A Computer Vision-based System for Surgical Waste Detection

Md. Ferdous, Sk. Md. Masudul Ahsan
Dept. of Computer Science and Engineering (CSE)
Khulna University of Engineering & Technology (KUET)
Khulna, Bangladesh

***Abstract*—The world population is going through a difficult time due to the pandemic of COVID-19 while other disasters prevail. However, a new environmental catastrophe is coming because surgical masks and gloves are putting down anywhere, leading to the massive spreading of COVID-19 and environmental disasters. A significant number of masks and gloves are not properly managed. They are scattered around us such as roads, rivers, beaches, oceans and other places. Since these types of waste are turned into microplastics and chemicals are deadly harmful to the environment, human health and other species, especially for the aquatic animals on this planet. During the outbreaks of corona pandemic, surgical waste in the open place or seawater can create a fatal contagious environment. Putting them in a particular area can protect us from spreading infectious diseases. This study proposed a system that can detect surgical masks, gloves and infectious/biohazard symbols to put down infectious waste in a specific place or a container. Among the various types of surgical waste, this study prefers mask and gloves since it is currently the most widely used element due to the COVID-19. A novel dataset is created named MSG (Mask, Biohazard Symbol and Gloves), containing 1153 images and their corresponding annotations. Different versions of the You Only Look Once (YOLO) are applied as the architecture of this study; however, the YOLOX model outperforms.**

***Keywords*—COVID-19; You Only Look Once (YOLO); surgical waste; deep learning; image dataset; real-time detection**

I. Introduction

Plastics have become a severe hazard to natural habitats and human health. Moreover, some of them are recyclable e.g. PET bottles. During the COVID-19 pandemic, surgical masks and gloves have increased extensively to reduce coronavirus spread. They are not reusable for being medical waste and infectious. People throw masks and gloves everywhere as a general waste due to the lack of planning and unconsciousness. Therefore, it is our prime duty to manage them properly. Otherwise, we will have to face extreme catastrophes. This hazard will likely be accelerated because of excessive use and exhaustion of plastic for example surgical masks, surgical gloves, face shields and personal protective equipment (PPE). A disposal system that can accurately identify masks, gloves, and biohazard symbols may participate in managing such type of terrible waste safely. This study preferred the biohazard symbol because symbols are unique all over the world than language. The outline is such that the system will identify masks and gloves as waste and keep these waste in a particular place/container/waste bin where the biohazard symbol is drawn. In this case, Computer Vision (CV) may help us a lot. There are different object detection models available right now. The YOLO models have the highest popularity because of their speed and auspicious performance. Although, in recent years many anchor-free object detection models [1] [2], Non-maximum Suppression (NMS) free i.e. end to end models [3] [4] [5] has been deployed. Training the model based on anchor creates a problem at the time of initialization of anchors. Rather, the anchor-free strategy does not face such types of problems. YOLO family always tries to execute the latest technology (e.g. YOLOv2 [6] anchor mechanism, YOLOv3 [7] Residual Net) to enhance speed and optimized implementation within a desirable time. Additionally, YOLO architecture has two mechanism-based models, one is anchor-based and the other is anchor-free training strategy. YOLOv3, YOLOv4, and YOLOv5 depend on anchor technology, whereas YOLOX relies on an anchor-free training mechanism. YOLOv3 is a mainly used model for industrial purposes and still exists many versions. YOLOv3 [7] focuses on layer-wise feature extraction and does not pay any attention to the sequential impact among the layers. Feature extraction performs using Darknet-53. YOLOv3-spp is another version of YOLOv3, which uses the spatial pyramid pooling (spp) module into the model and produces better performance than the other. YOLOv4 [8] and YOLOv5 [1] are two newly published architectures and both of them show comparable performance in many applications. There are different versions of YOLOv5 based on the model size such as small, medium, large and extra-large. The model architecture for each version remains the same. However, the only difference is the model depth and width. This theory also applies to YOLOX's different version. In this study, the authors explore a novel dataset and apply different CV models to determine which model yields the best performance to achieve the goal. In particular, four models are presented to identify surgical waste and biohazard symbols accurately. Finally, one of them is selected as the proposed model. Additionally, collecting relevant images, creating annotations and preparing the dataset are also discussed. The remaining sections of the paper are organized as follows: Section II describes the literature review of surgical waste detection. Dataset preparation and methods are discussed in Section III. The model's training process is expressed in Section IV. The performance measurement of the architecture is illustrated in Section V. Section VI represents the experimental result of the architectures. Deliberation of this study and limitations are enlightened in Section VII. Section VIII consists of concluding remarks with the future direction.

[1] https://github.com/ultralytics/yolov5

II. Literature Review

Scientists and Researchers have strongly advised us to wear masks for preventing coronavirus. However, the widespread use of these protective gear makes a terrible situation to the environmental system due to human insensibility. Many people consciously or unconsciously put down such dangerous waste in our surroundings which may cause a severe health hazard for any species on this glove. As a result, infectious waste (surgical masks and gloves) increases day by day. In 2020, about 150 million masks will go to the sea. Meanwhile, some countries face problems with this type of waste. Approximately the demand for surgical masks is 28 million per day all over the world [9]. Furthermore, every day 1.6 million tons of waste are generated due to the corona pandemic [10]. It is terrifying that this horrible rubbish is scattered around us. Hence, there is a possibility to spread coronavirus rather than resistance. Our primary purpose is to develop a system to detect infectious waste and infectious symbols as if it can detect and manage such dangerous malicious material from our environment. Object detection is a well-known research area. AquaVision [11] represents an automatic detection system that can detect waste bodies from the water. The authors try to use different transfer learning models to conduct their work. Floating plastic liter detection using Sentinel-2 imagery from space illustrated in [12]. A system that can detect marine life and plastic waste in underwater environment is shown in [13]. Different deep learning methods e.g. Single Shot Detector (SSD), MobileNet are used to detect aquatic animals and waste. In this study, we have dealt with surgical waste detection, which has rarely been done before. A disposal system that can identify waste from the environment and biohazard symbol for keeping the waste in a particular place. From this motivation, a novel dataset is created and named after the MSG dataset to detect the surgical mask, gloves and biohazard symbol. Surgical masks and gloves are detected as waste and biohazard symbols to detect place/container/waste bin to put these types of waste there. Several CV models are trained and tested using the MSG dataset to conduct this work as if more precise detection is generated within an acceptable time. YOLO models (YOLOv3-spp, YOLOv4, YOLOv5 and YOLOX) are selected as the detection architecture.

III. Materials and Methods

A. Dataset Preparation (MSG Dataset)

Realistic criteria are applied to the model at the training and testing time to swear the model's robustness. Among the criteria are taken into account is:

1) Real-time condition.
2) Lighting variations.
3) Multiclass.
4) Underwater condition.
5) Waste floating on the water.

