MEKATRONIKA
e-ISSN: 2637-0883
VOL. 2, ISSUE 2, 55 – 61
DOI: https://doi.org/10.15282/mekatronika.v2i2.6768

ORIGINAL ARTICLE

Deep Learning Based Human Presence Detection

Venketaramana Balachandran[1], Muhammad Nur Aiman Shapiee[1], Ahmad Fakhri Ab. Nasir[1], Mohd Azraai Mohd Razman[1] and Anwar P.P. Abdul Majeed[1]

[1]Faculty of Manufacturing Engineering, Universiti Malaysia Pahang, 26600 Pahang, Malaysia.

ABSTRACT – Human detection and tracking have been progressively demanded in various industries. The concern over human safety has inhibited the deployment of advanced and collaborative robotics, mainly attributed to the dimensionality limitation of present safety sensing. This study entails developing a deep learning-based human presence detector for deployment in smart factory environments to overcome dimensionality limitations. The objective is to develop a suitable human presence detector based on state-of-the-art YOLO variation to achieve real-time detection with high inference accuracy for feasible deployment at TT Vision Holdings Berhad. It will cover the fundamentals of modern deep learning based object detectors and the methods to accomplish the human presence detection task. The YOLO family of object detectors have truly revolutionized the Computer Vision and object detection industry and have continuously evolved since its development. At present, the most recent variation of YOLO includes YOLOv4 and YOLOv4 - Tiny. These models are acquired and pre-evaluated on the public CrowdHuman benchmark dataset. These algorithms mentioned are pre-trained on the CrowdHuman models and benchmarked at the preliminary stage. YOLOv4 and YOLOv4 – Tiny are trained on the CrowdHuman dataset for 4000 iterations and achieved a mean Average Precision of 78.21% at 25FPS and 55.59% 80FPS, respectively. The models are further fine-tuned on a Custom CCTV dataset and achieved significant precision improvements up to 88.08% at 25 FPS and 77.70% at 80FPS, respectively. The final evaluation justified YOLOv4 as the most feasible model for deployment.

ARTICLE HISTORY
Received: 2nd Nov 2020
Revised: 25th Nov 2020
Accepted: 16th Dec 2020

KEYWORDS
Deep Learning
Computer Vision
YOLOv4
CrowdHuman
FPS

INTRODUCTION

A glimpse at modern technological developments, we can infer that human-machine collaboration is progressively demanded across various applications, including manufacturing, healthcare, agriculture, and transportation, solely to integrate robotic advantages such as prowess and precision with human talent. However, human safety is an essential aspect to be considered for such collaboration to be employed. Present-day industrial sectors implement various guidelines and standards for the essential health and safety of the employees. Consequently, human-machine applications employed within the physical domain will be required to adhere to the established essential health and safety requirements [1]. Human detection and tracking are well-known problems that could not be accomplished through conventional sensors and systems. Computer Vision, however, made it possible to computerise human detection. Advancements in computer vision technology such as facial recognition and object classification have significantly contributed to object detection advancement and have since become a bedrock for various physical world applications, such as smart photo galleries, sports activity, video surveillance, autonomous vehicles and electronic devices [2], [3]. Contemporary object detectors progressively benefited from deep learning technologies, particularly convolutional neural networks, to improve detection accuracy in generic object detection. Detection systems are required to satisfy the attributes of high accuracy and real-time speed for it to be applicable in heavy-duty and high-risk applications such as smart factories, assembly lines, production lines, conventional robots and autonomous transportation. Detectors should give fact and entirely accurate identification with systems of limited computing power [4].

However, deep learning provides us with various unique advantages to further improve on the existing models. Several external factors can also be improved by emphasising the features that are essential for human detection. Thermal imaging can be considered as it helps emphasise objects purely of the energy emission of the recorded object and helps detection in hard weather conditions. It is because regular RGB cameras could only produce poor results, such as rain and fog or are not useful at all, such as in total darkness [5]. Improved Mask R-CNN [6] is a novel algorithm that combines pre-existing and pre-trained models to improve the models' feature extraction capabilities further, contributing to improved detection accuracy. You Only Look Once (YOLO) is a state-of-the-art algorithm that is the fastest to date [7]. It offers the best accuracy in scenarios with a significant number of people in a frame, and the computing time taken makes it a suitable algorithm in real-time applications and cost can be compromised in favour of

