



## Early Stopping Effectiveness for YOLOv4

Afif Rana Muhammad<sup>1)\*</sup>, Hamzah Prasetyo Utomo<sup>2)</sup>, Priyanto Hidayatullah<sup>3)</sup> ,  
Nurjannah Syakrani<sup>4)</sup> 

<sup>1)2)3)4)</sup>Politeknik Negeri Bandung, Indonesia

Jl. Gegerkalong Hilir, Ds. Ciwaruga, Bandung

<sup>1)</sup> [afif.rana.tif417@polban.ac.id](mailto:afif.rana.tif417@polban.ac.id), <sup>2)</sup> [hamzah.prasetyo.tif417@polban.ac.id](mailto:hamzah.prasetyo.tif417@polban.ac.id), <sup>3)</sup> [priyanto@polban.ac.id](mailto:priyanto@polban.ac.id), <sup>4)</sup> [nurjannahsy@jtk.polban.ac.id](mailto:nurjannahsy@jtk.polban.ac.id)

---

### Abstract

**Background:** YOLOv4 is one of the fastest algorithms for object detection. Its methods, i.e., bag of freebies and bag of specials, can prevent overfitting, but this can be combined with early stopping as it could also prevent overfitting.

**Objective:** This study aims to identify the effectiveness of early stopping in preventing overfitting in the YOLOv4 training process.

**Methods:** Four datasets were grouped based on the training data size and object class, These datasets were tested in the experiment, which was carried out using three patience hyperparameters: 2, 3, and 5. To assess the consistency, it was repeated eight times.

**Results:** The experimental results show that early stopping is triggered more frequently in training with data below 2,000 images. Of the three patience hyperparameters used, patience 2 and 3 were able to halve the training duration without sacrificing accuracy. Patience 5 rarely triggers early stopping. There is no pattern of correlation between the number of object classes and early stopping.

**Conclusion:** Early stopping is useful only in training with data below 2,000 images. Patience with a value of 2 or 3 are recommended.

**Keywords:** Early Stopping, Overfitting, Training data, YOLOv4

**Article history:** Received 20 September 2021, decision after peer review 14 November 2021, accepted 24 December 2021, available online 28 April 2022

---

### I. INTRODUCTION

Overfitting is when a model is too fixated on a training dataset pattern. In machine learning, this could be a major problem. Early stopping is one strategy to prevent overfitting [1]. As the name suggests, it avoids overfitting by stopping the training process earlier than the targeted [1] so that the process is faster, but the results remain accurate. In the telecommunications industry, early stopping has been successfully implemented in the polar code decoding process. The hardware complexity can be reduced from  $O(N \log N)$  to  $O(N)$  by using the belief-propagation based early stopping criterion [2]. Effland et al. have shown that the optimal application of early stopping time affects the obtained peak signal to noise ratio (PSNR) [3].

YOLO is a real-time object detection algorithm. Currently, YOLO is continuously developed and has reached the fourth version, referred to as YOLOv4—the fastest object detection algorithm in the current Ms COCO Dataset. YOLOv4 allows real-time object detection, even on a video. It is also easy to train and use on conventional devices that do not require multiple GPUs [4]. Like those of other object detectors, the architecture is complex due to the introduction of a data augmentation method called Bag of Freebies (BoF). This method can improve object detection accuracy and speed [4], and avoid overfitting [1].

However, as described in the official YOLOv4 documentation, there is still an overfitting possibility, so double-checking each model against the previous iteration is recommended. As such, we will know whether there is a better model than the training's final model [5]. Another solution is by applying early stopping, that is, to stop training early when the next training is predicted not to produce a better model. This method has been widely used in other machine learning experiments thanks to its efficiency and simplicity [6], but YOLOv4 has not implemented it.

This study aims to test the effectiveness of early stopping in assisting the YOLOv4 training process to improve accuracy. YOLOv4 was chosen as the experimental object because it is currently the most effective tool in the object detection category. An optimizer and learning rate adjustment were used in conjunction with early stopping. As such,

---

\* Corresponding author

the training process would be accelerated. If the training is completed faster, the load will be lighter.

Most machine learning model training processes, including YOLOv4, need powerful tools and a lot of energy. In completing the YOLOv4 training, a personal computer's hardware will be pushed to the limits, so the device's power consumption will rise. An Intel i7 CPU, for example, has a maximum power consumption per hour of around 95W. The training process uses 2.28 kWh of electricity for 24 hours. Nvidia 1080 Ti GPU has a maximum power per hour of around 250W. The training process lasts for 24 hours and consumes an electric power of 6 kWh. Power consumption and research costs can be reduced if the training process can be shortened through early stopping.

