HKIE THE HONG KONG INSTITUTION OF ENGINEERS 香港工程師學會

Effective real-time face mask detection on NVIDIA edge devices

Ho Chuen Kam[1,2], Nathan Chung Kan Lui[1], Ethan Yi Chen Du[1], Marcus Ngai Kan Lui[1], Chun Sing Chan[1], Yan Ting Cheung[1], Ka Chun Cheung[2] and Monica Sze Man Leung[1]

[1] Blueinno Technology Limited, Hong Kong, People's Republic of China
[2] NVIDIA AI Technology Centre, Hong Kong, People's Republic of China

ABSTRACT

During these difficult times of COVID-19, people are struggling to return to their normal routines, including going back to schools and workspaces. To prevent the spread of the disease, wearing face masks is essential for everyone to protect themselves and the ones around them. However, challenges arise in regard to enforcement of wearing masks in large crowds such as at educational centres and public transportation. This paper proposes a robust automatic system for face mask detection using transfer learning kits from NVIDIA. Based on the backbone of Resnet-18, the model results in high accuracy in the distinguishing of persons who do and do not wear masks. Leveraged by the NVIDIA edge accelerator, the system can run in real-time environments, making it applicable in various venues. Its feasibility was demonstrated by deploying the approach in an education centre in Hong Kong.

KEYWORDS Mask detection; transfer learning; convolutional neural network; edge AI; computer vision
CONTACT Ho Chuen Kam calvin@blueinnotechnology.com
Received 7 September 2021

1. Introduction

In light of the newly discovered coronavirus disease (COVID-19) and its variants, the World Health Organization (WHO) declared it as a global pandemic in the WHO Director-General's opening remarks at the media briefing on COVID-19 in 2020. At the time of writing, there were nearly 200 million confirmed cases worldwide according to the WHO (World Health Organization, 2021). The contagion is transmitted via tiny liquid particles when an infected person coughs, sneezes, or speaks (Taurus Protection Inc., 2020). Based on these findings, the WHO has advised that citizens of all countries should wear face masks in public areas where most of the cases originate from (World Health Organization, 2020a). It is considered to be a preventive measure combating the epidemic to stop the virus from spreading.

Face masks are essential in order to return to a normal lifestyle, such as daily commute and working in an office. Due to the outbreak of COVID-19, people have been left with no choice but to stay at home, impacting work performances of the employees for large projects (Bao et al., 2020). Surveys show that 85% of people in Hong Kong are required to work in an office during regular work hours (Vyas and Butakhieo, 2021), which increases the risk of virus transmission. In order to minimise the risk, face masks are required to be worn properly at all times outdoors to prevent the possible chains of infection. Face mask detection thus has become a new demand in the market for monitoring people going in and out of venues.

The proposed face mask detection model can be applied in a wide range of occasions, including public transportation, workspaces, and educational institutes. As the model is compatible with cameras of all sorts, it could be installed inside trains and buses. As cited by the US National Library of Medicine (Goscé and Johansson, 2018), inside these types of transportation are confined areas and passengers are closely packed together, failing to comply with mandatory social distancing rules. This universal situation hence necessitates the use of face masks.

Workplaces are one of the hotbeds for this global contagion. The proposed model can be installed at the entrance to ensure that everyone entering is wearing a mask. Indoor spaces, especially poorly ventilated buildings, pose an increased risk of infection. However, an analysis conducted by Liangzhu Wang (2021) demonstrates that wearing a mask is even more effective than improving ventilation.

Schools and tutorial centres are the most vulnerable venues for pandemics. During the beginning of the pandemic, they were locked down to deter potential threats to children and their teaching staff. Turning to online learning, studies show that this poses negative impacts for their knowledge acquisition and social skills. With the decline of the pandemic transmission rate, schools and tutorial centres are gradually reopening and returning to their normal operation. Enforcing the wearing of face masks has become essential in the school environment for better protection in schools. Tutorial centres are also applying the same measures to keep students safe.

In populated open areas, security guards are posted to enforce mask-wearing rules (Rana and Kathirvel, 2020). However, this has proven to be perilous in some parts of America. Reports show that customers who are reluctant to wear masks are posing security risks to guards and employees because they are not permitted to enter (Shepherd et al., 2021). In addition, security guards utilise human resources and are manpower intensive. Disadvantages include lack of motivation and experience of employees, as well as insufficient financial support and violations against social distancing rules.

