DOI: 10.37943/12TXQS9259

Saya Sapakova
Cand. of ph. and math. sc., Associate Professor of the Department of Computer Engineering
s.sapakova@iitu.edu.kz, orcid.org/0000-0001-6541-6806
International University of Information Technology, Kazakhstan

Yelidana Yilibule
Master student of the Department Information Systems
13899886566danaxi@gmail.com, orcid.org/0000-0002-6686-8967
Al-Farabi Kazakh National University, Kazakhstan

DEEP LEARNING-BASED FACE MASK DETECTION USING YOLOV5 MODEL

Abstract: Based on the background of rapid transmission of novel coronavirus and various pneumonia, wearing masks becomes the best solution to effectively reduce the probability of transmission. For a series of problems arising from crowded public places and collective units, where face recognition is difficult to increase target density, a deep convolutional neural network is used for real-time mask detection and recognition.

This paper presents the method based on YOLOv5 model for deep learning and mask detection in image recognition as well as a live camera to label the pedestrians without masks in time. This experiment will use LabelImg software to preprocess 5003 images and make lightweight improvements based on the original YOLOv5 model to generate the final face mask recognition model. The Mosaic method is added to merge the images effectively and process the images in batch, and secondly, the GIoU loss function is selected to calculate the bounding box regression loss by comparison, which improves the localization accuracy even more. According to the experimental detection results, analogized with the original model YOLOv5, the recall and accuracy are effectively improved. In this paper, YOLOv3, SSD, Fast-R-CNN detection algorithms are used for comparison, the detection results of this model have a high mAP value which is equal to 92.9, which are higher than the detection results of other models.

Real-time target recognition based on this model combined with practical applications can be applied in hospitals and crowd-gathering places to achieve effective reduction of epidemic transmission probability in a short period of time.

Keywords: YOLOv5, CNN, covid-19, target detection, deep learning

Introduction

With the outbreak of novel coronavirus [1] and related pneumonia, any public place, medical place, and any place where people gather need to wear masks to protect themselves and others, and the efficiency of people wearing masks reflect the progress and results of outbreak prevention and control. Therefore, it is very important to detect in real-time whether people are wearing masks in crowded places. In the environment of respiratory class virus infection, the requirements related to the wearing of masks will be specifically reminded and supervised by the relevant person in charge, in this process, although it effectively improves the chances of the crowd wearing masks, but consumes a huge amount of time and costs, and there is also a considerable degree of risk of infection from close contact with the detectors, so the creation

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

of automatic face mask identification system for the epidemic and daily protective measures is quite important.

As face mask detection belongs to the category of object detection, target detection is for the computer to automatically detect objects in a video or image and learn their features. The more popular method in today's society is based on deep learning, which is further divided into two types due to the different detection steps: two-stage and single-stage. The representative algorithms belonging to the two-stage are: R-CNN, SPP-NET, fast R-CNN, Faster R-CNN and others; single-stage algorithms are YOLOV1-V4, SSD, RetinaNet, FPN. In the process of recognizing faces, the mask obscures the face recognition features to some extent and poses a great challenge to this technique based on the certain influence of the detection environment.

M.D. Pramita [2] proposed a deep learning-based image classification model to detect whether a person is wearing a mask in real time by learning Indonesian face features from a private dataset using convolutional neural networks. Chengshuo Cao [3] introduced the attentional feature network SENet into the YOLO algorithm and proposed a YOLO-Mask algorithm to improve the detection capability of mask wearing target in different scenarios. The improved Retinaface algorithm is proposed to determine whether the mask is correctly worn based on the detection of mask wear.

In order to achieve better usage results, this paper conducts a study on a real-time mask-wearing detection algorithm in complex scenarios using YOLOv5 network model based on the above study. It is verified that this algorithm takes into account the accuracy and speed in the real-time detection of mask-wearing, which indicates good practicality.

YOLOv5 model

YOLO (You Only Look Once) is a high-performance general-purpose object detection model, YOLOv1 [4] used a first-order structure to complete the two tasks of classification and target localization, followed by YOLOv2 [5] and YOLOv3 [6] to achieve improvements in speed and accuracy, mainly in the change of the network structure, Backbone changed to Darknet-53, Darknet-53, ResNet-101 or ResNet-152 with similar accuracy, which is faster and further promotes the application of object detection in industry, and YOLOv4 [7] achieves the model training on an ordinary GPU (1080Ti), which can be considered as a clever combination to the architecture of YOLOv4 includes the CSPDarknet53 backbone, the SPP add-on module, the path – aggregation neck and the YOLOv3 (anchor based) header.

