Mathematics and Computer Science
2023; 8(5): 104-111
http://www.sciencepublishinggroup.com/j/mcs
doi: 10.11648/j.mcs.20230805.11
ISSN: 2575-6036 (Print); ISSN: 2575-6028 (Online)

Real-Time Object Identification Through Convolution Neural Network Based on YOLO Algorithm

Muhammad Saleem*, Naveed Sheikh, Abdul Rehman, Muhammad Rafiq, Shah Jahan

Department of Mathematics, University of Balochistan, Quetta, Pakistan

Email address:

Saleem.eptsc@gmail.com (Muhammad Saleem), naveed_maths@hotmail.com (Naveed Sheikh), abdul_maths@yahoo.com (Abdul Rehman), profrafiqkhan1987@gmail.com (Muhammad Rafiq), shahjahansharif001@gmail.com (Shah Jahan)

*Corresponding author

To cite this article:

Muhammad Saleem, Naveed Sheikh, Abdul Rehman, Muhammad Rafiq, Shah Jahan. (2023). Real-Time Object Identification Through Convolution Neural Network Based on YOLO Algorithm. *Mathematics and Computer Science*, *8*(5), 104-111. https://doi.org/10.11648/j.mcs.20230805.11

Received: November 13, 2023; **Accepted**: December 1, 2023; **Published**: December 28, 2023

Abstract: A widely utilized object detection technique in computer vision involves Convolutional Neural Networks (CNN) due to their simplicity and efficiency. The effectiveness of CNN-based object detection relies significantly on the choice of loss function, with localization precision being a critical determinant. In order to improve localization accuracy, we have made changes inside CIoU loss function resulting in the development of a new loss function known as Area-CIoU (ACIoU). This new loss function specifically adopts a comprehensive approach by taking into account the alignment of bounding boxes between predictions and ground truth, combining the relationship between aspect ratio and area for both bounding boxes. When both bounding boxes have the same aspect ratio, we take into account how the prediction box may affect localization accuracy. As a result, the penalty function is strengthened, which improves the network model's localization precision. Experimental results on a custom dataset of vehicles including car, person, motorcycle, truck and bus, affirm the efficacy of ACIoU in enhancing the localization accuracy of network models, as demonstrated through its application in the one-stage object detector YOLOv4. Experiments also show that the network's accuracy was enhanced but its FPS dropped due to the new penalty term composition in the loss function. We achieved AP of 88.48% and average recall rate of 86.37% with 41 frames per second.

Keywords: Object Detection, Loss Function, Real-Time, YOLOv4

1. Introduction

Humans recognise the objects in an image immediately as they see an image but the machines take a lot of time to learn to recognise the objects in an image. This is a difficult task for machines. Many researchers are working to solve this problem but they are only able to achieve 65% accuracy. It is a difficult task for machines to categorize and recognise objects like humans can Kumar and Srivastava [1]. Humans look at a picture and immediately understand what objects are in the picture, what is their location and how they are connected to each other Redmon et al. [2].

Computer vision is an area of study for enabling computers to recognize things. Nowadays computer vision is a hot topic for research with the use of artificial intelligence systems to extract information from images, such as intelligent traffic surveillance systems Tao et al. [3]. Humans can quickly recognize and distinguish objects, without knowing their circumstances, no matter what place they are, what is their orientation, difference in colour or shape, partially visible etc. Therefore, for humans object detection is so easy. The same task of object identification with a machine requires lot of processing in order to obtain any information on the shapes and objects in an image. Object detection in computer vision refers to recognizing and identifying an object in a picture or video Ahmad et al. [4]. Object detection is related to numerous applications, including image classification, human behavior analysis, facial recognition, and self-driving automobiles. It can give important information for semantic interpretation of pictures and videos Zhao et al. [5]. The invention of deep learning has greatly boosted computer vision. Deep learning tech learn and imitate the cognitive abilities of the

human brain to analyse and evaluate features, which has a significant impact on object detection. Unlike conventional methods of extraction of features, deep convolutional neural networks use multilayer convolution operations to extract the features thus can attain a high degree of accuracy. Among deep learning approaches, the most widely used object detection algorithms are RCNN, Faster RCNN, YOLO, and SSD Lu et al. [6]. The deep learning-based object detection method primarily involves two types: the two-stage region proposal-based method and the one-stage regression-based method. R-CNN, Fast R-CNN, Faster R-CNN, R-FCN method, light head R-CNN method, and other improved methods based on convolution neural network are among the conventional two-stage approaches. While the two-stage method is more accurate than the one-stage method, the one-stage method detects more quickly. In certain conditions that require higher real-time detection, the one-stage approach is more suitable for use Algabri et al. [7].