Computer science researchers hence started to investigate the use of AI to detect face masks automatically. Various mask detection algorithms have been constructed and introduced to the market.

The survey of the recent face mask detection algorithms can be found here (Balaji et al., 2021). Convolutional neural networks such as MobileNet v2 (Yadav, 2020) and YOLO (Du, 2018) were used for finding face masks in camera images. However, these methods usually require high computational power which depends on the Graphical Processing Unit (GPU) for inference acceleration. There will be threats posed to privacy breaching if the images are uploaded to the cloud for processing.

The initiative of this project is to create a fast, and yet, accurate deep learning approach to detect whether a person is wearing a face mask.

The use of the NVIDIA Transfer Learning Toolkit (TLT) is proposed for the model's development and deployment as the close integration of the edge accelerator and optimisation can be leveraged. In particular, Resnet-18 was used as the backbone of the detection. The result shows that the proposed system achieves high accuracy in detecting human faces with and without face masks.

To make the model operate in real time and preserve privacy, it was deployed to NVIDIA Jetson series edge devices for inference. The model can retain the inference time to 30 frames per second (FPS). In addition, ordinary webcams are compatible with the system, which further reduces the operational cost. The system was tested at a tutorial centre in Hong Kong for a trial run, where manual face mask checking is not efficient during rush hours. It was installed near the entrance of the centre. Young students were required to pass the system testing for face mask compliance. If they were not wearing a mask, or were wearing it improperly, alerts would be issued to alert the teachers. The system was proved to be sufficiently robust and accurate to replace the manpower required to guard the entrances.

The remainder of the paper is divided into four sections as outlined below. Section 2 evaluates the previously related works in the past and the motivation contributing to this model. Section 3 describes the dataset, image augmentation techniques applied, and details of the proposed model. Section 4 gives the details of the system setup and the analysis of the experimental results.

Section 5 concludes the work of the mask detection system deployed in a tutorial centre in Hong Kong.

2. Background study

In the past decades, artificial intelligence has greatly improved its ability to detect and classify objects. Using neural networks, algorithms such as YOLOv3 can analyse images and videos in real time. At a high speed of up to 155 frames per second (Du, 2018) , it runs faster than other convolutional neural network-based algorithms. It is capable of outputting bounding frames around the objects detected, as well as the probabilities of what the object could be. However, the model, which uses NVIDIA TLT, runs at a higher speed and efficiency than the YOLO architecture.

Due to the outbreak of the novel coronavirus pandemic, the study of face mask detection has been a subject of interest. A paper on face mask detection (Qin, and Li, 2020) uses SRCNet, a network implementing image sharpening, to identify and distinguish images of people into three classes: wearing, incorrectly wearing, and not wearing face masks. A dataset from Kaggle on medical masks was used, as well as a network to find and crop faces before being fed into the algorithm. Their method can classify images with 98.70% accuracy. Yet, it is of low efficiency with the capability of analysing 10 images per second. Another paper (Jiang et al., 2020) using several convolutional networks looks at a similar topic. Several datasets containing normal, covered and masked faces were used in the training process. This network can classify images in two scenarios of with and without a mask, scoring an accuracy 1.5% and 2.3% better than the baseline. However, they only explored the possibility of integrating their model with mobile devices. Yadav (2020) used MobileNet v2 architecture integrated with Raspberry Pi 4 which continuously detects whether people are covering their faces with a mask and abiding by social distancing. This model scores an accuracy of 91.7%. However, this model does not classify if a mask is worn improperly and only detects whether a mask is present or not. Yet, Rahaman et al. (2021) describe a technique to determine whether a person is masked, and propose to integrate it with public CCTV networks. If a person without a face mask is identified, the authorities of the network will be notified. A Convolutional Neural Network (CNN) is used as the architecture of deep learning. This proposition has an accuracy of 98.7% in detecting face masks. However, this proposed model has difficulties detecting face masks which are not properly worn. Furthermore, the system may encounter errors when transmitting data from the surveillance cameras to the network. Deore et al. (2016) used four different aspects for mask detection - distance from the camera, eye line detection, facial part detection and eye detection. When a face is not detected, but a person is detected, it is classified as a mask. Each of the four aspects has a different level of confidence. Distance from the camera, eye line detection, facial part detection, and eye detection contribute to an accuracy of 90%, 69.8%, 46.6% and 40% respectively. If a person is covering the majority of their face with an object other than a mask, face detection does not occur and this will result in a false-positive case.