Regarding cloud servers for the training process, rental fees are generally charged every hour of use. While free machine learning cloud servers are available, such as Google Colab, they have limitations. For example, Google Colab has a 24-hour lifetime limit and an idle time limit (between 30-90 minutes). If we exceed this limit, the virtual machine will shut down automatically. Early stopping could help deal with this situation and reduce the costs at the same time.

## II. LITERATURE REVIEW

Overfitting is a major issue in supervised machine learning [1]. A neural network model performs well on training data but fails on other datasets [7]. Early stopping prevents overfitting by stopping the training process early [6]. The optimal stopping point could be calculated using the Lagrangian approach [3] or patience hyperparameters. Previous studies [10, 11] implemented early stopping with the hyperparameters values of 3 and 5.

Past research has focused on creating a complex neural network architecture [10] to create a model with good generalization. While this complex architecture is more resistant to overfitting [11], it also requires complex data. Therefore, other researchers developed an alternative for good generalization through the learning rate tuning [10]. Research by Mishra et al. [10] implement cosine rate annealing technique with stochastic gradient descent with restart (SGD-R), differential learning rate (DLR), and cycle length multiplication (CLM). Using the above methods, high accuracy could be achieved without lengthy training time. Meanwhile, to tune the learning rate, an optimizer could be used. Research by Chaudhury [12] compares the performance of optimizer stochastic gradient descent (SGD), Nesterov accelerated gradient (NAG), Adam, and RmsProp. The result shows that SGD and NAG methods perform better in preventing overfitting, especially with noisy data.

YOLOv4 is the fourth version of the object detection model. The original 'you only look once' (YOLO) already uses darknet for its framework [13], and both convolutional layers and a fully connected layer. However, it has a major drawback: the inability to detect small objects. The second version of YOLO was named YOLO9000 [14], which uses batch normalization layers and introduces anchor boxes that replace the fully connected layer in YOLO version one. It also uses the darknet-19 classification network for feature extraction. In YOLO version three, detection could be done on three scales, from small to very large [15]. It also uses an independent logistic classifier, which solves a classification issue when an object falls into two classes. The prediction score of this kind of object will remain high in multiple classes. Finally, YOLOv4 uses CSPDarknet53 as its backbone and a spatial pyramid pooling block. It also uses two techniques called bag of freebies and bag of specials, making the training process more effective with little to no extra costs.

## III. METHODS

This section outlines the method used in the current research, i.e., the research variables, research data, experimental setup, and experimental scenarios.

### A. Research Variables

This study has three independent variables: the number of samples in datasets, the number of object classes, and the patience hyperparameter values. These independent variables will affect the results of the dependent variables: the accuracy and the number of iterations.

The size of datasets is an independent variable since the outcome of the neural network process is affected by the number of samples in datasets [1]. The dataset is where the neural network learns from in the training process. The bigger the dataset is, the more capable the neural network model will be. If the neural network is complex, the data required is larger and more varied. The neural network's various parameters will need to be fine-tuned to deliver good training accuracy. That is to say, the number of samples in datasets is limited not only by quantity but also by variety. For example, in the case of object detection, the dataset must truly represent the object's point of view from all conditions to allow the neural network to familiarize itself with the object [1].

The number of object classes is an independent variable because it affects learning complexity; hence, the performance of the YOLOv4 model in detecting an object. More classes mean the model must take more burden to determine the correct class. The model performance directly affects the accuracy value (mAP).

Patience hyperparameter is a value used in early stopping. It is when the signs of overfitting appear, which is usually known after the training is completed. However, overfitting can be predicted.

In this study, if mAP decreases or remains constant compared to the previous iteration, there is a waiting period to see if mAP changes. If it does not change after many iterations (as determined by the patience hyperparameter), early stopping will be commanded to prevent overfitting. If, on the other hand, the mAP rises during the waiting period, the training will continue until overfitting symptoms reappear.

Following the training process, the model's accuracy was assessed. This accuracy was compared to the model accuracy in the same dataset when early stopping was not used. In the YOLOv4 model training process for detecting an object, this comparison determines whether early stopping is effective or not. The mean average precision (mAP) formula was used to determine the accuracy of this model. In this study, the mAP is the average precision (AP) of all point interpolation calculations for each class. AP calculates the area under the precision-recall curve for each recall shown in (1). Then, we calculate the mean for AP in each class.

$$AP = \sum_{n=0}^1 (r_{n+1} - r_n) P_{interp}(r_{n+1}) \quad (1)$$

If early stopping does not happen, the YOLOv4 training process continues until it reaches a point where the process needs to be stopped. The number of iterations became a dependent variable because the training time was not always constant due to the unstable training servers. This number of iterations is linked to accuracy, indicating whether early stopping leads to better or worse accuracy.

