Video-Based Object Detection in Security Monitoring System

by

Chao Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Object detection technology has been widely used in many real world applications. With the development of the deep learning method, the accuracy and speed of object detection method have been improved significantly, demonstrating great promises to increase the efficiency of security-related business activities. Nevertheless, the robustness of the existing object detection methods on security video datasets is still lacking. This could substantially reduce performance in complex application scenarios, such as changeable target size, target occlusion and bad weather. This cannot be solved perfectly by image-based object detection because a single image's information is limited. On the other hand, the video dataset consists of a series of still images of rich temporal and spatial information, which could be used as supplements for the detection methods. Based on this idea, this thesis proposes an incremental optimization method that solves the existing problems of the object detection method. We first improve the accuracy of the image-based object detection method by adding new features, and then aggregate the temporal and spatial information of the target to enhance the performance of the video-based object detection method. Furthermore, a multi-layer feature cascade aggregation pyramid structure is adopted based on the traditional Faster-RCNN model. The Faster-RCNN is one of the most famous convolution neural networks used in object detection and recognition tasks, which was firstly proposed in 2016. It replaced the traditional selective search method with the region proposal network (RPN), which improved detection speed significantly. Because of its excellent detection performance, many recent proposed approaches still selected it as the backbone network. The new multi-layer feature cascade aggregation feature pyramid network (MCA-FPN) combines the deep and shallow semantic feature information to optimize feature utilization and improve the feature representation ability of any size. In order to address the negative effects generated by the imbalanced distribution of samples, a sample asymmetric weighted loss function (SAW-Loss) is proposed, which improves the efficiency of the network training. Experimental results show that the proposed MCA-FPN and SAW-Loss modules can improve the mAP of traditional FPN by 2.4% and 1.5% respectively, and the final improved object detection algorithm with both of two modules obtains a mAP of 86.0% on Pascal VOC dataset which is higher than the mAP of 82.1% tested from FPN. The proposed method performs significantly better than most of the existing method, such as FCOS with a mAP of 78.7%, RFBNet with an mAP of 82.2% and PFPNet with a mAP of 84.1%.

Video-based methods may make use of two types of information: local information which is obtained from adjacent frames and global information which is extracted from whole video series. We propose two types of information aggregation methods, namely local information aggregation and global information aggregation based on the feature similarity and the attention mechanism, and so as to aggregate features selectively by including more of the correlated feature information and less of the uncorrelated feature information. As such, the network could extract and learn more useful target features and abandon the interfered features. The accuracy of the proposed local global information aggregation methods could be improved by 0.9% and 1.1%, respectively compared with one of the most advanced video-based object detection methods MEGA. By adding both two modules, the mAP of the proposed method reaches 84.6% on the public dataset ImageNet VID, which is 1.7% higher than the mAP of MEGA. The proposed method also demonstrates potentials to detect occluded targets with high confidence.

Acknowledgements

I would like to thank those who supported and helped me during my postgraduate program at the University of Waterloo. First, I would like to gratefully thank my supervisor, professor Dayan Ban and Zhou Wang, who offered me the chance to study here and gave me lots of important advice on my research. Besides, I would also like to thank all lab mates of Prof. Dayan's group for sharing new ideas with me. Second, I would like to thank professor Guangqiang Yin, who encouraged me and shared his computation power with me. Also, I would like to thank all group members of Prof. Yin's group who proposed a lot of useful suggestions about my experiment work. Finally, I would like to acknowledge my parents and my girlfriend Xinru, who supported me with their encouragement and selfless care.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AP Average Precision 17

ATSS Adaptive Training Sample Selection 14

CNN Convolution Neural Network 6

DFF Depth Feature Flow 31

Faster-RCNN Faster Regions with CNN 32, 33

FGFA Flow-Guided Feature Aggregation 31, 53

FN False Negative 17

FP False Positive 17

FPN Feature Pyramid Network 12, 26–29, 32, 37–39, 46, 47

FPS Frames Per Second 18

FSAF Feature Selective Anchor-Free 14

ILSVRC ImageNet Large Scale Visual Recognition Challenge 18, 19

IoU Intersection over Union 13–15, 17, 35, 44, 45, 48, 65

mAP mean Average Precision 17

MCA-FPN Multi-layer Feature Cascade Aggregation Feature Pyramid Network 32, 33, 45–48, 50

MEGA Memory Guided Attention for Category 54–56, 62–64

NMS Non-Maximum Suppression 15, 17, 23

P-R Precision-Recall 17

PASCAL VOC PASCAL Visual Object Classes 18, 19, 45, 47, 48

R-CNN Regions with CNN 23, 24

ROI Region of Interest 23, 24, 33, 41, 56, 59, 60

RPN Region Proposal Network 26, 33–36, 41, 43, 54, 56

SAW-Loss Sample Asymmetric Weighted Loss Function 44, 45, 47, 48, 50

SFN Sample Fusion Network 13

SPP Spatial Pyramid Pooling 24

SSD Single Shot MultiBox Detector 28

SVM Support Vector Machine 23, 24

TN True Negative 17

TP True Positive 17

Chapter 1

Introduction

1.1 Background

Artificial intelligence has attracted great attention from both academic and engineering fields in the past few decades, especially with the exponentially growing interest in deep learning methods since 2012, when the deep learning-based method [33] got an impressive achievement in the ImageNet [13] picture recognition competition which is one of the most authoritative competitions in computer vision. This method performed much better than other traditional methods and attracted many researchers' interests to explore it further [35]. Since then, deep learning technologies have been rapidly applied to many real world applications, including medical test, auto-driving, security monitoring and et al, where it often delivers great results. As one of the most popular and mature field in artificial intelligence, computer vision can be found everywhere in our lives and have a great development prospect. In particular, many researchers have made many major contributions to its basic subdomain object detection, which constructs a strong foundation for other subdomain research.

There are two main tasks in object detection: image classification and object localization. Image classification aims to focus on the feature description of a detected object and classify which specific categories the object belongs to. Object localization aims to confirm the specific location of objects in the original image and make sure that all objects are labelled by different sizes of bounding boxes, as demonstrated in Figure 1.1.

In practical application, the object detection method needs to confirm the size, location and categories of the object from input images. However, in realistic scenarios, the size,

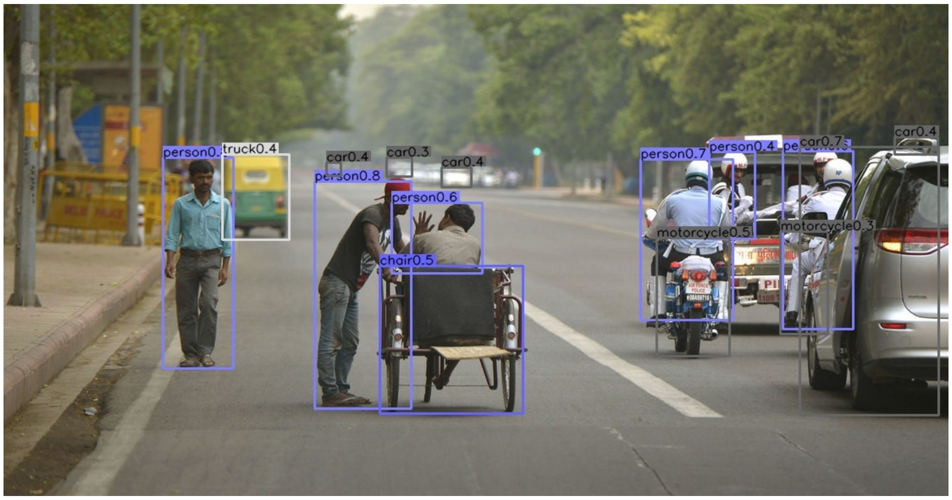


Figure 1.1: Examples of the bounding box

number, location and categories of objects are all variable, and it is more likely to contain other influential factors such as weather change, too weak or strong lightness, occlusion and et al. All of these internal and external causes make object detection more difficult and complex.

Smart security technology improves the application of object detection, which could be explained by two aspects. Firstly, the fundamental information construction provides object detection technology with adequate data resources from different kinds of security monitoring devices, which is crucial for the recent data-driven object detection algorithm. Second, the appearance of the deep learning method, especially the convolution neural network, enhances the performance of object detection algorithms and has advanced ahead of other classical algorithms [33]. By introducing the deep learning method, the algorithm could increase the accuracy and efficiency of object detection, which is a perfect fit for the current demand for real-time and precise detection.

Despite the progress made by object detection, there still exist limitations in its application. Firstly, as shown in Figure 1.2, it is common to find blurred or occlusion objects in footages, which are not included in the normal dataset used for algorithm training. Therefore, because of the lack of such "hard examples," the algorithm could not learn the feature of similar objects hence reducing their performance on them. Secondly, although the object detection algorithm has proved that it can perform very well on objects that are large in size, it still struggles in detecting small-sized objects. The results of the mainstream object detection algorithm on the multi-scale object are shown in Table 1.1. It can be easily seen that the accuracy is very low.

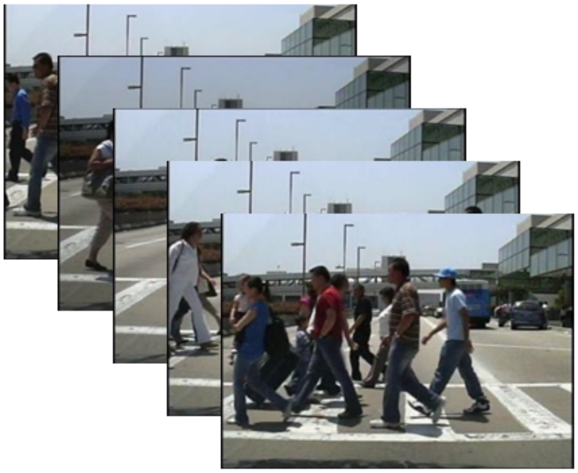


Figure 1.2: Occlusion and blur examples

Table 1.1: The performance of mainstream object detection algorithm on multi-scaled object

Mothods	Backbone	AP	AP50	AP75	APS	APM	APL
Fast-RCNN	ResNet-101	27.2	48.4	-	6.6	28.3	45
R-FCN	ResNet-101	30.5	52.9	31.2	12	33.9	43.8
YOLOv2	ResNet-101	21.6	44	19.2	5	22.4	35.5
YOLOv3	ResNet-101	28.2	51.5	29.7	11.9	30.6	43.4
SSD	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8

In fact, the solution to these problems could be the core point for how object detection can be better applied in the field of smart security. The traditional object detection method based on the image could maximize the use of feature information from a single image. But the information extracted from a single image is limited, and the performance of the algorithm on the hard examples will not improve, regardless of how effective the algorithm is since the information on occlusion and blurred objects is not available for training. Therefore, researchers have gradually started to transfer their attention from still image-based detection to video-based detection. Video data is composed of a large number of images with temporal, spatial relationships and content relevance, and such different dimensional relationships could be used as a supplementary to support the algorithm in solving the problems described. In the meantime, with the improvement in the hardware, the smoothness of video is getting higher, and over hundreds or even thousands of images can be produced from a tensecond video frame. These also allow the model to learn more detailed feature changes and improve its performance.

1.2 Limitations of Existing Target Detection Algorithm

The current image-based target detection has made relatively good achievements and performs well on most of the public datasets. Besides, such technology has also been well-applicated in the field of smart security, which helps to alleviate the pressure from the government and some organizations. However, the performance of these algorithms could decrease when they face some hard examples, which are commonly seen in realistic security scenarios, such as the case of target occlusion and obscure. This is because the amount of information included in a still image is too limited for such hard examples, and the network cannot extract and learn enough features for detection tasks. In comparison, video data contains more information than still image data does, which could be an ideal direction to solve these problems. However, if we apply the image-based target detection algorithm directly to video data, it is still hard to extract some specific feature information that is only contained in video data.

Therefore, the purpose of this thesis is to find an efficient and accurate target detection method which could be used to solve the existing problem of target detection in smart security. The problem could be attributed to the following points:

- The limited feature information extracted from hard examples is far from enough for algorithm training.
- Although there are algorithms that focus on the extraction of different dimensional relationship among nearby frames, they just simply organize the pixel data without providing enough semantic and correlation analysis. This leads to a relatively low data usage efficiency and inaccurate result.
- The most dataset contains a large proportion of easy samples and a limited proportion of difficult samples, and the hard examples could be easily treated as difficult training samples by the network. If we want to detect these hard examples, the network training process should pay more attention to them. However, in most cases, network

could easily ignore such unbalanced distribution of easy and difficult examples and lead to slower model convergence speed and lower detection accuracy.

1.3 Related Work

1.3.1 Convolution Neural Network

Computer vision aims to give the computer the ability to observe and analyze like animals; one of the representative fields is image processing. Traditional image feature extraction algorithms have poor accuracy and limited robustness. Gradually, as inspired by the process of human vision, researchers propose to find a similar way for the computer to process images. Therefore, the convolutional neural network was proposed, which refers to the transfer mode of vision never cell and provides computers with the function of cognition and processing of images.

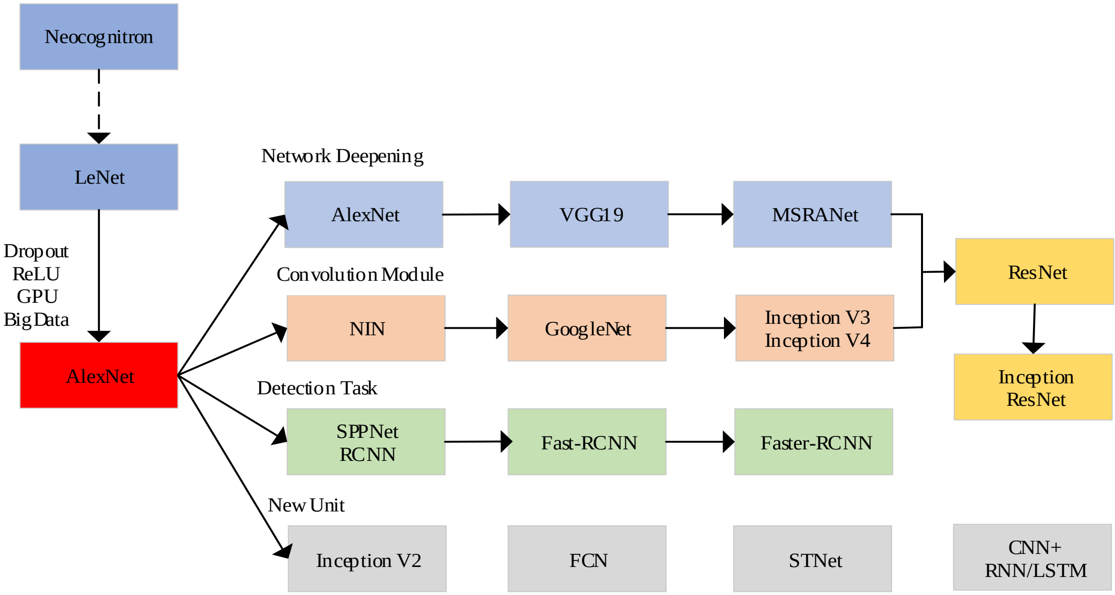


Figure 1.3: The brief history of convolution neural network

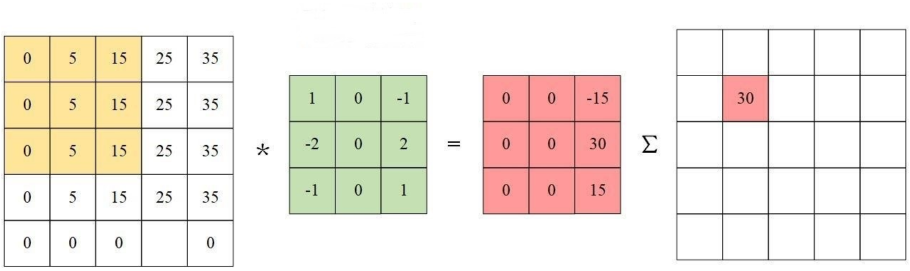
As shown in Figure 1.3, convolutional neural networks have been improved by the expansion of the depth and width of the network. From the first proposal of the Neo-cognitron neural network to the classical LeNet, it improves recognition performance under supervised learning. Besides, the proposal of the LeNet model firstly unified the basic backbone structure of a convolutional neural network with multi-layer stacked and fully-connected layers for classification. With the development of computation power and the increase of the amount of dataset, the advantage of convolution neural network is emphasized, which attracts more researchers to study it. Until now, based on the contributions from researchers, the basic structure of Convolution Neural Network (CNN) has been confirmed, and the improvement of CNN tries to change some modules without changing the whole structure. When an image is sent to the network as input, it first passes through the multi-layer convolution operations and pooling operation to extract different dimensional features with suitable sizes. The network then sends the extracted feature to the fully-connected layer for detection and classification, and finally, it generates probability results as an output layer.

Input Layer

The concept of the design of the input layer in convolutional neural networks is based on the operations in machine learning to optimize images by image preprocessing operations. The main operations include: averaging, normalization and PCA-SVD dimensionality reduction. The averaging method refers to the summation averaging method, which eliminates the effect of extreme value. Normalization is to restrict the range of data to a specific value domain. The dimensionality reduction operation is to remove the influence of too many image dimensions and ensure low correlation in each dimension of features.

Convolution Layer

The essence of the convolution layer is to use a certain size of convolutional kernel sliding on the original image or feature maps, and the convolution layer calculation could be attributed as a type of matrix calculation. The entire process of the convolution operation is explained in Figure 1.4. The convolution kernel is a type of weight matrix, and in this example, the size of the convolution matrix is 3x3. The yellow input matrix is multiplied by the convolution kernel, which generates the output matrix highlighted in red. After that, the value from the output matrix will be sent to the corresponding location of the feature map. This establishes the connection between the feature map and the input image. This example illustrates the simpler process of the convolution operation where the deeper convolution layer is just to repeat this process to obtain different dimensional features.



$$\begin{array}{|c|c|c|c|c|}\hline 0 & 5 & 15 & 25 & 35 \\ \hline 0 & 5 & 15 & 25 & 35 \\ \hline 0 & 5 & 15 & 25 & 35 \\ \hline 0 & 5 & 15 & 25 & 35 \\ \hline 0 & 0 & 0 & & 0 \\ \hline \end{array} * \begin{array}{|c|c|c|}\hline 1 & 0 & -1 \\ \hline -2 & 0 & 2 \\ \hline -1 & 0 & 1 \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 0 & 0 & -15 \\ \hline 0 & 0 & 30 \\ \hline 0 & 0 & 15 \\ \hline \end{array} \ \Sigma \ \begin{array}{|c|c|c|c|c|}\hline & & & & \\ \hline & 30 & & & \\ \hline & & & & \\ \hline & & & & \\ \hline & & & & \\ \hline \end{array}$$

Figure 1.4: The matrix calculation of the convolution layer

Figure 5 is another example that illustrates the convolution operation from another aspect. The bottom part represents the input image, and the top part represents the output matrix. At the start of the convolution operation, the input image expands its size by padding operation to ensure that the convolution kernel could slide on it with an integer result. Besides, the connection between the bottom part and top part represents the convolution operation and the corresponding location relationship. It is obvious that the size of the input matrix and the output matrix could be different, where such difference is generated by a different set of padding and step length since people believe that in this way, the convolution layer could extract the different types of feature.