Currently, there are several different algorithms for object detection, each yielding different speeds (fps) and accuracies. These include RCNN, R-FCN, SSD, RetinaNet and YOLO. Among these algorithms, YOLO (Redmon et al., 2016) currently has the fastest speed and highest accuracy (Hui, 2019). YOLO, known as You Only Look Once, is an object detection algorithm that is capable of running in real time. It has the ability to output the identities of the objects, bounding boxes around them, as well as confidence values. It uses two networks (Li et al., 2020), a backbone network called "Darknet-53" with 53 convolutional layers, and a detection network. YOLO has several advantages. First of all, its latency of 22 ms allows it to run in real time without noticeable lag, which is three times faster than SSD (Liu et al., 2016). It is also highly accurate with good learning efficiency.

Its original author created a new version called YOLOv3, has improvements over the previous versions. For example, its accuracy decreases with object size, where large object detection may be up to 10% below the best accuracy (Sonawane, 2018). In recent years researchers have explored different strategies to improve the YOLO. YOLOv4 (Bochkovskiy et al., 2020) introduced improvements based on feature aggregation and YOLOv5 (Thuan, 2021) proposed integration of the anchor box selection into the learning processes to increase the accuracy of the network. To examine the speed and accuracy of YOLO, a comparison of the algorithm to other algorithms will be made in a later section.

3. Theory and methodology

3.1. Datasets

In this work, three various datasets were used: MAFA (Ge et al., 2017), with 1,846 mask photos and 232 no mask photos; Kaggle Medical Mask (Waghe, 2020), with 4,154 mask photos and 790 no mask photos; FDDB (Jain and Learned-Miller, 2010), with 0 mask photos and 4,978 no mask photos. In total, 12,000 photos were used with 6,000 mask photos and 6,000 no mask photos. Figure 1 shows an example of the datasets.

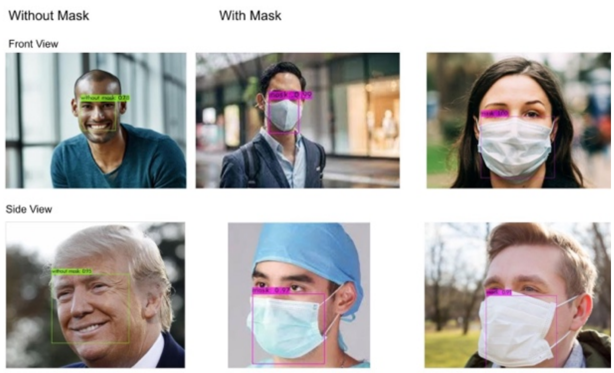


Figure 1. Example of the datasets.

3.2. Data augmentation

Data augmentation is a technique used to expand the size of a dataset. This is achieved by editing the image properties in different ways, such as changing the size, colour, rotation, etc. Resizing was used in this model, with each image being changed to 960 by 544 pixels. Flipping, which means reversing the order of pixels in rows and columns, was also used, both vertically and horizontally. Translation, which is shifting the pixels in the X and Y directions, was also used. Colour, which can be described by the hue, saturation, and contrast, was used to lower the model's reactivity to colour.

3.3. DetectNet v2

Figure 2 shows the proposed neural network architecture. In this work, a novel object detection framework developed by NVIDIA called DetectNet was adapted. It detects objects by predicting the objectness (coverage) within grid squares. With this information, bounding box corners can be found as they are related to the centres of the occupied grid squares. The architecture of DetectNet is summarised below:

- Data Ingestion and Augmentation Layer
- Fully Conventional Network (FCN)
- Loss Function
- Bounding Box Clustering

First, photos labelled with bounding boxes are required to train the neural networks. In this study, both positive (person wearing a face mask) and negative (person not wearing a face mask) were included in the datasets. They were first converted into the format of KITTI (Geiger et al., 2013), then the photos were cropped into the dimension of 940 × 544. Then they were fed into the data layer for further processing. To avoid overfitting and give more training data to the AI, online image augmentation was performed after the data feeding. Resnet-18 (He et al., 2016) was employed to be the feature extraction backbone. It was chosen because the network balances the accuracy and inference time, which fits the use case herein. Before training it was initialised by pretraining using the subset of the Google OpenImages dataset (Kuznetsova et al., 2020).

Second, the training data was passed to the fully convolutional network for feature extraction and the prediction of bounding box locations and object classes. This was done by coverage maps, which contain a series of grid cells to determine whether an object is located in patches on them.