### B. Research Data

Following the instructions on the YOLOv4 paper's GitHub page, researchers can determine the criteria for the datasets based on the number of images and their object class. The recommended number of images per object class is 2,000 [5]. In addition, each dataset has three object classes. TABLE 1 below lists the four dataset criteria used in the current experiment. Each dataset that meets these criteria is treated separately, which indicates that the datasets are independent of one another. In other words, the criteria are not met by only one dataset and then modified. The boundary value analysis approach was not used for the number of images. This means that the obtained datasets do not have to be above and below 2,000 images but far above and far below 2,000.

TABLE 1  
 DATASET CRITERIA

Dataset Criteria	Number of Images $\geq 2000$	Number of Object Class $= 3$
Dataset 1	Fulfilled	Not Fulfilled
Dataset 2	Fulfilled	Fulfilled
Dataset 3	Not Fulfilled	Not Fulfilled
Dataset 4	Not Fulfilled	Fulfilled

Four datasets were obtained from the Internet based on these criteria. The four datasets are called Laptop, Eye-Glasses-Goggles, Ball, and Mask in accordance to the object class within the datasets. The dataset Mask contain the object class for person with mask, without mask, and mask worn incorrectly. The datasets were sourced from Google Open Image Dataset and Kaggle [16][17]. As for adjusting the parameter of YOLOv4, to obtain faster training speed and higher accuracy, we used the default learning rate of 0.001 [18]. The training, validation, and test dataset ratio were 68:17:15 following the previous research [9]. Table 2 shows additional information on each dataset that meets the criteria.

A sample of the data used in the study is shown in Fig. 1. An example of an image used by YOLOv4 is shown on the left, and a sample annotation text file for each image is on the right. These annotations have a unique format that YOLOv4 understands.

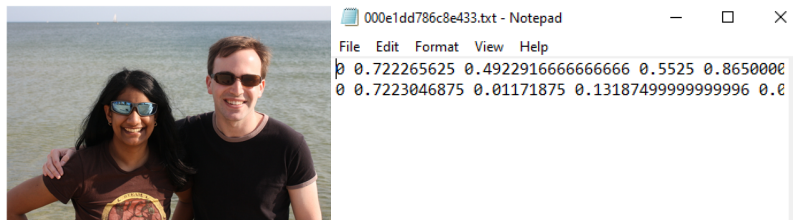


Fig. 1 Examples of data used in experiments

TABLE 2  
 EXPERIMENTAL DATASET

Dataset Name	Number of Object Class	Dataset Split	Number of Images
Laptop	1	Train set	3,913
		Validation set	977
		Test set	862
		Total	5,752
Eye-Glasses-Goggles	3	Train set	5,111
		Validation set	1,278
		Test set	1,126
		Total	7,515
Ball	1	Train set	609
		Validation set	151
		Test set	133
		Total	893
Mask	3	Train set	600
		Validation set	210
		Test set	210
		Total	1,020

### C. Experimental Setup

We use free Google Colab as the experimental equipment in this study. TABLE 3 presents the Google Colab hardware specifications for the experiment.

TABLE 3  
 GOGGLE COLAB HARDWARE SPECIFICATION

Hardware	Specification
Hard Disk	68.4 GB
GPU	Tesla P4
GPU Memory	8 GB
CPU	Intel(R) Xeon(R) CPU @ 2.30GHz
RAM	12.72 GB

### D. Experimental Scenarios

Experiments were conducted by calculating the accuracy of YOLOv4 training. The MAP threshold value was 0.5 (mAP@.5), a common accuracy calculation metric in object detection algorithms [19]. The training's duration was also recorded. This is to compare early stopping and no early stopping in training. The length of the training is also linked to the accuracy of the results in each scenario.

Four YOLOv4 training scenarios were prepared for each dataset. One of them was YOLOv4 training without implementing early stopping. This scenario was intended to gain information on each dataset's accuracy and training duration without early stopping restraints. The other three scenarios are YOLOv4 training with early stopping, with each scenario using the patience hyperparameters 2, 3, and 5.

For the training, since running four scenarios separately would result in a different neural network initial weight—which will affect the result—all scenarios were run within a single training. Training will not stop when early stopping is triggered during the training duration, training will not stop. But the model still saves the neural network current model when the early stopping is triggered. This way, we solved the different initial neural network weights and biases while still running the four different scenarios. Additionally, each scenario was re-enacted eight times in total [20]. This is to see if the occurrence of early stopping is consistent or simply a stroke of luck. This is necessary because YOLOv4 will continue to perform random initialization every time the training begins.

## IV. RESULTS

The results of the experiments are explained in this section: the laptop dataset, eye-glasses-goggles dataset, mask dataset, and ball dataset.