*CORRESPONDING AUTHOR | Anwar P.P. Abdul Majeed | ✉ amajeed@ump.edu.my

detection accuracy. Deep Learning still offers the potential to push Human Detection technology further. It is improbable that a single algorithm could be used for all cases of human detecti on. Therefore, the development of the algorithm should be catered toward the specific system to whereby requirements such as accuracy, detection speed, the computational cost could be optimised based on the necessity of the system. The present study would evaluate the proposed model and fine-tune the model for deployment in a factory at TT Vision Technologies, Penang, Malaysia, to accomplish the human presence detection task.

RELATED WORK

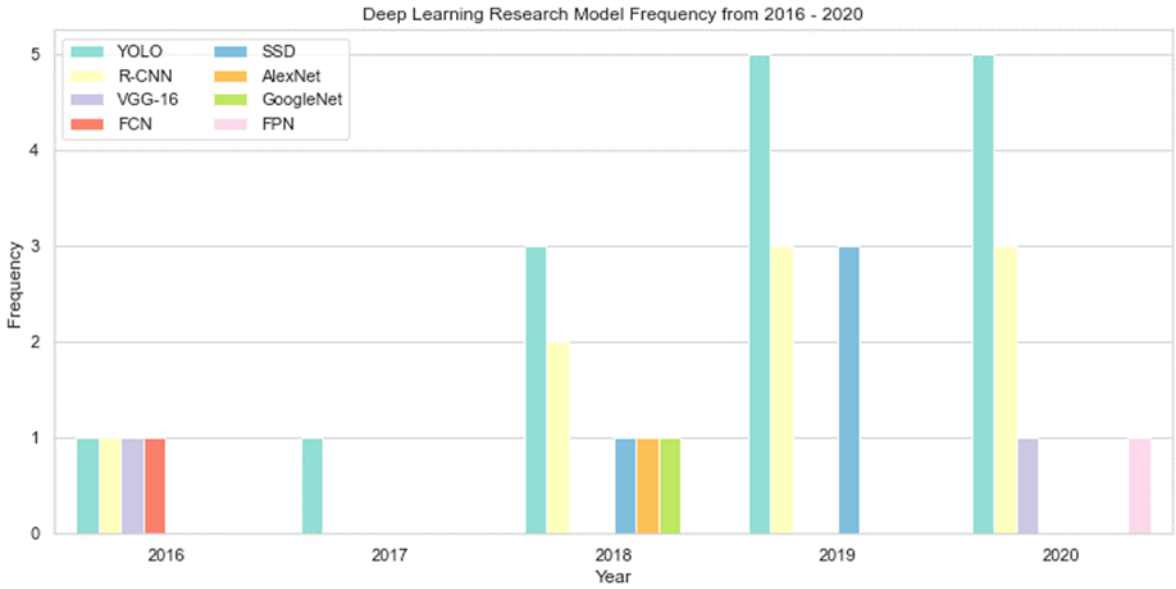

Figure 1. Deep Learning Research Model Frequency from 2016 - 2020

Through the literature reviewed, we can observe that the deep learning approach has gained popularity to accomplish person detection. Various models were used, but YOLO, Region Based Convolutional Neural Networks (R–CNN) and Single Shot Detector (SSD) stand out significantly. The models mentioned above are primary choices for object detection, which are constrained by the speed-accuracy tradeoff. R–CNN achieves high precision while YOLO offers fast inference speeds, and SSD falls between those other models. Moreover, various methods and techniques have been employed to complement the strength of each architecture to solve the person detection task. The application of deep learning technology has revolutionised human detection with state-off-the-art one stage detectors and multiple-stage detectors with their respective pros and cons. Most significantly, we can see the favourable results produced by the YOLO family one-stage detectors in real-time human detection. Based on the results obtained by [8]–[10], given the speed-accuracy tradeoff, YOLO seems to produce significant performance in terms of speed while still capable of delivering state-of-the-art accuracy. Ivašić-Kos approached their research by training the model on a similar high volume public dataset before fine-tuning on a custom and low volume dataset to maximise the model's generalizability and reduce overfitting when training on the custom dataset [5]. Joseph Redmon obtained improved results by manipulating the custom anchors relative to the dataset to further optimise the dataset for the YOLO model [8]. This research will evaluate the latest YOLOv4 family of detectors through pre-training on the CrowdHuman dataset and fine-tuning on a custom dataset.

