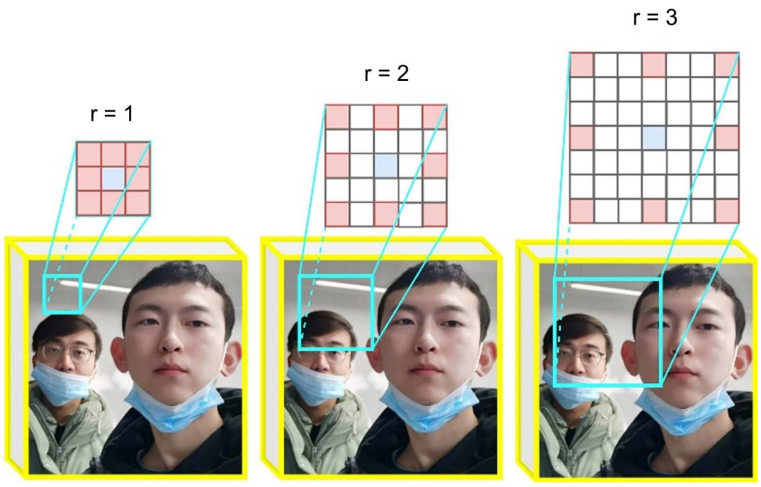

Figure 4. The pixel points are involved in the computation of the convolution kernel when the dilation factor is 1, 2 and 3.

Figure 5 shows the perceptual field area size and the number of times each pixel point is involved in the convolution calculation when the convolution kernel size is 3 × 3 for the conventional convolution and the hybrid dilation convolution rate of [1, 2, 3]. When the convolution kernel size is 3 × 3, the perceptual field area size of conventional convolution is 7 × 7. When the hybrid inflated convolution method is used, the hybrid inflated convolution rate of [1, 2, 3] increases the perceptual field area to 13 × 13, which increases the perceptual field area by 344.9% and preserves more image feature information.

Another benefit of using hybrid dilation convolution is that it can use arbitrary expansion rates in the process, thus it can naturally expand the receptive domain of the network without adding additional modules, which is important for identifying relatively large objects. The key difference between the hybrid inflationary convolution approach and the Atrous Space Pyramid Pool (ASPP) module [28] or the contextual aggregation module [29] [30] is that the hybrid dilation convolution approach utilizes an expansion factor with a common factor no greater than 1. In addition, hybrid dilation convolution can naturally integrate with the original layers of the network without adding additional modules.

4.2. Network Architecture

Faster R-CNN is a deep convolutional network used for object detection that appears to the user as a single, end-to-end, unified network [31]. The input image is extracted by a convolutional neural network for image features. The region proposal network generates proposal frames, and the proposal frames are

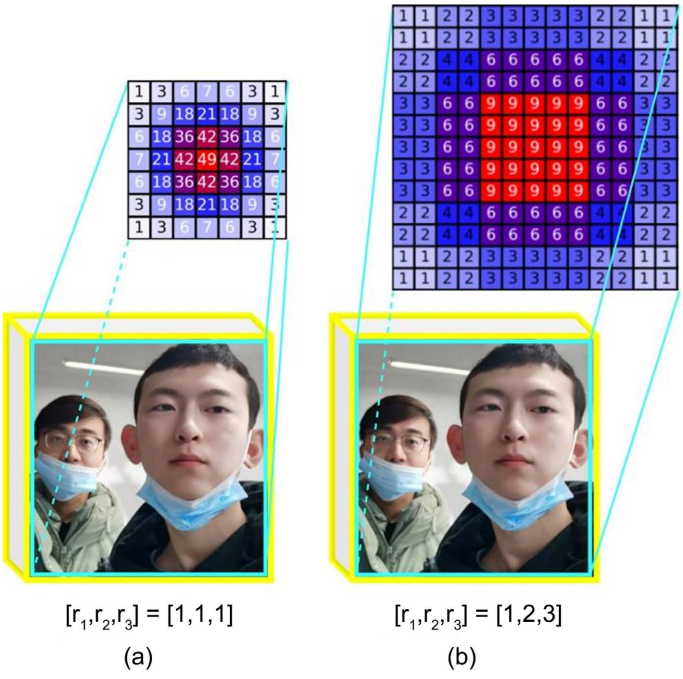

Figure 5. Comparison of perceptual field area size between traditional convolution and hybrid dilation convolution rate of [1, 2, 3].