The MSG dataset is built based on real-time images from our surroundings including roads, beaches, water, maintenance holes and so on. Several images of the dataset are synthetic. Moreover, most of them are natural. Some images are taken using the Samsung Galaxy A51 smartphone camera and the rest of the images are taken from internet mining. Images are chosen from close range and distance range to make the dataset a distance variant. The angle variation left, right, back and top angle images are taken. The dataset comprises diverse gesture conditions such as curling and kneeling. At the time of image collection, this study tries to take different types of colored masks and gloves. The color variation of the mask is white, sky blue, pink, black and others. Different types of masks are included surgical, N95, Cone-style, KN95 and so on. Surgical gloves also have blue, white, black and pink colors. Transparent gloves are included with more eagerness to make the system as robust and reliable underwater as well as an object floating on the water condition. According to the above criteria, 1153 images are collected from different internet sources and smartphones camera. Completing the collection of the images, our next step is to annotate the collected images. All the image annotations are handcrafted. The annotations process are done in a graphical image annotation tool called LabelImg [14]. Three types of annotation classes are there:

- Surgical mask as mask.
- Surgical gloves as gloves.
- Biohazard symbol as biohazard.

The dataset is available at https://github.com/Md-Ferdous/Surgical-waste-dataset. The MSG dataset contains 1153 images and 1990 instances where 80% of them (923 images) are selected as the training dataset, 8% (92 images) for validation and the remaining 12% (138 images) for the test dataset. There are three combinations of the MSG dataset keeping the same amount of images into the training, testing and validation set. A validation dataset is provided into the model for an unbiased evaluation and fine-tuning during training. Moreover, the validation process at the training time tells us about the model's training condition such as whether the model is going on the right path? The test dataset is used to evaluate the model's performance. The testing dataset contains ambiguous images for example a paper looks relative to a mask, a plastic/polythene similar to gloves. Creating this ambiguity is to see how well the model performs in real-world conditions. Fig. 1 shows the number of images of every class. There are 568 masked images and 251 images that contain both masks and gloves instances. 10 images contain three classes altogether. The rest is the same. The MSG dataset consists of 1133, 598 and 259 instances of mask, gloves and biohazard symbols, respectively.

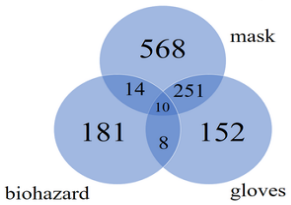


Fig. 1. Number of Images of every Class.

Fig. 2, exhibits the number of instances in the training, testing and validation set of a combination of the dataset. The training set contains 831, 492 and 217 instances of mask, gloves and biohazard symbols, respectively. The total number of masks, gloves and biohazard symbols in the validation set

are 117, 46 and 16. At last, the testing set consists of 184, 60 and 26 instances of the mask, gloves and biohazard symbol, respectively.

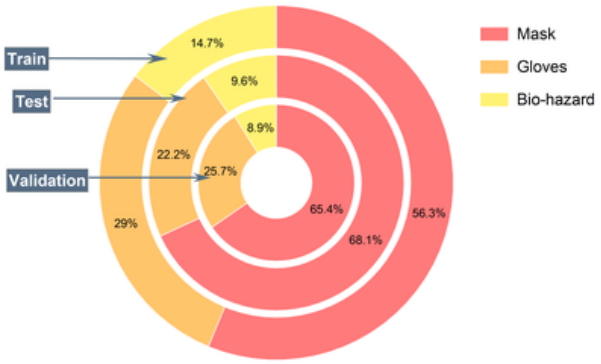


Fig. 2. Number of Instances in the Training, Testing and Validation Set.

B. Framework

A pictorial is shown in Fig. 3, which provides an overview of how the objects are detected from an image. First, an image is fed into the YOLO architecture; differential features are extracted from the network's backbone. Next, the backbone network uses the extracted features and emits a feature pyramid to the head network. After that, the head regresses the bounding boxes and classifies objects. Output from the prediction portion could be any combination of the desired three classes (mask, gloves, biohazard). Moreover, a novel dataset is created to detect and manage infectious waste from our surroundings. Finally, different variations, angles, states and textured images are selected from the real-time condition to accelerate the system robustness.

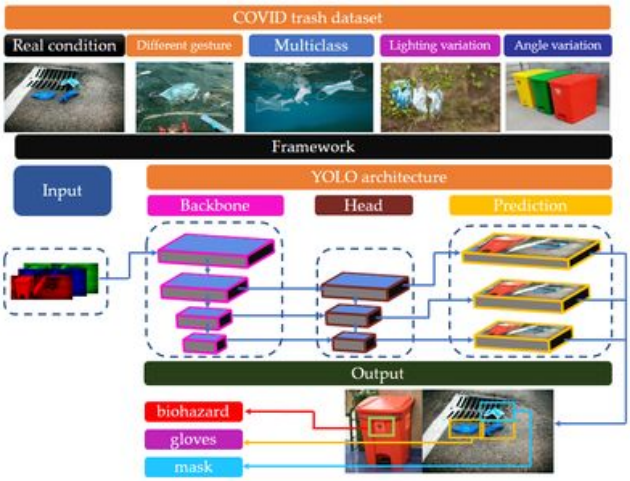


Fig. 3. Framework.

C. Objective of Experiment

This study aims to detect surgical waste and bio-hazard symbols correctly within a reasonable time interval. Therefore, different types of YOLO architectures are examined to achieve different objectives. Additionally, two types of YOLO models are chosen, one is anchor-based and the other is anchor-free training mechanism. Three types of anchor-based and one anchor-free object detection models are shown in Table I.

There are four different versions of YOLOv5 (YOLOv5-s, YOLOv5-m, YOLOv5-l and YOLOv5-x) and YOLOX (YOLOX-s, YOLOX-m, YOLOX-l and YOLOX-x) architecture. Here s, m, l and x denote small, medium, large and extra-large, respectively. According to the theory, a greater model size has better accuracy than a smaller one. On the other hand, a smaller model size is faster than a larger one. It is a dilemma and the solution lies in the perspective or application type you want to build. Therefore, it is essential to evaluate all versions of the YOLOv5 and YOLOX for giving a comprehensive description of their performance. Another notable subject is to see the performance between the anchor-based and anchor-free detectors.

TABLE I. Objective of Experiments

Training mechanism	Model
Anchor-based	YOLOv3-spp YOLOv4 YOLOv5-s YOLOv5-m YOLOv5-l YOLOv5-x
Anchor-free	YOLOX-s YOLOX-m YOLOX-l YOLOX-x

D. YOLOX

1) Decoupled Head: In object detection, two egregious tasks are performed in parallel, one is object classification and the other is bounding box regression. These two tasks share almost the same parameters [15] [16]. Hence, there is an impingement between two tasks and it is a sensitive issue in computer vision. Classification confidence is the probabilities of class levels. At the same time, localization confidence is inexistent. In repetitive regression or even non-maximum suppression, correctly aligned bounding boxes are misaligned. This issue was initially discovered by IoU-Net [17]. They discovered that the characteristic that produces a high classification score for object classification invariably predicts a coarse bounding box. The IoU calculation between ground truth and predicted bounding boxes might be a solution to solve this problem. To forecast the IoU, they first add head to the network as if it can emit the localization score. Then, the final classification score is calculated by combining the localization confidence score and the classification confidence score. This method does help to alleviate the problem of dislocation. Based on this formula, two certain branches, one for classification and the other for localization i.e. double-headed network, was proposed in Double head R-CNN [18] to untangle the head of siblings. In a double-headed network, classification is performed using a fully connected head network and box regression is performed using another convolution head network [15]. Due to having facilities in a double-headed network for object classification and localization many one-stage and two-stage object detection models follow dual-headed architecture [19] [20] [16] [18]. If we divide YOLO families architecture, it has three portions:

backbone, head, and prediction. Backbone (e.g. PAN [20] and FPN [21]) continuously emit feature pyramids to the head. The head classifies objects and localizes the bounding boxes using this feature. Still (YOLOv3-spp, YOLOv4 and YOLOv5) no YOLO family model used double-headed architecture where YOLOX uses double-headed architecture. YOLOX [22] shows that coupled heads may destroy the performance of object detection. The YOLOX architecture is shown in Fig. 4.

