# Deep Learning Method for Building Image Localization in Smart City

**Yuanpeng Long**
Southwestern University of Finance and Economics
**guo zhang** ( guo758881@foxmail.com )
Chongqing University of Posts and Telecommunications
**Yu Pang**
Chongqing University of Posts and Telecommunications
**Huiqian Wang**
Chongqing University of Posts and Telecommunications

**Research Article**

**Keywords:** Smart Cities, Building Detection, BFPN-RCNN, Geographic Position ,

**Posted Date:** June 29th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-557782/v1

# Deep Learning Method for Building Image Localization in Smart City

**Yuanpeng Long[1], Guo Zhang [2,3*], Yu Pang[4], , Huiqian Wang[4]**

[1] School of Economic Information Engineering, Southwestern University Of Finance And Economics, Chengdu 611130, China；longpeng302@163.com

[2] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; guo758881@foxmail.com

[3] School of Medical Information and Engineering, Southwest Medical University, Luzhou, 646000, China

[4] Chongqing Key Laboratory of Photoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; pangyu@cqupt.edu.cn (Y.P.); wanghq@cqupt.edu.cn (H.W.)

* Correspondence: guo758881@foxmail.com (G.Z.)

**Abstract:** In recent years, with the construction and development of smart cities, the rapid popularization of the Internet of Things and the sharp increase of Internet users, a large amount of multimedia data with geo-location information shared by users has emerged.However, only a small part of the image data is used correctly.Therefore, building detection can not only realize geographic positioning, but also has guiding significance for GIS mapping and automatic updating.With the extensive application of convolutional neural network and cyclic neural network in image processing, this paper proposes that the BFPN-RCNN algorithm can be used to recognize the curved buildings in the image.By comparing with other image detection algorithms on different data sets, it is proved that the proposed algorithm can effectively locate curved images in natural scene images.

**Keywords:** Smart Cities; Building Detection; BFPN-RCNN; Geographic Position ;

## 1.Introduction

With the continuous development of information technology and application of a new generation of Internet, cloud computing, smart sensing, communication, remote sensing, satellite positioning, geographic information system technology such as the combination of the all things will be able to realize intelligent identification, location, tracking and monitoring and management [1], so that the city reach the state of the "wisdom" in the construction of wise city technically possible [2].Smart city [3] is the application of new generation information technologies such as the Internet of Things, cloud computing and geospatial infrastructure, as well as tools and methods such as social network, comprehensive integration method, network-dynamic and all-media integrated communication terminal [4] [5]. It is the inevitable trend of current urban construction to change the urban development mode and improve the quality of urban development. By building urban wisdom, deliver timely integration of city economy and culture, and public resources, management, service, public life, ecological environment [6] and other kinds of

information, improve the content and content, objects and people, people of connectivity, comprehensive perception and use of information ability, can greatly improve the government management service ability and the level of people's material and cultural life [7].

Image positioning is one of the important components of smart city [8]. With the development of electronic devices, many electronic hand-held devices (such as digital cameras, smart phones, drones, aerial cameras, etc.) are integrated with GPS function, and these products can obtain the geographic location information of the photographing ground while taking pictures [9]. At the same time,Social software and websites such as QQ, WeChat, Douyin, Baidu and Google Earth provide tools specifically used to mark their geographical locations [10]. This has led to a rapid increase in the number of images with geo-location information on the Internet. At present, image location positioning mainly relies on a large number of ground perspective images with GPS information as a reference to determine the location information of queried images [11]. The reference images of ground perspective [12] are mainly concentrated in important cities, popular tourist attractions and other human gathering places. Buildings [13] are the main component of the ground perspective and also the main component of topographic map mapping. Recognition and extraction of buildings are of great significance for feature extraction and feature matching as the reference body of other targets [14]. It is generally believed that the urban landmark landscape [15] should refer to a specific area in the city, which is used to concentrate, condense, reflect and reflect the overall characteristics of the city. Landmark buildings in cities have spatial identification, which is used to calibrate distance, height and azimuth, and to determine the spatial relationship between location and landmark buildings. Therefore, image feature positioning of urban buildings has an important influence and significance on the development of smart cities [16].


**Figure 1.** landmark building in the city

In smart city, landmark buildings and landscape occupy a very important position. However, the appearance of more and more city landmarks also brings some trouble to the recognition. Image recognition technology is an effective way to solve these problems. With the advent of the era of big data and the substantial improvement of computer

computing power, image recognition technology based on deep learning can not only identify the content of images, but also Realizing Geographic Positioning in images. The most important network structure in the deep learning algorithm is the CNN (Convolutional Neural Network) [3] structure, which has the advantage of enabling the computer to automatically extract feature information. However, the convolutional neural network can automatically extract features in images after training [4]. New York university [5] proposed convolutional neural network structure for the first time in 1998, which is a milestone in the history of deep learning, and Lenet-5 network laid the foundation for the following structure of deep learning convolutional neural network. In 2006, Hinton[6] put forward the concept of deep learning. The emergence of big data improves the size of the data set and alleviates the problem of over-fitting by training. The rapid development of computer hardware makes the performance of computer greatly improved and the training speed of neural network is accelerated [7]. With the great improvement of computer performance and the rapid development of the algorithm, deep learning has achieved excellent results in image recognition. Various kinds of convolutional neural networks emerge one after another, AlexNet[8], VGG[9], InceptionNet[10], ResNet[11] and DenseNet[12] proving that the change of network structure can affect the final performance of the network to a certain extent. Meanwhile, the deep learning model with better and better performance has been widely applied in image recognition.

