報告題名：

鞋類分割的分析 YOLOv3 在不同的背景圖像上

An Analysis on Footwear Segmentation using YOLOv3 on Different Background Images

作者：許竣傑 Jun-Jie Xu、陳冠霖 Guan-Lin Chen

系級：電子碩一

學號：M0822886、M0807294

開課老師：梁詩婷

課程名稱：計算機視覺

開課系所：電子工程學系碩士班

開課學年：108　學年度 第 1 學期

中文摘要

機器學習技術已經廣泛應用於多種計算機視覺中，具備標準一致、高精度的優勢。本文旨在開發一種通過自動提取有意義的信息的方法，來對鞋類進行分割圖像。

文中提出已鞋類為對象的精確識別位置、標記類型的演算法。透過 YOLOv3 的網路框架實現，主要用於訓練鞋的特徵並驗證圖中未顯示的區塊。為證明有效性，採用兩不相同數據庫，共包含 700 個鞋類圖像。特別的是數據庫採用由乾淨背景及混雜背景的圖像組成的兩個數據庫，試圖已複雜的數據集模擬真實情境，使模型對於真實環境的答辯度提升。

在圖像上進行測試時，平均準確度分別為乾淨背景 95%和復雜背景 68%。另一方面，為使複雜背景的準確率上升，本實驗藉由乾淨背景數據集的權重轉移，來增強鞋的特徵。將來，這種方法可以進一步擴展，用作訓練其他數據集的基準，並結合更多樣的網絡體系、結構類型。

關鍵字：yolov3、物件偵測，深度學習

Abstract

Machine learning technology has been widely used in a variety of computer vision applications, with the ability to produce consistent and highly accurate performance. This article aims to develop an automatic method for segmenting images of footwear by suggesting ways to extract meaningful information from it.

The proposed algorithm generates the exact location of the footwear object and identifies the type of shoe. The method implemented in the experiment is called YOLOv3, which is mainly used to train the characteristics of shoes and verify invisible objects. To prove the validity, the two databases of the proposed method hold a total of 700 footwear images. In particular, the database consists of images with clean and complex backgrounds. In particular, complex datasets try to mimic reality, i.e.,It is more practical when implemented in practical applications.

When tested on clean and tidy images, the average accuracy was 95% clean background and 68% complex background, respectively. On the other hand, to improve the accuracy of complex data sets, the model must first be made to learn the features of a clean data set. In the future, this method can be extended to further train other benchmark databases and combine network architecture types.

Keyword：yolov3, object detection, deep learning

目錄

I. INTRODUCTION

Object detection is a process to identify and locate multiple objects in an image or video.Owing to its superior practical advantages, it has been successfully implemented on many applications, such as agriculture [1], medical fields [2], transportation [3] and others. Specifically, object detection predicts the position of the object through the bounding box, meanwhile, classifies the object's type in an image. In the era of artificial intelligence and the Internet of Things, the issues of realizing the detection applications in the real world and integrating with mobile phones have become interesting research directions nowadays.

There are some popular and publicly available object detection models proposed in the community. They can be categorized into the R-CNN (region-convolutional neural network)[4] family and the YOLO (you only look once)[5] family. The former approach includes R-CNN[4], Fast R-CNN [6], Mask R-CNN [7], Faster RCNN [8], whereas the latter incorporates YOLO[5], YOLOv2[9], YOLOv3[10]. In brief,R-CNN finds the region proposal of the object then performs a classification task. On the other hand, YOLOv3 utilizes a regression method to perform detection and classification simultaneously. Nevertheless, both families rely on deep learning architecture in the model training which requires a relatively large amount of calculations, high execution time and computational memory usage.

The example of exploiting YOLOv3 is to recognize the different styles of wagon numbers appearance under complex background with uneven lighting variations [11]. The proposed algorithm first identifies and crop the region of interest using a detector, then another detector is trained to recognize the object from the cropped region. Without the image preprocessing and character segmentation processes, the recognition rate achieved was 96% on more than 1000 images. As the identification of wagon numbers is demanding especially when applying on railway transportation, it may require manual efforts to rearrange the digits and rectify the errors.

Recent work is carried out by Benjdira et al [12] to detect an unmanned aerial vehicles (UAV) imagery dataset that collected from a drone, that consists of _300 images and _4000 instances of car objects. Both the Faster R-CNN and YOLOv3 approaches are tested on the images. The precision scores obtained were higher than 99%. However, the time taken for the Faster R-CNN is approximately 25 thousand times more than that of YOLOv3.

