Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

Performance Comparison of YOLOv10, YOLOv11, and YOLOv12 Models on Human Detection Datasets

Viky Hendriko[1*], Dedy Hermanto[2]
[1,2]Multi Data Palembang University, Indonesia
[1]vikyhendriko12@gmail.com, [2]dedy@mdp.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 09-07-2025
Accepted: 12-07-2025
Published: 21-07-2025

Keywords:
Dataset; Detection; Model; Performance; You Only Look Once.

ABSTRACT

One popular of object detection model for object detection is You Only Look Once (YOLO) with humans are among the most often utilized for detection objects. Despite the various of human datasets, just a few research that compared the datasets performance against various versions of the YOLO algorithm. This study compares the performance of YOLOv10, YOLOv11, and YOLOv12 on eight different datasets, such as CrowdHuman, CityPersons, Wider Person, Mall Dataset, INRIA, Microsoft Common Object (MS COCO), PASCAL VOC, and MOT17. Precision, recall, mAP@50, and mAP@50-95 are used to measure the YOLO model version's performance on each dataset. The results indicate that each datasets have different perfomance on each version of YOLO, so the performance on model depends on the variation of the dataset. The best results on the MOT17 dataset are obtained by YOLOv12, with 0.909 in precision, 0.775 in recall, 0.88 in mAP@50, and 0.695 in mAP@50-95. On the City Person dataset. However, YOLOv11 performs best result, with 0.782 in precision, 0.529 in recall, 0.694 in mAP@50, and 0.476 in mAP@50-95. Therefore, choosing a YOLO version that is appropriate for the dataset's complexity is essential to creating the best detection model Therefore, selecting the appropriate YOLO version according to the dataset complexity is crucial to obtain the most optimal detection model.

INTRODUCTION

One of the most widely applied fields of Artificial Intelligence was Deep Learning. Deep Learning utilized information from an image to produce predictions or classifications. The Convolutional Neural Network (CNN) method in Deep Learning had become the most favored technique among developers due to its capability to extract information from images effectively, resulting in high accuracy (Fang, Wang, and Ren 2019). In the subject of computer vision, CNN's progress has produced a variety of architectures and methods, particularly for real-time object detection.

Computer Vision provided a wide range of algorithms that were used to classify human actions in videos, track camera movement, follow moving objects, and more (Surbakti and Eka Putri 2022). Object detection algorithms such as You Only Look Once (YOLO), Single Shot Multibox Detector (SSD), and Faster Region-based Convolutional Neural Network (Faster R-CNN) were among the most popular algorithms frequently used in object detection due to their high accuracy (Qiu et al. 2024). One of the most commonly targeted objects in detection algorithms was the human figure. Human detection applications enabled algorithms to perform behavioral analysis, counting, or to serve as AI-based security systems. The performance of these algorithms heavily depended on the datasets used, which served as a key factor in enhancing object detection capabilities.

Although many studies had been conducted to develop object detection algorithms, there remained a need to understand how different dataset characteristics affected the performance of detection models. Popular datasets containing human objects, such as Microsoft Common Objects in Context (MS COCO), PASCAL VOC, Open Images, and others (Sanchez, Romero, and Morales 2020), significantly influenced object detection research. However, each dataset contained varying data conditions in terms of environment and quantity, which impacted the performance of object detection algorithms in recognizing human objects under different circumstances, such as crowd density, lighting, and viewpoint. Therefore, further discussion was needed to explore how dataset variations can effect the performance of algorithm, especially You Only Look Once (YOLO) algorithm, which had evolved through multiple versions.

This study aimed to compare the performance of YOLO versions 10, 11, and 12 across different datasets focused on human detection. The evaluation was conducted using metrics evaluation such as mean Average Precision (mAP), precision, recall, and inference speed (Qiu et al. 2024) to determine the effectiveness of each dataset in supporting the models to achieve optimal performance. By comparing various YOLO versions on different datasets, this study aimed to provide insights into how specific dataset characteristics influenced object detection outcomes based on environmental conditions, thereby assisting in the selection of the most suitable datasets for human object detection.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

LITERATURE REVIEW

Ranjan Sakopta et al. in 2024 conducted a study to thoroughly evaluate the performance of various configurations of YOLOv8, YOLOv9, and YOLOv10 models in detecting fruitlets on trees in commercial orchards, focusing on evaluation metrics such as precision, recall, mAP@50, and computational speed (pre-processing, inference, and post-processing). In addition, this research also aims to validate the detection results with live fruitlet counting in the field using iPhone devices and vision sensors, to support automation and efficiency in early fruit load management in the agricultural sector (Sapkota et al. 2024).

A study conducted by Chenjie Zhao et al. in 2024 provided a comprehensive comparison of datasets in the context of object detection algorithms applied to images captured by Unmanned Aerial Vehicles (UAVs) in maritime environments. The study aimed to compare the performance of their custom dataset, M2Ships, with other datasets in terms of object feature diversity, hardware limitations, and environmental variability. The datasets used included UAV123, DTB70, DAC-SBC, VisDrone2022, AFO, Tiny Person, among others, each containing different class distributions but generally focusing on objects such as humans, ships, and speedboats. The object detection algorithms employed were YOLO, Faster R-CNN, SSD, RetinaNet, CenterNet, NanoDet, and PP-PicoDet-L, with performance comparisons conducted using the mean Average Precision (mAP) metric (Zhao et al. 2024).

