Human Detection By Improved YOLOv8 Model

Nadillia Sahputra
nadillia.sahputra@studenti.unipd.it

Qiqi Zhang
qiqi.zhang@studenti.unipd.it

Abstract

*We adopt model YOLOv8 which is the newest state-of-the-art YOLO model that can be used for object detection, image classification, and instance segmentation tasks as our pre-trained model for human detection. To fine-tune the model, we use Pascal VOC and Human Detection Dataset to adapt the model to our specific domain. We also try to freeze part of the basic structure, fine-tune the hyper-parameters and increase the resolution of training set to achieve the best performance of the model. We find the highest accuracy is achieved when we freeze the backbone, set epoch to be 25 and use 1024*1024 input. Our improved models also indicate better performance in crowd situation including detection of part of the human body. All the work done by this task can also be extended to other classes of objects detection .*

1. Introduction

The human detection is a fundamental task in computer vision that involves identifying the presence of humans in images or video streams. This task is essential for various applications and has garnered significant attention in research and industry. Improving safety in public spaces is facilitated by its application, which proves beneficial in real-time surveillance for detecting individuals and managing crowds. Furthermore, it contributes to city planning and the efficient allocation of resources. The other promising application could be used in robotics to help robots recognize and interact with human beings.

In this work, we adopt YOLOv8 as our pre-trained model. YOLO is a family of real-time object detection algorithms known for their efficiency and speed. YOLOv8 is the newest state-of-the-art YOLO model which was developed by Ultralytics and has some significant improvements compared to former versions.

We apply the Pascal VOC dataset and Human Detection Dataset to retrain YOLOv8 and we try several experiments including freezing part of the original structure, fine-tuning the hyper-parameters, inputting higher resolution dataset and combining two datasets. Our results show that best performance is achieved by freezing the backbone, choosing epochs to be 25 with 1024*1024 training photos. And the improved model also perform better under super crowd situation including human who only shown part of the body in the photo. Combining the two datasets VOC and HDD shows result is more resistant to double identification with increased confidence.

2. Related Work

Early human detection methods were often based on handcrafted features and traditional computer vision techniques [1]. In the early 2010s, the advent of deep learning, especially convolutional neural networks (CNNs), brought a significant breakthrough in human detection. The R-CNN model [2], improved object detection, including humans, by proposing regions of interest. Building upon R-CNN, the Faster R-CNN architectures were introduced later [3]. These models improved the speed and accuracy of human detection tasks. Published in 2016, the YOLO architecture [4], revolutionized real-time human detection. SSD is another real-time object detection model that performs well in detecting humans among other objects [5].

A lot of models based on CNN also have been conducted for crowded human counting [6], as an efficient way to avoid security issues in public places like stampedes.

In our work, we adopt the newest version YOLO published in January 2023 as our pre-trained model. And apply our selected VOC and HDD datasets to improve the model performance for a specific task (human detection).

3. Dataset

The original dataset used for training and validating YOLOv8 is the Common Objects in Context (COCO) dataset.

We take two datasets as our selected datasets. The first one is the PASCAL Visual Object Classes (VOC) dataset [7], a widely used benchmark dataset in computer vision, specifically designed for object recognition and detection tasks. It has been instrumental in advancing the development and evaluation of algorithms for various computer vision challenges. The VOC dataset is known for its diversity of object classes, annotations, and image contexts. For our

specific task, we extract the class for the person and convert the annotation type from VOC format to YOLO format. We take 1239 pictures to be the training set and 910 pictures as the validation set.

The second dataset, Human Detection Dataset (HDD), is taken from GitHub repository [1]. This dataset is collected in Vietnam by the author of the Github page. It focuses only on human detection for people walking on the street. The dataset contains 3862 photos in total, we select 2220 images for the training set, 1232 images for the validation set and 410 for testing set.

We selected these two datasets due to their different characteristics that might complement each other. The first, PASCAL VOC dataset, is known for its diverse range of human images. In contrast, the second dataset is more specific for human detection on the street. This selection might allow us to enrich the model and lead to better outcomes.