On the other hand, YOLOv3 has been applied to detect the occurrence of fire outbreaks over different forest areas in real-time [13]. A recognition rate of 83% is exhibited with the frame rate of up to 3.2 fps when evaluated on the video that has a resolution of 1920_1080. However, the experiments were tested on large-area forest fires, whereas the recognition accuracy for the small-scale forest fires is less

satisfactory.

Tian et al [14] employ YOLOv3 on agri-food system. Succinctly, they monitor the growth stages of different types of apples. The image of the object is taken in orchards and has fluctuating illumination,
complex background (may contain branches and leaves) and overlapping apples. Data augmentation is utilized as the preprocessing step to increase the diversity and amount of data. Then YOLOv3 is modified
by replacing the input size to allow the model to train on higher resolution images. The average detection
rate is more than 80%. The reason that causes inaccurate detection might due to partial occlusion of branches and leaves, as well as the overlapped apples.

Inspired by the promising detection performance achieved by adopting YOLOv3 architecture on the various application, this paper attempts to detect the position of the shoes on both clean and complex backgrounds. The three contributions of this article are briefly listed as follows:
Collection of a shoe database, that comprises about 700 different types of shoe mages with distinct backgrounds. All the images are publicly searchable online.
Adoption of state-of-the-arts pre-trained neural networks (YOLOv3) to extract discriminant features and perform the detection task.
Comprehensive experimentation on the dataset to verify the robustness of the algorithms evaluated. Both the qualitative and quantitative results are presented**.**

II. DATABASE COLLECTION

There are a total of 756 footwear images downloaded from the Internet, which contains an equal amount of images with clean and complex backgrounds, viz, 378 images each. The sample images of the dataset is illustrated in Figure 1 and Figure 2. The definition of clean background is there is a white background and sometimes there is some shadow beneath the footwear. In contrast, the complex background normally presents in the form of real-world exemplars, such as the existence of outdoor scenes behind the target object. Note that the spatial resolution of the footwear is standardized to 512 x512 pixels by performing cropping and resizing operation. An overview of the micro-expression datasets information that used in the experiment is shown in Table I. Note that, the footwear images in both the training and testing set are completely non-overlapping.

TABLE I
DETAILED INFORMATION OF THE FOOTWEAR DATABASES USED IN THE EXPERIMENT

	Description		
Resolution	512 × 512 pixels		
Bit depth	24 bits		
	Train set	Test set	Total
Number of clean background image	344	34	378
Number of complex background image	344	34	378
Total	688	68	756

Fig. 1. Example of the clean training set

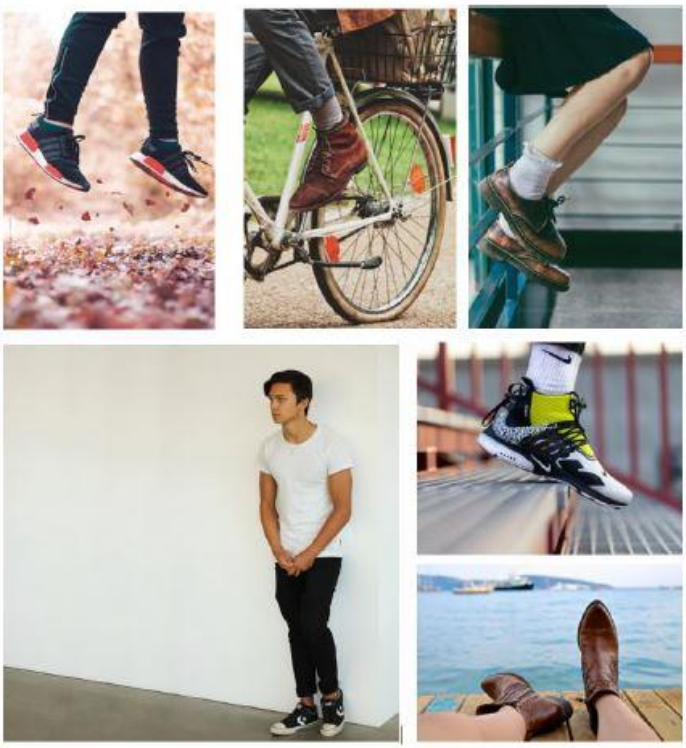

Fig. 2. Example of the complex training set

After preprocessing the footwear images, a ground-truth annotation of the region of footwear is performed.The labeling tool utilized in the experiment is Labelme, which allows us to annotate the bounding box regions of the target object in an image.