Since YOLOv1, the YOLO series has evolved to YOLOv5, which is more flexible than YOLOv4. The YOLOv5 project is created and maintained by Ultralytics Inc. The YOLOv5s model in YOLOv5 model version 6.0 is used for the experiments, and its network model structure (Figure 1). It is mainly divided into four parts: Input, Backbone, Neck and Prediction. It provides four versions, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, with increasing size and accuracy, and adopts the channel and layer control factors similar to EfficientNet [8] to realize the version change according to the Bottleneck number distinction. In practical applications, the appropriate model size can be selected according to the specific scenario.

DOI: 10.37943/12TXQS9259

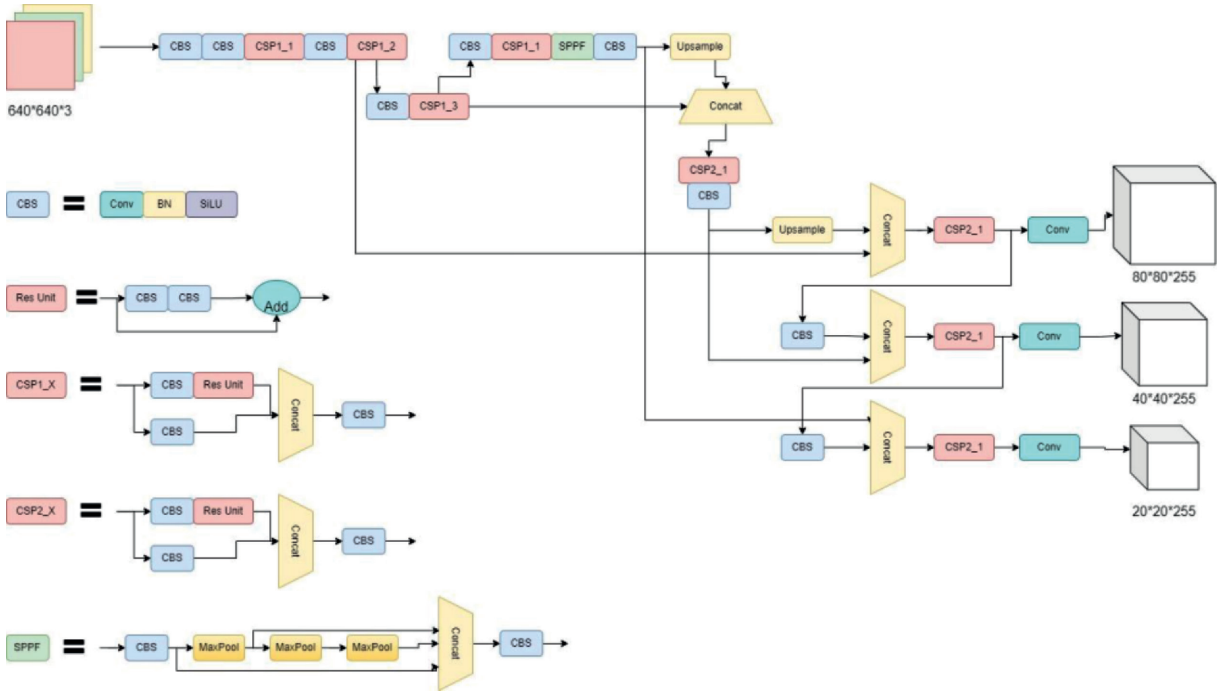


Figure 1. structure of YOLOv5 model

- Input

The Input part consists of image size scaling, adaptive anchor frame calculation, and Mosaic data augmentation. In general, the algorithm will scale down the images in the dataset to a uniform size before training the model, so the Mosaic data enhancement method randomly selects four images arbitrarily scaled, cropped, and arranged in a way to stitch, and then send them to the network to start training after processing the label information, which can effectively improve the detection effect for small targets.

- Backbone

The Backbone part mainly consists of the Focus structure, SPP (Spatial Pyramid Pooling) [9] structure and CSP structure [10], where the network extracts the features of high, medium and low layers of the image. the SPP is pooled and then the features of the output layer are combined and superimposed using Concat. The CSP structure involves two parameters, depth_multiple and width_multiple, which make the whole Backbone design more flexible.

- Head

The Head part mainly uses the combination of FPN (Feature Pyramid Network) + PAN (Path Aggregation Network) for downsampling and upsampling, which can pass features and perform feature fusion to improve the efficiency of prediction.

