# Comparison of YOLO v3, Faster R-CNN, and SSD for Real-Time Pill Identification

**Lu Tan**
The Third Affiliated Hospital of Southern Medical University

**Tianran Huangfu**
The Third Affiliated Hospital of Southern Medical University

**Liyao Wu**
The Third Affiliated Hospital of Southern Medical University

**Wenying Chen** ( ✉ chenwenying2016@163.com )
The Third Affiliated Hospital of Southern Medical University

**Research Article**

**Keywords:** convolutional neural network, Faster R-CNN, YOLO v3, SSD, pill identification

**Posted Date:** July 30th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-668895/v1

# Comparison of YOLO v3, Faster R-CNN, and SSD for Real-Time Pill Identification

Lu Tan[1], Tianran Huangfu[1], Liyao Wu[1], Wenying Chen[1*]

[1]Department of Pharmacy, The Third Affiliated Hospital of Southern Medical University, Guangzhou 510000,China.

[*]Correspondence: TEL: (020)62784810; chenwenying2016@163.com

# Comparison of YOLO v3, Faster R-CNN, and SSD for Real-Time Pill Identification

## Abstract

**Background:** The correct identification of pills is very important to ensure the safe administration of drugs to patients. We used three currently mainstream object detection models, respectively Faster R-CNN, Single Shot Multi-Box Detector (SSD), and You Only Look Once v3(YOLO v3), to identify pills and compare the associated performance.

**Methods:** In this paper, we introduce the basic principles of three object detection models. We trained each algorithm on a pill image dataset and analyzed the performance of the three models to determine the best pill recognition model. Finally, these models are then used to detect difficult samples and compare the results.

**Results:** The mean average precision (MAP) of Faster R-CNN reached 87.69% but YOLO v3 had a significant advantage in detection speed where the frames per second (FPS) was more than eight times than that of Faster R-CNN. This means that YOLO v3 can operate in real time with a high MAP of 80.17%. The YOLO v3 algorithm also performed better in the comparison of difficult sample detection results. In contrast, SSD did not achieve the highest score in terms of MAP or FPS.

**Conclusion:** Our study shows that YOLO v3 has advantages in detection speed while maintaining certain MAP and thus can be applied for real-time pill identification in a hospital pharmacy environment.

**Keywords:** convolutional neural network; Faster R-CNN; YOLO v3; SSD; pill identification

## Introduction

In China, most hospitals are required by existing medical insurance policies to control costs and usually sell pills in separate packages, whereas oral pills for inpatients are dispensed individually by the inpatient pharmacy according to the prescribed dosage. These cases usually require unpacking the pills from their original labeled containers. However, in contrast to management systems in countries such as the United States and Japan, the China Food and Drug Administration (CFDA) does not mandate that pills have an imprint code. Therefore, as some of the solid oral dosage forms may not be clearly distinguishable from each other in terms of size, shape, or color, when the packaging is removed it may be difficult for hospital pharmacists to distinguish between pills. Similar looking pills that cannot be identified must be discarded, which results in a waste of medical resources. Solving this problem requires not only long-term knowledge and experience of pharmacists but also sufficient focus on their work. However, with China's growing and aging population, the demand for medical care is gradually increasing**Error! Reference source not found.**, and more patients require more inpatient pharmacies, which places considerable pressure on the limited medical resources[1]. In most primary care hospitals, many pharmacists still dispense drugs and check them manually. Although some large hospitals have now adopted the expensive Automatic Tablet Dispensing Machine, it seems that filling errors, accidental dropping of medication into the machine, and other human errors remain unavoidable[2][3]. 'Err is Human'[4], and even experienced pharmacists can make mistakes under constant high intensity work. Dispensing the wrong drug will seriously compromise the safety of treatment[6][7].

Wit the phenomenal development of machine learning in recent years, machine learning has been widely applied to computer vision, medical image processing, and many other fields. Some progress has been made in drug discovery[8], drug production[9] and semiquantification[10], but very little research has been done on pill identification. As sophisticated algorithms continue to emerge, it seems likely that it will be possible to apply image processing research to pill identification. The accuracy of the model is the basic indicator that determines whether this technology can assist the pharmacist's work. In addition, the efficiency of the model is also important. If the model calculation takes too long, it will be difficult to play a practical role in busy work. To investigate this possibilty, we trained current mainstream object recognition algorithms, including Faster R-CNN, Single Shot Multi-Box Detector (SSD), and You Only Look Once (YOLO v3), on a newly created pill dataset and compared the results in terms of accuracy and detection rate, to determine the best pill identification strategies to assist pharmacists and other healthcare workers dispense and check drugs affordably, ultimately to better protect the lives of patients.

## Related Work

Early related research was mainly based on traditional machine learning. Lee et al. proposed a Canny edge detection and invariant moments method to extract the feature vector from pill imprint images[11]. Morimoto et al. used images captured from both-sides of tablets to identify them by matching distinctive marks[12]. Suntronsuk et al. used Otsu's thresholding with noise elimination to extract the imprint from pills as a vector, achieving precision and recall scores on the recognition of text on imprints of over 57%[13]. Neto et al. proposed a feature extractor based on shape and color in 1,000 images of 100 different classes of pills, obtaining an accuracy of over 99% using various classifiers[14]. Dhivya et al. used a support vector machine to recognize text imprinted on tablets[15].