The first deep learning object detection model using Convolutional Neural Networks (CNNs) with a sliding window approach was introduced by Sermanet and Eigen [8]. R-CNN, which is an effective and flexible detection algorithm that significantly improved Mean Average Precision (mAP) Girshick et al. [9]. R-CNN was further enhanced as Fast R-CNN, improving both training and testing speed while maintaining detection accuracy Girshick [10]. Faster R-CNN, which introduced the Region Proposal Network (RPN) for more efficient region proposal generation Ren et al. [11]. YOLO(You Only Look Once) a new and improved method, which is a single neural network and here object detection is defined as a regression problem [2]. Single Shot MultiBox Detector(SSD) for object detection, which used a single deep neural network. This model is easy because it fully removes the processing of proposals and feature resampling phases, combining all computing in a single network Liu et al. [12]. YOLOv2 which was considered best for detection tasks of standard data sets like PASCAL VOC and COCO at that time was introduced by Redmon and Farhadi [13]. The updated version of YOLOv2, which was made after a bunch of little design changes to make it better and presented the new version as YOLOv3 Redmon and Farhadi [14]. Many new features were introduced in YOLOv3 to improve the performance of YOLO, the new version is called YOLOv4 Bochkovskiy et al. [15].

In CNN-based object identification, loss functions are essential because they are used to measure the difference between models predicted values and actual values, which has an effect on detection accuracy. These procedures are crucial in the localization of objects. For tasks like object identification, instance segmentation, and object tracking, precise bounding box localization is essential Wang and Song [16]. Regression loss functions play a critical role in various machine learning tasks, particularly in object detection and localization tasks. Intersection over union (IoU) loss is the most preferred evaluation measurement technique for bounding box regression Yu et al. [17], Rezatofighi [18], but it struggles when the prediction's and ground truth's bounding boxes don't coincide. Generalized IoU (GIoU), distance-IoU (DIoU), and complete IoU (CIoU) were introduced to address this issue. GIoU addresses bounding box regression issues when there's no overlap between boxes Qian et al. [19], while DIoU expedites convergence by reducing separation between box centers Zheng et al. [20]. CIoU algorithm considers three geometric metrics simultaneously including center distance Wang and Song [16].

We propose a real-time object detection and identification system which will identify the objects in real-time in a video. We will use the advanced architecture of YOLOv4 with some changes in its loss function.

2. Materials and Methods

2.1. Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of artificial neural network designed specifically for processing structured grid data, such as images and videos. CNNs are particularly well-suited for tasks like image recognition, object detection, and image classification due to their ability to automatically learn and extract features from input data.

2.2. YOLOv4

The YOLOv4 neural network model is currently among the most widely used ones. When compared to YOLOv3, YOLOv4 significantly increases model detection accuracy while maintaining speed. YOLOv4 is mostly made up of three parts: the head, neck, and backbone. Yolov3's Darknet53 is replaced with CSPDarknet53 in the Backbone of YOLOv4. SPP and PANet were used in YOLOv4's Neck instead of the feature pyramid networks (FPN) of YOLOv3. The Head of YOLOv3 may be found inside the Head of YOLOv4. Strict adherence to the Darknet training protocol was used; Darknet416 served as the backbone network, and a maximum of 6K iterations were allowed.