METHODOLOGY

Research Flow

Generally, a Deep Learning-based object detector's typical development process consists of five phases: image data acquisition, object detection model selection, data pre-processing, model - training, and model evaluation. This research utilises a similar development method with added evaluation and refinement procedures. The methodology is distinguished into four (4) phases—data acquisition phase, preliminary benchmarking phase, fine-tuning, and finally, the evaluation phase.

The data acquisition phase involves acquiring the *CrowdHuman* public benchmark dataset and *Custom CCTV* dataset acquired from TT Vision Technologies Sdn. Bhd. The following phase entails the preliminary benchmarking of the proposed YOLOv4 models selected based on the reviewed literature considering aspects such as performance, development practicality and deployability. The preliminary benchmarking evaluates the YOLOv4 variants on the benchmark CrowdHuman dataset to ascertain that the performance meets up to the state–of–the art performance benchmarks.

The next phase is initiated once the performance of the YOLOv4 models is evaluated to be suitable to accomplish the person detection task. The fine-tuning phase is carried out based on research findings in [8], whereby the models' inferences are influenced by dataset environment, and change in environment may skew the inference results. This phase refines the TT Vision Technologies production environment models to ensure the models are suitable for deployment in TT Vision Technologies.

CrowdHuman Dataset

The CrowdHuman Dataset serves as a benchmark dataset to evaluate detection in crowded scenarios. It is a large, richly annotated dataset with 15,000 images which portrays 339,565 persons with a rate of 22.64 persons/image. Each person instance is annotated with a human full-body bounding box, head bounding box and human visible-region bounding box. This dataset is very suitable for training object detectors to detect people in crowded scenarios.

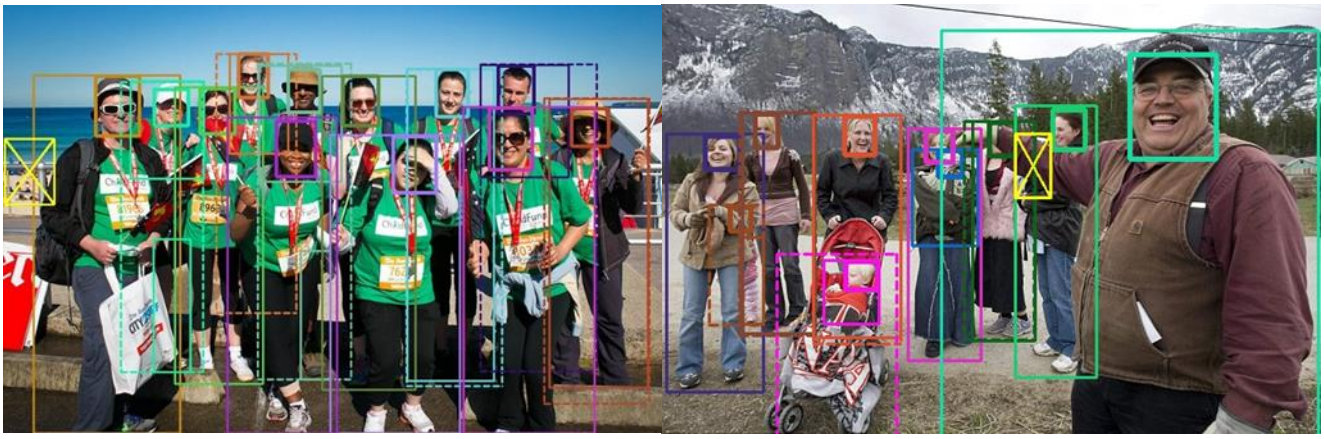

Figure 2. Sample images from CrowdHuman Dataset.

Custom CCTV Dataset

The custom CCTV dataset serves to fine-tune and evaluate the efficacy of the proposed models to perform real-time inference on live CCTV footage. The dataset consists of 1500 images acquired from the production areas at TT Vision Technologies Sdn. Bhd and 1500 augmented images. Each image consists of a rate of 2 persons/image. Each person instance are annotated with a human visible-region bounding – box and head bounding – box.

