MDPI

Article

Deformable ConvNet with Aspect Ratio Constrained NMS for Object Detection in Remote Sensing Imagery

Zhaozhuo Xu [1], Xin Xu [1]*, Lei Wang [1], Rui Yang [1], and Fangling Pu [1,2]

[1] School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China; xinxu@whu.edu.cn; xuzhaozhuo@whu.edu.cn
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, Hubei 430079; flpu@whu.edu.cn
* Correspondence: xinxu@whu.edu.cn; Tel.: +86-27-6875-2836

Academic Editor: name
Version September 4, 2017 submitted to Entropy

Abstract: Development of very high resolution remote sensors provide us detailed geo-spatial objects information. Recently, convolutional neural networks (CNNs) have shown its power on object detection in remote sensing images. However, CNNs have obvious limitations in modeling geometric variations of remote sensing targets. In this paper, we first introduced a new CNN structure,namely deformable convolutional networks to tackle geometric modeling in object recognition. By adding offsets to convolution layers, feature mapping of CNN can be spread to unfixed locations. We develop an efficient transferring mechanism. Deformable ConvNet is constructed via substituting regular convolution to deformable convolution. Then the new established CNN will be fine-tuned based on pre-trained natural image model. To tackle the increase of outlier region proposals, we develop aspect ratio constrained non maximum suppression (NMS). The precision of deformable ConvNet in objects detection is improved. The deformable CNN, fine-tuning strategy and aspect ratio constrained NMS form a workflow, which has shown benchmark beating performance in very high resolution objects detection experiments.

Keywords: Convolutional neural networks (CNNs); remote sensing images; object detection

1. Introduction

Object detection is one of the main task in remote sensing. Accurate identification and localization of land targets provide structural spatial information of complex geographical scenes, which enables a branch of following applications in urban planning[1], environmental management[2] and military uses[3]. With the development of very high resolution (VHR) optical sensors, we are access to more detailed remote sensing imagery. In this case, "detailed" means not only the increasing spatial resolution, but also complicated geometric variations. Different from bikes, persons and flowers in natural images , objects such as airplane, ship, harbor and vehicle in VHR images have more possibilities in scale, orientation and deformation. Therefore, a key challenge in VHR visual recognition is to build geometric aware model for high level object understanding.

In general, the object detection framework can be summarized as three major components: feature representation, classification and localization. In the past, researchers have two ways to model geometric variations, both on feature representation. The first way is to add geometric priors in training samples, which is usually done by manually rotate training objects in 2D or 3D space[4]. The second is to extract rotation invariance features in images and use them as input to classifiers. This

method contains many well known methods which dominate the past decades in computer vision, such as HOG[5] and SIFT[6] features that extract statistical result of windows sliding on the images.

Obvious weakness can be found in two modeling ways above. For the first way, all the geometric diversities learned in classifier come from hand-craft operations, which makes the prior less reliable. Meanwhile, it is a strong supervision model that the classifier cannot recognize unknown transformations. With complex rotation procedure and costing computation time taken into consideration, the first way in inefficient and limited in performance. The second way also shares hand-craft problems, both SIFT and HOG features are build based on artificial selection of gradients between pixels, which lacks representation of invariance in other characteristics such as color and intensity.

Recently, the popularity of convolutional neural networks (CNNs) triggers the end-to-end model design and benchmark beating in object detection. As the strongest classifiers ever, CNN uses convolution as a sampling approach and generate massive weights through connected convolution and fully connected network structure. Massive weights mean massive information containers. CNN is able to store abundant structural information through training, which gives it power in complex object classification. In remote sensing images, AlexNet[7] is first introduced to VHR optical images and beats Bag-of-Words (BoW)[8], spatial sparse coding BoW (SSCBoW)[9], Fisher discrimination dictionary learning (FDDL)[10] and the collection of part detectors (COPD)[11] model in object detection[4]. The success of AlexNet opens the deep learning world for remote sensing imagery. More and more CNN structures such as region-based CNN (R-CNN), Fast R-CNN and Faster R-CNN are introduced to improve the precision in VHR object detection. Nowadays, R-P-Faster R-CNN[12] with region proposal networks (RPN) adding to Faster R-CNN beats the benchmark left by AlexNet and stands out as the state-of-art object recognition approach in VHR images. CNN's continuous benchmark beating and structural innovation inspires us, making us to wonder if geometric variation or transformations can be learned end-to-end through structural augment in CNN. After Deformable Parts Model is officially regarded as special CNN structure in 2015[13], CNN is able to represent 2D layout of object parts, which provide a theoretical milestone for its geometric modeling.

However, CNN has inherent limitations in modeling geometric variations shown in visual appearance. As described in[4], CNN is problematic while being applied to VHR object detection directly. Therefore, [4] developed rotation invariant CNN (RICNN), which augment training objects by rotating them in 360 degrees as same as first way introduced in the second paragraph. These approach achieve better performance than pure AlexNet and almost close to R-P-Faster CNN, but it add strong supervision to training process and break the end-to-end CNN framework. In fact, RICNN still does not solve the inherent limitation in CNN and adopt the most insufficient way to model geometric variation. But why? From my perspective, I think the fundamental problem of CNN modeling geometric transformation lies in CNN's convolution. Because all convolution operations conducted within CNN only sample input feature map in fixed locations, which forbid most possibilities happen in object geometrically. As long as we use traditional CNN structure such as AlexNet, the only thing we could do is adjust training samples artificially. Therefore, there is no denying that feature mapping in convolution must be reformed to fit the geometric proprieties of VHR targets.

To break the mapping limitations in CNN, [14] establishes deformable convolution. By adding 2D offsets to regular convolution grid, deformable convolution samples feature from flexible locations instead of fixed locations, which allows free deformation in forms of sampling grid. In other words, deformable convolution refined traditional convolution via adding preceding offsets layers. The deformable convolution modules substitute part of convolution layers in CNN and form deformable ConvNets (DCN), which contain massive internal transform parameters to model geometric propriety of objects. Different from RICNN or previous method such as DPM, deformable ConvNets build a deep and end-to-end geometric aware CNN model. Once utilized in VHR images, it may have surprising performance.

In this work, we proposed an end-to-end workflow to tackle geometric modeling problem in object detection for VHR remote sensing imagery. Deformable ConvNet with deformable convolution layers embedded on Region-based Fully Convolutional Networks (R-FCN) is first introduced in remote sensing area for object detection. After experiment on several VHR annotated images, problem has been found that deformable R-FCN has false positive bounding boxes in distorted aspect ratio. Some of the boxes has limited width, which more lines like than boxes. To solve this problem, we proposed an aspect ratio constrained non maximum suppression (NMS) to eradicate false results and improve precision. The deformable R-FCN, together with aspect ratio constrained NMS (arcNMS) form our complete solution to geometric variant object detection in VHR remote sensing images. Out major contribution is as follows:

1. A clean, comprehensive and end-to-end CNN structure for geometric variant VHR remote sensing imagery object detection. While geometric transformation modeling is completed within the convolution sampling layers, feature maps extracted by deformable ConvNets may contain more abundant information about objects' visual appearance in remote sensing views. Meanwhile, structurally speaking the proposed deformable ConvNets obtain the raw images as input directly and output bounding boxes and labels without artificial operations, which proves its end-to-end proprieties.
2. An efficient strategy of network training for geometric variant VHR remote sensing imagery object detection. Because the expense of VHR remote sensing techniques is relatively high compared to other optical remote sensing approaches, the resources of annotated VHR optical imagery are limited. To tackle this problem, researches have been focused on fine-tuning parameters from well-known CNN in natural images to obtain accurate object detection in VHR remote sensing imagery. Our work proposed a time and computation saving fine-tuning approach. The deformable ConvNet substitutes several convolution layers in R-FCN with deformable convolution layers and then learn VHR imagery based on parameters obtained by R-FCN in natural images. In this way objects in VHR remote sensing images are better understood by deformable ConvNet in lesser time and lighter computation devices.
3. A precision improvement approach for deformable ConvNet in VHR remote sensing imagery object detection. Our work proposed an aspect ratio constrained NMS to improve deformable ConvNet's performance in remote sensing area. By modeling the logarithm of bounding boxes' aspect ratio in training annotations, arcNMS detect anomaly bounding boxes generated by deformable ConvNet and delete them, which restrain the number of false positive proposals in VHR remote sensing images and improve the precision of deformable ConvNet.