2) Exceptional Data Enhancement: Many data augmentation techniques have been proposed in recent years mosaic is one of them. Mosaic is an effective and efficient augmentation technique proposed by a company named ultralytics in their YOLOv3[2]. For boosting the performance of the YOLOX mosaic and MixUp augmentation strategies were applied in [22]. The YOLOv4 [8], YOLOv5 and other object detectors [23] used the mosaic augmentation technique. MixUp [24] is another augmentation strategy that is primarily designed for the image classification task. Modifying it into BoF [23] it is used for the training of objection detection tasks also. In YOLOX, mosaic augmentation is accomplished. A random affine transformation is performed where rotation is done on both axis using a value in the range of -10 deg to +10 deg, translation is done a value between (0.4, 0.6), scaling is attained on both axis within a value of (0.1, 2), the same amount of shear is done on both axis by a value of (-2 to +2). All values are taken from the author's perspective.

3) Anchor-free Mechanism: Although anchor-based training mechanism is well known and famous for object detection model, it has some handicaps. The drawback can be categorized as follows:

1) Selection: Before training an anchor-based model needs to choose an optimal set of anchors for the optimal performance. A clustering analysis needs to be conducted on anchors to choose an optimal set.
2) Complexity: Anchors may create complexity on detection heads for prediction. Besides, it may increase perplexity for the number of predictions for an image.
3) Memory inefficient: For edge computation, in terms of total latency, this might constitute a bottleneck for transferring a massive number of predictions (anchors) across devices (e.g. NPU to CPU) [22]. Recently published YOLO family (YOLOv3-spp, YOLOv4 and YOLOv5) follows the anchor-based training mechanism.

Megvii company [3] released a version named YOLOX, which is based on anchor-free training mechanism. Although YOLOv1 [25] may be the most common anchor-free detector. YOLOv1 predicts bounding boxes at points near the center of objects rather than utilizing anchor boxes. This strategy was done to achieve high performance. Furthermore, low recall problems suffered from YOLOv1. For this reason, YOLOv2 [6] went back to the anchor-based mechanism. In the last two years, Anchor-free detectors have grown at a breakneck pace e.g. FCOS [26], CornerNet [2], object as points [1]. Another type of anchorless detector is the [27] which is adopted on DenseBox [28]. The number of design parameters that require heuristic tuning and the number of tricks such as anchor clustering [6] and grid sensitive [29] requires less amount when using an anchor-free system. Anchor-free detectors enhance the model performance and simplify it especially at the training and decoding phase.

4) Multi Positives: Another aspect of YOLOX is it selects a center point and considers it as positive while ignoring the other predictions, although there is a high probability of being positive [22]. A 3×3 area is chosen around the center point, which is called center sampling in FCOS [26].

IV. Training Process

The entire training and testing process is carried out on the Google cloud platform. The entire training process can be divided into three stages:

1) Prefer a model for training and refashioning its corresponding configuration file as the target.
2) Set initial parameters into the network using pretrained weights.
3) Start the learning process by setting the training parameters. For training, the Stochastic Gradient Descent (SGD) approach was employed.

The learning rate is adopted using lr×batchsize/64 [30]. The initial learning rate is set to 0.01 and it changes over time according to the cosine learning rate schedule. The cosine learning rate can be calculated using Equation 1. Generalized Intersection over Union (GIoU) [31] is calculated for bounding box regression loss is shown in Equation 2. The weight decay is 0.0005 and the SGD momentum is 0.9. NMS threshold is set to 0.65. The first five epochs were warm-up epochs. These warm-up epochs help the network train gradually, making a basic sense of the dataset. Training is done up to 180 epochs. Input image size was 640×640. YOLOX is trained according to Megvii company's GitHub repository in the PyTorch environment. YOLOv5 is also trained in the PyTorch environment and trained according to the construction of a company named Ultralytics.

$$lr = 0.5 \times \left(1.0 + \cos\left(pi \times \frac{\text{iteration}}{\text{total iteration}}\right)\right) \quad (1)$$

$$IoU = \frac{\text{Intersection}}{\text{Union}} = \frac{G \cap D}{G \cup D} \quad (2)$$

$$GIoU = IoU - \frac{|\ C \backslash \text{ Union }\ |}{|C|} \quad (3)$$

Where G and D are the prediction and ground truth bounding boxes respectively. C is the smallest convex object for G and D. All models are trained in the PyTorch environment and an SGD optimizer is used. Table II represents the training hyperparameters.

TABLE II. Training Hyperparameters

Model	Initial learning rate	Decay	Momentum	Batch size
YOLOX	0.01	0.0005	0.9	32/16/12
YOLOv5	0.01	0.0005	0.9	32
YOLOv3-spp	0.001	0.000484	0.937	32
YOLOv4	0.001	0.0005	0.9	2

[2]https://github.com/ultralytics/yolov3
[3]https://github.com/Megvii-BaseDetection/YOLOX