Jaskirat Kaur and Williamjeet Singh, in a 2022 study, compared datasets using both two-stage and one-stage object detection algorithms. They employed a systematic review method by analyzing previous research to address seven questions related to object detection. The datasets reviewed included ImageNet, PASCAL VOC, Open Images, and MS COCO. The study compared the performance of these datasets using traditional object detection algorithms such as Viola-Jones Detectors (VJ) and Histogram of Oriented Gradients (HOG), as well as two-stage algorithms like R-CNN, Fast R-CNN, and Faster R-CNN, and one-stage algorithms like YOLO, SSD, RetinaNet, and LADet. Performance was measured using metrics such as mean Average Precision (mAP), frames per second (fps), precision, recall, and F1 score (Kaur and Singh 2022).

In 2023, Haitong Lou et al. developed a new model by modifying the YOLOv8 architecture by replacing the C2f block with a DC module. This DC module was inspired by DenseNet and VoVNet architectures, where the module collected information from previous blocks while reducing redundancy, but essential information was preserved to reduce model loss. The DC-YOLO model was developed using various datasets to assess its performance under different conditions, including VisDrone, PASCAL VOC, and TinyPerson (Lou et al. 2023).

Shrey Srivastava et al., in their 2021 study, enhanced models such as SSD, Faster R-CNN, and YOLOv3 using widely adopted datasets like MS COCO and PASCAL VOC. They noted that MS COCO delivered strong precision and accuracy in object detection and had been extensively used in YOLO and SSD models. On the other hand, PASCAL VOC was more commonly applied in older two-stage detection algorithms such as R-CNN, Fast R-CNN, and Faster R-CNN (Srivastava et al. 2021).

Licheng Jiao et al., in 2020, compared the performance of several object detection models across various datasets. Their study involved models such as Faster R-CNN, YOLO, SSD, RetinaNet, RetinaDet, CornerNet, and NAS-FPN, with modifications such as Haar Cascade integration, different Feature Pyramid Networks (FPN), and diverse backbones. The study aimed to compare the performance of each model and its modifications using evaluation metrics like mAP and inference time. Additionally, the study explored how different datasets affected detection quality. The datasets included MS COCO, PASCAL VOC, ImageNet, VisDrone2018, Open Images V3, and several pedestrian datasets like Caltech, KITTI, CityPerson, TDC, and EuroCity Person. To optimize performance, the researchers made several adjustments, such as improving localization accuracy, addressing detection imbalance, combining one-stage and two-stage approaches, and avoiding the use of pretrained models (Jiao et al. 2020).

Based on previous research, they compared the performance of object detection models with various datasets that have a variety of objects. These studies did not compare the performance of object detection model versions, especially the You Only Look Once (YOLO) model. In some studies, they used YOLOv8, which was released several years ago. In the last 2 years, Ultralytics as a YOLO model provider, has released 3 versions of the YOLO model that use different approaches from the previous versions, namely YOLOv10, YOLOv11, and YOLOv12 (Jocher, Qiu, and Chaurasia 2023). In Fig. 1, shows the difference in interference speed of each YOLO, where the difference in speed and mAP@50 metrics between YOLOv8 and YOLOv10 is significant.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

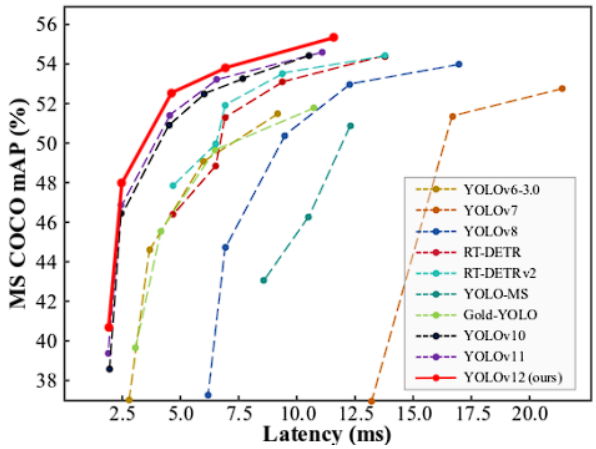


Fig. 1 Latency and mAP@50 Comparison for Each YOLO Version

YOLOv10, YOLOv11, and YOLOv12 will be utilized in this study to see how well each version performs on the different datasets. The three most recent versions of YOLO were chosen because they improved detection accuracy, inference efficiency, and detection capability (Wang et al. 2024). Meanwhile, YOLO version 12 is the most recent version, which focuses on improving object detection in real-world scenarios with low latency.

YOLOv10

The YOLOv10 model has undergone changes from previous YOLO versions by implementing a Dual-label Assignment approach, where both one-to-one and one-to-many approaches are used to produce more accurate bounding box predictions without the need for Non-Max Suppression (NMS) (Wang et al. 2024). With this approach, the post-inference process is not applied, and bounding boxes are predicted directly during the inference process, thereby reducing processing time. YOLO version 10 also employs new blocks to improve efficiency with the Parallel Split-Attention (PSA) module and Compact Inverted Bottleneck (CIB), enabling more efficient multi-scale processing and effective use of Attention blocks (Sapkota et al. 2024).