To make the model more robust to variations, we applied augmentation. This is done before the training. The dataset is augmented using the automatic feature provided by Ultralytics. The modifications are applied randomly to the dataset.

4. Method

In our work, we use YOLOv8 as our pre-trained model and explore some modified methods for improving the model performance.

4.1. YOLOv8 Structure

YOLO stands as a single-shot algorithm that efficiently classifies an object in a single pass. It accomplishes this by employing a singular neural network to predict bounding boxes and class probabilities, utilizing the entire image as input. The YOLOv8 represents the latest addition to the YOLO algorithm series.

The figure 1depicts the fundamental mechanics of an object detection model. The structure comprises a backbone, neck, and head. The backbone, a pre-trained Convolutional Neural Network (CNN), extracts low, medium, and high-level feature maps from an input image. The neck integrates these feature maps through path aggregation blocks, such as the Feature Pyramid Network (FPN). Subsequently, the combined feature maps are forwarded to the head for object classification and bounding box prediction. For YOLO, the head only consist of one-stage or dense prediction models. For models like R-CNN, the head contains two-stage detector: dense prediction and sparse prediction. The main features of YOLOv8 include mosaic data augmentation, anchor-free detection, a C2f module, a decoupled head, and a modified loss function.

[1] https://github.com/truong11062002/yolov8_for_human_detection/blob/main/YOLOv8_training_human_detection.ipynb

Frozen Layers	mAP50-95	Improved
-	0.44	Baseline
10	0.487	Yes
15	0.473	Yes
20	0.426	No

Table 1: Results for Different Number of Frozen Layers

4.2. Training Method

To optimize the pre-trained model for our specific task, we first selectively freeze different layers for the Human Detection Dataset to look for the best model. We vary the number of frozen layers for this dataset to be 10, 15, and 20 layers. We utilize mAP50-95 metric to select the best model.

As mentioned before, the automatic augmentation algorithm randomly applies the augmentation. It stops at the last 10 epochs. The gradient descent and the initial learning rate are also automatically adjusted based on the dataset size.

We also explore the possibility of developing a model utilizing two datasets. We use Pascal VOC as the first dataset to be trained and additionally retrain again with the Human Detection Dataset. We use 10 frozen layers to train VOC and for the retaining of the Human Detection Dataset, we use the number of frozen layers that gives the best result in the previous step.

Initially, this retraining involves running the model for 25 epochs. Subsequently, we introduce another variation by extending the training to 40 epochs.

We chose to freeze a minimum of 10 layers to preserve the backbone of the pre-trained YOLO model. This approach ensures that the model's fundamental strengths are retained, and also makes the training process more efficient.

Additionally, we also look for the impact of changing the image size on the model performance. We increase the size from 640x640 pixels to 700x700 and 1024x1024 pixels.

5. Experiments

5.1. Number of Frozen Layers

The comparison between the Pre-trained Model (original) and from the models trained using Human Detection Dataset with a varied number of frozen layers can be seen in the table 1. The highest value is the model utilizing 10 frozen layers to do the training. The dataset in this evaluation is the Human Detection Dataset. This is expected since the model is evaluated using the same dataset.

5.2. Number of Epochs

In this section, the starting point is the best model from the previous frozen layers (10 frozen layers model). This