Finally, two separate loss functions were aggregated to calculate the whole system loss, namely the coverage loss and the bounding box loss. Equation (1) shows the coverage loss calculated from the least square error (L2) between the output coverage map from the fully convolutional neural network and that generated from the training data (ground truth).

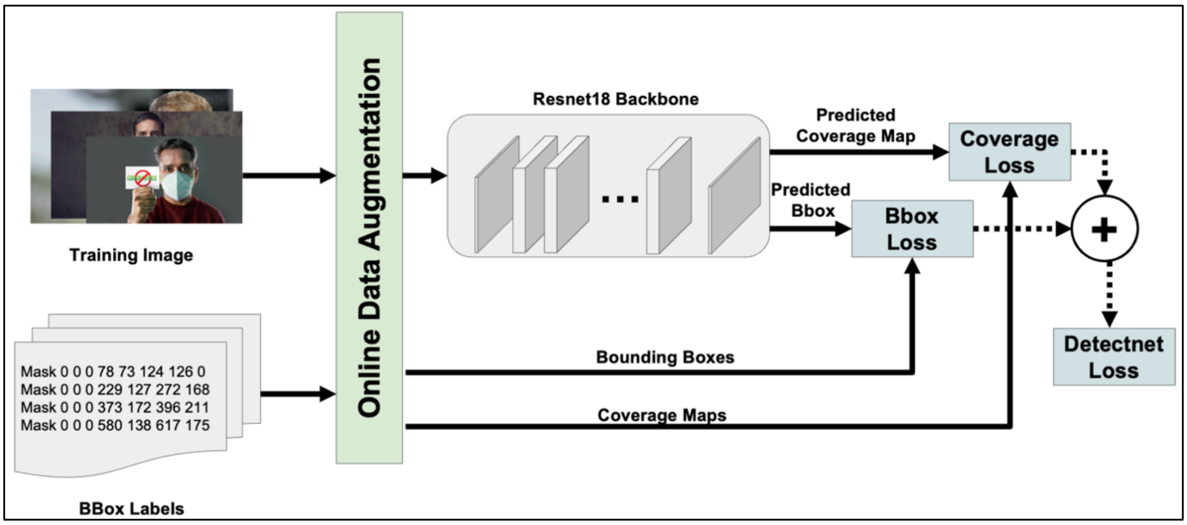


Figure 2. The training architecture of DetectNet V2.

$$\text{Coverage Loss} = \frac{1}{2N}\sum_{i=1}^{N}\left|\text{coverage}_i^t - \text{coverage}_i^p\right|^2, \quad (1)$$

where *coverage*^{t} means the true value from the ground truth and *coverage*^{p} means the predicted coverage value of the system; N denotes the batch size.

The second loss is related to the bounding boxes. Equation (2) shows that the bounding box losses were calculated by the mean absolute difference (L1) between the ground truth and the output-predicted bounding boxes.

$$\text{Bbox Loss} = \frac{1}{2N}\sum_{i=1}^{N} \begin{matrix} |x_{min}^t - x_{\min}^p| + |y_{\min}^t - y_{\min}^p| + \\ |x_{\max}^t - x_{\max}^p| + |y_{\max}^t - y_t^p| \end{matrix}. \quad (2)$$

In the last layer, DetectNet clustered and filtered those bounding boxes produced by the predictor. These bounding boxes with similar locations and sizes were then grouped together.

Inference speed is one of the important focuses as the mask detection system is a real-time application. Some parameters of the trained model have been pruned to reduce the size while preserving the model structure. As there may have been some useful weights that were removed during the pruning, it was retrained to recover the accuracy.

4. Experimental results

4.1. System setup

Figure 3 is the implementation of the model. To showcase the setup, it was placed at the entrance of a tutorial centre in Hong Kong. The monitor was placed on top of a shelf. As a person entered the view of the webcam, the monitor displayed a bounding box including the result and its degree of confidence. The webcam (C922 Pro HD Stream cam) was placed on a tripod, two metres away from and facing the entrance. It captured images of the entrance of the tutorial centre. The NVIDIA Edge Device (Jetson Xavier NX) was placed on top of a shelf, but it could be repositioned and hidden if necessary. The NVIDIA Jetson Xavier NX is a system-on-module supporting popular AI frameworks capable of running multiple neural networks simultaneously. It offers real-time inference of the mask detection model. The speaker will play an alarmed voice when a person without a mask is recognised.