projected onto the feature map to obtain the corresponding feature matrices. Each feature matrix is scaled to a 7 × 7 sized feature map through a region of interest pooling layer, followed by spreading the feature map through a series of fully connected layers to obtain the prediction results [32]. We then train the Faster R-CNN target detection model and use the detection result accuracy as the baseline model, focusing on the data augmentation application effect, and the integration application effect of hybrid expanded convolution and residual network. **Figure 6** shows the object detection algorithm Faster R-CNN.

The residual network [33] proposed by KaiMing He *et al.* in 2015 still has a wide influence today, in which the model with ResNet as the main backbone network continues to set new records for major datasets.

The residual network follows the complete convolutional layer design of VGG [34]. It starts with two convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function. Then we skip these 2 convolutional operations through the cross-layer datapath and add the input directly before the final ReLU activation function. Such a design requires that the outputs of the two convolutional layers have the same shape as the inputs so that they can be added together. If you want to change the number of channels, you need to introduce an additional convolutional layer to transform the input into the desired shape before doing the summation operation.

ResNet50 uses four modules consisting of residual blocks, each using several residual blocks with the same number of output channels. The first module has the same number of channels as the input channels. Since a maximum convergence

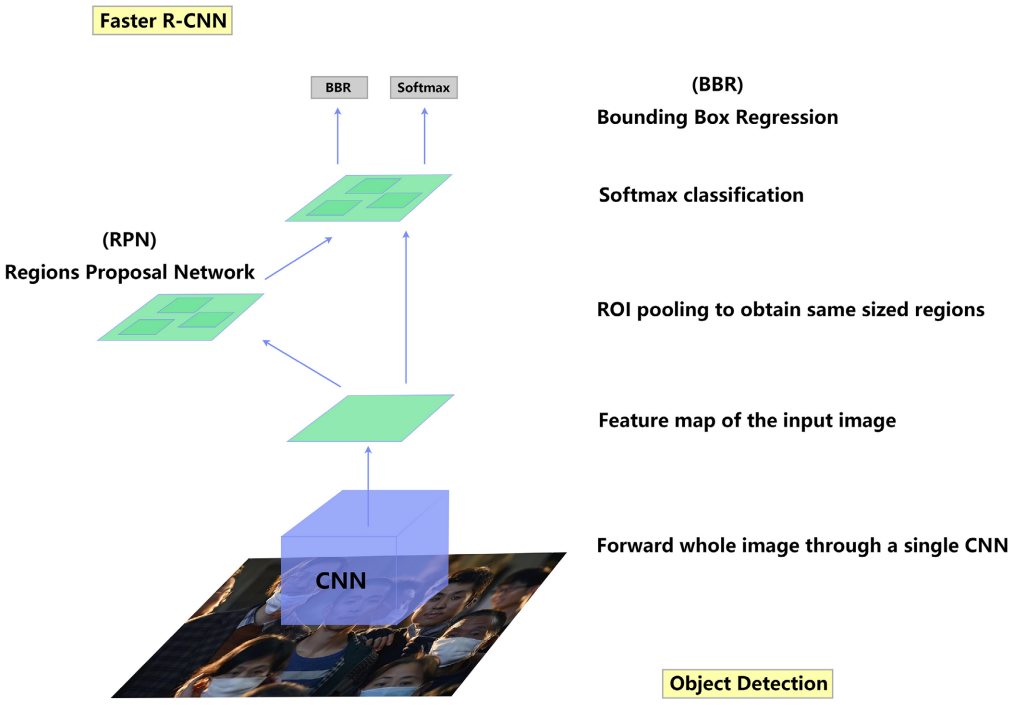

Figure 6. Faster R-CNN object detection algorithm.

layer with a step size of 2 has been used previously, there is no need to reduce the height and width. Each subsequent module doubles the number of channels of the previous module in the first residual block and halves the height and width. Figure 7 shows the network structure of the deep learning algorithm called RHF.