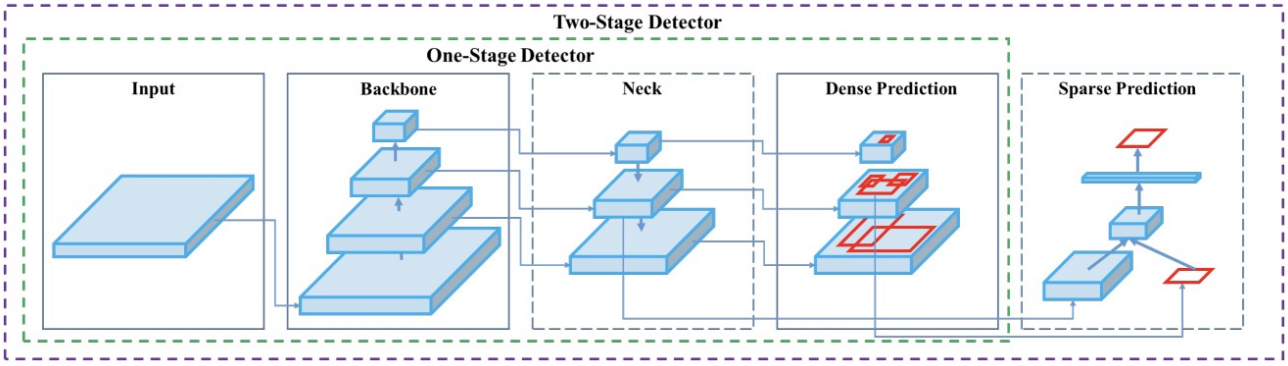


Figure 1: The essential mechanics of an object detection model

Model	Epochs	mAP50-95
HDD (Prev. Best)	25	0.487
VOC+HDD	25	0.484
VOC+HDD	40	0.486

Table 2: Results for Different Number of Epochs

Resolution	VOC+HDD mAP50-95	HDD mAP50-95
640*640	0.484	0.487
700*700	0.505	0.506
1024*1024	0.565	0.595

Table 3: Results for Different Resolution

is then compared to the combined model of Pascal VOC dataset and Human Detection Dataset with 2 variations of epoch (25 and 40). From the table 2, we see no improvement in the value of mAP50-95 in using an additional dataset. This might due to the final testing is using the Human Detection Dataset. To further evaluate the two model, we compare it to random images from neither dataset. This is explored in the section 5.4.

It is also apparent that increasing the number of epochs from 25 to 40 gives little improvement to the model's mAP50-95 value. As seen in fig . 2, the mAP values and validation loss start to plateau while the training loss is still decreasing. To avoid overfitting, for the next evaluation, we use the number of epochs to be 25.

5.3. Different Input Image Resolution

In this section, we compare different input image resolutions to the model of the combined Pascal VOC and Human Detection Dataset. It is clear from the mAP50-95 values in table 3 that increasing the resolution yields better results. However, this does come with the cost of higher memory and computing power needed. Same with the previous result, the models trained on HDD perform slightly better.

5.4. Evaluation on Additional Pictures

To further evaluate the performance of the model, we test it on several images and do a comparison. The images picked is to represent several conditions which are up close (Fig. 3), crowded (Fig. 4), and dense from a distance, (Fig. 5).

From the three different types of pictures, it is apparent that increasing the resolution improves the model significantly. This is in accordance with the results obtained above.

Generally, we see also an improvement from the original pre-trained model especially in a crowded and dense situation where only some parts of the body are shown. The model using Human Detection Dataset with 1024 resolution detects more people however it is prone to double identification. However the combined VOC and HDD is less prone to double identification.

In the upclose scenario, the model of the combined Pascal VOC and Human Detection Dataset performs better than all models. It detects the human correctly without double identification and has higher confidence.

We also test the model performance for videos of walking people, and we put the results in the GoogleDrive.[2]

6. Conclusion

In this paper, we improve the performance of YOLOv8 model for our specific task - detecting humans on the street by retraining parts of the model. Utilizing Pascal VOC and Human Detection Dataset, we find that it is best to freeze only 10 layers (backbone), set the number of epochs to 25 and increase the resolution is a sure way to improve the

[2] https://drive.google.com/drive/folders/1mpYnFRv4yvoo2QJS8MgNeecqMAj6Arf6?usp=drive_link