Traditional machine learning methods achieve the detection of targets by manually designing feature learning methods, and the characteristics of the feature extraction design and classifier selection often largely determine the final detection accuracy. Hence, the corresponding characteristic parameters need to be set manually for different tablets. However, because of China's Centralized Drug Bidding and Purchase Mechanism, the same drug will be centrally tendered each year, which means that it may be supplied by different pharmaceutical companies. Therefore, due to the annual variation in the types of pills chosen, a manual approach to feature design generates a significant amount of work. This approach may lack robustness to the diversity of the pills and cannot handle large volumes. In particular, when there is no imprint code, the similar appearance of pills and the lack of the corresponding parameters can degrade recognition accuracy. Also, the traditional object detection approach uses a computationally intensive sliding window method, which makes it difficult to achieve real-time performance. Therefore, an improved solution is desirable.

Convolutional neural networks (CNN) are the most common deep learning algorithm, applying multiple convolutional layers and convolutional computation. They have efficient feature extraction capability and provide a better problem-solving method for object detection. Wong et al. used the improved AlexNet-based algorithm, which won the ILSVRC 2012 championship, and compared it with two traditional machine learning methods, k-nearest neighbors and random forests, for pill feature extraction, ultimately demonstrating the superiority of AlexNet. The results showed that the top-1 pill recognition by the AlexNet-based network performed better than those with manually designed features, reaching 95.35%[16]. However, AlexNet, as a light network with only a few layers, can only implement simple applications, and as the complexity of the task increases, it is not flexible enough to train a robust neural network for this task. Swastika et al. proposed using three LeNet or AlexNet models to extract the three main features of pills, shape, color, and imprint, and combine three CNNs into an integrated network for pill identification. The network was trained on 24,000 images of eight types of pill, achieving a recognition accuracy of up to 99.16%[17]. Ou et al. proposed a drug pill detection system similar to a two-stage target detection algorithm based on ResNet for localization detection and Xception for classification. The training set included 131 categories and a total of 1,680 images for training. The top-1 accuracy rate for the trained network was up to 79.4%[18]. Based on these studies, deep learning has gradually replaced manual design extraction in pill feature extraction, and deep learning algorithms, such as LeNet, AlexNet, and ResNet, are able to address the problem of pill image classification. The CNNs used for target detection, such as Faster R-CNN, SSD, and YOLO architectures, incorporate the structure of the above-mentioned CNNs used for image classification, and can accomplish both image classification and target localization, but they have not been applied to pill identification. In addition, in practical applications, especially in places with high workloads such as pharmacies, there is a need to consider accuracy while also focusing on preforming the task in real-time, and there are no relevant studies focusing on real-time pill identification.

## Object Recognition Technology based on Deep Learning

Current approaches using deep learning methods for target classification and regression can be divided into two categories. One is the two-stage algorithm represented by architectures such as R-CNN, Fast R-CNN, and Faster R-CNN. This type of algorithm is usually carried out in two steps. First use selective search or Region Proposal Net (RPN) to generate Region Proposal, and then complete classification and regression on Region Proposal. This method has high accuracy but also limits the detection speed. Another algorithm is the one stage algorithm represented by SDD, YOLO, etc. This class is a regression-based end-to-end object detection and recognition algorithm that uses a single network to predict the object boundary box and category probability score directly from the image. As this algorithm does not use RPN, the detection speed is improved. However, the detection rate for small targets is not as good as the two-stage-algorithm. The detection accuracy and detection speed of the model directly affect the feasibility of pill recognition.

Faster R-CNN is an object detection algorithm proposed by Ren et al. in 2015**Error! Reference source not found.**, consisting of four parts: feature extraction network, region proposal network, ROI Pooling, and a fully connected layer. The overall detection process is shown in Figure 1. Faster R-CNN is a modified version of R-CNN and Fast R-CNN algorithms. The difference between the two is that the Faster R-CNN algorithm avoids the computationally expensive selective search algorithm and uses the RPN to generate candidate regions instead. This algorithm calculates the features of the whole image at once and thus does not involve repeated calculations, which greatly improves the detection speed of Faster R-CNN.

SSD[19] was proposed by Wei Liu et al. and draws on the anchor mechanism of Faster R-CNN and the end-to-end one-step structure of the YOLO algorithm in which object classification and location regression are performed directly in the convolution stage. The main network of the SSD algorithm is shown in Figure 2. SSD uses the VGG-16 network as a backbone and modifies it by replacing the last two fully connected layers with convolutional layers while also adding another four convolutional layers later to finally form the feature extraction network as Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2, whose sizes are (38, 38), (19, 19), (10, 10), (5, 5), (3, 3), and (1, 1), respectively. SSD is trained to obtain a set of fixed-sized bounding boxes and the class prediction scores of the targets in the bounding boxes. Then, redundant bounding boxes are filtered out and the final detection results are generated by the non-maximum suppression (NMS) algorithm, which has good results both in terms of speed and accuracy of detection.

YOLO[20] proposes a new idea for target detection by transforming the task into a regression problem. The whole framework only needs to use a relatively simple structure of CNN to directly complete the regression of target detection to predict the position of the bounding box and the class of the candidate box. The YOLO v3[21] backbone network structure does not have the pooling and fully connected layers, as shown in Figure 3, and the convolutional transformation of the image is achieved by changing the step size of the convolutional core. YOLO v3 uses Darknet-53 as the network skeleton, which makes the network structure deeper and better at extracting features, as demonstrated by its improved accuracy compared with YOLO v1 and YOLO v2. Darknet-53 makes extensive use of the ResNet residual structure, which can avoid the vanishing gradient problem even when the network structure is deep.