2.3. Loss Function

There are three parts of loss function in YOLOv4, which are Bounding box regression loss, confidence loss and classification loss as shown in equation (1).

$$Loss = \mathcal{L}_{ciou} + \mathcal{L}_{confidence} + \mathcal{L}_{class} \tag{1}$$

In our proposed approach we consider the bounding box regression loss which is $\mathcal{L}_{ciou}$ as illustrated by equation (2).

$$\mathcal{L}_{ciou} = 1 - IoU + \frac{\rho^2(B^p, B^{gt})}{c^2} + \alpha v \tag{2}$$

Complete-IoU(CIoU) was proposed to overcome the shortcomings in previous loss functions. CIoU is used in YOLOv4 as the loss function and it gave very good results. The overlap area, separation between center points of the boxes, and aspect ratio are the three geometric measurements that CIoU takes into account. Thus it converges faster than GIoU and DIoU. It improves average precision (AP) and average recall (AR) for object detection and segmentation. It is defined in equation (3).

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(B^p, B^{gt})}{c^2} + \alpha v \quad (3)$$

Where

c = Diagonal length of smallest enclosing box that covers both the boxes.

B^p and B^{gt} = center points of predicted and ground truth bounding box

v = shows the constancy of the aspect ratio as shown in equation (5)

α = positive trade-off parameter as shown in equation (4)

$\rho(\cdot)$ = Euclidean distance

$$\alpha = \frac{v}{(1-IoU)+v} \quad (4)$$

In equation (4), α is a hyperparameter that balances the importance of the different components in the loss function.

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w^p}{h^p})^2 \quad (5)$$

In equation (5), w^{gt}, h^{gt}, w^p and h^p are the width and height of bounding boxes of ground truth and prediction respectively. The performance of $\mathcal{L}_{CIoU}$ is better than $\mathcal{L}_{DIoU}$ as it considers aspect ratio. In equation (5), when $\frac{w^{gt}}{h^{gt}} \neq \frac{w^p}{h^p}$, then $v > 0, \alpha v > 0$, penalty term αv has a positive role in the calculation of loss. If $\frac{w^{gt}}{h^{gt}} = \frac{w^p}{h^p}$ in equation (5), then $v = 0\ and\ \alpha v = 0$, which causes $\mathcal{L}_{CIoU}$ to degenerate into $\mathcal{L}_{DIoU}$ and slows down convergence.

2.4. Relationship Between Loss Functions IoU, GIoU, DIoU, CIoU

Three important geometrical elements are often taken into account when computing regression loss. The overlap area, the separation between the center points, and the aspect ratios of the bounding boxes. IOU loss within these various loss functions concentrates on the overlap area, whereas GIOU loss tackles the issue when bounding boxes do not overlap. The center point distance is taken into consideration by DIOU loss, and CIOU loss also considers the aspect ratio. Theoretical research reveals that, in comparison to models utilizing various bounding box regression-based loss functions, a model adopting CIOU loss demonstrates faster convergence and greater detection accuracy. However, when there is a difference in the sizes of GT boxes, there will be certain unusual situations, as illustrated in Figure 1, where each pair of bounding boxes satisfies the requirements listed below:

$$\frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \approx \frac{|B' \cap B^{gt}|}{|B' \cup B^{gt}|} \quad (6)$$

Equation (6) shows that the pair of bounding boxes as shown in Figure 1 approximately have the same overlap area, thus their IoU loss is same.

$$\frac{\rho^2(B, B^{gt})}{c^2} \approx \frac{\rho^2(B', B^{gt})}{c'^2} \quad (7)$$

Equation (7) illustrates that the Euclidean distance between the centers of the bounding boxes shown in Figure 1 is also approximately same.

$$arctan\frac{w}{h} \approx arctan\frac{w'}{h'} \quad (8)$$

Equation (8) shows that the aspect ratios of the bounding boxes is also same in the pair of boxes given in Figure 1.