### A. Laptop Dataset

The result of the experiment using the laptop dataset is shown in TABLE 4, showing that in the experiment using the laptop dataset, early stopping with a patience value of 2 always triggers and achieves similar accuracy to training without early stopping. For early stopping with a patience value of 3, it only triggers two out of eight training

repetitions. The accuracy is slightly below average than training without early stopping. A patience value of 5 does not trigger early stopping on the laptop dataset.

TABLE 4  
 SUMMARY OF mAP OBTAINED FROM THE RESULTS OF THE LAPTOP DATASET EXPERIMENT

Repetition	Patience 2			Patience 3			Patience 5			Best			Last		
	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration
1	89.55	90.00	1976	89.80	90.23	2952	-	-	-	90.33	90.33	5148	90.32	90.32	6000
2	89.83	90.38	5148	-	-	-	-	-	-	90.52	90.52	5636	90.10	90.09	6000
3	89.97	90.40	6000	-	-	-	-	-	-	90.12	90.24	5636	89.97	90.40	6000
4	89.61	89.97	4904	-	-	-	-	-	-	90.21	89.75	4416	89.97	90.01	6000
5	89.08	89.46	2220	-	-	-	-	-	-	90.40	90.40	5392	90.02	90.04	6000
6	90.28	90.30	5392	90.04	90.50	5636	-	-	-	90.46	90.50	6000	90.46	90.50	6000
7	89.47	90.17	2708	-	-	-	-	-	-	90.40	90.17	5148	90.32	89.84	6000
8	89.17	89.23	2220	-	-	-	-	-	-	90.28	90.28	5880	90.28	90.06	6000

Furthermore, mAP log data for each iteration was also obtained. This information is required to observe the actual training process in detail. The average mAP for each repetition in the laptop dataset is shown in Fig. 2.

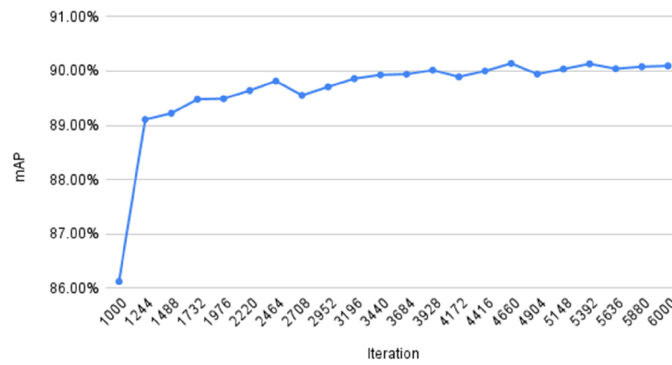


Fig. 2 Average mAP of all laptop dataset experimental repetitions

### B. Eye-Glasses-Goggles Dataset

The result of the experiment using the eye-glasses-goggles dataset is shown in TABLE 5, showing that on the experiment using the eye-glasses-goggles dataset, early stopping with a patience value of 2 triggers two times on the 7<sup>th</sup> and 8<sup>th</sup> repetition. It has similar accuracy to training without early stopping. On eye-glasses-goggles, patience values of 3 and 5 do not trigger early stopping. The average mAP for each repetition for the eye-glasses-goggles dataset is shown in Fig. 3.

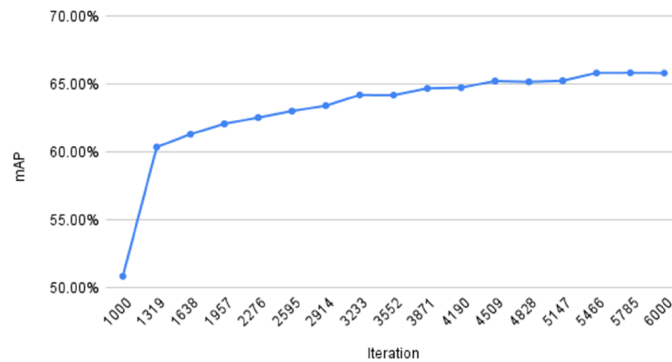


Fig. 3 Average mAP of all eye-glasses-goggles dataset experimental repetitions

TABLE 5  
 SUMMARY OF MAP OBTAINED FROM THE RESULTS OF THE EYE-GLASSES-GOGGLES DATASET EXPERIMENT

Repetition	Patience 2			Patience 3			Patience 5			Best			Last		
	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration
1	-	-	-	-	-	-	-	-	-	66.02	61.38	5466	66.03	61.77	6000
2	-	-	-	-	-	-	-	-	-	66.49	62.27	5785	65.61	62.08	6000
3	-	-	-	-	-	-	-	-	-	65.98	62.43	5785	65.29	61.96	6000
4	-	-	-	-	-	-	-	-	-	65.77	62.26	5785	65.62	61.42	6000
5	-	-	-	-	-	-	-	-	-	66.15	61.85	6000	66.15	61.85	6000
6	-	-	-	-	-	-	-	-	-	65.76	62.68	6000	65.76	62.68	6000
7	64.92	65.32	4828	-	-	-	-	-	-	66.06	66.06	5785	65.97	65.97	6000
8	63.88	64.29	3552	-	-	-	-	-	-	66.22	66.22	6000	66.22	66.22	6000

