Research on an Improved Neural Network Model for Film Text Image Segmentation in Film Internet of Things

Yangfan Tong
Wuhan University
Shuo Feng (shuofeng20201102@163.com)
Pingdingshan University
Ruiqing Zhang
Hunan University of Technology and Business,

Research

Keywords: Film Internet of Things, Film text, ResNet101, MASK RCNN, Film document image segmentation model, algorithm model

Posted Date: November 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-106013/v1

Abstract

In order to solve the problem that Film text is difficult to recognize and difficult to handle in Film Internet of Things, a method that can effectively identify the content in Film text is sought. This paper uses the Mask RCNN algorithm with ResNet101 as the backbone network to establish a Film document image segmentation model.The optimal hyperparameters are: the shape ratio of the anchor frame is [0.5, 1, 3], the threshold for non-maximum suppression is 0.15, and the confidence level is 0.85. The F1 score obtained at this time is 0.8951. When these hyperparameters are substituted into the IOU of 0.8, the F1 score is 0.7417. According to the results of the Pattern Recognition Laboratory of the Chinese Academy of Sciences, this algorithm model ranked first with an IOU of 0.6. Under the premise that IOU is 0.8, it is ranked second, and the first is a non-end-to-end model with a single task. It can be seen that the adjustment of the hyperparameters and the training of the algorithm model are relatively successful.The experimental results show that the MASK RCNN can accurately identify all the formulas in the Film Text. MASK RCNN is significantly better at identifying small objects such as formulas in Film Text images than traditional fast cnn and faster cnn.

1. Introduction

As an effective carrier of information, Film Texts have very important meaning in our daily life. In general, we will refer to everything that is readable on a computer or paper, with text, images, formulas, charts, etc. as a Film Text. In order to use and manage these Film Text information about a concise and efficient manner, scientists have conducted a lot of research on Film Text processing methods since the 1960s. For example, these Film Texts can be processed into images using a scanner, camera or mobile phone and then imported into a computer and organized into systematic Film Text images that can be stored, managed and transmitted more efficiently and conveniently. There are also many new problems to be solved when extracting Film Text images taken with a digital camera: how to extract the meaning of the text, how to extract the images and formulas and the information of the chart and how to judge the structure of a text image. In response to these problems most of the current traditional methods can only segment and recognize plain text images but this is only a part of the vast amount of information.

The traditional method roughly estimates the edge region of the text by using the Sobel operator and then expands the region by using a morphological method such as expansion and then classifies the separated single characters to obtain the final text information. The target detection algorithm is the YOLO[1] series of the one-stage method, the SSD[2] series and the RCNN[3] series of the two-stage method. The former is biased toward the operation speed and the latter is biased toward higher accuracy Currently more prosperous is the use of neural network algorithms for text image processing [4–11]..

For the application scene of text image segmentation it is often not necessary to have too high real-time performance so the more mature algorithm MASK R-CNN[11] in RCNN[12–18] series is used for image segmentation.

The application scene for text image segmentation often does not need too high real-time performance, so the more mature algorithm MASK R-CNN[19–21] in RCNN series is used for image segmentation which has the advantage of high accuracy and simultaneously perform four tasks: instance segmentation, semantic segmentation, target detection, and image classification[22–23].

2. Method

2.1 MASK R-CNN algorithm architecture

MASK RCNN is an algorithm that can perform Instance Segmentation. It can achieve various complex tasks such as target detection, semantic segmentation, instance segmentation and human pose estimation by adding different branches. It is flexible and powerful. It is a representative algorithm at the current stage. The abstract process architecture of the algorithm is shown in Fig. 1:

Here, first input the preprocessed image data to the network,Image data enters FPN (Pyramid Feature ExtractionNetwork) with ResNet101 as the backbone network□thereby obtaining the feature map after feature extraction. Then a sliding window is obtained for each point on the extracted feature map to obtain a generated candidate anchor point. These anchor points are combined with the feature map information into the regional recommendation network to obtain scores and correction information for each anchor point. Then since the predicted coordinates are floating point numbers and are not aligned with the feature map so the ROI Align layer is used for adjustment. Finally, different branches are connected to the recommended areas, MASK generation, regression of frame coordinates, and generation of categories are performed. In particular, the introduction of the MASK branch allows the network to perform instance segmentation. Another noteworthy issue is that the network abandoned the original softmax loss function and adopted the sigmod loss function which to some extent avoids competition between the same categories and puts more focus on optimizing the MASK's Pixel results.

2.2 Pyramid feature extraction network

The pyramid network is a feature extraction method that improves the accuracy by extracting multi-scale feature information, especially in the detection of small objects. Its essence is a component of ResNet or other backbone neural network that is a common feature extraction network that can be combined with different classic networks to improve the network.

The pyramid feature extraction network structure can be roughly divided into three parts: a top-down deep convolutional neural network, a bottom-up upsampling process and a side-to-side connection between feature maps for information fusion. The detailed process structure is shown in Fig. 2:

The top-down part on the left side of the above picture is actually the forward process of the neural network. In the forward process, the feature map changes in size after passing through certain layers, so the layer where the feature size change does not occur is divided into one stage and the feature map size changes five times in total, so there are a total of In five stages, the characteristics of each stage output correspond to {C1, C2, C3, C4, C5}, thereby forming a pyramid of features. In general, the residual feature

extraction structure of ResNet101 [12] or ResNet 50 [12] is used as the convolutional neural network structure in five stages.

The bottom-up process uses a bottom-up process using upsampling. The principle is an interpolation method that inserts new points by interpolating between two pixels in order to magnify the original image. The upsampled feature map used has the same size as the previous image.

The role of the side join is to fuse the same level of featureeich from top to bottom and from bottom to top. The role of the side join is to fuse the same level of feature maps from top to bottom and from bottom to top. First, a 1*1 convolution operation is performed on the feature map in the top-down process, the purpose of which is to adjust the number of channels of the feature map so that the feature map has the same number of channels as the upsampled feature map and at the same time Keep the original feature map information. Then, the two left and right feature maps of the same level are merged and the operation is a simple numerical value corresponding addition operation and the merged feature map has richer information. In this process, a total of four merged feature maps {P2, P3, P4, P5} are generated and then the three feature maps are subjected to a 3*3 convolution operation which can eliminate the aliasing effect of the upsampling. Thus, a new feature map {P2, P3, P4, P5} is obtained. In order to extract more detailed overall information, the P5 is downsampled and a new feature map P6 is generated. So far, the feature pyramid network has completed the extraction operations of different scale feature maps {P2, P3, P4, P5, P6}.

3. Model Establishment

3.1 Data preprocessing

For the task of Film Text image segmentation, this paper uses the icdar2017 dataset, which is a batch of text image data, which is a dataset used to segment instances of different image types in the text. There are four types of data, respectively, formulas, charts, and other backgrounds. The labeling effect is roughly as shown in Fig. 3.

There are two files in the dataset of icdar2017, which are folders for storing original images and MASK images. There is no file that stores the coordinate information of the border, which requires manual generation of the border coordinates. First select different suitable thresholds and the image is binarized by 0/255 to obtain different types of MASK images at first. After that, if it is a background MASK, directly generate an external matrix as a calibration frame, otherwise, the MASK image is first used to generate a set of coordinate points for the instance target contour in the MASK using the contour detection algorithm [13]. Then generate a circumscribed matrix according to the set of coordinate points, that is, the coordinates of the instance calibration frame. And then the MASKs of coall instances are then binarized to 0/1, thus completing the preliminary pre-processing of the data.

Second, the form needed to process the data into an algorithmic model□The MASK RCNN algorithm model requires a total of three inputs: the original image, MASK, category code list, and calibration box.

First, you need to perform the dimensioning process for the binary MASKs of different instances in an image and add a dimension at the end and these MASKs are then merged according to the last dimension, which results in a MASK matrix that is one dimension higher than the original, where each pixel is a uniquely One-hot encoding data. For the code list of the category, it only needs to correspond to the order of the categories that are uniquely encoded in the MASK. The calibration box is also merged in the order of the instances in the uniquely encoded pool and merged into an ordered matrix list.

3.2 Model framework and environment construction

This paper chooses Tensorflow as the framework of MASK RCNN. It has a combination of lower-level symbolic computing libraries and higher-level network specification libraries and is developed and maintained by Google. It has a huge community and therefore has stable support and performance. In addition, it has Tensorboard which is a powerful network model visualization tool. It can support GPU parallel training and support the development of multiple languages such as python and java. The most important thing is that it takes advantage of the method of building static graphs and then recalculating, it makes many calculations that affect computational performance compiled and implemented in C + + and speeds up the calculation

Secondly, Anaconda is used to build and manage the python environment and numerous third-party packages. Because Python has a very active and huge community, many extremely useful packages are developed by third parties and the version of various software becomes an extremely cumbersome process. Use Anaconda to build a development environment that automatically adapts the relationships between packages to automatically install third-party packages and greatly simplifying the process of setting up the environment. In this project, the following development packages are mainly required, as shown in the following Table 1:

Table 1
The primary environmental development package

Development package name	Specified version
numpy	1.15.4
scipy	1.1.0
Pillow	5.3.0
matplotlib	2.2.2
scikit-image	0.13.0
tensorflow-gpu	1.10.0
keras-gpu	2.1.3
IPython	7.2.0
opencv-python	-
h5py	2.8.0

3.3 setting hyperparameters

Most of the deep learning models are an end-to-end network training mode and accompanied by a lot of hyperparameters that requires people to manually adjust. how to set and adjust these hyperparameters is critical to the structure and performance of a model. Therefore, it is necessary to design a reasonable model training strategy to adjust these hyperparameters. In the MASK RCNN algorithm, this paper takes the structure and naming of hyperparameters in the code as an example and lists the following important hyperparameters and their values, as shown in Table 2 below.

Table 2
The important hyperparameter list

Hyperparameter name	Parameter explanation	Hyperparameter value
BACKBONE	Feature extraction backbone	"resnet101"
BACKBONE_ STRIDES	Step size of each layer of the FPN pyramid	[4, 8, 16, 32, 64]
RPN_ANCHOR _SCALES	Square anchor length (in pixels)	(32, 64, 128, 256, 512)
RPN_ ANCHOR _RATIOS	The anchor ratio (width/height) value of each unit is 1 for the square anchor point and 0.5 for the wide anchor point.	[0.5, 1, 2]
RPN_NMS _THRESHOLD	Threshold for non-maximum suppression used to filter the region of the RPN	0.7
RPN_TRAIN _ANCHORS_PER _IMAGE	Used to train the RPN network in each picture Number of anchors	256
POST _NMS_ROIS _TRAINING	Number of the largest recommended areas retained by NMS filtering during training	2000
POST_NMS _ROIS _INFERENCE	Number of the largest recommended areas retained by NMS filtering during the reasoning process	1000
LEARNING_RATE	Optimizer learning step	0.001
WEIGHT_DECAY	Regularized attenuation weight	0.0001
LOSS_WEIGHTS	Weight of each loss function	{"rpn_class_loss":1, "rpn_bbox_loss":1, "mRCNN_class_loss":1, "mRCNN_bbox_loss": 1, "mRCNN_MASK_loss": 1}

3.4 Select Optimizer

In the process of training, for faster training data and better non-convex optimization of parameters, this paper chooses the method of Stochastic Gradient Descent as the optimizer and here is a brief introduction to its principle.

Assume $h(\theta)$ is the function to be fitted and $J(\theta)$ is the corresponding loss function:

$$h(\theta) = \sum_{j=0}^{n} \theta_j x_j \quad (1)$$

$$J(\theta) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{2} [h_\theta(x^j) - y^j]^2 = \frac{1}{m} \sum_{i=1}^{m} \cos t(\theta, (x^i, y^i)) \quad (2)$$

$$\cos t(\theta, (x^i, y^i)) = \frac{1}{2} [h_\theta(x^i) - y^i]^2 \quad (3)$$

Because the random gradient descent is iteratively updated by each sample so it has a large sample size and it only needs a small amount of data to iterate to a relative optimal solution. The rate of random gradient descent is faster than the gradient descent and because the direction of the gradient's descent is constantly changing, it is relatively easy to jump out of the local optimal solution. The solution to the stochastic gradient descent is as follows:

First, the loss function $\cos t(\theta)$ is used to find the partial derivative of θ :

$$\frac{\partial J(\theta)}{\partial \theta_j} = -(y^i - h_\theta(x^i)) x_j^i \quad (4)$$

Since the loss function is usually minimized, each one is updated in the negative direction of each parameter:

This gradient update is complete. Random gradient descent, while having a faster speed, sacrifices a lot of precision and is more susceptible to noise and each optimization is not necessarily a positive optimization.

Over-fitting under-fitting is always an unavoidable problem during the training of the model, unfitting will make the model not learn enough knowledge, However, over-fitting can cause the model to form a bias which makes the generalization performance of the model worse. Therefore, it is especially important to set the appropriate training strategy to control the under-fitting and over-fitting of the model.

3.5 model training

First, the data set is divided into a total of about 2400 data in the data set of icdar2017, which is divided into three data sets according to the ratio of 2:1:1. They are a training set with about 1200 data, a validation set with about 600 data and a test set with about 600 data. There is no intersection between the three data sets. The training set is used to calculate the parameters of the neural network and to reduce the gradient, that is, the model draws knowledge from the training set. The distribution of training set data has the greatest impact on the model. The verification set is used to verify the performance and rationality of the model, thereby manually adjusting the hyper parameters of a training process, such as the number of model iterations, the threshold selection of non-maximum suppression and and other hyper parameters. Because the hyper parameters are adjusted based on the validation set, the model also learns a portion of the knowledge from the validation set's data, so the distribution of the validation set data has a minor impact on the model. The test set is used after all the steps of the model training are completed and it is used to make an objective evaluation of the model of the final result produced throughout the process. Because it is completely isolated from the validation set and the training set, the model does not generate any a priori bias on it, which makes it possible to objectively and fairly rely on the performance of a model, especially the generalization performance of the model.