The momentum method [35] is an accelerated gradient descent method that simulates momentum in physics, accumulating velocity vectors, and for a given objective function $f(\theta)$ to be minimized, the momentum method is given by:

$$v_{t+1} = \mu v_t - \varepsilon \nabla f(\theta_t) + \lambda \theta_t \tag{4}$$

$$\theta_{t+1} = \theta_t + v_{t+1} \tag{5}$$

where $\varepsilon > 0$ is the learning rate, $\mu \in [0,1]$ is the momentum coefficient and $\nabla f(\theta_t)$ is the gradient at θ_t.

5. Experiment

5.1. Experiment Setup

In the model training, various combinations of parameters are tested. We finally used the following parameters in our model based on the size of the device memory and the magnitude of the model parameters, taking into account the model training time and the model training effect. Higher learning rate will cause the model never converge, and the model cannot learn characteristics of the datasets at a lower learning rate.

The models with ResNet50 and RHF are developed on the Pytorch deep learning library [36]. The datasets used are LMFD and MFD. The model was trained for 20 epochs with a stochastic gradient descent (SGD) algorithm of learning rate ε = 0.005 and momentum μ = 0.9. The parameters used in the experiments include a batch size of 8, and weight decay λ is 0.0005. The learning rate decreases by 33% every three generations. An NVIDIA GeForce RTX 3090 GPU was employed.

5.2. Results

COCO metrics were used to validate the model effects. The mAP and IoU are

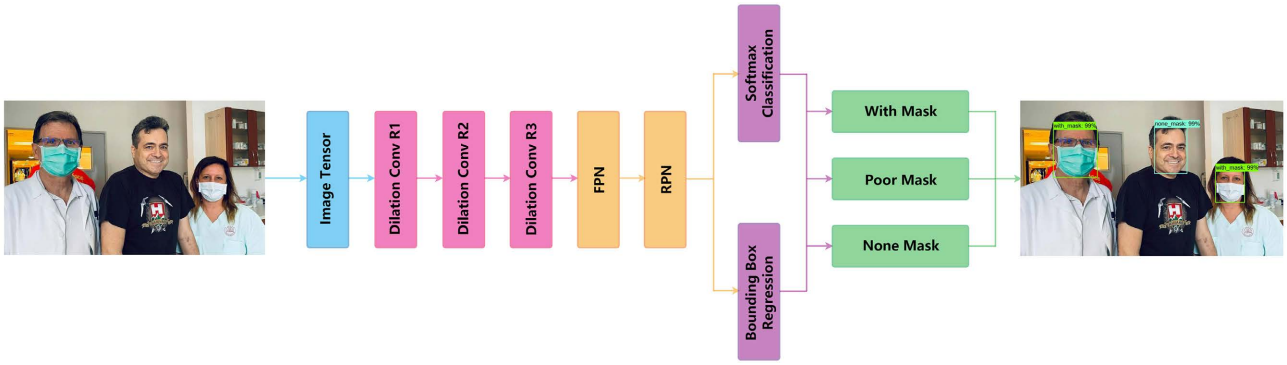

Figure 7. The network structure of the deep learning algorithm is called RHF.

used in this paper. The mAP is the average accuracy, and IoU is the intersection over union of prediction frames to object frames. The threshold of IoU is set to 0.6 because there are many small objects in datasets.

The performance and training time cost of ResNet50 and RHF models on the LMFD and the MFD dataset be shown in Table 3. Through observational studies the detection accuracy and the model on each label and the average precision of the detection on the dataset, it can be found that model training on RestNet50 costs more time than model training on RHF. The reduction in training time will allow the model to be trained faster on large datasets.

It took 1.5769 hours to train the ResNet50 model on LMFD, and only 1.3743 hours to train the RHF model. It took 23.5589 hours to train ResNet50 model on MFD, and only 13.9982 hours to train RHF model on LMFD.

By using the MFD dataset to train the RHF model, RHF shows good performance on the completely new images that never appeared in the original dataset before. The model inference 200 new images on NVIDIA GPU3090. The average testing time of each image was 0.0424 seconds, and the real-time frame rate reached 23.6FPS.