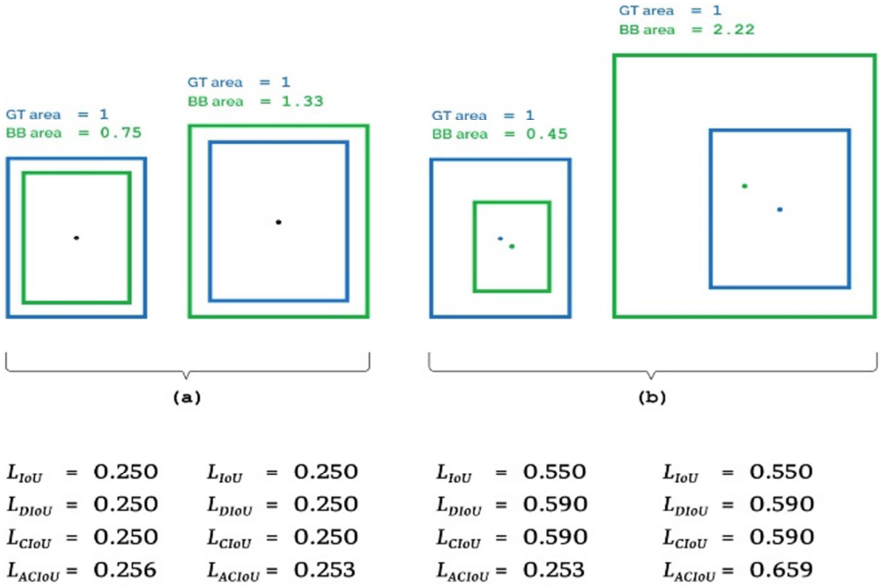


Figure 1. *Loss calculated by previous methods and by our custom loss function. Ground truth(GT) boxes are shown in blue while predicted bounding(BB) are shown in green color.*

Let's consider the boxes in Figure 1(a). The difference in area between the target box and left bounding box is 0.25 times the area of the ground truth box (A^{gt}), and the intersection over union (IoU) is 0.75. Similar to this, the difference in area of right boxes is 0.33 times A^{gt}. Even if the scale differences are not considerable, it is clear that the right

bounding box contains more details about the target. Thus right bounding box is preferable than the left one. Let's consider Figure 1(b), The difference in area between the target box and left bounding box is 0.55 times the area of the ground truth box (A^{gt}), the difference in area between the target box and right bounding box is a sizeable 1.2 times A^{gt}, and the IoU is 0.45. The area of the right bounding box in this case is significantly larger, raising questions about whether it contains simply the target or also other possibly irrelevant and disruptive elements. Therefore, even if the left bounding box has a smaller area than the ground truth bounding box and does not completely include the target information, it is both logical and natural to draw that conclusion. Interestingly, current loss functions would provide the same regression loss values in these circumstances despite the huge differences, making it potentially impossible to distinguish between them. Thus, a key factor affecting the computation of regression losses is the area difference between the boxes.

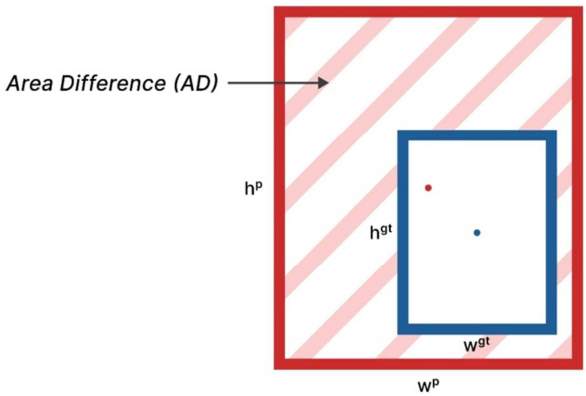


Figure 2. *In the above figure the aspect ratio is the same but area is different.*

2.5. Proposed Method

In the current study, an enhanced loss function Area-CIoU (ACIoU) method based on the CIoU algorithm is suggested, which takes into account the area difference between the two bounding boxes as the factor of geometric measurement. This approach successfully prevents the CIoU algorithm from changing into the DIoU algorithm when the aspect ratios of the predicted and ground truth bounding boxes are the same, thus localization accuracy is improved.Let us consider Figure 2, the aspect ratio of ground truth and predicted bounding box are same thus the CIoU loss degenerates into DIoU loss. To avoid this issue, we consider the area difference between the boxes. The penalty term in CIoU is changed to overcome the problem of degenerating of DIoU. When computing the regression loss, CIOU takes three geometric factors into consideration. However, the ground truth box areas in the same image change significantly when using multi-scale detection. It is suggested that the penalty function include a factor for the area difference between the predicted and ground truth bounding boxes in order to solve the problem of penalty function deterioration caused by the CIoU method when $\frac{w^{gt}}{h^{gt}} = \frac{w^{p}}{h^{p}}$. This method effectively prevents the penalty function from degrading, improving the accuracy of loss function estimates for a variety of bounding box sizes. At the same time, it improves the accuracy of the loss function computation by taking geometric elements like aspect ratio and area difference into account. The loss function created in this study is ACIoU, and its formulation is shown in equation (9).