3.6 Model Evaluation

In deep learning, after completing the construction of a model, the model must be effectively evaluated and then the model's hyper parameters and model training strategies are adjusted according to the results of this evaluation to obtain a fully trained high performance model. So how to set a suitable performance evaluation index is very important, the commonly used evaluation indicators have Accuracy and Recall and F1 scores. First introduce the concepts of TP (True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives), The meanings are: It is divided into positive samples and is paired correctly, divided into negative samples and paired correctly, divided into positive samples but divided into errors, divided into negative samples but divided into errors. Refer to Figure 4 for details.

In the field of target detection, the mean Average Precision is extended based on the accuracy and recall rate. Because in the task of target detection, the model needs to evaluate the classification and location of objects. Each image has a different type and location of the target and the metrics used in normal image classification cannot be directly applied to the target detection problem, so this evaluation method is based on the method of Intersection over Union. The concept is shown in Figure 5.

$$F1 = \frac{2 * precision * recall}{precision + recall} \qquad (6)$$

4. Graphic Film Text Segmentation Experiment

4.1 Experimental process

To evaluate the performance of a model, its evaluation index is only one aspect. If you really go deep into the model and observe and analyze the performance of each level, you can more clearly and in detail

reflect the performance of the model. The MASK RCNN algorithm mainly includes a pyramid feature network, a regional recommendation network, a coordinate area fine-tuning, and a head network. The feature pyramid network is a feature extraction network and the structure is clear and simple. Moreover, the recommendation of the region plays a very important role in the generation of the calibration frame. Therefore, the analysis results of the regional recommendation network, the coordinate area fine adjustment and the head network are mainly analyzed.

First select an image and make a prediction for this image. The original image of the selected picture is shown in Fig. 6.

The Regional Recommendation Network (RPN) runs a lightweight binary classifier on many boxes (anchors) on the image and returns scores with or without objects. Usually, even a positive anchor cannot completely cover an object. Therefore, the RPN also regresses the refinement (increment of position and size) to be applied to the anchor to shift it and adjust it slightly to the correct boundary of the object. The comparison of the refined effects is shown in Fig. 7.

Next, enter the interior of the regional recommendation network, run each of the operational graphs and dissect its predictions. Its main computing nodes are RPN network output, pre_nms_anchors and refined_anchors in the ROI, and refined_anchors_clipped. The output of the RPN network is mainly to predict whether all classes are in the anchor box and output the forward anchor frame. Take the anchor frame of the first one of the scores as an example. The output is shown in Fig. 8.

It can be seen that the generated anchor frame has a certain calibration effect but it is still relatively messy. Next, enter the coordinate fine-tuning phase, first generate the anchor frame that has not undergone non-maximum suppression and its deviation from the modified anchor frame, as shown in Figure (9)a below. At the same time, the effect picture after the cut-off to prevent the image from crossing the boundary is displayed, as shown in the following figure (9)b.

As can be seen from Fig. 9(a) above, the modified anchor frame is more accurate and performs better. In addition, due to the display of the anchor frame selected according to the score of the top fifty,it happens that there is no anchor frame beyond the scope of the image,so the shearing effect of anchor frame in Fig. 9 (b) is not particularly obvious. Next, it is necessary to perform non-maximum suppression on the anchor frame to avoid repeated frame recommendation, the idea is that the anchor frame with the highest IOU threshold is selected as the output with the highest confidence, the output after non-maximum suppression is shown in Fig. 10.

After the non-maximum suppression, the final recommended anchor frame of the regional recommendation network is output. It can be seen that the anchor frame has shifted from the repeated focus to a broader focus, that is, attention to all possible objects and its output. The anchor frame is calibrated for a wider range of content, avoiding the neglect of information.

Next, the three branches of the head of the network are entered, that is, the tasks of classification, calibration, and MASK generation are combined with the generated anchor frame and original image information. Firstly, the image of the anchor frame calibration is classified and the frame is predicted. Each predicted frame correction information is combined with the anchor frame to be corrected again and then the confidence level for each category is generated and then the threshold is filtered according to the confidence threshold to generate the final frame It is shown in Fig. 11 below.

Here's a step-by-step analysis of the process of border prediction and classification. The border effect of this part of the network input is shown in Fig. 12, it is a recommended area that has not been adjusted

The input and recommended borders are quite rough and cluttered. Here, the offset value of the frame predicted by the frame prediction network needs to be corrected. The comparison effect of several frames before and after the correction is as shown in Fig. 13.

After the correction, the calibration of the frame is more accurate and has excellent performance. Then, the non-maximum value suppression of the corrected frame is performed again and the corrected borders are prevented from overlapping and a repeated recommended area is formed. The frame after non-maximum suppression is shown in Fig. 14.

At this point, the resulting prediction frame and classification is the final result of the network and it can be seen that the model has quite good predictive power. In addition, the paper also statistics the predicted deviation values and obtains the deviation of the four predicted calibration frames. As shown in Fig. 15 below.

It can be seen that the degree of correction for the anchor point is small, and the degree of correction of the frame and the height of the frame is large, which is highly likely to be related to the value of the initialization of the hyperparameter (the aspect ratio of the recommended anchor frame), it can be considered to improve the result by changing the aspect ratio of the anchor frame.

The predictive branch of the MASK is a new feature of the MASK RCNN algorithm, which generates a MASK for instance prediction through a full convolutional neural network for pixel-level instance segmentation. For this image, the MASK image generated by the algorithm model is shown in Fig. 16.

Relative to the calibration frame, the MASK can express more views of the network model on the image, such as the focus and ignore points of the network. The partial output of the activation layer of the full convolution network is explained in more detail below. See Fig. 17 below.

It can be seen that the network has a higher response to images such as formulas and relatively low response to larger images. In addition, it also has a certain response to some of the numbers and formulas in the article, which forms a certain noise. In general, the model can produce an excellent activation response for the target you want to predict.

4.2 Model evaluation

Combined with the above analysis results, different Super-parameters are used to adjust the model and the evaluation method of F1 score is used to evaluate the model. Finally, the relatively appropriate Super-parameters are debugged. In the MASK RCNN, the hyperparameters that have a greater influence on the model results in the inference phase and have the shape ratio of the anchor frame, the threshold of the non-maximum suppression and the confidence, so that in the case of two IOUs (0.6, 0.8) Multiple experiments to explore the optimization performance of hyper parameters. In the experiment, the results of the hyperparameters when the IOU at the time of evaluation was 0.6 are shown in Table 3 below.

Combined with the above analysis results, different models are used to adjust the model and the F1 score evaluation method is used to evaluate the model and finally the corresponding hyperparameters are debugged.

Table 3
The experimental results of the hyperparameters(IOU = 0.6)

Anchor frame shape ratio	Threshold for non-maximum suppression	Confidence	F1 score
[0.5,1,2]	0.3	0.7	0.7272
[0.5,1,2]	0.1	0.7	0.7439
[0.5,1,2]	0.05	0.7	0.7428
[0.5,1,2]	0.15	0.7	0.7451
[0.5,1,3]	0.15	0.7	0.7495
[0.5,1,4]	0.15	0.7	0.7299
[0.5,1,3]	0.15	0.75	0.7616
[0.5,1,3]	0.15	0.8	0.7707
[0.5,1,3]	0.15	0.85	0.7751
[0.5,1,3]	0.15	0.9	0.7721

As can be seen from Table 3, the optimal hyper parameter is: the shape ratio of the anchor frame is [0.5, 1, 3], the threshold of non-maximum suppression is 0.15, the confidence is 0.85, and the F1 obtained at this time. The score is 0.7751. Substituting these hyper parameters into the IOU of 0.8 yielded an F1 score of 0.5417. The verification diagrams of fast cnn, faster CNN, and MASK RCNN are shown in Fig. 18.

The optimal hyper parameters are: the shape ratio of the anchor frame is [0.5, 1, 3], the threshold for non-maximum suppression is 0.15, and the confidence level is 0.85. The F1 score obtained at this time is 0.8951. When these hyper parameters are substituted into the IOU of 0.8, the F1 score is 0.7417. According to the results of the Pattern Recognition Laboratory of the Chinese Academy of Sciences [15], this algorithm model ranked first with an IOU of 0.6. Under the premise that IOU is 0.8, it is ranked second,

and the first is a non-end-to-end model with a single task. It can be seen that the adjustment of the hyper parameters and the training of the algorithm model are relatively successful.

5. Results And Discussion

According to the results of the Pattern Recognition Laboratory of the Chinese Academy of Sciences, this algorithm model ranked first with an IOU of 0.6. Under the premise that IOU is 0.8, it is ranked second, and the first is a non-end-to-end model with a single task. It can be seen that the adjustment of the hyper parameters and the training of the algorithm model are relatively successful.The experimental results show that the MASK RCNN can accurately identify all the formulas in the Film Text. MASK RCNN is significantly better at identifying small objects such as formulas in Film Text images than traditional fast CNN and faster CNN.

Abbreviations

WHO: World Health Organization; FN:False Negative; FP:False Positive;TN:True Negative;TP:True Positive; TPR:True Positive Rate; FPR:False Positive Rate

Declarations

Acknowledgements

The authors acknowledged the anonymous reviewers and editors for their efforts in valuable comments and suggestions.

Funding

None

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions

Y.F. Tong proposes the innovation ideas and theoretical analysis, and S. Feng carries out experiments and data analysis. R.Q. Zhang conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Art, Wuhan University , Wuhan 430072, China

[2]School of Literature, Pingdingshan University , PingDingshan 467000, China

[3]School of Literature and Journalism, Hunan University of Technology and Business, Changsha 410205, China

References

1. Andrés Villa-Henriksen,Gareth T.C. Edwards,Liisa A. Pesonen,Ole Green,Claus Aage Grøn Sørensen. Internet of Things in arable farming: Implementation, applications, challenges and potential[J]. Biosystems Engineering,2020,191.
2. Wariston Fernando Pereira,Leonardo da Silva Fonseca,Fernando Ferrari Putti,Bruno César Góes,Luciana de Paula Naves. Environmental monitoring in a poultry farm using an instrument developed with the internet of things concept[J]. Computers and Electronics in Agriculture,2020,170
3. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
4. McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.
5. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.
6. Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
7. Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.
8. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
9. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
10. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

11. He K, Gkioxari G, Dollár P, et al. MASK r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
12. Muhammad Aasim Rafique,Witold Pedrycz,Moongu Jeon. Vehicle license plate detection using region-based convolutional neural networks[J]. Soft Computing,2018,22(19).
13. [2]Zhang Jinpeng,Zhang Jinming,Yu Shan. Hot Anchors: A Heuristic Anchors Sampling Method in RCNN-Based Object Detection.[J]. Sensors (Basel, Switzerland),2018,18(10).
14. [3]Malhotra Kumar Rohit,Davoudi Anis,Siegel Scott,Bihorac Azra,Rashidi Parisa. Autonomous detection of disruptions in the intensive care unit using deep MASK RCNN.[J]. Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops,2018,2018.
15. [4]Nida Nudrat,Irtaza Aun,Javed Ali,Yousaf Muhammad Haroon,Mahmood Muhammad Tariq. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering.[J]. International journal of Film informatics,2019,124.
16. [5]Xudong Sun,Pengcheng Wu,Steven C.H. Hoi. Face detection using deep learning: An improved faster RCNN approach[J]. Neurocomputing,2018,299.
17. [6]Mehmet Turan,Yasin Almalioglu,Helder Araujo,Ender Konukoglu,Metin Sitti. Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots[J]. Neurocomputing,2018,275.
18. [7]Nudrat Nida,Aun Irtaza,Ali Javed,Muhammad Haroon Yousaf,Muhammad Tariq Mahmood. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering[J]. International Journal of Film Informatics,2019.
19. [9]Weishan Zhang,Xia Liu,Jiangru Yuan,Liang Xu,Haoyun Sun,Jiehan Zhou,Xin Liu. RCNN-based foreign object detection for securing power transmission lines (RCNN4SPTL)[J]. Procedia Computer Science,2019,147.
20. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
21. Suzuki S. Topological structural analysis of digitized binary images by border following[J]. Computer vision, graphics, and image processing, 1985, 30(1): 32-46.
22. Bottou L. Large-scale machine learning with stochastic gradient descent[M]//Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010: 177-186.
23. Gao L, Yi X, Jiang Z, et al. Icdar2017 competition on page object detection[C]//2017 14th IAPR International Conference on Film Text Analysis and Recognition (ICDAR). IEEE, 2017, 1: 1417-1422.