## Methods

### Dataset Preparation

The training of deep learning models typically requires many data samples to obtain reliable parameters and models. In 2016, the U.S. National Library of Medicine published an algorithm challenge competition on pill recognition, and publicly released the pill image dataset[23]. However, considering our particular situation in which there are some kinds of pills without an imprint code, this dataset was not considered suitable. Therefore, we decided to create our own dataset for use in this experiment.

The appearance of our existing oral solid dosage forms was analyzed by observation, and images were taken using a 12MP high-speed photographic apparatus connected to a computer. The pills were placed at a random location on the shooting board. Since the height of the high-speed photographic apparatus is fixed, the distance of each pill shot is also relatively fixed. Each pill shot includes both front and back images, for a total of 5,131 images. The statistics of the dosage form, printing, shape, color, and manufacturer of the pills are shown in Table 1. There are a total of 261 varieties of oral solid drugs commonly used in inpatient pharmacy, including 70 capsules and 191 tablets. We observed that some pills have a special code, manufacturer's trademark image, or several of them were printed at the same time after removing the packaging, which aids identification. However, there are still some tablets that are difficult to distinguish after removing the outer packaging. Representative images of the tablets are shown in Figure 4.

## Object image annotation

Since the object recognition method used in this experiment is a type of supervised learning, it is necessary to obtain the labeling information of the pill to be detected in the image; this includes the pill category information and the pill border location information. LabelImg is written in Python. Since the labeling format of LabelImg is consistent with PASCAL VOC and has a good graphical interactive interface with a rich array of shortcut keys, it was used to improve the labeling efficiency in our experiment. The image annotation process is shown in Figure 5. After labeling the tablets with LabelImg, the information of each image is saved in an "xml" file with the same name. The xml file contains all the information needed for training the network, including the class of the object and the location of the object in the image.

## Training Models

The experimental platform configuration for this paper is: OS: Win10, GPU: NVIDIA GeForce GTX 1080Ti, CPU: Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz. The experimental platform is built based on the Python programming language and the pytorch framework.All three models were trained on this configuration. The specific parameters are shown in Table 2.

Evaluation Indicators

To compare the results of the three different deep learning-based models for pill classification, we applied a range of standard metrics commonly used to evaluate machine learning models. There are four possible outcomes based on the output categories of the test samples compared with the categories of the true labels, as follows: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). If the target type is detected correctly, the center coordinates of the detection frame and the dimensions of the detection frame are within tolerable limits, then the detection result is recorded as TP. FP refers to a target category recognition error or the detection frame is not within the preset threshold. The predicted result of a target that is not detected is recorded as FN. As we did not predict the absence of a pill, the category of TN was not used.

The observed counts are combined into standard metrics including recall, precision, F1 score (F1), mean average precision (MAP), and frames per second (FPS). In the process of target detection, precision is the ratio of correctly detected targets to the number of all detected targets; recall is the ratio of the number of detected targets to all targets in the sample set. The definition of precision and recall are shown in Formulas 1 and 2, respectively:

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

$$\mathrm{Re}call = \frac{TP}{TP + FN} \qquad (2)$$

F1 is the weighted harmonic average of precision and recall. Since the amount of data for each pill is not equal, the F1 score is used to evaluate performance. The F1 score can be calculated from the precision and recall rates, as defined in Formula 3:

$$F_1 = \frac{2PR}{P+R} \tag{3}$$

Average precision (AP) is the precision across all elements of a category of pills, as defined in Formula 4:

$$AP = \int_0^1 p(r)dr \tag{4}$$

MAP is numerically equal to the average value of the AP sum across all categories, and this value is used to evaluate the overall performance of the model. The definition is shown in Formula 5:

$$\mathrm{MAP} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} AP_i \tag{5}$$

FPS is a common indicator for evaluating the speed of model detection. This refers to the number of images that can be processed per second. In general, FPS over 30 is considered to have achieved real-time detection.

## Results and Discussion

### Comparison of algorithm detection results

After training, the different algorithms were used for pill identification on the test set. The results are shown in Figure 6 and Table 3. Faster R-CNN has the highest MAP among the three algorithms. Compared with YOLO v3 and SSD, the MAP is 7.52% and 5.28% higher, and the overall F1 value is 8.09% and 6.10% higher than that of YOLO v3 and SSD, respectively. This demonstrates that the two-stage algorithm has advantages in terms of detection accuracy compared with the algorithms that complete their processing in one stage. YOLO v3 can predict multiple bounding boxes and their categories simultaneously, and the detection speed is faster than the other network model structures. As shown in Figure 7, YOLO v3 detects 51 images per second, and SSD detects 32 images per second. The detection speed of these two algorithms exceeds 30FPS, which is much faster than the case for Faster R-CNN. If detection efficiency is considered, YOLO v3 performs best among the three models, while Faster R-CNN does not meet the real-time requirement. This limits its potential applications and demonstrates the advantage of the end-to-end one-stage algorithm in detection speed. Based on the analysis of the above experimental results, Faster R-CNN is more suitable if the higher MAP of pill recognition is required, but YOLO v3 may be more suitable for use when the priority is real-time performance and it is feasible to accept a slightly lower MAP. Therefore, we believe that YOLO v3 has the potential to be applied to assist pharmacists to identify pills in a hospital dispensary environment.