The proposed deformable ConvNet with arcNMS is evaluated and compared with established VHR object detection methods in NWPU VHR 10 dataset[15]. The experiment results confirm our assumptions and prove deformable ConvNets outperform state-of-art CNN models in object detection tasks.

The rest of our paper is described as follows: Section 2 describes the proposed deformable ConvNet and arcNMS method, Section 3 presents the dataset and experimental settings for object recognition performance evaluation and comparison on VHR remote sensing imagery. The results of deformable ConvNets and other approaches in NWPU VHR 10 dataset are presented in Section 4. Section 5 draws the conclusion.

2. Proposed Method

The architecture of CNN model for object detection can be summarized as three major components: layers, network structure and NMS. In this section we introduce our proposed deformable ConvNet in the following of this three components. First, we presents the fundamental concepts of deformable convolution and its illustration in image feature sampling. Then we officially introduced our deformable ConvNet architecture for object detection in VHR imagery.

Including CNN layers, used pre-trained parameters and fine-tuning mechanism. In the end arcNMS post-processing step is described to illustrate how outlier bounding boxes can be eradicated.

2.1. Deformable Convolution

For image processing, convolution can be regarded as 2D spatial sampling. Given the weights n_1, n_2 in grid Ψ, convolution to input feature map $f(x, y)$ is

$$y\left[x, y\right] = \sum_{(n_1, n_2) \in \Psi} w\left[n_1, n_2\right] f\left[x - n_1, y - n_2\right] \tag{1}$$

where $y\left[x, y\right]$ represents the output of convolution between weights grid and input, in tasks such as denoising, smoothing and edge detection[16], grid Ψ is usually recognized as kernels.

In paper [14], deformable convolution is achieved by augmenting input feature map with 2D offsets while convolution. For better understanding in image processing perspective, we formulize deformable convolution as follows

$$y\left[x, y\right] = \sum_{(n_1, n_2) \in \Psi} w\left[n_1, n_2\right] f\left[x - n_1 - \Delta n_1, y - n_2 - \Delta n_2\right] \tag{2}$$

where convolution becomes weighted summation of unfixed locations of the input feature grid, which generate its diversities.

However, as the offsets are non-integer, the value of $f\left[x - n_1 - \Delta n_1, y - n_2 - \Delta n_2\right]$ remains to be determined. In our work we adopted bilinear interpolation to obtain the fractional location between integer pixels. If points clique (x, y) fall into a 2×2 pixels area with x ranging from x_1 to $x_1 + 1$ and y from from y_1 to $y_1 + 1$. The value of $f\left[x, y\right]$ by bilinear interpolation is approximate to

$$\begin{aligned} f\left[x, y\right] \approx\ & f\left[x_1, y_1\right](1 - x + x_1)(1 - y + y_1) + f\left[x_1 + 1, y_1\right](x - x_1)(1 - y + y_1) \\ & + f\left[x_1, y_1 + 1\right](1 - x + x_1)(y - y_1) + f\left[x_1 + 1, y_1 + 1\right](x - x_1)(y - y_1) \end{aligned} \tag{3}$$

To express this function generally, we have

$$f\left[x, y\right] \approx \sum_{(x_t, y_t) \in \Phi} g(x_t, x) \cdot g(y_t, y) \cdot f\left[x_t, y_t\right] \tag{4}$$

where Φ enumerates all integer points clique near $f\left[x, y\right]$. The interpolation kernel is decomposed into 2 kernels in x and y dimension. Both kernels are defined as

$$g(a, b) = max(0, 1 - |a - b|) \tag{5}$$

With deformable convolution function determined, the next step is to learn the offsets through training. In Figure 1 we illustrate how to obtain offsets by augmenting convolution layers with additional offsets field. With offsets vectors having the same spatial resolution as input feature maps, original sampling points spread outward to focus on the objects. While offsets and convolution weights are fully learned, CNN will have better geometric sampling features. Therefore, in the test procedure, deformable ConvNets are able to generate more accurate bounding boxes according to irregular feature mapping.

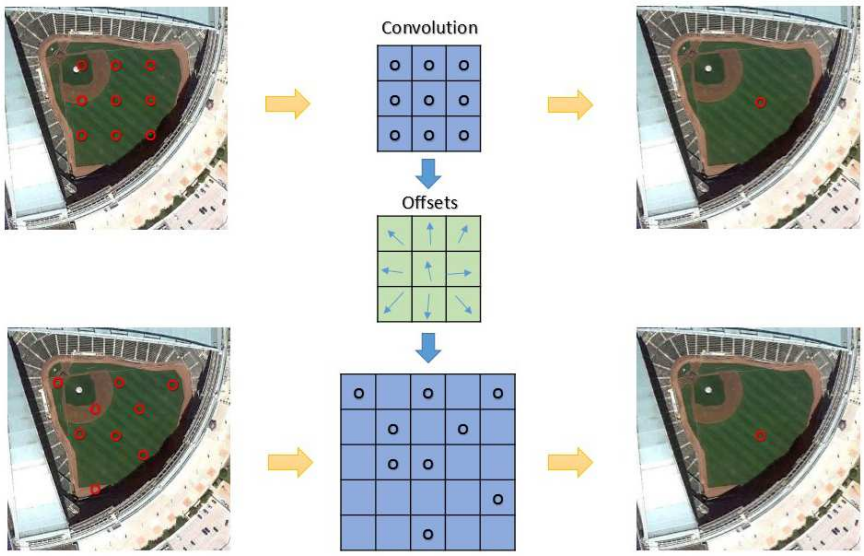


Figure 1. Illustration of deformable convolution on VHR remote sensing imagery

2.2. Deformable R-FCN

For image recognition, convolutional Neural Networks have been continuously enhanced since AlexNet[7] and VGG Nets[17]. In contrast to these two famous nets which consist of convolutional subnetwork with Region-of-Interest (RoI) pooling and fully connected layers, recently designed image classification CNN such as GoogeLeNets[18] and ResNets[19] are fully convolutional. However, in object detection area, state-of-art architectures like Faster R-CNN[20] are still using RoI subnetworks with hidden, computational unshared layers. To remedy this issue, Dai and his workmates developed Region-based Fully Convolutional Networks (R-FCN)[21] for object detection. The key parts of R-FCN are position-sensitive score maps that encode positioning information using banks of specialized convolutional layers and position-sensitive RoI pooling that gather all information contained in feature score maps. With these two parts, R-FCN realize an end-to-end fully convolutional network with all computation share on entire image instead of hundreds of region-based subnetworks. Results have shown that using ResNet-101 model pre-trained on ImageNet, R-FCN outperform Faster R-CNN and other network structures on both VOC and COCO object detection datasets.