YOLOv11

The YOLOv11 model uses the C3k2 block to replace the C2f block from its predecessor, implementing a more efficient and faster Bottleneck Cross Stage Partial (CSP) that enhances model performance. Additionally, the addition of the Cross Stage Partial with Spatial Attention (C2PSA) block after the Spatial Pyramid Pooling - Fast (SPPF) block enables the model to focus more on important areas of the image (Sapkota et al. 2024).

YOLOv12

The YOLOv12 model is the latest YOLO model that currently has the fastest performance and highest accuracy. YOLOv12 uses 7×7 separable convolutional layers to reduce computation, and the use of the Residual Efficient Layer Aggregation Network (R-ELAN) ensures stable calculations on each tensor, which improves efficiency in information extraction (Sapkota et al. 2024), (Tian, Ye, and Doermann 2025). The addition of the Flash Attention block before entering the neck and after the neck allows the model to focus on important parts of the image (Alif and Hussain 2025).

METHOD

Start from collect datasets that will be used to develop the YOLO model. Next, the datasets are split into training data and validation data and both data annotations are formatted according to the YOLO annotation standards. The following stage includes data preprocessing *along* with determining the hyperparameter tuning for train the model. The model is evaluated utilizing the evaluation metrics once all datasets have been used for training. Fig. 2 shows the complete research procedure.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

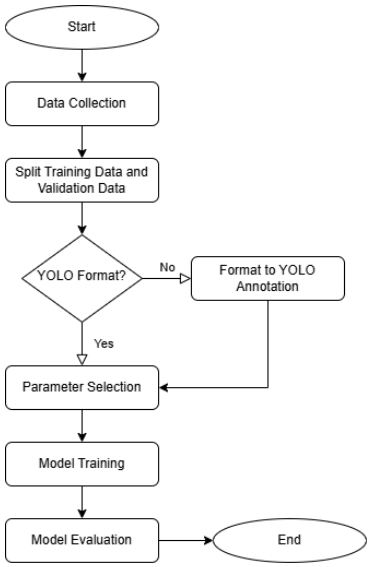


Fig. 2 Research Procedure

Data Collection

This research utilizes the following datasets: Microsoft Common Objects (MS COCO), CrowdHuman, CityPersons, Pattern Analysis, Statistical Modelling, and Computational Learning Visual Object Classes (PASCAL VOC) 2007 + 2012, WiderPerson, Mall Dataset, INRIA Person, and MOT17. Each dataset contains a varied number of data samples and objects. Each dataset was divided with 8 : 2 division ratio for training and validation data. Table 1 shows the quantity of training and validation data from each dataset.

Table 1. Dataset Comparison

Dataset	Total Object	Training	Validation	Total Image
MS COCO	250.000	51.315	12.800	64.515
PASCAL VOC 2007 + 2012	16.000	9.342	2.336	11.678
CrowdHuman	470.000	15.000	4.370	24.370
CityPerson	19.000	2.780	695	5.000
WiderPerson	386.000	7.200	1.800	13.382
Mall Dataset	60.000	1.800	200	2.000
INRIA Person	1.800	702	200	902
MOT17	292.733	4.252	1.064	11.235

Microsoft Common Object (MS COCO)

The MS COCO dataset is widely regarded as one of the best datasets for object detection due to its wide range of images and comprehensive categories. This dataset is useful for a variety of tasks, including object segmentation, detection, and classification, and it covers 91 commonly found object categories. Overall, there are approximately 2,500,000 object annotations distributed across 328,000 images, with the human object category dominating with 250,000 annotations spread over 64,515 images (Lin et al. 2014).

PASCAL VOC 2007 + 2012

The PASCAL VOC dataset is one of the object detection datasets that is known as a competitor to MS COCO due to its wide range of classes and large number of objects. In the 2007 version, this dataset contains 20 classes with a total of 5,032 images and 12,608 objects, where the human category dominates with 4,690 objects. Meanwhile, the 2012 edition consists of 17,125 images with similar 20 classes (Everingham et al. 2007), (Everingham et al. 2012), (Everingham et al. 2010).

CrowdHuman

The CrowdHuman dataset provides data specifically designed for detecting human objects in images with complex backgrounds and diverse scenarios. Each image depicts a crowd of people with conditions such as close proximity, partial occlusion, variations in lighting, and complicated backgrounds. This dataset includes 15,000 training images, 4,370 validation images, and 5,000 testing images, with total of 470,000 human object annotations (Shao et al. 2018).

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

CityPerson

The CityPerson dataset is a collection specifically designed for detecting pedestrians in urban environments. This dataset was compiled from image captures in 27 cities across 3 countries, with each image provided with high-quality annotations for every pedestrian. The photos were taken during three different seasons, spring, autumn, and winter, offering variations in lighting conditions and backgrounds. Overall, this dataset includes 5,000 images with a total of 35,016 human objects engaged in various activities such as walking, riding motorcycles or bicycles, sitting, and others. However, only 3,475 images have complete annotations (Zhang, Benenson, and Schiele 2017).

WiderPerson

The WiderPerson dataset is a collection that features images with high object density, complete with accurate annotations for each individual. The main focus of this dataset is human detection, where each image depicts various situational conditions such as marathons, traffic jams, dynamic activities, dancing, and crowds. In total, this dataset consists of 13,382 images, with 4,382 of them used as test data without annotations, and a total of 386,353 identified human objects (Zhang et al. 2020).