### Difficult samples detection comparison

In order to more effectively reflect the effect of the model in identifying tablets with very similar colors and shapes, the experiment selected the types of tablets. As can be seen from Figure 8(a), since the tablets are small and have no obvious printing codes, they are visually more Difficult to distinguish. Taking the YOLO v3 as an example, the results are shown in Figure 8(b). As shown, we can see that the algorithm still performs well on difficult samples. For this group of difficult-to-recognize samples are shown in Table 4. The three algorithms have little difference in the MAP of difficult-to-recognize samples, but YOLO v3 has obvious advantages in FPS and model size. Features that cannot be distinguished by vision can be learned through training (backpropagation ) through the convolution kernel in the CNN. The features learned by the network can then be used as the basis for correct judgment of the type of pills, which greatly speeds up manual dispensing and check the efficiency. In the pharmacy, we can set the confidence threshold to assist the pharmacist in taking the medicine. When the probability (confidence) that the network judges that the current pill belongs to a certain category is lower than our set value, we can think that the network model is difficult to judge the current pill, at this time, pharmacists can participate artificially to ensure correctness.

## Conclusion

We collected pill images and used LabelImg to make a standard PASCAL VOC format image database. Three currently dominant object detection methods, Faster R-CNN, YOLO v3 and SSD, were trained using our pill database and their performance was compared experimentally. The results show that each of the three models has its own advantages and disadvantages. The Faster R-CNN model has a high MAP (87.69%), but the detection speed (FPS : 7) is not fast enough for real-time application. SSD is intermediate in performance, with scores between the other two networks on both speed (FPS : 32) and MAP (82.41%). Although YOLO v3 does not have the highest MAP (80.17%), it can greatly improve the detection speed and achieve real-time performance (FPS : 51). In busy hospital pharmacies, pill identification requires not only a high enough MAP, but also detection speed. YOLOv3 may be the best compromise. This method can quickly the pharmacists to identify drugs, reduce the probability of dispensing the wrong drug, and can help improve patient safety. The YOLO v3 algorithm can meet the conditions of operating on low performance platforms, in environments with requirements for high speed of detection, and has broad development prospects and practical application value.

There are some shortcomings in our study, such as limitations in the experimental dataset, as we have only collected images of split pills from one hospital. A larger dataset would make the results more robust. Another important factor is that some different types of oral solid dosage forms currently in clinical use have a very similar appearance, which will reduce the MAP of model recognition. In future work, we will build larger datasets and keep testing new algorithms to further optimize the model and improve the MAP and speed of detection.

## Abbreviations

AP: Average Precision
CFDA: China Food and Drug Administration
CNN: Convolutional Neural Networks
F1: F1 score
FN: False Negatives
FP: False Positives
FPS: Frames Per Second
MAP: Mean Average Precision
RPN: Region Proposal Net
TN: True Negatives
TP: True Positives
SSD: Single Shot Multi-Box Detector
YOLO v3: You Only Look Once v3

# Declarations

## Ethics approval and consent to participate

Not applicable

## Consent for publication

Not applicable

## Competing interests

The authors declare that they have no conflict of interests.

## Funding

Not applicable

## Authors' contributions

TL and CW were involved in the design of the study; TL and HT collected the data; TL analyzed the data and wrote the manuscript; TL, CW and WL revised the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

Not applicable

## Authors' information

[1]Department of Pharmacy, The Third Affiliated Hospital of Southern Medical University, Guangzhou , China.

## Availability of data and materials

The dataset used in the current study is available from the corresponding author upon reasonable request.

# Reference

[1] Yu W, Li M, Ge Y, et al. Transformation of potential medical demand in China: A system dynamics simulation model[J]. Journal of Biomedical Informatics, 2015, 57: 399-414.

[2] Duan J, Jiao F, Zhang Q, et al. Predicting urban medical services demand in China: an improved grey Markov chain model by taylor approximation[J]. International journal of environmental research and public health, 2017, 14(8): 883.

[3] Rodriguez‐Gonzalez C G, Herranz‐Alonso A, Escudero‐Vilaplana V, et al. Robotic dispensing improves patient safety, inventory management, and staff satisfaction in an outpatient hospital pharmacy[J]. Journal of evaluation in clinical practice, 2019, 25(1): 28-35.

[4] Chang C H, Lai Y L, Chen C C. Implement the RFID position based system of automatic tablets packaging machine for patient safety[J]. Journal of medical systems, 2012, 36(6): 3463-3471.

[5] Mansur J M. Medication safety systems and the important role of pharmacists[J]. Drugs & aging, 2016, 33(3): 213-221. doi: 10.1007/s40266-016-0358-1. PMID: 26932714.

[6] James K L, Barlow D, McArtney R, et al. Incidence, type and causes of dispensing errors: a review of the literature[J]. International journal of pharmacy practice, 2009, 17(1): 9-30. PMID: 20218026.

[7] Tranchard F, Gauthier J, Hein C, et al. Drug identification by the patient: Perception of patients, physicians and pharmacists[J]. Therapies, 2019, 74(6): 591-598. doi: 10.1016/j.therap.2019.03.003. Epub 2019 Apr 2. PMID: 31014975.

[8] Aliper A, Plis S, Artemov A, et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data[J]. Molecular pharmaceutics, 2016, 13(7): 2524-2530. doi: 10.1021/acs.molpharmaceut.6b00248. Epub 2016 Jun 8. PMID: 27200455; PMCID: PMC4965264.