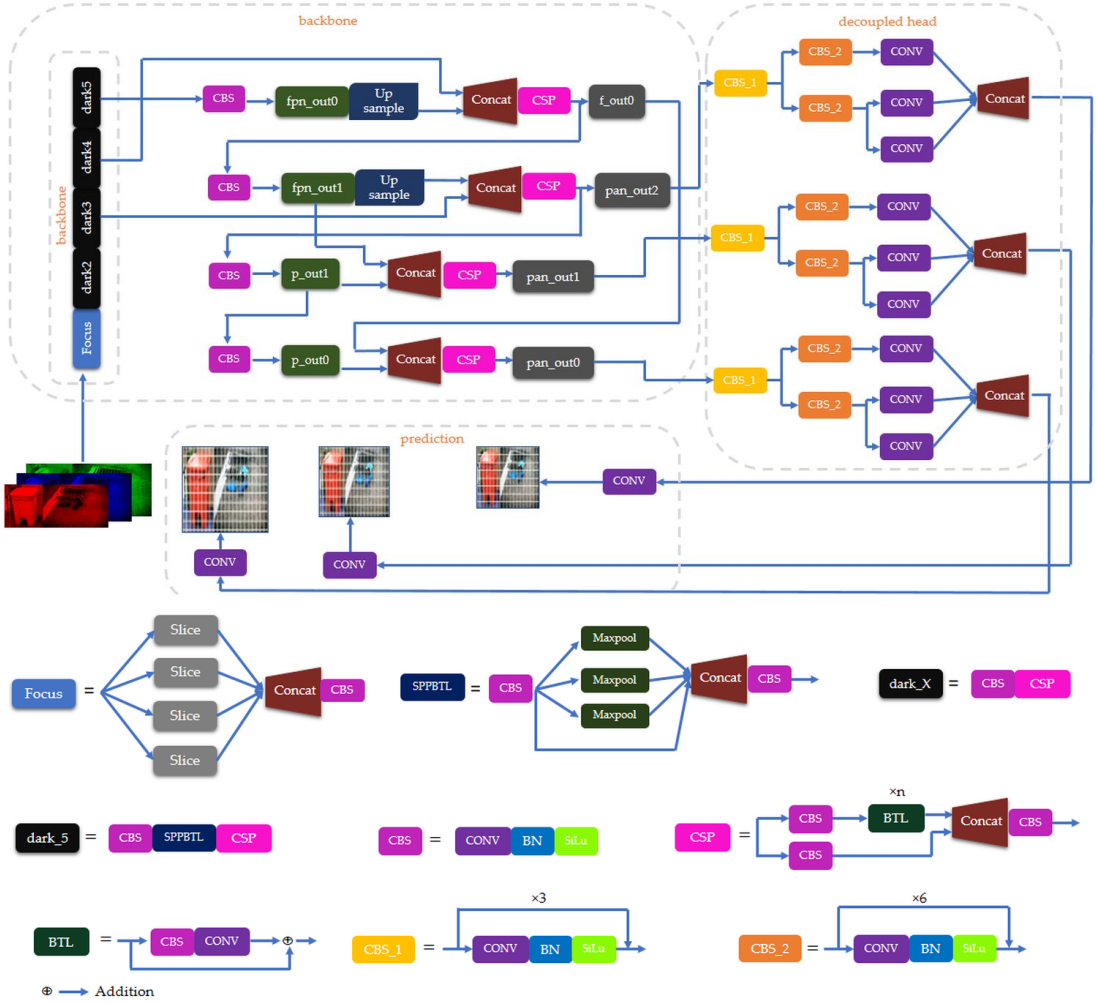


Fig. 4. Architecture of YOLOX.

V. Evaluation Metrics

Intersection over Union (IoU) is shown in Fig. 5. Though the IoU is used in non-maximum suppression, it can also be used in evaluating a model as a performance metric. When evaluating an object detection model, we try to determine whether the predicted class is like the desired class—simultaneously investigating the perfect alignment of the bounding boxes that exist around the object into the image. Using the ground truth bounding boxes (G) and predicted/detected bounding boxes (D), one can calculate the IoU. The IoU calculation tells us how much the predicted bounding boxes are related to the ground truth bounding boxes i.e. the percentage of overlap between two bounding boxes. The bigger the overlap area, the higher the IoU. Equation 2 [32] is used to calculate the IoU between two the bounding boxes. True Positive (TP), False Positive (FP) and False Negative (FN) are calculated according to [33] [34] [35]. TP, FP and FN are the confusion matrix criteria. Table III represents if the model predicts true class and its IoU is greater than 50%, then the detection would be considered as TP. Inversely, FP is considered if the IoU is smaller than 50% and detection tell the right class according to ground truth. FP detects a ground truth class, but its bounding box position is not correct like the ground truth box. FP yields an improper detection case. In the case of FN, the system will not be able to detect any class where ground truth boxes exist.

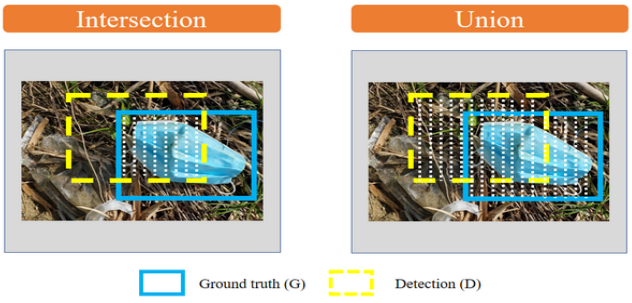


Fig. 5. IoU Calculation Mechanism

If the TP is omitted from the total positives, we will get the FN. Therefore, the object confidence score is set to greater than 50% for calculating the TP, FP and FN. In object detection, every position of an image will be True Negative (TN) without

desired class. Hence, TN is not so important to measure the performance of an object detection model.

TABLE III. CALCULATION PROCESS OF THE TP, FP AND FN

IoU (%)	Class	TP	FP	FN
>=50	true	✓	✗	✗
>=50	false	✗	✓	✗
<50	true/false	✗	✗	✓

Precision is a metric that depicts the capability of a model to identify its relative objects correctly. It is the proportion of positive predictions that are correct across all detections. The recall is a metric that delineates the power of a model to find all relevant examples. It is the proportion of positive predictions subject to the entire ground truth. Precision and recall are calculated as [33] [36]. This study evaluates the model's performance using the Average Precision (AP). AP is a scheme to encapsulate the Precision-Recall (PR) curve. Higher precision is clear evidence that a model is confident when classifying examples among the detections. On the other hand, higher recall is an indicator of the power of a model. It tells us how many correct detections are performed among all the ground truths. Moreover, precision and recall are primary metrics of an object detection model. If a model has high recall yet low precision is an obvious referential that the model emits maximum positive example truly, but it has many false positives i.e. classify many negative examples as positive. On the contrary, low recall yet high precision indicates that the model appropriately classifies positive examples; however, it may contain only a few positive examples. Hence, it is necessary to choose a threshold, as if both the precision and recall will be maximized. The PR curve helps us to select the appropriate threshold among the different threshold values. Using the precision and recall value, the PR curve can be plotted [33] [36]. AP is the Area Under the Curve (AUC) of the PR curve. AP can be calculated using the Equation 4. According to Equation 4, n is the number of thresholds. For every value of recall or precision, find the difference between the current and next to recall value, then multiply the difference with the Interpolated Precision (IP) value. IP are the maximum precision value at a recall value R where the corresponding recall value is greater than or equal to R. At each threshold, AP is the weighted sum of all precision where the recall value accelerates the weight.

$$AP = \sum_{k=0}^{k=n-1} [R(k) - R(k+1)] \times IP(k) \tag{4}$$

Different IoU marginal values are used to test the model in object detection. Therefore, different IoU values yield different performances. Furthermore, Mean Average Precision (mAP) is another metric that is calculated using the AP's of every class shown in Equation 5 where n is the number of classes.

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{5}$$

Evaluation metrics are calculated according to [33] [35] [36]. This study tests the model in both cloud-based and local environments. The local environment with CPU is not capable enough for training. We test the model in a local environment with a CPU to see its performance on user label equipment. The training phase is conducted on a cloud-based platform. A 15GB sized Tesla T4 GPU is used for high speed and performance for training and testing. Parameter description of two platforms is given in Table IV. At the time of speed calculation of the model, only the processing time is considered except the loading time of an image. The bottom-most speed is the mean of all testing images.