Mall Dataset

The Mall Dataset is a collection focused on indoor environments, such as shopping malls. This dataset consists of a video with a length of 2,000 frames that has been converted into image format with a resolution of 640 × 480 and a framerate of less than 2 Hz. The total number of human objects in this dataset reaches approximately 60,000, with unique annotations provided only on the head region (Loy, Gong, and Xiang 2013).

INRIA Person

The INRIA Person dataset is a collection used to develop new detection models using Histogram of Oriented Gradients (HOG), which is expected to produce better results compared to using Support Vector Machines (SVM). This dataset contains pedestrians from various locations with different poses to introduce diversity in the data. It consists of 614 training images and 288 testing images, with a total of 1,800 human objects (Dalal and Triggs 2005).

MOT17

MOT17 or Multi-Object Tracking 2017, is a dataset redeveloped from MOT16 with improved ground truth precision and the addition of detection results using Faster R-CNN, Deformable Part-based Model (DPM), and Spatial Dependency Perception (SDP). MOT17 was derived from 14 pedestrian-dense videos in both indoor and outdoor environments, converted into images from each frame, with a total of 292,733 human objects. This dataset provide 7 of videos for training data and 7 of videos for test data, totaling 5,316 training images and 5,919 testing images (Milan et al. 2016).

Parameter Selection

Several key parameters are used during the YOLO model's training process to ensure optimal optimization and consistency. The parameters used are the initial and final learning rates, momentum, batch size, image size, optimizer, epoch, IoU threshold, and confidence threshold. All of these parameters will be applied to all datasets containing YOLOv10, YOLOv11, and YOLOv12. Table 2 shows the parameters used to train YOLO model

Table 2. Parameter for Model Training

Paramater	Value
Initial learning rate	0.01
Final learning rate	0.1
Momentum	0.937
Batch size	16
Optimizer	SGD
Epochs	100
Intersection over union	0.6
Confidence threshold	0.25
Mosaic	0.5
Warmup epochs	10
Close Mosaic	20
Cosine learning rate	True

Metric Evalution

Models are assessed based on measures such as accuracy, recall, and mean Average Precision (mAP) at Intersection over Union (IoU) 50 and 50-95. Precision is a statistic that assesses how precise the model is in making correct predictions. Precision is measured by dividing the number of True Positives (TP) by the total number of True

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

Positives and False Positives (FP), as seen in equation (1). Meanwhile, recall is a statistic that indicates how well the model catches all genuine positive examples. Equation (2) shows how recall is computed by dividing the number of True Positives by the number of True Positives and False Negatives (FN).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

The mean Average Precision (mAP) in equation (4) is calculated using the average precision (AP) in equation (3), which includes precision and recall. AP is calculated based on the precision-recall curve by considering various thresholds to assess the performance of the object detection model (Lou et al. 2023). By adding the AP values of each class and dividing by the number of classes, mAP offers an overview of the model's performance across classes.

$$\text{Average Precision (AP) }) = \sum_{k=0}^{n}(R_k - R_{k-1}) \times P_k \qquad (3)$$

$$\text{mean Average Precision (mAP)} = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (4)$$

The mean average precision (mAP) metric is classified into two types, mAP@50 and mAP@50-95, both based on intersection over union (IoU). IoU calculates the ratio between the intersection and union areas of two bounding boxes, the prediction box and the ground truth, to determine how much overlap exists (Pahlevi 2024). In mAP50, the IoU threshold is set to 0.5 (50%), therefore the prediction is judged right if the overlap is 50% or more. In contrast, mAP50-95 calculates mAP gradually throughout a range of IoU thresholds from 0.5 to 0.95, allowing for a more rigorous and extensive investigation of various degrees of detection accuracy.

RESULT

The Ultralytics library version 8.3.91, which provides the model, is used to train the model on the Kaggle platform (Jocher et al. 2023) and utilizes PyTorch version 2.5.1+cu121 as the Deep Learning framework (Paszke et al. 2019). The model development uses Python version 3.10.12 and hardware consisting of NVIDIA TESLA T4 with CUDA version 12.1. The YOLO model variant used is YOLO nano because it has fewer parameters, resulting in faster computation processes.