C. Mask Dataset

The result of the experiment using the mask dataset is shown in TABLE 6, indicating that in the experiment using the mask dataset, early stopping with a patience value of 2 and 3 always triggers and achieve similar accuracy to the training without early stopping. Patience value of 5 triggers early stopping once. The average mAP for each repetition for the mask dataset is shown in Fig. .

TABLE 6  
 SUMMARY OF MAP OBTAINED FROM THE RESULTS OF THE MASK DATASET EXPERIMENT

Repetition	Patience 2			Patience 3			Patience 5			Best			Last		
	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration
1	85.70	86.17	1600	85.70	86.17	1700	-	-	-	86.39	85.57	3000	85.49	85.17	6000
2	85.50	85.26	1800	85.12	85.87	3400	-	-	-	85.95	84.94	5300	84.77	84.24	6000
3	84.17	84.37	1400	84.30	84.75	2600	-	-	-	85.28	84.55	5300	84.33	84.79	6000
4	84.30	85.94	1800	86.03	85.77	4700	86.12	86.03	5400	86.39	86.08	4200	85.32	85.84	6000
5	84.07	84.37	2300	83.77	84.98	2800	-	-	-	84.99	84.21	2500	84.23	84.98	6000
6	84.33	84.88	2100	84.03	85.53	3900	-	-	-	85.91	85.89	5000	85.37	85.37	6000
7	84.20	84.73	1700	84.04	84.13	1800	-	-	-	85.92	85.93	5700	85.16	85.16	6000
8	84.50	84.87	2100	84.54	85.23	2600	-	-	-	85.54	85.56	6000	85.54	85.56	6000

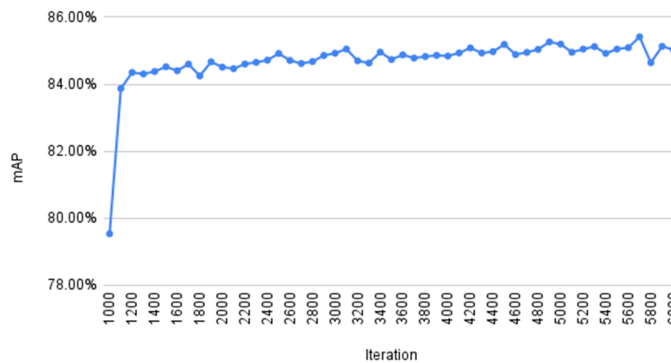


Fig. 4 Average mAP of all mask dataset experimental repetitions

D. Ball Dataset

The result of the experiment using the ball dataset is shown in TABLE 7, suggesting that in the experiment using the ball dataset, early stopping with a patience value of 2 always triggers and achieves, on average, a slightly lower

accuracy than the training without early stopping. Early stopping with a patience value of 3 triggers seven out of eight repetitions with similar accuracy to the training without early stopping. Patience value of 5 did not trigger early stopping. The average mAP for each repetition for the ball dataset is shown in Fig. 4.

TABLE 7  
 SUMMARY OF MAP OBTAINED FROM THE RESULTS OF THE BALL DATASET EXPERIMENT

Repetition	Patience 2			Patience 3			Patience 5			Best			Last		
	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration	Val set (%)	Test set (%)	Iteration
1	79.74	77.08	1400	-	-	-	-	-	-	80.27	78.24	2000	78.70	80.01	6000
2	81.01	76.72	1400	80.79	79.63	4500	-	-	-	81.96	80.06	2300	80.47	79.43	6000
3	79.34	78.24	1600	78.91	79.37	2800	-	-	-	80.09	77.58	4500	78.22	78.60	6000
4	82.60	77.74	1800	82.79	77.26	2400	-	-	-	82.79	77.26	2100	80.30	78.18	6000
5	78.93	79.19	1500	78.53	78.52	1500	-	-	-	79.89	79.89	1500	78.89	78.88	6000
6	79.00	79.33	1300	80.11	80.82	1300	-	-	-	81.46	81.46	1300	79.39	79.59	6000
7	81.16	82.12	1800	80.24	80.35	1800	-	-	-	82.52	82.53	1800	79.55	79.54	6000
8	82.13	82.28	1600	82.00	82.21	1600	-	-	-	83.94	83.76	1600	83.11	83.94	6000