The sample annotated bounding box is illustrated in Figure 3 with a rectangle closed boundary. In brief, a bounding box is the smallest rectangle that contains the footwear. This tool is simple and easy to use where the user can use the mouse to draw a rectangle box to indicate the boundary of the target object. Specifically, to create the smallest rectangle that can enclose all the footwear context, the user just has to click on a point on the top left corner and drag to the bottom right corner. This annotation process is the only stage that requires high human effort as all the 788 images are needed to be labeled manually.

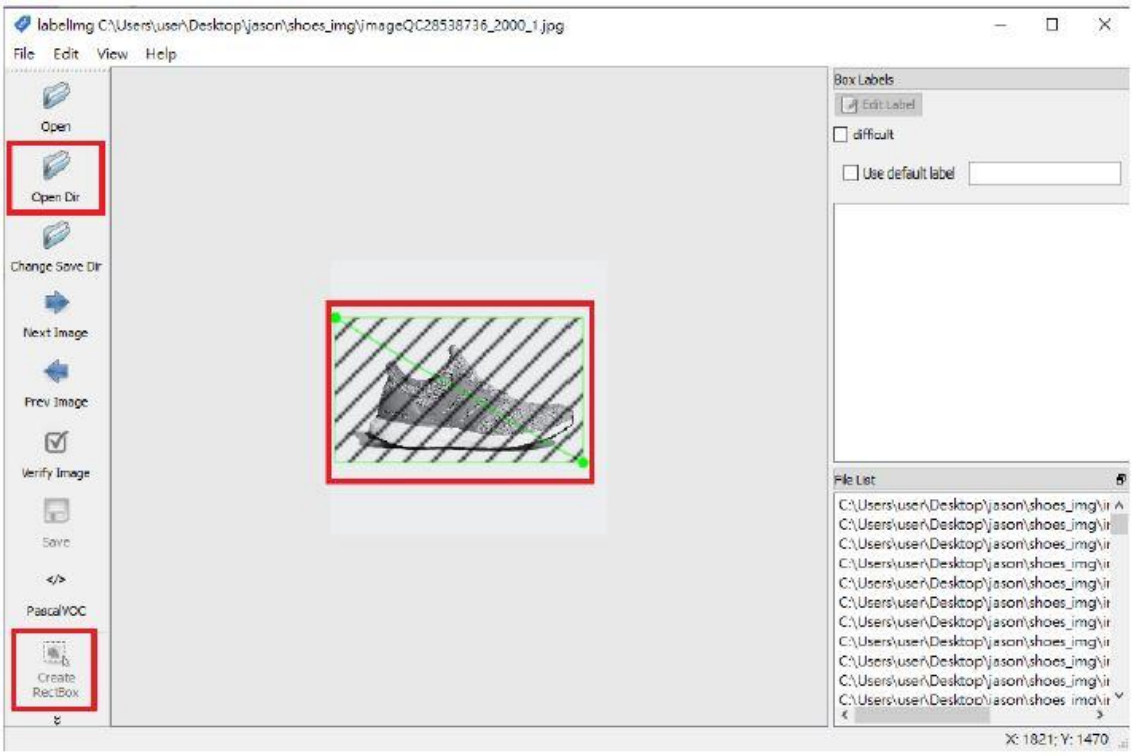

Fig. 3. Sample image to annotate the bounding box using Labelme software

III. PROPOSED METHOD

The purpose of this paper is to investigate the optimal neural network model that is capable to handle the footwear detection task with different kinds of background. In general, an object detection framework involves two basic steps, include: (1) Model training — identification of the important features from the image; (2) Model testing — localization and recognition of the object based on the features extracted. Figure 4 illustrates the basic flowchart of the recognition process. Succinctly, since there are two types of datasets collected, viz, clean and complex, four models are trained independently:

Model 1 - The model is trained on the clean dataset.

Model 2 - The model is trained on the clean dataset, then trained on the complex dataset.
Model 3 - The model is trained on the complex dataset.
Model 4 - The model is trained on the composite dataset that contains clean and complex data.
Note that, the object detection architecture used in the experiment is YOLOv3. In brief, the architecture divides the image into different regions, then predict the bounding box and probability of each region.