$$\mathcal{L}_{\mathrm{ACIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{B^p},\mathrm{B^{gt}})}{c^2} + \alpha\nu + 0.01(\mathrm{w^{gt}} * \mathrm{h^{gt}} - \mathrm{w^p} * \mathrm{h^p})^2 \quad (9)$$

Where $(1 - IoU + \frac{\rho^2(B^p,B^{gt})}{c^2} + \alpha\nu)$ shows the CIoU loss function and the factor which shows the area difference is defined by equation (10).

$$\mathrm{AD} = 0.01(\mathrm{w^{gt}} * \mathrm{h^{gt}} - \mathrm{w^p} * \mathrm{h^p})^2 \quad (10)$$

In equation (10) the area difference term is multiplied with a small constant 0.01 to give it lower weight than the main loss terms. If the aspect ratio of the ground truth and predicted bounding box become equal then the term αv becomes zero in equation (9), but the area term does not become zero so the loss function becomes as under, explained by equation (11).

$$\mathcal{L}_{\mathrm{ACIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{B^p},\mathrm{B^{gt}})}{c^2} + 0.01(\mathrm{w^{gt}} * \mathrm{h^{gt}} - \mathrm{w^p} * \mathrm{h^p})^2 \quad (11)$$

But when the aspect ratio and the area of the ground truth and predicted bounding box equals then the loss function degenerates to DIoU.

Algorithm: Calculating ACIoU Loss
Input: Bounding box of Ground truth $B^{gt} = (w^{gt}, h^{gt}, x^{gt}, y^{gt})$
Input: Bounding box of prediction $B^p = (w^p, h^p, x^p, y^p)$
Output: $\mathcal{L}_{ACIoU}$
1: If $(B^{gt} \neq 0) \cup (B^p \neq 0)$ do
2: If $\frac{w^{gt}}{h^{gt}} \neq \frac{w^p}{h^p}$ and $w^{gt} * h^{gt} \neq w^p h^p$ then
3: $\alpha\nu \neq 0$
4: $\mathcal{L}_{ACIoU} = 1 - IoU + \frac{\rho^2(B^p,B^{gt})}{c^2} + \alpha\nu + 0.01(w^{gt} * h^{gt} - w^p * h^p)^2$
5: If $\frac{w^{gt}}{h^{gt}} = \frac{w^p}{h^p}$ then
6: $\mathcal{L}_{ACIoU} = 1 - IoU + \frac{\rho^2(B^p,B^{gt})}{c^2} + 0.01(w^{gt} * h^{gt} - w^p * h^p)^2$
7: if $w^{gt} * h^{gt} = w^p * h^p$ then
8: $\mathcal{L}_{ACIoU} = \mathcal{L}_{DIoU} = 1 - IoU + \frac{\rho^2(B^p,B^{gt})}{c^2}$

In the above Algorithm, if B^{gt} *and* B^p are not zero and exist, When $\frac{w^{gt}}{h^{gt}} \neq \frac{w^p}{h^p}$ and $w^{gt} * h^{gt} \neq w^p h^p$, $\mathcal{L}_{ACIoU}$ is used, however, if aspect ratio of the ground truth and predicted boxes become equal the just AD term is added in the penalty term, and if aspect ratio and area both are equal then L_{ACIoU} degenerates to L_{DIoU}.

3. Result and Discussion

Standard YOLOv4 network trained on google colab

environment after changing its loss function with our custom loss function. The training and testing was done on my custom dataset and compared with other results.