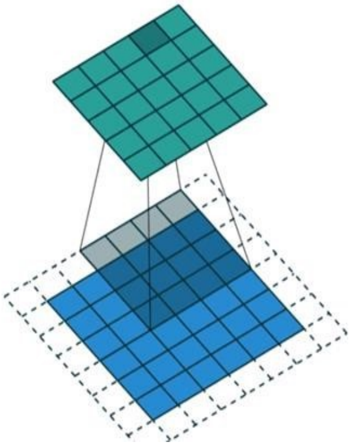


Figure 1.5: The process of convolution kernal sliding on the feature map

Activate Function

In general, functional relationships could be divided into two categories: linear and non-linear. Most traditional regression models are limited to modelling linear relationships. However, most target detection and recognition tasks require models to have strong non-linear expression ability, and therefore, the activate function is produced.

In fact, the activate function plays a very important role in a convolution neural network which is used to ensure the strong non-linear expression ability of the network and build the connection between the input data and the final classification task. Some examples of mainstream activation functions are presented in the following sections.

Sigmoid activate function: The shape of the sigmoid activate function is like the letter S, which has monotonically increasing properties in its independent variable interval. As shown in Equation 1.1 and Figure 1.6, the value of the sigmoid could range from 0 to 1 and have both top bound and bottom bound. However, since the derivation value of the sigmoid function will be 0 at both the bottom bound and top bound, it may generate a problem that we call 'gradient vanishment' which could pose a negative effect on the efficiency and convergence speed of the network.

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{1.1}$$

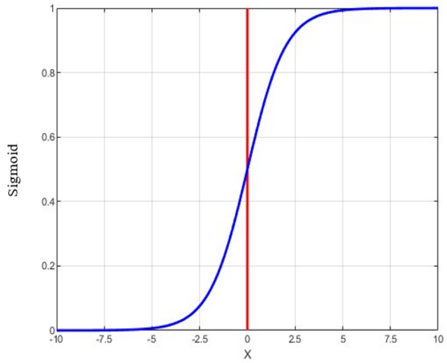


Figure 1.6: Sigmoid activate function image

Besides, as the Sigmoid activate function is not zero-centered, it may also cause the slow speed of network training convergence. To improve the training covergence speed, the tanh activate function is proposed, which is improved based on the sigmoid function.

It expands the value range of the dependent variable to the negative value region, which ensures the fast and smooth convergence process of the model. However, the problem of gradient vanishment in sigmoid function still exists. The equation and shape of the tanh activate function are shown in Equation 1.2 and Figure 1.7, respectively.

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1.2}$$

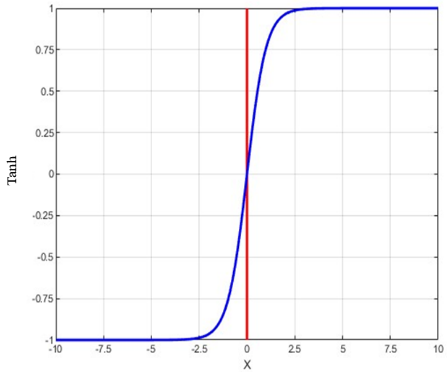


Figure 1.7: Tanh activate function image

ReLU activate function and its transformation: ReLU activate function solves the gradient vanishment problem by a type of design of judgement based on the input value. The equation and shape are shown in Equation 1.3 and Figure 1.8, respectively. It gives a constant gradient value when the input value is greater than 0 and gives a 0 gradient value when the input value is no more than 0. In this way, for those inputs whose value is no more than 0, the network will not activate these nerve cells, and thus they will not be calculated.

$$ReLU = \max(x, 0) \tag{1.3}$$

But the input with a negative value is still likely to contain some useful information, and such a rough approach will lead to inefficient network performance. Therefore, in the improved Leaky ReLU function, it gives relatively small values to those negative inputs instead of 0 so that the problem of neuronal inactivation could be addressed. The detail of the Leaky ReLU activate function is shown below:

$$LeakyReLU = \max(x, 0) + leak \times \min(0, x) \tag{1.4}$$

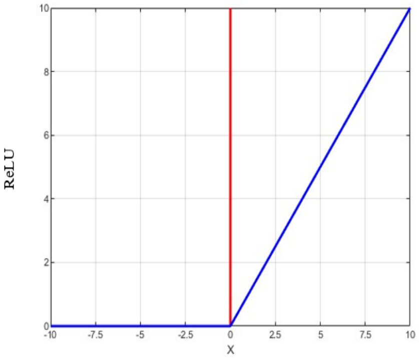


Figure 1.8: ReLU activate function image

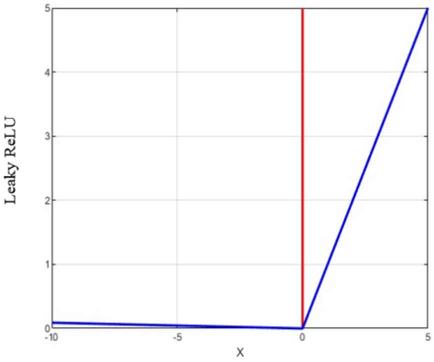
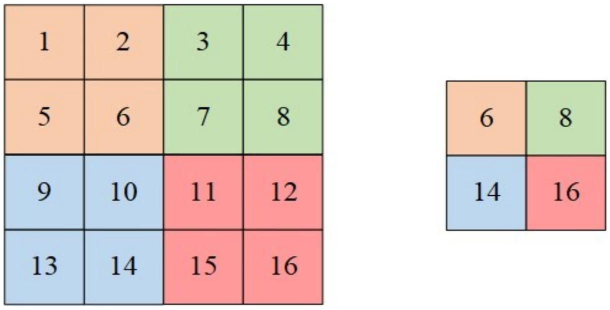


Figure 1.9: Leaky ReLU activate function image

Pooling Layer

The Pooling layer is a form of down-sampling. There are various forms of non-linear pooling functions, and ”Max pooling” is the most common one (shown in Figure 1.10). It involves dividing the input image into several rectangular regions and outputting the maximum values for each subregion. Intuitively, this mechanism is effective because after a feature is found, its exact position is much less important than its relative position to other features. The pooling layer continuously reduces the spatial size of the data, so the number of parameters and the computational effort decreases, which also controls overfitting to some extent. In general, pooling layers are periodically inserted between the convolutional layers of CNNs.



1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

6	8
14	16

Figure 1.10: The process of max pooling layer

Fully-connected Layer

The fully-connected layer aims to transfer the two-dimensional feature map output from the previous layer into a one-dimensional vector for the final classification task, and each node of it is connected to the node of the previous layer. The structure of the fully-connected layer is shown in Figure 1.11.

The main role of the fully connected layer is to compress the features extracted from the input image after convolution and pooling operations and to complete the classification function of the model according to the compressed features. It plays the role of ”classifier” in the whole convolutional neural network. If the convolutional layer, pooling layer and activation function operations aim to map the original data to the hidden feature space, then the fully connected layer maps the learned feature representation to the sample labelling space, which could reduce the influence of feature location on classification results and improve the robustness of the whole network.

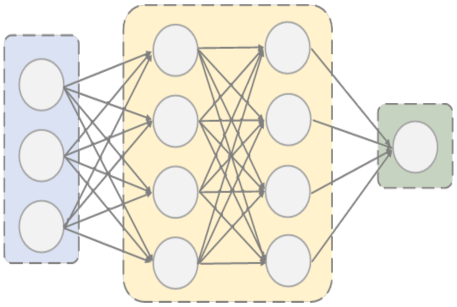


Figure 1.11: The structure of fully-connected layer

1.3.2 Feature Pyramid

As mentioned earlier, after the convolution and pooling operation, the network will generate a lot of feature maps of different size, which contains different dimensional features. In fact, it is key to understand how to use these feature maps well for a better performance of the network. Based on the above premise, the feature pyramid network was proposed by Lin et al. in 2017, which uses bilinear interpolation up-sampling method to expand the resolution and then stacked downward layer by layer. In this way, the deep feature map could be fused into the shallow feature maps, and the network could achieve the fusion of multi-scale features. Such design can significantly improve the detection accuracy and efficiency of the network on multi-scale targets.

The structure of the Feature Pyramid Network (FPN) is shown in Figure 1.12(a). The left pyramid is a traditional convolution process on the input image that is used to extract features from different sizes of feature maps. During this process, the operation of down-samplings will decrease the resolution value and extract more semantic feature information. In other words, more texture and localization information will be extracted at the low pyramid layers, and more semantic and related information will be extracted at the top layers. As for the up-sampling operation, the detail of this process is shown in Figure 1.12(b). There are two sources for each new generation of the feature map. The corresponding feature map on the left side is firstly passed through by a 1x1 convolution kernel, which is used to ensure the channel of the convolution layer from different sources is the same. Then, the feature pyramid fuses the two different feature maps by adding up the pixel value from each point.

Woo et al. [70] proposed a deconvolution method to replace the bilinear interpolation for upsampling, which was shown to have better results. Therefore, de-convolution operation gradually became the most commonused operation for up-sampling, and such method was

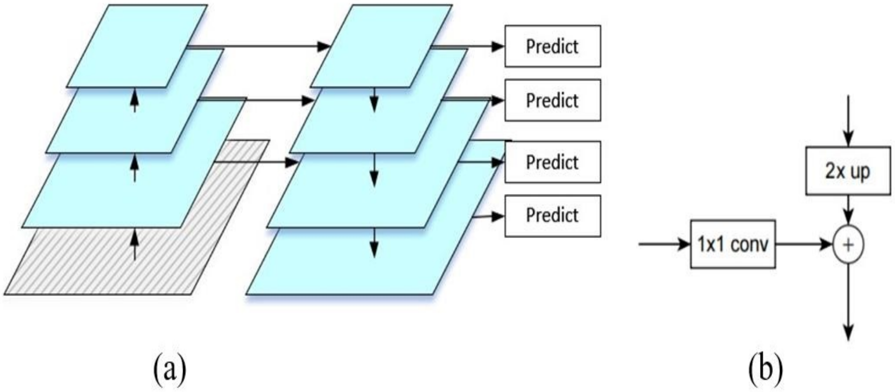


Figure 1.12: The Feature Pyramid (a) the structure of feature pyramid (b) the detail of feature fusion

also used by RefineDet [77], and HyperNet [32] used deconvolution to up-sample deep feature maps and constructed a network model that could generate high-resolution feature maps for feature map enhancement and multi-scale fusion.

The anchor frame mechanism was firstly proposed by Fast RCNN in RPN networks and has been widely utilized in various two-stage and single-stage detection algorithms since then. Unlike the traditional sliding window method, which uses a fixed size window for target detection, the anchor frame mechanism proposed by Fast RCNN uses nine predefined anchor frames with different sizes, lengths and widths as prior information and ensures all targets could be included in these anchors. As shown in Figure 1.13, there are three different anchors, which represent the big size, middle size and small size, respectively. Besides, with the different sized anchors, they also have different length-to-width ratios of 1:1, 1:2 and 2:1, respectively. During the detection period, the network trains anchors based on the feedback result of Intersection over Union (IoU), which is calculated by the ratio of the coincident area between predicted anchors and the ground truth box.

To improve the performance of the anchor mechanism on target detection tasks, researchers have started to focus on it and produce some good results. Zhang [78] proposed a scale-compensated anchor frame matching strategy in Sample Fusion Network (SFN) networks by assigning anchor frames of different scales to different detection layers for targets of different shapes and scales, which improves, to a certain extent, the problem of manually designed anchor frames being ineffective for small-scale face detection.

Wang concluded that the manually designed anchor frame mechanism cannot cover those targets with unusual shapes, such as particularly thin and wide targets, so they

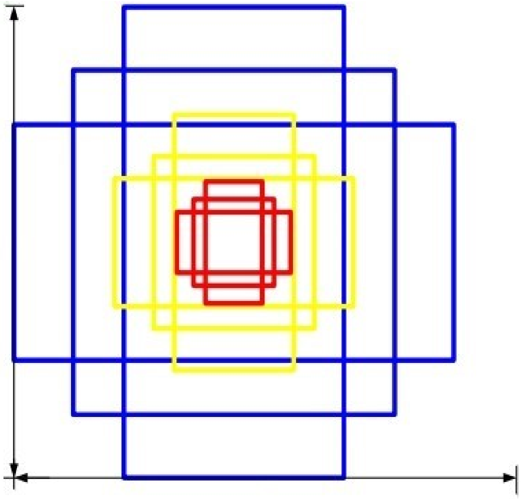


Figure 1.13: The region proposal of anchors mechanism

proposed the GA-RPN [67] network. A dynamic anchor frame generation strategy is designed in GA-RPN to learn the location and shape of anchor frames from the semantic information of the image itself. As a result, the GA-RPN gets some improvement on both two-stage and single-stage detection algorithms and reduces the number of anchor frames and model parameters.

Zhu proposed a Feature Selective Anchor-Free (FSAF) module [83] based on RetinaNet to address the sample imbalance problem in the anchor frame mechanism algorithm. By adding a branch of anchor-free detection to the anchor frame mechanism and combining the respective advantages of anchor-free and anchor-based detection, the ability of the model to detect hard examples is increased.

However, such a method introduces too many extra parameters, which decreases the efficiency of network training. Zhang [76] found a limitation concerning the anchor-frame mechanism. The anchor-frame-based algorithm recognizes the positive samples based on the value of IoU, which may result in having those anchor frames with low IoU values recognized as negative examples even if their locations fall into the part of the ground truth box.

Due to the above problem, they propose an Adaptive Training Sample Selection (ATSS) strategy, which introduces parameters to supervise the quality of anchor frames. This increases the number of positive samples selected and improves the detection performance of the model.

Ke [30] pointed out that the common target detection algorithm is to perform classification and localization tasks based on fixed candidate anchors, but this setup limits the optimization effect. Therefore, they propose a multi-anchor learning strategy which selects representative anchor frames based on the confidence of classification and localization to generate anchor bags. Then they use the anchor from the anchor bags to update parameters for a better optimization performance.

1.3.3 Assessment of Target Detection Algorithm

Evaluation of Localization Performance

IoU is used to measure the localization performance of the detection algorithm, defined as the ratio of the intersection area and concurrence area, which is shown in Equation 1.5 and Figure 1.14. Its value is obtained by the overlapping area of the bounding box divided by the union area. It calculates the intersection and union area of the predicting and ground truth boxes. A high value of IoU represents a great coincidence degree of the model result and ground truth, which means the outstanding detection performance of model.

$$IoU = \frac{AreaofOverlap}{AreaofUnion} \tag{1.5}$$

Bounding Box and Non-maximum Suppression

Object detection algorithms that rely on bounding boxes record image areas to obtain the classification and localization of each object, where the features in the areas are either predicted or actual target features. Correspondingly, the ground-truth labels determine the coordinates of the object in the images and attach the object classification within that coordinate range. These bounding boxes can distinguish objects from the background or other objects.

The prediction process of the network models will not be accurate, so for a target to be detected, there may be multiple prediction bounding boxes. The design of the Non-Maximum Suppression (NMS) is to solve the problem of repeated bounding boxes, as illustrated in Figure 1.15. By calculating the size of the IoU value, a certain number of high-quality bounding boxes with the best confidence scores are reserved to reduce the computational redundancy caused by too many low-quality bounding boxes.

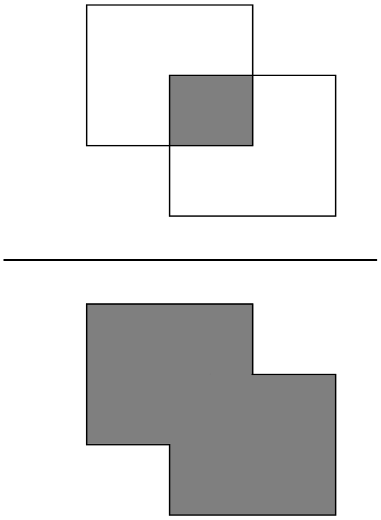


Figure 1.14: The vaule of Intersection-over-Union is obtained by the ratio of the intersection area and union area

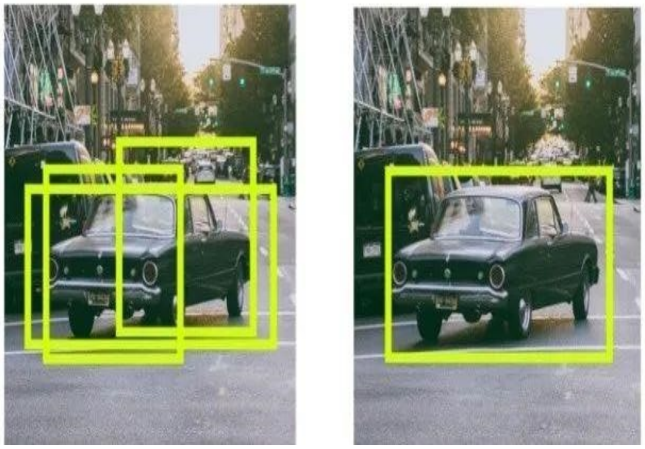


Figure 1.15: An example of Non-Maximum Suppression

Object detection networks compute a likelihood classification confidence score for each bounding box that belongs to a certain category. In the screening process of the NMS, the bounding boxes corresponding to samples with lower scores are discarded first. Second, the bounding box with the highest IoU value with the current likelihood score is discarded, this step is to reduce repetitions. Finally, the remaining bounding box scores are sorted, and the maximum value is taken. This process is repeated until all bounding boxes that need to be preserved are found.

Evaluation Metrics for Object Detection

In the classification task, the output result of the classification model is used as a category of its classification, and the output is compared with the real category label to determine whether the prediction is correct. The prediction can be divided into True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), which represent examples of positive and negative samples being correctly or incorrectly classified, respectively. The output result of the object detection task is different from the classification. The result contains the information that a certain position in the image belongs to a certain category, and it is likely to contain multiple targets. In fact, it is unrealistic for the output results to be completely consistent with the annotation labels. To determine the correct result in object detection, it is necessary to preset the IoU threshold and then determine the detection frame by category.

For object detection tasks, precision and recall can be calculated for the category to be detected. Precision represents the proportion of the part that the classifier considers to be a positive class and is indeed a positive class to the proportion that the classifier considers to be a positive class; Recall represents the proportion of the part that the classifier considers to be a positive class and is indeed a positive class to all the positive classes. The calculation formulas are shown in Equations 1.6 and 1.7 , respectively:

$$Precesion = \frac{TP}{TP + FP} \tag{1.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{1.7}$$

Through reasonable calculations, a Precision-Recall (P-R) curve can be obtained for each class, and the area under the curve is the value of Average Precision (AP). The key metric mean Average Precision (mAP) is used to measure the average accuracy of the network under various types. In addition, some metrics are usually not used as general regulations, including network detection speed metrics Frames Per Second (FPS), missed detection rate and false detection rate.