In our work, the CNN architecture is close to [14], but different in training strategy. As shown in Figure 2, on the basis of R-FCN which contains fully convolutional feature maps, RoI pooling and RPN, we use ResNet101 ImageNet pre-trained parameters as the initial values and substitute res5, res4b22, res4b21 and res4b20 layers by deformable convolution layers. Then the deformable R-FCN architecture will be fine-tuned by NWPU VHR 10 images as input.

This procedure provides and fine-tuned strategy to make full use of powerful CNN architecture and pre-trained natural image classification models. Once has a newly developed networks with better performance in ImageNet competition, we can alternate several convolution layers by deformable convolution layers and train them on limited remote sensing object imagery. This strategy help CNN learn the basic region invariance of both objects and background from well-established natural image models, which saves both time and computation resources. Then geometric proprieties that distinguished remote sensing objects will be modeled by deformable convolution layers, which enhance CNN's locality on remote sensing imagery. Therefore, higher precision object detection with be realized when VHR remote sensing imagery in limited.

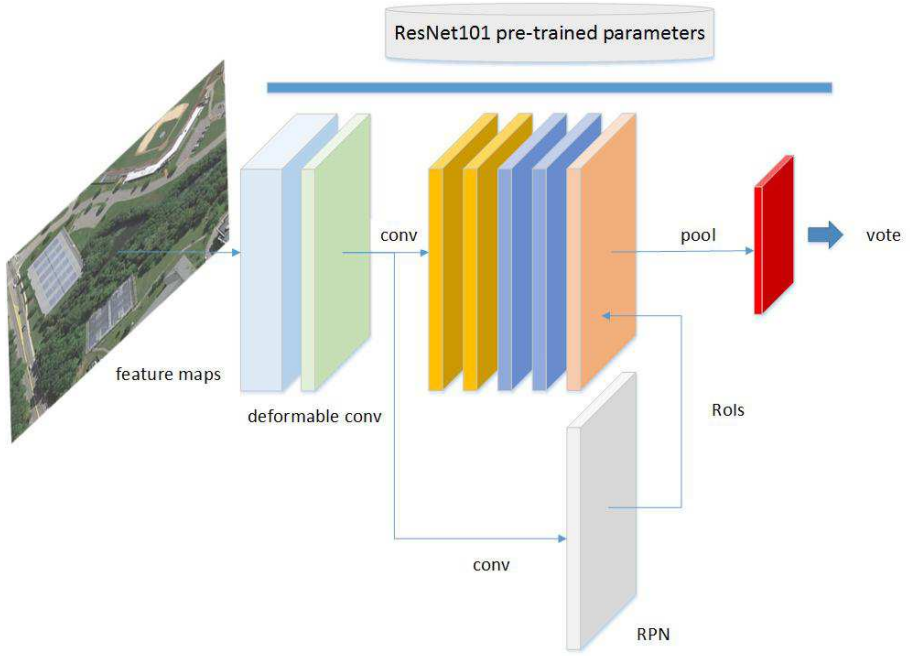


Figure 2. Deformable ConvNets based on R-FCN architecture

2.3. *Aspect Ratio Constrained NMS*

2.3.1. Lines like False Region Proposals

To test developed deformable ConvNets architecture , we conduct experiments on NWPU VHR 10 dataset with non maximum suppression. While inspecting the detection results, interesting phenomenon has been founded. More false region proposals have lines like shape instead of a box after. Because the proposals are processed by regular NMS, it is reasonable for us to assume that all the proposed regions are lines. After reexamine the concepts of deformable convolution, we found a possible explanation to this problem. Because deformable convolution samples features from flexible locations, sets of irrelevant points may be recognized as objects if they are exactly moved to object points by offsets field. This coincidence is illustrated in Figure 3. For a VHR remote sensing imagery containing baseball diamond, during the test procedure the deformable convolution layers in deformable ConvNet samples a line on the bottom of the image pieces with learned offsets vectors. If, however, the samples after deviation is actually falling into the baseball diamond area, then in the following layers the original pixels in this line will be recognized as objects no matter how the CNN changed.

There are already work published trying to remedy strange region proposals generated by CNN. In paper [22], the bounding boxes in unnatural aspect ratios are explained as confusion of global contexts. For instance, if the context of baseball diamond is similar to harbor or road, the false ship or car regions will be proposed. And most of these proposals will have thin and long shapes. Paper [22] summarized this and related problems as challenges easily overcame by traditional object detection method such as DPM, but usually ignored by region-based CNN. Therefore, an aspect ratio and context aware fully convolutional network (ARCFCN) is developed to add aspect ratio and context aware part-based models into FCN. Experiments have shown that ARCFCN benefits FCN and Faster R-CNN frameworks in precision.

However, augmenting FCN with context aware part-based models increase the computation complexities of CNN architecture, making it time-consuming. Moreover, the ARCFCN reduces the false positive region proposals, which will only improve the precisions but has no benefits to the recall rates. To remedy lines like false proposed boxes in simpler way, we focus on refining non

maximum suppression instead of CNN architecture. Non maximum suppression (NMS) is one of the key post-processing step in computer vision applications including edge detection, object detection and semantic segmentation[23]. NMS provides a Intersection over Union (IoU) based workflow to select region proposals to get the accurate object localization. Once aspect ratio are taken into NMS's consideration while selecting, lines like false boxes may be eliminated.

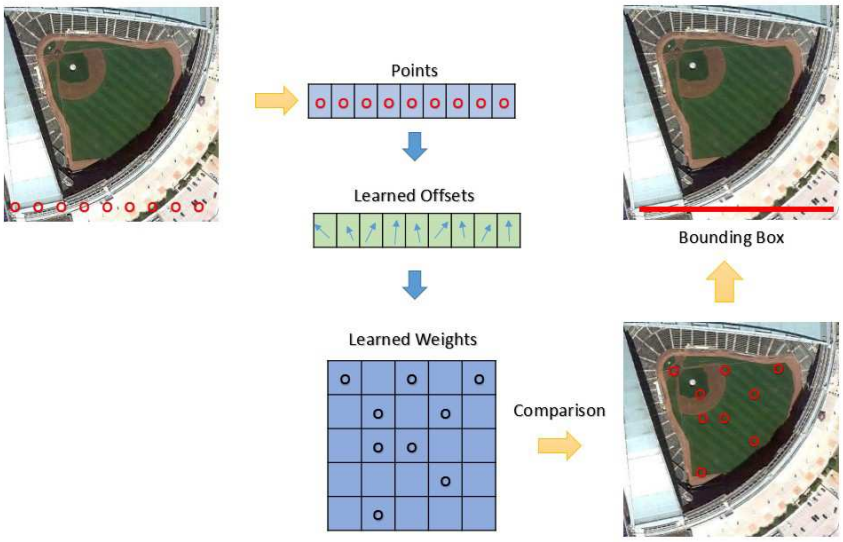


Figure 3. Illustration of line outlier bounding boxes generated by deformable ConvNets

2.3.2. Aspect Ratio Constrained NMS

In our work, a non maximum suppression with aspect ratio constraints is developed and applied to augment deformable ConvNet in object detection of VHR remote sensing imagery. To construct appropriate aspect ratio constraints, we calculate the logarithm of aspect ratios found both training samples and test results as

$$AR = log(\frac{length}{width + \delta_t}) \tag{6}$$

where $length$ and $width$ represent distance in x and y dimension between lower left and upper right vertexes of bounding boxes. δ_t is the fractional coefficient in case the denominator becomes zero. In our experiment we set δ_t as 10^{-46}. Then we derive the distribution of AR in both train and test sets. As Figure 4 shows, the logarithm of annotated bounding boxes appear to be normal distribution except large volume of samples have zero ARs, which means these boxes have same length and width. However, in the distribution shown in Figure 5, the ARs of regions proposed by deformable ConvNets have three peaks in locations other than zero. Certainly these three peaks represent outlier bounding boxes generated in deformable ConvNet. Once these peaks are eradicated in NMS, precision of deformable ConvNet will be improved.