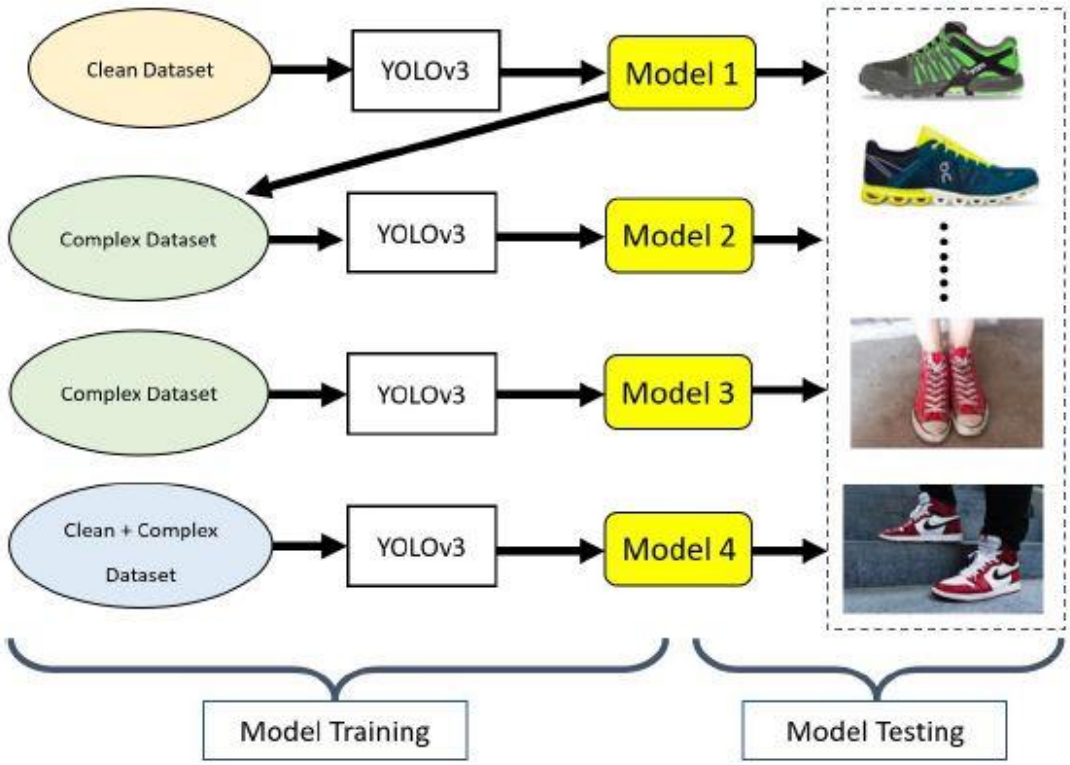

Fig. 4. Block diagram of the proposed footwear segmentation system

The backbone architecture of YOLOv3 is Darknet-53, which consists of 53 convolutional layers to capture deep features. The advantage of YOLOv3 is it enables end-to-end object detection and can directly feed the entire input image to the architecture. The trained model then predicts the coordinate position of the bounding box, which produces the confidence level of the prediction, as well as the category to which the object belongs. There are several studies demonstrate that YOLOv3 is robust to the complex background and thus generates a low false detection rate.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this experimental environment, the software platform was compiled in Python, and the hardware platform was the processor Intel Core (TM) i9-9900kf CPU @ 3.60GHz. The graphics card was NVIDIA GeForce RTX 2080 Ti. The configuration parameters of the YOLOv3 architecture are set to:

Learning rate = 0.0001

Epoch = [50, 300]

_ MiniBatchSize = 512

_ Backbone architecture = Darknet-53

The performance metric to evaluate this segmentation task is mAP. Conceptually, it mAP is derived from IoU, which indicates the ratio of the overlap of ground-truth to the segmented result:

$$IoU = \frac{\text{Ground-truth} \cap \text{Predicted}}{\text{Ground-truth} \cup \text{Predicted}} \quad (1)$$

Practically, there may exist more than one bounding box for each footwear image. When IoU is larger than 0.5, it denotes that the footwear is correctly localized (marked as TP), otherwise, it is a fail (denoted as FP). Consequently, since there's only one type of target object (i.e., the footwear), the mean average precision (mAP) for all the testing images can be computed as:

$$mAP = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

The proposed network models are tested on clean and complex datasets separately. Referring to Table I,the total number of the testing data in the clean dataset and complex dataset contains 34 images each.The performance results of evaluating the effectiveness of the four models are tabulated in summarized in Table II, Table III, and can be graphically visualized in Figure 5, where model 1 was trained on clean dataset; model 2 was trained on the clean dataset, then trained on the complex dataset; model 3 was trained on the complex dataset, and; model 4 was trained on the clean + complex datasets. Besides, Figure 6 depicts the overall performance of the four proposed models when tested on both datasets.