Deep learning has been studied very quickly by scholars, who have realized the importance and influence of this field. Zeiler M et al. [13] introduced a novel visualization technology, which can deeply understand the functions of the intermediate feature layer and the operation of the classifier. This technology is particularly sensitive to the local information in the image. Ma C et al. [14] used features extracted from the deep convolutional neural network trained on the object recognition data set to improve tracking accuracy and robustness. Cinbis R et al. [15] proposed a window refinement method to prevent the training from locking the wrong object position too early and improve the positioning accuracy. Dai J et al. [16] proposed a position-sensitive score graph to solve the dilemma between translation invariance in image classification and translation variance in object detection. Bell S et al. [17] used spatial repetition to integrate contextual information outside the region of interest. Zhang S et al. [18] designed a transmission link block to predict the location, size and category labels of objects in the object detection module by features in the transmission anchor frame module. Wang X et al. [19] proposed an alternative solution by learning a kind of adversarial network to generate examples of occlusion and deformation, while the adversarial target generates the examples which the object detector difficult to classify.

In 1990s, Irvin [20] and Liow [21] put forward a new idea of building extraction with shadow. And some scholars put forward methods based on artificial intelligence in recent years. The image was segmented and the regional features were extracted by the method combining multi-scale segmentation and Canny edge detection, and buildings were extracted by combining Bayesian network and other imaging conditions in paper [22]. In paper [23], the image edges were firstly extracted and the spatial relation diagram was

constructed, then Markov model was introduced to construct Markov random field, and finally the buildings were extracted by calculating the minimum energy function to set the threshold.

In this paper, the method of deep learning [24] is used to carry out feature matching between the captured images of scenic spots and landmark buildings and the database images, and to automatically obtain the real-time geographic position of the images [25], so as to realize image geographic location positioning.

## 2. Faster R-CNN

Faster R-CNN[26] (Region-based Convolutional Neural Network) is a relatively classical deep learning algorithm, which has high recognition accuracy, efficiency, and good recognition rate for the large target area. Faster R-CNN algorithm mainly consists of two modules: Fast R-CNN detection module and RPN [27] (Region Proposal Network) extraction module. The RPN is used to generate high-quality region from the basic feature map, and the Fast R-CNN directly detects and identifies the targets in the extracted suggestions region.

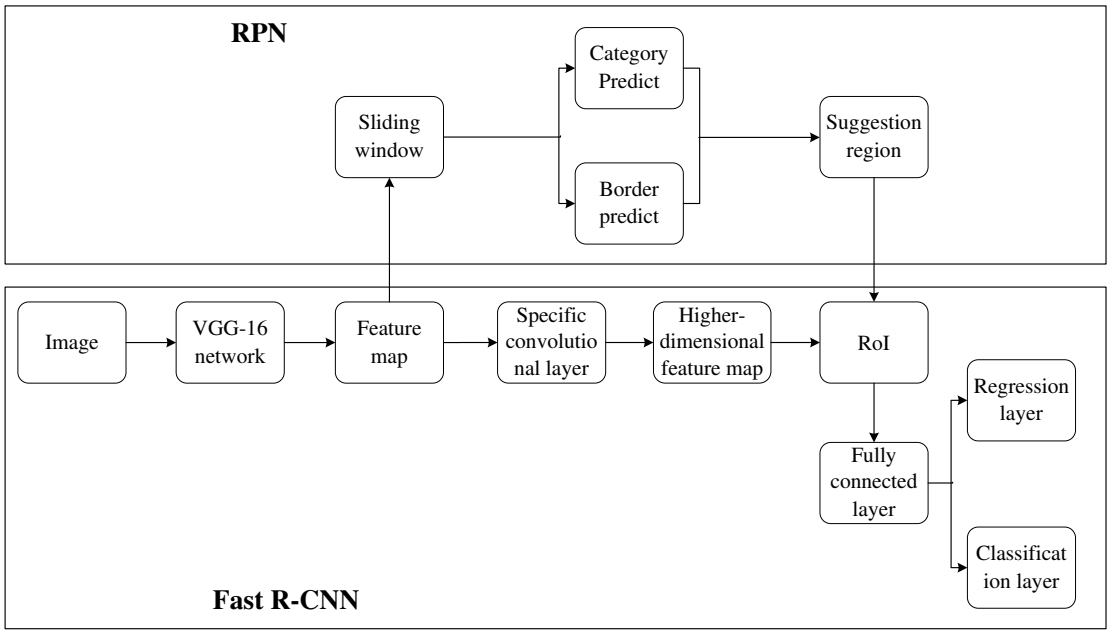


**Figure 2.** Faster R-CNN

As shown in figure 2, the processing diagram of Faster R-CNN. Firstly, the images of any size input to the VGG-16 (visual geometry group-16) network. Secondly, the CNN network generate the shared convolutional layer and get the feature map, on the one hand, the feature map input to the RPN network; on the other hand, it propagates further to the specific convolutional layer, and generates the higher-dimensional feature map. Thirdly, the higher-dimensional feature map and the suggestion region is input to the RoI (Region of Interest) pooling, and extract the features of the suggestion region. Then the features are entered into the following regression layer and classification layer. NMS[28] (Non-Maximum Suppression) algorithm was used to remove similar results from the predicted

target. Finally, the algorithm output the object category of target and the coordinates of the region.

Faster R-CNN algorithm has achieved excellent results in the field of target detection and recognition, and the performance of deep learning has been greatly improved. But Faster R-CNN algorithm is still lacking in some respects. Along with the network, the edge texture information of lower layer is filtered out slowly in the convolution process, and the feature maps of the network in the extraction are not particularly accurate. However, the edge texture information of the building is particularly important in the recognition for the building distinguish from other categories of buildings. At the same time, the candidate boxes are quantized twice in the RoI pooling. There is a problem of mismatch between the actual candidate boxes and the obtained candidate boxes.