(a) (b)

(c) (d)

Figure 3. *Sample dataset images representing each class: (a) bus, (b) person (c) truck (d) motorcycle.*

3.1. Custom Dataset

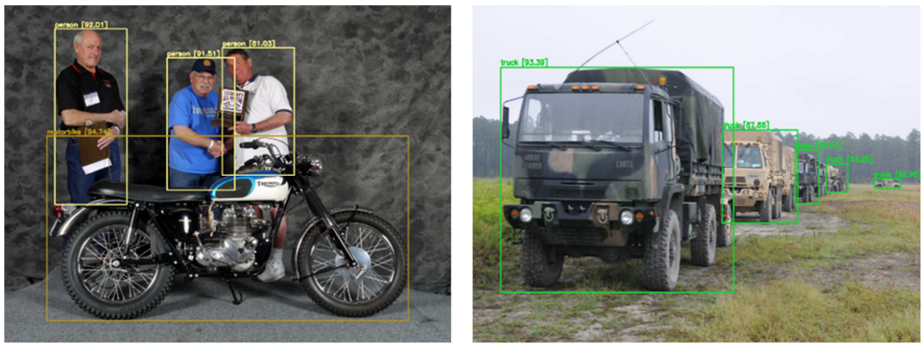
My custom dataset consists of 1704 images of Car, Person, Motorcycle, Truck, and Bus. Which are divided into train and test sets of 1460 images and 244 images respectively. In the training dataset, five classes including people, buses, trucks, cars, and motorcycles were assigned labels. Some sample images of my data set are given in Figure 3, where (a) show a Bus, (b) show a person, (c) show a Truck and (d) show a Motorcycle.

The images with labels are used to train the model and fine-tune the Custom YOLO algorithm's parameters. Each image frame in the video is captured during input for preprocessing before being added to the Custom YOLO algorithm. As seen in Figures 4 for image data and 5 for Video data, we can acquire the location and class information from the video by using the YOLO network to extract the picture characteristics from it.

In Figure 4(a) a person and Motorcycle is detected as shown by predicted bounding box with predicted class, similarly in Figure 4(b) a Bus is detected, in Figure 4(c) a Motorcycle and three persons are detected and in Figure 4(d) Trucks are detected and predicted by bounding box with class labels with confidence level.

Figure 5 shows the detection result of our custom YOLO on video and shows that cars are detected with bounding box and class labels with confidence level.

(a) (b)



(c) (d)

Figure 4. *Vehicles detection Image of (a)Motorcycle, (b)Bus, (c)Person, (d)Truck.*

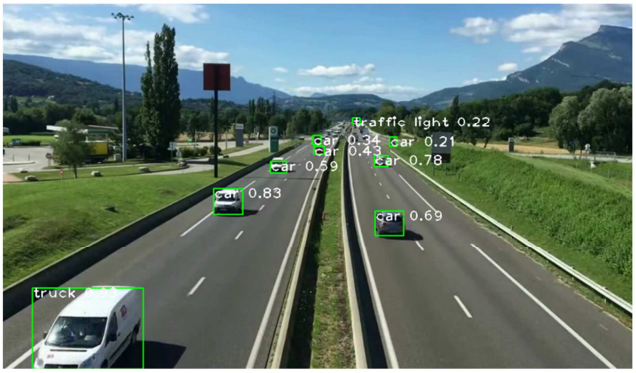


Figure 5. *Vehicle detection in video.*

3.2. Evaluation Protocol

For comparison, we use metrics such as precision (P), recall (R), and mAP. The PR-curve is computed using the model's confidence threshold. The recall is the percentage of all positive samples found over the 50% confidence threshold, and the precision is the percentage of all positive samples detected at the same level of confidence. Equations (12) and (13) illustrate how to compute P and R.

$$Precision(P) = \frac{TP}{TP+FP} \tag{12}$$

$$Recall(R) = \frac{TP}{TP+FN} \tag{13}$$

here the terms TP, FP, and FN denote the sums of true positives, false positives, and false negatives, respectively. The average precision (AP) of the entire number of classes is described by the mAP.

We tested the effectiveness of these algorithms using our custom dataset in the experiments. For the proposed approach and the standard techniques, we measured the precision rate, recall rate, and frames recognized per second (FPS). Table 1 shows the average precision and recall rates for different algorithms. From Table 1 it is cleared that Fast YOLO and our proposed algorithm outperformed other algorithms, particularly in terms of recall rate. Figures 6 illustrate that when the number of samples increase from 2000 to 6000 the precision rate of our proposed method and other standard approaches tend to rise and become stable. Similarly Figure 7 illustrate that the recall rate tends to rise when the number of samples increase from 2000 to 6000. When there are not enough training samples, the model is under-fit. As the number of samples increases, the model may more closely reflect the sample distribution. The model effectively starts to converge once there are 6000 samples. The YOLO algorithm consistently produces the best outcomes in various scenarios.