From Figure 5(a), it is noticed that model 1 produces the best result (mAP > 90%) when tested the images with a clean background for all the epochs. This is because model 1 has been optimized the features training by capturing the footwear characteristics. Model 4 ranked the second (mAP =_ 85%), as half of the training dataset consists of images with a clean background. In contrast, model 2 and model 3 exhibit poor performance, mAP < 40%. This may due to the weights and biases in the YOLOv3 network that have been confused by the complex background, conduced by

the complex dataset. Some of the visualizations that successfully detected the footwear is illustrated in Figure 7.

On the contrary, the segmentation performance when tested on complex datasets behaves completely differently from that of the clean dataset. The result is shown in Figure 5(b). Model 1, exhibited to the highest result on clean data, produces the worst performance (mAP < 30%) when testing on the complex dataset. This is because the network architecture does not know the complicated background. The model that generates the best result is model 2 (mAP =_ 65%). This is due to model 2 first learns the features of the footwear, then the complex background is introduced to the network for further network training. Model 3 and model 4 produce the average result of mAP=50%, which is significantly better than that of model 1. Figure 8 shows that the algorithm is capable to localize the footwear for the unseen image with complex background.

TABLE II
SEGMENTATION PERFORMANCE OF THE PROPOSED ALGORITHM WHEN TESTED ON THE IMAGES ON CLEAN DATASET

Epoch	Model 1	Model 2	Model 3	Model 4
50	0.9301	**0.3677**	0.1965	**0.8835**
100	0.9296	0.3671	0.2224	0.8309
150	0.9322	0.3652	**0.2241**	0.8142
200	0.9433	0.3654	0.2028	0.8433
250	0.9328	0.3644	0.2182	0.833
300	**0.9497**	0.3491	0.218	0.848
Max	**0.9497**	**0.3677**	**0.2241**	**0.8835**

TABLE III
SEGMENTATION PERFORMANCE OF THE PROPOSED ALGORITHM WHEN TESTED ON THE IMAGES ON COMPLEX DATASET

Epoch	Model 1	Model 2	Model 3	Model 4
50	**0.2921**	**0.675**	0.4729	**0.5544**
100	0.2377	0.6392	0.4848	0.5094
150	0.2442	0.6295	0.4601	0.4165
200	0.272	0.6008	**0.4903**	0.4089
250	0.2914	0.6091	0.4721	0.3875
300	0.2814	0.626	0.4674	0.3547
Max	**0.2921**	**0.675**	**0.4903**	**0.5544**

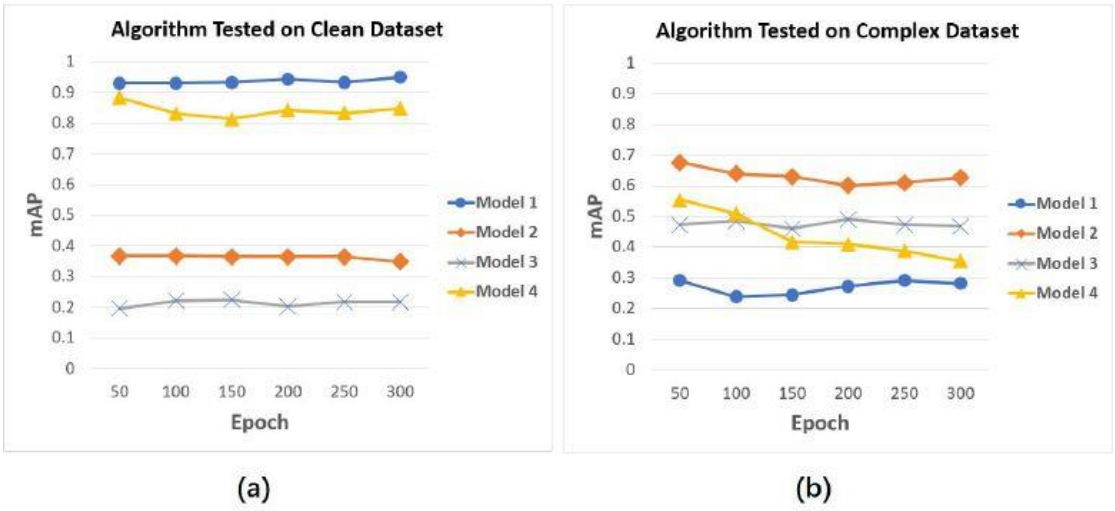

Fig. 5. Segmentation performance of the proposed algorithm when tested on the images in: (a) clean dataset, and;(b) complex dataset.