## 3.Image Recognition Algorithm Based On BFPN-RCNN

In natural scene images, there are many types of curved shape and irregular shape besides oblique image.Existing image detection methods based on quadrilateral bounding box are difficult to accurately detect images with irregular shapes, and it is difficult to completely enclose the image in quadrilateral, which will reduce the accuracy of image detection and affect the final recognition effect.On the other hand, most pixel-based segmentation detectors cannot separate features that are very close to each other when the building images are irregular in shape and the distance between the images is relatively close.To solve these problems in natural scene pictures, this paper proposes a image detection algorithm based on BFPN-RCNN. The algorithm model is shown in Figure 3.It is a kind of detector based on segmentation. First, the bottom-up path of FPN is expanded to enhance the transmission of shallow layer feature information, and the adaptive feature pooling is adopted to extract features from all levels, and then the fusion is made for prediction and multiple predictions are made for each image instance.These predictions correspond to different "cores" produced by scaling down the original image instance to different scales.The final detection result can be obtained by the progressive scaling algorithm, which can gradually expand the smallest size of the kernel into a full shape image instance.Due to the large geometric edges between these minimal cores, the proposed algorithm can effectively distinguish the adjacent image instances and is robust to arbitrary shapes.

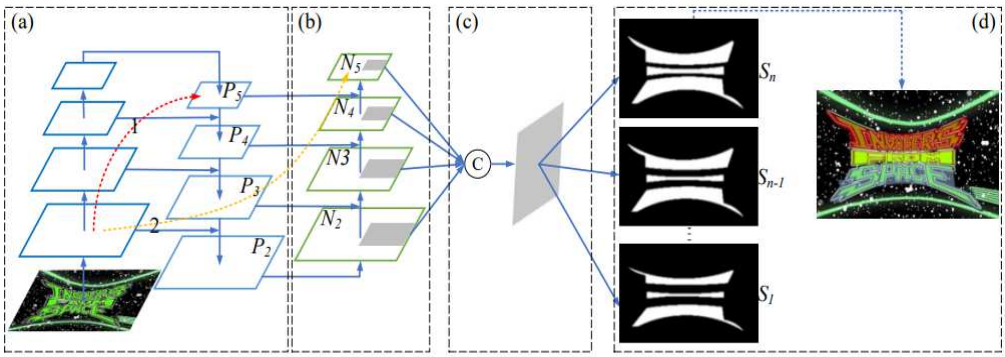


**Figure 3**.Image detection model based on BFPN-RCNN

### 3.1 Feature extraction stage

FPN mainly increases the top-down path with horizontal connection to enhance the transmission of semantic information, and integrates the high and low level features to improve the effect of target detection, especially to improve the detection effect of small-size targets.High-level neurons correspond to the overall target response, while lower-level neurons are more likely to be activated by local images.This indicates the need to enhance top-down propagation to obtain strong semantic information.FPN has proved that adding a bypass connection from top to bottom can increase the semantic nature of features and facilitate classification. However, FPN mainly enhances the transmission of high-level semantic information and has no effect on the transmission of target positioning information, while the most important aspect of image detection is the positioning of the target position of building images.Figure 4 shows the FPN network structure diagram.For this purpose, this article adds a bottom-up pyramid network behind the FPN.

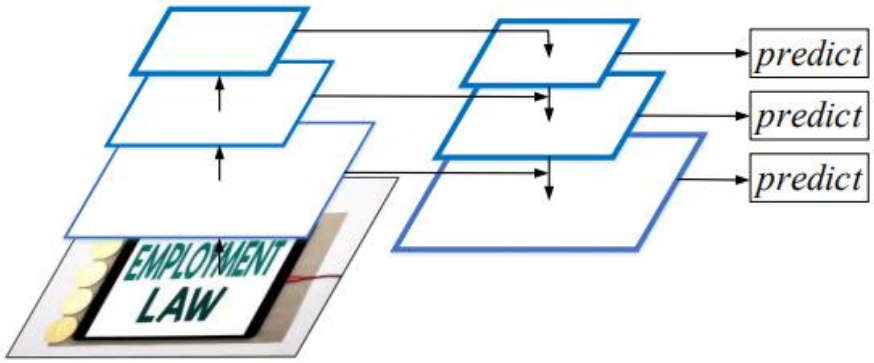


**Figure 4.** FPN

The introduction of bottom-up path enhancement mainly considers that the shallow layer feature information of network is very important for instance segmentation. The shallow layer feature is mostly the edge shape and other features, and the instance segmentation is pixel level classification.The red dotted arrow 1 in Fig. 3 represents the bottom-up process in the FPN algorithm, in which the features of the shallow layer are transferred to the top layer.The four blue rectangular blocks in Figure 3 are the output of RES2, RES3, RES4 and RES5 layers of RESNET from bottom to top, and the number of layers ranges from dozens to more than 100. Obviously, the loss of characteristic information at the shallow layer will be more serious after transmission at so many layers.The yellow dotted arrow 2 represents a bottom-up expansion path, which itself is less than 10 levels.When the shallow layer features are connected from the side of the original FPN to $p_2$, and then transferred from $p_2$ to the top layer along the expansion path, the number of layers passed through is less than 10, so the shallow layer feature information can be retained well.The detailed design for bottom-up path enhancement is shown in Figure 5.The characteristic layers obtained by fusion are $N_2$, $N_3$, $N_4$and $N_5$, of which $N_2$ is the same as $p_2$. These characteristic layers will be used in the subsequent classification and regression of prediction frame.