- Prediction

The structure of Prediction consists of classification loss, bounding box regression loss and NMS (Non-Maximum Suppression).GIoU focuses not only on the overlapping region but also on other non-overlapping regions, which can better reflect the overlap between the two and can keep the prediction frame and the real frame close to each other by minimizing GIOU_loss. The binary cross-entropy loss is used to calculate the classification loss. The role of NMS is to ensure that the candidate box with the highest prediction probability becomes the final prediction box.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

Detection algorithm optimization

In order to better apply the YOLOv5 model to the large mask detection environment, appropriate improvements and training are needed to accomplish the mask detection task. First, the GIoU loss function is chosen to calculate the bounding box regression [11] loss to speed up the convergence of the model.

Mosaic Data Enhancement

In the common case, if the images in the dataset are of different sizes, these materials will be scaled to a uniform size during training and testing before feeding into the network. However, this approach does not yield satisfactory results for small targets, so this paper will use Mosaic data augmentation on the input side, adaptive image scaling, and other methods.

The Mosaic implementation is located in the load_mosaic() function in datasets.py, and four images are randomly selected, and the parts of them are taken into this image, as shown in the following figure, four colors represent four sample images, and the parts beyond will be discarded. This method of data augmentation can effectively detect small target ranges and image backgrounds, and efficiently achieve real-time detection [12] of solid targets.

The specific process of Mosaic data enhancement: in the case of algorithmic processing, the images in the dataset are randomly stitched and a certain degree of the gray border is added, and finally the images are scaled to a standard size and sent to the network for training (Figure 2).

This figure shows the detection results based on combining multiple images for target recognition, the target recognition status is essentially correct, the face/mask status has different recognition rates based on the blurred degree of the images and external influencing factors, and the high recognition rate of 0.9 can be achieved when the features are complete.

Figure 2. Mosaic data enhancement

Bounding Box Regression loss function

The commonly used loss functions for target frame regression in the field of target detection are IoU loss, GIoU loss, DIoU loss (Distance IoU), CIoU loss [13] (Complete IoU), etc. Among them, the CIoU loss calculation method takes into account the overlap area, centroid distance and aspect ratio between the predicted frame and the real frame, which theoretically works best.

DOI: 10.37943/12TXQS9259

IoU [14] (Intersection over Union) is known as the intersection ratio and is the most commonly used index in target detection. In anchor-based methods, it is not only used to determine positive and negative samples but also to evaluate the distance between the predicted box and the ground-truth due to its scale invariant property, which is display in (1).

$$\mathrm{IoU} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{1}$$

GIoU [15] (Generalized IoU), on the other hand, introduces penalties on top of IoU to reflect more accurately the interaction between the detection frame and the real frame, as in (2).

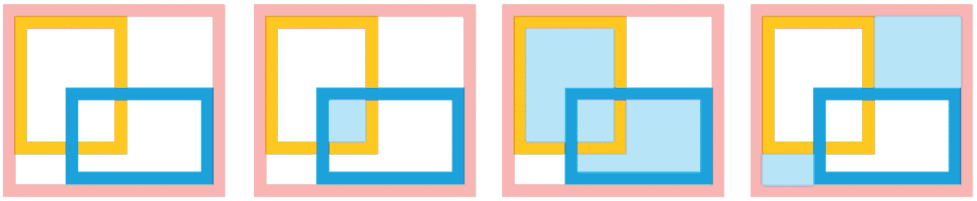
$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|\mathrm{C} - (\mathrm{A} \cup \mathrm{B})|}{|\mathrm{C}|} \tag{2}$$

In the formula, A is the detection box, B is the real box, and C represents the minimum external rectangular box between them. During the experiment, it was found that the indicators of GIoU loss were better than CIoU, and GIoU compared with IoU not only focused on the overlapping area of the predicted box and the real box, but also on the non-overlapping area of the two, whose calculation formula is shown in (3), and the meaning of the symbols in the formula (Figure 3).

$$L_{GIoU} = 1 - GIoU = 1 - \left(IoU - \frac{|C - (A \cup B)|}{|C|} \right) = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{|C - (A \cup B)|}{|C|} \right) \tag{3}$$

The blue box in the figure is the prediction box A, the red box is the ground truth box B, and the yellow box C is the smallest box that can completely surround A and B.