Therefore, for all deformable ConvNet proposed regions, the aspect ratio constraint is developed as

$$c_t = \begin{cases} 1 & \text{if} \quad |AR_t - \mu| < 3\sigma \\ 0 & \text{if} \quad |AR_t - \mu| > 3\sigma \end{cases} \tag{7}$$

where μ and σ are mean value and standard deviation of ARs in training annotations. For each bounding boxes proposed by deformable ConvNets, if the difference between its AR and μ exceeds 3σ, the region proposal is recognized as outlier with C_t becomes zero.

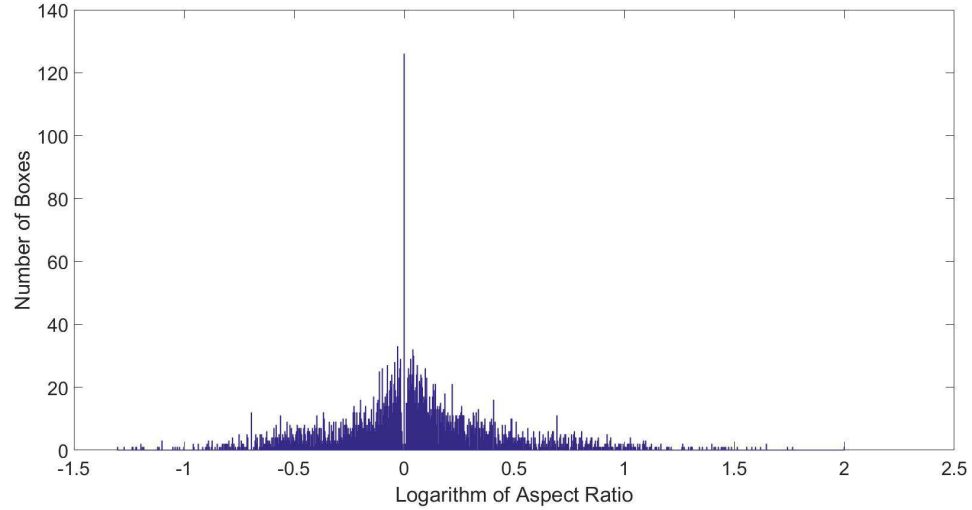


Figure 4. Aspect ratio of training annotations

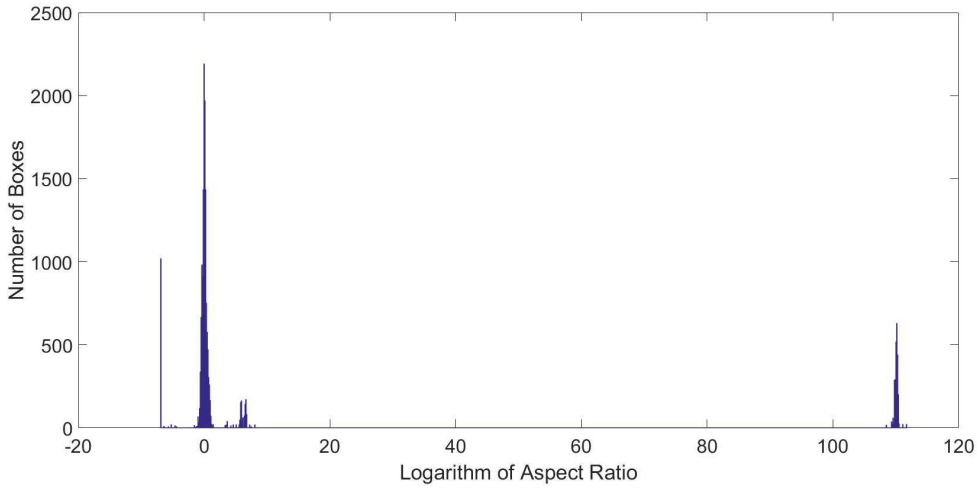


Figure 5. Aspect ratio of region proposed by deformable ConvNet

Then, the input to NMS algorithm is refined as

$$(b_t, s_t) = c_t \cdot (b_t, s_t) \tag{8}$$

where $B = \{b_1, b_2, \ldots b_n\}$ are the list of initial detection boxes. $S = \{s_1, s_2, \ldots s_n\}$ represent the corresponding scores of proposed boxes. For each $b_t \in B$ and $s_t \in S$, constraint c_t is added to (b_t, s_t), which delete all the outlier cliques by making it zero. Then the lists of region proposals and corresponding scores will be selected by NMS according to there IoU. Based on the arcNMS post-processing step, the precision of deformable ConvNet in object localization will be increased. The deformable network structure will provide powerful benchmark beating capacities in object detection.

3. Dataset and Experimental Settings

In order to evaluate and validate the effectiveness of deformable ConvNet and arcMNS on HSR remote sensing imagery, the utilized dataset, experimental settings, and the corresponding evaluation indicators of the experimental results are described in this section.

3.1. Dataset and Implementation Details

To compare the performance of various approaches developed for object detection in remote sensing images, many datasets are available for researchers to conduct further investigations. For vehicles detection, vehicle detection in aerial imagery (VEDAI) dataset[24], which is comprised of 1268 RGB tiles (1024×1024 px) and the associated infrared (IR) image at 12.5 cm spatial resolution, is broadly used as benchmarks[25]. VEDAI dataset contains 3687 annotated vehicles with types varying from home vehicles such as car, van and camping to specialized vehicles including tractor and boats. For airplanes detection, paper [3] presents an annotated airport dataset with pictures of Sydney International Airport, Tokyo Haneda Airport and Berlin Tegel Airport downloaded from Google Earth. The airport dataset is also state-of-art benchmark with extra works building on it[26]. Besides two famous object detection benchmarks above, according to survey [27], datasets such as SZTAKI-INRIA building detection dataset[28], ITM road extraction dataset[29] and TAS aerial car detection dataset[30] are applied in some works for training or validation uses. These datasets promote the development of object detection methods in remote sensing imagery, but still have obvious drawbacks. First, the volume of annotations in these datasets are limited, which constrain the power of CNN for it requires large scale training samples. Then, datasets are specialized as certain types of object such as vehicles planes or roads, but lacks a comprehensive benchmark for remote sensing imagery. To remedy this issue, Paper [31] collected 2326 images from Google Earth and then annotated oil tanks, aircrafts, overpasses and playgrounds on them. However, only training image sets are released. We cannot compare different CNN's performance on it.