The loss rate of the training process is shown in Figure 8. The mAP of the training process is shown in Figure 9. Results indicate that the RHF model's loss rate starts to converge after the 15th epoch. The ResNet50 model only starts to converge in MFD dataset after the 18th epoch. The loss rate curve shows that the hybrid dilation convolution accelerates the convergence of the model.

Figure 10 shows the detection results in multiple complex environments of the four models trained in this paper. The Original Image indicates the images without any image processing and data enhancement. Due to the length of the displayed images, only the detection results of the three data enhancement methods in multiple scenarios are used. The "GaussianBlur" means the images use a size of 3 × 7 Gaussian convolution kernel to blur. The "ColorJitter" method uses the brightness jitter factor in [0.5, 1.5], the hue jitter factor in [−0.3, 0.3]. The "Solarize" method inverts the pixel values exceeding 108 in each color channel.

6. Discussion

It is observed that both the data augmentation method and the hybrid dilation convolution method used in this paper can improve the accuracy of face mask

Table 3. Detection average precision of the model.

Model	Datasets	With Mask AP	Poor Mask AP	Poor Mask AP	mAP	Time (hour)
ResNet50	LMFD	79.57%	96.39%	77.54%	84.50%	1.5769
RHF	LMFD	**92.60%**	**97.32%**	**81.71%**	**90.50%**	**1.3743**
ResNet50	MFD	95.57%	99.00%	85.48%	93.30%	23.5589
RHF	MFD	**95.73%**	**99.10%**	**85.62%**	**93.45%**	**13.9982**

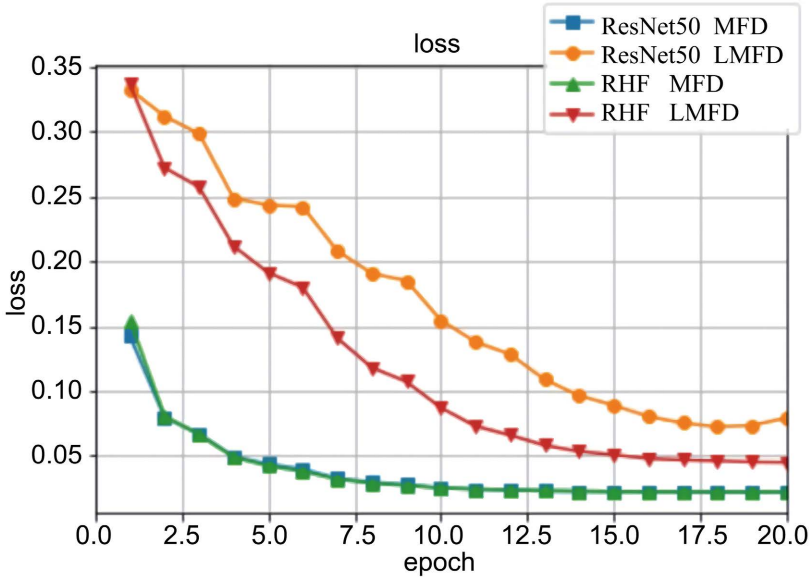

Figure 8. Loss during the training.