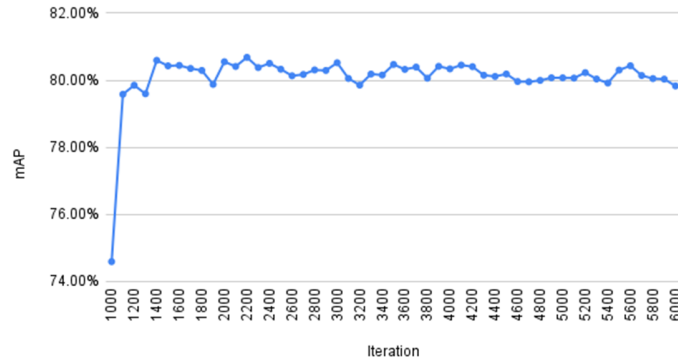


Fig. 4 Average mAP of all ball dataset experimental repetitions

## V. DISCUSSION

The results show that early stopping was more frequently triggered in the mask and ball datasets than laptops and eye-glasses-goggles. This indicates early stopping is sensitive to the types of datasets used. Patience hyperparameter values also influence early stopping occurrences. Meanwhile, there was no significant impact on the number of object classes. The following subsections discuss each point in more depth.

### A. The Impact of Data Size

In each object class, YOLOv4 requires training data of at least 2,000 images [5], but the data available for training may not always be sufficient. This was tested in the current study. The effect of data size on early stopping is presented in TABLE 8. A smaller training set tends to trigger early stopping—more than half of all early stopping scenarios. Thus there is a relationship between the number of samples in datasets and early stopping. In large training sets, early stopping is ineffective because it is rarely triggered.

TABLE 8  
 ANALYSIS OF THE EFFECT OF THE NUMBER OF SAMPLES IN DATASETS ON EARLY STOPPING

Dataset	Number of Training Images	Triggered Experiment
Laptop	3913	10/24
Eye-Glasses-Goggles	5111	2/24
Mask	600	17/24
Ball	609	15/24

### B. The Impact of Patience Hyperparameter

Three patience hyperparameters were used in this study's experiments: 2, 3, and 5, following [8] and [9]. The effect of the patience hyperparameter on early stopping is presented in TABLE 9. A smaller patience hyperparameter value triggers early stopping more frequently. This means the YOLOv4 training curve fluctuates quickly. A larger patience hyperparameter value is not very useful. The selection of the patience hyperparameter can be one of the considerations for using early stopping.

TABLE 9  
 ANALYSIS OF THE EFFECT OF PATIENCE HYPERPARAMETER ON EARLY STOPPING

Dataset	Number of Triggered Experiment		
	Patience 2	Patience 3	Patience 5
Laptop	8 times	2 times	0 times
Eye-Glasses-Goggles	2 times	0 times	0 times
Mask	8 times	8 times	1 time
Ball	8 times	7 times	0 times

### C. The Impact of the Number of Object Classes

There are two types of experimental datasets in this study: a single class object and a multiclass object. Determining single class and multiclass is to see if the number of object classes influences early stopping. The number of classes used in a multiclass category is limited to three. The effect of the number of object classes on early stopping is presented in TABLE 10. From the data below, it can be seen that early stopping was triggered in the four datasets, although with a different number of triggers. With that, it can be concluded that early stopping can be used for single or multiple object classes.

TABLE 10  
 ANALYSIS OF THE EFFECT OF THE NUMBER OF OBJECT CLASSES ON EARLY STOPPING

Dataset	Number of Object Class	Number of Triggered Experiment
Laptop	1	10/24
Eye-Glasses-Goggles	3	2/24
Mask	3	17/24
Ball	1	15/24

### D. Comparison of Accuracy and Speed

The average weight results of cases where early stopping was frequently triggered are shown in TABLE 11. The data presentation only includes two datasets to keep the comparison focused. The patience hyperparameter shown is also those frequently cause early stopping.

TABLE 11  
 AVERAGE MAP AND ITERATIONS IN CASES THAT OFTEN TRIGGER EARLY STOPPING

Dataset	Patience 2			Best			Last		
	Val set	Test set	Iteration	Val set	Test set	Iteration	Val set	Test set	Iteration
Mask	84.60%	85.15%	1,850	85.80%	85.34%	4625	85.03%	85.14%	6000
Ball	80.49%	79.09%	1,550	81.62%	80.10%	2136	79.83%	79.77%	6000

Early stopping does not always lead to higher accuracy. However, when measured by the number of training iterations, the difference is significant. In more detail, TABLE 12 provides information about the accuracy difference between the weight results from early stopping with the best weight and the last weight.