TABLE IV. PLATFORM PARAMETERS

Purpose	Platform	GPU	GPU size	CPU core	RAM
Training and testing	Cloud	Tesla T4	15GB	2	12GB
Testing	Local	None	None	4	32GB

VI. RESULTS

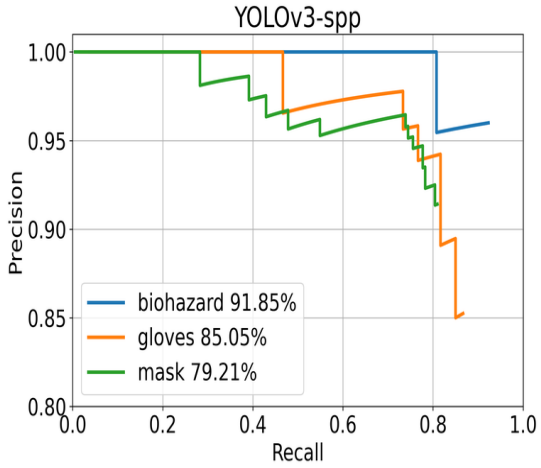
In the PR curve, precision and recall are plotted in Y-axis and X-axis respectively. Moreover, detection ability would be better when the precision is higher with the increase of the recall value. Therefore, according to this theory, better performance would be in the right up corner of the PR curve. To check the robustness of the model, training and testing are done using three different dataset combinations.

A. Anchor-based Method

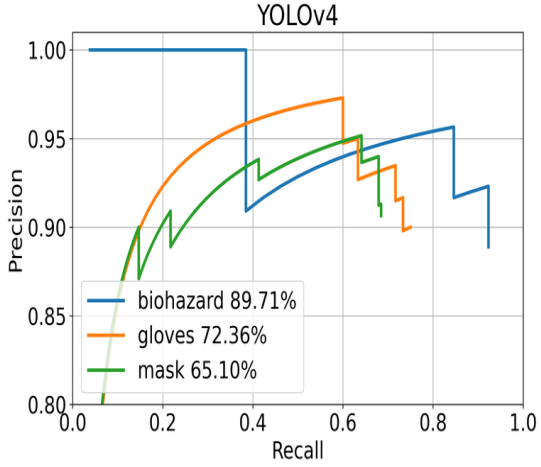
The average precision of the three classes for the anchor-based method is shown in Table V. YOLOv3-spp emits the highest mAP of 85.37% among the other methods. YOLOv5-l generates a better mAP of 84.78% among different versions of YOLOv5. For a dataset combination, for mask objects, 149 are truly detected, 14 objects are false detection and 35 are undetected. For the gloves object, 52 gloves are detected accurately, 9 are false and 8 are undetected. The number of true positive, false positive, and false negative are 24, 1 and 2 for the biohazard symbol. Fig. 6 represents the PR curves of several anchor-based methods for a dataset combination. Well performed models PR curve is shown here. Fig. 7 exhibits some qualitative measures of the anchor-based methods. YOLOv3-spp produces better generalization than the other tested models. YOLOv4 depicts some under-detected and false detection results. YOLOv5 yields better generalization even after having many false detection outputs than YOLOv4.

TABLE V. PERFORMANCE OF ANCHOR BASED MODEL

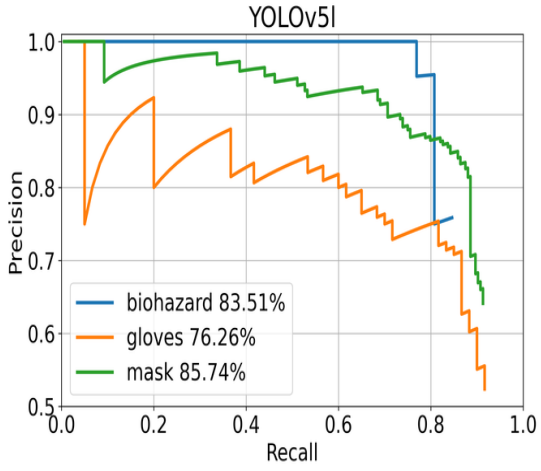
Model	mask AP	gloves AP	biohazard AP	mAP
YOLOv3-spp	79.21	**85.05**	**91.85**	**85.37**
YOLOv4	64.69	73.51	89.11	75.77
YOLOv5-s	84.21	79.36	82.19	81.92
YOLOv5-m	83.29	80.42	84.45	82.72
YOLOv5-l	**86.84**	77.69	89.81	84.78
YOLOv5-x	85.36	76.12	88.75	83.41



(a)



(b)



(c)

Fig. 6. Precision-recall Curve of the Anchor-based Model.

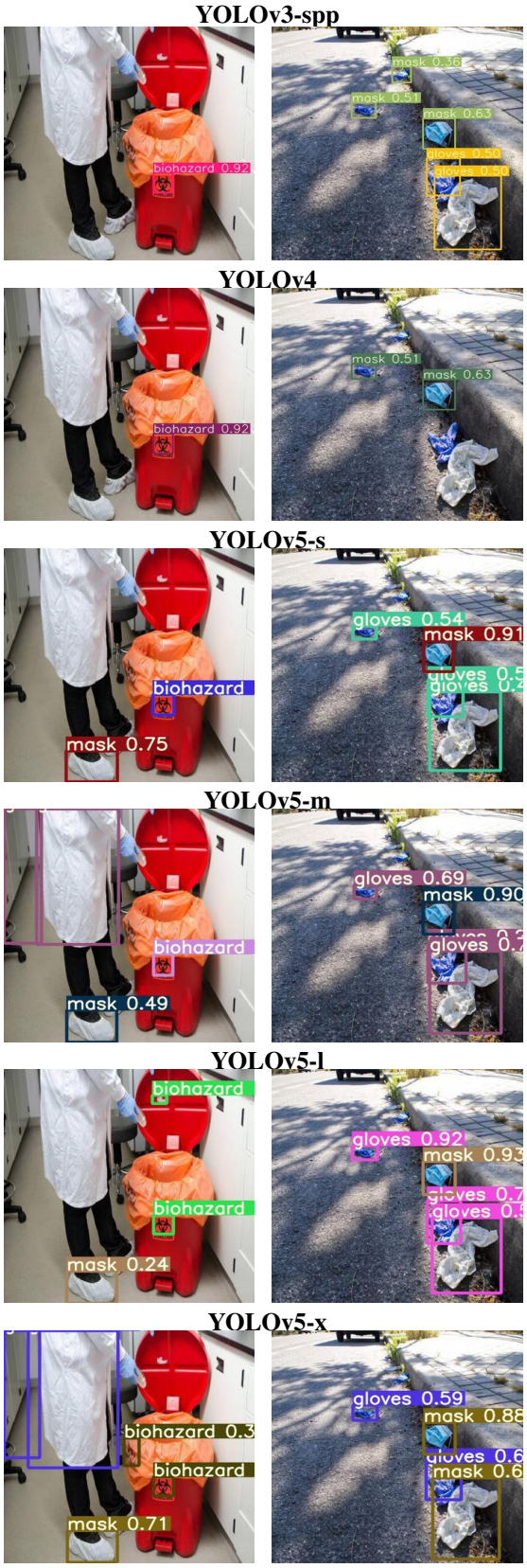


Fig. 7. Qualitative Measures of Anchor-based Methods.