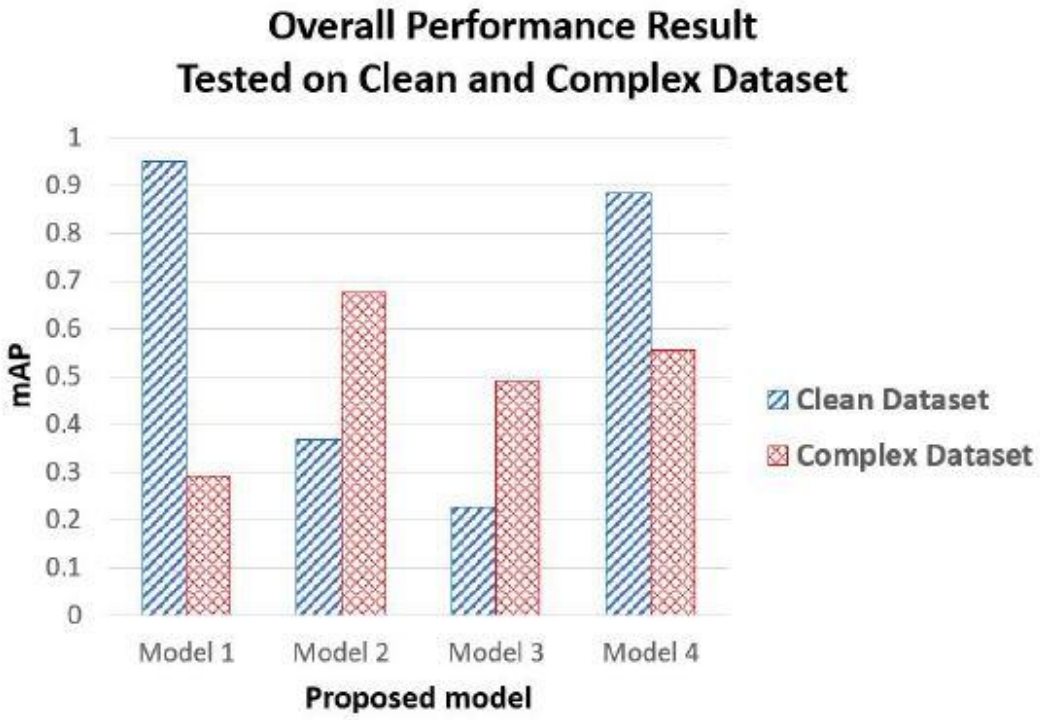

Fig. 6. Overall performance of the proposed models when tested on the images in clean and complex datasets

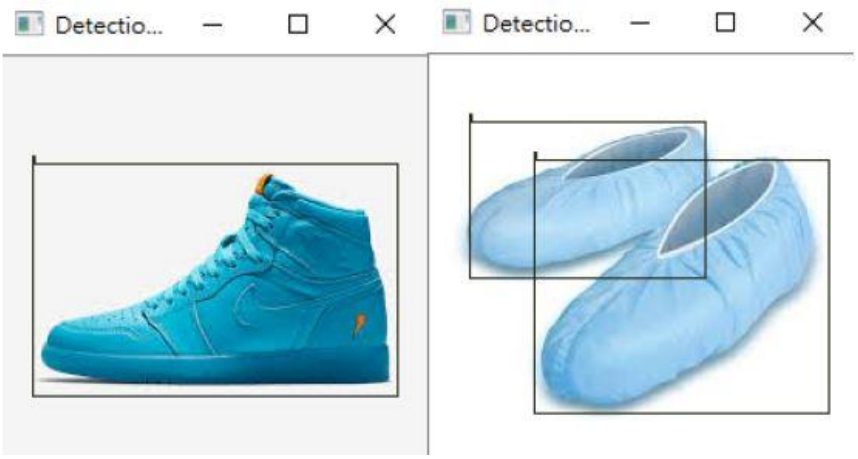

Fig. 7. The visualization of the output when testing on the image in clean dataset.