[9] Zheng S, Zhang W, Wang L, et al. Special shaped softgel inspection system based on machine vision[C]//2015 IEEE 9th International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE, 2015: 124-127. doi: 10.1109/ICASID.2015.7405675.

[10] Ju L, Lyu A, Hao H, et al. Deep learning-assisted three-dimensional fluorescence difference spectroscopy for identification and semiquantification of illicit drugs in biofluids[J]. Analytical chemistry, 2019, 91(15): 9343-9347. doi: 10.1021/acs.analchem.9b01315. Epub 2019 Jun 13. PMID: 31184116.

[11] Lee Y B, Park U, Jain A K. Pill-id: Matching and retrieval of drug pill imprint images[C]//2010 20th International Conference on Pattern Recognition. IEEE, 2010: 2632-2635. doi: 10.1109/ICPR.2010.645.

[12] Morimoto M, Fujii K. A visual inspection system for drug tablets[C]//2011 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2011: 1106-1110. doi: 10.1109/ICSMC.2011.6083822.

[13] Suntronsuk S, Ratanotayanon S. Automatic text imprint analysis from pill images[C]//2017 9th International Conference on Knowledge and Smart Technology (KST). IEEE, 2017: 288-293. doi: 10.1109/KST.2017.7886081.

[14] Neto M A V, de Souza J W M, Reboucas Filho P P, et al. CoforDes: An invariant feature extractor for the drug pill identification[C]//2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2018: 30-35.doi: 10.1109/CBMS.2018.00013.

[15] Dhivya A B, Sundaresan M. Tablet identification using support vector machine based text recognition and error correction by enhanced n-grams algorithm[J]. IET Image Processing, 2020, 14(7): 1366-1372.doi: 10.1049/iet-ipr.2019.0993.

[16] Wong Y F, Ng H T, Leung K Y, et al. Development of fine-grained pill identification algorithm using deep convolutional network[J]. Journal of biomedical informatics, 2017, 74: 130-136. doi: 10.1016/j.jbi.2017.09.005. PMID: 28923366.

[17] Swastika W, Prilianti K, Stefanus A, et al. Preliminary Study of Multi Convolution Neural Network-Based Model To Identify Pills Image Using Classification Rules[C]//2019 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE, 2019: 376-380. doi: 10.1109/ISITIA.2019.8937272.

[18] Ou Y Y, Tsai A C, Wang J F, et al. Automatic drug pills detection based on convolution neural network[C]//2018 International Conference on Orange Technologies (ICOT). IEEE, 2018: 1-4.doi:10.1109/ICOT.2018.8705849.

[19] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[20] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[21] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[22] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018. Available: https://arxiv.org/abs/1804.02767v1.

[23] Yaniv Z, Faruque J, Howe S, et al. The national library of medicine pill image recognition challenge: An initial report[C]//2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2016: 1-9. doi: 10.1109/AIPR.2016.8010584.

Figure 1. Faster R-CNN structure

Figure 2. SSD Network structure

Figure 3. YOLO v3 Network structure

Figure 4. Example images of solid oral dosage forms

Figure 5. LabelImg tool for image labeling

Figure 6. Graph of model performance measures

Figure 7. Performance of Deep Learning Model

Figure 8. The actual detection effect of the model (a) difficult samples (b) the detection results of YOLO v3

Table 1. Appearance of the Pills

| Dosage form | Printing | Non-round shape | Non-round appearance | Total number of pill varieties |
|---|---|---|---|---|
| Naked tablet | 2 | 0 | 7 | 21 |
| Sugar coated tablet | 1 | 0 | 8 | 14 |
| Film-coated tablet | 111 | 66 | 66 | 156 |
| Capsule | 34 | - | 55 | 61 |
| Soft capsule | 1 | - | 8 | 9 |
| Total | 149 | 66 | 144 | 261 |

Table 2. Parameter Configuration

| Parameters | Parameter values |
|---|---|
| Batch | 64 |
| Subdivisions | 16 |
| Learning rate | 0.001 |
| Momentum | 0.9 |
| Decay | 0.0001 |

Table 3. Evaluation of Deep Learning Models

| Algorithm | Precision/% | Recall/% | F1/% | MAP/% |
|---|---|---|---|---|
| YOLO v3 | 69.13 | 80.19 | 70.14 | 80.17 |
| Faster R-CNN | 62.19 | 94.24 | 78.23 | 87.69 |
| SSD | 63.17 | 88.69 | 72.13 | 82.41 |

Table 4. The Indicators of Models in Identifying Difficult Samples

| Algorithm | MAP/% | FPS | Model Size |
|---|---|---|---|
| YOLO v3 | 78.52 | 69 | 89M |
| Faster R-CNN | 79.63 | 3 | 426M |
| SSD | 78.69 | 41 | 149M |