Figures

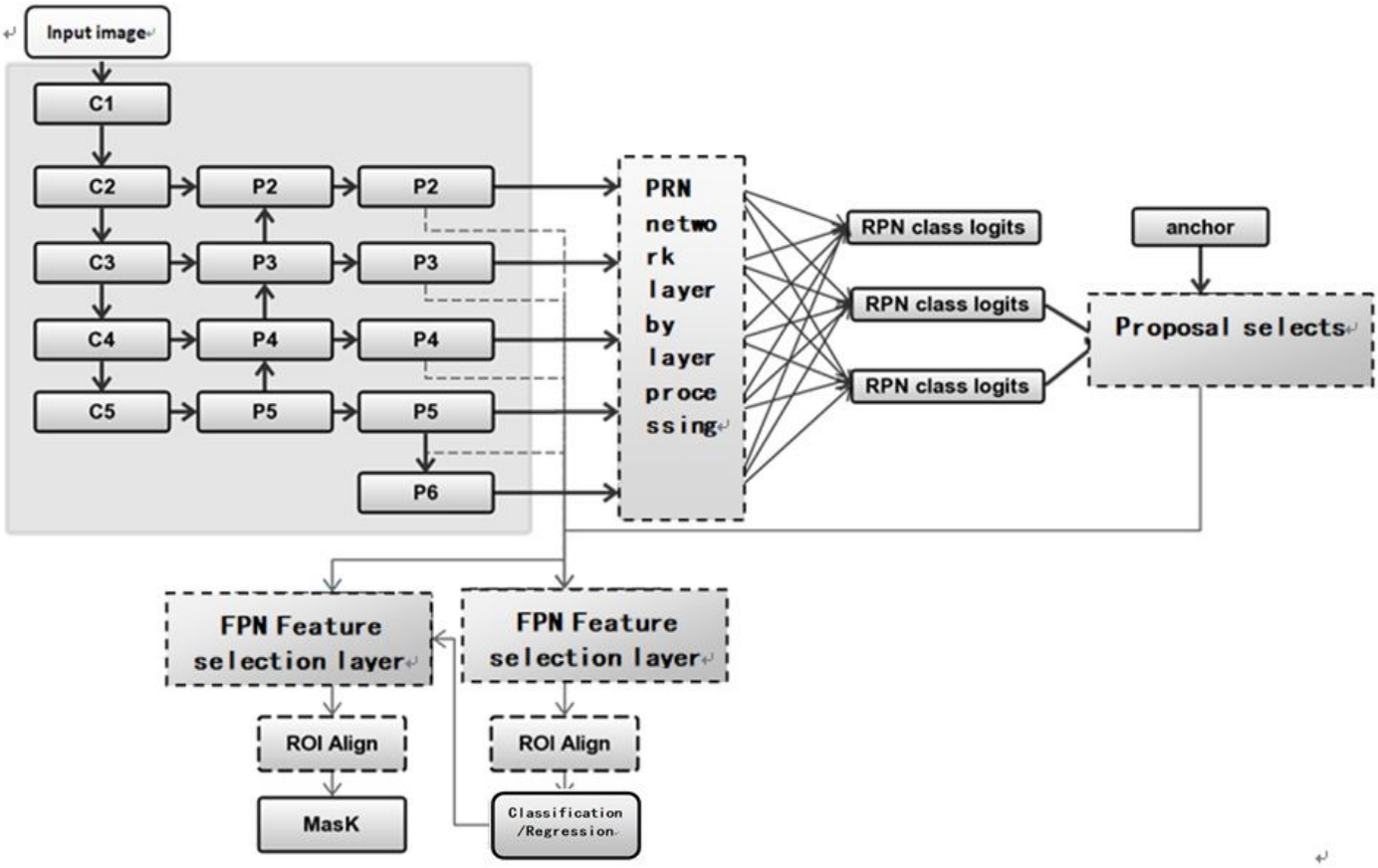

Figure 1

The structure diagram of the MASK RCNN

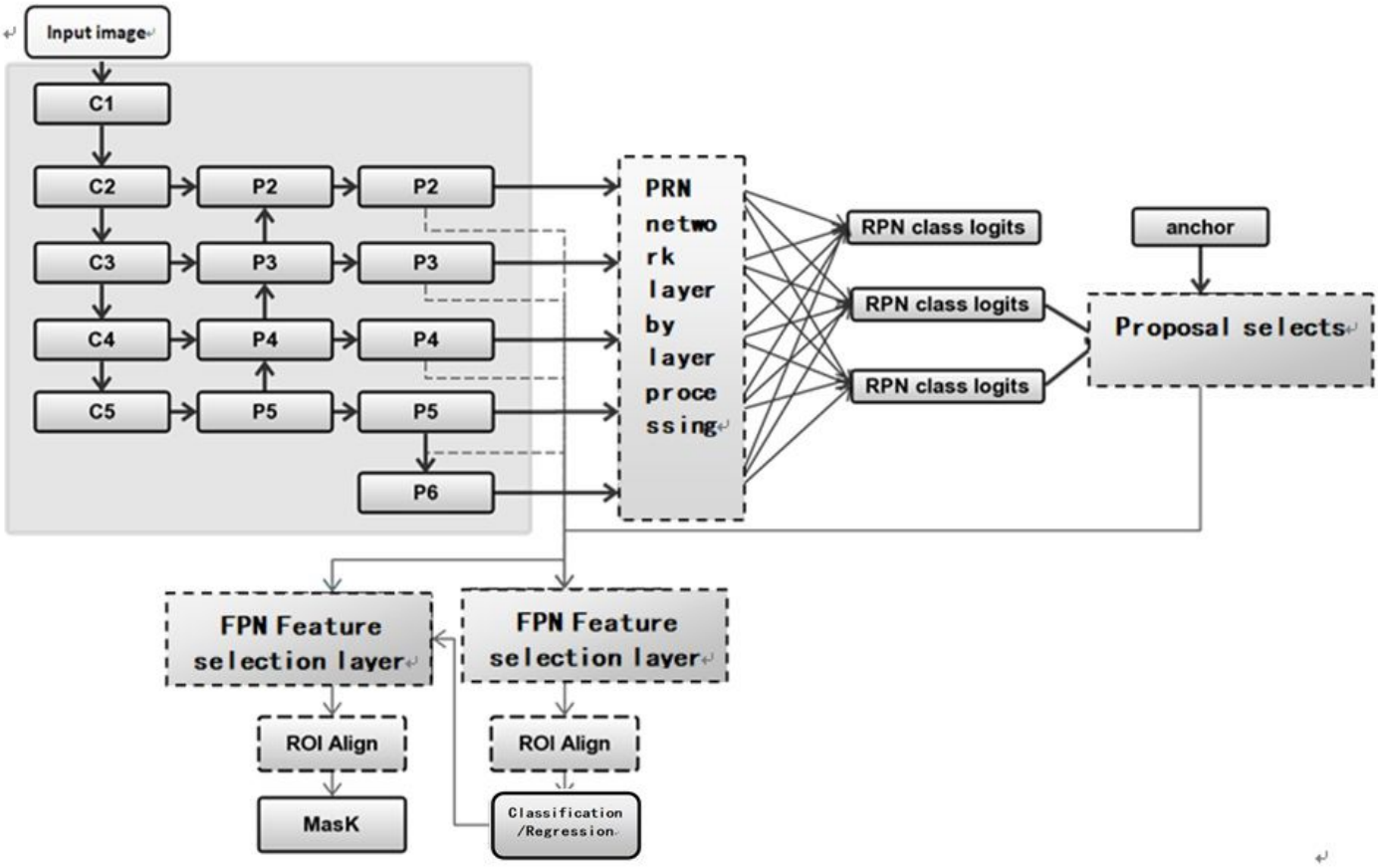

Figure 1

The structure diagram of the MASK RCNN

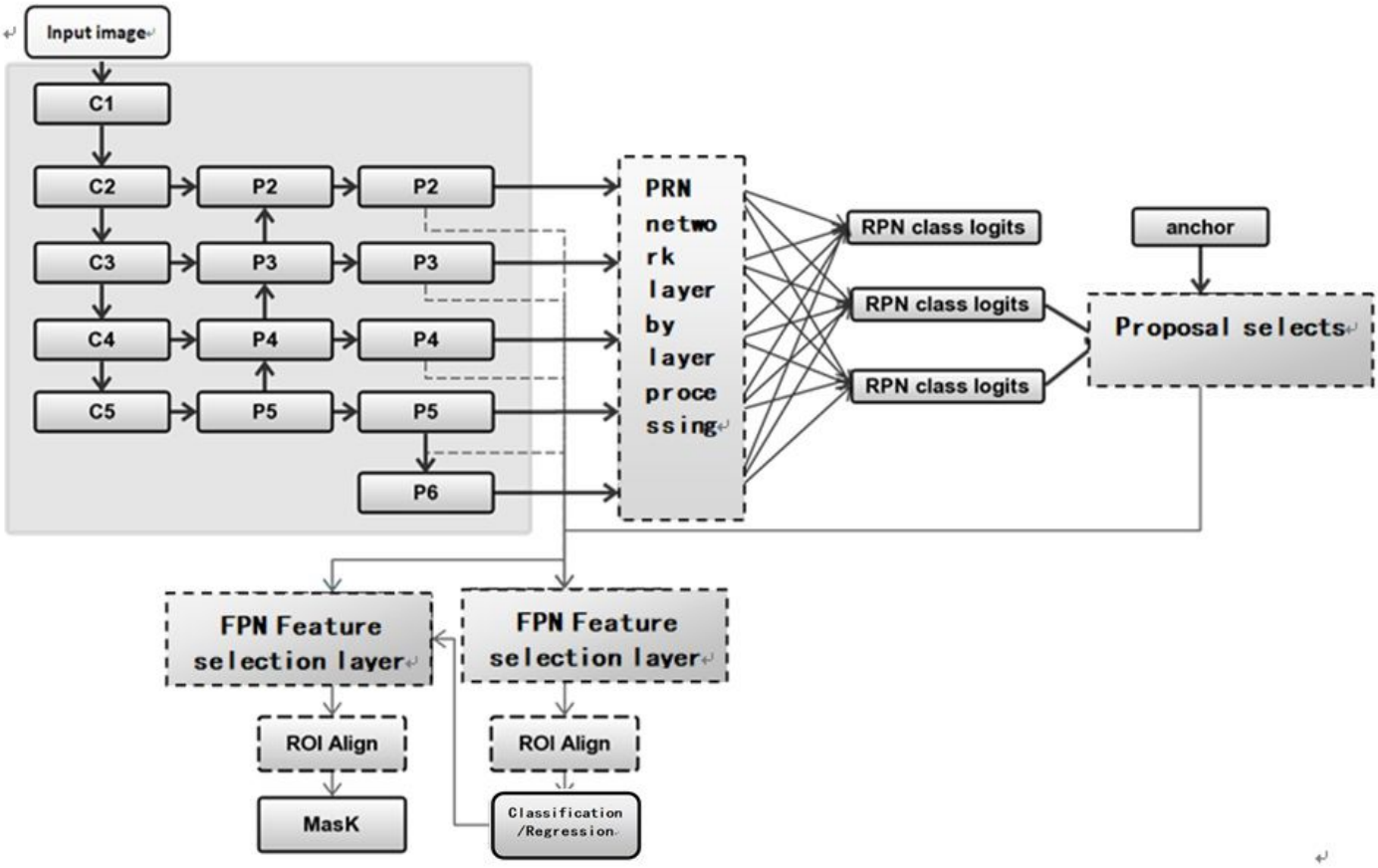

Figure 1

The structure diagram of the MASK RCNN

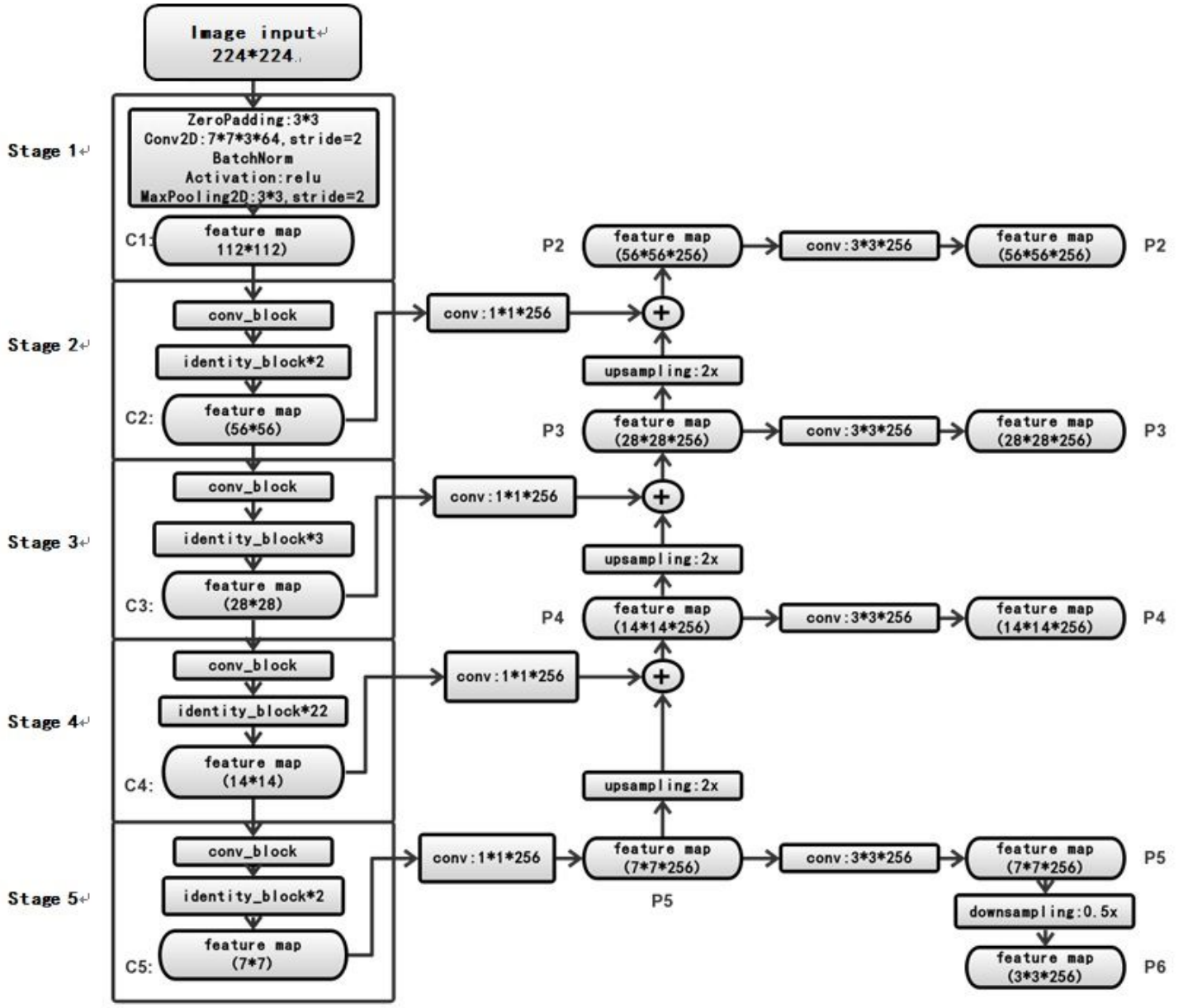

Figure 2

The structure diagram of the MASK RCNN

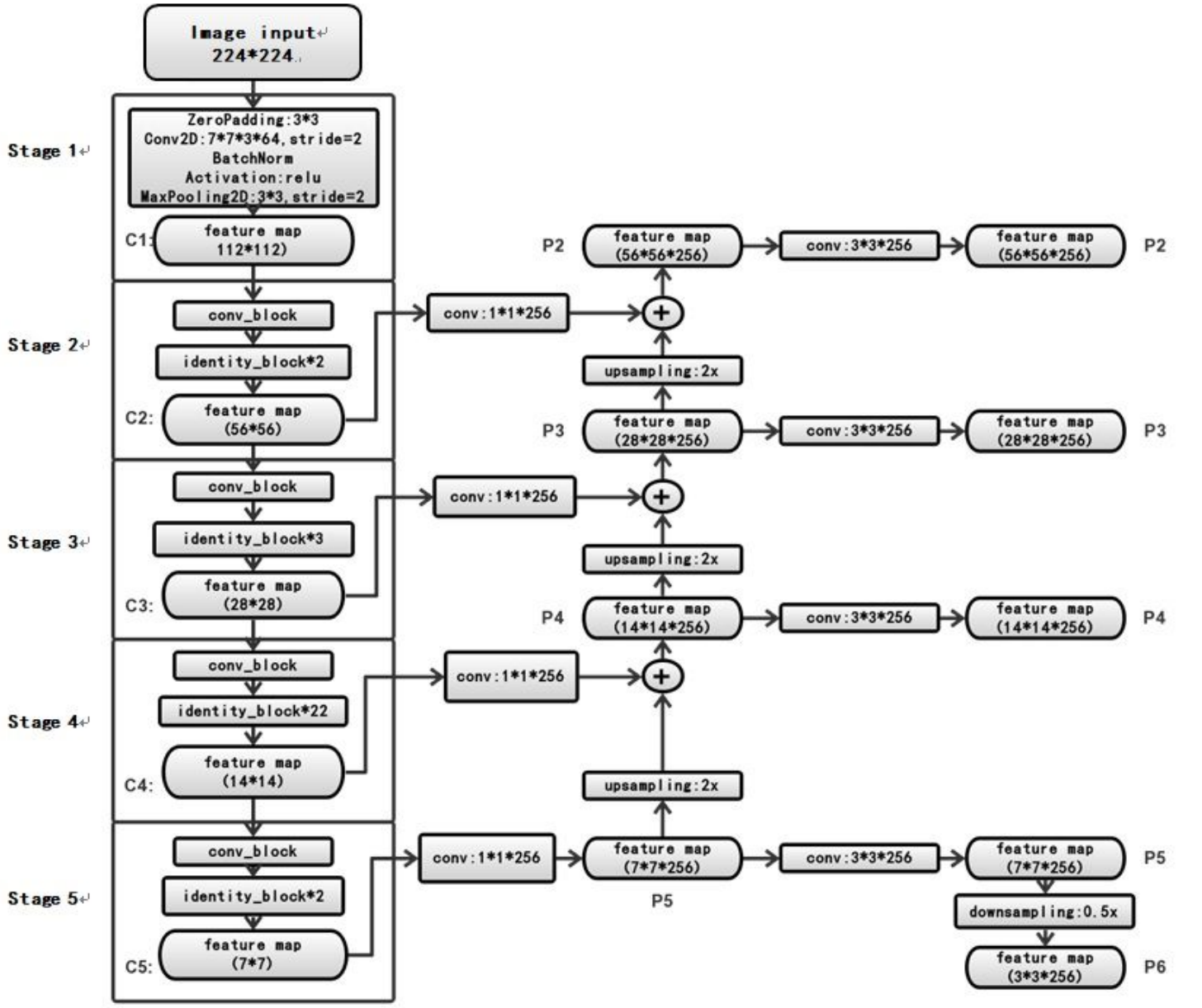

Figure 2

The structure diagram of the MASK RCNN

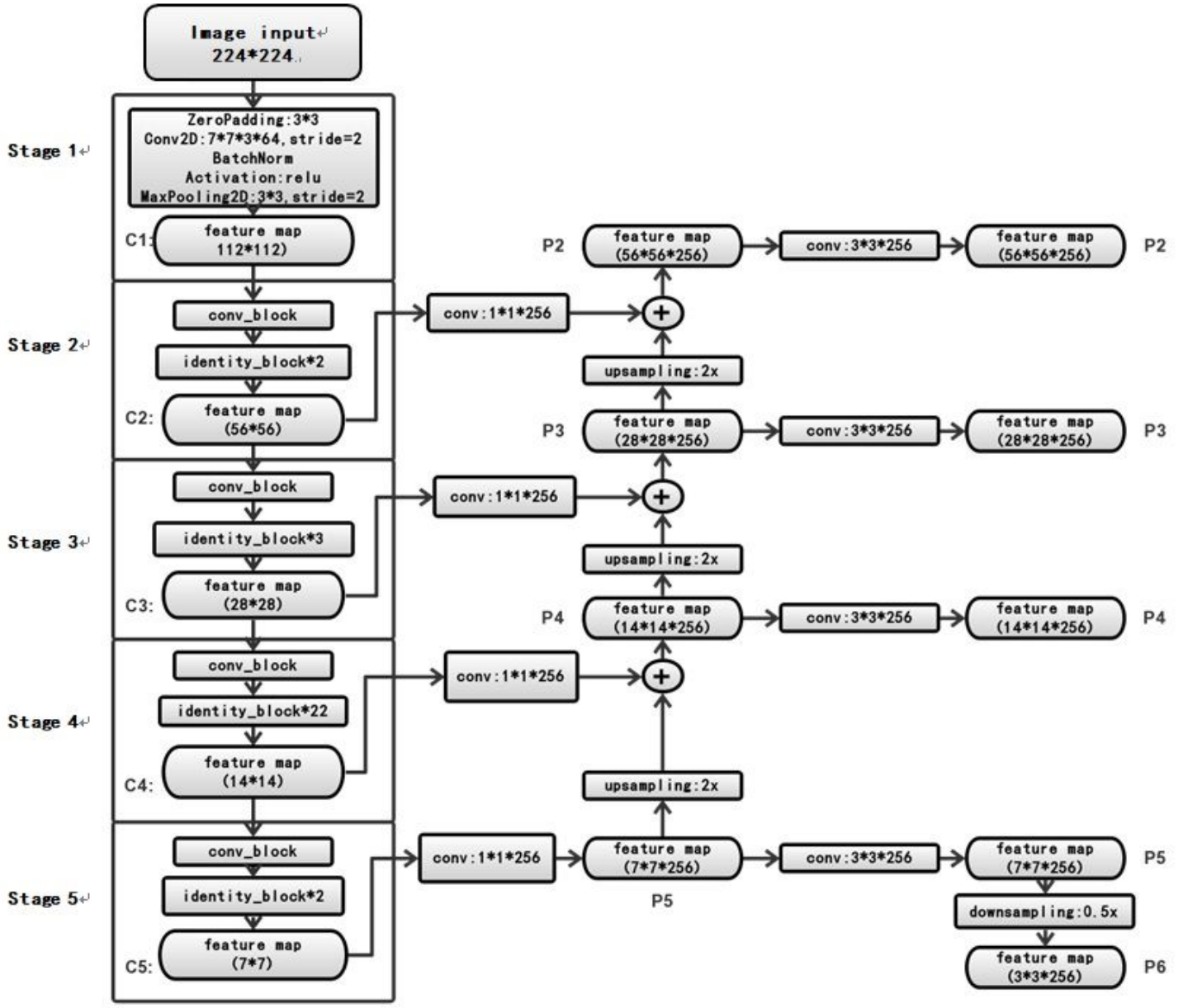

Figure 2

The structure diagram of the MASK RCNN

Figure 3

The structure diagram of the Feature Pyramid Networks

Figure 3

The structure diagram of the Feature Pyramid Networks