1.3.4 Training Datasets for Object Detection

The ImageNet challenge task ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition mentioned above formally established the standard for training and evaluating the target detection task using a unified public dataset. The work of this thesis is based on a single image and multi-frame video, and some mainstream datasets in the detection field are selected for testing. The effectiveness of the work can be shown by comparing it with algorithms that have achieved excellent results. However, general datasets can only represent the ideal situation. How to carry out the actual detection task for the security monitoring scene is also one of the key issues of this thesis.

PASCAL VOC Dataset

The PASCAL Visual Object Classes (PASCAL VOC) is a large-scale image processing task challenge that provides standard graphics datasets for testing computing and learning performance, as well as an international standard scoring system. A large amount of good computer vision models are based on this standard and metric, such as RCNN, YOLO series and SSD object detection models.

At first, PASCAL VOC was mainly designed for classification tasks, and in subsequent years, tasks such as instance segmentation and object detection were introduced. The standard and public datasets used in the challenge are mainstream in the industry. The PASCAL VOC datasets provide 20 categories for training and testing, covering the most common objects in daily life. All data is divided into three parts with training, validation and test subgroups, respectively.

The advantage of PASCAL VOC dataset could be attributed as the following points:

1. Firstly, it contains the sufficient quantity and categories of objects with great labelling accuracy.

2. The efficiency of this dataset has been proved by most outstanding object detection model so the performance of proposed models obtained from PASCAL VOC dataset could also be accepted.

3. There is a professional group to modify and update the content of dataset.

However, it also has two main limitations. The first is that it has not produced new version of dataset since 2012. Therefore, it is hard to meet the training demands of some new proposed object detection models. The second one is that the test set of PASCAL VOC 2012 is not public so anyone who wants to test the model performance have to submit their code to PASCAL VOC Evaluation Server, which is not very convenient.

ImageNet Dataset

The ILSVRC is used to evaluate common and massive mainstream datasets for object detection and image classification. The ImageNet dataset is a large-scale labeled image dataset organized according to the WordNet architecture. ILSVRC is built on part of the ImageNet dataset, which is divided into challenging tasks. The dataset can be used for a variety of detection tasks in the field of object detection, corresponding to object detection of static images and video images. There are around 20 thousand categories of objects with 14 million of pictures in total, which is one of the largest datasets for object detection task.

As a lot of advanced object detection models are trained based on the ImageNet dataset, the efficiency and accuracy of this dataset has been already proved. Besides, the classification of objects is very clear and detailed, which is very convenient for using.

ImageNet dataset is a robust dataset, however, there tend to be mistakes in the data labelling classification due to its enormous data variety and quantity. So there would be revisions each year to correct these mistakes, it is therefore advised all users to down-loaded the latest version and keep an eye on any revision it might update.

1.4 Objectives

The objectives of this thesis are as follows:

- To evaluate the advantage and disadvantages of existing image-based object detection methods and find out potential improvements.
- To improve the efficiency of feature extraction of image-based object detection method by optimizing the feature pyramid architecture with multi kinds of feature map connections.

- To balance the distribution of training samples by improving the focus on positive hard examples.

- To evaluate the advantage and disadvantages of existing video-based object detection methods and find out potential improvements.

- To enhance the efficiency and accuracy of local information aggregation by calculating feature similarity between the target frame and its adjacent frames.

- To enhance the efficiency and accuracy of global information aggregation by introducing the correlation coefficient between the target frame and global frames calculated by the attention mechanism.

1.5 Thesis Overview

The structure of the thesis is shown as follows:

Chapter 1 introduces the research background and the significance of the selected topic, demonstrating the practical application value. Besides, it discusses the existing problem of target detection algorithms when they are being applied to smart security and the potential improvements. Finally, it introduces the related academic definitions and theories involved in this thesis, which could elicit the value and significance of the subsequent innovative improvement work. It also reviews the relevant references and organizes them as part of the literature review.

Chapter 2 reviews the relevant references from three different aspects and organizes them as part of the literature review. Moreover, it discusses the existing problem of target detection algorithms when they are being applied to smart security and the potential improvements.

Chapter3 introduces the target detection algorithm for still images. Based on the foundation of one mainstream detection algorithm Faster RCN, it proposes two strategies with multilayer feature cascaded aggregation pyramid network and sample asymmetric weighted loss function. By testing on the mainstream dataset, the result shows that our proposed methods are efficient. Meanwhile, it also demonstrates the performance of an improved detection algorithm by a real detection image example.

Chapter 4 introduces the video-based target detection algorithm. By processing the content of video frames, relevance attention learning and feature fusion algorithms, the image-based detection algorithm is combined with continuous video frame features for feature enhancement, localization and fusion, which could obtain more expressive current frame features. The effectiveness of the proposed work is proved by testing on a public dataset.

Chapter 5 summarizes the overall work and provides an overview of future direction based on existing progress.

Chapter 2

Literature Review

2.1 Traditional Target Detection Algorithm

According to different feature extraction methods, target detection algorithms can be divided into traditional and deep learning-based methods.

Unlike deep learning-based algorithms, the traditional target detection algorithm uses manually designed feature vectors to represent the target features, and the whole process can be divided into three steps: generation of the regional proposal, extraction of feature information and category classification.

The sliding window is one of the most representative methods of the regional proposal [63]. It uses the sliding window with a predefine size and aspect ratio to traverse all potential locations that the target may appear and define all regions of interest as regional proposal containing the target. In addition, there are some developed methods in this field, such as graph segmentation [62], the most OTSU [47], and Selective Search [60]. The extraction of feature information is the most critical step of traditional target detection algorithms, which requires the manual design of suitable feature vectors to describe their semantic information in the regional proposal generated in the previous step. These kinds of designs require the designers to possess a great level of experience, especially when they aim to solve different problems under different application scenarios. The Classical approaches include Haar features [37] used in VJ [64] face monitor, HOG features [11] used in pedestrian detection, and the improved DPM [18] algorithm based on the ”divide and conquer” idea which divides the overall prediction task into a collection of local predic-tion results.

Regarding category classification, traditional target detection algorithms often use statistics classifiers to determine the category to which the regional proposal belongs. The common classification methods include Support Vector Machine [8], Bagging [4], Adaboost [24], and Cascade Learning [17].

Although the traditional target detection algorithm has slowly disappeared in recent research, it still has great historical significance. The methods and ideas proposed by it have provided important directions and references for the development of deep learning algorithms, such as the idea of the regional proposal and the NMS algorithm [45], which is used to remove redundant prediction anchor boxes. However, the traditional target detection method also generates a large number of redundant and invalid proposal anchor boxes, which increases time, complexity and computational cost to some extent. Such failed anchor boxes will have a further negative effect on the accuracy of the final classification. What is more, the manually designed features are easily affected by the surrounding environment, lightness, and scales. It is also over-depended on designers' personal experience, which has certain limitations and cannot be well-applicated in different scenarios. Therefore, the accuracy improvement of the traditional target detection methods is very limited.

2.2 Target Detection Based on Deep-Learning

The Deep learning method is a mathematical modelling approach to simulate the neural structure of the human brain and build a deep convolutional neural network model. It learns the feature representation vector of an image by sending a large amount of training data with the ability to automatically adjust parameters. With the development of computation power, convolution neural network shows their great performance on feature extraction, which attracts many more researchers into the field. In 2014, the Regions with CNN (R-CNN) algorithm [22] proposed by Ross Girshich officially opened the door to the application of deep learning methods in object detection algorithms. R-CNN is the first two-stage target detection method that follows the traditional idea of the regression-prediction frame on target, which divides the detection task into three steps: generating regional proposals, e-tracting image features and classification, as shown in Figure 2.1.

Firstly, thousands of regional target proposals Region of Interest (ROI) are generated on the input image by using the Selective Search method, then every feature vector from each candidate region is extracted by a deep convolution neural network and used by the Support Vector Machine (SVM) classifier to make a prediction on the location and category of the target. For the R-CNN network, it uses a fully-connected neural network, which requires a fixed size of the input image and a large number of feature vectors of the candidate regions. Such a structure will pose a negative effect on detection accuracy and efficiency.

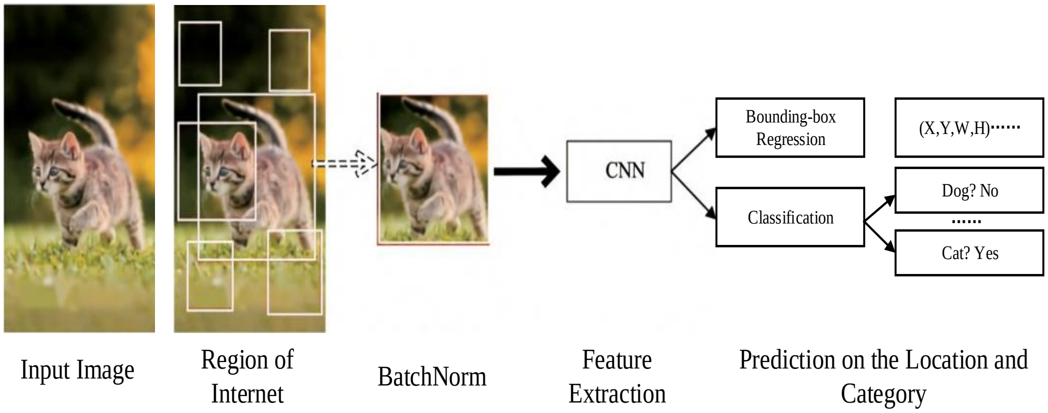


Figure 2.1: The basic structure of two-stage target detection algorithm

To improve the detection speed, in the same year, He et al. proposed a Spatial Pyramid Pooling (SPP) method [25], which introduced a SPP layer. The pyramid pooling layer accepts feature map inputs with different sizes, normalizes the feature map to a feature vector with length, then inputs them into the fully connected layer, which could improve the compatibility of the model.

The structure of the SPPNet is shown in Figure 2.2. SPPNet firstly makes a convolution operation on the whole image to extract the feature map and generates the target candidate regions on the feature map, it then performs subsequent computation operations on the region of interest. This process avoids repeated convolutional feature extraction for each candidate region and improves the operational efficiency of the network.

Inspired by the idea of the SPPNet model, Ross Girshich furtherly proposed the Fast RCNN [21] based on R-CNN in 2015, which adopts a shared feature map approach that only convolves the input image as a whole for just once to extract the feature map, then selects regions of interest on the feature map. The structure is shown in Figure 2.3(a).

The Fast RCNN proposes a new ROI pooling layer to unify different sizes of inputs into fixed-length feature vectors. This reduces the need for cropping and scaling inputs with different sizes to the same size. The fully connected layer is used for detection. The detector is divided into two branches, one branch uses a softmax classifier instead of SVM for classification, and the other branch uses the SmoothL1 Loss function for regression of the bounding box, which improves the detection speed and accuracy.

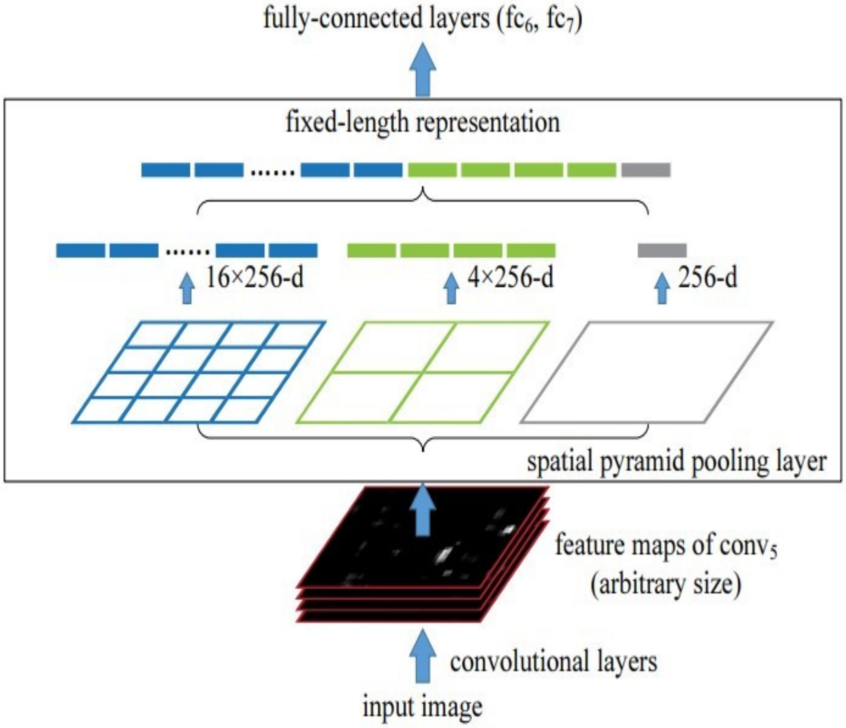

fully-connected layers (fc6, fc7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of conv5
(arbitrary size)
convolutional layers
input image

Figure 2.2: The structure of SPPNet

Shortly after fast RCNN was proposed, He et al. proposed Faster RCNN [53] in the same year, which replaced the method of Selective Search in fast RCNN with Region Proposal Network (RPN) to generate candidate regions. The structure is shown in Figure 2.3(b). The RPN module has nine anchors, and the sizes of them are different. Such anchors could produce a set of candidate regions on any size of input images and also share the convolution layer with the FasterRCNN backbone network, which could simultaneously perform extraction from the both region of interest and feature map. Designs like this can increase speed and efficiency.

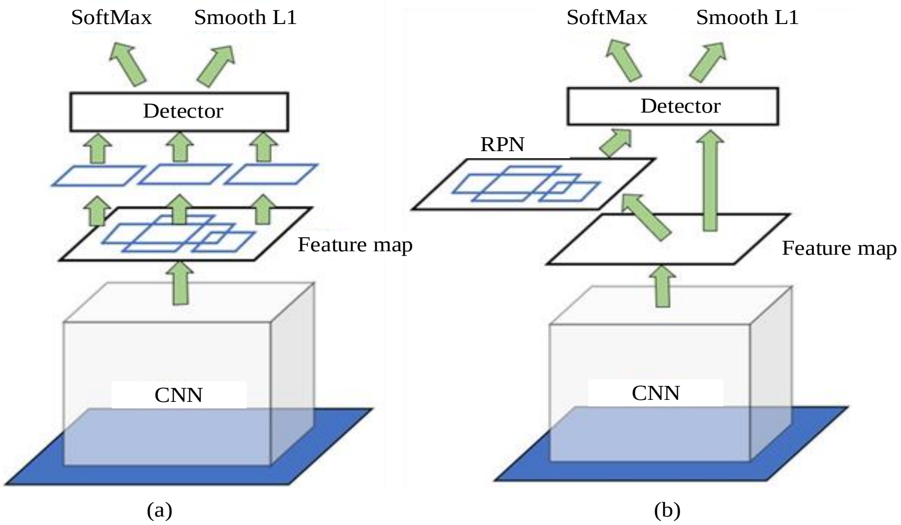


Figure 2.3: The structure of Fast RCNN and Faster RCNN (a) Fast RCNN (b) Faster RCNN

A key drawback of the deep learning-based detection algorithm is that only the last extracted layer of features could be used for detection, which reduces accuracy. In 2017, to resolve this drawback, Lin et al. proposed to replace the RPN network with the FPN [38]. FPN structure organizes the feature maps from different layers to produce feature maps with different scales and performs independent detection on different feature maps. One of the most significant advantages is that FPN could detect the different sizes of the target based on different feature maps, which could match the target and its corresponding feature map more accurately and, hence, increasing the detection performance of the network.

All of the networks described above divide the regional proposal extraction and classification recognition into separate stages. Therefore, they are defined as two-stage target detection algorithms. Although the two-stage algorithm has good performance on detection accuracy, the computation takes a long time to process, which is insufficient to meet the requirement of real-time detection in the application. In contrast, if a network could organize the location prediction and classification prediction into the same task, the target detection task could be considered as one regression task, which allows two regression tasks to be carried out simultaneously. As a result, this saves half of the time of the regression task and increases the speed of the detection algorithm, which makes it more suitable for real-time detection. This detection method is known as a one-stage target detection algorithm, and its structure is shown in Figure 2.4.

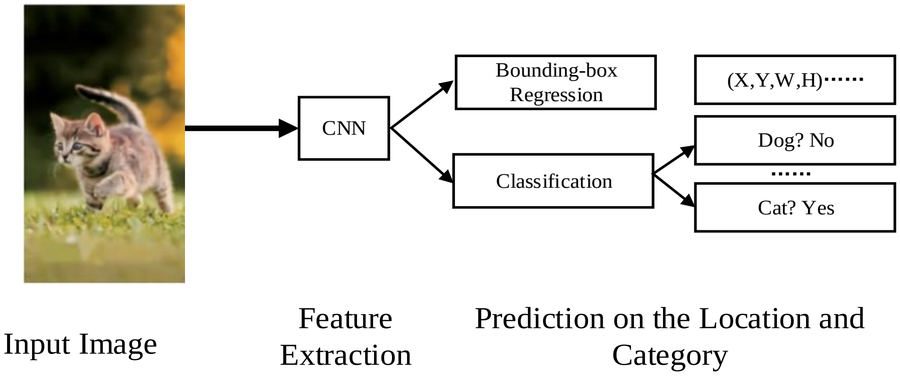


Figure 2.4: The basic structure of the one-stage target detection algorithm

In 2015, Redmon et al. proposed the Yolo framework [50], which removes the step of pregenerating candidate regions of the two-stage algorithm and treats detection as a regression problem. In this way, the algorithm only needs to operate the regression computation once for target detection. The Yolo algorithm divides the original image into N × N grid cells and predicts whether the grid contains a target. It then estimates the probability of target occurrence in this grid. By removing the preregion proposal step, the detection speed of the Yolo algorithm is significantly improved to meet the requirement of practical applications. Subsequently, in 2017, Redmon et al. proposed an improved YOLOv2 [51] which was inspired by the idea of an anchor mechanism from Faster RCNN. The improved Yolov2 network uses kmeans clustering to filter redundant and unqualified anchors, which improves efficiency. Besides, it also introduces DarkNet as a backbone network, which increases the average recall rate by 7% that further increases the accuracy of localization and detection (top 10). Moreover, in 2018, Redmon et al. proposed the improved Yolov3 algorithm [52], which combines the idea of multiscale feature fusion from FPN with the residual structure to improve the recall rate of detection on a small target.

In 2020, Bochkovskiy continued the research on the Yolo detection network and proposed the Yolov4 algorithm [3]. The version of Yolov4 introduces several advanced deep learning techniques, such as Mosaic (Bochkovskiy et al., 2020) data augmentation methods, CSP module[65], Mish activation function [43], PAN structure [74] and DIoU loss function [80]. Fortunately, these modules collaborate with each other very well and improve the performance of target detection.