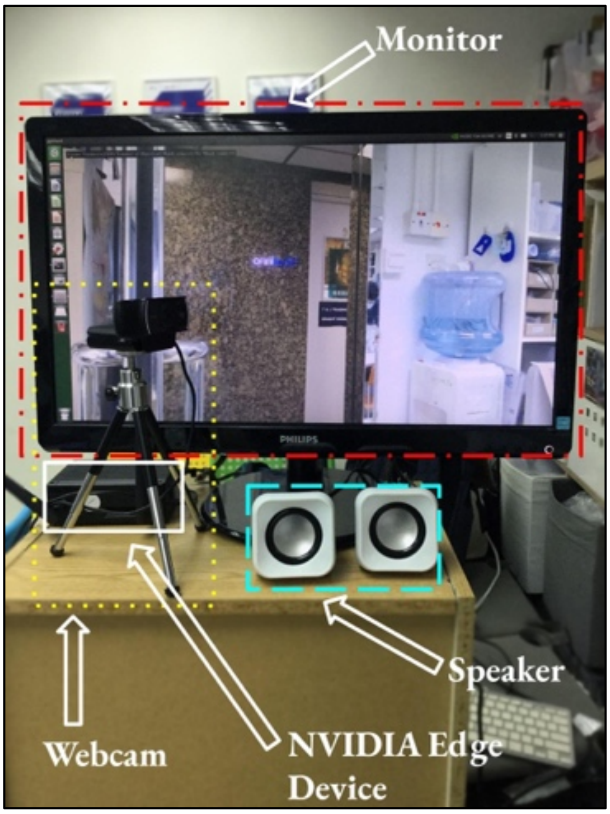


Figure 3. Overview of the system setup.

Figure 4 shows the outputs of the system under different scenarios. Figure 4(a) shows two people with masks directly facing the camera. Figure 4(b) demonstrates what will happen if one person is using proper mask etiquette or not. A red bounding box will appear on an unmasked face. Figure 4(c) shows the side view of one person with proper mask wearing etiquette and one without. Figure 4(d) includes the side view of one person without a mask and one person without proper mask etiquette. This proves that the system also works with different orientations of human faces to the camera.

(a)

(b)

(c)

(d)

Figure 4. System outputs in different scenarios. (a) Both students wearing masks; (b) one student wearing a mask while the other is not; (c) output of the side view containing one student wearing a mask while the other is not; and (d) output of the side view containing both students not wearing masks.

4.2. Implementation details

The proposed DetectNet v2 was implemented using the NVIDIA Tao toolkit, which is based on TensorFlow. In comparison, YOLOv3 tiny was developed based on the Darknet architecture. In total, there were 6,000 samples of mask-wearing pictures and 6,000 samples of ordinary human faces for the negative class. The datasets were further split into 80:20 for training and evaluation and augmented using online approaches while passing into the network for training. The training was performed on a NVIDIA A6000 GPU, and evaluated on a GPU and NVIDIA Xavier NX for accuracy and inference time respectively.

4.3. Metrics and comparisons

In the evaluation, two metrics were utilised for simple and direct comparison in the system case - (mean) average precision and inference time on the edge. The average precision per class and the overall mean Average Precision follows the standard in Pascal VOC Challenge (Everingham et al., 2014), which measures the AP from the results with the Intersection Over Union (IoU) greater than 0.5. Various models were evaluated by the validation datasets, and the APs were calculated based on the precision and recalls of two classes - wearing masks and not wearing masks. The camera streaming experiments were implemented on the same Jetson Xavier NX Developer Kit. It simulates a real-life application of face mask detection. The respective rates of frames per second (fps) of the application were considered the criteria to evaluate the ability of different models in real applications. The ideal model should have high average precision to prevent false positives and false negatives and low inference time / high frame rates for smooth and real-time monitoring. As DetectNet offers various choices on its feature extraction backbones, an extra two common backbones were chosen to be compared in the experiments, namely Resnet-50 and MobileNet (Howard et al., 2017). In the evaluation experiment the average precision was used as the parameter to compare the accuracy of the YOLOv3 tiny backbone and DetectNet v2 with various backbones. As DetectNet offers various choices of its feature extraction backbones, an extra two common backbones were chosen to compare in the experiments, namely Resnet-50 and MobileNet (Howard et al., 2017). In the evaluation experiment, the average precision was used as the parameter to compare the accuracy of the YOLOv3 tiny backbone and DetectNet v2 with various backbones.