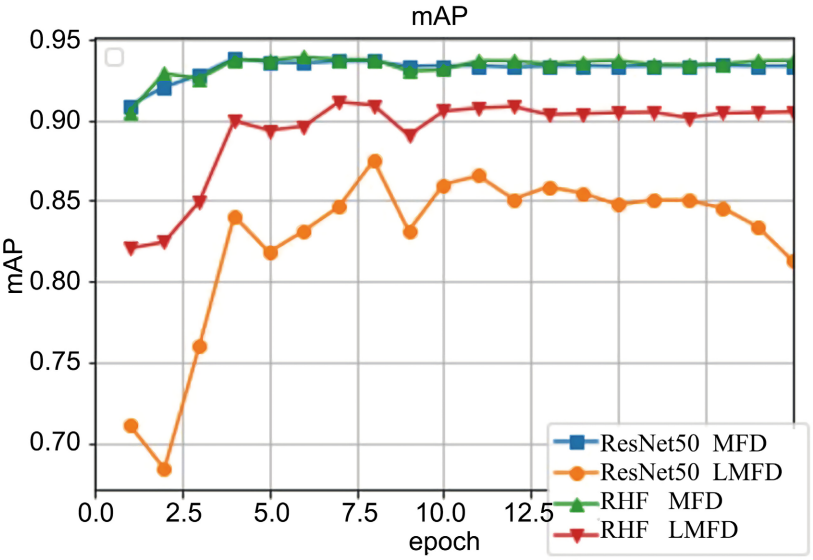

Figure 9. The average precision during the training.

detection. In particular, the hybrid dilation convolution method can improve the average precision by about 6% in LMFD dataset without using data augmentation. The average precision can be improved by 8.95% using only the data augmentation method. Meanwhile, the models which use hybrid dilation convolution also have a lower time cost.

The training time is another important factor for considering the performance of CV work. It is found that by using data augmentation and dilation method, the training time will increase accordingly based on the choice of augmentations used and the number of dilation factors tested. However, in order to have a better detection model for mask wearing status, the time is worth it since we are trading the volume of images needed for a little bit longer of time for training.

Figure 10. The detection results in multiple complex environments of the four models. The left header indicates the model used for that row of images, and the top header indicates the data augmentation method used for that column of images.

Observing the detection results in **Figure 10**, the RHF model training on MFD performance is better than other models. It can detect the right class and object while other models cannot work. When every model detects all of the right classes and objects, the RHF model will have higher performance and accuracy. It is also found that, parameters such as image brightness, size of objects are still the main contributors for object detection accuracy. Though the dilation factors can solve most of the small object features lost problem, when the mask wearing people is far away from the camera with bad angles and brightness, it can still affect the performance of the model.

Also, based on the experiment result, it is found that clustering methods can also be used as a pre-processing method to increase the whole framework's efficiency. For example, by applying the Twin Contrastive Learning (TCL) clustering methods clustering methods [37] [38] [39] proposed by Peng Xi, only one dataset with good image quality and labels is needed for training by TCL. Then TCL can be used to automatically cluster the three types of masks wearing status, which can not only save a lot of time in preparing image data but also increase the overall effectiveness of the whole framework.

7. Data Availability

The datasets and code are already open sources on GitHub. The GitHub repository link is https://github.com/shiningxy/RHF. This repository provides a download link to the dataset and the model weight file of this paper. In order to facilitate the researchers reproducing the project library, the structure and usage of all documents are indicated at the same time. And it provides some precau-

tions attentively. Now, this project has earned some Stars.

8. Conclusions and Future Work

In this paper, we develop an object detection model for identifying three wearing states of mask-wearing to ensure the personal health and safety of people in the context of a new coronavirus outbreak. We collected a considerable number of real mask-wearing images and manually annotated the mask-wearing information in these images. The total number of these images is 55,905, including 131,422 human faces, with three labels: "with mask", "poor mask", and "none mask".

Despite the proposed Face Mask Detection model's good performance, a lot of work still needs to be done. Firstly, the dataset provided in this paper still has a certain gap compared with the number of large-scale datasets such as WIDER FACE [12]. Secondly, the recognition of extreme poses, special masks, blurred targets and obscured targets still has room for improvement. Lastly, the proposed model has a large number of parameters that make it hard to be deployed on the mobile device. Thus, the future work will focus on 2 ways:

1) By combining clustering methods like TCL, the dataset constructed in this work can be expanded to improve the generalization and efficiency of the proposed work.

2) Light weighted model can be used to replace CNN model in this work to better suit the mobile devices.

The main innovation of this work is the use of dilation method in our mask detection model to solve the real-world problem: for using cameras to detect the mask wearing status, often it is hard for every person to face the camera directly with a good viewing angle. Besides, for most of time, it is far more important to be able to detect people not properly wearing the mask rather than not wearing or properly wearing the mask during the pandemic. Thus, it is very important for the construction of a dataset with enough and balanced number of images and labels for each of the mask wearing status in order to truly improve the performance of the mask detection network.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Khan, J.Y. and Alamin, M.A.A. (2021) A Comparative Analysis of Machine Learning Approaches for Automated Face Mask Detection during COVID-19.