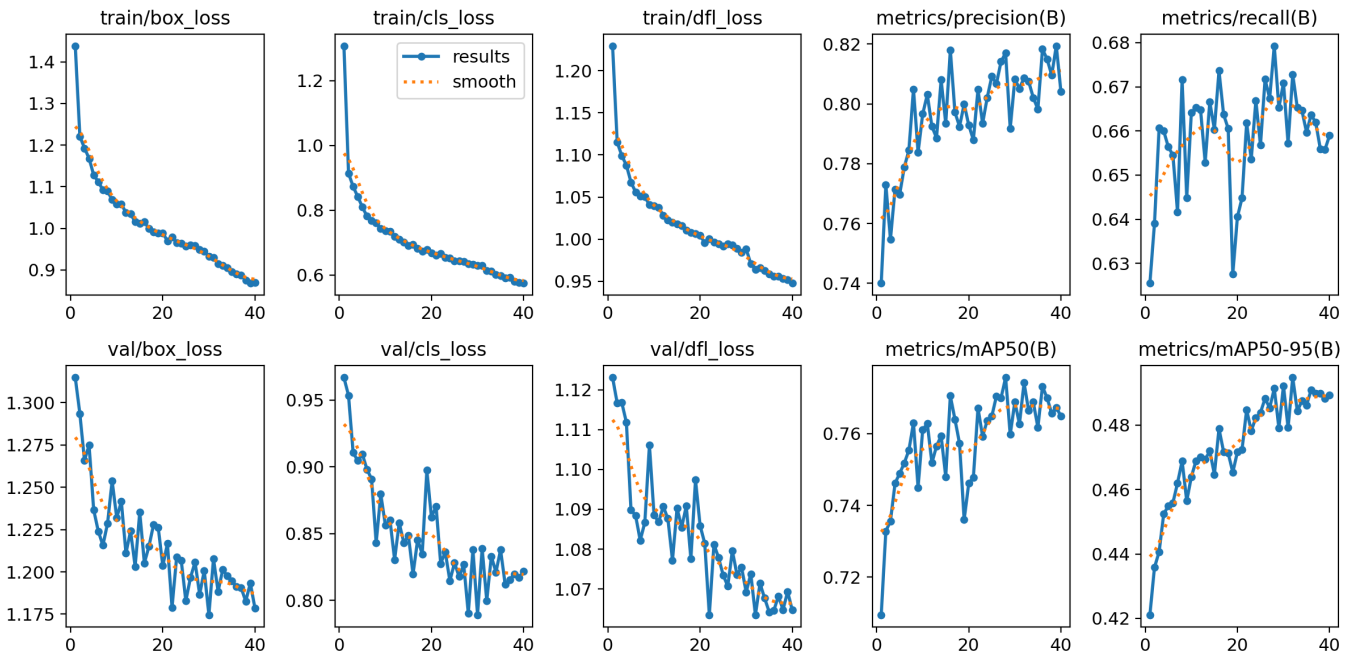


Figure 2: Results for HDD 40 epochs

model. The mAP50-95 value is improved from 0.44 to 0.59. With the best one being trained on only HDD. From additional tests, the VOC + HDD model is less prone to double identification and has higher confidence. The retrained models also behave better in very crowded situations where only parts of the body of human are shown in the photo, which would be an important application to manage the crowd.

This experiment has not evaluated the impact of changing hyperparameters such as weight, types of augmentation, and more. It could be useful to evaluate the hyperparameters more to have a better fine-tuning of the model. It is also useful to have an additional test dataset for a more thorough evaluation. Additionally, to extend the analysis to the video format so it will be closer to the real-life scenarios.

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

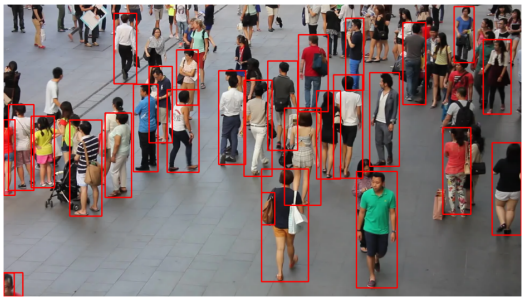
[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37, 2016.

[6] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge, 2010.