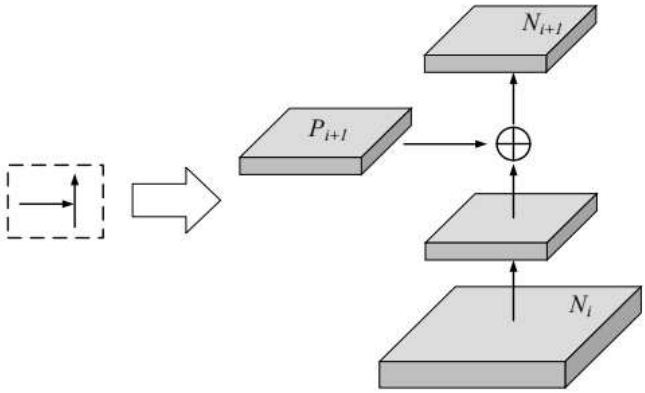


**Figure 5.** Image recognition algorithm structure diagram

In the fusion part, a new feature figure $N_{i+1}$ is generated by adopting a higher resolution feature figure $N_i$ and a coarser feature figure $P_{i+1}$ through horizontal connection.Each feature figure $N_i$ first reduces the size of the feature figure by half through a convolution layer of size 3×3 and step size 2.Then add each element of Feature Figure $P_{i+1}$ and the subsample graph by horizontal connection.The fused feature map is processed through another convolution layer 3×3 to generate feature layer $N_{i+1}$ for the subsequent network.This is an iterative process that terminates after $P_5$.In these building blocks, the number of channels is always 256. All convolutional layers are followed by a ReLU activation function.Then, the features of each candidate region are pooled from the newly acquired feature graphs N2-N5.The advantage of this new branch is to shorten the distance between the features with large size at the bottom and the features with small size at the top, which makes feature fusion more effective.

### 3.2 Dynamic feature pooling

In FPN, candidate regions are assigned to different feature levels according to their size.In this way, small candidate regions are assigned to low-level features, while large candidate regions are assigned to high-level features, which is simple but effective, but may also produce non-optimal results.For example, two candidate regions with a difference of 10 pixels may be assigned to different feature levels, but in fact these two candidate regions are very similar, and the importance of features may not have much to do with the feature levels they belong to. High-level features have a large receptive field and capture rich context information.Allowing small candidate regions to capture these features makes better use of contextual information for prediction.The low level features many small details and high positioning accuracy.Allowing large candidate regions to obtain these features is beneficial to improving the effect of text detection.Therefore, both high-level and low-level features have an impact on the text detection effect.For each candidate region, features from all levels are pooled and then fused to make predictions, a process known as adaptive feature pooling.Therefore, the main task of adaptive feature pooling is feature fusion.In the target detection or segmentation algorithm of the Faster RCNN series, Region of Interest alignment (ROI) is extracted from RPN network to extract ROI features. In this step, the features corresponding to each ROI are single-layer features, and the same

is true for FPN.For example, the output of res5 is commonly used in ResNet network.Adaptive feature pooling is to convert single-layer features into multi-layer features, that is, each ROI needs to perform ROI Align operations with multi-layer features, and then the resulting ROI features at different levels are fused together, so that each ROI feature is integrated with multi-layer features.

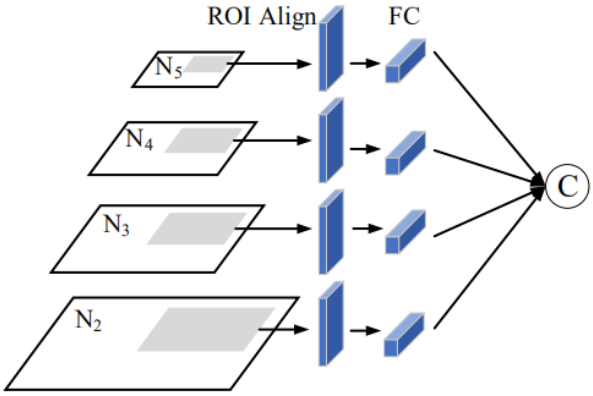


**Figure 6.** Adaptive feature pooling

### 3.3 Progressive expansion algorithm

Based on (BFS) algorithm, the algorithm firstly distributes each image to multiple predicted segmentation areas.These segmented regions are represented as "cores" in this paper.For each image, there are several corresponding cores.Each kernel shares a similar shape to the original entire image instance, all at the same central point, but with a different scale.Because the minimum-scale core boundaries are far away from each other, they can be easily separated.As shown in Figure 7, the kernel size is only 1/2 of the complete area, but the smallest kernel size cannot cover the complete area of the image instance.Then gradually add more pixels to the enlarged core to expand their region, until the largest core, the complete image region is found.

（a）Ground Truth　　（b）Kernal Scal=0.5

**Figure 7.** Image instance

There are four main reasons for selecting the progressive scaling algorithm in this

paper.First, nuclei with extremely small scales are easily separated because their boundaries are far from each other.Therefore, it overcomes the main shortcomings of the previous method based on segmentation.Second, the largest core or complete region of the image instance is essential for achieving the final accurate detection;Third, cores grow from small to large, so smooth monitoring will make the network easier to learn.Finally, the asymptotic scaling algorithm can ensure the exact location of the image instance because its boundaries are expanded in a careful and asymptotic manner.As shown in Figure 8, this is the specific process of the progressive scaling algorithm.