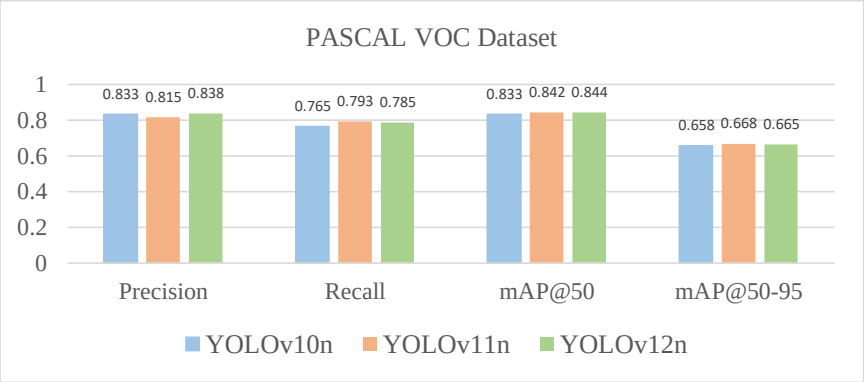


Fig 3. Model Training Results with PASCAL VOC Dataset

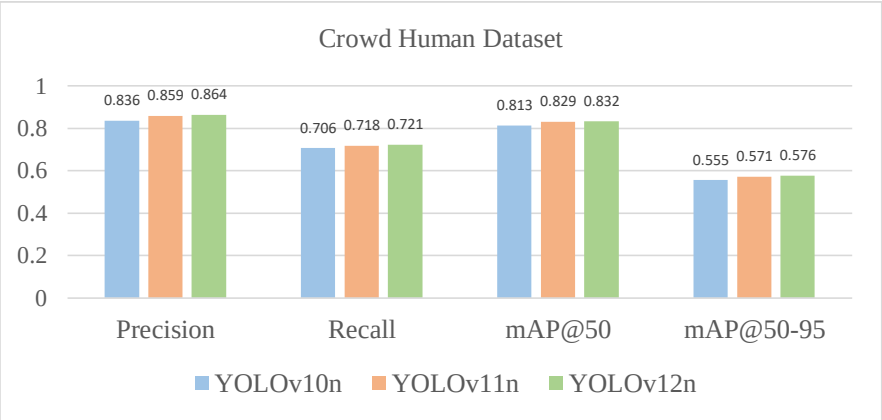


Fig 4. Model Training Results with CrowdHuman Dataset

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

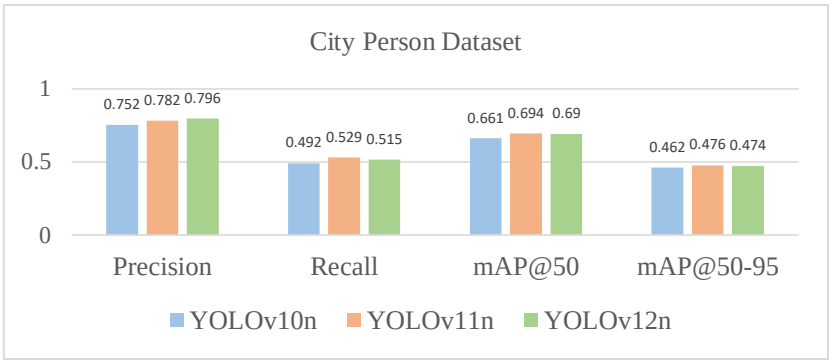


Fig 5. Model Training Results with City Person Dataset

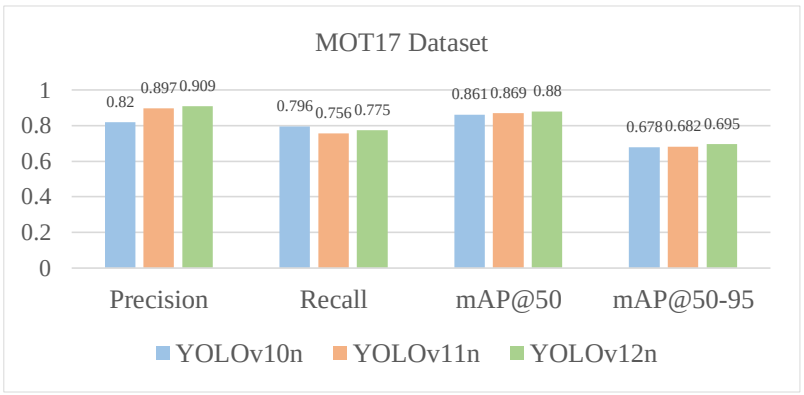


Fig 6. Model Training Results with MOT17 Dataset

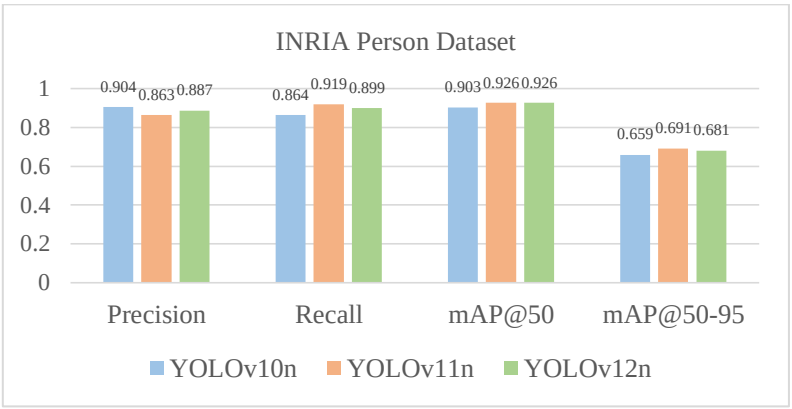


Fig 7. Model Training Results with INRIA Person Dataset

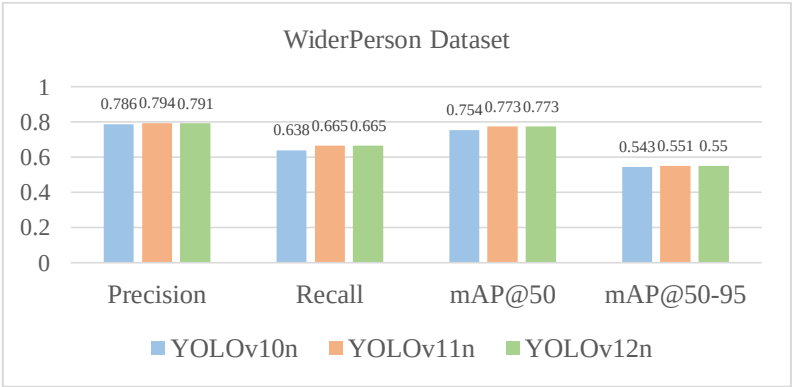


Fig 8. Model Training Results with WiderPerson Dataset

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

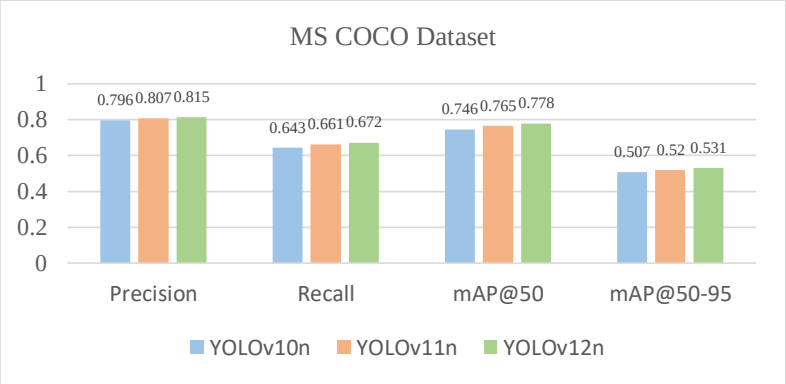


Fig 9. Model Training Results with MS COCO Dataset

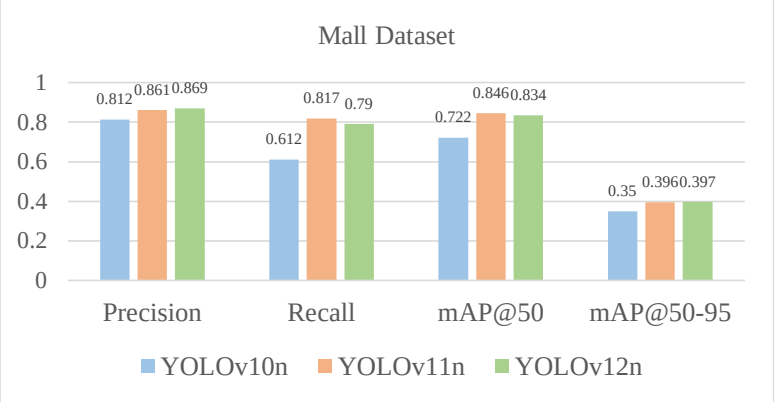


Fig 10. Model Training Results with Mall Dataset