Figure 3

The structure diagram of the Feature Pyramid Networks

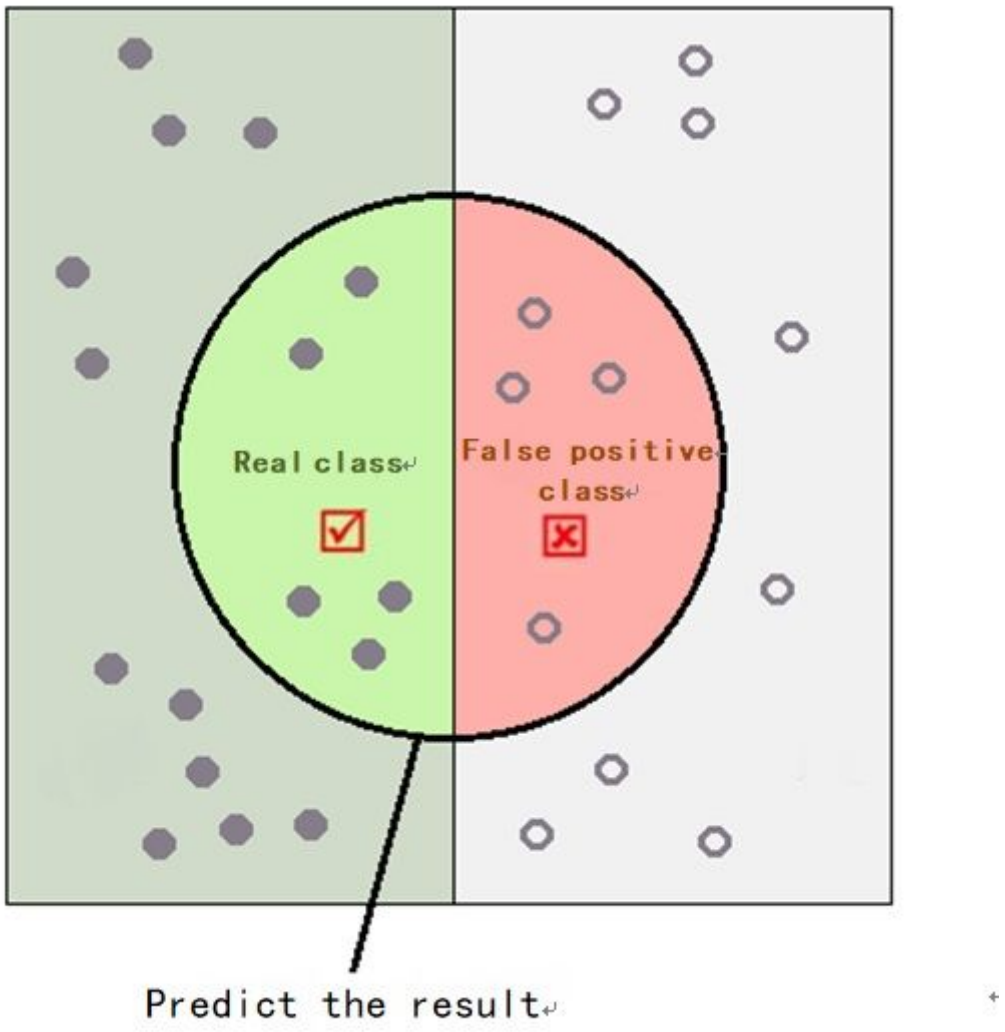

Figure 4

The schematic diagram of TP, TN, FP, FN

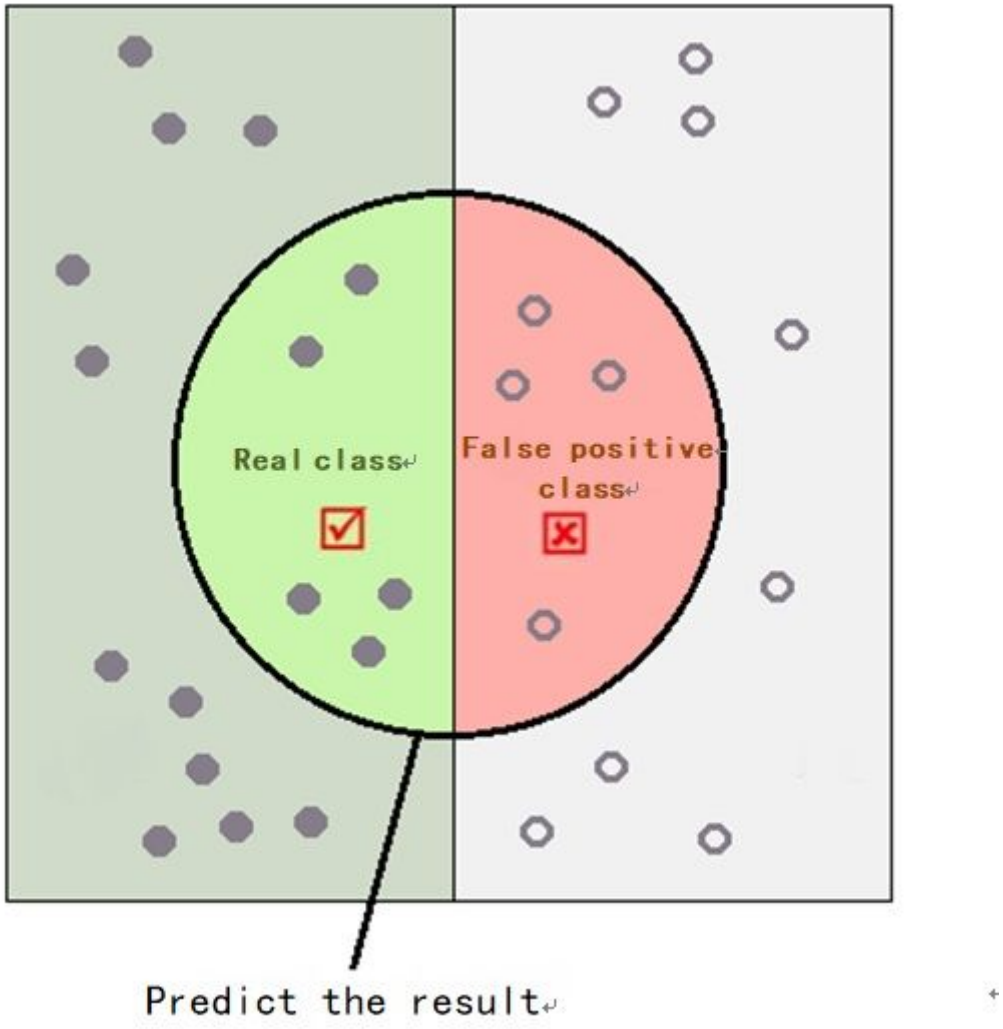

Figure 4

The schematic diagram of TP, TN, FP, FN

Figure 4

The schematic diagram of TP, TN, FP, FN

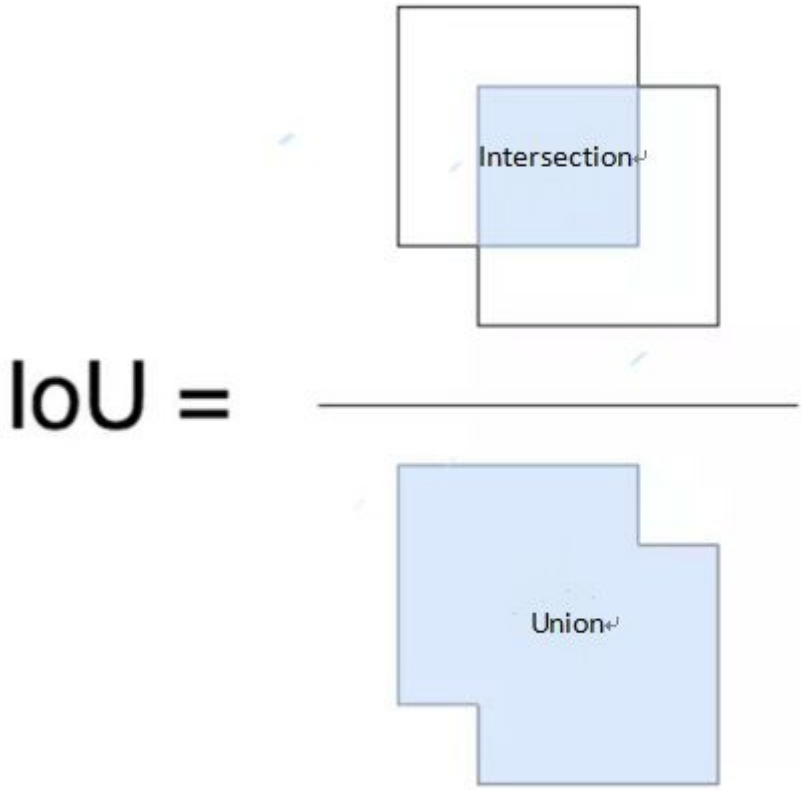

Figure 5

The schematic diagram of Intersection over Union

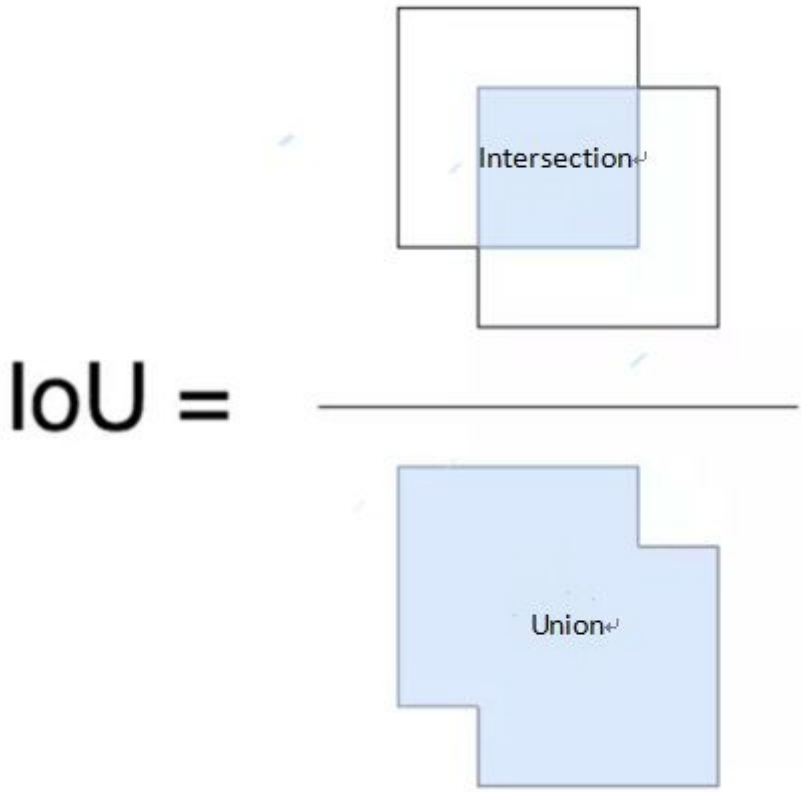

Figure 5

The schematic diagram of Intersection over Union

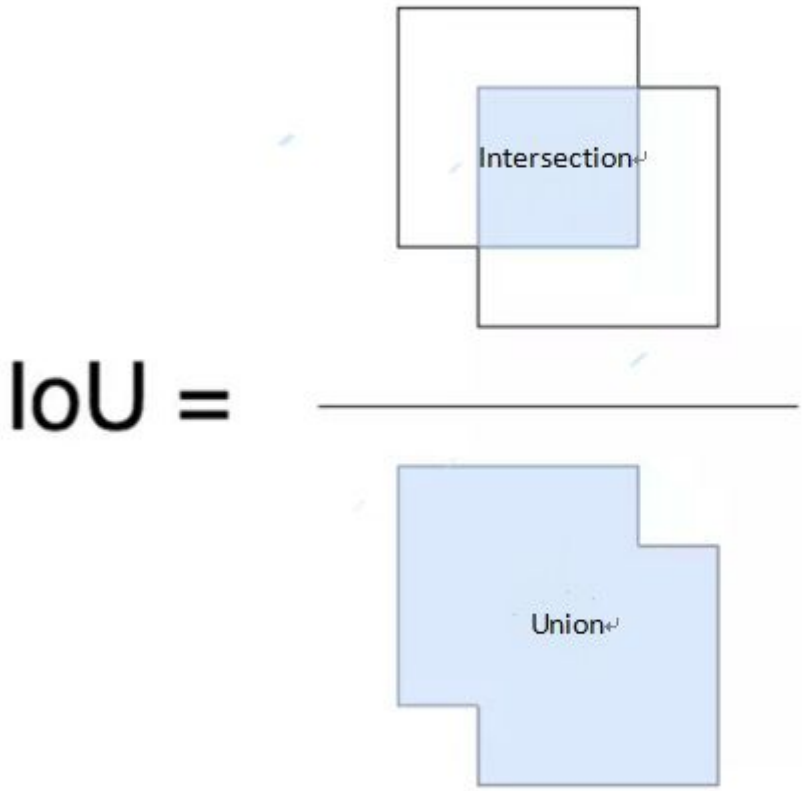

Figure 5

The schematic diagram of Intersection over Union

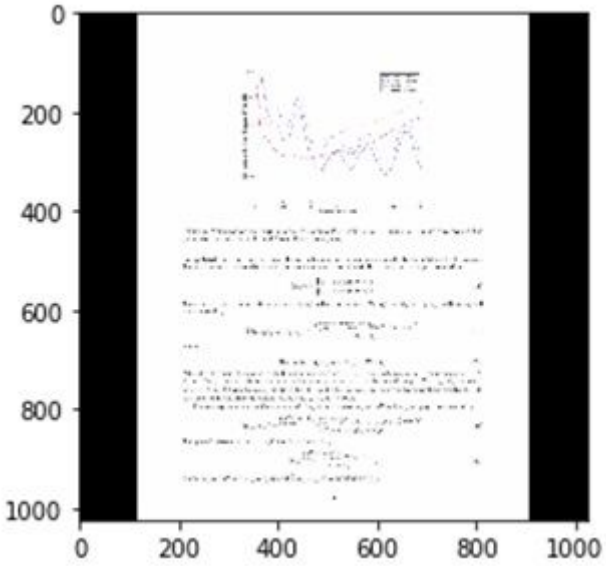

Figure 6