TABLE 12  
 ACCURACY DIFFERENCE BETWEEN EARLY STOPPING WEIGHT COMPARED TO BEST WEIGHT AND LAST WEIGHT

Dataset	Best			Last		
	Val set	Test set	Iteration	Val set	Test set	Iteration
Mask	-1.13%	-1.01%	588	0.66%	-0.68%	4450
Ball	-1.20%	-0.27%	2775	-0.43%	-0.07%	4150

The table shows that the accuracy of early stopping has only been superior to one in eight mAP tests: at the validation set mask. The largest difference in mAP weight between the validation set of the ball dataset and its best weight is 1.20%. Meanwhile, when compared to the last weight, the largest weight difference is 0.68% between the



mask dataset's test set. The biggest difference in the validation set of the mask dataset with its last weight was 0.66%. The overall interval is 1.86%, which indicates that the mAP obtained from early stopping is relatively close to the best and last weights.

Early stopping is superior in all cases in terms of the number of training iterations. Compared to the last weight, the difference reaches more than 3,000 iterations. The mAP obtained was not significantly different. However, if early stopping is not used, there is still the possibility of getting the best mAP from training. Early stopping may still be a viable option if minimizing the time is the primary goal.

#### E. Visualization of Training Load

As described in the introduction, early stopping is expected to lower the cost of training devices. There are two types of costs associated with using the device: electricity usage and cloud rental. When training on a personal computer, the cost of electricity can be a factor to consider. If cloud service is used, however, the cost of renting a service may be a factor to consider.

TABLE 13  
 WORKLOAD VISUALIZATION

Comparison Variables	Early Stopping	Without Early Stopping	Difference
Iteration (+/- 6 seconds)	+/-2,991	6,000	3,009
Estimated time	4.67	10.00	5.02
AWS Cost Assumptions (USD 2.147/hour)	USD10.70	USD21.47	USD10.77
Electricity Charge Assumption (Nvidia 1080 Ti GPU)	1,24kWh	2,5kWh	1,25kWh

TABLE 13 shows that there are numerous benefits when early stopping is triggered. The early stopping iterations listed are the average of the early stopping iterations in this study. The cloud service used as an example here is one of the Amazon Web Service (AWS) services as it has specifications similar to the Google Colab used in this study. Meanwhile, the GPU example was obtained from Pogančić.

## VI. CONCLUSIONS

The effectiveness of early stopping on YOLOv4 was investigated in this study. Sixteen different experimental scenarios were run to see how effective early stopping is. The test was run on four different datasets and three different patience hyperparameters. The four dataset variations used are laptops, eye-glasses-goggles, masks, and balls. The three values of patience hyperparameters used are 2, 3, and 5. Each scenario was run eight times to ensure that the sequence of events was consistent. Because the weight was randomly initialized every time training on YOLOv4 began, repetition was required. So, it would be more visible how often the occurrence of early stopping was triggered in each iteration.

The early stopping implementation experiment results on YOLOv4 were observed based on the accuracy value (mAP) and the length of the training process (number of iterations). In terms of accuracy, the weights produced with early stopping is not much different from the weights produced without early stopping. However, in terms of the number of iterations, the difference is significant. The training process could be shortened. Early stopping is not always an option to get the best mAP from training but can be a solution if the training time is reduced.

The findings reveal an effect on the number of samples in datasets on early stopping. The training set with a small number of images under 2,000 pictures (mask and ball) has a fluctuating training curve. As a result, early stopping is more frequently triggered. Meanwhile, training sets with images over 2,000 (laptops and eyeglasses-goggles) have a training curve that improves over time. This does not trigger early stopping easily. It can be concluded that early stopping is better for YOLOv4 when the number of samples in datasets is small, such as less than 2,000 images.

The patience hyperparameter selection also has an impact. Patience with a value of 2 and 3 triggered early stopping more frequently than 5. The YOLOv4 training curve tends to fluctuate. Using a patience hyperparameter with large values is not recommended in this case. The patience hyperparameter should be used with a value of less than 5. This is because the patience value of 5 is rarely triggered in this study.

The effect of the number of object classes, on the other hand, is not apparent. The ball dataset, which has only one object class, has a fluctuating curve on the one hand. As a result, early stopping is frequently triggered. The training curve in the laptop dataset, on the other hand, continues to improve. This makes it difficult to initiate early stopping. A similar phenomenon happens on datasets with multiple object classes. As a result, early stopping can be applied to a wide range of object classes.

Intuitively, early stopping has benefits. However, we recommend further investigating the effectiveness of early stopping on YOLOv4. For example, by increasing the variation of the dataset of various characteristics and increasing the patience hyperparameter used. We can also investigate how early stopping occurs in unbalanced and noisy datasets.