Based on the results of training the YOLO model on 8 datasets shown from fig. 3 to fig. 10, each model has different performance. On the majority of the datasets, including PASCAL VOC, CrowdHuman, MOT17, and MS COCO, YOLOv12 generally performed the best in terms of precision and mAP@50 measures. On the other hand, YOLOv11 performs exceptionally well in certain recall and mAP@50-95 metrics, particularly on the CityPersons and INRIA Person datasets, demonstrating its capacity to identify items that are difficult to detect or are missed.

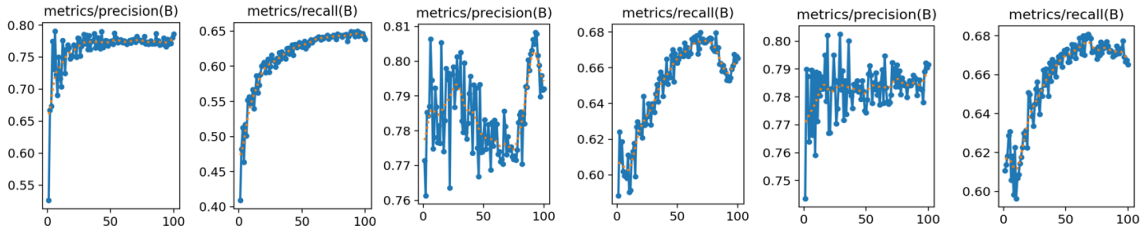
The results indicate a gradual improvement in precision and detection accuracy from YOLOv10 to YOLOv12, even though the differences between versions are not always noteworthy. YOLOv10 typically performs the worst across nearly all metrics and datasets. On the Mall dataset, however, head-based detection alone and crowded conditions caused all models to struggle, as evidenced by extremely low mAP50-95 values.