The original image of the predicted image

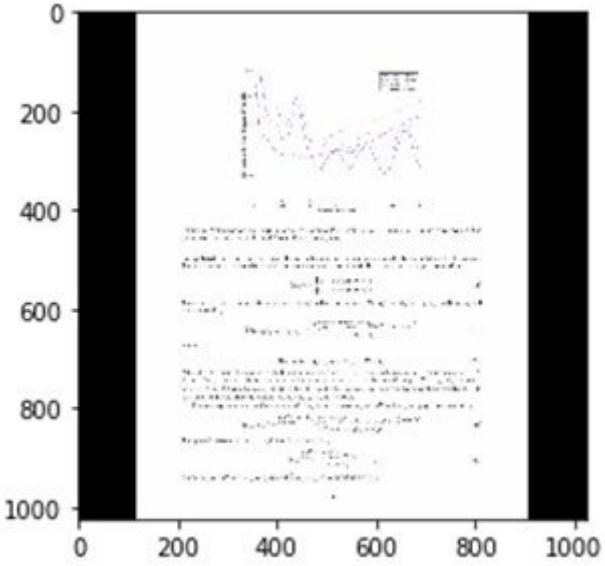

Figure 6

The original image of the predicted image

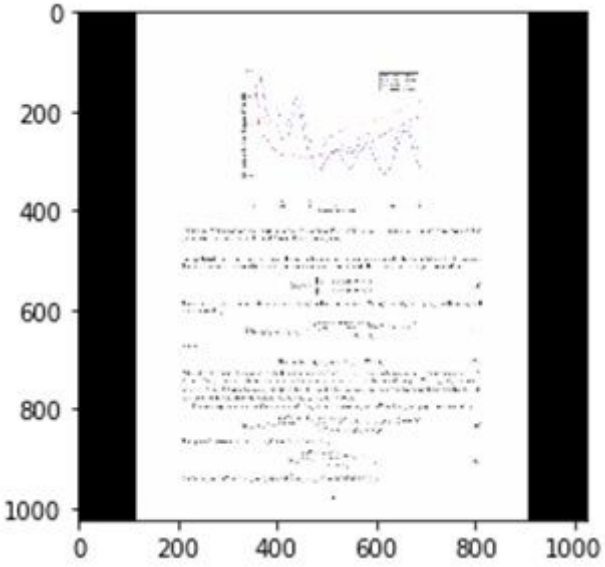

Figure 6

The original image of the predicted image

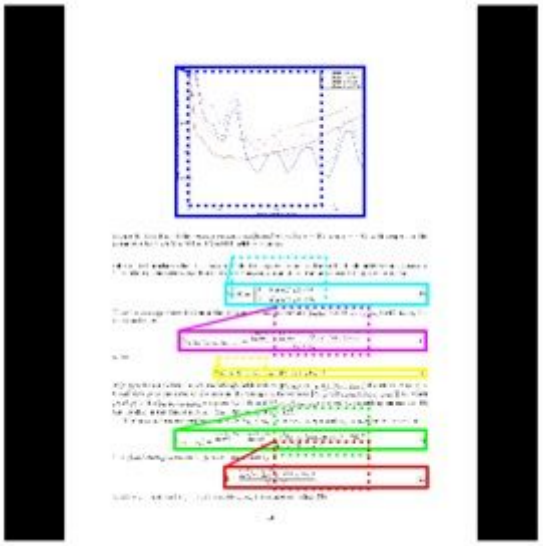

Figure 7

The effect diagram of the Anchor refinemen

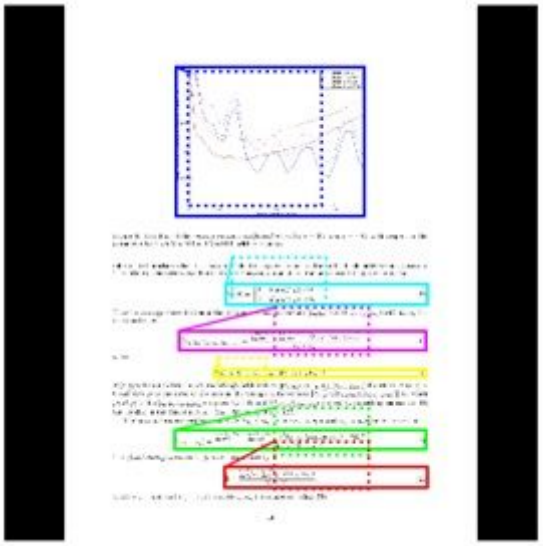

Figure 7

The effect diagram of the Anchor refinemen

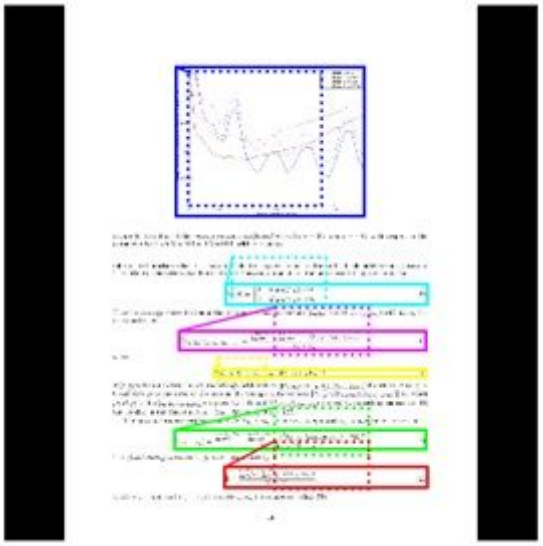

Figure 7

The effect diagram of the Anchor refinemen

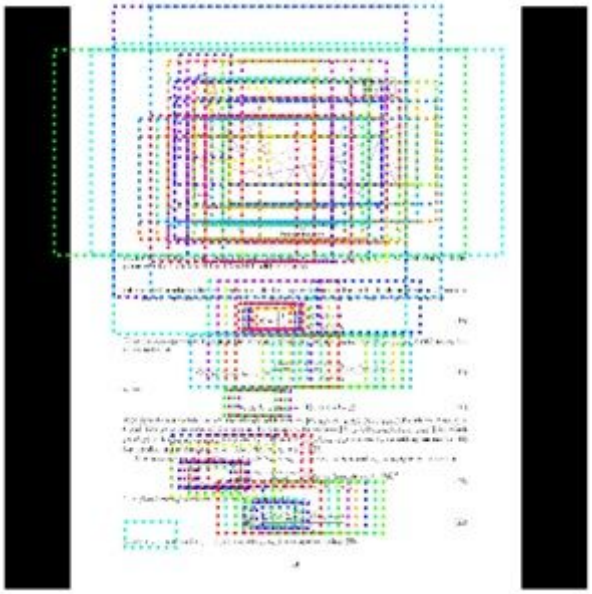

Figure 8

The diagram of the Anchor refinement

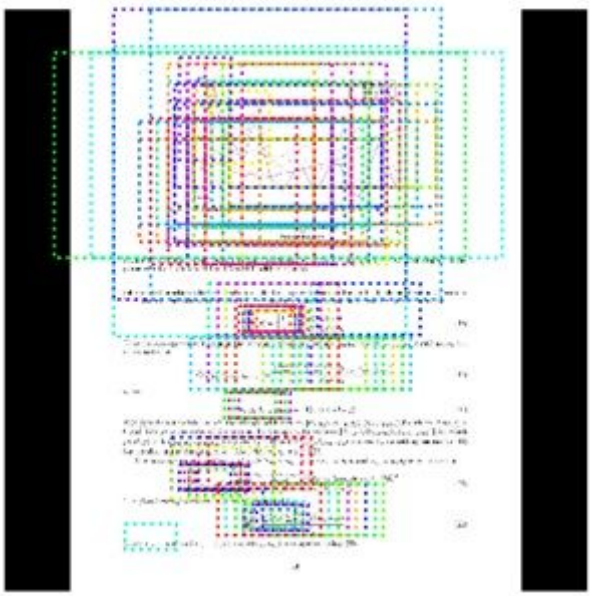

Figure 8