In our work we select NWPU VHR 10 dataset as benchmarks based on all the considerations above. The advantages of NWPU VHR 10 dataset can be summarized as:

1. Source and resolution diversity. NWPU VHR 10 dataset not only contains optical remote sensing images, but also includes pan-sharped color infrared images. 715 images were downloaded from Google Earth with spatial resolutions from 0.5 m–2.0 m. Meanwhile 85 pan-sharpened color infrared images were acquired from the Vaihingen data with a 0.08 m spatial resolution.
2. Comprehensive object types. NWPU VHR 10 dataset contains 10 different types of objects, including airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.
3. Abundant object annotations.The NWPU VHR-10 dataset contains 650 annotated images, within each image containing at least one target to be recognized. For the image set in VOC 2007 formula, 757 airplanes, 302 ships, 655 storage tanks, 390 baseball diamonds, 524 tennis courts, 150 basketball courts, 163 ground track fields, 224 harbors, 124 bridges, and 477 vehicles have been manually annotated with rectangular bounding boxes, and were utilized as the training samples and testing ground truth.

To evaluate deformable ConvNet with comparison to RICNN and R-P-Faster R-CNN, we establish same experiment as [12]. The ratios of training, validation and testing dataset are set as 20%, 20% and 60%. Then, we randomly select 130, 130 and 390 images in NWPU VHR 10 dataset to fill the three subsets correspondingly. Deformable ConvNets and comparison models are implemented on server with Nvidia GTX 1080Ti GPU. In the training of deformable ConvNet, we set learning rate as 0.0005 and fine-tuned them based on ResNet-101 pre-trained models. The RPN parameters in deformable ConvNet are same as paper[14]. To compare deformable ConvNet's fine-tuned efficiency, we implemented transfered AlexNet, newly trained AlexNet, RICNN with and without fine-tuning on ImageNet, and R-P-Faster R-CNN with Zeiler and Fergus (ZF) model or the visual geometry group

(VGG) model fine-tuned on ImageNet. While deformable ConvNet's performance is evaluated with other CNN structure, obvious improvement will be observed easily.

3.2. Evaluation Indicators

We adopt the precision–recall curve (PRC) and average precision (AP) to quantitatively evaluate the performance of different CNNs in object detection. They are two well-know and widely applied standard measures approaches comparisons[10] [27].

3.2.1. Precision–Recall Curve

PRC came from four well-established evaluation components in information retrieval, namely, true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP and FP represent the correctly and falsely detected objects' ratio in all region proposals. FN is the sum of regions not proposed. Based on this four component we provide the definition of precision and recall rate as

$$Precision = \frac{TP}{(TP + FP)} \tag{9}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{10}$$

PRC is built as a detection map developed on *Precision* and *Recall*. If the ratio of overlap between proposals and ground truth exceeds 0.5, the proposals are recognized as TP, otherwise they are FP. Typically, Precision and Recall are inversely related, ie. as Precision increases, recall falls and vice-versa. A balance between these two needs to be achieved by the IR system, and to achieve this and to compare performance, the precision-recall curves come in handy. The relationship between *Precision* and *Recall* is displayed in PRC by many times experiments. The existence of PRC provide detailed inspection of a model's performance in object detection.

3.2.2. Average Precision

The AP computes the average value of Precision over the interval from Recall = 0 to Recall = 1, i.e., the area under the PRC. In addition, mean AP (mAP) computes the average value of all the AP values of AP value is widely used as the quantitative indicators in object detection[8] [6] [27]. Most paper recognize higher AP values as the explicit proof of benchmark beating. Except AP value, paper [12] utilized accuracy form image classification as follows for performance comparison.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{11}$$

However, *Accuracy* is questionable in object detection. As object detection is different from image classification or retrieval, detector will not draw boxes on regions without objects, which means there will be no true negative samples in train and test. It is meaningless taking *Accuracy* as indicator. Moreover, the *Accuracy* in paper [12] is not match the equation when TN is zero. Therefore, *Accuracy* is not applied in our work.

4. Results

Visualization of objects detected by deformable R-FCN in NWPU VHR 10 dataset is shown in Figure 6. From the figure deformable R-FCN shows a better detection results in orientation variant targets such as airplanes (top left), ships (top right), bridges (bottom left) and vehicles (bottom right). Meanwhile, Figure 6 also shows deformable R-FCN has better performance in recognize adjacent or overlapped objects such as harbors, storage tanks and ball courts. In the following subsection, quantitative and qualitative analysis are presented to evaluate deformable R-FCN's performance.

Figure 6. Visualization of objects detected by deformable R-FCN in NWPU VHR 10 dataset

4.1. Quantitative Evaluation

Quantitative comparisons measured by AP values and average running time per image are displayed in Tables 1 and 2. Because the NMS process has no impact on TP and FN, recall rate of deformable R-FCN with and without arcNMS are the same. The recall rate value is shown in Figure 7. The proposed deformable R-FCN uses ResNet-101 pre-trained on ImageNet. Besides ZF model, the R-P-Faster R-CNN adopts VGG16 training mechanism, namely single fine-tuning and double fine-tuning[17]. The RICNN with fine-tuning utilizes AlexNet pre-trained on ImageNet. In Table 1, it can be seen that R-FCN outperforms other approaches in precision of storage tank, ground track field and vehicle, but suffers in other and overall precision. Deformable R-FCN enhances R-FCN in the detection of ship, tennis court, bridge and vehicle, thus pushing the benchmark into 78.4%. After arcNMS is added to deformable R-FCN, APs among objects such as airplane, baseball diamond, tennis court, basketball court, ground track field, harbor and bridge are all increased. The best mean AP value among all objects, thus, is obtained by deformable R-FCN with arcNMS fine-tuned on ResNet-101 ImageNet pre-trained model. Table 2 shows the average running time of all CNN architecture and previous approaches. For object detection in remote sensing, deformable R-FCN's running time per image is slightly slower than R-FCN and R-P-Faster R-CNN for its additional offsets layers. But generally speaking deformable R-FCN's time cost is acceptable compared to traditional methods and AlexNet. Figure 7 the recall values of the proposed deformable R-FCN, deformable R-FCN obtained 87.96% overall recall rate. Compared to R-P-Faster R-CNN[12] and RICNN[4], the average recall rate is lower than R-P-Faster R-CNN with VGG model and RICNN, but higher than R-P-Faster R-CNN with ZF model. Generally speaking, deformable R-FCN is proved to increase precision in object detection of VHR remote sensing images, especially geometric variant objects such as bridge, vehicle and baseball diamond. arcNMS is also validated on its enhancing effect on deformable R-FCN. Obviously there is a trade-off between precision and recall, and the following PRC will go deep into this problem to evaluate performance of various CNN structures.

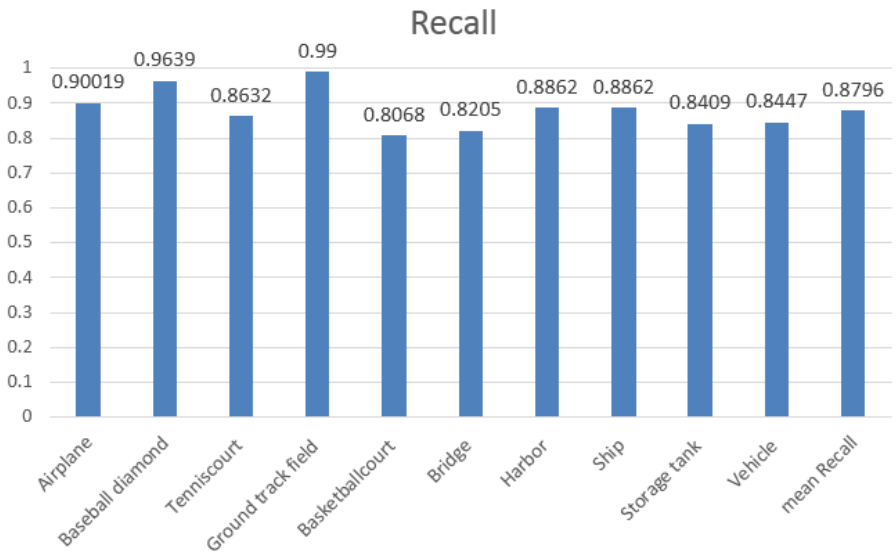


Figure 7. Aspect ratio of training annotations

Table 1. The AP values of the object detection methods.