[2] Lo, J.Y., Tsang, T.H., Leung, Y.-H., *et al.* (2005) Respiratory Infections during SARS Outbreak, Hong Kong, 2003. *Emerging Infectious Diseases*, **11**, 1738. https://doi.org/10.3201/eid1111.050729

[3] Cheng, V.C.-C., *et al.* (2020) The Role of Community-Wide Wearing of Face Mask for Control of Coronavirus Disease 2019 (COVID-19) Epidemic Due to SARS-CoV-2.

Journal of Infection, **81**, 107-114. https://doi.org/10.1016/j.jinf.2020.04.024

[4] Farfade, S.S., Saberian, M.J. and Li, L.-J. (2015) Multi-View Face Detection Using Deep Convolutional Neural Networks. *Proceedings of the* 5*th ACM on International Conference on Multimedia Retrieval*, Shanghai, 23-26 June 2015, 643-650. https://doi.org/10.1145/2671188.2749408

[5] Razavi, M., Alikhani, H., Janfaza, V., Sadeghi, B. and Alikhani, E. (2022) An Automatic System to Monitor the Physical Distance and Face Mask Wearing of Construction Workers in COVID-19 Pandemic. *SN Computer Science*, **3**, 1-8. https://doi.org/10.1007/s42979-021-00894-0

[6] Hendrycks, D., Mazeika, M., Kadavath, S. and Song, D. (2019) Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 1-13.

[7] Ge, S., Li, J., Ye, Q. and Luo, Z. (2017) Detecting Masked Faces in the Wild with LLE-CNNs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2682-2690. https://doi.org/10.1109/CVPR.2017.53

[8] Fan, X. and Jiang, M. (2021) RetinaFaceMask: A Single Stage Face Mask Detector for Assisting Control of the COVID-19 Pandemic. 2021 *IEEE International Conference on Systems*, *Man*, *and Cybernetics* (*SMC*), Melbourne, 17-20 October 2021, 832-837. https://doi.org/10.1109/SMC52423.2021.9659271

[9] Nguyen, K.-D., Nguyen, H.H., Le, T.-N., *et al.* (2021) Effectiveness of Detection-Based and Regression-Based Approaches for Estimating Mask-Wearing Ratio. 2021 16*th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG* 2021), Jodhpur, 15-18 December 2021, 1-8. https://doi.org/10.1109/FG52635.2021.9667046

[10] Cabani, A., Hammoudi, K., Benhabiles, H. and Melkemi, M. (2021) Masked-Face-Net—A Dataset of Correctly/Incorrectly Masked Face Images in the Context of COVID-19. *Smart Health*, **19**, Article ID: 100144. https://doi.org/10.1016/j.smhl.2020.100144

[11] https://github.com/AIZOOTech/FaceMaskDetection

[12] Yang, S., Luo, P., Loy, C.-C. and Tang, X. (2016) Wider Face: A Face Detection Bench-Mark. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 5525-5533. https://doi.org/10.1109/CVPR.2016.596

[13] Kazemi, V. and Sullivan, J. (2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1867-1874. https://doi.org/10.1109/CVPR.2014.241

[14] Wang, Z., *et al.* (2020) Masked Face Recognition Dataset and Application.

[15] Huang, G.B., Mattar, M., Berg, T. and Learned-Miller, E. (2008) Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *Workshop on Faces in "Real-Life" Images*: *Detection*, *Alignment*, *and Recognition*, Marseille, August 2008.

[16] Yi, D., Lei, Z., Liao, S. and Li, S.Z. (2014) Learning Face Representation from Scratch.