Table 1. *Results for evaluation metrics for different Algorithms.*

Algorithm Type	Precision (%)	Recall (%)
Sliding Window	70.58	72.90
CNN	80.81	78.37
RCNN	84.18	83.68
Faster R-CNN	83.95	82.64
YOLO	83.95	82.64
Fast YOLO	88.44	86.63
SSD	86.81	84.09
Custom YOLO	88.48	86.37

Table 2 shows the precision rate for different classes across all algorithms. The YOLO, Fast YOLO, and our Custom YOLO algorithms have a much greater precision for all the classes when compared to the other algorithms.

Table 2. *The precision rates for different Classes by using various Algorithms in percentage.*

Algorithm	Car	Person	Motorcycle	Truck	Bus
Sliding Window	65.34	70.25	68.64	68.33	72.15
CNN	81.87	75.37	79.97	80.20	72.30
RCNN	75.37	80.20	85.35	82.51	85.32
Faster R-CNN	76.23	84.31	85.31	74.80	80.20
YOLO	87.63	84.57	89.98	81.20	88.41
Fast YOLO	87.87	84.75	90.11	80.86	88.63
SSD	80.24	84.28	83.57	76.81	80.15
Custom YOLO	87.51	84.60	90.25	81.57	88.83

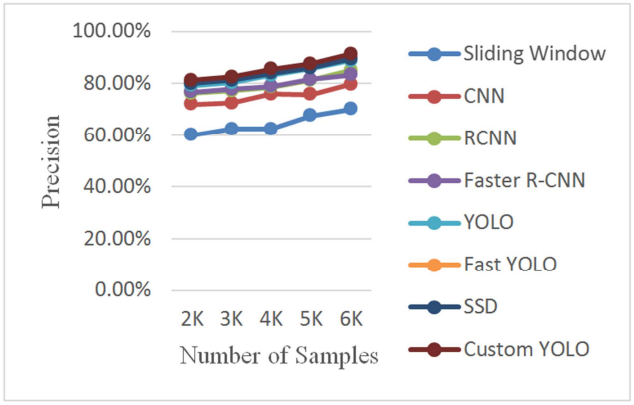


Figure 6. *The Precision for different algorithms based on number of samples.*

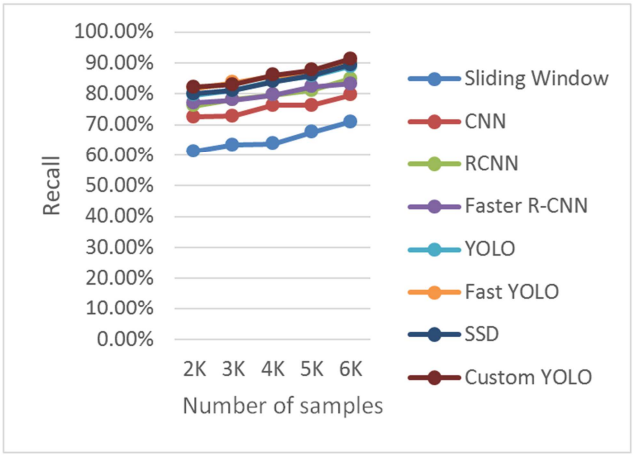


Figure 7. *The Recall for different algorithms based on number of samples.*

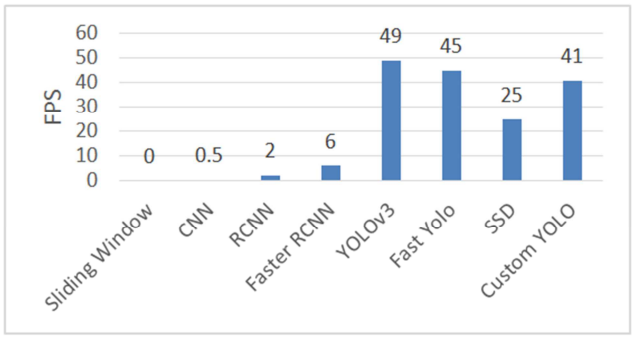


Figure 8. *Histogram of FPS.*

Figure 8 demonstrates that our approach, which uses the YOLO network, can detect at the rate of 41 frames per second.