Table 1 shows the average precision results of wearing a mask (AP_{mask}) and those without wearing a mask (AP_{no_mask}). The performance of YOLOv3 tiny is similar to that of DetectNet v2. Among various DetectNet v2 backbones, Resnet18 outperformed Resnet-50 and MobileNet. Table 2 shows the comparison of the frames per second (FPS) in the real-time video streaming with the YOLOv3 tiny detector and DetectNet v2 detector with various backbones that was performed. YOLOv3 tiny has a lower number of frames per second than DetectNet v2. For all DetectNet v2 backbones, fps stayed at 30 fps, which is the frame rate of the webcam used.

It was observed that DetectNet v2 demonstrates a greater ability to implement the detector in real time. From the results shown in Table 2, it was found that the DetectNet v2 detector is six times faster than the YOLOv3 tiny detector in the camera streaming experiments. The net result is that the DetectNet v2 is superior to the faster detector in terms of speed. In other words, the YOLOv3 tiny detector is time-consuming and costly in practical applications. Based on the limited resources, the DetectNet v2 detector performs better as its frame rate is significantly higher than that of the YOLOv3 tiny architecture with the same amount of work. Within the Resnet, it was observed that Resnet18 performs slightly better than Resnet50. This shows that YOLOv3 tiny is accurate but not quite feasible in real-time application as 5 fps is not acceptable for a quality face mask analysis. DetectNet v2 is the next choice, exhibiting nearly the same speed as the camera. Within the DetectNet, although MobileNet contains less trainable parameters than Resnet, which is faster in training and inference, it is less accurate. High false alarms will jeopardise the system's useability. Therefore, the remaining choices are Resnet-18 and Resnet-50. Resnet-18 exhibits a slighter higher accuracy in the case herein, and has lower parameters for training, and hence less training time.

To conclude, Resnet-18 is the best choice in the mask detection deployed in the NVIDIA edge devices.

Table 1. The average precision of YOLOv3 tiny and DetectNet v2 with various backbones.

Architecture	*Backbone*	AP_{mask}	$AP_{no\ mask}$	*mAP*
Darknet	YOLOv3 tiny	85.77	74.37	80.07
DetectNet v2	MobileNet v2	74.9859	79.7492	77.3675
DetectNet v2	Resnet50	82.3037	82.9887	82.6462
DetectNet v2	Resnet18 (Ours)	**83.5824**	**83.3965**	**83.4895**

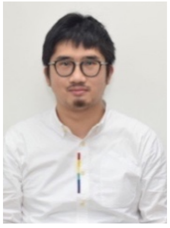
Table 2. The fps of YOLOv3 tiny and DetectNet v2 with various backbones in camera streaming.

Architecture	*Backbone*	*Frames per second (fps) in the real-time video inference system*
Darknet	YOLOv3 tiny	5
DetectNet v2	MobileNet v2	30.9
DetectNet v2	Resnet50	29.0
DetectNet v2	Resnet18 (Ours)	30.4

5. Conclusion

Throughout the past few years, the world has been battling against the global COVID-19 pandemic. This disease is transmitted more vigorously in crowded areas and to combat the coronavirus, it is of paramount importance for people to be masked. Governments all around the world have enforced mandatory rules stating that citizens should always wear a mask in public areas, especially in crowded places. It would pose a manpower challenge to monitor day and night. The purpose of the proposed model is to detect if people are wearing masks properly and reduce the dependency on a human workforce. The proposed model uses DetectNetv2 trained with three datasets (MAFA, Kaggle Medical Mask, and FDDB), and is paired with a monitor, webcam, and NVIDIA Edge Device. The proposed model using DetectNet v2 with a Resnet-18 backbone achieved a high accuracy of 83.49 in mAP while achieving 30.4 fps in camera streaming. During the trial phase, the model detected people from different camera angles accurately. This means that the system can achieve a high accuracy while detecting face masks in real time, with no data being transmitted over the internet, which protects the privacy of citizens in the field of view of the camera. It is currently placed and implemented at the entrance of a tutorial centre in Hong Kong. In the future, it is planned to include temperature screening, as well as integrating it into an active CCTV system.

Notes on contributors



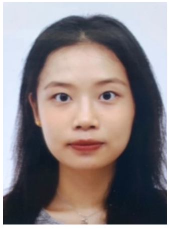
Dr Ho Chuen Kam received his Ph.D. (2021) in Computer Science and Engineering from The Chinese University of Hong Kong, and B.Sc. (2013) in Computer Science from the Hong Kong Baptist University. He is a data scientist and software engineer at Blueinno Technology, and a research intern at NVIDIA Corporation. His research interests include smart healthcare, human understanding, computer vision and artificial intelligence.