The diagram of the Anchor refinement

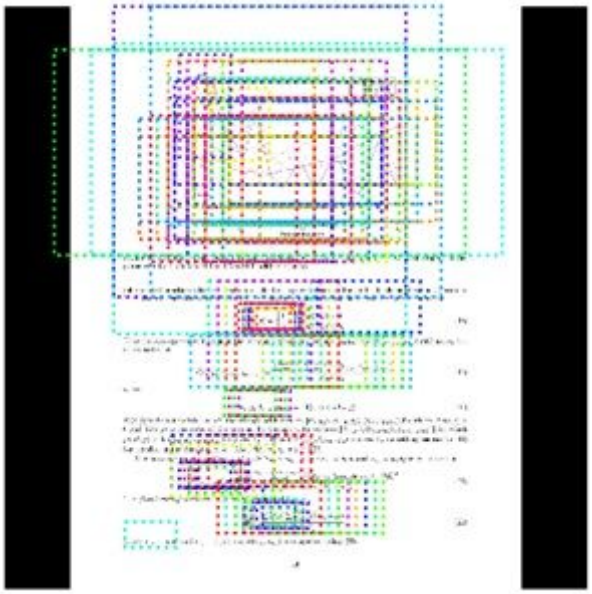

Figure 8

The diagram of the Anchor refinement

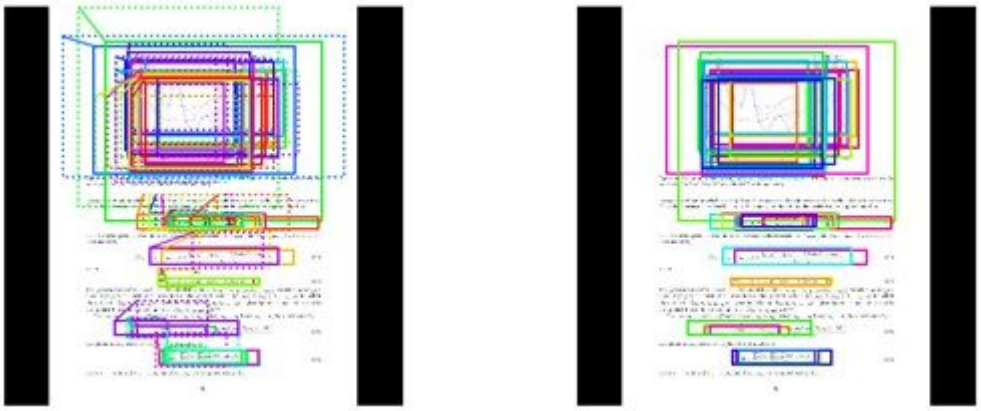

Figure 9

(a)The comparison of anchor frame and its modified anchor frame without non-maximum suppression (b)The diagram of the clipped anchor

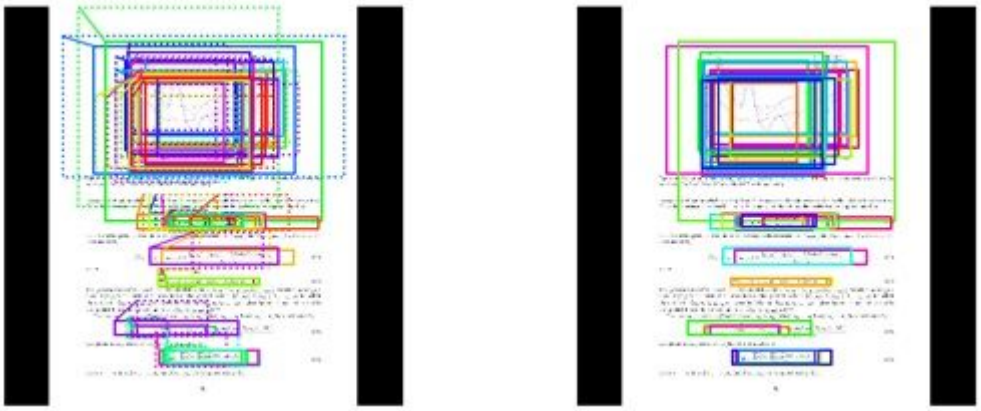

Figure 9

(a)The comparison of anchor frame and its modified anchor frame without non-maximum suppression (b)The diagram of the clipped anchor

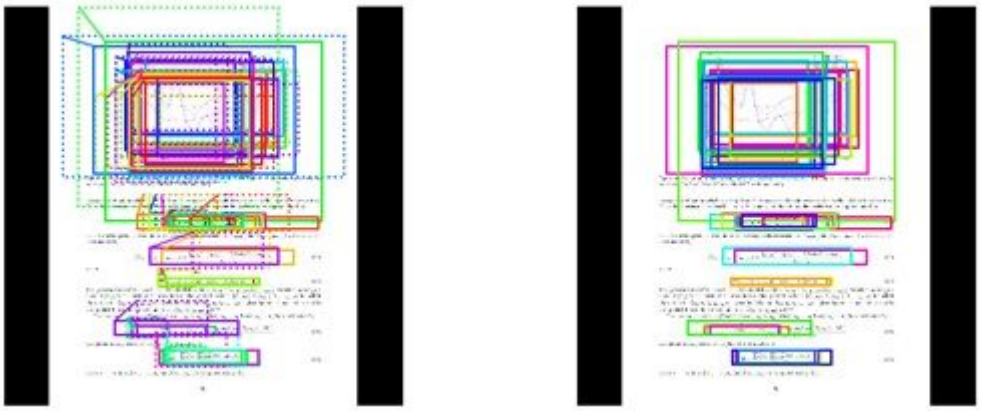

Figure 9

a The comparison of anchor frame and its modified anchor frame without non-maximum suppression
b The diagram of the clipped anchor