The video used for traffic monitoring typically has a frame rate of 25. Therefore, using the provided equipment, our approach may accomplish real-time traffic video detection.

4. Conclusions

In this paper, we present a custom YOLO network for fast object detection in real-time. On our custom dataset, we demonstrate how quickly and accurately our algorithm performed. Building upon the YOLOv4 network, we have enhanced the penalty term within the regression loss function. The revised penalty term now incorporates both area differences and aspect ratio relationships between predicted and ground truth bounding boxes. The network's accuracy was enhanced, but its FPS dropped due to the new penalty term's composition in the loss function, known as $\mathcal{L}_{ACIoU}$. We conduct tests on videos and images from our custom dataset using precision and recall rates as well as frames per second (FPS) as assessment criteria. Our approach can detect more images than the ones present in video frames. Therefore, with suitable hardware, it can enable real-time object detection for videos.

Our approach is considerably advanced based on the findings of the $\mathcal{L}_{ACIoU}$ function's calculations and theoretical analysis. The main limitation of our study, as it is described in this publication, is that it can only now be validated in the context of the YOLOv4 network. In order to more accurately evaluate the overall performance of the suggested approach, it is necessary to assess it across multiple neural networks in the future.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] A. Kumar and S. Srivastava, "Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector," Procedia Comput. Sci., vol. 171, no. 2019, pp. 2610–2617, 2020, doi: 10.1016/j.procs.2020.04.283.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[3] J. Tao, H. Wang, X. Zhang, X. Li, and H. Yang, "An object detection system based on YOLO in traffic scene," Proc. 2017 6th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2017, vol. 2018-Janua, pp. 315–319, 2018, doi: 10.1109/ICCSNT.2017.8343709.

[4] T. Ahmad et al., "Object Detection through Modified YOLO Neural Network," Sci. Program., vol. 2020, 2020, doi: 10.1155/2020/8403262.

[5] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[6] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, "A real-time object detection algorithm for video," Comput. Electr. Eng., vol. 77, pp. 398–408, 2019, doi: 10.1016/j.compeleceng.2019.05.009.

[7] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Towards Deep Object Detection Techniques for Phoneme Recognition," IEEE Access, vol. 8, pp. 54663–54680, 2020, doi: 10.1109/ACCESS.2020.2980452.

[8] P. Sermanet and D. Eigen, "OverFeat : Integrated Recognition, Localization and Detection using Convolutional Networks arXiv : 1312. 6229v4 [cs. CV] 24 Feb 2014".

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 580–587, 2014, doi: 10.1109/CVPR.2014.81.

[10] R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[12] W. Liu et al., "SSD: Single shot multibox detector," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[13] J. Redmon and A. Farhadi, "Yolo V2.0," Cvpr2017, no. April, pp. 187–213, 2017, [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/9789812771728_0012

[14] J. Redmon and A. Farhadi, "YOLO v.3, An incremental improvement" Tech Rep., pp. 1–6, 2018, [Online]. Available: https://pjreddie.com/media/files/papers/YOLOv3.pdf

[15] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv. 2020.

[16] X. Wang and J. Song, "ICIoU: Improved Loss Based on Complete Intersection over Union for Bounding Box Regression," IEEE Access, vol. 9, pp. 105686–105695, 2021, doi: 10.1109/ACCESS.2021.3100414.

[17] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox," pp. 516–520, 2016, doi: 10.1145/2964284.2967274.

[18] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 658–666, 2019, doi: 10.1109/CVPR.2019.00075.

[19] X. Qian, S. Lin, G. Cheng, X. Yao, H. Ren, and W. Wang, "Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion," Remote Sens., vol. 12, no. 1, 2020, doi: 10.3390/RS12010143.

[20] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," AAAI 2020 - 34th AAAI Conf. Artif. Intell., no. 2, pp. 12993–13000, 2020, doi: 10.1609/aaai.v34i07.6999.