B. Anchor-free Method

The YOLOX is trained and tested using three different combinations of the dataset shown in Table VI. Averaging of the mAP of each combination of datasets, the YOLO-l architecture delivers the highest mAP than the other models. For example, the ground truth of the mask, gloves and bio-hazard symbol are 184, 60, and 26, respectively, in a dataset combination. Accurate detection of masks, gloves and symbols is 169, 55 and 26; false detection is 8, 6 and 0; under detected objects are 15, 5 and 0, respectively for YOLOX-l. Due to easy examples, the model generates 100.00% AP for bio-hazard objects for a combination of the dataset.

TABLE VI. PERFORMANCE OF YOLOX IN THE DIFFERENT DATASET COMBINATIONS

Combi-nation	mask AP	gloves AP	biohazard AP	mAP	Avg. of the mAP
YOLOX-s					
1	85.99	81.60	**100.00**	89.20	90.1
2	93.00	92.25	98.36	94.54	
3	88.57	82.35	88.00	86.30	
YOLOX-m					
1	89.30	86.55	**100.00**	91.95	91.38
2	**96.74**	94.60	96.72	95.36	
3	87.91	83.69	90.99	87.53	
YOLOX-l					
1	89.47	87.05	**100.00**	92.17	**92.49**
2	95.64	**96.23**	96.72	**96.20**	
3	92.15	81.01	94.20	89.12	
YOLOX-x					
1	86.44	79.75	**100.00**	88.73	90.86
2	95.64	93.97	95.03	94.88	
3	90.56	85.47	90.88	88.97	

Fig. 8 represents the PR curve of tested anchor-free architecture of different versions for a dataset combination. Per image inference time on both GPU and CPU versus mAP is shown in Fig. 9, where it is clearly seen that CPU inference time is greater than the GPU time. Models mAP differs a little amount according to the model size. In GPU, the YOLOX-s needs 0.06s time to infer an image, whereas YOLOX-x takes 0.12s time, two times more than YOLOX-s. On the other hand, in CPU the YOLOX-s needs 0.70s time to infer an image, whereas YOLOX-x takes 4.01s time, which is approximately five times more than YOLOX-s. Hence, it is said that increasing model size CPU requires more time to infer an image than GPU. Additionally, GPU time maybe differ on the criteria of the GPU architecture. Fig. 10 displays several qualitative measures of anchor-free architecture—the different versions of the YOLOX architecture yield approximately the same results. The input combinations of the dataset are responsible for the performance. In this case, remarkable improvement may happen for the different training and testing dataset combinations, although their averaging produces the conventional results. Since different dataset combinations have been used, YOLOX-l architecture provides the highest individual mAP than others. Additional testing is conducted using the test dataset of the second combination because, among the three combinations, these combination generates the highest mAP. Furthermore, the test dataset of the second combination is divided into two portions, one is relatively easy to guess and the other is relatively hard to guess for the model. A relatively complex sub-division is created using several criteria that are listed below. These criteria are done from the author's perspective.

1) Objects and the image background color are approximately the same.
2) Excess or inadequate light into the images.
3) Complex image background compares to objects.
4) Occlusion and small objects into the images.

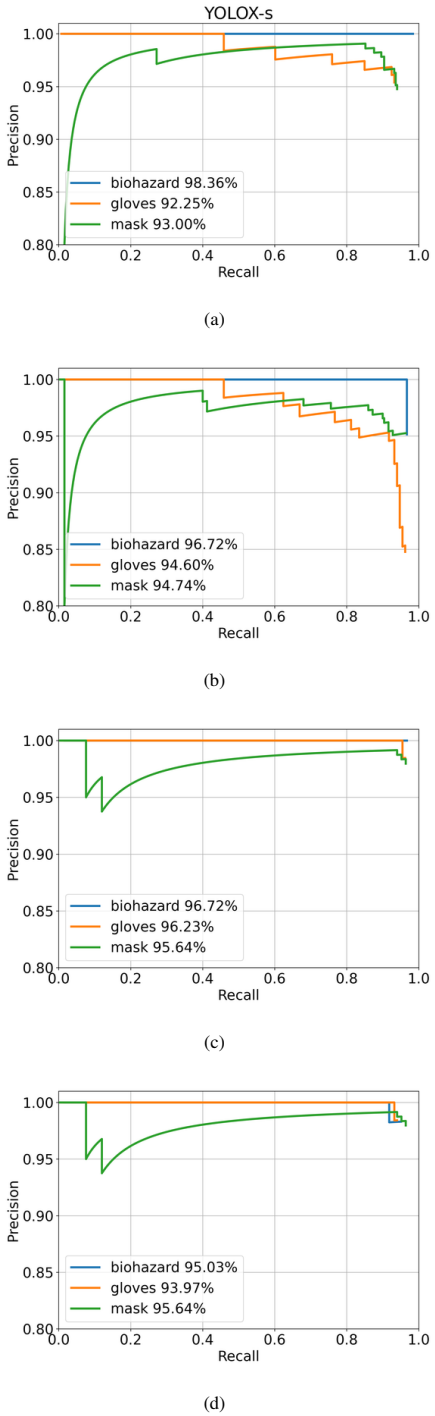


Fig. 8. Precision-recall Curve of the Anchor-free Model.

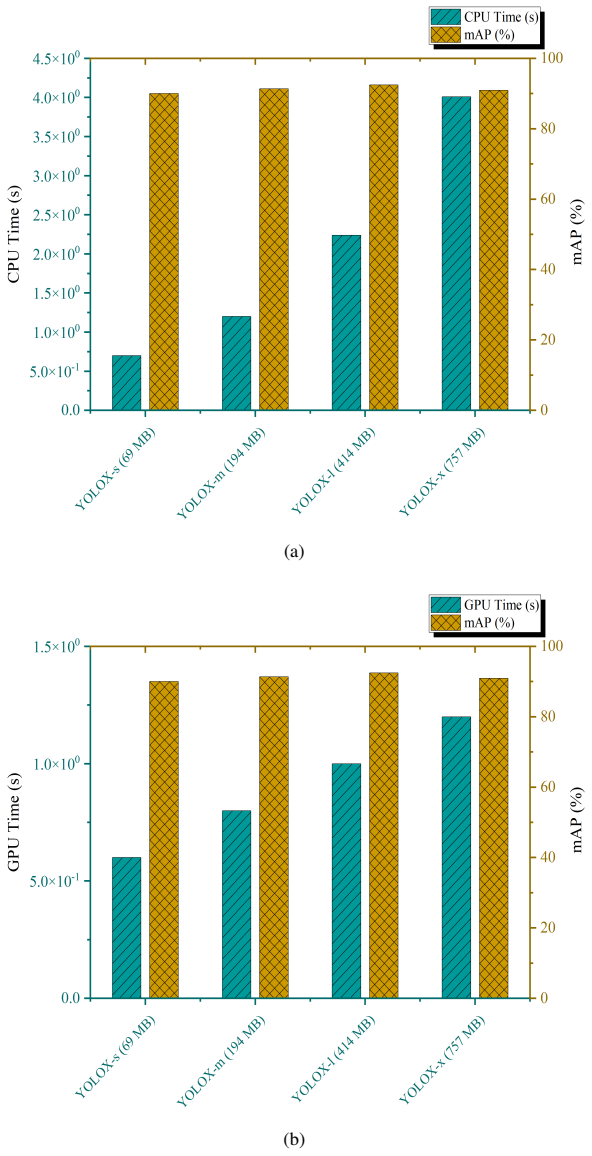


Fig. 9. Inference Time of GPU vs mAP in (a) and CPU vs mAP in (b) of all Version of the YOLOX.