However, while the appearance of the Yolo series detection network made some breakthroughs in the accuracy and speed of the target detection algorithm, it still has some problems that need to be solved, and small target detection is one of the most serious problems among them. To address this problem, Liu et al. proposed the SSD algorithm [41], which uses convolutional layers instead of fully connected layers for the final target detection operation. The structure of the SSDSingle Shot MultiBox Detector (SSD) network is shown in Figure 2.5, which uses feature maps with different scales for prediction. It also introduces an anchor mechanism to reduce the complexity of training. However, the SSD network has repeated computational problems when it detects the feature maps with different scales, which increases the computation cost of the network. To deal with this issue, Jeong et al. proposed the ESSD algorithm [27] that uses feature fusion for detection. Furthermore, Fu et al. proposed the DSSD algorithm [19], which replaced the original backbone network VGG16 with ResNet101 to extract deeper features and fuse deep features with shallow features by using deconvolution operation. Li et al. proposed the FSSD algorithm [36], which learns the idea of FPN to fuse feature maps with different scales. All of these proposed algorithms help to increase the detection accuracy of the model on a small target.

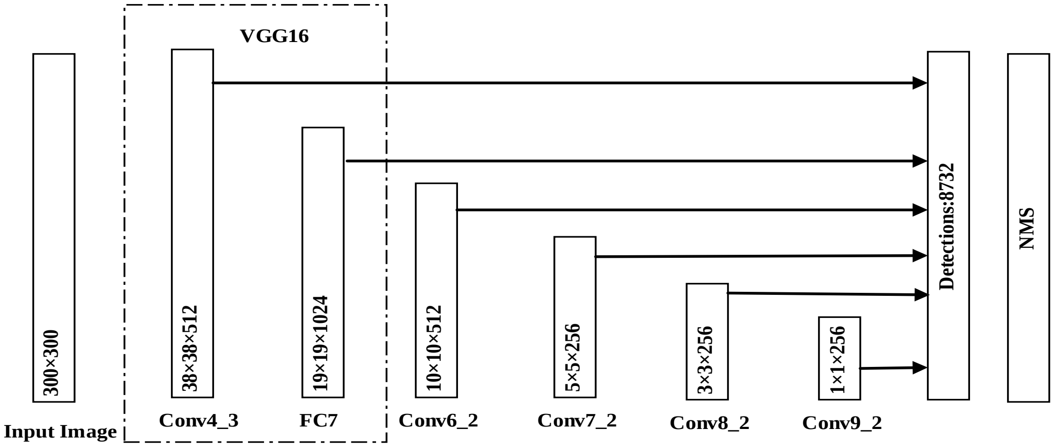


Figure 2.5: The structure of SSD network

During the research, Li et al. found that, because of the unbalanced number of positive and negative candidate anchor boxes from all samples, the detection accuracy of the singlestage target detection algorithm was lower than that of the two-stage algorithm. Due to this issue, they proposed the Retina algorithm [39], which uses ResNet as the backbone network for feature extraction and FPN for feature fusion.

Meanwhile, they design a Focal Loss loss function which suppresses the training gradient of the easy negative samples. More specifically, they dynamically adjust the distribution of positive and negative samples from all samples and balance the weight during the training process. In this way, it removes the negative effect of the unbalanced distribution of the samples. In summary, compared to another one-stage detection algorithm, RetinaNet improves the detection accuracy, but at the expense of detection speed.

Most of the target detection algorithms mentioned above are based on the anchor frame mechanism except for the YOLOv1 algorithm. In recent years, researchers have started to consider the drawbacks of the anchor frame mechanism. Firstly, the predefin anchor frame mechanism mainly depends on the prior knowledge of the researcher, and the hyperparameters (e.g. numbers and aspect ratios) could also easily impact the model performance. This means that such a mechanism lacks generalizability. Secondly, if we want to make the predefine anchor boxes more accurate than real bounding boxes, we need to set a large number of potential anchors, which will produce a lot of parameters. Such anchors and parameters will pose a great burden on computation. Besides, the redundant anchors that only contain the background information will be classified as negative samples, which will influence the convergence speed and efficiency of the network negatively. Lastly, due to the different shapes, postures and sizes of the target, the anchor-based detection algorithm will find it difficult to detect these hard examples.

To solve above problems, anchor-free target detection algorithms are proposed. CornerNet [34], proposed by Law et al. in 2018, uses the idea of heatmap and changes the way of prediction by predicting the coordinates of the upperleft and lower-right corner points of the bounding box instead of predicting the whole bounding box. Meanwhile, they propose the corner pooling method, which transfers the focus of attention from the center of the feature map to the border of the feature map, which makes the calculation of corner points more accurate. However, the possibility of error detection will also increase since the corner point matching just consider two points of the bounding box while other methods consider the entire box. Besides, since such a method focuses on the border area, it could also ignore the important information at the center of the bounding box. This limits the improvement of detection accuracy and speed.

Based on the similar idea of the heatmap, Zhou et al. proposed the CenterNet algorithm

[82], which no longer predicts based on corner points, but predicts the localization of the target based on the coordinates of the center point, length and width of the bounding box. It no longer uses non-maximum suppression to remove redundant bounding boxes, which leads to a faster network. However, there is a possibility that the center point of different targets locates on the same point, and, in this case, the CenterNet could only detect a single target while missing others. A similar centroid-based target detection algorithm was also proposed by Tian et al. in 2019, known as the FCOS algorithm [58] [57]. It predicts the different sizes of bounding based on a different level of feature information to avoid missing targets from the detection of overlapping targets. In summary, although the target detection algorithm without the anchor frame has been studied for a short period of time, it is still one of the potential future directions for target detection.

2.3 Video-based Target Detection

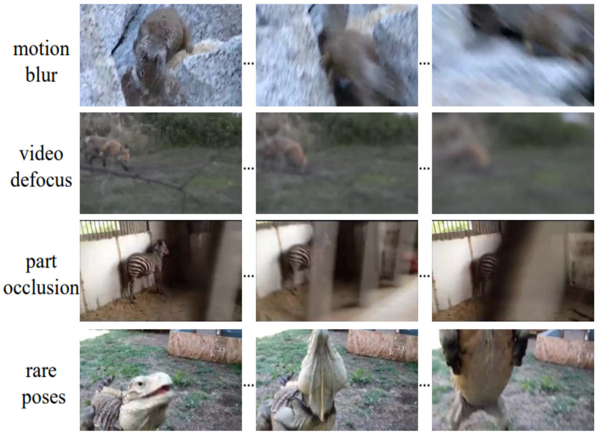


Figure 2.6: Some exsiting problems in video-based target detection

Traditional image-based target detection algorithms could maximize the use of feature information from a single picture. However, the extracted information from the still image is limited and is still far from solving some real problems. For example, the detection algorithm cannot confirm the target, which is completely obscured. One way to solve this problem is to transfer the dataset from still images to video since the videobased data contains more spatiotemporal information and semantic relationship. However, the performance of the image-based target detection algorithm is likely to drop when it processes video data (as shown in Figure 2.6), and it becomes harder to obtain good results [44].

The common video-based target detection algorithms can be classified into bounding-box-based algorithms and pixelbased algorithms using different levels of feature processing. TCNN is a type of bounding-box-based detection algorithm which could complete the object detection task on a video dataset by fusing contextual information of bounding box from a series of continuous images. TCNN proposes a feature tracking algorithm which uses a tree structure to manage and update the feature of the target appearance. Such design could lead to a smooth learning process for the model and, thus, has relatively high stability. Meanwhile, because of the idea of parameter sharing, the network could save storage space and increase the efficiency of the training process.

Regarding the pixel-based algorithms, they mainly use optical flow networks to calculate the differences between the motions of image features, such as depth feature-flow-based algorithms and optical flow feature-fusion-based algorithms. Both methods combine optical flow networks with deep learning methods to solve video target detection challenges from different perspectives.

The Depth Feature Flow (DFF) algorithm [85] firstly divides the current frame into different categories based on the different attributes. It then extracts the feature of the target based on the deep learning method and calculates the corresponding optical flow of the same target generated by movement changes from other frames. Finally, it combines both two types of features and sends them to the next module of recognition.

The Flow-Guided Feature Aggregation (FGFA) algorithm is a type of optical-flow-based algorithm [15], which aims to get a better recognition accuracy by enhancing the feature of the current frame. It combines the feature information of the same target from the front and back frames and sends them to the current frame feature map. Besides, by introducing an attention mechanism, the FGFA algorithm proposes a dynamic weight update mechanism which gives more weights to those frames that are highly related to the current frame. According to the experiment result, such a method could increase the quality of the target feature and improve the ability of the network to deal with occlusion and obscure cases.

Chapter 3

Target Detection Based on Feature Optimization and Training Sample Equalization

Both one-stage and two-stage target detection algorithms could be used on target detection based on a still image dataset. Typically, a onestage target detection algorithm has a faster detection speed, which directly passes through the feature map, and a two-stage target detection algorithm has better accuracy performance, which could be represented by the classical detection algorithm Faster Regions with CNN (Faster-RCNN). However, although both one-stage and two-stage detection algorithm have their own advantages and have improved substantially over time, there are still some problems needed to be addressed. First of all, recent research shows that the existing detection network performs badly when they have an imbalanced division of positive and negative samples [39]. Besides, by balancing the distribution of positive and negative examples, the performance of the detection network could be improved to some extent but still has areas that have the potential to be improved further. Secondly, the FPN network [38] could definitely increase the ability of feature extraction for the network, which combines the deep and shallow features greatly. But the information will still be lost during the process of every feature extraction operation, which is the main hurdle for the improvement of the FPN network. In this chapter, in order to address the above problems, this thesis proposed a target detection algorithm based on feature optimization and sample equalization. Details are as follows:

- A new Multi-layer Feature Cascade Aggregation Feature Pyramid Network (MCA-FPN) is proposed based on the structure of the feature pyramid. This module can fully combine the deep and shallow feature semantic information and improve the representation ability of each feature.

- Optimize the positive and negative sample matching mechanism, and design the sample equilibrium loss function by asymmetrically weighing the positive and negative samples to solve the imbalance that existed in the training process. The experiment result shows that the optimized network model can converge quickly.
- The designed MCA-FPN can be easily applied to other detection algorithms and help to achieve the algorithm migration on the basis of better performance.

3.1 Backbone Network Architecture

The backbone network selected in this chapter is traditional Faster-RCNN which has been introduced in previous chapters. It firstly uses RPN to generate different anchor boxes with different sizes and ratios based on different input still images. Then it uses the ROI pooling layer to do a down-sampling operation and get the necessary features. Finally, it uses two separated fully-connection layers for classification and localization. The whole structure of Faster-RCNN is shown in Figure 3.1.

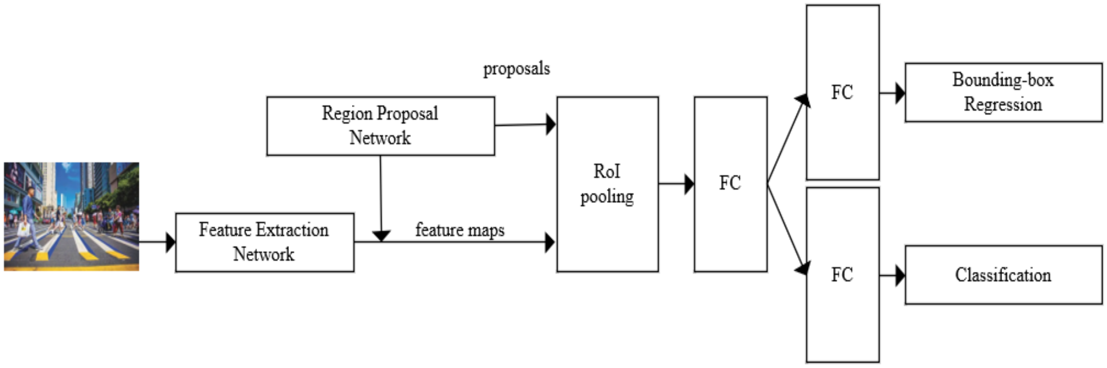


Figure 3.1: The structure of Faster-RCNN

For traditional detection algorithms which are based on bounding boxes, such as window sliding and selective search methods, one of the problems is that such methods would easily produce a large number of useless bounding boxes with noting being included. Besides, since such methods are doing supervised learning based on the coordinate and classification of ground-truth bounding boxes, when the number of target categories increases, the label of ground-truth with the information of the target will randomly and excessively be located on the input image, this makes it harder for the network to select the most suitable information for training. Based on the above problem, the RPN network was proposed.

As shown in Figure 3.2, for the feature map produced by the backbone network, the RPN will generate nine different anchors with three different sizes and three different shapes.

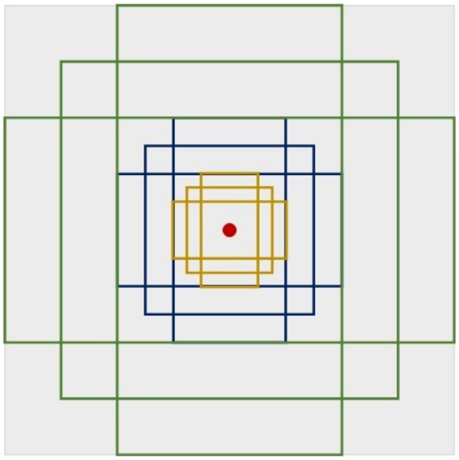


Figure 3.2: Anchor boxes produced by RPN

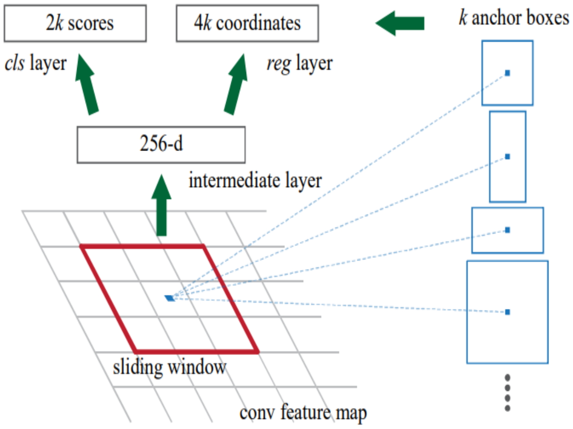


Figure 3.3: The structure of RPN

The whole process of extraction operation by the RPN network is shown in Figure

3.3. Firstly, RPN slides on the feature map by using predefined convolutional layers to extract corresponding features and aggregates features by a convolutional layer with 512-dimensional. Then, the result produced by RPN will be sent to different branches, which are classified by the different functions of classification and regression. The classification branch is used to produce the result of the possibility of foreground or background captured by the anchor box, and the regression branch is used to produce the coordinate of the selective bounding box. Finally, the output of two branches from RPN is sent to the fully-connected layer, which combines the information of two branches and produces the organized result. After the above sub-processes, RPN will have a series of highly-correlated anchor boxes for each target, and these anchor boxes could be used for the prediction of the location of ground truth.

Besides the way the RPN extracts the feature from the feature map, the design of the loss function will also be another important factor for network performance, which is the feedback signal for training. Typically, the value of the loss function could be produced by the difference between the true value and prediction value of the target. Therefore, the higher value of the loss function means a larger amount difference between the true value and prediction value and, as a result, the network will focus on those examples which have a large loss function value until their value decreases to a minimum level. Specifically, there are two types of value for prediction value, one is classification result, and the other is regression result. The value of the classification result is a kind of probability, and its range is from [0,1]. The value of the regression result is a series combination of target coordinates which shows the top left and bottom right coordinates as (x_1, y_1, x_2, y_2). As we have the prediction value of the target, the next step for the network is to confirm what types of examples the target could be attributed as a positive example or negative example, which is based on the value of IoU, and the detail of judgement is shown in Table 3.1.

Based on a suitable set of IoU standards, the network could find and define categories and location coordinates of all anchors. However, when the targets need to be detected become dense, the number of potential anchors generated around the target will be increased a lot, which could decrease the efficiency of the network dramatically. Worse yet, when the detected image contains a lot of background information, those anchors which also have background information will be easily defined as negative examples and easily lead an imbalanced distribution between positive and negative examples, such phenomenon could decrease the convergence speed and sometimes even lead no convergence. Even if the distribution of positive and negative examples could be balanced by the set of hyperparameters to some extent, such a pre-define setting can not be appropriate for all scenarios.

For loss function in RPN structure, it is composed of two parts with the loss function

Table 3.1: The performance of mainstream object detection algorithm on multi-scaled object

Label	Description
Label of Negative Example	The largest IoU valueof all anchors from this target is less than threshold of IoUmin
Label of Posotive Example	The anyone IoU value of anchors from this target is more than threshold of IoUmax
Incalid Label	The IoU value of target ranges from IoUmin to IoUmax, all these targets would be defined as invalid label

for classification and regression, respectively. The overall loss function used is proposed by Lin et al. in 2017 [39] and is shown as equation 3.1:

$$L(p_i, t_i) = \frac{1}{L_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{L_{reg}} \sum_i L_{reg}(p_i, p_i^*) \tag{3.1}$$

Where L_{cls} represents the loss of classification based on the cross-entropy loss function [39], p_i represents the true value and p_i^* represents the prediction value. The detail of the classification loss function is shown as equation 3.2:

$$L_{cls}(p_i, p_i^*) = -log[p_i p_i^* + (1 - p_i)(1 - p_i^*)] \tag{3.2}$$

Where L_{reg} represents the loss of regression based on the $smooth_{L1}$ function [39] , t_i represents the true value and t_i^* represents the prediction value. The detail of the regression loss function is shown as equation 3.3:

$$\sum_i L_{reg}(t_i, t_i^*) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, \text{otherwise} \end{cases} \tag{3.3}$$

Here, the gradient of $smooth_{L1}$ function is shown as equation 3.4, the absolute value of x exceeds 1, which could avoid the phenomenon of gradient vanishment or catastrophe to some extent during the training process. In this way, the loss of function of RPN could be produced and combined with other loss values for future backward feedback.

$$\frac{dSmooth_{L1}(x)}{dx} = \begin{cases} x, \ |x| < 1 \\ \pm 1, \text{otherwise} \end{cases} \tag{3.4}$$

3.2 Multi-layer Feature Cascade Aggregation

An important feature of convolutional networks, represented by ResNet, is that they produce feature maps hierarchically based on the different layers of the network. In fact, the information included in different feature maps is different. Those shallow feature maps are more likely to contain information about original image features, such as texture and color. Besides, as the network becomes deeper, the feature map produced by them will be operated by serval pooling layers and only leave the most representative pixels (or information), which leads the feature more abstract and shows more semantic information.

The core point of the target detection algorithm is to do different kinds of extraction and analysis operations on feature maps. However, as the size of the feature map decreases with the increasing network operations, some texture information is lost during this process and only left more semantic information, which poses a greatly negative effect on the detection performance of the network for blur or small targets. Therefore, for better detection performance, the network needs to find a method which could combine the texture (shallow) information and semantic (deep) information greatly. Based on the above technical demand, the feature pyramid was proposed.

3.2.1 Feature Pyramid Analysis

Based on the property that the backbone network produces different feature maps hierarchically, the feature pyramid model is proposed. The first research on the pyramid structure was conducted by SPPNet, which is the predecessor of the FPN and the purpose of which is to make detection network well-adapted to different sizes of targets. However, the traditional SPPNet has the problem of low utilization efficiency of samples and feature maps. To solve this problem, the FPN is designed, and the structure is shown in Figure 3.4, which uses two pyramid structures with bottom-up and top-down connections, respectively. Such structure could organize different dimensions of feature maps efficiently and increases the applicability of the network on different sizes of targets.