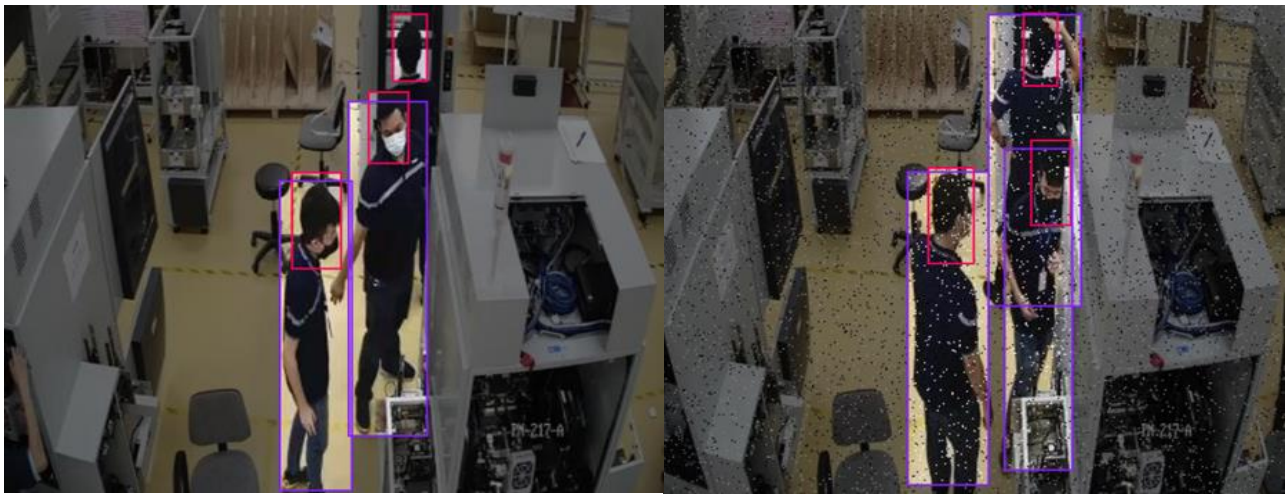

Figure 3. Sample images from Custom CCTV Dataset.

Data Pre - Processing

Data pre-processing is essential for data analytics. Data acquisition tends to be loosely controlled, which results in outliers and noise. Analysing such data will significantly penalise the accuracy of the pipeline. In the case of machine vision, we should consider occlusions and overexposure. Several data preprocessing methods will be applied to refine the dataset to ensure that the model does not expend unnecessary computing costs on unrequired data. The dataset has to be prepared relative to the proposed modes concerning the model's framework requirements and input size. The model requires the images to comply with its input size requirements. Therefore a 416 x 416 model will only accept images of size 416 x 416. Some image augmentation is required to consider the Darknet framework requirements of the label files required for each image in the dataset. Fortunately, the CrowdHuman dataset comes readily available with the annotation's files in odgt format. However, it is not compatible with the model and framework, which requires the annotations to be in YOLO text files. Some annotation conversion must convert the incompatible label format to .txt label format before the dataset is ready for training.

Performance Metrics: Precision

Precision is the proportion of the instances correctly predicted positive **(**+ve) by the model to all positive (+ve) instances as in Equation 1.

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{1}$$

Performance Metrics: Recall

As depicted in Equation 2, a recall is the proportion of the instances correctly predicted positive (+ve) by the model to all actual positive (+ve) instances.

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{2}$$

Performance Metrics: F1–Score

F1–Score represents the harmonic mean between the Precision and Recall. A higher F1–Score represents a more robust model as in Equation 3.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision\ +\ Recall} \tag{3}$$

Performance Metrics: Intersection over Union (IoU)

Intersection over Union (IoU) applied for accuracy evaluation in object detectors as shown in Equation 4. It is a commonly used evaluation metric for state-of-the-art object classifiers. Fundamentally, the concept involves a ground–truth bounding box, B^{gt}, which is hand labelled that specifies where our object is located in the image and the bounding box, B, predicted by the classifier model. Intersection over Union can be applied over the two sets of bounding boxes. The IoU equations are described as:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{4}$$

$$\mathcal{L}_{IoU} = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{5}$$

Deep Learning

Deep Learning is a subset of Machine Learning which is modelled after the way human brains filter information. It is metaphorically represented as learning from examples. It helps computers segment the input data through layers and performs predictions and classification accordingly. Deep Learning architectures are generally categorised into Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN) and Recursive Neural Networks (RvNN). The present study will entail the deployment of Deep Learning using convolutional neural networks for feature extraction.