Table VII displays the performance of YOLOX-l models for relatively easy and hard samples of the second combination of test dataset where, for easy and challenging examples, 95.71% and 94.04% mAP are encountered, respectively. The average precision of both the anchor-free and anchor-based architecture is shown in Fig. 11, where anchor-free models mAP is larger than anchor-based models.

TABLE VII. PERFORMANCE OF YOLOX-L FOR RELATIVELY EASY AND HARD CASES

case	mask AP	gloves AP	biohazard AP	mAP
Relatively easy	97.91	94.79	100.00	97.57
Relatively hard	93.78	92.87	97.62	94.76

1) Where is the Milestone?: Fig. 12 shows several satisfactory results. According to the Fig. 12, images are natural and most wanted phenomena for waste. Fig. 12a, 12b, 12c, 12d and 12f objects are correctly detected even after complex background; despite being small objects in Fig. 12e, 12g, and 12i objects are well-identified; into the Fig. 12h are some underwater images where obscurity, different lighting condition may happen, still objects are recognized properly.

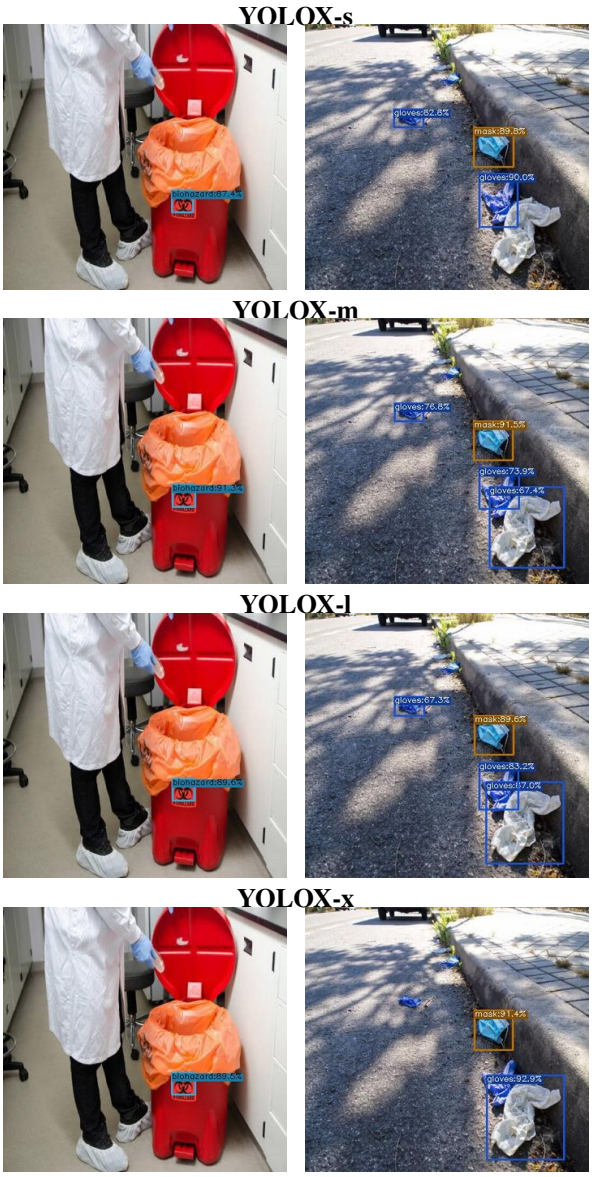


Fig. 10. Several Qualitative Measures of Anchor-free Methods.

2) Where is Incorrect Detection?: According to the dataset combination, the FP rate is 4.34%, 10%, and 0.0% for masks, gloves, and biohazards, respectively. FN occurs when the model is unable to detect an object despite its presence in the image. FN rate of mask, gloves, and biohazard class is 8.15%, 8.33%, and 0.0%, respectively. Fig. 13 shows some false detection of the YOLOX-l model. FP detection occurs when objects have crowded situations into the images. When the objects are blurred, they are under-detected by the model in

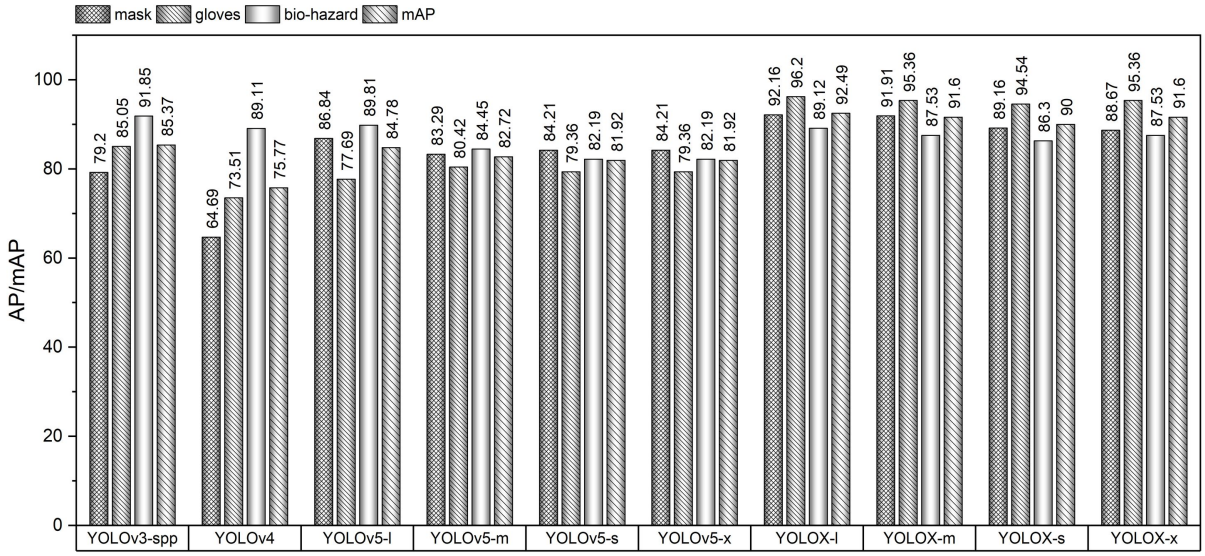
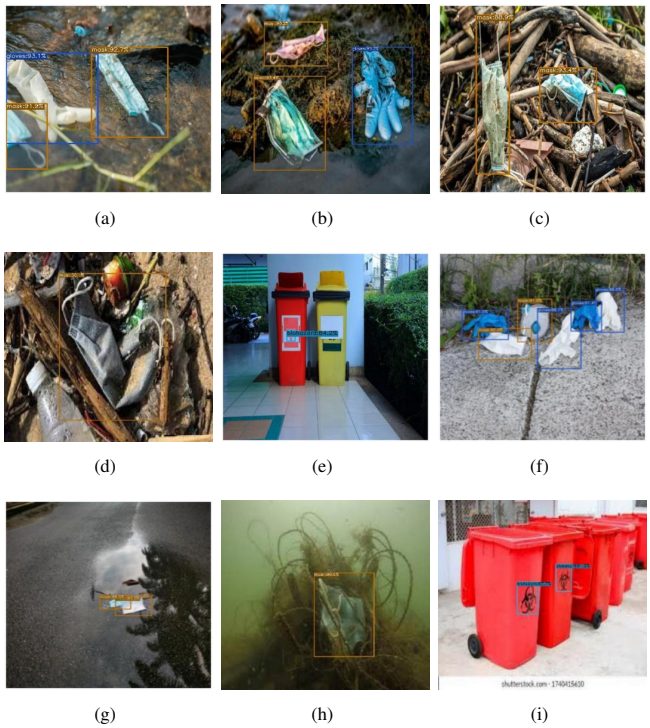


Fig. 11. Average Precision of Each Model.

most cases. Fig. 13a additional gloves and masks are detected. Crowed masks and gloves are there. In Fig. 13b contains small and occluded gloves and mask objects perhaps for this reason they are under detected, in Fig. 13c perhaps objects may not be detected because there are exists some blurriness on the gloves and mask.