But there are still some limitations for FPN, one is horizontal single direction information transmission, and the other is longitudinal single direction information transmission, both of them could limit the information sharing efficiency and amount.

For longitudinal single-direction information transmission, it happens on the right pyramid, which only has a single-directed enhancement operation from the top layer to the bottom layer. Even if this structure could enhance part of feature maps and provide more information for prediction, it still lacks the reverse connection from low dimensional feature

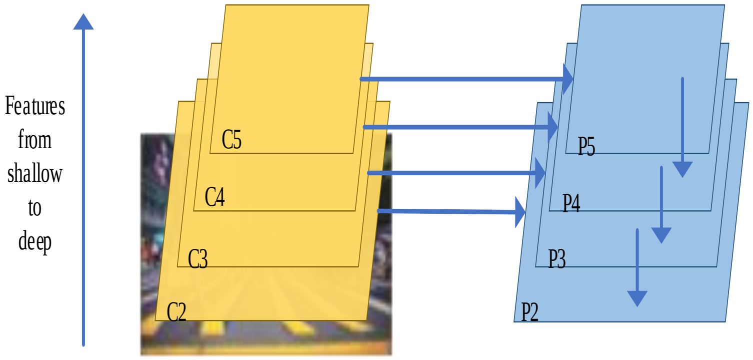


Figure 3.4: The structure of feature pyramid

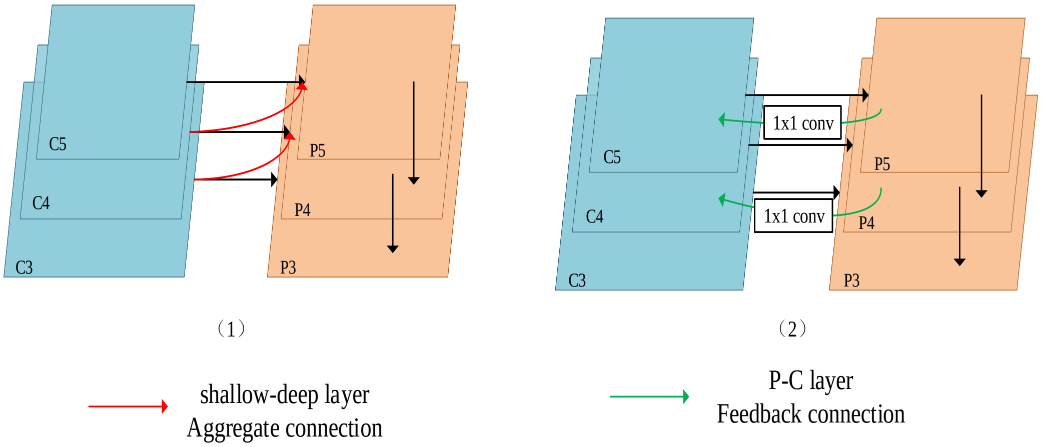
map to high dimensional feature map. Take the P5 layer as an example: this deep feature map will be used to enhance other feature maps but not be enhanced by other shallow feature information, which will lead to the loss of low dimensional features and, thus, decrease the accuracy of final regression and prediction. For horizontal single-direction information transmission, it has a similar kind of problem, which lacks further feedback enhancement from the right to the left side.

To solve the above problems, some researchers proposed some improvements based on the original FPN structure, PANet [40] added a kind of crosslayer connection to improve the utilization of bottom feature information. STDN [81] proposed a scale transfer module which higher information transmission efficiency. G-FRNet [26] designed new receptive fields with different scales to combine different scales of information successfully. Both NAS-FPN [20] and Auto FPN [73] use the finetuning network method to improve the accuracy of FPN. EfficientDet [16] proposed a BiFPN layer which repeats the process of feature extraction and combination. However, all of these methods did not solve all mentioned limitations perfectly, some of them just solved one of the limitations, and some of them improved the performance but with too much extra computation.

3.2.2 Feature Pyramid Network

Based on the analysis result, this thesis proposes a type of feature pyramid structure increase the information transmission efficiency and network performance without extra computation consumption, the structure of the new network is shown in Figure 3.5.

Take the P4 layer as an example: the original P4 layer is produced by the feature



（1）

（2）

Figure 3.5: Proposed improved Feature pyramid network with two more feature connections

aggregation from the P5 layer and C4 layer, which only contains the information from the current layer (C4) and deep layer (P5).

Our improvement is to add an extra connection from the C3 layer to the P4 layer, which helps the P4 layer to learn the information from the shallow layer(C3). Such design will not change the scale of the original feature map but will make the network more sensitive to the location and category information of the target.

Meanwhile, inspired by the improvement of ResNet backbone from Wide Residual Networks [75], another feedback connection from feature pyramid P to C is added (as shown in Figure 3.5(2). The ResNet network will allow original input x and feedback input R(f) to be calculated in the regional proposed network. Significantly, the feedback connection here is only operated on the first residual blocks, which has a limited negative effect on the whole network structure.

The input of the traditional FPN module is shown below:

$$F^{i}_{BB} = C_i(x_{i-1}) \tag{3.5}$$

$$F^{i}_{FPN} = P_i(f_{(i+1)},x_i) \tag{3.6}$$

Where C_i represents the i^{th} operation from the bottom-up pyramid (left side), P_i represents the i^{th} operation from the up-bottom pyramid (right side), $\{F^{i}_{BB}|i = 1, ..., S\}$

represents the input feature from backbone network to FPN module $\{F_{BB}^i | i = 1, \ldots, S\}$, F_{FPN}^i represents the output feature from FPN module $\{F_{FPN}^i | i = 1, \ldots, S\}$ and the S represents the number of stages.

After improving the shallow to deep aggregation link and P-C feedback link, the input and output of the model are expressed as follows:

$$F_{FPN}^i = P_i(F_{FPN}^{i+1}, F_{BB}^i, F_{BB}^{i-1}) \tag{3.7}$$

$$F_{BB}^i = C_i(F_{BB}^i, R_i(F_{FPNi}^i)) \tag{3.8}$$

$$R_i(F_{FPNi}^i) = Conv(F_{FPNi}^i) \tag{3.9}$$

Where equation 3.7 3.8 3.9 represents the connection from the shallow feature map to the deep feature map and is shown as the red connection line in Figure 3.5, the output of the feature map F_{FPN}^i is obtained from three different kinds of input with the deeper feature map from P pyramid F_{FPN}^{i+1}, the same level of feature map from C pyramid $F_B B^i$ and the shallower feature map from C pyramid F_{BB}^{i-1}. Equation 3.8 represents the feedback connection from P pyramid to C pyramid and is shown as the green connection line in Figure 3.5. More details about this structure are illustrated in equation 3.9, which adds an extra convolution layer with the kernel size of 1x1 to add feedback information from the feature map in the P pyramid to F_{FPNi}^i to the feature map in the C pyramid F_{BB}^i.

3.3 Sample Asymmetric Weighted Loss Function (SAW-Loss)

Although the development of the target detection algorithm has made great progress with many innovations in structure, parameter settings and modules, basically, the target detection algorithm is still in the data-driven stage, which needs a high-quality dataset to ensure the performance of algorithms. However, the quality and distribution of the dataset obtained from our real-life are not as good as most public training datasets, which causes low network training efficiency and detection accuracy. This is because the dataset collected from our real life has a large proportion of hard examples instead of easy examples. For such distributed dataset, the network should transfer its attention from easy examples to hard examples to ensure performance. Based on this phenomenon, some research has been proposed to focus on solving the sample-related problem.

3.3.1 Sample Imbalance Analysis

The example used for network training could be divided into several categories, which are shown in Table 3.2. Besides, there is also an example shown in Figure 3.6. The red box is defined as ground truth which is handle-labelled correctly. The green box represents the easy positive sample since it includes most facial information of the target and would be easy to identify as a positive sample by the network, the pink box represents the easy negative sample since it only contains background information but not any target information and the network would also be easy to identify it as negative samples. The black box is defined as the hard negative sample, as it contains some insignificant feature information of target and background information. The network may recognize it as a positive sample, but, in fact, it is far from being a successful detection. The blue box is defined as a hard positive sample since the feature information of the target it contains seems to be relatively sufficient for a successful detection, but the background information it contains will pose a negative effect on the final prediction result.

The basic explanation of sample imbalance is the extremely uneven distribution among the above categories with too many easy examples of positive examples. As shown from the research [46], the network is hard to learn useful feature information from too many easy examples, and hard examples could also have a positive effect on network training. For example, the feature from easy examples could be easily learned by the detection network; when the dataset has too many easy examples, the network will learn them continuously and have great performance on them. But such continuously learning on easy examples can not help network to learn the feature of hard examples, so the accuracy of the network on hard examples can not be improved. As a result, the loss value of the training process will not decrease as expected, and the network will be hard to be converged.

The basic idea to solve the problem of sample imbalance could be summarized as data preprocessing method and model finetuning method [56] [49] [48] [55]. Take the Faster RCNN as an example; the network will ensure the balance between positive and negative examples by setting appropriate hyperparameter during the RPN and ROI stages.

However, such a predefined hyperparameter setting mainly depends on the researcher's experience, which is too subjective for academic research. Besides, this type of setting just balances the distribution of positive and negative samples but can not solve the imbalance between hard examples and easy examples.

Table 3.2: Categories of example

Category	Definition	Note
Positive example	The area which is located in correct-labeled anchors, usually are targets	Usually are targets
Negative example	The area which is not located in correct-labeled anchors, usually are backgrounds	Usually are backgrounds
Easy positive example	Those positive examples which are easy to be classified. They usually occupy a large proportion of all samples	The value of loss function for each example would be small but the cumulative value would be large which leads the trend of total loss function
Easy negative example	Those negative examples which are easy to be classified	The value of loss furcation for each example would be small but the cumulative value would be large which leads the trend of total loss function
Hard positive example	Those positive examples which are classified as negative examples	The value of loss function for each example would be larger but the cumulative value would be small
Hard negative example	Those negative examples which are classified as positive examples	The value of loss function for each example would be larger but the cumulative value would be small

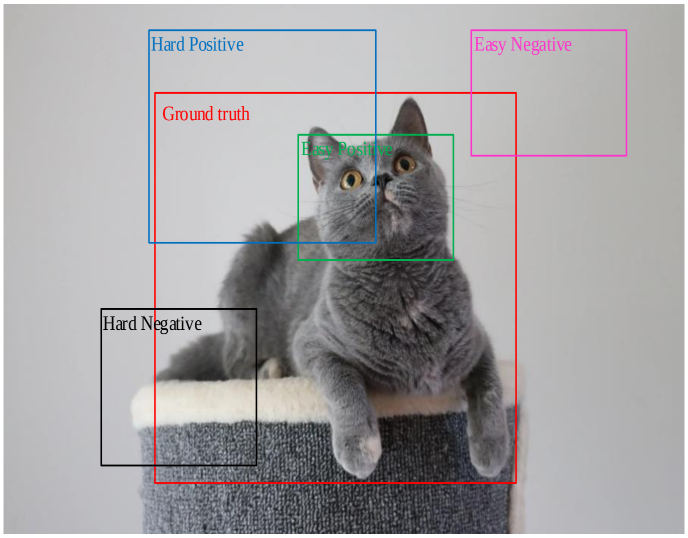


Figure 3.6: An example of different sample labels

3.3.2 SAW-Loss

The positive and negative samples that participated in the target detection model should be firstly ensured that their quantity and distribution are appropriate for training; Secondly, the proportion of positive samples which have a limited continuously positive effect on model training should be reduced. Thirdly, the importance of difficult examples should be emphasized. In other words, the proportion of difficult examples should be increased during the training process.

The fast RCNN loss function is composed of two parts, in which the classification branch of the RPN network selects BCE cross-entropy loss:

$$L = -y \log y^* - (1-y)log(1-y^*) \tag{3.10}$$

Where y represents the label of the target, which takes a value of 1 when the sample is positive, and 0 when the sample is negative, y^* represents the predicted probability of network for target y, and the range of y and y^* are both from 0 to 1. Obviously, for positive samples, a higher prediction probability the y^* will lead to a lower value of loss function. However, for negative samples, the lower the predicted probability, the lower the y^* will be in the loss function. This design is more likely to lead to a slow training literation phenomenon when the model is trained on the dataset with extreme distribution (too many easy or difficult samples), sometime it is even hard to be optimal.

To solve the above problems, this thesis adds controlling factors into the BCE cross-entropy loss function to control the distribution of samples for an appropriate ratio. The details are as follows:

Firstly, it introduces k as the controlling factor, which balances the negative effect of an unbalanced distribution of positive and negative samples. The value of k ranges from 0 to 1. By this step, the order of training has changed from positive samples to negative samples, which means that the network will focus on the positive samples first, then equation 3.11 changes to:

$$L = \begin{cases} -k \log y^*, y = 1 \\ -(1-k) log(1-y^*), y = 0 \end{cases} \tag{3.11}$$

Secondly, it introduces b as the controlling factor which balance the negative effect of an unbalanced distribution of easy and difficult samples, the value of b ranges from 0 to 1. Based on the design of Focal Loss (Liu et al., 2016), by introducing positive exponential factor b, it decreases the value of $(1-y^*)^b$, which allows the model to focus more on difficult and incorrect classification samples. Combined with equation 3.11, the new loss function becomes:

$$L = \begin{cases} -k(1-y^*)^b \log y^*, y = 1 \\ -(1-k) y^{*b} log(1-y^*), y = 0 \end{cases} \tag{3.12}$$

Lastly, based on Singh's [55] idea of an unbalanced weighted mechanism and Cao's [5] idea of an importance-based dynamic weight mechanism, this thesis proposes a Sample Asymmetric Weighted Loss Function (SAW-Loss), which is shown as equation 3.13:

$$SAWLoss = \begin{cases} -\frac{1}{K_{IoU}}(1-y^*)^b \log y^*, y = 1 \\ -(1-a) y^{*b} log(1-y^*), y = 0 \end{cases} \tag{3.13}$$

On the basis of equation 3.12, it uses the IoU value as the value of controlling factor k when the sample is positive. By doing this, the samples that have high IoU values are more likely to be classified as positive, easy samples, and the value of loss function obtained from these samples will be smaller. Therefore, the loss function will focus less on such samples. In comparison, the attention of loss function will focus more on positive difficult samples, which is more important than positive easy samples for model training. Moreover, because the IoU value calculation is compulsory for the Faster RCNN backbone network, such a design will not generate any extra calculations. As a result, the samples concerned in the training process will become positive-difficult, negative-difficult, positive-easy and negative-easy.

The final loss function for network training is shown as equation 4.1, where the λ is used to control the loss difference between two branches.

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i NLoss(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg} \tag{3.14}$$

3.4 Experiment Result

3.4.1 Training Configuration

In this thesis, the experimental setting of Faster RCNN is strictly followed. The size of the image input to the model is reset to 600*800, and the training data are expanded by means of flipping and splicing. At the same time, the ResNet network adopts the ResNet-101 pre-training model in order to ensure the comparative significance of the training results. The momentum and weight attenuation parameters are set as 0.9 and 0.0001, respectively. The change of learning rate adopts preheating strategy. And the IoU crossover ratio threshold is set to 0.7. For parameter setting, refer to the setting of b and λ in Focal Loss, where $b = 2$ and $\lambda = 2$.

The total loss curve in the training process of the network model is shown in Figure 3.7. Compared to the loss function in the Faster RCNN, the improved classified SAW-Loss can converge to a stable state faster, which significantly speeds up the training process and reduces the loss. This effectively improves the learning ability of the network model for positive and difficult samples.

3.4.2 Ablation Experiment

In this thesis, the improved multi-layer feature cascade polymerization pyramid network MCA-FPN and the improved sample equalization loss function SAW-Loss are introduced into the original Faster RCNN. Through training and testing the PASCAL VOC dataset, the accuracy and the number of parameters of different improved models were obtained, and the effectiveness of the proposed scheme in this thesis is demonstrated.

We used ResNet-101 as the backbone network to extract image features, and the mAP of Faster RCNN on the PASCAL VOC dataset reached 79.8%. After adding feature pyramid

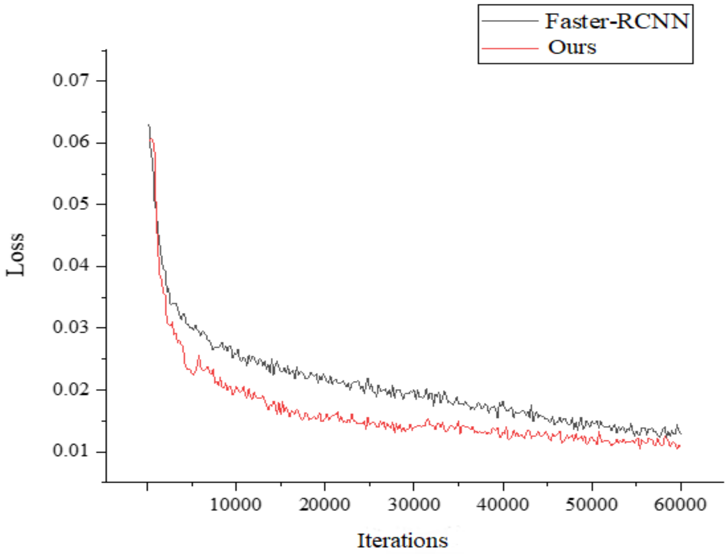


Figure 3.7: The comparison of training process between proposed method and Faster-RCNN

Table 3.3: Experimental results on Pascal VOC dataset

Steps	Backbone Network	Added Scheme	mAP(%)	Parameters (M)
1	ResNet-101	/	79.8	160.2
2	ResNet-101	FPN	82.1	163.5
3	ResNet-101	SAW-Loss	83.6	163.6
4	ResNet-101	MCA-FPN	84.5	165.4
5	ResNet-101	MCA-FPN + SAW	86.0	165.5

FPN, the mAP was improved to 82.1%. When FPN is not added, the loss function changes to the SAW-Loss function and the mAP improves to 83.6%. Without modifying the original loss function, the FPN of the feature pyramid was changed to the MCA-FPN of the multi-stage feature cascade aggregation pyramid module, and the mAP reached 84.5%. Finally, when we used both the MCA-FPN module and the sample equalizing loss function SAW-Loss, the mAP improved to 86.0%. As the network complexity is smaller, the model is more suitable for industrial landing and application. Compared with Faster RCNN+ResNet-101, the accuracy of the proposed improvement increased by 6.2%, and the model parameters only increased by 5.3M, indicating that the proposed improvement did not introduce too much extra calculation.

3.4.3 Plug and Play Experiment

MCA-FPN, a multi-layer feature cascade aggregation pyramid module, optimizes features after feature extraction from the backbone network and improves the ability of the target detection model to deal with scale problems. In this thesis, four typical target detection networks are selected for the multi-layer feature cascade polymerization pyramid network MCA-FPN, and MCA-FPN is embedded into the target detection network. Experiments are carried out in the PASCAL VOC dataset, and the experimental results are compared with the original structure. The specific results are shown in Table 3.4.