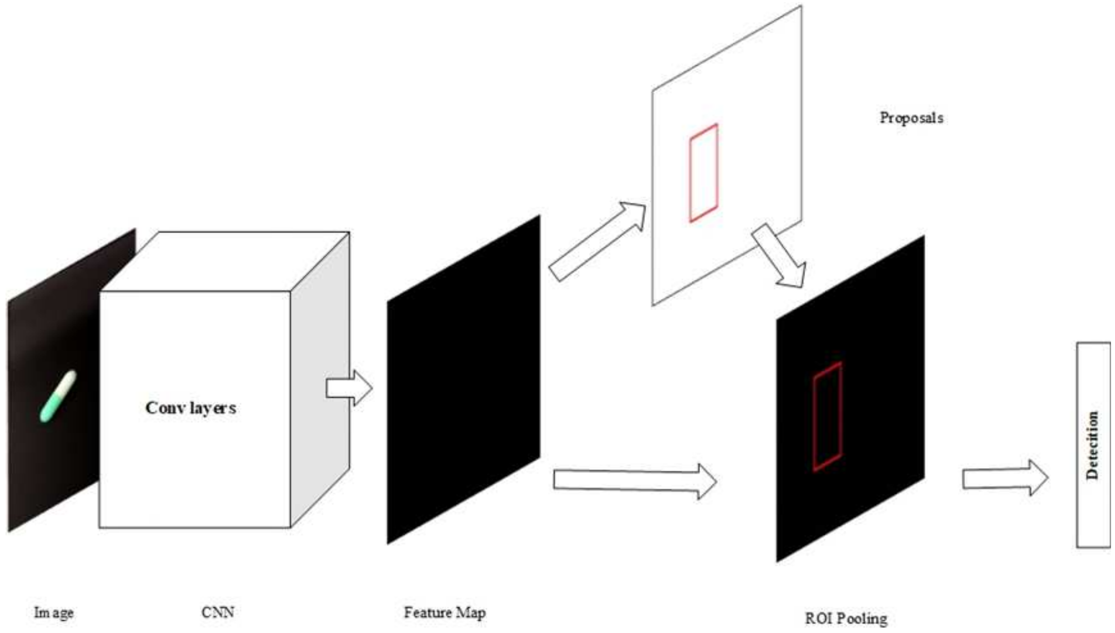

Proposals
Conv layers
Detecition
Im age
CNN
Feature Map
ROI Pooling

Figure 1.

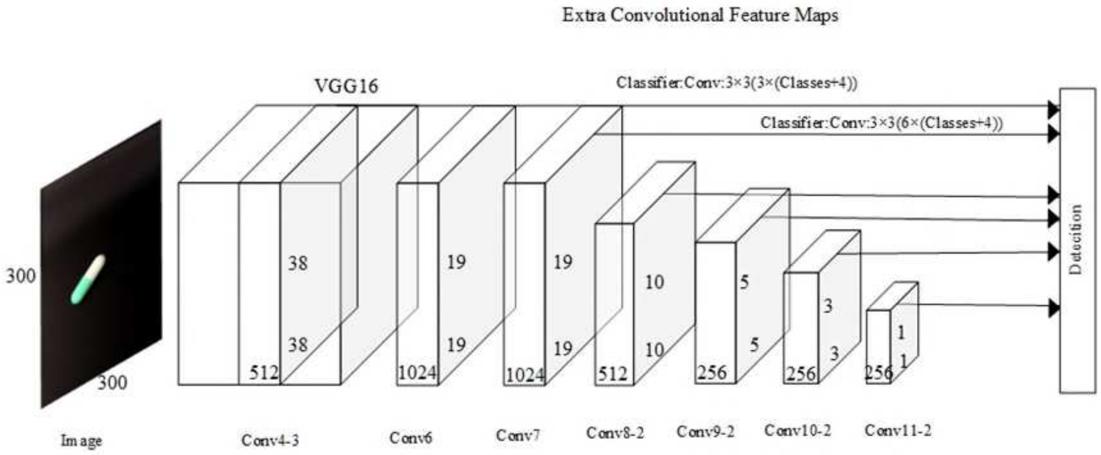
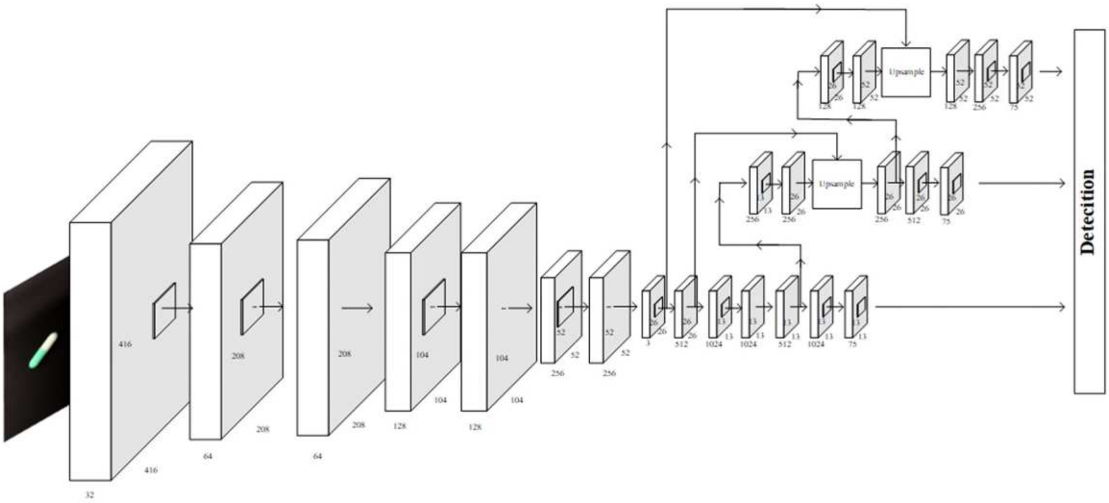

Extra Convolutional Feature Maps
VGG16
Classifier:Conv:3×3(3×(Classes+4))
Classifier:Conv:3×3(6×(Classes+4))
Detection
300
300
38
38
512
19
19
1024
19
19
1024
10
10
512
5
5
256
3
3
256
1
1
256
Image
Conv4-3
Conv6
Conv7
Conv8-2
Conv9-2
Conv10-2
Conv11-2

Figure 2.