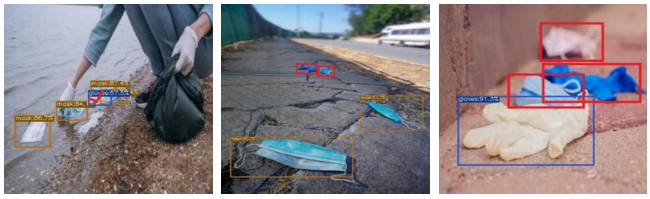


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 12. Satisfactory Result of YOLOX-l Architecture.



(a) (b) (c)

Fig. 13. Incorrect Detection of YOLOX-l Architecture.

VII. Discussion

This study examined four different types of object detectors from the YOLO family. The models are trained, tested, and evaluated to attain different goals. The gist contribution and findings are listed below:

1) A novel dataset is created to detect and manage surgical waste (mask, gloves and biohazard symbol) from our surroundings including roads, beaches, underwater as well as floating on the water condition.
2) Experiment is conducted using two different training mechanisms, one is anchor-based and the other is an anchor-free mechanism. Anchor-based models (YOLOv3-spp, YOLOv4 and YOLOv5) highest mAP 85.27% is lower than the anchor-free models (YOLOX-l) mAP 92.49%. Among the anchor-based model, YOLOv3-spp (mAP 85.27%) performs better than others.
3) Additional exploration is performed using all versions of the YOLOX to find TP, FP and FN. In most cases, satisfactory results are yielded by the model. False detection produces when objects are occluded, blurred and crowded. False-negative (misdetection) generates when objects are small and complex to attain minimum feature selection.

Additionally, several limitations exist in this study that can be explored in future work. First, the dataset size may be increased for system robustness. This study detects only three types of classes; hence integrating more classes would be a good exploration. The YOLO architectures are fast and accurate in detecting objects from images. Additionally, YOLO models have the sensitiveness of small objects. Therefore, other anchor-free object detection models may be investigated to evaluate the dataset.

VIII. Conclusion

Plastic waste is going to be a threat to humankind and other species day by day. This type of waste is scattered worldwide only for the unconsciousness of humankind. If the proper steps are not taken, it will be dangerous for us in the days ahead. Surgical waste is a kind of plastic waste. Nowadays, the number of this type of waste is also increasing alarmingly, which is responsible for serious health hazards. In this study, authors try to create a novel dataset with surgical waste (mask and gloves) and biohazard symbols to detect such waste from our surroundings and appropriately manage them in certain places. Two types of training mechanism-based YOLO models are chosen to conduct this work. One is anchor-based (YOLOv3-spp, YOLOv4 and YOLOv5) and the other is anchor-free (YOLOX) architecture. This study found that anchor-free architecture performs better generalization than anchor-based architectures. More clearly, YOLOX yields the highest mAP of 92.49% and YOLOv3-spp generates the highest mAP of 85.27%. Satisfactory performance comes up even if there are some limitations. Dataset explorations would be reasonable for future work. Applying the other anchor-free architecture in this area will be a future direction.

References

[1] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[2] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[4] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15849–15858, 2021.

[5] Qiang Zhou, Chaohui Yu, Chunhua Shen, Zhibin Wang, and Hao Li. Object detection made simpler by eliminating heuristic nms. *arXiv preprint arXiv:2101.11782*, 2021.

[6] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[9] Organisation for Economic Co-operation and Development. *The Face Mask Global Value Chain in the COVID-19 Outbreak: Evidence and Policy Lessons*. OECD Publishing, 2020.

[10] Nsikak U Benson, David E Bassey, and Thavamani Palanisami. Covid pollution: impact of covid-19 pandemic on global plastic waste footprint. *Heliyon*, 7(2):e06343, 2021.

[11] Harsh Panwar, PK Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, Sudhansh Sharma, and Iqbal H Sarker. Aquavision: Automating the detection of waste in water bodies using deep transfer learning. *Case Studies in Chemical and Environmental Engineering*, 2:100026, 2020.

[12] Kyriacos Themistocleous, Christiana Papoutsa, Silas Michaelides, and Diofantos Hadjimitsis. Investigating detection of floating plastic litter from space using sentinel-2 imagery. *Remote Sensing*, 12(16):2648, 2020.

[13] Rahul Hegde, Sanobar Patel, Rosha G Naik, Sagar N Nayak, KS Shivaprakasha, and Rekha Bhandarkar. Underwater marine life and plastic waste detection using deep learning and raspberry pi. In *Advances in VLSI, Signal Processing, Power Electronics, IoT, Communication and Embedded Systems*, pages 263–272. Springer, 2021.

[14] TzuTa Lin. Labelimg, 2015.

[15] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020.

[16] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018.

[18] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[21] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018.

[22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[23] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13039–13048, 2021.

[24] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[26] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[27] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019.

[28] L Huang, Y Yang, Y Deng, and Y Densebox Yu. Unifying landmark localization with end to end object detection. arxiv 2015. *arXiv preprint arXiv:1509.04874*.

[29] Xin Huang, Xinxin Wang, Wenyu Lv, Xiaying Bai, Xiang Long, Kaipeng Deng, Qingqing Dang, Shumin Han, Qiwen Liu, Xiaoguang Hu, et al. Pp-yolov2: A practical object detector. *arXiv preprint arXiv:2104.10419*, 2021.

[30] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[31] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[32] Rafael Padilla, Sergio L Netto, and Eduardo AB da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242. IEEE, 2020.

[33] Rafael Padilla, Wesley L Passos, Thadeu LB Dias, Sergio L Netto, and Eduardo AB da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3):279, 2021.

[34] Md Ferdous and Sk Md Masudul Ahsan. Multi-scale safety hardhat wearing detection using deep learning: A top-down and bottom-up module. In *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6. IEEE, 2021.

[35] Jimin Yu and Wei Zhang. Face mask wearing detection algorithm based on improved yolo-v4. *Sensors*, 21(9):3263, 2021.

[36] Haidi Zhu, Haoran Wei, Baoqing Li, Xiaobing Yuan, and Nasser Kehtarnavaz. A review of video object detection: Datasets, metrics and methods. *Applied Sciences*, 10(21):7834, 2020.