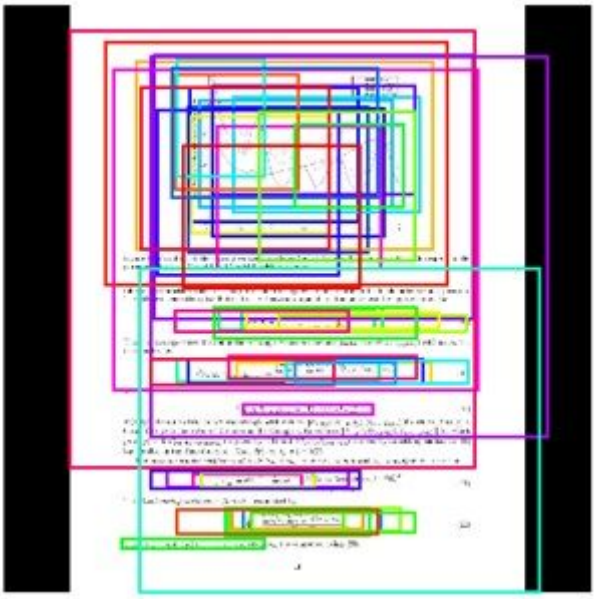

Figure 10

The Output reult of the anchors after NMS

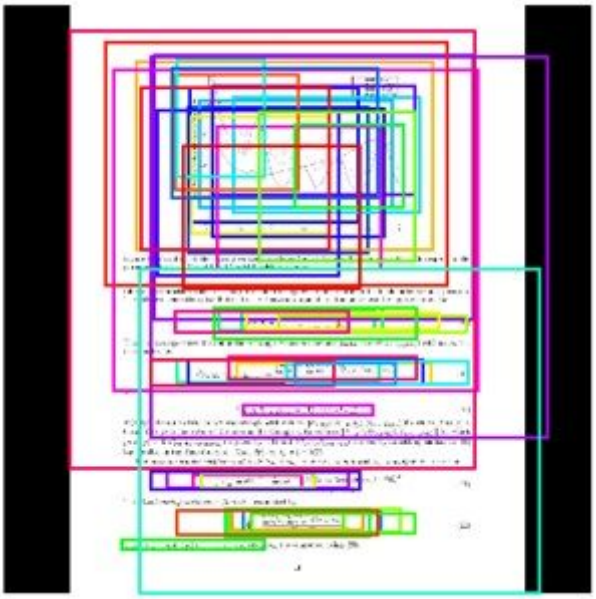

Figure 10

The Output reult of the anchors after NMS

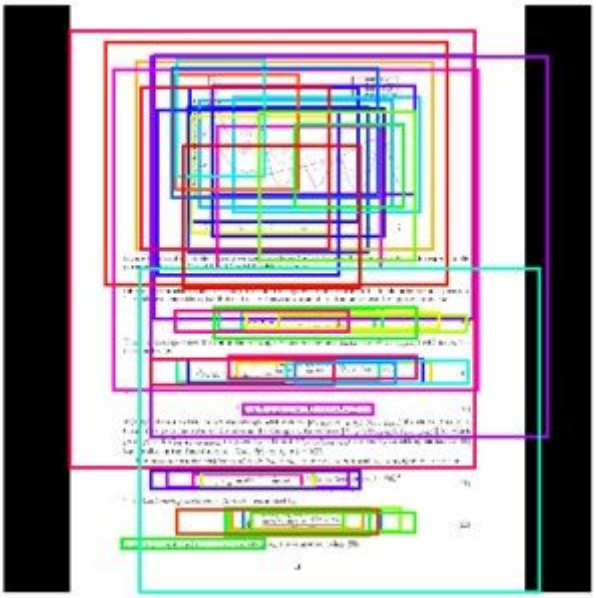

Figure 10

The Output reult of the anchors after NMS

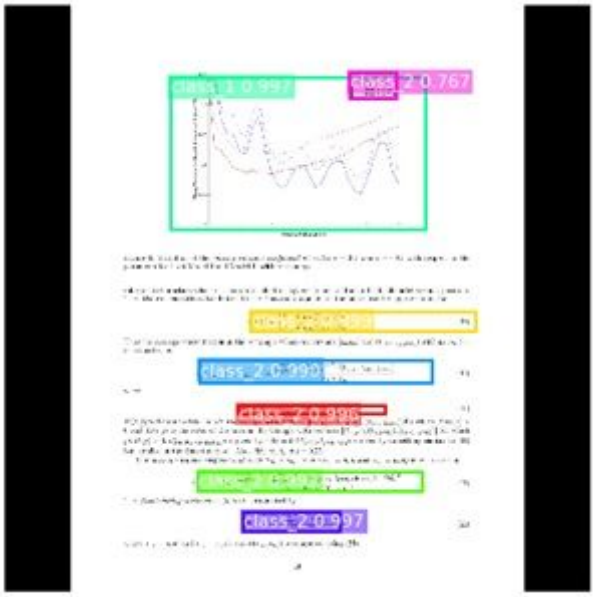

Figure 11

The Predicted borders and classification renderings

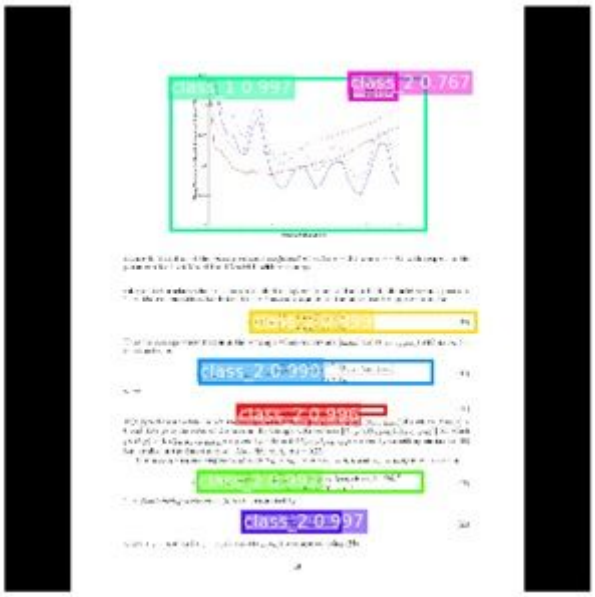

Figure 11

The Predicted borders and classification renderings

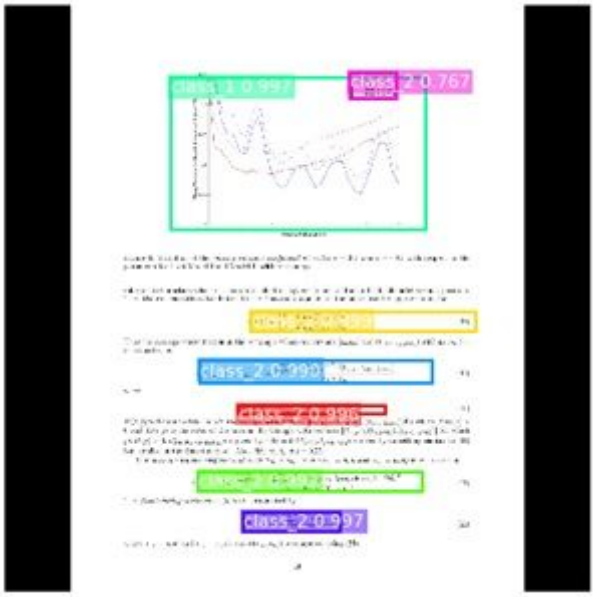

Figure 11

The Predicted borders and classification renderings

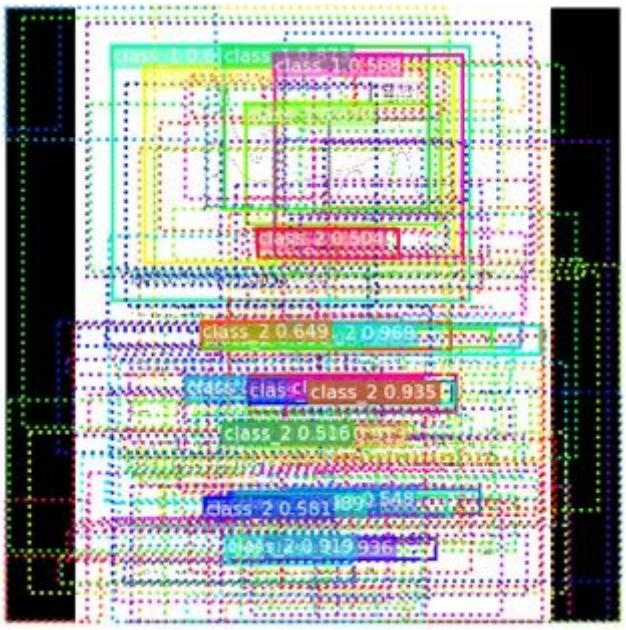

Figure 12

The ROIs before refinement

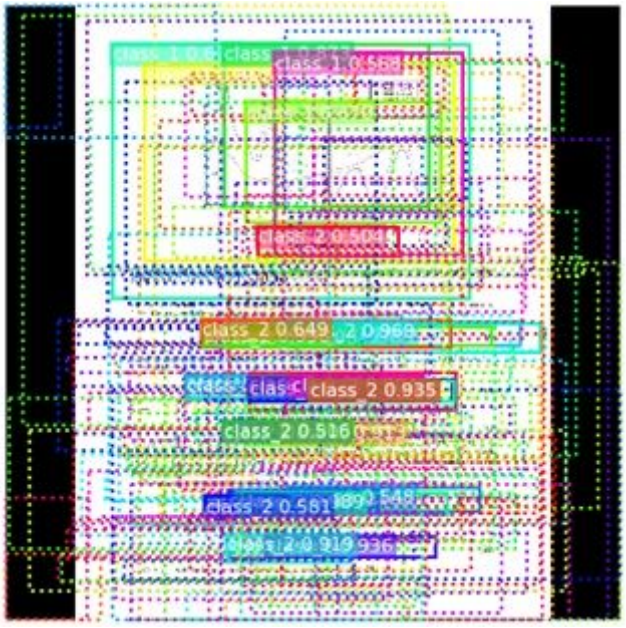

Figure 12

The ROIs before refinement

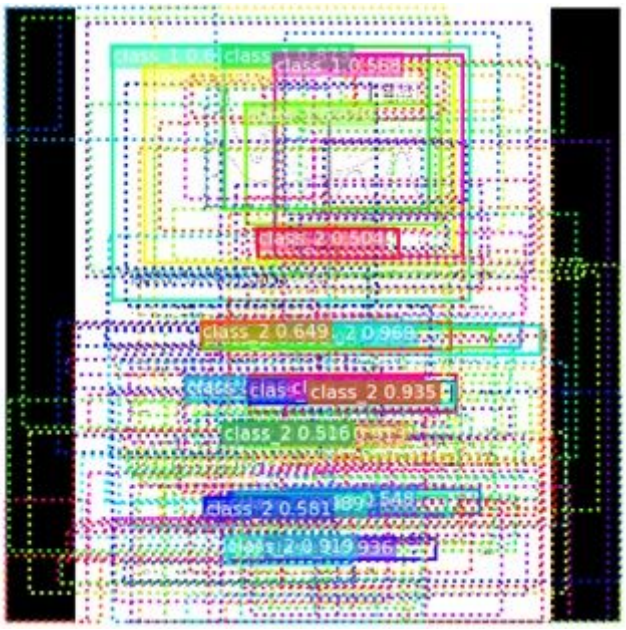

Figure 12

The ROIs before refinement

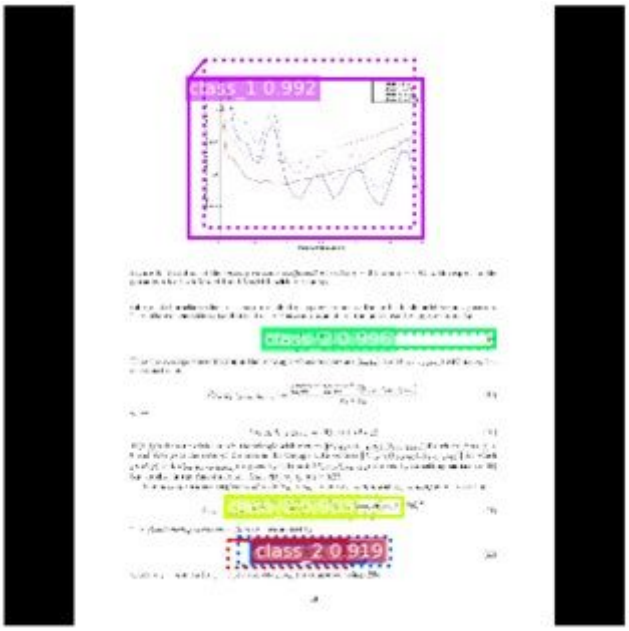

Figure 13

The ROIs after second refinement

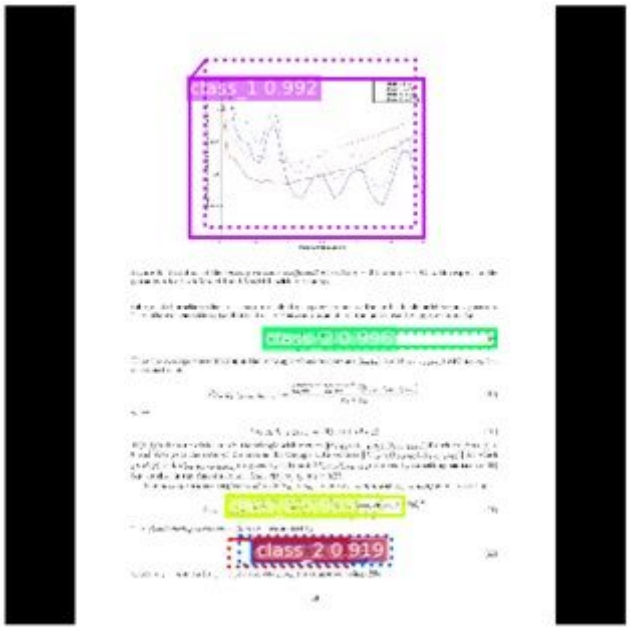

Figure 13

The ROIs after second refinement

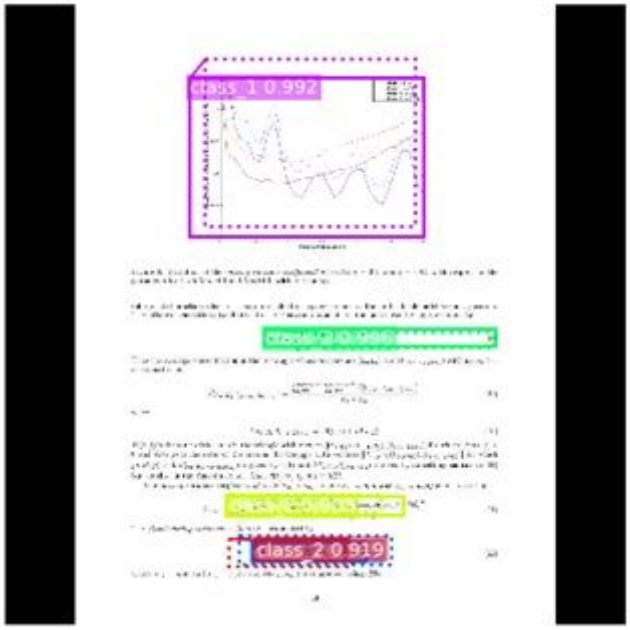

Figure 13

The ROIs after second refinement

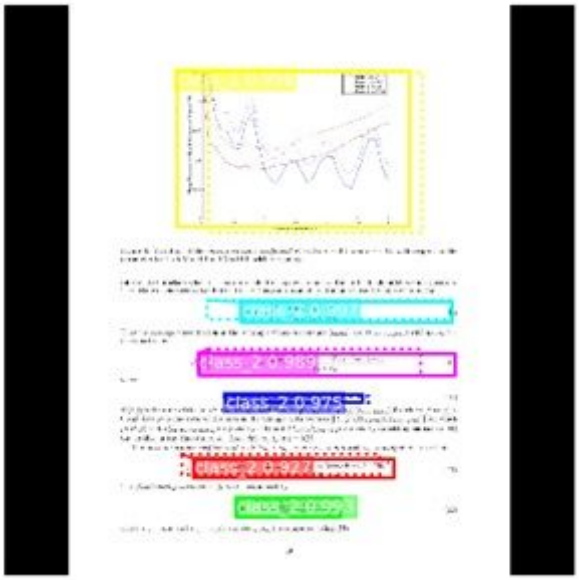

Figure 14

The ROIs after second NMS

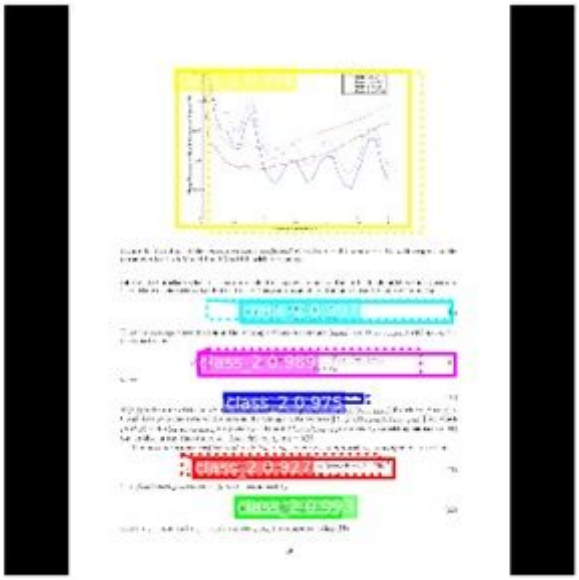

Figure 14

The ROIs after second NMS

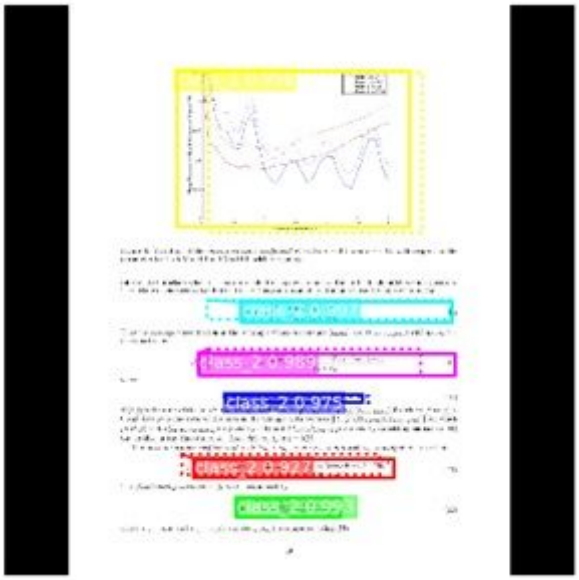

Figure 14

The ROIs after second NMS