	BoW	SSC BoW	FDDL	CPOD	Transfered AlexNet	Newly Trained AlexNet	RICNN without Fine-Tuning	RICNN with Fine-Tuning
Airplane	0.025	0.506	0.292	0.623	0.661	0.701	0.860	0.884
Ship	0.585	0.508	0.376	0.689	0.569	0.637	0.760	0.773
Storage tank	0.632	0.334	0.770	0.637	0.843	0.843	0.850	0.853
Baseball diamond	0.090	0.435	0.258	0.833	0.816	0.836	0.873	0.881
Tennis court	0.047	0.003	0.028	0.321	0.350	0.355	0.396	0.408
Basketball court	0.032	0.150	0.036	0.363	0.459	0.468	0.579	0.585
Ground track field	0.078	0.101	0.201	0.853	0.800	0.812	0.855	0.867
Harbor	0.530	0.583	0.254	0.553	0.620	0.623	0.665	0.686
Bridge	0.122	0.125	0.215	0.148	0.423	0.454	0.585	0.615
Vehicle	0.091	0.336	0.045	0.440	0.429	0.448	0.680	0.711
mean AP	0.246	0.308	0.245	0.546	0.597	0.618	0.710	0.726

	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (Double)(VGG16)	R-P-Faster R-CNN (Single)(VGG16)	R-FCN (ResNet-101)	Deformable R-FCN (ResNet-101)	Deformable R-FCN (ResNet-101) with arcNMS
Airplane	0.803	0.906	0.904	0.817	0.861	**0.873**
Ship	0.681	0.762	0.750	0.806	**0.816**	0.814
Storage tank	0.359	0.403	0.444	**0.662**	0.626	0.636
Baseball diamond	0.906	0.908	0.899	0.903	0.904	**0.904**
Tennis court	0.715	0.797	0.79	0.802	**0.816**	**0.816**
Basketball court	0.677	0.774	0.776	0.697	0.724	**0.741**
Ground track field	0.892	0.880	0.877	**0.898**	0.898	**0.903**
Harbor	0.769	0.762	0.791	0.786	0.722	**0.753**
Bridge	0.572	0.575	0.682	0.478	**0.714**	**0.714**
Vehicle	0.646	0.666	0.732	**0.783**	**0.757**	0.755
mean AP	0.702	0.743	0.765	0.763	**0.784**	**0.791**

Table 2. Computation time comparisons for objects detection methods

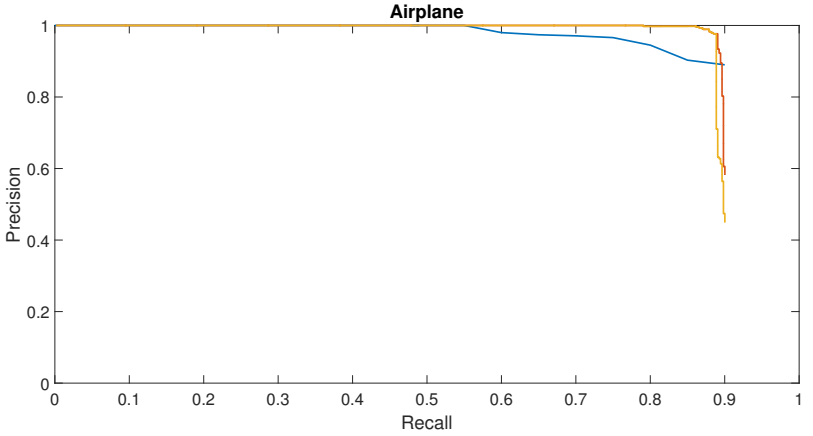
	BoW	SSC BoW	FDDL	CPOD	Transfered CNN	Newly Trained CNN	RICNN without Fine-Tuning	RICNN with Fine-Tuning
Average running time per image (second)	5.32	40.32	7.17	1.06	5.24	8.77	8.77	8.77

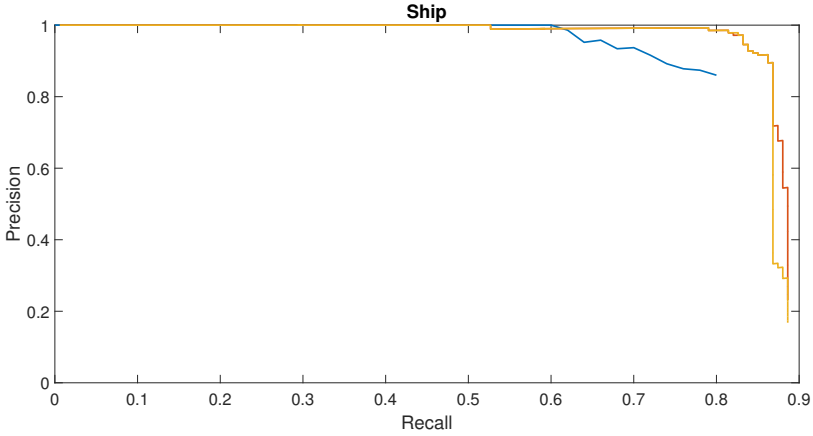
	R-P-Faster R-CNN (ZF)	R-P-Faster R-CNN (Double)(VGG16)	R-P-Faster R-CNN (Single)(VGG16)	R-FCN (ResNet-101)	Deformable R-FCN (ResNet-101)	Deformable R-FCN (ResNet-101) with arcNMS
Average running time per image (second)	0.005	0.155	0.155	0.156	0.201	0.201