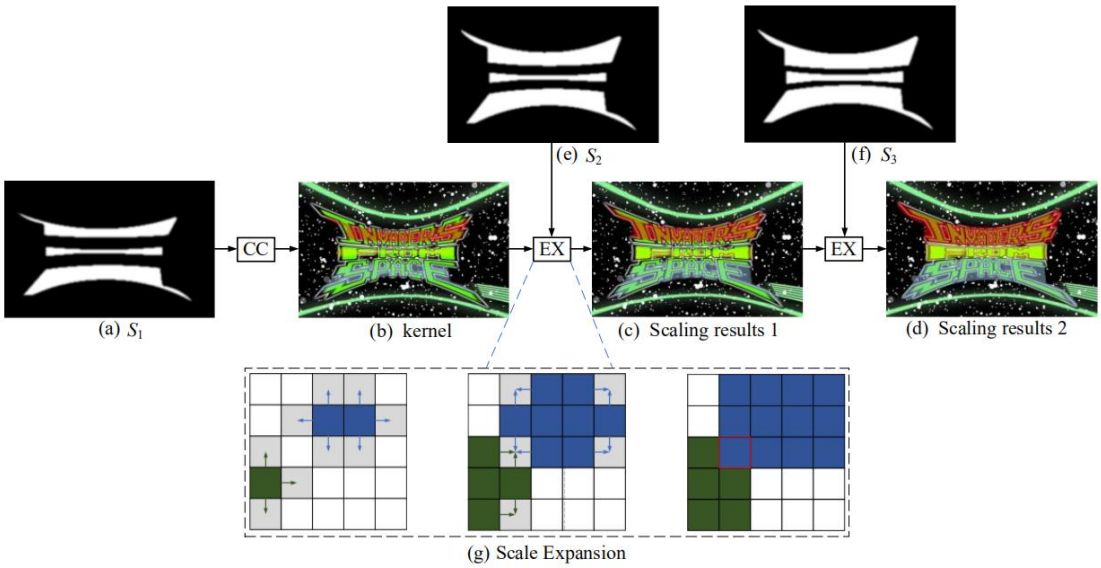


**Figure 8.** Asymptotic scaling algorithm

In the example of the progressive scaling algorithm in Figure 8, the segmentation results of three image regions are $S_1$ ,$S_2$and $S_3$.First, based on $S_1$of the minimum kernel feature graph, three different connected components $c = \{c_1, c_2, c_3\}$ can be initialized.In Figure 8(b), different colors are used to represent different connected domains. the central part of all image instances, the minimum kernel, has been detected.Then by merging the pixels in$S_2$, the detected kernel is gradually expanded in $S_3$.These two scaling results are shown in Figure 8(c) and (d) respectively.Finally, the connected domains marked with different colors are extracted from Figure 8(d) as the final prediction region of the image instance.

The specific process of progressive scaling is shown in Figure 8 (g).Since the expansion process is based on a breadth-first search algorithm, the algorithm iterates and merges adjacent image pixels starting from pixels with multiple cores.However, there may be conflicting pixels during the expansion, as shown in the red box in Figure 8(g).The principle of handling conflicts in image detection practice is that confused pixels can only be merged by a single kernel.Because of the "progressive" extender, these boundary conflicts do not affect final detection and performance.

Each image, image detection process instance assigned to multiple prediction region segmentation is $S_1$ ……$S_n$. These segmented regions are then represented as "cores" in the image, and there are several corresponding cores for an image instance.Each kernel shares a similar shape to the original entire image instance, and they are all at the same central point but differ in scale. The generation process of Ground Truth corresponding to these "cores" is shown in Figure 9.

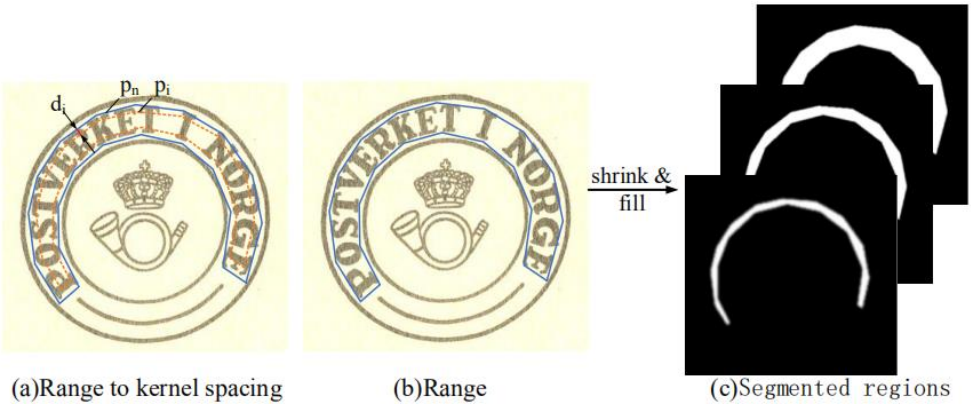


**Figure 9 .**Ground Truth Produce process

In Figure 9 (a), $p_i$ is the ith core, $p_n$ is the nth core, and $d_i$ is the distance between the edge of $p_i$ and$p_n$;(b) is the original image area;(c) is a plurality of segmented regions generated.In order to obtain the reduced mask in sequence such as in Figure 9, this paper uses the Vatti clipping algorithm to reduce the $d_i$ pixels of the polygon $p_n$ and obtain the reduced polygon $p_i$.Subsequently, each reduced polygon $p_i$ is converted into a 0/1 binary mask, which is used to segment the label Ground Truth.These Ground truths are respectively expressed as $G_1$ ……$G_n$.If the ratio between $p_n$ and $p_i$ is considered to be $r_i$, then the range $d_i$ between $p_n$ and $p_i$ can be expressed by the formula:

$$d_i = \frac{\text{Area}\left(P_i\right)\times\left(1-r_i^2\right)}{\text{Perimeter}\ \left(P_n\right)} \tag{1}$$

Area represents the Area of a polygon, and Perimeter is the Perimeter of the polygon.The calculation formula of $r_i$ is:

$$r_i = 1-\frac{(1-m)\times(n-i)}{n-1} \tag{2}$$

Where $m$ is the reduced scale, and its value range is (0,1);$n$ is the number of image segmentation instances, that is, the number of "cores".Under normal conditions, the image area is much larger than the detection area, so the cross entropy loss of dichotomy will make the result more biased to the detection area. Therefore, the formula of DICE coefficient is used in this model as follows:

$$D\left(S_i,G_i\right) = \frac{2\sum_{x,y}\left(S_{i,x,y}*G_{i,x,y}\right)}{\sum_{x,y}S_{i,x,y}^2+\sum_{x,y}G_{i,x,y}^2} \tag{3}$$