Table 3.4: Plug and play performance comparison table

Model structure	Original mAP(%)	MCA-FPN(%)	Promote(%)
YOLOv4 [3]	60.6	68.9	13.6
ResNet-FOCS [68]	78.7	85.2	8.2
ResNet-RetinaNet [55]	80.7	84.7	4.9
ResNet-Faster RCNN[54]	79.8	84.5	5.8

As shown in Table 3.4, the MCA-FPN proposed in this thesis can be introduced into other target detection models as a plug-and-play module, and all of them have certain performance improvements and universality. At the same time, the detection effect of the single-stage target is more significant, indicating that the structure has an obvious improvement effect on multi-scale problems.

3.4.4 Performance Comparison

As shown in Table 3.5, this thesis lists the accuracy results of the target detection-related models on VOC datasets in the last five years. It can be seen that the mAP of the network model proposed in this thesis is 1.9% higher than the PFPNet but 0.5% lower than NAS Yolo, the competition model. The effectiveness and superiority of the proposed method are fully proved.

Table 3.5: Comparison table of algorithm accuracy

Algorithm	mAP(%)	Algorithm	mAP(%)	Algorithm	mAP(%)
MLKP [66]	80.6	RefineDet [77]	83.8	FCOS [57]	78.7
RDAD [1]	81.2	PFPNet [31]	84.1	HKRM [28]	78.8
RFBNet [29]	82.2	NAS Yolo	86.5	**MCA+SAW**	86.0

PASCAL VOC dataset contains 20 detection categories. This thesis compares the accuracy of the improved detection model with that of the original baseline model, and the accuracy of each category is shown in Figure 3.8. By introducing the MCA-FPN module, it can effectively deal with the multi-scale problems of similar targets and is more friendly to the detection of small-size targets. The part of chairs and tables in the picture often occupies most of the space, which will be hard for the network to distinguish the difference between foreground and background, positive and negative samples. This thesis designed a SAW-Loss function which gives different loss function structures to positive and negative samples and uses the IoU value as the asymmetric weight of loss function for easy samples. Such a method effectively improves the network learning ability for positive samples and difficult samples.

The method designed in this thesis is intended to improve the ability of the object detection model to deal with size and sample division problems. Combined with the above experiments, various accuracy diagrams clearly showed the effectiveness of the improved method. Among them, the improvement of the detection accuracy of some typical categories truly reflects that the method can better solve the problems faced by the existing detection model. The detection effect diagram of some categories is shown in Figure 3.9.

The above pictures show the difference between the result of the Faster RCNN+ResNet structure (left side) and our proposed method (right side). This original image contains some difficult detection tasks, such as multi-scaled targets detection and similar background and targets detection, so the result in this picture could illustrate clearly how the proposed method improves performance. Firstly, our proposed method decreases the missing rate,

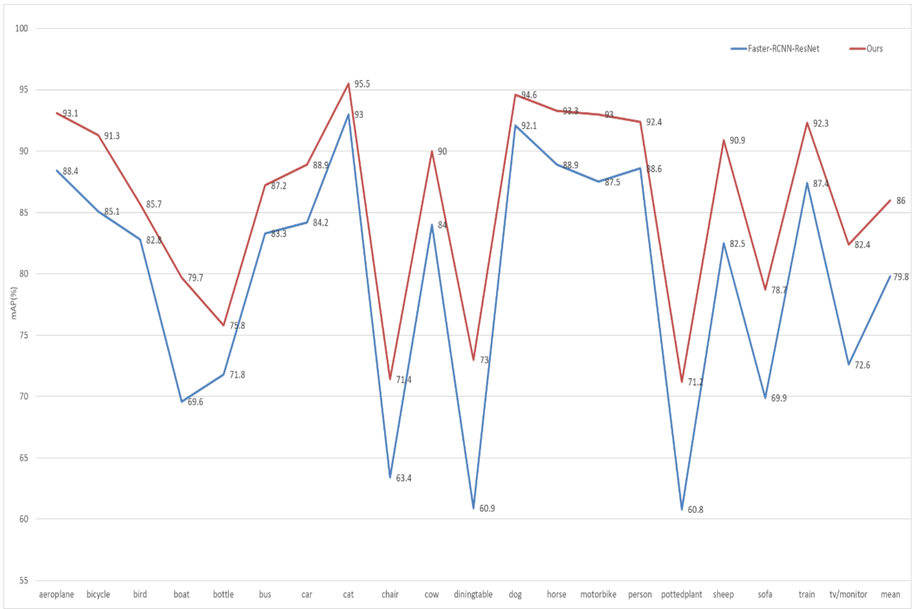


Figure 3.8: The accuracy of the proposed method on different categories

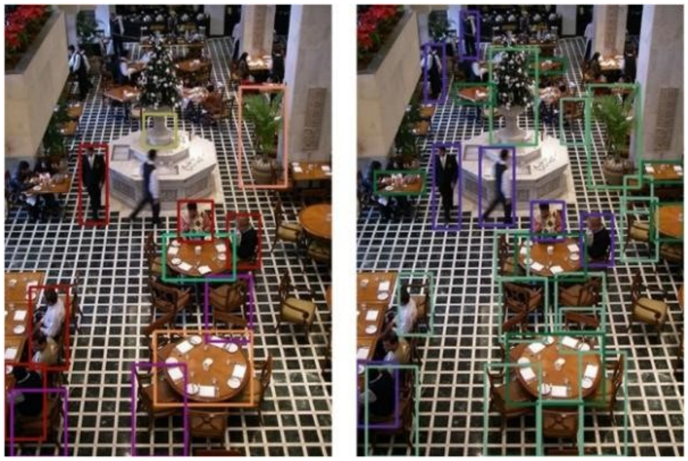


Figure 3.9: The visualization comparison between proposed method and Faster-RCNN

such as the successful detection of the waiters in the left top of the image, which fails to be detected by Faster RCNN. Secondly, our proposed method increases the detection accuracy, such as the accurate detection of the tree in the middle of the image. Moreover, our proposed method could even detect the waiter in the middle of the image even if the target is a blur.

3.5 Experiment Result

To solve the problem of fast RCNN being insensitive to feature scale and the problem of imbalance in sample distribution, this thesis proposes a multi-level feature cascade aggregation pyramid module MCA-FPN and sample equalization loss function SAW Loss. The MCA-FPN module adds two different direction connections between the feature maps, which helps the model extract features more efficiently and analyze information more accurately. The new sample equalization loss function SAW-Loss helps the network to use data with a different distribution more efficiently and converge faster by adding different types of controlling factors. As shown from the ablation experiment result, our proposed MCA-FPN module and SAW loss function could improve the mAP of the traditional ResetNet -101 model by 2.3% and 3.8%, respectively, and by adding both of these two modules, the final mAP could increase to 86.0% which is much higher than the mAP of original ResetNet-101 with 79.8%. Meanwhile, the result of the plug-and-play experiment proves that our proposed modules could be applied to another backbone network with different levels of promotions. Finally, this thesis also makes a performance comparison between our proposed method and other mainstream methods. Our proposed method performs better than most of the published methods except NAS Yolo, which is a competition algorithm and not published.

Chapter 4

Target Detection Based on Feature Similarity and Attention Weighted Information Aggregation

In the previous chapter, our proposed method has been shown to improve the performance of an image-based target detection algorithm, but it still struggles to solve problems such as target occlusion and blur target from the image dataset because the information extracted from a single image is limited. Such a situation could be worse when the data collection equipment is not qualified. In practice, people could confirm the targets using similar targets from other related images. Fortunately, it is possible for deep learning methods to solve this type of problem by imitating human operations.

In this chapter, we proposed a type of video-based target detection algorithm to extract and combine the feature information from a series of continuous frames, which provides important supplementary information for target localization and classification. The detail of the proposed method are as follows:

- A local information aggregation method - this method designs a type of simi-larity screening mechanism to select the most related feature information to combine with the current frame for network prediction.
- A global information aggregation method based on attention weight correla-tion co-efficient - this module improves the existing global information aggre-gation method by calculating the correlation coefficient among all potential proposals as a way of enhancing for the current frame.

4.1 Motivation of Video-based Target Detection

One of the main differences between the image-based target detection algorithm and the video-based target detection algorithm is that the latter method could extract more supplementary timing and spatial information than the former. This thesis classifies this type of supplementary information into two groups; one is local localization information which is used to confirm the location of the target. Only adjacent frames could be used because the location difference between two global frames is too large to confirm the target location. Another one is global semantic information which makes use of all frames to confirm the category of targets. The reason for deploying this method is to take into account the fact that targets from adjacent frames could still be unclear and that the category of the target remains unchanged in video data. The process of how to combine these two types of information is shown in Figure 4.1.

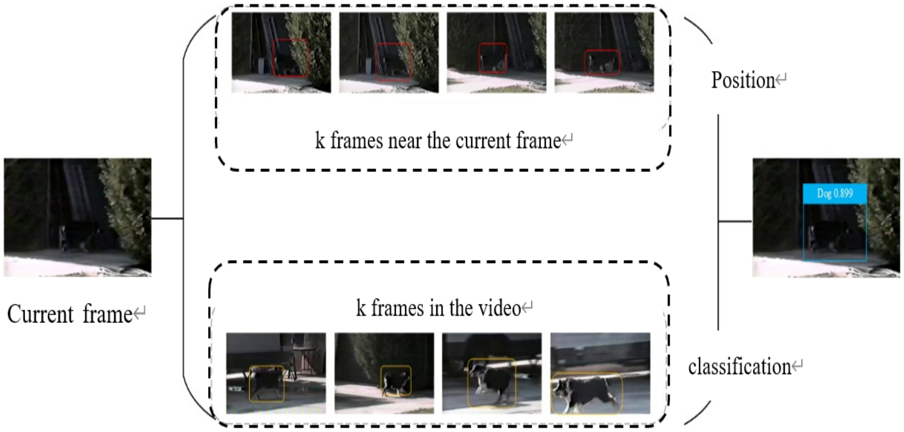


Figure 4.1: The process of information combination

If a network can confirm both category and location of the target, it would have a successful detection. Based on this idea, researchers proposed two methods that are based on local information aggregation and global information aggregation. The two kinds of methods are shown in Figure 4.2.

Besides, a summary about the limitation of these methods are also shown in Table 4.1 below.

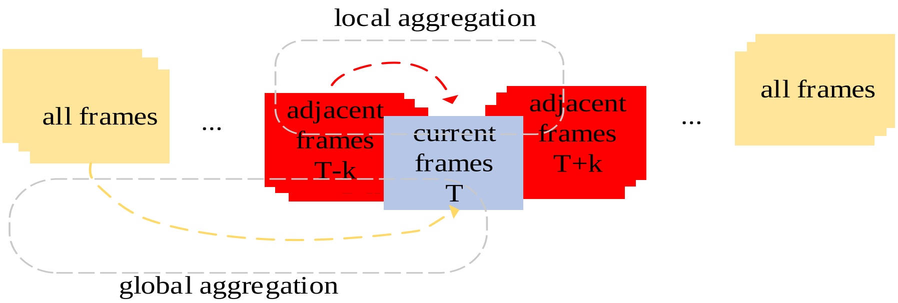


Figure 4.2: Local information aggregation and global information aggregation

Table 4.1: Summary of video-based methods

Methods	Limitations
Full connection	Too much information and hard to process
Local aggregation	Lack of global semantic information and hard to confirm category
Global aggregation	Lack of localization information and hard to confirm location

4.2 Analysis of The Feature Aggregation Method

As we have introduced the two types of video-based target detection methods in the previous sections, this section discusses the details of these methods.

Regarding the local aggregation method, it aggregates the feature information from adjacent frames to the current frame. For example, both FGFA [84] and MANet [69] use FlowNet [2] to calculate and utilize the movement difference of the target for local feature enhancement. The basic idea for such a method is that the optical flow could reflect the target movement from a series of continuous frames. However, optical flows need to be trained independently and supported by a large amount of computation power. In order to address such a need, STMN [72] uses RNN [59] network and DCN [9] network to learn the alignment feature from the current frame without the need for the optical flow calculation. In all, this method only focuses on the current frame and its adjacent frames, which has a limited amount of information aggregation.

The global aggregation method analyzes the semantic feature correlation between the current frame and other global frames, and features are enhanced based on such a correlation relationship [71] [12] [23]. This method, on the one hand, breaks the spatial limitation in the local aggregation method, but on the other hand, it loses the advantage of target localization based on local information.

By combining the pros and cons of the two methods, Memory Guided Attention for Category (MEGA) [6] is proposed. It uses RPN to generate some potential proposals from adjacent frames of the current frame and other global frames. A training tip used by authors here is that they use the shuffled video frames with limited numbers. Such a design could distinguish global information from local information with a slight increase in the number of parameters. Then it uses the relation module to aggregate the corresponding feature of global information to the proposal of local frames. In this way, the local information is combined with global information successfully.

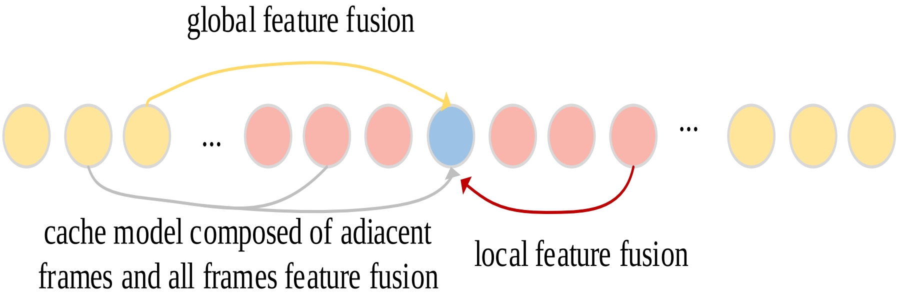

global feature fusion
...
...
cache model composed of adiacent
frames and all frames feature fusion
local feature fusion

Figure 4.3: The basic idea of MEGA

Besides, as inspired by Transformer-XL [10], MEGA designs a memory module to increase the amount of information sent for classification and regression, which is shown in Figure 4.4.

There are two main operations for MEGA with global aggregation and local aggregation, respectively. It starts with the global aggregation stage, which aggregates the information from shuffled video frames to some specifically ordered video frames based on a type of relation module. Such an operation could enhance the local information by its corresponding global information and provide the network with more correlated target feature information. Then that aggregated information will be passed through three local aggregation stages to obtain more enhanced target feature information. In this way, the local information and global information of the target could be aggregated successfully and used for final regression and classification tasks. Besides these basic two operations, MEGA also designs a type of memory module to provide the model with more correlated information from previous global aggregation stages and local aggregation stages. For example, assuming that MEGA conserves the three frames (i.e. shuffled video frames, ordered video frames and keyframes) as a memory module, with two local aggregation stages, the network could have extra six frames of information. Besides, as every local frame is enhanced by its corresponding global frame, there will be another six extra global frames added. As

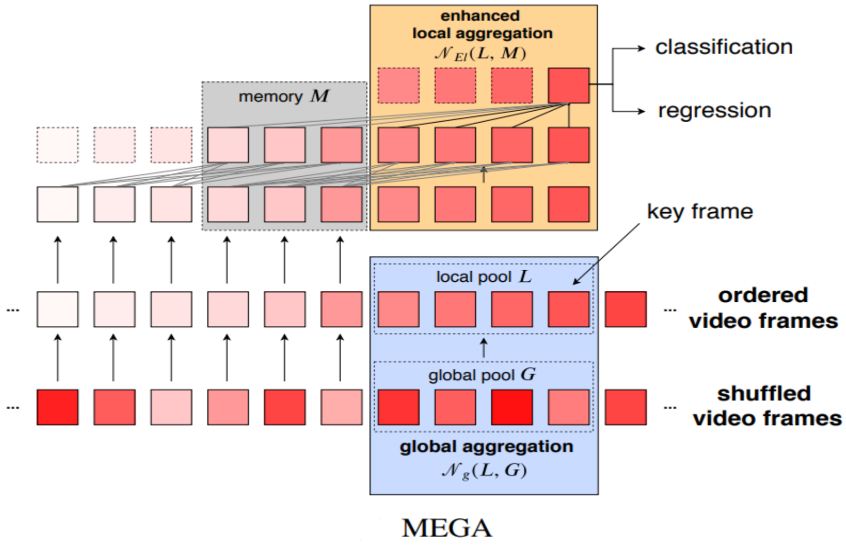


Figure 4.4: The structure of MEGA

a result, a three-frame memory module could add information of 12 extra frames without any extra calculation other than just additional storage spaces.

In summary, MEGA improves the efficiency and the amount of feature information utilization for the network without introducing too many extra modules and computation demands. The idea of MEGA is similar to the attention mechanism that has been introduced in many other networks, which results in a great performance [7] [42] [61]. The core concept of the attention mechanism is to allow the network to imitate human's selective attention, which can transfer the network's attention from the whole image to the specific interesting area. Such a mechanism could decrease the detection time consumed by nontarget area detection.

However, compared to the attention mechanism, MEGA still has its limitations when it aggregates the related feature information. Firstly, the aggregation stage from global information to local information is a type of random selection from shuffled video frames, which lacks correlation analysis. Secondly, the memory module proposed by MEGA sets a fixed size of storage for local frames. Such a design makes it hard to aggregate correlated frames to the keyframe and may lead to limited information efficiency.

As a result, this thesis proposes two corresponding methods to improve the performance of the existing video-based target detection algorithm.

4.3 Local Information Aggregation Based on Feature Similarity

Regarding the "hard" examples which are occluded in the current frame from a series of video datasets, it is highly likely that the clear information of this target could be extracted from its adjacent frames if the target is continuously moving. Therefore, MEGA proposed a method to aggregate feature information from adjacent frames to the current frame, which helps solve the above problem. Moreover, if we add an extra correlation analysis mechanism to select more correlated adjacent frames, the algorithm could learn more correlated feature information of the target, and the performance of the algorithm would improve.

Based on this idea, this thesis proposes a type of correlation filter mechanism, which is shown in Figure 4.5.

In the fast RCNN network, the region proposal represents the region which is most likely to include the target in a feature map, and such region proposals could be obtained after being passed through by the ROI pooling layer and RPN network. Compared to aggregating the feature information of the whole image, such region proposal aggregation will be more efficient due to less consumption of time and computation power. Therefore, this thesis selects the region proposal from the current frame generated by the RPN network as the target and iteratively searches for the most correlated region proposals from its adjacent frames. Then, it produces a table of correlation enhancement for feature aggregation.