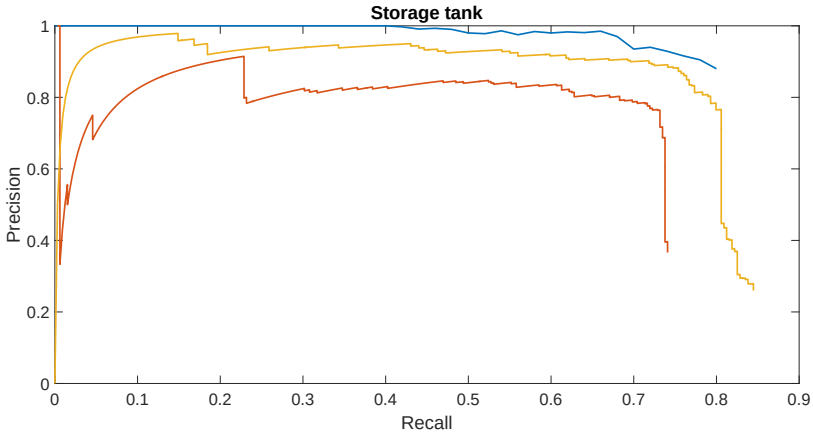
4.2. PRC evaluation

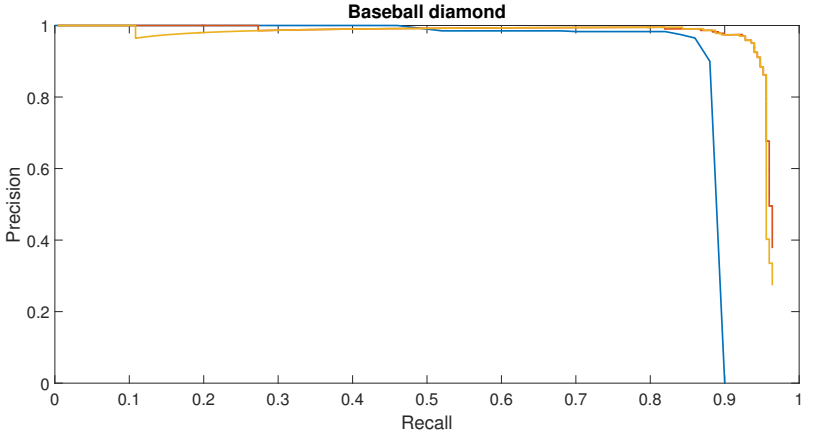
For object detection approaches, PRC is one of the primary indicators of robustness and effectiveness. In PRCs, precision vector generated in experiments is measured in Y axis and the recall rates are in X axis. The curve at the top of the PRCs indicates a better performance. In previous papers[4] [12], COPD, FDDL, SSCBoW, BoW, AlexNet are all displayed for comparison. Too many curves make it hard to recognize new established approaches and compare it with state-of-art CNNs. As traditional approaches' performance are fully evaluated.

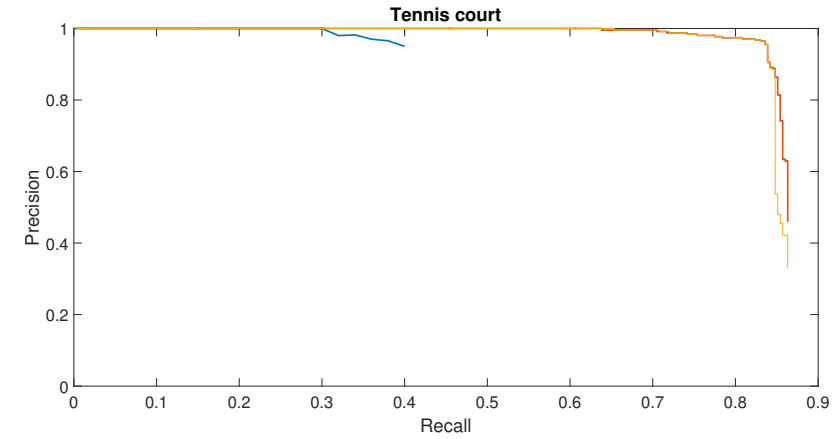
In this paper we focus on top three object detection methods in Table 1, which are deformable R-FCN with arcNMS, deformable R-FCN and R-P-Faster R-CNN trained on VGG16 model in single