Where, $S_{x,y}$ is the value of the pixel point (x,y) in the predicted instance, and $\mathrm{G}_{x,y}$ is the value of the pixel point (x,y) in the label.The model also defines a new loss function with the formula as follows:

$$L = \lambda L_c + (1-\lambda) L_s \tag{4}$$

Where $L_c$ is the image region classification loss, $L_s$ is the image shrinkage loss, and its formula is as follows:

$$L_c = 1 - D\left(S_n \cdot M, G_n \cdot M\right) \tag{5}$$

$$L_s = 1 - \frac{\sum_{1}^{n-1} D\left(S_i \cdot W, G_i \cdot W\right)}{n-1}, W_{x,y} = \begin{cases} 1, & S_{n,x,y} \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The value of M is generated by online hard example miniing （OHEM）algorithm. OHEM and Focal Loss have similar functions, but they are different.When Focal loss is applied to the one-stage image detection model, positive and negative samples cannot be combined freely, so it can only suppress negative samples and simple samples by calculating the loss value, and mining difficult samples.The OHEM algorithm is applied to the two-stage image detection model. The positive and negative samples are controllable in the operation process, and the operation process of OHEM algorithm is the process of hard case mining.The core idea is to filter out input samples that have a greater impact on the image detection process according to the loss, and then apply these samples to carry out model training through the Stochastic Gradient Descent (SGD) algorithm.Specifically in the image detection model, all positive samples and difficult samples were selected, and simple negative samples were filtered out.The selected pixel has a value of 1, and the unselected pixel has a value of 0.

**4. Experiment and result analysis**

In order to verify the image detection effect of the BFPN-RCNN image detection algorithm proposed in this paper in a complex environment, training and testing were carried out on different data sets, and the detection effect of different algorithms on curved images in natural scenes was compared.This paper uses the RESNET network as the backbone of the text detection model.All the networks are optimized using stochastic gradient descent.1000 SCUT-CTW1500 images were used to train the model, and the image detection results on SCUT-CTW1500 were obtained.The data enhancement of the training image is as follows :(1) The random scaling ratio of the image is;(2) The image rotates randomly horizontally within the range of [-10,10];(3) Cut the image of the same size randomly from the converted image.Calculate the minimum rectangular region and extract the text bounding box.For the curvilinear image dataset, the final result is generated by extending the network at an asymptotic scale.Detailed training parameter Settings are shown in Table 1.

**Table 1.** Training parameter setting

| Type | setting |
|---|---|
| Batch size | 16 |
| learning rate | $10^{-3}$ |
| Focal Loss | $\alpha = 0.25, \gamma = 3$ |
| OHEM | 3 |

| | |
|---|---|
| Nesterov momentum | 0.99 |
| Learning decay rate | 0.9/10000 |
| iterations | 100000 |

The effect of image detection.As shown in Figure 10, the blue curve and the green curve are the experimental results of the model on the SCUT-CTW1500J and ICDAR2017 datasets respectively.When m is too large or too small, the value of f-measures on the test set decreases.When m is too large, it is difficult for the detection model to segment the image instances that are close to each other.When m value is small, the detection model often divides the whole image area into different parts wrongly, and the training cannot converge well.In addition, when the kernel scale is set to 1, only the image segmentation graph is applied as the final result, and no progressive scaling algorithm is used.The performance of the image detection network is not ideal at this time, because the network cannot distinguish the images that are close together.

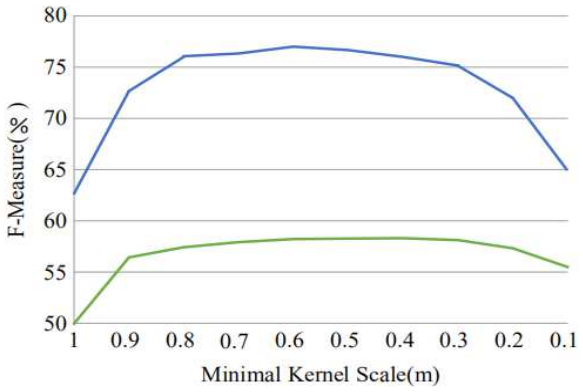


**Figure 10.** The effect of minimum kernel scale on image recognition

This paper also verifies the effect of kernel number n on the performance of the text detection model.The minimum kernel scale m is kept unchanged, and different kernel number n is used to train the model.Set $m$ = 0.4 on the ICDAR2017 dataset and $m$ = 0.6 on the CTW1500 to increase n from 2 to 10.As shown in Fig. 11, the blue curve and the green curve are the experimental results of the model on the SCUT-CTW1500J and ICDAR2017 datasets respectively.Thus, it can be found that with the increase of $n$, F-measure on the test set also keeps rising and starts to stabilize when $n$≥5.The advantage of multi-kernel is that it allows you to reconstruct two closely spaced image instances with a large gap.

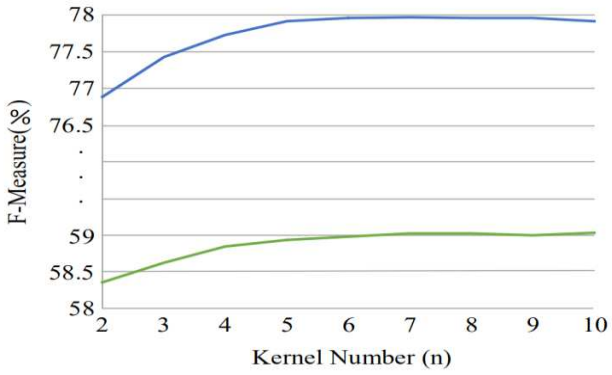

**Figure 11.** The influence of kernel number on image recognition effect

By comparing with other existing image detection algorithms on SCUT-CTW1500 and ICDAR2017 data sets, the proposed detection model for curved images is proved to have good results, and the proposed algorithm is proved to have good applicability through experiments on two different types of data sets.The experimental comparison results are shown in Table 2 and Table 3.