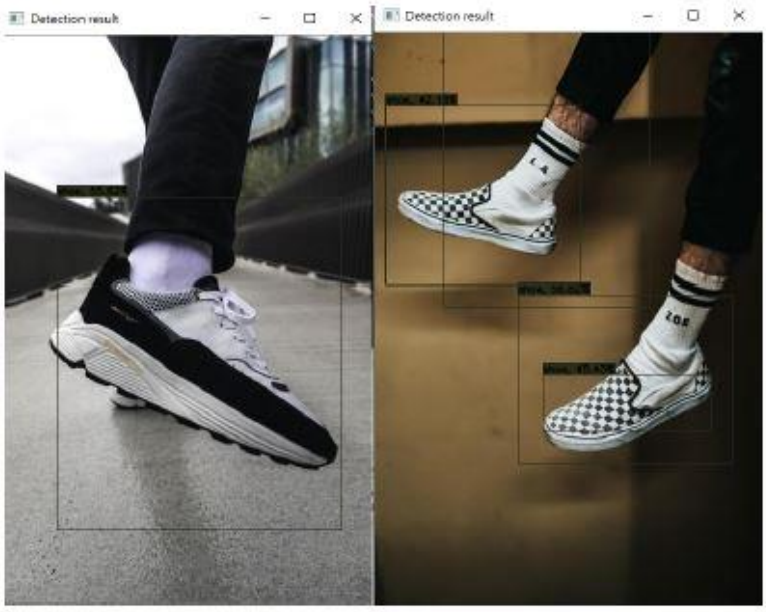

Fig. 8. The visualization of the output when testing on the image in complex dataset.

V. CONCLUSION

This paper provides a comprehensive analysis to investigate footwear detection and localization when it is placed on a variety of different backgrounds. The segmentation approach employed is YOLOv3, with the backbone network of Darknet-53. There are two types of the dataset collected, with the image background of clean and complex. Particularly, complex dataset attempts to mimic the real-world condition, which is more practical when implementing in real-life applications. The performance results demonstrate that the YOLOv3 model of training solely the clean

dataset and test on unseen clean background image produces more than 90% of the mAP. On the other hand, to tackle the complex dataset, the model that first learns the features of the clean dataset then the complex dataset is more satisfactory. In the future, this method can be extended to further train other benchmark databases and combining with types of network architecture.

參考文獻

[1] H. Yang, L. Chen, M. Chen, Z. Ma, F. Deng, M. Li, and X. Li, “Tender tea shoots recognition and positioning for picking robot using
improved yolo-v3 model,” IEEE Access, vol. 7, pp. 180 998–181 011, 2019.
[2] A. Lemay, “Kidney recognition in ct using yolov3,” arXiv preprint arXiv:1910.01268, 2019.
[3] A. Rasouli and J. K. Tsotsos, “Joint attention in driver-pedestrian interaction: from theory to practice,” arXiv preprint arXiv:1802.02522,
2018.
[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,”
in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the
IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[6] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[7] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision,
2017, pp. 2961–2969.
[8] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances
in neural information processing systems, 2015, pp. 91–99.
[9] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in Proceedings of the IEEE conference on computer vision and pattern
recognition, 2017, pp. 7263–7271.
[10] ——, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.
[11] Z. Liu, Z. Wang, and Y. Xing, “Wagon number recognition based on the yolov3 detector,” in 2019 IEEE 2nd International Conference
on Computer and Communication Engineering Technology (CCET). IEEE, 2019, pp. 159–163.
[12] B. Benjdira, T. Khursheed, A. Koubaa, A. Ammar, and K. Ouni, “Car detection using unmanned aerial vehicles: Comparison between
faster r-cnn and yolov3,” in 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS). IEEE, 2019, pp. 1–6.

[13] Z. Jiao, Y. Zhang, J. Xin, L. Mu, Y. Yi, H. Liu, and D. Liu, “A deep learning based forest fire detection approach using uav and
yolov3,” in 2019 1st International Conference on Industrial Artificial Intelligence (IAI). IEEE, 2019, pp. 1–5.
[14] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, “Apple detection during different growth stages in orchards using the
improved yolo-v3 model,” Computers and electronics in agriculture, vol. 157, pp. 417–426, 2019.