- Prediction box: the box calculated from the output of the target detection model.
- Ground truth box: is the location of the manual annotation, stored in the annotation file.
- Bounding box: used to identify the position of the object, the common format is top-left and bottom-right coordinates.



(a) A, B, C (b) $A \cap B$ (c) $A \cup B$ (d) $C - (A \cup B)$

Figure 3. Bonding regression loss of symbolic meaning

Experiments and analysis of results

Experimental environment and data set

The dataset used in this study is RMFD (Real-World Masked Face Dataset) from the National Multimedia Software Engineering Technology Research Center of Wuhan University. 5003 images were collected for experiment and training, and the images were divided into training and validation sets in the ratio of 8:2. In the face mask detection system, there are two types of target entities to be located: faces with masks and faces without masks. Therefore, the process of face-mask detection system is defined: select a picture, or real-time camera detection, output the entity detected as face in that picture and give the corresponding detection data (0: face, 1: mask). The experimental steps are as follows:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

1) Import the dataset into the computer and use the LabelImg software to manually create a marker box for the pictures, label for people wearing masks as mask, and label for people without masks as face, all the pictures will generate the corresponding xml files, and the pictures in the dataset can start training directly.
2) The images outside of dataset need to be converted to txt files to generate the corresponding labels and target box locations.
3) Check the results, each generated txt file has relevant information data, there are 0 and 1 for no mask and mask, and other parameters represent the x, y coordinate position of the recognition frame.

Before starting the model learning training, the relevant learning values need to be given. We scale the image size to 640*640*3, run batch as 16, epoch equal to 200, and initial learning rate equal to 0.01 to start the training.

Evaluation Indicators

The evaluation metrics of YOLOv5 model include Precision, Recall, Average Precision, mAP(Mean Average Precision), P-R curve [16] and so on. The precision of the model mainly depends on AP, mAP, and P-R curve, which can effectively demonstrate the performance of the model in detecting and learning images.

Average Precision is calculated from the area enclosed by the P-R curve and the coordinates, and the horizontal axis is the recall rate and the vertical axis is the precision rate, when the PR curve is closer to the upper right, the better the model performance. Where recall is the probability that the correct category in the sample is predicted correctly use (4).

$$\mathrm{Recall} = \frac{\mathrm{T_P}}{\mathrm{T_P} + \mathrm{F_N}} \tag{4}$$

The T_P in the formula represents the true positive in the confusion matrix, that is, the number of categories that will predict the correct rate as positive, and F_N represents the number of categories that are false negative. In this experiment, the result is approximately equal to 0.9.

Precision rate, which means how many of the samples in the dataset that are predicted to be positive are truly positive samples. In general, a high precision rate leads to a lower recall rate, so we need the P-R curve to accurately evaluate the improved model according (5).

$$Precision = \frac{T_P}{T_P + F_P} \tag{5}$$

The mAP [17] value represents a measure of recognition accuracy, and the P-R curve can be plotted by calculating the highest accuracy rate at different recall rates, and the area enclosed by this curve is the AP value for that category. For a binary classification problem, a threshold is often set, and when the prediction value is greater than this threshold, the prediction is a positive sample, and when it is less than this threshold, the prediction is a negative sample. If Recall is the horizontal axis and Precision is the vertical axis, then when setting a threshold, a point is drawn on the axis, and when setting multiple thresholds, a curve can be drawn, which is the PR curve.

If the requirement is to predict the face state as accurately as possible, then it is to improve precision; and if it is to predict as much as possible the potential population of whether or not to wear a mask, then it is to improve recall. In general, increasing the threshold of binary classification can improve the precision, and decreasing the threshold can improve the recall, then we can observe the PR curve and get the optimal threshold.

DOI: 10.37943/12TXQS9259

The result shows the P-R curve of the improved YOLOv5 model after learning training through 5003 images of the dataset (Figure 4). In the figure show that the average value of detected faces is 0.965, the average value of detected faces with masks is 0.894, and Mean Average Precision is equal to 0.929. These parameters represent this experiment for identifying the unmasked state over the masked state, theoretically, the improvement in precision decreases the recall rate, so the experimental show better detection results.