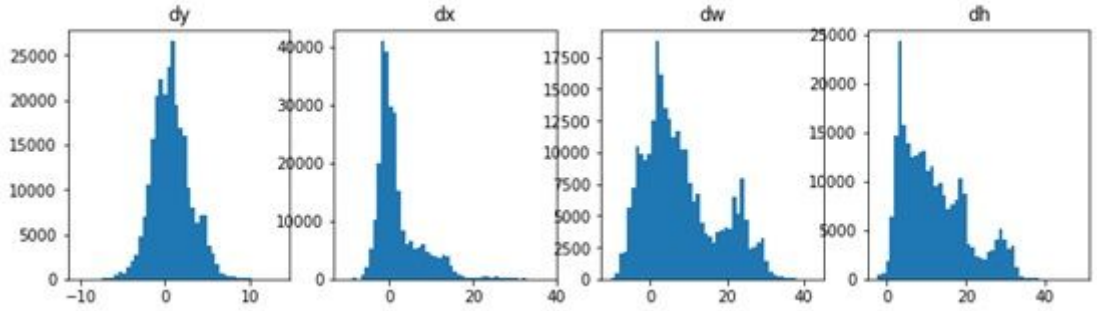

Figure 15

The Frequency chart of prediction results

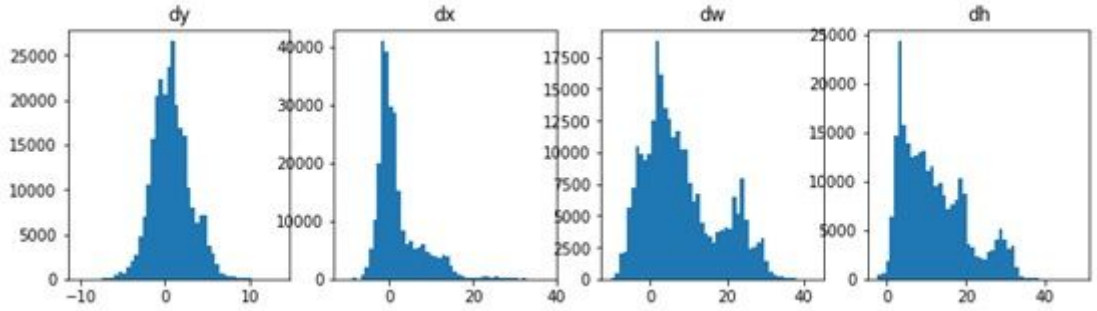

Figure 15

The Frequency chart of prediction results

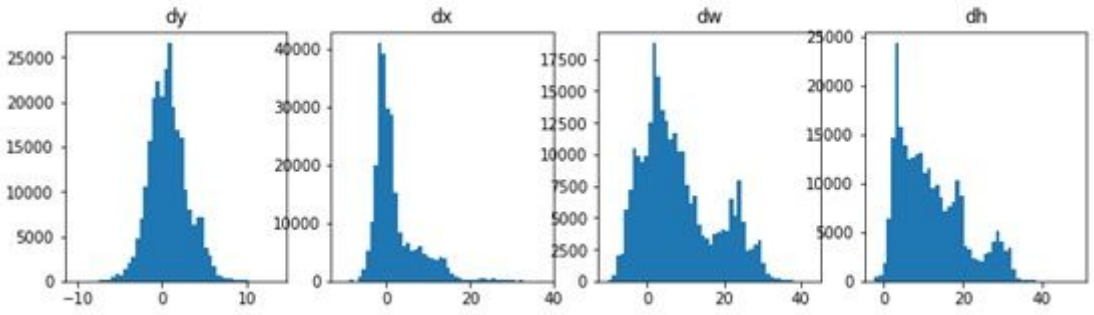

Figure 15

The Frequency chart of prediction results

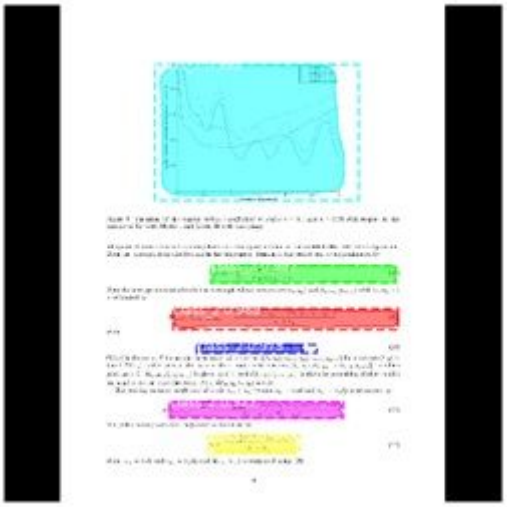

Figure 16

The diagram of the MASK

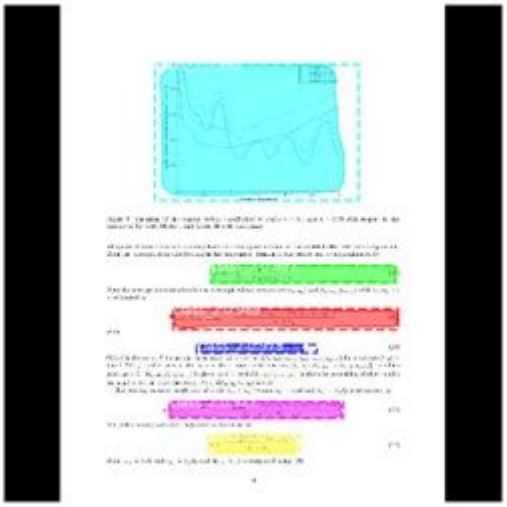

Figure 16

The diagram of the MASK

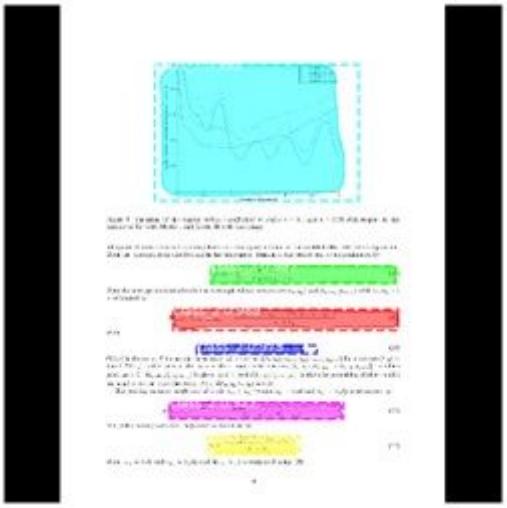

Figure 16

The diagram of the MASK

Figure 17

The Partial output of the full convolutional neural network activation layer

Figure 17

The Partial output of the full convolutional neural network activation layer

Figure 17

The Partial output of the full convolutional neural network activation layer

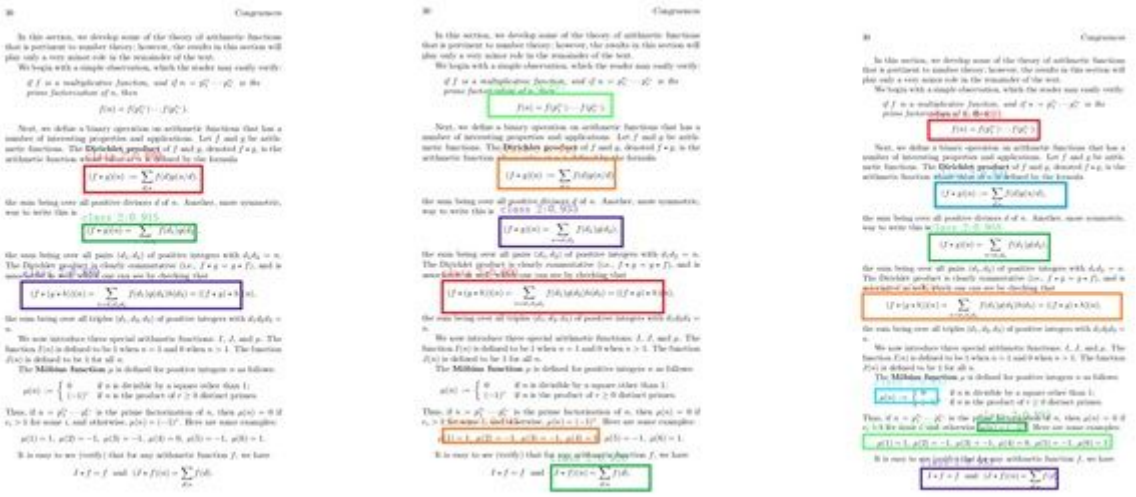

Figure 18

(a) (left 1) fast CNN verification diagram(b) (left 2) faster CNN verification diagram(c) (left 3) MASK RCNN verification diagram

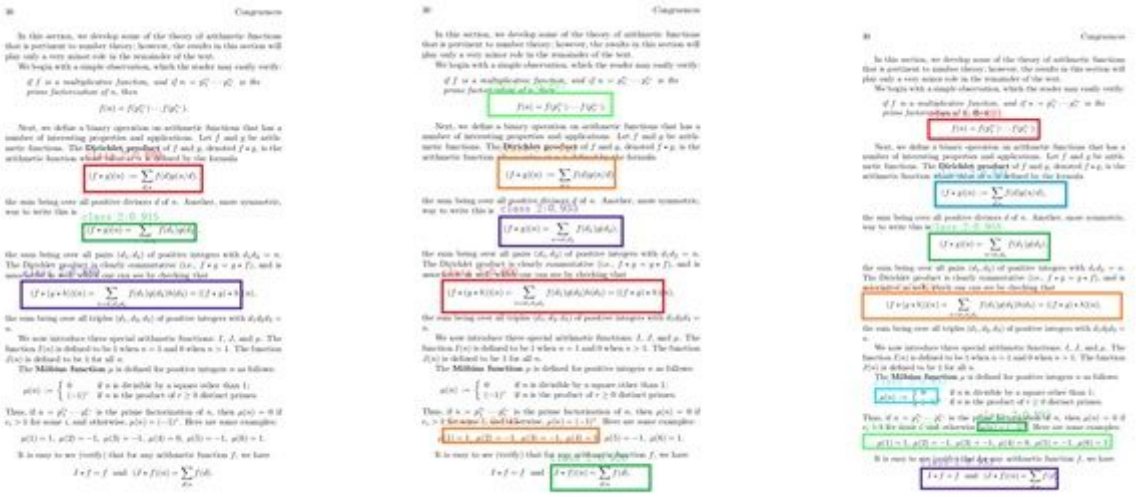

Figure 18

(a) (left 1) fast CNN verification diagram(b) (left 2) faster CNN verification diagram(c) (left 3) MASK RCNN verification diagram

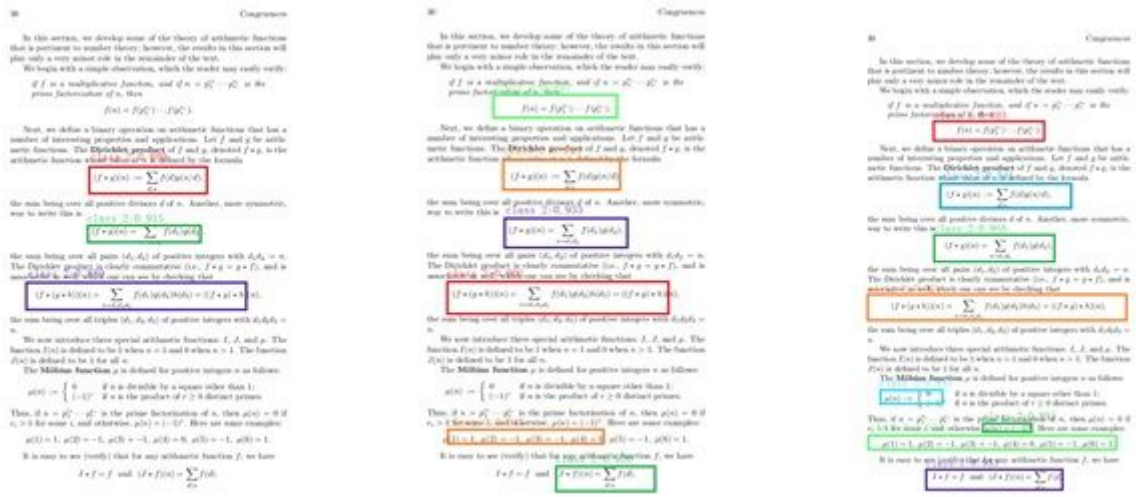

Figure 18

(a) (left 1) fast CNN verification diagram(b) (left 2) faster CNN verification diagram(c) (left 3) MASK RCNN verification diagram