(a) Original model

(b) HDD (640x640 input)

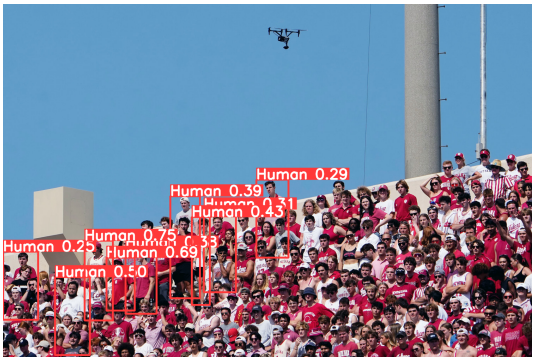
(c) VOC+HDD (1024x1024 input)

(d) HDD (1024x1024 input)

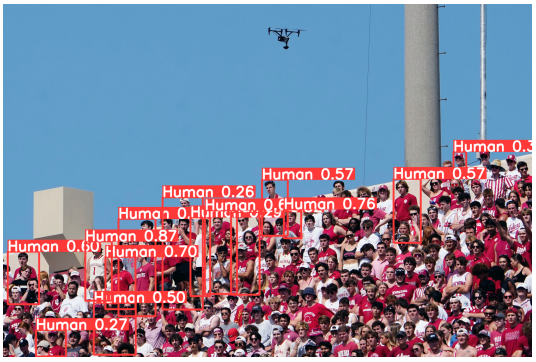
Figure 3: Comparison between 4 models in a crowded slightly sparse condition

(a) Original model

(b) HDD (640x640 input)



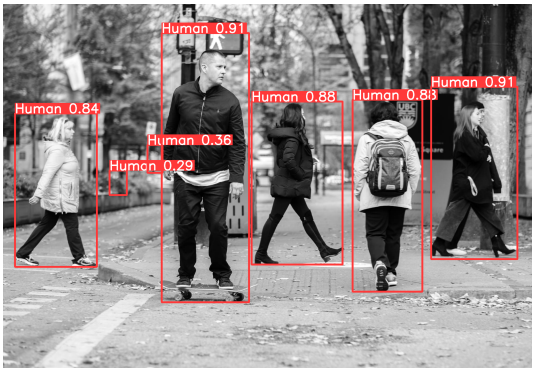
(c) VOC+HDD (1024x1024 input)



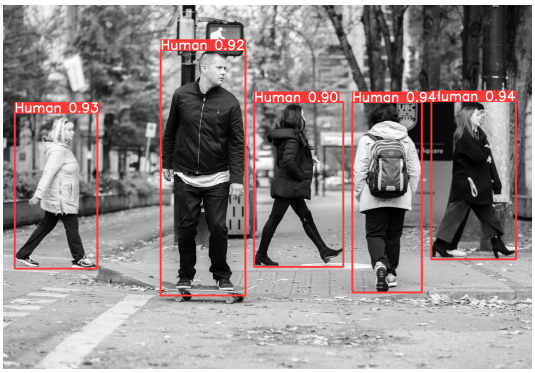
(d) HDD (1024x1024 input)

Figure 4: Comparison between 4 different models in dense condition

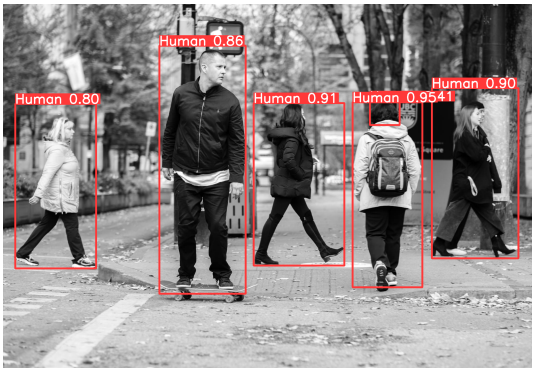
(a) Original model



(b) HDD (640x640 input)



(c) VOC+HDD (1024x1024 input)



(d) HDD (1024x1024 input)

Figure 5: Comparison between 4 different models upclose