Feature Extraction and Classification

Deep CNN is extensively applied in state-of-the-art object detectors. They are considered modern multi-layer perceptrons and have shown promising results in image classification and object detection. Advancements in computing hardware such as GPU and large datasets and large bandwidths have paved the way for the AI revolution and has allowed researchers to tinker around in the prospects of machine learning and further improve the capabilities of deep learning algorithms through deeper network models. Expanding from the traditional dual hidden layer CNN, it is now capable of numerous hidden layers.

The development of a CNN consists of several parts starting from the hidden layers responsible for feature extraction. The neural network performs a sequence of convolutions and pooling operations to detect and extract features. The next part would be the classification part. The fully connected outer layers will serve as a classifier above the extracted features. A probability is assigned to the object in the image depending on the classes predicted by the algorithm. State-of-the-art CNN utilises fully connected layers (classification), convolutional layers (feature extraction), pooling layer (reduction of dimensionality) and nonlinearity (sigmoid and ReLU).

YOLOv4 : Optimal Speed and Accuracy of Object Detection

YOLOv4 architecture consists of CSPDDarknet53 as the backbone, spatial pyramid pooling layer (SPP) and Path Aggregation Network (PAN) as the Neck and YOLOv3 (anchor-based) one stage detector as the head. YOLOv4 is developed to be more suitable for GPU training with additional improvements utilising data augmentation Mosaic and Self Adversarial Training (SAT). YOLOv4 was improved with optimal hyperparameter tuning and modification of the existing architecture for efficient training and detection. YOLOv4 managed to achieve state-of-the-art results with 65.7%AP on the MS COCO dataset at a real-time speed of 65FPS on a Tesla V100 GPU [9]. The model used is the YOLOv4, with an input size of 416.

YOLOv4 - Tiny : Improved Object Detection and Recognition

YOLOv4 Tiny is a variant of YOLOv4 with decreased convolutional layer depth designed for low-end Graphics processing unit (GPU) devices. The backbone utilises CSPOSANet with Printed circuit board (PCB) architecture. The inferences speed is dramatically increased to approximately 2800% faster than its predecessors at the time of its development. However, the increase in detection speed comes at the cost of decreased inference accuracy at 2/3 relative to the MS COCO dataset. YOLOv4 Tiny incorporates a pooling layer to reduce the figure for the convolution. It yields a 3-dimensional tensor prediction that comprises objectness score, bounding box and class predictions at multiple scales. The final layers ignore the bounding boxes with poor objectness scores and utilise convolution layers and max pooling layer for feed forward arrangement of its architecture. The model used is the YOLOv4 Tiny with an input size of 416.

RESULTS AND DISCUSSION

Preliminary Benchmarking

Table 1. Summary of Preliminary Benchmarking

Model Specification		mAP(%) @ 0.5 IoU		FPS		
Models	Input Size	CrowdHuman	Pre – Custom	TensorFlow	Darknet	F1 - Score
YOLOv4	416	78.21	58.65	15	25	0.76
YOLOv4 - Tiny	416	55.59	54.45	80	80	0.63

Appraising the results respective to the CrowdHuman dataset, we can observe that the model with the highest mean average precision (mAP) is the YOLOv4 with 78.21%. The higher mAP can be attributed to the larger input size, enabling the model to detect refined features in the dataset. YOLOv4 – Tiny recorded an mAP of 55.59%. The lower mAP by YOLOv4 Tiny is due to its decreased network depth with reduced convolutional layers, which corresponds to fewer feature extraction layers. However, although the models performed well on the CrowdHuman dataset, the inference accuracy showed a significant decrease at 58.65% and 54.45%, respectively, when tested on the Custom CCTV dataset. This decrease can be attributed to the compatibility of the dataset to the use case. Moreover, YOLOv4 Tiny achieved the highest average FPS of 80 compared to 25 FPS by YOLOv4 on the Darknet framework.