#### ACKNOWLEDGEMENTS

**Author Contributions:** *Afff Rana Muhammad*: conceptualization, data curation, formal analysis, investigation, project administration, software, visualization, writing. *Hamzah Prasetyo Utomo*: conceptualization, formal analysis, investigation, project administration, software, visualization, writing. *Priyanto Hidayatullah*: funding acquisition, resources, supervision, validation, visualization, writing – review & editing. *Nurjannah Syakrani*: funding acquisition, resources, supervision, validation, visualization, writing – review & editing.

**Funding:** This work was supported by Politeknik Negeri Bandung with SK number B/492/PL1/HK.02.00/2021.

**Acknowledgments:** The authors would like to thank Mr Ade Chandra Nugraha and Mr Jonner Hutahaean for the guidance and inputs on this research and Sophia Gianina Daeli and Atika Khoirunnisa for helping with discussions regarding machine learning and YOLOv4.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### REFERENCES

- [1] X. Ying, "An Overview of Overfitting and its Solutions," in *Journal of Physics: Conference Series*, Mar. 2019, vol. 1168, no. 2, doi: 10.1088/1742-6596/1168/2/022022.
- [2] Y. Yan, X. Zhang, and B. Wu, "Simplified early stopping criterion for belief-propagation polar code decoder based on frozen bits," *IEEE Access*, vol. 7, pp. 134691–134696, 2019, doi: 10.1109/ACCESS.2019.2940135.
- [3] A. Effland, E. Kobler, K. Kunisch, and T. Pock, "Variational Networks: An Optimal Control Approach to Early Stopping Variational Methods for Image Restoration," *J. Math. Imaging Vis.*, vol. 62, no. 3, pp. 396–416, 2020, doi: 10.1007/s10851-019-00926-8.
- [4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv*, 2020.
- [5] A. Bochkovskiy, "AlexeyAB/darknet: YOLOv4v / Scaled-YOLOv4 - Neural Networks for Object Detection (Windows and Linux version of Darknet)," *GitHub*, 2020. <https://github.com/AlexeyAB/darknet>. accessed on 12 Jan2021.
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [7] D. Jones, E. Nn, and M. L. P. Nn, "Neural Networks," 2012, pp. 29–35.
- [8] D. Kadish, S. Risi, and A. S. Løvlie, "Improving Object Detection in Art Images Using Only Style Transfer," Feb. 2021, Accessed: May 21, 2021. [Online]. Available: <http://arxiv.org/abs/2102.06529>.
- [9] M. S. Salekin, A. Babaeian Jelodar, and R. Kushol, "Cooking state recognition from images using inception architecture," *1st Int. Conf. Robot. Electr. Signal Process. Tech. ICREST 2019*, pp. 163–168, 2019, doi: 10.1109/ICREST.2019.8644262.
- [10] S. Mishra, T. Yamasaki, and H. Imaizumi, "Improving image classifiers for small datasets by learning rate adaptations," *arXiv*, 2019.
- [11] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks," *arXiv*, Mar. 2019, Accessed: May 21, 2021. [Online]. Available: <http://arxiv.org/abs/1903.11680>.
- [12] S. Chaudhury and T. Yamasaki, "Robustness of Adaptive Neural Network Optimization under Training Noise," *IEEE Access*, pp. 37039–37053, 2021, doi: 10.1109/ACCESS.2021.3062990.
- [13] 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-December, pp. 779–788, Jun. 2015, doi: 10.1109/CVPR.2016.91.
- [14] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6517–6525, Dec. 2016, doi: 10.1109/CVPR.2017.690.
- [15] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, Accessed: Dec. 04, 2021. [Online]. Available: <https://arxiv.org/abs/1804.02767v1>.
- [16] A. Kuznetsova et al., "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020, doi: 10.1007/s11263-020-01316-z.
- [17] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 426–434, doi: 10.1109/CVPR.2017.53.
- [18] P. Hidayatullah et al., "DeepSperm: A robust and real-time bull sperm-cell detection in densely populated semen videos," *Comput. Methods Programs Biomed.*, vol. 209, p. 106302, Sep. 2021, doi: 10.1016/j.cmpb.2021.106302.
- [19] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electron.*, vol. 10, no. 3, pp. 1–28, 2021, doi: 10.3390/electronics10030279.
- [20] A. Lodwich, Y. Rangoni, and T. Breuel, "Evaluation of robustness and performance of Early Stopping Rules with Multi Layer Perceptrons," *Proc. Int. Jt. Conf. Neural Networks*, pp. 1877–1884, 2009, doi: 10.1109/IJCNN.2009.5178626.

**Publisher's Note:** Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.