The detailed process of this correlation filter method is shown as follows:

Firstly, this thesis confirms the current frame as F_t and the range of its corresponding adjacent frames. For the k^{th} region proposal in current fame RoI_t^k, we assumed that it firstly searches the most correlated region proposal from the previous frame F_{t-1}. Normally speaking, the information of the target's location and feature would not be changed significantly. Therefore, by calculating the feature similarity $S(\bullet)$ and location similarity $L(\bullet)$ between the current frame F_t and the previous frame F_{t-1}, we can find the region proposal RoI_{t-1}^k that is best matched with RoI_t^k. Then the iterative search is adopted based on the above process to find the best-matched RoI_{t-2}^k. This process is repeated until we find the best-matched region proposal from $F_{t-\tau}$ and $F_{t+\tau}$. Such iterative search process based on the feature similarity and location similarity, we can find the best-matched region proposals from adjacent frames. The equation of the similarity selection is shown in Equation 4.1:

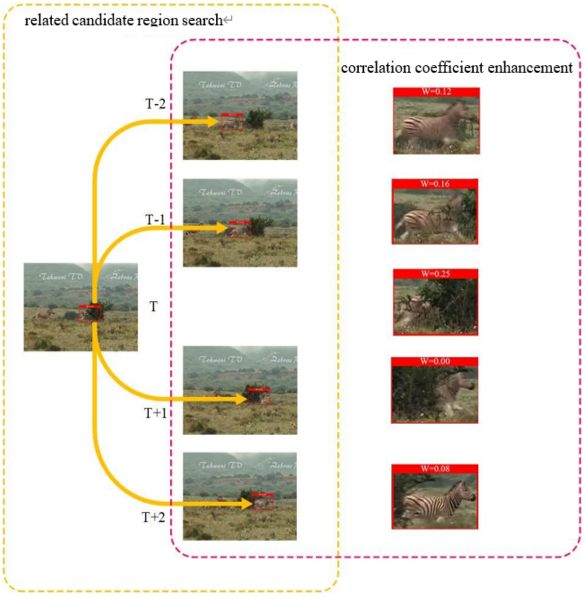

related candidate region search↵
correlation coefficient enhancement
T-2
T-1
T
T+1
T+2
W=0.12
W=0.16
W=0.25
W=0.00
W=0.08

Figure 4.5: The process of proposed local information aggregation method

$$k_{t-1} = max(S(Fea_t^{k_t}, Fea_{t-1}), L(RoI_t^{k_t}, RoI_{t-1})) \tag{4.1}$$

For the feature similarity $S(\bullet)$, this thesis selects Euclidean distance to calculate the similarity. Based on the definition of Euclidean distance, for a n-dimensional space, the distance d between two points is shown in Equation 4.2 [64]:

$$d(x, y) = \sqrt{(x - y)^2} \tag{4.2}$$

What is more, as the feature of two region proposals is represented as Fea_1 and Fea_2 in a matrix form, every point similarity from these features could be calculated by above equation. Finally, the sum of all point distances could be defined as the result of the possibility that two features belongs to the same category, which is show in Equation 4.3:

$$s(Fea_1, Fea_2) = \sum d(Fea_1, Fea_2) \tag{4.3}$$

Equation 4.4 is the equation of location similarity which consists of two parts, one is used to calculate the scale similarity and the other one is used to calculate the coordinate similarity. X and Y represent the width and height of proposal respectively, d_x and d_y are produced by the regression branch of Faster RCNN which represents the center and location bias of region proposal respectively:

$$L(RoI_1, RoI_2) = L1(RoI_1, RoI_2) + L2(RoI_1, RoI_2) \tag{4.4}$$

$$L1(RoI_1, RoI_2) = \min(\frac{w^1}{w^2}, \frac{w2}{w^1}) \times \min(\frac{h^1}{h^2}, \frac{h^2}{h^1}) \tag{4.5}$$

$$L2(RoI_1, RoI_2) = \exp(-\frac{\left\|(d_x^1, d_{\mathrm{y}}^1) - (d_x^2, d_{\mathrm{y}}^2)\right\|_2}{\sigma^2}) \tag{4.6}$$

After the above process, the network could find a series of region proposals which are highly correlated with the specific region proposal in the current frame. The next step is to aggregate those correlated region proposals from the adjacent frames to the region proposal from the current frame. As this thesis introduces the attention mechanism to improve the aggregation efficiency, the aggregation equations [79] are shown in Equation 4.7 and 4.8. The features from adjacent frames are aggregated by the weighted sum of their corresponding matrix, as shown in Equation 4.8, where the function $S(\bullet)$ is obtained from Equation 4.3:

$$Fea_t^{k_t} = \sum_{i=t-\tau}^{t+\tau} w_i^{k_i} Fea_i^{k_i} \tag{4.7}$$

$$w_i^{k_t} = \frac{\exp(s(Fea_t^{k_t}, Fea_i^{ki})}{\sum_{i=t-\tau}^{i=t+\tau} \exp(s(Fea_t^{k_t}, Fea_i^{ki})} \tag{4.8}$$

As the essence of region proposal is a series of features surrounded by a bounding box. Therefore, by calculating the feature similarity included in these region proposals and the location similarity of these bounding box, we can obtain a series of correlated region proposals and use them to enhance the current feature, which can significantly increase the information utilization efficiency and correlation.

4.4 Global Information Aggregation Based on Attention Weighted Correlation Coefficient

The proposed method from the previous chapter could aggregate the feature in an efficient way between current frames and adjacent frames. But if we apply this method to select global correlated frames from all video sequences, the process will be complex, and the time and computation cost will be extremely high. This will not be acceptable even if the performance may be improved. Therefore, in this chapter, we propose a weighted correlation aggregation method to enhance the current feature with other global correlated features. The proposed method from the previous chapter could aggregate the feature in an efficient way between current frames and adjacent frames. But if we apply this method to select global correlated frames from all video sequences, the process will be complex, and the time and computation cost will be extremely high. This will not be acceptable even if the performance may be improved. Therefore, this chapter proposes a weighted correlation aggregation method to enhance the current feature with other global correlated features. First, this method builds a type of correlation coefficient matrix to allocate different weights to different global proposal features based on their correlation to the current proposal frame. Then it uses the feature regularization operation to derivate the relationship between the weighted correlation coefficient matrix and feature similarity of the current proposal feature and global proposal features. Besides, this method also proposes a supervision loss function to decrease the distance between similar targets and increase the distance between different targets, which can obtain a better training result.

The process is shown in Figure 4.6:

Assume that all features are included in a D-dimensional space, there are N proposals from current frame and M proposals from all global frames, where $X_{current} \in V \times D$represents the proposal feature in current frame and $X_{global} \in M \times D$ represents the proposal feature from all global frames. As inspired by the attention mechanism, we compute the attention weights between the proposals from detected frame and the proposals from global frames and use these weights as correlation coefficient. We use X_i to represent the value of i^{th} column in $X_{current}$ matrix and X_j to represent the value of j^{th} row in X_{global}. The details of the algorithm are illustrated below.

The current proposal feature $X_{current}$ and global proposal feature X_{global} are firstly sent to a fully connected layer to aggregate them together (Step 1).

Then the network makes a feature regularization operation on the output produced by step 1, the essence of which is to do regularization operation on multiple ROI matrixes (Step 2). In fact, those ROI matrixes are obtained by different type of ROI pooling operation

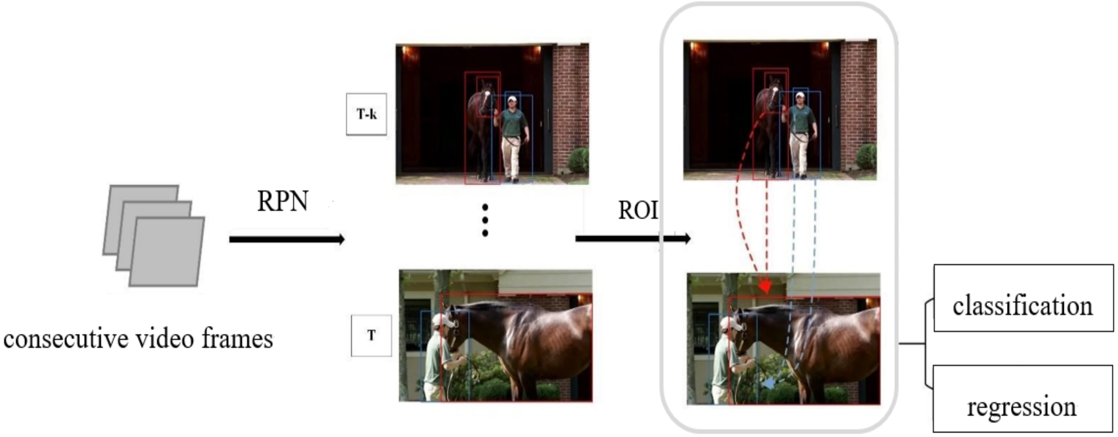


Figure 4.6: The process of weighted correlation aggregation method

which may insert some values in original matrixes to size changing. However, these values are not correlated with original features so such ROI pooling operation may pose negative effect on final classification. By using regularization on ROI matrix, the network could decrease such negative effect generated by the inserted features [71].

After step 2, the network starts to set up correlation relationship between current frame and global frames by building a correlation coefficient G with the size of $N \times M$:

$$G = G(X_{current}, X_{global}) : R^{N\times D} \times R^{D\times M} \to R^{N\times M} \tag{4.9}$$

The matrix is used to calculate the attention weight of X_{global}, the higher the value of $G_{i,j}$, the higher the correlation relationship between X_i and X_j.

If X_i appears in some global frames, the network should use its corresponding global proposal to enhance the current frame. Besides, this thesis introduces another supervision loss function which decreases the distance between similar targets and increases the distance between different targets. This supervision loss function is then added it into the loss function of whole network (see Equation 4.10):

$$L(x_i, x_j, y_{i,j}) = \frac{1-y_{i,j}}{2} d(x_i, x_j)^2 + \frac{y_{i,j}}{2}[max(0, \mu - d(x_i, x_j))]^2 \tag{4.10}$$

where $y_{i,j}$ represents the irrelevance degree between two proposals, $y_{i,j} = 1$ represents different proposals and $y_{i,j} = 0$ represents the same proposals.

For feature regularization which leads to $\|x_i\| = \|x_j\| = 1$, we use L_2 scalar product to obtain the relationship between G and d:

$$d^2(x_i, x_j) = \|x_i\|_2^2 + \|x_j\|_2^2 - 2(x_i, x_j) = 2(1 - G_{i,j}) \tag{4.11}$$

Finally, the matrix G is used as the attention weight of X_{global}, each row of data would be regularized by Softmax function to derive the regularized function $G^{'}$. Here, the method of feature aggregation is the same as Equation 4.7 but the weight is replaced by matrix $G^{'}$.

4.5 Experiment Result

4.5.1 Training Configuration

This thesis has tested the performance of the proposed algorithm based on two public datasets, namely, ImageNet DET and VID. It selects 38 hundred video footage for training, five hundred for validation and nine hundred for testing. What is more, for a faster convergence trend, this thesis selects the RestNet-101 network and uses the pretrained module given by the official author of this network. As for the detail of input image processing, Table 4.2 shows some setting parameters:

Table 4.2: Input image processing

Item	Setting
Max/Min training batch size	1000/600
Max/Min testing batch size	1000/600
Horizontal turnover rate	0.5
Mean value of pixel	[102.9801, 115.9465, 122.7717]

For the backward feedback, this thesis selects the SGD method where the parameter of momentum and weight attenuation is set to 0.9 and 0.0001, respectively. The total training batch is 120000, and the learning rate is changed based on the linear warmup strategy. The learning rate will increase to 0.001 in the first 500 batches and decrease to 0.0001 after 80000 bathes of training. The relationship between the learning rate and the loss function is shown in Figure 4.7:

The loss value shows a downward trend which decreases rapidly in the first 500 batches. After that, the learning rate is stabilized at 0.001 until the 80000 batches and the loss value of network decreased steadily with a final value of 0.35.

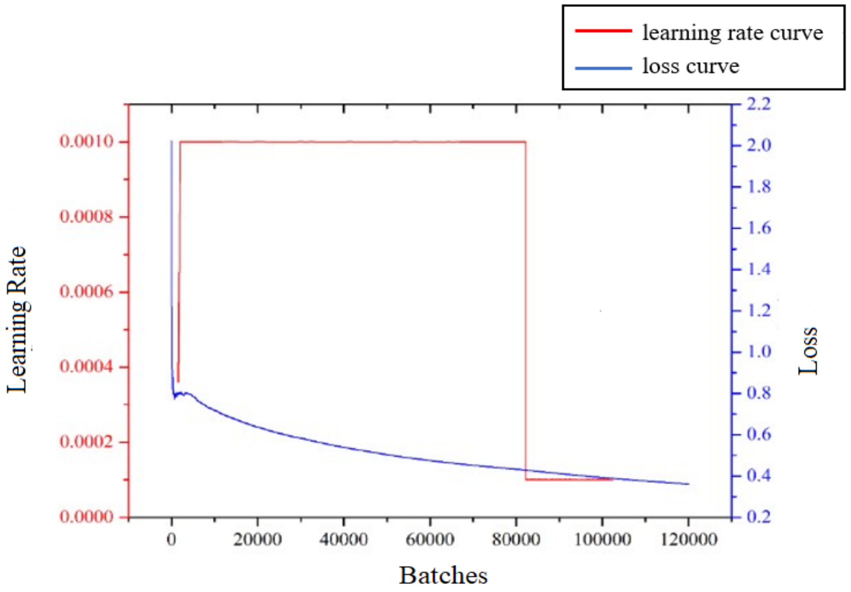


Figure 4.7: The trend of loss value with the change of learning rate from 0.001 to 0.0001 at the batches of 80000

4.5.2 Ablation Experiment

This thesis tests the performance of the proposed local information aggregation method and global information aggregation method. For convenience, we set M1 to represent the proposed local information aggregation method and M2 to represent the proposed global information aggregation method. All algorithms are trained on ImageNet DET&VID dataset and tested on the VID dataset. The detail of the experiment result is shown in Table 4.3:

Firstly, this thesis trains the basic Faster RCNN and Faster RCNN New networks to obtain their performance accuracy. As we can see from the table 4.3, both networks are not performing very well, with the mAP of 73.4% and 75.4%, respectively. Since both Faster RCNN and Faster RCNN New network are image-based target detection algorithms, such a result demonstrates that the traditional still image-based target detection algorithms are not fitted for video data.

We also test the performance of the video-based target detection algorithm MEGA which improves on Faster RCNN with a mAP of 82.9%. Moreover, it does another experiment which still uses MEGA as a basic algorithm but replaces Faster RCNN with Faster

Table 4.3: Ablation experiment result

Network	MEGA	Module M1	Module M2	mAP(%)
Faster RCNN	-	-	-	73.4
Faster RCNN_New	-	-	-	75.4
Faster RCNN_New	✓	-	-	82.9
Faster RCNN_New	✓	✓	-	83.8
Faster RCNN_New	✓	-	✓	84.0
Faster RCNN_New	✓	✓	✓	84.6

RCNN New and obtains higher accuracy of 84.0%. Such a result shows that by appropriately aggregating global information and local information, the algorithm could learn more correlated features and increase the accuracy of detection.

Based on MEGA, this thesis introduces a new local information aggregation method as the M1 module, which aggregates correlated proposals from adjacent frames to the current frame. In this way, the network could learn more correlated information to enhance target features. As a result, the M1 module improves the accuracy to 83.8%, which is 0.9 percentage points higher than MEGA. Such limited improvement of 0.9 percentage point is explained by the fact that most parts of the dataset only contain a few visible targets, so M1 could only extract limited correlated information from adjacent frames. If the adjacent frames that are used to aggregate the current frame contain more irrelative features to target, the accuracy difference between our proposed method and MEGA would be larger.

The global information aggregation method (M2) is also introduced to basic MEGA, which increases the mAP by 1.1 percentage points compared to the performance of the original MEGA.

Lastly, this thesis introduces both M1 and M2 modules to the network and obtains the highest mAP value of 84.6%, which is 11.2 percentage points higher than Faster RCNN alone.

Figure 4.8 shows how the mAP changes with the training batch in MEGA and our proposed method. In comparison, our proposed method performs better than MEGA at every training size, and finally, its mAP is 1.7% higher than MEGA.

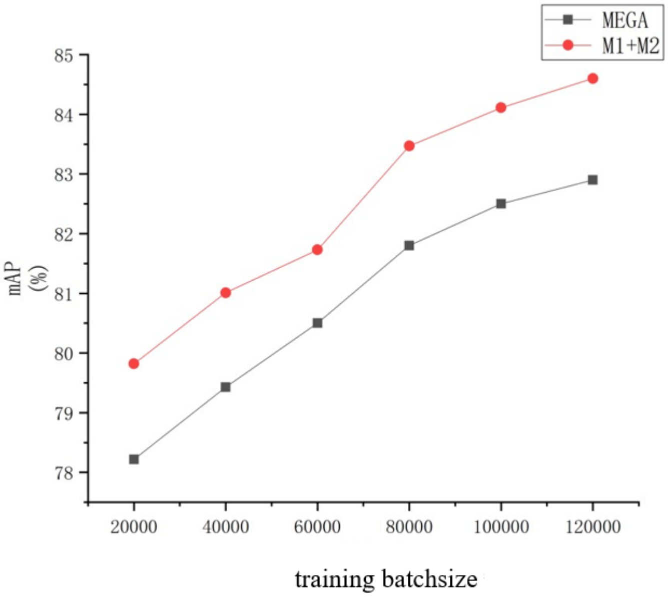


Figure 4.8: The comparison of mAP between the MEGA and the proposed method

4.5.3 Objective Performance Comparison

As shown in Table 4.3, we compared the performance of our proposed methods with a list of existing mainstream algorithms, and all results are obtained under the same testing environment.

The previous section shows that our proposed method performs the best among all ResNet101-based algorithms, which proves the efficiency and advancement of our method. In addition, some of these methods use the DCN module, which is a type of variable convolution to increase the acceptance area of the feature map, but this module is not the key point of our related work, so we do not discuss it here.

For the algorithm reasoning speed, this thesis constructs Figure 4.9 to demonstrate the relationship between mAP and its corresponding processing time.

Typically, algorithms with high accuracy require more time to process, so the ideal algorithm should balance the accuracy and time consumption in an efficient way. As shown in Figure 4.9, our proposed method has the highest mAP value among all six algorithms and has a similar processing time with MEGA for around 9FPS.

Besides, this thesis also evaluates the performance of algorithms on targets with slow

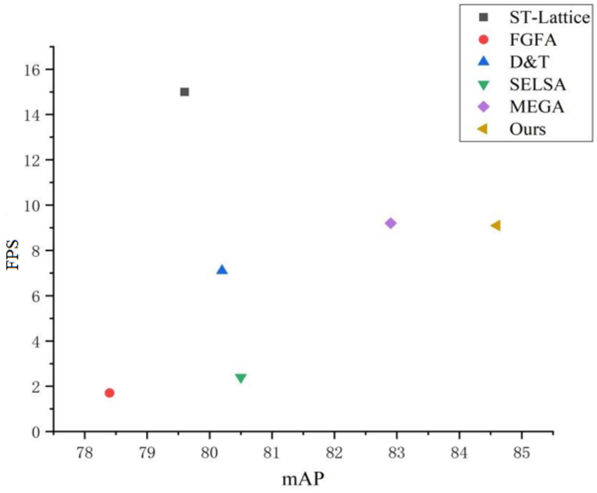


Figure 4.9: The comparison of overall performance between the proposed method and others

speed, medium and fast speed. We use the IoU value to distinguish three categories, and the criterias are shown in Table 4.4:

Table 4.4: The criteria for slow-speed, medium-speed and fast speed target

	Slow-speed target	Medium-speed target	Module Fast-speed target
IoU	$IoU > 0.9$	$0.9 > IoU > 0.7$	$0.7 > IoU$

The performance on different kinds of target is shown in Table 4.5:

Our proposed method has the best performance on all three types of targets with an AP of 89.9% for slow-speed target, 83.1% for medium-speed target and 69.3% for fast-speed target.