The factor of speed-accuracy tradeoff obtained suggests that the results are feasible. The conclusive factor to consider here would be the Precision representing the proportion of Positives that are True Positives. YOLOv4 achieved the highest Precision at 82% and Recall of 70%, indicating that the model correctly predicts a person correctly 70% of

the time compared to 53% by YOLOv4 Tiny. YOLOv4 achieved an F1–Score of 0.6 which is higher when compared to 0.63 by YOLOv4 Tiny. We can conclude that the most suitable model thus far would be the YOLOv4 given the speed-accuracy, which achieves up to 78.21% (CrowdHuman) and 58.65% (Custom CCTV) mAP with close to real-time inference speed of 25 FPS on Darknet. The results only partially satisfy the research objective and scope, which is to develop different deep learning models for person detection with high accuracy and fast inference speeds. However, higher mAP is desirable. Bextens' research findings showed that the models' mAP could be further improved by fine-tuning the models to the local environment [1]. Therefore, the next subsection will entail the findings of the fine-tuning process.

Fine – Tuning

Table 2. Summary of Fine – Tuning

Model Specification		mAP(%) @ 0.5 IoU		FPS		
Models	Input Size	Pre – Csutom	Post – Custom	TensorFlow	Darknet	F1 - Score
YOLOv4	416	58.65	80.08	15	25	0.78
YOLOv4 - Tiny	416	54.45	77.70	80	80	0.73

Table 2 illustrates the obtained results, and the fine-tuning procedure has successfully improved the performance of the models respective to the standard of the architecture. The YOLOv4 and YOLOv4 Tiny achieved mAP of 88.08% and 77.70%, respectively. The preliminary benchmark is comparable with the YOLOv4 achieving the highest mAP and YOLOv4 – Tiny achieving the highest inference speed. However, in terms of practicality and considering the speed-accuracy tradeoff, the most suitable model would be the YOLOv4 which achieves a good mAP of 88.08% and close to real-time inference speed of 25 FPS. YOLOv4 achieved a Precision of 70%, which falls compared to the other variant but makes up for it with a Recall of 88%, which indicates that generally, the model is 88% likely to classify a person correctly. Finally, on average, YOLOv4 achieved a F1–Score of 78%, which leads to the YOLOv4 Tiny variant by 0.05.