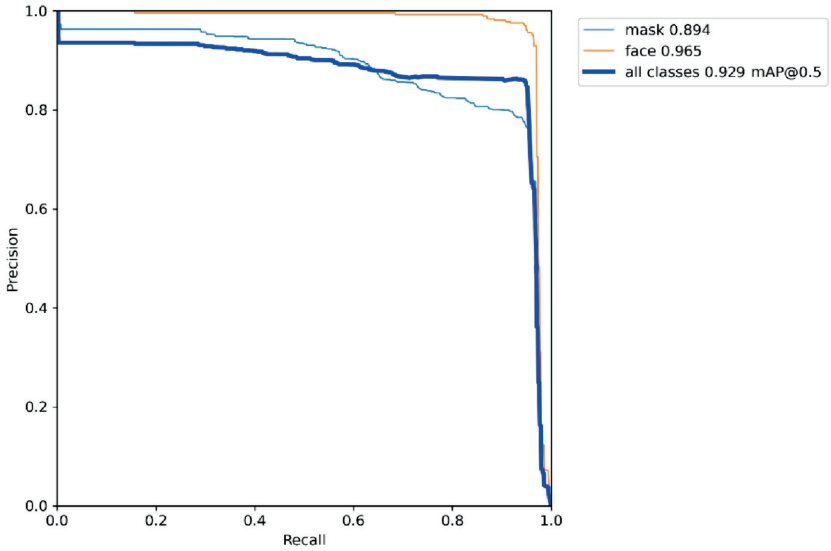


Figure 4. P-R curve and Mean Average Precision

FPS measures the number of frames transmitted per second. Due to the special structure of the human senses, the brain perceives a picture as coherent when the frame rate seen by the naked eye is higher than 16 FPS, a phenomenon also known as visual transience [18]. It has been verified that the maximum number of frames per second detected by the algorithm in this paper can reach up to 30.0 FPS.

Analysis of results

After the training of the model, in order to apply this model to the reality, by using the UI image interface in the PyCharm compiled environment, you can manually select the detection images from the folder or use the camera for real-time detection, this model can effectively complete the face mask detection. And the output result is labeled with mask and predicted values for face with mask and without mask.

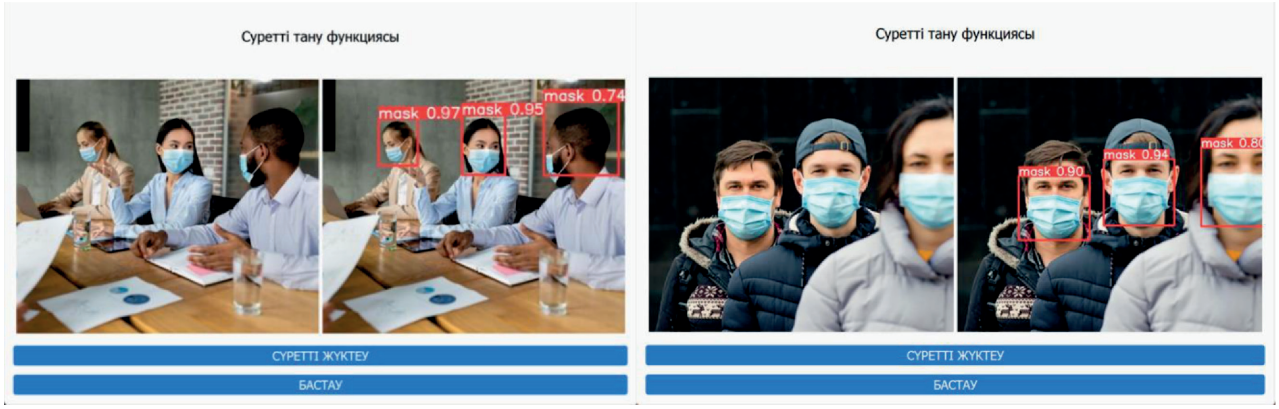


Figure 5. Detection result

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

For the case of multiple people in the picture, it also works well, there are no missed detections in the multi-target detection. The results of the YOLOv5 face mask detection show a good result in the experiment (Figure 5). In the figure depending on the blurred degree of the picture or the position of the person, in the experimental results, a value of approximately 0.9 or more is given for the case of specific facial features of the object in the left figure, while a mask determination status of 0.74 is given for the side face and the state of incomplete facial features. The mask recognition rate will also rise or decrease, based on whether to wear a mask and the given mask or face state is basically judged correctly.

Comparison of algorithms

In this research paper, an algorithm comparison is used to demonstrate the usefulness of the YOLOv5 improved model. Other common target detection algorithms including SSD, Fast-R-CNN, and YOLOv3 are assembled, and the comparison of different algorithms significantly reveals the scientific and effectiveness of this model (Table 1).