[17] W. Xiangyu. https://github.com/shiningxy/RHF

[18] Silveira, P., Teixeira, A. and Soares, C.G. (2013) Use of AIS Data to Characterise Marine Traffic Patterns and Ship Collision Risk off the Coast of Portugal. *The*

Journal of Navigation, **66**, 879-898. https://doi.org/10.1017/S0373463313000519

[19] Ho, K.-F., Lin, L.-Y., Weng, S.-P. and Chuang, K.-J. (2020) Medical Mask versus Cotton Mask for Preventing Respiratory Droplet Transmission in Micro Environments. *Science of the Total Environment*, **735**, Article ID: 139510. https://doi.org/10.1016/j.scitotenv.2020.139510

[20] Gallo, O., Locatello, L.G., Mazzoni, A., Novelli, L. and Annunziato, F. (2021) The Central Role of the Nasal Microenvironment in the Transmission, Modulation, and Clinical Progression of SARS-CoV-2 Infection. *Mucosal Immunology*, **14**, 305-316. https://doi.org/10.1038/s41385-020-00359-2

[21] Perez, L. and Wang, J. (2017) The Effectiveness of Data Augmentation in Image Classification Using Deep Learning.

[22] Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P. and Srinivas, A. (2020) Reinforcement Learning with Augmented Data. *Advances in Neural Information Processing Systems*, **33**, 19884-19895.

[23] Gedraite, E.S. and Hadad, M. (2011) Investigation on the Effect of a Gaussian Blur in Image Filtering and Segmentation. *Proceedings ELMAR*-2011, *IEEE*, Zadar, 14-16 September 2011, 393-396.

[24] Shaked, D. and Tastl, I. (2005) Sharpness Measure: Towards Automatic Image Enhancement. *IEEE International Conference on Image Processing*, Vol. 1, I-937. https://doi.org/10.1109/ICIP.2005.1529906

[25] Kwon, O.-Y. and Chien, S.-I. (2011) Improved Posterized Color Images Based on Color Quantization and Contrast Enhancement. *Proceedings International Conference Machine Vision, Image Processing, and Pattern Analysis*, Vol. 5, 1203-1206.

[26] Soroush, M., Wessel-Berg, D., Torsaeter, O. and Kleppe, J. (2014) Investigating Residual Trapping in CO_2 Storage in Saline Aquifers—Application of a 2D Glass Model, and Image Analysis. *Energy Science & Engineering*, **2**, 149-163. https://doi.org/10.1002/ese3.32

[27] Wang, P., *et al.* (2018) Understanding Convolution for Semantic Segmentation. 2018 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Lake Tahoe, 12-15 March 2018, 1451-1460. https://doi.org/10.1109/WACV.2018.00163

[28] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2017) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[29] Fan, X., Jiang, M. and Yan, H. (2021) A Deep Learning Based Light-Weight Face Mask Detector with Residual Context Attention and Gaussian Heatmap to Fight against COVID-19. *IEEE Access*, **9**, 96964-96974. https://doi.org/10.1109/ACCESS.2021.3095191

[30] Yu, F. and Koltun, V. (2015) Multi-Scale Context Aggregation by Dilated Convolutions.

[31] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[32] Wang, S., Zargar, S.A. and Yuan, F.-G. (2021) Augmented Reality for Enhanced Visual Inspection through Knowledge-Based Deep Learning. *Structural Health Monitoring*, **20**, 426-442. https://doi.org/10.1177/1475921720976986

[33] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image

Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[34] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[35] Sutskever, I., Martens, J., Dahl, G. and Hinton, G. (2013) On the Importance of Initialization and Momentum in Deep Learning. *International Conference on Machine Learning*, *PMLR*, Atlanta, 17-19 June 2013, 1139-1147.

[36] Paszke, A., *et al*. (2019) An Imperative Style, High-Performance Deep Learning Library. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 8026.

[37] Li, Y., Yang, M., Peng, D., Li, T., Huang, J. and Peng, X. (2022) Twin Contrastive Learning for Online Clustering. *International Journal of Computer Vision*, **130**, 2205-2221. https://doi.org/10.1007/s11263-022-01639-z

[38] Peng, X., Li, Y., Tsang, I.W., Zhu, H., Lv, J. and Zhou, J.T. (2022) XAI Beyond Classification: Interpretable Neural Clustering. *Journal of Machine Learning Research*, **23**, 6:1-6:28.

[39] Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T. and Peng, X. (2021) Contrastive Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 8547-8555. https://doi.org/10.1609/aaai.v35i10.17037