Table.2 Comparison with other methods in the CTW1500 dataset

| Algorithm | Year | Recall | Precision | F-score |
|---|---|---|---|---|
| SegLink[54] | 2017 | 0.40 | 0.42 | 0.41 |
| EAST[35] | 2017 | 0.49 | 0.79 | 0.60 |
| CTD+TLOC[55] | 2017 | 0.70 | 0.77 | 0.73 |
| SLPR[56] | 2018 | 0.70 | 0.80 | 0.75 |
| TextSnake[57] | 2018 | 0.85 | 0.68 | 0.76 |
| Ours | — | 0.75 | 0.81 | 0.78 |

Table.3 Comparison with other methods in the ICDAR 2017 dataset

| Algorithm | Year | Recall | Precision | F-score |
|---|---|---|---|---|
| Linkage-ER-Flow[58] | 2017 | 0.26 | 0.44 | 0.32 |
| TH-DL[58] | 2017 | 0.35 | 0.68 | 0.46 |
| TDN-SJTU[58] | 2017 | 0.47 | 0.64 | 0.54 |
| SARI FDU RRPN[58] | 2017 | 0.55 | 0.71 | 0.62 |
| SCUT DLVClab1[58] | 2018 | 0.54 | 0.80 | 0.65 |
| Lyu et al.[59] | 2018 | 0.55 | 0.84 | 0.67 |
| FOTS[60] | 2018 | 0.58 | 0.81 | 0.67 |
| Ours | — | 0.59 | 0.83 | 0.69 |

Deep neural networks have been shown to improve the performance of large-scale image classification and target detection.In order to better analyze the image detection performance of the proposed BFPN-RCNN algorithm, three RESNET networks with depths of 50, 101 and 152 were used as the backbone network of the image detection algorithm, and the tests were carried out on the large-scale data set SCUT-CTW1500. Under the same external conditions, increasing the depth of the backbone network from 50 to 152 can significantly improve the performance from 76.8% to 78.0%, and the comprehensive index F can be improved by 1.2%.Part of the test images are shown in the follow figures 9:

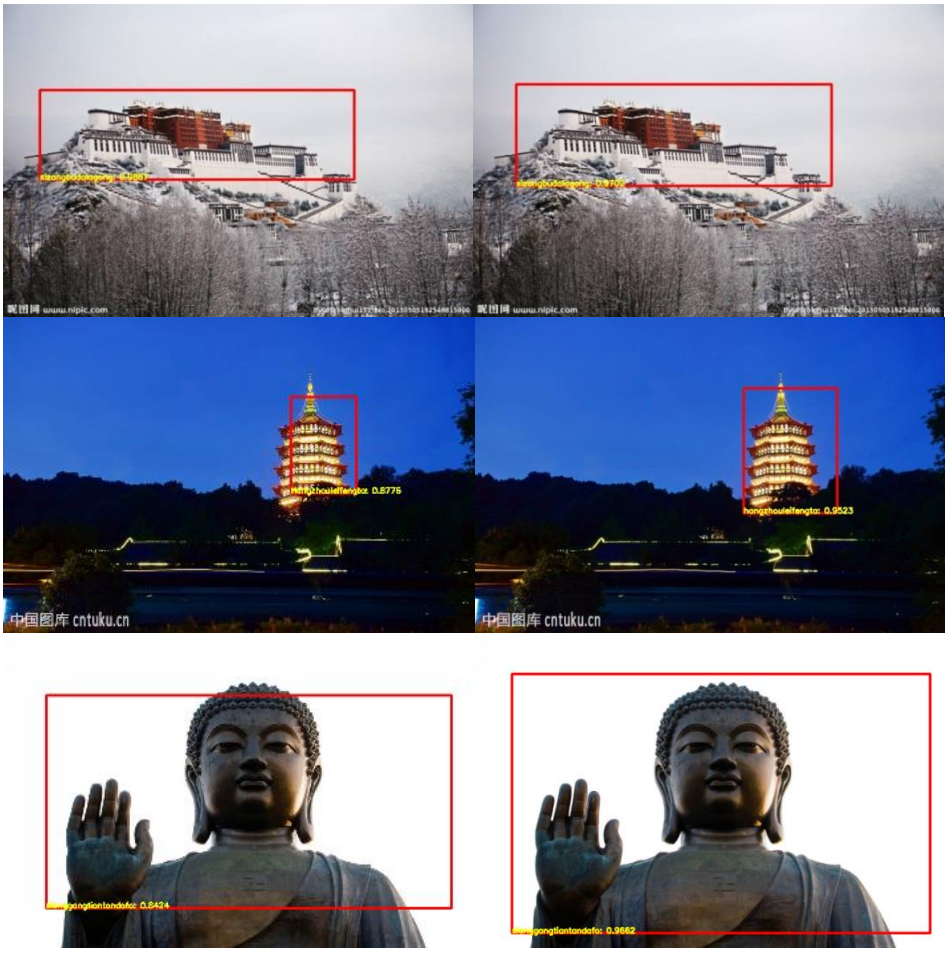

**Figure 12.** Experiment results of the test images

The left column is the test results for the model in this article, and the right column is the test results for Resnet101.Obviously, the model in this paper can be used to predict the area of buildings in the image more accurately and in a wider range, and the target buildings are basically within the predicted area.At the same time, it also has a good identification effect for buildings in complex environment (local buildings, buildings in rain and fog weather, and buildings at night), as shown in Fig. 10.

Figure 13. Test images in the complex environment

In this paper, the curved images of natural scenes are tested and compared with the image detection algorithm based on NEAST under the same conditions.The natural scene images selected by experiment have the characteristics of high background complexity and near curved distribution of image feature interval.