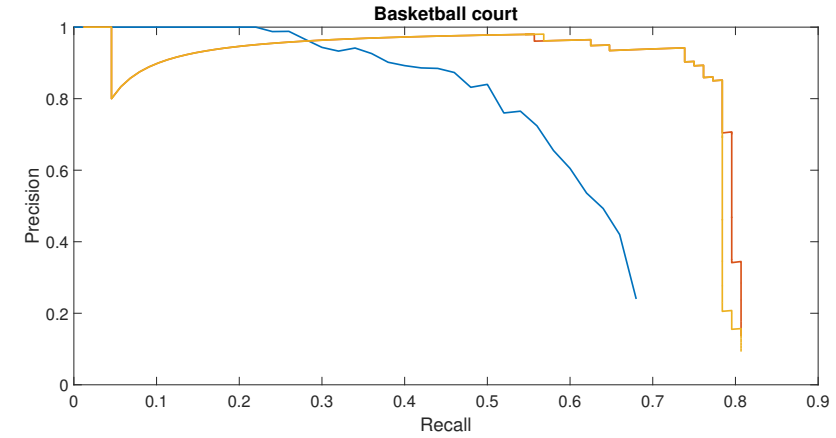

Airplane
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9

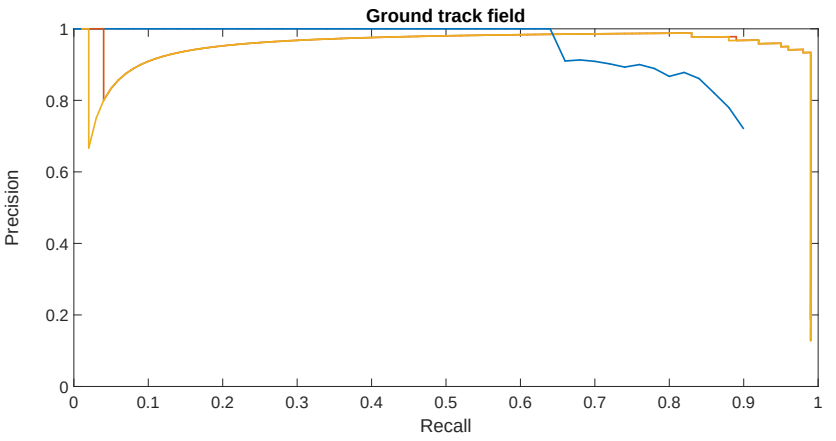

Ship
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9

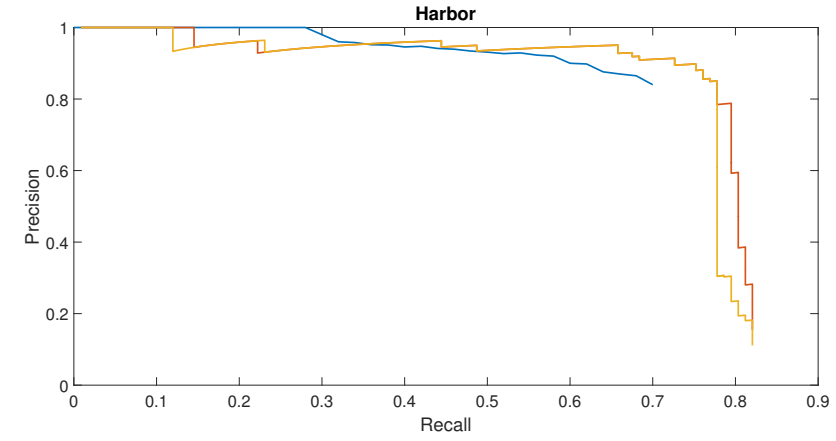

Storage tank
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


Baseball diamond
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


Tennis court
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


Basketball court
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


Ground track field
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


Harbor
Precision
Recall
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9

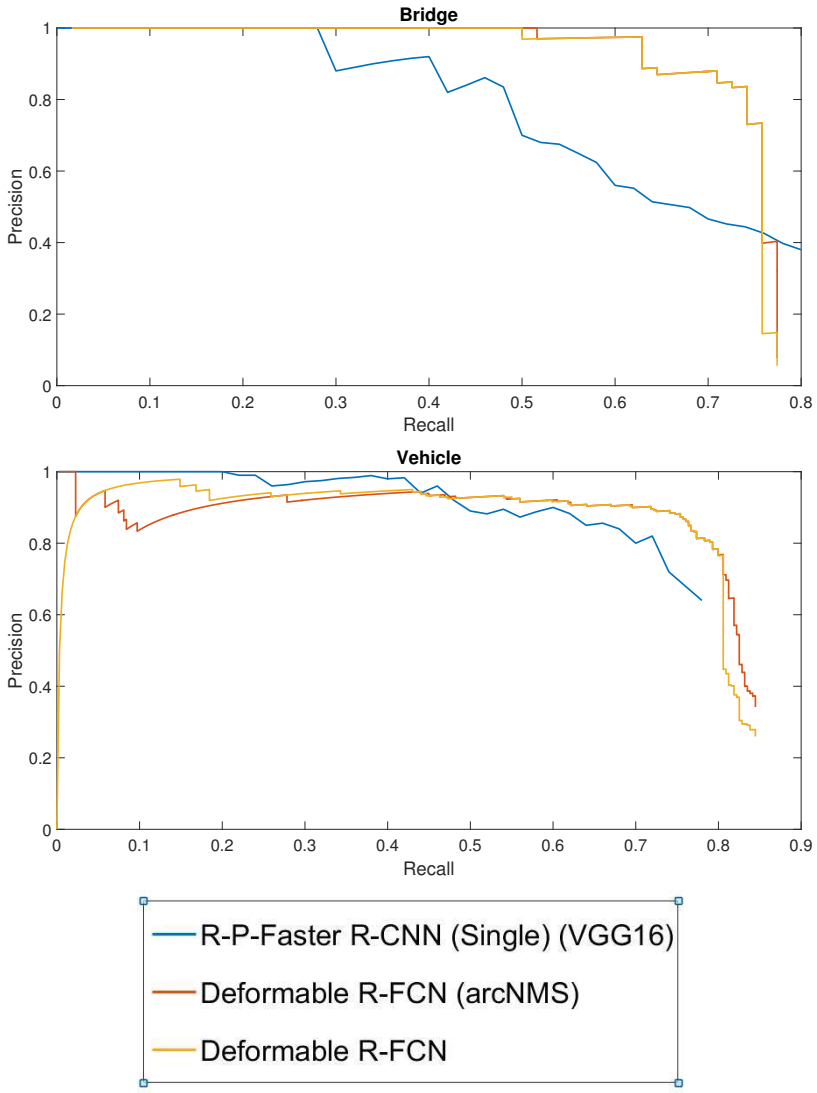


Figure 8. PRCs of top three object detection approaches in mean AP

fine-tuning mode. Figure 8 shows the PRCs of these three methods. For deformable R-FCN, it can be seen that most of the classes show a better detection performance than R-P-Faster R-CNN. But in the classes of storage tank, basketball court, vehicle and harbor, deformable R-FCN requires improvements. Moreover, arcNMS is proved to be effective in improving AP value by resisting PRC from falling to quickly. By jointly analyzing the AP values, the recall rate, and the PRCs, it can be seen that the proposed deformable R-FCN with arcNMS algorithm shows a superior detection performance for VHR remote sensing objects.

5. Conclusions

In this paper, an end-to-end deformable convolutional neural network structure is presented to model geometric variations in VHR remote sensing object. While standard convolution sampling is substituted by 2D offsets flexible feature mapping, the CNN is capable of recognizing remote sensing objects in more complicated visual appearance. We also proposed a transfer mechanism of VHR remote sensing objects. Deformable ConvNets become more effective when fine-tuning on pre-trained natural image CNN models. Finally, a post-processing arcNMS is developed to delete outlier region proposals and improve precision.

Our workflow has been evaluated under NWPU-VHR-10 dataset. Results show that proposed deformable R-FCN with arcNMS approach outperforms state-of-art benchmarks in object detection. Detailed investigation proves that deformable R-FCN has better performance in geometric diverse objects such as bridge, harbor and baseball diamond. And arcNMS is also proved to enhance deformable ConvNets. Experiment on running time and PRC confirm that deformable ConvNets are efficient and effective. Deformable CNN explanation and visualization may be researched in future works.

Acknowledgments: This work was supported by

Author Contributions: All the authors made significant contributions to the work. Zhaozhuo Xu, Lei Wang, Rui Yang designed the research and analyzed the results. Xin Xu and Fangling Pu provided advice for the preparation and revision of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Bibliography

1. Gui, R.; Xu, X.; Dong, H.; Song, C.; Pu, F. Individual Building Extraction from TerraSAR-X Images Based on Ontological Semantic Analysis. *Remote Sensing* **2016**, *8*, 708.
2. Zhong, P.; Wang, R. A Multiple Conditional Random Fields Ensemble Model for Urban Area Detection in Remote Sensing Optical Images. *IEEE Transactions on Geoscience and Remote Sensing* **2007**, *45*, 3978–3988.
3. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54*, 5553–5563.
4. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54*, 7405–7415.
5. Cheng, G.; Han, J.; Guo, L.; Liu, T. Learning coarse-to-fine sparselets for efficient object detection and scene classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1173–1181.
6. Yuan, Y.; Hu, X. Bag-of-Words and Object-Based Classification for Cloud Extraction From Satellite Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2015**, *8*, 4197–4205.
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90.
8. Xu, S.; Fang, T.; Li, D.; Wang, S. Object Classification of Aerial Images With Bag-of-Visual Words. *IEEE Geoscience and Remote Sensing Letters* **2010**, *7*, 366–370.

9. Sun, H.; Sun, X.F.; Wang, H.; Li, Y.; Li, X. Automatic Target Detection in High-Resolution Remote Sensing Images Using Spatial Sparse Coding Bag-of-Words Model. *IEEE Geoscience and Remote Sensing Letters* **2012**, *9*, 109–113.
10. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Bu, S.; Wu, J. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *Isprs Journal of Photogrammetry and Remote Sensing* **2014**, *89*, 37–48.
11. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing* **2014**, *98*, 119 – 132.
12. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sensing* **2017**, *9*, 666.
13. Girshick, R.; Iandola, F.; Darrell, T.; Malik, J. Deformable Part Models are Convolutional Neural Networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
14. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. *arXiv preprint arXiv:1703.06211* **2017**.
15. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *Isprs Journal of Photogrammetry and Remote Sensing* **2016**, *117*, 11–28.
16. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing (3rd Edition)*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2006.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* **2014**, *abs/1409.1556*.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149.
21. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 379–387.
22. Li, B.; Wu, T.; Shao, S.; Zhang, L.; Chu, R. Object Detection via End-to-End Integration of Aspect Ratio and Context Aware Part-based Models and Fully Convolutional Networks. *CoRR* **2016**, *abs/1612.00534*.
23. Rothe, R.; Guillaumin, M.; Van Gool, L. Non-Maximum Suppression for Object Detection by Passing Messages between Windows. *asian conference on computer vision* **2014**, *9003*, 290–306.
24. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation* **2016**, *34*, 187 – 203.
25. Audebert, N.; Le Saux, B.; Lefevre, S. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing* **2017**, *9*, 368.
26. Wang, G.; Wang, X.; Fan, B.; Pan, C. Feature Extraction by Rotation-Invariant Matrix Representation for Object Detection in Aerial Image. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 851–855.
27. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *{ISPRS} Journal of Photogrammetry and Remote Sensing* **2016**, *117*, 11 – 28.
28. Benedek, C.; Descombes, X.; Zerubia, J. Building Development Monitoring in Multitemporal Remotely Sensed Image Pairs with Stochastic Birth-Death Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2012**, *34*, 33–50.
29. Das, S.K.; Mirnalinee, T.T.; Varghese, K. Use of Salient Features for the Design of a Multistage Framework to Extract Roads From High-Resolution Multispectral Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing* **2011**, *49*, 3906–3931.
30. Forsyth, D.A.; Torr, P.H.S.; Zisserman, A., Eds. *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, Vol. 5302, *Lecture Notes in Computer Science*. Springer, 2008.

31. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 2486–2498.

Sample Availability: Samples of the compounds ..... are available from the authors.