Summary and Final Evaluation

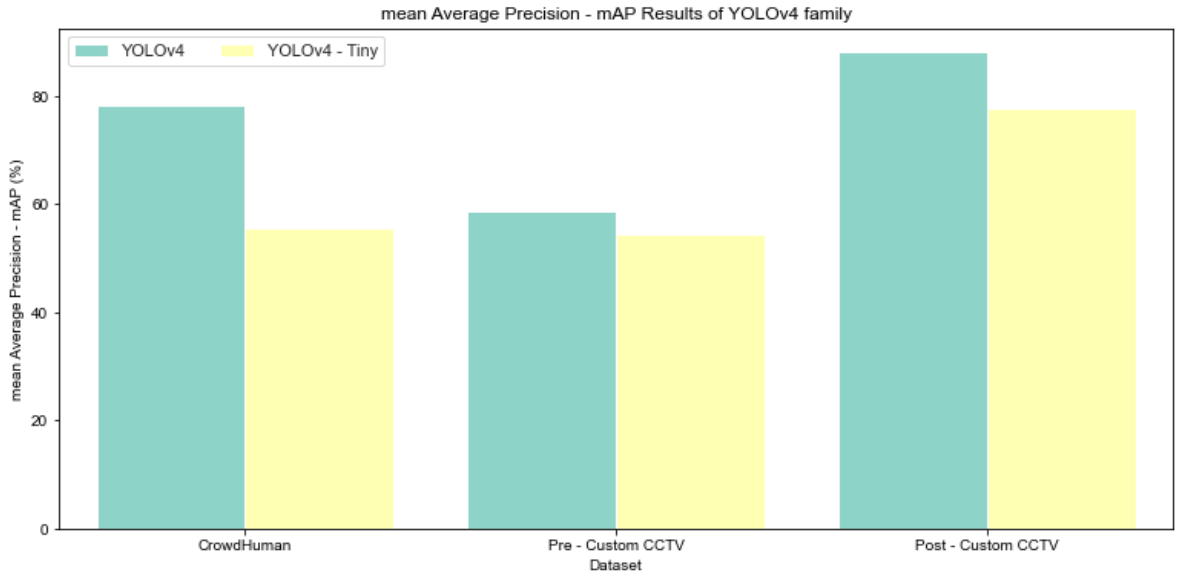

Figure 4. Summary of YOLOv4 model performance.

To sum up the evaluation of the YOLOv4 variations, it is apparent that the YOLOv4 is the most suitable model. Although YOLOv4 did not match up to the inference speed of the YOLOv4 Tiny variation, it achieved satisfactory and feasible performance with an mAP of 88.08%, F1-Score of 0.78 and inference speed of 25 FPS on the Custom CCTV dataset, which satisfies the requirement for real-time inference. YOLOv4 Tiny achieved the fastest inference speed in all the evaluations by sacrificing feature extraction and downscaling the architecture, which placed a big hit on the Precision of the model. In conclusion, the YOLOv4 variant is the most practical choice in contrast to the YOLOv4 Tiny for real-time person detection.

CONCLUSION

The research involves benchmarking the proposed YOLOv4 variants by pre-training the models on the benchmark *CrowdHuman* public dataset. The benchmarking evaluates the Precision of the YOLO models on person detection and the inference speeds on TensorFlow and Darknet frameworks. The Precision of the models is evaluated based on the mAP performance metric, and the inference speed is evaluated based on the inference FPS on the different frameworks. The research produced three (2) different YOLOv4 variant models namely, YOLOv4 and YOLOv4 Tiny. The models mentioned achieved satisfactory results for the person detection use case in industrial environments with mAP of 88.08% and 77.70%, respectively. The models achieved F1-Scores of 0.78 and 0.73 and respectively. In terms of inference speeds, YOLOv4 achieved 25FPS while YOLOv4 Tiny achieved 80FPS on the GTX1650 GPU. The final robustness evaluation considers the speed-accuracy tradeoff of deep learning models. Therefore the most suitable model should run inferences with high Precision and close to real-time inference speed. The YOLOv4 variant that meets this criterion is YOLOv4, capable of running inference with a precision of 88.08% and 25 FPS on Darknet.

REFERENCES

[1] S. Bexten, C. Walter, J. Scholle, and N. Elkmann, "Discussion of using Machine Learning for Safety Purposes in Human Detection," pp. 1587–1593, 2020.

[2] J. A. M. Jizat, A. F. A. Nasir, A. P. P. A. Majeed, and E. Yuen, "Effect of Image Compression using Fast Fourier Transformation and Discrete Wavelet Transformation on Transfer Learning Wafer Defect Image Classification," *MEKATRONIKA*, vol. 2, no. 1, pp. 16–22, Jun. 2020.

[3] M. N. A. Shapiee, M. A. R. Ibrahim, M. A. M. Razman, M. A. Abdullah, R. M. Musa, and A. P. P. Abdul Majeed, "The Classification of Skateboarding Tricks by Means of the Integration of Transfer Learning and Machine Learning Models," *Lect. Notes Electr. Eng.*, vol. 678, pp. 219–226, 2020.

[4] A. Angelova *et al.*, "Real-Time Pedestrian Detection With Deep Network Cascades," pp. 1–12.

[5] M. Ivašić and M. Pobar, "Human detection in thermal imaging using YOLO," pp. 2–7.

[6] Y. Wang, J. Wu, and H. Li, "Human Detection Based on Improved Mask R-CNN," 2020.

[7] S. Ghosh, "Comparative Analysis and Implementation of Different Human Detection Techniques," pp. 443–447, 2020.

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

[9] C. Wang and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection."

[10] S. S. Sumit, J. Watada, A. Roy, and D. R. A. Rambli, "In object detection deep learning methods , YOLO shows supremum to Mask R-CNN In object detection deep learning methods , YOLO shows supremum to Mask R-CNN," pp. 0–8, 2020.