**Figure 14.** Image recognition effect of NEAST


**Figure 15.** Image recognition effect of BFPN-RCNN

As shown in Figure 14 and Figure 15, the image detection effects of the image detection model proposed in this paper and the image detection algorithm based on Neast on the SCUT-CTW1500 data set are compared.As can be seen from the figure, the image detection model based on BFPN-RCNN proposed in this paper can effectively detect the features of curved images, and the detection effect of irregular shape images is also more ideal.Therefore, the image detection model proposed in this paper can improve the accuracy of the final detection results through targeted training.

## 5. Conclusions

The architectural image in the city is complex and changeable. The buildings in the image have different directions and shapes.In this paper, we use the RESNET method to improve the faster R-CNN algorithm, which not only relieves the problem of gradient disappearance, deepens the depth of network model, but also strengthens the reuse of the underlying network characteristic information.In complex environment, more characteristic information can be extracted to identify buildings more accurately.In addition, a curved image recognition algorithm based on deep learning is proposed in this paper. This algorithm enhances the extraction of shallow features of the network by extending the path of pyramid network.Then, the ROI features of different layers are fused together through adaptive feature pooling. Finally, the target object is effectively identified through the progressive expansion algorithm.In addition, the way of generating labels and

the design of loss function are also introduced.The experimental results show that the algorithm has a good effect on the recognition of building images from different angles and different states.

**References**


[1] Harrison, C., et al., Foundations for Smarter Cities. Ibm Journal of Research and Development, 2010. 54(4).

[2] Batty, M., et al., Smart cities of the future. European Physical Journal-Special Topics, 2012. 214(1): p. 481-518.

[3] Batty, M., Big data, smart cities and city planning. Dialogues in human geography, 2013. 3(3): p. 274-279.

[4] Perera, C., et al., Sensing as a service model for smart cities supported by Internet of Things. Transactions on Emerging Telecommunications Technologies, 2014. 25(1): p. 81-93.

[5] Sanchez, L., et al., SmartSantander: IoT experimentation over a smart city testbed. Computer Networks, 2014. 61: p. 217-238.

[6] Zanella, A., et al., Internet of Things for Smart Cities. Ieee Internet of Things Journal, 2014. 1(1): p. 22-32.

[7] Botta, A., et al., Integration of Cloud computing and Internet of Things: A survey. Future Generation Computer Systems-the International Journal of Escience, 2016. 56: p. 684-700.

[8] Centenaro, M., et al., LONG-RANGE COMMUNICATIONS IN UNLICENSED BANDS: THE RISING STARS IN THE IOT AND SMART CITY SCENARIOS. Ieee Wireless Communications, 2016. 23(5): p. 60-67.

[9] Meijer, A. and M.P. Rodriguez Bolivar, Governing the smart city: a review of the literature on smart urban governance. International Review of Administrative Sciences, 2016. 82(2): p. 392-408.

[10] Akpakwu, G.A., et al., A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges. Ieee Access, 2018. 6: p. 3619-3647.

[11] Xiao Xiang , et al. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." IEEE Geoscience and Remote Sensing Magazine 5.4(2017):8-36.

[12] Pandit, Vaibhav R. , and R. J. Bhiwani . "Image Fusion in Remote Sensing Applications: A Review." International Journal of Computer Applications 120.10(2015):22-32.

[13] Irvin, R. Bruce, and David M. McKeown. "Methods for exploiting the relationship between buildings and their shadows in aerial imagery." IEEE Transactions on Systems, Man, and Cybernetics 19.6 (1989): 1564-1575.

[14] Liow, Yuh-Tay, and Theo Pavlidis. "Use of shadows for extracting buildings in aerial images." Computer Vision, Graphics, and Image Processing 49.2 (1990): 242-277.

[15] Stassopoulou, Athena, Terry Caelli, and R. Ramirez. "Automatic extraction of building statistics from digital orthophotos." International Journal of Geographical

Information Science 14.8 (2000): 795-814.
[16] Katartzis, A., et al. "Detection of buildings from a single airborne image using a Markov random field model." IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217). Vol. 6. IEEE, 2001.
[17] Lecun, Yann . "Deep learning & convolutional networks." 2015 IEEE Hot Chips 27 Symposium (HCS) IEEE, 2015.
[18] Krizhevsky, Alex , I. Sutskever , and G. Hinton . "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems 25.2(2012).
[19] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[20] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. (2015): 1-9.
[21] Neumann L, Matas J. A method for text localization and recognition in real-world images[C]// Asian conference on computer vision. Berlin: Springer, 2010: 770-783.
[22] Chen H Z , Tsai S S , Schroth G , et al. Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions[C]// 18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium: IEEE, 2011: 11-14.
[23] Neumann L , Matas J . Real-time scene text localization and recognition[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence RI: IEEE, 2012: 16-21.
[24] Gang S Z , Shen Y. New algorithm for text segmentation based on Stroke Filter[J]. Computer Science, 2010:4347 - 4350.
[25] Zayene O , Seuret M , Touj S M , et al. Text detection in Arabic News Video based on SWT operator and Convolutional Auto-Encoders[C]// 2016 12th IAPR Workshop on Document Analysis Systems (DAS). Santorini: IEEE, 2016: 11-14.
[26] Grzegorzek M, Li C, Raskatow J, et al. Texture-Based text detection in digital images with wavelet features and Support Vector Machines[C]// Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013. Heidelberg: Springer, 2013: 857-866.
[27] Hu H, Zhang C Q, Luo Y X, et al. WordSup: Exploiting word annotations for character based text detection[J]. International conference on computer vision, 2017: 4950-4959.
[28] Zhong Z Y, Jin L W, Huang S P, et al. DeepText: A new approach for text proposal generation and text detection in natural images[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans :IEEE, 2017: 1208-1212.