4.5.4 Human Visual Evaluation

In the previous chapter, we have introduced the efficiency of our proposed method by figures, and in this chapter, we will visualize the performance of our proposed method based on the best-trained model.

We select the VID dataset as a validation set, which includes over 30 categories of

Table 4.5: Comparison table of algorithm accuracy

Algorithm	AP(Slow)	AP(medium)	AP(fast)
DFF [85]	84.5	73.5	48.3
FGFA [84]	85.6	76.9	55.3
MANet [69]	86.9	76.8	56.7
RDN [14]	87.7	79.4	60.2
MEGA [6]	89.4	81.6	62.7
HVRNet [23]	88.7	82.3	66.6
Ours	**89.9**	**83.1**	**69.3**

targets and because of the space limitation of this thesis, we select three classical scenarios based on different moving speeds.

Firstly, we use the experiment result of a giant panda to demonstrate the algorithm performance on a slow-speed target. As shown in Figure 4.10, our proposed method could detect pandas successfully even if, in some images, the feature panda is insufficient as they are cuddled together, and the accuracy for such cases reaches around 80%.

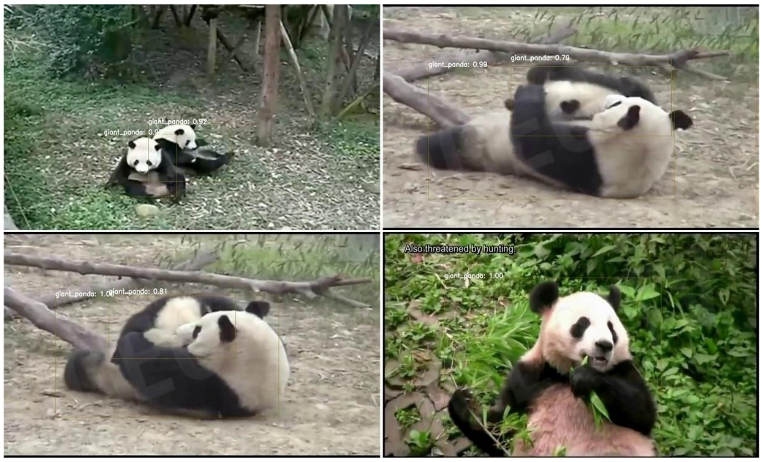


Figure 4.10: The visualization result of proposed method on slow-speed target

Then we use the experiment result of a moving panda and whale to demonstrate the algorithm performance on fast-speed targets. As shown in Figure 4.11, even if most part of the moving panda is occluded from stone, our proposed method could still detect and recognize it successfully with a precision of 87% based on correlated information extracted from other frames. As for the whale, the top-left image shows that our proposed method could still detect it correctly based on the feature of its mouth only.

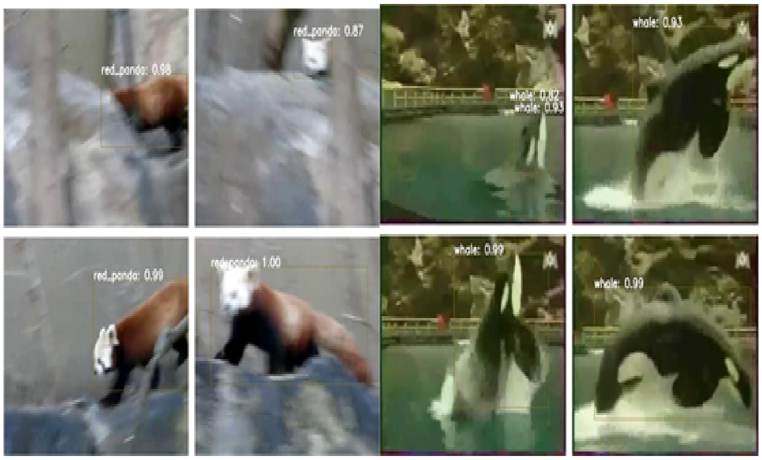


Figure 4.11: The visualization result of proposed method on fastw-speed target

4.6 Conclusion

As the performance of the traditional image-based detection algorithm decreases significantly when they are applied to video datasets, this thesis proposed two video-based target detection algorithms, which increase the efficiency of feature extraction and utilization. One is the local information aggregation method based on feature similarity, and another is the global information aggregation method based on the attention weight correlation coefficient. Both of them focus on the correlated relation-ship between the proposal of the current frame and the proposal of other frames. Then, based on these correlation relationships, our proposed method produces the feature similarity and attention correlation coefficient, which are used as the weight for feature aggregation.

To validate the efficiency of our proposed method, this thesis tests the performance of the public dataset ImageNet VID. According to the comparisons with other video-based mainstream detection algorithms, our proposed method shows to perform better in terms of accuracy and reasoning speed.

Chapter 5

Conclusion and Future Plan

5.1 Conclusion

With the development of the machine learning method, target detection technology has been widely applicated to various fields, and smart security is one of the most common fields. However, the traditional target detection algorithm does not perform well on the practical dataset, and the reasons can be attributed to the following points: (1) The scale of feature is changed greatly for different types of targets, which makes it harder for the network to extract different scales of feature efficiently. (2) The distribution of the dataset may not fit the existing structure of loss function, which may cause low training efficiency and detection accuracy. (3) The traditional image-based target detection does not address the problem of target occlusion and shape change very well because the feature information provided by still image is too limited for detection. In order to address the above problems, this thesis proposes a series of methods to improve the performance of the target detection algorithm from various aspects.

Firstly, I propose an improvement based on the traditional feature pyramid structure for better feature extraction and combination. Secondly, I propose a type of sample asymmetric weighted loss function to balance the distribution of different types of samples. Then, while addressing the existing problems of the image-based target detection method, we also aggregate the global feature information and the local feature information of the target based on the location and semantic correlation. Finally, as backed by experiment results, it can conclude that the proposed method performs better than other mainstream target detection algorithms based on the public dataset and the proposed method improves the accuracy and efficiency of the target detection algorithm. The details of novelty are described below.

This thesis firstly improves the basic Faster RCNN algorithm by proposing a multi-layer feature cascade aggregation pyramid network which is inspired by the hierarchical feature extraction structure from ResNet. The proposed method could aggregate different scales of semantic feature information horizontally and vertically based on the feature pyramid structure to enhance the performance of the algorithm on multi-size target detection tasks. What is more, the proposed improved feature pyramid can also be applicated in other backbone networks and achieve algorithm migration. Secondly, a sample asymmetric weighted loss function is designed to address the problem of sample distribution imbalance. Such loss function allows the network to focus more on those positive hard examples and helps the network to learn more useful feature information. As shown by the experiment results, the proposed methods increase the performance of the basic network.

Thirdly, based on the above improvements on the image-based target detection algorithm, this thesis furtherly reviewed the problems of the video-based target detection algorithm. As most of the existing methods lack correlation analysis when aggregating features from different frames, I proposed a type of video-based feature aggregation method based on feature similarity. By aggregating correlated local information from adjacent frames to the current frame, the network could extract and analyze more target-related information for the final target localization and classification.

Finally, I proposed a global feature aggregation method based on the correlation coefficient to be used as the weight of an attention mechanism. The network finally aggregates the global feature information to the current frame based on its corresponding attention weight, which aggregates more correlated information instead of random aggregation. Our experiment results show that our proposed method is competitive in both accuracy and reasoning speed.

5.2 Future Plan

While our proposed method proves to solve some existing problems of the traditional target detection algorithm and improves the accuracy and processing time of the video-based target detection algorithm, it still has some limitations which can be improved. The following points illustrate my future research:

- To construct a sufficient and more comprehensive public dataset. As our deep learning method is data-driven, a more comprehensive dataset which contains more features from different scenarios could provide the network with more useful information for training. Besides, as our purpose is to extend the algorithm application in smart security, the existing public dataset may not meet some specific demands due to confidential reasons. Therefore, if we could construct more security-related datasets, our algorithm would be more suitable for the application.

- The basic network we used is one of the mainstream algorithms in computer vision, and some people have proposed other kinds of improved methods which generate better results. Therefore, I plan to add the proposed modules to other advanced methods and test whether they will have better performance.

- To make a series of finetuning operations based on different demands. For example, in this paper, we use ResNet-101 as the feature extraction backbone network. The deeper the network has stronger fitting ability and the larger the perceptual field, the larger the perceptual field is before a certain amount of higher order semantic information exists, as we can replace it with the ResNet-152 backbone network to improve the accuracy of target detection. However, the increase in network depth will bring about a larger computational effort and affect the inference speed of the model, which can be replaced by the MobileNet lightweight backbone network to reduce the model parameters and improve the target detection speed.

References

[1] Seung Hwan Bae. *Object Detection Based on Region Decomposition and Assembly.* AAAI Press, 2019.

[2] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. *Object Detection in Video with Spatiotemporal Sampling Networks*, volume 11216 of *Lecture Notes in Computer Science*. Springer, 2018.

[3] Alexey Bochkovskiy, ChienYao Wang, and HongYuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*, volume abs/2004.10934. 2020.

[4] Leo Breiman. *Bagging Predictors*, volume 24. 1996.

[5] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. *Prime Sample Attention in Object Detection.* Computer Vision Foundation / IEEE, 2020.

[6] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. *Memory Enhanced Global-Local Aggregation for Video Object Detection.* Computer Vision Foundation / IEEE, 2020.

[7] Jianpeng Cheng, Li Dong, and Mirella Lapata. *Long Short-Term Memory-Networks for Machine Reading.* The Association for Computational Linguistics, 2016.

[8] Corinna Cortes and Vladimir Vapnik. *Support-Vector Networks*, volume 20. 1995.

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. *Deformable Convolutional Networks.* IEEE Computer Society, 2017.

[10] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Association for Computational Linguistics, 2019.

[11] Navneet Dalal and Bill Triggs. *Histograms of Oriented Gradients for Human Detection.* IEEE Computer Society, 2005.

[12] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Martin Robertson, and Haibing Guan. *Object Guided External Memory Network for Video Object Detection*. IEEE, 2019.

[13] Jia Deng, Wei Dong, and Richard Socher. *ImageNet: A large-scale hierarchical image database*. IEEE Computer Society, 2009.

[14] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. *Relation Distillation Networks for Video Object Detection*. IEEE.

[15] Alexey Dosovitskiy, Philipp Fischer, and Eddy Ilg. *FlowNet: Learning Optical Flow with Convolutional Networks*. IEEE Computer Society, 2015.

[16] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. *Neural Architecture Search: A Survey*, volume 20. 2019.

[17] Scott E. Fahlman and Christian Lebiere. *The Cascade-Correlation Learning Architecture*. Morgan Kaufmann, 1989.

[18] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. *Object Detection with Discriminatively Trained Part-Based Models*, volume 32. 2010.

[19] Cheng Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. *DSSD*.

[20] Golnaz Ghiasi, Tsung Yi Lin, and Quoc V. Le. *NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection*. Computer Vision Foundation / IEEE, 2019.

[21] Ross B. Girshick. *Fast R-CNN*. IEEE Computer Society, 2015.

[22] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*. IEEE Computer Society, 2014.

[23] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. *Mining Inter-Video Proposal Relations for Video Object Detection*, volume 12366 of *Lecture Notes in Computer Science*. Springer, 2020.

[24] T. Hastie, S. Rosset, Z. Ji, and Z. Hui. *Multi-class AdaBoost*. Number 3. 2009.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, volume 37. 2015.

[26] Md. Amirul Islam, Mrigank Rochan, Neil D. B. Bruce, and Yang Wang. *Gated Feedback Refinement Network for Dense Image Labeling*. IEEE Computer Society, 2017.

[27] Jisoo Jeong, Hyojin Park, and Nojun Kwak. *Enhancement of SSD by concatenating feature maps for object detection*. BMVA Press, 2017.

[28] Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. *Hybrid Knowledge Routed Modules for Large-scale Object Detection*. 2018.

[29] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. *New Generation Deep Learning for Video Object Detection: A Survey*. 2021.

[30] Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, and Dong Huang. *Multiple Anchor Learning for Visual Object Detection*. Computer Vision Foundation / IEEE, 2020.

[31] Seung Wook Kim, Hyong Keun Kook, Jee Young Sun, Mun Cheon Kang, and Sung Jea Ko. *Parallel Feature Pyramid Network for Object Detection*, volume 11209 of *Lecture Notes in Computer Science*. Springer, 2018.

[32] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. *HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection*. IEEE Computer Society, 2016.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. *ImageNet classification with deep convolutional neural networks*, volume 60. Commun. ACM, 2017.

[34] Hei Law and Jia Deng. *CornerNet: Detecting Objects as Paired Keypoints*, volume 11218 of *Lecture Notes in Computer Science*. Springer, 2018.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. *Deep learning*, volume 521. 2015.

[36] Zuoxin Li and Fuqiang Zhou. *FSSD: Feature Fusion Single Shot Multibox Detector*, volume abs/1712.00960. 2017.

[37] R. Lienhart and J. Maydt. *An extended set of Haar-like features for rapid object detection*, volume 1. 2002.

[38] Tsung Yi Lin, Piotr Doll, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. *Feature Pyramid Networks for Object Detection*. IEEE Computer Society, 2017.

[39] TsunYi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Doll. *Focal Loss for Dense Object Detection*. IEEE Computer Society, 2017.

[40] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. *Path Aggregation Network for Instance Segmentation*. Computer Vision Foundation / IEEE Computer Society, 2018.

[41] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, volume 9905 of *Lecture Notes in Computer Science*. Springer, 2016.

[42] Thang Luong, Hieu Pham, and Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*. The Association for Computational Linguistics, 2015.

[43] Diganta Misra. *Mish: A Self Regularized Non-Monotonic Activation Function*.

[44] Hyeonseob Nam, Mooyeol Baek, and Bohyung Han. *Modeling and Propagating CNNs in a Tree Structure for Visual Tracking*, volume abs/1608.07242. 2016.

[45] Alexander Neubeck and Luc Van Gool. *Efficient Non-Maximum Suppression*. IEEE Computer Society, 2006.

[46] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. *Imbalance Problems in Object Detection: A Review*, volume 43. 2021.

[47] journal= Otsu, Nobuyuki.

[48] W. Ouyang, X. Wang, Z. Cong, and X. Yang. *Factors in Finetuning Deep Model for Object Detection with Long-Tail Distribution*. 2016.

[49] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. *Libra R-CNN: Towards Balanced Learning for Object Detection*. Computer Vision Foundation / IEEE, 2019.

[50] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. *You Only Look Once: Unified, Real-Time Object Detection*. IEEE Computer Society, 2016.

[51] Joseph Redmon and Ali Farhadi. *YOLO9000: Better, Faster, Stronger*. IEEE Computer Society, 2017.

[52] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*, volume abs/1804.02767. 2018.

[53] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2015.

[54] Olga Russakovsky, Jia Deng, and Hao Su. *ImageNet Large Scale Visual Recognition Challenge*, volume 115. 2015.

[55] Bharat Singh and Larry S. Davis. *An Analysis of Scale Invariance in Object Detection SNIP*. Computer Vision Foundation / IEEE Computer Society, 2018.

[56] Bharat Singh, Mahyar Najibi, and Larry S. Davis. *SNIPER: Efficient Multi-Scale Training*. 2018.

[57] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. *FCOS: Fully Convolutional One-Stage Object Detection*. IEEE, 2019.

[58] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. *FCOS: A Simple and Strong Anchor-Free Object Detector*, volume 44. 2022.

[59] Subarna Tripathi, Zachary C. Lipton andSerge J. Belongie, and Truong Q. Nguyen. *Context Matters: Refining Object Detection in Video with Recurrent Neural Networks*. BMVA Press, 2016.

[60] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. *Selective Search for Object Recognition*, volume 104. 2013.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention is All you Need*. 2017.

[62] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. *Graph cut based image segmentation with connectivity priors*. IEEE Computer Society, 2008.

[63] Paul A. Viola and Michael J. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features*. IEEE Computer Society, 2001.

[64] Paul A. Viola and Michael J. Jones. *Robust Real-Time Face Detection*, volume 57. 2004.

[65] Chien Yao Wang, Hong Yuan Mark Liao, Yueh Hua Wu, Ping Yang Chen, Jun Wei Hsieh, and I Hau Yeh. *CSPNet: A New Backbone that can Enhance Learning Capability of.*

[66] Hao Wang, Qilong Wang, Mingqi Gao, Peihua Li, and Wangmeng Zuo. *Multi-Scale Location-Aware Kernel Representation for Object Detection.* Computer Vision Foundation / IEEE Computer Society, 2018.

[67] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. *Region Proposal by Guided Anchoring.* Computer Vision Foundation / IEEE, 2019.

[68] Robert J. Wang, Xiang Li, Shuang Ao, and Charles X. Ling. *Pelee: A Real-Time Object Detection System on Mobile Devices.* OpenReview.net, 2018.

[69] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. *Fully Motion-Aware Network for Video Object Detection*, volume 11217 of *Lecture Notes in Computer Science.* Springer, 2018.

[70] Sanghyun Woo, Soonmin Hwang, and In So Kweon. *StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection.* IEEE Computer Society, 2018.

[71] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhao Xiang Zhang. *Sequence Level Semantics Aggregation for Video Object Detection.* IEEE, 2019.

[72] Fanyi Xiao and Yong Jae Lee. *Video Object Detection with an Aligned Spatial-Temporal Memory*, volume 11212 of *Lecture Notes in Computer Science.* Springer, 2018.

[73] H. Xu, L. Yao, Z. Li, X. Liang, and W. Zhang. *Auto-FPN: Automatic Network Architecture Adaptation for Object Detection Beyond Classification.* 2020.

[74] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John W. Paisley. *PanNet: A Deep Network Architecture for Pan-Sharpening.* IEEE Computer Society, 2017.

[75] Sergey Zagoruyko and Nikos Komodakis. *Wide Residual Networks.* BMVA Press, 2016.

[76] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. *Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection.* Computer Vision Foundation / IEEE, 2020.

[77] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. *Single-Shot Refinement Neural Network for Object Detection.* Computer Vision Foundation / IEEE Computer Society, 2018.

[78] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. *S3FD: Single Shot Scale-invariant Face Detector*, volume abs/1708.05237. 2017.

[79] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. *Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-Based Person Re-Identification.* Computer Vision Foundation / IEEE, 2020.

[80] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. *Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression.* AAAI Press, 2020.

[81] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. *Scale-Transferrable Object Detection.* Computer Vision Foundation / IEEE Computer Society, 2018.

[82] Xingyi Zhou and Dequan Wang. *Objects as Points*, volume abs/1904.07850. 2019.

[83] Chenchen Zhu, Yihui He, and Marios Savvides. *Feature Selective Anchor-Free Module for Single-Shot Object Detection.* Computer Vision Foundation / IEEE, 2019.

[84] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. *Flow-Guided Feature Aggregation for Video Object Detection.* IEEE Computer Society, 2017.

[85] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. *Deep Feature Flow for Video Recognition.* IEEE Computer Society, 2017.