DISCUSSION

The training of the YOLOv10, YOLOv11, and YOLOv12 models using eight different datasets was conducted to measure the performance and effectiveness of each model under various conditions and data variations presented by each dataset. From the training results, it was found that the three model versions, YOLOv10n, YOLOv11, and YOLOv12n exhibited different performance outcomes across each dataset. Under certain conditions, a particular version of the model may have advantages over the others. The varying complexity of the datasets necessitates more suitable adaptations for each model version. For example, on the WiderPerson dataset, YOLOv11n demonstrated better evaluation metrics compared to other versions. However, when observing the precision and recall curves per epoch as shown in Figure 8, YOLOv10 shows more stability during model training compared to YOLOv11n and YOLOv12n, which are less stable in each epoch and thus require more epochs to converge. Therefore, the YOLOv10 model is more capable of achieving convergence better and faster than the other versions.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

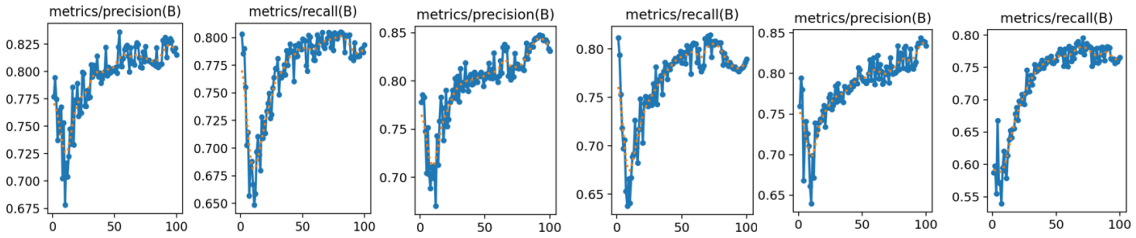


(a) (b) (c)

Fig. 11 Precision and Recall for Each Epoch with WiderPerson Dataset
(a) YOLOv10n (b) YOLOv11n (c) YOLOv12n

In the INRIA Person dataset, YOLOv10n has an advantage in the precision metric compared to YOLOv11n and YOLOv12n. This indicates that YOLOv10n can be used to predict objects with a lower detection error rate on the INRIA dataset. Meanwhile, in the MOT17 dataset, YOLOv10n shows superiority in the recall evaluation metric, which allows the model to detect more of the objects that should be detected. However, compared to the newer YOLO versions, the performance of YOLOv10 still lags behind, especially in the mAP metric.

YOLOv11n demonstrates fairly consistent performance across all evaluation metrics, particularly in recall, as tested on the PASCAL VOC, MOT17, and INRIA Person datasets. This indicates that the model has fewer false negatives or can successfully identify the items that need to be identified. However, when viewed from the precision and recall curves in each epoch, YOLOv11n exhibits considerable fluctuations in precision and recall across epochs, indicating that YOLOv11n actually requires more training epochs to achieve convergence. On the other hand, YOLOv12n shows relatively stable performance in precision and recall across epochs compared to YOLOv11n and YOLOv10n on the PASCAL VOC dataset, as seen in Figure 9.



(a) (b) (c)

Fig. 12 Precision and Recall for Each Epoch with PASCAL VOC Dataset (a) YOLOv10n (b) YOLOv11n (c) YOLOv12n

YOLOv12n demonstrates better performance compared to YOLOv10n and YOLOv11n across almost all datasets, such as MS COCO, the Mall Dataset, and CrowdHuman. YOLOv12n achieves higher in precision, mAP@50, and mAP@50-95 than the previous versions. This balanced and stable model performance makes it suitable for more complex scenarios, such as those found in the Mall Dataset. The Mall Dataset features relatively small head bounding boxes, requiring the model to effectively adapt to such objects. YOLOv12n is able to produce better evaluation metrics on this dataset compared to others, making it a model that adapts well to data variation and complexity.

Performance analysis across the eight datasets also shows that the differences in mAP50 and mAP50-95 values between models vary depending on the dataset. Dataset complexity plays a major role in determining model performance. For instance, PASCAL VOC contains a moderate object density, allowing YOLOv10 to remain relevant for use with this dataset. However, in more complex datasets such as MOT17 and CrowdHuman, YOLOv10 becomes less suitable compared to YOLOv11 and YOLOv12, which deliver almost equal performance. Complexity is also influenced by lighting conditions and variations in camera angles within the dataset. Thus, selecting an appropriate dataset for a specific application depends on the user's needs. In this context, YOLOv11 and YOLOv12 tend to provide more robust results compared to YOLOv10, which experiences a significant performance drop in mAP metrics.

CONCLUSION

The results of the evaluation of the YOLOv10, YOLOv11, and YOLOv12 model performance on eight distinct datasets indicate that each model's performance is significantly impacted by the complexity and various datasets. The dataset's complexity and environmental factors, including lighting, viewing angle, and item density, present various obstacles for each YOLO model to achieve optimal performance. In general, YOLOv11 and YOLOv12 outperform YOLOv10, particularly when it comes to datasets with high complexity. Nevertheless, YOLOv10 can still be applied to datasets with low levels of complexity, depending on the dataset's requirements.

The performance of the YOLO models also showed that version upgrades not only provided better results overall

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

but also offered improved adaptability to varying data conditions. YOLOv12 was a better option in some situations because it continuously produced better results in measures like mAP and recall. Nonetheless, YOLOv11 continued to outperform YOLOv12 in several circumstances. Consequently, when it comes to comprehensive and precise detection under various data settings, both YOLOv11 and YOLOv12 can be regarded as choices. However, since no single model is infallibly better in every situation, the best object detection model should still be chosen after taking into account the particulars of the dataset being utilized as well as the requirements of the application as a whole. These results highlight how crucial it is to have a thorough grasp of the data context while creating and using YOLO-based object detection models.

REFERENCES

Alif, Mujadded Al Rabbani, and Muhammad Hussain. 2025. "YOLOv12: A Breakdown of the Key Architectural Features." *ArXivLabs*.