416
416
32
208
208
64
208
208
64
104
104
128
104
104
128
52
52
256
52
52
256
26
26
3
26
26
512
13
13
1024
13
13
1024
13
13
512
13
13
1024
13
13
75
13
13
256
26
26
256
Upsample
26
26
256
26
26
512
26
26
75
26
26
128
52
52
128
Upsample
52
52
128
52
52
256
52
52
75
Detecition

Figure 3.

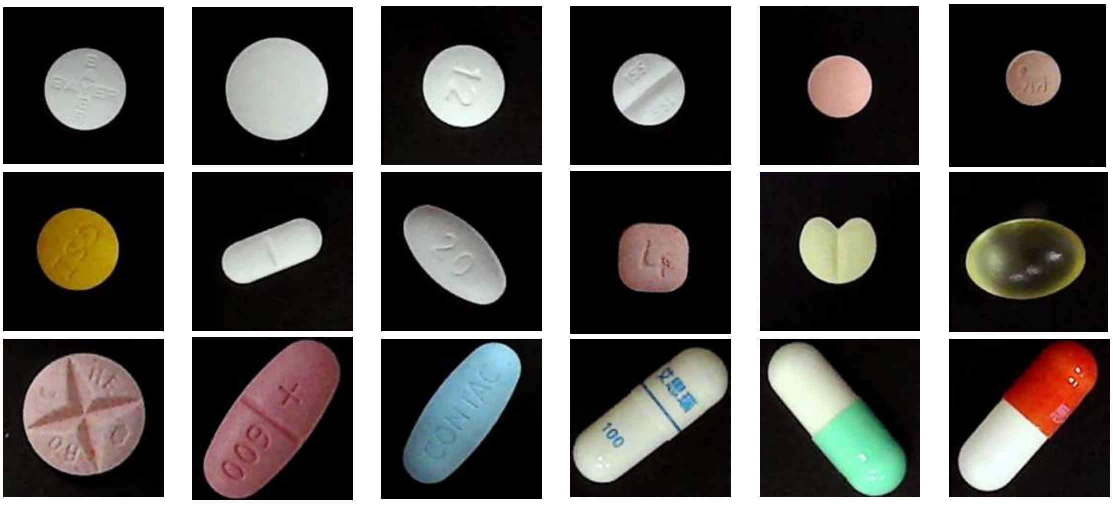
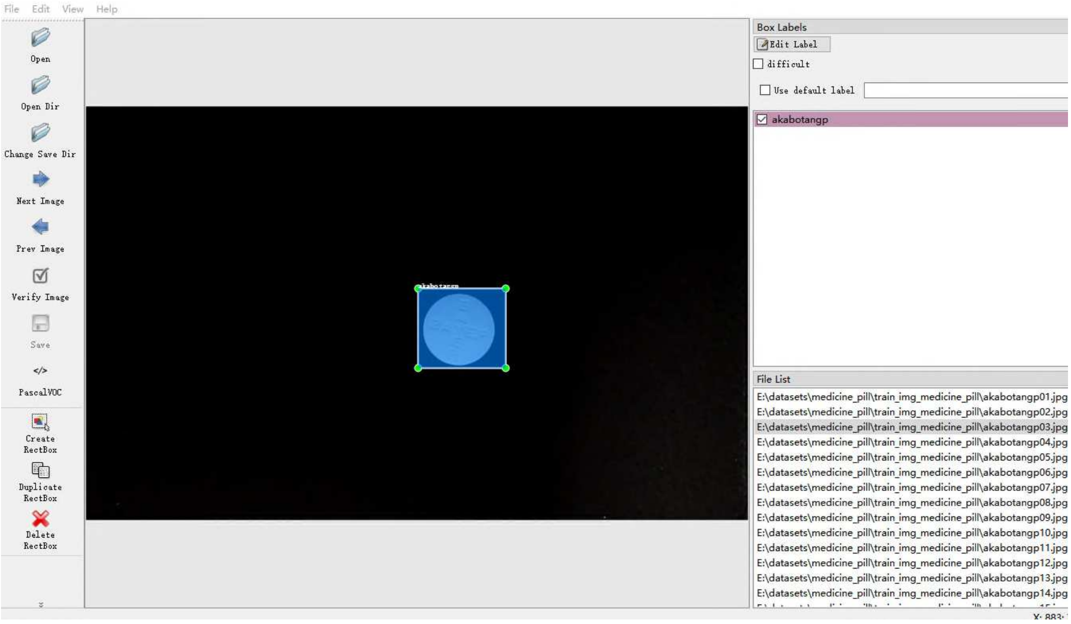


Figure 4.



```
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp01.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp02.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp03.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp04.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp05.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp06.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp07.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp08.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp09.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp10.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp11.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp12.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp13.jpg
E:\datasets\medicine_pill\train_img_medicine_pill\akabotangp14.jpg
```

Figure 5.