Mr Nathan Chung Kan Lui is currently a Year 11 student at German Swiss International School in Hong Kong. He is pursuing a career in engineering, with interests in computer science and physics, and multiple programming languages. He is also fluent in English, Mandarin, Cantonese, and German. Currently, he is a research mentee at Blueinno Technology, providing assistance to the A.I. projects.

Mr Ethan Yi Chen Du has experience in programming. He is interested in computer science, visual arts, as well as physics and engineering. He is currently a Year 10 student at German Swiss International School in Hong Kong. He is a research mentee at Blueinno Technology, and is contributing to A.I. research projects.

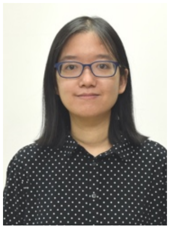
Mr Marcus Ngai Kan Lui has five years of experience in programming in different languages. He is passionate about computer science and has interests in a career in software engineering. Currently, he is a Year 9 student at German Swiss International School in Hong Kong.

Mr Chun Sing Chan received a bachelor's degree from The Hong Kong Polytechnic University in Electronic & Information Engineering. He is currently a software engineer at Blueinno Technology and focuses his research and development on the AI medical field.



Ms Yan Ting Cheung was raised in Hong Kong and is enthusiastic about artificial intelligence and finance. She is currently pursuing a bachelor's degree in computer science at The University of Hong Kong and is working as an intern for Blueinno Technology.

Dr Ka Chun Cheung received his B.Sc. (2011) and Ph.D. (2016) in applied Mathematics from Hong Kong Baptist University focusing on meshless collocation methods for solving partial differential equations. He was a technical manager in the Hong Kong Applied Science and Technology Research Institute leading a team to develop a machine vision project. He joined NVIDIA in 2018 as a deep learning solution architect. His research interests include numerical analysis, partial differential equations and machine learning.



Ms Monica Sze Man Leung is the Founder and Director of Blueinno Technology Limited, which focuses on STEM and AI Education. Monica holds a Master of Science and a Bachelor of Engineering (Hon) major in Electronic Engineering from The Chinese University of Hong Kong. Monica has 15+ years of research and development experience in software development and engineering of product solution design. She is currently an Executive Committee Member of the Hong Kong Electronics & Technologies Association (HKETA) and is also a Committee Member of the Computer Science Advisory Committee (CSAC) of The University of Hong Kong (HKU).

References

[1] Anand Sonawane (2018). YOLOv3: *A Huge Improvement*. [online]. Available at: <https://medium.com/@anand_sonawane/yolo3-a-huge-improvement-2bc4e6fc44c5>. [Accessed on 6 August. 2021].

[2] Balaji S, Balamurugan B, Ananth KT, Rajmohan R and Praveen KP (2021). *A brief survey on AI based face mask detection system for public places. Irish Interdisciplinary Journal of Science & Research (IIJSR)*.

[3] Bao L, Li T, Xia X, Zhu K, Li H and Yang X (2020). *How does working from home affect developer Productivity?-A case study of baidu during COVID-19 pandemic. arXiv preprint arXiv:2005.13167*.

[4] BBC (2020). Coronavirus France: Cameras to monitor masks and social distancing. *BBC News*. [online]. Available at: <https://www.bbc.com/news/world-europe-52529981>. [Accessed on 6 August 2021].

[5] Bochkovskiy A, Wang C-Y and Liao H-YM (2020). YOLOv3: Optimal Speed and Accuracy of Object Detection. *arXiv:2004.10934 [cs, eess]*. [online]. Available at: <https://arxiv.org/abs/2004.10934>.

[6] Du J (2018). Understanding of Object Detection Based on CNN Family and YOLO. Journal of *Physics: Conference Series*, 1004, p.012029. doi:10.1088/1742-6596/1004/1/012029.

[7] Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J and Zisserman A (2014). The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1), pp. 98-136.

[8] Ge S, Li J, Ye Q and Luo Z (2017). Detecting masked faces in the wild with lle-cnns. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2682-2690.

[9] Geiger A, Lenz P, Stiller C and Urtasun R (2013). Vision meets robotics: The kitti dataset. The *International Journal of Robotics Research*, 32, pp. 1231-1237.