Dalal, N., and B. Triggs. 2005. "Histograms of Oriented Gradients for Human Detection." Pp. 886–93 vol. 1 in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2007. "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." Retrieved March 16, 2025 (http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html).

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results." Retrieved March 16, 2025 (http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html).

Everingham, Mark, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. "The Pascal Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision* 88(2):303–38. doi: 10.1007/s11263-009-0275-4.

Fang, Wei, Lin Wang, and Peiming Ren. 2019. "A Novel Squeeze YOLO-Based Real-Time People Counting Approach." *International Journal of Bio-Inspired Computation* 14(4):1. doi: 10.1504/ijbic.2019.10024002.

Jiao, Licheng, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. 2020. "A Survey of Deep Learning-Based Object Detection." *IEEE Access* 7(3):128837–68. doi: 10.1109/ACCESS.2019.2939201.

Jocher, Glenn, Jing Qiu, and Ayush Chaurasia. 2023. "Ultralytics YOLO."

Kaur, Jaskirat, and Williamjeet Singh. 2022. "Tools, Techniques, Datasets and Application Areas for Object Detection in an Image: A Review." *Multimedia Tools and Applications* 81(27):38297–351. doi: 10.1007/s11042-022-13153-y.

Lin, Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. "Microsoft COCO: Common Objects in Context." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS(PART 5):740–55. doi: 10.1007/978-3-319-10602-1_48.

Lou, Haitong, Xuehu Duan, Junmei Guo, Haiying Liu, Jason Gu, Lingyun Bi, and Haonan Chen. 2023. "DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor." *Electronics (Switzerland)* 12(10):1–14. doi: 10.3390/electronics12102323.

Loy, Chen Change, Shaogang Gong, and Tao Xiang. 2013. "From Semi-Supervised to Transfer Counting of Crowds." *Proceedings of the IEEE International Conference on Computer Vision* 2256–63. doi: 10.1109/ICCV.2013.270.

Milan, Anton, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. 2016. "MOT16: A Benchmark for Multi-Object Tracking." *ArXivLabs* 1–12.

Pahlevi, Said Mirza. 2024. *Kecerdasan Buatan Dengan Deep Computer Vision*. Jakarta: PT Elex Media Komputindo.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *Advances in Neural Information Processing Systems* 32(NeurIPS).

Qiu, Xiaoyang, Yajun Chen, Wenhao Cai, Meiqi Niu, and Jianying Li. 2024. "LD-YOLOv10: A Lightweight Target Detection Algorithm for Drone Scenarios Based on YOLOv10." *Electronics (Switzerland)* 13(16). doi: 10.3390/electronics13163269.

Sanchez, S. A., H. J. Romero, and A. D. Morales. 2020. "A Review: Comparison of Performance Metrics of Pretrained Models for Object Detection Using the TensorFlow Framework." *IOP Conference Series: Materials Science and Engineering* 844(1). doi: 10.1088/1757-899X/844/1/012024.

Sapkota, Ranjan, Zhichao Meng, Martin Churuvija, Xiaoqiang Du, Zenghong Ma, and Manoj Karkee. 2024. "Comprehensive Performance Evaluation of YOLOv10, YOLOv9 and YOLOv8 on Detecting and Counting Fruitlet in Complex Orchard Environments." (March).

Shao, Shuai, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. "CrowdHuman: A

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6447

Benchmark for Detecting Human in a Crowd." *Artikel Ilmiah* 1–9.

Srivastava, Shrey, Amit Vishvas Divekar, Chandu Anilkumar, Ishika Naik, Ved Kulkarni, and V. Pattabiraman. 2021. "Comparative Analysis of Deep Learning Image Detection Algorithms." *Journal of Big Data* 8(1). doi: 10.1186/s40537-021-00434-w.

Surbakti, Agung Wibowo Ardiyanta, and Rahmi Eka Putri. 2022. "Penghitung Pengunjung Dan Deteksi Masker Menggunakan OpenCV Dan YOLO." *Chipset* 3(02):83–93. doi: 10.25077/chipset.3.02.83-93.2022.

Tian, Yunjie, Qixiang Ye, and David Doermann. 2025. "YOLOv12: Attention-Centric Real-Time Object Detectors." *ArXivLabs*.

Wang, Ao, Hui Chen, Lihao Liu, Kai Chen, and C. V May. 2024. "YOLOv10 : Real-Time End-to-End Object Detection." *ArXivLabs* 1–18.

Zhang, Shanshan, Rodrigo Benenson, and Bernt Schiele. 2017. "CityPersons: A Diverse Dataset for Pedestrian Detection." *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua:4457–65. doi: 10.1109/CVPR.2017.474.

Zhang, Shifeng, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, and Guodong Guo. 2020. "WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild." *IEEE Transactions on Multimedia* 22(2):380–93. doi: 10.1109/TMM.2019.2929005.

Zhao, Chenjie, Ryan Wen Liu, Jingxiang Qu, and Ruobin Gao. 2024. "Deep Learning-Based Object Detection in Maritime Unmanned Aerial Vehicle Imagery: Review and Experimental Comparisons." *Engineering Applications of Artificial Intelligence* 128:1–32. doi: 10.1016/j.engappai.2023.107513.