The results of the comparison are as follows:

Table 1. Comparison of target detection algorithms

Detection algorithm	mAP/%	FPS	Mask/%	Face/%
YOLOv3	83.97	43.92	82.71	85.23
SSD	79.23	15.6	77.26	81.20
Fast-R-CNN	80.45	6.7	82.35	78.56
YOLOv5	92.90	30	89.40	96.50

The experimental results prove that the improved YOLOv5 face mask detection algorithm used in this paper can distinguish masked faces from unmasked faces with high accuracy, and the Mean Average Precision is significantly higher than other target detection algorithms, which verifies the superiority of this model.

Conclusion

To summarize, to balance the accuracy and speed of mask wear detection in complex scenes, this paper proposes a real-time mask wear detection algorithm based on YOLOv5 in complex scenes. The algorithm has good performance in terms of accuracy, recall and other evaluation indexes, and can meet the real-time requirements of real-time camera detection, so it shows that this method has certain advantages. The detection rate is low in the video sample with too many targets, and there is a certain degree of missing detection in the case of obscured targets, unclear targets and small targets. In addition, the algorithm needs to be further investigated and improved by combining mask-wearing features with other associated features to achieve a more applicable safety and health surveillance system to better meet the actual needs of society and daily life.

References

1. Zhao, W. M., Song, S. H., Chen, M. L., Zou, D., Ma, L. N., Ma, Y. K., ... & Bao, Y. M. (2020). The 2019 novel coronavirus resource. *Hereditas*, *42*(2), 212-221.
2. Pramita, M. D., Kurniawen, B., & Utame, N P. (2020). Mast wearing classification using CNN, *7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE, 1-4.
3. CAO, C.S., & YUAN, J. (2021). YOLO-Mask algorithm based mask wear detection method. *Advances in Excitation and Optoelectronics*, *58*(8), 211-218

4. Review: YOLOv1 – You Only Look Once (Object Detection). (2022). Retrieved from https://towardsdatascience.com/yolov1-you-only-look-once-object-detection-e1f3ffec8a89
5. Sharma, A. (2022). A Better, Faster, and Stronger Object Detector (YOLOv2) – PyImageSearch. Retrieved from https://pyimagesearch.com/2022/04/18/a-better-faster-and-stronger-object-detector-yolov2/
6. Papers with Code – YOLOv3 Explained. (2022). Retrieved from https://paperswithcode.com/method/yolov3
7. Papers with code – yolov4 explained. Explained | Papers With Code. (n.d.). Retrieved from https://paperswithcode.com/method/yolov4
8. Jiang, W. (2022). Study on resnet and EfficientNet Remote Sensing Image Scene Classification. *Computer Science and Application*, *12*(5), 1301–1313. https://doi.org/10.12677/CSA.2022.125130
9. Shi, L., Zhou, Z., & Guo, Z. (2021). Face anti-spoofing using Spatial Pyramid pooling. 2020 *25th International Conference on Pattern Recognition (ICPR)*. https://doi.org/10.1109/ICPR48806.2021.9412407
10. Brailsford, S. C., Potts, C. N., & Smith, B. M. (1999). Constraint satisfaction problems: Algorithms and applications. *European journal of operational research*, *119*(3), 557-581.
11. Wang, X., & Song, J. (2021). ICIOU: Improved loss based on complete intersection over union for bounding box regression. *IEEE Access*, *9*, 105686–105695. https://doi.org/10.1109/ACCESS.2021.3100414
12. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.91
13. Alexey Bochkovskiy, Chien-Yao Wang, & Hong-Yuan Mark Liao. (2020). YOLOv4: Optimal speed and accuracy of object detection. *ArXiv: Computer vision and pattern recognition*.
14. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(07), 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999
15. Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, & Dongwei Ren. (2019). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. ArXiv: Computer Vision and Pattern Recognition. Retrieved from http://export.arxiv.org/pdf/1911.08287
16. Khan, S. A., & Ali Rana, Z. (2019). Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR). 2019 *2nd International Conference on Advancements in Computational Sciences (ICACS)*. https://doi.org/10.23919/ICACS.2019.8689135
17. Pereira, N. (2022). PereiraASLNet: ASL letter recognition with Yolox taking mean average precision and inference time considerations. 2022 *2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. https://doi.org/10.1109/AISP53593.2022.9760665
18. Aagten-Murphy, D., & Bays, P. M. (2019). Independent working memory resources for egocentric and allocentric spatial information. *PLOS Computational Biology*, *15*(2), e1006563. https://doi.org/10.1371/journal.pcbi.1006563