[10] Goscé L and Johansson A (2018). Analysing the link between public transport use and airborne transmission: mobility and contagion in the London underground. *Environmental Health*, 17, pp. 1-11.

[11] He K, Zhang X, Ren S and Sun J (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770-778.

[12] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M and Adam H (2017). MobileNets: *Efficient Convolutional Neural Networks for Mobile Vision Applications*. [online]. arXiv.org. Available at: <https://arxiv.org/abs/1704.04861>.

[13] Hui J (2019). *Object detection: speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3)*. [online] Medium. Available at: <https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359>. [Accessed on 6 August 2021].

[14] Jain V and Learned-Miller E (2010). *Fddb: A benchmark for face detection in unconstrained settings*. UMass Amherst technical report.

[15] Katal A, Albettar M and Wang LL (2021). *City reduced probability of infection (CityRPI) for indoor airborne transmission of SARS-CoV-2 and urban building energy impacts. medRxiv*. Cold Spring Harbor Laboratory Press.

[16] Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M and Kolesnikov A (2018). *The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982*.

[17] Lewis D (2021). Why indoor spaces are still prime COVID hotspots. *Nature*, 592, pp. 22-25.

[18] Li Z, Zhao L, Han X and Pan M (2020). Lightweight ship detection methods based on YOLOv3 and DenseNet. *Mathematical Problems in Engineering*, 2020.

[19] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y and Berg AC (2016). Ssd: Single shot multibox detector. In: *Springer*. pp. 21-37.

[20] Qin B and Li D (2020). Identifying Facemask-Wearing Condition Using Image Super-Resolution with Classification Network to Prevent COVID-19. *Sensors*, 20(18), pp. 5236. doi:10.3390/s20185236.

[21] Rana K and Kathirvel S (2020). Unsung Heroes in Managing COVID 19 Pandemic in India: the changed role of security guards in hospitals. *International Journal of Health Systems and Implementation Research*, 4, pp. 5-10.

[22] Redmon J, Divvala S, Girshick R and Farhadi A (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [online]. Available at: <https://arxiv.org/pdf/1506.02640.pdf>.

[23] Shepherd JP, Moore SC, Long A, Mercer LM. and Sumner SA (2021). Association between COVID-19 lockdown measures and emergency department visits for violence-related injuries in Cardiff, Wales. *JAMA*, 325, pp. 885-887.

[24] Shorten C and Khoshgoftaar TM (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, pp. 1-48.

[25] Taurus Protection Inc. (2020). *Security Risking Their Lives for Mask Enforcement*. [online] www.securitybytaurus.com. Available at: <https://www.securitybytaurus.com/post/security-risking-their-lives-for-mask-enforcement>. [Accessed on 3 August 2020].

[26] Thuan D. (2021). *Evolution of YOLO algorithm and YOLOv5: The State-of-the-Art object detention algorithm*. [online] www.theseus.fi. Available at: <https://www.theseus.fi/handle/10024/452552>. [Accessed on 27 January 2022].

[27] Vyas L and Butakhieo N (2021). The impact of working from home during COVID-19 on work and life domains: an exploratory study on Hong Kong. *Policy Design and Practice*, 4, pp. 59-76.

[28] Waghe S (2020). *Medical Mask Dataset*. [online] kaggle.com. Available at: <https://www.kaggle.com/shreyashwaghe/medical-mask-dataset>. [Accessed on 5 August 2020].

[29] World Health Organization (2020a). *Coronavirus disease (COVID-19): How is it transmitted?* [online]. Available at: <https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted>. [Accessed on 2 August 2021].

[30] World Health Organization (2020b). *When and how to use masks*. [online]. Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks>. [Accessed on 2 August 2021].

[31] World Health Organization (2020c). *WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020*. [online]. Available at: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. [Accessed on 3 August 2021].

[32] World Health Organization (2021). *WHO COVID-19 dashboard*. [online]. Available at: <https://covid19.who.int>. [Accessed on 27 June 2021].

[33] Yadav S (2020). Deep Learning based Safe Social Distancing and Face Mask Detection in Public Areas for COVID-19 Safety Guidelines Adherence. *International Journal for Research in Applied Science and Engineering Technology*, 8(7), pp.1 368-1375. doi:10.22214/ijraset.2020.30560.

[34] Yang S, Luo P, Loy C-C and Tang X (2016). Wider face: A face detection benchmark. In: *roceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5525-5533