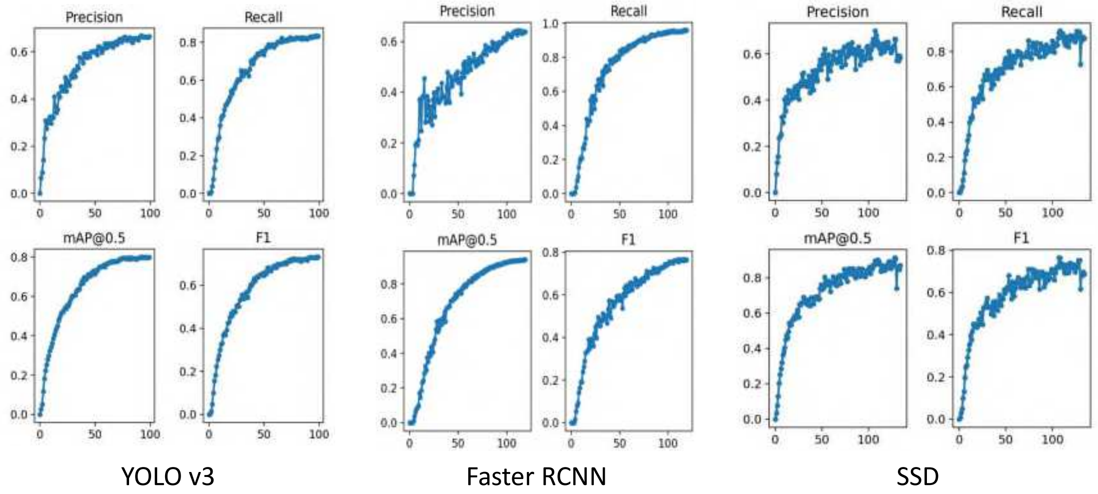

Precision
Recall
mAP@0.5
F1
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
YOLO v3
Faster RCNN
SSD

Figure 6.

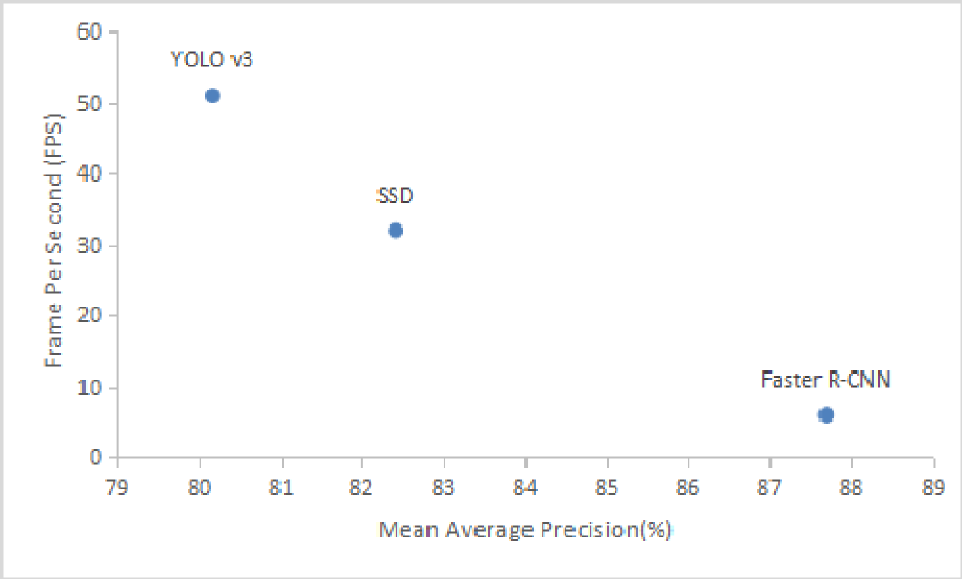

60
50
40
30
20
10
0
Frame Per Se cond (FPS)
YOLO v3
SSD
Faster R-CNN
79
80
81
82
83
84
85
86
87
88
89
Mean Average Precision(%)

Figure 7.

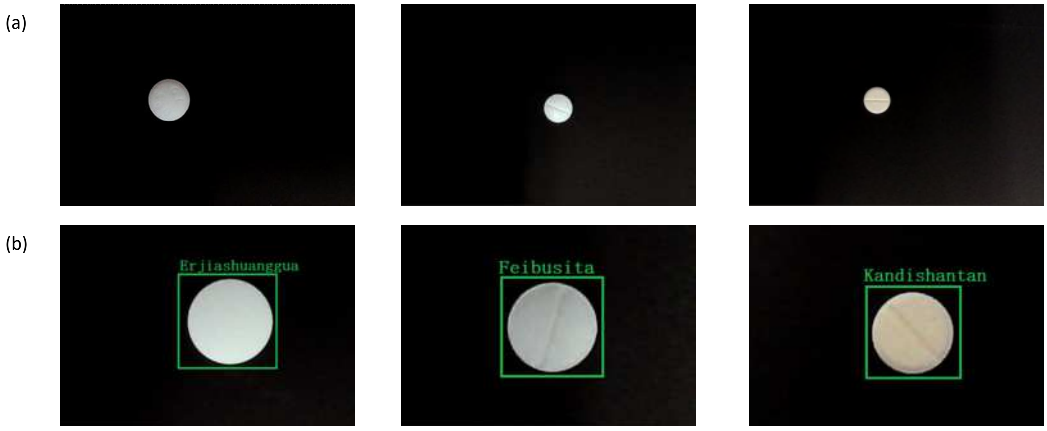


Figure 8.